Semismooth Newton Methods and Applications[1]

by

Michael Hintermüller

Department of Mathematics
Humboldt-University of Berlin

hint@math.hu-berlin.de

[1] Oberwolfach-Seminar on "Mathematics of PDE-Constrained Optimization" at Mathematisches Forschungsinstitut in Oberwolfach, November 2010.

CHAPTER 1

Newton's method for smooth systems

In many practical applications one has to find a zero of a system of nonlinear equations:

$$\text{Given } F:\mathbb{R}^n\to\mathbb{R}^n,\quad \text{find } x_*\in\mathbb{R}^n \text{ such that } F(x_*)=0. \tag{1}$$

Throughout this first section we assume that F is (at least) once continuously differentiable. This is our requirement for calling $F(x)=0$ a *smooth system*. Our aim is to derive Newton's method for solving (1), discuss its requirements and local as well as some aspects of its global convergence characteristics. Further, this section also serves the purpose of highlighting characteristics of Newton's method and of pointing to assumptions and techniques which cannot be used in the case where F is not differentiable in the classical sense. The latter situation, however, is the one we will be most interested in later on.

The traditional approach for defining Newton's method for solving (1) is based on replacing the complicated nonlinear map F by its linearization about a given estimate $x_k\in\mathbb{R}^n$ of x_*:

$$F(x_k+d)\approx F(x_k)+\nabla F(x_k)d=:m_k(x_k+d),\quad d\in\mathbb{R}^n, \tag{2}$$

where $\nabla F:\mathbb{R}^n\to\mathbb{R}^{n\times n}$ denotes the Jacobi-matrix of F. Now we replace the complicated problem (1) by solving a sequence of simpler problems. In fact, with the aim of improving the current estimate x_k one computes $d_k\in\mathbb{R}^n$ such that

$$m_k(x_k+d_k)=0. \tag{3}$$

Obviously, this step is only well-defined if $\nabla F(x_k)$ is non-singular. Our improved estimate is set to be

$$x_{k+1}:=x_k+d_k. \tag{4}$$

This allows to define our basic Newton algorithm.

ALGORITHM 1 (Newton's method for smooth systems.).
Given $F:\mathbb{R}^n\to\mathbb{R}^n$ continuously differentiable and $x_0\in\mathbb{R}^n$, $k:=0$:

(1) *Unless a stopping rule is satisfied, solve (for d_k)*

$$\nabla F(x_k)d_k=-F(x_k).$$

(2) *Set $x_{k+1}:=x_k+d_k$, $k:=k+1$, and go to (1).*

For Algorithm 1 to be well-defined we have to guarantee that, for each k, $\nabla F(x_k)$ is invertible. This is addressed in Lemma 1.1 below. Concerning an appropriate stopping rule for numerical realizations we only mention that a criterion like

$$\begin{aligned} \|F(x_k)\| &\leq \epsilon^F_{\text{rel}}\|F(x_0)\| + \epsilon^F_{\text{abs}}, \\ \|d_k\| &\leq \epsilon^d_{\text{rel}}\|x_0\| + \epsilon^d_{\text{abs}}, \end{aligned}$$

with sufficiently small positive tolerances ϵ^z_{rel}, ϵ^z_{abs}, $z \in \{d, F\}$, is suitable.

Already now we emphasize that it will turn out that, unless x_k is already sufficiently close to x_*, we might not be allowed to take the full step along d_k. Rather we have to reduce the *step size* for obtaining a convergent method. In this case (4) is replaced by

$$x_{k+1} = x_k + \alpha_k d_k, \tag{5}$$

where $\alpha_k \in (0, 1]$ is a suitable chosen step size. We will specify appropriate selection criteria later.

1. Local convergence of Newton's method

In this section we are interested in studying local convergence properties of Newton's method (Algorithm 1). Here, the term *local* refers to the fact that the results only hold true if x_0 is chosen sufficiently close to x_*. The results discussed here may be already well-known to the reader. Thus, let us point out that we are merely interested in the technique of proof for later reference.

Let $B(x, r) = \{y \in \mathbb{R}^n : \|y-x\| < r\}$ denote the open ball of radius $r > 0$ about x. We start by addressing the non-singularity issue in connection with ∇F. We write $\nabla F(x)^{-1}$ for $(\nabla F(x))^{-1}$.

Lemma 1.1. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in the open set $D \subset \mathbb{R}^n$ with ∇F Lipschitz continuous in D (with constant $L > 0$). Let $z \in D$ be fixed, and assume that $\nabla F(z)^{-1}$ exists. Further assume that there exists $\beta > 0$ such that $\|\nabla F(z)^{-1}\| \leq \beta$. Then for all $x \in B(z, \eta)$, with $0 < \eta < \frac{c}{L\beta}$ and $0 < c < 1$ fixed, $\nabla F(x)$ is nonsingular and satisfies*

$$\|\nabla F(x)^{-1}\| \leq \frac{\beta}{1-c}.$$

The proof makes use of (91) of Theorem A.1.

Remark 1.1. *If we require ∇F to be only Hölder continuous with exponent γ (instead of Lipschitz), with $0 < \gamma < 1$ and $L > 0$ still denoting the constant, then the assertion of Lemma 1.1 holds true for $x \in B(z, \eta)$ with $0 < \eta < (\frac{c}{L\beta})^{1/\gamma}$ with $0 < c < 1$ fixed.*

Now we can prove the local convergence result for Algorithm 1.

THEOREM 1.1. *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in an open convex set $D \subset \mathbb{R}^n$. Assume that there exist $x_* \in \mathbb{R}^n$ and $r, \beta > 0$ such that $B(x_*, r) \subset D$, $F(x_*) = 0$, $\nabla F(x_*)^{-1}$ exists with $\|\nabla F(x_*)^{-1}\| \leq \beta$, and ∇F Lipschitz continuous with constant L on $B(x_*, r)$. Then there exists $\epsilon > 0$ such that for all $x_0 \in B(x_*, \epsilon)$ the sequence $\{x_k\}$ generated by Algorithm 1 is well-defined, converges to x_*, and satisfies*

$$\|x_{k+1} - x_*\| \leq \frac{L\beta}{2(1-c)}\|x_k - x_*\|^2, \quad k = 0, 1, 2, \ldots, \tag{6}$$

for some fixed $0 < c \leq \frac{2}{3}$.

Proof. The proof is by induction. For $0 < c \leq \frac{2}{3}$ fixed, let

$$\epsilon = \min\left\{r, \frac{c}{L\beta}\right\}.$$

Then Lemma 1.1 with $z := x_*$ and $x := x_0$ yields that $\nabla F(x_0)$ is nonsingular and satisfies

$$\|\nabla F(x_0)^{-1}\| \leq \frac{\beta}{1-c}.$$

Therefore, x_1 is well-defined and satisfies

$$\begin{aligned} x_1 - x_* &= x_0 - x_* - \nabla F(x_0)^{-1} F(x_0) \\ &= x_0 - x_* - \nabla F(x_0)^{-1}(F(x_0) - F(x_*)) \\ &= \nabla F(x_0)^{-1}\left(F(x_*) - F(x_0) - \nabla F(x_0)(x_* - x_0)\right). \end{aligned}$$

Application of Theorem A.2 yields

$$\begin{aligned} \|x_1 - x_*\| &\leq \|\nabla F(x_0)^{-1}\| \, \|F(x_*) - F(x_0) - \nabla F(x_0)(x_* - x_0)\| \\ &\leq \frac{L\beta}{2(1-c)}\|x_0 - x_*\|^2 \end{aligned}$$

which proves (6) for $k = 0$. Since $\|x_0 - x_*\| \leq \frac{c}{L\beta}$ we have

$$\|x_1 - x_*\| \leq \frac{c}{2(1-c)}\|x_0 - x_*\| \leq \|x_0 - x_*\| \leq \epsilon.$$

The proof of the induction step proceeds identically. □

Again, we may reduce the Lipschitz continuity of ∇F to Hölder continuity.

REMARK 1.2. *If ∇F is assumed to be only Hölder continuous with exponent γ, with $0 < \gamma < 1$ and L still denoting the constant, then we obtain the estimate*

$$\|x_{k+1} - x_*\| \leq \frac{L\beta}{2(1-c)}\|x_k - x_*\|^{1+\gamma}, \quad k = 0, 1, 2, \ldots. \tag{7}$$

This essentially means that $\{x_k\}$ approaches x_* at a slower rate as in the case of ∇F being Lipschitz.

2. Convergence rates

The remark concerning the convergence speed is made more precise in this section.

Definition 1.1.

(a) *Let* $\{x_k\} \subset \mathbb{R}^n$ *denote a sequence with limit* $x_* \in \mathbb{R}^n$, *and let* $p \in [1, +\infty)$. *Then*

$$Q_p\{x_k\} := \begin{cases} \limsup_{k\to\infty} \frac{\|x_{k+1}-x_*\|}{\|x_k-x_*\|^p}, & \text{if } x_k \neq x_* \; \forall k \geq k_0, \\ 0, & \text{if } x_k = x_* \; \forall k \geq k_0, \\ +\infty, & \text{else}, \end{cases}$$

for some $k_0 \in \mathbb{N}$, *is called the quotient factor (Q-factor) of* $\{x_k\}$.

(b) *The quantity*

$$O_Q\{x_k\} := \inf\{p \in [1, +\infty) : Q_p\{x_k\} = +\infty\}$$

is called the Q-order of $\{x_k\}$.

We now collect several important properties.

Remark 1.3.

(1) *The Q-factor depends on the used norms, the Q-order does not!*

(2) *There always exists a value* $p_0 \in [1, +\infty)$ *such that*

$$Q_p\{x_k\} = \begin{cases} 0 & \text{for } p \in [1, p_0), \\ +\infty & \text{for } p \in (p_0, +\infty). \end{cases}$$

(3) *The Q-orders 1 and 2 are of special interest. We call*

$Q_1\{x_k\} = 0$	*Q-superlinear convergence*
$0 < Q_1\{x_k\} < 1$	*Q-linear convergence*
$Q_2\{x_k\} = 0$	*Q-superquadratic convergence*
$0 < Q_2\{x_k\} < +\infty$	*Q-quadratic convergence*

With this definition we see that Theorem 1.1 proves Newton's method, *i.e.* Algorithm 1, to converge locally at a Q-quadratic rate. In the case where ∇F is assumed to be only Hölder continuous with exponent γ, Newton's method converges locally at a superlinear rate. The Q-order is $1 + \gamma$.

For checking the Q-convergence of a sequence, criteria like

$$\|x_k - x_*\| \leq \epsilon \tag{8}$$

are unrealistic since they require knowledge of the solution x_*. The following result allows to replace (8) by the practicable criterion

$$\|x_{k+1} - x_k\| \leq \epsilon. \tag{9}$$

THEOREM 1.2. *An arbitrary sequence* $\{x_k\} \subset \mathbb{R}^n$ *with* $\lim_k x_k = x_*$ *satisfies*

$$\left|1 - \frac{\|x_{k+1} - x_k\|}{\|x_k - x_*\|}\right| \le \frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} \quad \text{if } x_k \neq x_*.$$

If $\{x_k\}$ *converges Q-superlinearly to* x_* *and* $x_k \neq x_*$ *for* $k \ge k_0$*, then*

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x_k\|}{\|x_k - x_*\|} = 1.$$

Besides the Q-convergence, the R-convergence is of interest.

DEFINITION 1.2.

(a) *Let* $\{x_k\} \subset \mathbb{R}^n$ *denote a sequence with limit* $x_* \in \mathbb{R}^n$*, and let* $p \in [1, +\infty)$*. Then*

$$R_p\{x_k\} := \begin{cases} \limsup_{k\to\infty} \|x_k - x_*\|^{1/k}, & \text{if } p = 1, \\ \limsup_{k\to\infty} \|x_k - x_*\|^{1/p^k}, & \text{if } p > 1 \end{cases}$$

is called root (convergence) factor (R-factor) of $\{x_k\}$*.*

(b) *The quantity*

$$O_R\{x_k\} := \inf\{p \in [1, +\infty) : R_p\{x_k\} = 1\}$$

is called the R-order of $\{x_k\}$*.*

REMARK 1.4.

(1) *In contrast to the Q-factor, the R-factor is independent of the norm.*

(2) *There always exists a value* $p_0 \in [1, +\infty)$ *such that*

$$R_p\{x_k\} = \begin{cases} 0 \text{ for } p \in [1, p_0), \\ 1 \text{ for } p \in (p_0, +\infty). \end{cases}$$

(3) *The Q and R quantities are related as follows:*

$$O_Q\{x_k\} \le O_R\{x_k\} \quad \text{and} \quad R_1\{x_k\} \le Q_1\{x_k\}.$$

Very often it is convenient to use the *Landau symbols* $\mathcal{o}$ and $\mathcal{O}$ for describing the convergence behavior of a sequence.

DEFINITION 1.3. *Let* $f, g : \mathbb{R}^n \to \mathbb{R}^m$ *and* $x_* \in \mathbb{R}^n$ *be given. We write*

(a) $f(x) = \mathcal{O}(g(x))$ *for* $x \to x_*$ *iff there exists a uniform constant* $\lambda > 0$ *and a neighborhood* U *of* x_* *such that for all* $x \in U \setminus \{x_*\}$ *it holds that*

$$\|f(x)\| \le \lambda \|g(x)\|.$$

(b) $f(x) = \mathcal{o}(g(x))$ *for* $x \to x_*$ *iff for all* $\epsilon > 0$ *there exists a neighborhood* U *of* x_* *such that for all* $x \in U \setminus \{x_*\}$ *it holds that*

$$\|f(x)\| \le \epsilon \|g(x)\|.$$

REMARK 1.5. *For* $\lim_k x_k = x_*$*, the sequence* $\{x^k\}$ *converges to* x_* *(at least)*

(1) *Q-superlinearly if* $\|x_{k+1} - x_*\| = \mathcal{o}(\|x_k - x_*\|)$*;*

(2) *Q-quadratically if* $\|x_{k+1} - x_*\| = \mathcal{O}(\|x_k - x_*\|^2)$*.*

The R-convergence plays a role in the important Newton-Kantorovich result.

Theorem 1.3 (Kantorovich). *Let $r > 0$, $x_0 \in \mathbb{R}^n$, $F : \mathbb{R}^n \to \mathbb{R}^n$, and assume that F is continuously differentiable in $B(x_0, r)$. Assume for a vector norm and the induced operator (matrix) norm that ∇F is Lipschitz continuous on $B(x_0, r)$ with constant L and with $\nabla F(x_0)$ nonsingular, and that there exist constants β, η such that*

$$\|\nabla F(x_0)^{-1}\| \leq \beta, \quad \|\nabla F(x_0)^{-1} F(x_0)\| \leq \eta.$$

Define $\gamma_r = \beta L$, $\alpha = \gamma_r \eta$. If $\alpha \leq \frac{1}{2}$ and $r \geq r_0 = (1 - \sqrt{1-2\alpha})/\gamma_r$, then the sequence $\{x_k\}$ produced by Algorithm 1 is well-defined and converges to x_*, a unique zero of F in the closure of $B(x_0, r_0)$. If $\alpha < \frac{1}{2}$, then x_* is the unique zero of F in $B(x_0, r_1)$, where $r_1 = \min\{r, (1 + \sqrt{1-2\alpha})/\gamma_r\}$ and

$$\|x_k - x_*\| \leq (2\alpha)^{2^k} \frac{\eta}{\alpha}, \quad k = 0, 1, \ldots \tag{10}$$

Notice that from (10) we conclude that the sequence $\{x_k\}$ converges R-quadratically to x_*.

3. Global convergence

As announced earlier, unless one is able to find a good initial guess x_0, one needs a suitable globalization strategy for Newton's method to convergence. Here global convergence means that we are free to choose x_0. Since our overall emphasis is on local convergence, we present this material only for completeness. Many extensions are possible.

In order to obtain a convergent method, rather than accepting the full Newton step, *i.e.*, setting

$$x_{k+1} = x_k + d_k,$$

where d_k solves

$$\nabla F(x_k) d_k = -F(x_k),$$

one has to determine a step size (damping parameter) $\alpha_k \in (0, 1]$ and set

$$x_{k+1} = x_k + \alpha_k d_k.$$

There are many different techniques for picking α_k. We highlight one of them: Given d_k, let α_k denote the first element of the sequence $\{\omega^l\}_{l=0}^{\infty}$, $\omega \in (0, 1)$ fixed, such that

$$\|F(x_k + \omega^l d_k)\| \leq (1 - \nu\omega^l)\|F(x_k)\|, \tag{11}$$

where $\nu \in (0, 1)$ denotes a fixed parameter. The condition (11) is called a *sufficient decrease condition.* The particular realization considered here is known as *Armijo step size rule.* A condition such as

$$\|F(x_k + \alpha_k d_k)\| < \|F(x_k)\|$$

instead of (11) may cause $\{x_k\}$ to stagnate at a non-zero $\bar{x}$, and it is, thus, not suitable for proving convergence.

4. Newton's method and unconstrained minimization

In many applications the nonlinear mapping F is the gradient mapping of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$. Then, typically the aim is to find x_* such that f is (locally) minimized. Here x_* is called a (strict) local minimizer of f if there exists $r > 0$ such that $f(x) \geq f(x_*)$ $(f(x) > f(x_*))$ for all $x \in B(x, r)$. We have

THEOREM 1.4. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable in an open convex set $D \subset \mathbb{R}^n$. Then $x \in D$ can be a local minimizer of f only if $\nabla f(x) = 0$.*

Obviously, also local maximizers, which are defined analogously to local minimizers by just reverting the inequality signs, satisfy $\nabla f(x) = 0$. Therefore we need an additional criterion to decide whether a point is a local minimizer or not.

THEOREM 1.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable in the open convex set $D \subset \mathbb{R}^n$, and assume that there exists $x \in D$ such that $\nabla f(x) = 0$. If $\nabla^2 f(x)$ is positive definite[1], then x is a local minimizer of f. If $\nabla^2 f(x)$ is Lipschitz continuous at x, then x can be a local minimizer of f only if $\nabla^2 f(x)$ is positive semidefinite.*

To achieve the link to the task of finding the zero of a nonlinear system, we set, for f twice continuously differentiable,

$$F(x) = \nabla f(x) \quad \text{and} \quad \nabla F(x) = \nabla^2 f(x).$$

Now let us assume that x_* is a local minimizer of f which satisfies the *second order sufficient condition*, *i.e.*, $\nabla^2 f(x_*)$ is positive definite. Then by continuity we have that there exists $r > 0$ such that $\nabla^2 f(x)$ is positive definite for $x \in B(x_*, r)$ as well. For our local convergence regime of Newton's method we assume that $x_0 \in B(x_*, r)$. Then d_k with $\nabla F(x_k) d_k = -F(x_k)$ satisfies

$$d_k^\top \nabla f(x_k) = d_k^\top F(x_k) = -d_k^\top \nabla F(x_k) d_k = -d_k^\top \nabla^2 f(x_k) d_k < 0, \tag{12}$$

if $d_k \neq 0$. In general, a direction d satisfying

$$d^\top \nabla f(x) < 0$$

is called a *descent direction* of f at x. In our case, it guarantees that there exists $\alpha_k > 0$ such that

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \nu \alpha_k \nabla f(x_k)^\top d_k, \tag{13}$$

with $0 < \nu < 1$ fixed. Now (13) can be used instead of (11) for determining a suitable step size. Condition (13) is also called *Armijo condition.*

[1] A matrix $M \in \mathbb{R}^{n \times n}$ is positive definite iff there exists $\epsilon > 0$ such that $d^\top M d \geq \epsilon \|d\|^2$ for all $d \in \mathbb{R}^n$. For M positive semidefinite we have $\epsilon \geq 0$.

Maintaining (13) when using Newton's method for minimizing a function f guarantees that

$$f(x_*) \leq f(x_0).$$

Observe that the descent property hinges on the positive-definiteness of $\nabla^2 f$. For poor initial choices x_0 this positive definiteness cannot be guaranteed in general. Then, in order to achieve global convergence to a local minimizer, one either has to combine (13) with a possible positive definite modification of the Hessian of f (*e.g.*, quasi-Newton techniques), or one employs a trust region approach. For details see, *e.g.*, [**12**].

CHAPTER 2

Generalized Newton methods. The finite dimensional case

Very often the nonlinear mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ is not necessarily differentiable in the classical (Fréchet) sense. This is fact coins the name "nonsmooth" in connection with analytical (nonsmooth analysis) as well as numerical issues (nonsmooth or generalized Newton's method). Of course, one is still interested in computing x_* such that $F(x_*) = 0$. Also, one would like to use a Newton-type approach due to its favorable convergence properties for smooth F. Hence, the following questions arise naturally:

- How does a generalization of the classical differential look like and what are its properties?
- Is it possible to use this generalized derivative for defining a generalized Newton procedure?
- What about the local convergence speed of this generalized Newton method?

1. Generalized differentials and generalized Newton methods

The construction of a generalized differential of a (in the classical sense) nondifferentiable function goes back to [**11**] and it is based on Rademacher's theorem.

Theorem 2.1 (Rademacher). *Suppose $F : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz continuous. Then F is almost everywhere differentiable.*

Now let $D_F \subset \mathbb{R}^n$ denote the set of points at which F is differentiable. Our aim is now to introduce several objects from nonsmooth analysis which provide generalizations of the classical differentiability concept.

We start by defining the B-subdifferential, the generalized Jacobian, and the C-subdifferential of F. Here "B" stands for "Bouligand", who introduced the concept.

Definition 2.1. *Let $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$, with U open, be locally Lipschitz continuous at $x \in U$.*

(a) *The set*

$$\partial_B F(x) := \{G \in \mathbb{R}^{n\times m} : \exists \{x_k\} \subset D_F \text{ with } x_k \to x, \nabla F(x_k) \to G\}$$

is called B-subdifferential *of F at x.*

(b) *Clarke's* generalized Jacobian *is defined as*

$$\partial F(x) = \mathrm{co}(\partial_B F(x)),$$

where co *denotes the convex hull.*

(c) *The set*

$$\partial_C F(x) = \partial F_1(x) \times \ldots \times \partial F_m(x)$$

is called Qi's C-subdifferential.

In the case $m = 1$, $\partial F(x)$ is called the *generalized gradient.* Next we study properties of the respective generalized derivative.

Theorem 2.2. *Let $U \subset \mathbb{R}^n$ be open and $F : U \to \mathbb{R}^m$ be locally Lipschitz continuous. Then for $x \in U$ there holds:*

(a) *$\partial_B F(x)$ is nonempty and compact.*

(b) *$\partial F(x)$ and $\partial_C F(x)$ are nonempty, compact, and convex.*

(c) *The set-valued mappings $\partial_B F$, ∂F, and $\partial_C F$ are locally bounded and upper semicontinuous*[1]*.*

(d) *The following inclusions hold true:*

$$\partial_B F(x) \subset \partial F(x) \subset \partial_C F(x).$$

(e) *If F is continuously differentiable in a neighborhood of x, then*

$$\partial_C F(x) = \partial F(x) = \partial_B F(x) = \{\nabla F(x)\}.$$

The upper semicontinuity has an important implication which will be used later when generalizing Lemma 1.1.

Remark 2.1. *The upper semicontinuity of $\partial_B F$, ∂F, and $\partial_C F$ implies that for $x_k \to x$, $G_k \in \partial F(x_k)$ and $G_k \to G$, then $G \in \partial F(x)$; analogously for $\partial_B F$ and $\partial_C F$. This fact is also referred to as* closedness *of $\partial_B F$, ∂F, and $\partial_C F$, respectively, at x.*

The generalized Jacobian (gradient) satisfies a mean-value property.

Theorem 2.3. *Let $U \subset \mathbb{R}^n$ be convex and open, and let $F : U \to \mathbb{R}^m$ be locally Lipschitz continuous. Then, for any $x, y \in U$,*

$$F(y) - F(x) \in \mathrm{co}\left(\partial F([x,y])(y-x)\right),$$

where $[x,y]$ represents the line segment joining x and y, and the right hand side denotes the convex hull of all points of the form $G(u)(y-x)$ with $G(u) \in \partial F(u)$ for some point u in $[x,y]$.

