Numerical Analysis and Algorithms for Optimal Control of Partial Differential Equations with Control and State Constraints

Karl Kunisch

Numerical Analysis and Algorithms for Optimal Control of Partial Differential Equations with Control and State Constraints

Karl Kunisch[1]

July 28, 2006

[1]Institute of Mathematics and Scientific Computing, Karl-Franzens-Universität Graz, A-8010 Graz, Austria.

1 Model Problems and their Optimality Systems

The purpose of these notes is to introduce an approach for solving optimization problems which contain expressions which are Lipschitz continuous but not C^1 by Newton-type methods. In spite of the lack of C^1 regularity, "high rate" of convergence is our goal. More precisely we aim for a super-linear convergence.

In the remainder of this section we show by means of examples, where such problems arise in practice. Optimal control problems with control and state constraints are two of our generic model problems. The treatment of these problems here is infinite dimensional, i.e. we post these methods in properly chosen function spaces. Clearly for numerical realisation a disretisation is required. This is not addressed in these notes.

For each of the follwing problems we also state the first order optimality system. The derivation of these systems rely on Lagrange multiplier theorems. For convenience of the reader, we recall a prototype multiplier theorem in the Appendix of this section. The readers will notice that the regularity of the Lagrange multipliers, which are functional quantities in our cases, differ significantly from one problem to the other. These regularity properties have a tremendously strong influence on the proper numerical treatment.

Throughout these notes Ω denotes a bounded domain in $\mathbb{R}^n$, with boundary denoted by Γ or $\partial\Omega$, assumed to be sufficiently smooth.

1.1 Optimal Control with Control Constraints

We consider the optimal control problem with distributed control u, state variable y and unilateral control constraints:

$$(\mathcal{P}1) \qquad \min \quad J(y,u) = \frac{1}{2}\int_\Omega (y-z)^2\,dx + \frac{\beta}{2}\int_\Omega u^2\,dx\ ,$$

$$(1.1) \qquad -\Delta y = u \text{ in } \Omega\ , \quad y = 0 \quad \text{on } \Gamma\ ,$$

$$(1.2) \qquad u \in L^2(\Omega)\ ,\ u(x) \le \psi(x) \text{ for a.e. } x \in \Omega,$$

where $z \in L^2,\ \beta > 0$ and , $\psi \in L^\infty(\Omega)$.

For every $u \in L^2$ system (1.1) has a unique solution y in $H^2 \cap H_0^1$. It is standard that problem $(\mathcal{P}1)$ has a unique solution (y^*, u^*) characterized by the following optimality system :

$$\begin{cases} -\Delta y^* = u^* \text{ in } \Omega, \quad y^* \in H_0^1(\Omega), \\ -\Delta p^* = z - y^* \text{ in } \Omega, \quad p^* \in H_0^1(\Omega), \\ (\beta u^* - p^*, u - u^*)_{L^2} \geq 0 \qquad \text{for all } u \leq \psi. \end{cases}$$

Here p^* is referred to as the adjoint state. Let us give an equivalent formulation for this optimality system which is essential to motivate the forthcoming algorithm:

Theorem 1.1. *The unique solution (y^*, u^*) to problem $(\mathcal{P}1)$ is characterized by*

$$(\mathcal{S}) \qquad \begin{cases} -\Delta y^* = u^* \ \textit{in } \Omega, \quad y^* \in H_0^1(\Omega) \ , \\ -\Delta p^* = z - y^* \ \textit{in } \Omega, \quad p^* \in H_0^1(\Omega) \ , \\ \beta u^* = p^* - \lambda^*, \\ \lambda^* = c[u^* + \dfrac{\lambda^*}{c} - \Pi(u^* + \dfrac{\lambda^*}{c})] = c \max(0, u^* + \dfrac{\lambda^*}{c} - \psi) \ , \end{cases}$$

for every $c > 0$. Here Π denotes the projection of L^2 onto $U_{ad} = \{u \in L^2 : u \leq \psi\}$.

The proof can be given by inspection. Here and throughout, order relations like "max" and" $\leq$ "between elements of L^2 are understood in the pointwise almost everywhere sense.

We point out that the last equation in $(\mathcal{S})$

$$\lambda^* = c[u^* + \frac{\lambda^*}{c} - \Pi(u^* + \frac{\lambda^*}{c})] \tag{1.3}$$

is equivalent to

$$\lambda^* \in \partial I_{U_{ad}}(u^*) \ , \tag{1.4}$$

where ∂I_C denotes the subdifferential of the indicator function I_C of a a convex set C. This follows from general properties of convex functions (see

[IK1] for example) and can also easily be verified directly for the convex function $I_{U_{ad}}$. The replacement of the well known differential inclusion (1.4) in the optimality system for $(\mathcal{P}1)$ by (1.3) is an essential ingredient of the algorithm that we shall discuss.

It is also useful to consider the constraint $u \leq \psi$ in L^2 in abstract terms, expressing it as

$$G(u) = u - \psi \leq 0, \qquad \text{where } G : L^2 \to L^2.$$

Clearly G is surjective and hence existence of a Lagrange multiplier λ^* **in** $\mathbf{L^2}$ with the specified properties follows from abstract Lagrange multiplier theory, c.f. the Appendix of the chapter.

Let us also note that (1.3) can be expressed as

$$\lambda^* \geq 0, \quad u^* \leq \psi, \quad (u^* - \psi)\lambda^* = 0. \tag{1.5}$$

The ensemble of inequalities in (1.5) is called a complementarity system. Equation (1.3) is an equivalent formulation for (1.5) by means of nonlinear equation. In this context, the max operation is referred to as complementarity function.

For further treatment of $(\mathcal{P}1)$ we refer to [BIK], which is contained in these notes.

1.2 Obstacle Problems

We consider

$$(\mathcal{P}2) \qquad \begin{cases} \min \ \frac{1}{2}\, a(y,y) - (f,y) \\ \\ y \in H_0^1(\Omega) \\ \\ y \leq \psi \ \text{ a.e. in } \Omega, \end{cases}$$

where $a\,(\cdot,\cdot)$ is a bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$ satisfying

$$a(v,v) \geq \nu |v|^2_{H_0^1}, \quad a(w,z) \leq \mu |w|_{H^1} |z|_{H^1}, \tag{1.6}$$

for some $\nu > 0$ and $\mu > 0$ independently of $v \in H_0^1(\Omega)$ and $w, z \in H^1(\Omega)$. For example,

$$a(v,w) = \int_\Omega \nabla v \nabla w \, dx \tag{1.7}$$

satisfies these requirements. Further it is assumed that $f \in L^2(\Omega)$, $\psi \in H^1(\Omega)$ with $\psi|_{\partial\Omega} \geq 0$. Since $\psi \in H^1(\Omega)$ the trace $\psi|_{\partial\Omega}$ is well-defined. The assumption $\psi|_{\partial\Omega} \geq 0$ implies that the set of admissible functions y for ($\mathcal{P}2$) is nonempty. For our discussion the weak maximum principle, i.e. for all $v \in H_0^1(\Omega)$

$$a(v, v^+) \leq 0 \text{ implies } v^+ = 0, \tag{1.8}$$

where $v^+ = \max(0, v)$, will be important. It is satisfied for (1.7)

It is standard to argue that $\mathcal{P}2$ admits a unique solution $y^* \in H_0^1(\Omega)$. From subsection 1.8, for example, it follows that there exists Lagrange multiplier $\lambda^* \in H^{-1}(\Omega)$ associated to the inequality constraint $y \leq \psi$. In fact,

$$G(y) = y - \psi, \quad \text{where } G : H_0^1 \to H_0^1$$

is surjective, so that the Lagrange multiplier is in $(H_0^1)^* = H^{-1}$. Under well-known regularity assumptions on the problem data it can be shown that the Lagrange multiplier satisfies additional regularity in the sense that $\lambda^* \in L^2(\Omega)$, and that the following optimality system holds:

$$\begin{cases} a(y^*, v) + (\lambda^*, v) = (f, v), \text{ for all } v \in H_0^1(\Omega) \\ \\ (\lambda^*, y^* - \psi) = 0,\ y^* \leq \psi,\ (\lambda^*, v) \geq 0 \text{ for all } v \in K, \end{cases} \tag{1.9}$$

where $K = \{v \in H_0^1(\Omega) \colon v \geq 0 \text{ a.e.}\}$ and the inner products are taken in $L^2(\Omega)$. By inspection (1.9) can equivalently be expressed as

$$\begin{cases} a(y^*, v) + (\lambda^*, v) = (f, v) \text{ for all } v \in H_0^1(\Omega) \\ \\ \lambda^* = \max(0, \lambda^* + c(y^* - \psi)), \end{cases} \tag{1.10}$$

for arbitrary $c > 0$. (More precisely, (1.9) implies (1.10) for every c, and (1.10) for some $c > 0$ implies (1.9)). The extra Lagrange multiplier regularity does not follow from nonlinear programming arguments but rather by variational or pde techniques.

1.3 Optimal Control with State Constraints

This problem is related to $\mathcal{P}1$ but the constraint acts on the state-variable rather than the control:

$$(\mathcal{P}3) \qquad \begin{cases} \min J(y,u) = \frac{1}{2}|y-z|^2_{L^2} + \frac{\beta}{2}|u|^2_{L^2} \\ \text{subject to} \\ -\Delta y = u \ \text{ in } \ \Omega, \\ y = 0 \text{ on } \partial\Omega, \\ y \leq \psi \ \text{ a.e. in } \ \Omega \\ (y,u) \in H^1_0(\Omega) \times L^2(\Omega), \end{cases}$$

where $z \in L^2(\Omega)$, $\psi \in C(\Omega)$, $\psi > 0$ on $\partial\Omega$ and $\beta > 0$. It will be convenient to set $\mathcal{W} = H^1_0(\Omega) \cap H^2(\Omega)$. Under appropriate regularity requirements every solutions to $-\Delta y = u$, with $u \in L^2(\Omega)$ and $y = 0$ on $\partial\Omega$ satisfies $y \in \mathcal{W} \subset C(\overline{\Omega})$, $n \leq 3$. It is standard to argue the existence of a solution $(y^*, u^*) \in \mathcal{W} \times L^2(\Omega)$ to $(\mathcal{P}3)$. It is also straightforward to argue the existence of a Lagrange multiplier $\lambda^* \in \mathcal{W}^*$ since

$$G(y) = y - \psi, \qquad \text{where } G : \mathcal{W} \to \mathcal{W}$$

is surjective. Let $\langle \cdot, \cdot \rangle_{C^*,C}$ denote the duality pairing between $C(\overline{\Omega})$ and its topological dual $C^*(\Omega)$. The proof to the following characterization of the solution to $\mathcal{P}3$ with some extra regularity for the Lagrange multiplier is found in [BK], for example.

Theorem 1.2. *The pair $(y^*, u^*) \in \mathcal{W} \times L^2(\Omega)$ is a solution to (P) if and only if there exists $p^* \in L^2(\Omega)$ and $\lambda^* \in C^*(\Omega)$ such that*

$$-\Delta y^* = u^* \ \text{ in } \ \Omega, \ y^* = 0 \ \text{ on } \partial\Omega,$$
$$(p^*, -\Delta y) + \langle \lambda^*, y \rangle_{C^*,C} = (z - y^*, y) \ \text{ for all } \ y \in \mathcal{W}$$

$$\beta u^* = p^*$$

$$y^* \leq \psi, \ \langle \lambda^*, y^* - \psi \rangle_{C^*,C} = 0,$$

$$\langle \lambda^*, y \rangle_{C^*,C} \geq 0, \ \text{ for all } \ y \in C(\Omega) \text{ with } y \geq 0.$$

In general the regularity of λ^* is no better than specified in Theorem1.2. In fact if the active set at the solution

$$\mathcal{A} = \{x : y^*(x) = \psi(x)\}$$

is a connected domain in Ω then it can be shown that $\lambda^* = \lambda_d^* + \lambda_b^*$, where $\lambda_d^* \in L^2(\Omega)$ and $\lambda_b^* \in L^2(\partial\mathcal{A})$. In particular λ^* in not in $L^2(\Omega)$ in general.

1.4 L^1 and BV Functionals

In recent years, the space BV (functions of total bounded variation) and of the use of L^1 functionals receives a considerable amount of attention. We consider two such cases here.

The relationship to the previous subsection stems from the fact that indicator functions (describing the inequality constraints) and norm functions are dual to each other in the sense of Fenchel duality.

We consider the **image denoising problem with L^1-fitting** and smooth regularization

$$(\mathcal{P}4) \qquad \min_{u \in H_0^1(\Omega)} \int_\Omega \left[\frac{\beta}{2} |\nabla u|^2 + |u - z| \right] dx$$

for the given function $z \in L^1(\Omega)$. Note that the cost functional in $(\mathcal{P}4)$ is nondifferentiable in the classical sense. We see that problem $(\mathcal{P}4)$ is equivalent to

$$(1.11) \qquad \min_{u \in H_0^1(\Omega)} \max_{\lambda \in C} \int_\Omega \left[\frac{\beta}{2} |\nabla u|^2 + \lambda(u - z) \right] dx = \min_{u \in H_0^1(\Omega)} \max_{\lambda \in C} \; l(u, \lambda),$$

where $l : H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ is the Lagrange-functional and the set C is defined as

$$C := \left\{ \lambda \in L^2(\Omega) \; : \; |\lambda(x)| \leq 1 \text{ a.e. } x \in \Omega \right\}.$$

The equivalence of the two formulations results from the following identity

$$\max_{\lambda \in C} \int_\Omega \lambda(u - z) \, dx = \int_\Omega |u - z| \, dx.$$

Hence, λ represents the Lagrange smoothing of the subdifferential $sign(u-z)$ [IK1].

The ptimality conditions for the solution u^* is formally easily found to be

(1.12)
$$\begin{cases} -\beta\,\Delta u^* + \lambda^* = 0, \quad \text{in } \Omega, \\ \lambda^*(x) = \dfrac{\lambda^*(x) + c\,(u^* - z)(x)}{\max\{1, |\lambda^*(x) + c\,(u^* - z)(x)|\}}, \quad \text{a. e. } x \in \Omega, \quad \text{for each } c > 0. \end{cases}$$

The second condition realizing the complementarity condition is equivalent to the actual definition of the subdifferential

$$\begin{aligned} \lambda^*(x) &= \frac{(u^* - z)(x)}{|(u^* - z)(x)|}, && \text{in } I^*(x) = \{x \in \Omega \,|\, (u^* - z)(x) \neq 0\}, \\ |\lambda^*(x)| &\leq 1, && \text{in } J^*(x) = \{x \in \Omega \,|\, (u^* - z)(x) = 0\}. \end{aligned} \tag{1.13}$$

From the optimality conditions it follows that $u^* \in H^2$.

The corresponding image denoising problem with **quadratic** L^2-fitting reads as

$$\min_{u \in H_0^1(\Omega)} \int_\Omega \left[\frac{\beta}{2} |\nabla u|^2 + |u - z|^2 \right] dx \tag{1.14}$$

with the linear optimality condition

$$-\beta \Delta \tilde{u}^* + 2\,(\tilde{u}^* - z) = 0 \tag{1.15}$$

for the unique solution $\tilde{u}^*$. Comparing (1.15) to (1.12) and (1.13), one notices that for the L^2-formulation the distance between $\tilde{u}^*$ and z plays an important role in the optimality condition (1.15), whereas only the (regularized) sign-function for $(u^* - z)$ appears in (1.12). This illustrates the relative insensitivity of the L^1-formulation towards **outliers** compared to the L^2-formulation.

We recall the Fenchel duality theorem in infinite dimensional spaces in a form that is convenient here. Let V and Y be Banach spaces with topological duals denoted by V^* and Y^*, respectively. Further let $\Lambda \in \mathcal{L}(V, Y)$ and let $\mathcal{F} : V \to \mathbb{R} \cup \{\infty\}$, $\mathcal{G} : Y \to \mathbb{R} \cup \{\infty\}$ be convex, lower semi-continuous functionals not identically equal to ∞, and assume that there exists $v_0 \in V$ such that $\mathcal{F}(v_0) < \infty$, $\mathcal{G}(\Lambda v_0) < \infty$ and $\mathcal{G}$ is continuous at Λv_0. Then we have

$$\inf_{u \in V} \mathcal{F}(u) + \mathcal{G}(\Lambda u) = \sup_{p \in V^*} -\mathcal{F}^*(\Lambda^* p) - \mathcal{G}^*(-p), \tag{1.16}$$

where $\mathcal{F}^* : V^* \to \mathbb{R} \cup \{\infty\}$ denotes the conjugate of $\mathcal{F}$ defined by

$$\mathcal{F}^*(v^*) = \sup_{v \in V} \langle v, v^* \rangle_{V,V^*} - \mathcal{F}(v).$$

Under the conditions imposed on $\mathcal{F}$ and $\mathcal{G}$ it is known that the problem on the right hand side of (1.16) admits a solution. Moreover, $(\bar{u}, \bar{p})$ are solutions to the two optimization problems in (1.16) if and only if

$$\Lambda^* \bar{p} \in \partial \mathcal{F}(\bar{u}), \tag{1.17a}$$

$$-\bar{p} \in \partial \mathcal{G}(\Lambda \bar{u}), \tag{1.17b}$$

where $\partial \mathcal{F}$ denotes the subdifferential of the convex functional $\mathcal{F}$.

Using the Fenchel duality theorem the dual of ($\mathcal{P}4$) is formally found to be

$$\min_{p \in L^2(\Omega)^2} F^*(-div\, p) + \frac{1}{2\beta} \int_\Omega |p|^2, \tag{1.18}$$

where

$$F^*(v) = \begin{cases} \infty, & |v| > 1 \\ vz, & |v| \leq 1. \end{cases}$$

.

The relation between the solution to (1.18) and $\mathcal{P}4$ is given by

$$p^* = -\beta \nabla u^*$$

.

A typical **image reconstruction problem** based on the BV semi-norm is given by

$$(\mathcal{P}5) \qquad \begin{cases} \min & \frac{1}{2} \int_\Omega |Ku - f|^2 dx + \frac{\alpha}{2} \int_\Omega |u|^2 dx + \beta \int_\Omega |Du| \\ \text{over} & u \in BV, \end{cases}$$

where $\beta > 0$, $\alpha \geq 0$ are given and $\mathrm{K} \in \mathcal{L}(\mathrm{L}^2(\Omega))$. We assume that constant functions are not in the kernel of K or $\alpha > 0$. Further $\mathrm{BV}(\Omega)$ denotes the space of functions of bounded variation. A function u is in $\mathrm{BV}(\Omega)$ if the BV semi-norm defined by

$$\int_\Omega |Du| = \sup \left\{ \int_\Omega u \ \mathrm{div}\, \vec{v} : \vec{v} \in (C_0^\infty(\Omega))^2,\ |\vec{v}(x)|_{\ell^\infty} \leq 1 \right\}$$

is finite.

The great advantage of BV-regularization over regularization involving $|\nabla u|^2$ lies in the fact that the former perserves corners and edges in the image significantly better than the latter.

Formally (as a reasonably simple exercise) the Fenchel dual of this problem is found to be:

$$\text{(1.19)} \qquad \begin{cases} \inf \ \frac{1}{2}|\operatorname{div} \vec{p} + \mathrm{K}^* f|_B^2 \\ \text{s.t.} \ \ -\beta\vec{1} \leq \vec{p}(x) \leq \beta\vec{1} \text{ for a.e. } x \in \Omega, \end{cases}$$

where $|v|_B^2 = (v, B^{-1}v)$, and the relationship between solutions to the original and the dual problem is given by

$$\text{(1.20)} \qquad \operatorname{div} \vec{p} = Bu - \mathrm{K}^* f, \quad \vec{p} = \beta \frac{\nabla u}{|\nabla u|} \text{ on } \{x : \nabla u(x) \neq 0\}.$$

Note that (1.19) is a bilaterally constrained optimization problem.

Rigorously we have the following result, where $H_0(\operatorname{div}) = \{\vec{v} \in \mathbb{L}^2(\Omega) : \operatorname{div} \vec{v} \in \mathrm{L}^2(\Omega), \vec{v} \cdot n = 0 \text{ on } \partial\Omega\}$, and n is the outer normal to $\partial\Omega$.

Theorem 1.3. *Consider the problem*

$$\text{(1.21)} \qquad \begin{cases} \min \quad \frac{1}{2}|\operatorname{div} \vec{p} + \mathrm{K}^* f|_B^2 \quad \textit{for } \vec{p} \in H_0(\operatorname{div}) \\ \textit{s.t.} \ \ -\beta\vec{1} \leq \vec{p} \leq \beta\vec{1}, \end{cases}$$

Its dual is given by $(\mathcal{P}5)$.

1.5 Miscellanies

There are still many related problems of non-differentiable optimization problems in function spaces, for example friction and contact problems and Bingham fluid problems (two phase fluids).

Interesting and, in part open problems, are related to considering optimization problems subject to variational inequalities, as treated in 1.3 as constraints. Such problems are referred to as control of variational inequalities. Equally interesting are problems, where the obstacle itself acts as a control. From the point of view of mathematical programming all these problems are nested complementarity problems in function spaces.

1.6 Comments on the attached papers,[BIK, HIK, IK3, IK5, HK2]

Let me give a short introduction to the five selected papers which are attached to this file.

In [BIK] the primal dual active set strategy for solving control constrained optimal control problems is introduced and some global convergence properties are obtained. The starting point for this algorithm is the last equation in $(\mathcal{S})$:

$$\lambda^* = \max(0, \lambda^* + c(u^* - \psi)$$

. In the course of an iterative algorithm we decide, at iteration-level k to define the updated 'active' set to be

$$\mathcal{A}_k = \{x : \lambda_k + c(u_k - \psi) > 0\}.$$

In the following iteration the control is fixed to be ψ on $\mathcal{A}_k$, and is considered unconstrained on the inactive set $\mathcal{I}_k = \Omega \setminus \mathcal{A}_k$. We ask the readers, who just want to glimpse into these notes to read sections 1 and 2 of [BIK]. The primal-dual active set strategy may appear to be a fixed point iteration - but this would be the wrong way to think of it. In fact, it is a Newton method, where the max-operation is teated as if it was differentiable.

While the max-operation is not differentiable in the classical sense, it is **Newton-differentiable**, as operator between appropriate spaces, as described in [HIK]. Actually, Newton-differentiable is called 'slant-differentiability' in [HIK], for reason that I will explain in the course. Suffice it to say here that Newton differentiability implies local super-linear convergence of the Newton algorithm. Moreover, it is shown in [HIK] that the primal dual active set strategy is equivalent to a Newton step (we now call it 'semi-smooth' Newton step), applied to the max-operation. In [HIK] it is shown that max is Newton differentiable between finite-dimensional spaces, and as operator between L^p and L^q, provided that $q < p$. Looking back over the examples in subsections 1.1.-1.4. we need to address the question, when this case of Newton-differentiability occurs. It holds, for quite generally for control constrained optimal control problems. This is also covered in [HIK]. But it is not true generically. (We ask the readers not to skip too much from [HIK].)

In [IK3] we focus on the semi-smooth Newton method for obstacle-type problems, as considered in section 1.2 above. Recall, that this is the case where the Lagrange multiplier has L^2 regularity, but this does not follow directly from a Lagrange multiplier theorem. We define a family of regularized

problems which are semi-smooth and which converge to the original problem asymptotically. Moreover we analyse monotonicity type problems, which are related to the weak maximum principle, satisfied by this class of examples - or the M-matrix property, if properly discretized. I ask the readers to read the theorems - the proofs are not essential for what follows.

In [IK5] we treat the even less regular case of state constrained optimal control problems. Here the Lagrange multipliers are generically not L^2, but rather only measures. Again we introduce a family of approximating problems which are semi-smooth, and which converge to the original problem asymptotically. The reader may want to consider the numerical section in both [IK3] and [IK5] and note that , if we knew how to tune the parameter, which defines the regularization, and let it tend to infinity in a clever way, then this would be very efficient.

This point is also addressed in [HK1]. We introduce a "path" which describes the behavior of the regularized problems as a function of the regularization parameter. Mathematically the path is a consequence of sensitivity analysis. Intuitively we may think that on the path the problems are better behaved than far off the path. - But of course this path is not available to us quantitatively. However, some intricate manipulations allow to get **models** for the path on the basis of just two evaluations of the regularized problems with two different regularization parameters. On the basis of these models an approximate path is available, and strategies can be developed for systematically updating the regularization parameter.

1.7 Appendix: A Lagrange Multiplier Theorem

To derive first order necessary optimality conditions for constrained optimization problems the following Lagrange multiplier theorem is useful, c.f.[C]. Below DG denotes the Gateaux differential of the mapping G. By definition the Gateaux derivative is a continuous linear mapping.

Theorem 1.4. *Let U and Z be Banach spaces, and $K \subset U$, $C \subset Z$ be convex subsets, with C having a nonempty interior. Let $\bar{u} \in K$ ba a solution of the optimization problem*

$$\begin{cases} \min J(u), \\ u \in K, \quad G(u) \in C, \end{cases}$$

where $J : U \to (-\infty, \infty]$ and $G : U \to Z$ are two Gateaux differentiable

mappings at $\bar{u}$. Then there exist a real number $\bar{\mu} \geq 0$ and an element $\bar{\lambda} \in Z^*$ such that

$$\bar{\mu} + |\bar{\lambda}|_{Z^*} > 0,$$
$$\langle \bar{\lambda}, z - G(\bar{u}) \rangle \leq 0, \text{ for all } z \in C,$$
$$\langle \bar{\mu} J'(\bar{u}) + [DG(\bar{u})]^* \bar{\lambda}, u - \bar{u} \rangle \geq 0 \text{ for all } u \in K.$$

Moreover, $\bar{\mu}$ can be taken equal to 1 if the following condition of Slater type is satisfied:

$$\text{there exists } u_0 \in K \text{ such that } G(\bar{u}) + DG(\bar{u})(u_0 - \bar{u}) \in \text{int}\, C.$$

References

[BHHK] M. BERGOUNIOUX, M. HADDOU, M. HINTERMÜLLER and K. KUNISCH: A Comparison of a Moreau-Yosida Based Active Strategy and Interior Point Methods for Constrained Optimal Control Problems, SIAM J. on Optimization, 11(2000), 495–521.

[BIK] M. BERGOUNIOUX, K. ITO and K. KUNISCH: Primal-dual Strategy for Constrained Optimal Control Problems, SIAM J. Control and Optimization, 37(1999), 1176–1194.

[BK] M. BERGOUNIOUX und K. KUNISCH: On the Structure of the Lagrange Multiplier for State-Constrained Optimal Control Problems, Systems and Control Letters, 48(2002), 16-176.

[C] E. Casas: Boundary Control of Semilinear Elliptic Equations with Pointwise State Contraints, SIAM J. Control and Optim., 31(1993), 993-1006.

[DK1] J. C. de los REYES and K. KUNISCH: A semi-smooth Newton method for control constrained optimal control of the Navier Stokes equations, Nonlinear Analysis, 62(2005),1289-1316.

[DK2] J. C. de los REYES and K. KUNISCH: A comparison of algorithms for control constrained optimal control of the Burgers equation, 41(2004), 203-225.

[HHKOS] W. HINTERBERGER, M. HINTERMÜLLER, K. KUNISCH, M. v. OEHSEN and O. SCHERZER: Tube methods for BV regularization, J. Math. Imaging and Vision, 19(2003), 219-236.

[HIK] M. HINTERMÜLLER, K. ITO and K. KUNISCH: The primal–dual active set strategy as a semi–smooth Newton method, SIAM Journal on Optimization, 13(2002), 865-888.

[HK1] M. HINTERMÜLLER and K. KUNISCH: Total bounded variation regularization as bilaterally constrained optimization problem, SIAM J. Appl. Mathematics 64(2004), 1311-1333.

[HK2] M. HINTERMÜLLER and K. KUNISCH: Path-following methods for a class of constrained minimization problems in function space, to appear in SIAM J. on Optimization.

[IK1] K. ITO and K. KUNISCH: Augmented Lagrangian Methods for Nonsmooth Convex Optimization in Hilbert Spaces, Nonlinear Analysis, Theory, Methods and Applications, 41(2000), 591–616.

[IK2] K. ITO and K. KUNISCH: Optimal Control of Elliptic Variational Inequalities, Applied Mathematics and Optimization, 41(2000), 343–364.

[IK3] K. ITO and K. KUNISCH: Semi-smooth Newton methods for variational inequalities of the first kind, Mathematical Modelling and Numerical Analysis, 37(2002), 41-62.

[IK4] K. ITO and K. KUNISCH: The primal-dual active set method for nonlinear optimal control problems with bilateral constraints, SIAM J. on Control and Optimization, 43(2004), 357-376.

[IK5] K. ITO and K. KUNISCH: Semi-smooth Newton methods for state-constrained optimal control problems, Systems and Control Letters, 50(2003), 221-228.

[IK6] K. ITO and K. KUNISCH: Parabolic variational inequalities: the Langrange multiplier approach, Journal de Math. Pures et Appl., to appear.

bibitem[IK7]IK7 K. ITO and K. KUNISCH: Optimal Control of Obstacle Problems by H^1-Obstacles, Appl. Math. and Optim., to appear.

[KKM] T. KÄRKKÄINEN, K. KUNISCH and K. MAJAVA: Denoising using L^1-fitting, Computing, 74(2005), 353-376.

[KKT] T. KARKKAINEN, K. KUNISCH und P. TARVAINEN: Primal-Dual Active set Methods for Obstacle Problems, J. Optimization Theory and Appl., 119(2003), 499-533.

[KRe] K. KUNISCH and F. RENDL: An Infeasible Active Set Method for Quadratic Problems with Simple Bounds, SIAM Journal on Optimization, 14(2003), 35-52.

[KRo] K. KUNISCH and A. RÖSCH: Primal-dual active set strategy for a general class of constrained optimal control problems, SIAM Journal on Optimization, 13(2002), 321-334.

[KS] K. KUNISCH and G. STADLER: Generalized Newton methods for the 2D-Signorini contact problem with friction in function space, ESIAM: M2AN, 39(2005), 827-854.

Primal-dual Strategy for Constrained Optimal Control Problems

MAÏTINE BERGOUNIOUX[1] KAZUFUMI ITO[2] KARL KUNISCH[3]

Abstract

An algorithm for efficient solution of control constrained optimal control problems is proposed and analyzed. It is based on an active set strategy involving primal as well as dual variables. For discretized problems sufficient conditions for convergence in finitely many iterations are given. Numerical examples are given and the role of strict complementarity condition is discussed.

Keywords: Active Set, Augmented Lagrangian, Primal-dual method, Optimal Control.
AMS subject classification. 49J20, 49M29

1 Introduction and formulation of the problem

In the recent past significant advances have been made in solving efficiently nonlinear optimal control problems. Most of the proposed methods are based on variations of the sequential quadratic programming (SQP) technique, see for instance [HT, KeS, KuS, K, T] and the references given there. The SQP-algorithm is sequential and each of its iterations requires the solution of a quadratic minimization problem subject to linearized constraints. If these auxiliary problems contain inequality constraints with infinite dimensional image space then their solution is still a significant challenge.

In this paper we propose an algorithm for the solution of infinite dimensional quadratic problems with linear equality constraints and pointwise affine inequality constraints. It is based on an active set strategy involving primal and dual variables. It thus differs significantly from conventional active set strategies that involve primal variables only, see [Sch] for example. In practice the proposed algorithm behaves like an infeasible one. The iterates of

[1]UMR-CNRS 6628, Université d'Orléans, U.F.R. Sciences, B.P. 6759, F-45067 Orléans Cedex 2, France. E-mail: Maitine.Bergounioux@labomath.univ-orleans.fr. This work was supported in part by EEC, HCM Contract CHRX-CT94-0471

[2]Department of Mathematics, North Carolina State University, Raleigh, NC27695, USA.

[3]Institut für Mathematik, Universität Graz, A-8010 Graz, Austria, E-mail: Kunisch@kfunigraz.ac.at. Work supported in part by EEC, HCM Contract CHRX-CT94-0471 and Fonds zur Förderung der wissenschaftlichen Forschung, UF8,"Optimization and Control".

the algorithm violate the constraints up to the next-to-the-last iterate. The algorithm stops at a feasible and optimal solution.

Within this paper we do not aim for generality but rather we treat as a model problem an unilateral control constraint optimal control problem related to elliptic partial differential equations. The distributed nature of this problem, which is reflected in the fact that it behaves like an obstacle problem for the biharmonic equation, makes it difficult to analyze.

Let us briefly outline the contents of the paper. The algorithm will be presented in Section 2. We prove that if the algorithm produces the same active set in two consecutive iterates then the optimal solution has been obtained. In Section 3 we shall give sufficient conditions which guarantee that an augmented Lagrangian functional behaves as a decreasing merit function for the algorithm. In practice this implies finite step convergence of the discretized problem. Section 4 is devoted to showing that for a minor modification of the algorithm the cost functional is increasing until the feasible optimal solution is reached. In Section 5 several numerical examples are given. For most examples the algorithm behaves extremely efficient and typically converges in less than five iterations. Thus, to present interesting cases the majority of the test examples is in some sense extreme: Either the strict complementarity condition is violated or the cost of the control is nearly zero.

To describe the problem, let Ω be an open, bounded subset of $\mathbb{R}^N$, N ≤ 3, with smooth boundary Γ and consider the following distributed optimal control problem :

$$\min \quad J(y,u) = \frac{1}{2}\int_\Omega (y-z_d)^2\,dx + \frac{\alpha}{2}\int_\Omega\,(u-u_d)^2\,dx\ , \tag{$\mathcal{P}$}$$

$$-\Delta y = u \text{ in } \Omega\ , \quad y = 0 \quad \text{on } \Gamma\ , \tag{1.1}$$

$$u \in U_{ad} \subset L^2(\Omega)\ , \tag{1.2}$$

where $z_d,\ u_d \in L^2(\Omega),\ \alpha > 0$ and $U_{ad} = \{\ u \in L^2(\Omega) \mid u(x) \leq b(x) \text{ a.e. in } \Omega\},\ b \in L^\infty(\Omega)$. It is well known that, for every $u \in L^2(\Omega)$ system (1.1) has a unique solution $y = \mathcal{T}(u)$ in $H^2(\Omega) \cap H_o^1(\Omega)$.

Remark 1.1 *To emphasis the basic ideas of the proposed approach we treated the rather simple problem* ($\mathcal{P}$)*. Many generalizations are possible. In particular,* $-\Delta$ *in (1.1) can be replaced by any strictly elliptic second order differential operator.*

It is standard that problem ($\mathcal{P}$) has a unique solution (y^*, u^*) characterized by the following optimality system :

$$\begin{cases} -\Delta y^* = u^* \text{ in } \Omega, \quad y^* \in H_o^1(\Omega)\ , \\ -\Delta p^* = z_d - y^* \text{ in } \Omega, \quad p^* \in H_o^1(\Omega)\ , \\ (\alpha(u^* - u_d) - p^*, u - u^*) \geq 0 \qquad \text{for all } u \in U_{ad}, \end{cases}$$

where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$-inner product.

Let us give an equivalent formulation for this optimality system which is essential to motivate the forthcoming algorithm:

Theorem 1.1 *The unique solution (y^*, u^*) to problem $(\mathcal{P})$ is characterized by*

$$(\mathcal{S}) \qquad \begin{cases} -\Delta y^* = u^* \text{ in } \Omega, \quad y^* \in H_o^1(\Omega) \ , \\ -\Delta p^* = z_d - y^* \text{ in } \Omega, \quad p^* \in H_o^1(\Omega) \ , \\ u^* = u_d + \dfrac{p^* - \lambda^*}{\alpha} \ , \\ \lambda^* = c[u^* + \dfrac{\lambda^*}{c} - \Pi(u^* + \dfrac{\lambda^*}{c})] = c\max(0, u^* + \dfrac{\lambda^*}{c} - b) \ , \end{cases}$$

for every $c > 0$. Here Π denotes the projection of $L^2(\Omega)$ onto U_{ad}.

Proof - We refer to [IK]. ■

We point out that the last equation in $(\mathcal{S})$

$$\lambda^* = c[u^* + \frac{\lambda^*}{c} - \Pi(u^* + \frac{\lambda^*}{c})] \tag{1.3}$$

is equivalent to

$$\lambda^* \in \partial I_{U_{ad}}(u^*) \ , \tag{1.4}$$

where ∂I_C denotes the subdifferential of the indicator function I_C of a a convex set C. This follows from general properties of convex functions (see [IK] for example) and can also easily be verified directly for the convex function $I_{U_{ad}}$. The replacement of the well known differential inclusion (1.4) [B] in the optimality system for $(\mathcal{P})$ by (1.3) is an essential ingredient of the algorithm that we shall propose.

Here and below, order relations like "max" and" $\leq$ "between elements of $L^2(\Omega)$ are understood in the pointwise almost everywhere sense.

Let us interpret the optimality system $(\mathcal{S})$. From $-\Delta y^* = u_d + \dfrac{p^* - \lambda^*}{\alpha}$ it follows that $p^* = \alpha[-\Delta y^* - u_d] + \lambda^*$ and hence

$$-\alpha\Delta y^* - \Delta^{-1} y^* + \lambda^* = \alpha \ u_d - \Delta^{-1} z_d \ .$$

It follows that

$$\begin{aligned} \alpha u^* + \Delta^{-2} u^* + \lambda^* &= \alpha u_d - \Delta^{-1} z_d \ , \\ \lambda^* &= c\max(0, u^* + \frac{\lambda^*}{c} - b) \qquad \text{for all } c > 0 \end{aligned}$$

which implies the highly distributed nature of the optimal control. Setting $\mathcal{H} = \alpha I + \Delta^{-2}$ and $f = \alpha u_d - \Delta^{-1} z_d$, system $(\mathcal{S})$ can be expressed as

$$(\mathcal{S})_1 \qquad \begin{cases} \mathcal{H} u^* + \lambda^* = f \ , \\ \lambda^* = c\max(0, u^* + \dfrac{\lambda^*}{c} - b) \qquad \text{for all } c > 0 \end{cases}$$

We observe that by setting $u = -\Delta y$, system $(\mathcal{S})$ constitutes an optimality system for the variational inequality

$$\begin{cases} \min & \dfrac{\alpha}{2}\displaystyle\int_\Omega |\Delta y|^2 dx + \dfrac{1}{2}\int_\Omega |y - (z_d - \alpha\ \Delta u_d)|^2 dx \\ y \in H_o^1(\Omega) \cap H^2(\Omega) & \\ -\Delta y \leq b & \end{cases}$$

the regularity of which was studied in [BS].

2 Presentation of the Algorithm

In this section we present the primal-dual active set algorithm and discuss some of its basic properties. Let us introduce the active and inactive sets for the solution to $(\mathcal{P})$ and define

$$\mathcal{A}^* = \{\ x \mid u^*(x) = b \ \text{ a.e. }\} \text{ and } \mathcal{I}^* = \{\ x \mid u^*(x) < b \ \text{ a.e. }\}\ .$$

The proposed strategy is based on (1.3). Given (u_{n-1}, λ_{n-1}) the active set for the current iterate is chosen as

$$\mathcal{A}_n = \{\ x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} > b \ \text{ a.e. }\}\ .$$

We recall that $\lambda^* \geq 0$ and in the case of strict complementarity $\lambda^* > 0$ on $\mathcal{A}^*$. The complete algorithm is specified next

Algorithm

1. Initialization : choose $y_o,\ u_o$ and λ_o and set $n = 1$.
2. Determine the following subsets of Ω :

$$\mathcal{A}_n = \{\ x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} > b\ \}\ ,\quad \mathcal{I}_n = \{\ x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} \leq b\ \}\ .$$

3. If $n \geq 2$ and $\mathcal{A}_n = \mathcal{A}_{n-1}$ then STOP.
4. Else, find $(y_n, p_n) \in H_o^1(\Omega) \times H_o^1(\Omega)$ such that

$$-\Delta y_n = \begin{cases} b & \text{in } \mathcal{A}_n \\ u_d + \dfrac{p_n}{\alpha} & \text{in } \mathcal{I}_n\ , \end{cases}$$
$$-\Delta p_n = z_d - y_n \ \text{ in } \Omega\ .$$

and set

$$u_n = \begin{cases} b & \text{in } \mathcal{A}_n \\ u_d + \dfrac{p_n}{\alpha} & \text{in } \mathcal{I}_n\ , \end{cases}$$

5. Set $\lambda_n = p_n - \alpha(u_n - u_d)$, update $n = n + 1$ and goto 2.

The existence of the triple (y_n, u_n, p_n) satisfying the system of step 4 of the Algorithm follows from the fact that it constitutes the optimality system for the auxiliary problem

$$(\mathcal{P}_{aux}) \qquad \min \ \{ \ J(y,u) \mid y \in H_o^1(\Omega), \ -\Delta y = u \text{ in } \Omega, \ u = b \text{ on } \mathcal{A}_n \ \}$$

which has (y_n, u_n) as unique solution.

We may use different initialization schemes. The one that was used most frequently is the following one:

$$\begin{cases} u_o = b \ , \\ -\Delta y_o = u_o, \ y_o \in H_o^1(\Omega) \ , \\ -\Delta p_o = z_d - y_o \ , \ p_o \in H_o^1(\Omega) \ , \\ \lambda_o = \max(0, \alpha(u_d - b) + p_o) \ . \end{cases} \tag{2.1}$$

This choice of initialization has the property of feasibility. Alternatively, we tested the algorithm with initialization as the solution of the unconstrained problem, i.e.

$$\begin{cases} \lambda_o = 0 \\ -\Delta y_o = u_d + \dfrac{p_o}{\alpha} \ , \quad y_o \in H_o^1(\Omega) \ , \\ -\Delta p_o = z_d - y_o \ , \quad p_o \in H_o^1(\Omega) \ , \\ u_o = u_d + \dfrac{p_o}{\alpha} \ . \end{cases} \tag{2.2}$$

For all examples the first initialization behaved better or equal to the second.
The initialization process (2.1) has the property that the first set $\mathcal{A}_1$ is always included in the active set $\mathcal{A}^*$ of problem $(\mathcal{P})$. More precisely we have

Lemma 2.1 *If (u_o, y_o, λ_o) are given by (2.1) with $u_o \geq u^*$; then $\lambda_o \leq \lambda^*$.*
In addition, if $u_o = b$ then $\mathcal{A}_1 \subset \mathcal{A}^*$.

Proof - By construction

$$\lambda_o = \max(0, \alpha(u_d - u_o) + p_o) = \max(0, \alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d)) \ ,$$

and as a consequence of $(\mathcal{S})$

$$\lambda^* = \alpha(u_d - u^*) + p^* = \alpha(u_d - u^*) + \Delta^{-1}(y^* - z_d) = \alpha(u_d - u^*) - \Delta^{-2}u^* - \Delta^{-1}z_d \ \geq 0.$$

It follows that
$\lambda^* - \lambda_o = \lambda^* \geq 0$ if $\alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d) \leq 0$, and
$\lambda^* - \lambda_o = \alpha(u_o - u^*) + \Delta^{-2}(u_o - u^*) + \alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d)$ else .

If $u_o \geq u^*$ the maximum principle yields $\Delta^{-2}(u_o - u^*) \geq 0$ and

$$\lambda^* - \lambda_o \begin{cases} = \lambda^* \geq 0 & \text{if } \alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d) \leq 0 \\ \geq \alpha(u_d - u_o) + \Delta^{-1}(y_o - z_d) \geq 0 & \text{else .} \end{cases}$$

Therefore $\lambda_o \leq \lambda^*$.
In addition, if $u_o = b$ then $u_o + \dfrac{\lambda_o}{c} = b + \dfrac{\lambda_o}{c} > b$ on $\mathcal{A}_1$. Consequently $\lambda_o > 0$ on $\mathcal{A}_1$ and $\lambda^* > 0$. It follows that $\mathcal{A}_1 \subset \mathcal{A}^*$ and the proof is complete. ■

A first convergence result which also justifies the stopping criterion in Step 3 is given in the following theorem.

Theorem 2.1 *If there exists $n \in \mathbb{N} - \{0\}$ such that $\mathcal{A}_n = \mathcal{A}_{n+1}$ then the algorithm stops and the last iterate satisfies*

$$(\mathcal{S}_n) \qquad \begin{cases} -\Delta y_n = u_n = \begin{cases} b & \text{in } \mathcal{A}_n \\ u_d + \dfrac{p_n}{\alpha} & \text{in } \Omega - \mathcal{A}_n \ , \end{cases} \\ -\Delta p_n = z_d - y_n \quad \text{in } \Omega \ . \\ \lambda_n = p_n - \alpha(u_n - u_d) \ , \ u_n \in U_{ad} \end{cases}$$

with

$$\lambda_n = 0 \text{ on } \mathcal{I}_n \text{ and } \lambda_n > 0 \text{ on } \mathcal{A}_n \ . \tag{2.3}$$

Therefore, the last iterate is the solution of the original optimality system $(\mathcal{S})$.

Proof - If there exists $n \in \mathbb{N} - \{0\}$ such that $\mathcal{A}_n = \mathcal{A}_{n+1}$, then it is clear that algorithm stops and the last iterate satisfies $(\mathcal{S}_n)$ by construction except possibly for $u_n \in U_{ad}$.
Thus we have to prove $u_n \in U_{ad}$ and (2.3).

- On $\mathcal{I}_n$ we have $\lambda_n = 0$ by step 5 of the Algorithm. Moreover $u_n + \dfrac{\lambda_n}{c} = u_n \leq b$, since $\mathcal{I}_n = \mathcal{I}_{n+1}$.
- On $\mathcal{A}_n$ we get $u_n = b$ and $u_n + \dfrac{\lambda_n}{c} > b$ since $\mathcal{A}_n = \mathcal{A}_{n+1}$.Therefore $\lambda_n > 0$ on $\mathcal{A}_n$ and $u_n \in U_{ad}$.

To prove that the last iterate is a solution of the original optimality system $(\mathcal{S})$, it remains to show that

$$\lambda_n = c[u_n + \frac{\lambda_n}{c} - \Pi(u_n + \frac{\lambda_n}{c})] \ .$$

- On $\mathcal{I}_n$ we have $\lambda_n = 0$ and $u_n + \dfrac{\lambda_n}{c} = u_n \leq b$. It follows that

$$u_n + \frac{\lambda_n}{c} - \Pi(u_n + \frac{\lambda_n}{c}) = u_n - \Pi(u_n) = 0 = \lambda_n \ .$$

- On $\mathcal{A}_n$ we get $u_n = b$, $\lambda_n > 0$ and therefore

$$c[u_n + \frac{\lambda_n}{c} - \Pi(u_n + \frac{\lambda_n}{c})] = c[b + \frac{\lambda_n}{c} - b] = \lambda_n \ .$$

■

Now we give a structural property of the algorithm :

Lemma 2.2 *If u_n is feasible for some $n \in \mathbb{N} - \{0\}$ (i.e. $u_n \leq b$) then $\mathcal{A}_{n+1} \subset \mathcal{A}_n$.*

Proof - On $\mathcal{I}_n$ we get $\lambda_n = 0$ by construction, so that $u_n + \frac{\lambda_n}{c} = u_n \leq b$ (because of feasibility). This implies $\mathcal{I}_n \subset \mathcal{I}_{n+1}$ and consequently $\mathcal{A}_{n+1} \subset \mathcal{A}_n$. ■

Note that Theorem 2.1 and in particular (2.3) does not utilize or imply strict complementarity. In fact, if (2.3) holds, then the set of x for which $u_n(x) = b$ and $\lambda_n(x) = 0$ is contained in $\mathcal{I}_n$.

We end this section with "simple cases" where we may conclude easily that the algorithm is convergent.

Theorem 2.2 *For initialization (2.1), the Algorithm converges in one iteration in the following cases*

1. $z_d \leq 0,\ u_d = 0\ , b \geq 0$ and the solution to $-\alpha\Delta u - \Delta^{-1}u = z_d$ is nonpositive.

2. $z_d \geq 0,\ b \leq 0,\ u_d > b$ or

$z_d \geq 0,\ b \leq 0,\ u_d \geq b$ and $z_d + \Delta^{-1}b$ is not zero as element in $L^2(\Omega)$.

Proof - Let us first examine case 1. The maximum principle implies that $-\Delta^{-1}u_o \geq 0$. Consequently $z_d + \Delta^{-1}u_o \leq 0$ and by a second application of the maximum principle

$$-\Delta^{-1}(z_d + \Delta^{-1}u_o) \leq 0 \ .$$

Together with the fact that $u_d - b = -b \leq 0$, this implies

$$\lambda_o = \max(0, \alpha(u_d - b) - \Delta^{-1}(z_d + \Delta^{-1}u_o)) = 0 \ .$$

Therefore $\mathcal{A}_1 = \emptyset$ and $\mathcal{I}_1 = \Omega$.
Using the first iteration we obtain $u_1 = \frac{p_1}{\alpha}$ in Ω. Moreover $-\Delta y_1 = u_1$ and $-\Delta p_1 = z_d - y_1$ imply that

$$-\alpha\Delta u_1 - \Delta^{-1}u_1 = z_d \ .$$

By assumption u_1 is feasible. Therefore $\mathcal{A}_2 = \mathcal{A}_1 = \emptyset$ and by Theorem 2.1 the algorithm stops at the solution to $(\mathcal{P})$.
Now we consider case 2. By assumption and due to (2.1) we have $z_d \geq 0,\ b \leq 0\ , \lambda_o \geq 0$ and $\mathcal{A}_1 = \{\ \lambda_o > 0\ \}$. Due to the maximum principle $-\Delta^{-1}u_o \leq 0$ and

$$p_o = -\Delta^{-1}(z_d - y_o) = -\Delta^{-1}[z_d - (-\Delta^{-1}u_o)] \geq 0 \ .$$

Moreover if $z_d+\Delta^{-1}b$ is not the zero element in $L^2(\Omega)$, then $p_o > 0$ in Ω and $\alpha(u_d-b)+p_o > \alpha(u_d-b)$.
If $u_d > b$ or ($u_d = b$ and $z_d + \Delta^{-1}b \neq 0$) then $\lambda_o = \max(0, \alpha(u_d-b)+p_o) > 0$ in Ω (and $\lambda_o = 0$ on $\partial\Omega$). Consequently $\mathcal{A}_1 = \Omega$ and $u_1 = b$, $\lambda_1 = -\Delta^{-1}(z_d+\Delta^{-1}b)+\alpha(u_d-b) > 0$. This yields $\mathcal{A}_2 = \mathcal{A}_1 = \Omega$ and the algorithm stops. ■

3 Convergence analysis

3.1 The Continuous Case

The convergence analysis of the Algorithm is based on the decrease of appropriately chosen merit functions. For that purpose we define the following augmented Lagrangian functions

$$L_c(y,u,\lambda) = J(y,u) + (\lambda, \hat{g}_c(u,\lambda)) + \frac{c}{2}\|\hat{g}_c(u,\lambda)\|^2 \quad \text{and} \quad \hat{L}_c(y,u,\lambda) = L_c(y,u,\lambda^+) \ ,$$

where $(\cdot,\cdot)$ is the $L^2(\Omega)$-inner product, $\|\cdot\|$ is the $L^2(\Omega)$-norm, $\lambda^+ = \max(\lambda, 0)$ and $\hat{g}_c(u,\lambda) = \max(g(u), -\frac{\lambda}{c})$ with $g(u) = u - b$. Further $(\cdot,\cdot)_{|S}$ and $\|\cdot\|_{|S}$ denote the L^2-inner product and norm on a measurable subset $S \subset \Omega$. Note that the mapping

$$u \mapsto (\lambda, \hat{g}_c(u,\lambda)) + \frac{c}{2}\|\hat{g}_c(u,\lambda)\|^2 \ ,$$

is $\mathcal{C}^1$, which is not the case for the function given by

$$u \mapsto (\lambda, g(u)) + \frac{c}{2}\|\max(g(u), 0)\|^2 \ .$$

The following relationship between primal and dual variables will be essential.

