The Primal-Dual Active Set Strategy as a Semismooth Newton Method

M Hintermüller, K. Ito, K. Kunisch

THE PRIMAL-DUAL ACTIVE SET STRATEGY AS A SEMISMOOTH 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 semismooth 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.

Key words. complementarity problems, function spaces, semismooth Newton method

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 semismooth 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 semismooth 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 therein. Typically, under appropriate semismoothness and regularity assumptions locally superlinear convergence rates of semismooth 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 semismoothness in infinite dimensional spaces were introduced. In our work we primarily use the notion of slant differentiability from [CNQ] which we recall for the

[†]Institute of Mathematics, University of Graz, A-8010 Graz, Austria (michael.hintermueller@uni-graz.at, karl.kunisch@uni-graz.at).

[‡]North Carolina State University, Raleigh, NC 27695 (kito@unity.ncsu.edu).

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 semismooth 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., [Ca]) 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 semismooth 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 a 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. While this discussion is perhaps more philosophical than mathematical, we mention it because 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 semismooth 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 the case where X is infinite dimensional.

Let us put the concept of slant differentiability into perspective with the notion of semismoothness as introduced in [Mi] for real-valued functions and extended in [QS] to finite dimensional vector-valued functions. Semismoothness 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 semismooth 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 the case where A is symmetric positive definite (2.1) is the optimality system for

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

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 componentwise.

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

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

This leads to the following algorithm.

PRIMAL-DUAL ACTIVE SET ALGORITHM.

(i) Initialize y^0, λ^0. Set $k = 0$.

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

(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 semismooth 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

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

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. Semismooth Newton methods for generalized Jacobians in Clarke's sense were considered, e.g., in [Q1, QS].

The choice G_m suggests a semismooth 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 semismooth 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 semismooth 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

$$\left\{ \begin{array}{l} Ay + \lambda = f, \\ \lambda - \max(0, \lambda + c(y - \psi)) = 0, \end{array} \right. \tag{3.1}$$

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 componentwise. 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

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

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

$$g_i(z) = \left\{ \begin{array}{lll} 0 & \text{if} & z < 0\,, \\ 1 & \text{if} & z > 0\,, \\ \delta_i & \text{if} & z = 0\,. \end{array} \right.$$

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 semismooth 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

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

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 semismooth Newton method converge superlinearly to $x^* = (y^*, \lambda^*)$, provided that $\|x^0 - x^*\|$ is sufficiently small.*

The boundedness requirement of $(G_F)^{-1}$ according to Theorem 1.1 can be derived analogously to the infinite dimensional case; see the proof of Theorem 4.1.

In our finite dimensional setting this result can be obtained alternatively by observing that G_m corresponds to a generalized Jacobian in Clarke's sense combined with the convergence results for semismooth 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 semismoothness of F at x^*. Hence, Theorem 3.2 of [QS] yields the locally superlinear convergence property. For a discussion of the semismoothness concept in finite dimensions we refer the reader to [Q1, QS].

Furthermore, since (3.1) is strongly semismooth, 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 semismoothness we refer the reader 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 an M-matrix, if it is nonsingular, $(m_{ij}) \le 0$, for $i \ne j$, and $M^{-1} \ge 0$. Our notion of an M-matrix coincides with that of nonsingular M-matrices as defined in [BP].

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

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. □

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 one-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 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 a 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 the case where A is an 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 an 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

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

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)

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

Note that $A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} \leq M_{\mathcal{I}}^{-1} K_{\mathcal{I}\mathcal{A}} - M_{\mathcal{I}}^{-1} K_{\mathcal{I}} (M+K)_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}}$. 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 $M_{\mathcal{I}}^{-1} M_{\mathcal{I}\mathcal{A}} \leq 0$. Since $y^k \geq \psi$ 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, Thm. 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 a *good* prediction, 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 semismooth Newton method, then the convergence analysis is based on the identity

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

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

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

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 $\Psi(x^0) = 0$. ∎

Finally, we shall point out that Theorems 3.2–3.4 establish global convergence of the primal-dual active set strategy or, equivalently, semismooth 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 semismooth 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 pointwise 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 self-adjoint, 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 pointwise 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 semismooth 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 semismooth Newton updates coincide, provided that the slanting function of the max-function is taken according to

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

which we henceforth assume.

