ANZIAM J. 49(2007), 1–38

MESH INDEPENDENCE AND FAST LOCAL CONVERGENCE OF A PRIMAL-DUAL ACTIVE-SET METHOD FOR MIXED CONTROL-STATE CONSTRAINED ELLIPTIC CONTROL PROBLEMS

M. HINTERMÜLLER[1]

(Received 28 April, 2007)

Abstract

A class of mixed control-state constrained optimal control problems for elliptic partial differential equations arising, for example, in Lavrentiev-type regularized state constrained optimal control is considered. Its numerical solution is obtained via a primal-dual active-set method, which is equivalent to a class of semi-smooth Newton methods. The locally superlinear convergence of the active-set method in function space is established, and its mesh independence is proved. The paper contains a report on numerical test runs including a comparison with a short-step path-following interior-point method and a coarse-to-fine mesh sweep, that is, a nested iteration technique, for accelerating the overall solution process. Finally, convergence and regularity properties of the regularized problems with respect to a vanishing Lavrentiev parameter are considered.

2000 *Mathematics subject classification*: primary 65K05; secondary 90C33.
Keywords and phrases: Active-set strategy, mesh-independence, mixed control-state constraints, PDE constraints, primal-dual methods, semi-smooth Newton methods.

1. Introduction

Recently, there has been significant interest in studying the optimal control (model) problem

$$\left.\begin{array}{lll} \text{minimize} & J(y,u) := \dfrac{1}{2}\|y-y_d\|_{L^2}^2 + \dfrac{\alpha}{2}\|u\|_{L^2}^2 & \\ \text{over} & (y,u) \in H_0^1(\Omega)\times L^2(\Omega), & \\ \text{subject to} & Ay = u + f \qquad \text{in } \Omega, & \\ & a \le c\,u + y \le b \quad \text{almost everywhere (a.e.) in } \Omega, & \end{array}\right\} \tag{1.1}$$

[1] Department of Mathematics and Scientific Computing, University of Graz, Heinrichstr. 36, A-8010 Graz, Austria; e-mail: michael.hintermueller@uni-graz.at.

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

where $\Omega \subset \mathbb{R}^n$ is a bounded domain with sufficiently smooth boundary $\Gamma = \partial\Omega$, and $y_d \in L^2(\Omega)$, $f \in H^{-1}(\Omega)$, $a, b \in L^q(\Omega)$, for some $q > 2$ and with $a < b$, $c \in L^\infty(\Omega)$, with $c \geq \epsilon_c > 0$ a.e.in Ω, and $\alpha > 0$ are given data. Further, A denotes a second-order linear elliptic differential operator. Throughout we use $\|\cdot\|_{L^2} = \|\cdot\|_{L^2(\Omega)}$ and similarly for other function space norms. In what follows, we call y the state and u the control (variable), respectively.

The recent focus on the model problem (1.1) is mainly due to its importance in the context of (purely) state-constrained optimal control problems, which is $c \equiv 0$ in (1.1) with $f \in L^2(\Omega)$, that is,

$$\left.\begin{array}{lll} \text{minimize} & J(y,u) := \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 & \\ \text{over} & (y,u) \in H_0^1(\Omega) \times L^2(\Omega), & \\ \text{subject to} & Ay = u + f \quad \text{in } \Omega, & \\ & a \leq y \leq b \quad \text{almost everywhere (a.e.) in } \Omega. & \end{array}\right\} \tag{1.2}$$

For this problem, it is well known that the Lagrange multiplier associated with the pointwise almost everywhere state constraints is a Borel measure only; see [7, 9]. Consequently, numerical algorithms such as projected Newton or semismooth Newton techniques suffer from a mesh-dependent behaviour and typically admit no function space analysis. As a remedy, in [17] a Lavrentiev-type regularization of pointwise state constraints is proposed. The resulting regularized problem is of the type (1.1) for some small, but fixed, $c(\mathbf{x}) = \epsilon > 0$. For its numerical solution a short-step primal-dual path-following interior-point method is applied.

An alternative path-following concept for the solution of (1.2) can be found in [13]. It is based on a generalized Moreau-Yosida-type regularization, that is, it replaces (1.2) by the approximate problem

$$\left.\begin{array}{ll} \text{minimize} & J(y,u) + \frac{1}{2\gamma}\left(\left\|\max\left(0, \tilde{\lambda}_b + \gamma(y-b)\right)\right\|_{L^2}^2 \right. \\ & \qquad \left. + \left\|\max\left(0, \tilde{\lambda}_a + \gamma(a-y)\right)\right\|_{L^2}^2\right) \\ \text{over} & (y,u) \in H_0^1(\Omega) \times L^2(\Omega), \\ \text{subject to} & Ay = u + f \quad \text{in } \Omega, \end{array}\right\} \tag{1.3}$$

where $\gamma > 0$ represents a regularization parameter, and $\tilde{\lambda}_a, \tilde{\lambda}_b \geq 0$ are fixed shift-parameters in $L^2(\Omega)$. The regularized problem (1.3) is solved efficiently by a semismooth Newton method (SSN). Note that the objective function of (1.3) is related to an augmented Lagrangian penalization technique (of the pointwise inequality constraints) with γ representing the corresponding penalty parameter and with $\tilde{\lambda}_a, \tilde{\lambda}_b$ being approximations of the Lagrange multipliers associated with the pointwise inequality constraints. In our function space context, in addition to the penalization

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

aspect, replacing (1.2) by (1.3) induces a regular approximation of the Lagrange multipliers of the inequality constraints in (1.2). In fact, let $0 \leq \bar{\lambda}_a \in \mathscr{M}(\Omega)$, with $\mathscr{M}(\Omega)$ representing the set of regular Borel measures on Ω, denote the Lagrange multiplier associated with $a \leq \bar{y}$ at the solution $\bar{y} \in H_0^1(\Omega) \cap H^2(\Omega)$ of (1.2). Then, in the context of (1.3), the quantity

$$\lambda_a^\gamma = \max\left(0, \tilde{\lambda}_a + \gamma(a - y^\gamma)\right) \in L^2(\Omega)$$

is a regular approximation of $\bar{\lambda}_a$. Here y^γ denotes the optimal solution of (1.3). It can be shown that $\lambda_a^\gamma \rightharpoonup \bar{\lambda}_a$ as $\gamma \to \infty$ in $\left(H_0^1(\Omega) \cap H^2(\Omega)\right)^*$. For details on this approach we refer the reader to [13].

We point out that replacing (1.2) by the Lavrentiev-regularized problem (1.1) also induces a regularization of the Lagrange multipliers associated with the pointwise inequality constraints. In the next section we shall see that the multipliers for the inequality constraints in (1.1), like the ones for (1.3), exist as $L^2(\Omega)$-functions, respectively.

The present research is motivated by our numerical experience which shows that the SSN, or equivalently the primal-dual active-set method (pdAS), is typically superior to path-following interior point algorithms [6, 13]. This claim relies on the fact that the convergence of the SSN can be proved in function space. In this case, the convergence rate of the SSN, respectively the pdAS, is typically locally q-superlinear. Accordingly, one goal of this paper is to show that the SSN can be used as a solver for (1.1) (instead of the short-step path-following method in [17]) and that it converges locally superlinearly in the appropriate function spaces.

For Newton's method applied to smooth operator equations it is known that it exhibits a mesh-independent behaviour [2]. In the presence of pointwise inequality constraints, in [3] the mesh independence of Newton's method for generalized equations is shown. In fact, based on Robinson's generalized equations technique [21–23], for the numerical solution of constrained nonlinear optimal control problems by a sequential quadratic programming (SQP) method, the analysis in [3] establishes the mesh independence of the SQP-iteration (outer iteration). This, however, does not include the corresponding mesh-independence result for the inner iteration for solving the quadratic programming (QP) sub-problem of every SQP-iteration. In the context of optimization problems with partial differential equation (PDE) constraints and pointwise control constraints, in the recent paper [14] this gap is closed by proving a mesh-independence result when using semismooth Newton methods as QP-solvers. As the method we are proposing for solving (1.1) is of SSN-type and in view of the above results, a second focus of the present paper is on proving the mesh independence of our semismooth Newton iteration, or equivalently of the primal-dual active-set method for solving (1.1).

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

In the case where (1.1) comes from a Lavrentiev-type regularization of (1.2) with $c \equiv \epsilon_n$ and $\epsilon_n > 0$, in [17] it is shown that for $\epsilon_n \downarrow 0$ the sequence $\{(\bar{y}, \bar{u})\} = \{(\bar{y}(\epsilon_n), \bar{u}(\epsilon_n))\}$ converges strongly in $L^2(\Omega) \times L^2(\Omega)$ to $(y^*, u^*) \in H_0^1(\Omega) \times L^2(\Omega)$, the optimal solution of (1.2). In the present paper we improve this result and establish a rate of convergence for $\epsilon_n \downarrow 0$. Moreover, the Hölder continuity of $\|\bar{y}(\epsilon^1) - \bar{y}(\epsilon^2)\|_{L^2}$ and $\|\bar{u}(\epsilon^1) - \bar{u}(\epsilon^2)\|_{L^2}$ with respect to $\epsilon^i > 0$, $i = 1, 2$, is argued. These latter findings are of interest in connection with path-following or homotopy approaches to the solution of (1.2) based on the Lavrentiev regularization concept.

The rest of the paper is organized as follows: In Section 2 we study the first-order optimality system associated with (1.1). The following Section 3 concentrates on the primal-dual active-set method, or equivalently the SSN, as a solution technique for (1.1). Section 4 is devoted to the mesh-independence analysis of the primal-dual active-set method. Section 5 contains a report on numerical results. Finally, for $c \equiv \epsilon > 0$ and $\epsilon \downarrow 0$ in Section 6 the convergence behaviour of the regularized solution to the solution of the state constrained problem is studied and a Hölder continuity result is established.

2. First-order optimality system

In [17] it was observed that, by a simple transformation, the problem (1.1) can be cast as a control constrained optimal control problem. For this purpose let ι_0 denote the compact embedding operator of $H_0^1(\Omega)$ into $L^2(\Omega)$, and define $\iota_{-1} : L^2(\Omega) \to H^{-1}(\Omega)$ by $\iota_{-1} = \iota_0^*$. Strictly speaking, we have $Ay = f + \iota_{-1}u$ as the state equation in (1.1) with $A : H_0^1(\Omega) \to H^{-1}(\Omega)$. Now let $A^{-1} : H^{-1}(\Omega) \to H_0^1(\Omega)$ denote the solution operator of $Ay = f + \iota_{-1}u$ in $H^{-1}(\Omega)$, that is, $y = \mathfrak{y}(u) = A^{-1}(\iota_{-1}u + f)$. Define $T = \iota_0 A^{-1} \iota_{-1}$, then $T : L^2(\Omega) \to L^2(\Omega)$ is a compact operator. Further observe that by Riesz-Schauder theory we infer that the Fredholm-operator $F := (c \operatorname{id} + T)$ admits a continuous inverse $F^{-1} : L^2(\Omega) \to L^2(\Omega)$.

As a consequence, by defining the transformed control variable

$$v = Fu,$$

Problem (1.1) becomes

$$\left.\begin{array}{ll} \text{minimize} & \hat{J}(v) := \dfrac{1}{2}\left\|TF^{-1}v - \tilde{y}_d\right\|_{L^2}^2 + \dfrac{\alpha}{2}\left\|F^{-1}v\right\|_{L^2}^2 \\ \text{over} & v \in L^2(\Omega), \\ \text{subject to} & \tilde{a} \le v \le \tilde{b} \quad \text{a.e. in } \Omega, \end{array}\right\} \tag{2.1}$$

where

$$\tilde{y}_d = y_d - A^{-1}f, \quad \tilde{a} = a - A^{-1}f \quad \text{and} \quad \tilde{b} = b - A^{-1}f$$

neglecting embedding operators. Note that the objective function in (2.1) is uniformly convex and continuously Fréchet-differentiable in $L^2(\Omega)$. Further, the feasible set is closed and convex. Thus, standard arguments guarantee the existence of a unique solution $\bar{v} \in L^2(\Omega)$ of (2.1). Given $\bar{v}$ we can reconstruct the unique solution $(\bar{y}, \bar{u})$ of (1.1) by

$$\bar{u} = F^{-1}\bar{v} \quad \text{and} \quad \bar{y} = A^{-1}(\iota_{-1}\bar{u} + f).$$

Our algorithmic considerations in the subsequent sections will be based on the transformed problem (2.1), more specifically, on its first-order optimality system, which we state next. Its proof follows from standard arguments; see, for example, [17, 24]. For the formulation, for $w, z \in L^2(\Omega)$, we use

$$0 \leq w \perp z \geq 0 \quad \Longleftrightarrow \quad w \geq 0,\ z \geq 0,\ w\,z = 0 \quad \text{a.e. in } \Omega,$$

and F^{-*} for $(F^*)^{-1}$, which is the inverse of the adjoint operator of F.

THEOREM 2.1. *The optimal solution* $\bar{v} \in L^2(\Omega)$ *of* (2.1) *is characterized by the existence of* $(\bar{\lambda}_a, \bar{\lambda}_b) \in L^2(\Omega) \times L^2(\Omega)$ *satisfying*

$$F^{-*}(T^*T + \alpha\,\mathrm{id})F^{-1}\bar{v} + \bar{\lambda}_b - \bar{\lambda}_a = F^{-*}T^*\tilde{y}_d, \tag{2.2}$$

$$0 \leq \bar{\lambda}_a \perp (\bar{v} - \tilde{a}) \geq 0, \tag{2.3}$$

$$0 \leq \bar{\lambda}_b \perp (\tilde{b} - \bar{v}) \geq 0. \tag{2.4}$$

Applying F^* to (2.2) and inserting $\bar{u} = F^{-1}\bar{v}$ yields

$$(T^*T + \alpha\,\mathrm{id})\bar{u} + F^*(\bar{\lambda}_b - \bar{\lambda}_a) = T^*\tilde{y}_d. \tag{2.5}$$

Expanding F^* gives

$$T^*(T\bar{u} - \tilde{y}_d + \bar{\lambda}_b - \bar{\lambda}_a) + \alpha\,\bar{u} + c\,(\bar{\lambda}_b - \bar{\lambda}_a) = 0. \tag{2.6}$$

Next we define

$$\bar{p} := A^{-*}\iota_0^*(y_d - \iota_0\bar{y} - \bar{\lambda}_b + \bar{\lambda}_a) \tag{2.7}$$

then we have

$$\iota_{-1}^*\bar{p} = T^*(y_d - \iota_0\bar{y} - \bar{\lambda}_b + \bar{\lambda}_a)$$

where we used $\iota_0\bar{y} = T\bar{u} + \iota_0 A^{-1}f$. Equation (2.7) implies

$$A^*\bar{p} + \bar{y} - \bar{\lambda}_a + \bar{\lambda}_b = y_d, \tag{2.8}$$

which is the adjoint equation, and $\bar{p} \in H_0^1(\Omega) \cap H^2(\Omega)$ denotes the adjoint state associated with $(\bar{y}, \bar{v})$. Note that we have neglected the embedding operators in (2.8) as we shall do in general from now on. From (2.6) it follows

$$\alpha\,\bar{u} - \bar{p} + c\,(\bar{\lambda}_b - \bar{\lambda}_a) = 0. \tag{2.9}$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

Now we study the complementarity system (2.3)–(2.4). First we condense the Lagrange multipliers $\bar\lambda_a$ and $\bar\lambda_b$ into one multiplier

$$\bar\lambda = \bar\lambda_b - \bar\lambda_a. \tag{2.10}$$

Then we utilize nonlinear complementarity problem (NCP) functions to reformulate (2.3)–(2.4) as a single equality. For this purpose, based on numerical experience [12] we use

$$\bar\lambda - \min\left(0, \bar\lambda + \sigma\,(\bar v - \tilde a)\right) - \max\Big(0, \bar\lambda + \sigma\,(\bar v - \tilde b)\Big) = 0 \tag{2.11}$$

for some $\sigma \in L^\infty(\Omega)$, $\sigma > 0$. Here the max (respectively min) operations are performed pointwise. It is straightforward to prove that (2.11) and (2.3)–(2.4) are equivalent.