Lemma 3.1 *For all $n \in \mathbb{N} - \{0\}$ and $(y,u) \in H_o^1(\Omega) \times L^2(\Omega)$ satisfying $-\Delta y = u$ we have*

$$J(y_n,u_n) - J(y,u) = -\frac{1}{2}\|y-y_n\|^2 - \frac{\alpha}{2}\|u-u_n\|^2 + (\lambda_n, u-u_n)_{|\mathcal{A}_n} \tag{3.1}$$

Proof - Using $\|a\|^2 - \|b\|^2 = -\|a-b\|^2 + 2\,(a-b,a)$ and Steps 4 and 5 of the Algorithm, we find that

$$\begin{aligned} J(y_n,u_n) - J(y,u) &= -\frac{1}{2}\|y-y_n\|^2 - \frac{\alpha}{2}\|u-u_n\|^2 + (y_n-y, y_n-z_d) + \alpha\,(u_n-u, u_n-u_d) \\ &= -\frac{1}{2}\|y-y_n\|^2 - \frac{\alpha}{2}\|u-u_n\|^2 + (\Delta(y_n-y), p_n) + \alpha\,(u_n-u, u_n-u_d) \\ &= -\frac{1}{2}\|y-y_n\|^2 - \frac{\alpha}{2}\|u-u_n\|^2 + (u_n-u, -p_n + \alpha(u_n-u_d)) \\ &= -\frac{1}{2}\|y-y_n\|^2 - \frac{\alpha}{2}\|u-u_n\|^2 + (u-u_n, \lambda_n) \ . \end{aligned}$$

As $\lambda_n = 0$ on $\mathcal{I}_n$ the result follows. ■

Let us define

$$\mathcal{S}_{n-1} = \{\ x \in \mathcal{A}_{n-1} \mid \lambda_{n-1}(x) \leq 0\ \} \text{ and } \mathcal{T}_{n-1} = \{\ x \in \mathcal{I}_{n-1} \mid u_{n-1}(x) > b(x)\ \}\ .$$

These two sets can be paraphrased by calling $\mathcal{S}_{n-1}$ the set of elements that the active set strategy predicts to be active at level $n-1$ but the Lagrange multiplier indicates that they should be inactive, and by calling $\mathcal{T}_{n-1}$ the set of elements that was predicted to be inactive but the $n-1$st iteration level corrects it to be active. We note that

$$\Omega = (\mathcal{I}_{n-1} \backslash \mathcal{T}_{n-1}) \cup \mathcal{T}_{n-1} \cup \mathcal{S}_{n-1} \cup (\mathcal{A}_{n-1} \backslash \mathcal{S}_{n-1}) \tag{3.2}$$

defines a decomposition of Ω in mutually disjoint sets. Moreover we have the following relation between these sets at each level n:

$$\mathcal{I}_n = (\mathcal{I}_{n-1} \backslash \mathcal{T}_{n-1}) \cup \mathcal{S}_{n-1}\ ,\ \mathcal{A}_n = (\mathcal{A}_{n-1} \backslash \mathcal{S}_{n-1}) \cup \mathcal{T}_{n-1}\ . \tag{3.3}$$

In fact, as $\Omega = \mathcal{I}_n \cup \mathcal{A}_n$ is is sufficient to prove that

$$(\mathcal{I}_{n-1} \backslash \mathcal{T}_{n-1}) \cup \mathcal{S}_{n-1} \subset \mathcal{I}_n \text{ and } (\mathcal{A}_{n-1} \backslash \mathcal{S}_{n-1}) \cup \mathcal{T}_{n-1} \subset\ \mathcal{A}_n\ ,$$

that is

$$\mathcal{S}_{n-1} \subset \mathcal{I}_n \text{ and } \mathcal{T}_{n-1} \subset\ \mathcal{A}_n\ .$$

Since $\mathcal{S}_{n-1} \subset \mathcal{A}_{n-1}$ we find $u_{n-1} = b$ on $\mathcal{S}_{n-1}$. From the definition of $\mathcal{S}_{n-1}$ we conclude that $\lambda_{n-1} \leq 0$ so that $u_{n-1} + \dfrac{\lambda_{n-1}}{c} \leq b$. This implies $\mathcal{S}_{n-1} \subset \mathcal{I}_n$. The verification of $\mathcal{T}_{n-1} \subset\ \mathcal{A}_n$ is quite similar.

For the convenience of the reader we present these sets in Figure 1.

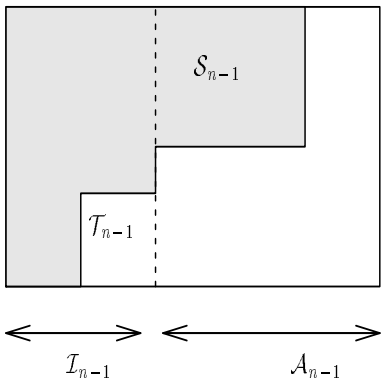

Figure 1: Decomposition of Ω at levels $n-1$ and n

In Figure 1 the shaded region depicts $\mathcal{I}_n$ and the white region is $\mathcal{A}_n$. The following table depicts the signs of primal and dual variables for two consecutive iteration levels.

	λ_{n-1}	λ_n	u_{n-1}	u_n
$\mathcal{T}_{n-1} = \mathcal{I}_{n-1} \cap \mathcal{A}_n$	0		$> b$	$= b$
$\mathcal{S}_{n-1} = \mathcal{A}_{n-1} \cap \mathcal{I}_n$	≤ 0	0	$= b$	
$\mathcal{I}_{n-1} \backslash \mathcal{T}_{n-1}$ $(\subset \mathcal{I}_n)$	0	0	$\leq b$	
$\mathcal{A}_{n-1} \backslash \mathcal{S}_{n-1}$ $(\subset \mathcal{A}_n)$	> 0		$= b$	$= b$

Table 1

Below $\|\Delta^{-1}\|$ will denote the operator norm of Δ^{-1} in $\mathcal{L}(L^2(\Omega))$.

Theorem 3.1 *If $\mathcal{A}_n \neq \mathcal{A}_{n-1}$ and*

$$\alpha + \gamma \leq c \leq \alpha - \frac{\alpha^2}{\gamma} + \frac{\alpha^2}{\|\Delta^{-1}\|^2} \tag{3.4}$$

for some $\gamma > 0$, then $\hat{L}_c(y_n, u_n, \lambda_n) \leq \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})$. In addition, if the second inequality of (3.4) is strict then either $\hat{L}_c(y_n, u_n, \lambda_n) < \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})$ or the Algorithm stops at the solution to $(\mathcal{S})$.

Proof - A short computation gives

$$\begin{aligned}
&(\lambda, \hat{g}_c(u, \lambda)) + \frac{c}{2}\|\hat{g}_c(u, \lambda)\|^2 \\
&= \left(\frac{1}{\sqrt{c}}\lambda, \sqrt{c}\ \hat{g}_c(u, \lambda)\right) + \frac{1}{2}\left(\sqrt{c}\ \hat{g}_c(u, \lambda), \sqrt{c}\ \hat{g}_c(u, \lambda)\right) \\
&= \frac{1}{2}\ \|\sqrt{c}\max(g(u), -\frac{\lambda}{c}) + \frac{1}{\sqrt{c}}\ \lambda\|^2 - \frac{1}{2c}\|\lambda\|^2 \\
&= \frac{1}{2}\|\max(\sqrt{c}\ g(u), -\frac{\lambda}{\sqrt{c}}) + \frac{1}{\sqrt{c}}\ \lambda\|^2 - \frac{1}{2c}\|\lambda\|^2 \\
&= \frac{1}{2c}\|\max(c\ g(u) + \lambda, 0)\|^2 - \frac{1}{2c}\|\lambda\|^2.
\end{aligned}$$

Moreover for all (y, u, λ) we find

$$L_c(y, u, \lambda) = J(y, u) + \frac{1}{2c}\|\max(c\, g(u) + \lambda, 0)\|^2 - \frac{1}{2c}\|\lambda\|^2\ . \tag{3.5}$$

By assumption $\mathcal{A}_n \neq \mathcal{A}_{n-1}$ and hence $\mathcal{S}_{n-1} \cup \mathcal{T}_{n-1} \neq \emptyset$. Using (3.5) we get

$$\begin{aligned}
&\hat{L}_c(y_n, u_n, \lambda_n) - \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1}) = \\
&J(y_n, u_n) - J(y_{n-1}, u_{n-1}) \\
&+\frac{1}{2c}\left[\|\max(c\ g(u_n) + \lambda_n^+, 0)\|^2 - \|\lambda_n^+\|^2 - \|\max(c\ g(u_{n-1}) + \lambda_{n-1}^+, 0)\|^2 + \|\lambda_{n-1}^+\|^2\right]
\end{aligned}$$

and by (3.1)

$$\hat{L}_c(y_n,u_n,\lambda_n)-\hat{L}_c(y_{n-1},u_{n-1},\lambda_{n-1})=$$
$$-\frac{1}{2}\|y_{n-1}-y_n\|^2-\frac{\alpha}{2}\|u_{n-1}-u_n\|^2+(u_{n-1}-u_n,\lambda_n)_{\mathcal{T}_{n-1}}+ \tag{3.6}$$
$$\frac{1}{2c}\left[\|\max(c\ g(u_n)+\lambda_n^+,0)\|^2-\|\lambda_n^+\|^2-\|\max(c\ g(u_{n-1})+\lambda_{n-1}^+,0)\|^2+\|\lambda_{n-1}^+\|^2\right]\ .$$

It will be convenient to introduce $d(x)=$

$$|\max(c\,g(u_n(x))+\lambda_n^+(x),0)|^2-|\lambda_n^+(x)|^2-|\max(c\,g(u_{n-1}(x))+\lambda_{n-1}^+(x),0)|^2+|\lambda_{n-1}^+(x)|^2.$$

Let us estimate d on the four distinct subsets of Ω according to (3.2).

• On $\mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1}$ we have $\lambda_n(x)=\lambda_{n-1}(x)=0,\ u_{n-1}(x)\le b(x)\ (g(u_{n-1}(x))\le 0)$ and

$$d(x)=|\max(c\,g(u_n(x)),0)|^2-|\max(c\,g(u_{n-1}(x)),0)|^2\le c^2|u_n(x)-u_{n-1}(x)|^2\ .$$

Moreover as $\lambda_n=p_n-\alpha(u_n-u_d)=0$ and $\lambda_{n-1}=p_{n-1}-\alpha(u_{n-1}-u_d)=0$ we have $u_n(x)-u_{n-1}(x)=\dfrac{p_n(x)-p_{n-1}(x)}{\alpha}$ so that

$$|u_n(x)-u_{n-1}(x)|\le\ \frac{1}{\alpha}|p_n(x)-p_{n-1}(x)|\quad\text{ on }\mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1}$$

• On $\mathcal{S}_{n-1},\ \lambda_n(x)=0,\ \lambda_{n-1}(x)\le 0,\ g(u_{n-1}(x))=0$, so that $d(x)=|\max(c\,g(u_n(x)),0)|^2$. Here we used the positivity of λ^+ to get $\lambda_{n-1}^+(x)=0$. To estimate $d(x)$ in detail we consider first the case where $u_n(x)\ge b(x)$. Since $x\in\mathcal{S}_{n-1}\subset\mathcal{I}_n$ we obtain $\lambda_n(x)=p_n(x)-\alpha[u_n(x)-u_d(x)]=0$ and hence $u_n(x)=\dfrac{p_n(x)}{\alpha}+u_d(x)$. Moreover, $\lambda_{n-1}(x)=p_{n-1}(x)-\alpha[u_{n-1}(x)-u_d(x)]\le 0$ so that $u_d(x)-b(x)\le-\dfrac{p_{n-1}(x)}{\alpha}$ where we used $u_{n-1}(x)=b(x)$. Since by assumption $u_n(x)\ge b$ these estimates imply

$$|u_n(x)-u_{n-1}(x)|=u_n(x)-b(x)=\frac{p_n(x)}{\alpha}+u_d(x)-b(x)\le\frac{p_n(x)}{\alpha}-\frac{p_{n-1}(x)}{\alpha}=\frac{1}{\alpha}\,|p_n(x)-p_{n-1}(x)|\,.$$

In addition it is clear that on the set $\mathcal{I}_n$:

$$d(x)=|\max(c\,g(u_n(x)),0)|^2\le c^2|u_n(x)-u_{n-1}(x)|^2\ .$$

In the second case, $u_n(x)<b(x)$ so that $\max(c\,g(u_n(x)),0)=0$ and $d(x)=0$.

Finally we have a precise estimate on the whole set $\mathcal{I}_n$. Let us denote

$$\mathcal{I}_n^*=\mathcal{I}_{n-1}\backslash\mathcal{T}_{n-1}\cup\{x\in\mathcal{S}_{n-1}\mid u_n(x)\ge b(x)\}\ =\mathcal{I}_n\backslash\{x\in\mathcal{S}_{n-1}\mid u_n(x)<b(x)\}\ ;$$

then

$$\int_{\mathcal{I}_n}d(x)\ dx=\int_{\mathcal{I}_n^*}d(x)\ dx=c^2\|\max(g(u_n),0)\|_{\mathcal{I}_n^*}^2\le c^2\ \|u_n-u_{n-1}\|_{\mathcal{I}_n^*}^2\ . \tag{3.7}$$

We note that we have proved in addition that

$$\|u_n-u_{n-1}\|_{\mathcal{I}_n^*}\le\frac{\|\Delta^{-1}\|}{\alpha}\|y_n-y_{n-1}\|\ . \tag{3.8}$$

• On $\mathcal{A}_{n-1}\backslash\mathcal{S}_{n-1}$, we have $g(u_{n-1}(x)) = g(u_n(x)) = 0,\ \lambda_{n-1}(x) > 0$ and hence

$$d(x) = |\max(\lambda_n^+(x), 0)|^2 - |\lambda_n^+(x)|^2 \leq 0\ . \tag{3.9}$$

• On $\mathcal{T}_{n-1}$ we have $\lambda_{n-1}(x) = 0,\ g(u_n(x)) = 0,\ g(u_{n-1}(x)) > 0$ and thus

$$d(x) = -c^2|g(u_{n-1}(x))|^2 = -c^2|u_n(x) - u_{n-1}(x)|^2\ . \tag{3.10}$$

Next we estimate the term $(\lambda_n, u_{n-1} - u_n)_{\mathcal{T}_{n-1}}$ in (3.6):

$$\begin{aligned}(\lambda_n, u_{n-1} - u_n)_{\mathcal{T}_{n-1}} &= (\lambda_n - \lambda_{n-1}, u_{n-1} - u_n)_{\mathcal{T}_{n-1}} \\ &= (p_n - p_{n-1}, u_{n-1} - u_n)_{\mathcal{T}_{n-1}} + \alpha\|u_n - u_{n-1}\|^2_{\mathcal{T}_{n-1}}\ .\end{aligned}$$

and therefore

$$(\lambda_n, u_{n-1} - u_n)_{\mathcal{T}_{n-1}} \leq \|\Delta^{-1}\|\ \|y_n - y_{n-1}\|_\Omega\|u_n - u_{n-1}\|_{\mathcal{T}_{n-1}} + \alpha\|u_n - u_{n-1}\|^2_{\mathcal{T}_{n-1}}\ . \tag{3.11}$$

Inserting (3.7-3.11) into (3.6) we find

$$\begin{aligned}&\hat{L}_c(y_n, u_n, \lambda_n) - \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1}) \leq \\ &-\frac{1}{2}\|y_{n-1} - y_n\|^2 - \frac{\alpha}{2}\|u_{n-1} - u_n\|^2_{\mathcal{I}_n^*} - \frac{\alpha}{2}\|u_{n-1} - u_n\|^2_{\mathcal{I}_n\backslash\mathcal{I}_n^*} - \frac{\alpha}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}} \\ &+\|\Delta^{-1}\|\ \|y_n - y_{n-1}\|_\Omega\|u_n - u_{n-1}\|_{\mathcal{T}_{n-1}} + \alpha\|u_n - u_{n-1}\|^2_{\mathcal{T}_{n-1}} \\ &+\frac{c}{2}\|u_{n-1} - u_n\|^2_{\mathcal{I}_n^*} - \frac{c}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}}\ .\end{aligned} \tag{3.12}$$

Using $ab \leq \frac{1}{2}(\frac{a^2}{\rho} + \rho b^2)$ for every $\rho > 0$ and relation (3.8), we get for $c \geq \alpha$

$$\begin{aligned}&\hat{L}_c(y_n, u_n, \lambda_n) - \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1}) \leq \\ &-\frac{1}{2}\|y_{n-1} - y_n\|^2 + \frac{(c-\alpha)}{2}\|u_{n-1} - u_n\|^2_{\mathcal{I}_n^*} + \frac{(\alpha - c)}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}} \\ &+\frac{\|\Delta^{-1}\|}{2\rho}\|y_{n-1} - y_n\|^2 + \frac{\rho\|\Delta^{-1}\|}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}} \quad \leq \\ &-\frac{1}{2}\|y_{n-1} - y_n\|^2 + \frac{(c-\alpha)\|\Delta^{-1}\|^2}{2\alpha^2}\|y_{n-1} - y_n\|^2 \\ &+\frac{\alpha - c + \rho\|\Delta^{-1}\|}{2}\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}} + \frac{\|\Delta^{-1}\|}{2\rho}\|y_{n-1} - y_n\|^2 \ = \\ &\frac{1}{2}\left[(c-\alpha)\frac{\|\Delta^{-1}\|^2}{\alpha^2} + \frac{\|\Delta^{-1}\|}{\rho} - 1\right]\|y_{n-1} - y_n\|^2 + \frac{1}{2}(\alpha + \rho\|\Delta^{-1}\| - c)\|u_{n-1} - u_n\|^2_{\mathcal{T}_{n-1}}\ .\end{aligned}$$

Setting $\gamma = \rho\|\Delta^{-1}\|$ then $\hat{L}_c(y_n, u_n, \lambda_n) \leq \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})$ provided that

$$\left[[\frac{(c-\alpha)}{\alpha^2} + \frac{1}{\gamma}]\|\Delta^{-1}\|^2 - 1\right] \leq 0 \quad \text{and} \quad \alpha + \gamma - c \leq 0\ .$$

The latter condition is equivalent to

$$\alpha + \gamma \leq c \leq \alpha - \frac{\alpha^2}{\gamma} + \frac{\alpha^2}{\|\Delta^{-1}\|^2} \ . \tag{3.4}$$

If the second inequality is strict then $\hat{L}_c(y_n, u_n, \lambda_n) < \hat{L}_c(y_{n-1}, u_{n-1}, \lambda_{n-1})$ except if $y_n = y_{n-1}$. In this latter case $u_n = u_{n-1}$ as well and the Algorithm stops at the solution to $(\mathcal{S})$. ■

Remark 3.1 *Note that for the choice $\gamma = \alpha$ condition (3.4) is equivalent to*

$$2\ \alpha \leq c \leq \frac{\alpha^2}{\|\Delta^{-1}\|^2} \ . \tag{3.13}$$

Remark 3.2 *If there exists γ such that (3.4) holds, then necessarily*

$$c > \alpha \geq 2\|\Delta^{-1}\|^2$$

holds. Indeed, assume that $\alpha < 2\|\Delta^{-1}\|^2$. Then

$$\alpha + \gamma < \alpha - \frac{\alpha^2}{\gamma} + 2\alpha \ ,$$

that is

$$\gamma^2 - 2\alpha\gamma + \alpha^2 = (\gamma - \alpha)^2 < 0 \ ,$$

which is a contradiction.

3.2 The Discrete Case

So far we have given a sufficient condition for $\hat{L}_c$ to act as a merit function for which the Algorithm has a strict descent property. In particular this eliminates the possibility of chattering of the Algorithm: it will not return to the same active set a second time. If the control and state spaces are discretized then the descent property can be used to argue convergence in a finite number of steps. More precisely, assume that a finite difference or finite element based approximation to $(\mathcal{P})$ results in

$$(\mathcal{P}^{N,M}) \qquad \begin{array}{ll} \min & J^{N,M}(Y,U) = \dfrac{1}{2}\|M_1^{\frac{1}{2}}(Y - Z_d)\|^2_{\mathbb{R}^N} + \dfrac{\alpha}{2}\|M_2^{\frac{1}{2}}(U - U_d)\|^2_{\mathbb{R}^M} \ , \\ & S\ Y = M_3\ U \ , \\ & U \leq B \ . \end{array}$$

Here Y and Z_d denotes vectors in $\mathbb{R}^N$ corresponding to the discretization of y and z_d, and $U,\ U_d$ and B denote vectors in $\mathbb{R}^M$, corresponding to the discretizations of $u,\ u_d$ and b. Further $M_1,\ S$ and M_2 are respectively $N \times N,\ N \times N$ and $M \times M$ positive definite matrices while M_3 is an $N \times M$ matrix. The norms in $(\mathcal{P}^{N,M})$ denote Euclidian norms and the inequality is understood coordinatewise. Finally, it is assumed that M_2 is a diagonal matrix.

It is simple to argue the existence of a solution (Y^*, U^*) to $(\mathcal{P}^{N,M})$. A first order optimality system is given by

$$\left\{ \begin{array}{rcl} S\ Y^* & = & M_3\ U^* \\ S\ P^* & = & -M_1(Y^* - Z_d) \\ U^* & = & U_d + \dfrac{1}{\alpha} M_2^{-1}(M_3^\top P^* - \Lambda^*) \\ \Lambda^* & = & c \max(0, U^* + \dfrac{1}{c}\Lambda^* - B)\ , \end{array} \right. \tag{3.14}$$

with $(P^*, \Lambda^*) \in \mathbb{R}^N \times \mathbb{R}^M$, for every $c > 0$. Here max is understood coordinatewise. The algorithm for the discretized problem is given next.

Discretized Algorithm

1. Initialization : choose $Y^o,\ U^o$ and Λ^o, and set $n = 1$.

2. Determine the following subsets of $\{1, \ldots, M\}$:

$$A_n = \{\ i \mid U_i^{n-1} + \frac{1}{c}\Lambda_i^{n-1} > B_i\ \}\ ,\quad I_n = \{1, \ldots, M\} \backslash A_n\ .$$

3. If $n \geq 2$ and $A_n = A_{n-1}$ then STOP.

4. Else, find $(Y^n, P^n) \in \mathbb{R}^N \times \mathbb{R}^N$ such that

$$\begin{array}{l} S\, Y^n = M_3 \left\{ \begin{array}{ll} B & \text{in } A_n \\ U_d + \dfrac{1}{\alpha} M_2^{-1} M_3^\top\ P^n & \text{in } I_n\ , \end{array} \right. \\ S\, P^n = -M_1(Y^n - Z_d) \end{array}$$

   and set

$$U^n = \left\{ \begin{array}{ll} B & \text{in } A_n \\ U_d + \dfrac{1}{\alpha} M_2^{-1} M_3^\top\ P^n & \text{in } I_n\ , \end{array} \right.$$

5. Set $\Lambda^n = M_3^\top P^n - \alpha M_2(U^n - U_d)$, update $n = n + 1$ and goto 2.

The following corollary describing properties of the Discretized Algorithm can be obtained with techniques analogous to those utilized above for analysing the continuous Algorithm. We shall denote

$$\underline{m_2} = \min_i (M_2)_{i,i},\ \overline{m_2} = \max_i (M_2)_{i,i} \text{ and } K = \|M_2^{-1} M_3^\top\|\ \|S^{-1} M_1\|\ .$$

Corollary 3.1 *If*

$$\overline{m_2}\,(\alpha + \gamma) \leq c < \alpha\, \underline{m_2} - \frac{\alpha^2}{\gamma} + \frac{\alpha^2 \|M_1\|}{K} \tag{3.15}$$

holds for some $\gamma > 0$ then the Discretized Algorithm converges in finitely many steps to the solution of $(\mathcal{P}^N)$.

Proof - First we observe that if the Discretized Algorithm stops in Step 3 then the current iterate gives the unique solution. Then we show with an argument analogous to that of the proof of Theorem 3.1 that with (3.15) holding, we have $L_c^{N,M}(Y_n, U_n, \Lambda_n) < L_c^{N,M}(Y_{n-1}, U_{n-1}, \Lambda_{n-1})$ or $(Y_n, U_n) = (Y_{n-1}, U_{n-1})$, where the discretized merit function is given by

$$L_c^{N,M}(Y, U, \Lambda) = \frac{1}{2}\|M_1^{\frac{1}{2}}\,(Y - Z_d)\|^2_{\mathbb{R}^N} + \frac{\alpha}{2}\|M_2^{\frac{1}{2}}\,(U - U_d)\|^2_{\mathbb{R}^M} + (\,\Lambda, \hat{g}_c(U, \Lambda))_{\mathbb{R}^M} + \frac{c}{2}\|\hat{g}_c(U, \Lambda)\|^2_{\mathbb{R}^M}\,,$$

with $\hat{g}_c(U, \Lambda) = (\max(U_1 - B_1, -\frac{\Lambda_1}{c}), \ldots, \max(U_M - B_M, -\frac{\Lambda_M}{c}))^\top$. If $(Y_n, U_n) = (Y_{n-1}, U_{n-1})$ then $A_{n+1} = A_n$ and the Discretized Algorithm stops at the solution. The case $L_c^{N,M}(Y_n, U_n, \Lambda_n) < L_c^{N,M}(Y_{n-1}, U_{n-1}, \Lambda_{n-1})$ cannot occur for infinitely many n since there are only finitely many different combinations of active index sets. In fact, assume that there exists $p < n$ such that $A_n = A_p$ and $I_n = I_p$. Since (Y_n, U_n) is a solution of the optimality system of Step 4 if and only if (Y_n, U_n) is the unique solution of

$$\min\{\; J^{N,M}(y, u) \mid S\; Y = M_3\; U,\; U = B \;\text{ in } A_n\;\}\;,$$

it follows that $Y_n = Y_p$, $U_n = U_p$ and $\Lambda_n = \Lambda_p$. This contradicts $L_c^{N,M}(Y_n, U_n, \Lambda_n) < L_c^{N,M}(Y_p, U_p, \Lambda_p)$ and ends the proof. ■

4 Ascent properties of Algorithm

In the previous section sufficient conditions for convergence of the Algorithm in terms of α, c and $\|\Delta^{-1}\|$ were given. Numerical experiments showed that the Algorithm converges also for values of α, c and $\|\Delta^{-1}\|$ which do not satisfy the conditions of Theorems 3.1. In fact the only possibility of constructing an example for which the Algorithm has some difficulties (which will be made precise in the following section) is based on violating the strict complementarity condition.

Thus one is challenged to further justify theoretically the efficient behavior of the Algorithm. In the tests that were performed it was observed that the cost functional was always increasing so that in practice the Algorithm behaves like an infeasible algorithm. To parallel theoretically this behavior of the Algorithm as far as possible, we slightly modify the Algorithm. For the modified Algorithm an ascent property of the cost J will be shown.

Modified Algorithm

1. Initialization : choose u_o, y_o and λ_o; set $n = 1$.

2. (a) Determine the following subsets of Ω :

$$\mathcal{A}_n = \{\; x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} > b\;\}\;,\quad \mathcal{I}_n = \{\; x \mid u_{n-1}(x) + \frac{\lambda_{n-1}(x)}{c} \le b\;\}\;,$$

(b) and find $(\tilde{y}, \tilde{p}) \in H_o^1(\Omega) \times H_o^1(\Omega)$ such that

$$-\Delta\tilde{y} = \begin{cases} b & \text{in } \mathcal{A}_n \\ u_d + \dfrac{\tilde{p}}{\alpha} & \text{in } \mathcal{I}_n \end{cases},$$
$$-\Delta\tilde{p} = z_d - \tilde{y} \text{ in } \Omega .$$

and set

$$\tilde{u} = \begin{cases} b & \text{in } \mathcal{A}_n \\ u_d + \dfrac{\tilde{p}}{\alpha} & \text{in } \mathcal{I}_n \end{cases},$$

3. $\tilde{\lambda} = \tilde{p} - \alpha(\tilde{u} - u_d)$.

4. Set

$$\widetilde{\mathcal{A}} = \{ x \mid \tilde{u}(x) + \frac{\tilde{\lambda}(x)}{c} > b \} .$$

If $\widetilde{\mathcal{A}} = \mathcal{A}_n$ then STOP, else goto 5.

5. Check for $J(\tilde{y}, \tilde{u}) > J(y_{n-1}, u_{n-1})$.

(a) If $J(\tilde{y}, \tilde{u}) > J(y_{n-1}, u_{n-1})$ then
$n = n + 1$, $y_n = \tilde{y}$, $u_n = \tilde{u}$, $\lambda_n = \tilde{\lambda}$ and goto 2a.

(b) Otherwise, determine

$$\mathcal{T}_{n-1} = \{ x \in \mathcal{I}_{n-1} \mid u_{n-1}(x) > b \} .$$

- If measure of $\mathcal{T}_{n-1}$ is null then STOP;
- else set

$$\mathcal{A}_n = \mathcal{A}_{n-1} \cup \mathcal{T}_{n-1} \quad , \quad \mathcal{I}_n = \mathcal{I}_{n-1} \backslash \mathcal{T}_{n-1} ,$$

then goto 2b.

Theorem 4.1 *If the Modified Algorithm stops in Step 4, then $(\tilde{u}, \tilde{y}, \tilde{\lambda})$ is the solution to $(\mathcal{S})$. If it never stops in Step 5b, then the sequence $J(y_n, u_n)$ $(n \geq 2)$ is strictly increasing and converges to some J^*.*

Proof - Let us first assume that the algorithm stops in Step 4. In case $\mathcal{A}_n$ is calculated from 2a then $(\tilde{u}, \tilde{y}, \tilde{\lambda})$ is the solution to $(\mathcal{S})$ by Theorem 2.1. If $\mathcal{A}_n$ is determined from 5b then an argument analogous to that used in the proof of Theorem 2.1 allows to argue that again $(\tilde{u}, \tilde{y}, \tilde{\lambda})$ is the solution to $(\mathcal{S})$.

Next we assume that algorithm never stops in Step 4. Let us consider an iteration level, where the check for ascent in Step 5a is not passed. Consequently $\mathcal{A}_n$ and $\mathcal{I}_n$ are redefined

according to step 5b and $(\tilde{y}, \tilde{u})$ are recalculated from 2b. We have already noticed that $(\tilde{y}, \tilde{u})$ is a solution of the optimality system of Step 2b if and only if $(\tilde{y}, \tilde{u})$ is the unique solution of

$$(\mathcal{P}_{aux}) \qquad \min\{\ J(y,u) \mid -\Delta y = u \text{ in } \Omega\ , y \in H_o^1(\Omega),\ u = b \text{ in } \mathcal{A}_n\ \}\ .$$

Since $\mathcal{A}_n = \mathcal{A}_{n-1} \cup \mathcal{T}_{n-1}$ strictly contains $\mathcal{A}_{n-1}$ it necessary follows that

$$J(y_{n-1}, u_{n-1}) \leq J(\tilde{y}, \tilde{u})\ . \tag{4.1}$$

It will next be shown that equality in (4.1) is impossible. In fact if $J(\tilde{y}, \tilde{u}) = J(y_{n-1}, u_{n-1})$ then due to uniqueness of the solution to $(\mathcal{P}_{aux})$ it follows that $(\tilde{y}, \tilde{u}) = (y_{n-1}, u_{n-1})$ and consequently $\tilde{\lambda} = \lambda_{n-1}$. On $\mathcal{A}_n = \mathcal{A}_{n-1} \cup \mathcal{T}_{n-1}$, we get $\tilde{u} = b = u_{n-1}$. This implies that $u_{n-1} = b$ on $\mathcal{T}_{n-1}$ and gives a contradiction to the assumption that the measure of $\mathcal{T}_{n-1}$ is non null. Hence $J(y_{n-1}, u_{n-1}) = J(\tilde{y}, \tilde{u})$ is impossible. Together with (4.1) it follows that $J(y_{n-1}, u_{n-1}) < J(\tilde{y}, \tilde{u})$ and thus the sequence $\{J(y_n, u_n)\}$ generated by the Modified Algorithm is strictly increasing. The pair (y_b, b) with $-\Delta y_b = b$ in Ω is feasible for all $(\mathcal{P}_{aux})$ so that $J(y_n, u_n) \leq J(y_b, b)$. It follows that $J(y_n, u_n)$ is convergent to some J^*.
We note, in addition that $\tilde{u}$ is feasible since $\tilde{u} = u_{n-1} = u_{n-1} + \frac{\lambda_{n-1}}{c} \leq b$ on $\mathcal{I}_n$ ($\lambda_{n-1} = \tilde{\lambda} = 0$ on $\mathcal{I}_n$). ■

The previous result can be strengthened in the case where $(\mathcal{P})$ is discretized as in subsection 3.1.

Corollary 4.1 *If the Modified Algorithm is discretized as described in the previous section and if it never stops in Step 5b, then the (discretized) solution is obtained in finitely many steps.*

Proof - Unless the algorithm stops in Step 4, the values of $J^N(Y_n, U_n)$ $(n \geq 2)$ are strictly increasing. As argued in the proof of Corollary 3.1 at each level of the iteration the minimization is carried out over an active set different from all those that have been computed before. As there are only finitely many different possibilities for active sets, the Modified Algorithm terminates in Step 4 at the unique solution of $(\mathcal{S})$. ■

We have not found a numerical example in which the Modified Algorithm terminates in Step 5b.

5 Numerical Experiments

In this section we report on numerical tests with the proposed Algorithm. For these tests we chose $\Omega =]0,1[\times]0,1[$ and the five-point finite difference approximation of the Laplacian. Unless otherwise specified the discretization was carried out on a uniform mesh with grid size 1/50.

For the chosen dimension $\|\Delta^{-1}\| = \dfrac{1}{2\pi^2}$ so that $\dfrac{1}{\|\Delta^{-1}\|^2} = 4\pi^4 \simeq 390$. Relation (3.13) which is required for the applicability of Theorem 3.1 is satisfied if $\alpha \geq 5.\ 10^{-3}$ to get the convergence via Theorem 3.1. Nevertheless we have also tested the method for smaller values of α.

The tests were performed on an HP Work station using the MATLAB© package.

5.1 Example 1

We set

$$z_d(x_1, x_2) = \sin\ (2\pi x_1)\ \sin\ (2\pi x_2)\ \exp(2x_1)/6\ ,\quad b \equiv 0\ .$$

Several tests for different values for $\alpha,\ c$ and u_d were performed. We present two of them. For the first one (3.13) is satisfied with strict inequalities.

Table 2

Example 1.a: $u_d \equiv 0\ ,\ \alpha = 10^{-2}\ ,\ c = 10^{-1}$

Iteration	$\max(u_n - b)$	size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$	$\hat{L}_c(y_n, u_n, \lambda_n)$
1	4.8708e-02	1250	4.190703e-02	4.190785e-02	4.190785e-02
2	5.8230e-05	1331	4.190712e-02	4.190712e-02	4.190712e-02
3	0.0000e+00	1332	4.190712e-02	4.190712e-02	4.190712e-02
4	0.0000e+00	1332	4.190712e-02	4.190712e-02	4.190712e-02

Let us give plots of the optimal control and state.

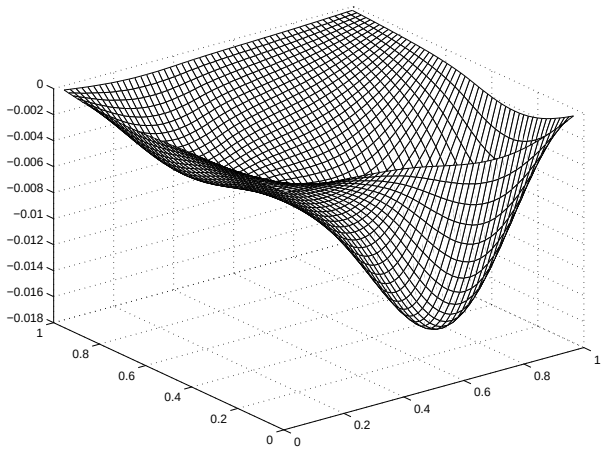

Figure 2: Optimal State

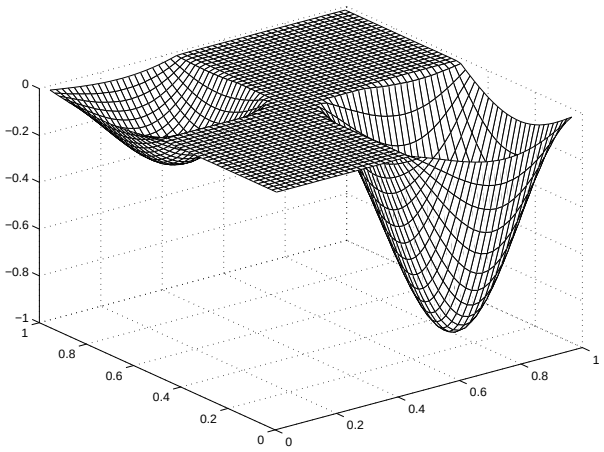

Figure 3: Optimal control

We present below a second example where (3.13) is not fulfilled because α is too small; in addition u_d has been chosen infeasible.

Table 3

Example 1.b: $u_d \equiv 1$, $\alpha = 10^{-6}$, $c = 10^{-2}$

Iteration	$\max(u_n - b)$	size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$	$\hat{L}_c(y_n, u_n, \lambda_n)$
1	5.0986e+02	1250	1.734351e-02	9.858325e+00	9.858325e+00
2	4.4728e+02	1487	2.089663e-02	7.688683e+00	7.688683e+00
3	3.6796e+02	1677	2.375001e-02	5.612075e+00	5.612075e+00
4	5.8313e+02	1831	2.603213e-02	4.526200e+00	4.526200e+00
5	6.7329e+02	1944	2.782111e-02	3.657995e+00	3.657995e+00
6	5.3724e+02	2039	2.911665e-02	2.402021e+00	2.402021e+00
7	3.6175e+02	2098	2.981378e-02	1.191161e+00	1.191161e+00
8	1.5071e+02	2146	3.011540e-02	3.678089e-01	3.678089e-01
9	6.5928e+01	2178	3.018832e-02	7.796022e-02	7.796022e-02
10	2.3420e+01	2196	3.019715e-02	3.344241e-02	3.344241e-02
11	3.4889e+00	2208	3.019762e-02	3.022994e-02	3.022994e-02
12	0.0000e+00	2210	3.019762e-02	3.019762e-02	3.019762e-02
13	0.0000e+00	2210	3.019762e-02	3.019762e-02	3.019762e-02

Though the size of the set $\mathcal{A}_n$, in the sense of number of grid points in $\mathcal{A}_n$ is increasing, the sequence $\mathcal{A}_n$ does not increase monotonically. More precisely points in $\mathcal{A}_n$ at iteration n may not belong to $\mathcal{A}_{n+1}$ at iteration $n+1$.

We observe numerically that the algorithm stops as soon as an iterate is feasible. So the sequence of iterates is not feasible until it reaches the solution. We could say that we have an "outer" method. We must also underline that differently from classical primal active set methods, the primal-dual method that we propose can move a lot of points from one iteration to the next.

We compared the new Algorithm to an Uzawa method for the augmented Lagrangian with Gauss-Seidel splitting. For convenience we recall that algorithm.

Algorithm : UGS

- Step 1. Initialization : Set $n = 1$ and choose $\gamma > 0$.

Choose $q_o \in L^2(\Omega)$ and $u_{-1} \in L^2(\Omega)$.

- Step 2. Choose $k_n \in \mathbb{N}$, set $u_n^{-1} = u_{n-1}$ and for $j = 0, \ldots, k_n$

$$\begin{aligned} y_n^j &= \text{ Arg min } \{ \; L_\gamma(y, u_n^{j-1}, q_n) \mid y \in H^2(\Omega) \cap H_o^1(\Omega) \; \} \\ u_n^j &= \text{ Arg min } \{ \; L_\gamma(y_n^j, u, q_n) \mid u \in U_{ad} \; \} \; . \end{aligned}$$

  End of the inner loop : $y_n = y_n^{k_n}$, $u_n = u_n^{k_n}$.

- Step 3.

$$q_{n+1} = q_n + \frac{\rho}{k_n + 1} \sum_{j=0}^{k_n} (A y_n^j - u_n^j), \qquad \text{where } \rho \in (0, 2\gamma] \; ,$$

where

$$L_\gamma(y, u, q) = J(y, u) + (q, Ay - u)_{L^2(\Omega)} + \frac{\gamma}{2} \|Ay - u\|^2_{L^2(\Omega)}.$$

For this algorithm a detailed convergence analysis was given in [BK]. Due to the splitting technique the second constrained minimization in Step 2 can be carried out by a simple algebraic manipulation. Algorithm UGS is an iterative algorithm that approximates the solution (y^*, u^*), whereas the new Algorithm obtains the exact (discretized) solution. For Example 1a. the computing time was 61 secs whereas the Algorithm UGS with accuracy set at 10^{-3} was stopped after 105 min. At that moment the difference between the Algorithm and Algorithm UGS was

$$|J_{ugs} - J(y^*, u^*)| \approx \; 4.10^{-8}, \; \|y_{ugs} - y^*\|_{L^\infty} \approx \; 8.10^{-7} \text{ and } \; \|u_{ugs} - u^*\|_{L^\infty} \approx \; 4.10^{-6} \; ,$$

where the index "ugs" refers to the result from Algorithm UGS. For Example 1.b the Algorithm took 191 secs whereas Algorithm UGS was stopped after 120 min.

5.2 Example 2

The desired state z_d, b are set as in the previous example and $\alpha = 10^{-2}$, $c = 10^{-1}$. This example has been constructed such that there is no strict complementarity at the solution. More precisely we have set $u_d = b - \frac{1}{\alpha}[-\Delta^{-1} z_d + \Delta^{-2} b]$ so that the exact solution of problem $(\mathcal{P})$ is $u^* = b = 0$ and $\lambda^* = 0$ and hence λ^* is not positive where the constraint is active. This example was considered by means of the optimality system $(\mathcal{S})$ of Theorem 1.1.

Table 4

$u_o \equiv 0 \ (\equiv b)$

Iteration	$\max(u_n - b)$	size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$	$\hat{L}_c(y_n, u_n, \lambda_n)$
1	4.4409e-15	1385	4.296739e-02	4.296739e-02	4.296739e-02
2	1.2546e-14	160	4.296739e-02	4.296739e-02	4.296739e-02
3	3.2752e-15	2078	4.296739e-02	4.296739e-02	4.296739e-02
4	4.5519e-15	2308	4.296739e-02	4.296739e-02	4.296739e-02
5	4.5242e-15	1613	4.296739e-02	4.296739e-02	4.296739e-02
6	4.3299e-15	1787	4.296739e-02	4.296739e-02	4.296739e-02

Here the canonical initial guess u_o coincides with the solution u^*. From the Table 3 we observe that $u_n,\ J(y_n, u_n),\ L_c(y_n, u_n)$ and $\hat{L}_c(y_n, u_n)$ remain constant while the active sets $\mathcal{A}_n$ chatter. For different initial guesses for u_o the same type of behavior is observed, the Algorithm always reaches the optimal value for u and J in one iteration, and if the stopping criterion of the Algorithm was based on the coincidence of two consecutive values of J it would stop after one iteration. The chattering of active sets is due to lack of strict complementarity and machine precision. Let us briefly consider this phenomenon and note at first that the signs in the Algorithm are set such that at the limit we should have $\Omega = \mathcal{I}^*$ (all inactive with $\lambda^* = u^* = 0$). If $x \in \mathcal{A}_{n-1}$ then $u_{n-1}(x) = 0$ by Step 4 and $\lambda_{n-1}(x) = \pm\varepsilon$, with ε equal to the computer epsilon, will decide whether $x \in \mathcal{A}_n$ or $\mathcal{I}_n$, although for numerical purposes the exact pair for (u, λ) is already obtained. If $x \in \mathcal{I}_{n-1}$ then $\lambda_{n-1} = 0$ and $u_{n-1}(x) = \pm\varepsilon$ will decide whether $x \in \mathcal{A}_n$ or $\mathcal{I}_n$, while the influence of this choice on J or L_c is of the order of ε^2 i.e. it is numerically zero. Therefore we decided to replace "$> b$" in the definition of $\mathcal{A}_n$ by "$> b - \varepsilon$ " (and $\mathcal{I}_n = \Omega \backslash \mathcal{A}_n$): the algorithm behaves now as expected and stops after 2 iterations.

5.3 Example 3

We have seen with Example 1. that the augmented Lagrangian function decreases during iterations. We show with this example that the augmented Lagrangian function may not decrease though the method is convergent and provides the exact solution. Let us precise the data :

$$z_d = \begin{cases} 200\ x_1 x_2\ (x_1 - \frac{1}{2})^2\ (1 - x_2) & \text{if } 0 < x_1 \leq 1/2\ , \\ 200\ x_2\ (x_1 - 1)(x_1 - \frac{1}{2})^2\ (1 - x_2) & \text{if } 1/2 < x_1 \leq 1\ , \end{cases}$$

$$u_d \equiv 0\ ,\ b \equiv 1,\ c = 10^{-2}\ .$$

Table 5

Example 3.a: $\alpha = 10^{-6}$, $u_o \equiv 1$ $(\equiv b)$

Iteration	$\max(u_n - b)$	size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$	$\hat{L}_c(y_n, u_n, \lambda_n)$
1	4.1995e+02	1100	3.314755e-02	9.645226e+00	9.645226e+00
2	3.8057e+02	1370	3.672870e-02	7.943326e+00	7.943326e+00
3	3.6453e+02	1300	3.963515e-02	7.393744e+00	7.393744e+00
4	3.7512e+02	1400	4.249987e-02	7.809205e+00	7.809205e+00
5	3.8952e+02	1500	4.555558e-02	8.300084e+00	8.300084e+00
6	3.9452e+02	1600	4.880515e-02	8.320358e+00	8.320358e+00
7	3.8004e+02	1700	5.203947e-02	7.485445e+00	7.485445e+00
8	3.3858e+02	1800	5.490267e-02	5.699382e+00	5.699382e+00
9	2.6458e+02	1898	5.701220e-02	3.286759e+00	3.286759e+00
10	1.5311e+02	1986	5.811845e-02	1.093548e+00	1.093548e+00
11	8.3048e+01	2040	5.834162e-02	3.099587e-01	3.099587e-01
12	1.5809e+01	2086	5.839423e-02	5.959874e-02	5.959874e-02
13	0.0000e+00	2098	5.839438e-02	5.839438e-02	5.839438e-02
14	0.0000e+00	2098	5.839438e-02	5.839438e-02	5.839438e-02

The solution was obtained in 210 secs.

The following plot shows the influence of α on the behavior of the Lagrangian function L_c.

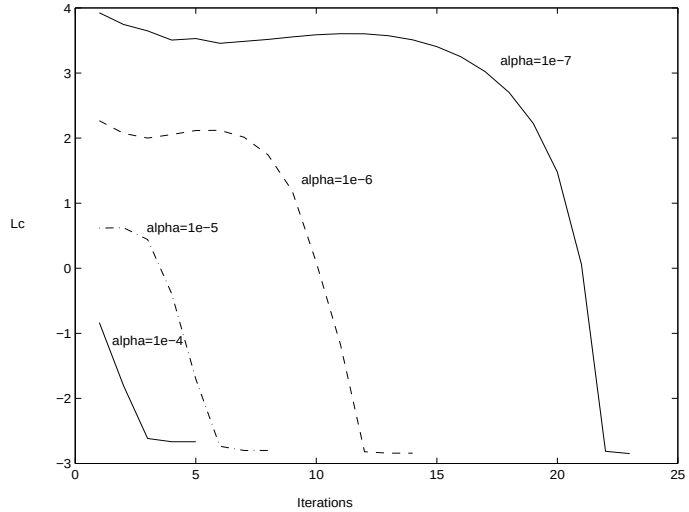

Figure 4: Influence of α on the behavior of L_c (Logarithmic scale)

We see that during the first iterations the augmented Lagrangian function does not decrease if α is too small.

However, if the initialization point is close enough the solution then this function becomes decreasing. We have tested initialization points different from b which were closer to the solution and obtained decrease of L_c. As an example we give in Table 6 the results for $\alpha = 10^{-10}$ with an initialization according to (2.1) but with u_o the solution for $\alpha = 10^{-5}$

Table 6

Example 3.b: $\alpha = 10^{-10}$, u_o given by the solution to $(\mathcal{P})$ for $\alpha = 10^{-5}$

Iteration	$\max(u_n - b)$	size of $\mathcal{A}_n$	$J(y_n, u_n)$	$L_c(y_n, u_n, \lambda_n)$
1	1.6605e+03	1986	5.696032e-02	4.889158e+01
2	1.4741e+03	2034	5.750110e-02	2.948470e+01
3	1.1542e+03	2082	5.781067e-02	1.299992e+01
4	6.8931e+02	2130	5.793424e-02	2.631407e+00
5	1.6713e+02	2168	5.795024e-02	2.198494e-01
6	1.1931e+02	2172	5.795048e-02	1.276798e-01
7	7.0091e+01	2176	5.795058e-02	7.857522e-02
8	2.0618e+01	2180	5.795061e-02	5.958497e-02
9	0.0000e+00	2182	5.795061e-02	5.795061e-02
10	0.0000e+00	2182	5.795061e-02	5.795061e-02

Note that the total number of iterations including the initialization with $\alpha = 10^{-5}$ to obtain the solution corresponding for $\alpha = 10^{-10}$ is equal to 18. If one computes the solution with initialization $u_o = b$, the number of iterations is 27 and L_c decreases after iteration 12. Thus a good initial guess can decrease the number of iterations to obtain the solution. This process was repeated successfully for smaller values of α up to $\alpha = 10^{-15}$ as well.

References

[B] **V. Barbu**, *Analysis and Control of Non Linear Infinite Dimensional Systems*, Math in Science and Engineering, Vol. 190, Academic Press 1993.

[BK] **M. Bergounioux - K. Kunisch**, *Augmented Lagrangian Techniques for Elliptic State Constrained Optimal Control Problems,* SIAM Journal on Control and Optimization, Vol.35, n^o 5 (1997).

[BS] **H. Brézis - G. Stampacchia**, *Remarks on some fourth order variational inequalities*, Annali Scuola Norm. Sup. Pisa, 4 (1977), 363-371.

[HT] **M. Heinkenschloss - F. Tröltzsch**, *Analysis of the Lagrange-SQP-Newton method for the control of a phase field equation*, Preprint, Virginia Tech.

[IK] **K. Ito - K. Kunisch**, *Augmented Lagrangian formulation of nonsmooth convex optimization in Hilbert spaces* , Lecture Notes in Pure and Applied Mathematics, Control of Partial Differential Equations and Applications, E. Casas, Ed., Marcel Dekker, Vol. 174 (1995), 107-117.

[KeS] **C.T. Kelley - E. Sachs**, *Approximate quasi-Newton methods*, Mathematical Programming 48 (1990), 41-70.

[KuS] **K. Kunisch - E. Sachs**, *Reduced SQP-methods for parameter identification problems*, SIAM Journal Numerical Analysis, 29 (1992), 1793-1820.

[K] **F.S. Kupfer**, *An infinite-dimensional convergence theory for reduced SQP-methods in Hilbert spaces*, SIAM Journal on Optimization, 6 (1996), 126-163.

[Sch] **K. Schittkowski**, *On the convergence of a sequential quadratic programming method with an augmented Lagrangian line search function*, Math. Operations forsch. u. Statist., Ser Optimization 14(1983), 197-216.

[T] **F. Tröltzsch**, *An SQP-method for optimal control of a nonlinear heat equation*, Control & Cybernetics, 23 (1994), 268-288.

THE PRIMAL-DUAL ACTIVE SET STRATEGY AS A SEMI-SMOOTH NEWTON METHOD

M. HINTERMÜLLER, K. ITO, AND K. KUNISCH

ABSTRACT. This paper addresses complementarity problems motivated by constrained optimal control problems. It is shown that the primal-dual active set strategy, which is known to be extremely efficient for this class of problems, and a specific semi-smooth Newton method lead to identical algorithms. The notion of slant differentiability is recalled and it is argued that the max-function is slantly differentiable in L^p-spaces when appropriately combined with a two-norm concept. This leads to new local convergence results of the primal-dual active set strategy. Global unconditional convergence results are obtained by means of appropriate merit functions.

1. Introduction

This paper is motivated by linearly constrained quadratic problems of the type

$$\text{(P)} \qquad \begin{cases} \min J(y) = \frac{1}{2}(y, Ay) - (f, y) \\ \text{subject to } y \leq \psi \end{cases}$$

where A is positive definite and f, ψ are given. In previous contributions [IK1, IK2, BIK, BHHK] we proposed a primal-dual active set strategy as an extremely efficient method to solve (P). We shall show in the present work that the primal-dual active set method can be interpreted as a semi-smooth Newton method. This opens up a new interpretation and perspective of analyzing the primal-dual active set method. Both the finite dimensional case with $y \in \mathbb{R}^n$ and the infinite dimensional case with $y \in L^2(\Omega)$ will be considered. While our results are quite generally applicable the main motivation arises from infinite dimensional constrained variational problems and their discretization. Frequently such problems have a special structure which can be exploited. For example, in the case of discretized obstacle problems A can be an M–matrix, and for constrained optimal control problems A is a smooth additive perturbation of the identity operator.

The analysis of semi-smooth problems and the Newton–algorithm to solve such problems has a long history for finite dimensional problems. We refer to selected papers [Q1, Q2, QS] and the references given there. Typically, under appropriate semi-smoothness and regularity assumptions locally superlinear convergence rates of semi-smooth Newton methods are obtained. Since many definitions used in the above papers depend on Rademacher's theorem, which has no analogue in infinite dimensions, very recently e.g. in [CNQ, U] new concepts for generalized derivatives and semi-smoothness in infinite dimensional spaces were introduced. In our work we use primarily the notion of slant differentiability from [CNQ] which we recall for the reader's convenience at the end of this section. For the problem under consideration it coincides with the differentiability concept in [U]. This will be explained in Section 4.