In applications typically F is obtained as the composition of mappings. Therefore, for computing the generalized derivative of F we have to study the chain rule in the context of the generalized derivative.

[1] A set-valued mapping Θ is upper semicontinuous if for every $\epsilon > 0$ there exists a $\delta > 0$ such that, for all $y \in B(x,\delta)$,

$$\Theta(y) \subseteq \Theta(x) + B(0,\epsilon).$$

THEOREM 2.4. *Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^l$ be nonempty open sets, let $g : U \to V$ be locally Lipschitz continuous at $x \in U$, and let $h : W \to \mathbb{R}^m$ be locally Lipschitz continuous at $g(x)$. Then, $F = h \circ g$ is locally Lipschitz at x and for all $v \in \mathbb{R}^n$ it holds that*

$$\begin{aligned} \partial F(x)v &\subset \operatorname{co}(\partial h(g(x))\partial g(x)v) \\ &= \operatorname{co}\{G_h G_g v : G_h \in \partial h(g(x)),\ G_g \in \partial g(x)\}. \end{aligned}$$

If, in addition, h is continuously differentiable near $g(x)$, then, for all $v \in \mathbb{R}^n$,

$$\partial F(x)v = \nabla h(g(x))\partial g(x)v.$$

If h is real-valued (i.e., if $m = 1$), then in both chain rules the vector v can be omitted.

Sometimes, one has to deal with the special case where $h(y) = e_i^\top y = y_i$, where e_i denotes the ith unit vector, and $g = F$. Then we have

COROLLARY 2.1. *Let $U \subset \mathbb{R}^n$ be open, and let $F : U \to \mathbb{R}^m$ be locally Lipschitz at $x \in U$. Then*

$$\partial F_i(x) = e_i^\top \partial F(x) = \{G_{i,\bullet} : G_{i,\bullet} \text{ is the } i\text{th row of some } G \in \partial F(x)\}.$$

In the case $m = 1$, Clarke's generalized gradient can be characterized by a generalized directional derivative.

DEFINITION 2.2. *Let $F : U \subset \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz on the open set U. The generalized directional derivative of F at $x \in U$ in direction $d \in \mathbb{R}^n$ is given by*

$$F^\circ(x; d) := \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + td) - F(y)}{t}.$$

Note that due to the lim sup in the above definition of the generalized directional derivative is well-defined and finite (the latter due to the local Lipschitz property of F). It holds that

$$\partial F(x) = \{\xi \in \mathbb{R}^n : \xi^\top d \le F^\circ(x; d)\ \forall d \in \mathbb{R}^n\}. \tag{14}$$

We have the following properties of F°.

PROPOSITION 2.1. *Let $F : U \subset \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz on the open set U. Then the following statements hold true.*

(i) *For every $x \in U$, $F^\circ(x; \cdot)$ is Lipschitz continuous, positively homogeneous, and sublinear.*

(ii) *For every $(x, d) \in U \times \mathbb{R}^n$ we have*

$$F^\circ(x; d) = \max\{\xi^\top d : \xi \in \partial F(x)\}.$$

(iii) *$F^\circ : U \times \mathbb{R}^n \to \mathbb{R}$ is upper semicontinuous.*

Next we introduce the notion of a directional derivative of F.

Definition 2.3. *Let $F : U \subset \mathbb{R}^n \to \mathbb{R}^m$ be locally Lipschitz on the open set U. We call $F'(x,d)$ defined by*

$$F'(x,d) = \lim_{t \downarrow 0} \frac{F(x+td) - F(x)}{t}$$

the directional derivative of F at x in direction d.

For $m = 1$, we have for $(x,d) \in U \times \mathbb{R}^n$

$$F'(x;d) \leq F^\circ(x;d)$$

provided the object on the left hand side exists.

Definition 2.4. *Let $F : U \subset \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz on the open set U. F is said to be C(larke)-regular at $x \in U$ iff*

(i) *f is directionally differentiable at x, and*
(ii) *$g^\circ(x;d) = g'(x;d)$ for all $d \in \mathbb{R}^n$.*

We summarize a few calculus rules in the following theorem. Further calculus rules can be found in [**11**].

Theorem 2.5. *Let $F_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \ldots, m$, be a family of Lipschitz functions at x.*

(i) *For any coefficients a_i we have*

$$\partial \left(\sum_{i=1}^m a_i F_i \right) \subseteq \sum_{i=1}^m a_i \partial F_i(x)$$

with equality holding if all functions are C-regular at x and the $a_i^\prime s$ are non-negative.

(ii) *$\partial(F_1 F_2)(x) \subseteq F_2(x)\partial F_1(x) + F_1(x)\partial F_2(x)$ with equality holding if F_1 and F_2 are C-regular at x and $\min(F_1(x), F_2(x)) \geq 0$.*

(iii) *If $g_2(x) \neq 0$, then*

$$\partial \left(\frac{F_1}{F_2} \right)(x) \subseteq \frac{F_2(x)\partial F_1(x) - F_1(x)\partial F_2(x)}{F_2^2(x)}$$

with equality holding if F_1 and $-F_2$ are C-regular at x, $F_1(x) \geq 0$, and $F_2(x) > 0$.

(iv) *For $F(x) = \max\{F_i(x) : i = 1, \ldots, m\}$ it holds that*

$$\partial F(x) \subseteq \mathrm{co}\{\partial F_i(x) : i \in \mathcal{A}(x)\},$$

where $\mathcal{A}(x) := \{i : F_i(x) = F(x)\}$. Equality holds if all F_i are C-regular at x.

Before we start our discussion of semismoothness of a mapping F, we state two more results from nonsmooth analysis. The first result, Theorem 2.6, represents an implicit function theorem for locally Lipschitz continuous functions. The second result, Theorem 2.7, is a generalization of the inverse function theorem. For the implicit function theorem we introduce the following projection: For a function $F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ of two variables

$x \in \mathbb{R}^n$ and $y \in \mathbb{R}^p$, we denote by $\Pi_x \partial F(x,y)$ the set of $n \times n$ matrices G for which there exists a $n \times p$ matrix H such that $[G\,H]$ belongs to $\partial F(x,y)$.

THEOREM 2.6. *Let $F : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ be Lipschitz continuous in a neighborhood of a point $(\bar{x}, \bar{y}) \in \mathbb{R}^n \times \mathbb{R}^p$ for which $F(\bar{x}, \bar{y}) = 0$. Assume that all matrices in $\Pi_x \partial F(\bar{x}, \bar{y})$ are nonsingular. Then there exist open neighborhoods $V_{\bar{x}}$ and $V_{\bar{y}}$ of $\bar{x}$ and $\bar{y}$, respectively, such that, for every $y \in V_{\bar{y}}$, the equation $F(x,y) = 0$ has a unique solution $x \equiv f(y) \in V_{\bar{x}}$, $f(\bar{y}) = \bar{x}$, and the map $f : V_{\bar{x}} \to V_{\bar{y}}$ is Lipschitz continuous.*

The generalized inverse function theorem is stated next.

THEOREM 2.7. *Let $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^n$ be locally Lipschitz at $x \in \Omega$. If the generalized Jacobian $\partial F(x)$ is nonsingular, then F is a locally Lipschitz homeomorphism at x.*

Let us point out that while the nonsingularity of the Jacobian is necessary and sufficient for continuously differentiable mappings, Theorem 2.7 only gives a sufficient condition for F to be a locally Lipschitz homeomorphism.

Since we introduced ∂F as a generalization of ∇F in the case where F is not (continuously) differentiable, we may attempt to define a generalized version of our local Newton method, Algorithm 1.

ALGORITHM 2 (Newton's method for nonsmooth systems.).
Given $F : \mathbb{R}^n \to \mathbb{R}^n$ locally Lipschitz continuous and $x_0 \in \mathbb{R}^n$, $k := 0$:

(1) *Unless a stopping rule is satisfied, solve*

$$G(x_k)d_k = -F(x_k)$$

for d_k, where $G(x_k)$ is an arbitrary element of $\partial F(x_k)$.

(2) *Set $x_{k+1} := x_k + d_k$, $k = k + 1$, and go to (1).*

As in the case of Algorithm 1, one has to guarantee the nonsingularity of $G(x_k)$. Note that now this is a more involved task: (i) The generalized Jacobian ∂F can be set-valued; (ii) we cannot argue that $\|G(y) - G(x)\|$ becomes small as y approaches x.

Note further that we more or less only copied Algorithm 1 without recalling its motivation. Remember, our key motivation for defining Algorithm 1 was (3), where we argued that locally the linearization $m(x_k + d)$ of F about x_k gives a reasonably good approximation of F. This need not be the case for F being only locally Lipschitz. Consequently the question arises: "What type of convergence do we expect"? Without requiring further properties of F, at best one can only expect Q-linear convergence of Algorithm 2 (presumed that the iteration is well-defined) if convergence at all.

We finish this section by proving that nonsingularity of the generalized Jacobian at x allows to argue nonsingularity of the generalized Jacobians in a sufficiently small neighborhood of x.

THEOREM 2.8. *If $\partial F(x)$ is nonsingular, i.e., all $G \in \partial F(x)$ are nonsingular, then there exists $\beta, r > 0$ such that, for any $y \in B(x,r)$ and any $G(y) \in \partial F(y)$, $G(y)$ is nonsingular and satisfies*

$$\|G(y)^{-1}\| \leq \beta.$$

Proof. Assume that the conclusion is not true. Then there exists a sequence $x_k \to x$, $G_k \in \partial F(x_k)$ such that either all G_k are singular or $\|G_k^{-1}\| \to +\infty$. By Theorem 2.2 (c) there exists a subsequence $\{x_{k(l)}\}$ such that $G_{k(l)} \to G$. Then G must be singular. From the upper semicontinuity of ∂F we obtain $G \in \partial F(x)$ which is a contradiction to the nonsingularity of $\partial F(x)$. □

In general, this result is not enough for proving well-definedness of the generalized Newton iteration, Algorithm 2. In fact, without further assumptions on F it cannot be guaranteed that $x_{k+1} \in B(x_*, \epsilon)$ when $x_k \in B(x_*, \epsilon)$, where x_* satisfies $F(x_*) = 0$ and $\epsilon > 0$ is sufficiently small (compare our local convergence result, Theorem 1.1, in the smooth case).

2. Semismoothness and semismooth Newton methods

In this section we narrow the class of generalized differentiable functions such that we finally obtain well-definedness and local superlinear convergence of the generalized version of Newtons' method, *i.e.*, Algorithm 2.

DEFINITION 2.5. *Let $U \subset \mathbb{R}^n$ be nonempty and open. The function $F : U \to \mathbb{R}^m$ is* semismooth *at $x \in U$, if it is locally Lipschitz at x and if*

$$\lim_{\substack{G \in \partial F(x+t\tilde{d}) \\ \tilde{d} \to d, t \downarrow 0}} G\tilde{d} \tag{15}$$

exists for all $d \in \mathbb{R}^n$. If F is semismooth at all $x \in U$, we call F semismooth *(on* U*).*

The concept of semismoothness of a function was introduced in [**26**] and extended in [**29**]. The characterization of semismoothness in Definition 2.5 is cumbersome to handle; especially in connection with the generalized Newton method. The following theorem provides equivalent characterizations which are more convenient.

THEOREM 2.9. *Let $F : U \to \mathbb{R}^m$ be defined on the open set $U \subset \mathbb{R}^n$. Then, for $x \in U$, the following statements are equivalent:*

(a) *F is semismooth at x.*

(b) *F is locally Lipschitz continuous at x, $F'(x;\cdot)$ exists, and, for any $G \in \partial F(x+d)$,*

$$\|Gd - F'(x,d)\| = \mathcal{O}(\|d\|) \quad \text{as } d \to 0.$$

(c) *F is locally Lipschitz continuous at x, $F'(x;\cdot)$ exists, and, for any $G \in \partial F(x+d)$,*

$$\|F(x+d) - F(x) - Gd\| = \mathcal{o}(\|d\|) \quad \textit{as } d \to 0.$$

Assertion (c) of Theorem 2.9 is particularly useful when proving local superlinear convergence of Newton's method. We also mention that semismoothness is closed under scalar multiplication, summation, and composition. Further a vector-valued function F is semismooth iff its component functions are semismooth.

Theorem 2.10. *Let $U \subset \mathbb{R}^n$ be open. If $F : U \to \mathbb{R}^m$ is continuously differentiable in a neighborhood of $x \in U$, then F is semismooth at x.*

Examples for semismooth functions are:

- $F : \mathbb{R}^n \to \mathbb{R}$, $x \mapsto \|x\|_{\ell_2}^2$.
- $\phi_{\mathrm{FB}} : \mathbb{R}^2 \to \mathbb{R}$, $x \mapsto x_1 + x_2 - \|x\|_{\ell_2}$. This function is called *Fischer-Burmeister function.* It has the following nice property:

$$a \geq 0,\ b \geq 0,\ ab = 0 \quad \Leftrightarrow \quad \phi_{\mathrm{FB}}(a,b) = 0. \tag{16}$$

- Another semismooth function satisfying a relation like (16) is $\phi_{\max} : \mathbb{R}^2 \to \mathbb{R}$, $x \mapsto x_1 - \max\{x_1 - cx_2, 0\}$ with $c > 0$ arbitrarily fixed.

For semismooth functions F we can prove the following local convergence result for the generalized Newton method, Algorithm 2. Whenever F is semismooth we call Algorithm 2 the *semismooth Newton method.*

Theorem 2.11 (local convergence). *Suppose that $x_* \in \mathbb{R}^n$ satisfies $F(x_*) = 0$, F is locally Lipschitz continuous and semismooth at x_*, and $\partial F(x_*)$ is nonsingular. Then there exists $\epsilon > 0$ such that for $x_0 \in B(x_*, \epsilon)$ the sequence $\{x_k\}$ generated by the semismooth Newton method (Algorithm 2) is well-defined, converges to x_*, and satisfies*

$$\|x_{k+1} - x_*\| = \mathcal{o}(\|x_k - x_*\|), \quad \textit{as } k \to +\infty.$$

Proof. First observe that there exists $r > 0$ such that $\partial F(x)$ is nonsingular for any $x \in B(x_*, r)$ by Theorem 2.8. Let $\beta > 0$ denote the bound on $\|G(x)^{-1}\|$ for all $x \in B(x_*, r)$, and choose $\epsilon \leq r$. The rest of the proof is by induction. We have, for some $G(x_0) \in \partial F(x_0)$,

$$\begin{aligned} x_1 - x_* &= x_0 - x_* - G(x_0)^{-1}F(x_0) \\ &= G(x_0)^{-1}\left(F(x_*) - F(x_0) - G(x_0)(x_* - x_0)\right) \end{aligned}$$

Due to the semismoothness of F at x_*, for arbitrary $0 < \eta \leq \frac{c}{\beta}$, with $0 < c < 1$ fixed, there exists $\tilde{r} > 0$ such that

$$\|F(x_*) - F(x_0) - G(x_0)(x_* - x_0)\| \leq \eta\|x_0 - x_*\|$$

for $x_0 \in B(x_*, \tilde{r})$. By possibly reducing ϵ such that $\epsilon \leq \min\{r, \tilde{r}\}$ we obtain

$$\|x_1 - x_*\| \leq \beta\eta\|x_0 - x_*\| \leq c\|x_0 - x_*\| < \|x_0 - x_*\|.$$

Thus, if $x_0 \in B(x_*, \epsilon)$ then $x_1 \in B(x_*, \epsilon)$ as well. The induction step is similar. From this we also obtain $x_k \to x_*$ since $0 < c < 1$. But then, due to the semismoothness of F at x_*, we find

$$\|x_{k+1} - x_*\| \leq \beta \|F(x_*) - F(x_k) - G(x_k)(x_* - x_k)\| = o(\|x_k - x_*\|)$$

as $k \to \infty$. □

In other words, the semismooth Newton method converges locally at a Q-superlinear rate. We can sharpen this result by requiring F to be semismooth of order γ.

DEFINITION 2.6. *Let $F : U \to \mathbb{R}^n$ be defined on the open set $U \subset \mathbb{R}^n$. Then, for $0 < \gamma \leq 1$, F is called* γ-order semismooth *at $x \in U$ if F is locally Lipschitz at x, $F'(x, \cdot)$ exists, and, for any $G \in \partial F(x + d)$,*

$$\|Gd - F'(x, d)\| = \mathcal{O}(\|d\|^{1+\gamma}) \quad \text{as } d \to 0.$$

If F is γ-order semismooth at all $x \in U$, then we call F γ-order semismooth *(on U).*

We have a similar characterization to Theorem 2.9 (c).

THEOREM 2.12. *Let $F : U \to \mathbb{R}^m$ be defined on the open set $U \subset \mathbb{R}^n$. Then, for $x \in U$ and $0 < \gamma \leq 1$, the following statements are equivalent:*

(a) *F is γ-order semismooth at x.*
(b) *F is locally Lipschitz continuous at x, $F'(x, \cdot)$ exists, and, for any $G \in \partial F(x + d)$ there holds*

$$\|F(x + d) - F(x) - Gd\| = \mathcal{O}(\|d\|^{1+\gamma}) \quad \text{as } d \to 0.$$

Similar to before, we can connect γ-order semismoothness to continuity of the derivative of F.

THEOREM 2.13. *Let $U \subset \mathbb{R}^n$ be open. If $F : U \to \mathbb{R}^m$ is continuously differentiable in a neighborhood of $x \in U$ with γ-Hölder continuous derivative, $0 < \gamma \leq 1$, then F is γ-order semismooth at x.*

Our examples of semismooth functions in fact turn out to be 1-order semismooth.

The local convergence result for the semismooth Newton algorithm now reads as follows.

THEOREM 2.14 (local convergence, γ-order semismoothness). *Suppose that $x_* \in \mathbb{R}^n$ satisfies $F(x_*) = 0$, F is locally Lipschitz continuous and γ-order semismooth at x_*, with $0 < \gamma \leq 1$, and $\partial F(x_*)$ is nonsingular. Then there exists $\epsilon > 0$ such that for $x_0 \in B(x_*, \epsilon)$ the sequence $\{x_k\}$ generated by the semismooth Newton method (Algorithm 2) is well-defined, converges to x_*, and satisfies*

$$\|x_{k+1} - x_*\| = \mathcal{O}(\|x_k - x_*\|^{1+\gamma}), \quad \text{as } k \to \infty.$$

Note that for $\gamma \in (0,1)$ we obtain locally Q-superlinear convergence with Q-order $1+\gamma$. If $\gamma = 1$, then we obtain a locally Q-quadratic convergence rate.

Finally, we point out that a result inspired by the Kantorovich theorem, Theorem 1.3, is available.

THEOREM 2.15 (global convergence). *Suppose that F is locally Lipschitz continuous and semismooth on U_0, the closure of $B(x_0, r)$. Also suppose that for any $G(x) \in \partial F(x)$, $x \in U_0$, $G(x)$ is nonsingular, and, with $y \in U_0$,*

$$\|G(x)^{-1}\| \leq \beta, \quad \|G(x)(y-x) - F'(x; y-x)\| \leq L\|y-x\|,$$
$$\|F(y) - F(x) - F'(x; y-x)\| \leq \eta\|y-x\|,$$

where $\alpha = \beta(L+\eta) < 1$ and $\beta\|F(x_0)\| \leq r(1-\alpha)$. Then the iterates $\{x_k\}$ of the semismooth Newton algorithm, Algorithm 2, remain in U_0 and converge to the unique solution x_* of $F(x) = 0$ in U_0. Moreover, there holds

$$\|x_k - x_*\| \leq \frac{\alpha}{1-\alpha}\|x_k - x_{k-1}\|, \quad k = 1, 2, \ldots$$

Notice the difference to the Kantorovich theorem. Now, the requirement $\alpha < 1$, which induces the smallness of L and η, cannot be achieved by locality arguments. Rather it represents a limitation on F. For instance, small L is obtained if the diameter of $\partial F(x)$ is small for all $x \in U_0$.

3. Inexact semismooth Newton methods

In practice, the exact solution of the Newton system $G(x_k)d_k = -F(x_k)$ is often expensive (e.g. when the system $F(x) = 0$ results from discretizing a (system of) partial differential equation(s)). In this case one seeks to solve the system only approximately. The following algorithm realizes inexact step computations.

ALGORITHM 3 (Inexact Newton's method for nonsmooth systems.).
Given $F : \mathbb{R}^n \to \mathbb{R}^n$ locally Lipschitz continuous, a sequence $\{\eta_k\}$ of nonnegative scalars, $x_0 \in \mathbb{R}^n$, $k := 0$:

(1) *Unless a stopping rule is satisfied, compute d_k such that*

$$G(x_k)d_k = -F(x_k) + r_k,$$

where $G(x_k)$ is an arbitrary element of $\partial F(x_k)$ and r_k satisfies

$$\|r_k\| \leq \eta_k\|G(x_k)\|.$$

(2) *Set $x_{k+1} := x_k + d_k$, $k = k+1$, and go to (1).*

One can prove the following convergence result.

THEOREM 2.16. *Suppose that $x_* \in \mathbb{R}^n$ satisfies $F(x_*) = 0$, F is locally Lipschitz continuous and semismooth at x_*, and $\partial F(x_*)$ is nonsingular. Then there exists a positive number $\bar{\eta} > 0$ such that, if $\eta_k \leq \bar{\eta}$ for all $k \in \mathbb{N}$, then there exists $\epsilon > 0$ such that for $x_0 \in B(x_*, \epsilon)$ the sequence $\{x_k\}$ generated by the semismooth Newton method (Algorithm 2) is well-defined* *and converges q-linearly to x_*. If furthermore $\eta_k \downarrow 0$, then the convergence rate is q-superlinear.*

CHAPTER 3

Generalized differentiability and semismoothness in infinite dimensions

The notion of generalize derivatives, which we introduced for functions on $\mathbb{R}^n$, may be readily extended to mappings $F : U \subset X \to \mathbb{R}$, where X denotes a Banach space and U a subset of X. Assuming that F is locally Lipschitz at x, the definition of the generalized directional derivative of F at x in Definition 2.2 immediately extends to the Banach space case. Also, the properties stated in Proposition 2.1 remain valid for F defined over a Banach space X. Note, however, that Proposition 2.1(ii) requires some modification as the generalized derivatives need to be defined differently. For this purpose observe first that Rademacher's Theorem cannot be extended readily to infinite dimensional Banach spaces. Hence, the definition of the generalized derivative needs to be re-visited. When X is a (finite or infinite dimensional) Banach space one may rely on the following construction:

- First of all we write $\langle \xi, d\rangle_{\mathbb{R}^n} := \xi^\top d$, where $\langle \cdot, \cdot\rangle_{\mathbb{R}^n}$ denotes the duality pairing in $\mathbb{R}^n$, i.e., upon identifying ξ with its dual we have $\langle \xi, \cdot\rangle_{\mathbb{R}^n} = \xi^\top \cdot$. In this sense, we generalize and use the duality pairing $\langle \cdot, \cdot\rangle_{X^*,X}$ whenever $F : U \subset X \to \mathbb{R}$. Here X^* denotes the dual space of X.
- Based on (14) and the Hahn-Banach Theorem, which states that any positively homogeneous and subadditive functional on X (such as F°) majorizes some linear functional on X (with the latter being an element of X^*). Thus, when F is locally Lipschitz at x and due to the generalization of Proposition 2.1, there exists at least one element $\xi \in X^*$ such that
$$F^\circ(x; d) \geq \langle \xi, d\rangle_{X^*,X} \quad \forall d \in X.$$
- Based on this, (14) can be extended to F defined on a general Banach space X, i.e.,
$$\partial F(x) := \{\xi \in X^* : \langle \xi, d\rangle_{X^*,X} \leq F^\circ(x; d) \quad \forall d \in X\}.$$

In the extension of Proposition 2.2(b) the compactness property has to be replaced by weak*-compactness of $\partial F(x)$ as a subset of X^* and $\|\xi\|_{X^*} \leq L$ for all $\xi \in \partial F(x)$, where $L > 0$ denotes the local Lipschitz constant of F at x. Note that the weak*-compactness is a direct consequence of Alaoglu's Theorem, which states that the closed unit ball of the dual space of a normed vector space is compact in the weak* topology. The upper semicontinuity of

∂F usually requires that X is finite dimensional. The notion of C-regularity of F and the calculus rules readily carry over to F defined on a Banach space X.