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

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

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

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

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

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

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

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

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

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

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

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

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

At this point we remark that the primal-dual active set strategy has no straightforward 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 the reader to [Ca]. Theorem 3.2 proves global convergence of the primal-dual active set strategy or, equivalently, semismooth 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

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

Again the multiplier associated with 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 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 semismooth 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 semismoothness of composite mappings as developed in [U]. For this purpose we first recall the notion of semismoothness 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 semismooth 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

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

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 semismooth in the sense of [QS]. Consequently, $\tilde{\Psi}$ is semismooth 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

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

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

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

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 semismooth 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

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

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 semismoothness 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 us 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

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

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

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

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

$$\left(\int_{\Omega_0(h)} |w(x)|^{2\sigma} dx\right)^{1/2\sigma} \leq \left(\int_{\Omega_0(h)} |w(x)|^{\frac{2\sigma q}{r}} |w(x)|^{\frac{2\sigma(r-q)}{r}} dx\right)^{1/2\sigma}$$
$$\leq \left(\int_{\Omega_0(h)} |w(x)|^{\frac{2\sigma q}{r}\frac{r}{2\sigma}}\right)^{1/r} \left(\int_{\Omega_0(h)} |w(x)|^{\frac{2\sigma(r-q)}{r}\frac{r}{r-2\sigma}}\right)^{(r-2\sigma)/2r\sigma}$$
$$= \left(\int_{\Omega_0(h)} |w(x)|^q dx\right)^{1/r} \left(\int_{\Omega_0(h)} \frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}} dx\right)^{(r-2\sigma)/2r\sigma},$$

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

$$\frac{1}{\|h\|_{L^q}^{1+\alpha}} \|D_{w,h}\|_{L^2}$$
$$\leq (1+|\delta|)(|\Omega_\epsilon(h)|^{1/2\tau} + |\Omega_\epsilon(w)|^{1/2\tau}) \left(\int_{\Omega_0(h)} \frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}} dx\right)^{(r-2\sigma)/2r\sigma}.$$

Suppose that

$$\int_{\{x:|w(x)|\neq 0\}} \frac{1}{|w(x)|^{\frac{2\sigma(q-r)}{r-2\sigma}}} dx < +\infty\,. \tag{4.16}$$

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

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

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 most 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 $\sigma > 1$

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

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

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

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 semismooth 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 to 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 involve only the system matrix A, whereas $\mathcal{M}(y^k)$ also depends on the specific choice of f and ψ. □

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). □

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 were used for second order sufficient optimality conditions and the analysis of SQP-methods in [M, I, IK3], for example, and also for semismooth Newton methods in [U]. □

In view of the fact that (P) consists 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 eight 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.

Table 1

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

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 iteration. Increasing β results in a further reduction of three iterations; see Tables 1 and 2.

TABLE 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	–	–	–

Appendix A.

Proof of Theorem 3.2. The assumption that A is an 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 $\delta y_{\mathcal{A}_k} = \psi_{\mathcal{A}_k} - y_{\mathcal{A}_k}^k \leq 0$. For $i \in \mathcal{I}_k$ we have $\lambda_i^k + c(y_i^k - \psi_i) \leq 0$ which implies $\delta\lambda_{\mathcal{I}_k} \geq 0$ by (2.4) and (A.1). Since $\delta y_{\mathcal{I}_k} = -A_{\mathcal{I}_k}^{-1} A_{\mathcal{I}_k \mathcal{A}_k} \delta y_{\mathcal{A}_k} - A_{\mathcal{I}_k}^{-1} \delta\lambda_{\mathcal{I}_k}$ 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 $V = \{i : y_i^1 > \psi_i\}$. For $i \in V$ we have $\lambda_i^1 = 0$ by (A.1), and hence $\lambda_i^1 + c(y_i^1 - \psi_i) > 0$ 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 $\lambda_i^{\bar{k}} < 0$. Then necessarily $i \in \mathcal{A}_{\bar{k}-1}$, $y_i^{\bar{k}} = \psi_i$, and $\lambda_i^{\bar{k}} + c(y_i^{\bar{k}} - \psi_i) < 0$. It follows that $i \in \mathcal{I}_{\bar{k}}$, $\lambda_i^{\bar{k}+1} = 0$, and $\lambda_i^{\bar{k}+1} + c(y_i^{\bar{k}+1} - \psi_i) \leq 0$, since $y_i^{k+1} \leq \psi_i$, $k \geq 1$. 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 $\lambda_i^{\bar{k}} < 0$ for some coordinate i at iteration $\bar{k}$, then $\lambda_i^k = 0$ and $y_i^k \leq \psi_i$ 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