We define the a-active, b-active and inactive sets by

$$\begin{aligned}
\mathscr{A}_a(v,\lambda) &:= \{\mathbf{x} \in \Omega : \lambda(\mathbf{x}) + \sigma(v(\mathbf{x}) - a(\mathbf{x})) < 0\},\\
\mathscr{A}_b(v,\lambda) &:= \{\mathbf{x} \in \Omega : \lambda(\mathbf{x}) + \sigma(v(\mathbf{x}) - b(\mathbf{x})) > 0\},\\
\mathscr{I}(v,\lambda) &:= \Omega \setminus (\mathscr{A}_a(v,\lambda) \cup \mathscr{A}_b(v,\lambda)).
\end{aligned}$$

Further we shall frequently use the active set $\mathscr{A}(v,\lambda) = \mathscr{A}_a(v,\lambda) \cup \mathscr{A}_b(v,\lambda)$ and $\bar{\mathscr{A}}_a = \mathscr{A}_a(\bar v, \bar\lambda)$, and similarly for the active, b-active and inactive sets. Observe that by definition we have

$$\bar\lambda|_{\bar{\mathscr{A}}_a} \le 0, \quad \bar\lambda|_{\bar{\mathscr{A}}_b} \ge 0, \quad \bar\lambda|_{\bar{\mathscr{I}}} = 0.$$

Next we replace $\bar v$ by $F\bar u$ in (2.11) and, considering the composition of F, we get

$$\bar\lambda - \min\left(0, \bar\lambda + \sigma\,(c\,\bar u + \bar y - a)\right) - \max\left(0, \bar\lambda + \sigma\,(c\,\bar u + \bar y - b)\right) = 0. \tag{2.12}$$

Collecting (2.8)–(2.10) and (2.12) we obtain the following characterization.

THEOREM 2.2. *The optimal solution* $(\bar y, \bar u) \in H_0^1(\Omega) \times L^2(\Omega)$ *is characterized by the adjoint state* $\bar p \in H_0^1(\Omega) \cap H^2(\Omega)$ *and the Lagrange multiplier* $\bar\lambda \in L^2(\Omega)$ *satisfying*

$$A^\star \bar p + \bar\lambda + \bar y = y_d, \tag{2.13}$$

$$\alpha\,\bar u - \bar p + c\,\bar\lambda = 0, \tag{2.14}$$

$$\bar\lambda - \min\left(0, \bar\lambda + \sigma\,(c\,\bar u + \bar y - a)\right) - \max\left(0, \bar\lambda + \sigma\,(c\,\bar u + \bar y - b)\right) = 0 \tag{2.15}$$

for arbitrarily fixed $\sigma \in L^\infty(\Omega)$ *with* $\sigma > 0$ *a.e. in* Ω. *The multipliers* $\bar\lambda_a$ *and* $\bar\lambda_b$ *in* (2.3)–(2.4) *can be reconstructed by*

$$\bar\lambda_a = -\chi_{\bar{\mathscr{A}}_a}\bar\lambda \quad \text{and} \quad \bar\lambda_b = \chi_{\bar{\mathscr{A}}_b}\bar\lambda,$$

where χ_S *denotes the characteristic function of a set* $S \subset \Omega$.

Our algorithmic development in the subsequent section is based on the system (2.13)–(2.15) together with the state equation $A\bar y = \bar u + f$.

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

3. Primal-dual active-set method

Now we focus on a numerical technique for computing the solution of (1.1). Due to the complementarity system and its reformulation (2.12) it has to cope with the non-differentiable max- and min-terms. Since our goal is to apply a fast solution technique based on appropriate linearization of the first-order optimality system, we hence have to work with generalized derivatives when linearizing (2.15). This is done by employing the differentiability concept developed in [10] and [12]. We first recall the general notion, and then we apply it to our specific context.

DEFINITION 1. *Let $\mathscr{X}$ and $\mathscr{Y}$ be Banach spaces, and let $\mathscr{D} \subset \mathscr{X}$ be an open set. A mapping $\mathscr{F} : \mathscr{D} \to \mathscr{Y}$ is called* generalized differentiable in the open set $\mathscr{U} \subset \mathscr{D}$ *if there exists a family of mappings $\mathscr{G} : \mathscr{U} \to \mathscr{L}(\mathscr{X}, \mathscr{Y})$ such that*

$$\lim_{s \to 0} \frac{1}{\|s\|_{\mathscr{X}}} \|\mathscr{F}(x+s) - \mathscr{F}(x) - \mathscr{G}(x+s)s\|_{\mathscr{Y}} = 0 \quad \textit{for every } x \in \mathscr{U}. \tag{3.1}$$

Notice that the generalized derivative need not be unique. In [15] a notion similar to the one in Definition 1 is introduced and the name *Newton map* is coined for an element of the generalized derivative. Here we adopt this notion for operators $\mathscr{G}$ satisfying (3.1).

REMARK 1. In an L^p-setting it was shown in [12] that $\max(0, \cdot) : L^r(\Omega) \to L^s(\Omega)$ is generalized differentiable if and only if $r > s \geq 1$. The mapping

$$\mathscr{G}^0_{\max}(w)(\mathbf{x}) = \begin{cases} 1 & \text{if } w(\mathbf{x}) > 0, \\ 0 & \text{if } w(\mathbf{x}) \leq 0, \end{cases}$$

that is, $\mathscr{G}^0_{\max}(w) = \chi_{\{w>0\}}$, is a particular Newton map. A more general class of Newton maps for the max-operation is given by

$$\mathscr{G}^{m_l,m_u}_{\max}(w)(x) \in \begin{cases} \{1\} & \text{if } w(\mathbf{x}) > 0, \\ \{0\} & \text{if } w(\mathbf{x}) < 0, \\ [m_l, m_u] & \text{if } w(\mathbf{x}) = 0 \end{cases}$$

with arbitrary fixed $m_l, m_u \in \mathbb{R}$, $m_l \leq m_u$. The choice $(m_l, m_u) = (0, 1)$ yields the subgradient of convex analysis. The analogous result holds true for the min-operator with $\mathscr{G}^0_{\min}(w) = \chi_{\{w<0\}}$ being a particular Newton map, and with

$$\mathscr{G}^{m_l,m_u}_{\min}(w)(x) \in \begin{cases} \{1\} & \text{if } w(\mathbf{x}) < 0, \\ \{0\} & \text{if } w(\mathbf{x}) > 0, \\ [m_l, m_u] & \text{if } w(\mathbf{x}) = 0 \end{cases}$$

representing a more general class.

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

Let us assume we are interested in finding $\bar{x} \in \mathscr{X}$ such that

$$\mathscr{F}(\bar{x}) = 0. \tag{3.2}$$

This can be achieved by employing a Newton iteration, that is, given $x^k \in \mathscr{X}$, a sufficiently good approximation of $\bar{x}$, one linearizes (3.2) in the generalized sense and computes the next iterate x^{k+1} such that

$$\mathscr{F}(x^k) + \mathscr{G}(x^k)(x^{k+1} - x^k) = 0.$$

Note that x^{k+1} is uniquely defined whenever $\mathscr{G}(x^k)$ is invertible. In fact, we have the following result; see [10, 12].

THEOREM 3.1. *Suppose $\bar{x}$ is a solution of* (3.2) *and that $\mathscr{F}$ is generalized differentiable in an open neighbourhood $\mathscr{U}$ containing $\bar{x}$ with a Newton map $\mathscr{G}$. If $\mathscr{G}(x)$ is non-singular for all $x \in \mathscr{U}$ and $\{\|\mathscr{G}(x)^{-1}\| : x \in \mathscr{U}\}$ is bounded, then the generalized Newton iteration*

$$x^{k+1} = x^k - \mathscr{G}(x^k)^{-1}\mathscr{F}(x^k), \qquad \textit{with } x^0 \in \mathscr{U} \textit{ given}, \tag{3.3}$$

is well defined and converges locally at a superlinear rate to $\bar{x}$ provided that x^0 is sufficiently close to $\bar{x}$.

In finite dimensional space, in [20] the concept of *semismoothness* of a scalar-valued function (see [18] for its definition) is extended to the vector-valued case. It is shown that semismoothness of a mapping $\mathscr{F} : \mathbb{R}^n \to \mathbb{R}^m$ is equivalent to (3.1) with $\mathscr{G}$ replaced by an element of the generalized Jacobian in Clarke's sense at $x + s$. Hence, whenever $\mathscr{G}$ satisfies (3.1) we call (3.3) a *semismooth Newton method.*

Now we turn to the solution of (1.1). From (2.14) we infer

$$\bar{\lambda} = c^{-1}\,\bar{p} - \alpha\, c^{-1}\,\bar{u}.$$

Inserting this identity in (2.13) yields

$$\left(A^\star + c^{-1}\,\mathrm{id}\right)\bar{p} + \left(A^{-1} - \alpha\, c^{-1}\,\mathrm{id}\right)\bar{u} = \tilde{y}_d. \tag{3.4}$$

Here we used $\bar{y} = \mathfrak{y}(\bar{u}) := A^{-1}(\bar{u} + f)$. Since A is a second-order linear elliptic partial differential operator and $c \geq \epsilon_c > 0$ a.e. in Ω, we conclude that, given $\bar{u} \in L^2(\Omega)$, (3.4) admits a unique solution $\bar{p} \in H_0^1(\Omega)$, that is, $(A^\star + c^{-1}\,\mathrm{id}) : H_0^1(\Omega) \to H^{-1}(\Omega)$ is a continuously invertible linear operator. In addition, under a regularity assumption on the coefficients of the operator $A^\star + c^{-1}$ id, elliptic regularity theory yields $\bar{p} \in H^2(\Omega)$; see [11]. Hence we have

$$\bar{p} = \mathfrak{p}(\bar{u}) = (A^\star + c^{-1}\ \mathrm{id})^{-1}(\tilde{y}_d + \alpha\, c^{-1}\bar{u} - A^{-1}\bar{u}).$$

As a consequence, we obtain

$$\bar{\lambda} = \mathfrak{l}(\bar{u}) = c^{-1}\mathfrak{p}(\bar{u}) - \alpha\, c^{-1}\bar{u}.$$

The last identity is now used in studying $\bar{\lambda} + \sigma\,(c\,\bar{u} + \bar{y})$. In fact, we find

$$\begin{aligned} \bar{\lambda} + \sigma\,(c\,\bar{u} + \bar{y}) &= c^{-1}\mathfrak{p}(\bar{u}) - \alpha\, c^{-1}\bar{u} + \sigma\,\left(c\,\bar{u} + A^{-1}(\bar{u} + f)\right) \\ &= c^{-1}\mathfrak{p}(\bar{u}) + (\sigma\, c - \alpha\, c^{-1})\bar{u} + \sigma\, A^{-1}(\bar{u} + f). \end{aligned} \tag{3.5}$$

Choosing $\sigma = \alpha\, c^{-2} \geq (\alpha/\|c\|^2_{L^\infty}) > 0$ a.e., we can further simplify (3.5):

$$\bar{\lambda} + \sigma\,(c\,\bar{u} + \bar{y}) = c^{-1}\mathfrak{p}(\bar{u}) + \alpha\, c^{-2}\, A^{-1}(\bar{u} + f). \tag{3.6}$$

We adopt this choice for σ from now on. The Sobolev embedding theorem [1] yields

$$H_0^1(\Omega) \hookrightarrow L^s(\Omega), \quad \text{with} \quad s \begin{cases} = \infty & \text{if } n = 1, \\ \in [1, \infty) & \text{if } n = 2, \\ \in \left[1, 2 + \frac{4}{n-2}\right] & \text{if } n \geq 3. \end{cases} \tag{3.7}$$

Hence, from $\mathfrak{p}(\bar{u}) \in H_0^1(\Omega)$, $A^{-1} : H^{-1}(\Omega) \to H_0^1(\Omega)$, $c \in L^\infty(\Omega)$, with $c \geq \epsilon_c > 0$ a.e. in Ω, (3.6) and (3.7) we conclude

$$\begin{aligned} \ell(\bar{u}) :&= \bar{\lambda} + \sigma\,(c\,\bar{u} + \bar{y}) \\ &= c^{-1}\mathfrak{p}(\bar{u}) + \alpha\, c^{-2}\, A^{-1}(\bar{u} + f) \in L^{2+\kappa}(\Omega), \quad \text{with } \kappa > 0. \end{aligned}$$

We have shown the first part of the following result.

PROPOSITION 3.2. *The mapping*

$$\ell(u) = c^{-1}\mathfrak{p}(u) + \alpha\, c^{-2}\, A^{-1}(u + f)$$

is continuous from $L^2(\Omega)$ to $L^{2+\kappa}(\Omega)$ with

$$\kappa \begin{cases} = \infty & \text{if } n = 1, \\ \in [0, \infty) & \text{if } n = 2, \\ \in \left[0, \frac{4}{n-2}\right] & \text{if } n \geq 3. \end{cases}$$

Moreover, $\ell : L^2(\Omega) \to L^{2+\kappa}(\Omega)$ is continuously Fréchet differentiable.

PROOF. The proof of the continuity result lies in the discussion before the proposition. The continuous Fréchet differentiability is then immediate due to the affine linear nature of the operators involved in the definition of ℓ. □

Next we use the results obtained so far to reformulate the first-order system in Theorem 2.2. In fact, since $\bar{y} = \mathfrak{y}(\bar{u})$ and $\bar{p} = \mathfrak{p}(\bar{u})$, we can condense (2.13)–(2.15) into

$$c^{-1}\mathfrak{p}(\bar{u}) - \alpha c^{-1}\bar{u} - \min\left(0, \ell(\bar{u}) - \alpha c^{-2}a\right) - \max\left(0, \ell(\bar{u}) - \alpha c^{-2}b\right) = 0. \tag{3.8}$$

Setting $\bar{x} := \bar{u}$ and $\mathscr{F} : L^2(\Omega) \to L^2(\Omega)$, with

$$\mathscr{F}(\bar{x}) := c^{-1}\mathfrak{p}(\bar{x}) - \alpha c^{-1}\bar{x} - \min\left(0, \ell(\bar{x}) - \alpha c^{-2}a\right) - \max\left(0, \ell(\bar{x}) - \alpha c^{-2}b\right), \tag{3.9}$$

(3.8) is equivalent to the non-differentiable equation $\mathscr{F}(\bar{x}) = 0$. Thus, we are back at (3.2) with $\mathscr{X} = L^2(\Omega)$, that is, finding a solution of (1.1) is equivalent to finding the root of $\mathscr{F}$. In what follows, we prefer to keep $\bar{u}$ (or u) instead of $\bar{x}$ (or x).

Given some guess u^0 of $\bar{u}$, our goal is to find $\bar{u}$ using a semismooth Newton method (SSN). For a successful application of a SSN in function space we have to verify property (3.1). Remark 1 yields generalized differentiability of $\max : L^r(\Omega) \to L^s(\Omega)$ (and also $\min : L^r(\Omega) \to L^s(\Omega)$) if $r > s$. Note that in (3.9) we consider $\mathscr{F}$ from $L^2(\Omega)$ to $L^2(\Omega)$, only. A generalized differentiability result, however, still holds true.

PROPOSITION 3.3. *The mapping $\mathscr{F} : L^2(\Omega) \to L^2(\Omega)$ defined in* (3.9) *is generalized differentiable in the sense of Definition* 1. *A particular Newton map is given by*

$$\begin{aligned}\langle \mathscr{G}(u), \varphi\rangle = \left\langle c^{-1}\left(\mathfrak{p}'(u) - \alpha\,\mathrm{id}\right), \varphi\right\rangle - \left\langle \mathscr{G}^0_{\min}\left(\ell(u) - \alpha c^{-2}a\right)\ell'(u), \varphi\right\rangle \\ - \left\langle \mathscr{G}^0_{\max}\left(\ell(u) - \alpha c^{-2}b\right)\ell'(u), \varphi\right\rangle \quad \forall \varphi \in L^2(\Omega).\end{aligned} \tag{3.10}$$

PROOF. We argue only for the max-operator. The proof for the min-operator follows from analogous arguments.

First notice that $\ell(\cdot)$ is a continuous affine linear operator from $L^2(\Omega)$ to $L^{2+\kappa}(\Omega)$. Hence, there exist a continuous linear operator $L : L^2(\Omega) \to L^{2+\kappa}(\Omega)$ and $g \in L^{2+\kappa}(\Omega)$ such that

$$\ell(u) = L\,u + g. \tag{3.11}$$

Next we study the relevant difference quotient

$$\begin{aligned}&\frac{1}{\|s\|_{L^2}}\Big\|\max\left(0, \ell(u+s) - \alpha c^{-2}b\right) - \max\left(0, \ell(u) - \alpha c^{-2}b\right) \\ &\quad - \mathscr{G}^a_{\max}\left(\ell(u+s) - \alpha c^{-2}b\right)\ell'(s)\Big\|_{L^2} \\ &\qquad \le C\,\frac{1}{\|L\,s\|_{L^{2+\kappa}}}\Big\|\max\left(0, L(u+s) + g - \alpha c^{-2}b\right) - \max\left(0, L\,u + g - \alpha c^{-2}b\right) \\ &\qquad\quad - \mathscr{G}^a_{\max}\left(L(u+s) + g - \alpha c^{-2}b\right)L\,s\Big\|_{L^2},\end{aligned} \tag{3.12}$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

where $\mathcal{G}^a_{\max}$ denotes an arbitrary Newton map of the max-operator. Here we have used Proposition 3.2, which yields $\|L\,s\|_{L^{2+\kappa}}/\|s\|_{L^2} \le C$ for some positive constant C. By Remark 1, the quotient in (3.12) tends to zero for $\|s\|_{L^2} \to 0$. As a result $\max(0, \ell(\cdot) - \alpha\, c^{-2}b)$ is generalized differentiable, and $\mathcal{G}^a_{\max}(L\,\cdot + g - \alpha\, c^{-2}b)L\cdot$ provides a Newton map fulfilling (3.1). □

Now we have all the ingredients at hand for defining a semismooth Newton method for solving (1.1), or equivalently (2.1).