Let us briefly outline the structure of the paper. In Section 2 the relationship between the primal-dual active set method and semi-smooth Newton methods is explained. Local as well as global convergence for finite dimensional problems which is unconditional with respect to initialization in certain cases is addressed in Section 3. The global convergence results depend on properties of the matrix A. For instance, the M-matrix property required in Theorem 3.2 is typically obtained when

discretizing obstacle problems (see e.g. [H, KNT]) by finite differences or finite elements. Theorem 3.3 can be connected to discretizations of control constrained optimal control problems. Some relevant numerical aspects of the conditions of Theorem 3.3 are discussed at the end of Section 4. An instance of the perturbation result of Theorem 3.4 is given by discretized optimal control problems with sufficiently small cost parameter. Perturbations of M-matrices resulting from discretized obstacle problems and state constrained optimal control problems (see e.g. [C]) fit into the framework of Theorem 3.4. In Section 4 slant differentiability properties of the max-function between function spaces are analyzed. Superlinear convergence of semi-smooth Newton methods for optimal control problems with pointwise control constraints is proved. Several alternative methods were analyzed to solve optimal control problems with pointwise constraints on the controls. Among them are the projected Newton method, analyzed e.g. in [HKT, KS], and affine scaling interior point Newton methods [UU]. We plan to address nonlinear problems in future work. Let us stress, however, that nonlinear iterative methods frequently rely on solving auxiliary problems of the type (P) and solving them efficiently is important.

To briefly describe some of the previous work in the primal dual active set method, we recall that this method arose as a special case of generalized Moreau-Yosida approximations to nondifferentiable convex functions [IK1]. Global convergence proofs based on a modified augmented Lagrangian merit function are contained in [BIK]. In [BHHK] comparisons between the primal-dual active set method and interior point methods are carried out. In [IK2] the primal-dual active set method was used to solve optimal control of variational inequalities problems. For this class of problems convergence proofs are not yet available.

We now turn to the notion of differentiability which will be used in this paper. Let X and Z be Banach spaces and consider the nonlinear equation

$$F(x) = 0\,, \tag{1.1}$$

where $F\colon D \subset X \to Z$, and D is an open subset of X.

Definition 1. The mapping $F\colon D \subset X \to Z$ is called slantly differentiable in the open subset $U \subset D$ if there exists a family of mappings $G\colon U \to \mathcal{L}(X, Z)$ such that

$$\lim_{h\to 0} \frac{1}{\|h\|} \|F(x+h) - F(x) - G(x+h)h\| = 0, \tag{A}$$

for every $x \in U$.

We refer to G as a slanting function for F in U. Note that G is not required to be unique to be a slanting function for F in U. The definition of slant differentiability in an open set is a slight adaptation of the terminology introduced in [CNQ], where in addition it is required that $\{G(x) : x \in U\}$ is bounded in $\mathcal{L}(X, Z)$. In [CNQ] also the term slant differentiability at a point is introduced. In applications to Newton's method this presupposes knowledge of the solution, whereas slant differentiability of F in U requires knowledge of a set which contains the solution. Under the assumption of slant differentiability in an open set Newton's method converges superlinearly for appropriate choices of the initialization. Thus the assumption of slant differentiability in an open set parallels the hypothesis of knowledge of the domain within which a second order sufficient optimality condition is satisfied for smooth problems.

Kummer [K2] introduced a notion similar to slant differentiability at a point and coined the name Newton map. He also pointed out the discrepancy between the requirements needed for numerical realization and for the proof of superlinear convergence of the semi-smooth Newton method.

The following convergence result is already known [CNQ].

Theorem 1.1. *Suppose that x^* is a solution to (1.1) and that F is slantly differentiable in an open neighborhood U containing x^* with slanting function $G(x)$. If $G(x)$ is nonsingular for all $x \in U$ and $\{\|G(x)^{-1}\| : x \in U\}$ is bounded, then the Newton–iteration*

$$x^{k+1} = x^k - G(x^k)^{-1}F(x^k)$$

converges superlinearly to x^* provided that $\|x^0 - x^*\|$ is sufficiently small.

We provide the short proof since it will be used to illustrate the subsequent discussion.

Proof. Note that the Newton iterates satisfy

$$\|x^{k+1} - x^*\| \le \|G(x^k)^{-1}\| \, \|F(x^k) - F(x^*) - G(x^k)(x^k - x^*)\|, \tag{1.2}$$

provided that $x^k \in U$. Let $B(x^*, r)$ denote a ball of radius r centered at x^* contained in U and let M be such that $\|G(x)^{-1}\| \le M$ for all $x \in B(x^*, r)$. We apply (A) with $x = x^*$. Let $\eta \in (0, 1]$ be arbitrary. Then there exists $\rho \in (0, r)$ such that

$$\|F(x^* + h) - F(x^*) - G(x^* + h)h\| < \frac{\eta}{M} \, \|h\| \le \frac{1}{M} \, \|h\| \tag{1.3}$$

for all $\|h\| < \rho$. Consequently, if we choose x^0 such that $\|x^0 - x^*\| < \rho$, then by induction from (1.2), (1.3) with $h = x^k - x^*$ we have $\|x^{k+1} -$

$x^*\| < \rho$ and in particular $x^{k+1} \in B(x^*, \rho)$. It follows that the iterates are well-defined. Moreover, since $\eta \in (0, 1]$ is chosen arbitrarily $x^k \to x^*$ converges superlinearly. □

Note that replacing property (A) by a condition of the type

$$\lim_{h\to 0} \frac{1}{\|h\|} \|F(x) - F(x-h) - G(x)h\| = 0$$

would require a uniformity assumption with respect to $x \in U$, for Theorem 1.1 to remain valid in case X is infinite dimensional.

Let us put the concept of slant differentiability into a perspective with the notion of semi-smoothness as introduced in [QS] in finite dimensions. Semi-smoothness of $F : U \subset \mathbb{R}^n \to \mathbb{R}^m$ in the sense of Qi and Sun [QS] implies

$$\|F(x+h) - F(x) - Vh\| = o(\|h\|)\,, \tag{1.4}$$

for $x \in U$, where V is an arbitrary element of the generalized Jacobian $\partial F(x+h)$ in the sense of Clarke [C, Prop. 2.6.2]. Thus, slant differentiability introduced in Definition 1 is a more general concept. In fact, the slanting functions according to Definition 1 are not required to be elements of $\partial F(x+h)$. On the other hand, if (1.4) holds for $x \in U \subset \mathbb{R}^n$, then a single-valued selection $V(x) \in \partial F(x)$, $x \in U$, serves as a slanting function in the sense of Definition 1.

We shall require the notion of a P-matrix which we recall next.

Definition 2. An $n \times n$-matrix is called a P-matrix if all its principal minors are positive.

It is well-known [BP] that A is a P-matrix if and only if all real eigenvalues of A and of its principal submatrices are positive. Here B is called a principal submatrix of A if it arises from A by deletion of rows and columns from the same index set $\mathcal{J} \subset \{1, \dots, n\}$.

2. The primal-dual active set strategy as semi-smooth Newton method

In this section we consider complementarity problems of the form

$$\begin{cases} Ay + \lambda = f, \\ y \le \psi,\ \lambda \ge 0,\ (\lambda, y - \psi) = 0\,, \end{cases} \tag{2.1}$$

where $(\cdot, \cdot)$ denotes the inner product in $\mathbb{R}^n$, A is an $n \times n$-valued P-matrix and f, $\psi \in \mathbb{R}^n$. The assumption that A is a P-matrix guarantees

the existence of a unique solution $(y^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^n$ of (2.1) [BP]. In case A is symmetric positive definite (2.1) is the optimality system for

$$\text{(P)} \qquad \begin{cases} \min J(y) = \dfrac{1}{2}(y, Ay) - (f, y) \\ \text{subject to } y \leq \psi. \end{cases}$$

Note that the complementarity system given by the second line in (2.1) can equivalently be expressed as

$$\mathcal{C}(y, \lambda) = 0, \text{ where } \mathcal{C}(y, \lambda) = \lambda - \max(0, \lambda + c(y - \psi)), \tag{2.2}$$

for each $c > 0$. Here the max–operation is understood component-wise.

Consequently (2.1) is equivalent to

$$\begin{cases} Ay + \lambda = f \\ \mathcal{C}(y, \lambda) = 0. \end{cases} \tag{2.3}$$

The primal-dual active set method is based on using (2.2) as a prediction strategy, i.e. given a current primal-dual pair (y, λ) the choice for the next active and inactive sets is given by

$$\mathcal{I} = \{i \colon \lambda_i + c(y - \psi)_i \leq 0\}, \text{ and } \mathcal{A} = \{i \colon \lambda_i + c(y - \psi)_i > 0\}.$$

This leads to the following algorithm.

Primal-dual active set algorithm.

(i) Initialize y^0, λ^0. Set $k = 0$.
(ii) Set $\mathcal{I}_k = \{i \colon \lambda_i^k + c(y^k - \psi)_i \leq 0\}$, $\mathcal{A}_k = \{i \colon \lambda_i^k + c(y^k - \psi)_i > 0\}$.
(iii) Solve

$$Ay^{k+1} + \lambda^{k+1} = f$$

$$y^{k+1} = \psi \text{ on } \mathcal{A}_k, \lambda^{k+1} = 0 \text{ on } \mathcal{I}_k.$$

(iv) Stop, or set $k = k + 1$ and return to (ii).

Above we utilize $y^{k+1} = \psi$ on $\mathcal{A}_k$ to stand for $y_i^{k+1} = \psi_i$ for $i \in \mathcal{A}_k$. Let us now argue that the above algorithm can be interpreted as a semi-smooth Newton method. For this purpose it will be convenient to arrange the coordinates in such a way that the active and inactive ones occur in consecutive order. This leads to the block matrix representation of A as

$$A = \begin{pmatrix} A_{\mathcal{I}_k} & A_{\mathcal{I}_k \mathcal{A}_k} \\ A_{\mathcal{A}_k \mathcal{I}_k} & A_{\mathcal{A}_k} \end{pmatrix},$$

where $A_{\mathcal{I}_k} = A_{\mathcal{I}_k \mathcal{I}_k}$ and analogously for $A_{\mathcal{A}_k}$. Analogously the vector y is partitioned according to $y = (y_{\mathcal{I}_k}, y_{\mathcal{A}_k})$ and similarly for f and ψ. In Section 3 we shall argue that $v \to \max(0, v)$ from $\mathbb{R}^n \to \mathbb{R}^n$

is slantly differentiable with a slanting function given by the diagonal matrix $G_m(v)$ with diagonal elements

$$G_m(v)_{ii} = \begin{cases} 1 & \text{if} \quad v_i > 0, \\ 0 & \text{if} \quad v_i \leq 0. \end{cases}$$

Here we use the subscript m to indicate particular choices for the slanting function of the max-function. Note that G_m is also an element of the generalized Jacobian (see [C, Definition 2.6.1]) of the max-function. Semi-smooth Newton methods for generalized Jacobians in Clarke's sense were considered e.g. in [Q1, QS].

The choice G_m suggests a semi-smooth Newton step of the form

$$\begin{pmatrix} A_{\mathcal{I}_k} & A_{\mathcal{I}_k\mathcal{A}_k} & I_{\mathcal{I}_k} & 0 \\ A_{\mathcal{A}_k\mathcal{I}_k} & A_{\mathcal{A}_k} & 0 & I_{\mathcal{A}_k} \\ 0 & 0 & I_{\mathcal{I}_k} & 0 \\ 0 & -cI_{\mathcal{A}_k} & 0 & 0 \end{pmatrix} \begin{pmatrix} \delta y_{\mathcal{I}_k} \\ \delta y_{\mathcal{A}_k} \\ \delta\lambda_{\mathcal{I}_k} \\ \delta\lambda_{\mathcal{A}_k} \end{pmatrix} = - \begin{pmatrix} (Ay^k + \lambda^k - f)_{\mathcal{I}_k} \\ (Ay^k + \lambda^k - f)_{\mathcal{A}_k} \\ \lambda^k_{\mathcal{I}_k} \\ -c(y^k - \psi)_{\mathcal{A}_k} \end{pmatrix} \tag{2.4}$$

where $I_{\mathcal{I}_k}$ and $I_{\mathcal{A}_k}$ are identity matrices of dimensions $\mathrm{card}(\mathcal{I}_k)$ and $\mathrm{card}(\mathcal{A}_k)$. The third equation in (2.4) implies that

$$\lambda^{k+1}_{\mathcal{I}_k} = \lambda^k_{\mathcal{I}_k} + \delta\lambda_{\mathcal{I}_k} = 0 \tag{2.5}$$

and the last one yields

$$y^{k+1}_{\mathcal{A}_k} = \psi_{\mathcal{A}_k}. \tag{2.6}$$

Equations (2.5) and (2.6) coincide with the conditions in the second line of step (iii) in the primal-dual active set algorithm. The first two equations in (2.4) are equivalent to $Ay^{k+1} + \lambda^{k+1} = f$, which is the first equation in step (iii).

Combining these observations we can conclude that the semi-smooth Newton update based on (2.4) is equivalent to the primal-dual active set strategy.

We also note that the system (2.4) is solvable since the first equation in (2.4) together with (2.5) gives

$$(A\ \delta y)_{\mathcal{I}_k} + (A\ y^k)_{\mathcal{I}_k} = f_{\mathcal{I}_k},$$

and consequently by (2.6)

$$A_{\mathcal{I}_k}\ y^{k+1}_{\mathcal{I}_k} = f_{\mathcal{I}_k} - A_{\mathcal{I}_k\mathcal{A}_k}\ \psi_{\mathcal{A}_k}. \tag{2.7}$$

Since A is a P-matrix $A_{\mathcal{I}_k}$ is regular and (2.7) determines $y^{k+1}_{\mathcal{I}_k}$. The second equation in (2.4) is equivalent to

$$\lambda^{k+1}_{\mathcal{A}_k} = f_{\mathcal{A}_k} - (Ay^{k+1})_{\mathcal{A}_k}. \tag{2.8}$$

In Section 4 we shall consider (P) in the space $L^2(\Omega)$. Again one can show that the semi-smooth Newton update and the primal-dual active set strategy coincide.

3. Convergence analysis: the finite dimensional case

This section is devoted to local as well as global convergence analysis of the primal-dual active set algorithm to solve

$$\begin{cases} Ay + \lambda = f \\ \lambda - \max(0, \lambda + c(y - \psi)) = 0, \end{cases} \tag{3.1}$$

where $f \in \mathbb{R}^n$, $\psi \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ is a P-matrix and the max-operation is understood component-wise. To discuss slant differentiability of the max-function we define for an arbitrarily fixed $\delta \in \mathbb{R}^n$ the matrix-valued function $G_m \colon \mathbb{R}^n \to \mathbb{R}^{n\times n}$ by

$$G_m(y) = \text{diag } (g_1(y_1), \cdots, g_n(y_n)), \tag{3.2}$$

where $g_i \colon \mathbb{R} \to \mathbb{R}$ is given by

$$g_i(z) = \begin{cases} 0 & \text{if} \quad z < 0\,, \\ 1 & \text{if} \quad z > 0\,, \\ \delta_i & \text{if} \quad z = 0\,. \end{cases}$$

Lemma 3.1. *The mapping $y \to \max(0, y)$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ is slantly differentiable on $\mathbb{R}^n$ and G_m defined in (3.2) is a slanting function for every $\delta \in \mathbb{R}^n$.*

Proof. Clearly $G_m \in \mathcal{L}(\mathbb{R}^n)$ and $\{\|G_m(y)\| \colon y \in \mathbb{R}^n\}$ is bounded. We introduce $D \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$D(y, h) = \| \max(0, y + h) - \max(0, y) - G_m(y + h)h\|.$$

It is simple to check that

$$D(y, h) = 0 \ \text{ if } \ \|h\|_\infty < \min \{|y_i| \colon y_i \neq 0\} =: \beta.$$

Consequently the max-function is slantly differentiable. □

Remark 3.1. Note that the value of the generalized derivative G_m of the max-function can be assigned an arbitrary value at the coordinates satisfying $y_i = 0$. The numerator D in Definition 1 satisfies $D(y, h) = 0$ if $\|h\|_\infty < \beta$. Moreover, for every $\gamma > \beta$ there exists h satisfying

$$D(y, h) \geq \beta \ \text{ and } \ \|h\|_\infty = \gamma.$$

Here we assume that $\beta := 0$ whenever $\{i | y_i \neq 0\} = \emptyset$. Consequently, for $\beta > 0$ the mapping

$$\gamma \mapsto \sup \{\| \max(0, y + h) - \max(0, y) - G_m(y + h)h\|_\infty \colon \|h\|_\infty = \gamma\}$$

is discontinuous at $\gamma = \beta$ and equals zero for $\gamma \in (0, \beta)$. ⋄

Let us now turn to the convergence analysis of the primal-dual active set method or, equivalently, the semi-smooth Newton method for (3.1). Note that the choice G_m for the slanting function in Section 2 corresponds to a slanting function with $\delta = 0$. In view of (2.5)–(2.8), for $k \geq 1$ the Newton update (2.4) is equivalent to

$$\begin{pmatrix} A_{\mathcal{I}_k} & 0 \\ A_{\mathcal{A}_k\mathcal{I}_k} & I_{\mathcal{A}_k} \end{pmatrix} \begin{pmatrix} \delta y_{\mathcal{I}_k} \\ \delta\lambda_{\mathcal{A}_k} \end{pmatrix} = - \begin{pmatrix} A_{\mathcal{I}_k\mathcal{A}_k}\delta y_{\mathcal{A}_k} + \delta\lambda_{\mathcal{I}_k} \\ A_{\mathcal{A}_k}\delta y_{\mathcal{A}_k} \end{pmatrix} \tag{3.3}$$

and

$$\delta\lambda_i = -\lambda_i^k,\ i \in \mathcal{I}_k, \quad \text{and} \quad \delta y_i = \psi_i - y_i^k,\ i \in \mathcal{A}_k. \tag{3.4}$$

Let us introduce $F\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ by

$$F(y,\lambda) = \begin{pmatrix} Ay + \lambda - f \\ \lambda - \max(0, \lambda + c(y - \psi)) \end{pmatrix},$$

and note that (3.1) is equivalent to $F(y,\lambda) = 0$. As a consequence of Lemma 3.1 the mapping F is slantly differentiable and the system matrix of (2.4) is a slanting function for F with the particular choice G_m for the slanting function of the max-function. We henceforth denote the slanting function of F by G_F.

Let (y^*, λ^*) denote the unique solution to (3.1) and $x^0 = (y^0, \lambda^0)$ the initial values of the iteration. From Theorem 1.1 we deduce the following fact:

Theorem 3.1. *The primal-dual active set method or equivalently the semi-smooth Newton method converge superlinearly to $x^* = (y^*, \lambda^*)$, provided that $\|x^0 - x^*\|$ is sufficiently small.*

In our finite dimensional setting this result can be derived alternatively by observing that G_m corresponds to a generalized Jacobian in Clarke's sense combined with the convergence results for semi-smooth Newton methods in [Q1, QS]. In fact, from (2.4) we infer that $G_F(x^*)$ is a nonsingular generalized Jacobian, and Lemma 3.1 proves the semi-smoothness of F at x^*. Hence, Theorem 3.2 of [QS] yields the locally superlinear convergence property. For a discussion of the semi-smoothness concept in finite dimensions we refer to [Q1, QS].

Furthermore, since (3.1) is strongly semi-smooth, by utilizing Theorem 3.2 of [QS] the convergence rate can even be improved. Indeed, the primal-dual active set strategy converges locally with a q-quadratic rate. For the definition of strong semi-smoothness we refer to [FFKP].

We also observe that if the iterates $x^k = (y^k, \lambda^k)$ converge to $x^* = (y^*, \lambda^*)$ then they converge in finitely many steps. In fact, there are only finitely many choices of active/inactive sets and if the algorithm would determine the same sets twice then this contradicts convergence

of x^k to x^*. We refer to [FK] for a similar observation for a nonsmooth Newton method of the types discussed in [Q1, QS, K1], for example.

Let us address global convergence next. In the following two results sufficient conditions for convergence for arbitrary initial data $x^0 = (y^0, \lambda^0)$ are given. We recall that A is referred to as M-matrix, if it is nonsingular, $(m_{ij}) \leq 0$, for $i \neq j$, and $M^{-1} \geq 0$. Our notion of an M-matrix coincides with that of nonsingular M-matrices as defined in [BP].

Theorem 3.2. *Assume that A is a* M*-matrix. Then $x^k \to x^*$ for arbitrary initial data. Moreover, $y^* \leq y^{k+1} \leq y^k$ for all $k \geq 1$ and $y^k \leq \psi$ for all $k \geq 2$.*

For a proof of Theorem 3.2 we can utilize the proof of Theorem 1 in [H], where a (primal) active set algorithm is proposed and analyzed. However, we provide a proof in appendix A since in contrast to the algorithm in [H] the primal-dual active set strategy makes use of the dual variable λ and includes arbitrarily fixed $c > 0$. From the proof in Appendix A it can be seen that for unilaterally constrained problems c drops out after the first iteration. We point out that, provided the active and inactive sets coincide, the linear systems that have to be solved in every iteration of both algorithms coincide. In practice, however, λ and c play a significant role and make a distinct difference between the performance of the algorithm in [H] and the primal-dual active set strategy. In fact, the primal-dual active set strategy fixes $\lambda_i^{k+1} = 0$ for $i \in \mathcal{I}_k$. The decision whether an inactive index $i \in \mathcal{I}_k$ becomes an active one, i.e. whether $i \in \mathcal{A}_{k+1}$, is based on

$$\lambda_i^{k+1} + c(y_i^{k+1} - \psi_i) > 0\,.$$

In contrast, the (primal) active set algorithm in [H] uses the criterion

$$f_i - (Ay^{k+1})_i + (y_i^{k+1} - \psi_i) > 0$$

instead. Clearly, if the linear system of both algorithms are solved approximately (e.g. by some iterative procedure) then the numerical behavior may differ.

Remark 3.2. Concerning the applicability of Theorem 3.2 we recall that many discretizations of second order differential operators give rise to M-matrices. $\diamond$

For a rectangular matrix $B \in \mathbb{R}^{n\times m}$ we denote by $\|\cdot\|_1$ the subordinate matrix norm when both $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the 1-norms. Moreover, B_+ denotes the $n \times m$-matrix containing the positive parts of the elements of B. The following result can be applied

to discretizations of constrained optimal control problems. We refer to the end of Section 4 for a discussion of the conditions of the following Theorem 3.3 in the case of control constrained optimal control problems.

Theorem 3.3. *If A is a* P*-matrix and for every partitioning of the index set into disjoint subsets $\mathcal{I}$ and $\mathcal{A}$ we have $\|(A_{\mathcal{I}}^{-1}A_{\mathcal{I}\mathcal{A}})_+\|_1 < 1$ and $\sum_{i\in\mathcal{I}}(A_{\mathcal{I}}^{-1}y_{\mathcal{I}})_i \geq 0$ for $y_{\mathcal{I}} \geq 0$, then $\lim_{k\to\infty} x^k = x^*$.*

Proof. From (3.3) we have

$$(y^{k+1}-\psi)_{\mathcal{I}_k} = (y^k-\psi)_{\mathcal{I}_k} + A_{\mathcal{I}_k}^{-1}A_{\mathcal{I}_k\mathcal{A}_k}(y^k-\psi)_{\mathcal{A}_k} + A_{\mathcal{I}_k}^{-1}\lambda_{\mathcal{I}_k}^k$$

and upon summation over the inactive indices

$$\begin{aligned}\sum_{\mathcal{I}_k}(y_i^{k+1}-\psi_i) = \sum_{\mathcal{I}_k}(y_i^k-\psi_i) + \sum_{\mathcal{I}_k}\left(A_{\mathcal{I}_k}^{-1}A_{\mathcal{I}_k\mathcal{A}_k}(y^k-\psi)_{\mathcal{A}_k}\right)_i \\ + \sum_{\mathcal{I}_k}(A_{\mathcal{I}_k}^{-1}\lambda_{\mathcal{I}_k}^k)_i\end{aligned} \tag{3.5}$$

Adding the obvious equality

$$\sum_{\mathcal{A}_k}(y_i^{k+1}-\psi_i) - \sum_{\mathcal{A}_k}(y_i^k-\psi_i) = -\sum_{\mathcal{A}_k}(y_i^k-\psi_i)$$

to (3.5) implies

$$\sum_{i=1}^n(y_i^{k+1}-y_i^k) \leq -\sum_{\mathcal{A}_k}(y_i^k-\psi_i) + \sum_{\mathcal{I}_k}(A_{\mathcal{I}_k}^{-1}A_{\mathcal{I}_k\mathcal{A}_k}(y^k-\psi)_{\mathcal{A}_k})_i\,. \tag{3.6}$$

Here we used the fact $\lambda_{\mathcal{I}_k}^k = -\delta\lambda_{\mathcal{I}_k} \leq 0$, established in the proof of Theorem 3.2. There it was also argued that $y_{\mathcal{A}_k}^k \geq \psi_{\mathcal{A}_k}$. Hence it follows that

$$\sum_{i=1}^n(y_i^{k+1}-y_i^k) \leq -\|y^k-\psi\|_{1,\mathcal{A}_k} + \|(A_{\mathcal{I}_k}^{-1}A_{\mathcal{I}_k\mathcal{A}_k})_+\|_1\,\|y^k-\psi\|_{1,\mathcal{A}_k} < 0\,, \tag{3.7}$$

unless $y^{k+1} = y^k$. Consequently

$$y^k \to \mathcal{M}(y^k) = \sum_{i=1}^n y_i^k$$

acts as a merit function for the algorithm. Since there are only finitely many possible choices for active/inactive sets there exists an iteration index $\bar{k}$ such that $\mathcal{I}_{\bar{k}} = \mathcal{I}_{\bar{k}+1}$. Moreover, $(y^{\bar{k}+1}, \lambda^{\bar{k}+1})$ is solution to (3.1). In fact, in view of (iii) of the algorithm it suffices to show that $y^{\bar{k}+1}$ and $\lambda^{\bar{k}+1}$ are feasible. This follows from the fact that due to

$\mathcal{I}_{\bar{k}} = \mathcal{I}_{\bar{k}+1}$ we have $c(y_i^{\bar{k}+1} - \psi_i) = \lambda_i^{\bar{k}+1} + c(y_i^{\bar{k}+1} - \psi_i) \leq 0$ for $i \in \mathcal{I}_{\bar{k}}$ and $\lambda_i^{\bar{k}+1} + c(y_i^{\bar{k}+1} - \psi_i) > 0$ for $i \in \mathcal{A}_{\bar{k}}$. Thus the algorithm converges in finitely many steps. □

Remark 3.3. Let us note as a corollary to the proof of Theorem 3.3 that in case A is a M-matrix then $\mathcal{M}(y^k) = \sum_{i=1}^n y_i^k$ is always a merit function. In fact, in this case the conditions of Theorem 3.3 are obviously satisfied.◇

A perturbation result: We now discuss the primal-dual active set strategy for the case where the matrix A can be expressed as an additive perturbation of an M-matrix.

Theorem 3.4. *Assume that $A = M + K$ with M an* M*-matrix and with K an $n \times n$-matrix. Then, if $\|K\|_1$ is sufficiently small, (3.1) admits a unique solution $x^* = (y^*, \lambda^*)$, the primal-dual active set algorithm is well-defined and $\lim_{k\to\infty} x^k = x^*$.*

Proof. Recall that as a consequence of the assumption that M is a M-matrix all principal submatrices of M are nonsingular M-matrices as well [BP]. Let $\mathcal{S}$ denote the set of all subsets of $\{1, \ldots, n\}$, and define

$$\rho = \sup_{\mathcal{I}\in\mathcal{S}} \|M_{\mathcal{I}}^{-1} K_{\mathcal{I}}\|_1 .$$

Let K be chosen such that $\rho < \frac{1}{2}$. For every subset $\mathcal{I} \in \mathcal{S}$ the inverse of $A_{\mathcal{I}}$ exists and can be expressed as

$$A_{\mathcal{I}}^{-1} = (I_{\mathcal{I}} + \sum_{i=1}^{\infty} \left(-M_{\mathcal{I}}^{-1} K_{\mathcal{I}}\right)^i) M_{\mathcal{I}}^{-1} .$$

As a consequence the algorithm is well-defined. Proceeding as in the proof of Theorem 3.3 we arrive at

$$\begin{aligned} \sum_{i=1}^{n} (y_i^{k+1} - y_i^k) = &- \sum_{i\in\mathcal{A}} (y_i^k - \psi_i) + \sum_{i\in\mathcal{I}} \left(A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} (y^k - \psi)_{\mathcal{A}}\right)_i \\ &+ \sum_{i\in\mathcal{I}} (A_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k)_i , \end{aligned} \tag{3.8}$$

where $\lambda_i^k \leq 0$ for $i \in \mathcal{I}$ and $y_i^k \geq \psi_i$ for $i \in \mathcal{A}$. Here and below we drop the index k with $\mathcal{I}_k$ and $\mathcal{A}_k$. Setting $g = -A_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k \in \mathbb{R}^{|\mathcal{I}|}$ and since $\rho < \frac{1}{2}$ we find

$$\begin{aligned} \sum_{i\in\mathcal{I}} g_i &\geq \|M_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k\|_1 - \sum_{i=1}^{\infty} \|M_{\mathcal{I}}^{-1} K_{\mathcal{I}}\|_1^i \|M_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k\|_1 \\ &\geq \frac{1 - 2\rho}{1 - \rho} \|M^{-1} \lambda_{\mathcal{I}}^k\|_1 \geq 0 , \end{aligned}$$

and consequently by (3.8)

$$\sum_{i=1}^{n}(y_i^{k+1}-y_i^k) \le -\sum_{i\in\mathcal{A}}(y_i^k-\psi_i)+\sum_{i\in\mathcal{I}}(A_{\mathcal{I}}^{-1}A_{\mathcal{I}\mathcal{A}}(y^k-\psi)_{\mathcal{A}})_i\,.$$

Note that $A_{\mathcal{I}}^{-1}A_{\mathcal{I}\mathcal{A}} \le M_{\mathcal{I}}^{-1}K_{\mathcal{I}\mathcal{A}} - M_{\mathcal{I}}^{-1}K_{\mathcal{I}}(M+K)_{\mathcal{I}}^{-1}A_{\mathcal{I}\mathcal{A}}$. Here we have used $(M+K)_{\mathcal{I}}^{-1}-M_{\mathcal{I}}^{-1} = -M_{\mathcal{I}}^{-1}K_{\mathcal{I}}(M+K)_{\mathcal{I}}^{-1}$ and $M_{\mathcal{I}}^{-1}M_{\mathcal{I}\mathcal{A}} \le 0$. Since $y^k \ge \psi$ on $\mathcal{A}$, it follows that $\|K\|_1$ can be chosen sufficiently small such that $\sum_{i=1}^n(y_i^{k+1}-y_i^k) < 0$ unless $y^{k+1}=y^k$, and hence

$$y^k \mapsto \mathcal{M}(y^k) = \sum_{i=1}^{n} y_i^k$$

is a merit function for the algorithm. The proof is now completed in the same manner as that of Theorem 3.3 □

The assumptions of Theorem 3.4 do not require A to be a P-matrix. From its conclusions existence of a solution to (3.1) for arbitrary f follows. This is equivalent to the fact that A is a P-matrix [BP, Theorem 10.2.15]. Hence, it follows that Theorem 3.4 represents a sufficient condition for A to be a P-matrix.

Observe further that the M-matrix property is not stable under arbitrarily small perturbations since off-diagonal elements may become positive. This implies certain limitations of the applicability of Theorem 3.2. Theorem 3.4 guarantees that convergence of the primal-dual active set strategy for arbitrary initial data is preserved for sufficiently small perturbations K of an M-matrix. Therefore, Theorem 3.4 is also of interest in connection with numerical implementations of the primal-dual active set algorithm.

Remark 3.4. The primal-dual active set strategy can be interpreted as a prediction strategy which, on the basis of (y^k,λ^k) predicts the *true* active and inactive sets, i.e.

$$\mathcal{A}^* = \{i : \lambda_i^* + c(y_i^*-\psi_i) > 0\} \quad \text{and} \quad \mathcal{I}^* = \{1,\dots,n\}\setminus\mathcal{A}^*\,.$$

To further pursue this point we define the following partitioning of the index set at iteration level k:

$$\mathcal{I}_G = \mathcal{I}_k\cap\mathcal{I}^*, \quad \mathcal{I}_B = \mathcal{I}_k\cap\mathcal{A}^*, \quad \mathcal{A}_G = \mathcal{A}_k\cap\mathcal{A}^*, \quad \mathcal{A}_B = \mathcal{A}_k\cap\mathcal{I}^*\,.$$

The sets $\mathcal{I}_G$, $\mathcal{A}_G$ give *good*, the sets $\mathcal{I}_B$ and $\mathcal{A}_B$ a *bad* prediction. Let us denote by $G_F(x^k)$ the system matrix of (2.4) and let $\Delta y = y^{k+1}-y^*$, $\Delta\lambda = \lambda^{k+1}-\lambda^*$. If the primal-dual active set method is interpreted as a semi-smooth Newton method then the convergence analysis is based

on the identity

(3.9)
$$G_F(x^k)\begin{pmatrix}\Delta y_{\mathcal{I}_k}\\ \Delta y_{\mathcal{A}_k}\\ \Delta\lambda_{\mathcal{I}_k}\\ \Delta\lambda_{\mathcal{A}_k}\end{pmatrix} = -\big(F(x^k)-F(x^*)-G_F(x^k)(x^k-x^*)\big) =: \Psi(x^k)\,.$$

Without loss of generality we can assume that the components of the equation $\lambda-\max\{0,\lambda+c(y-\psi)\}=0$ are ordered as $(\mathcal{I}_G,\mathcal{I}_B,\mathcal{A}_G,\mathcal{A}_B)$. Then the right hand side of (3.9) has the form

$$\Psi(x^k) = -\mathrm{col}\left(0_{\mathcal{I}_k}, 0_{\mathcal{A}_k}, 0_{\mathcal{I}_G}, \lambda^*_{\mathcal{I}_B}, 0_{\mathcal{A}_G}, c(\psi-y^*)_{\mathcal{A}_B}\right)\,, \tag{3.10}$$

where $0_{\mathcal{I}_k}$ denotes a vector of zeros of length $|\mathcal{I}_k|$, $\lambda^*_{\mathcal{I}_B}$ denotes a vector of λ^* coordinates with index set $\mathcal{I}_B$, and analogously for the remaining terms. Since $y^k\geq\psi$ on $\mathcal{A}_k$ and $\lambda^k\leq 0$ on $\mathcal{I}_k$ we have

$$\|\psi-y^*\|_{\mathcal{A}_B}\leq\|y^k-y^*\|_{\mathcal{A}_B}\quad\text{and}\quad\|\lambda^*\|_{\mathcal{I}_B}\leq\|\lambda^k-\lambda^*\|_{\mathcal{I}_B}\,. \tag{3.11}$$

Exploiting the structure of $G_F(x^k)$ and (3.10) we find

(3.12)
$$\Delta y_{\mathcal{A}_G}=0,\quad \Delta y_{\mathcal{A}_B}=(\psi-y^*)_{\mathcal{A}_B},\quad \Delta\lambda_{\mathcal{I}_G}=0,\quad \Delta\lambda_{\mathcal{I}_B}=-\lambda^*_{\mathcal{I}_B}\,.$$

On the basis of (3.9)-(3.12) we can draw the following conclusions:

(i) If $x^k\to x^*$ then there exists an index $\bar{k}$ such that $\mathcal{I}_B=\mathcal{A}_B=\emptyset$ for all $k\geq\bar{k}$. Consequently $\Psi(x^{\bar{k}})=0$ and, as we noted before, if $x^k\to x^*$ then convergence occurs in finitely many steps.

(ii) By (3.9)–(3.11) there exists a constant $\kappa\geq 1$ independent of k such that

$$\|\Delta y\|+\|\Delta\lambda\|\leq\kappa\left(\|(y^k-y^*)_{\mathcal{A}_B}\|+\|(\lambda^k-\lambda^*)_{\mathcal{I}_B}\|\right)\,.$$

Thus if the incorrectly predicted sets are small in the sense that

$$\|(y^k-y^*)_{\mathcal{A}_B}\|+\|(\lambda^k-\lambda^*)_{\mathcal{I}_B}\|\leq\tfrac{1}{2\kappa-1}\Big(\|(y^k-y^*)_{\mathcal{A}_{B,c}}\|+\|(\lambda^k-\lambda^*)_{\mathcal{I}_{B,c}}\|\Big)\,,$$

where $\mathcal{A}_{B,c}$ ($\mathcal{I}_{B,c}$) denotes the complement of the indices $\mathcal{A}_B$ ($\mathcal{I}_B$), then

$$\|y^{k+1}-y^*\|+\|\lambda^{k+1}-\lambda^*\|\leq\tfrac{1}{2}\left(\|y^k-y^*\|+\|\lambda^k-\lambda^*\|\right)\,,$$

and convergence follows.

(iii) If $y^*<\psi$ and $\lambda^0+c(y^0-\psi)\leq 0$ (e.g. $y^0=\psi$, $\lambda^0=0$), then the algorithm converges in one step. In fact, in this case $\mathcal{A}_B=\mathcal{I}_B=\emptyset$ and $\Psi(x^0)=0$.$\diamond$

Finally, we shall point out that Theorems 3.2–3.4 establish global convergence of the primal-dual active set strategy or, equivalently, semi-smooth Newton method without the necessity of a line search. The rate of convergence is locally superlinear. Moreover, it can be observed from (2.4) that if $\mathcal{I}_k = \mathcal{I}_{k'}$ for $k \neq k'$, then $y^k = y^{k'}$ and $\lambda^k = \lambda^{k'}$. Hence, in case of convergence no cycling of the algorithm is possible, and termination at the solution of (2.1) occurs after finitely many steps.

4. The infinite dimensional case

In this section we first analyze the notion of slant differentiability of the max-operation between various function spaces. Then we turn to the investigation of convergence of semi-smooth Newton methods applied to (P). We close the section with a numerical example for superlinear convergence.

Let X denote a space of functions defined over a bounded domain or manifold $\Omega \subset \mathbb{R}^n$ with Lipschitzian boundary $\partial\Omega$, and let $\max(0, y)$ stand for the point-wise maximum operation between 0 and $y \in X$. Let $\delta \in \mathbb{R}$ be fixed arbitrarily. We introduce candidates for slanting functions G_m of the form

$$G_m(y)(x) = \begin{cases} 1 & \text{if} \quad y(x) > 0\,, \\ 0 & \text{if} \quad y(x) < 0\,, \\ \delta & \text{if} \quad y(x) = 0\,, \end{cases} \tag{4.1}$$

where $y \in X$.

Proposition 4.1.

(i) *G_m can in general not serve as a slanting function for* $\max(0, \cdot)\colon L^p(\Omega) \to L^p(\Omega)$, *for* $1 \leq p \leq \infty$.
(ii) *The mapping* $\max(0, \cdot)\colon L^q(\Omega) \to L^p(\Omega)$ *with* $1 \leq p < q \leq \infty$ *is slantly differentiable on* $L^q(\Omega)$ *and G_m is a slanting function.*

The proof is deferred to Appendix A.

We refer to [U] for a related investigation of the *two-norm problem* involved in Proposition 4.1 in the case of superposition operators. An example in [U] proves the necessity of the norm-gap for the case in which the complementarity condition is expressed by means of the Fischer-Burmeister functional.

We now turn to (P) posed in $L^2(\Omega)$. For convenience we repeat the problem formulation

$$\text{(P)} \qquad \begin{cases} \min J(y) = \dfrac{1}{2}(y, Ay) - (f, y) \\ \text{subject to } y \le \psi, \end{cases}$$

where $(\cdot,\cdot)$ now denotes the inner product in $L^2(\Omega)$, f and $\psi \in L^2(\Omega)$, $A \in \mathcal{L}(L^2(\Omega))$ is selfadjoint and

$$\text{(H1)} \qquad (Ay, y) \ge \gamma \|y\|^2 ,$$

for some $\gamma > 0$ independent of $y \in L^2(\Omega)$. There exists a unique solution y^* to (P) and a Lagrange multiplier $\lambda^* \in L^2(\Omega)$, such that (y^*, λ^*) is the unique solution to

$$\text{(4.2)} \qquad \begin{cases} Ay^* + \lambda^* = f, \\ \mathcal{C}(y^*, \lambda^*) = 0, \end{cases}$$

where $\mathcal{C}(y, \lambda) = \lambda - \max(0, \lambda + c(y - \psi))$, with the max–operation defined point-wise a.e. and $c > 0$ fixed. The primal-dual active set strategy is analogous to the finite dimensional case. We repeat it for convenient reference:

Primal-dual active set algorithm in $L^2(\Omega)$.

(i) Choose y^0, λ^0 in $L^2(\Omega)$. Set $k = 0$.
(ii) Set $\mathcal{A}_k = \{x \colon \lambda^k(x) + c(y^k(x) - \psi(x)) > 0\}$ and $\mathcal{I}_k = \Omega \backslash \mathcal{A}_k$.
(iii) Solve

$$\begin{aligned} &Ay^{k+1} + \lambda^{k+1} = f \\ &y^{k+1} = \psi \ \text{ on } \ \mathcal{A}_k, \lambda^{k+1} = 0 \ \text{ on } \ \mathcal{I}_k. \end{aligned}$$

(iv) Stop, or set $k = k + 1$ and return to (ii).

Under our assumptions on A, f and ψ it is simple to argue the solvability of the system in step (iii) of the above algorithm.

For the semi-smooth Newton step as well we can refer back to Section 2. At iteration level k with $(y^k, \lambda^k) \in L^2(\Omega) \times L^2(\Omega)$ given, it is of the form (2.4) where now $\delta y_{\mathcal{I}_k}$ denotes the restriction of δy (defined on Ω) to $\mathcal{I}_k$ and analogously for the remaining terms. Moreover $A_{\mathcal{I}_k\mathcal{A}_k} = E^*_{\mathcal{I}_k} A\, E_{\mathcal{A}_k}$, where $E_{\mathcal{A}_k}$ denotes the extension-by-zero operator for $L^2(\mathcal{A}_k)$ to $L^2(\Omega)$–functions, and its adjoint $E^*_{\mathcal{A}_k}$ is the restriction of $L^2(\Omega)$–functions to $L^2(\mathcal{A}_k)$, and similarly for $E_{\mathcal{I}_k}$ and $E^*_{\mathcal{I}_k}$. Moreover $A_{\mathcal{A}_k\mathcal{I}_k} = E^*_{\mathcal{A}_k} A\, E_{\mathcal{I}_k}$, $A_{\mathcal{I}_k} = E^*_{\mathcal{I}_k} A\, E_{\mathcal{I}_k}$ and $A_{\mathcal{A}_k} = E^*_{\mathcal{A}_k} A\, E_{\mathcal{A}_k}$. It can be argued precisely as in Section 2 that the primal-dual active set strategy and the semi-smooth Newton updates coincide, provided that the

slanting function of the max-function is taken according to

$$G_m(u)(x) = \begin{cases} 1 & \text{if} \quad u(x) > 0 \\ 0 & \text{if} \quad u(x) \leq 0, \end{cases} \tag{4.3}$$

which we henceforth assume.

Proposition 4.1 together with Theorem 1.1 suggest that the semi-smooth Newton algorithm applied to (4.2) may not converge in general. We therefore restrict our attention to operators A of the form

(H2) $A = C + \beta I$, with $C \in \mathcal{L}(L^2(\Omega), L^q(\Omega))$, where $\beta > 0$, $q > 2$.

We show next that a large class of optimal control problems with control constraints can be expressed in the form (P) with (H2) satisfied.

Example 1. We consider the optimal control problem

$$\begin{cases} \text{minimize} & \frac{1}{2}\|y - z\|^2_{L^2} + \frac{\beta}{2}\|u\|^2_{L^2} \\ \text{subject to} & -\Delta y = u \text{ in } \Omega,\ y = 0 \text{ on } \partial\Omega\,, \\ & u \leq \psi,\ u \in L^2(\Omega)\,, \end{cases} \tag{4.4}$$

where $z \in L^2(\Omega)$, $\psi \in L^q(\Omega)$, and $\beta > 0$. Let $B \in \mathcal{L}(H^1_o(\Omega), H^{-1}(\Omega))$ denote the operator $-\Delta$ with homogeneous Dirichlet boundary conditions. Then (4.4) can equivalently be expressed as

$$\begin{cases} \text{minimize} & \frac{1}{2}\|B^{-1}u - z\|^2_{L^2} + \frac{\beta}{2}\|u\|^2_{L^2} \\ \text{subject to} & u \leq \psi,\ u \in L^2(\Omega)\,. \end{cases} \tag{4.5}$$

In this case $A \in \mathcal{L}(L^2(\Omega))$ turns out to be $Au = B^{-1}\mathcal{J}B^{-1}u + \beta u$, where $\mathcal{J}$ is the embedding of $H^1_o(\Omega)$ into $H^{-1}(\Omega)$, and $f = B^{-1}z$. Condition (H2) is obviously satisfied.

In (4.4) we considered the distributed control case. A related boundary control problem is given by

$$\begin{cases} \text{minimize} & \frac{1}{2}\|y - z\|^2_{L^2(\Omega)} + \frac{\beta}{2}\|u\|^2_{L^2(\partial\Omega)} \\ \text{subject to} & -\Delta y + y = 0 \text{ in } \Omega,\ \frac{\partial y}{\partial n} = u \text{ on } \partial\Omega\,, \\ & u \leq \psi,\ u \in L^2(\partial\Omega)\,, \end{cases} \tag{4.6}$$

where n denotes the unit outer normal to Ω along $\partial\Omega$. This problem is again a special case of (P) with $A \in \mathcal{L}(L^2(\partial\Omega))$ given by $Au = B^{-*}\mathcal{J}B^{-1}u + \beta u$ where $B^{-1} \in \mathcal{L}(H^{-1/2}(\Omega), H^1(\Omega))$ denotes the solution operator to

$$-\Delta y + y = 0 \text{ in } \Omega,\ \frac{\partial y}{\partial n} = u \text{ on } \partial\Omega\,,$$

and $f = B^{-*}z$. Moreover, $C = B^{-*}\mathcal{J}B^{-1}_{|L^2(\Omega)} \in \mathcal{L}(L^2(\partial\Omega), H^{1/2}(\partial\Omega))$ with $\mathcal{J}$ the embedding of $H^{1/2}(\Omega)$ into $H^{-1/2}(\partial\Omega)$ and hence (H2) is satisfied as a consequence of the Sobolev embedding theorem.

For the sake of illustration it is also worthwhile to specify (2.5)–(2.8), which were found to be equivalent to the Newton-update (2.4) for the case of optimal control problems. We restrict ourselves to the case of the distributed control problem (4.4). Then (2.5)–(2.8) can be expressed as

(4.7)
$$\begin{cases} \lambda^{k+1}_{\mathcal{I}_k} = 0, \quad u^{k+1}_{\mathcal{A}_k} = \psi_{\mathcal{A}_k}, \\ E^*_{\mathcal{I}_k}\left[(B^{-2}+\beta I)E_{\mathcal{I}_k}u^{k+1}_{\mathcal{I}_k} - B^{-1}z + (B^{-2}+\beta I)E_{\mathcal{A}_k}\psi_{\mathcal{A}_k}\right] = 0\,, \\ E^*_{\mathcal{A}_k}\left[\lambda^{k+1} + B^{-2}u^{k+1} + \beta u^{k+1} - B^{-1}z\right] = 0\,, \end{cases}$$

where we set $B^{-2} = B^{-1}\mathcal{J}B^{-1}$. Setting $p^{k+1} = B^{-1}z - B^{-2}u^{k+1}$, a short computation shows that (4.7) is equivalent to

(4.8)
$$\begin{cases} -\Delta y^{k+1} = u^{k+1} \quad \text{in } \Omega\,, \quad y^{k+1} = 0 \quad \text{on } \partial\Omega\,, \\ -\Delta p^{k+1} = z - y^{k+1} \quad \text{in } \Omega\,, \quad p^{k+1} = 0 \quad \text{on } \partial\Omega\,, \\ \quad p^{k+1} = \beta u^{k+1} + \lambda^{k+1} \quad \text{in } \Omega\,, \\ \quad u^{k+1} = \psi \quad \text{in } \mathcal{A}_k\,, \lambda^{k+1} = 0 \quad \text{in } \mathcal{I}_k\,. \end{cases}$$

This is the system in the primal variables (y, u) and adjoint variables (p, λ), previously implemented in [BHHK, BIK] for testing the algorithm.$\diamond$

At this point we remark that the primal-dual active set strategy has no straight-forward infinite dimensional analogue for state constrained optimal control problems and obstacle problems [H]. For state constrained optimal control problems the Lagrange multiplier is only a measure in general and hence the core steps (ii) and (iii) of our algorithm are no longer meaningful. For details on the regularity issue we refer to [C]. Theorem 3.2 proves global convergence of the primal-dual active set strategy or, equivalently, semi-smooth Newton method for discretized obstacle problems. However, no comparable result can be expected in infinite dimensions. The main reason comes from the fact that the systems that would have to be solved in step (iii) are the first order conditions related to the problems

$$\min \quad \tfrac{1}{2}(Ay, y)_{L^2(\Omega)} - (f, y)_{L^2(\Omega)} \quad \text{s.t.} \quad y = \psi \quad \text{a.e. on } \mathcal{A}_k\,.$$

Again the multiplier associated to the equality constraint is only a measure in general.

Our main intention is to consider control constrained problems as in Example 1. To prove convergence under assumptions (H1), (H2) we utilize a reduced algorithm which we explain next.

The operators $E_{\mathcal{I}}$ and $E_{\mathcal{A}}$ denote the extension by zero and their adjoints are restrictions to $\mathcal{I}$ and $\mathcal{A}$, respectively. The optimality system (4.2) does not depend on the choice of $c > 0$. Moreover, from the discussion in Section 2 the primal-dual active set strategy is independent of $c > 0$ after the initialization phase. For the specific choice $c = \beta$ system (4.2) can equivalently be expressed as

$$\beta y^* - \beta\psi + \max(0, Cy^* - f + \beta\psi) = 0\,, \tag{4.9}$$

$$\lambda^* = f - Cy^* - \beta y^*\,. \tag{4.10}$$

We shall argue in the proof of Theorem 4.1 below that the primal-dual active set method in $L^2(\Omega)$ for (y, λ) is equivalent to the following algorithm for the reduced system (4.9)– (4.10), which will be shown to converge superlinearly.

Reduced algorithm

(i) *Choose* $y^0 \in L^2(\Omega)$ *and set* $k = 0$.
(ii) *Set* $\mathcal{A}_k = \{x : (f - Cy_k - \beta\psi)(x) > 0\}$, $\mathcal{I}_k = \Omega \setminus \mathcal{A}_k$.
(iii) *Solve*

$$\beta y_{\mathcal{I}_k} + (C(E_{\mathcal{I}_k} y_{\mathcal{I}_k} + E_{\mathcal{A}_k}\psi_{\mathcal{A}_k}))_{\mathcal{I}_k} = f_{\mathcal{I}_k}$$

and set $y^{k+1} = E_{\mathcal{I}_k} y_{\mathcal{I}_k} + E_{\mathcal{A}_k}\psi_{\mathcal{A}_k}$.
(iv) *Stop, or set* $k = k + 1$ *and return to (ii).*

Theorem 4.1. *Assume that (H1), (H2) hold and that* ψ *and* f *are in* $L^q(\Omega)$. *Then the primal-dual active set strategy or equivalently the semi-smooth Newton method converge superlinearly if* $\|y^0 - y^*\|$ *is sufficiently small and* $\lambda^0 = \beta(y^0 - \psi)$.

The proof is given in Appendix A. It consists essentially of two steps. In the first equivalence between the reduced algorithm and the original one is established and in the second one slant differentiability of the mapping $\hat{F} : L^2(\Omega) \to L^2(\Omega)$ given by $\hat{F}(y) = \max(0, Cy - f + \beta\psi)$ is shown. With respect to the latter we can alternatively utilize the theory of semi-smoothness of composite mappings as developed in [U]. For this purpose we first recall the notion of semi-smoothness as introduced in [U]. Suppose, we are given the superposition operator

$$\tilde{\Psi} : Y \to L^r(\Omega), \quad \tilde{\Psi}(y)(x) = \tilde{\psi}(H(y)(x)),$$

where $\tilde{\psi} : \mathbb{R}^m \to \mathbb{R}$ and $H : Y \to \prod_{i=1}^m L^{r_i}(\Omega)$, with $1 \le r \le r_i < \infty$, and Y is a Banach space. Then, $\tilde{\Psi}$ is called semi-smooth at $y \in Y$ if

$$\sup_{G \in \partial_s \tilde{\Psi}(y+h)} \|\tilde{\Psi}(y+h) - \tilde{\Psi}(y) - Gh\|_{L^r} = o(\|h\|_Y) \quad \text{as } h \to 0 \text{ in } Y. \tag{4.11}$$

Here $\partial_s \tilde{\Psi}$ denotes the generalized differential

(4.12)
$$\partial_s \tilde{\Psi}(y) = \left\{ G \in \mathcal{L}(Y, L^r) \,\middle|\, \begin{array}{l} G : v \mapsto \sum_i d_i(y)(H_i'(y)v), \text{ where } d(y) \\ \text{is a measurable selection of } \partial\tilde{\psi}(H(y)) \end{array} \right\},$$

where $\partial\tilde{\psi}$ is Clarke's generalized Jacobian [C], and prime denotes the Fréchet derivative. In our context $Y = L^r(\Omega) = L^2(\Omega)$, $m = 2$, $r_i = 2$, $H(y) = (0, Cy - f + \beta\psi)$, and $\tilde{\psi}(a, b) = \max(a, b)$. Clearly H is affine with respect to the second component. By (H2), and since $\psi \in L^q(\Omega)$, $f \in L^q(\Omega)$, it follows that H is Lipschitz from $L^2(\Omega)$ to $(L^q(\Omega))^2$, with $q > 2$. Moreover, $\tilde{\psi}$ is semi-smooth in the sense of [QS]. Consequently, $\tilde{\Psi}$ is semi-smooth in the sense of (4.11) by [U, Thm. 5.2].

