

# Chapter 2

## Optimization Methods in Banach Spaces

Michael Ulbrich

**Abstract** In this chapter we present a selection of important algorithms for optimization problems with partial differential equations. The development and analysis of these methods is carried out in a Banach space setting. We begin by introducing a general framework for achieving global convergence. Then, several variants of generalized Newton methods are derived and analyzed. In particular, necessary and sufficient conditions for fast local convergence are derived. Based on this, the concept of semismooth Newton methods for operator equations is introduced. It is shown how complementarity conditions, variational inequalities, and optimality systems can be reformulated as semismooth operator equations. Applications to constrained optimal control problems are discussed, in particular for elliptic partial differential equations and for flow control problems governed by the incompressible stationary Navier-Stokes equations. As a further important concept, the formulation of optimality systems as generalized equations is addressed. We introduce and analyze the Josephy-Newton method for generalized equations. This provides an elegant basis for the motivation and analysis of sequential quadratic programming (SQP) algorithms. The chapter concludes with a short outline of recent algorithmic advances for state constrained problems and a brief discussion of several further aspects.

### 2.1 Synopsis

The aim of this chapter is to give an introduction to selected optimization algorithms that are well-suited for PDE-constrained optimization. For the development and analysis of such algorithms, a functional analytic setting is the framework of choice. Therefore, we will develop optimization methods in this abstract setting and then return to concrete problems later.

Optimization methods are iterative algorithms for finding (global or local) solutions of minimization problems. Usually, we are already satisfied if the method can be proved to converge to *stationary* points. These are points that satisfy the first-order necessary optimality conditions. Besides global convergence, which will not be the main focus of this chapter, fast local convergence is desired. All fast converging optimization methods use the idea of Newton's method in some sense. Therefore, our main focus will be on Newton-type methods for optimization problems in Banach spaces.

---

M. Ulbrich (✉)

Lehrstuhl für Mathematische Optimierung, TU München, Garching, Germany

e-mail: [mulbrich@ma.tum.de](mailto:mulbrich@ma.tum.de)

Optimization methods for minimizing an objective function  $f : W \rightarrow \mathbb{R}$  on a feasible set  $W_{\text{ad}} \subset W$ , where  $W$  is a Banach space, generate a sequence  $(w^k) \subset W$  of iterates. Essentially, as already indicated, there are two desirable properties an optimization algorithm should have:

1. Global convergence:

There are different flavors to formulate global convergence. Some of them use the notion of a stationarity measure. This is a function  $\Sigma : W \rightarrow \mathbb{R}_+$  with  $\Sigma(w) = 0$  if  $w$  is stationary and  $\Sigma(w) > 0$ , otherwise. In the unconstrained case, i.e.,  $W_{\text{ad}} = W$ , a common choice is  $\Sigma(w) := \|f'(w)\|_{W^*}$ . The following is a selection of global convergence assertions:

- (a) Every accumulation point of  $(w^k)$  is a stationary point.
- (b) For some continuous stationarity measure  $\Sigma(w)$  there holds

$$\lim_{k \rightarrow \infty} \Sigma(w^k) = 0.$$

- (c) There exists an accumulation point of  $(w^k)$  that is stationary.
- (d) For the continuous stationarity measure  $\Sigma(w)$  there holds

$$\liminf_{k \rightarrow \infty} \Sigma(w^k) = 0.$$

Note that (b) implies (a) and (c) implies (d).

2. Fast local convergence:

These are local results in a neighborhood of a stationary point  $\bar{w}$ :

There exists  $\delta > 0$  such that, for all  $w^0 \in W$  with  $\|w^0 - \bar{w}\|_W < \delta$ , we have  $w^k \rightarrow \bar{w}$  and

$$\|w^{k+1} - \bar{w}\|_W = o(\|w^k - \bar{w}\|_W) \quad (\text{q-superlinear convergence}),$$

or even, for  $\alpha > 0$ ,

$$\|w^{k+1} - \bar{w}\|_W = O(\|w^k - \bar{w}\|_W^{1+\alpha})$$

(q-superlinear convergence with order  $1 + \alpha$ ).

The case  $1 + \alpha = 2$  is called q-quadratic convergence.

We begin with a discussion of globalization concepts. Then, in the rest of this chapter, we present locally fast convergent methods that all can be viewed as Newton-type methods.