ALGORITHM 1 (Semismooth Newton method). (i) Choose $u^0 \in L^2(\Omega)$, and set $k = 0$.

(ii) Unless some stopping rule is satisfied, compute $\mathcal{G}(u^k)$ according to (3.10) and solve for δu^k:

$$\mathcal{G}(u^k)\delta u^k = -\mathcal{F}(u^k), \tag{3.13}$$

with $\mathcal{F}$ given by (3.9).

(iii) Set $u^{k+1} = u^k + \delta u^k$, and $k := k + 1$. Return to (ii).

We start our convergence analysis of Algorithm 1 by showing that (3.13) admits a unique solution for every $k \in \mathbb{N}$. For this purpose observe that (3.13) is equivalent to

$$\begin{aligned} 0 =& c^{-1}\left(\mathfrak{p}(u^k) + \mathfrak{p}'(u^k)\delta u^k\right) - \alpha c^{-1}(u^k + \delta u^k) - \chi_{\mathcal{A}^k_a}\left(L(u^k + \delta u^k) + g - \alpha\, c^{-2}a\right) \\ & - \chi_{\mathcal{A}^k_b}\left(L(u^k + \delta u^k) + g - \alpha\, c^{-2}b\right), \end{aligned} \tag{3.14}$$

where we have used

$$\mathcal{A}^k_a := \left\{\mathrm{x} \in \Omega : \lambda^k(\mathrm{x}) + \alpha\, c^{-1}u^k(\mathrm{x}) + \alpha\, c^{-2}(y^k - a)(\mathrm{x}) < 0\right\}, \tag{3.15}$$

$$\mathcal{A}^k_b := \left\{\mathrm{x} \in \Omega : \lambda^k(\mathrm{x}) + \alpha\, c^{-1}u^k(\mathrm{x}) + \alpha\, c^{-2}(y^k - b)(\mathrm{x}) > 0\right\}, \tag{3.16}$$

$$\mathcal{I}^k := \Omega \setminus \mathcal{A}^k, \quad \text{with} \quad \mathcal{A}^k := \mathcal{A}^k_a \cup \mathcal{A}^k_b.$$

Next recall that $\mathfrak{y}(w) = A^{-1}(w + f)$ for $w \in L^2(\Omega)$, with A^{-1} a linear continuous operator, and $\mathfrak{p}(w) = (A^* + c^{-1}\,\mathrm{id})^{-1}(y_d + \alpha c^{-1}w - \mathfrak{y}(w))$. This yields

$$A^*\mathfrak{p}(w) + c^{-1}(\mathfrak{p}(w) - \alpha w) = y_d - \mathfrak{y}(w).$$

Since $\mathfrak{l}(w) = c^{-1}\mathfrak{p}(w) - \alpha c^{-1}w$, we obtain

$$A^*\mathfrak{p}(w) + \mathfrak{l}(w) = y_d - \mathfrak{y}(w).$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

For $w = u^k + \delta u^k =: u^{k+1}$ we set $y^{k+1} = \mathfrak{y}(u^k + \delta u^k)$, and similarly for p^{k+1} and λ^{k+1}. Hence (3.14) becomes

$$\begin{aligned}
0 &= \lambda^{k+1} - \chi_{\mathscr{A}_a^k}\left(\lambda^{k+1} + \alpha\, c^{-1} u^{k+1} + \alpha\, c^{-2}(y^{k+1} - a)\right) \\
&\quad - \chi_{\mathscr{A}_b^k}\left(\lambda^{k+1} + \alpha\, c^{-1} u^{k+1} + \alpha\, c^{-2}(y^{k+1} - b)\right), && (3.17) \\
0 &= A y^{k+1} - u^{k+1} - f, && (3.18) \\
0 &= A^\star p^{k+1} + \lambda^{k+1} + y^{k+1} - y_d, && (3.19) \\
0 &= \alpha u^{k+1} - p^{k+1} + c\,\lambda^{k+1}. && (3.20)
\end{aligned}$$

A further analysis of (3.17) yields

$$\begin{aligned}
\lambda^{k+1} &= 0 \quad \text{on } \mathscr{I}^k, && (3.21) \\
c\, u^{k+1} + y^{k+1} &= a \quad \text{on } \mathscr{A}_a^k, && (3.22) \\
c\, u^{k+1} + y^{k+1} &= b \quad \text{on } \mathscr{A}_b^k. && (3.23)
\end{aligned}$$

Combining (3.18)–(3.23) we conclude that in every iteration of our Algorithm 1 the following system has to be solved:

$$\begin{aligned}
A y^{k+1} &= u^{k+1} + f, && (3.24) \\
A^\star p^{k+1} + \lambda^{k+1} &= y_d - y^{k+1}, && (3.25) \\
\alpha u^{k+1} - p^{k+1} + c\,\lambda^{k+1} &= 0, && (3.26) \\
\lambda^{k+1} &= 0 \quad \text{on } \mathscr{I}^k, && (3.27) \\
c\, u^{k+1} + y^{k+1} &= a \quad \text{on } \mathscr{A}_a^k, && (3.28) \\
c\, u^{k+1} + y^{k+1} &= b \quad \text{on } \mathscr{A}_b^k. && (3.29)
\end{aligned}$$

For proceeding with our arguments that Step (ii) of Algorithm 1 is well defined, we rewrite (3.24)–(3.26). In fact, solving (3.24) for y^{k+1}, inserting the result in (3.25), solving for p^{k+1}, utilizing the result in (3.26), and taking into account the various embedding operators, we get

$$(T^\star T + \alpha\, \mathrm{id}) u^{k+1} + F^\star \lambda^{k+1} = T^\star \tilde{y}_d. \qquad (3.30)$$

Using $u^{k+1} = F^{-1} v^{k+1}$ and the invertibility of $F^\star$, (3.30) is equivalent to

$$F^{-\star}(T^\star T + \alpha\, \mathrm{id}) F^{-1} v^{k+1} + \lambda^{k+1} = F^{-\star} T^\star \tilde{y}_d.$$

Further recall that

$$c\, u^{k+1} + y^{k+1} = F\, u^{k+1} + A^{-1} f = v^{k+1} + A^{-1} f.$$

Thus, (3.27)–(3.29) become

$$\begin{aligned}
\lambda^{k+1} &= 0 \quad \text{on } \mathscr{I}^k, \\
v^{k+1} &= \tilde{a} \quad \text{on } \mathscr{A}_a^k, \\
v^{k+1} &= \tilde{b} \quad \text{on } \mathscr{A}_b^k.
\end{aligned}$$

PROPOSITION 3.4. *The system*

$$F^{-\star}(T^\star T + \alpha\,\mathrm{id})F^{-1}w + \mu = a_1, \tag{3.31}$$

$$\mu = 0 \quad \text{on } \mathscr{I}, \tag{3.32}$$

$$w = a_2 \quad \text{on } \mathscr{A}_a, \tag{3.33}$$

$$w = a_3 \quad \text{on } \mathscr{A}_b, \tag{3.34}$$

with $a_i \in L^2(\Omega)$, $i \in \{1, 2, 3\}$, and $(\mathscr{I}, \mathscr{A}_a, \mathscr{A}_b)$ a partitioning of Ω, admits a unique solution $(\bar{w}, \bar{\mu}) \in L^2(\Omega) \times L^2(\Omega)$.

PROOF. First note that

$$\begin{aligned}\left(F^{-\star}(T^\star T + \alpha\,\mathrm{id})F^{-1}\varphi, \varphi\right)_{L^2} &= \left\|TF^{-1}\varphi\right\|_{L^2}^2 + \alpha\left\|F^{-1}\varphi\right\|_{L^2}^2\\ &\geq \frac{\alpha}{\|F\|^2}\|\varphi\|_{L^2}^2 \quad \forall\varphi \in L^2(\Omega).\end{aligned} \tag{3.35}$$

Hence, for given $\mu, a_1 \in L^2(\Omega)$, (3.31) admits a unique solution $w = w(\mu) \in L^2(\Omega)$.

Let $E_{\mathscr{I}}$ denote the extension-by-zero operator from $\mathscr{I}$ to Ω, and analogously for $E_{\mathscr{A}}$ with $\mathscr{A} = \mathscr{A}_a \cup \mathscr{A}_b$. By $E_{\mathscr{I}}^\star$ and $E_{\mathscr{A}}^\star$ we denote the respective restriction operators. Then, considering (3.32)–(3.34) in (3.31) we obtain

$$E_{\mathscr{I}}^\star C E_{\mathscr{I}} w_{\mathscr{I}} = E_{\mathscr{I}}^\star a_1 - E_{\mathscr{I}}^\star C E_{\mathscr{A}} \hat{a}_{\mathscr{A}}, \tag{3.36}$$

where $w_{\mathscr{I}} \in L^2(\mathscr{I})$,

$$\hat{a}\,\big|_{\mathscr{A}_a} := a_2\,\big|_{\mathscr{A}_a}, \qquad \hat{a}\,\big|_{\mathscr{A}_b} := a_3\,\big|_{\mathscr{A}_b} \tag{3.37}$$

and

$$C := F^{-\star}\left(T^\star T + \alpha\,\mathrm{id}\right)F^{-1}. \tag{3.38}$$

If $\mathscr{I} \neq \emptyset$ is measurable, then, from the properties of C, we conclude that Equation (3.36) admits a unique solution $\bar{w}_{\mathscr{I}} \in L^2(\mathscr{I})$. From this we construct a solution of (3.31)–(3.34) in the following way:

$$\bar{w}\,|_{\mathscr{I}} := \bar{w}_{\mathscr{I}}, \qquad \bar{w}\,\big|_{\mathscr{A}_a} := a_2\,\big|_{\mathscr{A}_a}, \qquad \bar{w}\,\big|_{\mathscr{A}_b} := a_3\,\big|_{\mathscr{A}_b}. \tag{3.39}$$

Then $\bar{w} \in L^2(\Omega)$. Further, $\bar{\mu}|_{\mathscr{I}} = 0$ and

$$\bar{\mu}_{\mathscr{A}} = E_{\mathscr{A}}^\star a_1 - E_{\mathscr{A}}^\star C\bar{w}, \quad \bar{\mu}_{\mathscr{A}} \in L^2(\mathscr{A}),$$

which defines $\bar{\mu}|_{\mathscr{A}} = \bar{\mu}_{\mathscr{A}}$ uniquely. This ends the proof. □

From Proposition 3.4 we immediately infer that Step (ii) of Algorithm 1 is well defined.

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

COROLLARY 3.5. *For every $k \in \mathbb{N}$, Step* (ii) *of Algorithm* 1 *admits a unique solution $\delta u^k \in L^2(\Omega)$. Further, there exists a constant $K_{\mathscr{G}} > 0$ independent of k such that $\|\mathscr{G}(u^k)^{-1}\| \leq K_{\mathscr{G}}$ for all $k \in \mathbb{N}_0$.*

PROOF. By our discussion before Proposition 3.4 the Newton system (3.13) is equivalent to (3.31)–(3.34) with $\mathscr{I} = \mathscr{I}^k$, $\mathscr{A}_a = \mathscr{A}_a^k$, $\mathscr{A}_b = \mathscr{A}_b^k$, $w = F(u^k + \delta u^k)$, $\mu = \lambda^{k+1}$, $a_1 = F^{-\star}T^\star \tilde{y}_d$, $a_2 = \tilde{a}$ and $a_3 = \tilde{b}$. Now, Proposition 3.4 yields that (3.13) admits a unique solution $\delta u^k = u^{k+1} - u^k$. Further λ^{k+1} is the multiplier associated with u^{k+1}.

The uniform boundedness of $\left\{\left\|\mathscr{G}(u^k)^{-1}\right\|\right\}_{k\geq 0}$ follows from the fact that C (see (3.38) in the proof of Proposition 3.4) and its corresponding coercivity constant, see, for example, (3.35), are independent of k. □

The locally superlinear convergence of the semismooth Newton Algorithm 1 is the subject of our next result.

THEOREM 3.6. *Let $\{u^k\}$ be the sequence generated by Algorithm* 1, *and define the corresponding states by $y^k = A^{-1}(u^k + f) \in H_0^1(\Omega)$. Then $\{(y^k, u^k)\}$ converges to the solution $(\bar{y}, \bar{u}) \in H_0^1(\Omega) \times L^2(\Omega)$ of* (1.1) *at a superlinear rate provided that $u^0 \in L^2(\Omega)$ is sufficiently close to $\bar{u}$.*

PROOF. Proposition 3.3 shows that Algorithm 1 is a semismooth Newton method for solving (3.9) with $x = u$. Hence, by our general result, Theorem 3.1, the sequence $\{u^k\}$ converges superlinearly to $\bar{u}$ provided that u^0 is sufficiently close to $\bar{u}$. The locally superlinear convergence of $\{y^k\}$ is then an immediate consequence. □

We end this section by establishing a relation between the semismooth Newton method, Algorithm 1, and a primal-dual active-set method. We already showed that computing δu^k such that (3.13) is satisfied is equivalent to solving (3.24)–(3.29). Hence we may restate Algorithm 1 as follows.

ALGORITHM 2 (Primal-dual active-set method). (i) Choose $u^0 \in L^2(\Omega)$ and compute (y^0, p^0, λ^0) such that

$$\begin{aligned} Ay^0 &= u^0 + f, \\ A^\star p^0 + \lambda^0 + y^0 &= y_d, \\ \alpha\, u^0 - p^0 + c\,\lambda^0 &= 0. \end{aligned}$$

Set $k := 0$.

(ii) Unless some stopping rule is satisfied, determine

$$\begin{aligned}
\mathscr{A}_a^k &:= \left\{\mathbf{x} \in \Omega : \left(\lambda^k + \alpha c^{-1} u^k + \alpha c^{-2}(y^k - a)\right)(\mathbf{x}) < 0\right\}, \\
\mathscr{A}_b^k &:= \left\{\mathbf{x} \in \Omega : \left(\lambda^k + \alpha c^{-1} u^k + \alpha c^{-2}(y^k - b)\right)(\mathbf{x}) > 0\right\}, \\
\mathscr{I}^k &:= \Omega \setminus \left(\mathscr{A}_a^k \cup \mathscr{A}_b^k\right).
\end{aligned}$$

(iii) Solve for $(u^{k+1}, y^{k+1}, p^{k+1}, \lambda^{k+1})$:

$$\begin{aligned}
A y^{k+1} - u^{k+1} &= f, \\
A^* p^{k+1} + \lambda^{k+1} + y^{k+1} &= y_d, \\
\alpha u^{k+1} - p^{k+1} + c\lambda^{k+1} &= 0, \\
\lambda^{k+1} &= 0 \quad \text{on } \mathscr{I}^k, \\
c u^{k+1} + y^{k+1} &= a \quad \text{on } \mathscr{A}_a^k, \\
c u^{k+1} + y^{k+1} &= b \quad \text{on } \mathscr{A}_b^k.
\end{aligned}$$

Set $k := k + 1$, and continue with (ii).

4. Mesh independence

In this section we establish a mesh-independence result for our semismooth Newton method. It states that for any q-linear rate of convergence θ, there exists a radius $\rho > 0$ such that, for all h sufficiently small, the convergence basin of the primal-dual active-set method, Algorithm 2, or equivalently the semismooth Newton method, Algorithm 1, and the discrete counterparts contain the ρ-balls about their respective solutions. A similar result was proven in [14] for control constrained semilinear elliptic control problems. This type of mesh independence is in contrast to the strong mesh-independence principle like the one in [2] for smooth operator equations.

We consider a finite element discretization of (3.9). Here we only provide a brief description and refer to, for example, [4] for more details on appropriate discretizations in constrained optimal control of PDEs. In fact, let $\mathscr{T}_h$ be a sufficiently regular subdivision (triangulation) of Ω into subdomains $T \in \mathscr{T}_h$ such that

$$\bar{\Omega} = \bigcup_{T \in \mathscr{T}_h} T, \quad T_1, T_2 \in \mathscr{T}_h, \quad T_1 \neq T_2 \quad \Rightarrow \quad T_1 \cap T_2 \subset \partial T_1 \cup \partial T_2.$$

The subscript h refers to the maximal diameter of all elements. Motivated by (3.9) we next define the space

$$U_h = \{u_h : \Omega \to \mathbb{R} : u_h|_{\operatorname{int} T} = \text{constant } \forall T \in \mathscr{T}_h\}$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

which we endow with the L^2-norm, that is, $\|\cdot\|_{U_h} = \|\cdot\|_{L^2}$. An appropriate discretization of (3.9) yields $\mathscr{F}_h : U_h \to U_h$ and in particular $\ell_h : U_h \to U_h$ and $\mathfrak{p}_h : U_h \to U_h$, the discrete versions of $\mathfrak{p}$ and ℓ, respectively. We denote by $\bar{u}_h \in U_h$ the unique solution of $\mathscr{F}_h(u_h) = 0$. Further, discrete Newton maps of $\mathscr{F}_h$ are denoted by $\mathscr{G}_h$. This allows us to define a discrete version of Algorithm 1, the discrete semismooth Newton iteration:

ALGORITHM 3. (i) Choose $u_h^0 \in U_h$, and set $k = 0$.