In general, a slanting function G according to Definition 1 need not satisfy $G(y) \in \partial_s \tilde{\Psi}(y)$. However, the particular slanting function

$$\hat{G}(y)v = G_m(Cy - f + \beta\psi)Cv$$

with

$$G_m(u)(x) = \begin{cases} 1 & \text{if } u(x) \geq 0, \\ 0 & \text{if } u(x) < 0 \end{cases}$$

satisfies $\hat{G}(y) \in \partial_s \tilde{\Psi}(y)$. In fact, $d(y) = (d_1(y), d_2(y)) = (0, G_m(Cy - f + \beta\psi))$ is a measurable selection of $\partial \max(0, Cy - f + \beta\psi)$. Thus, (4.12) yields

$$\partial_s \tilde{\Psi}(y) \ni G(y)v = \sum_i d_i(y)(H_i'(y)v) = G_m(Cy - f + \beta\psi)Cv = \hat{G}(y)v.$$

Consequently, from the proof of Theorem 6.4 in [U] we infer that the reduced algorithm converges locally superlinearly.

Let us point out that the semi-smooth Newton method in [U] requires a smoothing step while our primal-dual active set strategy does not. To explain the difference of the two approaches, we note that with respect to (P) the following NCP-problem is considered in [U]: Find $y \in Y$ such that

$$y - \psi \leq 0, \quad Z(y) := Ay - f \geq 0, \quad (y - \psi)Z(y) = 0. \tag{4.13}$$

Then (4.13) is reformulated by utilizing an NCP-function. In our context, this yields

$$\max(y - \psi, f - Ay) = 0. \tag{4.14}$$

Following [U] one chooses $Y = L^p(\Omega)$, $p > 2$, and considers $y \mapsto \max(y - \psi, f - Ay)$ from $L^p(\Omega)$ to $L^2(\Omega)$ in order to introduce the norm gap which is required for semi-smoothness according to (4.11). In Algorithm 6.3 of [U] the Newton step first produces an update in $L^2(\Omega)$, which requires smoothing to obtain the new iterate in $L^p(\Omega)$

which is utilized in (4.14). In our formulation (4.13) is reformulated as (4.9) rather than (4.14). Here we can take advantage of the fact that (4.9) allows to directly exploit the smoothing property of the operator C. Consequently, we obtain a superlinearly convergent Newton method without the necessity of a smoothing step.

If an appropriate growth condition is satisfied then the superlinear convergence result of Theorem 4.1 can be improved to superlinear convergence with a specific rate. Let us suppose that there exists $\alpha > 0$ such that

$$\text{(A')} \qquad \lim_{h\to 0} \frac{1}{\|h\|^{1+\alpha}} \|F(x^* + h) - F(x^*) - G(x^* + h)h\| = 0\,.$$

Then an inspection of the proof of Theorem 1.1 shows that the rate of convergence of x^k to x^* is of q-order $1 + \alpha$, i.e. we have $\|x^{k+1} - x^*\| = \mathcal{O}(\|x^k - x^*\|^{1+\alpha})$ as $k \to \infty$. To investigate (A') for the specific F appearing in the proof of Theorem 4.1 one can apply the general theory in [U]. We prefer to give an independent proof adapted to our problem formulation. Let the assumptions of Theorem 4.1 hold and recall that $F : L^2(\Omega) \to L^2(\Omega)$ is given by $F(y) = \beta y - \beta\psi + \max(0, Cy - f + \beta\psi)$. First we consider the case $2 < q < +\infty$. The relevant difference quotient for the nonlinear term which must be analyzed for (A') to hold is given by

$$\begin{aligned}
&\frac{1}{\|h\|_{L^2}^{1+\alpha}} \| \max(0, C(y^* + h) - f + \beta\psi) - \max(0, Cy^* - f + \beta\psi) \\
&\qquad - G_m(Cy^* + Ch - f + \beta\psi)(Ch)\|_{L^2} \\
=&\frac{1}{\|Ch\|_{L^q}^{1+\alpha}} \| \max(0, w + Ch) - \max(0, w) - G_m(w + Ch)(Ch)\|_{L^2} \frac{\|Ch\|_{L^q}^{1+\alpha}}{\|h\|_{L^2}^{1+\alpha}}\,,
\end{aligned}$$

where we set $w = Cy^* - f + \beta\psi$. Utilizing the fact that $C \in \mathcal{L}(L^2(\Omega), L^q(\Omega))$ it suffices to consider

$$\frac{1}{\|h\|_{L^q}^{1+\alpha}} \|D_{w,h}\|_{L^2} = \frac{1}{\|h\|_{L^q}^{1+\alpha}} \| \max(0, w+h) - \max(0, w) - G_m(w+h)h\|_{L^2}\,.$$

Here and below we use the notation introduced in the proof of Proposition 4.1(ii). Proceeding as in the proof of Proposition 4.1(ii) we find

for $\frac{1}{\sigma}+\frac{1}{\tau}=1$, $\sigma\in(1,\infty)$,

$$\begin{aligned}\frac{1}{\|h\|_{L^q}^{1+\alpha}}\|D_{w,h}\|_{L^2} &\le \frac{1+|\delta|}{\|h\|_{L^q}^{1+\alpha}}\Big[|\Omega_\epsilon(h)|^{1/2\tau}\Big(\int_{\Omega_\epsilon(h)}|w(x)|^{2\sigma}dx\Big)^{1/2\sigma}\\ &\quad + |\Omega_\epsilon(w)|^{1/2\tau}\Big(\int_{\Omega_0(h)\setminus\Omega_\epsilon(h)}|w(x)|^{2\sigma}dx\Big)^{1/2\sigma}\Big]\\ &= \frac{1+|\delta|}{\|h\|_{L^q}^{1+\alpha}}(|\Omega_\epsilon(h)|^{1/2\tau}+|\Omega_\epsilon(w)|^{1/2\tau})\Big(\int_{\Omega_0(h)}|w(x)|^{2\sigma}dx\Big)^{1/2\sigma}.\end{aligned}\tag{4.15}$$

Let us set $r=\frac{q}{1+\alpha}$. We have

$$\begin{aligned}&\Big(\int_{\Omega_0(h)}|w(x)|^{2\sigma}dx\Big)^{1/2\sigma}\le\Big(\int_{\Omega_0(h)}|w(x)|^{\frac{2\sigma q}{r}}|w(x)|^{\frac{2\sigma(r-q)}{r}}dx\Big)^{1/2\sigma}\\ &\le\Big(\int_{\Omega_0(h)}|w(x)|^{\frac{2\sigma q}{r}\frac{r}{2\sigma}}\Big)^{1/r}\Big(\int_{\Omega_0(h)}|w(x)|^{\frac{2\sigma(r-q)}{r}\frac{r}{r-2\sigma}}\Big)^{(r-2\sigma)/2r\sigma}\\ &\Big(\int_{\Omega_0(h)}|w(x)|^{q}dx\Big)^{1/r}\Big(\int_{\Omega_0(h)}\frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}}dx\Big)^{(r-2\sigma)/2r\sigma},\end{aligned}$$

where it is assumed that $r=\frac{q}{1+\alpha}>2\sigma>2$. Since $|w(x)|\le|h(x)|$ for $x\in\Omega_0(h)$ we find

$$\begin{aligned}&\frac{1}{\|h\|_{L^q}^{1+\alpha}}\|D_{w,h}\|_{L^2}\le\\ &\quad\le(1+|\delta|)(|\Omega_\epsilon(h)|^{1/2\tau}+|\Omega_\epsilon(w)|^{1/2\tau})\Big(\int_{\Omega_0(h)}\frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}}dx\Big)^{(r-2\sigma)/2r\sigma}.\end{aligned}$$

Suppose that

$$\int_{\{x:|w(x)|\neq0\}}\frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}}dx<+\infty\,.\tag{4.16}$$

Then, following the argument in the proof of Proposition 4.1(ii) we have

$$\lim_{\|h\|_{L^q}\to0}\frac{1}{\|h\|_{L^q}^{1+\alpha}}\|D_{w,h}\|_{L^2}=0\,,$$

and hence (A') holds. Let us interpret the conditions on α and q. As already pointed out we must have $q>2(1+\alpha)$ which for $\alpha=0$ is consistent with the requirement that there must be a norm gap. The exponent in (4.16) can equivalently be expressed as $Q(\alpha,q)=\frac{2\sigma\alpha q}{q-2\sigma(1+\alpha)}$. Hence for fixed q, the quotient $Q(\alpha,q)$ is increasing with α and (4.16) is more likely to be satisfied for small rather than for large α. Similarly, for fixed α, $Q(\alpha,q)$ is decreasing with respect to q ($>2\sigma(1+\alpha)$) and

hence (4.16) has a higher chance to be satisfied for large rather than small q.

Convergence of q-order larger than 2 is possible, if $q > 2$ and (4.16) holds for the associated values of q and α. If w is Lipschitzian then it must be of at least linear growth across the boundary of the set $\{x : w(x) \neq 0\}$. For this reason it is of interest to consider the range of α-values satisfying $\frac{2\alpha q}{q-2(1+\alpha)} < 1$. This necessitates $\alpha < \frac{1}{2}$.

In the case $q = +\infty$ we have for every $\sigma > 1$

$$\Big(\int_{\Omega_0(h)} |w(x)|^{2\sigma} dx\Big)^{1/2\sigma} = \Big(\int_{\Omega_0(h)} |w(x)|^{2\sigma(1+\alpha)} |w(x)|^{-2\sigma\alpha} dx\Big)^{1/2\sigma}$$
$$\leq \|h\|_{L^\infty}^{1+\alpha} \Big(\int_{\Omega_0(h)} |w(x)|^{-2\sigma\alpha} dx\Big)^{1/2\sigma}.$$

This estimate and (4.15) for $q = +\infty$ yield

$$\frac{1}{\|h\|_{L^\infty}^{1+\alpha}} \|D_{w,h}\|_{L^2} \leq (1 + |\delta|)(|\Omega_\epsilon(h)|^{1/2\tau} + |\Omega_\epsilon(w)|^{1/2\tau}) \cdot \Big(\int_{\Omega_0(h)} \frac{1}{|w(x)|^{2\sigma\alpha}} dx\Big)^{1/2\sigma}.$$

Now suppose that for some $\sigma > 1$,

$$\int_{\{x:|w(x)|\neq 0\}} \frac{1}{|w(x)|^{2\sigma\alpha}} dx < +\infty\,. \tag{4.17}$$

Then, again following the arguments in the proof of Proposition 4.1, we obtain

$$\lim_{\|h\|_{L^\infty} \to 0} \frac{1}{\|h\|_{L^\infty}^{1+\alpha}} \|D_{w,h}\|_{L^2} = 0\,,$$

which shows that (A') is satisfied.

Example 1 (continued). As already observed Theorem 4.1 is directly applicable to problems (4.4) and (4.6) and confirms local superlinear convergence of the semi-smooth Newton algorithm.

Convergence for (4.4) was already analyzed in [BIK] where it was proved that a modified augmented Lagrangian acts as a merit function provided that

$$\beta + \gamma \leq c \leq \beta - \frac{\beta^2}{\gamma} + \frac{\beta^2}{\|\Delta^{-1}\|^2} \tag{4.18}$$

for some $\gamma > 0$. Here $\|\Delta^{-1}\|$ denotes the operator norm of Δ^{-1} in $\mathcal{L}(L^2(\Omega))$. This previous convergence result is unconditional with respect of the initial condition but it restricts the range of β. Theorem 4.1

is a local result with respect to initialization, but does not restrict the range of $\beta > 0$. Further, the discussion following Theorem 4.1 provides rate of convergence results.

Let us also comment on the discretized version of (4.4). To be specific we consider a two dimensional domain Ω endowed with a uniform rectangular grid, with Δ_h denoting the five-point-star discretization of Δ, and functions z, ψ, y, u discretized by means of grid functions at the nodal points. Numerical results for this case were reported in [BIK] and [BHHK] and convergence can be argued provided the discretized form of (4.18) holds. Let us consider to which extent Theorems 3.2–3.4 provide new insight on confirming convergence, which was observed numerically in practically all examples. Theorem 3.2 is not applicable since $A_h = \beta I + \Delta_h^{-2}$ is not an M-Matrix. Theorem 3.4 is applicable with $M = \beta I$ and $K = \Delta_h^{-2}$, and asserts convergence if β is sufficiently large. We also tested numerically the applicability of Theorem 3.3 and found that for $\Omega = (0,1)^2$ the norm condition was satisfied in all cases we tested with grid-size $h \in [10^{-2}, 10^{-1}]$ and $\beta \geq 10^{-4}$, whereas the cone condition $\sum_{i\in\mathcal{I}}(A_{\mathcal{I}}^{-1} y_{\mathcal{I}})_i \geq 0$ for $y_{\mathcal{I}} \geq 0$ was satisfied only for $\beta \geq 10^{-2}$, for the same range of grid-sizes. Still the function $y^k \to \mathcal{M}(y^k)$ utilized in the proof of Theorem 3.4 behaved as a merit function for the wider range of $\beta \geq 10^{-3}$. Note that the norm and cone condition of Theorem 3.4 only involve the system matrix A, wheres $\mathcal{M}(y^k)$ also depends on the specific choice of f and ψ.$\diamond$

Remark 4.1. Throughout the paper we used the function $\mathcal{C}$ defined in (2.2) as a complementarity function. Another popular choice of complementarity function is given by the Fischer-Burmeister function

$$\mathcal{C}_{FB}(y, \lambda) = \sqrt{y^2 + \lambda^2} - (y + \lambda)\,.$$

Note that $\mathcal{C}_{FB}(0, \lambda) = \sqrt{\lambda^2} - \lambda = 2\max(0, -\lambda)$, and hence by Proposition 4.1 the natural choices for slanting functions do not satisfy property (A).$\diamond$

Remark 4.2. Condition (H2) can be considered as yet another incidence, where a *two norm concept* for the analysis of optimal control problems is essential. It utilizes the fact that the control-to-solution mapping of the differential equation is a smoothing operation. Two norm concepts where used for second order sufficient optimality conditions and the analysis of SQP-methods in [M, I, IK3], for example, and also for semi-smooth Newton methods in [U].$\diamond$

In view of the fact that (P) consist of a quadratic cost functional with affine constraints the question arises whether superlinear convergence coincides with one step convergence after the active/inactive sets are identified by the algorithm. The following example illustrates the fact that this is not the case.

Example 2. We consider Example 1 with the specific choices

$$z(x_1, x_2) = \sin(5x_1) + \cos(4x_2), \quad \psi \equiv 0, \quad \beta = 10^{-5}, \text{ and } \Omega = (0,1)^2.$$

A finite difference based discretization of (4.4) with a uniform grid of mesh size $h = \frac{1}{100}$ and the standard five point star discretization of the Laplace operator was used. The primal-dual active set strategy with initialization given by solving the unconstrained problem and setting $\lambda_h^0 = 0$, was used. The exact discretized solution $(u_h^*, \lambda_h^*, y_h^*)$ was attained in 8 iterations. In Table 1 we present the values for

$$q_u^k = \frac{|u_h^k - u_h^*|}{|u_h^{k-1} - u_h^*|}, \quad q_\lambda^k = \frac{|\lambda_h^k - \lambda_h^*|}{|\lambda_h^{k-1} - \lambda_h^*|},$$

where the norms are discrete L^2-norms. Clearly these quantities indicate superlinear convergence of u_h^k and λ_h^k.

k	1	2	3	4	5	6	7
q_u^k	1.0288	0.8354	0.6837	0.4772	0.2451	0.0795	0.0043
q_λ^k	0.6130	0.5997	0.4611	0.3015	0.1363	0.0399	0.0026

TABLE 1.

We also tested whether the quantities appearing in the rate of convergence discussion are reflected in the numerical results. For this purpose note that for the problem under consideration w appearing in (4.16) and (4.17) is given by $w = \Delta^{-2}u^* + \Delta^{-1}z + \beta\psi$. Roughly, (4.16) and (4.17) have a higher chance to be satisfied with larger value for α, if w is not smooth across the boundary of the set $\{x : w(x) = 0\}$. In a numerical test we kept all problem data identical to those specified above except for changing ψ to $\psi(x_1, x_2) = x_1x_2 - 1$. Note that this new ψ increases the chance that (4.16) and (4.17) are satisfied. Moreover, increasing β (for the same ψ) results in an increase of the influence of ψ to w. Thus we expect an improved convergence as β is increased. For the new ψ and small β the algorithm finds the solution in one less iterations. Increasing β results in a further reduction of 3 iterations, see Tables 1 and 2.

	q_u^k					
k	1	2	3	4	5	6
$\beta = 10^{-5}$	1.0443	0.8359	0.6780	0.4679	0.2342	0.0614
$\beta = 10^{-3}$	0.1410	0.0455	0.0041	–	–	–

TABLE 2.

Acknowledgment. We appreciate many helpful comments by the referees.

REFERENCES

[BHHK] M. Bergounioux, M. Haddou, M. Hintermüller and K. Kunisch, A Comparison of a Moreau-Yosida Based Active Set Strategy and Interior Point Methods for Constrained Optimal Control Problems, SIAM J. on Optimization **11** (2000), pp. 495–521.

[BIK] M. Bergounioux, K. Ito and K. Kunisch, Primal-dual Strategy for Constrained Optimal Control Problems, SIAM J. Control and Optimization, **37** (1999), pp. 1176–1194.

[BP] A. Berman, R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Computer Science and Scientific Computing Series, Academic Press, New York, 1979.

[C] E. Casas, Control of an elliptic problem with pointwise state constraints, SIAM J. Control and Optimization, **24** (1986), pp. 1309–1318.

[CNQ] X. Chen, Z. Nashed and L. Qi, Smoothing Methods and semi-smooth methods for nondifferentiable operator equations, SIAM J. on Numerical Analysis, **38** (2000), pp. 1200–1216.

[C] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, Wiley, New York, 1983.

[FFKP] F. Facchinei, A. Fischer, C. Kanzow, and J.-M. Peng, A simply constrained optimization reformulation of KKT systems arising from variational inequalities, Applied Mathematics and Optimization, **40** (1999), pp.19–37.

[FK] A. Fischer, C. Kanzow, On finite termination of an iterative method for linear complementarity, Mathematical Programming, **74** (1996), pp. 279–292.

[HKT] M. Heinkenschloss, C.T. Kelley, and H. Tran, Fast algorithms for nonsmooth compact fixed point problems, SIAM J. Num. Analysis **29** (1992), pp. 1769–1792.

[H] R.H.W. Hoppe, Multigrid algorithms for variational inequalities, SIAM J. Numerical Analysis, **24** (1997), pp. 1046–1065.

[I] A. Ioffe, Necessary and sufficient conditions for a local minimum 3, SIAM J. Control and Optimization **17** (1979), pp. 266–288.

[IK1] K. Ito and K. Kunisch, Augmented Lagrangian Methods for Nonsmooth Convex Optimization in Hilbert Spaces, Nonlinear Analysis, Theory, Methods and Applications **41** (2000), pp. 573–589.

[IK2] K. Ito and K. Kunisch, Optimal Control of Elliptic Variational Inequalities, Applied Mathematics and Optimization **41** (2000), pp. 343–364.

[IK3] K. Ito and K. Kunisch, The SQP-Algorithm for a Class of Weakly Singular Optimal Control Problems, SIAM J. on Optimization **10** (2000), pp. 896–916.

[KS] C.T. Kelley and E. Sachs, Multilevel algorithms for constrained compact fixed point problems, SIAM J. Sci. Comput., **15** (1994), pp. 645–667.

[K1] B. Kummer, Newton's method for nondifferentiable functions, in: J. Guddat et. al., eds., *Mathematical Research, Advances in Optimization*, Akademie-Verlag Berlin, 1988, pp. 114–125.

[K2] B. Kummer, Generalized Newton and NCP methods: convergence, regularity, actions, Discuss.Math.Differ.Incl.Control Optim., **20** (2000), pp. 209-244.

[KNT] Y.A. Kuznetsov, P. Neittaanmäki, P. Tarvainen, Overlapping block methods for obstacle problems with convection-diffusion operators, in: *Complementarity and Variational Problems. State of the Art*, M.C. Ferris, J.-S. Pang, eds., SIAM, Philadelphia, 1997.

[M] H. Maurer, First and second order sufficient optimality conditions in mathematical programming and optimal control, Math. Prog. Study **14** (1981), pp. 43–62.

[Q1] L. Qi, Convergence analysis of some algorithms for solving nonsmooth equations, Math. of Operations Research **18** (1993), pp. 227–244.

[Q2] H. Qi, A regularized smoothing Newton method for box constrained variational inequality problems with P_0–functions, SIAM J. on Optimization **10** (1999), pp. 315–330.

[QS] L. Qi and J. Sun, A nonsmooth version of Newton's method, Mathematical Programming **58** (1993), pp. 353–367.

[UU] M. Ulbrich and S. Ulbrich, Superlinear convergence of affine-scaling interior-point Newton methods for infinite-dimensional nonlinear problems with pointwise bounds, SIAM J. on Control and Optimization, **38** (2000), pp. 1938–1984.

[U] M. Ulbrich, Semi-smooth Newton methods for operator equations in function spaces, report TR00-11, Department. of Comput. and Appl. Math. - MS134, Rice University, Houston, Texas, 2000, to appear in SIAM J. on Optimization.

Appendix A.

Proof of Theorem 3.2. The assumption that A is a M-matrix implies that for every index partition $\mathcal{I}$ and $\mathcal{A}$ we have $A_{\mathcal{I}}^{-1} \geq 0$ and $A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} \leq 0$, see [BP, p. 134]. Let us first show the monotonicity property of the y-component. Observe that for every $k \geq 1$ the complementarity property

$$\lambda_i^k = 0 \quad \text{or} \quad y_i^k = \psi_i \,, \quad \text{for all } i, \text{ and } k \geq 1 \,, \tag{A.1}$$

holds. For $i \in \mathcal{A}_k$ we have $\lambda_i^k + c(y_i^k - \psi_i) > 0$ and hence by (A.1) either $\lambda_i^k = 0$, which implies $y_i^k > \psi_i$, or $\lambda_i^k > 0$, which implies $y_i^k = \psi_i$.

Consequently $y^k \geq \psi = y^{k+1}$ on $\mathcal{A}_k$ and $\delta y_{\mathcal{A}_k} = \psi_{\mathcal{A}_k} - y^k_{\mathcal{A}_k} \leq 0$. For $i \in \mathcal{I}_k$ we have $\lambda^k_i + c(y^k_i - \psi_i) \leq 0$ which implies $\delta\lambda_{\mathcal{I}_k} \geq 0$ by (2.4) and (A.1). Since $\delta y_{\mathcal{I}_k} = -A^{-1}_{\mathcal{I}_k} A_{\mathcal{I}_k\mathcal{A}_k} \delta y_{\mathcal{A}_k} - A^{-1}_{\mathcal{I}_k}\delta\lambda_{\mathcal{I}_k}$ by (3.3) it follows that $\delta y_{\mathcal{I}_k} \leq 0$. Therefore $y^{k+1} \leq y^k$ for every $k \geq 1$.

Next we show that y^k is feasible for all $k \geq 2$. Due to the monotonicity of y^k it suffices to show that $y^2 \leq \psi$. Let $V = \{i : y^1_i > \psi_i\}$. For $i \in V$ we have $\lambda^1_i = 0$ by (A.1), and hence $\lambda^1_i + c(y^1_i - \psi_i) > 0$ and $i \in \mathcal{A}_1$. Since $y^2 = \psi$ on $\mathcal{A}_1$ and $y^2 \leq y^1$ it follows that $y^2 \leq \psi$.

To verify that $y^* \leq y^k$ for all $k \geq 1$ note that

$$\begin{aligned} f_{\mathcal{I}_{k-1}} &= \lambda^*_{\mathcal{I}_{k-1}} + A_{\mathcal{I}_{k-1}} y^*_{\mathcal{I}_{k-1}} + A_{\mathcal{I}_{k-1}\mathcal{A}_{k-1}} y^*_{\mathcal{A}_{k-1}} \\ &= A_{\mathcal{I}_{k-1}} y^k_{\mathcal{I}_{k-1}} + A_{\mathcal{I}_{k-1}\mathcal{A}_{k-1}} \psi_{\mathcal{A}_{k-1}} \,. \end{aligned}$$

It follows that

$$A_{\mathcal{I}_{k-1}} \left(y^k_{\mathcal{I}_{k-1}} - y^*_{\mathcal{I}_{k-1}} \right) = \lambda^*_{\mathcal{I}_{k-1}} + A_{\mathcal{I}_{k-1}\mathcal{A}_{k-1}} \left(y^*_{\mathcal{A}_{k-1}} - \psi_{\mathcal{A}_{k-1}} \right) .$$

Since $\lambda^*_{\mathcal{I}_{k-1}} \geq 0$ and $y^*_{\mathcal{A}_{k-1}} \leq \psi_{\mathcal{A}_{k-1}}$ the M-matrix properties of A imply that $y^k_{\mathcal{I}_{k-1}} \geq y^*_{\mathcal{I}_{k-1}}$ for all $k \geq 1$.

Turning to the feasibility of λ^k assume that for a pair of indices $(\bar{k}, i)$, $\bar{k} \geq 1$, we have $\lambda^{\bar{k}}_i < 0$. Then necessarily $i \in \mathcal{A}_{\bar{k}-1}$, $y^{\bar{k}}_i = \psi_i$, and $\lambda^{\bar{k}}_i + c(y^{\bar{k}}_i - \psi_i) < 0$. It follows that $i \in \mathcal{I}_{\bar{k}}$, $\lambda^{\bar{k}+1}_i = 0$, and $\lambda^{\bar{k}+1}_i + c(y^{\bar{k}+1}_i - \psi_i) \leq 0$, since $y^{k+1}_i \leq \psi_i$, $k \geq 1$. Consequently $i \in \mathcal{I}_{\bar{k}+1}$ and by induction $i \in \mathcal{I}_k$ for all $k \geq \bar{k}+1$. Thus, whenever a coordinate of λ^k becomes negative at iteration $\bar{k}$, it is zero from iteration $\bar{k}+1$ onwards, and the corresponding primal coordinate is feasible. Due to finite-dimensionality of $\mathbb{R}^n$ it follows that there exists k_o such that $\lambda^k \geq 0$ for all $k \geq k_o$.

Monotonicity of y^k and $y^* \leq y^k \leq \psi$ for $k \geq 2$ imply the existence of $\bar{y}$ such that $\lim y^k = \bar{y} \leq \psi$. Since $\lambda^k = Ay^k + f \geq 0$ for all $k \geq k_o$, there exists $\bar{\lambda}$ such that $\lim \lambda^k = \bar{\lambda} \geq 0$. Together with (A.1) it follows that $(\bar{y}, \bar{\lambda}) = (y^*, \lambda^*)$. □

Remark A.1. From the proof it follows that if $\lambda^{\bar{k}}_i < 0$ for some coordinate i at iteration $\bar{k}$, then $\lambda^k_i = 0$ and $y^k_i \leq \psi_i$ for all $k \geq \bar{k}+1$.

Proof of Proposition 4.1. (i) It suffices to consider the one dimensional case $\Omega = (-1, 1) \subset \mathbb{R}$. We show that property (A) does not hold at $y(x) = -|x|$. Let us define $h_n(x) = \frac{1}{n}$ on $(-\frac{1}{n}, \frac{1}{n})$ and $h_n(x) = 0$

otherwise. Then

$$\int_{-1}^{1} |\max(0, y+h_n)(x) - \max(0,y)(x) - (G_m(y+h_n)(h_n))(x)|^p \, dx$$
$$= \int_{\{x: y(x)+h_n(x)>0\}} |y(x)|^p dx = \int_{-\frac{1}{n}}^{\frac{1}{n}} |y(x)|^p dx = \frac{2}{p+1}\left(\frac{1}{n}\right)^{p+1},$$

and $\|h_n\|_{L^p} = \sqrt[p]{2/n^{p+1}}$. Consequently,

$$\lim_{n\to\infty} \tfrac{1}{\|h_n\|_{L^p}} \|\max(0, y+h_n) - \max(0,y) - G_m(y+h_n)h_n\|_{L^p} = \sqrt[p]{\tfrac{1}{p+1}} \neq 0\,,$$

and hence (A) is not satisfied at y for any $p \in [1,\infty)$.

To consider the case $p=\infty$ we choose $\Omega = (0,1)$ and show that (A) is not satisfied at $y(x) = x$. For this purpose define for $n = 2, \ldots$

$$h_n(x) = \begin{cases} -(1+\frac{1}{n})x & \text{on } (0, \frac{1}{n}]\,, \\ (1+\frac{1}{n})x - \frac{2}{n}(1+\frac{1}{n}) & \text{on } (\frac{1}{n}, \frac{2}{n}]\,, \\ 0 & \text{on } (\frac{2}{n}, 1]\,. \end{cases}$$

Observe that $E_n = \{x : y(x) + h_n(x) < 0\} \supset (0, \frac{1}{n}]$. Therefore

$$\lim_{n\to\infty} \tfrac{1}{\|h_n\|_{L^\infty([0,1])}} \|\max(0, y+h_n) - \max(0,y) - G_m(y+h_n)h_n\|_{L^\infty([0,1])}$$
$$= \lim_{n\to\infty} \tfrac{n^2}{n+1} \|y\|_{L^\infty(E_n)} \geq \lim_{n\to\infty} \tfrac{n}{n+1} = 1$$

and hence (A) cannot be satisfied.

(ii) Let $\delta \in \mathbb{R}$ be fixed arbitrarily and y, $h \in L^q(\Omega)$, and set

$$D_{y,h}(x) = \max(0, y(x)+h(x)) - \max(0, y(x)) - G_m(y+h)(x)h(x)\,.$$

A short computation shows that

$$|D_{y,h}(x)| \begin{cases} \leq |y(x)| & \text{if } (y(x)+h(x))y(x) < 0\,, \\ \leq (1+|\delta|)\,|y(x)| & \text{if } y(x)+h(x) = 0\,, \\ = 0 & \text{otherwise.} \end{cases} \tag{A.2}$$

For later use we note that from Hölder's inequality we obtain for $1 \leq p < q \leq \infty$

$$\|w\|_{L^p} \leq |\Omega|^r \|w\|_{L^q}\,, \quad \text{with } r = \begin{cases} \frac{q-p}{pq} & \text{if } q < \infty\,, \\ \frac{1}{p} & \text{if } q = \infty\,. \end{cases}$$

From (A.2) it follows that only

$$\Omega_o(h) \;=\; \{x \in \Omega : y(x) \neq 0,\, y(x)(y(x)+h(x)) \leq 0\}$$

requires further investigation. For $\epsilon > 0$ we define subsets of $\Omega_o(h)$ by

$$\Omega_\epsilon(h) = \{x \in \Omega : |y(x)| \geq \epsilon,\, y(x)(y(x)+h(x)) \leq 0\}\,.$$

Note that $|y(x)| \geq \epsilon$ a.e. on $\Omega_\epsilon(h)$ and therefore

$$\|h\|_{L^q(\Omega)} \geq \epsilon |\Omega_\epsilon(h)|^{1/q}, \quad \text{for } q < \infty\,.$$

It follows that

$$\lim_{\|h\|_{L^q(\Omega)} \to 0} |\Omega_\epsilon(h)| = 0 \quad \text{for every fixed } \epsilon > 0\,. \tag{A.3}$$

For $\epsilon > 0$ we further define sets

$$\Omega_\epsilon(y) = \{x \in \Omega : 0 < |y(x)| \leq \epsilon\} \subset \{x : y(x) \neq 0\}\,.$$

Note that $\Omega_\epsilon(y) \subset \Omega_{\epsilon'}(y)$ whenever $0 < \epsilon \leq \epsilon'$ and $\bigcap_{\epsilon>0} \Omega_\epsilon(y) = \emptyset$. As a consequence

$$\lim_{\epsilon \to 0^+} |\Omega_\epsilon(y)| = 0\,. \tag{A.4}$$

From (A.2) we find

$$\begin{aligned}
\frac{1}{\|h\|_{L^q}} \|D_{y,h}\|_{L^p} &\leq \frac{1+|\delta|}{\|h\|_{L^q}} \left(\int_{\Omega_o(h)} |y(x)|^p dx \right)^{1/p} \\
&\leq \frac{1+|\delta|}{\|h\|_{L^q}} \Big[\left(\int_{\Omega_\epsilon(h)} |y(x)|^p dx \right)^{1/p} + \left(\int_{\Omega_o(h)\setminus\Omega_\epsilon(h)} |y(x)|^p dx \right)^{1/p} \Big] \\
&\leq \frac{1+|\delta|}{\|h\|_{L^q}} \Big[|\Omega_\epsilon(h)|^{(q-p)/(qp)} \left(\int_{\Omega_\epsilon(h)} |y(x)|^q dx \right)^{1/q} + \\
&\qquad |\Omega_\epsilon(y)|^{(q-p)/(qp)} \left(\int_{\Omega_o(h)\setminus\Omega_\epsilon(h)} |y(x)|^q dx \right)^{1/q} \Big] \\
&\leq (1+|\delta|) \left(|\Omega_\epsilon(h)|^{(q-p)/(qp)} + |\Omega_\epsilon(y)^{(q-p)/(qp)}| \right)\,.
\end{aligned}$$

Choose $\eta > 0$ arbitrarily and note that by (A.4) there exists $\bar\epsilon > 0$ such that $(1+|\delta|)|\Omega_{\bar\epsilon}(y)|^{(q-p)/(qp)} < \eta$. Consequently

$$\frac{1}{\|h\|_{L^q}} \|D_{y,h}\|_{L^p} \leq (1+|\delta|)|\Omega_{\bar\epsilon}(h)|^{(q-p)/(qp)} + \eta$$

and by (A.3)

$$\lim_{\|h\|_{L^q} \to 0} \frac{1}{\|h\|_{L^q}} \|D_{y,h}\|_{L^p} \leq \eta\,.$$

Since $\eta > 0$ is arbitrary the claim holds for $1 \leq p < q < \infty$.

The case $q = \infty$ follows from the result for $1 \leq p < q < \infty$. □

Proof of Theorem 4.1. Let y^k, $k \geq 1$, denote the iterates of the reduced algorithm and define

$$\lambda^{k+1} = \begin{cases} 0 & \text{on } \mathcal{I}_k\,, \\ (f - Cy^{k+1} - \beta\psi)_{\mathcal{A}_k} & \text{on } \mathcal{A}_k\,, \end{cases} \qquad \text{for } k = 0, 1, \dots\,,$$

We obtain $\lambda^k + \beta(y^k - \psi) = f - Cy^k - \beta\psi$ for $k = 1, 2, \ldots$, and hence the active sets $\mathcal{A}_k$, the iterates y^{k+1} produced by the reduced algorithm and by the algorithm in the two variables (y^{k+1}, λ^{k+1}) coincide for $k = 1, 2, \ldots$, provided the initialization strategies coincide. This, however, is the case since due to our choice of λ^0 and $\beta = c$ we have $\lambda^0 + \beta(y^0 - \psi) = f - Cy^0 - \beta\psi$ and hence the active sets coincide for $k = 0$ as well.

To prove convergence of the reduced algorithm we utilize Theorem 1.1 with $F : L^2(\Omega) \to L^2(\Omega)$ given by $F(y) = \beta y - \beta\psi + \max(0, Cy - f + \beta\psi)$. From Proposition 4.1(ii) it follows that F is slantly differentiable. In fact, the relevant difference quotient for the nonlinear term in F is

$$\frac{1}{\|Ch\|_{L^q}} \big\| \max(0, Cy - f + \beta\psi + Ch) - \max(0, Cy - f + \beta\psi) - G_m(Cy - f + \beta\psi + Ch)(Ch) \big\|_{L^2} \frac{\|Ch\|_{L^q}}{\|h\|_{L^2}},$$

which converges to 0 for $\|h\|_{L^2} \to 0$. Here

$$G_m(Cy - f + \beta\psi + Ch)(x) = \begin{cases} 1 & \text{if } (C(y+h) - f + \beta\psi)(x) \geq 0, \\ 0 & \text{if } (C(y+h) - f + \beta\psi)(x) < 0, \end{cases}$$

so that in particular δ of (4.1) was set equal to 1 which corresponds to the '$\leq$' sign in the definition of $\mathcal{I}_k$. A slanting function G_F of F at y in direction h is therefore given by

$$G_F(y + h) = \beta I + G_m(Cy - f + \beta\psi + Ch)C.$$

It remains to argue that $G_F(z) \in \mathcal{L}(L^2(\Omega))$ has a bounded inverse. Since for arbitrary $z \in L^2(\Omega)$, $h \in L^2(\Omega)$

$$G_F(z)h = \begin{pmatrix} \beta I_{\mathcal{I}} + C_{\mathcal{I}} & C_{\mathcal{I}\mathcal{A}} \\ 0 & \beta I_{\mathcal{A}} \end{pmatrix} \begin{pmatrix} h_{\mathcal{I}} \\ h_{\mathcal{A}} \end{pmatrix},$$

where $\mathcal{I} = \{x : (Cz - f + \beta\psi)(x) \geq 0\}$ and $\mathcal{A} = \{x : (Cz - f + \beta\psi)(x) < 0\}$ it follows from (H1) that $G_F(z)^{-1} \in \mathcal{L}(L^2(\Omega))$. Above we denoted $C_{\mathcal{I}} = E_{\mathcal{I}}^* C E_{\mathcal{I}}$ and $C_{\mathcal{I}\mathcal{A}} = E_{\mathcal{I}}^* C E_{\mathcal{A}}$. □

(M. Hintermüller) Institute of Mathematics, University of Graz, A-8010 Graz, Austria

(K. Ito) North Carolina State University, Raleigh, NC 27695, USA

(K. Kunisch) Institute of Mathematics, University of Graz, A-8010 Graz, Austria

Semi–Smooth Newton Methods for Variational Inequalities of the First Kind.

Kazufumi Ito[1] and Karl Kunisch[2]

July 2002

[1]Center for Research in Scientific Computation, Department of Mathematics, North Carolina State University; supported in part by AFSOR under contract F-49620-95-1-0447.

[2]Institut für Mathematik, Universität Graz, Graz, Austria; supported in part by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03, Optimierung und Kontrolle".

Abstract

Semi–smooth Newton methods are analyzed for a class of variational inequalities in infinite dimensions. It is shown that they are equivalent to certain active set strategies. Global and local super-linear convergence are proved. To overcome the phenomenon of finite speed of propagation of discretized problems a penalty version is used as the basis for a continuation procedure to speed up convergence. The choice of the penalty parameter can be made on the basis of an L^∞ estimate for the penalized solutions. Unilateral as well as bilateral problems are considered.

1 Introduction

This paper is devoted to the convergence analysis of iterative algorithms for solving variational inequalities of the form

$$\begin{cases} \min \ \frac{1}{2}\, a(y,y) - (f,y) \\ \\ y \in H_0^1(\Omega) \\ \\ y \leq \psi \ \text{ a.e. in } \Omega, \end{cases} \tag{1.1}$$

where $a(\cdot,\cdot)$ is a coercive bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$, and $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$. The precise assumptions on the quantities appearing in (1.1) are given in Section 2. While iterative methods for solving finite dimensional discretization of (1.1) are extensively studied see e.g. [D, H, HK] and the references therein, little attention has been paid to the infinite–dimensional counter–parts. Our contribution will focus on the convergence of the infinite dimensional algorithms. More precisely we shall analyze primal–dual active set algorithms or - as we shall argue - equivalently semi–smooth Newton algorithms. To briefly describe this class of algorithms let y^* denote the solution to (1.1) and let λ^* be the associated Lagrange multiplier. As we shall recall in Section 2, the optimality system associated to (1.1) can be expressed as

$$\begin{cases} a(y^*,v) + (\lambda^*,v) = (f,v), \ \ \text{for all} \ \ v \in H_0^1(\Omega), \\ \\ \lambda^* = \max(0, \lambda^* + c(y^* - \psi)), \end{cases} \tag{1.2}$$

for each $c > 0$, where max denotes the pointwise a.e. maximum operation. The second order augmented Lagrangian method in [B, IK1] employs the primal–dual active set strategy based on the second equality in (1.2) and is given as the following iterative method: given a current pair (y_k, λ_k) of primal and dual variables, predict the active set $\mathcal{A}_{k+1}$ as

$$\mathcal{A}_{k+1} = \{x \colon \lambda_k(x) + c(y_k(x) - \psi(x)) > 0\}. \tag{1.3}$$

We arrive at the following formal algorithm:

Algorithm

(i) Choose $c > 0$, (y_o, λ_0), set $k = 0$.

(ii) Determine $\mathcal{A}_{k+1}$ according to (1.2).

(iii) Solve for $y_{k+1} = \arg\min\{\frac{1}{2}a(y,y) - (f,y) \colon y = \psi \text{ on } \mathcal{A}_{k+1}\}$.

(iv) Let λ_{k+1} be the Lagrange multiplier associated to the constraint in (iii) with $\lambda_{k+1} = 0$ on $\Omega \setminus \mathcal{A}_{k+1}$.

(v) Set $k = k + 1$ and goto (ii).

Let us observe that the optimality system for the variational problem in (iii) is given by

$$\text{(1.4)} \qquad \begin{cases} a(y_{k+1}, v) + \langle \lambda_{k+1}, v \rangle_{H^{-1}, H_0^1} = (f, v) \ \text{ for all } \ v \in H_0^1(\Omega), \\ \\ y_{k+1} = \psi \ \text{ on } \ \mathcal{A}_{k+1}, \ \lambda_{k+1} = 0 \ \text{ on } \ \mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}. \end{cases}$$

In particular, the Lagrange multiplier associated to the constraint $y = \psi$ on $\mathcal{A}_{k+1}$, is in general only a distribution in $H^{-1}(\Omega)$. This results from the fact that $\frac{\partial y_{k+1}}{\partial n}$ is not continuous across the boundaries between active and inactive sets. Rather λ_{k+1} contains jumps of magnitude $\frac{\partial y_{k+1}}{\partial n_J^+} - \frac{\partial y_{k+1}}{\partial n_J^-}$, where $n_J^{\pm}$ stands for the normal directions to either side of the boundary between active and inactive set. These jumps are not present in the solution of the limit–problem (1.1), since under mild assumptions [KS],[T] we have $y^* \in H^2(\Omega)$ and $\lambda^* \in L^2(\Omega)$. The fact that the Lagrange multipliers λ_{k+1} of the auxiliary problems in (iii) of the Algorithm are not contained in the pivot space $L^2(\Omega)$ between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$ presents a serious difficulty, both from the point of view of numerical implementation and convergence analysis. In order to remedy this difficulty we consider a one-parameter family of regularized problems based on smoothing of the complementarity condition by

$$\lambda = \alpha \ \max(0, \lambda + c(y - \psi)), \quad 0 < \alpha < 1$$

which replaces the second equation in (1.2). The motivation for this regularization is that it is a relaxation of the second equation in (1.2). We analyze (i) the convergence of the active set strategy to the regularized problem, (ii) the monotone convergence property and L^∞ rate of convergence of solutions

to the regularized problem to the original variational inequality and then (iii) develop and test a continuation method for the second order augmented Lagrangian method based on (i) and (ii).

The outline of the paper is as follows. In Section 2 we first introduce an equivalent but much more convenient form of the regularized problems and subsequently an iteration method based on the primal–dual active set strategy. We show that the method based on the active set strategy is equivalent to a semi-smooth Newton method [HIK]. Global as well as local super–linear convergence of the iteration method for the regularized problems is proven. The equivalence to the semi-smooth Newton is used to prove local super–linear convergence. Section 3 is devoted to the asymptotic analysis with respect to the regularization parameter. Monotone convergence properties of the solutions of the regularized problems towards the solution of the original problem are proven and an L^∞ error estimate for this convergence is obtained. It is important to note that the L^∞-error estimate can be used as a guideline for the choice of the penalty in terms of the mesh-size. In Section 5 we present our numerical examples to demonstrate the structural results obtained in this paper. Moreover we demonstrate that the algorithm allows to determine the boundary of the active set within grid–size accuracy. We also show that regularization can be used to overcome an essential drawback of active set strategies applied to (1.1), i.e., when the bilinear form a is discretized by finite differences (the five point stencil in dimension two) then changes from one iteration to the next occur along layers between active and inactive sets which have only the width of one mesh-size. For fine mesh-sizes this results in large iteration numbers. This difficulty can be overcome by multigrid methods, for example. Here we show that regularization techniques provide an alternative to deal with this shortcoming of active set strategies for (1.1). A regularized version of the above algorithm converges within a very few iteration due to its capability to change large sets of active indices to inactive ones and vice versa. We shall demonstrate that this property can advantageously be used in a continuation procedure with respect to the regularization parameter. The focus of our numerical test is not to compete with the most efficient implementations for this frequently tested class of obstacle problems, but rather to validate the structural results of the paper and to show the potential of a systematic use of regularization.

Our theoretical results provide a framework for an efficient second order iterative proccess for solving a regularized form of (1.2). It should also be noted that solving the regularized problem is equivalent to solving a single

step of the first order augmented Lagrangian method, e.g., see [IK2] and thus semi-smooth Newton methods should also improve the original implementation of the first order augmented Lagrangian method reported in [IK2]. This can be the focus of future investigations.

Beyond the motivation of overcoming the difficulty due to lack of regularity of the Lagrange multiplier our interest in analyzing primal–dual active set strategies for (1.1) also stems from our desire to investigate these algorithms separately for classes of problems which differ with respect to the regularity properties of the Lagrange multipliers. The abstract results are contained in [IK1]. In [BIK] we considered optimal control problems with control constraints. In this case the Lagrange multipliers of the original problem as well as those arising in the auxiliary problems of the primal–dual active set algorithm are in $L^2(\Omega)$ or $L^2(\partial\Omega)$, depending on whether distributed or boundary control problems are considered. For such problems large sets of active and inactive indices are moved from one iteration to the next. In [HIK] we established the strong relationship of these methods with superlinearly convergent semi–smooth Newton methods. For variational inequalities of the form (1.1) the Lagrange multipliers of the limit problem are L^2 but those of the auxiliary problems are not. Finally, for state constrained optimal control problems as well as for control of variational inequalities the Lagrange multipliers of the limit–problems themselves are not L^2 smooth but are in general only measures. Numerical results for these classes of problems are contained in [BHHK, IK2]. Convergence results for the latter are only available in the case of discretized state constrained optimal control problems.

We briefly summarize those facts on semi–smooth Newton methods which are relevant for our analysis in Section 2. Let X and Z be Banach spaces and let $F\colon D \subset X \to Z$ be a nonlinear mapping with open domain D.

Definition 1.1 *The mapping $F\colon D \subset X \to Z$ is called generalized–differentiable on the open subset $U \subset D$ if there exists a family of generalized derivatives $G\colon U \to \mathcal{L}(X, Z)$ such that*

$$\text{(A)} \qquad \lim_{h\to 0} \frac{1}{\|h\|} \|F(x+h) - F(x) - G(x+h)h\| = 0,$$

for every $x \in U$.

We shall refer to mappings F which allow a generalized derivative on U in the sense of Definition 1.1 as Newton–differentiable.

Theorem 1.1 *Suppose that $x^* \in D$ is a solution to $F(x) = 0$ and that F is Newton–differentiable in an open neighborhood U containing x^* and that $\{\|G(x)^{-1}\| : x \in U\}$ is bounded. Then the Newton–iteration $x_{k+1} = x_k - G(x_k)^{-1}F(x_k)$ converges superlinearly to x^* provided that $\|x_0 - x^*\|$ is sufficiently small.*

Let us consider Newton–differentiability of the max–operator. For this purpose X denotes a function space of real–valued functions on $\Omega \subset \mathbb{R}^n$ and $\max(0, y)$ is the pointwise max–operation. For $\delta \in \mathbb{R}$ we introduce candidates for the generalized derivative of the form

$$G_{m,\delta}(y)(x) = \begin{cases} 1 & \text{if} \quad y(x) > 0 \\ 0 & \text{if} \quad y(x) < 0 \\ \delta & \text{if} \quad y(x) = 0, \end{cases}$$

where $y \in X$.

Proposition 1.1 *The mapping $\max(0, \cdot) : L^q(\Omega) \to L^p(\Omega)$ with $1 \leq p < q < \infty$ is Newton differentiable on $L^q(\Omega)$ and $G_{m,\delta}$ is a generalized derivative.*

For the proofs of Theorem 1.1 and Proposition 1.1 we refer to [HIK]. Related results can be found in [U]. The following chain rule will be utilized in Section 2. We utilize a third Banach space Y.

Proposition 1.2 *Let $F_2 : Y \to X$ be an affine mapping with $F_2 y = By + b$, $B \in \mathcal{L}(Y, X)$, $b \in X$, and assume that $F_1 : D \subset X \to Z$ is Newton–differentiable on the open subset $U \subset D$ with generalized derivative G. If $F_2^{-1}(U)$ is nonempty then $F = F_1 \circ F_2$ is Newton–differentiable on $F_2^{-1}(U)$ with generalized derivative given by $G(By + b)B \in \mathcal{L}(Y, Z)$, for $y \in F_2^{-1}(U)$.*

Proof. By assumption $F_2^{-1}(U)$ is nonempty and due to continuity of F_2 the set $F_2^{-1}(U)$ is open. Note that $G(By + b)B \in \mathcal{L}(Y, Z)$ for each $y \in F_2^{-1}(U)$ since $G(x) \in \mathcal{L}(X, Z)$ for each $x \in U$. For $y \in F_2^{-1}(U)$ we find

$$\begin{aligned} &\lim_{\substack{h \to 0 \\ h \in Y}} \frac{1}{\|h\|} \|(F_1 \circ F_2)(y + h) - (F_1 \circ F_2)(y) - G(F_2(y + h))Bh\| \\ &= \lim_{\substack{h \to 0 \\ h \in Y}} \frac{1}{\|Bh\|} \|F_1(By + b + Bh) - F_1(By + b) - G(By + b + Bh)Bh\| \frac{\|Bh\|}{\|h\|} \\ &= 0, \end{aligned}$$

and hence the claim follows.

2 Global and local convergence of the iterative method for the regularized problems

We consider

$$\text{(2.1)} \qquad \begin{cases} \min\ \frac{1}{2}\, a(y,y) - (f,y) \\ \\ y \in H_0^1(\Omega) \\ \\ y \leq \psi \ \text{ a.e. in } \Omega, \end{cases}$$

where $a\,(\cdot,\cdot)$ is a bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$ satisfying

$$\text{(2.2)} \qquad a(v,v) \geq \nu |v|^2_{H_0^1}, \quad a(w,z) \leq \mu |w|_{H^1} |z|_{H^1},$$

for some $\nu > 0$ and $\mu > 0$ independently of $v \in H_0^1(\Omega)$ and $w, z \in H^1(\Omega)$. Further it is assumed that $f \in L^2(\Omega)$, $\psi \in H^1(\Omega)$ with $\psi|_{\partial\Omega} \geq 0$. Throughout Ω is a bounded domain in R^n with Lipschitzian boundary $\partial\Omega$. Since $\psi \in H^1(\Omega)$ the trace $\psi|_{\partial\Omega}$ is well-defined. The assumption $\psi|_{\partial\Omega} \geq 0$ implies that the set of admissible functions y for (2.1) is nonempty. We shall also require that a satisfies the weak maximum principle, i.e. for all $v \in H_0^1(\Omega)$

$$\text{(2.3)} \qquad a(v, v^+) \leq 0 \ \text{ implies } \ v^+ = 0,$$

where $v^+ = \max(0, v)$. We set $K = \{v \in H_0^1(\Omega) \colon v \geq 0 \ \text{ a.e.}\}$. It will further be convenient to introduce the representation operator

$$A \colon H_0^1(\Omega) \to H^{-1}(\Omega)$$

associated to $a(\cdot,\cdot)$. Utilizing (2.2) it is well-known [KS] that (2.1) admits a unique solution $y^* \in H_0^1(\Omega)$ and an associated Lagrange multiplier $\lambda^* \in H^{-1}(\Omega)$. Under well-known additional regularity assumptions which we recall in Remark 2.3 below $\lambda^* \in L^2(\Omega)$ and the following optimality system characterizes y^*:

$$\text{(2.4)} \qquad \begin{cases} a(y^*, v) + (\lambda^*, v) = (f, v), \ \text{ for all } \ v \in H_0^1(\Omega) \\ \\ (\lambda^*, y^* - \psi) = 0, \ y^* \leq \psi, \ (\lambda^*, v) \geq 0 \ \text{ for all } \ v \in K. \end{cases}$$

By inspection (2.4) can equivalently be expressed as

$$\text{(2.5)} \qquad \begin{cases} a(y^*, v) + (\lambda^*, v) = (f, v) \ \text{ for all } \ v \in H_0^1(\Omega) \\ \\ \lambda^* = \max(0, \lambda^* + c(y^* - \psi)), \end{cases}$$

for arbitrary $c > 0$. (More precisely, (2.4) implies (2.5) for every c, and (2.5) for some $c > 0$ implies (2.4)). Next we turn to the regularization of the max–function in (2.5). We have motivated the necessity for regularization for the primal–dual active set method by the abstract Algorithm in Section 1. Concerning the semi–smooth Newton approach we have from Proposition 1.1 that the max operation is Newton differentiable from $L^p(\Omega)$ to $L^2(\Omega)$ if $p > 2$. If we were to consider both y and λ as independent variables in a semi–smooth Newton approach to (2.5), then we can expect to obtain the necessary smoothing for the y component due to the first equation in (2.5) but we lack the smoothing property with respect to λ.