In optimization, convex functions play a special role since local minimizers of convex functions are also global minimizers and first order optimality conditions are both, necessary and sufficient. Now, let U be an open convex subset of X and $F : U \to \bar{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$ be convex, i.e., for all $u, v \in U$ and $\mu \in [0, 1]$ we have

$$F(\mu u + (1-\mu)v) \leq \mu F(u) + (1-\mu)F(v).$$

Proposition 3.1. *Let F be bounded from above on some neighborhood of a point in U. Then F is locally Lipschitz at x for any $x \in U$.*

The fact that tangent (hyper)planes support the graph of a convex function from below immediately yields the following definition.

Definition 3.1. *Let $F : U \subset X \to \bar{\mathbb{R}}$ be a convex function. Then the* subdifferential *of F at $x \in U$ is given by*

$$\partial F(x) = \{\xi \in X^* : F(x) + \langle \xi, y - x\rangle_{X^*,X} \leq F(y) \quad \forall y \in U\}.$$

The next result guarantees that there is no ambiguity between the subdifferential as defined above and the earlier notion of a generalized derivative of a locally Lipschitz mapping.

Proposition 3.2. *Let F be convex on U and locally Lipschitz at x, then $F^\circ(x; d) = F'(x; d)$ for all d and the generalized derivative of F at x coincides with the subdifferential of F at x.*

A convex function is called *proper* if it is not identically $+\infty$. The *domain* of the convex function $F : U \to \bar{\mathbb{R}}$ is given by $\mathrm{dom}(F) := \{x \in U : F(x) < +\infty\}$. We end this section by stating a chain rule for convex functions. Below $\mathcal{L}(X, Y)$ denotes the space of continuous linear operators between the Banach spaces X and Y. For $\Lambda \in \mathcal{L}(X, Y)$, $\Lambda^* \in \mathcal{L}(Y^*, X^*)$ denotes the dual operator of Λ.

Theorem 3.1. *Let X and Y denote Banach spaces, let $F_1 : X \to \bar{\mathbb{R}}$ and $F_2 : Y \to \bar{\mathbb{R}}$ be proper, lower semicontinuous, convex functions, and let $\Lambda \in \mathcal{L}(X, Y)$. Suppose that $0 \in \mathrm{int}(\Lambda\,\mathrm{dom}(F_1) - \mathrm{dom}(F_2))$. Then we have that*

$$\partial\left(F_1 + F_2 \circ \Lambda\right)(x) = \partial F_1(x) + \Lambda^* \partial F_2(\Lambda x).$$

1. Semismooth Newton methods in function space

Our aim is now to generalize the semismoothness concept to operator equations in function space. First note that the notion of semismoothness in $\mathbb{R}^n$ (see Definition 2.5) is based on Clarke's generalized Jacobian. The latter object, in turn, is based on Rademacher's theorem, Theorem 2.1. Unfortunately, Rademacher's theorem has no analogue in the infinite dimensional function space setting. Thus, the whole construction for proving locally superlinear convergence of our generalized Newton method fails. At this point, the reader may question the importance of studying Newton methods in infinite dimensional spaces, since every numerical realization only makes sense if we are able to *discretize* the problem. Then, after discretization, the problem is finite dimensional, and finding $x_*^h \in \mathbb{R}^{n^h}$ such that $F^h(x^h) = 0$ with $F^h : \mathbb{R}^{n^h} \to \mathbb{R}^{n^h}$ can serve as a prototype. Here we use superscript h to indicate that the nonlinear system arises as a discretization (with parameter h) of a general operator equation. For instance, in the case where F involves (nonlinear partial) differential equations, one may interpret h as the mesh size of the finite element or finite difference discretization. Now, once the problem can be written as an equation between $\mathbb{R}^{n^h}$ the methods and techniques of the previous chapter apply. However, by this approach the infinite dimensional properties of the problem are covered up and further numerical analysis is hardly possible. On the other hand, if we know that there exists a well-defined semismooth Newton method in function space, then this can be the basis for further investigations such as the proof of mesh independence of Newton's method. By the latter notion we refer to the fact that, for sufficiently small mesh sizes h, either the number of iterations of the discretized method is essentially constant w.r.t. h, or that the discretized method achieves essentially the same convergence rate independently of h. In general this need not be the case. A detailed discussion along these lines would go far beyond the scope of this presentation. We refer the interested reader to [**18**].

Studying the proof of Theorem 2.11 we can see that the characterization in Theorem 2.9 (c) is crucial. The idea is now to use this notion as the defining property for generalized derivatives of mappings between Banach spaces (finite as well as infinite dimensional ones). Consequently, we would automatically obtain semismoothness and locally superlinear convergence of the corresponding semismooth Newton method. For the following discussion let X, Z denote Banach spaces.

Definition 3.2. *The mapping* $F : D \subset X \to Z$ *is* generalized (or Newton) differentiable *on the open set* $U \subset D$ *if there exists a family of mappings* $G : U \to \mathcal{L}(X, Z)$ *such that*

$$\lim_{d \to 0} \frac{1}{\|d\|} \|F(x+d) - F(x) - G(x+d)d\| = 0. \tag{17}$$

for every $x \in U$.

Here $\mathcal{L}(X, Z)$ denotes the set of continuous linear operators from X to Z. We refer to such an operator G as a generalized (or Newton) derivative of F. Note that G need not be unique. However it is unique if F is Fréchet differentiable. Equation (17) resembles Theorem 2.9 (c).

Next we consider the problem

$$\text{Find } x_* \in X : \quad F(x_*) = 0. \tag{18}$$

Based on Definition 3.2 we define the following semismooth Newton algorithm.

Algorithm 4 (Newton's method for semismooth operator equations.). *Given $F : D \to Y$ generalized differentiable in $U \subset D$, and $x_0 \in U$, $k := 0$:*

(1) *Unless a stopping rule is satisfied, solve*

$$G(x_k)d_k = -F(x_k)$$

for d_k, where $G(x_k)$ is an arbitrary generalized derivative of F at x_k.

(2) *Set $x_{k+1} = x_k + d_k$, $k = k + 1$, and go to (1).*

We immediately obtain the following local convergence result.

Theorem 3.2. *Suppose that x_* is a solution to $F(x) = 0$, and that F is Newton differentiable in an open neighborhood U containing x_*, $G(x)$ is nonsingular for all $x \in U$ and $\{\|G(x)^{-1}\| : x \in U\}$ is bounded. Then the Newton iteration*

$$x_{k+1} = x_k - G(x_k)^{-1}F(x_k),$$

i.e., Algorithm 4, is well-defined and converges superlinearly to x_* provided that $\|x_0 - x_*\|$ is sufficiently small.

The proof technique of Theorem 3.2 is identical to the one of Theorem 2.11. Details can be found in [**13**]. Related concepts were introduced and analysed in [**10, 36**].

CHAPTER 4

Applications

1. A class of finite dimensional complementarity problems

We start by considering complementarity problems of the form

$$\left\{\begin{array}{l} Ay + \lambda = f, \\ y \le \psi,\ \lambda \ge 0,\ (\lambda, y - \psi) = 0\,, \end{array}\right. \tag{19}$$

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

Definition 4.1. *A $n \times n$-matrix is called a* P*-matrix if all its principal minors are positive.*

It is well-known [**5**] that A is a P-matrix if and only if all real eigenvalues of A and of its principal submatrices are positive. Here B is called a principal submatrix of A if it arises from A by deletion of rows and columns from the same index set $\mathcal{J} \subset \{1, \ldots, n\}$.

The assumption that A is a P-matrix guarantees the existence of a unique solution $(y^*, \lambda^*) \in \mathbb{R}^n \times \mathbb{R}^n$ of (19) [**5**]. In case A is symmetric positive definite (19) is the optimality system for

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

Note that the complementarity system given by the second line in (19) can equivalently be expressed as

$$\mathcal{C}(y, \lambda) = 0, \text{ where } \mathcal{C}(y, \lambda) = \lambda - \max(0, \lambda + c(y - \psi)), \tag{20}$$

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

Consequently (19) is equivalent to

$$\left\{\begin{array}{l} Ay + \lambda = f \\ \mathcal{C}(y, \lambda) = 0. \end{array}\right. \tag{21}$$

Applying a semismooth Newton step to (21) gives the following algorithm, which we shall frequently call the *primal-dual active set stratey.* From now on the iteration counter k is denoted as a superscript since we frequently will use subscripts for components of vectors.

Algorithm 5.

(i) *Initialize y^0, λ^0. Set $k = 0$.*

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

(iii) *Solve*

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

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

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

Above we utilize $y^{k+1} = \psi$ on $\mathcal{A}_k$ to stand for $y_i^{k+1} = \psi_i$ for $i \in \mathcal{A}_k$. We call $\mathcal{A}_k$ the estimate of the active set $\mathcal{A}^* = \{i : y_i^* = \psi_i\}$ and $\mathcal{I}_k$ the estimate of the inactive set $\mathcal{I}^* = \{i : y_i^* < \psi_i\}$. Hence the name of the algorithm.

Let us now argue that the above algorithm can be interpreted as a semismooth Newton method. For this purpose it will be convenient to arrange the coordinates in such a way that the active and inactive ones occur in consecutive order. This leads to the block matrix representation of A as

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

where $A_{\mathcal{I}_k} = A_{\mathcal{I}_k\mathcal{I}_k}$ and analogously for $A_{\mathcal{A}_k}$. Analogously the vector y is partitioned according to $y = (y_{\mathcal{I}_k}, y_{\mathcal{A}_k})$ and similarly for f and ψ. In Section 2 we shall argue that $v \to \max(0, v)$ from $\mathbb{R}^n \to \mathbb{R}^n$ is generalized differentiable in the sense of Definition 3.2 with a particular generalized derivative given by the diagonal matrix $G_m(v)$ with diagonal elements

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

Here we use the subscript m to indicate particular choices for the generalized derivative of the max-function. Note that G_m is also an element of the generalized Jacobian of the max-function. Semismooth Newton methods for generalized Jacobians in Clarke's sense were considered e.g. in [**29**].

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

$$\begin{pmatrix} A_{\mathcal{I}_k} & A_{\mathcal{I}_k\mathcal{A}_k} & I_{\mathcal{I}_k} & 0 \\ A_{\mathcal{A}_k\mathcal{I}_k} & A_{\mathcal{A}_k} & 0 & I_{\mathcal{A}_k} \\ 0 & 0 & I_{\mathcal{I}_k} & 0 \\ 0 & -cI_{\mathcal{A}_k} & 0 & 0 \end{pmatrix} \begin{pmatrix} \delta y_{\mathcal{I}_k} \\ \delta y_{\mathcal{A}_k} \\ \delta\lambda_{\mathcal{I}_k} \\ \delta\lambda_{\mathcal{A}_k} \end{pmatrix} = - \begin{pmatrix} (Ay^k + \lambda^k - f)_{\mathcal{I}_k} \\ (Ay^k + \lambda^k - f)_{\mathcal{A}_k} \\ \lambda^k_{\mathcal{I}_k} \\ -c(y^k - \psi)_{\mathcal{A}_k} \end{pmatrix} \tag{22}$$

where $I_{\mathcal{I}_k}$ and $I_{\mathcal{A}_k}$ are identity matrices of dimensions $\text{card}(\mathcal{I}_k)$ and $\text{card}(\mathcal{A}_k)$. The third equation in (22) implies that

$$\lambda^{k+1}_{\mathcal{I}_k} = \lambda^k_{\mathcal{I}_k} + \delta\lambda_{\mathcal{I}_k} = 0 \tag{23}$$

and the last one yields

$$y^{k+1}_{\mathcal{A}_k} = \psi_{\mathcal{A}_k}. \tag{24}$$

Equations (23) and (24) coincide with the conditions in the second line of step (iii) in the primal-dual active set algorithm. The first two equations in (22) are equivalent to $Ay^{k+1} + \lambda^{k+1} = f$, which is the first equation in step (iii).

Combining these observations we can conclude that the semi-smooth Newton update based on (22) is equivalent to the primal-dual active set strategy.

We also note that the system (22) is solvable since the first equation in (22) together with (23) gives

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

and consequently by (24)

$$A_{\mathcal{I}_k}\ y^{k+1}_{\mathcal{I}_k} = f_{\mathcal{I}_k} - A_{\mathcal{I}_k\mathcal{A}_k}\ \psi_{\mathcal{A}_k}. \tag{25}$$

Since A is a P-matrix $A_{\mathcal{I}_k}$ is regular and (25) determines $y^{k+1}_{\mathcal{I}_k}$. The second equation in (22) is equivalent to

$$\lambda^{k+1}_{\mathcal{A}_k} = f_{\mathcal{A}_k} - (Ay^{k+1})_{\mathcal{A}_k}. \tag{26}$$

In Section 3 we shall consider (P) in the space $L^2(\Omega)$. Again one can show that the semi-smooth Newton update and the primal-dual active set strategy coincide.

2. Convergence analysis: the finite dimensional case

This section is devoted to local as well as global convergence analysis of the semismooth Newton algorithm to solve

$$\begin{cases} Ay + \lambda = f \\ \lambda - \max(0, \lambda + c(y - \psi)) = 0, \end{cases} \tag{27}$$

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

$$G_m(y) = \operatorname{diag}\ (g_1(y_1), \cdots, g_n(y_n)), \tag{28}$$

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

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

Lemma 4.1. *The mapping $y \to \max(0, y)$ from $\mathbb{R}^n$ to $\mathbb{R}^n$ is generalized differentiable on $\mathbb{R}^n$ and G_m defined in (28) is a particular element of the generalized derivative for every $\delta \in \mathbb{R}^n$.*

Let us now turn to the convergence analysis of the semi-smooth Newton method for (27). Note that the choice G_m for in Section 1 corresponds to a generalized derivative with $\delta = 0$. In view of (23)–(26), for $k \geq 1$ the Newton update (22) is equivalent to

$$\begin{pmatrix} A_{\mathcal{I}_k} & 0 \\ A_{\mathcal{A}_k\mathcal{I}_k} & I_{\mathcal{A}_k} \end{pmatrix} \begin{pmatrix} \delta y_{\mathcal{I}_k} \\ \delta\lambda_{\mathcal{A}_k} \end{pmatrix} = - \begin{pmatrix} A_{\mathcal{I}_k\mathcal{A}_k}\delta y_{\mathcal{A}_k} + \delta\lambda_{\mathcal{I}_k} \\ A_{\mathcal{A}_k}\delta y_{\mathcal{A}_k} \end{pmatrix} \tag{29}$$

and

$$\delta\lambda_i = -\lambda_i^k,\ i \in \mathcal{I}_k, \quad \text{and} \quad \delta y_i = \psi_i - y_i^k,\ i \in \mathcal{A}_k. \tag{30}$$

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

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

and note that (27) is equivalent to $F(y, \lambda) = 0$. As a consequence of Lemma 4.1 the mapping F is generalized differentiable and the system matrix of (22) is an element of the generalized derivative for F with the particular choice G_m for the max-function. We henceforth denote the generalized derivative of F by G_F.

Let (y^*, λ^*) denote the unique solution to (27) and $x^0 = (y^0, \lambda^0)$ the initial values of the iteration. From Theorem 3.2 we deduce the following fact:

Theorem 4.1. *Algorithm 5 converges superlinearly to $x^* = (y^*, \lambda^*)$, provided that $\|x^0 - x^*\|$ is sufficiently small.*

The boundedness requirement of $(G_F)^{-1}$ according to Theorem 3.2 can be derived analogously to the infinite dimensional case; see the proof of Theorem 4.6.

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

Let us address global convergence next. In the following two results sufficient conditions for convergence for arbitrary initial data $x^0 = (y^0, \lambda^0)$ are given. We recall that A is referred to as M-matrix, if it is nonsingular, $(m_{ij}) \le 0$, for $i \neq j$, and $M^{-1} \ge 0$. Our notion of an M-matrix coincides with that of nonsingular M-matrices as defined in [**5**].

Theorem 4.2. *Assume that A is a* M-*matrix. Then $x^k \to x^*$ for arbitrary initial data. Moreover, $y^* \le y^{k+1} \le y^k$ for all $k \ge 1$ and $y^k \le \psi$ for all $k \ge 2$.*

Proof. The assumption that A is a M-matrix implies that for every index partition $\mathcal{I}$ and $\mathcal{A}$ we have $A_{\mathcal{I}}^{-1} \ge 0$ and $A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} \le 0$, see [**5**, p. 134]. Let us first show the monotonicity property of the y-component. Observe that for every $k \ge 1$ the complementarity property

$$\lambda_i^k = 0 \quad \text{or} \quad y_i^k = \psi_i\,, \quad \text{for all } i, \text{ and } k \ge 1\,, \tag{31}$$

holds. For $i \in \mathcal{A}_k$ we have $\lambda_i^k + c(y_i^k - \psi_i) > 0$ and hence by (31) either $\lambda_i^k = 0$, which implies $y_i^k > \psi_i$, or $\lambda_i^k > 0$, which implies $y_i^k = \psi_i$. Consequently $y^k \ge \psi = y^{k+1}$ on $\mathcal{A}_k$ and $\delta y_{\mathcal{A}_k} = \psi_{\mathcal{A}_k} - y_{\mathcal{A}_k}^k \le 0$. For $i \in \mathcal{I}_k$ we have $\lambda_i^k + c(y_i^k - \psi_i) \le 0$ which implies $\delta\lambda_{\mathcal{I}_k} \ge 0$ by (22) and (31). Since $\delta y_{\mathcal{I}_k} =$

$-A_{\mathcal{I}_k}^{-1} A_{\mathcal{I}_k \mathcal{A}_k} \delta y_{\mathcal{A}_k} - A_{\mathcal{I}_k}^{-1} \delta\lambda_{\mathcal{I}_k}$ by (29) it follows that $\delta y_{\mathcal{I}_k} \le 0$. Therefore $y^{k+1} \le y^k$ for every $k \ge 1$.

Next we show that y^k is feasible for all $k \ge 2$. Due to the monotonicity of y^k it suffices to show that $y^2 \le \psi$. Let $V = \{i : y_i^1 > \psi_i\}$. For $i \in V$ we have $\lambda_i^1 = 0$ by (31), and hence $\lambda_i^1 + c(y_i^1 - \psi_i) > 0$ and $i \in \mathcal{A}_1$. Since $y^2 = \psi$ on $\mathcal{A}_1$ and $y^2 \le y^1$ it follows that $y^2 \le \psi$.

To verify that $y^* \le y^k$ for all $k \ge 1$ note that

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

It follows that

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

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

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

Monotonicity of y^k and $y^* \le y^k \le \psi$ for $k \ge 2$ imply the existence of $\bar{y}$ such that $\lim y^k = \bar{y} \le \psi$. Since $\lambda^k = Ay^k + f \ge 0$ for all $k \ge k_o$, there exists $\bar{\lambda}$ such that $\lim \lambda^k = \bar{\lambda} \ge 0$. Together with (31) it follows that $(\bar{y}, \bar{\lambda}) = (y^*, \lambda^*)$. □

REMARK 4.1. *Concerning the applicability of Theorem 4.2 we recall that many discretizations of second order differential operators give rise to* M-*matrices.*

For a rectangular matrix $B \in \mathbb{R}^{n \times m}$ we denote by $\| \cdot \|_1$ the subordinate matrix norm when both $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the 1-norms. Moreover, B_+ denotes the $n \times m$-matrix containing the positive parts of the elements of B. The following result can be applied to discretizations of constrained optimal control problems. We refer to the end of Section 3 for a discussion of the conditions of the following Theorem 4.3 in the case of control constrained optimal control problems.

THEOREM 4.3. *If A is a* P-*matrix and for every partitioning of the index set into disjoint subsets $\mathcal{I}$ and $\mathcal{A}$ we have $\|(A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}})_+\|_1 < 1$ and $\sum_{i \in \mathcal{I}} (A_{\mathcal{I}}^{-1} y_{\mathcal{I}})_i \ge 0$ for $y_{\mathcal{I}} \ge 0$, then $\lim_{k \to \infty} x^k = x^*$.*

Proof. From (29) we have

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

and upon summation over the inactive indices

$$\begin{aligned}\sum_{\mathcal{I}_k}(y_i^{k+1}-\psi_i) = \sum_{\mathcal{I}_k}(y_i^k-\psi_i) + \sum_{\mathcal{I}_k}\left(A^{-1}_{\mathcal{I}_k}A_{\mathcal{I}_k\mathcal{A}_k}(y^k-\psi)_{\mathcal{A}_k}\right)_i \\ + \sum_{\mathcal{I}_k}(A^{-1}_{\mathcal{I}_k}\lambda^k_{\mathcal{I}_k})_i\end{aligned} \tag{32}$$

Adding the obvious equality

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

to (32) implies

$$\sum_{i=1}^n(y_i^{k+1}-y_i^k) \le -\sum_{\mathcal{A}_k}(y_i^k-\psi_i) + \sum_{\mathcal{I}_k}(A^{-1}_{\mathcal{I}_k}A_{\mathcal{I}_k\mathcal{A}_k}(y^k-\psi)_{\mathcal{A}_k})_i\,. \tag{33}$$

Here we used the fact $\lambda^k_{\mathcal{I}_k} = -\delta\lambda_{\mathcal{I}_k} \le 0$, established in the proof of Theorem 4.2. There it was also argued that $y^k_{\mathcal{A}_k} \ge \psi_{\mathcal{A}_k}$. Hence it follows that

$$\sum_{i=1}^n(y_i^{k+1}-y_i^k) \le -\|y^k-\psi\|_{1,\mathcal{A}_k} + \|(A^{-1}_{\mathcal{I}_k}A_{\mathcal{I}_k\mathcal{A}_k})_+\|_1\,\|y^k-\psi\|_{1,\mathcal{A}_k} < 0\,, \tag{34}$$

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

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

acts as a merit function for the algorithm. Since there are only finitely many possible choices for active/inactive sets there exists an iteration index $\bar{k}$ such that $\mathcal{I}_{\bar{k}} = \mathcal{I}_{\bar{k}+1}$. Moreover, $(y^{\bar{k}+1},\lambda^{\bar{k}+1})$ is solution to (27). In fact, in view of (iii) of the algorithm it suffices to show that $y^{\bar{k}+1}$ and $\lambda^{\bar{k}+1}$ are feasible. This follows from the fact that due to $\mathcal{I}_{\bar{k}} = \mathcal{I}_{\bar{k}+1}$ we have $c(y_i^{\bar{k}+1}-\psi_i) = \lambda_i^{\bar{k}+1} + c(y_i^{\bar{k}+1}-\psi_i) \le 0$ for $i \in \mathcal{I}_{\bar{k}}$ and $\lambda_i^{\bar{k}+1}+c(y_i^{\bar{k}+1}-\psi_i) > 0$ for $i \in \mathcal{A}_{\bar{k}}$. Thus the algorithm converges in finitely many steps. □

REMARK 4.2. *Let us note as a corollary to the proof of Theorem 4.3 that in case A is a* M-*matrix then $\mathcal{M}(y^k) = \sum_{i=1}^n y_i^k$ is always a merit function. In fact, in this case the conditions of Theorem 4.3 are obviously satisfied.*

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

THEOREM 4.4. *Assume that $A = M + K$ with M an* M*-matrix and with K an $n \times n$-matrix. Then, if $\|K\|_1$ is sufficiently small, (27) admits a unique solution $x^* = (y^*, \lambda^*)$, the primal-dual active set algorithm is well-defined and* $\lim_{k\to\infty} x^k = x^*$.