*Notation* If  $W$  is a Banach space, we denote by  $W^*$  its dual space. The Fréchet-derivative (F-derivative) of an operator  $G : X \rightarrow Y$  between Banach spaces is denoted by  $G' : X \rightarrow \mathcal{L}(X, Y)$ , where  $\mathcal{L}(X, Y)$  are the bounded linear operators  $A : X \rightarrow Y$ . In particular, the derivative of a real-valued function  $f : W \rightarrow \mathbb{R}$  is denoted by  $f' : W \rightarrow W^*$ . In case of a Hilbert space  $W$ , the gradient  $\nabla f : W \rightarrow W$  is the Riesz representation of  $f'$ , i.e.,

$$\langle \nabla f(w), v \rangle_W = \langle f'(w), v \rangle_{W^*, W} \quad \forall v \in W.$$

Here  $\langle f'(w), v \rangle_{W^*, W}$  denotes the dual pairing between the dual space  $W^* = \mathcal{L}(W, \mathbb{R})$  and  $W$  and  $(\cdot, \cdot)_W$  is the inner product. Note that in Hilbert space we can do the identification  $W^* = W$  via  $\langle \cdot, \cdot \rangle_{W^*, W} = (\cdot, \cdot)_W$ , but this is not always done.

## 2.2 Globally Convergent Methods in Banach Spaces

### 2.2.1 Unconstrained Optimization

For understanding how global convergence can be achieved, it is important to look at unconstrained optimization first:

$$\min_{w \in W} f(w)$$

with  $W$  a real Banach space and  $f : W \rightarrow \mathbb{R}$  continuously  $F$ -differentiable.

The first-order optimality conditions for a local minimum  $\bar{w} \in W$  are well-known:

$\bar{w} \in W$  satisfies

$$f'(\bar{w}) = 0.$$

We develop a general class of methods that is globally convergent: *Descent methods*.

The idea of descent methods is to find, at the current ( $k$ th) iterate  $w^k \in W$ , a direction  $s^k \in W$  such that  $\phi_k(t) \stackrel{\text{def}}{=} f(w^k + ts^k)$  is decreasing at  $t = 0$ :

$$\phi'_k(0) = \langle f'(w^k), s^k \rangle_{W^*, W} < 0.$$

Of course, this descent can be very small. However, from the (sharp) estimate

$$\phi'_k(0) = \langle f'(w^k), s^k \rangle_{W^*, W} \geq -\|f'(w^k)\|_{W^*} \|s^k\|_W$$

it is natural to derive the following quality requirement (“angle” condition)

$$\langle f'(w^k), s^k \rangle_{W^*, W} \leq -\eta \|f'(w^k)\|_{W^*} \|s^k\|_W \quad (2.1)$$

for the descent direction. Here  $\eta \in (0, 1)$  is fixed.

A second ingredient of a descent method is a step size rule to obtain a step size  $\sigma_k > 0$  such that

$$\phi_k(\sigma_k) < \phi_k(0).$$

Then, the new iterate is computed as  $w^{k+1} := w^k + \sigma_k s^k$ . Overall, we obtain:

**Algorithm 2.1** (General descent method)

0. Choose an initial point  $w^0 \in W$ .

For  $k = 0, 1, 2, \dots$ :

1. If  $f'(w^k) = 0$ , STOP.
2. Choose a descent direction  $s^k \in W$ :  $\langle f'(w^k), s^k \rangle_{W^*, W} < 0$ .
3. Choose a step size  $\sigma_k > 0$  such that  $f(w^k + \sigma_k s^k) < f(w^k)$ .
4. Set  $w^{k+1} := w^k + \sigma_k s^k$ .

In this generality, it is not possible to prove global convergence. We need additional requirements on the quality of the descent direction and the step sizes:

1. Admissibility of the search directions:

$$\frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W} \xrightarrow{k \rightarrow \infty} 0 \implies \|f'(w^k)\|_{W^*} \xrightarrow{k \rightarrow \infty} 0.$$

2. Admissibility of the step sizes:

$$f(w^k + \sigma_k s^k) < f(w^k) \quad \forall k \quad \text{and} \\ f(w^k + \sigma_k s^k) - f(w^k) \xrightarrow{k \rightarrow \infty} 0 \implies \frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W} \xrightarrow{k \rightarrow \infty} 0.$$

These conditions become more intuitive by realizing that the expression  $\frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W}$  is the slope of  $f$  at  $w^k$  in the direction  $s^k$ :

$$\left. \frac{d}{dt} f\left(w^k + t \frac{s^k}{\|s^k\|_W}\right) \right|_{t=0} = \frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W}.$$

Therefore, admissible step sizes mean that if the  $f$ -decreases become smaller and smaller then the slopes along the  $s^k$  have to become smaller and smaller. And admissible search directions mean that if the slopes along the  $s^k$  become smaller and smaller then the steepest possible slopes have to become smaller and smaller.