In our first attempt to regularize the max–function in (2.5) we are tempted to use the well–known smoothing

$$\max_\sigma(x) = \begin{cases} 0 & \text{for} \quad x < -\frac{\sigma}{2} \\ \\ \frac{1}{2\sigma}(x + \frac{\sigma}{2})^2 & \text{for} \quad -\frac{\sigma}{2} \le x \le \frac{\sigma}{2} \\ \\ x & \text{for} \quad x > \frac{\sigma}{2}, \end{cases}$$

with $\sigma > 0$, see e.g. [B]. After a short computation we obtain an explicit expression for $\lambda = \lambda_\sigma(z)$ satisfying $\lambda = \max_\sigma(0, \lambda + c\,z)$ as

$$\lambda_\sigma(z) \begin{cases} = 0 & \text{for} \quad \lambda + c\,z < -\frac{\sigma}{2} \\ \\ = \frac{\sigma}{2} - c\,z - \sqrt{-2c\sigma z} & \text{for} \quad -\frac{\sigma}{2} \le \lambda + c\,z \le \frac{\sigma}{2} \\ \\ \in [\frac{\sigma}{2}, \infty) & \text{for} \quad \lambda + c\,z > \frac{\sigma}{2}. \end{cases}$$

Thus we obtain an equation $Ay + \lambda_\sigma(y - \psi) = f$ for $y \in H_0^1(\Omega)$, where λ_σ is a multi–valued function defined above. This smoothing has some nice properties but it is much less convenient than penalty–type smoothing that we turn to next.

As stated in introduction we shall use

$$\text{(2.6)} \qquad \lambda = \alpha\ \max(0, \lambda + c(y^* - \psi)), \quad 0 < \alpha < 1$$

to regularize the second equation in (2.5). This is equivalent to

$$\lambda = \max(0, \bar{\lambda} + \gamma(y - \psi)),\ \gamma \in (0, \infty), \tag{2.7}$$

where $\bar{\lambda} \in L^2(\Omega)$, if we set $\bar{\lambda} = 0$ and $\gamma = c\alpha/(1-\alpha)$. Note that $\gamma \to \infty^+$ as $\alpha \to 1^-$. This type of regularization will allow us to prove global monotone convergence of the primal–dual active set method. The introduction of $\bar{\lambda}$ in (2.7), which does not appear in the original regularization, was motivated by augmented Lagrangians, [IK1], [IK2]. We shall see in Section 3 that depending on its choice the feasibility of the approximations can be controlled. Note that if $\bar{\lambda} = 0$ on $\{x\colon y(x) \geq \psi(x)\}$, then (2.7) can be regarded as a penalty–type formulation of the complementarity condition

$$y - \psi \leq 0,\ \lambda \geq 0,\ (y - \psi, \lambda) = 0,$$

as $\gamma \to \infty$. In the remainder of this subsection $\gamma > 0$ is a fixed constant and we consider an active set strategy or alternatively a semi–smooth Newton method to solve

$$\begin{cases} a(y, v) + (\lambda, v) = (f, v) \ \text{ for all } \ v \in H_0^1(\Omega) \\ \\ \lambda = \ \max(0, \bar{\lambda} + \gamma(y - \psi)). \end{cases} \tag{2.8}$$

Monotone operator theory provides the existence of a unique solution $(y_\gamma, \lambda_\gamma) \in H_0^1(\Omega) \times L^2(\Omega)$. An independent existence proof will follow from the results of this section.

We turn to the description of the algorithm.

Primal–Dual Active Set (PDAS) Algorithm

(i) Choose y_0 , set $k = 0$.

(ii) Set $\mathcal{A}_{k+1} = \{x\colon (\bar{\lambda} + \gamma(y_k - \psi))(x) > 0\},\ \mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}$.

(iii) Solve for $y_{k+1} \in H_0^1(\Omega)$:

$$a(y, v) + (\bar{\lambda} + \gamma(y - \psi), \chi_{\mathcal{A}_{k+1}} v) = (f, v) \ \text{ for all } \ v \in H_0^1(\Omega). \tag{2.9}$$

(iv) Set

$$\lambda_{k+1} = \begin{cases} 0 & \text{on} \quad \mathcal{I}_{k+1} \\ \\ \bar{\lambda} + \gamma(y_{k+1} - \psi) & \text{on} \quad \mathcal{A}_{k+1}. \end{cases}$$

(v) Stop or $k = k + 1$, goto (ii).

Remark 2.1 Here we establish the relationship between the above algorithm and a semi–smooth Newton method applied to (2.8). Recall the definition $A\colon H_0^1(\Omega) \to H^{-1}(\Omega)$ and introduce the nonlinear mapping $F\colon H_0^1(\Omega) \times L^2(\Omega) \to H^{-1}(\Omega) \times L^2(\Omega)$, by

$$F(y, \lambda) = \begin{pmatrix} A\, y + \lambda - f \\ \\ \lambda - \max(0, \bar{\lambda} + \gamma(y - \psi)) \end{pmatrix}.$$

A generalized derivative G of F in the sense of Definition 1.1 and Proposition 1.1 with $\delta = 0$ is given by

$$G(y_k, \lambda_k)h = \begin{pmatrix} A\, h_1 + h_2 \\ \\ h_2 - \gamma \chi_{\mathcal{A}_{k+1}} h_1 \end{pmatrix}.$$

where $h = (h_1, h_2) \in H_0^1(\Omega) \times L^2(\Omega)$.
The resulting semi–smooth Newton–update is thus given by

$$\text{(2.10)} \qquad \begin{cases} A\, \delta y + \delta\lambda = -A y_k - \lambda_k + f \\ \\ \delta\lambda = -\lambda_k \ \text{ on } \ \mathcal{I}_{k+1}, \\ \\ \delta\lambda - \gamma \delta y = -\lambda_k + \bar{\lambda} + \gamma(y_k - \psi) \ \text{ on } \ \mathcal{A}_{k+1} \end{cases}$$

where $\delta y = y_{k+1} - y_k$ and $\delta\lambda = \lambda_{k+1} - \lambda_k$, and coincides with step (iii)–(iv) of the primal–dual active set algorithm.

Remark 2.2 The semi-smooth Newton can be applied to (2.6) without reformulation as (2.7). Based on (2.6) it coincides with the one we specified above, with $\bar{\lambda} = 0$, except for the initialization phase, where now y_0 and λ_0 must be prescribed. In case of (2.6) the active set in step (ii) of the algorithm would be set $\tilde{\mathcal{A}}_{k+1} = \{x\colon (\lambda_k + \gamma(y_k - \psi))(x) > 0\}$ and the update on the basis of (2.6) for λ_{k+1} coincides with the one of step (iv) in the algorithm. Note that sgn $(\lambda_k + \gamma(y_k - \psi))(x) =$ sgn $(y_k - \psi)(x)$ for all $x \in \Omega$, and $k \geq 1$, and hence $\tilde{\mathcal{A}}_{k+1} = \mathcal{A}_{k+1}$ for all $k \geq 2$. A similar remark applies in case $\bar{\lambda} \neq 0$.

Properties of the semi–smooth Newton algorithm or equivalently the PDAS are analyzed next.

Proposition 2.1 *If $\mathcal{A}_{k+1} = \mathcal{A}_k$ for $k \geq 1$, then (y_k, λ_k) is the solution to (2.9).*

Proof. Since for given $\mathcal{A}_{k+1}$ the solution to (2.9) is unique it follows from $\mathcal{A}_k = \mathcal{A}_{k+1}$ that $y_k = y_{k+1}$ and consequently $\lambda_{k+1} = \lambda_k$.

Proposition 2.2 *The sequence $\{y_k\}_{k=1}^{\infty}$ is monotonically decreasing, i.e. $y_{k+1} \leq y_k$, a.e. on Ω for all $k \geq 1$.*

Proof. Let $\delta y = y_{k+1} - y_k$ for $k \geq 1$ and observe that

$$a(\delta y, \delta y^+) + (\lambda_{k+1} - \lambda_k, \delta y^+) = 0. \tag{2.11}$$

We have

$$\lambda_{k+1}(x) - \lambda_k(x) \begin{cases} = 0 & \text{for} \quad x \in \mathcal{I}_{k+1} \cap \mathcal{I}_k, \\ = \gamma(y_{k+1}(x) - y_k(x)) & \text{for} \quad x \in \mathcal{A}_{k+1} \cap \mathcal{A}_k, \\ = -\bar{\lambda} - \gamma(y_k - \psi)(x) \geq 0 & \text{for} \quad x \in \mathcal{I}_{k+1} \cap \mathcal{A}_k, \\ > \gamma(y_{k+1} - y_k)(x) & \text{for} \quad x \in \mathcal{A}_{k+1} \cap \mathcal{I}_k. \end{cases}$$

It follows that $(\lambda_{k+1} - \lambda_k, \delta y^+) \geq 0$ and by (2.11)

$$a(\delta y, \delta y^+) \leq 0.$$

Consequently $\delta y^+ = 0$ by (2.3) and $y_{k+1} \leq y_k$ follows. □

Proposition 2.3 *For every $k \geq 1$ we have $\mathcal{I}_k \subset \mathcal{I}_{k+1}$.*

Proof. If not, then there exists a set $S \subset \Omega$ of positive measure and $x \in \mathcal{I}_k \cap \mathcal{A}_{k+1}$ for every $x \in S$. It follows that $(\bar{\lambda} + (y_{k-1} - \psi))(x) \leq 0$ and by Proposition 2.2 $(\bar{\lambda} + (y_k - \psi))(x) \leq 0$. On the other hand $x \in \mathcal{A}_{k+1}$, and hence $(\bar{\lambda} + (y_k - \psi)(x)) > 0$. This gives the desired contradiction. □

Proposition 2.4 *For every $k \geq 1$ we have $y_\gamma \leq y_k$.*

Proof. We consider the sign of $\lambda_k - \lambda_\gamma$. Let $\mathcal{A}_\gamma = \{x\colon (\bar{\lambda}+\gamma(y_\gamma-\psi))(x) > 0\}$, and $\mathcal{I}_\gamma = \Omega \setminus A_\gamma$. For $x \in \mathcal{I}_\gamma \cap \mathcal{I}_k$ we have $(\lambda_k - \lambda_\gamma)(x) = 0$, and for $x \in \mathcal{A}_\gamma \cap \mathcal{A}_k$ we find $(\lambda_k - \lambda_\gamma)(x) = \gamma(y_k - y_\gamma)(x)$. If $x \in \mathcal{I}_\gamma \cap \mathcal{A}_k$ then $(\lambda_k - \lambda_\gamma)(x) = (\bar{\lambda} + \gamma(y_k - \psi))(x) \leq \gamma(y_k - y_\gamma)$. Finally, if $x \in \mathcal{A}_\gamma \cap \mathcal{I}_k$ then $(\lambda_k - \lambda_\gamma)(x) = -(\bar{\lambda} + \gamma(y_\gamma - \psi))(x) \leq 0$. We find

$$a(y_\gamma - y_k, (y_\gamma - y_k)^+) = -(\lambda_\gamma - \lambda_k, (y_\gamma - y_k)^+) \leq 0.$$

By (2.2) it follows that $(y_\gamma - y_k)^+ = 0$ and hence $y_\gamma \leq y_k$. □

Proposition 2.5 *For every $k \geq 1$ we have $0 \leq \lambda_{k+1} \leq \lambda_k$.*

Proof. The claim follows from Propositions 2.2 and 2.3. □

Note that Propositions 2.2–2.5 hold for $k \geq 1$ and are in general not valid for the initialization step with $k = 0$.

Theorem 2.1 *For every $\gamma > 0$ we have $\lim_{k\to\infty}(y_k, \lambda_k) = (y_\gamma, \lambda_\gamma)$ in $H_0^1(\Omega)\times L^2(\Omega)$.*

Proof. The sequences $\{y_k\}_{k=1}^\infty$ and $\{\lambda_k\}_{k=1}^\infty$ are decreasing pointwise almost everywhere and are uniformly bounded by $L^2(\Omega)$ functions. By (2.2) and (2.9) moreover, $\{y_k\}_{k=1}^\infty$ is bounded in $H_0^1(\Omega)$. Hence there exist $\hat{y} \in H_0^1(\Omega)$ and $\hat{\lambda} \in L^2(\Omega)$ such that a subsequence of y_k converges weakly in $H_0^1(\Omega)$ to $\hat{y}$, and $\lim_{k\to\infty} y_k = \hat{y}$ a.e. and $\lim_{k\to\infty} \lambda_k = \hat{\lambda}$ a.e.. Since $\mathcal{I}_k \subset \mathcal{I}_{k+1}$ and $\lambda_k = 0$ on $\mathcal{I}_k$ it follows that $\hat{\lambda} = 0$ on $\mathcal{I} = \bigcup_{k=1}^\infty \mathcal{I}_k$ and $\hat{\lambda} = \bar{\lambda} + \gamma(\hat{y} - \psi)$ on $\mathcal{A} = \bigcap_{k=1}^\infty \mathcal{A}_k$. Moreover, if $x \in \mathcal{I}$ then $(\bar{\lambda} + (\hat{y} - \psi))(x) \leq 0$, and for $x \in \mathcal{A}$ we have $(\bar{\lambda} + \gamma(y_k - \psi))(x) > 0$ for all k and hence $(\bar{\lambda} + (\hat{y} - \psi))(x) \geq 0$. Consequently $\hat{\lambda} = \max(0, \bar{\lambda}+\gamma(\hat{y}-\psi))$. By Lebesgue's bounded convergence theorem $\lim_{k\to\infty} \lambda_k = \hat{\lambda}$ in $L^2(\Omega)$. Taking the limit in

$$a(y_k, v) + (\lambda_k, v) = (f, v),$$

we arrive at

$$a(\bar{y}, v) + (\hat{\lambda}, v) = (f, v) \quad \text{for all} \quad v \in H_0^1(\Omega)$$

$$\hat{\lambda} = \max(0, \bar{\lambda} + \gamma(\hat{y} - \psi)).$$

Since the solution to this system is unique we have $(\hat{y}, \hat{\lambda}) = (y_\gamma, \lambda_\gamma)$. Finally, setting $v = y_k$ in (2.6) and using (2.2) we find $|y_k|_{H_0^1} \to |y_\gamma|_{H_0^1}$. Together with weak convergence of y_k to y_γ in H_0^1 this implies $\lim_{k\to\infty} y_k = y_\gamma$ in $H_0^1(\Omega)$. □

Remark 2.3 Under additional regularity assumptions the above result can be strengthened. We shall repeatedly refer to these assumptions which we now summarize. The bilinear form has the form

$$a(v, w) = \int_\Omega [a_{ij}\partial_{x_i} v \partial_{x_j} w + dw]dx,$$

for $v, w \in H^1(\Omega)$, where we use the summation convention, the leading differential operator is uniformly elliptic and $a_{ij} \in C^1(\bar{\Omega})$, $d \in L^\infty(\Omega)$, $d \geq 0$. Moreover $\psi \in H^2(\Omega)$, $\partial\Omega$ is $C^{1,1}$ or Ω is a polyhedron.

Under these requirements the representation operator A is a homeomorphism from $H^2(\Omega) \cap H_0^1(\Omega)$ to $L^2(\Omega)$. The solution to (2.1) satisfies $(y^*, \lambda^*) \in (H^2(\Omega) \cap H_0^1(\Omega)) \times L^2(\Omega)$, see e.g. [KS], [T], [IK2], or as corollary to the results of section 3. Moreover $\lim_{k\to\infty} y_k = y_\gamma$ in $H_0^1(\Omega) \cap H^2(\Omega)$ in the statement of Theorem 2.1. □

Theorem 2.1 guarantees global convergence of the semi–smooth Newton method, i.e. the algorithm converges from any initial condition. Next we establish that once the iterates are sufficiently close to the solution, then the convergence is superlinear.

For this purpose we introduce the mapping $F\colon L^2(\Omega) \to L^2(\Omega)$ by

$$F(\lambda) = \lambda - \max(0, \bar{\lambda} + \gamma(A^{-1}(f - \lambda) - \psi)). \tag{2.12}$$

Note that $F(\lambda) = 0$ is equivalent to system (2.8). We consider the following reduced algorithm in the variable λ. It arises from applying the quasi–Newton method to $F(\lambda) = 0$. It turns out that the reduced algorithm is equivalent to the primal-dual active set algorithm.

Reduced Algorithm

(i) Choose $y_0 \in H_0^1(\Omega)$, set $\lambda_0 = f - Ay_0$ and $k = 0$.

(ii) Set $\mathcal{A}_{k+1} = \{x\colon [\bar{\lambda} + \gamma A^{-1}(f - \lambda_k) - \gamma\,\psi](x) > 0\}$, $\mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}$.

(iii) Set $\delta\lambda = \lambda_k$ on $\mathcal{I}_{k+1}$, and solve for $\delta\lambda \in L^2(\Omega)$

$$(\delta\lambda + \gamma A^{-1}(\delta\lambda))(x) = [-\lambda_k + \bar{\lambda} - \gamma\,\psi + \gamma A^{-1}(f - \lambda_k)](x), \quad x \in \mathcal{A}_{k+1}.$$

(iv) Set $\lambda_{k+1} = \lambda_k + \delta\lambda$ and goto (ii).

In fact (iii)–(iv) of the reduced algorithm is equivalent to

$$(\lambda_{k+1} - \bar{\lambda} + \gamma\,\psi)(x) = \gamma(A^{-1}(f - \lambda_{k+1}))(x) \;\text{ for }\; x \in \mathcal{A}_{k+1},\; \lambda_{k+1} = 0 \;\text{ in }\; \mathcal{I}_{k+1}.$$

and thus it is equivalent to (iii)–(iv) of the primal–dual active set algorithm with $y_{k+1} = A^{-1}(f - \lambda_{k+1})$. Since the initializations for both algorithms are the same the two algorithms give identical iterates. Note that while λ_0 may only be in $H^{-1}(\Omega)$, the iterates satisfy $\lambda_k \in L^2(\Omega)$ for $k \geq 1$.

Theorem 2.2 *If $\lambda_0 \in L^2(\Omega)$ and $|\lambda_0 - \lambda_\gamma|_{L^2(\Omega)}$ is sufficiently small then $(y_k, \lambda_k) \to (y_\gamma, \lambda_\gamma)$ superlinearly in $H_0^1(\Omega) \times L^2(\Omega)$.*

Proof. First we show superlinear convergence of λ_k to λ_γ by applying Theorem 1.1 to F defined in (2.12). Let $q = \frac{1}{2} - \frac{1}{n}$, if $n \geq 3$ and $q \in (2, \infty)$ if $n = 2$. Then $H^1(\Omega)$ is continuously injected into $L^q(\Omega)$, and $q > 2$ for each n. From Propositions 1.1 and 1.2 it follows that F is Newton–differentiable on $L^2(\Omega)$. For this purpose we set $B = \gamma A^{-1}$ and $b = \bar{\lambda} + \gamma(A^{-1}f - \psi)$ in Proposition 1.2. To apply Theorem 1.1 it remains to verify that the generalized derivatives $G(\lambda) \in \mathcal{L}(L^2(\Omega))$ of F have uniformly bounded inverses. We define

$$\mathcal{A} = \{x\colon [\bar{\lambda} + \gamma(A^{-1}(f - \lambda) - \psi)](x) > 0\}, \quad \mathcal{I} = \Omega \setminus \mathcal{A}.$$

Further let $E_{\mathcal{A}}\colon L^2(\mathcal{A}) \to L^2(\Omega)$ and $E_{\mathcal{I}}\colon L^2(\mathcal{I}) \to L^2(\Omega)$ denote the extension - by - zero operators from $\mathcal{A}$ and $\mathcal{I}$ to Ω, respectively. Their adjoints $E_{\mathcal{A}}^*\colon L^2(\Omega) \to L^2(\mathcal{A})$ and $E_{\mathcal{I}}^*\colon L^2(\Omega) \to L^2(\mathcal{I})$ are restriction operators. The mapping $(E_{\mathcal{A}}^*, E_{\mathcal{I}}^*)\colon L^2(\Omega) \to L^2(\mathcal{A}) \times L^2(\mathcal{I})$ determines an isometric isomorphism and every $\lambda \in L^2(\Omega)$ can uniquely be expressed as $(E_{\mathcal{A}}^*\lambda, E_{\mathcal{I}}^*\lambda)$. A generalized derivative of F in the sense of Definition 1.1 is obtained by setting $\delta = 0$ in the definition $G_{m,\delta}$ for generalized derivatives of the max–operation. We obtain

$$G(\lambda) = I + \gamma\chi_{\mathcal{A}}A^{-1}.$$

This operator can equivalently be expressed as

$$G(\lambda) = \begin{pmatrix} I_{\mathcal{A}} & 0 \\ 0 & I_{\mathcal{I}} \end{pmatrix} + \gamma \begin{pmatrix} E_{\mathcal{A}}^* A^{-1} E_{\mathcal{A}} & E_{\mathcal{A}}^* A^{-1} E_{\mathcal{I}} \\ 0 & 0 \end{pmatrix},$$

where $I_{\mathcal{A}}$ and $I_{\mathcal{I}}$ denote the identity operators on $L^2(\mathcal{A})$ and $L^2(\mathcal{I})$. Let $(g_{\mathcal{A}}, g_{\mathcal{I}}) \in L^2(\mathcal{A}) \times L^2(\mathcal{I})$ be arbitrary and consider the equation

$$G(\lambda)((\delta\lambda)_{\mathcal{A}}, (\delta\lambda)_{\mathcal{I}}) = (g_{\mathcal{A}}, g_{\mathcal{I}}). \tag{2.13}$$

Then necessarily $(\delta\lambda)_{\mathcal{I}} = g_{\mathcal{I}}$ and (2.13) is equivalent to

$$(\delta\lambda)_{\mathcal{A}} + \gamma E_{\mathcal{A}}^* A^{-1} E_{\mathcal{A}} (\delta\lambda)_{\mathcal{A}} = g_{\mathcal{A}} - \gamma E_{\mathcal{A}}^* A^{-1} E_{\mathcal{I}} g_{\mathcal{I}}. \tag{2.14}$$

The Lax–Milgram theorem and positivity of A^{-1} imply the existence of a unique solution $(\delta\lambda)_{\mathcal{A}}$ to (2.14) and consequently (2.13) has a unique solution for every $(g_{\mathcal{A}}, g_{\mathcal{I}})$ and every λ. Moreover these solutions are uniformly bounded w.r.t. $\lambda \in L^2$. This follows from $(\delta\lambda)_{\mathcal{I}} = g_{\mathcal{I}}$ and

$$|\delta\lambda_{\mathcal{A}}|_{L^2(\mathcal{A})} \le |g_{\mathcal{A}}|_{L^2(\Omega)} + \gamma \|A^{-1}\|_{\mathcal{L}(L^2(\Omega))} \, |g_{\mathcal{I}}|_{L^2(\mathcal{I})}.$$

This proves superlinear convergence $\lambda_k \to \lambda_\gamma$ in $L^2(\Omega)$. Superlinear convergence of y_k to y_γ in $H_0^1(\Omega)$ follows from $Ay_k + \lambda_k = f$ and the fact that $A : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a homeomorphism. □

If the problem data are sufficiently regular as specified in Remark 2.3 such that $A\colon H_0^1(\Omega) \cap H^2(\Omega) \to L^2(\Omega)$ is a homeomorphism, then $y_k \to y_\gamma$ in $H^2(\Omega)$ under the assumptions of Theorem 2.2.

3 Convergence of regularized problems

First we establish a general convergence result with respect to the penalty parameter γ. For related results we refer to [GLT], for example.

Theorem 3.1 *The solutions $(y_\gamma, \lambda_\gamma)$ to the regularized problem (2.8) converge to (y^*, λ^*) in the sense that $y_\gamma \to y^*$ strongly in $H_0^1(\Omega)$ and $\lambda_\gamma \rightharpoonup \lambda^*$ weakly in $H^{-1}(\Omega)$ as $\gamma \to \infty$.*

Proof. From (2.4) and (2.8) we have for every $\gamma > 0$

$$a(y_\gamma, y_\gamma - y^*) + (\lambda_\gamma, y_\gamma - y^*) = (f, y_\gamma - y^*).$$

where $\lambda_\gamma = \max(0, \bar{\lambda} + \gamma(y_\gamma - \psi))$. Since $\lambda_\gamma \ge 0$ and $\psi - y^* \ge 0$ we have

$$(\lambda_\gamma, y_\gamma - y^*) = (\lambda_\gamma, \frac{\bar\lambda}{\gamma} + y_\gamma - \psi + \psi - y^* - \frac{\bar\lambda}{\gamma}) \geq \frac{1}{\gamma}(\lambda_\gamma, \bar\lambda + \gamma(y_\gamma - \psi)) - \frac{1}{\gamma}(\lambda_\gamma, \bar\lambda),$$

and hence

$$(\lambda_\gamma, y_\gamma - y^*) \geq \frac{1}{\gamma}|\lambda_\gamma|^2_{L^2} - \frac{1}{\gamma}(\lambda_\gamma, \bar\lambda). \tag{3.1}$$

Using this inequality and the equation derived from (2.4) and (2.8) we have

$$a(y_\gamma, y_\gamma) + \frac{1}{\gamma}|\lambda_\gamma|^2_{L^2} \leq a(y_\gamma, y^*) + (f, y_\gamma - y^*) + \frac{1}{\gamma}(\bar\lambda, \lambda_\gamma).$$

It thus follows from (2.2) that

$$\nu\,|y_\gamma|^2_{H^1_0} + \frac{1}{\gamma}|\lambda_\gamma|^2_{L^2}$$

is uniformly bounded with respect to $\gamma \geq 1$ and hence by (2.8) the family $\{\lambda_\gamma\}_{\gamma\geq 1}$ is bounded in $H^{-1}(\Omega)$. Consequently there exist $(\hat y, \hat\lambda) \in H^1_0(\Omega) \times H^{-1}(\Omega)$ and a sequence $\{(y_{\gamma_n}, \lambda_{\gamma_n}\}$ with $\lim \gamma_n = \infty$ such that

$$w - \lim(y_{\gamma_n}, \lambda_{\gamma_n}) = (\hat y, \hat\lambda) \ \text{ in } \ H^1_0(\Omega) \times H^{-1}(\Omega).$$

Henceforth we drop the subscript n with γ_n. Note that

$$\frac{1}{\gamma}|\lambda_\gamma|^2_{L^2} = \gamma\,|\max(0, \frac{\bar\lambda}{\gamma} + y_\gamma - \psi)|^2_{L^2}.$$

Since $H^1_0(\Omega)$ is embedded compactly into $L^2(\Omega)$, we can assume without loss of the generality that y_γ converges to $\hat y$ a.e. in Ω. From the above equality and Fatou's lemma we conclude that $|(\hat y - \psi)^+| = 0$ and therefore $\hat y \leq \psi$. From (2.4) and (2.8) we also have

$$a(y_\gamma - y^*, y_\gamma - y^*) + \langle \lambda_\gamma - \lambda^*, y_\gamma - y^*\rangle_{H^{-1}, H^1_0} = 0,$$

and by (3.1)

$$(\lambda_\gamma, y_\gamma - y^*) \geq -\frac{1}{4\gamma}|\bar\lambda|^2_{L^2}$$

Hence

$$0 \le \overline{\lim_{\gamma\to\infty}}\, \nu\, |y_\gamma - y^*|^2_{H^1_0} \le \lim_{\gamma\to\infty} \langle \lambda^*, y_\gamma - y^*\rangle_{H^{-1},H^1_o}$$

$$= \langle \lambda^*, \hat{y} - \psi\rangle_{H^{-1},H^1_0} \le 0,$$

where we used the complementarity condition $\langle \lambda^*, y^* - \psi\rangle_{H^{-1},H^1_0} = 0$ and $\hat{y} \le \psi$. It follows that $\lim_{\gamma\to\infty} y_\gamma = y^*$ in $H^1_0(\Omega)$ and hence $\hat{y} = y^*$. Taking the limit in

$$a(y_\gamma, v) + (\lambda_\gamma, v) = (f, v) \ \text{ for all } \ v \in H^1_0,$$

we find

$$a(y^*, v) + \langle \hat{\lambda}, v\rangle_{H^{-1},H^1_0} = (f, v) \ \text{ for all } \ v \in H^1_0.$$

This equation is also satisfied with $\hat{\lambda}$ replaced by λ^* and consequently $\lambda^* = \hat{\lambda}$. Since (y^*, λ^*) is the unique solution to (2.5) the whole family $\{(y_\gamma, \lambda_\gamma)\}$ converges in the sense given in the statement of the theorem. □

In the next two sections we establish monotonicity for the family $\{y_\gamma\}_{\gamma\ge 0}$ and the rate of convergence to y^* in $L^\infty(\Omega)$ for two specific selections of $\bar{\lambda}$. We believe that such results are new and they play an important role in developing a fast algorithm in Section 5.

3.1 Infeasible case

Here we choose $\bar{\lambda} = 0$. For $\gamma > 0$ we set

$$\mathcal{A}_\gamma = \{x \in \Omega\colon (y_\gamma - \psi)(x) > 0\} \ \text{ and } \ \mathcal{I}_\gamma = \Omega \setminus \mathcal{A}_\gamma.$$

Proposition 3.1 *If $0 < \alpha < \beta$ then*

$$y^* \le y_\beta \le y_\alpha, \quad \textit{a.e. in } \ \Omega.$$

Proof. By (2.8) we have

$$\lambda_\alpha - \lambda_\beta = \max(0, \alpha(y_\alpha - \psi)) - \max(0, \beta(y_\beta - \psi)).$$

It follows that

$$\begin{aligned} &(\lambda_\alpha - \lambda_\beta)(x) = 0 \ \text{ for } \ x \in \mathcal{I}_\alpha \cap \mathcal{I}_\beta \\ &(\lambda_\alpha - \lambda_\beta)(x) \le \beta(y_\alpha - y_\beta)(x) \ \text{ for } \ x \in \mathcal{A}_\alpha \cap \mathcal{A}_\beta. \end{aligned} \tag{3.2}$$

For $x \in I_\beta \cap \mathcal{A}_\alpha$ we find $(\lambda_\alpha - \lambda_\beta)(x) = \alpha(y_\alpha - \psi)(x) \leq \beta(y_\alpha - \psi)(x) - \beta(y_\beta - \psi)(x) = \beta(y_\alpha - y_\beta)(x)$, and thus

$$(\lambda_\alpha - \lambda_\beta)(x) \leq \beta(y_\alpha - y_\beta)(x) \quad \text{for} \quad x \in \mathcal{I}_\beta \cap \mathcal{A}_\alpha. \tag{3.3}$$

Moreover

$$(\lambda_\alpha - \lambda_\beta)(x) \leq 0 \quad \text{for} \quad x \in \mathcal{A}_\beta \cap \mathcal{I}_\alpha. \tag{3.4}$$

For (3.2)–(3.4) and (2.8) we find

$$a(y_\beta - y_\alpha, (y_\beta - y_\alpha)^+) = (\lambda_\alpha - \lambda_\beta, (y_\beta - y_\alpha)^+) \leq 0$$

and hence $y_\beta \leq y_\alpha$. The verification that $y^* \leq y_\alpha$ is quite similar. □

Proposition 3.2 *For $0 < \alpha < \beta$ we have*

$$\mathcal{I}^* \supset \mathcal{I}_\beta \supset \mathcal{I}_\alpha.$$

Proof. If $x \in \mathcal{A}_\beta \cap \mathcal{I}_\alpha$ then $(y_\alpha - \psi)(x) \leq 0$ and $(y_\beta - \psi)(x) > 0$. Hence $y_\alpha(x) < y_\beta(x)$ which contradicts Proposition 3.1 and therefore $\mathcal{I}_\beta \supset \mathcal{I}_\alpha$. □

Our next objective is to prove convergence of y_γ to y^* in $L^\infty(\Omega)$ with rate γ^{-1}, provided certain regularity conditions are satisfied. We require a technical lemma which we describe first. For this purpose let ω denote a subdomain of Ω with Lipschitzian boundary $\partial\omega$. The restriction of a to $H^1(\omega) \times H^1(\omega)$ will be denoted by the same symbol.

Lemma 3.1 *Assume that $g \in L^\infty(\omega)$ and that $a(1, v) \geq 0$, for all $v \in H^1(\omega)$ with $v \geq 0$. For $c > 0, c \in \mathbb{R}$, let $y_c \in H_0^1(\omega)$ denote the solution to*

$$a(y, v) + c(y, v)_{L^2(\omega)} = (g, v)_{L^2(\omega)} \quad \textit{for all} \quad v \in H_0^1(\omega). \tag{3.5}$$

Then $y_c \in L^\infty(\omega)$ and $|y_c|_{L^\infty(\omega)} \leq \frac{1}{c}|g|_{L^\infty(\omega)}$.

Proof. For the sake of completeness we include the proof which can be obtained with known techniques. Let $\bar{g} = \max(0, \sup\{g(x) \colon x \in \omega\})$.
For all $v \in H_0^1(\omega)$ we have

$$a(y_c - \tfrac{1}{c}\bar{g}, v) + (\bar{g} - g, v)_{L^2(\omega)} = (\bar{g} - cy_c, v)_{L^2(\omega)} - a(\tfrac{\bar{g}}{c}, v). \tag{3.6}$$

Set $v = (y_c - \frac{1}{c}\bar{g})^+$ and observe that $v \in H_0^1(\omega)$ since $\bar{g} \geq 0$. Since $a(1, v) \geq 0$ for all $v \in H^1(\omega)$ and $v \geq 0$, it follows from (3.6) that

$$a(y_c - \tfrac{1}{c}\,\bar{g}, (y_c - \tfrac{1}{c}\,\bar{g})^+) \leq 0$$

and consequently $y(x) \leq \frac{1}{c}|g|_{L^\infty(\omega)}$ for a.e. $x \in \omega$. Analogously it can be verified that $-\frac{1}{c}|g|_{L^\infty(\omega)} \leq y(x)$ for a.e. $x \in \omega$. $\square$

Let us introduce the active and inactive sets associated to the solution y^* of (1.1):

$$\mathcal{A}^* = \{x \in \Omega \colon y^*(x) = \psi(x)\}, \quad \mathcal{I}^* = \{x \in \Omega \colon y^*(x) < \psi(x)\},$$

with boundaries $\partial\mathcal{A}^*$ and $\partial\mathcal{I}^*$ respectively.

Theorem 3.2 *Let the regularity requirements of Remark 2.3 be satisfied and assume that $f \in L^\infty(\Omega)$ and $A\psi \in L^\infty(\Omega)$. If, moreover, the boundary $\partial\mathcal{A}^*$ of the active set is $C^{1,1}$ manifold in R^{n-1} and for every $\gamma > 0$ the boundary $\partial\mathcal{A}_\gamma$ of $\mathcal{A}_\gamma$ is Lipschitzian manifold in R^{n-1} , then*

$$|y_\gamma - y^*|_{L^\infty(\Omega)} \leq \tfrac{1}{\gamma}|f - A\psi|_{L^\infty(\Omega)}.$$

Proof. The regularity assumption imply that $y^* \in W^{2,p}(\Omega)$ and $y_\gamma \in W^{2,p}(\Omega)$ with $p > n$. Recall that $W^{2,p}(\Omega) \subset C(\bar{\Omega})$ if $p > n$. From Proposition 3.2 it follows that $\mathcal{A}^* \subset \mathcal{A}_\gamma$ for every $\gamma > 0$. From the definition of $\mathcal{A}_\gamma$ we have

$$\begin{cases} Ay_\gamma + \gamma(y_\gamma - \psi) = f & \text{in } \mathcal{A}_\gamma \\ y_\gamma - \psi = 0 & \text{on } \partial\mathcal{A}_\gamma. \end{cases}$$

Consequently

$$\begin{cases} A(y_\gamma - \psi) + \gamma(y_\gamma - \psi) = f - A\psi & \text{in } \mathcal{A}_\gamma \\ y_\gamma - \psi = 0 & \text{on } \partial\mathcal{A}_\gamma. \end{cases}$$

From Lemma 3.1 with $\omega = \mathcal{A}_\gamma$ and $g = f$ we find

$$|y_\gamma - \psi|_{L^\infty(\mathcal{A}_\gamma)} \leq \tfrac{1}{\gamma}|f - A\psi|_{L^\infty(\Omega)}$$

and in particular

$$|y_\gamma - \psi|_{L^\infty(\mathcal{A}^*)} \leq \tfrac{1}{\gamma}|f - A\psi|_{L^\infty(\Omega)}. \tag{3.7}$$

Note further that on $\mathcal{I}^*$ we have

$$\begin{cases} A(y_\gamma - y^*) = \lambda^* - \lambda_\gamma \leq 0 & \text{in } \mathcal{I}^* \\ \\ y_\gamma - y^* = y_\gamma - \psi \geq 0 & \text{on } \partial\mathcal{I}^*. \end{cases} \tag{3.8}$$

From the maximum principle applied to (3.8), and (3.7) it follows that

$$|y_\gamma - y^*|_{L^\infty(\mathcal{I}^*)} \leq |y_\gamma - \psi|_{L^\infty(\partial\mathcal{I}^*)} \leq \tfrac{1}{\gamma}|f - A\psi|_{L^\infty(\Omega)}, \tag{3.9}$$

see e.g. [T], pg 191. Combining (3.7) and (3.9) gives the desired conclusion. □

To justify the terminology to refer to $\bar{\lambda} = 0$ as the infeasible case note that if $y_\gamma < \psi$ for some $\gamma > 0$ then $\mathcal{I}_\gamma = \Omega$, $\lambda_\gamma = 0$ and $(y_\gamma, \lambda_\gamma)$ satisfy the optimality system (2.4). Consequently $(y^*, \lambda^*) = (y_\gamma, \lambda_\gamma)$ and y^* is also a solution of the unconstrained problem. Thus unless y^* is also a solution to the unconstrained problem, $y_\gamma \leq \psi$ for some finite γ is impossible. In the following subsection it will be shown that proper choice of $\bar{\lambda}$ guarantees feasibility of the solutions y_γ to (2.8).

3.2 Feasible case

Here we choose $\bar{\lambda} \in L^2(\Omega)$ such that

$$\begin{cases} \bar{\lambda} \geq 0 \ \text{ and } \ \bar{\lambda} - (f - A\psi) \geq 0, \ \text{ a.e. in } \Omega \\ \\ \langle \bar{\lambda} - (f - A\psi), v \rangle_{H^{-1}, H_0^1} \geq 0 \ \text{ for all } \ v \in K. \end{cases} \tag{3.10}$$

Note that if $\psi \in H^2(\Omega)$ then for the choice $\bar{\lambda} = \max(0, f - A\psi)$ (3.10) is satisfied.

Proposition 3.3 *If (3.10) holds and* $0 < \alpha < \beta$ *then*

$$y_\alpha \leq y_\beta \leq \psi \ \ \textit{a.e. in } \Omega.$$

In particular y_α *is feasible for every* $\alpha > 0$.

Proof. From (2.8) we have by (3.10)

$$a(y_\alpha - \psi, (y_\alpha - \psi)^+) = (f - \lambda_\alpha, (y_\alpha - \psi)^+) - a(\psi, (y_\alpha - \psi)^+)$$

$$= \langle f - A\psi - \max(0, \bar{\lambda} + \alpha(y_\alpha - \psi)), (y_\alpha - \psi)^+ \rangle$$

$$= \langle f - A\psi - \bar{\lambda}, (y_\alpha - \psi)^+ \rangle - \alpha(y_\alpha - \psi, (y_\alpha - \psi)^+)$$

$$\leq -\alpha |\, (y_\alpha - \psi)^+|^2 \leq 0$$

and hence by (2.3)

$$y_\alpha \leq \psi.$$

It follows that y_α is feasible for every $\alpha > 0$.

Next let $0 < \alpha < \beta$. By (2.8)

$$a(y_\alpha - y_\beta, (y_\alpha - y_\beta)^+) = (\lambda_\beta - \lambda_\alpha, (y_\alpha - y_\beta)^+). \tag{3.11}$$

We introduce the set

$$S = \{x \colon y_\alpha(x) - y_\beta(x) > 0\}$$

and decompose this set as $S = S_1 \cap S_2 \cup S_3$, where

$$S_1 = \{x \in S \colon (\bar{\lambda} + \beta(y_\beta - \psi))(x) \leq 0\}$$

$$S_2 = \{x \in S \colon (\bar{\lambda} + \beta(y_\beta - \psi))(x) > 0, (\bar{\lambda} + \alpha(y_\alpha - \psi))(x) \leq 0\}$$

$$S_3 = \{x \in S \colon (\bar{\lambda} + \beta(y_\beta - \psi))(x) > 0, (\bar{\lambda} + \alpha(y_\alpha - \psi))(x) > 0\}.$$

To estimate the right hand side of (3.11) recall that

$$\lambda_\beta - \lambda_\alpha = \max(0, \bar{\lambda} + \beta(y_\beta - \psi)) - \max(0, \bar{\lambda} + \alpha(y_\alpha - \psi)).$$

We find

$$(\lambda_\beta - \lambda_\alpha, (y_\alpha - y_\beta)^+) = (\lambda_\beta - \lambda_\alpha, y_\alpha - y_\beta)_{L^2(S_1)} + (\lambda_\beta - \lambda_\alpha, y_\alpha - y_\beta)_{L^2(S_2)}$$

$$+(\lambda_\beta - \lambda_\alpha, y_\alpha - y_\beta)_{L^2(S_3)}$$

$$\leq (\beta(y_\beta - \psi) - \alpha(y_\alpha - \psi), (y_\alpha - y_\beta)_{L^2(S_2)} + (\beta(y_\beta - y_\alpha), y_\alpha - y_\beta)_{L^2(S_3)}$$

$$+(\beta(y_\alpha - \psi) - \alpha(y_\alpha - \psi), y_\alpha - y_\beta)_{L^2(S_3)}.$$

Utilizing the fact that $y_\alpha \leq \psi$ and $y_\beta \leq \psi$ we find

$$(\lambda_\beta - \lambda_\alpha, (y_\alpha - y_\beta)^+) \leq \beta(y_\beta - y_\alpha, y_\alpha - y_\beta)_{L^2(S_2)}$$

$$+(\beta - \alpha)(y_\alpha - \psi, y_\alpha - y_\beta)_{L^2(S_3)} \leq 0.$$

Inserting this estimate into (3.11) and using the weak maximum principle implies that $y_\alpha \leq y_\beta$. $\square$

Corollary 3.1 *If (3.10) holds and* $0 < \alpha < \beta$ *then*

$$0 \leq \lambda_\alpha \leq \lambda_\beta \leq \max(0, \bar{\lambda})$$

and $\mathcal{I}_\alpha \supset \mathcal{I}_\beta$.

Proof. From the representation $\lambda_\gamma = \max(0, \bar{\lambda} + \gamma(y_\gamma - \psi))$ and the fact that $\gamma \to \gamma(y_\gamma - \psi)(x)$ is increasing with respect to γ for a.e. $x \in \Omega$, it follows that λ_γ is increasing and $\mathcal{I}_\gamma$ is decreasing with respect to γ. The estimate $\lambda_\gamma \leq \max(0, \bar{\lambda})$ is a consequence of the feasibility of y_γ for every γ. $\square$

As in the infeasible case we can consider the question of rate of convergence with respect to γ if additional regularity requirements are satisfied.

Remark 3.1 From Theorem 3.1, Corollary 3.1 and Lebesgue's monotone convergence theorem it follow that $\lambda_\gamma \to \lambda^*$ strongly in $L^2(\Omega)$.

Theorem 3.3 *Assume that* $f \in L^\infty(\Omega)$, $A\psi \in L^\infty(\Omega)$, $\bar{\lambda} \in L^\infty(\Omega)$ *and that the assumptions of Remark 2.3 hold. If in addition* $\mathcal{A}_\gamma$ *is a domain with a* $C^{1,1}$ *boundary, then*

$$|y_\gamma - y^*|_{L^\infty(\Omega)} \leq \tfrac{1}{\gamma} |\bar{\lambda}|_{L^\infty(\Omega)}.$$

Proof. By the assumptions of the theorem y^* and $y_\gamma \in W^{2,p}(\Omega)$, $p > n$. On $\overline{\mathcal{A}_\gamma}$ we have $\bar{\lambda} + \gamma(y_\gamma - \psi) \geq 0$ and $y_\gamma \leq \psi$, and hence

$$|y_\gamma - \psi|_{L^\infty(\mathcal{A}_\gamma)} \leq \tfrac{1}{\gamma} |\bar{\lambda}|_{L^\infty(\Omega)}.$$

Since $\mathcal{A}_\gamma \subset \mathcal{A}^*$ by Corollary 3.1 this implies that

$$|y_\gamma - y^*|_{L^\infty(\mathcal{A}_\gamma)} \leq \tfrac{1}{\gamma} |\bar{\lambda}|_{L^\infty(\Omega)}.$$

Moreover we have

$$\begin{cases} A(y^* - y_\gamma) = \lambda_\gamma - \lambda^* \leq 0 & \text{in } \mathcal{I}_\gamma \\ \\ y^* - y_\gamma \leq 0 & \text{on } \partial\mathcal{I}_\gamma. \end{cases}$$

From the maximum principle and the regularity assumption on $\partial\mathcal{I}_\gamma$ it follows that

$$|y^* - y_\gamma|_{L^\infty(\mathcal{I}_\gamma)} \leq \tfrac{1}{\gamma}\, |\bar{\lambda}|_{L^\infty(\Omega)}.$$

□

4 Bilateral constraints

The treatment of bilateral constraints gives rise to some additional difficulties. Here we consider

$$\begin{cases} \min\ \frac{1}{2}\, a(y,y) - (f,y) \\ \\ \text{over } y \in H_0^1(\Omega) \\ \\ \varphi \leq y \leq \psi \ \text{ in } \Omega. \end{cases} \tag{4.1}$$

Throughout this section we assume that

$$a(\varphi,\psi) = (\nabla\varphi, \nabla\psi) \text{ for all } \varphi,\psi \in H_0^1(\Omega),$$

that φ and ψ are in $H^1(\Omega)$, that $\partial\Omega$ is $C^{1,1}$ and

$$\begin{aligned} \varphi \leq 0 \leq \psi \text{ on } \partial\Omega, \quad & \max(0, \Delta\psi + f) \in L^2(\Omega), \\ & \min(0, \Delta\varphi + f) \in L^2(\Omega), \end{aligned} \tag{4.2}$$

$$S_1 := \{x \in \Omega\colon \Delta\psi + f > 0\} \cap S_2 := \{x \in \Omega\colon \Delta\varphi + f < 0\} = \emptyset, \tag{4.3}$$

and that there exists $c_0 > 0$ such that

$$-\Delta(\psi - \varphi) + c_0(\psi - \varphi) \geq 0 \ \text{ a.e. in } \Omega. \tag{4.4}$$

Under these assumptions it was shown in [IK1] that there exists a solution $y^* \in H_0^1(\Omega) \cap H^2(\Omega)$ to (4.1) with associated Lagrange multiplier $\lambda^* \in L^2(\Omega)$. This was verified by passing to the limit $\gamma \to \infty$ in

$$(4.5)\quad -\Delta y_\gamma + \lambda_\gamma = f, \quad \lambda_\gamma = \begin{cases} \bar\lambda + \gamma\,(y_\gamma - \psi) & \text{if } \bar\lambda + \gamma\,(y_\gamma - \psi) > 0, \\ \bar\lambda + \gamma\,(y_\gamma - \varphi) & \text{if } \bar\lambda + \gamma\,(y_\gamma - \varphi) < 0, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$(4.6)\qquad \bar\lambda = \begin{cases} \Delta\psi + f & \text{on } S_1 \\ \Delta\varphi + f & \text{on } S_2 \\ 0 & \text{otherwise.} \end{cases}$$

The weak limit of $(y_\gamma, \lambda_\gamma)$ in $H^2(\Omega) \times L^2(\Omega)$ satisfies

$$-\Delta y^* + \lambda^* = f, \quad \lambda^* = \begin{cases} \lambda^* + c\,(y^* - \psi) & \text{if } \lambda^* + c\,(y^* - \psi) > 0, \\ \lambda^* + c\,(y^* - \varphi) & \text{if } \lambda^* + c\,(y^* - \varphi) < 0, \\ 0 & \text{otherwise,} \end{cases}$$

for every $c > 0$. The latter equation can be equivalently expressed as

$$(4.7)\qquad \begin{cases} -\Delta y^* + \lambda^* = f, \\ \\ \lambda^* = \max(0, \lambda^* + c(y^* - \psi)) + \min(0, \lambda^* + c(y^* - \varphi)), \end{cases}$$

for every arbitrary fixed $c > 0$.

Primal–Dual Active Set–Algorithm

(i) Choose y_0 , set $k = 0$.

(ii) Set

$$\mathcal{A}^\psi_{k+1} = \{x\colon (\bar\lambda + \gamma(y_k - \psi))(x) > 0\},$$

$$\mathcal{A}^\varphi_{k+1} = \{x\colon (\bar\lambda + \gamma(y_k - \varphi))(x) < 0\}$$

$$\mathcal{I}_{k+1} = \Omega \setminus (\mathcal{A}^\psi_{k+1} \cup \mathcal{A}^\varphi_{k+1}).$$

(iii) Solve for $y_{k+1} \in H^1_0(\Omega)$.

(4.8) $a(y,v) + ((\bar\lambda + (y-\psi))\chi_{\mathcal{A}^\psi_{k+1}}, v) + ((\bar\lambda + (y-\varphi))\chi_{\mathcal{A}^\varphi_{k+1}}, v) = (f,v),$

for all $v \in H^1_0(\Omega)$. Set

$$\lambda_{k+1} = \begin{cases} 0 & \text{on } \mathcal{I}_{k+1} \\ \bar\lambda + \gamma(y_{k+1} - \psi) & \text{on } \mathcal{A}^\psi_{k+1} \\ \bar\lambda + \gamma(y_{k+1} - \varphi) & \text{on } \mathcal{A}^\varphi_{k+1}. \end{cases}$$

(iv) Stop, or $k = k+1$ and goto (ii).

(y_k, λ_k) to $(y_\gamma, \lambda_\gamma)$.

For the following local convergence result the choice of $\bar\lambda$ as in (4.6) is not essential.

Proposition 4.1 *If $|\lambda_0 - \lambda_\gamma|_{L^2(\Omega)}$ is sufficiently small then $(y_k, \lambda_k) \to (y_\gamma, \lambda_\gamma)$ superlinearly in $H^1_0(\Omega) \times L^2(\Omega)$.*

Proof. The proof is quite similar to that of Theorem 2.2 and we therefore only give a brief outline. Again the algorithm is expressed in the variable λ only. The resulting iteration map $F\colon L^2(\Omega) \to L^2(\Omega)$ is given by

$$F(\lambda) = \lambda - \max(0, \bar\lambda + \gamma(A^{-1}(f-\lambda) - \psi)) - \min(0, \bar\lambda + \gamma(A^{-1}(f-\lambda) - \varphi)),$$

and (4.7) is equivalent to $F(\lambda) = 0$. Steps (ii) and (iii) of the reduced algorithm are replaced by:

(ii') Set

$$\mathcal{A}^\psi_{k+1} = \{x\colon (\bar\lambda + \gamma A^{-1}(f - \lambda_k) - \gamma\,\psi((x) > 0\},$$

$$\mathcal{A}^\varphi_{k+1} = \{x\colon (\bar\lambda + \gamma A^{-1}(f - \lambda_k) - \gamma\,\varphi)(x) < 0\},$$

$$\mathcal{I}_{k+1} = \Omega \setminus (\mathcal{A}^\psi_{k+1} \cup \mathcal{A}^\varphi_{k+1}).$$

(iii') Set

$$\delta\lambda = -\lambda_k \ \text{ on } \ \mathcal{I}_{k+1} \ \text{ and solve for } \ \delta\lambda \in H^{-1}$$

$$\delta\lambda + \gamma A^{-1}(\delta\lambda) = -\lambda_k + \bar{\lambda} - \gamma\,\psi + \gamma A^{-1}(f - \lambda_k) \ \text{ in } \ \mathcal{A}^{\psi}_{k+1}$$

$$\delta\lambda + \gamma A^{-1}(\delta\lambda) = -\lambda_k + \bar{\lambda} - \gamma\,\varphi + \gamma A^{-1}(f - \lambda_k) \ \text{ in } \ \mathcal{A}^{\varphi}_{k+1}.$$

As in the proof of Theorem 2.2 one argues that F is Newton–differentiable. To characterize the generalized derivative we set

$$c_\psi = \bar{\lambda} + \gamma(A^{-1}(f - \lambda) - \psi), \ c_\varphi = \bar{\lambda} + \gamma(A^{-1}(f - \lambda) - \varphi),$$

and

$$\mathcal{A}_\psi = \{x\colon c_\psi(x) > 0\}, \ \mathcal{A}_\varphi = \{x\colon c_\varphi(x) < 0\}, \ \mathcal{I} = \Omega \setminus (\mathcal{A}_\psi \cup \mathcal{A}_\varphi).$$

A generalized derivative is given by

$$G(\lambda) = I + \gamma\,\chi_{\mathcal{A}_\psi} A^{-1} + \gamma\,\chi_{\mathcal{A}_\varphi} A^{-1} = I + \gamma\,\chi_{\mathcal{A}} A^{-1},$$

where $\mathcal{A} = \mathcal{A}_\varphi \cup A_\psi$. Existence and uniform boundedness of the inverses of $G(\lambda)$ is verified as in the proof of Theorem 2.2. □

5 Numerical experiments

In this section we describe some numerical experiments to illustrate and confirm our results. The problem under consideration is

$$\text{(5.1)} \qquad \begin{cases} -\Delta y + \lambda = f & \text{in } \ \Omega, \\ y = 0 & \text{on } \ \partial\Omega \\ y \le \psi, \ \lambda \ge 0, \quad (\lambda, y - \psi)_{L^2(\Omega)} = 0 \end{cases}$$

which is discretized by means of node–based finite differences. In the one-dimensional case $\Omega = (0,1)$ with grid $\{x_i\} = \{\frac{i}{m}\}_{i=0}^{m}$ and in the two–dimensional case $\Omega = (0,1)\times(0,1)$ with grid $\{x_{i,j}\} = \{(\frac{i}{m}, \frac{j}{m}\}_{i,j=0}^{m}$. The

functions y, λ, f and ψ are discretized by grid functions denoted by the same symbol and $-\Delta$ is discretized by the three-, respectively five-point finite difference stencil. It is well-known that the resulting discretized operator $-\Delta_h$ satisfies the discrete maximum principle. Unless specified otherwise the algorithms are initialized with the unconstrained solution to (5.1), i.e. $\psi = \infty$.