$$\Omega_0(h) = \{x \in \Omega : y(x) \neq 0,\ y(x)(y(x)+h(x)) \leq 0\}$$

requires further investigation. For $\epsilon > 0$ we define subsets of $\Omega_0(h)$ 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

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

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

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

Note that $\Omega^\epsilon(y) \subset \Omega^{\epsilon'}(y)$ whenever $0 < \epsilon \le \epsilon'$ 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

$$\begin{aligned}
\frac{1}{\|h\|_{L^q}}\|D_{y,h}\|_{L^p} &\le \frac{1+|\delta|}{\|h\|_{L^q}} \left(\int_{\Omega_0(h)} |y(x)|^p dx\right)^{1/p} \\
&\le \frac{1+|\delta|}{\|h\|_{L^q}} \left[\left(\int_{\Omega_\epsilon(h)} |y(x)|^p dx\right)^{1/p} + \left(\int_{\Omega_0(h)\setminus\Omega_\epsilon(h)} |y(x)|^p dx\right)^{1/p}\right] \\
&\le \frac{1+|\delta|}{\|h\|_{L^q}} \left[|\Omega_\epsilon(h)|^{(q-p)/(qp)} \left(\int_{\Omega_\epsilon(h)} |y(x)|^q dx\right)^{1/q} \right. \\
&\qquad \left. + |\Omega^\epsilon(y)|^{(q-p)/(qp)} \left(\int_{\Omega_0(h)\setminus\Omega_\epsilon(h)} |y(x)|^q dx\right)^{1/q}\right] \\
&\le (1+|\delta|)\left(|\Omega_\epsilon(h)|^{(q-p)/(qp)} + |\Omega^\epsilon(y)^{(q-p)/(qp)}|\right)\,.
\end{aligned}$$

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,

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

and by (A.3)

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

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

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

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

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

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 $C_{\mathcal{I}\mathcal{A}} = E_{\mathcal{I}}^* C E_{\mathcal{A}}$. □

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. Optim., 11 (2000), pp. 495–521.

[BIK] M. Bergounioux, K. Ito, and K. Kunisch, *Primal-dual strategy for constrained optimal control problems*, SIAM J. Control Optim., 37 (1999), pp. 1176–1194.

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

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

[CNQ] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.

[C] F.H. Clarke, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, 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*, Appl. Math. Optim., 40 (1999), pp. 19–37.

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

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

[H] R.H.W. Hoppe, *Multigrid algorithms for variational inequalities*, SIAM J. Numer. Anal., 24 (1987), pp. 1046–1065.

[I] A.D. Ioffe, *Necessary and sufficient conditions for a local minimum.* 3: *Second order conditions and augmented duality*, SIAM J. Control Optim., 17 (1979), pp. 266–288.

[IK1] K. Ito and K. Kunisch, *Augmented Lagrangian methods for nonsmooth convex optimization in Hilbert spaces*, Nonlinear Anal., 41 (2000), pp. 573–589.

[IK2] K. Ito and K. Kunisch, *Optimal control of elliptic variational inequalities*, Appl. Math. Optim., 41 (2000), pp. 343–364.

[IK3] K. Ito and K. Kunisch, *Newton's method for a class of weakly singular optimal control problems*, SIAM J. Optim., 10 (2000), pp. 896–916.

[KS] C.T. Kelley and E.W. 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 Advances in Optimization Math. Res. 45, J. Guddat et al., eds., 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, and P. Tarvainen, *Overlapping block methods for obstacle problems with convection-diffusion operators*, in Complementarity and Variational Problems, M.C. Ferris and J.-S. Pang, eds., SIAM, Philadelphia, 1997.

[Mi] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[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. Oper. Res., 18 (1993), pp. 227–244.

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

[QS] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. 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. Control Optim., 38 (2000), pp. 1938–1984.

[U] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–842.