Proof. Recall that as a consequence of the assumption that M is a M-matrix all principal submatrices of M are nonsingular M-matrices as well [**5**]. Let $\mathcal{S}$ denote the set of all subsets of $\{1, \dots, n\}$, and define

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

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

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

As a consequence the algorithm is well-defined. Proceeding as in the proof of Theorem 4.3 we arrive at

$$\begin{aligned}\sum_{i=1}^{n}(y_i^{k+1} - y_i^k) = -\sum_{i\in\mathcal{A}}(y_i^k - \psi_i) + \sum_{i\in\mathcal{I}} \left(A_{\mathcal{I}}^{-1} A_{\mathcal{I}\mathcal{A}} (y^k - \psi)_{\mathcal{A}}\right)_i \\ + \sum_{i\in\mathcal{I}} (A_{\mathcal{I}}^{-1} \lambda_{\mathcal{I}}^k)_i \,,\end{aligned} \tag{35}$$

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

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

and consequently by (35)

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

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

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

is a merit function for the algorithm. The proof is now completed in the same manner as that of Theorem 4.3. □

The assumptions of Theorem 4.4 do not require A to be a P-matrix. From its conclusions existence of a solution to (27) for arbitrary f follows. This is equivalent to the fact that A is a P-matrix [**5**, Theorem 10.2.15]. Hence, it follows that Theorem 4.4 represents a sufficient condition for A to be a P-matrix.

Observe further that the M-matrix property is not stable under arbitrarily small perturbations since off-diagonal elements may become positive. This implies certain limitations of the applicability of Theorem 4.2. Theorem 4.4 guarantees that convergence of the primal-dual active set strategy for arbitrary initial data is preserved for sufficiently small perturbations K of an M-matrix. Therefore, Theorem 4.4 is also of interest in connection with numerical implementations of the primal-dual active set algorithm.

Finally, we shall point out that Theorems 4.2–4.4 establish global convergence of the primal-dual active set strategy or, equivalently, semi-smooth Newton method without the necessity of a line search. The rate of convergence is locally superlinear. Moreover, it can be observed from (22) that if $\mathcal{I}_k = \mathcal{I}_{k'}$ for $k \neq k'$, then $y^k = y^{k'}$ and $\lambda^k = \lambda^{k'}$. Hence, in case of convergence no cycling of the algorithm is possible, and termination at the solution of (19) occurs after finitely many steps.

3. The infinite dimensional case

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

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

$$G_m(y)(x) = \begin{cases} 1 & \text{if} \quad y(x) > 0\,, \\ 0 & \text{if} \quad y(x) < 0\,, \\ \delta & \text{if} \quad y(x) = 0\,, \end{cases} \tag{36}$$

where $y \in X$.

Theorem 4.5.

(i) *G_m can in general not serve as a slanting function for* $\max(0, \cdot)\colon L^p(\Omega) \to L^p(\Omega)$, *for* $1 \le p \le \infty$.

(ii) *The mapping* $\max(0, \cdot)\colon L^q(\Omega) \to L^p(\Omega)$ *with* $1 \le p < q \le \infty$ *is slantly differentiable on* $L^q(\Omega)$ *and* G_m *is a slanting function.*

Proof. (i) It suffices to consider the one dimensional case $\Omega = (-1, 1) \subset \mathbb{R}$. We show that property (17) does not hold at $y(x) = -|x|$. Let us define

$h_n(x) = \frac{1}{n}$ on $(-\frac{1}{n}, \frac{1}{n})$ and $h_n(x) = 0$ otherwise. Then

$$\int_{-1}^{1} |\max(0, y + h_n)(x) - \max(0, y)(x) - (G_m(y + h_n)(h_n))(x)|^p \, dx$$

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

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

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

and hence (17) is not satisfied at y for any $p \in [1, \infty)$.

To consider the case $p = \infty$ we choose $\Omega = (0, 1)$ and show that (17) is not satisfied at $y(x) = x$. For this purpose define for $n = 2, \ldots$

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

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

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

and hence (17) cannot be satisfied.

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

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

A short computation shows that

$$|D_{y,h}(x)| \begin{cases} \leq |y(x)| & \text{if } (y(x) + h(x))y(x) < 0\,, \\ \leq (1 + |\delta|)\, |y(x)| & \text{if } y(x) + h(x) = 0\,, \\ = 0 & \text{otherwise.} \end{cases} \tag{37}$$

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

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

From (37) it follows that only

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

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

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

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

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

It follows that

$$\lim_{\|h\|_{L^q(\Omega)}\to 0} |\Omega_\epsilon(h)| = 0 \quad \text{for every fixed } \epsilon > 0\,. \tag{38}$$

For $\epsilon > 0$ we further define sets

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

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

$$\lim_{\epsilon\to 0^+} |\Omega^\epsilon(y)| = 0\,. \tag{39}$$

From (37) we find

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

Choose $\eta > 0$ arbitrarily and note that by (39) there exists $\bar\epsilon > 0$ such that $(1+|\delta|)|\Omega^{\bar\epsilon}(y)|^{(q-p)/(qp)} < \eta$. Consequently

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

and by (38)

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

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

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

We refer to [**36**] for a related investigation of the *two-norm problem* involved in Proposition 4.5 in the case of superposition operators. An example in [**36**] proves the necessity of the norm-gap for the case in which the complementarity condition is expressed by means of the Fischer-Burmeister functional.

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

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

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

$$(Ay, y) \geq \gamma \|y\|^2 \,, \tag{H1}$$

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

$$\begin{cases} Ay^* + \lambda^* = f, \\ \mathcal{C}(y^*, \lambda^*) = 0, \end{cases} \tag{40}$$

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

Algorithm 6.

(i) *Choose* y^0, λ^0 *in* $L^2(\Omega)$. *Set* $k = 0$.

(ii) *Set* $\mathcal{A}_k = \{x \colon \lambda^k(x) + c(y^k(x) - \psi(x)) > 0\}$ *and* $\mathcal{I}_k = \Omega \backslash \mathcal{A}_k$.

(iii) *Solve*

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

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

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

For the semi-smooth Newton step we can refer back to Section 1. At iteration level k with $(y^k, \lambda^k) \in L^2(\Omega) \times L^2(\Omega)$ given, it is of the form (22) where now $\delta y_{\mathcal{I}_k}$ denotes the restriction of δy (defined on Ω) to $\mathcal{I}_k$ and analogously for the remaining terms. Moreover $A_{\mathcal{I}_k \mathcal{A}_k} = E^*_{\mathcal{I}_k} A \, E_{\mathcal{A}_k}$, where $E_{\mathcal{A}_k}$ denotes the extension-by-zero operator for $L^2(\mathcal{A}_k)$ to $L^2(\Omega)$–functions, and its adjoint $E^*_{\mathcal{A}_k}$ is the restriction of $L^2(\Omega)$–functions to $L^2(\mathcal{A}_k)$, and similarly for $E_{\mathcal{I}_k}$ and $E^*_{\mathcal{I}_k}$. Moreover $A_{\mathcal{A}_k \mathcal{I}_k} = E^*_{\mathcal{A}_k} A \, E_{\mathcal{I}_k}$, $A_{\mathcal{I}_k} = E^*_{\mathcal{I}_k} A \, E_{\mathcal{I}_k}$ and $A_{\mathcal{A}_k} = E^*_{\mathcal{A}_k} A \, E_{\mathcal{A}_k}$. It can be argued precisely as in Section 1 that the above primal-dual active set strategy, *i.e.*, Algorithm 6, and the semi-smooth Newton updates coincide, provided that the generalized derivative of the max-function is taken according to

$$G_m(u)(x) = \begin{cases} 1 & \text{if} \quad u(x) > 0 \\ 0 & \text{if} \quad u(x) \leq 0, \end{cases} \tag{41}$$

which we henceforth assume.

Proposition 4.5 together with Theorem 3.2 suggest that the semi-smooth Newton algorithm applied to (40) may not converge in general. We therefore restrict our attention to operators A of the form

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

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

EXAMPLE 4.1. *We consider the optimal control problem*

$$\text{(42)}\qquad \begin{cases} \textit{minimize} & \frac{1}{2}\|y-z\|^2_{L^2}+\frac{\beta}{2}\|u\|^2_{L^2} \\ \textit{subject to} & -\Delta y = u \textit{ in } \Omega,\ y=0 \textit{ on } \partial\Omega\,, \\ & u\le\psi,\ u\in L^2(\Omega)\,, \end{cases}$$

where $z\in L^2(\Omega)$, $\psi\in L^q(\Omega)$, *and* $\beta>0$. *Let* $B\in\mathcal{L}(H^1_o(\Omega),H^{-1}(\Omega))$ *denote the operator* $-\Delta$ *with homogeneous Dirichlet boundary conditions. Then (42) can equivalently be expressed as*

$$\text{(43)}\qquad \begin{cases} \textit{minimize} & \frac{1}{2}\|B^{-1}u-z\|^2_{L^2}+\frac{\beta}{2}\|u\|^2_{L^2} \\ \textit{subject to} & u\le\psi,\ u\in L^2(\Omega)\,. \end{cases}$$

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

In (42) *we considered the distributed control case. A related boundary control problem is given by*

$$\text{(44)}\qquad \begin{cases} \textit{minimize} & \frac{1}{2}\|y-z\|^2_{L^2(\Omega)}+\frac{\beta}{2}\|u\|^2_{L^2(\partial\Omega)} \\ \textit{subject to} & -\Delta y+y=0 \textit{ in } \Omega,\ \frac{\partial y}{\partial n}=u \textit{ on } \partial\Omega\,, \\ & u\le\psi,\ u\in L^2(\partial\Omega)\,, \end{cases}$$

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

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

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

For the sake of illustration it is also worthwhile to specify (23)–(26), which were found to be equivalent to the Newton-update (22) for the case of optimal control problems. We restrict ourselves to the case of the distributed control problem (42). Then (23)–(26) can be expressed as

(45)

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

where we set $B^{-2} = B^{-1}\mathcal{J}B^{-1}$. Setting $p^{k+1} = B^{-1}z - B^{-2}u^{k+1}$, a short computation shows that (45) is equivalent to

$$
\begin{cases}
-\Delta y^{k+1} = u^{k+1} \quad \text{in } \Omega\,, \quad y^{k+1} = 0 \quad \text{on } \partial\Omega\,,\\
-\Delta p^{k+1} = z - y^{k+1} \quad \text{in } \Omega\,, \quad p^{k+1} = 0 \quad \text{on } \partial\Omega\,,\\
\quad p^{k+1} = \beta u^{k+1} + \lambda^{k+1} \quad \text{in } \Omega\,,\\
\quad u^{k+1} = \psi \quad \text{in } \mathcal{A}_k\,, \lambda^{k+1} = 0 \quad \text{in } \mathcal{I}_k\,.
\end{cases}
\tag{46}
$$

This is the system in the primal variables (y,u) and adjoint variables (p,λ), previously implemented in [**2**] *for testing the algorithm.*

Our main intention is to consider control constrained problems as in Example 4.1. To prove convergence under assumptions (H1), (H2) we utilize a reduced algorithm which we explain next.

The operators $E_{\mathcal{I}}$ and $E_{\mathcal{A}}$ denote the extension by zero and their adjoints are restrictions to $\mathcal{I}$ and $\mathcal{A}$, respectively. The optimality system (40) does not depend on the choice of $c > 0$. Moreover, from the discussion in Section 1 the primal-dual active set strategy is independent of $c > 0$ after the initialization phase. For the specific choice $c = \beta$ system (40) can equivalently be expressed as

$$\beta y^* - \beta\psi + \max(0, Cy^* - f + \beta\psi) = 0\,, \tag{47}$$

$$\lambda^* = f - Cy^* - \beta y^*\,. \tag{48}$$

We shall argue in the proof of Theorem 4.6 below that the primal-dual active set method in $L^2(\Omega)$ for (y,λ) is equivalent to the following algorithm for the reduced system (47)– (48), which will be shown to converge superlinearly.

Algorithm 7 (Reduced algorithm).

(i) *Choose $y^0 \in L^2(\Omega)$ and set $k = 0$.*

(ii) *Set $\mathcal{A}_k = \{x : (f - Cy_k - \beta\psi)(x) > 0\}$, $\mathcal{I}_k = \Omega \setminus \mathcal{A}_k$.*

(iii) *Solve*

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

and set $y^{k+1} = E_{\mathcal{I}_k} y_{\mathcal{I}_k} + E_{\mathcal{A}_k}\psi_{\mathcal{A}_k}$.

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

Theorem 4.6. *Assume that (H1), (H2) hold and that ψ and f are in $L^q(\Omega)$. Then the primal-dual active set strategy or equivalently the semismooth Newton method converge superlinearly if $\|y^0 - y^*\|$ is sufficiently small and $\lambda^0 = \beta(y^0 - \psi)$.*

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

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

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

To prove convergence of the reduced algorithm we utilize Theorem 3.2 with $F : L^2(\Omega) \to L^2(\Omega)$ given by $F(y) = \beta y - \beta\psi + \max(0, Cy - f + \beta\psi)$. From Proposition 4.5(ii) it follows that F is slantly differentiable. In fact, the relevant difference quotient for the nonlinear term in F is

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

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

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

so that in particular δ of (36) was set equal to 1 which corresponds to the '$\leq$' sign in the definition of $\mathcal{I}_k$. A slanting function G_F of F at y in direction h is therefore given by

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

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

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

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

Let us also comment on the discretized version of (42). To be specific we consider a two dimensional domain Ω endowed with a uniform rectangular grid, with Δ_h denoting the five-point-star discretization of Δ, and functions z, ψ, y, u discretized by means of grid functions at the nodal points. Numerical results including convergence considerations for this case were reported in [**3**] and [**2**]. Let us consider to which extent Theorems 4.2–4.4 provide new insight on confirming convergence, which was observed numerically in practically all examples. Theorem 4.2 is not applicable since $A_h = \beta I + \Delta_h^{-2}$ is not an M-Matrix. Theorem 4.4 is applicable with $M = \beta I$ and $K = \Delta_h^{-2}$, and asserts convergence if β is sufficiently large. We also tested numerically the applicability of Theorem 4.3 and found that for $\Omega = (0, 1)^2$ the norm condition was satisfied in all cases we tested with grid-size $h \in [10^{-2}, 10^{-1}]$

and $\beta \geq 10^{-4}$, whereas the cone condition $\sum_{i\in\mathcal{I}}(A_{\mathcal{I}}^{-1}y_{\mathcal{I}})_i \geq 0$ for $y_{\mathcal{I}} \geq 0$ was satisfied only for $\beta \geq 10^{-2}$, for the same range of grid-sizes. Still the function $y^k \to \mathcal{M}(y^k)$ utilized in the proof of Theorem 4.4 behaved as a merit function for the wider range of $\beta \geq 10^{-3}$. Note that the norm and cone condition of Theorem 4.4 only involve the system matrix A, whereas $\mathcal{M}(y^k)$ also depends on the specific choice of f and ψ.

REMARK 4.3. *Throughout the paper we used the function $\mathcal{C}$ defined in (20) as a complementarity function. Another popular choice of complementarity function is given by the Fischer-Burmeister function*

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

Note that $\mathcal{C}_{FB}(0,\lambda) = \sqrt{\lambda^2}-\lambda = 2\max(0,-\lambda)$, and hence by Proposition 4.5 the natural choices for generalized derivatives do not satisfy property (17).

REMARK 4.4. *Condition (H2) can be considered as yet another incidence, where a two norm concept for the analysis of optimal control problems is essential. It utilizes the fact that the control-to-solution mapping of the differential equation is a smoothing operation. Two norm concepts where used for second order sufficient optimality conditions and the analysis of SQP-methods in* [**24, 19**]*, for example, and also for semi-smooth Newton methods in* [**36**].

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

EXAMPLE 4.2. *We consider Example 4.1 with the specific choices*

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

A finite difference based discretization of (42) *with a uniform grid of mesh size $h = \frac{1}{100}$ and the standard five point star discretization of the Laplace operator was used. The primal-dual active set strategy with initialization given by solving the unconstrained problem and setting $\lambda_h^0 = 0$, was used. The exact discretized solution $(u_h^*, \lambda_h^*, y_h^*)$ was attained in 8 iterations. In Table 1 we present the values for*

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

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

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

Table 1.

CHAPTER 5

Moreau-Yosida path-following for problems with low multiplier regularity

In the late 1980s and early 1990s a number of research efforts focused on the existence of Lagrange multipliers for pointwise state constraints in optimal control of partial differential equations (PDEs); see, for instance, [**7**] in the case of zero-order state constraints, i.e. $\varphi \leq y \leq \psi$, and [**8**] for constraints on the gradient of y such as $|\nabla y| \leq \psi$, as well as the references therein. Here, y denotes the state of an underlying (system of) partial differential equation(s) and φ, ψ represent suitably chosen bounds. While [**7, 8**] focus on second order linear elliptic differential equations and tracking-type objective functionals, subsequent work such as, e.g., [**30, 31**] considered parabolic PDEs and/or various types of nonlinearities. Moreover, investigations of second order optimality conditions in the presence of pointwise state constraints can be found in [**32**] and the references therein. In many of these papers, for guaranteeing the existence of multipliers it is common to rely on the Slater constraint qualification, which requires that the feasible set contains an interior point.

Concerning the development of numerical solution algorithms for PDE-constrained optimization problems subject to pointwise state constraints significant advances were obtained only in comparatively recent work. In [**21, 14, 15**], for instance, Moreau-Yosida-based inexact primal-dual path-following techniques are proposed and analysed, and in [**25, 28, 35**] Lavrentiev-regularization is considered which replaces $y \leq \psi$ by the mixed constrained $\epsilon u + y \leq \psi$ with u denoting the control variable and $\epsilon > 0$ a small regularization parameter. In [**16, 17**] a technique based on shape sensitivity and level set methods is introduced. These works do not consider the case of combined control and state constraints and the case of pointwise constraints on the gradient of the state. Concerning the optimal control of ordinary differential equations with control as well as state constraints we mention [**6, 23**] and references given there. Control problems governed by PDEs with states and controls subject to pointwise constraints can be found, e.g., in [**1, 9, 22, 27**] and the refreneces therein.

In the present paper we investigate the case where point-wise constraints on the control and the state variable appear simultaneously and independently, i.e. not linked as in the mixed case, which implies a certain extra regularity of the Lagrange multipliers. First- and second-order state constraints are admitted. To obtain efficient numerical methods, regularization of the state-constraints is required. Here we investigate the Moreau-Yosida technique which turns out to be very flexible with respect to various types of pointwise state constraints and can combined with pointwise constraints on the control variable, which need not be regularized. This flexibility makes it an ideal candidate for a unifying approach to a wide range of PDE-constrained minimization problems subject to pointwise constraints of controls and states with respect to both, the proof of existence of Lagrange multipliers and the design of algorithms. Concerning the latter we show in this paper that for the numerical solution of the associated subproblems semismooth Newton solvers are available which allow a function space analysis and converge locally at a q-superlinear rate. In addition, the path-following technique of [**14**] (see also [**15**]) provides an update tool for the regularization parameter leading to efficient inexact path-following iterations. Further, for the proof of existence of multipliers the Moreau-Yosida approach is based on a constraint qualification which is weaker than the usually invoked Slater condition. In [**27**] such a condition is used for point-wise zero-order state constraints.

The remainder of the paper is organized as follows. In section 1 we introduce the underlying rather general problem class, a constraint qualification of range-space-type and the Moreau-Yosida regularized problem. Moreover, the existence of multipliers for the unregularized problem is guaranteed and an associated first order optimality characterization is derived. Section 2 is concerned with the semismooth Newton method for solving the regularized problems. It turns out that for a certain subclass of the underlying general problem a lifting step is necessary in order to bridge an L^2-L^r-norm gap with $r > 2$. The gap occurs due to the fact that the natural function space for the regularization is L^2 whereas the existence of multipliers requires L^r-regularity of the associated control variable. Here "lifting" refers to the fact that the standard semismooth Newton iteration has to be equipped with an additional step lifting the Newton updated from L^2 to L^r; see [**36**] for a related concept. Lifting, in the context of the present paper is used for pointwise zero-order state constraints if the spatial dimension is larger than three, and for pointwise constraints on the gradient. Section 3 ends the paper by a report on numerical test. The appendix contains a chain rule result for the composition of two Newton differentiable functions, which is of interest in its own right.

1. Moreau-Yosida regularization and first order optimality

In this section we derive, in a rather general setting and under a weak constraint qualification, first order optimality conditions for the problem

$$\text{(P)} \qquad \begin{cases} \text{minimize } J(y,u) = J_1(y) + \frac{\alpha}{2}\,|u - u_d|^2_{L^2(\tilde{\Omega})} \\ \text{subject to } Ay = E_{\tilde{\Omega}}u, \quad u \in C_u, \quad y \in C_y, \end{cases}$$

where the control domain $\tilde{\Omega}$ is an open subset of Ω, and the constraints on the control variable u and the state variable y are defined by

$$C_u = \{u \in L^2(\tilde{\Omega}) : \underline{\varphi} \le u \le \bar{\varphi} \text{ a.e. in } \tilde{\Omega}\}, \quad C_y = \{y \in W : |Gy| \le \psi \text{ in } \Omega\}.$$

Here $A \in \mathcal{L}(W, L)$ with W and L reflexive Banach spaces of functions defined on the bounded domain $\Omega \subset \mathbb{R}^d$, satisfying $L^r(\Omega) \subset L$, with dense embedding, $2 \le r < \infty$ and

$$\langle v_1, v_2 \rangle_{L^*, L} = (v_1, v_2)_{L^{r'}(\Omega), L^r(\Omega)} \text{ for all } v_1 \in L^*, \; v_2 \in L^r(\Omega), \tag{49}$$

with $\frac{1}{r} + \frac{1}{r'} = 1$. Further $E_{\tilde{\Omega}} : L^r(\tilde{\Omega}) \to L^r(\Omega)$ is the extension-by-zero operator with adjoint $E^*_{\tilde{\Omega}}$, the restriction to $\tilde{\Omega}$ operator. The quantifiers characterising the constraint sets C_u and C_y satisfy $G \in \mathcal{L}(W, \mathcal{C}(\bar{\Omega})^l)$ for some $1 \le l \le d$,

$$\underline{\varphi}, \, \bar{\varphi} \in L^{2(r-1)}(\tilde{\Omega}), \text{ and } \psi \in \mathcal{C}(\bar{\Omega}), \;\; 0 < \underline{\psi} \le \psi, \;\text{ for some } \underline{\psi} \in \mathbb{R}, \tag{50}$$

$|\cdot|$ denotes the Euclidean-norm in $\mathbb{R}^l$ and the inequalities are interpreted in the pointwise almost everywhere (a.e.) sense. The minor extra regularity that is assumed by requiring that $\underline{\varphi}, \bar{\varphi} \in L^{2(r-1)}(\tilde{\Omega})$ rather than $\underline{\varphi}, \bar{\varphi} \in L^2(\tilde{\Omega})$ will be used in two ways: First the intermediate extra regularity $\underline{\varphi}, \bar{\varphi} \in L^r(\tilde{\Omega})$ is used for the sake of consistency with assumption (H4) below and, secondly, the $L^{2(r-1)}(\tilde{\Omega})$ bound on the admissible controls will be required for passing to the limit in a sequence of approximating problems to (P) in Theorem 5.1 below.