(ii) Unless some stopping rule is satisfied, compute $\mathscr{G}_h(u_h^k)$, a Newton map of $\mathscr{F}_h$ at u_h^k, and solve for δu_h^k:

$$\mathscr{G}_h(u_h^k)\,\delta u_h^k = -\mathscr{F}_h(u_h^k)\,. \tag{4.1}$$

(iii) Set $u_h^{k+1} = u_h^k + \delta u_h^k$, and $k := k+1$. Return to (ii).

For $\mathscr{G}_h$ based on (3.10), by a similar reasoning as in the continuous case, one can show that (4.1) is equivalent to the discrete version of (3.24)–(3.29) and Algorithm 3 is equivalent to the discrete analogue of the active-set method, Algorithm 2.

For the proof of our main assertion we need an auxiliary result concerning the mesh independence of (3.1). For this purpose recall that $\ell(u) = L\,u + g$ with $L \in \mathscr{L}(L^2(\Omega), L^q(\Omega))$ and $g \in L^q(\Omega)$ with

$$q \in \begin{cases} [1,\infty] & n = 1, \\ [1,\infty) & n = 2, \\ [1, 2+\frac{4}{n-2}] & n \geq 3. \end{cases}$$

We also suppose that our discretization yields $\ell_h(u_h) = L_h u_h + g_h$ such that the following assumption holds true.

ASSUMPTION 1. *There exist some $q > 2$ and some positive constant K such that*

$$\lim_{h\to 0^+} \max\left(\|g - g_h\|_{L^q}, \left\|c^{-2}a - c_h^{-2}a_h\right\|_{L^q}, \left\|c^{-2}b - c_h^{-2}b_h\right\|_{L^q}\right) = 0, \tag{4.2}$$

$$\lim_{h\to 0^+} \|\bar{u}_h - \bar{u}\|_{L^2} = 0,$$

$$\lim_{h\to 0^+} \|L_h\bar{u}_h - L\,\bar{u}\|_{L^q} = 0, \tag{4.3}$$

$$\|L_h\|_{L^2\to L^q} \leq K.$$

Note that due to (3.1), for given $\gamma \in (0,1)$, there exists $\delta_0 > 0$ such that

$$\|\mathscr{F}(\bar{u}+s) - \mathscr{F}(\bar{u}) - \mathscr{G}(\bar{u}+s)s\|_{L^2} \leq \gamma\|s\|_{L^2} \quad \forall s \in L^2(\Omega), \quad \|s\|_{L^2} \leq \delta_0. \tag{4.4}$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

Here $\mathscr{G}(\bar{u}+s)$ denotes an arbitrary Newton map of $\mathscr{F}$ at $\bar{u}+s$. For our main result we need the mesh independence of

$$\|\mathscr{F}_h(u_h)-\mathscr{F}_h(\bar{u}_h)-\mathscr{G}_h(u_h)(u_h-\bar{u}_h)\|_{U_h} \quad \text{as } h \to 0 \tag{4.5}$$

for arbitrary $\mathscr{G}_h(u_h)$ satisfying the discrete analogue of (3.1). Due to the structure of $\mathscr{F}_h$, that is,

$$\begin{aligned}\mathscr{F}_h\,(u_h) = c_h^{-1}\mathrm{p}_h\,(u_h) - \alpha\, c_h^{-1}u_h - \min\left(0, \ell_h\,(u_h) - \alpha\, c_h^{-2}a_h\right)\\ - \max\left(0, \ell_h\,(u_h) - \alpha\, c_h^{-2}b_h\right),\end{aligned} \tag{4.6}$$

with p_h affine linear, we only have to focus on the max- and min-terms, respectively, when proving the mesh independence of (4.5). We define

$$\mathscr{F}_{\max}(u) := \max\left(0, \ell(u) - \alpha\, c^{-2}b\right).$$

In what follows, the Newton maps of $\mathscr{F}_{\max}$ are denoted by $\mathscr{G}_{\max}$. Their discrete counterparts are $\mathscr{F}_{\max,h}$ and $\mathscr{G}_{\max,h}$, respectively.

LEMMA 4.1. *Suppose that Assumption* 1 *holds true, and*

$$\left|\left\{\ell(\bar{u}) - \alpha\, c^{-2}b = 0\right\}\right| = \left|\left\{L\,\bar{u} + g - \alpha\, c^{-2}b = 0\right\}\right| = 0 \tag{4.7}$$

is satisfied. Further assume that

$$\mathscr{G}_{\max}(u) = \mathscr{G}_{\max}^{m_l,m_u}\left(L\,u + g - \alpha\, c^{-2}b\right)L$$

(see Remark 1 *for the definition of* $\mathscr{G}_{\max}^{m_l,m_u}$*) and its discrete analogue are chosen as the Newton maps of* $\mathscr{F}_{\max}$ *and* $\mathscr{F}_{\max,h}$*, respectively. Then, for* $\gamma \in (0,1)$*, there exists* $\bar{\delta} > 0$ *and* $\bar{h} > 0$ *such that* $\forall u \in L^2(\Omega)$, $\|u-\bar{u}\|_{L^2} \le \bar{\delta}$,

$$\|\mathscr{F}_{\max}(u) - \mathscr{F}_{\max}(\bar{u}) - \mathscr{G}_{\max}(u)(u-\bar{u})\|_{L^2} \le \gamma\, \|u-\bar{u}\|_{L^2}$$

as well as

$$\begin{gathered}\left\|\mathscr{F}_{\max,h}(u_h) - \mathscr{F}_{\max,h}(\bar{u}_h) - \mathscr{G}_{\max,h}(u_h)(u_h-\bar{u}_h)\right\|_{U_h} \le \gamma\, \|u_h-\bar{u}_h\|_{U_h}\\ \forall u_h \in U_h,\ \|u_h-\bar{u}_h\|_{U_h} \le \bar{\delta}, \quad \forall h \in (0,\bar{h}].\end{gathered}$$

PROOF. For $\epsilon > 0$ and $0 < \eta \le \epsilon$ define the sets

$$\begin{aligned}\Omega(\epsilon) &:= \left\{\left|L\,\bar{u} + g - \alpha\, c^{-2}b\right| < \epsilon\right\}, & \Omega_h(\epsilon) &:= \left\{\left|L_h\,\bar{u}_h + g_h - \alpha\, c_h^{-2}b_h\right| < \epsilon\right\},\\ \Omega_h^1(\epsilon) &:= \{|L_h(u_h-\bar{u}_h)| < \eta\} \setminus \Omega_h(\epsilon), & \Omega_h^2(\epsilon) &:= \Omega_h(\epsilon) \cup \{|L_h(u_h-\bar{u}_h)| \ge \eta\}.\end{aligned}$$

Note that $\Omega_h^1(\epsilon) \cup \Omega_h^2(\epsilon) = \Omega$. Further define the remainder term

$$R_h(u_h;\bar{u}_h) := \mathscr{F}_{\max,h}(u_h) - \mathscr{F}_{\max,h}(\bar{u}_h) - \mathscr{G}_{\max,h}(u_h)(u_h-\bar{u}_h).$$

On $\Omega_h^1(\epsilon)$ we have $R_h(u_h;\bar{u}_h) = 0$. Indeed:

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

(1) For $\mathbf{x} \in \Omega$ with $\left(L_h \bar{u}_h + g_h - \alpha\, c_h^{-2} b_h\right)(\mathbf{x}) \geq \epsilon$ and $|L_h(u_h - \bar{u}_h)(\mathbf{x})| < \eta$, we obtain

$$\left(L_h \bar{u}_h + g_h - \alpha\, c_h^{-2} b_h\right)(\mathbf{x}) \geq \epsilon \geq \eta > \left(L_h(\bar{u}_h - u_h)\right)(\mathbf{x})$$

and, hence,

$$\left(L_h u_h + g_h - \alpha\, c_h^{-2} b_h\right)(\mathbf{x}) > 0.$$

As a consequence, we have

$$\begin{aligned} R_h(u_h; \bar{u}_h)(\mathbf{x}) &= \left(L_h u_h + g_h - \alpha\, c_h^{-2} b_h\right)(\mathbf{x}) - \left(L_h \bar{u}_h + g_h - \alpha c_h^{-2} b_h\right)(\mathbf{x}) \\ &\qquad - \left(L_h(u_h - \bar{u}_h)\right)(\mathbf{x}) \\ &= \left(L_h\, (u_h - \bar{u}_h)\right)(\mathbf{x}) - \left(L_h\, (u_h - \bar{u}_h)\right)(\mathbf{x}) = 0. \end{aligned}$$

(2) For $\mathbf{x} \in \Omega$ with $\left(L_h \bar{u}_h + g_h - \alpha\, c_h^{-2} b_h\right)(\mathbf{x}) \leq -\epsilon$ and $\left|L_h(u_h - \bar{u}_h)(\mathbf{x})\right| < \eta$ we get

$$(L_h \bar{u}_h + g_h - \alpha\, c_h^{-2} b_h)(\mathbf{x}) \leq -\epsilon \leq -\eta < \left(L_h(\bar{u}_h - u_h)\right)(\mathbf{x})$$

and, hence,

$$\left(L_h u_h + g_h - \alpha\, c_h^{-2} b_h\right)(\mathbf{x}) < 0.$$

Consequently, we infer $\mathscr{G}_{\max,h}(u_h)(\mathbf{x}) = 0$ and $R_h(u_h; \bar{u}_h)(\mathbf{x}) = 0$.

Summarizing both cases we have

$$R_h\,(u_h; \bar{u}_h)\,(\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \Omega_h^1(\epsilon).$$

The estimate on $\Omega_h^2(\epsilon)$ is more delicate. We make use of the following fact: For $v \in L^q(\Omega)$, $q > 2$ and $\eta > 0$ we have

$$|\{|v| \geq \eta\}| \leq \frac{1}{\eta^q} \|v\|_{L^q}^q. \tag{4.8}$$

For $\mu > 0$, this estimate implies

$$|\{|L_h(u_h - \bar{u}_h)| \geq \eta\}| \leq \frac{1}{\eta^q} \|L_h(u_h - \bar{u}_h)\|_{L^q}^q \leq \frac{1}{\eta^q} \|L_h\|_{L^2 \to L^q}^q \|u_h - \bar{u}_h\|_{U_h}^q \leq \mu \tag{4.9}$$

for $\|u_h - \bar{u}_h\|_{U_h} \leq \bar{\delta}$ with

$$\bar{\delta} := \min\left(\delta_0, \frac{\eta \sqrt[q]{\mu}}{\|L_h\|_{L^2 \to L^q}}\right) > 0$$

and δ_0 from (4.4). Observe that $|\max(0, r) - \max(0, s)| \leq |r - s|$. We, thus, conclude

$$\begin{aligned} \|R_h(u_h; \bar{u}_h)\|_{L^q} &= \Big\|\max\left(0, L_h u_h + g_h - \alpha\, c_h^{-2} b_h\right) - \max\left(0, L_h \bar{u}_h + g_h - \alpha\, c_h^{-2} b_h\right) \\ &\qquad - \mathscr{G}_{\max,h}^{m_l, m_u}\left(L_h u_h + g_h - \alpha\, c_h^{-2} b_h\right) L_h(u_h - \bar{u}_h)\Big\|_{L^q} \\ &\leq \|L_h(u_h - \bar{u}_h)\|_{L^q} + \left\|\mathscr{G}_{\max,h}^{m_l, m_u}\left(L_h u_h + g_h - \alpha\, c_h^{-2} b_h\right) L_h(u_h - \bar{u}_h)\right\|_{L^q} \\ &\leq (1 + \max\left(|m_l|, |m_u|\right)) \|L_h\|_{L^2 \to L^q}\, \|u_h - \bar{u}_h\|_{U_h}. \end{aligned} \tag{4.10}$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

Further note that our assumption (4.7) implies

$$|\Omega(\epsilon)| = \left|\left\{\left|L\,\bar u - g - \alpha\, c^{-2}b\right| < \epsilon\right\}\right| \to 0 \quad \text{as } \epsilon \downarrow 0. \tag{4.11}$$

We next use the last two equations to establish the desired estimate of $R_h(u_h; \bar u_h)$ on $\Omega_h^2(\epsilon)$.

(3) Consider $\mathbf{x} \in \Omega$ such that

$$d_h(\bar u, \bar u_h)(\mathbf{x}) := (L\,\bar u + g - \alpha\, c^{-2}b)(\mathbf{x}) - (L_h\bar u_h + g_h - \alpha c_h^{-2}b_h)(\mathbf{x}) > \epsilon. \tag{4.12}$$

Now assume that $\mathbf{x} \in \Omega(2\epsilon)$ satisfies (4.12). Then

$$\begin{aligned}2\epsilon >&(L\,\bar u + g - \alpha\, c^{-2}b)(\mathbf{x})\\ =&(L\bar u + g - \alpha\, c^{-2}b - L_h\bar u_h - g_h + \alpha\, c_h^{-2}b_h)(\mathbf{x}) + (L_h\bar u_h + g_h - \alpha\, c_h^{-2}b_h)(\mathbf{x})\\ >&\epsilon + \left(L_h\bar u_h + g_h - \alpha c_h^{-2}b_h\right)(\mathbf{x}),\end{aligned}$$

which implies

$$\left(L_h\bar u_h + g_h - \alpha\, c_h^{-2}b_h\right)(\mathbf{x}) < \epsilon.$$

Similarly we obtain

$$\left(L_h\bar u_h + g_h - \alpha\, c_h^{-2}b_h\right)(\mathbf{x}) > -\epsilon.$$

We hence conclude

$$\Omega_h(\epsilon) \subset \Omega(2\epsilon) \cap \{d_h(\bar u, \bar u_h) > \epsilon\}. \tag{4.13}$$

Since $d_h(\bar u, \bar u_h) \in L^q(\Omega)$, (4.8) yields

$$|\{d_h(\bar u, \bar u_h) > \epsilon\}| \le \frac{1}{\epsilon^q}\|d_h(\bar u, \bar u_h)\|_{L^q}^q. \tag{4.14}$$

The estimate

$$\|d_h(\bar u, \bar u_h)\|_{L^q} \le \|g - g_h\|_{L^q} + \alpha\|c^{-2}b - c_h^{-2}b_h\|_{L^q} + \|L\bar u - L_h\bar u_h\|_{L^q},$$

(4.2), (4.3) and (4.14) yield

$$|\{d_h(\bar u, \bar u_h) > \epsilon\}| \to 0 \quad \text{as } h \to 0^+. \tag{4.15}$$

Using (4.11) and (4.15), for $\mu > 0$ there exist $\epsilon = \epsilon(\mu) > 0$ and $h_1 = h_1(\epsilon) > 0$ such that for all $h \le h_1$ we obtain

$$|\Omega(2\epsilon)| \le \frac{\mu}{2} \quad \text{and} \quad |\{d_h(\bar u, \bar u_h) > \epsilon\}| \le \frac{\mu}{2}$$

and therefore

$$|\Omega_h(\epsilon)| \le \mu \tag{4.16}$$

by (4.13).

(4) Since $q > 2$, Hölder's inequality and the estimates (4.10), (4.16) and (4.9) yield

$$\begin{aligned}\|R_h\,(u_h;\bar{u}_h)\|_{L^2(\Omega_h^2(\epsilon))} &\le \left|\Omega_h^2(\epsilon)\right|^{\frac{1}{2}-\frac{1}{q}}\,\|R_h(u_h;\bar{u}_h)\|_{L^q(\Omega_h^2(\epsilon))}\\ &\le (1+\max(|m_l|,|m_u|))\,(2\mu)^{\frac{q-2}{2q}}\,\|L_h\|_{L^2\to L^q}\|u_h-\bar{u}_h\|_{U_h}\end{aligned}$$

for $\|u_h - \bar{u}_h\|_{U_h} \le \bar{\delta}$.

(5) For $\gamma \in (0, 1)$ we choose

$$\mu = \frac{1}{2}\left(\frac{\gamma}{(1+\max(|m_l|,|m_u|))\,\|L_h\|_{L^2\to L^q}}\right)^{2q/(q-2)}$$

and $\epsilon = \epsilon(\mu)$, $\bar{h} = h_1(\epsilon)$, $\bar{\delta} = \bar{\delta}(\mu, \epsilon)$. Then, for all $h \le \bar{h}$, we have

$$\|R_h(u_h;\bar{u}_h)\|_{U_h} \le \gamma\,\|u_h-\bar{u}_h\|_{U_h} \quad \forall u_h \in U_h,\ \|u_h-\bar{u}_h\|_{U_h} \le \bar{\delta}.$$

This proves the assertion. □

The same proof technique yields a similar result for the min-operation in (3.9), that is,

$$\mathscr{F}_{\min}(u) := \min\left(0, \ell(u) - \alpha\, c^{-2} a\right)$$

and its discrete version $\mathscr{F}_{\min,h}$ in (4.6).