Example 5.1 Here $\Omega = (0, 1)$, $f = \frac{1}{4} \times \sin(5x)$, $\psi = \frac{1}{4}$, and $m = 100$. For all runs that we report upon the primal–dual active set algorithm converges in finitely many steps, i.e. the situation discussed in Proposition 2.1 occurs. We denote the number of iterations that are required until the algorithm reaches the solution of the discretized problem by iter. For $\gamma > 0$ the iterates of the algorithm are denoted by y_k, the solution by y_γ. Similarly $\mathcal{A}_k$ stands for the active sets of the iterates, $\mathcal{A}_\delta$ for the active set corresponding to solution y_δ. In this example as well as in Examples 5.2 and 5.3 below the algorithm was terminated when in two successive iterations the active sets coincide. The current variables then give the solution of the discretized problem.

Let us start with some general observations for the numerical solution:

- $y_{\gamma_2} \leq y_{\gamma_1}$ for $\gamma_1 \leq \gamma_2$ and $\bar{\lambda} = 0$.
- $y_{\gamma_1} \leq y_{\gamma_2}$ for $\gamma_1 \leq \gamma_2$ and $\bar{\lambda} = \max(0, f + \Delta_k \psi)$.
- $\mathcal{A}_{k+1} \subset \mathcal{A}_k$ for $\gamma > 0$ and $\bar{\lambda} = 0$ or $\bar{\lambda} = \max(0, f + \Delta_h \psi)$.
- iter $(\gamma_1) \geq$ iter (γ_2) if $\gamma_1 \leq \gamma_2$
- for large γ changes after the initialization phase from active to inactive occur only along the boundary of $\mathcal{A}_k$. This is not the case for small γ.

In Table 1 we report the required number of iterations and the cardinality of the active set $\mathcal{A}$ as a function of γ, for $\bar{\lambda} = 0$.

γ	2.5	5	10	20	100	1000	10000
iter	4	4	6	6	15	20	20
$\#(\mathcal{A}_\gamma)$	29	26	23	22	19	18	18

Table 1:

The results of Table 1 suggest to combine the primal–dual active set strategy with a continuation procedure with respect to γ: Thus we start with small γ and use the solution as initialization for the algorithm with larger γ. Table 2 shows that this continuation method is effective.

γ	5	20	10^4
iter	5	3	4

Table 2:

Concerning superlinear convergence of the algorithm for fixed $\gamma \in (1, \infty)$ it is not obvious whether the continuous result can be used as indicator for the discrete one, due to finite speed of propagation of the discrete Laplacian. In Table 3 we report the results for the quotients

$$q_k = (y_{k+1} - y_{20})^T \Delta_h (y_{k+1} - y_{20}) \diagup (y_k - y_{20}) \Delta_h (y_k - y_{20}),$$

for selected values of k, where $\bar{\lambda} = 0$, $\gamma = 10^4$.

k	2	6	10	13	14	15	16	17	18
q_k	.84	.80	.72	.62	.58	.51	.43	.31	.13

Table 3:

It is quite typical for the runs that we tested that the quotients decrease approximately by one power of 10, between initialization and final result.

Example 5.2 For this example $\Omega = (0,1) \times (0,1)$, $f = 500\, x\, \sin(5x) \cos(2y)$, $\psi = 1$ on the annulus, $R = \{(x,y)\colon .2 < \sqrt{x^2 + y^2} < .4\}$, $\psi = 10$ on $\Omega \setminus R$, and $m = 200$. Again the algorithm with $\bar{\lambda} = 0$ and $\bar{\lambda} = \max(0, f + \Delta_h \psi)$ terminates after finitely many iterations. The same observations can be made as for the one-dimensional example above. Typical results for $\bar{\lambda} = 0$ and various values of γ are given in Table 4.

For $\bar{\lambda} = \max(0, f + \Delta_h \psi)$ the number of required iterations is comparable and the final active sets for $\gamma \geq 10^8$ is the same. For $\gamma = 10^3$ changes

γ	10^3	10^4	10^5	10^6	10^7	10^8	10^9
iter	5	8	12	27	35	36	36
$\#(\mathcal{A}_\gamma)$	3117	2530	2348	2306	2302	2301	2301

Table 4:

from active to inactive sets take place along the boundaries of these sets in layers up to the depth of 16 pixels. Continuation procedures with respect to γ as explained in Example 5.1 again reduce the total number of iterations significantly, see Table 5.

γ	10^4	10^6	10^8
iter	8	5	1

Table 5:

We carried out computations with the same specifications as in Table 5 with a series of mesh-sizes characterized by $m = (100, 200, 300, 400)$. The resulting number of total iterations are $(11, 14, 16, 20)$.– Again superlinear convergence of the iterates can be observed. In Table 6 we give selected results for the quotients q_k with $m = 200$, $\gamma = 10^8$ and $\bar{\lambda} = 0$. Since in this case the algorithm terminates in 36 iterations we set $q_f = q_{36}$.

k	1	8	15	22	29	31	33	35
q_k	.86	.82	.79	.75	.55	.42	.21	.17

Table 6:

Tests with the smooth obstacle $\psi = 8((x - \frac{1}{2})^2 + (y - \frac{1}{2})^2) - 1$ give quite similar results. The iteration procedure with the same values for γ as in Table 5 requires 16 iterations to obtain the solution, without continuation procedure 44, for $\gamma = 10^8$.

Example 5.3 This is an example with lack of strict complementarity. The choice for Ω, and f is as in Example 5.2. We set m=40. Let y_h^* denote

the solution to the unconstrained problem $-\Delta_h y_h = f$, and define $\psi = 10$ except on $S = (\frac{3}{8}, \frac{5}{8})$, where it is set equal to y_h^*. The algorithm with both $\bar{\lambda} = 0$ and $\bar{\lambda} = \max(0, f + \Delta_h \psi)$ terminates in 1 iteration for a large range of γ–values. In a similar experiment with $m = 30$ and S replaced by $(\frac{1}{3}, \frac{2}{3}) \times (\frac{1}{3}, \frac{2}{3})$ the algorithm starts to chatter if $\bar{\lambda} = 0$, while it converges in finitely many iterations comparable to those in Table 4 for $\bar{\lambda} = \max(0, f + \Delta_h \psi)$. Due to finite precision arithmetic and the fact that the active/inactive set structure and the stopping rule are determined by commands involving machine zero, chattering in the case of lack of strict complementarity comes as no surprise. There are various remedies to avoid chattering based on stopping rules involving machine epsilon. The alternative choice of using $\bar{\lambda} = \max(0, f + \Delta_h \psi)$ rather than $\bar{\lambda} = 0$ has consistently eliminated chattering in this and other examples. For instance, again with $m = 30$, we chose $\psi = 10$ on $\Omega \backslash S$ and $\psi = y_h^* - 1$. In the interior of the active set we have lack of strict complementarity and for $\bar{\lambda} = 0$ and $\gamma > 10^6$ the iterates chatter. With $\bar{\lambda} = \max(0, f + \Delta_h \psi)$ no chattering occurs.

In Examples 5.1 and 5.2 we investigated the case when the penalty parameter tends to ∞. For a specific application it may be desirable to compute with a fixed penalty parameter. For this purpose the penalty parameter should be chosen such that the error due to penalization is of the same order as that due to discretization. Theorems 3.2 and 3.3 then suggest to choose γ proportional to h^{-2}. The success of this procedure is illustrated in the following example.

Example 5.4 The following example from [G], p. 44–45. It represents an elasto–plastic torsion problem formulated as obstacle problem on the unit disc $\widetilde{\Omega}$ with center at $(.5, .5)$. Let $r = \sqrt{(x_1 - .5)^2 + (x_2 - .5)^2}$, $d > 2$ be a constant and set

$$\psi(x) = 1 - r, \quad \text{and} \quad f(x) = d.$$

Then the solution to the obstacle problem on $\widetilde{\Omega}$ is given by

$$y(x) = \begin{cases} 1 - r & \text{if } \frac{2}{d} \le r \le 1 \\ \\ \frac{d}{4}[(1 - r^2) - (r - \frac{2}{d})^2] & \text{if } 0 \le r \le \frac{2}{d}. \end{cases}$$

In our calculation we chose $d = 5.123$.

To compute on the unit square Ω we used exact non–homogeneous boundary conditions assigned at the boundary. The regularization parameter was

$\gamma = |\bar{\lambda}|_\infty / h^2$ with $h = 0.01$. The exact interface Γ between the active and inactive sets is given by $r = 1 - \frac{2}{d}$. We use this example to demonstrate how the interface can be approximated by our proposed algorithms. In the following figures we show the estimated interface for both the infeasible (Γ_1) and feasible (Γ_2) method determined by means of

$$
\begin{aligned}
\Gamma_1 &= \{x \in R^2 \colon y - \psi = 0\} \\
\Gamma_2 &= \{x \in R^2 \colon \bar{\lambda} + \gamma(y - \psi) = 0\}
\end{aligned}
\tag{5.2}
$$

or alternatively by

$$
\begin{aligned}
\Gamma_1 &= \{x \in R^2 \colon y - \psi = \tfrac{1}{\gamma}\} \\
\Gamma_2 &= \{x \in R^2 \colon \bar{\lambda} + \gamma(y - \psi) = \tfrac{1}{\gamma}\}
\end{aligned}
\tag{5.3}
$$

Figure 1 shows a blow–up section of Γ and the Γ_1 's for the infeasible method and Figure 2 shows a blow–up section of Γ and the Γ_2 's for the feasible method. The most outer curve is for the exact interface Γ both in Figures 1 and 2. In this example the second estimates by (5.3) provide better and smoother estimates of the interface Γ both for the infeasible and feasible methods than (5.2).

References

[B] D. P. Bertsekas: *Constrained Optimization and Lagrange Mulitpliers*, Academic Press, New York, 1982.

[BHHK] M. Bergounioux, M. Haddou, M. Hintermüller and K. Kunisch: *A comparison of a Moreau–Yosida based active strategy and interior point methods for constrained optimal control problems*, SIAM J. on Optimization, **11** (2000), 495–521.

[BIK] M. Bergounioux, K. Ito and K. Kunisch: *Primal–dual strategy for constrained optimal control problems* , SIAM J. Control and Optimization, **37** (1999), 1176–1194.

[D] Z. Dostal: *Box constrained quadratic programming with proportioning and projections*, SIAM J. Optimization **7** (1997), 871–887.

[G] R. Glowinski: *Numerical Methods for Nonlinear Variational Problems*, Springer Verlag, New York, 1984.

[GLT] R. Glowinski, J.L. Lions and T. Tremolieres: *Analyse Numerique des Inequations Variationnelles, Vol. 1*, Dunod, Paris, 1976.

[HIK] M. Hintermüller, K. Ito and K. Kunisch: *The primal–dual active set strategy as semi–smooth Newton method*, preprint.

[H] R. Hoppe: Multigrid algorithms for variational inequalities, SIAM J. Numerical Analysis, **24** (1987), 1046–1065.

[HK] R. Hoppe and R. Kornhuber: *Adaptive multigrid methods for obstacle problems*, SIAM J. Numerical Analysis, **31** (1994), 301–323.

[IK1] K. Ito and K. Kunisch: *Augmented Lagrangian methods for nonsmooth convex optimization in Hilbert spaces*, Nonlinear Analysis, Theory, Methods and Applications **41** (2000), 573–589.

[IK2] K. Ito and K. Kunisch: *Optimal control of elliptic variational inequalities*, Applied Mathematics and Optimization, **41** (2000), 343–364.

[KS] D. Kinderlehrer and G. Stampacchia: *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[T] D. M. Troianello: *Elliptic Differential Equations and Obstacle Problems*, Plenum Press, New York, 1987.

[U] M. Ulbrich, *Semi-smooth Newton methods for operator equations in function space*, SIAM J. Optimization, to appear.

Semi-Smooth Newton Methods for State-Constrained Optimal Control Problems.

Kazufumi Ito [1] and Karl Kunisch[2]

July 2002

[1]Department of Mathematics, North Carolina State University, Raleigh, North Carolina, USA

[2]Institut für Mathematik, Karl-Franzens-Universität Graz, Heinrichstrasse 36, A-8010 Graz, Austria, supported by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03 ,,Optimierung und Kontrolle".

Abstract

A regularized optimality system for state-constrained optimal control problems is introduced and semi-smooth Newton methods for its solution are analyzed. Convergence of the regularized problems is proved. Numerical tests confirm the theoretical results and demonstrate the efficiency of the proposed methodology.

Keywords: State constrained optimal control problems, semi-smooth Newton methods, primal dual active set strategy, superlinear convergence.

1 Introduction

This paper is aimed at describing an approach for the numerical solution of optimal control problems with point-wise state constraints. The objective consists in obtaining a method that is as close as possible to a Newton scheme, with the property that super-linear rate of convergence of the iterates can be observed numerically and proved analytically. Two obstacles need to be overcome. First, due to the constraints, the underlying optimization problem in non-differentiable in the Frechet sense, and secondly as a consequence of the specific nature of state constraints, the infinite dimensional variables which describe the optimality system experience very little regularity. The Lagrange multiplier associated to the state constrained, for example, is only a measure, in general, and the adjoint state is typically only in L^2 (solution tres faible). The first of these two difficulties will be overcome by utilising results from semi-smooth Newton methods in infinite dimensional spaces. The second is approached by utilizing a regularization technique.

State constrained have presented a challenge for some time. Earlier work focused on the derivation of first order optimality conditions. From among the many contributions we mention [AR1] [AR2] [B1] [BC] [BK] [C] [T]. More recently second order necessary as well as sufficient optimality conditions for certain classes optimal control problems subject to elliptic partial differential equations were investigated in [CT] and [CTU]. Finite element approximations of the infinite dimensional optimality systems are considered in [AM] and in [TT] parabolic optimal control problems, for example. The literature on numerical methods for the treatment of state-constrained optimal control problems is less rich. In [B1] and [B2] Lagrangian and augmented Lagrangian methods are analysed for state-constrained optimal control problems. The Lagrangian formulation is utilized for decoupling the state equation. Within the resulting saddle point problems the point-wise constraints, however, remain as explicit constraints. A significantly different approach towards numerical realization of state constrained optimal control problems was followed in [HR], where level set methods are employed to determine the interface between active and inactive sets. - Of course, one can also take the point of view that after discretization of the optimality condition, see (1.2) below, one arrives at a finite dimensional complementarity problem. Such problems have been intensively studied, see [LPR], for example. Such an approach misses important features, however, as for example, the regularity of Lagrange multipliers and its consequences, and smoothing effects or

lack thereof of the partial differential equation. These properties significantly influence the behavior of numerical algorithms.

In our work we focus on the treatment of the constraints as infinite dimensional inequality conditions. The numerical algorithms are related to those already utilized in [BHHK]. There, however, no attempt towards a convergence analysis was made. The latter, in turn, suggests a modification of the algorithm utilized in [BHHK]. This modification results in a significant speed-up of the algorithm that we propose in this paper over the previous one in [BHHK]. Both algorithms are of iterative nature, but while the earlier one would typically only alter the active set along the interface between active and inactive sets, the new one has the capability of making changes on larger patches. The price for this advantage consists in the necessity to utilize a tuning parameter which determines the influence of the regularization procedure.

Within this paper we shall not aim at generality but rather we consider a model problem. Generalization to more complex cost-functionals, and differential equations are possible. We shall focus on the following problem:

$$
\left\{
\begin{array}{l}
\min J(y,u) = \frac{1}{2}|y-z|_{L^2}^2 + \frac{\alpha}{2}|u|_{L^2}^2 \\
\\
\text{subject to} \\
\\
-\Delta y = u \ \text{ in } \ \Omega, \\
\\
y = 0 \text{ on } \partial\Omega, \\
\\
y \leq \psi \ \text{ a.e. in } \ \Omega \\
\\
(y,u) \in H_0^1(\Omega) \times L^2(\Omega),
\end{array}
\right. \tag{1.1}
$$

where Ω is a bounded domain in $\mathbb{R}^d$, $d \in \{1,2,3\}$, with a $C^{1,1}$ boundary if $d = 2$ or 3. Further we assume that $z \in L^2(\Omega)$, $\psi \in C(\Omega)$, $\psi > 0$ on $\partial\Omega$ and $\alpha > 0$. The norm on $L^2(\Omega)$ is denoted by $|\cdot|_{L^2}$ or alternatively by $|\cdot|$, and the $L^2(\Omega)$-inner product by $(\cdot,\cdot)$. It will be convenient to set $\mathcal{W} = H_0^1(\Omega) \cap H^2(\Omega)$. Due to the regularity requirements every solutions to $-\Delta y = u$, with $u \in L^2(\Omega)$ and $y = 0$ on $\partial\Omega$ satisfies $y \in \mathcal{W} \subset C(\overline{\Omega})$. It is standard to argue the existence of a solution $(y^*, u^*) \in \mathcal{W} \times L^2(\Omega)$ to (P). Let $\langle \cdot, \cdot \rangle_{C^*,C}$ denote the duality pairing between $C(\overline{\Omega})$ and its topological dual

$C^*(\Omega)$. The proof to the following characterization of the solution is found in [BK], for example.

Proposition 1.1 *The pair $(y^*, u^*) \in \mathcal{W} \times L^2(\Omega)$ is a solution to (P) if and only if there exists $p^* \in L^2(\Omega)$ and $\lambda^* \in C^*(\Omega)$ such that*

$$-\Delta y^* = u^* \quad \text{in } \Omega,\ y^* = 0 \quad \text{on } \partial\Omega,$$
$$(p^*, -\Delta y) + \langle \lambda^*, y \rangle_{C^*,C} = (z - y^*, y) \quad \text{for all } y \in \mathcal{W}$$

$$\alpha u^* = p^*$$

$$y^* \leq \psi,\ \langle \lambda^*, y^* - \psi \rangle_{C^*,C} = 0,$$

$$\langle \lambda^*, y \rangle_{C^*,C} \geq 0, \quad \text{for all } y \in C(\Omega) \text{ with } y \geq 0.$$

In Section 2 we shall present a regularized optimality system and describe an algorithm for its solution. Local super-linear as well as global convergence for a certain range of regularization parameters will be proved. Section 3 is devoted to the convergence analysis with respect to the regularization parameter. The theoretical results are confirmed by numerical tests which are presented in Section 4. The nature of the algorithm changes with the size of regularization parameter. As the size of the parameter value is increased the widths of the sets which change from one iteration to the next decreases. Moreover for moderate values of the regularization parameter the behavior of the iterates is monotone. This is not the case for large parameter values. These different properties suggest to use the regularization parameter within a continuation procedure.

2 Semi-smooth Newton algorithm

In this section we describe and analyze a semi-smooth Newton algorithm for a regularized approximation to the optimality condition expressed in Proposition (1.1). To motivate the procedure note that formally the optimality system can be expressed as

$$\text{(2.1)} \qquad \begin{cases} \alpha A y = p \\ \\ Ap + \lambda = z - y \\ \\ \lambda = \max(0, \lambda + y - \psi), \end{cases}$$

where A denotes the negative Laplacian with Dirichlet boundary conditions. Eliminating p from these equations we find

$$\text{(2.2)} \qquad \begin{cases} (I + \alpha A^2) y = z - \lambda \\ \\ \lambda = \max(0, \lambda + y - \psi), \end{cases}$$

or

$$\text{(2.3)} \qquad \lambda = \max(0, \lambda + B(z - \lambda) - \psi)$$

where $B = (I + \alpha A^2)^{-1}$. We note that λ appears twice inside the max-operation, once with and once without smoothing operation. The latter is responsible for the fact that semi-smooth Newton techniques cannot be applied directly [HIK]. The structure of (2.3) distinguishes state constrained from control constrained problems, since for the latter the optimality system can be reformulated in such a way that inside the max-operation the relevant variable only appears under a smoothing operation. To circumvent this difficulty we replace $\lambda = \max(0, \lambda + y - \psi)$ in (2.1) by the two-parameter family of approximations [IK1, IK2]

$$\text{(2.4)} \qquad \lambda = \max(0, \bar{\lambda} + \gamma(y - \psi)),$$

where $\gamma > 0$ and $\bar{\lambda} \in L^2(\Omega)$. The role of $\bar{\lambda}$ will become evident in Section 4.

Algorithm

(i) Choose $y_0 \in L^2(\Omega)$, set $k = 0$.

(ii) Set $\mathcal{A}_{k+1} = \{x \colon (\bar{\lambda} + \gamma(y_k - \psi))(x) > 0\}$, $\mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}$.

(iii) Solve for $(y_{k+1}, p_{k+1}) \in \mathcal{W} \times \mathcal{W}$.

$$\begin{aligned} &\alpha\, Ay = p \\ &Ap + y + (\bar{\lambda} + \gamma(y - \psi))\chi_{\mathcal{A}_{k+1}} = z \end{aligned}$$

(iv) Set

$$\lambda_{k+1} = \begin{cases} 0 & \text{on} \quad \mathcal{I}_{k+1} \\ \\ \bar{\lambda} + \gamma(y_{k+1} - \psi) & \text{on} \quad \mathcal{A}_{k+1}. \end{cases}$$

(v) Stop or $k = k + 1$, and goto (ii).

Above $\chi_{\mathcal{A}_{k+1}}$ denotes the characteristic function of the set $\mathcal{A}_{k+1}$. Observe that (iii) is the necessary and sufficient optimality condition to the unconstrained problem

$$\begin{cases} \min\ J(y,u) + \frac{1}{2\gamma} \displaystyle\int_{\mathcal{A}_{k+1}} |\bar{\lambda} + \gamma(y - \psi)|^2 dx \\ \\ \text{subject to} \\ \\ -\Delta y = u \ \text{ in } \ \Omega,\ y = 0 \ \text{ on } \ \partial\Omega. \end{cases} \tag{2.5}$$

This problem clearly admits a unique solution (y_{k+1}, u_{k+1}), and (y_{k+1}, p_{k+1}), with $p_{k+1} = \alpha u_{k+1}$, gives the solution to the system in step (iii) of the Algorithm. In the remainder of this section we argue convergence of the iterates (y_k, p_k, λ_k) to the solution $(y_\gamma, p_\gamma, \lambda_\gamma)$ of

$$\begin{cases} \alpha\, Ay = p \\ \\ Ap + \lambda = z - y \\ \\ \lambda = \max(0, \bar{\lambda} + \gamma(y - \psi)). \end{cases} \tag{2.6}$$

Here we suppress the dependence of $(y_\gamma, p_\gamma, \lambda_\gamma)$ on $\bar{\lambda}$.

Lemma 2.1 *For every $\gamma > 0$ there exists a unique solution $(y_\gamma, p_\gamma, \lambda_\gamma) \in \mathcal{W} \times \mathcal{W} \times L^2(\Omega)$ to (2.6).*

Proof. Consider the operator $B\colon \mathcal{W} \to \mathcal{W}^*$ given by

$$B = \alpha\, A^2 + I + \max(0, \bar{\lambda} + \gamma(\cdot - \psi)).$$

Clearly $\alpha\, A^2$ is maximal monotone and $B - \alpha\, A^2$ is continuous and monotone, and thus B is maximal monotone as well. Moreover

$$(By - B\bar{y}, y - \bar{y})_{\mathcal{W}^*,\mathcal{W}} \geq \alpha |A(y - \bar{y})|^2 + |y - \bar{y}|^2 \tag{2.7}$$

for all $y, \bar{y} \in \mathcal{W}$. Hence B is coercive and consequently surjective. Thus $By = z$ admits a solution $y \in \mathcal{W}$. By (2.7) it is unique. Set $p = \alpha\, Ay \in L^2(\Omega)$, $\lambda = \max(0, \bar{\lambda} + \gamma(y - \psi)) \in L^2(\Omega)$ and note that $By = A\, p + y + \lambda = z$. This equation holds in $\mathcal{W}^*$. But $z - y - \lambda \in L^2(\Omega)$ and hence $p \in \mathcal{W}$. Thus $(y, p, \lambda) \in \mathcal{W} \times \mathcal{W} \times L^2(\Omega)$ is the desired unique solution to (2.6). For the relevant facts on monotone operators we refer to [BP], Chapter 1, for example. $\square$

Proposition 2.1 *If $\mathcal{A}_{k+1} = \mathcal{A}_k$, then (y_k, p_k, λ_k) is the solution to (2.6).*

Proof. Since, for given $\mathcal{A}_{k+1}$, the solution to (iii) is unique we have $(y_k, p_k) =$

(y_{k+1}, p_{k+1}). By assumption $\mathcal{A}_k = \{x\colon (\bar{\lambda} + \gamma(y_k - \psi)(x) > 0\}$. Consequently (y_k, p_k, λ_k) satisfies (2.6). $\square$

Theorem 2.1 *If y_0, ψ and $\bar{y} \in L^p(\Omega)$ for some $p > 2$ and $|y_0 - y_\gamma|_{L^p}$ is sufficiently small, then (y_k, p_k, λ_k) converges to $(y_\gamma, p_\gamma, \lambda_\gamma)$ super-linearly in $\mathcal{W} \times \mathcal{W} \times L^2(\Omega)$ as $k \to \infty$.*

Proof. (i) Observe that the iterate (y_{k+1}, p_{k+1}) can be expressed as

$$\begin{cases} \alpha\, Ay_{k+1} = p_{k+1} \\ \\ A\, p_{k+1} + y_{k+1} + \max(0, \bar{y} + \gamma(y_k - \psi)) \\ \\ +\gamma\, G(\bar{y} + \gamma(y_k - \psi))(y_{k+1} - y_k) = z, \end{cases} \tag{2.8}$$

where, for $g \in L^p(\Omega)$,

$$G(g)(x) = \begin{cases} 0 & \text{if} \quad g(x) \leq 0 \\ \\ g(x) & \text{if} \quad g(x) > 0. \end{cases}$$

It is wellknown, see e.g. [HIK, U], that for every $g \in L^p(\Omega)$ we have

$$|\max(0, g + h) - \max(0, g) - G(g + h)h|_{L^2} = o(|h|_{L^p}), \tag{2.9}$$

as $|h|_{L^p} \to 0$.

(ii) Let us set

$$\delta y = y_{k+1} - y_\gamma,\ \delta p = p_{k+1} - p_\gamma,\ \delta\lambda = \lambda_{k+1} - \lambda_\gamma.$$

From (2.5) and (2.8) we have

$$\text{(2.10)} \qquad \begin{cases} \alpha\, A\, \delta y = \delta p \\ \\ A\, \delta p + \delta\lambda + \delta y = 0 \\ \\ \delta\lambda = \gamma G(\bar\lambda + \gamma(y_k - \psi))\delta y + R, \end{cases}$$

where

$$R = \max(0, \bar\lambda + \gamma(y_\gamma + (y_k - y_\gamma) - \psi)) - \max(0, \bar\lambda + \gamma(y_\gamma - \psi))$$

$$-\gamma\, G(\bar\lambda + \gamma(y_\gamma + (y_k - y_\gamma) - \psi)(y_k - y_\gamma).$$

Taking the inner product with $A\,\delta y$ in the first equation of (2.10) we find

$$\alpha|A\,\delta y|^2 + |\delta y|^2 + \gamma(G(\bar\lambda + \gamma(y_k - \psi))\delta y, \delta y) = -(R, \delta y).$$

From (2.9) we have $|R|_{L^2(\Omega)} = o(|y_k - y_\gamma|_{L^p(\Omega)})$ and hence

$$|y_{k+1} - y_\gamma|_{\mathcal{W}} = o(|y_k - y_\gamma|_{L^p}).$$

From the last equation in (2.10) it follows that

$$|\lambda_{k+1} - \lambda_\gamma|_{L^2(\Omega)} = o(|y_k - y_\gamma|_{L^p}),$$

and finally the second equation leads to

$$|p_{k+1} - p_\gamma|_{\mathcal{W}} = o(|y_k - y_\gamma|_{L^p}). \qquad \square$$

We now address the convergence of the Algorithm to the solution of (2.6) from arbitrary initial conditions and an appropriate range of values for γ.

Theorem 2.2 *If* $\frac{\gamma}{\alpha}\|A^{-1}\|^2_{\mathcal{L}(L^2(\Omega))} < 1$ *then* $\lim_{k\to\infty}(y_k, p_k, \lambda_k) = (y_\gamma, p_\gamma, \lambda_\gamma)$ *in* $\mathcal{W} \times \mathcal{W} \times L^2(\Omega)$.

Proof. Let $\delta y = y_{k+1} - y_k$, $\delta\lambda = \lambda_{k+1} - \lambda_k$, and $\delta p = p_{k+1} - p_k$. From step (iii) of the Algorithm we have

$$\text{(2.11)} \qquad \begin{cases} \alpha\, A\, \delta y = \delta p \\ \\ A\, \delta p + \delta y + \chi_{\mathcal{A}_{k+1}}\, \delta y + R = 0, \end{cases}$$

where $\chi_{\mathcal{A}_{k+1}}$ is the characteristic function of the set $\mathcal{A}_{k+1}$ and

$$\text{(2.12)} \qquad R = \begin{cases} 0 & \text{on} \quad (\mathcal{A}_{k+1} \cap \mathcal{A}_k) \cup (\mathcal{I}_{k+1} \cap \mathcal{I}_k) \\ \\ \bar{\lambda} + \gamma(y_k - \psi) & \text{on} \quad \mathcal{A}_{k+1} \cap \mathcal{I}_k \\ \\ -\bar{\lambda} - \gamma(y_k - \psi) & \text{on} \quad \mathcal{I}_{k+1} \cap \mathcal{A}_k. \end{cases}$$

From (2.11) we deduce that

$$(\alpha A^2 + q)\, \delta y = -R,$$

where $q = 1 + \chi_{\mathcal{A}_{k+1}}$, which implies

$$\alpha |A\, \delta y|^2 + |q\, \delta y^2| = -(R, y)$$

and hence

$$\text{(2.13)} \qquad |A\, \delta y| \;\le\; \tfrac{1}{\alpha} \|A^{-1}\|_{\mathcal{L}(L^2(\Omega))}\, |R|.$$

On $\mathcal{I}_k$ we have $\bar{\lambda} + \gamma(y_{k-1} - \psi) \le 0$ and hence

$$\bar{\lambda} + \gamma(y_k - \psi) \le \gamma(y_k - y_{k-1}) \text{ on } \mathcal{A}_{k+1} \cap \mathcal{I}_k.$$

Similarly $\bar{\lambda} + \gamma(y_{k+1} - \psi) \ge 0$ on $\mathcal{A}_k$ and hence

$$-\bar{\lambda} - \gamma(y_k - \psi) \le \gamma(y_{k-1} - y_k) \text{ on } \mathcal{I}_{k+1} \cap \mathcal{A}_k.$$

From (2.12), (2.13) we find

$$|A\, \delta y| \;\le\; \tfrac{\gamma}{\alpha} \|A^{-1}\|_{\mathcal{L}(L^2(\Omega))}\, |y_k - y_{k-1}|,$$

and hence

$$|A(y_{k+1} - y_k)| \;\le\; \tfrac{\gamma}{\alpha} \|A^{-1}\|^2_{\mathcal{L}(L^2(\Omega))}\, |A(y_k - y_{k-1})|.$$

Since $\frac{\gamma}{\alpha}\|A^{-1}\|^2_{\mathcal{L}(L^2(\Omega))} < 1$ by assumption, it follows that $\{y_k\}$ is a Cauchy sequence in $\mathcal{W}$. Hence there exists $\hat{y} \in \mathcal{W}$ such that $\lim_{k\to\infty} y_k = \hat{y}$ in $\mathcal{W}$. Let us define $\hat{\lambda} = \max(0, \bar{\lambda} + \gamma(\hat{y} - \psi))$ and set

$$\hat{\mathcal{A}} = \{x\colon (\bar{\lambda} + \gamma(\hat{y} - \psi))(x) > 0\},\ \hat{\mathcal{I}} = \Omega \setminus \hat{\mathcal{A}}.$$

Observe that

$$\lambda_{k+1} - \hat{\lambda} = \begin{cases} \gamma(y_{k+1} - \hat{y}) & \text{on} \quad \mathcal{A}_{k+1} \cap \hat{\mathcal{A}} \\ \gamma(y_{k+1} - \hat{y}) + \hat{\lambda} + \gamma(\hat{y} - \psi) & \text{on} \quad \mathcal{A}_{k+1} \cap \hat{\mathcal{I}} \\ -(\bar{\lambda} + \gamma(\hat{y} - \psi)) & \text{on} \quad \mathcal{I}_{k+1} \cap \hat{\mathcal{A}} \\ 0 & \text{on} \quad \mathcal{I}_{k+1} \cap \hat{\mathcal{I}}, \end{cases}$$

which implies

$$\begin{aligned} |\lambda_{k+1} - \hat{\lambda}|_{L^2} \quad &\leq \gamma |y_{k+1} - \hat{y}|_{L^2} + \left(\int_\Omega |\bar{\lambda} + \gamma(\hat{\lambda} - \psi)|^2 \chi_{\mathcal{A}_{k+1} \cap \hat{\mathcal{I}}}\, dx \right)^{1/2} \\ &\quad + \left(\int_\Omega |\bar{\lambda} + \gamma(\hat{\lambda} - \psi)|^2 \chi_{\hat{\mathcal{A}} \cap \mathcal{I}_{k+1}}\, dx \right)^{1/2}. \end{aligned}$$

Lebesgue's bounded convergence theorem implies that $\lim_{k\to\infty} \lambda_k = \hat{\lambda}$ in $L^2(\Omega)$. Taking the limit with respect to k in

$$Ap_{k+1} + y_{k+1} + \lambda_{k+1} = z,$$

we find that $\{p_k\}_{k=1}^\infty$ converges in $\mathcal{W}$ and the limit $\hat{p}$ satisfies

$$A\hat{p} + \hat{y} + \hat{\lambda} = z.$$

By uniqueness of the solution to (2.6) we have $(\hat{y}, \hat{p}, \hat{\lambda}) = (y_\gamma, p_\gamma, \lambda_\gamma)$. □

3 Convergence of regularized problems

In this section convergence of the solutions to (2.6) as $\gamma \to \infty$ is analyzed. Our first result holds for arbitrary choices of $\bar{\lambda}$.

Theorem 3.1 *The solutions* $\{(y_\gamma, p_\gamma, \lambda_\gamma)\}_{\gamma>0}$ *to the regularized problem (2.6) converge to* (y^*, p^*, λ^*) *in the sense that* $y_\gamma \to y^*$ *strongly in* $\mathcal{W}$, $p_\gamma \to p^*$ *strongly in* $L^2(\Omega)$ *and* $\lambda_\gamma \rightharpoonup \lambda^*$ *weakly in* $\mathcal{W}^*$ *as* $\gamma \to \infty$.

Proof. As a preparatority step let us note that

$$\begin{aligned}(\lambda_\gamma, y_\gamma - y^*) &= \tfrac{1}{\gamma}(\lambda_\gamma, \bar\lambda + \gamma(y_\gamma - \psi) + \gamma(\psi - y^*) - \bar\lambda) \\ &\geq \tfrac{1}{\gamma}|\lambda_\gamma|^2 - \tfrac{1}{\gamma}(\lambda_\gamma, \bar\lambda),\end{aligned} \tag{3.1}$$

where we used that $\lambda_\gamma \geq 0$ and $\psi \geq y^*$. Consequently

$$(\lambda_\gamma, y_\gamma - y^*) \;\geq\; -\tfrac{1}{4\gamma}|\bar\lambda|^2, \tag{3.2}$$

for every $\gamma > 0$. From Proposition 1.1 and (2.6) we conclude that

$$\alpha(A\, y_\gamma, A(y_\gamma - y^*)) = (A\, p_\gamma, y_\gamma - y^*) = (z - y_\gamma - \lambda_\gamma, y_\gamma - y^*),$$

and therefore by (3.1)

$$\begin{aligned}&\alpha|A\, y_\gamma|^2 + |y_\gamma|^2 + \tfrac{1}{\gamma}|\lambda_\gamma|^2 \\ &\leq \alpha(A\, y_\gamma, A\, y^*) + (y_\gamma, y^*) + (z, y_\gamma - y^*) + \tfrac{1}{\gamma}(\lambda_\gamma, \bar\lambda).\end{aligned}$$

This inequality implies that

$$\left\{|y_\gamma|^2_{\mathcal{W}} \;+\; \tfrac{1}{\gamma}|\lambda_\gamma|^2\right\}_{\gamma\geq 1} \text{ is uniformly bounded,} \tag{3.3}$$

and from (2.6)

$$\{|p_\gamma|\}_{\gamma\geq 1} \;\text{ and }\; \{|\lambda_\gamma|_{\mathcal{W}^*}\}_{\gamma\geq 1} \;\text{ are uniformly bounded ,} \tag{3.4}$$

as well. From (3.3), (3.4) we conclude the existence of $(\hat y, \hat p, \hat\lambda) \in \mathcal{W}\times L^2(\Omega)\times \mathcal{W}^*$ and of a subsequence, again denoted by $(y_\gamma, p_\gamma, \lambda_\gamma)$, such that

$$(y_\gamma, p_\gamma, \lambda_\gamma) \rightharpoonup (\hat y, \hat p, \hat\lambda) \;\text{ weakly in }\; \mathcal{W}\times L^2(\Omega)\times \mathcal{W}^*.$$

Since $\lambda_\gamma \geq 0$ for every $\gamma > 0$ we have

$$\langle \hat\lambda, v\rangle_{\mathcal{W}^*,\mathcal{W}} \geq 0 \;\text{ for every }\; v \in \mathcal{W}.$$

Note that

$$\tfrac{1}{\gamma^2}|\lambda_\gamma|_{L^2}^2 = |\max(0,\ \tfrac{1}{\gamma}|\bar\lambda| + y_\gamma - \psi)|^2,$$

and therefore by (3.3)

$$\lim_{\gamma\to\infty} |\max(0,\ \tfrac{1}{\gamma}|\bar\lambda| + y_\gamma - \psi)|^2 = 0.$$

By Fatou's lemma this implies that $\max(0, \hat y - \psi) = 0$ and thus

$$\hat y \le \psi \ \text{ in } \ \Omega.$$

By Proposition 1.1 and (2.6)

$$\alpha(A(y_\gamma - y^*), A(y_\gamma - y^*)) = (p_\gamma - p^*, A(y_\gamma - y^*))$$

$$= -|y_\gamma - y^*|^2 + \langle \lambda^* - \lambda_\gamma, y_\gamma - y^*\rangle_{\mathcal{W}^*,\mathcal{W}},$$

which, using (3.2) implies

$$\alpha|A(y_\gamma - y^*)|^2 + |y_\gamma - y^*|^2 \le \langle \lambda^*, y_\gamma - y^*\rangle_{\mathcal{W}^*,\mathcal{W}} + \tfrac{1}{4\gamma}|\bar\lambda|^2.$$

Taking the limit $\gamma \to \infty$ we find

$$0 \le \overline{\lim}\ \alpha|A(y_\gamma - y^*)|^2 + |y_\gamma - y^*|^2 \le \langle \lambda^*, \bar y - y^*\rangle_{\mathcal{W}^*,\mathcal{W}}$$

$$= \langle \lambda^*, \hat y - \psi\rangle \le 0.$$

Consequently $\lim_{\gamma\to\infty} y_\gamma = y^*$ in $\mathcal{W}$ and $\lim_{\gamma\to\infty} p_\gamma = p^*$ in $L^2(\Omega)$, by (2.6). Taking the limit $\gamma \to \infty$ in

$$(p_\gamma, A\,v) + \langle \lambda_\gamma, v\rangle_{\mathcal{W}^*,\mathcal{W}} = (z - y_\gamma, v) \ \text{ for all } \ v \in \mathcal{W}$$

implies that

$$(p^*, A\,v) + \langle \hat\lambda, v\rangle_{\mathcal{W}^*,\mathcal{W}} = (z - y^*, v) \ \text{ for all } \ v \in \mathcal{W}.$$

Since this equation is also satisfied with $\hat\lambda$ replaced λ^* we have $\hat\lambda = \lambda^*$. Finally, since the solution to the optimality system given in Proposition 1.1 is unique the whole sequence $(y_\gamma, p_\gamma, \lambda_\gamma)$ converges to (y^*, p^*, λ^*). □

4 Numerical example

Utilizing a finite difference discretization with a five-point stencil approximation to the Laplace operator, the Algorithm was tested for several examples. The Algorithm was initialized with the solution to the unconstrained problem. Super-linear convergence could be observed for arbitrary fixed choices of γ. For relatively small values of γ a large number of grid points is moved from active to inactive an vice-versa from one iteration to the next. For large γ, on the other hand, one iteration of the algorithm tends to have an effect only along the current active/inactive set interface. This suggests to utilize a continuation procedure with respect to γ. Here we present numerical findings for an example already treated in [BHHK], where Ω is the unit square, $z(x, y) = \sin(2\pi xy)$ and $\psi = 0.1$. In [BHHK] a wide range of α values was tested. Among these values $\alpha = 0.001$ required most iterations for an active set algorithm without γ - continuation procedure. In Table 1 we depict the results with step-size $\frac{1}{60}$ and $\bar{\lambda} = 0$. The number of active mesh-points and the required number of iterations are shown as a function of γ. The Algorithm was stopped when two consecutive active sets coincide, i.e. the exact discretized solution was computed, see Proposition 2.1.

γ	10^3	10^4	10^5	10^6	10^8	10^9	10^{10}
iter	10	17	27	30	30	31	31
active	791	667	606	587	577	575	575

Table 1:

In Table 1 for each value of γ the Algorithm was initialized by the optimal control to the unconstrained problem. In Table 2 for $\gamma > 10^3$ the Algorithm was initialized with the result obtained before with the smaller γ - value.

γ	10^3	10^6	10^9	
iter	10	6	3	Σ =19

Table 2:

In this as well as in other examples we found that a reduction of at least

30 percent can be obtained by the continuation procedure. To document super-linear convergence which can be observed numerically we computed the ratios

$$r_k = \frac{|\Delta(y_{k+1}^h - y_*^h)|}{|\Delta(y_k^h - y_*^h)|},$$

where y_*^h denotes the solution to the descritized problem (2.6) for a fixed γ - value. We present the selected data corresponding to $\gamma = 10^4$ in Table 3.

i	9	10	11	12	13	14	15	16
r_i	.7049	.6468	.5475	.4554	.4131	.2780	.0562	0.0

Table 3:

In Table 4 we present results for another example with $\psi = (x - .5) + (y - .5) - .1$ and $\alpha = .1$ and all other specifications as before. For $\gamma \leq 10^4$ the behavior of the iterates is monotone in the sense that the sequences $\{y_k\}_{k=1}^{\infty}$, $\{\lambda_k\}_{k=1}^{\infty}$, $\{p_k\}_{k=1}^{\infty}$ and $\{u_k\}_{k=1}^{\infty}$ are monotonically decreasing and $\mathcal{I}_k \subset \mathcal{I}_{k+1}$ for every $k \geq 1$. For $\gamma > 10^2$ the Algorithm was initialized with the result obtained with the smaller γ - value.

γ	10^2	10^3	10^4	10^{10}
iter	4	5	5	11
active	857	416	183	34

Table 4:

Augmented Lagrangian method. System (2.6) with special choices for $\bar{\lambda}$ is precisely the system that arises as auxiliary problem in the classical first order augmented Lagrangian method. The Algorithm of Section 2 is an efficient method to solve such systems. This suggests to investigate the augmented Lagrangian method as an alternative to the continuation method with respect to γ to solve the original system (2.1). We specify the augmented Lagrangian method next:

ALM

(a) Choose $\gamma > 0$, λ_0; set $k = 0$.

(b) Solve for (y_k, p_k, λ_k):
$\alpha\, A\, y_k = p_k$
$A\, p_k + \lambda_k + y_k = z$
$\lambda_k = \max(0, \lambda_{k-1} + \gamma(y_k - \psi))$.

(c) $k = k + 1$ and goto (1).

Note that γ is not increased in ALM, rather λ is updated. To solve the system in (b) the semi-smooth Newton Algorithm of Section 2 with $\bar{\lambda} = \lambda_{k-1}$ is used. Below we present our numerical findings with ALM for the test problem presented at the beginning of this section. We carried out tests for $\lambda_0 = 0$ and $\lambda_0 = \max(0, z - \psi - \alpha\, A^2\psi)$. Since $\lambda_0 = 0$ gives better results we only report on them. The question arises concerning the precision to which the system in (b) is solved, i.e. how many iterates of the semi-smooth Newton Algorithm should be carried out. For the results below we solved the system exactly for $k = 0$. For $k \geq 1$ only one step of the semi-smooth Newton Algorithm was performed before λ was updated. We chose $h = \frac{1}{60}$, $\alpha = 0.001$ and $\psi = 0.1$. For $\gamma = 10^5$ we required 55 system solves to reach the exact solution, for $\gamma = 10^6$ it took 37 system solves. We tested several alternatives to the above procedure without success to decrease the number of system solves significantly. It appears that we can safely conclude that for the class of problems under consideration the continuation procedure with respect to γ is numerically more efficient than the augmented Lagrangian method.

References

[AM] W. Alt and U. Mackenroth: *Convergence of finite element approximations to state constrained convex parabolic boundary control problems*, SIAM J. Control and Optim. 22(1991), 83-98.

[AR1] N. Arada and J.P. Raymond: *State-constrained relaxed problems for semilinear elliptic equations*, J. Math. Anal. and Appl. 223(1998), 248-271.

[AR2] N. Arada and J.P. Raymond: *Optimal control problems with mixed control-state constraints*, preprint, Universite Paul Sabatier, Toulouse.

[B1] M. Bergounioux: *On boundary state constrained control problems*, Numer. Funct. Anal. and Optimiz. 14(1993), 515-543.

[B2] M. Bergounioux: *Augmented Lagrangian method for distributed optimal control problems with state constraints*, J. Optim.Theory and Appl. 78 (1993), 493-521.

[BC] F.J. Bonnans and E. Casas: *An extension of Pontryagin's maximum principle for state-constrained optimal control of semilinear elliptic equations and variational inequalities*, SIAM J. Control and Optim. 33(1995), 274-298.

[BHHK] M. Bergounioux, M. Haddou, M.Hintermüller and K. Kunisch: *A comparison of a Moreau-Yosida based active set strategy and interior point problems for constrained optimal control problems*, SIAM J. Optim 11(2000), 495-521.

[BK] M. Bergounioux and K. Kunisch: *On the structure of Lagrange multipliers for state-constrained optimal control problems*, System and Control Letters, to appear.

[BP] V. Barbu and Th. Percupanu : *Convexity and Optimization in Banach Spaces*, D. Reidel Publ. Comp., Dodrecht, 1986.

[C] E. Casas: *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control and Optim. 31(1993), 993-1006.

[CT] E. Casas and F. Tröltzsch: *Second order necessary optimality conditions for some state-constrained control problems of semilinear elliptic equations*, Appl. Math. Optimization, to appear.

[CTU] E. Casas, F. Tröltzsch and A. Unger: *Second order sufficient optimality conditions for nonlinear elliptic control problems*, J. Analysis and its Applications 15(1996), 687-707-

[HIK] M. Hintermüller, K. Ito and K. Kunisch: *The Primal-Dual Active Set Strategy as Semi-Smooth Newton Method*, submitted.

[HR] M. Hintermüller and W. Ring: *A level set approach for the solution of a state-constrained optimal control problem*, submitted.

[IK1] K. Ito and K. Kunisch: *Augmented Lagrangian Methods for Nonsmooth Convex Optimization in Hilbert Spaces*, Nonlinear Analysis, Theory, Methods and Applications, 41(2000), 573–589.

[IK2] K. Ito and K. Kunisch: *Semi–smooth Newton methods for variational inequalities of the first kind*, Mathematical Modelling and Numerical Analysis, to appear.

[LPR] Z.Q. Luo, J.S. Pang and D.Ralph: *Mathematical Programs With Equilibrium Constraints*, Cambridge University Press, 1966

[T] D. Tiba: *Optimal control for parabolic control problems and applications*, Lecture Notes in Mathematics 1459, Springer-Verlag, Berlin, 1990.

[TT] D. Tiba and F. Tröltzsch: *Error estimates for the discetization of state constrained convex control problems*, Numer. Funct. Anal. and Optimiz. 17(1996), 1005-1028.

[U] M. Ulbrich, *Semi-smooth Newton methods for operator equations in function spaces*, to appear in SIAM J. Optimization.

PATH-FOLLOWING METHODS FOR A CLASS OF CONSTRAINED MINIMIZATION PROBLEMS IN FUNCTION SPACE[◇]

M. HINTERMÜLLER[*,†] AND K. KUNISCH[†]

Abstract. Path-following methods for primal-dual active set strategies requiring a regularization parameter are introduced. Existence of a primal-dual path and its differentiability properties are analyzed. Monotonicity and convexity of the primal-dual path value function are investigated. Both feasible and infeasible approximations are considered. Numerical path following strategies are developed and their efficiency is demonstrated by means of examples.

1. Introduction

Primal-dual active set strategies or, in some cases equivalently, semi-smooth Newton methods, were proven to be efficient methods for solving constrained variational problems in function space [1, 9, 10, 11, 12, 13]. In certain cases regularization is required resulting in a family of approximating problems with more favorable properties than the original one, [12, 13]. In previous work [13] convergence, and in some cases rate of convergence, with respect to the regularization parameter was proved. In the numerical work the adaptation of these parameters was heuristic, however. The focus of the present investigation is on an efficient control of the regularization parameter in the primal-dual active set strategy for a class of constrained variational problems. To explain the involved issues we proceed mostly formally in this section

Date: May 10, 2005.

1991 *Mathematics Subject Classification.* 49M15,49M37,65K05,90C33.

Key words and phrases. semi-smooth Newton methods, path-following methods, active set strategy, primal-dual methods.

[*]Department of Computational and Applied Mathematics, Rice University, Houston, Texas.

[†]Institut für Mathematik, Karl-Franzens-Universität Graz, A-8010 Graz, Austria .