The cost-functional is supposed to satisfy

$$\begin{aligned} &J_1 \in \mathcal{C}^{1,1}(W, \mathbb{R}) \text{ is convex and } \; y_n \rightharpoonup y \text{ in } W \text{ implies that} \\ &J_1(y_n) \to J_1(y) \text{ and } J_1'(y_n) \rightharpoonup J_1'(y) \text{ in } W^*. \end{aligned} \tag{51}$$

Here and below '$\to$' and '$\rightharpoonup$' indicate strong and weak convergence, respectively. Moreover we fix

$$\alpha > 0 \text{ and } u_d \in L^2(\tilde{\Omega}). \tag{52}$$

In addition to the above technical assumptions we require the following hypotheses:

(H1) There exists a feasible point for the constraints in (P).

$$A : W \to L \text{ is a homeomorphism.} \tag{H2}$$

$$G : W \to \mathcal{C}(\bar{\Omega})^l \text{ is a compact linear operator.} \tag{H3}$$

(H4)

$$\left\{ \begin{aligned} &\text{There exists a bounded set } M \subset C_y \times C_u \subset W \times L^r(\tilde{\Omega}) \text{ such that} \\ &0 \in \operatorname{int}\{Ay - E_{\tilde{\Omega}} u : (y, u) \in M\} \subset L^r(\Omega), \text{where the interior is taken} \\ &\text{with respect to } L^r(\Omega). \end{aligned} \right.$$

Conditions (H1) and (H2) are needed for existence of a solution to (P) and the hypotheses (H3)–(H4) are used to establish an optimality system. In particular, (H4) guarantees the existence of a Lagrange multiplier, or an adjoint state, associated with the equality constraint. Condition (H4) is weaker than Slater-type conditions. This can be argued as in [**27**, pp. 113-122]; see also [**33**]. In fact, let $(\bar{y}, \bar{u}) \in W \times L^r(\Omega)$ satisfy

$$A\,\bar{y} = E_{\tilde{\Omega}}\,\bar{u}, \quad \underline{\varphi} \le \bar{u} \le \bar{\varphi}, \quad |G\bar{y}(x)| < \psi(x) \text{ for all } x \in \bar{\Omega}. \tag{53}$$

For $\rho > 0$ and $|\eta|_{L^r(\Omega)} \le \rho$ let y_η denote the solution to

$$A\,y_\eta = E_{\tilde{\Omega}}\bar{u} + \eta.$$

Then

$$|y_\eta - \bar{y}|_W \le \rho\, \|A^{-1}\|_{\mathcal{L}(L^r(\Omega),W)}.$$

Hence, if ρ is sufficiently small, then $y_\eta \in C_y$ and the set

$$M = \{(y_\eta, \bar{u}) : \eta \in L^r(\Omega), |\eta|_{L^r(\Omega)} \le \rho\}$$

serves the purpose required by condition (H4). Differently from the Slater condition (53) condition (H4) operates in the range space of the operator A. Note also that when arguing that (53) implies (H4) the freedom to vary $u \in C_u$ was not used. For the analysis of the proposed algorithm it will be convenient to introduce the operator $B = GA^{-1}E_{\tilde{\Omega}}$. Conditions (H2) and (H3) imply that $B \in \mathcal{L}(L^r(\tilde{\Omega}), \mathcal{C}(\bar{\Omega})^l)$. The compactness assumption in (H3) is needed to pass to the limit in an appropriately defined approximation to problem (P).

To argue existence for (P), note that any minimizing sequence $\{(u_n, y(u_n))\}$ is bounded in $L^r(\tilde{\Omega}) \times W$ by the properties of C_u and (H2). The properties of C_u as well as C_y, strict convexity of J together with a subsequential limit argument guarantee the existence of a unique solution (y^*, u^*) of (P).

More general state constraints of the form

$$C_y = \{\tilde{y} \in W : \, |(G\tilde{y})(\mathrm{x}) - g(\mathrm{x})| \le \psi \text{ for all } \mathrm{x} \in \bar{\Omega}\}$$

for some $g \in \mathcal{C}(\bar{\Omega})^l$ can be treated as well. In fact, if there exists $\tilde{y}_g \in W$ with $Gy_g = g$ and $A\,y_g \in L^r(\Omega)$, then the shift $y := \tilde{y} - y_g$ brings us back to the framework considered in (P) with a state equation of the form $A\,y = u - A\,y_g$, i.e., an affine term must be admitted and in (H4) the expression $Ay - E_{\tilde{\Omega}}u$ must be replaced by $Ay - E_{\tilde{\Omega}}u - Ay_g$.

Before we establish first order optimality, let us mention two problem classes which are covered by our definition of C_y in (P):

Example 5.1 (Pointwise zero-order state constraints). Let A denote the second order linear elliptic partial differential operator

$$Ay = -\sum_{i,j=1}^{d} \partial_{\mathrm{x}_j}(a_{ij}\partial_{\mathrm{x}_i} y) + a_0 y$$

with $C^{0,\delta}(\bar{\Omega})$-coefficients a_{ij}, $i,j = 1,\dots,d$, for some $\delta \in (0,1]$, which satisfy $\sum_{i,j=1}^{d} a_{ij}(\mathrm{x})\xi_i\xi_j \geq \kappa|\xi|^2$ for almost all $\mathrm{x} \in \Omega$ and for all $\xi \in \mathbb{R}^d$, and $a_0 \in L^\infty(\Omega)$ with $a_0 \geq 0$ a.e. in Ω. Here we have $\kappa > 0$. The domain Ω is assumed to be either polyhedral and convex or to have a $\mathcal{C}^{1,\delta}$-boundary Γ and to be locally on one side of Γ. We choose $W = W_0^{1,p}(\Omega)$, $L = W^{-1,p}(\Omega), p > d$, and $G = \mathrm{id}$, which implies $l = 1$. Then

$$|Gy| \leq \psi \text{ in } \Omega \quad \Longleftrightarrow \quad -\psi \leq y \leq \psi \text{ in } \Omega,$$

which is the case of zero order pointwise state constraints. Since $p > d$, condition (H3) is satisfied. Moreover, $A : W \to W^{-1,p}(\Omega)$ is a homeomorphism [**34**, p.179] so that in particular (H2) holds. Moreover there exists a constant C such that

$$|u|_L \leq C|u|_{L^2(\tilde{\Omega})}, \text{ for all } u \in L^2(\tilde{\Omega}),$$

provided that $2 \leq \frac{dp}{dp-d-p}$. Consequently we can take $r = 2$. Here we use the fact that $W^{1,\frac{p}{p-1}}(\Omega)$ embeds continuously into $L^2(\Omega)$, provided that $2 \leq \frac{dp}{dp-d-p}$, and hence $L^2(\Omega) \subset W^{-1,p}(\Omega)$. Note that $2 \leq \frac{dp}{dp-d-p}$ combined with $d < p$ can only hold for $d \leq 3$.

In case Γ is sufficiently regular so that A is a homeomorphism from $H^2(\Omega) \cap H_0^1(\Omega) \to L^2(\Omega)$, we can take $W = H^2(\Omega) \cap H_0^1(\Omega)$, $L = L^2(\Omega)$ and $r = 2$. In this case again (H2) and (H3) are satisfied if $d \leq 3$. □

Example 5.2 (Pointwise first-order state constraints). Let A be as in (i) but with $C^{0,1}(\bar{\Omega})$-coefficients a_{ij}, and let Ω have a $\mathcal{C}^{1,1}$-boundary. Choose $W = W^{2,r}(\Omega) \cap W_0^{1,r}(\Omega), L = L^r(\Omega)$, $r > d$ and, for example, $G = \nabla$, which yields $l = d$. Then

$$C_y = \{y \in W : |\nabla y(\mathrm{x})| \leq \psi(\mathrm{x}) \text{ for all } \mathrm{x} \in \bar{\Omega}\},$$

and (H2) and (H3) are satisfied due to the compact embedding of $W^{2,r}(\Omega)$ into $C^1(\bar{\Omega})$ if $r > d$.

An alternative treatment of pointwise first-order state constraints can be found in [**8**].

If, on the other hand, $G = \mathrm{id}$, as in Example 1, then it suffices to choose $r \geq \max(\frac{d}{2}, 2)$ for (H2) and (H3) to hold. □

We emphasize here that our notion of zero- and first-order state constraints does not correspond to the concept used in optimal control of ordinary differential equations. Rather, it refers to the order of the derivatives involved in the pointwise state constraints.

For deriving a first order optimality system for (P) we introduce the regularized problem

$$(\mathrm{P}_\gamma) \qquad \begin{cases} \text{minimize } J_1(y) + \frac{\alpha}{2}\,|u-u_d|^2_{L^2(\tilde{\Omega})} + \frac{\gamma}{2}\,|(|Gy|-\psi)^+|^2_{L^2(\Omega)} \\ \text{subject to } Ay = E_{\tilde{\Omega}}u, \ u \in C_u, \end{cases}$$

where $\gamma > 0$ and $(\cdot)^+ = \max(0, \cdot)$ in the pointwise almost everywhere sense. In the following sections we shall see that (P_γ) is also useful for devising efficient numerical solution algorithms.

Let $(y_\gamma, u_\gamma) \in W \times L^r(\tilde{\Omega})$ denote the unique solution of (P_γ). Utilizing standard surjectivity techniques and (H2) we can argue the existence of Lagrange multipliers

$$(p_\gamma, \bar{\mu}_\gamma, \underline{\mu}_\gamma) \in L^* \times L^{r'}(\tilde{\Omega}) \times L^{r'}(\tilde{\Omega}),$$

with $\frac{1}{r} + \frac{1}{r'} = 1$, such that

$$(\mathrm{OS}_\gamma) \qquad \begin{cases} Ay_\gamma = E_{\tilde{\Omega}} u_\gamma, \\ A^* p_\gamma + G^* \lambda_\gamma = -J_1'(y_\gamma), \\ \alpha(u_\gamma - u_d) - E_{\tilde{\Omega}}^* p_\gamma + \bar{\mu}_\gamma - \underline{\mu}_\gamma = 0, \\ \bar{\mu}_\gamma \geq 0, \quad u_\gamma \leq \bar{\varphi}, \quad \bar{\mu}_\gamma (u_\gamma - \bar{\varphi}) = 0, \\ \underline{\mu}_\gamma \leq 0, \quad u_\gamma \geq \underline{\varphi}, \quad \underline{\mu}_\gamma (u_\gamma - \underline{\varphi}) = 0, \\ \lambda_\gamma = \gamma(|Gy_\gamma| - \psi)^+ q_\gamma, \\ q_\gamma(\mathrm{x}) \in \begin{cases} \left\{ \frac{Gy_\gamma}{|Gy_\gamma|}(\mathrm{x}) \right\} & \text{if } |Gy_\gamma(\mathrm{x})| > 0, \\ \bar{B}(0,1)^l & \text{else,} \end{cases} \end{cases}$$

where $\bar{B}(0,1)^l$ denotes the closed unit ball in $\mathbb{R}^l$. Above, $A^* \in \mathcal{L}(L^*, W^*)$ is the adjoint of A and G^* denotes the adjoint of G as operator in $\mathcal{L}(W, L^2(\Omega)^l)$. Note that the expression for λ_γ needs to be interpreted pointwise for every $\mathrm{x} \in \Omega$ and $\lambda_\gamma(\mathrm{x}) = 0$ if $|(Gy_\gamma)(\mathrm{x})| - \psi(\mathrm{x}) \leq 0$. In particular this implies that λ_γ is uniquely defined for every $\mathrm{x} \in \Omega$. Moreover, we have $\lambda_\gamma \in L^2(\Omega)^l$, in fact $\lambda_\gamma \in \mathcal{C}(\Omega)^l$. The adjoint equation, which is the second equation in (P_γ), must be interpreted as

$$(54) \qquad \langle p_\gamma, Av \rangle_{L^*,L} + (\lambda_\gamma,\, Gv)_{L^2(\Omega)} = -\langle J_1'(y_\gamma),\, v \rangle_{W^*,W} \text{ for any } v \in W,$$

i.e., in the very weak sense. For later use we introduce the scalar factor of λ_γ defined by

$$(55) \qquad \lambda_\gamma^s := \frac{\gamma}{|Gy_\gamma|}\, (|Gy_\gamma| - \psi)^+ \text{ on } \{|Gy_\gamma| > 0\} \quad \text{and} \quad \lambda_\gamma^s := 0 \text{ else.}$$

This implies that

$$\lambda_\gamma = \lambda_\gamma^s\, G\, y_\gamma.$$

The boundedness of the primal and dual variables is established next.

Lemma 5.1. *Let* (H1)–(H4) *hold. Then the family*

$$\{(y_\gamma,\, u_\gamma,\, p_\gamma,\, \bar{\mu}_\gamma - \underline{\mu}_\gamma,\, \lambda_\gamma^s)\}_{\gamma \geq 1}$$

is bounded in $W \times L^r(\tilde{\Omega}) \times L^{r'}(\Omega) \times L^{r'}(\tilde{\Omega}) \times L^1(\Omega)$.

Proof. Since $u_\gamma \in C_u$ for all $\gamma \geq 1$ we have by (H2) that

$$\{(y_\gamma, u_\gamma)\}_{\gamma\geq 1} \text{ is bounded in } W \times L^r(\tilde\Omega).$$

By (H3) the family $\{Gy_\gamma\}_{\gamma\geq 1}$ is bounded in $\mathcal{C}(\bar\Omega)^l$ as well. Henceforth let C denote a generic constant independent of $\gamma \geq 1$. Let $(y,u) \in M$ be arbitrary. By (OS$_\gamma$)

$$\langle p_\gamma,\, A(y_\gamma - y)\rangle_{L^*,L} + (\lambda_\gamma,\ G(y_\gamma - y))_{L^2(\Omega)} = -\langle J_1'(y_\gamma),\, y_\gamma - y\rangle_{W^*,W} \tag{56}$$

and

$$\begin{aligned}(\lambda_\gamma, G(y_\gamma - y))_{L^2(\Omega)} &= \gamma((|Gy_\gamma| - \psi)^+\, q_\gamma,\ Gy_\gamma - Gy)_{L^2(\Omega)}\\ &= \gamma\int_\Omega (|Gy_\gamma| - \psi)^+(|Gy_\gamma| - \psi) + \gamma\int_\Omega (|Gy_\gamma| - \psi)^+(\psi - q_\gamma\cdot Gy)\\ &\geq \gamma\int_\Omega \left|(|Gy_\gamma| - \psi)^+\right|^2 + \gamma\int_\Omega (|Gy_\gamma| - \psi)^+(\psi - |Gy|)\\ &\geq \gamma\left|(|Gy_\gamma| - \psi)^+\right|^2_{L^2(\Omega)}.\end{aligned}$$

Therefore

$$\langle p_\gamma,\, A(y_\gamma - y)\rangle_{L^*,L} + \gamma\left|(|Gy_\gamma| - \psi)^+\right|^2_{L^2(\Omega)} \leq -\langle J_1'(y_\gamma),\, y_\gamma - y\rangle_{W^*,W}$$

and

$$\begin{aligned}&\langle p_\gamma, -Ay + E_{\tilde\Omega}u\rangle_{L^*,L} + \gamma\left|(|Gy_\gamma| - \psi)^+\right|^2_{L^2(\Omega)}\\ &\qquad\leq -\langle J_1'(y_\gamma),\, y_\gamma - y\rangle_{W^*,W} + (p_\gamma, E_{\tilde\Omega}(u - u_\gamma))_{L^{r'}(\Omega),L^r(\Omega)}.\end{aligned}\tag{57}$$

The first term on the right hand side is bounded since $\{y_\gamma\}_{\gamma\geq 1}$ is bounded in W and $J_1 \in \mathcal{C}^{1,1}(W,\mathbb{R})$, and the second term satisfies

$$\begin{aligned}&\big(p_\gamma, E_{\tilde\Omega}(u - u_\gamma)\big)_{L^{r'}(\Omega),L^r(\Omega)} = \big(\alpha(u_\gamma - u_d) + \bar\mu_\gamma - \underline{\mu}_\gamma,\ u - u_\gamma\big)_{L^{r'}(\tilde\Omega),L^r(\tilde\Omega)}\\ &\qquad\leq C + (\bar\mu_\gamma,\ u - \bar\varphi + \bar\varphi - u_\gamma)_{L^2(\tilde\Omega)} - (\underline{\mu}_\gamma,\ u - \underline{\varphi} + \underline{\varphi} - u_\gamma)_{L^2(\tilde\Omega)} \leq C.\end{aligned}$$

Inserting these estimates into (57) and utilizing (H4) and (49) we have the existence of a constant C independent of γ such that

$$\{|p_\gamma|_{L^{r'}(\Omega)}\}_{\gamma\geq 1} \text{ is bounded.}$$

Integrating the third equation of (OS$_\gamma$) over $\{\mathrm{x} : \bar\mu_\gamma(\mathrm{x}) > 0\}$ we deduce that $\{\bar\mu_\gamma\}_{\gamma\geq 1}$ is bounded in $L^{r'}(\tilde\Omega)$. Similarly $\{\underline{\mu}_\gamma\}_{\gamma\geq 1}$ is bounded in $L^{r'}(\tilde\Omega)$. Finally we turn to estimate the scalar factor λ^s_γ. We have

$$\begin{aligned}\int_\Omega \lambda^s_\gamma &= \int_\Omega \frac{\gamma}{|Gy_\gamma|}\left(|Gy_\gamma| - \psi\right)^+ \leq \int_\Omega \frac{\gamma}{\underline{\psi}^2}\left(|Gy_\gamma| - \psi\right)^+|Gy_\gamma|\\ &= -\frac{1}{\underline{\psi}^2}(p_\gamma,\, Ay_\gamma)_{L^{r'}(\Omega),L^r(\Omega)} - \frac{1}{\underline{\psi}^2}\langle J'(y_\gamma),\, y_\gamma\rangle_{W^*,W} \leq C,\end{aligned}$$

where we used that $\lambda^s_\gamma = 0$ on $\{|Gy_\gamma| \leq \underline{\psi}\}$ and $|Ay_\gamma|_{L^r(\Omega)} = |E_{\tilde\Omega}u_\gamma|_{L^r(\Omega)} \leq C$. Hence, $\{\lambda^s_\gamma\}_{\gamma\geq 1}$ is bounded in $L^1(\Omega)$. □

The preceding lemma implies that there exists

$$(y_*, u_*, p_*, \bar\mu_*, \underline{\mu}_*, \lambda^s_*) \in W \times L^r(\tilde\Omega) \times L^{r'}(\tilde\Omega) \times L^{r'}(\tilde\Omega) \times L^{r'}(\tilde\Omega) \times \mathcal{M}(\bar\Omega) =: X,$$

where $\mathcal{M}(\bar\Omega)$ are the regular Borel measures on $\bar\Omega$, such that on subsequence

$$(y_\gamma, u_\gamma, p_\gamma, \bar\mu_\gamma, \underline{\mu}_\gamma, \lambda^s_\gamma) \rightharpoonup (y_*, u_*, p_*, \bar\mu_*, \underline{\mu}_*, \lambda^s_*) \text{ in } X,$$

which, for the λ^s-component, means that

$$\int_\Omega \lambda^s_\gamma v \to \int_\Omega \lambda^s_* v \quad \text{for all} \quad v \in \mathcal{C}(\bar\Omega).$$

By (H3) this implies that

$$Gy_\gamma \to Gy_* \text{ in } \mathcal{C}(\bar\Omega)^l$$

and hence

$$\int_\Omega \lambda_\gamma v \to \langle \lambda^s_*(Gy_*), v\rangle_{\mathcal{M},\mathcal{C}(\bar\Omega)^l} \text{ for all } v \in \mathcal{C}^l(\bar\Omega)^l.$$

Passing to the limit in the second equation of (OS_γ) we find

$$(p_*,\, Av)_{L^{r'}(\Omega),L^r(\Omega)} = -\langle \lambda^s_*(Gy_*), Gv\rangle_{\mathcal{M}^l(\bar\Omega),\mathcal{C}^l(\bar\Omega)} - \langle J_1'(y_*),\, v\rangle_{W^*,W}$$

for all $v \in D_A$, where $D_A = \{v \in W : Av \in L^r(\Omega)\}$ and $\langle\cdot,\cdot\rangle_{\mathcal{M}^l(\bar\Omega),\mathcal{C}^l(\bar\Omega)}$ denotes the duality pairing between $\mathcal{C}(\bar\Omega)^l$ and $\mathcal{M}(\bar\Omega)^l$, the space of regular vector-valued Borel-measures on $\bar\Omega$. Since D_A is dense in W and since the right hand side uniquely defines a continuous linear functional on W a density argument implies that p_* can be uniquely extended to an element in L^*, and the left hand side can be replaced by $\langle p_*, Av\rangle_{L^*,L}$, for all $v \in W$.

We can now pass to the limit in the first three equations of (OS_γ) to obtain

$$Ay_* = E_{\tilde\Omega} u_* \quad \text{in } L^{r'}(\Omega), \tag{58}$$

$$\langle p_*,\, Av\rangle_{L^*,L} + \langle \lambda^s_*(Gy_*), Gv\rangle_{\mathcal{M}^l(\bar\Omega),\mathcal{C}^l(\bar\Omega)} = -\langle J_1'(y_*),\, v\rangle_{W^*,W} \text{ for all } v \in W, \tag{59}$$

$$\alpha(u_* - u_d) - E^*_{\tilde\Omega} p_* + (\bar\mu_* - \underline{\mu}_*) = 0 \quad \text{in } L^{r'}(\Omega). \tag{60}$$

Standard arguments yield

$$\underline{\mu}_* \geq 0,\ \bar\mu_* \geq 0,\ \underline{\varphi} \leq u_* \leq \bar\varphi \quad \text{a.e. in } \tilde\Omega. \tag{61}$$

Note that

$$J_1(y_\gamma) + \frac{\alpha}{2}\,|u_\gamma - u_d|^2_{L^2(\tilde\Omega)} + \frac{\gamma}{2}\,\left|(|Gy_\gamma| - \psi)^+\right|^2_{L^2(\Omega)} \leq J_1(y^*) + \frac{\alpha}{2}\,|u^* - u_d|^2_{L^2(\tilde\Omega)}. \tag{62}$$

This implies that $|Gy_*(\mathrm{x})| \le \psi(\mathrm{x})$ for all x $\in \Omega$, and hence (y_*, u_*) is feasible. Moreover by (51)

$$\begin{aligned} J_1(y_*) + \frac{\alpha}{2}\,|u_* - u_d|^2_{L^2(\tilde\Omega)} & \\ &\le \lim_{\gamma\to\infty} J_1(y_\gamma) + \limsup_{\gamma\to\infty} \frac{\alpha}{2}\,|u_\gamma - u_d|^2_{L^2(\tilde\Omega)} \\ &\le J_1(y^*) + \frac{\alpha}{2}\,|u^* - u_d|^2_{L^2(\tilde\Omega).} \end{aligned} \tag{63}$$

Since (y_*, u_*) is feasible and the solution of (P) is unique, it follows that $(y_*, u_*) = (y^*, u^*)$. Moreover, from (63) and weak lower semi-continuity of norms we have that $\lim_{\gamma\to\infty} u_\gamma = u^*$ in $L^2(\tilde\Omega)$. From $u_\gamma \in C_u$ for all γ, and $u^* \in C_u$, with $C_u \subset L^{2(r-1)}(\tilde\Omega)$ by (50), together with Hölder's inequality we obtain

$$|u_\gamma - u^*|_{L^r(\tilde\Omega)} \le |u_\gamma - u^*|^{1/r}_{L^2(\tilde\Omega)} |u_\gamma - u^*|^{(r-1)/r}_{L^{2(r-1)}(\tilde\Omega)} \le C|u_\gamma - u^*|^{1/r}_{L^2(\tilde\Omega)} \overset{\gamma\to\infty}{\longrightarrow} 0$$

with some positive constant C. This yields the strong convergence of u_γ in $L^r(\tilde\Omega)$.

Complementary slackness of u_*, $\bar\mu_*$ and $\underline{\mu}_*$, i.e.,

$$\bar\mu_*(u_* - \bar\varphi) = 0, \quad \underline{\mu}_*(u_* - \underline{\varphi}) = 0 \quad \text{a.e. in } \tilde\Omega \tag{64}$$

now follows from the forth and fifth equation of (OS_γ), respectively, the weak convergence of $(\bar\mu_\gamma, \underline{\mu}_\gamma)$ to $(\bar\mu_*, \underline{\mu}_*)$ in $L^{r'}(\tilde\Omega)^2$ and the strong convergence of u_γ to u_* in $L^r(\tilde\Omega)$.