LEMMA 4.2. *Suppose that Assumption* 1 *holds true, and*

$$\left|\left\{\ell(\bar{u}) - \alpha\, c^{-2} a = 0\right\}\right| = \left|\left\{L\,\bar{u} + g - \alpha\, c^{-2} a = 0\right\}\right| = 0$$

is satisfied. Further assume that

$$\mathscr{G}_{\min}(u) = \mathscr{G}_{\min}^{m_l,m_u}\left(L\,u + g - \alpha\, c^{-2} a\right) L$$

(see Remark 1 *for the definition of* $\mathscr{G}_{\min}^{m_l,m_u}$*) and its discrete analogue are chosen as the Newton maps of* $\mathscr{F}_{\min}$ *and* $\mathscr{F}_{\min,h}$*, respectively. Then, for* $\gamma \in (0, 1)$*, there exist* $\bar{\delta} > 0$ *and* $\bar{h} > 0$ *such that*

$$\begin{gathered}\|\mathscr{F}_{\min}(u) - \mathscr{F}_{\min}(\bar{u}) - \mathscr{G}_{\min}(u)(u-\bar{u})\|_{L^2} \le \gamma\,\|u-\bar{u}\|_{L^2}\\ \forall u \in L^2(\Omega), \quad \|u-\bar{u}\|_{L^2} \le \bar{\delta},\end{gathered}$$

as well as

$$\begin{gathered}\left\|\mathscr{F}_{\min,h}(u_h) - \mathscr{F}_{\min,h}(\bar{u}_h) - \mathscr{G}_{\min,h}(u_h)(u_h-\bar{u}_h)\right\|_{U_h} \le \gamma\,\|u_h-\bar{u}_h\|_{U_h}\\ \forall u_h \in U_h, \quad \|u_h-\bar{u}_h\|_{U_h} \le \bar{\delta}, \quad \forall h \in (0,\bar{h}].\end{gathered}$$

Combining Lemmas 4.1 and 4.2, we obtain the following mesh-independence result.

PROPOSITION 4.3. *Let the assumptions of Lemmas 4.1 and 4.2 hold true. Further assume that for $u \in L^2(\Omega)$ and $u_h \in U_h$ with $\|u - u_h\|_{L^2} \to 0$ as $h \to 0^+$ we have*

$$\left\|c^{-1}\mathfrak{p}(u) - c_h^{-1}\mathfrak{p}_h(u_h)\right\|_{L^2} \to 0 \quad as\ h \to 0^+.$$

Then, for $\gamma \in (0, 1)$, there exist $\bar{\delta} > 0$ and $\bar{h} > 0$ such that

$$\|\mathscr{F}(u) - \mathscr{F}(\bar{u}) - \mathscr{G}(u)(u - \bar{u})\|_{L^2} \leq \gamma \|u - \bar{u}\|_{L^2}$$
$$\forall u \in L^2(\Omega), \quad \|u - \bar{u}\|_{L^2} \leq \bar{\delta},$$

and in the discrete case

$$\|\mathscr{F}_h(u_h) - \mathscr{F}_h(\bar{u}_h) - \mathscr{G}_h(u_h)(u_h - \bar{u}_h)\|_{U_h} \leq \gamma \|u_h - \bar{u}_h\|_{U_h}$$
$$\forall u_h \in U_h, \quad \|u_h - \bar{u}_h\|_{U_h} \leq \bar{\delta}, \quad \forall h \in (0, \bar{h}].$$

PROOF. Note that $\mathfrak{p}$ and $\mathfrak{p}_h$ are affine linear. Hence

$$\mathfrak{p}(u_h) - \mathfrak{p}(\bar{u}_h) - \mathfrak{p}'(u_h)(u_h - \bar{u}_h) = 0.$$

This fact, together with Lemmas 4.1 and 4.2, yields the assertion. □

Our mesh-independence result now follows from [14, Theorem 3]. Here we only state the theorem and refer to [14] for the proof.

THEOREM 4.4. *Let the assumptions of Proposition 4.3 hold true. Then, for arbitrarily fixed $\theta \in (0, 1)$, there exists $\bar{\delta} > 0$ and $\bar{h} > 0$ such that for all $h \leq \bar{h}$ and $k \in \mathbb{N}_0$*

$$\left\|u^{k+1} - \bar{u}\right\|_{L^2} \leq \theta \left\|u^k - \bar{u}\right\|_{L^2},$$
$$\left\|u_h^{k+1} - \bar{u}_h\right\|_{L^2} \leq \theta \left\|u_h^k - \bar{u}_h\right\|_{U_h}$$

provided that $\max(\|u^0 - \bar{u}\|_{L^2}, \|u_h^0 - \bar{u}_h\|_{U_h}) \leq \bar{\delta}$.

In Section 5 we provide a validation of the above result. In fact, in our numerical tests we even observe strong mesh independence, that is, for h sufficiently small and for fixed $\epsilon > 0$ the algorithm stops with $\|u_h^k - \bar{u}_h\|_{U^h} \leq \epsilon$ after essentially the same number of iterations regardless of the mesh size of discretization.

5. Numerics

Now we report on the numerical behaviour of Algorithm 3. Among other aspects, we aim at numerically verifying our mesh-independence result, Theorem 4.4. Further we study our algorithm when solving degenerate problems.

Below we denote by $y_h, u_h, \ldots$ the coefficient vectors for the corresponding finite element representation. We assume that $y_h, p_h \in \mathbb{R}^{n_h^y}$ and $u_h, \lambda_h \in \mathbb{R}^{n_h^u}$, where n_h^y and n_h^u depend on the mesh size of discretization. Concerning the data a, b, c, f (which we assume to be in $L^2(\Omega)$ for simplicity), and y_d we employ a piecewise constant (over every triangle T) discretization. We therefore have $a_h, b_h, c_h, f_h, y_{d,h} \in \mathbb{R}^{n_h^u}$, respectively. For later use we also introduce the mass matrices $M_h^0 \in \mathbb{R}^{n_h^u \times n_h^y}$ and $M_h^1 := (M_h^0)^\top M_h^0 \in \mathbb{R}^{n_h^y \times n_h^y}$. Further, we denote by C_h the diagonal matrix $\operatorname{diag}(c_h)$ with entries $c_{h,i}$, $i = 1, \ldots, n_h^u$.

In all of our test runs reported on below, we initialize Algorithm 3 by setting $\lambda_h^0 = 0$ and computing (y_h^0, u_h^0, p_h^0) as the solution to

$$A_h y_h - \left(M_h^0\right)^\top u_h = \left(M_h^0\right)^\top f_h, \quad A_h^\top p_h + M_h^1 y_h = \left(M_h^0\right)^\top y_{d,h}, \quad \alpha u_h - M_h^0 p_h = 0.$$

This procedure provides our initial u_h^0. Note that $(y_h^0, u_h^0, p_h^0, \lambda_h^0)$ corresponds to the solution of the unconstrained version of (1.1). We stop the algorithm if the active sets $(\mathcal{A}_{a,h}^{k+1}, \mathcal{A}_{b,h}^{k+1})$ and $(\mathcal{A}_{a,h}^{k}, \mathcal{A}_{b,h}^{k})$ coincide for $k \geq 1$ or as soon as the norm of the residual of the first-order optimality system drops below a prescribed tolerance. Above, we denote by $\mathcal{A}_{a,h}^{k}, \mathcal{A}_{b,h}^{k}$ the discrete analogues of $\mathcal{A}_a^k, \mathcal{A}_b^k$ in (3.15) and (3.16). Assuming that the linear systems are solved exactly in Algorithm 3, then the first stopping rule yields the exact solution of the discrete problem.

THEOREM 5.1. *Let $\{u_h^k\}$ be computed by Algorithm 3. If for $k \geq 1$ we have $(\mathcal{A}_{a,h}^{k+1}, \mathcal{A}_{b,h}^{k+1}) = (\mathcal{A}_{a,h}^{k}, \mathcal{A}_{b,h}^{k+1})$, then $(u_h^{k+1}, y_h^{k+1}, p_h^{k+1}, \lambda_h^{k+1})$ solves*

$$A_h y_h - \left(M_h^0\right)^\top u_h = \left(M_h^0\right)^\top f_h, \tag{5.1}$$

$$A_h^\top p_h + \left(M_h^0\right)^\top \lambda_h + M_h^1 y_h = \left(M_h^0\right)^\top y_{d,h}, \tag{5.2}$$

$$\alpha\, u_h - M_h^0 p_h + C_h \lambda_h = 0, \tag{5.3}$$

and the complementarity system

$$a_h \leq C_h u_h + M_h^0 y_h \leq b_h, \tag{5.4}$$

$$\lambda_{h,i} < 0 \quad \text{for all } i \in \mathcal{A}_{a,h}^{k+1}, \tag{5.5}$$

$$\lambda_{h,i} > 0 \quad \text{for all } i \in \mathcal{A}_{b,h}^{k+1}, \tag{5.6}$$

$$\lambda_{h,i} = 0 \quad \text{for all } i \in \mathcal{I}_h^{k+1}. \tag{5.7}$$

PROOF. First notice that due to the equivalence of Algorithm 3 to the discrete analogue of Algorithm 2 the system (5.1)-(5.3) is satisfied in every iteration k. Hence, we only need to check (5.4)–(5.6). If $\mathcal{A}_{a,h}^{k+1} = \mathcal{A}_{a,h}^{k}$, then, for $i \in \mathcal{A}_{a,h}^{k+1}$, we have $c_{h,i}\, u_{h,i}^{k+1} + (M_h^0 y_h^{k+1})_i = a_{h,i}$ and

$$0 > \lambda_{h,i}^{k+1} + \alpha\, c_{h,i}^{-2}\left(c_{h,i}\, u_{h,i}^{k+1} + (M_h^0 y_h^{k+1})_i - a_{h,i}\right) = \lambda_{h,i}^{k+1}.$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

Similarly, if $\mathscr{A}_{b,h}^{k+1} = \mathscr{A}_{b,h}^{k}$, then we get

$$\lambda_{h,i}^{k+1} > 0 \quad \text{and} \quad c_{h,i}\, u_{h,i}^{k+1} + (M_h^0 y_h^{k+1})_i = b_{h,i} \quad \text{for all } i \in \mathscr{A}_{b,h}^{k+1}.$$

Finally, for $i \in \mathscr{I}_h^{k+1} = \mathscr{I}_h^k$ we have

$$\lambda_{h,i}^{k+1} = 0 \quad \text{and} \quad a_{h,i} \le c_{h,i} u_{h,i}^{k+1} + (M_h^0 y_h^{k+1})_i \le b_{h,i}.$$

This ends the proof. □

Subsequently we focus on the following test problems. In all cases we have $\Omega = (0, 1)^2$.

5.1. Problems Next we specify our test problems.

EXAMPLE 1. With $\mathbf{x} = (x_1, x_2)$, the data for this example are: $y_d(\mathbf{x}) = \sin(5\pi x_1) + \cos(10\pi x_2)$, $f(\mathbf{x}) = -\exp(x_1)$, $a(\mathbf{x}) = -0.1|x_1 + x_2 - 1|$, $b = 0.045 + a$, $c(\mathbf{x}) = 0.0025(1 + (x_1 - 0.5)^2 + (x_2 - 0.5)^2)$ and $\alpha = 0.001$. In Figure 1 we depict the optimal state $\bar{y}_h$, the optimal control $\bar{u}_h$, the corresponding optimal Lagrange multiplier $\bar{\lambda}_h$ and the active set (black regions) for $h = 1/256$. Note that due to the active regions next to the boundary in the lower left and upper right corners of the domain and the requirement $\bar{y}|_{\partial\Omega} = 0$, the Lagrange multiplier becomes large in these regions.

EXAMPLE 2. The data are as in Example 1 except for the upper bound which is now $b = -a$. Notice that we have $a = b$ along the diagonal $\{\mathbf{x} \in \Omega : |x_1 + x_2 - 1| = 0\}$ of the unit square ($= \Omega$). As a consequence $a < b$ is violated on a set of measure zero in $\mathbb{R}^2$. The discrete optimal state, control, multiplier and active set are shown in Figure 2.

EXAMPLE 3. The data are as in Example 1 except for c and α, which are now $c(\mathbf{x}) = 7.5\text{E-}4\big(1 + (x_1 - 0.5)^2 + (x_2 - 0.5)^2\big)$ and $\alpha = 0.1$. Notice that this choice increases the quotient α/c^2 when compared to the one in Example 1. This fact and the resulting active-set structure make this problem more challenging than Example 1. The discrete optimal state and the active set can be found in Figure 3. Due to the requirements $\bar{y}|_{\partial\Omega} = 0$ and $a \le c\,\bar{u} + \bar{y} \le b$, the control action close to the lower left and upper right corners (where we have $b < 0$) is more pronounced then the one in Example 1.

EXAMPLE 4. The construction of this test problem yields a highly degenerate, that is, very flat transition of $c\,\bar{u} + \bar{y}$ into the active set. As can be seen from Figure 4, $\bar{\lambda}$ is also highly degenerate. This usually poses severe difficulties for numerical algorithms

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

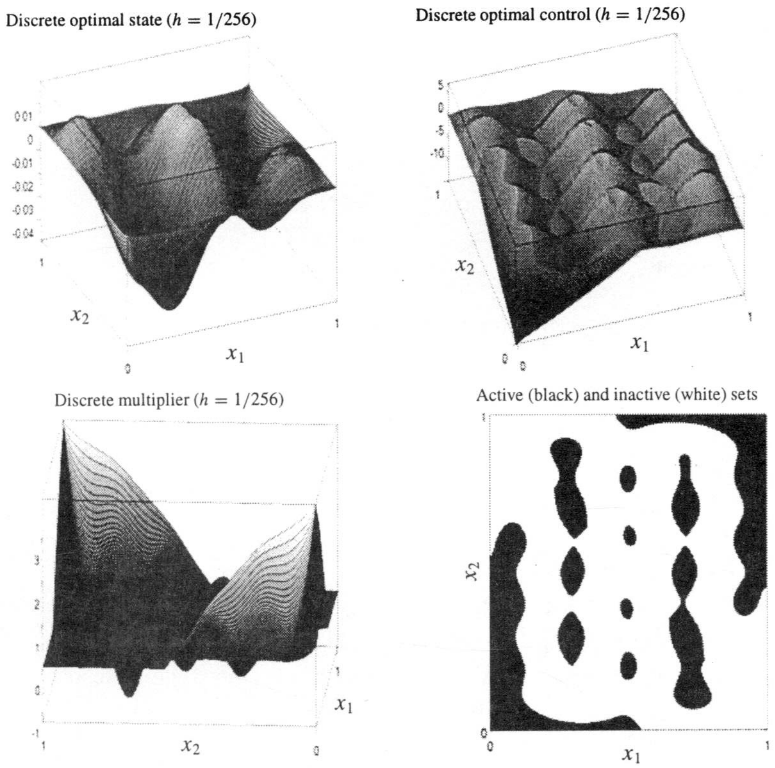

FIGURE 1. Example 1: Optimal state $\bar{y}_h$ (upper left), optimal control $\bar{u}_h$ (upper right), corresponding Lagrange multiplier $\bar{\lambda}_h$ (lower left), and active set (in black; lower right) for $h = 1/256$.

due to possible instabilities in the active set detection. For the problem formulation we slightly extend the objective function in (1.1) by considering

$$J(y, u) = \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u - u_d\|_{L^2}^2.$$

The remaining problem data are as follows: Set $g(\mathbf{x}) := x_1^8(1 - x_1)x_2(1 - x_2)$, $c \equiv 0.5$, and define the optimal control and state by setting $\bar{u} := g$, $\bar{y} := g - c\,\bar{u}$. The source term in the state equation is chosen as $f = A(g - c\,\bar{u}) - \bar{u}$. Next fix $t(\mathbf{x}) := (x_1 - 0.5)^2 + (x_2 - 0.3)^2 - 0.4$ and determine $\mathscr{A}_t := \{t \geq 0\}$ and $\mathscr{I}_t = \Omega \setminus \mathscr{A}_t$. Then the upper bound is defined by $b := \chi_{\mathscr{A}_t} g + \chi_{\mathscr{I}_t}(g + t^8)$, and the lower bound is $a \equiv -\infty$. The optimal multiplier is $\bar{\lambda} = \chi_{\mathscr{A}_t}|t|^4$, and the adjoint state is $\bar{p} = 10x_1(1 - x_1)x_2(1 - x_2)\sin(5\pi x_1)$. Then the desired state is computed as $y_d = A\bar{p} + \bar{\lambda} + \bar{y}$. Finally, we set $u_d = \bar{u} + \alpha^{-1}(c\,\bar{\lambda} - \bar{p})$ with $\alpha = 1\text{E-}5$. Figure 4 contains the discrete solution, the corresponding multiplier and the active set (black).