[◇]Research partially supported by the Fonds zur Förderung der wissenschaftlichen Forschung under SFB 03 ,,Optimierung und Kontrolle".

and consider the problem

$$\text{(1)} \qquad \begin{cases} \min \mathcal{J}(v) \quad \text{over } v \in X \\ \text{s.t. } G\,v \le \psi, \end{cases}$$

where $\mathcal{J}$ is a quadratic functional on a Hilbert space X, and $G\colon X \to Y$. It is assumed that $Y \subset L^2(\Omega)$ is a Hilbert lattice with ordering $\le$ induced by the natural ordering of $L^2(\Omega)$. We note that (1) subsumes problems of very different nature. For example, for the control constrained optimal control problem

$$\begin{cases} \min \frac{1}{2}|y-z|^2_{L^2} + \frac{\alpha}{2}|u|^2_{L^2} \\ \text{s.t. } -\Delta y = u \ \text{ in } \ \Omega,\ y = 0 \ \text{ on } \ \partial\Omega, \\ \qquad u \le \psi \text{ a.e. in } \Omega, \end{cases}$$

with Ω a bounded domain in $\mathbb{R}^n$, $z \in L^2(\Omega)$, $\alpha > 0$, one can use $y = (-\Delta)^{-1}u$, where Δ denotes the Laplacian with homogenous Dirichlet boundary conditions, and arrives at

$$\begin{cases} \min \frac{1}{2}|(-\Delta)^{-1}u - z|^2 + \frac{\alpha}{2}|u|^2 \\ \text{s.t. } u \le \psi \text{ a.e. in } \Omega, \end{cases}$$

which is clearly of the form (1). For $\mathcal{J}(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 dx - \int_\Omega f\,v$, $X = H^1_0(\Omega)$, and $G = I$ we obtain the classical obstacle problem. For state constrained control problems with $y \le \psi$ one has

$$\begin{cases} \min \frac{1}{2}|(-\Delta)^{-1}u - z|^2 + \frac{\alpha}{2}|u|^2 \\ \text{s.t. } (-\Delta)^{-1}u \le \psi \text{ a.e. in } \Omega, \end{cases}$$

which is also of the form (1). From the point of view of duality theory these three problems are very different. While it is straightforward to argue the existence of a Lagrange multiplier in $L^2(\Omega)$ for the control constrained optimal control problem, it is already more involved and requires additional assumptions to guarantee its existence in $L^2(\Omega)$ for obstacle problems, and for state constrained problems the Lagrange multiplier is only a measure. If we resort to a formal discussion, then in either of these cases we arrive at the optimality system of the form

$$\text{(2)} \qquad \begin{cases} \mathcal{J}'(v) + G^*\lambda = 0, \\ \lambda = \max(0, \lambda + c(G(v) - \psi)\,) \end{cases}$$

for any fixed $c > 0$. Here, G^* denotes the adjoint of G. The second equation in (2) is equivalent to $\lambda \ge 0$, $G(v) \le \psi$ and $\lambda(G(v) - \psi) = 0$.

Continuing formally, the primal-dual active set strategy determines the active set at iteration level k by means of

$$\mathcal{A}_{k+1} = \{x \in \Omega \colon \lambda_k(x) + c(G(v_k)(x) - \psi(x)\,) > 0\},$$

assigns the inactive set $\mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}$, and updates (v, λ) by means of

$$\text{(3)} \qquad \begin{cases} \mathcal{J}'(v_{k+1}) + G^*\lambda_{k+1} = 0, \\ \lambda_{k+1} = 0 \ \text{ on } \ \mathcal{I}_{k+1}, \ (G(v_{k+1}) - \psi)(x) = 0 \ \text{ for } \ x \in \mathcal{A}_{k+1}. \end{cases}$$

These auxiliary problems require special attention. For obstacle problems the constraint $v_{k+1} = \psi$ on $\mathcal{A}_{k+1}$ induces that the associated Lagrange multiplier λ_{k+1} is in general less regular than the Lagrange multiplier associated to $v \leq \psi$ for the original problem; see, *e.g.*, [13]. For problems with combined control and state constraints it may happen that due to the assignment on $\mathcal{I}_{k+1}$ and $\mathcal{A}_{k+1}$ (3) has no solution while the original problem does. For these reasons in, *e.g.*, [9, 12, 13] the second equation in (2) was regularized resulting in the family of equations

$$\text{(4)} \qquad \begin{cases} \mathcal{J}'(v) + G^*\lambda = 0, \\ \lambda = \max(0, \bar{\lambda} + \gamma(G(v) - \psi)), \end{cases}$$

where $\bar{\lambda}$ is fixed, possibly $\bar{\lambda} = 0$, and $\gamma \in \mathbb{R}^+$. In the above mentioned references it was shown that under appropriate conditions the solutions $(v_\gamma, \lambda_\gamma)$ to (4) exist, the quantity λ_γ enjoys extra regularity and $(v_\gamma, \lambda_\gamma)$ converge to the solution of (2) as $\gamma \to \infty^+$.

In previous numerical implementations the increase of γ to infinity was heuristic. As the system (4) becomes increasingly ill-conditioned as γ tends to ∞, in this paper a framework for a properly controlled increase of γ-values will be developed in order to cope with the conditioning problem. At the same time we aim at solving the auxiliary problems (3) only inexactly to keep the overall computational cost low. To this end we defined neighborhoods of the path which allow inexact solutions and which contract in a controlled way towards the path as the iteration proceeds. Our work is inspired by concepts from path-following methods in finite dimensional spaces [4, 5, 16, 18, 19]. We first guarantee the existence of a sufficiently smooth path $\gamma \to (v_\gamma, \lambda_\gamma)$, with $\gamma \in (0, \infty)$ in appropriately chosen function spaces. Once the path is available it can be used as the basis for updating-strategies of the path parameter. Given a current value γ_k, with associated primal and dual states $(v_{\gamma_k}, \lambda_{\gamma_k})$, the γ-update should be sufficiently large to make good progress towards satisfying the complementarity conditions. On

the other hand, since we are not solving the problems along the path exactly, we have to use safeguards against steps which would lead us too far off the path. Of course, these goals are impeded by the fact that the path is not available numerically. To overcome this difficulty we use qualitative properties of the value function, like monotonicity and convexity, which can be verified analytically. These suggest the introduction of model functions which will be shown to approximate very well the value functional along the path. We use these model functions for our updating strategies of γ. In the case of exact path following we can even prove convergence of the resulting strategy. In the present paper the program just described is carried out for a class of problems, corresponding to contact problems. State-constrained optimal control problems require a different approach that will be considered independently. As we shall see, the (infinite dimensional) parameter $\bar{\lambda}$ can be used to guarantee that the iterates of the primal variable are feasible. Further it turns out that the numerical behavior of infeasible approximations is superior to the feasible ones from the point of view of iteration numbers.

Interior point methods also require an additional parameter, which, however enters into (2) differently. For the problem under consideration here, the interior-point relaxation replaces the second equation in (2) by

$$\lambda(x)\,(\psi - G(v))(x) = \frac{1}{\gamma} \text{ for } x \in \Omega. \tag{5}$$

Path following interior point methods typically start strictly feasible, with iterates which are required to stay strictly feasible during the iterations while satisfying, or satisfying approximately, the first equation in (2) and (5). Path-following interior point methods have not received much attention for infinite dimensional problems yet. In fact, we are only aware of [17], were such methods are analyzed for optimal control problems related to ordinary differential equations. For the problem classes that we outlined at the beginning of this section, the primal-dual active set strategy proved to be an excellent competitor to interior point methods, as was demonstrated, for example, in [1] comparing these two methods.

This paper is organized as follows. Section 2 contains the precise problem formulation and the necessary background on the primal-dual active set strategy. The existence and regularity of the primal-dual path is discussed in Section 3. Properties of the primal-dual path value functional are analyzed in Section 4. Section 5 contains the derivation

of the proposed model functions for the primal-dual path value functional. Exact as well as inexact path-following algorithms are proposed in Section 6 and their numerical behavior is discussed there as well.

2. Problem statement, regularization and its motivation

We consider

$$\text{(P)} \qquad \begin{cases} \min \frac{1}{2}\, a(y,y) - (f,y) \quad \text{over} \quad y \in H_0^1(\Omega) \\ \text{s.t. } y \le \psi \end{cases}$$

where $f \in L^2(\Omega)$, $\psi \in H^1(\Omega)$, with $\psi_{|\partial\Omega} \ge 0$, where Ω is a bounded domain in $\mathbb{R}^n$ with Lipschitz continuous boundary $\partial\Omega$. Throughout $(\cdot,\cdot)$ denotes the standard $L_2(\Omega)$-inner product, and we assume that $a(\cdot,\cdot)$ is a bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$ satisfying

$$a(v,v) \ge \nu |v|^2_{H_0^1} \ \text{ and } \ a(w,z) \le \mu |w|_{H^1} |z|_{H^1} \tag{6}$$

for some $\nu > 0$, $\mu > 0$ independent of $v \in H_0^1(\Omega)$ and $w, z \in H^1(\Omega)$. Here and throughout we use $|v|_{H_0^1} = |\nabla v|_{L^2}$ for $v \in H_0^1(\Omega)$ which defines a norm on $H_0^1(\Omega)$ due to Friedrichs' inequality, and $|w|_{H^1} = (|w|^2_{L^2} + |\nabla w|^2_{L^2})^{1/2}$ denotes the standard H^1-norm; see, *e.g.*, [2]. Moreover let $A\colon H_0^1(\Omega) \to H^{-1}(\Omega)$ be defined by

$$a(v,w) = \langle A\, v, w \rangle_{H^{-1},H_0^1} \ \text{ for all } v, w \in H_0^1(\Omega).$$

It is well-known that (P) admits a unique solution $y^* \in H_0^1(\Omega)$ with associated Lagrange multiplier $\lambda^* = -Ay^* + f$, satisfying the optimality system

$$\begin{cases} a(y^*, v) + \langle \lambda^*, v \rangle_{H^{-1},H_0^1} = (f,v), \\ \langle \lambda^*, y^* - \psi \rangle_{H^{-1},H_0^1} = 0, y^* \le \psi, \langle \lambda^*, v \rangle \le 0 \ \text{ for all } \ v \le 0. \end{cases} \tag{7}$$

This also holds with $f \in H^{-1}(\Omega)$. Under well-known additional requirements on a, ψ and Ω, as for example

$$\begin{cases} a(v,w) = \int_\Omega (\sum a_{ij} v_{x_i} w_{x_j} + d\, v\, w), \ \text{ with } \ a_{ij} \in C^1(\bar\Omega), d \in L^\infty(\Omega), \\ d \ge 0, \ \psi \in H^2(\Omega), \partial\Omega \ \text{ is } \ C^{1,1} \ \text{ or } \ \Omega \ \text{ is a convex polyhedron,} \end{cases} \tag{8}$$

we have $(y^*, \lambda^*) \in H^2(\Omega) \times L^2(\Omega)$ and the optimality system can be expressed as

$$\begin{cases} A\, y^* + \lambda^* = f \ \text{ in } \quad L^2(\Omega), \\ \lambda^* = (\lambda^* + c(y^* - \psi))^+, \quad \text{for some } \ c > 0, \end{cases} \tag{9}$$

where $(v)^+ = \max(0, v)$; for details see, *e.g.*, [14].

Our aim is the development of Newton-type methods for solving (7) or (9) which is complicated by the system of inequalities in (7) and the non-differentiable max-operator in (9). In the recent past significant progress was made in the investigation of semi-smooth Newton methods and primal-dual active set methods to cope with non-differentiable functionals in infinite-dimensional spaces; see for instance [10, 15]. A direct application of these techniques to (9) results in the following algorithm.

Algorithm A

(i) Choose $c > 0$, (y_0, λ_0); set $k = 0$.
(ii) Set $\mathcal{A}_{k+1} = \{x \in \Omega\colon \lambda_k(x) + c(y_k(x) - \psi(x)) > 0\}$.
(iii) Compute $y_{k+1} = \arg\min \{\frac{1}{2}\, a(y,y) - (f,y)\colon y = \psi \text{ on } \mathcal{A}_{k+1}\}$
(iv) Let λ_{k+1} be the Lagrange multiplier associated to the constraint in (iii) with $\lambda_{k+1} = 0$ on $\Omega \setminus \mathcal{A}_{k+1}$.
(v) Set $k := k + 1$ and go to (ii).

The optimality system for the variational problem in (iii) is given by

$$\text{(10)} \qquad \begin{cases} a(y_{k+1}, v) + \langle \lambda_{k+1}, v\rangle_{H^{-1},H_0^1} = (f,v) \ \text{ for all } \ v \in H_0^1(\Omega), \\ y_{k+1} = \psi \ \text{ on } \ \mathcal{A}_{k+1}, \quad \lambda_{k+1} = 0 \ \text{ on } \ \mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}. \end{cases}$$

The Lagrange multiplier associated to the constraint $y = \psi$ on $\mathcal{A}_{k+1}$ is in general only a distribution in $H^{-1}(\Omega)$ and is not in $L^2(\Omega)$. In fact λ_{k+1} is related to the jumps in the normal derivatives of y across the interface between $\mathcal{A}_{k+1}$ and $\mathcal{I}_{k+1}$, [13]. This complicates the convergence analysis for Algorithm A since the calculus of Newton (or slant) differentiability [10] does not apply. We note that these difficulties are not present if (7) or (9) are discretized. However, they are crucial for the treatment of infinite dimensional problems and as such they are generic. Analogous difficulties arise for state constrained optimization problems, for inverse problems with BV-regularization, and for elasticity problems with contract and friction, to mention a few. This suggests the introduction of regularized problems, which in our case are chosen as

$$(P_\gamma) \ \ \min \frac{1}{2}\, a(y,y) - (f,y) + \frac{1}{2\gamma} \int_\Omega |(\bar\lambda + \gamma(y-\psi))^+|^2 \text{ over } y \in H_0^1(\Omega)$$

where $\gamma > 0$ and $\bar\lambda \in L^2(\Omega)$, $\bar\lambda \geq 0$ are fixed. For later use we denote the objective functional of (P_γ) by $J(y;\gamma)$. The choice of $\bar\lambda$ will be used to influence the feasibility of the solution y_γ of (P_γ). Using Lebesgue's bounded convergence theorem to differentiate the max under the integral in $J(y;\gamma)$, the first order optimality condition associated with

(P_γ) is given by

$$(\mathrm{OC}_\gamma) \qquad \begin{cases} a(y_\gamma, v) + (\lambda_\gamma, v) = (f, v) \ \text{ for all } \ v \in H_0^1(\Omega), \\ \lambda_\gamma = (\bar\lambda + \gamma(y_\gamma - \psi))^+, \end{cases}$$

where $(y_\gamma, \lambda_\gamma) \in H_0^1(\Omega) \times L^2(\Omega)$. With (8) holding, we have $y_\gamma \in H^2(\Omega)$. The primal-dual active set strategy, or equivalently the semi-smooth Newton method, for (P_γ) is given next. For its statement and for later use we introduce $\chi_{\mathcal{A}^{k+1}}$, the characteristic function of the set $\mathcal{A}^{k+1} \subseteq \Omega$.

Algorithm B

(i) Choose $\bar\lambda \geq 0$, (y_0, λ_0); set $k = 0$.
(ii) Set $\mathcal{A}_{k+1} = \{x \in \Omega\colon \bar\lambda(x) + \gamma(y_k(x) - \psi(x)) > 0\}$, $\mathcal{I}_{k+1} = \Omega \setminus \mathcal{A}_{k+1}$.
(iii) Solve for $y_{k+1} \in H_0^1(\Omega)$:
$a(y_{k+1}, v) + (\bar\lambda + \gamma(y_{k+1} - \psi)\chi_{\mathcal{A}_{k+1}}, v) = (f, v)$, for all $v \in H_0^1(\Omega)$.
(iv) Set

$$\lambda_{k+1} = \begin{cases} 0 & \text{on } \mathcal{I}_{k+1}, \\ \bar\lambda + \gamma(y_{k+1} - \psi) & \text{on } \mathcal{A}_{k+1}. \end{cases}$$

Algorithm B was analyzed in [13] where global as well as locally superlinear convergence for every fixed $\gamma > 0$ were established. However, the choice and adaptation (increase) of γ was heuristic in [13] and earlier work. The focus of the present investigation is the automatic adaptive choice of γ. We shall utilize the following two results which we recall from [13] where the proofs can also be found.

Proposition 2.1. *The solutions $(y_\gamma, \lambda_\gamma)$ to (OC_γ) converge to (y^*, λ^*) in the sense that $y_\gamma \to y^*$ strongly in $H_0^1(\Omega)$ and $\lambda_\gamma \rightharpoonup \lambda^*$ weakly in $H^{-1}(\Omega)$ as $\gamma \to \infty$.*

We say that a satisfies the weak maximum principle, if for any $v \in H_0^1(\Omega)$

$$a(v, v^+) \leq 0 \text{ implies } v^+ = 0. \tag{11}$$

Proposition 2.2. *Assume that (11) holds and let $0 < \gamma_1 \leq \gamma_2 < \infty$.*
a) In the infeasible case, *i.e., for $\bar\lambda = 0$, we have $y^* \leq y_{\gamma_2} \leq y_{\gamma_1}$.*
b) In the feasible case, *i.e., if*

$$\bar\lambda \geq 0 \ \text{ and } \ \langle \bar\lambda - f + A\psi, v\rangle_{H^{-1}, H_0^1} \geq 0 \ \text{ for all } \ v \in H_0^1(\Omega), \tag{12}$$

with $v \geq 0$, then $y_{\gamma_1} \leq y_{\gamma_2} \leq y^* \leq \psi$.

3. The primal-dual path

In this section we introduce the primal-dual path and discuss its smoothness properties.

Definition 3.1. *The family of solutions* $\mathcal{C} = \{(y_\gamma, \lambda_\gamma) \colon \gamma \in (0, \infty)\}$ *to* (OC_γ)*, considered as subset of* $H_0^1(\Omega) \times H^{-1}(\Omega)$*, is called the primal-dual path associated to (P).*

For $r \geq 0$ we further set $\mathcal{C}_r = \{(y_\gamma, \lambda_\gamma) \colon \gamma \in [r, \infty)\}$ and with some abuse of terminology we also refer to $\mathcal{C}_r$ as path. In the following lemma we denote by $\hat{y}$ the solution to the unconstrained problem

$$(\hat{P}) \qquad \min J(y) = \frac{1}{2}\, a(y, y) - (f, y) \text{ over } y \in H_0^1(\Omega).$$

Subsequently, in connection with convergence of a sequence in function space we use the subscript 'weak' together with the space to indicate convergence in the weak sense.

Lemma 3.1. *For each* $r > 0$ *the path* $\mathcal{C}_r$ *is bounded in* $H_0^1(\Omega) \times H^{-1}(\Omega)$*, with* $\lim_{\gamma \to \infty} (y_\gamma, \lambda_\gamma) = (y^*, \lambda^*)$ *in* $H_0^1(\Omega) \times H^{-1}(\Omega)_{weak}$*. For* $\bar{\lambda} = 0$ *the path* $\mathcal{C}_0$ *is bounded in* $H_0^1(\Omega) \times H^{-1}(\Omega)$*, with* $\lim_{\gamma \to 0^+} (y_\gamma, \lambda_\gamma) = (\hat{y}, 0)$ *in* $H_0^1(\Omega) \times L^2(\Omega)$*.*

Proof. From (OC_γ) we have for every $\gamma > 0$

$$a(y_\gamma, y_\gamma - y^*) + (\lambda_\gamma, y_\gamma - y^*) = (f, y_\gamma - y^*). \tag{13}$$

Since $\lambda_\gamma = \max(0, \bar{\lambda} + \gamma(y_\gamma - \psi)) \geq 0$ and $\psi - y^* \geq 0$ we have

$$\begin{aligned}(\lambda_\gamma, y_\gamma - y^*) &= (\lambda_\gamma, \frac{\bar{\lambda}}{\gamma} + y_\gamma - \psi + \psi - y^* - \frac{\bar{\lambda}}{\gamma}) \\ &\geq \frac{1}{\gamma}(\lambda_\gamma, \bar{\lambda} + \gamma(y_\gamma - \psi)) - \frac{1}{\gamma}(\lambda_\gamma, \bar{\lambda}) \\ &\geq \frac{1}{\gamma}\left[|\lambda_\gamma|_{L^2}^2 - (\lambda_\gamma, \bar{\lambda})\right].\end{aligned}$$

Combined with (13) this implies that

$$a(y_\gamma, y_\gamma) + \frac{1}{\gamma}\,|\lambda_\gamma|_{L^2}^2 \leq a(y_\gamma, y^*) + (f, y_\gamma - y^*) + \frac{1}{\gamma}\,(\bar{\lambda}, \lambda_\gamma). \tag{14}$$

This estimate, (6) and (OC_γ) imply that $\mathcal{C}_r$ is bounded in $H_0^1(\Omega) \times H^{-1}(\Omega)$ for every $r > 0$. In fact,

$$\begin{aligned}
&\nu|y_\gamma|_{H^1}^2 + \tfrac{1}{\gamma}\,|\lambda_\gamma|_{L^2}^2 \le a(y_\gamma, y_\gamma) + \tfrac{1}{\gamma}\,|\lambda_\gamma|_{L^2}^2 \\
&\le \mu|y_\gamma|_{H^1}\,|y^*|_{H^1} + |f|_{H^{-1}}\,(|y_\gamma|_{H^1} + |y^*|_{H^1}) + \tfrac{1}{\gamma}\,|\bar\lambda|_{L^2}\,|\lambda_\gamma|_{L^2} \\
&\le \tfrac{\nu}{4}\,|y_\gamma|_{H^1}^2 + \tfrac{\mu^2}{\nu}\,|y^*|_{H^1}^2 + \tfrac{\nu}{2}|y_\gamma|_{H^1}^2 + \tfrac{1}{2\nu}|f|_{H^{-1}}^2 \\
&+ \tfrac{1}{2\gamma}|\lambda_\gamma|_{L^2}^2 + \tfrac{1}{2\gamma}|\bar\lambda|_{L^2}^2 + |f|_{H^{-1}}\,|y^*|_{H^1},
\end{aligned}$$

and hence

$$\tfrac{\nu}{4}|y_\gamma|_{H^1}^2 + \tfrac{1}{2\gamma}\,|\lambda_\gamma|_{L^2}^2 \le \tfrac{\mu^2}{\nu}|y^*|_{H^1}^2 + \tfrac{1}{2\nu}\,|f|_{H^{-1}} + |f|_{H^{-1}}\,|y^*|_{H^1} + \tfrac{1}{2\gamma}\,|\bar\lambda|_{L^2}^2.$$

This estimate implies that $\{y_\gamma : \gamma \ge r\}$ is bounded in $H_0^1(\Omega)$ for every $r > 0$. The first equation of (OC_γ) implies that $\{\lambda_\gamma : \gamma \ge r\}$ is bounded in $H^{-1}(\Omega)$ as well. From Proposition 2.1 we have that $\lim_{\gamma\to\infty}(y_\gamma, \lambda_\gamma) = (y^*, \lambda^*)$ in $H_0^1(\Omega) \times H^{-1}(\Omega)_{weak}$. If $\bar\lambda = 0$, then from (14), (6) and (OC_γ) the path $\mathcal{C}_o$ is bounded in $H_0^1(\Omega) \times H^{-1}(\Omega)$ and $\lambda_\gamma \to 0$ in $L^2(\Omega)$ for $\gamma \to 0^+$. From (OC_γ) and the optimality condition for ($\hat P$) we have

$$a(y_\gamma - \hat y, y_\gamma - \hat y) + (\lambda_\gamma, y_\gamma - \hat y) = 0,$$

and hence $\lim_{\gamma\to 0^+} y_\gamma = \hat y$ in $H_0^1(\Omega)$. □

Proposition 3.1. *The path $\mathcal{C}_r$ is globally Lipschitz in $H_0^1(\Omega)\times H^{-1}(\Omega)$, for every $r > 0$. If $\bar\lambda = 0$, then $\mathcal{C}_0$ is globally Lipschitz continuous.*

Proof. Let $\gamma, \bar\gamma \in [r, \infty)$ be arbitrary. Then

$$A(y_\gamma - y_{\bar\gamma}) + (\bar\lambda + \gamma(y_\gamma - \psi))^+ - (\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+ = 0.$$

Taking the inner-product with $y_\gamma - y_{\bar\gamma}$ and using the monotonicity and Lipschitz continuity (with constant $L = 1$) of $x \mapsto \max(0, x)$ we find

$$\begin{aligned}
a(y_\gamma - y_{\bar\gamma},\ & y_\gamma - y_{\bar\gamma}) \\
&\le \left|\left((\bar\lambda + \gamma(y_\gamma - \psi))^+ - (\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+, y_\gamma - y_{\bar\gamma}\right)\right| \\
&\le |\gamma - \bar\gamma|\ |y_\gamma - \psi|_{L^2}|y_\gamma - y_{\bar\gamma}|_{L^2}.
\end{aligned}$$

By Lemma 3.1 the set $\{y_\gamma\}_{\gamma\ge r}$ is bounded in $H_0^1(\Omega)$. Hence there exists $K_1 > 0$ such that

$$\nu|y_\gamma - y_{\bar\gamma}|_{H_0^1}^2 \le K_1|\gamma - \bar\gamma| \cdot |y_\gamma - y_{\bar\gamma}|_{L^2}$$

and by Poincare's inequality there exists $K_2 > 0$ such that

$$|y_\gamma - y_{\bar\gamma}|_{H_0^1} \le K_2|\gamma - \bar\gamma| \ \text{ for all } \ \gamma \ge r, \bar\gamma \ge r.$$

Let us recall here that $|y|_{H_0^1} = |\nabla y|_{L^2}$. Lipschitz continuity of $\gamma \mapsto \lambda_\gamma$ from $[r, \infty)$ to $H^{-1}(\Omega)$ follows from the first equation in (OC_γ). For $\bar\lambda = 0$ the set $\{y_\gamma\}_{\gamma \geq 0}$ is bounded in $H_0^1(\Omega)$. The remainder of the proof remains identical. □

Lemma 3.2. *For every subset $I \subset [r, \infty)$, $r > 0$, the mapping $\gamma \mapsto \lambda_\gamma$ is globally Lipschitz from I to $L^2(\Omega)$.*

Proof. For $0 < \gamma_1 \leq \gamma_2$ we have by (OC_γ)

$$\begin{aligned} |\lambda_{\gamma_1} - \lambda_{\gamma_2}|_{L^2} &= |(\bar\lambda + \gamma_1(y_{\gamma_1} - \psi))^+ - (\bar\lambda + \gamma_2(y_{\gamma_2} - \psi))^+|_{L^2} \\ &\leq (K_3\gamma_1 + K_1 + |\psi|_{L^2})\,|\gamma_1 - \gamma_2| \end{aligned}$$

for some constant $K_3 > 0$. □

We shall use the following notation:

$$S_\gamma = \{x \in \Omega \colon \bar\lambda(x) + \gamma(y_\gamma - \psi)(x) > 0\}.$$

Further we set

$$g(\gamma) = \bar\lambda + \gamma(y_\gamma - \psi). \tag{15}$$

Since $\gamma \mapsto y_\gamma \in H_0^1(\Omega)$ is Lipschitz continuous by Proposition 3.1, there exists a weak accumulation point $\dot y (= \dot y_\gamma)$ of $\frac{1}{\bar\gamma - \gamma}(y_{\bar\gamma} - y_\gamma)$ as $\bar\gamma \to \gamma > 0$, which is also a strong accumulation point in $L^2(\Omega)$. Further $\frac{1}{\bar\gamma - \gamma}(g(\bar\gamma) - g(\gamma))$ has $\dot g(\gamma) \colon = y_\gamma - \psi + \gamma\,\dot y_\gamma$ as strong accumulation point in $L^2(\Omega)$ as $\bar\gamma \to \gamma$. In case $\bar\gamma$ approaches γ from above (or below) the associated accumulation points satisfy certain properties which are described next.

Proposition 3.2. *Let $\gamma > 0$ and denote by $\dot y_\gamma$ any weak accumulation point of $\frac{1}{\bar\gamma - \gamma}(y_{\bar\gamma} - y_\gamma)$ in $H_0^1(\Omega)$ as $\bar\gamma \downarrow \gamma$. Set*

$$S_\gamma^+ = S_\gamma \cup \{x \colon \bar\lambda(x) + \gamma(y_\gamma(x) - \psi(x)) = 0 \wedge \dot g(\gamma)(x) \geq 0\}.$$

Then $\dot y_\gamma$ satisfies

$$a(\dot y_\gamma, v) + ((y_\gamma - \psi + \gamma \dot y_\gamma)\chi_{S_\gamma^+}, v) = 0 \quad \text{for all} \quad v \in H_0^1(\Omega). \tag{16}$$

Proof. By (OC_γ) we have for every $v \in H_0^1(\Omega)$

$$a(y_{\bar\gamma} - y_\gamma, v) + ((\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+ - (\bar\lambda + \gamma(y_\gamma - \psi))^+, v) = 0 \tag{17}$$

We multiply (17) by $(\bar\gamma - \gamma)^{-1}$ and discuss separately the two terms in (17). Clearly, we have

$$\lim_{\bar\gamma \downarrow \gamma} (\bar\gamma - \gamma)^{-1} a(y_{\bar\gamma} - y_\gamma, v) = a(\dot y_\gamma, v).$$

Here and below the limit is taken on the sequence of $\bar\gamma$-values, which provides the accumulation point. Lebesgue's bounded convergence theorem allows to consider the pointwise limits of the integrands. Considering separately the cases $g(\gamma)(x) < 0$, $g(\gamma)(x) > 0$ and $g(\gamma)(x) = 0$ we have

$$\begin{aligned}(\bar\gamma-\gamma)^{-1}((g(\bar\gamma))^+ - (g(\gamma))^+, v)\\ \to ((y_\gamma - \psi + \gamma\,\dot y_\gamma)\chi_{S_\gamma^+}, v) \ \text{ as } \ \bar\gamma \downarrow \gamma.\end{aligned} \tag{18}$$

□

As a consequence of the proof we obtain

Corollary 3.1. *Let $\gamma > 0$ and denote by $\dot y_\gamma$ any weak accumulation point of $\frac{1}{\bar\gamma-\gamma}(y_{\bar\gamma} - y_\gamma)$ in $H_0^1(\Omega)$ as $\bar\gamma \uparrow \gamma$. Set $S_\gamma^- = S_\gamma \cup \{x\colon \bar\lambda(x) + \gamma(y_\gamma(x) - \psi(x)) = 0 \wedge \dot g(\gamma)(x) \geq 0\}$. Then $\dot y_\gamma$ satisfies*

$$a(\dot y_\gamma, v) + ((y_\gamma - \psi + \gamma\,\dot y_\gamma)\chi_{S_\gamma^-}, v) = 0 \ \text{ for all } \ v \in H_0^1(\Omega). \tag{19}$$

Another corollary of Proposition 3.2 treats the case $\bar\lambda = 0$.

Corollary 3.2. *Let $\bar\lambda = 0$ and assume that (11) holds. Then the right- and left derivatives $\dot y_\gamma^r$ and $\dot y_\gamma^l$ of $\gamma \mapsto y_\gamma$, $\gamma \in (0,\infty)$ exist and are given by*

$$a(\dot y_\gamma^r, v) + ((y_\gamma - \psi + \gamma\,\dot y_\gamma^r)\chi_{y_\gamma > \psi}, v) = 0 \ \text{ for all } \ v \in H_0^1(\Omega) \tag{20}$$

$$a(\dot y_\gamma^l, v) + ((y_\gamma - \psi + \gamma\,\dot y_\gamma^l)\chi_{y_\gamma \geq \psi}, v) = 0 \ \text{ for all } \ v \in H_0^1(\Omega). \tag{21}$$

Proof. Let $\bar\gamma \downarrow \gamma$. By Proposition 2.2 any accumulation point $\dot y_\gamma^r$ of $(\bar\gamma-\gamma)^{-1}(y_{\bar\gamma} - y_\gamma)$ satisfies $\dot y_\gamma^r \leq 0$ and hence

$$S_\gamma^+ = \{x \in \Omega\colon y_\gamma(x) > \psi(x)\} \cup \{x \in \Omega\colon y_\gamma(x) = \psi(x) \wedge \dot y_\gamma^r(x) = 0\}.$$

This implies that every accumulation point $\dot y_\gamma^r$ satisfies (20). Since the solution to (20) is unique, the directional derivative from the right exists.

Similarly, if $\bar\gamma \uparrow \gamma$, by Proposition 2.2 we have $S_\gamma^- = \{x \in \Omega\colon y_\gamma(x) \geq \psi(x)\}$, and (21) follows. □

Henceforth we set

$$S_\gamma^\circ = \{x \in \Omega\colon \bar\lambda(x) + \gamma(y_\gamma - \psi)(x) = 0\}.$$

Corollary 3.3. *If meas $(S_\gamma^\circ) = 0$ then $\gamma \mapsto y_\gamma \in H_0^1(\Omega)$ is differentiable at γ and the derivative $\dot y_\gamma$ satisfies*

$$a(\dot y_\gamma, v) + ((y_\gamma - \psi + \gamma\,\dot y_\gamma)\chi_{S_\gamma}, v) = 0 \ \text{ for all } \ v \in H_0^1(\Omega). \tag{22}$$

Proof. Let z denote the difference of two accumulation points of $(\bar\gamma - \gamma)^{-1}(y_{\bar\gamma} - y_\gamma)$ as $\bar\gamma \to \gamma$. As a consequence of (16) and (19)

$$a(z, v) + \gamma(z\chi_{S_\gamma}, v) = 0 \ \text{ for all } \ v \in H_0^1(\Omega).$$

This implies that $z = 0$ by (6). Consequently, accumulation points are unique and by (16), (19) they satisfy (22). □

4. The primal-dual path value functional

In this section we investigate the value function associated with (P_γ) and study its monotonicity as well as smoothness properties.

Definition 4.1. *The functional*

$$\gamma \mapsto V(\gamma) = J(y_\gamma; \gamma) = \frac{1}{2}a(y_\gamma, y_\gamma) - (f, y_\gamma) + \frac{1}{2\gamma}|(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2_{L^2}$$

defined on $(0, \infty)$ *is called the primal-dual-path value functional.*

Let us start by studying first order differentiability properties of V.

Proposition 4.1. *The value function V is differentiable with*

$$\dot V(\gamma) = -\frac{1}{2\gamma^2}\int_\Omega |(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 \ + \ \frac{1}{\gamma}\int_\Omega (\bar\lambda + \gamma(y_\gamma - \psi))^+(y_\gamma - \psi).$$

Corollary 4.1. *For $\bar\lambda = 0$ we have $\dot V(\gamma) = \frac{1}{2}\int_\Omega |(y_\gamma - \psi)^+|^2 \geq 0$, and $\dot V(\gamma) > 0$ unless y_γ is feasible. For $\bar\lambda$ satisfying (12) and with (11) holding we have $y_\gamma \leq \psi$ and hence $\dot V(\gamma) \leq 0$, for $\gamma \in (0, \infty)$.*

In either of the two cases $\dot V(\gamma) = 0$ implies that y_γ solves $(\hat P)$.

Proof. We only show that $\dot V(\gamma) = 0$ implies that y_γ solves $(\hat P)$. The rest of the assertion follows immediately from Proposition 4.1.

If $\bar\lambda = 0$, then $\dot V(\gamma) = 0$ yields $y_\gamma \leq \psi$. Thus, $\lambda_\gamma = 0$ and, hence, y_γ solves $(\hat P)$.

If (11) and (12) are satisfied, then $y_\gamma \leq \psi$ and $\dot V(\gamma) = 0$ implies $\gamma(y_\gamma - \psi) \leq \bar\lambda + \gamma(y_\gamma - \psi) \leq 0$. As a consequence $\lambda_\gamma = 0$, and y_γ solves $(\hat P)$. □

Proof (of Proposition 4.1). For $\bar\gamma, \gamma \in (0, \infty)$ we find

$$\begin{aligned} &\tfrac{1}{2}\, a(y_{\bar\gamma} + y_\gamma, y_{\bar\gamma} - y_\gamma) - (f, y_{\bar\gamma} - y_\gamma) + \\ &\tfrac{1}{2}((\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+ + (\bar\lambda + \gamma(y_\gamma - \psi))^+, y_{\bar\gamma} - y_\gamma) = 0, \end{aligned} \tag{23}$$

and consequently

$$\begin{aligned}
V(\bar\gamma) - V(\gamma) &= \frac{1}{2}a(y_{\bar\gamma}, y_{\bar\gamma}) - \frac{1}{2}a(y_\gamma, y_\gamma) - (f, y_{\bar\gamma} - y_\gamma) \\
&\quad + \frac{1}{2\bar\gamma}\int_\Omega |(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+|^2 - \frac{1}{2\gamma}\int_\Omega |(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 \\
&= \frac{1}{2\bar\gamma}\int_\Omega |(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+|^2 + \frac{1}{2\gamma}\int_\Omega -|(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 \\
&\quad + \frac{1}{2}\int_\Omega -((\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^+ + (\bar\lambda + \gamma(y_\gamma - \psi))^+)(y_{\bar\gamma} - y_\gamma) \\
&= \int_{P_{\bar\gamma}\cap P_\gamma} z + \int_{P_{\bar\gamma}\cap N_\gamma} z + \int_{P_\gamma \cap N_{\bar\gamma}} z = \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3,
\end{aligned}$$

where z stands for the sum of the kernels on the left of the above equalities,

$$P_\gamma = \{x\colon \bar\lambda + \gamma(y_\gamma - \psi) > 0\}, N_\gamma = \{x\colon \bar\lambda + \gamma(y_\gamma - \psi) < 0\},$$

and $P_{\bar\gamma}$, $N_{\bar\gamma}$ are defined analogously. For $\mathcal{I}_2$ we have

$$\begin{aligned}
|\mathcal{I}_2| &\le \frac{1}{2}\int_{P_{\bar\gamma}\cap N_\gamma} \frac{1}{\bar\gamma}(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^2 + |\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi)|\,|y_{\bar\gamma} - y_\gamma| \\
&\le \frac{1}{2}\int_\Omega \frac{1}{\bar\gamma}(\bar\gamma(y_{\bar\gamma} - \psi) - \gamma(y_\gamma - \psi))^2 + |y_{\bar\gamma} - y_\gamma|(|\bar\gamma y_{\bar\gamma} - \gamma\, y_\gamma| + |\bar\gamma - \gamma|\,|\psi|)
\end{aligned}$$

and hence by Proposition 3.1

$$\lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma - \gamma} \quad |\mathcal{I}_2| = 0. \tag{24}$$

Analogously one verifies that

$$\lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma - \gamma} \quad |\mathcal{I}_3| = 0. \tag{25}$$

On $P_{\bar\gamma} \cap P_\gamma$ we have

$$\begin{aligned}
z &= \tfrac{1}{2\bar\gamma}\,(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^2 \\
&\qquad - \frac{1}{2\gamma}\,(\bar\lambda + \gamma(y_\gamma - \psi))^2 - \tfrac{1}{2}\,(2\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi) + \gamma(y_\gamma - \psi))(y_{\bar\gamma} - y_\gamma) \\
&= \tfrac{\gamma - \bar\gamma}{2\bar\gamma\gamma}\,(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^2 \\
&\qquad + \tfrac{1}{2\gamma}\,\left[2\bar\lambda(\bar\gamma(y_{\bar\gamma} - \psi) - \gamma(y_\gamma - \psi)) + \bar\gamma^2(y_{\bar\gamma} - \psi)^2 - \gamma^2(y_\gamma - \psi)^2\right] \\
&\qquad - \tfrac{1}{2}\,(2\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi) + \gamma(y_\gamma - \psi))(y_{\bar\gamma} - y_\gamma) \\
&= \tfrac{\gamma - \bar\gamma}{2\bar\gamma\gamma}\,(\bar\lambda + \bar\gamma(y_{\bar\gamma} - \psi))^2 + \tfrac{\bar\lambda}{\gamma}\,[\bar\gamma(y_{\bar\gamma} - \psi) - \gamma(y_{\bar\gamma} - \psi)] \\
&\qquad + \tfrac{1}{2}\,\left[\tfrac{\bar\gamma^2}{\gamma}(y_{\bar\gamma} - \psi)^2 - \bar\gamma(y_{\bar\gamma} - \psi)^2 + (\bar\gamma - \gamma)(y_{\bar\gamma} - \psi)(y_\gamma - \psi)\right]
\end{aligned}$$

and thus on $P_{\bar\gamma} \cap P_{\bar\gamma}$

$$(\bar\gamma - \gamma)^{-1} z = \frac{-1}{2\bar\gamma\gamma}(\bar\lambda + \bar\gamma(y_\gamma - \psi))^2 + \frac{\bar\lambda}{\gamma}(y_{\bar\gamma} - \psi) + \frac{1}{2}\left[\frac{\bar\gamma}{\gamma}(y_{\bar\gamma} - \psi)^2 + (y_{\bar\gamma} - \psi)(y_\gamma - \psi)\right].$$

By Lebesgue's bounded convergence theorem

$$\lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma - \gamma} \quad \mathcal{I}_1 = \lim_{\bar\gamma\to\gamma} \frac{1}{\bar\gamma - \gamma} \int_\Omega z\, \chi_{P_{\bar\gamma}\cap P_\gamma}$$
$$= -\frac{1}{2\gamma^2}\int_\Omega ((\bar\lambda + \gamma(y_\gamma - \psi))^+)^2 + \frac{1}{\gamma}\int_\Omega (\bar\lambda + \gamma(y_\gamma - \psi))^+(y_\gamma - \psi).$$

Together with (24) and (25) this implies the claim. □

Remark 4.1. *Note that $\dot V$ is characterized without recourse to $\dot y_\gamma$.*

The boundedness of $\{\gamma^2 \dot V(\gamma)\}_{\gamma\geq 0}$ is established next. In the sequel we use $(v)^- = -\min(0, v)$.

Proposition 4.2. *If $\bar\lambda = 0$ and $a(v^+, v^-) = 0$ for all $v \in H_0^1(\Omega)$, then $\{\gamma^2\dot V(\gamma)\}_{\gamma\geq 0}$ is bounded. If* (11) *and* (12) *hold, then again $\{\gamma^2\dot V(\gamma)\}_{\gamma\geq 0}$ is bounded.*

Proof. In the case $\bar\lambda = 0$ we have

$$a(y_\gamma - \psi, v) + \gamma((y_\gamma - \psi)^+, v) = (f, v) - a(\psi, v) \text{ for all } v \in H_0^1(\Omega).$$

Since $(y_\gamma - \psi) \in H_0^1(\Omega)$ and $a((y_\gamma - \psi)^+, (y_\gamma - \psi)^-) = 0$ we have, using (6) with $v = (y_\gamma - \psi)^+$,

$$\nu|(y_\gamma - \psi)^+|^2_{H_0^1(\Omega)} + \gamma|(y_\gamma - \psi)^+|^2_{L^2} \leq |f|_{L^2}|(y_\gamma - \psi)^+|_{H_0^1} + \mu|\psi|_{H^1}|y_\gamma - \psi|_{H^1}.$$

This implies the existence of a constant K, depending on $|\psi|_{H^1}$ and $|f|_{L^2}$, but independent of $\gamma \geq 0$, such that $\gamma|(y_\gamma - \psi)^+|_{L^2} \leq K$. Since $\dot V(\gamma) = \frac{1}{2}\int_\Omega |(y_\gamma - \psi)^+|^2$ the claim follows.

Turning to the feasible case with (11) and (12) holding, we have that $y_\gamma \leq \psi$ for every $\gamma > 0$, and hence $(\bar\lambda + \gamma(y_\gamma - \psi))(x) > 0$ if and only if $\bar\lambda(x) > \gamma(\psi - y_\gamma)(x)$. Consequently,

$$|\dot V(\gamma)| \leq \frac{1}{2\gamma^2}\int_\Omega |(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 + \frac{1}{\gamma}\int_\Omega (\bar\lambda + \gamma(y_\gamma - \psi))^+(\psi - y_\gamma)$$
$$\leq \frac{3}{2\gamma^2}|\bar\lambda|^2_{L^2},$$

which again implies the claim. □

Before we investigate $\ddot V$, we state a result which connects $\gamma\dot V(\gamma)$, $|y^* - y_\gamma|_{H_0^1}$, and $V^* - V(\gamma)$, where $V^* = \lim_{\gamma\to\infty} V(\gamma)$.

Proposition 4.3. *In the feasible respectively infeasible case the following estimate holds true:*

$$|y^* - y_\gamma|^2_{H^1_0} \;\leq\; \frac{2}{\nu}\Big(V^* - V(\gamma) - \gamma \dot{V}(\gamma)\Big)$$

Proof. We have $V^* - V(\gamma) = J(y^*) - J(y_\gamma;\gamma)$ and

$$\begin{aligned} J(y^*) - J(y_\gamma;\gamma) \geq\; & \frac{\nu}{2}|y^* - y_\gamma|^2_{H^1_0} + a(y_\gamma, y^* - y_\gamma) - (f, y^* - y_\gamma) \\ & - \frac{1}{2\gamma}|(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2_{L^2} \end{aligned}$$

where we used (6). From (OC_γ) we have

$$a(y_\gamma, y^* - y_\gamma) - (f, y^* - y_\gamma) = -((\bar\lambda + \gamma(y_\gamma - \psi))^+, y^* - y_\gamma),$$

and hence

$$\begin{aligned} J(y^*) - J(y_\gamma;\gamma) \geq\; & \frac{\nu}{2}|y^* - y_\gamma|^2_{H^1_0} - ((\bar\lambda + \gamma(y_\gamma - \psi))^+, y^* - y_\gamma) \\ & - \frac{1}{2\gamma}|(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2_{L^2} \\ \geq\; & \frac{\nu}{2}|y^* - y_\gamma|^2_{H^1_0} - \frac{1}{2\gamma}|(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2_{L^2} \\ & + ((\bar\lambda + \gamma(y_\gamma - \psi))^+, y_\gamma - \psi) \\ =\; & \frac{\nu}{2}|y^* - y_\gamma|^2_{H^1_0} + \gamma \dot{V}(\gamma). \end{aligned}$$

This completes the prove. □

Below we shall assume that $y_\gamma - \psi \in C(\bar\Omega)$. Recall that for dimension $n \leq 3$ and with (6) and (8) holding, we have $y_\gamma \in H^2(\Omega) \subset C(\bar\Omega)$.

Proposition 4.4. *Let $\dot y_\gamma$ denote any accumulation point of $(\bar\gamma - \gamma)^{-1}(y_{\bar\gamma} - y_\gamma)$ as $\bar\gamma \to \gamma$.*

(a) *If $\bar\lambda = 0$, $y_\gamma - \psi \in C(\bar\Omega)$ and (8) is satisfied, then $\gamma \mapsto V(\gamma)$ is twice differentiable at γ with*

$$\ddot V(\gamma) = \int_\Omega (y_\gamma - \psi)^+ \dot y_\gamma. \tag{26}$$

(b) *For arbitrary* $\bar\lambda$, *if* $\mathrm{meas}(S_\gamma^\circ) = 0$, *then* $\gamma \mapsto V(\gamma)$ *is twice differentiable at* γ *with*

$$\begin{aligned}\ddot V(\gamma) = \tfrac{1}{\gamma^3}\int_\Omega |(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 - \\ \tfrac{2}{\gamma^2}\int_\Omega (\bar\lambda + \gamma(y_\gamma - \psi))^+(y_\gamma - \psi) + \\ \tfrac{1}{\gamma}\int_\Omega (y_\gamma - \psi)(y_\gamma - \psi + \gamma\,\dot y_\gamma)\chi_{S_\gamma}.\end{aligned} \tag{27}$$

Proof. (a) On the subsequence γ_n realizing the accumulation point, we have that $\lim_{n\to\infty}(\gamma_n-\gamma)^{-1}(\dot V(\gamma_n)-\dot V(\gamma))$ equals the right hand side of (26). The claim will be established by verifying that the accumulation points $\dot y_\gamma$ restricted to $S_\gamma = \{x\colon y_\gamma(x) - \psi(x) > 0\}$ are unique. Let z denote the difference of two accumulation points. By (16) and (19) we have

$$a(z,v) + \gamma(z,v) = 0 \ \text{ for all } \ v \in H_0^1(\Omega) \ \text{ with } \ v = 0 \ \text{ on } \ \Omega \setminus S_\gamma.$$

Using (8) and the fact that S_γ is an open set relative to Ω due to continuity of $y_\gamma - \psi$, we find that $z = 0$ in S_γ, as desired.
(b) Let $\dot y_\gamma$ denote any accumulation point of $(\bar\gamma-\gamma)^{-1}(y_{\bar\gamma}-y_\gamma)$ as $\bar\gamma \downarrow \gamma$, and recall the notation $g(\gamma) = \bar\lambda + \gamma(y_\gamma - \psi)$ and S_γ^+ from section 3. On the subsequence realizing the accumulation point we find

$$\begin{aligned}\lim_{\bar\gamma\to\gamma} \tfrac{1}{\bar\gamma-\gamma}\,(\dot V(\bar\gamma) - \dot V(\gamma)) = \ \tfrac{1}{\gamma^3}\int_\Omega |(\bar\lambda + \gamma(y_\gamma - \psi))^+|^2 \\ -\tfrac{2}{\gamma^2}\int_\Omega (\bar\lambda + \gamma(y_\gamma - \psi))^+(y_\gamma - \psi) \\ +\tfrac{1}{\gamma}\int_\Omega (y_\gamma - \psi)(y_\gamma - \psi + \gamma\,\dot y_\gamma)\chi_{S_\gamma^+}.\end{aligned} \tag{28}$$

By assumption, $\mathrm{meas}(S_\gamma^\circ) = 0$ and, hence, the right hand sides of (27) and (28) coincide. Since $\dot y_\gamma$ is unique by Corollary 3.3 the claim is established. □

5. Model functions

In this section we derive low-parameter families of functions which approximate the value functional V and share some of its qualitative properties. We will make use of these models in the numerics section when devising path following algorithms.

5.1. **Infeasible case.** Throughout this subsection we assume

$$\bar\lambda = 0, y_\gamma - \psi \in C(\bar\Omega) \ \text{ for all } \ \gamma \in (0,\infty), \ \text{ and (8).} \tag{29}$$

Observe that (8), together with the general assumption (6), imply (11). In fact, for any $v \in H_0^1(\Omega)$ we have $a(v, v^+) \geq \gamma |v^+|^2$ and hence $0 \geq a(v, v^+)$ implies $v^+ = 0$.

Proposition 5.1. *The value function V satisfies $\dot{V}(\gamma) \geq 0$ and $\ddot{V}(\gamma) \leq 0$ for $\gamma \in (0, \infty)$.*

Proof. Proposition 4.1 implies that $\dot{V}(\gamma) \geq 0$. Moreover $y_{\gamma_2} \leq y_{\gamma_1}$ for $\gamma_2 \geq \gamma_1 > 0$ and hence $\dot{y}_\gamma \leq 0$ a.e. on S_γ. Consequently $\ddot{V}(\gamma) \leq 0$ by Proposition 4.4. □

A model function m for the value function V should reflect the sign properties of V and its derivatives. Moreover $V(0)$ gives the value of $(\hat{P})$ and hence we shall require that $m(0) = V(0)$. Finally from Lemma 3.1 we conclude that V is bounded on $[0, \infty)$. All these properties are satisfied by functions of the form

$$m(\gamma) = C_1 - \frac{C_2}{E + \gamma} \tag{30}$$

with $C_1 \in \mathbb{R}$. $C_2 \geq 0$, $E > 0$ satisfying

$$m(0) = V(0) = C_1 - \frac{C_2}{E}. \tag{31}$$

Other choices for model functions are also conceivable, for example, $\gamma \to C_1 - \frac{C_1}{(E+\gamma)^r}$ with $r > 1$. Note, however, that the asymptotic behavior of the model in (30) is such that $\gamma^2 \dot{m}(\gamma)$ is bounded for $\gamma \to \infty$. This is consistent with the boundedness of $\gamma^2 \dot{V}(\gamma)$ for $\gamma \to \infty$ asserted in Proposition 4.2.

Another reason for choosing (30) is illustrated next. Choosing $v = (y_\gamma - \psi)^+$ in (OC_γ) we find

$$a(\dot{y}_\gamma, (y_\gamma - \psi)^+) + |(y_\gamma - \psi)^+|_{L^2}^2 + \gamma \int_\Omega (y_\gamma - \psi)^+ \dot{y}_\gamma = 0. \tag{32}$$

For the following discussion we

(33) replace $a(\cdot, \cdot)$ by $E(\cdot, \cdot)$ with $E > 0$ a constant, and V by m.

By Proposition 4.1 and (26) the following ordinary differential equation is obtained for m:

$$(E + \gamma)\, \ddot{m}(\gamma) + 2\, \dot{m}(\gamma) = 0. \tag{34}$$

The solutions to (34) are given by (30). To get an account for the quality of our model in (30) we refer to the left plot of Figure 4 in Section 6.

5.2. **Feasible case.** Throughout this subsection we assume

(35) (11), $\bar\lambda$ satisfies (12) and meas $(S_\gamma^\circ) = 0$ for all $\gamma \in (0,\infty)$.

Proposition 5.2. *The value function V satisfies $\dot V(\gamma) \le 0$ and $\ddot V(\gamma) \ge 0$ for $\gamma \in (0,\infty)$.*

Proof. By Proposition 2.2 we have $y_\gamma \le \psi$ and hence $\dot V(\gamma) \le 0$ by Proposition 4.1. A short computation based on (27) shows that

(36)
$$\ddot V(\gamma) = \frac{1}{\gamma^3}\int_\Omega \chi\bar\lambda^2 + \int_\Omega \chi(y_\gamma - \psi)\dot y_\gamma \ge \frac{1}{\gamma}\int_\Omega \chi(y_\gamma-\psi)^2 + \int_\Omega \chi\,(y_\gamma-\psi)\dot y_\gamma,$$

where χ is the characteristic function of the set $S_\gamma = \{\bar\lambda + \gamma(y_\gamma - \psi) > 0\}$. From (22) we have

$$\gamma|\dot y_\gamma|_{L^2(S_\gamma)} \le |\psi - y_\gamma|_{L^2(S_\gamma)},$$

and hence $\ddot V(\gamma) \ge 0$. □

An immediate consequence is stated next.

Lemma 5.1. *If the solution to the unconstrained problem is not feasible, then* $\lim_{\gamma\downarrow 0} V(\gamma) = \infty$.