Let $y \in C_y$. Then $\int_\Omega \lambda^s_\gamma(|Gy| - \psi) \le 0$ and hence

$$\int_\Omega \lambda^s_*(|G(y)| - \psi) \le 0.$$

Moreover, $\int_\Omega \lambda^s_*\varphi \ge 0$ for all $\varphi \in \mathcal{C}(\bar\Omega)$ with $\varphi \ge 0$.

For every accumulation point λ^s_*, the corresponding adjoint variable p_* and Lagrange multipliers $\bar\mu_*, \bar\mu_*$ are unique. In fact, since $y_* = y^*$ is unique, the difference δp of two accumulation points of p_γ satisfies

$$(\delta p, Av) = 0 \quad \text{for all } v \in W,$$

and since A is a homeomorphism we have that $\delta p = 0$.
From (60) we deduce that

$$\bar\mu_* = (E^*_{\tilde\Omega} p_* - \alpha(u^* - u_d))^+,\ \underline{\mu}_* = (E^*_{\tilde\Omega} p_* - \alpha(u^* - u_d))^-,$$

where $(\cdot)^- = \min(0, \cdot)$ in the pointwise almost everywhere sense.

We summarize our above findings in the following theorem which provides necessary and sufficient first order optimality conditions for (P).

THEOREM 5.1. *Let* (49)–(52) *and* (H1)–(H4) *hold. Then there exists*

$$(p_*, \bar\mu_*, \underline{\mu}_*, \lambda^s_*) \in L^* \times L^{r'}(\tilde\Omega) \times L^{r'}(\tilde\Omega) \times \mathcal{M}(\bar\Omega)$$

such that

$$\begin{array}{rcll} Ay^* & = & E_{\tilde{\Omega}}u^* & \text{in } L^r(\Omega), \\ A^*p_* + G^*(\lambda_*^s G y^*) & = & -J_1'(y^*) & \text{in } W^*, \\ \alpha(u^* - u_d) - E_{\tilde{\Omega}}^* p_* + (\bar{\mu}_* - \underline{\mu}_*) & = & 0 & \text{in } L^{r'}(\tilde{\Omega}), \\ \bar{\mu}_* \geq 0,\ u^* \leq \bar{\varphi},\ \bar{\mu}_*(u^* - \bar{\varphi}) & = & 0 & a.e. \text{ in } \tilde{\Omega}, \\ \underline{\mu}_* \geq 0,\ u^* \geq \underline{\varphi},\ \underline{\mu}_*(u^* - \underline{\varphi}) & = & 0 & a.e. \text{ in } \tilde{\Omega}, \end{array}$$

and $\int_\Omega \lambda_*^s \varphi \geq 0$ *for all* $\varphi \in \mathcal{C}(\bar{\Omega})$ *with* $\varphi \geq 0$. *Further* $(p_\gamma, \bar{\mu}_\gamma, \underline{\mu}_\gamma)$ *converges weakly in* $L^{r'}(\Omega) \times L^{r'}(\tilde{\Omega}) \times L^{r'}(\tilde{\Omega})$ *(along a subsequence) to* $(p_*, \bar{\mu}_*, \underline{\mu}_*)$, $\langle \lambda_\gamma^s, v \rangle \to \langle \lambda_*^s, v \rangle$ *(along a subsequence) for all* $v \in \mathcal{C}(\bar{\Omega})$, *and* $(y_\gamma, u_\gamma) \to (y^*, u^*)$ *strongly in* $W \times L^r(\tilde{\Omega})$ *as* $\gamma \to \infty$.

We briefly revisit the examples 5.1 and 5.2 in order to discuss the structure of the respective adjoint equation.

Example 5.1 (revisited). For the case of pointwise zero-order state constraints with $W = W_0^{1,p}(\Omega)$ the adjoint equation in variational form is given by

$$\langle p_*, Av \rangle_{W_0^{1,p'}(\Omega), W^{-1,p}(\Omega)} + \langle \lambda_*^s y^*, v \rangle_{\mathcal{M}(\bar{\Omega}), \mathcal{C}(\bar{\Omega})} = -\langle J_1'(u^*), v \rangle_{W^*, W}$$

for all $v \in W$.

Example 5.2 (revisited). Returning to pointwise gradient constraints expressed by $|\nabla y(\mathrm{x})| \leq \psi(\mathrm{x})$ the adjoint equation can be expressed as

$$\langle p_*, Av \rangle_{L^{r'}(\Omega), L^r(\Omega)} + \langle \lambda_*^s \nabla y^*, \nabla v \rangle_{\mathcal{M}(\bar{\Omega})^l, \mathcal{C}(\bar{\Omega})^l} = -\langle J_1'(u^*), v \rangle_{W^*, W}$$

for all $v \in W = W^{2,r}(\Omega) \cap W_0^{1,r}(\Omega)$.

Remark 5.1. Condition (H4) is quite general and also allows the case $\psi = 0$ on parts of Ω. Here we only briefly consider a special case of such a situation. Let $\tilde{\Omega} = \Omega, r = 2, L = L^2(\Omega), W = H^2(\Omega) \cap H_0^1(\Omega)$ and $G = I$, i.e. we consider zero-order state constraints without constraints on the controls, in dimensions $d \leq 3$. We assume that $A : W \to L^2(\Omega)$ is a homeomorphism and that (50) is replaced by

$$0 \leq \psi,\ \psi \in \mathcal{C}(\bar{\Omega}). \tag{2.2'}$$

In this case (H1), (H2), and (H4) are trivially satisfied and $C_y = \{y \in W : |y| \le \psi \text{ in } \Omega\}$. The optimality system is given by

$$(\mathrm{OS}'_\gamma) \qquad \begin{cases} Ay_\gamma = u_\gamma, \\ A^* p_\gamma + \lambda_\gamma = -J_1'(y_\gamma), \\ \alpha(u_\gamma - u_d) - p_\gamma = 0, \\ \lambda_\gamma = \gamma(|y_\gamma| - \psi)^+ q_\gamma, \\ q_\gamma(\mathrm{x}) \in \begin{cases} \left\{\frac{y_\gamma}{|y_\gamma|}(\mathrm{x})\right\} & \text{if } |y_\gamma(\mathrm{x})| > 0, \\ \bar{B}(0,1) & \text{else,} \end{cases} \end{cases}$$

with $(y_\gamma, u_\gamma, p_\gamma, \lambda_\gamma) \in W \times L^2(\Omega) \times L^2(\Omega) \times L^2(\Omega)$. As in Lemma 5.1 we argue, using (H4), that $\{(y_\gamma, u_\gamma, p_\gamma)\}_{\gamma \ge 1}$ is bounded in $W \times L^2(\Omega) \times W^*$. Since we do not assume that $\underline{\psi} > 0$ we argue differently than before to obtain a bound on $\{\lambda_\gamma\}_{\gamma \ge 1}$. In fact the second equation in (OS'_γ) implies that $\{\lambda_\gamma\}_{\gamma \ge 1}$ is bounded in W^*. Hence there exists $(y^*, u^*, p_*, \lambda_*) \in W \times L^2(\Omega) \times L^2(\Omega) \times W^*$, such that $(y_\gamma, u_\gamma) \to (y^*, u^*)$ in $W \times L^2(\Omega)$ and $(p_\gamma, \lambda_\gamma) \rightharpoonup (p_*, \lambda_*)$ weakly in $L^2(\Omega) \times W^*$, as $\gamma \to \infty$. Differently from the case with state and control constraints, we have convergence of the whole sequence, rather than subsequential convergence of $(p_\gamma, \lambda_\gamma)$ in this case. In fact, the third equation in (OS'_γ) implies the convergence of p_γ, and the second equation the convergence of λ_γ. Passing to the limit as $\gamma \to \infty$ we obtain from (OS'_γ)

$$(\mathrm{OS}') \qquad \begin{cases} Ay^* & = u^*, \\ A^* p_* + \lambda_* & = -J_1'(y^*), \quad , \\ \alpha(u^* - u_d) - p_* & = 0. \end{cases}$$

and λ_* has the additional properties as the limit of elements λ_γ. For example, if $\psi \ge \underline{\psi} > 0$ on a subset $\hat{\Omega} \subset \Omega$ and y^* is inactive on $\hat{\Omega}$, then $\lambda_* = 0$ as functional on continuous functions with compact support in $\hat{\Omega}$.

2. Semismooth Newton method

As mentioned earlier, (P_γ) is appealing as it can be solved with superlinearly convergent numerical methods. Combined with a suitable update strategy for γ, an overall solution algorithm for (P) is obtained. Here we analyse in detail the superlinear solution process of (P_γ), for a fixed value γ. The constants in this section therefore depend on γ. For the path-following strategy with respect to γ one may proceed as in [**14, 15**].

2.1. Newton differentiability/semismoothness. In [**13**], see also [**10**], for a mapping $F : \mathcal{X} \to \mathcal{Y}$, with $\mathcal{X}$ and $\mathcal{Y}$ Banach spaces, a generalized derivative is introduced in such a way that q-superlinear convergence of the Newton algorithm can be guaranteed without requiring that F is Frechet differentiable. In fact, F is called *Newton* (or *slant) differentiable* in an open set $U \subset \mathcal{X}$ if there exists a family of generalized derivatives $G_F(x) \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, $x \in U$, such that

$$\lim_{|h|_{\mathcal{X}} \to 0} |h|_{\mathcal{X}}^{-1} |F(x+h) - F(x) - G_F(x+h)h|_{\mathcal{Y}} = 0 \quad \text{for every } x \in U. \tag{65}$$

Note that F need not be Frechet-differentiable in order to have the property (65). In general, there exists a set of Newton derivatives at x which becomes a singleton whenever F is Frechet-differentiable at x. We also point out that (65) resembles the concept of *semismoothness* of a mapping which was introduced in [**26**] for scalar-valued functionals on $\mathbb{R}^n$ and extended to the vector-valued case in [**29**]. The concept of semi-smoothness in finite dimensions, however, is linked to Rademacher's theorem, which states, that locally Lipschitz continuous functions are almost everywhere differentiable. This concept is not available in infinite dimensions. But property (65) quantifies one of the essential ingredients for the Newton method to be locally superlinearly convergent. Consequently it is becoming customary now to refer to the Newton method, in infinite dimensions, as a *semismooth Newton method*, if (65) holds. As usual the Newton method for finding $x^* \in \mathcal{X}$ such that $F(x^*) = 0$ consists in the iteration:

Algorithm 8 (**Semismooth Newton method**).

(i) Choose $x^0 \in \mathcal{X}$.
(ii) Unless some stopping rule is satisfied, perform the update step

$$x^{k+1} = x^k - G_F(x^k)^{-1} F(x^k) \quad \text{for } k = 0, 1 \ldots. \tag{66}$$

This iteration is locally q-superlinearliy convergent to x^* within a neighborhood $U(x^*)$, if $x_0 \in U(x^*)$, and (65) as well as

$$\|G_F(x)^{-1}\|_{\mathcal{L}(\mathcal{Y},\mathcal{X})} \leq C, \text{ for a constant } C \text{ independently of } x \in U(x^*), \tag{67}$$

hold, [**10, 13**].

The remainder of this subsection is devoted to the analysis of the semismoothness property (65) of the mapping F_γ, which defines the Newton iteration associated with (OS$_\gamma$). This is done for $\mathcal{X} = \mathcal{Y} = L^r(\tilde{\Omega})$ where the choice of r is dictated by the need that $Gy(u) \in \mathcal{C}(\bar{\Omega})^l$ for $u \in L^r(\tilde{\Omega})$. In the subsequent subsection 3.3 we address (67) in $L^2(\tilde{\Omega})$. Superlinear convergence is investigated in the final subsection. In case $r > 2$ a lifting step is introduced to compensate the fact that (67) is only available in $L^2(\tilde{\Omega})$.

Throughout this section it will be convenient to utilize the operator

$$B\, u = GA^{-1}E_{\tilde{\Omega}} u,$$

which satisfies $B \in \mathcal{L}(L^r(\tilde{\Omega}), \mathcal{C}(\bar{\Omega})^l)$ if (H2) and (H3) are satisfied. In particular $B^* \in \mathcal{L}(L^s(\Omega)^l, L^{r'}(\tilde{\Omega}))$ for every $s \in (1, \infty)$. We shall require the following two additional hypotheses for some $\hat{r} > r$:

$$\text{(H5)} \qquad \begin{aligned} &u_d,\ \bar{\varphi},\ \underline{\varphi} \in L^{\hat{r}}(\tilde{\Omega}), \text{ and } u \mapsto A^{-*}J_1'(A^{-1}E_{\tilde{\Omega}}u) \text{ is continuously} \\ &\text{Frechet differentiable from } L^2(\tilde{\Omega}) \to L^{\hat{r}}(\Omega), \end{aligned}$$

and

$$\text{(H6)} \qquad B^* \in \mathcal{L}(L^r(\tilde{\Omega})^l, L^{\hat{r}}(\tilde{\Omega})) \quad \text{where } \tfrac{1}{\hat{r}} + \tfrac{1}{\hat{r}'} = 1.$$

We interpret the hypotheses (H5) and (H6) in view of the examples 5.1 and 5.2 in section 1.

Example 5.1 (revisited). We have $G = \text{id}$ and, hence, $B = A^{-1}E_{\tilde{\Omega}}$. Note that $A : W_0^{1,r'}(\Omega) \to W^{-1,r'}(\Omega)$ is a homeomorphism. Consequently, $A^{-*} \in \mathcal{L}(W^{-1,r}(\Omega), W^{1,r}(\Omega))$. For every d there exists $\hat{r} > r$ such that $W_0^{1,r}(\Omega)$ embeds continuously into $L^{\hat{r}}(\Omega)$. Therefore $A^{-*} \in \mathcal{L}(L^r(\Omega), L^{\hat{r}}(\Omega))$ and $B^* \in \mathcal{L}(L^r(\tilde{\Omega}), L^{\hat{r}}(\tilde{\Omega}))$. Hence, assumption (H6) is satisfied. The second part of hypothesis (H5) is fulfilled, e.g., for the tracking-type objective functional $J_1(y) = \frac{1}{2}|y - y_d|^2_{L^2(\Omega)}$ with $y_d \in L^2(\Omega)$ given.

Example 5.2 (revisited). Differently to example 5.1 we have $G = \nabla$ and, thus, $B = \nabla A^{-1}E_{\tilde{\Omega}}$. Since $G^* \in \mathcal{L}(L^r(\Omega), W^{-1,r}(\Omega))$ we have $A^{-*}G^* \in \mathcal{L}(L^r(\Omega), W^{1,r}(\Omega))$. As in example 5.1 there exists for every d some $\hat{r} > r$ such that $B^* \in E^*_{\tilde{\Omega}}A^{-*}G^* \in \mathcal{L}(L^r(\Omega)^l, L^{\hat{r}}(\Omega))$. For J_1 as in example 5.1 above (H5) is satisfied.

Next we note that $\bar{\mu}_\gamma$ and $\underline{\mu}_\gamma$ in (OS$_\gamma$) may be condensed into one multiplier $\mu_\gamma := \bar{\mu}_\gamma - \underline{\mu}_\gamma$. Then the fourth and fifth equation of (OS$_\gamma$) are equivalent to

$$\mu_\gamma = (\mu_\gamma + c(u_\gamma - \bar{\varphi}))^+ + (\mu_\gamma + c(u_\gamma - \underline{\varphi}))^- \tag{68}$$

for some $c > 0$. Fixing $c = \alpha$ and using the third equation in (OS$_\gamma$) results in

$$\alpha(u_\gamma - u_d) - E^*_{\tilde{\Omega}}p_\gamma + (E^*_{\tilde{\Omega}}p_\gamma + \alpha(u_d - \bar{\varphi}))^+ + (E^*_{\tilde{\Omega}}p_\gamma + \alpha(u_d - \underline{\varphi}))^- = 0. \tag{69}$$

Finally, using the state and the adjoint equation to express y_γ and p_γ in terms of u_γ, (OS$_\gamma$) turns out to be equivalent to

$$F_\gamma(u_\gamma) = 0, \quad F_\gamma : L^r(\tilde{\Omega}) \to L^r(\tilde{\Omega}),$$

with

$$F_\gamma(u_\gamma) := \alpha(u_\gamma - u_d) - \hat{p}_\gamma + (\hat{p}_\gamma + \alpha(u_d - \bar{\varphi}))^+ + (\hat{p}_\gamma + \alpha(u_d - \underline{\varphi}))^-. \tag{70}$$

and

$$\hat{p}_\gamma := p_\gamma(u_\gamma) = -\gamma B^*(|B\,u_\gamma| - \psi)^+ q(B\,u_\gamma) - E^*_{\tilde{\Omega}}A^{-*}J_1'(A^{-1}E_{\tilde{\Omega}}u_\gamma),$$

where

$$q(B\,u)(\mathrm{x}) = \begin{cases} \left(\frac{B\,u}{|B\,u|}\right)(\mathrm{x}) & \text{if } |B\,u(\mathrm{x})| > 0, \\ 0 & \text{otherwise.} \end{cases}$$

We further set

$$\mathfrak{p}_\gamma(u) := -\gamma B^*(|Bu| - \psi)^+ q(Bu), \qquad \text{where } \mathfrak{p}_\gamma : L^r(\tilde{\Omega}) \to L^{\hat{r}}(\tilde{\Omega}). \tag{71}$$

For the semismoothness of F_γ we first study the Newton differentiability of $\mathfrak{p}_\gamma(\cdot)$. For its formulation we need

$$G_{\max}(\omega)(\mathrm{x}) := \begin{cases} 1 & \text{if } \omega(\mathrm{x}) > 0, \\ 0 & \text{if } \omega(\mathrm{x}) \leq 0, \end{cases}$$

which was shown in [**13**] to serve as a generalized derivative for $\max(0, \cdot) : L^{s_1}(\Omega) \to L^{s_2}(\Omega)$ if $1 \leq s_2 < s_1 \leq \infty$. An analogous result holds true for $\min(0, \cdot)$. Further the norm-functional $|\cdot| : L^{s_1}(\Omega)^l \to L^{s_2}(\Omega)$, with s_1, s_2 as above, is Newton differentiable with generalized derivative $q(\cdot)$. This follows from Example 8.1 and Theorem 8.1 in [**20**]. There only the case $l = 1$ is treated, but the result can be extended in a straightforward way to $l > 1$.

We define

$$Q(Bv) := |Bv|^{-1} \left(\mathrm{id} - |Bv|^{-2} (Bv)(Bv)^\top \right).$$

Throughout this section, whenever we refer to (H3) it would actually suffice to have $G \in \mathcal{L}(W, \mathcal{C}(\bar{\Omega})^l)$.

Lemma 5.2. *Assume that* (H2)*,* (H3) *and* (H6) *hold true. Then the mapping* $\mathfrak{p}_\gamma : L^r(\tilde{\Omega}) \to L^{\hat{r}}(\tilde{\Omega})$ *is Newton differentiable in a neighborhood of every point* $u \in L^r(\tilde{\Omega})$ *and a generalized derivative is given by*

(72)
$$G_{\mathfrak{p}_\gamma}(u) = -\gamma B^* \left[G_{\max}(|Bu| - \psi) q(Bu) q(Bu)^\top + (|Bu| - \psi)^+ Q(Bu) \right] B.$$

Proof. By (H6) there exists a constant $C_1(\gamma)$ such that

$$\|\gamma B^*\|_{\mathcal{L}(L^r(\Omega),\, L^{\hat{r}}(\tilde{\Omega}))} \leq C_1(\gamma).$$

Let u and $h \in L^r(\tilde{\Omega})$. Then we have by the definition of $\mathfrak{p}_\gamma$ in (71) and the expression for $G_{\mathfrak{p}_\gamma}$ in (72)

$$\begin{aligned}
&|\mathfrak{p}_\gamma(u+h) - \mathfrak{p}_\gamma(u) - G_{\mathfrak{p}_\gamma}(u+h)h|_{L^{\hat{r}}(\tilde{\Omega})} \\
&\quad \leq C_1(\gamma) |\, (|B(u+h)| - \psi)^+\ q(Bu+h) - (|Bu| - \psi)^+\, q(Bu) \\
&\qquad - [G_{\max}(|B(u+h)| - \psi)\, q(B(u+h))\, q(B(u+h))^\top \\
&\qquad + (|B(u+h)| - \psi)^+\, Q(B(u+h))] Bh|_{L^r(\Omega)^l} \\
&\quad \leq C_1(\gamma) |(|B(u+h)| - \psi)^+ \big(q(B(u+h)) - q(Bu) - Q(B(u+h))Bh \big)|_{L^r(\Omega)^l} \\
&\qquad + C_1(\gamma)\, |((|B(u+h)| - \psi)^+ - (|Bu| - \psi)^+) q\,(Bu) - \\
&\qquad\qquad - [G_{\max}(|B(u+h)| - \psi)\, q\,(B(u))\, q\,(B(u+h))^\top] Bh|_{L^r(\Omega)^l} \\
&\qquad + C_1(\gamma) |G_{\max}(|B(u+h)| - \psi) \big(q\,(Bu) - q\,(B(u+h)) \big)\, q\,(B(u+h)^\top Bh|_{L^r(\Omega)} \\
&\qquad = I + II + III.
\end{aligned}$$

We now estimate separately the terms $I - III$. Let

$$S = \left\{ \mathrm{x} : |Bu(\mathrm{x})| \leq \frac{\psi(\mathrm{x})}{2} \right\}.$$

Then there exists $U(u) \subset L^r(\tilde{\Omega})$ and $\delta > 0$ such that

$$|B(u(\mathrm{x}) + h(\mathrm{x}))| \leq \psi(\mathrm{x}), \quad \text{for all} \quad \mathrm{x} \in S, \quad u \in U(u), \quad |h|_{L^r(\tilde{\Omega})} \leq \delta,$$

where we use that $B \in \mathcal{L}(L^r(\tilde{\Omega}),\, \mathcal{C}(\tilde{\Omega})^l)$ due to (H2) and (H3). Consequently

$$I \leq C|q(B(u+h)) - q(Bu) - Q(B(u+h))Bh|_{\mathcal{C}(\overline{\Omega \backslash S})^l},$$

where $C = C(u, \delta)$. We check that $H : v \to \frac{v}{|v|}$ from $\mathcal{C}(\overline{\Omega \backslash S})^l$ to itself is uniformly Fréchet differentiable with Fréchet derivative

$$H'(v) = \frac{1}{|v|}\,(\mathrm{id} - \frac{v\, v^\top}{|v|^2}),$$

provided that $v(\mathrm{x}) \neq 0$ for all $\mathrm{x} \in \overline{\Omega \backslash S}$. Moreover $v \to H'(v)$ is locally Lipschitz continuous from $\mathcal{C}(\overline{\Omega \backslash S})^l$ to $\mathcal{C}(\overline{\Omega \backslash S})^{l \times l}$. Together with $B \in \mathcal{L}(L^r(\Omega),\, \mathcal{C}(\bar{\Omega})^l)$ this implies that

$$I = \mathcal{O}(|h|^2_{L^r(\tilde{\Omega})}), \tag{73}$$

where $\mathcal{O}$ is uniform with respect to $u \in U$. We turn to estimate II and consider $u \to (|Bu| - \psi)^+$ in the neighborhood of U of u. As noted above $G : v \to |v|$ is Newton differentiable from $L^{s_1}(\Omega)^l$ to $L^{s_2}(\Omega)$ if $1 \leq s_2 < s_1 \leq \infty$ at every $v \in L^{s_1}(\Omega)^l$, with a generalized derivative $\frac{v}{|v|}$, if $|v| \neq 0$. This, together with the chain rule for Newton differentiable maps composed with Frechet differentiable maps, see e.g. [**20**], Lemma 8.1 or [**21**], and $B \in \mathcal{L}(L^r(\tilde{\Omega}),\, \mathcal{C}(\bar{\Omega})^l)$ (hence $B \in \mathcal{L}(L^r(\tilde{\Omega}),\, L^{r+2\epsilon}(\Omega)^l)$) implies that $u \to |Bu|$ is Newton differentiable from $L^r(\tilde{\Omega})$ to $L^{r+\varepsilon}(\Omega)$ for some $\varepsilon > 0$, with a generalized derivative given by $q(u)$. Newton differentiability of this mapping also follows from [**36**] Theorem 5.2. The chain rule for two superimposed Newton differentiable maps given in Proposition B.1 implies then that $u \to (|Bu| - \psi)^+$ is Newton differentiable from $L^r(\tilde{\Omega})$ to $L^r(\Omega)$ and hence

$$II = o\,(|h|_{L^r(\tilde{\Omega})}), \tag{74}$$

with o uniform with respect to $u \in U$. It is straightforward to argue that

$$III = \mathcal{O}\,(|h|^2_{L^r(\tilde{\Omega})}), \tag{75}$$

with $\mathcal{O}$ uniform in $u \in U$. Combining (73)–(75) we have shown that

$$|\mathfrak{p}_\gamma(u+h) - \mathfrak{p}_\gamma(u) - G_{\mathfrak{p}_\gamma}(u+h)h|_{L^{\hat{r}}(\tilde{\Omega})} = o\,(|h|_{L^r(\tilde{\Omega})}),$$

as $|h|_{L^r(\tilde{\Omega})}) \to 0$, with o uniform in $u \in U$. Hence, $\mathfrak{p}_\gamma$ is Newton differentiable in the neighborhood U of u. □

Newton differentiability of F_γ is established next.