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

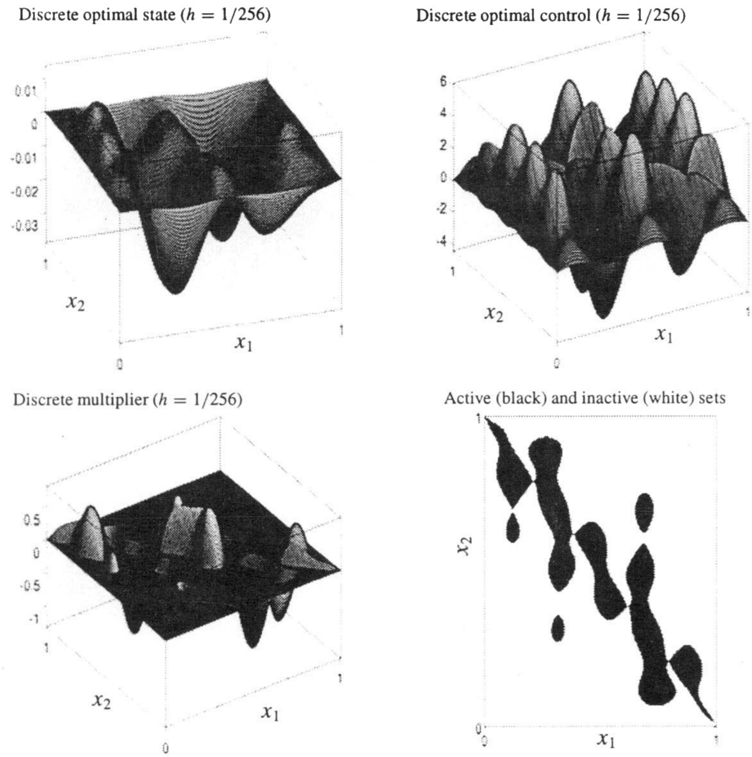

FIGURE 2. Example 2: Optimal state $\bar{y}_h$ (upper left), optimal control $\bar{u}_h$ (upper right), corresponding Lagrange multiplier $\bar{\lambda}_h$ (lower left), and active set (in black; lower right) for $h = 1/256$.

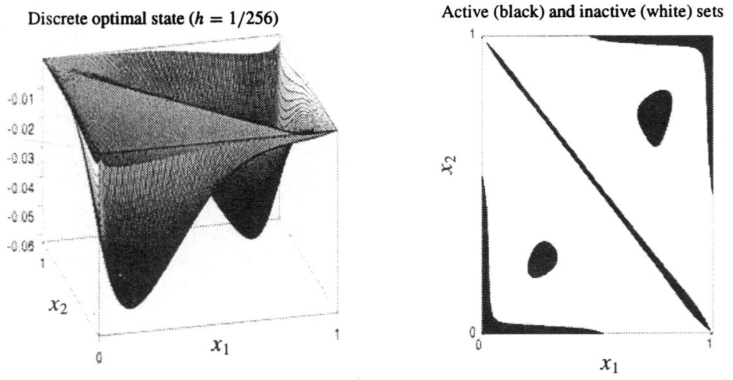

FIGURE 3. Example 3: Optimal state $\bar{y}_h$ (left) for $h = 1/256$ and the active set (in black; right) for $h = 1/512$.

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

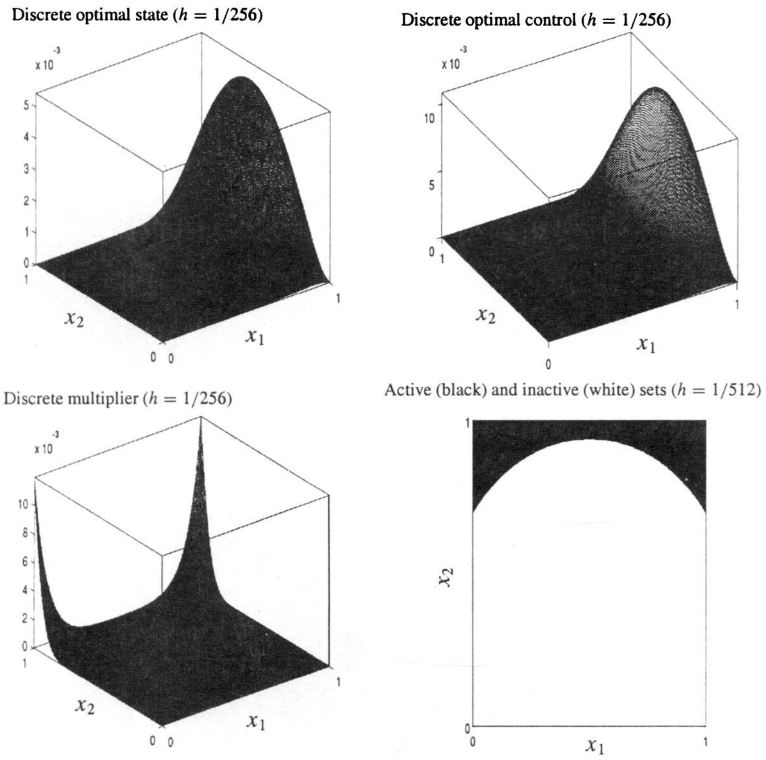

FIGURE 4. Example 4: Optimal state $\bar{y}_h$ (upper left), optimal control u_h (upper right), corresponding Lagrange multiplier $\bar{\lambda}_h$ (lower left), each for $h = 1/256$, and active set (in black; lower right) for $h = 1/512$.

5.2. Mesh independence In the following tables, for the numerical validation of Theorem 4.4 we provide the quotients

$$q_h^k = \frac{\|u_h^k - u_h^*\|_{U_h}}{\|u_h^{k-1} - u_h^*\|_{U_h}}, \qquad k = 1, 2, 3, \dots,$$

where u_h^* denotes the discrete optimal control which is obtained by restricting the discrete solution for the mesh size $1/512$ to the current mesh of mesh size h. Here we use the restriction operator corresponding to the nine-point-interpolation scheme. For Examples 1, 2 and 4, Tables 1–3 depict the behaviour of q_h^k for various mesh sizes (rows). The qualitative behaviour of q_h^k for Example 3 is similar.

Upon studying Tables 1-3 we can draw the following conclusions: For each fixed mesh size (row) Algorithm 3 converges superlinearly, that is, the quotients q_h^k tend to zero. The increase of q_h^k from the next to the last column, respectively in each table, is related to the restriction process for obtaining u_h^*. If we study the behaviour of

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

these critical quotients along the columns, we observe a stabilizing (even decreasing) behaviour. The mesh independence (validation of Theorem 4.4) of Algorithm 3 is reflected by q_h^k along the columns of each table. For each test example we clearly detect a stable behaviour of q_h^k as the mesh is refined. As a consequence, we obtain a mesh-independent superlinear convergence. We further observe that for each example the algorithm requires essentially the same number of iterations until successful termination. This effect, which is beyond our theoretical result, is known as strong mesh independence; see [2] for smooth operator equations. Also note that the mesh-independent convergence of our algorithm is not influenced by the degeneracy of the solution of Example 4. Although the optimal solutions, the adjoint states and the Lagrange multipliers on various meshes reflect this numerical stability, the active set detection is indeed affected. In Figure 5, we show the active sets upon termination of the algorithm when solving Example 4 for mesh sizes h ranging from $1/32$ to $1/256$. The active set for $h = 1/512$ is depicted in Figure 4 (lower right plot). It coincides with the true active set on the underlying mesh.

TABLE 1. Example 1: Mesh-independent behaviour of (convergence) quotient q_h^k.

h	q_h^k					
1/32	0.633	0.386	0.315	0.389	0.977	–
1/64	0.570	0.394	0.309	0.161	0.577	–
1/128	0.542	0.394	0.313	0.147	0.149	0.969
1/256	0.528	0.396	0.314	0.145	0.047	0.624
1/512	0.521	0.395	0.313	0.144	0.036	0.002

TABLE 2. Example 2: Mesh-independent behaviour of (convergence) quotient q_h^k.

h	q_h^k					
1/32	0.930	0.362	0.185	0.773	–	–
1/64	0.937	0.393	0.149	0.233	–	–
1/128	0.945	0.391	0.154	0.085	0.585	–
1/256	0.946	0.397	0.152	0.063	0.157	–
1/512	0.946	0.397	0.152	0.062	0.015	5E-5

5.3. Comparison with a short-step path-following interior-point method. The recent paper [19] establishes a convergence result for a short-step path-following interior-point method (SPF) in function space for the solution of the unilaterally constrained version of (1.1). We point out that for more progressive versions of path-following interior-point methods (such as long-step methods or predictor-corrector algorithms; see, for example, [25]) to date no function space analysis is available. It is therefore of interest to compare SPFs with our semismooth Newton (SSN), or equivalently primal-dual active-set, framework.

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

TABLE 3. Example 4: Mesh-independent behaviour of (convergence) quotient q_h^k.

h	q_h^k					
1/32	0.010E-1	0.403E-1	0.586E-1	0.540E-1	0.416E-1	–
1/64	0.013E-1	0.403E-1	0.728E-1	0.526E-1	0.286E-1	–
1/128	0.013E-1	0.365E-1	0.649E-1	0.521E-1	0.312E-1	3.151E-1
1/256	0.013E-1	0.357E-1	0.681E-1	0.499E-1	0.302E-1	3.282E-1
1/512	0.013E-1	0.358E-1	0.682E-1	0.503E-1	0.274E-1	0.979E-1

Consider the duality measure

$$\mu_h^k := \frac{1}{2n_h^u}\sum_{i=1}^{n_h^u} z_{h,i}w_{h,i},$$

with $z_h = ((\lambda_{a,h}^k)^\top, (\lambda_{b,h}^k)^\top)^\top \in \mathbb{R}^{2n_h^u}$ and $w_h = (w_{a,h}^\top, w_{b,h}^\top)^\top \in \mathbb{R}^{2n_h^u}$, $w \geq 0$, denotes a vector of slack variables such that

$$C_h u_h^k + M_h^0 y_h^k - a_h - w_{a,h} = 0,$$
$$C_h u_h^k + M_h^0 y_h^k - b_h + w_{b,h} = 0.$$

We stop the interior-point method as soon as μ_h^k drops below some prescribed tolerance $\epsilon_\mu > 0$. In Table 4 we report on the number of iterations required until successful termination for various mesh sizes h.

TABLE 4. Comparison of iteration numbers required by a short-step path-following interior-point method (SPF) and by our semismooth Newton method (SSN).

Alg.	h =1/32	1/64	1/128	1/256	1/512
	Example 1				
SPF	17	17	17	17	17
SSN	5	5	6	6	6
	Example 2				
SPF	19	19	19	19	19
SSN	4	4	5	5	6
	Example 3				
SPF	64	64	63	63	63
SSN	9	10	10	10	11
	Example 4				
SPF	17	17	18	18	18
SSN	5	5	6	6	6

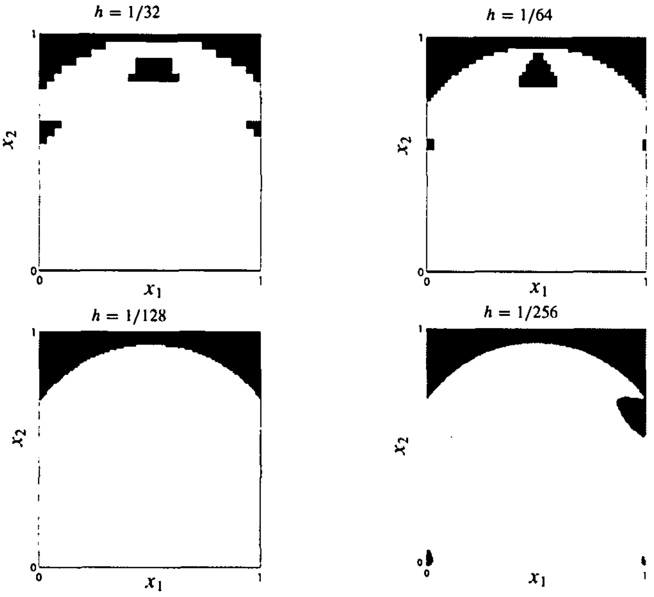

FIGURE 5. Example 4: Active (black) and inactive (white) sets upon termination of Algorithm 3 for various mesh sizes.

From the results in Table 4 we find that in all test cases the SSN requires a significantly smaller number of iterations than the SPF. In Figure 6, for the degenerate Example 4 we depict the active-set estimates upon termination of the SPF for various mesh sizes.

Comparing these results with the ones for the SSN in Figures 4 and 5, we see that the SSN yields better approximations of the true active sets than the SPF.

5.4. Coarse-to-fine sweep. Next we report on the speed-up of the solution process when combining Algorithm 3 with a coarse-to-fine sweep with respect to the underlying meshes. Starting on the coarsest mesh ($h = 1/4$ on our regular triangulation) Algorithm 3 is used for computing the numerical solution which is then prolongated to the next finer mesh. The prolongated coarse-mesh solution is taken as the starting point for Algorithm 3 on the fine mesh. This cycle is repeated until a desired mesh size is reached. In Table 5 we report on the results for Examples 1 and 3. In the penultimate column, in parenthesis we provide the number of iterations needed by Algorithm 3 on the finest mesh ($h = 1/512$ in our case) without the coarse-to-fine technique. The last column contains the ratio of the CPU-time consumed by Algorithm 3 with coarse-to-fine feature versus the CPU-time without the mesh refinement.

From the results in Table 5 we infer that the coarse-to-fine sweep speeds up the

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

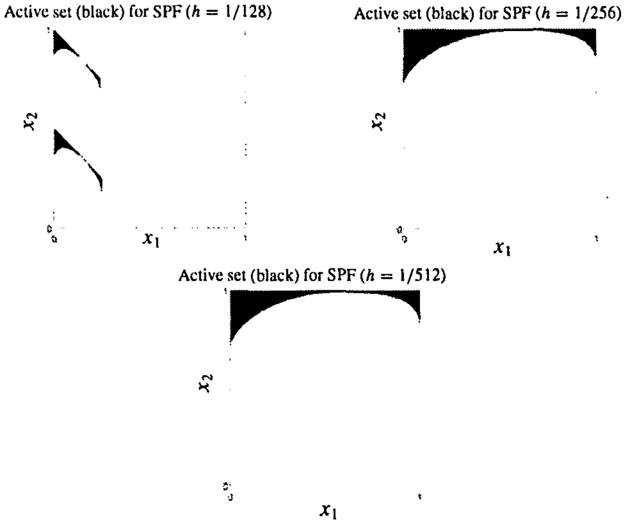

FIGURE 6. Short-step path-following interior-point method (SPF) for Example 4: Active sets upon termination of the SPF for mesh sizes $h = 1/128$, $h = 1/256$ and $h = 1/512$ (from left to right).

TABLE 5. Coarse-to-fine sweep combined with Algorithm 3 for $h_i = 2^{-i}$, $i = 2, 3, \dots, 9$ (from left to right).

Problem	#iterations/h								CPU-ratio
Example 1	3	3	3	2	2	1	2	1 (6)	0.41
Example 3	1	2	2	7	3	2	2	1 (11)	0.26

overall solution process (CPU-ratio < 1) considerably. It provides excellent initial points on the fine meshes ($h \le 1/128$ in our examples) such that Algorithm 3 requires at most two iterations until successful termination on these fine meshes. The results for Examples 2 and 4 are similar to the ones for Example 1. In the case of Example 3 we point out that only for $h = 1/32$ (and smaller) the problem features are resolved reasonably well (with increasing accuracy as h decreases). Therefore, Algorithm 3 requires seven iterations for $h = 1/32$ where it encounters these problem features for the first time in our mesh refinement process. Again, the subsequent applications of Algorithm 3 benefit from excellent initial points.