Proof. Assume that $\lim_{\gamma\downarrow 0} V(\gamma)$ is finite. Then, using (P_γ), there exists a sequence $\gamma_n \to 0$ and $\tilde y \in H_0^1(\Omega)$ such that $y_{\gamma_n} \rightharpoonup \tilde y$ weakly in $H_0^1(\Omega)$, with y_{γ_n} the solution to (P_{γ_n}), and $\lambda_{\gamma_n} = \max(0, \bar\lambda + \gamma_n(y_n - \psi)) \to 0$ in $L^2(\Omega)$. Consequently $\tilde y \le \psi$. Taking the limit with respect to n in (OC_{γ_n}) it follows that $\tilde y \le \psi$ is the solution to $(\hat P)$ which contradicts our assumption. □

From Lemmas 3.1 and 5.1, and Proposition 5.2 it follows that $\gamma \mapsto V(\gamma)$, $\gamma \in (0,\infty)$, is a monotonically strictly decreasing, convex function with $\lim_{\gamma\to 0^+} V(\gamma) = \infty$. All these properties are also satisfied by functions of the form

$$m(\gamma) = C_1 \;-\; \frac{C_2}{E+\gamma} \;+\; \frac{B}{\gamma}, \tag{37}$$

provided that $C_1 \in \mathbb{R}$, $C_2 \ge 0$, $E > 0$, $B > 0$ and $C_2 \le B$.

We now give the motivation for choosing the model function m for V as in (37). From (22) with $v = (y_\gamma - \psi)\chi$ we get

$$a(\dot y_\gamma, (y-\psi)\chi) + \gamma(\dot y_\gamma\chi, y_\gamma - \psi) + ((y_\gamma - \psi)\chi, y_\gamma - \psi) = 0,$$

where $\chi = \chi_{S_\gamma}$. As in the infeasible case we replace $a(\cdot,\cdot)$ by $E(\cdot,\cdot)$, with E a constant, and using (22) we arrive at

$$(E+\gamma)(\dot y_\gamma\chi, v) + ((y_\gamma - \psi)\chi, v) = 0.$$

The choice $v = y_\gamma - \psi$ implies

$$(E+\gamma)(\dot{y}_\gamma\chi, y_\gamma - \psi) + ((y_\gamma - \psi)\chi, y_\gamma - \psi) = 0. \tag{38}$$

Note that $\dot{V}(\gamma)$ can be expressed as

$$\dot{V}(\gamma) = -\frac{1}{2\gamma^2}\int_\Omega \bar{\lambda}^2\chi + \frac{1}{2}\int_\Omega (y_\gamma - \psi)^2\chi. \tag{39}$$

Using (36) and (39) in (38), and replacing V by m, due to the substitution for $a(\cdot,\cdot)$, we find

$$(E+\gamma)\ddot{m} + 2\dot{m} - E\,\gamma^{-3}\int_\Omega \chi\,\bar{\lambda}^2 = 0.$$

We further replace $\int_\Omega \chi\bar{\lambda}^2$, which is a bounded quantity depending on γ, by $2B$, and obtain, as the ordinary differential equation that we propose for the model function m in the feasible case,

$$(E+\gamma)\ddot{m} + 2\dot{m} - 2\gamma^{-3}E\,B = 0. \tag{40}$$

The family of solutions is given by (37). In the right plot of Figure 4 in Section 6 we depict the approximation quality of $m(\gamma)$.

6. Path-following algorithms

In this section we study the basic Algorithm B together with a variety of adjustment schemes for the path parameter γ. For this purpose recall that, depending on the shift parameter $\bar{\lambda}$, the elements y_γ along the primal-dual path are feasible or infeasible. As we have seen in the previous section, this implies different models for approximating the value function V. We will see, however, that for $\gamma > 0$ in both cases similar strategies for updating γ may be used. When referring to the infeasible or feasible case, (29), respectively, (35) is assumed to hold.

The subsequent discussion is based on the following two-dimensional test problems. We point out that the bound ψ in P1 below does not satisfy $\psi \in H^1(\Omega)$. However, as we shall see, the feasible and infeasible primal-dual path as well as the algorithms introduced subsequently still perform satisfactorily. We include this example since discontinuous obstacles are of practical relevance.

Test problem P1. We consider (8) with $a_{ij} = \delta_{ij}$, with δ_{ij} the Kronecker symbol, $d = 0$ and $\Omega = (0,1)^2$. We choose

$$f(x_1, x_2) = 500x_1\sin(5x_1)\cos(x_2)$$

and $\psi \equiv 10$ on $\Omega \setminus K$, and $\psi \equiv 1$ on K with $K = \{x \in \Omega : \frac{1}{5} \leq \|x - (\frac{1}{2}, \frac{1}{2})^\top\|_2 \leq \frac{2}{5}\}$. The solution y^*, the obstacle ψ, and the active set $\mathcal{A}^*$ at the solution are shown in Figure 1.

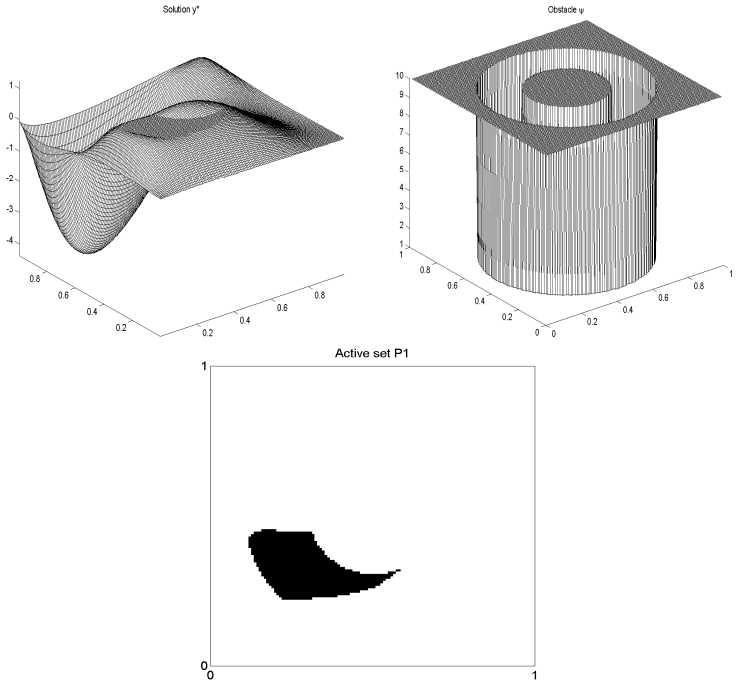

FIGURE 1. Optimal solution y^* (upper left plot), obstacle ψ (upper right plot), and the active set $\mathcal{A}^*$ (lower plot) for test problem P1.

Test problem P2. Again we consider (8) with a_{ij}, d and Ω as before, and define

$$y^\dagger := \begin{cases} x_1 & \text{on } T_1 := \{x \in \Omega : x_2 \le x_1 \wedge x_2 \le 1 - x_1\}, \\ 1 - x_2 & \text{on } T_2 := \{x \in \Omega : x_2 \le x_1 \wedge x_2 \ge 1 - x_1\}, \\ 1 - x_1 & \text{on } T_3 := \{x \in \Omega : x_2 \ge x_1 \wedge x_2 \ge 1 - x_1\}, \\ x_2 & \text{on } T_4 := \{x \in \Omega : x_2 \ge x_1 \wedge x_2 \le 1 - x_1\}. \end{cases} \tag{41}$$

The obstacle ψ is defined by $\psi \equiv y^\dagger$ on $S_1 := \{x \in \Omega : \|x - (\frac{1}{2}, \frac{1}{2})^\top\|_\infty \le \frac{1}{4}\}$, $\psi \equiv \frac{1}{4}$ on $S_2 \setminus S_1$, and

$$\psi := \begin{cases} 2x_1 & \text{on } T_1 \cap (\Omega \setminus S_2), \\ \frac{1}{4} - 2(x_2 - \frac{7}{8}) & \text{on } T_2 \cap (\Omega \setminus S_2), \\ \frac{1}{4} - 2(x_1 - \frac{7}{8}) & \text{on } T_3 \cap (\Omega \setminus S_2), \\ 2x_2 & \text{on } T_4 \cap (\Omega \setminus S_2), \end{cases}$$

with $S_2 := \{x \in \Omega : \|x - (\frac{1}{2}, \frac{1}{2})^\top\|_\infty \le \frac{3}{8}\}$. The forcing term is given by

$$(f, \phi)_{L^2} = \int_{\Omega^+} \phi(s)ds + (\chi_{S_1}, \phi)_{L^2} + \int_{S_1 \cap \Omega^+} \phi(s)ds \text{ for all } \phi \in H_0^1(\Omega),$$

where $\Omega^+ := \{x \in \Omega : x_2 = x_1\} \cup \{x \in \Omega : x_2 = 1 - x_1\}$. We recall that for $\phi \in H_0^1(\Omega)$, $\Omega \subset \mathbb{R}^2$, the traces along smooth curves are well-defined. The solution y^* is given by $y^* = y^\dagger$. The active or coincidence set at the solution is $\mathcal{A}^* = S_1$. The Lagrange multiplier $\lambda^* = f + \Delta y^*$ is in $H^{-1}(\Omega)$ and enjoys no extra regularity. In Figure 2 we display the optimal solution y^*, the obstacle ψ, and the active set $\mathcal{A}^*$.

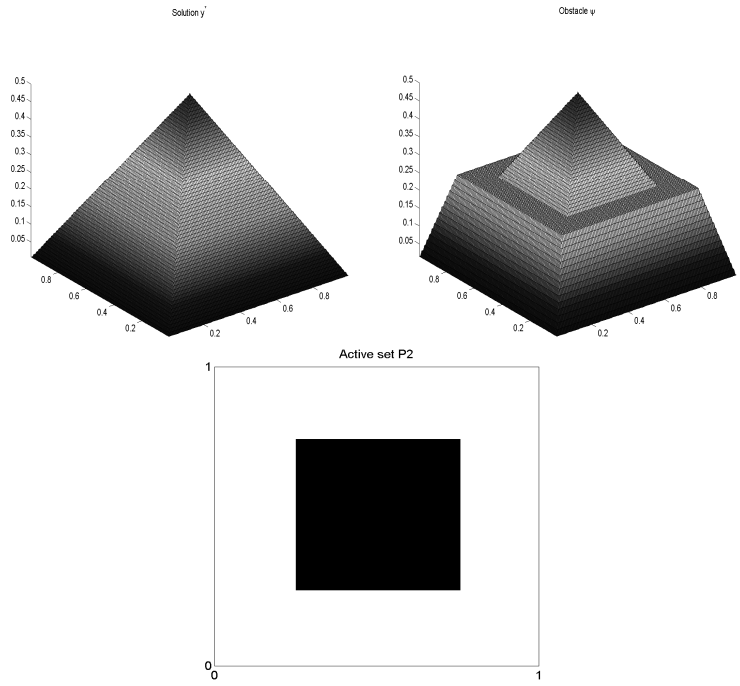

FIGURE 2. Optimal solution y^* (upper left plot), obstacle ψ (upper right plot), and the active set $\mathcal{A}^*$ (lower plot) for test problem P2.

Test problem P3. For this test problem (8) is satisfied. We therefore obtain $y^* \in H^2(\Omega)$ and $\lambda^* \in L^2(\Omega)$. The coefficients a_{ij} and d as well as Ω are as before. The volume force f is given by $f = -\Delta v$ with $v(x_1, x_2) = \sin(3\pi x_1)\sin(3\pi x_2)$. Further, we have $\psi = \frac{1}{4} - \frac{1}{10}\sin(\pi x_1)\sin(\pi x_2)$. The optimal solution y^*, the Lagrange multiplier λ^*, and the active set at y^* are displayed in Figure 3.

Unless specified otherwise, the subsequent algorithms are initialized by $y_0 = (-\Delta)^{-1} f$, where $-\Delta$ denotes the Laplacian with homogeneous Dirichlet boundary conditions. The initial Lagrange multiplier is chosen as $\lambda_0 = \gamma_0 \chi_{\{y_0 > \psi\}}(y_0 - \psi)$.

The discretization of $-\Delta$ is based on the classical five point finite difference stencil. By h we denote the mesh size which we occasionally

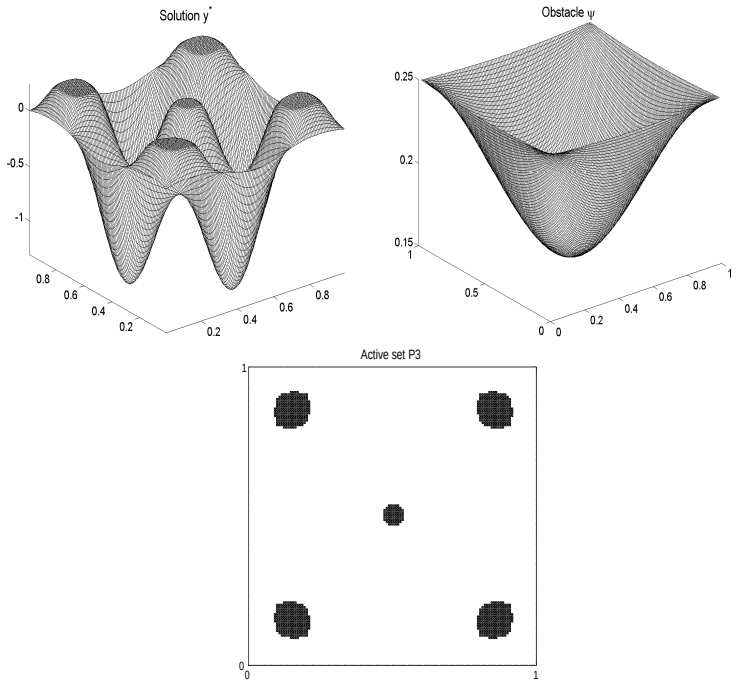

FIGURE 3. Optimal solution y^* (upper left plot), obstacle ψ (upper right plot), and the active set $\mathcal{A}^*$ (lower plot) for test problem P3.

drop for convenience. The forcing term f in $P2$ is discretized by $f = -\Delta y^\dagger + \chi_{S_1} e + \chi_{S_1}(-\Delta y^\dagger)$, where e is the vector of all ones, and χ_{S_1} represents a diagonal matrix with entry $(\chi_{S_1})_{ii} = 1$ for grid points $x_i \in S_1$ and $(\chi_{S_1})_{ii} = 0$ otherwise. Above $y^\dagger$ denotes the grid function corresponding to (41).

6.1. **A strategy based on model functions – exact path following.** As outlined in section 5 there are good reasons to trust our model functions (30) and (37) in the infeasible and feasible cases, respectively. Let us start by focusing on the infeasible case. The model is given by $m(\gamma) = C_1 - C_2(E+\gamma)^{-1}$. For determining the three parameters C_1, C_2 and E, we use the information $V(0), V(\gamma), \dot{V}(\gamma)$, which, by Proposition 4.1, is available from one solve of the unconstrained problem $(\hat{P})$ and one solve for (P_γ). The conditions

$$m(0) = V(0), \quad m(\gamma) = V(\gamma), \quad \dot{m}(\gamma) = \dot{V}(\gamma) \tag{42}$$

yield

$$
\begin{aligned}
E &= \gamma^2 \dot{V}(\gamma) \left(V(\gamma) - V(0) - \gamma \dot{V}(\gamma) \right)^{-1}, \\
C_2 &= \gamma^{-1} E (E + \gamma) \left(V(\gamma) - V(0) \right), \qquad (43) \\
C_1 &= V(0) + C_2 E^{-1}.
\end{aligned}
$$

We could have used an alternative reference value $\gamma_r \in (0, \gamma)$ and computed $m(\gamma_r) = V(\gamma_r)$ instead of $m(0) = V(0)$. In Figure 4 we compare $V(\gamma)$ to $m(\gamma)$ for different values of the coefficients (C_1, C_2, E). These coefficients depend on different values y_f for γ (in (42)) produced by Algorithm EPTS (see below) for problem $P1$. The solid line corresponds to $V(\gamma)$. The corresponding γ-values γ_f for (42) are depicted in the legend of the left plot in Figure 4. The dotted and dashed line belong to rather small γ-values and the dashed-dotted and the circled lines to large γ_f in (42). As we can see, the dotted line is accurate in the range for relatively small γ_f, while the other lines are more accurate for large γ_f. From now on we consider only the choices $\gamma_r = 0$ and $\gamma = \gamma_k$ in

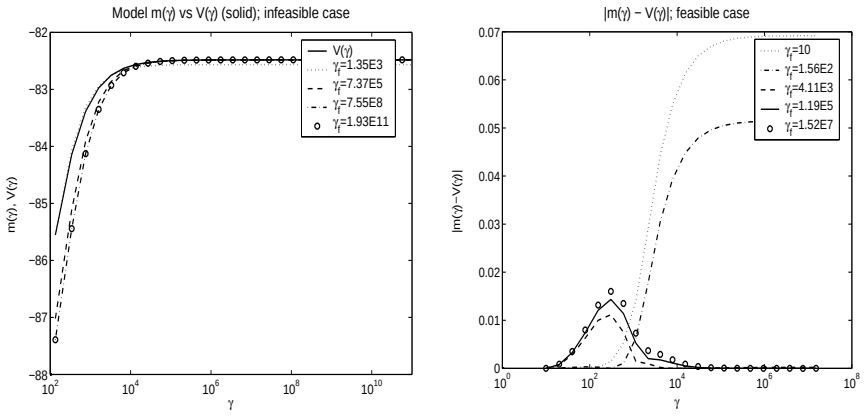

FIGURE 4. Left: Model $m(\gamma)$ vs. $V(\gamma)$ (solid) in the infeasible case for P1. Right: Model $m(\gamma)$ vs. $V(\gamma)$ in the feasible case.

(42) when updating γ_k.

Next we discuss properties of the model parameters E, C_1, C_2 according to (43). For this purpose assume that the solution $\hat{y}$ to $(\hat{P})$ is not feasible for (P). Then by Corollary 4.1 we have $\dot{V}(\gamma) > 0$ for all $\gamma > 0$. Consequently $V(\gamma) > V(0)$ and $V(\gamma) - V(0) - \gamma \dot{V}(\gamma) = -\int_0^\gamma \int_s^\gamma \ddot{V}(\sigma) d\sigma ds > 0$, and, hence, $E > 0$ and $C_2 > 0$ for all $\gamma \in (0, +\infty)$. This implies $m(\gamma) \leq C_1$ and $m(\gamma) \to C_1$ for $\gamma \to +\infty$.

We propose the following update strategy for γ: Let $\{\tau_k\}$ satisfy $\tau_k \in (0,1)$ for all $k \in \mathbb{N}$ and $\tau_k \downarrow 0$ as $k \to \infty$, and assume that $V(\gamma_k)$ is available. Then, given γ_k the updated value γ_{k+1} should ideally satisfy

$$|V^* - V(\gamma_{k+1})| \leq \tau_k |V^* - V(\gamma_k)|. \tag{44}$$

Since V^* and $V(\gamma_{k+1})$ are unknown, we use $C_{1,k}$ and our model $m_k(\gamma) = C_{1,k} - C_{2,k}/(E_k + \gamma)$ at $\gamma = \gamma_{k+1}$ instead. Thus, (44) is replaced by

$$|C_{1,k} - m_k(\gamma_{k+1})| \leq \tau_k |C_{1,k} - V(\gamma_k)| =: \beta_k. \tag{45}$$

Solving the equation $C_{1,k} - m_k(\gamma_{k+1}) = \beta_k$, we obtain

$$\gamma_{k+1} = \frac{C_{2,k}}{\beta_k} - E_k. \tag{46}$$

In Theorem 6.1 we shall show that $\gamma_{k+1} \geq \kappa \gamma_k$, with $\kappa > 1$, independently of $k \in \mathbb{N}$.

In the feasible case, *i.e.*, when $\bar{\lambda}$ satisfies (12), we use the model $m(\gamma) = C_1 - C_2(E+\gamma)^{-1} + B\gamma^{-1}$ with $C_2 \geq 0$ and $E, B > 0$; see (37). Let $\gamma_r > 0$, $\gamma_r \neq \gamma$, denote a reference γ-value, then we use the conditions

$$m(\gamma_r) = V(\gamma_r), \quad \dot{m}(\gamma_r) = \dot{V}(\gamma_r), \quad m(\gamma) = V(\gamma), \quad \dot{m}(\gamma) = \dot{V}(\gamma)$$

for fixing B, $C1$, $C2$, E. Solving the corresponding system of nonlinear equations, we get

$$E = \Big((\gamma_r - \gamma)(\dot{V}(\gamma_r)\gamma_r^2 + \dot{V}(\gamma)\gamma^2) + 2\gamma_r\gamma(V(\gamma) - V(\gamma_r)) \Big) / \\ \Big((\dot{V}(\gamma)\gamma + \dot{V}(\gamma_r)\gamma_r)(\gamma - \gamma_r) + (\gamma_r + \gamma)(V(\gamma_r) - V(\gamma)) \Big)$$

and

$$B = \gamma_r^2\gamma^2 \Big((V(\gamma) - V(\gamma_r))^2 - \dot{V}(\gamma)\dot{V}(\gamma_r)(\gamma - \gamma_r)^2 \Big) / \\ \Big((\gamma - \gamma_r)^2(\dot{V}(\gamma_r)\gamma_r^2 + \dot{V}(\gamma)\gamma^2) + 2(\gamma - \gamma_r)\gamma_r\gamma(V(\gamma_r) - V(\gamma)) \Big)$$

Then the parameters C_1 and C_2 are given by

$$\begin{aligned} C_2 &= (E+\gamma)^2 \left(\frac{B}{\gamma^2} + \dot{V}(\gamma) \right), \\ C_1 &= V(\gamma) + \frac{C_2}{E+\gamma} - \frac{B}{\gamma}. \end{aligned}$$

In the right plot of Figure 4 we show $|m(\gamma) - V(\gamma)|$ with $m(\gamma)$ produced by the iterates of Algorithm EPTS for $P1$ similar to the infeasible case. Again we can see that our model yields a close approximation of the value function V.

If we require that (45) is satisfied in the feasible case, then we obtain the following update strategy for γ:

$$\gamma_{k+1} = -\frac{D_k}{2} + \sqrt{\frac{D_k^2}{4} + \frac{B_k E_k}{\beta_k}}, \tag{47}$$

where $D_k = E_k + (C_{2,k} - B_k)/\beta_k$. In Theorem 6.1 we shall establish $\gamma_{k+1} \geq \kappa\gamma_k$ for all $k \in \mathbb{N}_0$ with $\kappa > 1$ independent of k.

Next we describe an **e**xact **p**ath-following version of Algorithm B which utilizes the update strategy (45) for updating γ.

Algorithm EP.

(i) Select γ_r. Compute $V(\gamma_r)$ and choose $\gamma_0 > \max(1, \gamma_r)$; set $k = 0$.
(ii) Apply Algorithm B to obtain y_{γ_k}.
(iii) Compute $V(\gamma_k)$, $\dot{V}(\gamma_k)$ and γ_{k+1} according to (46) in the infeasible case or (47) in the feasible case.
(iv) Set $k = k + 1$, and go to (ii).

Concerning the choice of γ_r note that in the infeasible case we have $\gamma_r \geq 0$, and in the feasible case $\gamma_r > 0$. Convergence of Algorithm EP is addressed next.

Theorem 6.1. *Assume that the solution to* $(\hat{P})$ *is not feasible for* (P). *Then the iterates* γ_k *of Algorithm EP tend to* ∞ *as* $k \to \infty$ *and consequently* $\lim_{k\to\infty}(y_{\gamma_k}, \lambda_{\gamma_k}) = (y^*, \lambda^*)$ *in* $H_0^1(\Omega) \times H^{-1}(\Omega)_{weak}$.

Proof. Let us consider the infeasible case. then (45) is equivalent to

$$0 < C_{1,k} - m_k(\gamma_{k+1}) < \tau_k(C_{1,k} - m_k(\gamma_k)). \tag{48}$$

Since $\gamma \mapsto m_k(\gamma)$ is strictly increasing and $\tau_k \in (0, 1)$, it follows that $\gamma_{k+1} > \gamma_k$ for every $k = 0, 1, \ldots$. If $\lim_{k\to\infty} \gamma_k = \infty$, then $\lim_{k\to\infty}(y_{\gamma_k}, \lambda_{\gamma_k}) = (y^*, \lambda^*)$. Otherwise there exists $\bar{\gamma}$ such that $\lim_{k\to\infty} \gamma_k = \bar{\gamma}$. Since $\gamma \mapsto V(\gamma)$ and $\gamma \mapsto \dot{V}(\gamma)$ are continuous on $(0, \infty)$, it follows from (42) and (43) that $\lim_{k\to\infty} E_k = E(\bar{\gamma})$, $\lim_{k\to\infty} C_{1,k} = C_1(\bar{\gamma})$, and $\lim_{k\to\infty} C_{2,k} = C_2(\bar{\gamma})$, where $E(\bar{\gamma})$, $C_1(\bar{\gamma})$, $C_2(\bar{\gamma})$ are given by (43) with γ replaced by $\bar{\gamma}$. Taking the limit with respect to k in (48) we arrive at

$$\frac{C_2(\bar{\gamma})}{E(\bar{\gamma}) + \bar{\gamma}} = 0,$$

which is impossible, since $C_2(\bar{\gamma}) > 0$ and $E(\bar{\gamma}) > 0$ if the solution to $(\hat{P})$ is not feasible for (P). Thus $\lim_{k\to\infty} \gamma_k = \infty$. The feasible case is treated analogously. □

Numerically we stop the algorithm as soon as $\|(r_k^{1,h}, r_k^{2,h}, r_k^{3,h})^\top\|_2 \leq \sqrt{\epsilon_M}$, where

$$\begin{aligned} r_k^{1,h} &= \|y_{\gamma_k}^h + (-\Delta^h)^{-1}(\lambda_{\gamma_k}^h - f^h)\|_{H^{-1,h}} / \|f^h\|_{H^{-1,h}}, \\ r_k^{2,h} &= \|\lambda_{\gamma_k}^h - \max(0, \lambda_{\gamma_k}^h + y_{\gamma_k}^h - \psi^h)\|_{H^{-1,h}}, \\ r_k^{3,h} &= \|\max(0, y_{\gamma_k}^h - \psi^h)\|_{L_2^h}, \end{aligned}$$

ϵ_M denotes the machine accuracy. Here $|\cdot|_{H^{-1,h}}$ denotes the discrete version of $|\cdot|_{H^{-1}}$. For some vector v it is realized as $|v|_{H^{-1}} = |\nabla^h(-\Delta^h)^{-1}v|_{L_2^h}$ with $|\cdot|_{L_2^h}$ the discrete L^2-norm and ∇^h a forward difference approximation of the gradient operator; see [8]. The inner iteration, *i.e.*, Algorithm B for $\gamma = \gamma^k$ is terminated if successive active sets coincide or

$$\|-\Delta^h y_{\gamma_k}^{h,l} + \lambda_{\gamma_k}^{h,l} - f^h\|_{H^{-1,h}} / \|f^h\|_{H^{-1,h}} \leq \sqrt{\epsilon_M}.$$

Here the superscript $l = l(k)$ denotes the iteration index of Algorithm B for fixed k. For a discussion and numerical results in the case where the approximation errors due to the discretization of the underlying function space problems is incorporated in the algorithmic framework, e.g., when stopping the algorithm, we refer to the next section 6.2.

The initialization of γ is as follows: In the infeasible case we propose a choice of γ_0 based on the deviation of the linearization of $V(\gamma)$ at $\gamma = \gamma_r$ from the objective value of the unconstrained problem $(\hat{P})$ at the projection of y_{γ_r} onto the feasible set. In our realization of this heuristic we choose $\gamma_r = 0$, compute $\hat{y}$, $V(0)$ and $\dot{V}(0)$. Then we set

$$\gamma_0 = \max\left\{1, \zeta \frac{J(y_b) - V(0)}{\dot{V}(0)}\right\}, \tag{49}$$

where $\zeta \in (0, 1]$ is some fixed constant, $y_b(x) = \min(\hat{y}, \psi(x))$, and J denotes the objective function of $(\hat{P})$. Note that $\hat{y}$ is the minimizer of the unconstrained problem $(\hat{P})$. For the examples below we used $\zeta = 1$. In the feasible case we choose a reference value γ_r, *e.g.*, $\gamma_r = 1$, and solve the path problem (P_γ). Then we choose

$$\gamma_0 = \gamma_r + \frac{J(\hat{y}) - V(\gamma_r)}{\dot{V}(\gamma_r)}, \tag{50}$$

where $\hat{y}$ denotes the minimizer of the discretized unconstrained problem $(\hat{P})$. If $\hat{y}$ is not feasible for (P), then one has $J(\hat{y}) < V(\gamma_r)$ and hence $\gamma_0 > \gamma_r$.

When applied to P1, P2 and P3 for $h = 1/128$ and with $\tau_k = 0.01^{k+1}$, we obtain the results shown in Figure 5 and Table 1.

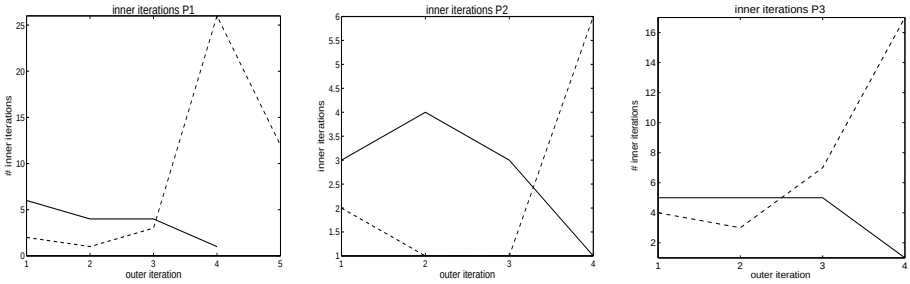

FIGURE 5. Number of inner iterations (vertical axis) per outer iteration for P1 (left plot), P2 (middle plot), and P3 (right plot); solid line – infeasible case; dashed line – feasible case.

	$P1$		$P2$		$P3$	
version	# out. it.	# in. it.	# out. it.	# in. it.	# out. it.	# in. it.
feasible	5	44	4	10	4	31
infeasible	4	15	4	11	4	16

TABLE 1. Comparison of iteration counts.

From our test runs, also for other test problems, we observe the following characteristics:

- For the feasible version the number of inner iterations exhibits an increasing tendency until a saturation value is reached and then, unless the algorithm stops at an approximate solution, it starts to decrease. For the infeasible version we typically observe that the first couple of iterations require several inner iterations. As the outer iterations proceed the number of inner iterations drops eventually to one. We also tested less aggressive γ-updates compared to the ones used here, like, *e.g.*, updates based on $\gamma_{k+1} = \xi\gamma_k$ with $\xi > 1$ fixed.
- The numerically observable convergence speed of y_{γ_k} towards y^* in $H_0^1(\Omega)$ is typically superlinear. This can be seen from Figure 6 where the plots for the discrete versions q_k^h of the quotients
$$q_k = \frac{|y_{\gamma_{k+1}} - y^*|_{H_0^1}}{|y_{\gamma_k} - y^*|_{H_0^1}}$$
are shown. Note the vertical axis uses a logarithmic scale. In the first row, for P1 we depict the behavior of q_k^h for $h = 2^{-i}$, $i = 5, 6, 7, 8$, for the infeasible case (left plot) and the feasible case

(right plot), respectively. We observe that the convergence rate is stable with respect to decreasing mesh size h. In the second row we see the behavior of q_k^h for P2 and P3, respectively, with $h = 2^{-7}$. Again, we observe a superlinear rate of convergence. With respect to decreasing h the same conclusion as for P1 holds true. These stability results provide a link between our function space theory and the numerical realization of the algorithms.

- In connection with the convergence speed it is of interest how the detection process of the correct active set works. For the rather aggressive γ-updates used in Algorithm EP the difference between two successive active sets is zero typically only in the last iteration. However, if a less aggressive strategy for updating γ is used, then it is to expect, that the difference of active sets might become zero already earlier along the iteration. In Figure 7, for the strategy $\gamma_{k+1} = 2\gamma_k$, we show the difference of successive active sets, *i.e.*, the vertical axis relates to the number of grid point which are in $\mathcal{A}_{k+1}$ but not in $\mathcal{A}_k$ and vice versa. We detect that for the infeasible case there exists an iteration index $\bar{k}$ after which the difference is constantly zero. This behavior is a strong indication that the correct active set was detected. It suggests to fix this set $\mathcal{A}_{\bar{k}}$, and to set $\bar{y}_{|\mathcal{A}_{\bar{k}}} = \psi_{|\mathcal{A}_{\bar{k}}}$, $\mathcal{I}_{\bar{k}} = \Omega \setminus \mathcal{A}_{\bar{k}}$ and $\bar{\lambda}_{\mathcal{I}_{\bar{k}}} = 0$. Then one computes $\bar{y}_{|\mathcal{I}_{\bar{k}}}$ and $\bar{\lambda}_{\mathcal{A}_{\bar{k}}}$ such that $a(\bar{y}, v) + \langle \bar{\lambda}, v \rangle_{H^{-1}, H_0^1} = (f, v)$ for all $v \in H_0^1(\Omega)$ and checks whether $(\bar{y}, \bar{\lambda})$ satisfies (7). If this is the case, then the solution is found; otherwise $\gamma_{\bar{k}}$ is updated and the iteration continued. If we apply this technique for $P1$ in the infeasible case, then the algorithm stops at iteration 15 (35 inner iterations) with the exact discrete solution as compared to 28 outer and 47 inner iterations without the additional stopping rule. There were four iterations where the additional system solve was necessary but without obtaining the numerical solution. Hence, w.r.t. system solves the amount of work drops from 47 solves to 39 $(= 35 + 4)$. A similar observation is true for $P2$ and $P3$. In the feasible case, however, this strategy yields no reduction of iterations. Here, typically the correct active set is determined in the last iteration (for large enough γ).
- The dependence of the iteration number on the mesh size of the discretization for $P1$ are depicted in Table 2 (the ones for $P2$ and $P3$ are similar). In parenthesis we show the number of inner iterations. The results clearly indicate that the outer

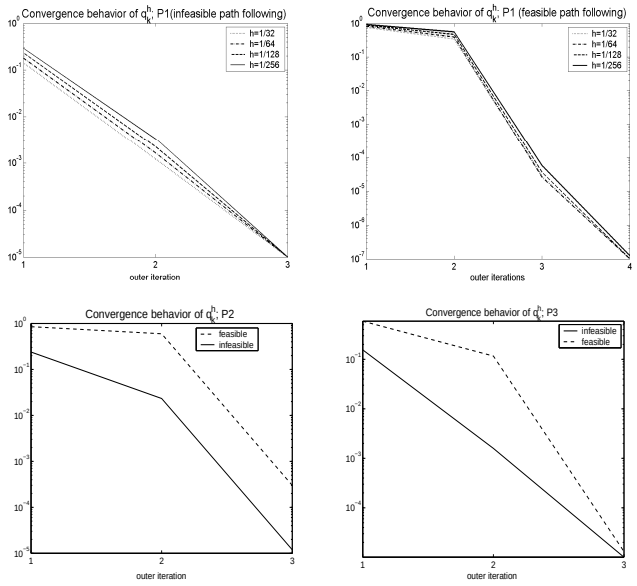

FIGURE 6. Discrete quotients q_k^h for $P1$ and various mesh sizes h (upper row) and for $P2$ (lower left) and $P3$ (lower right) for $h = 1/128$.

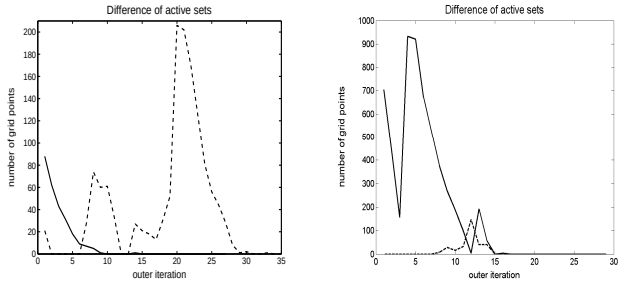

FIGURE 7. Difference in active sets for $P1$ and $P2$; solid line – infeasible case; dashed line – feasible case.

iterations are mesh independent, while the number of inner iterations increases as the mesh size decreases. In the third row we display the results obtained by applying Algorithm A for the solution of the unregularized problem (P) with data according to P1. If we compare these results with the ones of the infeasible exact path following algorithm, we find that for sufficiently small mesh sizes h the infeasible version of Algorithm EP requires significantly less iterations than Algorithm A, which is also an infeasible algorithm. Also, the number of

iteration required by Algorithm A exhibits a relatively strong dependence on h when compared to Algorithm EP in the infeasible case. Similar observations apply also to $P2$ and $P3$, respectively. This shows that taking into account the function space theoretic properties when regularizing problem (**??**) results in an algorithmic framework which performs stably with respect to decreasing mesh size of discretization.

	Mesh size h				
version	1/16	1/32	1/64	1/128	1/256
EP feasible	5(19)	5(23)	5(30)	5(44)	5(72)
EP infeasible	4(8)	4(11)	4(13)	4(15)	4(19)
Algorithm A	4	8	14	26	48

TABLE 2. Comparison of iteration counts for different mesh sizes.

- From the plots in Figure 8, where the y-axis again has a logarithmic scale, it can be seen that our strategy (45) produces a rapidly increasing sequence $\{\gamma_k\}$. The plots in Figure 8 depict the increase of γ_k as a function of the iteration number. The question arises, whether one could increase γ more rapidly. Numerical examples employing an ad-hoc strategy show that if γ is increased too quickly, then the numerical error may prevent the residuals of the first order system to drop below $\sqrt{\epsilon_M}$. This effect is due to the ill-conditioning of the linear systems for large γ. On the other hand, small increases in γ result in a slow convergence speed of Algorithm EP. Further, in our test runs and as can be seen from Figure 8 the feasible version of Algorithm EP was less aggressive in enlarging γ_k.

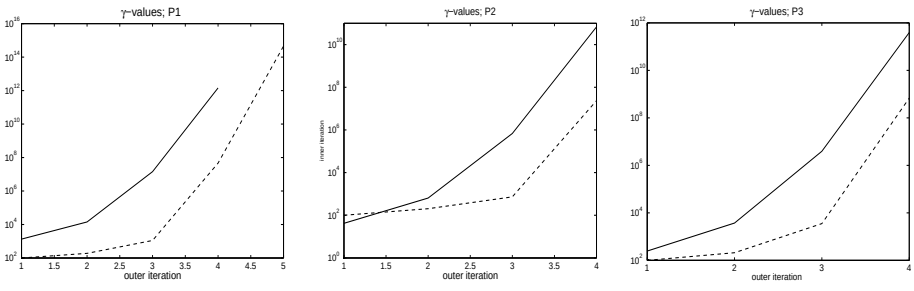

FIGURE 8. γ-updates; solid line – infeasible case; dashed line – feasible case.

6.2. **Inexact path following.** While exact path following is primarily of theoretical interest, the development of inexact path following techniques which keep the number of iterations as small as possible is of more practical importance. The strategy in the previous section relies on the fact that for every γ_k the corresponding point on the primal-dual path is computed. This, however, is not the case for inexact techniques and, as a consequence, a different update strategy for the path parameter γ is necessary. A common concept in inexact path-following methods is based on the definition of an appropriate neighborhood of the path; see, *e.g.*, [3] and the references therein for a non-interior neighborhood-based path-following method, or [5, 16, 18, 19] for path-following techniques related to interior point methods. It is typically required that the primal-dual iterates stay within the neighborhood of the path, with the goal to reduce the computational burden while still maintaining convergence of the method.

We define

$$r_\gamma^1(y,\lambda) = \|-\Delta y+\lambda-f\|_{H^{-1}}, \tag{51a}$$

$$r_\gamma^2(y,\lambda) = \|\lambda-\max(0,\lambda+\gamma(y-\psi))\|_{H^{-1}}, \tag{51b}$$

and the neighborhood:

$$\mathcal{N}(\gamma) := \{(y,\lambda)\in H_0^1(\Omega)\times L^2(\Omega) : \|(r_\gamma^1(y,\lambda), r_\gamma^2(y,\lambda))^\top\|_2 \le \frac{\tau}{\sqrt{\gamma}}\} \tag{52}$$

in the infeasible case and

$$\mathcal{N}(\gamma) := \{(y,\lambda)\in H_0^1(\Omega)\times L^2(\Omega) :\|(r_\gamma^1(y,\lambda), r_\gamma^2(y,\lambda))^\top\|_2 \le \frac{\tau}{\sqrt{\gamma}} \wedge \frac{\partial}{\partial\gamma}J(y;\gamma)\le 0\} \tag{53}$$

in the feasible case. Above $\tau>0$ denotes some fixed parameter. Note that adding the condition $\frac{\partial}{\partial\gamma}J(y;\gamma)\ge 0$ in (52) yields no further restriction, since this condition is automatically satisfied by the structure of $J(y;\gamma)$. We also point out that the conditions on the derivative of $J(y;\gamma)$ are included in (52) and (53), respectively, in order to qualitatively capture (up to first order) the analytical properties of the primal-dual path.

Next we specify our framework for an **i**nexact **p**ath following algorithm.

Algorithm IP.

(i) Initialize γ_0 according to (49) in the infeasible case or (50) in the feasible case; set $k := 0$.

(ii) Apply Algorithm B to find $(y_{k+1}, \lambda_{k+1}) \in \mathcal{N}(\gamma_k)$.
(iii) Update γ_k to obtain γ_{k+1}.
(iv) Set $k = k + 1$, and go to (ii).

Note that if in step (ii) the path-problem (P_γ) is solved, then $r^1_\gamma(y_\gamma, \lambda_\gamma) = r^2_\gamma(y_\gamma, \lambda_\gamma) = 0$.

As it is the case with primal-dual path-following interior point methods, the update strategy for γ in step (iii) of Algorithm IP is a delicate issue. If the increase of γ from one iteration to the next is rather small, then we follow the path closely and the convergence speed is slow. If the γ-update is too aggressive, then step (ii) requires many iterations of Algorithm B to produce iterates in the neighborhood. We propose the following strategy which performed very well in our numerical tests.

We introduce the *primal infeasibility measure* ρ^F, and the *complementarity measure* ρ^C as follows

$$\rho^F_{k+1} := \int_\Omega (y_{k+1} - \psi)^+ dx, \tag{54}$$

$$\rho^C_{k+1} := \int_{\mathcal{I}_{k+1}} (y_{k+1} - \psi)^+ dx + \int_{\mathcal{A}_{k+1}} (y_{k+1} - \psi)^- dx, \tag{55}$$

where $(\cdot)^- = -\min(0, \cdot)$ and $(\cdot)^+ = \max(0, \cdot)$. Note that at the optimal solution both measures vanish. Further we point out that ρ^C is related to the duality measure well-known from primal-dual path following interior point methods. These measures are used in the following criterion for updating γ:

$$\gamma_{k+1} \geq \max\left(\gamma_k \max\left(\tau_1, \frac{\rho^F_{k+1}}{\rho^C_{k+1}}\right), \frac{1}{(\max(\rho^F_{k+1}, \rho^C_{k+1}))^q}\right) \tag{56}$$

with $\tau_1 > 1$, and $q \geq 1$. The first term in the outermost max-expression is used because of our observation that $\rho^F_{k+1} \geq \rho^C_{k+1}$ in the infeasible case. If ρ^C is small compared to ρ^F we find that the iterates primarily lack feasibility as compared to complementarity. Therefore, a strong increase in γ which aims at reducing constraint infeasibility is favorable. If both measures are of almost the same size and rather small, then the second term in the outer max expression should yield a significant increase in γ. Typically $q \in [\frac{3}{2}, 2]$ is chosen which induces growth rates for γ.

If there is still a significant change in the active sets from one iteration to the next and the update γ_{k+1} based on (56) would be too large compared to γ_k, then many inner iterations would be necessary to

keep track of the path or very conservative γ-updates in the following iterations have to be chosen. We safeguard the γ-updates by utilizing our model function $m(\gamma)$, which was found to be a reliable tool. In fact, in updating γ large deviations from $m(\gamma)$ are prohibited by comparing the value of the tangent to $J(y;\gamma)$ at $\gamma = \gamma_k$ with the actual model value. If necessary and as long as γ_{k+1} is much larger than γ_k, we reduce the actual γ-value until

$$|t_k(\gamma_{k+1}) - m_k(\gamma_{k+1})| \leq \tau_3 |J(y_{k+1};\gamma_k) - J(y_k;\gamma_{k-1})| \tag{57}$$

with $0 < \tau_3 < 1$, $t_k(\gamma) = J(y_{k+1};\gamma_k) + \frac{\partial J}{\partial \gamma}(y_{k+1};\gamma_k)(\gamma - \gamma_k)$, and $m_k(\gamma)$ the model related to γ_k. Recall that $m_k(\gamma_k) = J(y_{k+1};\gamma_k)$. The motivation of this strategy utilizes the good approximation qualities of our models. Indeed, for small γ the distance between t_k and m_k might be large, but so is $|J(y_{k+1};\gamma_k) - J(y_k;\gamma_{k-1})|$ since the change in the function value is expected to be relatively large for small γ. For large γ, however, both difference measures tend to be small.

Concerning the numerical realization of Algorithm IP in the discrete setting we point out that by an a posteriori analysis of the discretization errors one finds that the norm of the residuals in (51a) and (51b) can be approximated typically to the order of h. This can be used as an upper bound for γ in the discrete versions of (52) and (53), respectively. However, since, on a fixed grid, our discrete versions of (P) and (P_γ) are consistent (as $\gamma \to \infty$) and admit unique solutions in $\mathbb{R}^{N_h}$, where $N_h \in \mathbb{N}$ depends on the mesh size of discretization h, it is of interest to consider $\gamma \to \infty$. On a fixed grid, this allows us also to study the behavior of our discretized algorithms as finite dimensional solvers for problems similar to the discrete versions of the ones under consideration. With respect to the latter aspect, below we report on test runs of Algorithm IP when applied to our test problems $P1$, $P2$, and $P3$. The parameters had values $q = 1.5$, $\tau_1 = 10$, $\tau_3 = 0.999$, $\tau = 1e6$. The stopping rule for the outer iteration is as before.

$P1$. The infeasible version of Algorithm IP requires 9 outer iterations and at most two inner iterations per outer iteration. In particular, in many iterations the criterion $(y_{k+1}, \lambda_{k+1}) \in \mathcal{N}(\gamma_k)$ was satisfied within one inner iteration. The feasible version of Algorithm IP stops after 11 iterations. With respect to inner iterations in the feasible case we note that more than one or two inner iterations were necessary only in the last three outer iterations with 3, 4, and 6 inner iterations, respectively. For both runs, the behavior of the measures ρ^F and ρ^C is shown in Figure 9. Note that the vertical scale is a logarithmic one. The left plot corresponds to the infeasible case. The feasibility measure ρ^F and the complementarity measure ρ^C are both convergent

at a superlinear rate. In the feasible case, which is depicted in the right plot, we observe that ρ^C is only linearly convergent. In some iterations we have $\rho_k^F > 0$. However, the constraint violation is of the order of the machine precision and, thus, negligible.

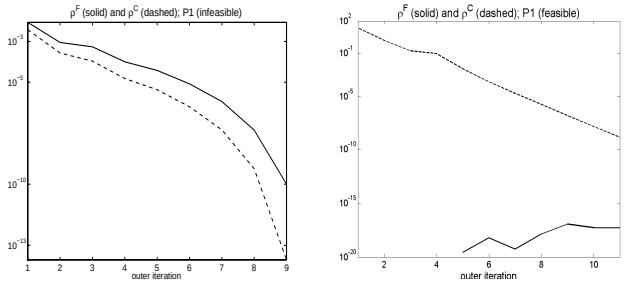

FIGURE 9. Behavior of the measure ρ^F (solid) and ρ^C (dashed) for $P1$; left plot – infeasible case; right plot – feasible case.

$P2$. For this test problem the infeasible version of Algorithm IP required 11 iterations with one inner iteration per outer iteration. The feasible version needed 6 outer iterations and 9 inner iterations.

$P3$. The behavior of Algorithm IP for solving $P3$ is comparable to its behavior for $P1$ and $P2$. In fact, the infeasible version required 11 outer iterations and 11 inner iterations for solving the discrete problem. The feasible variant of Algorithm IP stopped successfully after 9 outer and 19 inner iterations. For the latter run, in the next to the last iteration 5 inner iterations were necessary; otherwise at most 2 inner iterations were needed. With respect to the behavior of the decrease of the measures ρ^C and ρ^F a similar observation to the one obtained from Figure 9 for $P1$ holds true. We only remark that in the feasible case ρ^C exhibits an almost superlinear convergence behavior.

Compared to the exact path-following strategy of Algorithm EP, the inexact path-following concept of Algorithm IP is in many cases more efficient. In Table 3 we provide the number of outer and inner iterations for exact vs. inexact path following. In parenthesis we write the number of inner iterations.

Finally we address the issue of how to incorporate the approximation error due to the discretization of function space quantities; see [6, 7]. First note that with (8) holding (which is the case for $P3$) the discretization of the residual in the definition of the neighborhoods (52) respectively (53) approximates the original one to the order of h. Hence, in our discrete version of Algorithm IP the neighborhood

	Infeasible case			Feasible case		
	$P1$	$P2$	$P3$	$P1$	$P2$	$P3$
EP	4 (15)	4 (11)	4 (16)	5 (44)	4 (10)	4 (31)
IP	9 (12)	11 (11)	11 (11)	11 (25)	6 (9)	9 (19)

TABLE 3. Comparison of iteration counts between exact and inexact path following.

criterion

$$\|(r_\gamma^1(y,\lambda), r_\gamma^2(y,\lambda))^\top\|_2 \leq \frac{\tau}{\sqrt{\gamma}}$$

becomes

$$\|(r_\gamma^{1,h}(y,\lambda), r_\gamma^{2,h}(y,\lambda))^\top\|_2 \leq \max\left\{\sqrt{\epsilon_M}, \kappa_{\text{in}} h, \frac{\tau}{\sqrt{\gamma}}\right\},$$

with some constant $\kappa_{\text{in}} > 0$. We stop the outer iteration as soon as the discrete residual drops below $\max\{\kappa_{\text{out}} h, \sqrt{\epsilon_M}\}$ where $\kappa_{\text{out}} > 0$ is fixed. In our tests we use $\kappa_{\text{in}} = 1$ and $\kappa_{\text{out}} = 10$. Applying this strategy for the solution of $P3$, we obtain (outer) iteration numbers as displayed in Table 4. Here, in parenthesis we denote the total number of inner iterations.

version	mesh size					
	1/16	1/32	1/64	1/128	1/256	1/512
IP	1 (1)	4 (4)	5 (5)	8 (8)	9 (10)	10 (10)

TABLE 4. Inexact path following with h-dependent stopping of inner and outer iterations.

REFERENCES

[1] M. Bergounioux, M. Haddou, M. Hintermüller, 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(2):495–521, 2000.

[2] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002.

[3] X. Chen and P. Tseng. Non-interior continuation methods for solving semidefinite complementarity problems. *Math. Program.*, 95(3, Ser. A):431–474, 2003.

[4] A. V. Fiacco and G. P. McCormick. *Nonlinear Programming*, volume 4 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990. Sequential unconstrained minimization techniques.

[5] A. Forsgren, P. E. Gill, and M. H. Wright. Interior methods for nonlinear optimization. *SIAM Rev.*, 44(4):525–597, 2002.
[6] C. Großmann and A. A. Kaplan. On the solution of discretized obstacle problems by an adapted penalty method. *Computing*, 35(3-4):295–306, 1985.
[7] C. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen*. Teubner Studienbücher Mathematik. [Teubner Mathematical Textbooks]. B. G. Teubner, Stuttgart, second edition, 1994.
[8] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Teubner Verlag, Stuttgart, 1986.
[9] M. Hintermüller. Inverse coefficient problems for variational inequalities: optimality conditions and numerical realization. *M2AN Math. Model. Numer. Anal.*, 35(1):129–152, 2001.
[10] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set strategy as a semi-smooth Newton method. *SIAM J. Optim.*, 13(3):865–888, 2003.
[11] K. Ito and K. Kunisch. Optimal control of elliptic variational inequalities. *Appl. Math. Optim.*, 41(3):343–364, 2000.
[12] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems Control Lett.*, 50(3):221–228, 2003.
[13] K. Ito and K. Kunisch. Semi-smooth Newton methods for variational inequalities of the first kind. *M2AN Math. Model. Numer. Anal.*, 37(1):41–62, 2003.
[14] G. M. Troianiello. *Elliptic differential equations and obstacle problems*. The University Series in Mathematics. Plenum Press, New York, 1987.
[15] M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.*, 13(3):805–842, 2003.
[16] R. J. Vanderbei. *Linear Programming*. International Series in Operations Research & Management Science, 37. Kluwer Academic Publishers, Boston, MA, second edition, 2001. Foundations and extensions.
[17] M. Weiser. Interior point methods in function space. ZIB-Report 03-35, Konrad-Zuse Zentrum für Informationstechnik Berlin, 2003.
[18] S. J. Wright. *Primal-dual Interior-Point Methods*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.
[19] Y. Ye. *Interior Point Algorithms*. John Wiley & Sons Inc., New York, 1997. Theory and analysis, A Wiley-Interscience Publication.