Theorem 5.2. *Let* (H2), (H3), (H5) *and* (H6) *hold true. Then* $F_\gamma : L^r(\tilde{\Omega}) \to L^r(\tilde{\Omega})$ *is Newton differentiable in a neighborhood of every* $u \in L^r(\tilde{\Omega})$*.*

Proof. We consider the various constituents of F_γ separately. In terms of

$$\hat{p}_\gamma(u) := \mathfrak{p}_\gamma(u) - E^*_{\tilde{\Omega}} A^{-*} J_1'(A^{-1} E_{\tilde{\Omega}} u)$$

we have by (70)

$$F_\gamma(u) = \alpha(u - u_d) - \hat{p}_\gamma(u) + \big(\hat{p}_\gamma(u) + \alpha(u_d - \bar{\varphi})\big)^+ + \big(\hat{p}_\gamma(u) + \alpha(u_d - \underline{\varphi})\big)^-.$$

Lemma 5.2 and (H5) for J_1 yield the Newton differentiability of

$$u \mapsto \alpha(u - u_d) - \hat{p}_\gamma(u) \text{ from } L^r(\tilde{\Omega}) \text{ to } L^r(\tilde{\Omega}),$$

in a neighborhood $U(u)$ of u.

We further have by Lemma 5.2 that

$$\hat{p}_\gamma(\cdot) + \alpha(u_d - \bar{\varphi}) \text{ and } \hat{p}_\gamma(\cdot) + \alpha(u_d - \underline{\varphi})$$

are locally Lipschitz continuous and Newton differentiable from $L^r(\tilde{\Omega})$ to $L^{\hat{r}}(\tilde{\Omega})$, respectively. Then the results of Appendix B yield the Newton differentiability of

$$\big(\hat{p}_\gamma(\cdot) + \alpha(u_d - \bar{\varphi})\big)^+ + \big(\hat{p}_\gamma(\cdot) + \alpha(u_d - \underline{\varphi})\big)^-$$

from $L^r(\tilde{\Omega})$ to $L^r(\tilde{\Omega})$ in a, possibly smaller neighborhood $U(u)$ of u. Combining these results proves the assertion. □

The structure of a particular generalized derivative associated with F_γ immediately follows from a combination of the previous results.

Corollary 5.1. *Let the assumptions of Theorem 5.2 hold. Then a particular generalized derivative of* F_γ *at* $u \in L^r(\tilde{\Omega})$ *is given by*

$$\begin{aligned} G_{F_\gamma}(u) = \alpha\,\mathrm{id} - G_{\hat{p}_\gamma}(u) &+ G_{\max}(\hat{p}_\gamma(u) + \alpha(u_d - \bar{\varphi}))G_{\hat{p}_\gamma}(u) \\ &+ G_{\min}(\hat{p}_\gamma(u) + \alpha(u_d - \underline{\varphi}))G_{\hat{p}_\gamma}(u) \end{aligned}$$

with

$$G_{\hat{p}_\gamma}(u) = G_{\mathfrak{p}_\gamma}(u) - E^*_{\tilde{\Omega}} A^{-*} J_1''(A^{-1} E_{\tilde{\Omega}} u) A^{-1} E_{\tilde{\Omega}}.$$

2.2. Uniform boundedness of the inverse of the generalized derivative in $L^2(\tilde{\Omega})$. Next we study G_{F_γ} in more detail. For a well-defined semismooth Newton step we need its non-singularity on a particular subspace. Given an approximation u^k of u_γ, in our context the semismooth Newton update step is defined as

$$G_{F_\gamma}(u^k)\delta_u^k = -F_\gamma(u^k) \tag{76}$$

with $\delta_u^k = u^{k+1} - u^k$, compare (66) with $x = u$ and $F = F_\gamma$.

For our subsequent investigation we define the active and inactive sets

$$\bar{\mathcal{A}}^k := \{\mathrm{x} \in \tilde{\Omega} : \big(\hat{p}_\gamma(u^k) + \alpha(u_d - \bar{\varphi})\big)(\mathrm{x}) > 0\}, \tag{77}$$

$$\underline{\mathcal{A}}^k := \{\mathrm{x} \in \tilde{\Omega} : \big(\hat{p}_\gamma(u^k) + \alpha(u_d - \underline{\varphi})\big)(\mathrm{x}) < 0\}, \tag{78}$$

$$\mathcal{A}^k := \bar{\mathcal{A}}^k \cup \underline{\mathcal{A}}^k, \tag{79}$$

$$\mathcal{I}^k := \tilde{\Omega} \setminus \mathcal{A}^k. \tag{80}$$

Further we introduce $\chi_{\mathcal{I}^k}$, the characteristic function of the inactive set $\mathcal{I}^k$, and the extension-by-zero operators $E_{\bar{\mathcal{A}}^k}$, $E_{\underline{\mathcal{A}}^k}$, $E_{\mathcal{A}^k}$, and $E_{\mathcal{I}^k}$ with the properties $E_{\mathcal{A}^k}\chi_{\mathcal{I}^k} = 0$ and $E_{\mathcal{I}^k}\chi_{\mathcal{I}^k} = \chi_{\mathcal{I}^k}$.

Corollary 5.1 and the structure of $G_{\max}$ and $G_{\min}$, respectively, yield that

$$G_{F_\gamma}(u^k) = \alpha\,\mathrm{id} - \chi_{\mathcal{I}^k} G_{\hat{p}_\gamma}(u^k).$$

Hence, from the restriction of (76) to $\bar{\mathcal{A}}^k$ we find

$$\delta^k_{u|\bar{\mathcal{A}}^k} = E^*_{\bar{\mathcal{A}}^k}\delta^k_u = E^*_{\bar{\mathcal{A}}^k}(\bar{\varphi} - u^k) = \bar{\varphi}_{|\bar{\mathcal{A}}^k} - u^k_{|\bar{\mathcal{A}}^k} \tag{81}$$

and similarly

$$\delta^k_{u|\underline{\mathcal{A}}^k} = E^*_{\underline{\mathcal{A}}^k}\delta^k_u = E^*_{\underline{\mathcal{A}}^k}(\underline{\varphi} - u^k) = \underline{\varphi}_{|\underline{\mathcal{A}}^k} - u^k_{|\underline{\mathcal{A}}^k}. \tag{82}$$

Hence, $\delta^k_{u|\mathcal{A}^k}$ is obtained by a simple assignment of data according to the previous iterate only. Therefore, it remains to study (76) on the inactive set

$$E^*_{\mathcal{I}^k} G_{F_\gamma}(u^k) E_{\mathcal{I}^k} \delta^{\mathcal{I}^k}_u = -E^*_{\mathcal{I}^k}\big(F_\gamma(u^k) + G_{F_\gamma}(u^k) E_{\mathcal{A}^k}\delta^k_{u|\mathcal{A}^k}\big) \tag{83}$$

as equation in $L^2(\mathcal{I}^k)$.

LEMMA 5.3. *Let* (H2)*,* (H3)*,* (H5) *and* (H6) *hold and* $\mathcal{I} \subset \Omega$*. Then the inverse to the operator*

$$E^*_{\mathcal{I}} G_{F_\gamma}(u) E_{\mathcal{I}} : L^2(\mathcal{I}) \to L^2(\mathcal{I}),$$

with $G_{F_\gamma}(u) = \alpha\,\mathrm{id} - \chi_{\mathcal{I}} G_{\hat{p}_\gamma}(u)$*, exists and is bounded by* $\frac{1}{\alpha}$ *regardless of* $u \in L^r(\tilde{\Omega})$ *as long as* $\mathrm{meas}(\mathcal{I}) > 0$ *.*

PROOF. Note that we have

$$E^*_{\mathcal{I}} G_{F_\gamma}(u) E_{\mathcal{I}} = \alpha\,\mathrm{id}_{|\mathcal{I}} + \gamma E^*_{\mathcal{I}} B^* T(u) B E_{\mathcal{I}} + E^*_{\mathcal{I}} E^*_{\tilde{\Omega}} A^{-*} J_1''(A^{-1}E_{\tilde{\Omega}} u) A^{-1} E_{\tilde{\Omega}} E_{\mathcal{I}}$$

with

$$T(u) = G_{\max}(|Bu| - \psi) q(Bu) q(Bu)^\top + (|Bu| - \psi)^+ Q(Bu).$$

From $B \in \mathcal{L}(L^{\hat{r}'}(\tilde{\Omega}), L^{\hat{r}'}(\Omega)) \cap \mathcal{L}(L^r(\tilde{\Omega}), L^r(\Omega))$, by (H2), (H3) and (H6), we conclude by interpolation that $B \in \mathcal{L}(L^2(\tilde{\Omega}), L^2(\Omega))$. Moreover $T(u) \in \mathcal{L}(L^2(\Omega))$. Therefore

$$\gamma E^*_{\mathcal{I}} B^* T(u) B E_{\mathcal{I}} \in L^2(\tilde{\Omega}) \text{ and } E^*_{\mathcal{I}} E^*_{\tilde{\Omega}} A^{-*} J_1''(A^{-1}E_{\tilde{\Omega}} u) A^{-1} E_{\tilde{\Omega}} E_{\mathcal{I}} \in L^2(\tilde{\Omega}),$$

where we also use (H5). In conclusion, the operator $E_{\mathcal{I}}^{*}G_{F_\gamma}(u)E_{\mathcal{I}}$ is an element of $\mathcal{L}(L^2(\mathcal{I}))$. From the convexity of J we infer for arbitrary $z \in L^r(\mathcal{I})$ that

$$\big((\alpha \operatorname{id}_{|\mathcal{I}} + E_{\mathcal{I}}^{*}E_{\tilde\Omega}^{*}A^{-*}J_1''(A^{-1}E_{\tilde\Omega}u)A^{-1}E_{\tilde\Omega}E_{\mathcal{I}})z,\, z\big)_{L^2(\mathcal{I})} \geq \alpha\|z\|^2_{L^2(\mathcal{I})}. \tag{84}$$

Turning to $\gamma E_{\mathcal{I}}^{*}B^{*}T(u)BE_{\mathcal{I}}$ we observe that $T(u) = 0$ in $\{|Bu| - \psi \leq 0\}$ and $0 < \psi/|Bu| - 1 < 1$ in $\{|Bu| - \psi > 0\}$. Hence,

$$(T(u)w, w)_{L^2(\tilde\Omega)} = \int_{\{|Bu|-\psi>0\}} \Big(1 - \frac{\psi}{|Bu|}\Big)|w|^2 \geq 0$$

for all $w \in L^2(\tilde\Omega)$. From this and (84) we conclude that the inverse to $E_{\mathcal{I}}^{*}G_{F_\gamma}(u)E_{\mathcal{I}} : L^2(\mathcal{I}) \to L^2(\mathcal{I})$ is bounded by $\frac{1}{\alpha}$. □

Proposition 5.1. *If* (H2), (H3), (H5) *and* (H6) *hold, then the semismooth Newton update step* (76) *is well-defined and* $\delta_u^k \in L^r(\tilde\Omega)$.

Proof. Well-posedness of the Newton step with $\delta_u^k \in L^2(\tilde\Omega)$ follows immediately from (81), (82) and Lemma 5.3. Note that whenever $\mathcal{I}^k = \emptyset$, then δ_u^k is fully determined by (81) and (82). An inspection of (81), (82) and (83), using (H5) and the structure of $E_{\mathcal{I}}^{*}G_{F_\gamma}(u)E_{\mathcal{I}}$, moreover shows that $\delta_u^k \in L^r(\tilde\Omega)$. □

From Lemma 5.3 and the proof of Proposition 5.1 we conclude that $E_{\mathcal{I}}^{*}G_{F_\gamma}(u)E_{\mathcal{I}}\, v = f$ is solvable in $L^r(\tilde\Omega)$ if $f \in L^r(\tilde\Omega)$. It is not clear, however, whether $(E_{\mathcal{I}}^{*}G_{F_\gamma}(u)E_{\mathcal{I}})^{-1}$ is bounded as an operator in $\mathcal{L}(L^r(\tilde\Omega))$ uniformly with respect to u.

We are now prepared to consider (67) for G_{F_γ} specified in Corollary 5.1.

Proposition 5.2. *Let* (H2), (H3), (H5) *and* (H6) *hold. Then for each* $\hat u \in L^r(\tilde\Omega)$ *there exists a neighborhood* $U(\hat u) \subset L^r(\tilde\Omega)$ *and a constant* K *such that*

$$\|G_{F_\gamma}(u)^{-1}\|_{\mathcal{L}(L^2(\hat\Omega))} \leq K \textit{ for all } u \in U(\hat u). \tag{85}$$

Proof. Let $\mathcal{A}$ and $\mathcal{I}$ denote disjoint subsets of $\tilde\Omega$ such that $\mathcal{A} \cup \mathcal{I} = \tilde\Omega$. Then observe that every $v \in L^2(\tilde\Omega)$ can be uniquely decomposed in two components $(E_{\mathcal{I}}^{*}v,\ E_{\mathcal{A}}^{*}v)$. For $g \in L^2(\tilde\Omega)$ the equation

$$G_{F_\gamma}(u)v = g$$

is equivalent to

$$\begin{cases} E_{\mathcal{A}}^{*}\, v = E_{\mathcal{A}}^{*}\, g, \\ (E_{\mathcal{I}}^{*}\, G_{F_\gamma}(u)E_{\mathcal{I}})\, E_{\mathcal{I}}^{*}\, v = E_{\mathcal{I}}^{*}\, g - E_{\mathcal{I}}^{*}\, G_{F_\gamma}(u)E_{\mathcal{A}}\, E_{\mathcal{A}}^{*}\, g. \end{cases} \tag{86}$$

In the proof of Lemma 5.3 we argued that

$$G_{F_\gamma}(u) \in \mathcal{L}(L^2(\tilde\Omega)), \text{ for each } u \in L^2(\tilde\Omega).$$

Slightly generalizing this argument shows that for each $\hat{u} \in L^2(\tilde{\Omega})$ there exists a neighborhood $U(\hat{u})$ and $C_{\hat{u}}$ such that

$$\|G_{F_\gamma}(u)\|_{\mathcal{L}(L^2(\tilde{\Omega}))} \le C_{\hat{u}} \text{ for all } u \in U(\hat{u}).$$

From (86) and Lemma 5.3 it follows that (85) holds with $K = 1 + \frac{1}{\alpha}(1 + C_{\hat{u}})$. □

2.3. Local q-superlinear convergence of the semismooth Newton iteration without and with a lifting step. For $r = 2$ we can deduce the following result form the discussion at the beginning of Section 3, Lemma 5.2 and Proposition 5.2.

Theorem 5.3. *If* (H2), (H3), (H5) *and* (H6) *hold, then the semi-smooth Newton iteration* (66) *applied to F_γ given in* (70) *with generalized derivative G_{F_γ} given in Corollary 5.1, is locally q-superlinearly convergent in $L^2(\tilde{\Omega})$.*

In case $r > 2$ the semi-smooth Newton algorithm is supplemented by a lifting step.

Algorithm 9 (**Semi-smooth Newton method with lifting**).

(i) Choose $u^0 \in L^r(\tilde{\Omega})$.

(ii) Solve for $\tilde{u}^{k+1} \in L^r(\tilde{\Omega})$:

$$G_{F_\gamma}(u^k)(\tilde{u}^{k+1} - u^k) = -F_\gamma(u^k).$$

(iii) Perform a lifting step:

$$u^{k+1} = \frac{1}{\alpha}\big(u_d + p_\gamma - (p_\gamma + \alpha(u_d - \bar{\varphi}))^+ - (p_\gamma + \alpha(u_d - \underline{\varphi}))^-\big),$$

where $p_\gamma = p_\gamma(\tilde{u}^{k+1})$.

The case with $r > 2$ is addressed next.

Theorem 5.4. *If* (H2), (H3), (H5) *and* (H6) *hold, then the semismooth Newton method with lifting step is locally q-superlinearly convergent in $L^r(\tilde{\Omega})$.*

Proof. Let $U(u_\gamma)$ denote the neighborhood of u_γ according to Theorem 5.2. Proposition 5.2 implies the existence of a constant M and $\bar{\rho} > 0$ such that

$$\|G_{F_\gamma}^{-1}(u)\|_{\mathcal{L}(L^2(\tilde{\Omega}))} \le M$$

for all $u \in B_r(u_\gamma, \rho)$. Here $B_r(u_\gamma, \rho)$ denotes the open ball with radius ρ and center u_γ in $L^r(\tilde{\Omega})$, with ρ sufficiently small such that $B_r(u_\gamma, \rho) \subset U(u_\gamma)$. We recall the definition of $p_\gamma(u)$:

$$p_\gamma(u) = -\gamma B^*(|B\,u| - \psi)^+ q(B\,u) - E_{\tilde{\Omega}}^* A^{-*} J_1'(A^{-1} E_{\tilde{\Omega}} u). \tag{87}$$

A computation shows that $v \to (|v| - \psi)^+ q(v)$ is globally Lipschitz continuous from $L^r(\tilde{\Omega})^l$ to itself with Lipschitz constant 3. Since $B \in \mathcal{L}(L^r(\tilde{\Omega}), L^\infty(\Omega))$

by (H3) and $B \in \mathcal{L}(L^{\hat{r}'}(\tilde{\Omega}), L^{r'}(\Omega))$ by (H6), the Marcinkiewicz interpolation theorem then implies that $B \in \mathcal{L}(L^2(\tilde{\Omega}), L^r(\Omega))$. Moreover $B^* \in \mathcal{L}(L^r(\tilde{\Omega}), L^r(\Omega))$ again as a consequence of (H6), and hence the first summand in (87) is gobally Lipschitz continuous from $L^2(\tilde{\Omega})$ to $L^r(\Omega)$. This, together with (H5) shows that $p_\gamma(u)$ is locally Lipschitz continuous from $L^2(\tilde{\Omega})$ to $L^r(\Omega)$. Let L denote the Lipschitz constant of $p_\gamma(\cdot)$ in $B_2(u_\gamma, \bar{\rho} M) \subset L^2(\tilde{\Omega})$. Without loss of generality we assume that $\alpha < L\, M$.

With L, M and $\bar{\rho}$ specified the lifting property of F_γ implies the existence of a constant $0 < \rho < \bar{\rho}$ such that

$$|F_\gamma(u_\gamma + h) - F_\gamma(u_\gamma) - G_{F_\gamma}(u_\gamma + h)h|_{L^r(\tilde{\Omega})} \le \frac{\alpha}{3L\, M\, |\Omega|^{\frac{r-2}{2}}} |h|_{L^r(\tilde{\Omega})}$$

for all $|h|_{L^r(\tilde{\Omega})} < \rho$. Let u_0 be such that $u_0 \in B_2(u_\gamma, \bar{\rho}) \cap B_r(u_\gamma, \bar{\rho})$ and proceeding by induction, assume that $u_k \in B_2(u_\gamma, \bar{\rho}) \cap B_r(u_\gamma, \bar{\rho})$. Then

$$\begin{aligned} |\tilde{u}^{k+1} - u_\gamma|_{L^2(\tilde{\Omega})} &\le \|G_{F_\gamma}(u^k)^{-1}\|_{\mathcal{L}(L^2)}\, |\Omega|^{\frac{r-2}{r}} \cdot \\ &\qquad \cdot |F_\gamma(u^k) - F(u_\gamma) - G_{F_\gamma}(u^k)(u^k - u_\gamma)|_{L^r(\tilde{\Omega})} \\ &\le \frac{\alpha}{3LM} |u^k - u_\gamma|_{L^r(\tilde{\Omega})} < |u^k - u_\gamma|_{L^r(\tilde{\Omega})}, \end{aligned} \tag{88}$$

and, in particular, $\tilde{u}_{k+1} \in B_2(u_\gamma, \bar{\rho})$. We further investigate the implications of the lifting step:

$$\begin{aligned} u^{k+1} - u_\gamma = \frac{1}{\alpha}\big(&p_\gamma(\tilde{u}^{k+1}) - p_\gamma(u_\gamma) - (p_\gamma(\tilde{u}^{k+1}) + \alpha(u_d - \bar{\varphi}))^+ \\ &+ (p_\gamma(u_\gamma) + \alpha(u_d - \bar{\varphi}))^+ - (p_\gamma(\tilde{u}^{k+1}) + \alpha(u_d - \underline{\varphi}))^- \\ &+ (p_\gamma(u_\gamma) + \alpha(u_d - \underline{\varphi}))^+\big), \end{aligned}$$

which implies that

$$|u^{k+1} - u_\gamma|_{L^r(\tilde{\Omega})} \le \frac{3}{\alpha} |p_\gamma(\tilde{u}^{k+1}) - p_\gamma(u_\gamma)|_{L^r(\Omega)} \le \frac{3L}{\alpha} |\tilde{u}^{k+1} - u_\gamma|_{L^2(\tilde{\Omega})}. \tag{89}$$

Combining (88) and (89) implies that $u^{k+1} \in B_r(u_\gamma, \bar{\rho})$. Thus the iteration is well-defined. Moreover we find

$$\frac{|u^{k+1} - u_\gamma|_{L^r(\tilde{\Omega})}}{|u^k - u_\gamma|_{L^r(\tilde{\Omega})}} \le \frac{\frac{3L}{\alpha} M\, |\Omega|^{\frac{r-2}{r}} |F_\gamma(u^k) - F_\gamma(u^\gamma) - G_{F_\gamma}(u^k)(u^k - u_\gamma)|_{L^r(\tilde{\Omega})}}{|u^k - u_\gamma|_{L^r(\tilde{\Omega})}},$$

which by (65) implies q-superlinear convergence. □

Remark 5.2. If we had a uniform estimate on $\|G_{F_\gamma}(u^k)^{-1}\|_{\mathcal{L}(L^r(\tilde{\Omega}))}$, then the lifting step could be avoided. In fact, note that we are not using the full power of the semismooth estimate (65) in (88), since we overestimate the L^2-norm by the L^r-norm.

We note, however, that for each fixed $u \in L^r(\tilde{\Omega})$ the operator $G_{F_\gamma}(u)$ is continuously invertible from $L^r(\tilde{\Omega})$ to itself; see Proposition 5.1. Thus, if $u_k \mapsto G_{F_\gamma}(u_k)$ is continuous for all sufficiently large k then the desired uniform estimate $G_{F_\gamma}(u_k)^{-1} \in \mathcal{L}(L^r(\tilde{\Omega}))$ holds. This continuity cannot be expected in general since $G_{F_\gamma}(u)$ contains the operator

$$T(u) = G_{\max}(|Bu| - \psi)q(Bu)q(Bu)^\top + (|Bu| - \psi)^+ Q(Bu);$$

see Lemma 5.3. If, however, the measure of the set $\{\mathrm{x} : (|Bu^k| - \psi)(\mathrm{x}) > 0\}$, and similarly for the max-term changes continuously with k, then the uniform bound on the inverse holds and the lifting step can be avoided. In the numerical examples given in the following section this behavior could be observed.