6. Asymptotics and Hölder regularity

Whenever (1.1) is the result of a Lavrentiev-type regularization of the state-constrained problem (1.2) with $c \equiv \epsilon > 0$, it is of interest to study the asymptotic

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

to ϵ when solving (1.2) numerically, certain convergence and smoothness properties of the solution (y_ϵ, u_ϵ) of (1.1) with $c \equiv \epsilon$ with respect to ϵ are of interest. These issues are addressed next. For ease of presentation we restrict ourselves to the unilaterally constrained case with $f \equiv 0$, that is, we consider

$$\left.\begin{array}{lll} \text{minimize} & J(y,u) := \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 & \\ \text{over} & (y,u) \in H_0^1(\Omega) \times L^2(\Omega), & \\ \text{subject to} & Ay = u & \text{in } \Omega, \\ & \epsilon\, u + y \le b & \text{a.e. in } \Omega \end{array}\right\} \tag{6.1}$$

which is the Lavrentiev regularization of the state-constrained problem

$$\left.\begin{array}{lll} \text{minimize} & J(y,u) := \frac{1}{2}\|y - y_d\|_{L^2}^2 + \frac{\alpha}{2}\|u\|_{L^2}^2 & \\ \text{over} & (y,u) \in H_0^1(\Omega) \times L^2(\Omega), & \\ \text{subject to} & Ay = u & \text{in } \Omega, \\ & y \le b & \text{a.e. in } \Omega. \end{array}\right\} \tag{6.2}$$

We assume that the coefficients of A are sufficiently smooth and $b \in H^2(\Omega)$ with $b|_\Gamma > 0$. It is well known (see [8, 9]) that there exists a unique optimal solution $(y^*, u^*) \in H^2(\Omega) \cap H_0^1(\Omega) \times L^2(\Omega)$ of (6.2) which is characterized by the existence of $(p^*, \lambda^*) \in L^2(\Omega) \times \mathscr{C}^*(\Omega)$ such that

$$\begin{aligned} & Ay^* = u^*, \quad y^* \le b \quad \text{a.e. in } \Omega, \\ & \alpha u^* = p^*, \\ & (y^* - y_d, \varphi_1)_{L^2} + (p^*, A\varphi_1)_{L^2} + \langle \lambda^*, \varphi_1 \rangle_{\mathscr{C}^*,\mathscr{C}} = 0, \\ & \langle \lambda^*, \varphi_2 - y^* \rangle_{\mathscr{C}^*,\mathscr{C}} \le 0 \end{aligned}$$

for all $\varphi_1, \varphi_2 \in H^2(\Omega) \cap H_0^1(\Omega)$ with $\varphi_2 \le b$. Further, in [8] it is shown that under sufficient regularity of b and the active set $\mathscr{A}^* = \{x \in \Omega : y^*(x) = b(x)\}$ the multiplier λ^* can be decomposed into an absolutely continuous part $\lambda^*|_{\mathscr{A}^*} \in L^2(\Omega)$ and $\lambda^*|_{\mathscr{I}^*} = 0$, and a singular part $\lambda_s^* \in H^{1/2}(\Sigma^*)$ concentrated on the boundary Σ^* between the active set $\mathscr{A}^*$ and the inactive set $\mathscr{I}^* = \Omega \setminus \mathscr{A}^*$.

Our first goal is to show that the optimal state-control pair (y_ϵ, u_ϵ) associated with Problem (6.1) approaches (y^*, u^*) as ϵ tends to zero. In what follows we use $Y := H^2(\Omega) \cap H_0^1(\Omega)$ and the bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ defined by $a(v, w) = \langle Av, w \rangle_{H^{-1}, H_0^1(\Omega)}$. Note that a satisfies

$$a(v,v) \ge \omega \|v\|_{H_0^1}^2 = \omega \|\nabla v\|_{L^2}^2 \quad \text{and} \quad a(v,w) \le C\|v\|_{H_0^1}\|w\|_{H_0^1} \tag{6.3}$$

for constants $\omega > 0$ and $C > 0$.

PROPOSITION 6.1. *Let* $\{\epsilon_n\} \subset \mathbb{R}^{++}$ *denote a sequence with* $\epsilon_n \downarrow 0$ *for* $n \to \infty$ *and let* $(\hat{y}, \hat{u}) \in Y \times L^2(\Omega)$ *be a feasible point of* (6.2). *Further, assume that* $\epsilon_n \hat{u}_{\epsilon_n} + \hat{y}_{\epsilon_n} := \hat{y}$ *and* $A\hat{y}_{\epsilon_n} = \hat{u}_{\epsilon_n}$. *Then* $(y_{\epsilon_n}, u_{\epsilon_n}) \to (\hat{y}, \hat{u})$ *strongly in* $Y \times L^2(\Omega)$. *Moreover,*

$$\left\| \hat{y}_{\epsilon_n} - \hat{y} \right\|_{L^2} = \mathcal{O}(\epsilon_n) \quad \textit{and} \quad \left\| \nabla(\hat{y}_{\epsilon_n} - \hat{y}) \right\|_{L^2} = o(\sqrt{\epsilon_n}) \quad \textit{for } n \to \infty$$

and $\|\hat{u}_{\epsilon_n}\|_{L^2} \leq \|\hat{u}\|_{L^2}$ *for all* $n \in \mathbb{N}$.

PROOF. By definition we have

$$\epsilon_n \hat{u}_{\epsilon_n} + \hat{y}_{\epsilon_n} := \hat{y} \leq b \tag{6.4}$$

and $\hat{y}_{\epsilon_n} = (A + \epsilon_n^{-1}\,\mathrm{id})^{-1}(\epsilon_n^{-1}\hat{y})$. Then the state equation yields

$$\epsilon_n a(\hat{y}_{\epsilon_n}, \hat{y}_{\epsilon_n} - \hat{y})_{L^2} + (\hat{y}_{\epsilon_n}, \hat{y}_{\epsilon_n} - \hat{y})_{L^2} = (\hat{y}, \hat{y}_{\epsilon_n} - \hat{y})_{L^2}.$$

Since $\hat{y}, \hat{y}_{\epsilon_n} \in Y$ we get

$$\left\| \hat{y}_{\epsilon_n} - \hat{y} \right\|_{L^2}^2 \leq \epsilon_n (-A\hat{y}, \hat{y}_{\epsilon_n} - \hat{y})_{L^2} \leq \epsilon_n C \left\| \hat{y}_{\epsilon_n} - \hat{y} \right\|_{L^2} \tag{6.5}$$

with $C > 0$ independent of ϵ_n. This implies that $\hat{y}_{\epsilon_n} \to \hat{y}$ strongly in $L^2(\Omega)$ and that $\|\hat{y}_{\epsilon_n} - \hat{y}\|_{L^2} = \mathcal{O}(\epsilon_n)$ for $n \to \infty$. Further note that due to (6.4) and (6.5) we have

$$\|\hat{u}_{\epsilon_n}\|_{L^2} = \epsilon_n^{-1} \|\hat{y}_{\epsilon_n} - \hat{y}\|_{L^2} \leq \|A\hat{y}\|_{L^2} \leq C. \tag{6.6}$$

Hence, once again from the state equation, we obtain

$$\begin{aligned} \omega \|\nabla(y_{\epsilon_n} - \hat{y})\|_{L^2}^2 &\leq -\epsilon_n^{-1} \|\hat{y}_{\epsilon_n} - \hat{y}\|_{L^2}^2 + (A\hat{y}, \hat{y} - \hat{y}_{\epsilon_n})_{L^2} \\ &\leq (\|A\hat{y}\|_{L^2} - \epsilon_n^{-1} \|\hat{y}_{\epsilon_n} - \hat{y}\|_{L^2}) \|\hat{y}_{\epsilon_n} - \hat{y}\|_{L^2} \\ &= \mathcal{O}(\epsilon_n) \quad \text{for} \quad n \to \infty. \end{aligned} \tag{6.7}$$

As a consequence we have $\hat{y}_{\epsilon_n} \to \hat{y}$ strongly in $H_0^1(\Omega)$ and $\|\nabla(\hat{y}_{\epsilon_n} - \hat{y})\|_{L^2} = \mathcal{O}(\sqrt{\epsilon_n})$ for $n \to \infty$.

Concerning the convergence of $\left\{\hat{u}_{\epsilon_n}\right\}$ we observe that

$$a(\hat{y}, \varphi)_{L^2} = (\hat{u}, \varphi)_{L^2} \quad \text{and} \quad a(\hat{y}_{\epsilon_n}, \varphi)_{L^2} = (\hat{u}_{\epsilon_n}, \varphi)_{L^2}$$

for $\varphi \in H_0^1(\Omega)$. By subtraction, these two equations yield

$$a\left(\hat{y} - \hat{y}_{\epsilon_n}, \varphi\right)_{L^2} = \left(\hat{u} - \hat{u}_{\epsilon_n}, \varphi\right)_{L^2} \qquad \text{for } \varphi \in H_0^1(\Omega).$$

Since $\hat{y}_{\epsilon_n}$ converges strongly in $H_0^1(\Omega)$, by a density argument we get $\hat{u}_{\epsilon_n} \to \hat{u}$ weakly in $L^2(\Omega)$. From $A\hat{y}_{\epsilon_n} = \hat{u}_{\epsilon_n} = \epsilon_n^{-1}(\hat{y} - \hat{y}_{\epsilon_n})$ we infer that

$$\omega \left\| \nabla(\hat{y}_{\epsilon_n} - \hat{y}) \right\|_{L^2}^2 + \left(A\hat{y}, \hat{y}_{\epsilon_n} - \hat{y}\right)_{L^2} \leq -\epsilon_n^{-1} \left\| \hat{y}_{\epsilon_n} - \hat{y} \right\|_{L^2}^2$$

and hence

$$\epsilon_n \left(\hat{u}, \hat{u}_{\epsilon_n}\right)_{L^2} = \left(-A\hat{y}, \hat{y}_{\epsilon_n} - \hat{y}\right)_{L^2} \geq \epsilon_n \left\| \hat{u}_{\epsilon_n} \right\|_{L^2}^2 + \omega \left\| \nabla(\hat{y}_{\epsilon_n} - \hat{y}) \right\|_{L^2}^2 \geq 0.$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

Next consider

$$\left\|\hat u - \hat u_{\epsilon_n}\right\|_{L^2}^2 = \left\|\hat u\right\|_{L^2}^2 - 2\left(\hat u, \hat u_{\epsilon_n}\right)_{L^2} + \left\|\hat u_{\epsilon_n}\right\|_{L^2}^2 \le \left\|\hat u\right\|_{L^2}^2 - \left\|\hat u_{\epsilon_n}\right\|_{L^2}^2 .$$

Consequently, we obtain

$$\left\|\hat u_{\epsilon_n}\right\|_{L^2} \le \left\|\hat u\right\|_{L^2} \quad \text{for all } n \in \mathbb{N}.$$

Due to the weak lower semicontinuity of norms we get

$$J(\hat y, \hat u) \le \liminf_{n\to\infty} J(\hat y_{\epsilon_n}, \hat u_{\epsilon_n}) \le \liminf_{n\to\infty} J(\hat y_{\epsilon_n}, \hat u) = \lim_{n\to\infty} J(\hat y_{\epsilon_n}, \hat u) = J(\hat y, \hat u) .$$

This implies $\|\hat u_{\epsilon_n}\|_{L^2} \to \|\hat u\|_{L^2}$ for $n \to \infty$. Now the weak convergence of $\{\hat u_{\epsilon_n}\}$ in $L^2(\Omega)$ together with the convergence of norms yields $\hat u_{\epsilon_n} \to \hat u$ strongly in $L^2(\Omega)$.

Considering $A\hat y = \hat u, \hat u_{\epsilon_n} = \epsilon_n^{-1}(\hat y - \hat y_{\epsilon_n})$ and $\|\hat u_{\epsilon_n}\|_{L^2} \to \|\hat u\|_{L^2}$, we derive from (6.7)

$$\left\|\nabla(\hat y_{\epsilon_n} - \hat y)\right\|_{L^2} = o\left(\sqrt{\epsilon_n}\right) \quad \text{for } n \to \infty.$$

The strong convergence of y_{ϵ_n} in $H^2(\Omega)$ follows from standard elliptic regularity estimates; see, for example, [11, Theorem 8.13]. In fact, we have

$$A\left(\hat y_{\epsilon_n} - \hat y\right) = \hat u_{\epsilon_n} - \hat u \quad \text{in } \Omega.$$

Since $\|\hat y_{\epsilon_n} - \hat y\|_{L^2} = \mathcal{O}(\epsilon_n)$ and $u_{\epsilon_n} \to \hat u$ strongly in $L^2(\Omega)$, we obtain

$$\left\|\hat y_{\epsilon_n} - \hat y\right\|_{H^2} \le C\left(\left\|\hat y_{\epsilon_n} - \hat y\right\|_{L^2} + \left\|u_{\epsilon_n} - \hat u\right\|_{L^2}\right) \to 0 \quad \text{for } n \to \infty. \qquad \square$$

We note that due to the boundedness of $\{\|\hat u_{\epsilon_n}\|_{L^2}\}$ (see (6.6)), we have

$$\epsilon_n \left\|\hat u_{\epsilon_n}\right\|_{L^2} \to 0 \quad \text{for } n \to \infty.$$

Next we study the convergence behaviour of (y_ϵ, u_ϵ) to (y^*, u^*) as $\epsilon \downarrow 0$. For this purpose we invoke the assumption

$$J(y^*, u^*) \le \inf\left\{\mathcal{L}_{\mathcal{A}^*}(y, u, \lambda^*) : (y, u) \in Y \times L^2(\Omega) \text{ with } Ay = u\right\}, \qquad (6.8)$$

where we use

$$\mathcal{L}_{\mathcal{A}^*}(y, u, \lambda^*) = J(y, u) + \left(\lambda^*|_{\mathcal{A}^*}, (y - b)|_{\mathcal{A}^*}\right)_{L^2} .$$

Notice that (6.8) relates to some type of a saddle-point condition (see, for example, [16]) and we require the additional multiplier regularity addressed in connection with (6.2) at the beginning of this section.

THEOREM 6.2. *Let $\{\epsilon_n\} \subset \mathbb{R}^{++}$ denote a sequence with $\epsilon_n \downarrow 0$ for $n \to \infty$ and assume that assumption* (6.8) *holds true. Then the sequence $\{(y_{\epsilon_n}, u_{\epsilon_n})\}$ is uniformly bounded in $Y \times L^2(\Omega)$ and*

$$\left\|y_{\epsilon_n} - y^*\right\|_{L^2} = \mathcal{O}(\sqrt{\epsilon_n}) \quad \textit{and} \quad \left\|u_{\epsilon_n} - u^*\right\|_{L^2} = \mathcal{O}\left(\sqrt{\epsilon_n}\right) \quad \textit{for } n \to \infty.$$

PROOF. Let $\mathscr{L}_\epsilon : Y \times L^2(\Omega) \times H_0^1(\Omega) \times L^2(\Omega) \to \mathbb{R}$ with

$$\mathscr{L}_\epsilon(y, u, p, \lambda) = J(y, u) + \langle Ay - u, p\rangle_{H^{-1},H_0^1} + (\lambda, \epsilon u + y - b)_{L^2}$$

denote the Lagrange function associated with (6.1). For ease of notation we use $w = (y, u)$. Note that by first-order optimality we have

$$\nabla_w \mathscr{L}_{\epsilon_n}\big(y_{\epsilon_n}, u_{\epsilon_n}, p_{\epsilon_n}, \lambda_{\epsilon_n}\big) = 0.$$

Further, we observe that

$$\big\langle \nabla_{ww} \mathscr{L}_{\epsilon_n}\big(y_{\epsilon_n}, u_{\epsilon_n}, p_{\epsilon_n}, \lambda_{\epsilon_n}\big)(y, u), (y, u)\big\rangle = \|y\|_{L^2}^2 + \alpha\|u\|_{L^2}^2$$

for all $(y, u) \in Y \times L^2(\Omega)$. The point $(\hat{y}_{\epsilon_n}, \hat{u}_{\epsilon_n})$ of Proposition 6.1 is chosen as

$$\epsilon_n \hat{u}_{\epsilon_n} + \hat{y}_{\epsilon_n} =: y^* \quad \text{and} \quad A\hat{y}_{\epsilon_n} = \hat{u}_{\epsilon_n}.$$

Then $J(y_{\epsilon_n}, u_{\epsilon_n}) \le J(\hat{y}_{\epsilon_n}, \hat{u}_{\epsilon_n}) \le C$ implies $\max\{\|y_{\epsilon_n}\|_{L^2}, \|u_{\epsilon_n}\|_{L^2}\} \le C$. The uniform bound on $\{y_{\epsilon_n}\}$ in Y then follows from elliptic regularity theory. Further, we have $y_{\epsilon_n} - b \le -\epsilon_n u_{\epsilon_n}$ a.e. in Ω by feasibility. Hence, we infer that

$$\big\|\max(0, y_{\epsilon_n} - b)\big\|_{L^2} = \mathcal{O}(\epsilon_n).$$

Now, assumption (6.8) yields

$$J\left(y^*, u^*\right) - J\left(y_{\epsilon_n}, u_{\epsilon_n}\right) \le \left(\lambda^*|_{\mathscr{A}^*}, \left(y_{\epsilon_n} - b\right)\big|_{\mathscr{A}^*}\right)_{L^2} \le \mathcal{O}\left(\epsilon_n\right). \tag{6.9}$$

Subsequently we use $w_{\epsilon_n} = (y_{\epsilon_n}, u_{\epsilon_n})$, $\hat{w}_{\epsilon_n} = (\hat{y}_{\epsilon_n}, \hat{u}_{\epsilon_n})$ and $w^* = (y^*, u^*)$. From a second-order Taylor expansion of $\mathscr{L}_{\epsilon_n}$ we obtain

$$\begin{aligned}
&\frac{1}{2}\big\langle \nabla_{ww} \mathscr{L}_{\epsilon_n}\big(y_{\epsilon_n}, u_{\epsilon_n}, p_{\epsilon_n}, \lambda_{\epsilon_n}\big)\big(\hat{w}_{\epsilon_n} - w_{\epsilon_n}\big), \big(\hat{w}_{\epsilon_n} - w_{\epsilon_n}\big)\big\rangle \\
&\quad = J\big(\hat{w}_{\epsilon_n}\big) - J(w^*) + J(w^*) - J\big(w_{\epsilon_n}\big) + \big(\lambda_{\epsilon_n}, \epsilon_n \hat{u}_{\epsilon_n} + \hat{y}_{\epsilon_n} - b\big)_{L^2} \\
&\quad \le J\big(\hat{w}_{\epsilon_n}\big) - J(w^*) + J(w^*) - J\big(w_{\epsilon_n}\big),
\end{aligned} \tag{6.10}$$

where we also used $\lambda_{\epsilon_n} \ge 0$ and $\epsilon_n \hat{u}_{\epsilon_n} + \hat{y}_{\epsilon_n} - b \le 0$ by feasibility.

Next observe that due to $\|\hat{u}_{\epsilon_n}\|_{L^2} \le \|u^*\|_{L^2}^2$ by Proposition 6.1 we get

$$\begin{aligned}
J(\hat{w}_{\epsilon_n}) - J(w^*) &= \frac{1}{2}\left(\big\|\hat{y}_{\epsilon_n} - y_d\big\|_{L^2}^2 - \|y^* - y_d\|_{L^2}^2\right) + \frac{\alpha}{2}\left(\big\|\hat{u}_{\epsilon_n}\big\|_{L^2}^2 - \|u^*\|_{L^2}^2\right) \\
&\le \frac{1}{2}\left((y^* - y_d, \hat{y}_{\epsilon_n} - y^*)_{L^2} + \frac{1}{2}\big\|\hat{y}_{\epsilon_n} - y^*\big\|_{L^2}^2\right) \\
&\le C\left(1 + \big\|\hat{y}_{\epsilon_n} - y^*\big\|_{L^2}\right)\big\|\hat{y}_{\epsilon_n} - y^*\big\|_{L^2} = \mathcal{O}(\epsilon_n).
\end{aligned} \tag{6.11}$$

Now, (6.9)–(6.11) yield

$$\left\|\hat{y}_{\epsilon_n} - y_{\epsilon_n}\right\|_{L^2}^2 + \alpha \left\|\hat{u}_{\epsilon_n} - u_{\epsilon_n}\right\|_{L^2}^2 \leq \mathcal{O}(\epsilon_n). \tag{6.12}$$

Finally observe that due to Proposition 6.1 and (6.12) we have

$$\left\|y_{\epsilon_n} - y^*\right\|_{L^2} \leq \left\|y_{\epsilon_n} - \hat{y}_{\epsilon_n}\right\|_{L^2} + \left\|\hat{y}_{\epsilon_n} - y^*\right\|_{L^2} \leq \mathcal{O}(\sqrt{\epsilon_n})$$

for $\epsilon_n \leq 1$. An analogous assertion is true for $\|u_{\epsilon_n} - u^*\|_{L^2}$, which proves the claim. □

We end this section by showing that y_ϵ and u_ϵ are Hölder continuous with exponent 1/2 with respect to $\epsilon > 0$. For this purpose we recall the first-order optimality system of (6.1):

$$Ay_\epsilon - u_\epsilon = 0, \tag{6.13}$$

$$y_\epsilon - y_d + A^* p_\epsilon + \lambda_\epsilon = 0, \tag{6.14}$$

$$\alpha u_\epsilon - p_\epsilon + \epsilon\lambda_\epsilon = 0, \tag{6.15}$$

$$\epsilon u_\epsilon + y_\epsilon \leq b, \quad \lambda_\epsilon \geq 0, \quad \lambda_\epsilon(\epsilon u_\epsilon + y_\epsilon - b) = 0. \tag{6.16}$$

We start by proving an auxiliary result.

LEMMA 6.3. *Let* $\epsilon_1 > 0$ *and* $\epsilon_2 > 0$. *Then we have*

$$\left(u_{\epsilon_1} - u_{\epsilon_2}, \epsilon_1(\lambda_{\epsilon_1} - \lambda_{\epsilon_2})\right)_{L^2} + (\lambda_{\epsilon_1} - \lambda_{\epsilon_2}, y_{\epsilon_1} - y_{\epsilon_2})_{L^2} \geq (\epsilon_1 - \epsilon_2)(\lambda_{\epsilon_2} - \lambda_{\epsilon_1}, u_{\epsilon_2})_{L^2}.$$

PROOF. Using (6.16) we find that

$$\begin{aligned}
&\left(u_{\epsilon_1} - u_{\epsilon_2}, \epsilon_1(\lambda_{\epsilon_1} - \lambda_{\epsilon_2})\right)_{L^2} + (\lambda_{\epsilon_1} - \lambda_{\epsilon_2}, y_{\epsilon_1} - y_{\epsilon_2})_{L^2} \\
&\quad = \left(u_{\epsilon_1} - u_{\epsilon_2}, \epsilon_1(\lambda_{\epsilon_1} - \lambda_{\epsilon_2})\right)_{L^2} + (\lambda_{\epsilon_1}, b - \epsilon_1 u_{\epsilon_1})_{L^2} - (\lambda_{\epsilon_2}, y_{\epsilon_1})_{L^2} - (\lambda_{\epsilon_1}, y_{\epsilon_2})_{L^2} \\
&\qquad + (\lambda_{\epsilon_2}, b - \epsilon_2 u_{\epsilon_2})_{L^2} \\
&\quad = -(\epsilon_1 u_{\epsilon_2}, \lambda_{\epsilon_1})_{L^2} - (\epsilon_1 u_{\epsilon_1}, \lambda_{\epsilon_2})_{L^2} + (\lambda_{\epsilon_1}, b)_{L^2} - (\lambda_{\epsilon_2}, y_{\epsilon_1})_{L^2} - (\lambda_{\epsilon_1}, y_{\epsilon_2})_{L^2} \\
&\qquad + (\lambda_{\epsilon_2}, b)_{L^2} + (\epsilon_1 - \epsilon_2)(\lambda_{\epsilon_2}, u_{\epsilon_2})_{L^2} \\
&\quad = (\lambda_{\epsilon_1}, b - y_{\epsilon_2} - \epsilon_1 u_{\epsilon_2})_{L^2} + (\lambda_{\epsilon_2}, b - y_{\epsilon_1} - \epsilon_1 u_{\epsilon_1})_{L^2} + (\epsilon_1 - \epsilon_2)(\lambda_{\epsilon_2}, u_{\epsilon_2})_{L^2} \\
&\quad \geq \left(\lambda_{\epsilon_1}, (\epsilon_2 - \epsilon_1)u_{\epsilon_2}\right)_{L^2} + (\epsilon_1 - \epsilon_2)(\lambda_{\epsilon_2}, u_{\epsilon_2})_{L^2} \\
&\quad \geq (\epsilon_1 - \epsilon_2)(\lambda_{\epsilon_2} - \lambda_{\epsilon_1}, u_{\epsilon_2})_{L^2}
\end{aligned}$$

which proves the claim. □

THEOREM 6.4. *Let* $\min(\epsilon_1, \epsilon_2) \geq \underline{\epsilon} > 0$. *Then there exists a constant* $C > 0$ *such that*

$$\max\left\{\left\|y_{\epsilon_1} - y_{\epsilon_2}\right\|_{L^2}, \left\|u_{\epsilon_1} - u_{\epsilon_2}\right\|_{L^2}\right\} \leq C\sqrt{|\epsilon_1 - \epsilon_2|}.$$

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

PROOF. With $\epsilon = \epsilon_1$ and $\epsilon = \epsilon_2$ in (6.14), subtraction yields

$$y_{\epsilon_1} - y_{\epsilon_2} + A^*(p_{\epsilon_1} - p_{\epsilon_2}) + \lambda_{\epsilon_1} - \lambda_{\epsilon_2} = 0.$$

From this we obtain

$$\left\| y_{\epsilon_1} - y_{\epsilon_2} \right\|_{L^2}^2 + a(p_{\epsilon_1} - p_{\epsilon_2}, y_{\epsilon_1} - y_{\epsilon_2}) + (\lambda_{\epsilon_1} - \lambda_{\epsilon_2}, y_{\epsilon_1} - y_{\epsilon_2})_{L^2} = 0.$$

The state equation yields $a(y_{\epsilon_1} - y_{\epsilon_2}, p_{\epsilon_1} - p_{\epsilon_2}) = (u_{\epsilon_1} - u_{\epsilon_2}, p_{\epsilon_1} - p_{\epsilon_2})$. Hence we have

$$\left\| y_{\epsilon_1} - y_{\epsilon_2} \right\|_{L^2}^2 + (u_{\epsilon_1} - u_{\epsilon_2}, p_{\epsilon_1} - p_{\epsilon_2}) + (\lambda_{\epsilon_1} - \lambda_{\epsilon_2}, y_{\epsilon_1} - y_{\epsilon_2})_{L^2} = 0. \tag{6.17}$$

Using (6.15) in (6.17) we get

$$\begin{aligned}
0 &= \left\| y_{\epsilon_1} - y_{\epsilon_2} \right\|_{L^2}^2 + \alpha \left\| u_{\epsilon_1} - u_{\epsilon_2} \right\|_{L^2}^2 + (u_{\epsilon_1} - u_{\epsilon_2}, \epsilon_1 \lambda_{\epsilon_1} - \epsilon_2 \lambda_{\epsilon_2})_{L^2} \\
&\quad + (\lambda_{\epsilon_1} - \lambda_{\epsilon_2}, y_{\epsilon_1} - y_{\epsilon_2})_{L^2} \\
&= \left\| y_{\epsilon_1} - y_{\epsilon_2} \right\|_{L^2}^2 + \alpha \left\| u_{\epsilon_1} - u_{\epsilon_2} \right\|_{L^2}^2 + (u_{\epsilon_1} - u_{\epsilon_2}, (\epsilon_1 - \epsilon_2) \lambda_{\epsilon_2})_{L^2} \\
&\quad + (u_{\epsilon_1} - u_{\epsilon_2}, \epsilon_1 (\lambda_{\epsilon_1} - \lambda_{\epsilon_2}))_{L^2} + (\lambda_{\epsilon_1} - \lambda_{\epsilon_2}, y_{\epsilon_1} - y_{\epsilon_2})_{L^2} \\
&\geq \left\| y_{\epsilon_1} - y_{\epsilon_2} \right\|_{L^2}^2 + \alpha \left\| u_{\epsilon_1} - u_{\epsilon_2} \right\|_{L^2}^2 + (u_{\epsilon_1} - u_{\epsilon_2}, (\epsilon_1 - \epsilon_2) \lambda_{\epsilon_2})_{L^2} \\
&\quad + (\epsilon_1 - \epsilon_2) (\lambda_{\epsilon_2} - \lambda_{\epsilon_1}, u_{\epsilon_2})_{L^2}
\end{aligned} \tag{6.18}$$

where we used Lemma 6.3 for the last estimate. From (6.18) we infer that

$$\left\| y_{\epsilon_1} - y_{\epsilon_2} \right\|_{L^2}^2 + \alpha \left\| u_{\epsilon_1} - u_{\epsilon_2} \right\|_{L^2}^2 \leq |\epsilon_1 - \epsilon_2| \left| (u_{\epsilon_1}, \lambda_{\epsilon_2})_{L^2} - (u_{\epsilon_2}, \lambda_{\epsilon_1})_{L^2} \right|.$$

The boundedness of $u_{\epsilon_1}, u_{\epsilon_2}, \lambda_{\epsilon_1}, \lambda_{\epsilon_2}$ in $L^2(\Omega)$ for $\min(\epsilon_1, \epsilon_2) \geq \underline{\epsilon} > 0$ yields the existence of $C = C(\underline{\epsilon}) > 0$ such that

$$\left\| y_{\epsilon_1} - y_{\epsilon_2} \right\|_{L^2}^2 + \alpha \left\| u_{\epsilon_1} - u_{\epsilon_2} \right\|_{L^2}^2 \leq C |\epsilon_1 - \epsilon_2|$$

which ends the proof. □

7. Conclusions

In this paper we prove the locally superlinear convergence of a primal-dual active-set, or equivalently semismooth Newton, method in function space. We further establish a mesh-independence result proving the numerical stability of the fast local convergence of our algorithm under mesh refinements. This latter behaviour was observed in numerical practice before. In our report on numerical results we validate the mesh-independence theory and study the numerical behaviour of our algorithm in the case of primal as well as dual degeneracy. We also provide a comparison with a short-step path-following interior-point method. This latter comparison is of interest since the short-step version of interior-point methods is currently the only available path-following method with a function space convergence analysis. In our tests we find that our method is superior to the short-step path-following algorithm. Another advantage of our primal-dual active-set method when compared to the interior-point technique is related to its warm-start ability. In fact, our numerical results show that our algorithm benefits significantly from a coarse-to-fine mesh refinement. For interior-point methods, on the other hand, it was observed in [5] and [13] that such a warm-start property is much harder (if possible at all) to achieve.

The warm-start ability (respectively the speed-up) under coarse-to-fine mesh-refinements of our primal-dual active-set method is also of interest in the case of a vanishing Lavrentiev parameter. In this case (1.1) is used as a (regularizing) device for solving the state-constrained problem (1.2). From our numerical findings and the convergence results with respect to the Lavrentiev parameter $\epsilon > 0$ a combined tuning of the mesh size of the underlying discretization and the Lavrentiev parameter appears to be appealing and is the subject of future research.

Acknowledgements

This research was supported by the Austrian Science Fund FWF under START-program Y305-N12 "Interfaces and Free Boundaries" and by the Institute of Mathematics and its Applications (IMA) at the University of Minneapolis, Twin Cities, Minnesota.

References

[1] R. A. Adams, *Sobolev spaces* (Academic Press, New York-London, 1975).

[2] E. L. Allgower, K. Böhmer, F. A. Potra and W. C. Rheinboldt, "A mesh-independence principle for operator equations and their discretizations", *SIAM J. Numer. Anal.* **23** (1986) 160–169.

[3] W. Alt, *Discretization and mesh-independence of Newton's method for generalized equations*, Volume 195 of *Lecture Notes in Pure and Appl. Math.* (Dekker, New York, 1998) 1–30.

[4] N. Arada, E. Casas and F. Tröltzsch, "Error estimates for the numerical approximation of a semilinear elliptic control problem", *Comput. Optim. Appl* **23** (2002) 201–229.

[5] R. E. Bank, P. E. Gill and R. F. Marcia, *Interior methods for a class of elliptic variational inequalities*, Volume 30 of *Lect. Notes Comput. Sci. Eng.* (Springer, Berlin, 2003) 218–235.

[6] 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) 495–521.

[7] M. Bergounioux and K. Kunisch, "Primal-dual strategy for state-constrained optimal control problems", *Comput. Optim. Appl.* **22** (2002) 193–224.

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press

[8] M. Bergounioux and K. Kunisch, "On the structure of Lagrange multipliers for state-constrained optimal control problems", *Systems Control Lett.* **48** (2003) 169–176.

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

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

[11] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Volume 224 of *Grundlehren der mathematischen Wissenschaften* (Springer Verlag, Berlin, 1977).

[12] M. Hintermüller, K. Ito and K. Kunisch, "The primal-dual active set strategy as a semismooth Newton method", *SIAM J. Optim.* **13** (2003) 865–888.

[13] M. Hintermüller and K. Kunisch, "Feasible and non-interior path-following in constrained minimization with low multiplier regularity", *SIAM J. Control Optim.* **45** (2006) 1198–1221.

[14] M. Hintermüller and M. Ulbrich, "A mesh-independence result for semismooth Newton methods", *Math. Program.* **101** (2004) 151–184.

[15] B. Kummer, "Generalized Newton and NCP methods: convergence, regularity, actions", *Discuss. Math. Differ. Incl.* **20** (2000) 209–244.

[16] D. G. Luenberger, *Optimization by vector space methods* (John Wiley & Sons Inc., New York, 1969).

[17] C. Meyer, U. Prüfert and F. Tröltzsch, "On two numerical methods for state-constrained elliptic control problems", Technical Report 5-2005, Department of Mathematics, TU Berlin, 2005.

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

[19] U. Prüfert, F. Tröltzsch and M. Weiser, "The convergence of an interior point method for an elliptic control problem with mixed control-state constraints", Technical report, TU Berlin, 2004, Preprint 36-2004.

[20] L. Qi and J. Sun, "A nonsmooth version of Newton's method", *Math. Programming* **58 (3, Ser. A)** (1993) 353–367.

[21] S. M. Robinson, "Generalized equations and their solutions. I. Basic theory. Point-to-set maps and mathematical programming", *Math. Programming Stud.* **10** (1979) 128–141.

[22] S. M. Robinson, "Strongly regular generalized equations", *Math. Oper. Res.* **5** (1980) 43–62.

[23] S. M. Robinson, "Generalized equations and their solutions. II. Applications to nonlinear programming. Optimality and stability in mathematical programming.", *Math. Programming Stud.* **19** (1982) 200–221.

[24] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen* (Vieweg, Wiesbaden, Germany, 2005).

[25] S. J. Wright, *Primal-dual Interior-Point Methods* (Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997).

https://doi.org/10.1017/S1446181100012657 Published online by Cambridge University Press