3. Numerics

Finally, we report on our numerical experience with Algorithm 1. In our tests we are interested in solving (P_γ) with large γ, i.e., we aim at a rather accurate approximation of the solution of (P). Algorithmically this is achieved by pre-selecting a sequence $\gamma_\ell = 10^\ell$, $\ell = 0, \ldots, 8$, of γ-values and solving (P_γ) for γ_ℓ with the solution of the problem corresponding to $\gamma_{\ell-1}$ as the initial guess. For $\ell = 0$ we use $u^0 \equiv 0$ as the initial guess. Such a continuation strategy with respect to γ is well-suited for producing initial guesses for the subsequent problem (P_γ) which satisfy the locality assumption of Theorem 5.3, and it usually results in a small number of semismooth Newton iterations until successful termination. We point out that more sophisticated and automatic selection rules for (γ_ℓ) may be used. For instance, one may adapt the technique of [**14**] for zero-order state constraints without additional constraints on the control.

In the numerical tests we throughout consider $A = -\Delta$, $\tilde{\Omega} = \Omega = (0,1)^2$ and $J_1(y) = \|y - y_d\|^2_{L^2(\Omega)}$. Here we discuss results for the following two problems.

PROBLEM 5.1. *The setting for this problem corresponds to Example 5.1. In this case we have zero-order state constraints, i.e.* $G = \mathrm{id}$, *with* $\psi(x_1, x_2) = 5E\text{-}3 \cdot (1 + 0.25|0.5 - x_1|)$. *The lower and upper bounds for the control are* $\underline{\varphi} \equiv 0$ *and* $\bar{\varphi}(x_1, x_2) = 0.1 + |\cos(2\pi x_1)|$, *respectively. Moreover we set* $u_d \equiv 0$, $y_d(x_1, x_2) = \sin(2\pi x_1)\exp(2x_2)/6$ *and* $\alpha = 1E\text{-}2$.

PROBLEM 5.2. *The second example corresponds to first-order state constraints with* $G = \nabla$ *and* $\psi \equiv 0.1$. *The pointwise bounds on the control are*

$$\underline{\varphi}(x_1, x_2) = \begin{cases} -0.5 - |x_1 - 0.5| - |x_2 - 0.5| & \textit{if } x_1 > 0.5 \\ 0 & \textit{if } x_1 \leq 0.5, \end{cases}$$

and $\bar{\varphi}(x_1, x_2) = 0.1 + |\cos(2\pi x_1)|$. *The desired control* u_d, *the desired state* y_d *and* α *are as in Problem 5.1.*

For the discretization of the respective problem we choose a regular mesh with mesh size h. The discretization of A is based on the standard five-point stencil and the one of G in Problem 5.2 uses symmetric differences. For each γ-value the algorithm is stopped as soon as $\|A_h p_h + \gamma G_h^\top (|G_h y_h| -$

$\psi_h)^+ + J'_{1h}(y_h)\|_{-1}$ and $\|\mu_h - (\mu_h + \alpha(u_{d,h} - b_h)^+ - (\mu_h + \alpha(u_{d,h} - a_h)^- \|_2$ drop below `tol`= 1E-8. Here we use $\|w\|_{-1} = \|A_h^{-1} w\|_2$, and the subscript h refers to discretized quantities. Before we commence with reporting on our numerical results, we briefly address step (ii) of Algorithm 2. In our tests, the solution of the linear system is achieved by sparse (Cholesky) factorization techniques. Alternatively, one may rely on iterative solvers (such as preconditioned conjugate gradient methods) for the state and the adjoint equation, respectively, as well as for the linear system in step (ii) of Algorithm 2.

In Figure 1 we display the state, control and multiplier μ_h upon termination of Algorithm 1 when solving the discrete version of Problem 5.1 for γ = 1E8 and $h = 1/256$. The active and inactive set structure with respect to the pointwise constraints on u_h can be seen in the left plot of Figure 2. Here, the white region corresponds to the inactive set, the gray region represents the active set for the lower bound and the black set is the active set with respect to the upper bound. The graph in the middle shows the approximation of the active set for the zero-order state constraint. On the right we show the overlapping region where the pointwise state constraint and one of the bounds on the control are active simultaneously. In Table 1 we display the iteration numbers upon successful termination of Algorithm 1 for various mesh sizes and for each γ-value of our pre-selected sequence. We recall that these figures are based on our γ-continuation strategy.

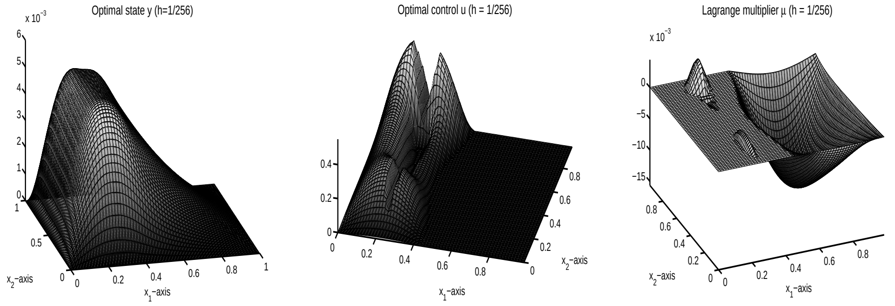

FIGURE 1. Problem 1 (γ = 1E8, $h = 1/256$). State y_h, control u_h and multiplier μ_h upon termination of Algorithm 1.

Upon studying the above results for Problem 5.1 we first note that Algorithm 1 with γ-continuation exhibits a mesh independent behavior. This can be seen from the stable iteration counts along the columns of Table 1. Moreover, for fixed h the number of iterations until termination is rather stable as well. This is due to the excellent initial guesses produced by our γ-continuation technique. In our tests we also found that Algorithm 1 without the γ-continuation for solving (P$_\gamma$) for large γ and with initial choice $u_h^0 = 0$ may fail to converge. Concerning the test example under investigation we note that the overlap of the active sets for the state and the controls

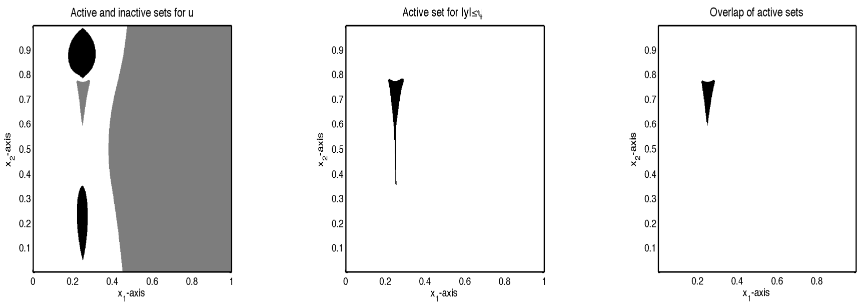

FIGURE 2. Problem 1 (γ = 1E8, $h = 1/256$). Inactive set (white), active set for the lower bound (gray) and for the upper bound (black) in the left plot. Approximation of the active set for the zero-order state constraint (black) in the middle plot. Overlap of active regions for control and state constraints in the right plot.

Iterations									
h/γ	1E0	1E1	1E2	1E3	1E4	1E5	1E6	1E7	1E8
$\frac{1}{64}$	3	3	4	5	5	5	4	4	2
$\frac{1}{128}$	3	3	4	6	5	5	5	4	4
$\frac{1}{256}$	3	3	5	6	5	5	5	5	4
$\frac{1}{512}$	4	3	5	6	6	6	5	5	5

TABLE 1. Problem 1. Number of iterations for various mesh sizes and γ-values.

is rather special. In this case, the bound on the state and the control satisfy the state equation in the region of overlapping active sets.

Next we report on our findings for Problem 5.2. In Figure 3 we show the state, control and multiplier μ_h upon termination for γ = 1E8 and $h = 1/256$. Figure 4 shows the active and inactive sets for the constraints

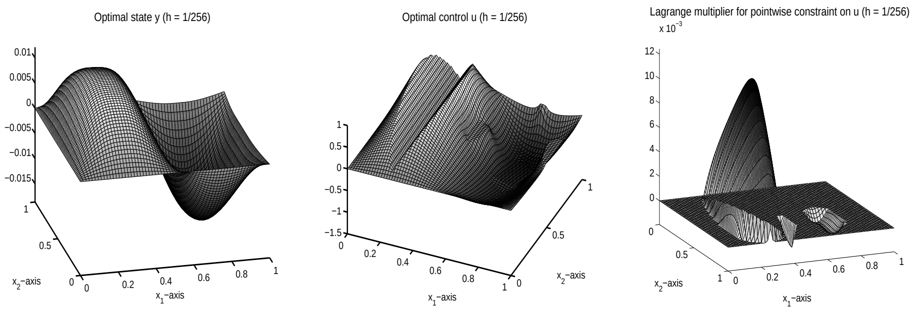

FIGURE 3. Problem 2 (γ = 1E8, $h = 1/256$). State y_h, control u_h and multiplier μ_h upon termination of Algorithm 1.

on the control in the left plot and the approximation for the active set for the pointwise gradient-constraint on the state on the right. As before, Table 2 provides the iteration numbers upon successful termination for various mesh sizes and γ-values.

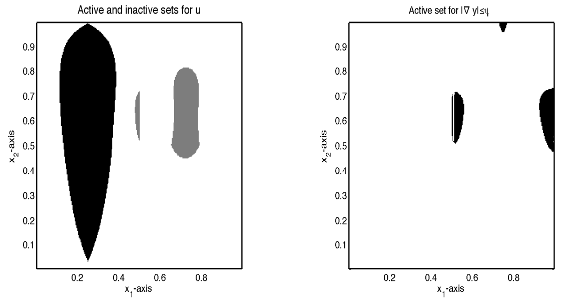

FIGURE 4. Problem 2 ($\gamma = 1\text{E}8, h = 1/256$). Inactive set (white), active set for the lower bound (gray) and for the upper bound (black) in the left plot. Approximation of the active set for the zero-order state constraint (black) in the right plot.

Iterations									
h/γ	1E0	1E1	1E2	1E3	1E4	1E5	1E6	1E7	1E8
$\frac{1}{32}$	6	6	6	4	3	3	2	2	2
$\frac{1}{64}$	7	7	6	4	4	3	3	2	2
$\frac{1}{128}$	7	7	6	5	5	4	3	2	2
$\frac{1}{256}$	7	7	6	6	5	5	4	3	2

TABLE 2. Problem 2. Number of iterations for various mesh sizes and γ-values.

Concerning the mesh independence and the stability of the iteration counts due to the employed γ-continuation scheme, the results of Table 2 support the same conclusions as for Problem 5.1. Again, without the continuation technique Algorithm 1 may fail to converge for the simple initial guess $u_h^0 = 0$. We observe that a stable (with respect to h and γ) and a superlinear convergence behavior of the semismooth Newton method is obtained without utilizing the lifting step.

Next we demonstrate the difference in regularity between λ_γ of Problem 5.1 (see Figure 5; left plot) and λ_γ^s of Problem 5.2 (see Figure 5; right plot). Note that for visualization purposes we linearly interpolate the multiplier values at the grid points. The approximate multiplier $\lambda_{\gamma,h}$ reflects the structural result obtained in [**4**]. According to this result, under sufficient regularity the multiplier is L^2-regular on the active set, zero on the inactive set and measure-valued on the boundary between the active and inactive set. Such a structure can be observed by inspecting the left plot of Figure 5.

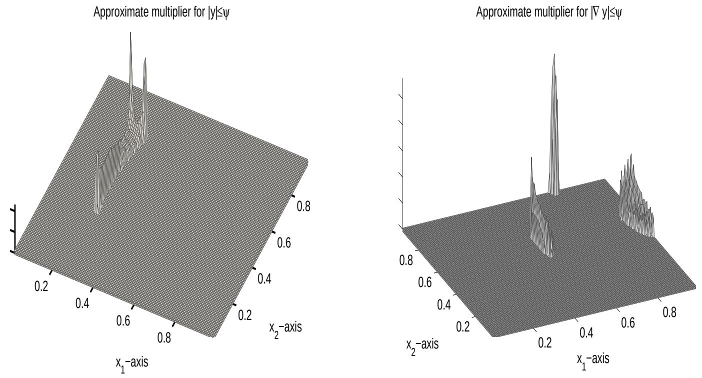

FIGURE 5. $\gamma = 1\text{E}8$, $h = 1/256$. Approximate multiplier $\lambda_{\gamma,h}$ for Problem 5.1 (left) and $\lambda^s_{\gamma,h}$ for Problem 5.2 (right).

On the other hand, for pointwise state constraints of gradient-type (first-order constraints) additional regularity on the active set appears not to be available for the example under investigation. Indeed, $\lambda^s_{\gamma,h}$ in the right plot of Figure 5 exhibits low regularity in the interior of the (smooth) active set.

Finally we note that the rather small value for `tol` and the rather large values for γ in our numerics reflect our interest of studying Algorithm 1 as a solver for a given discrete problem. In view of the error in discretization, however, when solving (P) by a sequence of approximating problems (P_γ) one would be interested in estimating the overall error in terms of the discretization and the γ-relaxation error, respectively. Such a result would allow a γ-choice such that both errors are balanced on a given mesh. This kind of numerical analysis is important in its own right, but goes beyond the scope of the present paper and is the subject of future research.

APPENDIX A

Some useful results

Given a vector norm $\|\cdot\|$ in $\mathbb{R}^n$ we assume throughout that for matrices $A, B \in \mathbb{R}^{n\times n}$ the corresponding norms satisfy the *consistency relation*

$$\|AB\| \le \|A\|\,\|B\|.$$

This condition is in particular satisfied if the matrix norm is *induced* by a given vector norm, such as the ℓ_p-norms:

$$\|A\| = \max_{v\in\mathbb{R}^n, v\neq 0} \left\{ \frac{\|Av\|}{\|v\|} \right\}.$$

THEOREM A.1. *Let $\|\cdot\|$ denote any norm on $\mathbb{R}^{n\times n}$ which satisfies the consistency relation above, and let $\|I\| = 1$, $E \in \mathbb{R}^{n\times n}$. If $\|E\| < 1$, then $(I-E)^{-1}$ exists and*

$$\|(I-E)^{-1}\| \le \frac{1}{1-\|E\|}. \tag{90}$$

If $A \in \mathbb{R}^{n\times n}$ is nonsingular and $\|A^{-1}(B-A)\| < 1$, then B is nonsingular and

$$\|B^{-1}\| \le \frac{\|A^{-1}\|}{1-\|A^{-1}(B-A)\|}. \tag{91}$$

The next result is concerned with the approximation quality of a sufficiently smooth nonlinear mapping.

THEOREM A.2. *Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable in the open convex set $D \subset \mathbb{R}^n$, $x \in D$, and let ∇F be Lipschitz continuous at x in the neighborhood D (with constant $L > 0$). Then, for any $x + d \in D$,*

$$\|F(x+d) - F(x) - \nabla F(x)d\| \le \frac{L}{2}\|d\|^2.$$

Proof. By using a mean value type result in integral form, we find

$$\begin{aligned} F(x+d) - F(x) - \nabla F(x)d &= \int_0^1 \nabla F(x+td)d\,dt - \nabla F(x)d \\ &= \int_0^1 (\nabla F(x+td) - \nabla F(x))\,d\,dt. \end{aligned}$$

Using the Lipschitz property of the Jacobian, we get

$$\begin{aligned}
\|F(x+d)-F(x)-\nabla F(x)d\| &\leq \int_0^1 \|\nabla F(x+td)-\nabla F(x)\| dt\, \|d\| \\
&\leq L\|d\|^2 \int_0^1 t\, dt \qquad (92)\\
&= \frac{L}{2}\|d\|^2.
\end{aligned}$$

□

Notice, if ∇F is Hölder continuous with exponent γ, with $0 < \gamma < 1$ and L still denoting the constant, then we obtain

$$\|F(x+d)-F(x)-\nabla F(x)d\| \leq \frac{L}{2}\|d\|^{1+\gamma}.$$

APPENDIX B

Auxiliary result

The following proposition establishes the Newton differentiability of a superposition of Newton differentiable maps.

PROPOSITION B.1. *Let $f : \mathcal{Y} \to \mathcal{Z}$ and $g : \mathcal{X} \to \mathcal{Y}$ be Newton differentiable in open sets V and U, respectively, with $U \subset \mathcal{X}$, $g(U) \subset V \subset \mathcal{Y}$. Assume that g is locally Lipschitz continuous and that there exists a Newton map $G_f(\cdot)$ of f which is bounded on $g(U)$. Then the superposition $f \circ g : \mathcal{X} \to \mathcal{Z}$ is Newton differentiable in U with a Newton map $G_f G_g$.*

PROOF. Let $x \in U$ and consider

$$\begin{aligned} &|f(g(x+h)) - f(g(x)) - G_f(g(x+h))G_g(x+h)h|_{\mathcal{Z}} \\ &\qquad = |f(w+k) - f(w) - G_f(w+k)k + R(x,h)|_{\mathcal{Z}}, \end{aligned} \tag{93}$$

where $w = g(x)$, $k = k(h) = g(x+h) - g(x)$ and $R(x,h) = G_f(g(x+h))\big(g(x+h) - g(x)\big) - G_f(g(x+h))G_g(x+h)h$. Observe that

$$\begin{aligned} |R(x,h)|_{\mathcal{Z}} &= |G_f(g(x+h))\big(g(x+h) - g(x) - G_g(x+h)h\big)|_{\mathcal{Z}} \\ &\leq C|g(x+h) - g(x) - G_g(x+h)h|_{\mathcal{Y}} = o(|h|_{\mathcal{X}}) \end{aligned}$$

as $|h|_{\mathcal{X}} \to 0$ by Newton differentiability of g at x. Further, owing to the local Lipschitz continuity of g there exists a constant $L > 0$ such that $|g(x+h) - g(x)|_{\mathcal{Y}} \leq L|h|_{\mathcal{X}}$ for all h sufficiently small. Hence, $|k(h)|_{\mathcal{Y}} = \mathcal{O}(|h|_{\mathcal{X}})$ as $|h|_{\mathcal{X}} \to 0$. Now we continue (93) by

$$\begin{aligned} &|f(w+k) - f(w) - G_f(w+k)k + R(x,h)|_{\mathcal{Z}} \\ &\qquad \leq |f(w+k) - f(w) - G_f(w+k)k|_{\mathcal{Z}} + o(|h|_{\mathcal{X}}) \\ &\qquad = o(|k|_{\mathcal{Y}}) + o(|h|_{\mathcal{X}}) = o(|h|_{\mathcal{X}}) \end{aligned}$$

as $|h|_{\mathcal{X}} \to 0$, where we use Newton differentiability of f at $g(x)$. This proves the assertion. □

Bibliography

[1] N. Arada and J.-P. Raymond. Dirichlet boundary control of semilinear parabolic equations. I. Problems with no state constraints. *Appl. Math. Optim.*, 45(2):125–143, 2002.

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

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

[4] M. Bergounioux and K. Kunisch. On the structure of Lagrange multipliers for state-constrained optimal control problems. *Systems Control Lett.*, 48(3-4):169–176, 2003. Optimization and control of distributed systems.

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

[6] C. Büskens and H. Maurer. SQP-methods for solving optimal control problems with control and state constraints: adjoint variables, sensitivity analysis and real-time control. *J. Comput. Appl. Math.*, 120(1-2):85–108, 2000. SQP-based direct discretization methods for practical optimal control problems.

[7] E. Casas. Control of an elliptic problem with pointwise state constraints. *SIAM J. Control Optim.*, 24(6):1309–1318, 1986.

[8] E. Casas and L. A. Fernandez. Optimal control of semilinear equations with pointwise constraints on the gradient of the state. *Appl. Math. Optim.*, 27:35–56, 1993.

[9] E. Casas, F. Tröltzsch, and A. Unger. Second order sufficient optimality conditions for a class of elliptic control problems. In *Control and partial differential equations (Marseille-Luminy, 1997)*, volume 4 of *ESAIM Proc.*, pages 285–300 (electronic). Soc. Math. Appl. Indust., Paris, 1998.

[10] X. Chen, Z. Nashed, and L. Qi. Smoothing methods and semismooth methods for nondifferentiable operator equations. *SIAM J. Numer. Anal.*, 38(4):1200–1216 (electronic), 2000.

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

[12] J.E. Dennis Jr., and R. B. Schnabel. *Numerical Methods for Unconstrained optimization and Nonlinear Equations.* Series *Classics in Applied Mathematics*, vol. 16, SIAM, Philadelphia, 1996.

[13] M. Hintermüller, K. Ito, and K. Kunisch. The primal-dual active set method as a semismooth Newton method. SIAM J. Optimization, 13 (2003), pp. 865-888.

[14] M. Hintermüller and K. Kunisch. Feasible and non-interior path-following in constrained minimization with low multiplier regularity. *SIAM J. Control Optim.*, 45(4):1198–1221, 2006.

[15] M. Hintermüller and K. Kunisch. Path-following methods for a class of constrained minimization problems in function space. *SIAM J. Optim.*, 17(1):159–187, 2006.

[16] M. Hintermüller and W. Ring. Numerical aspects of a level set based algorithm for state constrained optimal control problems. *Computer Assisted Mechanics and Eng. Sci.*, 3, 2003.

[17] M. Hintermüller and W. Ring. A level set approach for the solution of a state constrained optimal control problem. *Num. Math.*, 2004.

[18] M. Hintermüller, and M. Ulbrich. A mesh independence result for semismooth Newton methods. Mathematical Programming, 101 (2004), pp. 151-184.

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

[20] K. Ito and K. Kunisch. *On the Lagrange Multiplier Approach to Variational Problems and Applications*. SIAM, Philadelphia. forthcoming.

[21] K. Ito and K. Kunisch. Semi-smooth Newton methods for state-constrained optimal control problems. *Systems Control Lett.*, 50(3):221–228, 2003.

[22] K. Krumbiegel and A. Rösch. On the regularization error of state constrained Neumann control problems. *Control Cybernet.*, 37(2):369–392, 2008.

[23] P. Lu. Non-linear systems with control and state constraints. *Optimal Control Appl. Methods*, 18(5):313–326, 1997.

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

[25] C. Meyer, A. Rösch, and F. Tröltzsch. Optimal control of PDEs with regularized pointwise state constraints. *Comput. Optim. Appl.*, 33(2-3):209–228, 2006.

[26] R. Mifflin, Semismooth and semiconvex functions in constrained optimization. SIAM J. Control and Optimization, 15 (1977), pp. 957-972.

[27] P. Neittaanmäki, J. Sprekels, and D. Tiba. *Optimization of Elliptic Systems.* Springer, Berlin, 2006.

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

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

[30] J. P. Raymond. Optimal control problem for semilinear parabolic equations with pointwise state constraints. In *Modelling and optimization of distributed parameter systems (Warsaw, 1995)*, pages 216–222. Chapman & Hall, New York, 1996.

[31] J. P. Raymond. Nonlinear boundary control of semilinear parabolic problems with pointwise state constraints. *Discrete Contin. Dynam. Systems*, 3(3):341–370, 1997.

[32] J.-P. Raymond and F. Tröltzsch. Second order sufficient optimality conditions for nonlinear parabolic control problems with state constraints. *Discrete Contin. Dynam. Systems*, 6(2):431–450, 2000.

[33] D. Tiba and C. Zalinescu. On the necessity of some constraint qualification conditions in convex programming. *J. Convex Anal.*, 11:95–110, 2004.

[34] G. M. Troianiello. *Elliptic Differential Equations and Obstacle Problems*. The University Series in Mathematics. Plenum Press, New York, 1987.

[35] F. Tröltzsch. Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints. *SIAM J. Optim.*, 15(2):616–634 (electronic), 2004/05.

[36] M. Ulbrich. Semismooth Newton methods for operator equations in function spaces. SIAM J. Optimization, 13 (2003), pp. 805-842.