With these two conditions at hand, we can prove global convergence.

**Theorem 2.2** *Let  $f$  be continuously  $F$ -differentiable and  $(w^k)$ ,  $(s^k)$ ,  $(\sigma_k)$  be generated by Algorithm 2.1. Assume that  $(\sigma_k)$  and  $(s^k)$  are admissible and that  $(f(w^k))$  is bounded below. Then*

$$\lim_{k \rightarrow \infty} f'(w^k) = 0. \tag{2.2}$$

*In particular, every accumulation point of  $(w^k)$  is a stationary point.*

*Proof* Let  $f^* = \inf_{k \geq 0} f(w^k) > -\infty$ . Then, using  $f(w^k + \sigma_k s^k) - f(w^k) < 0$ , we see that  $f(w^k) \rightarrow f^*$  and

$$f(w^0) - f^* = \sum_{k=0}^{\infty} (f(w^k) - f(w^{k+1})) = \sum_{k=0}^{\infty} |f(w^k + \sigma_k s^k) - f(w^k)|.$$

This shows  $f(w^k + \sigma_k s^k) - f(w^k) \rightarrow 0$ . By the admissibility of  $(\sigma_k)$ , this implies

$$\frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W} \xrightarrow{k \rightarrow \infty} 0.$$

Now the admissibility of  $(s^k)$  yields

$$\|f'(w^k)\|_{W^*} \xrightarrow{k \rightarrow \infty} 0.$$

Next, consider the situation where  $\bar{w}$  is an accumulation point of  $(w^k)$ . Then there exists a subsequence  $(w^k)_K \rightarrow \bar{w}$  and due to monotonicity of  $f(w^k)$  we conclude  $f(w^k) \geq f(\bar{w})$  for all  $k$ . Hence, we can apply the first part of the theorem and obtain (2.2). Now, by continuity,

$$f'(\bar{w}) = \lim_{k \rightarrow \infty} f'(w^k) = 0.$$

There are two questions open:

- (a) How can we check in practice if a search direction is admissible or not?
- (b) How can we compute admissible step sizes?

An answer to question (a) is provided by the following Lemma:

**Lemma 2.1** *If the search directions  $(s^k)$  satisfy the angle condition (2.1) then they are admissible.*

*Proof* The angle condition yields

$$\|f'(w^k)\|_{W^*} \leq -\frac{1}{\eta} \frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W}.$$

A very important step size rule is the

### 2.2.1.1 Armijo Rule

Given a descent direction  $s^k$  of  $f$  at  $w^k$ , choose the maximum  $\sigma_k \in \{1, 1/2, 1/4, \dots\}$  for which

$$f(w^k + \sigma_k s^k) - f(w^k) \leq \gamma \sigma_k \langle f'(w^k), s^k \rangle_{W^*, W}.$$

Here  $\gamma \in (0, 1)$  is a constant. The next result shows that Armijo step sizes exist.

**Lemma 2.2** *Let  $f'$  be uniformly continuous on  $N_0^\rho = \{w + s : f(w) \leq f(w^0), \|s\|_W \leq \rho\}$  for some  $\rho > 0$ . Then, for every  $\varepsilon > 0$ , there exists  $\delta > 0$  such that for all  $w^k \in W$  with  $f(w^k) \leq f(w^0)$  and all  $s^k \in W$  that satisfy*

$$\frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W} \leq -\varepsilon,$$

there holds

$$f(w^k + \sigma s^k) - f(w^k) \leq \gamma \sigma \langle f'(w^k), s^k \rangle_{W^*, W} \quad \forall \sigma \in [0, \delta / \|s^k\|_W].$$

*Proof* We have, with appropriate  $\tau_\sigma \in [0, \sigma]$ ,

$$\begin{aligned} f(w^k + \sigma s^k) - f(w^k) &= \sigma \langle f'(w^k + \tau_\sigma s^k), s^k \rangle_{W^*, W} \\ &\leq \sigma \langle f'(w^k), s^k \rangle_{W^*, W} + \sigma \|f'(w^k + \tau_\sigma s^k) \\ &\quad - f'(w^k)\|_{W^*} \|s^k\|_W \\ &= \gamma \sigma \langle f'(w^k), s^k \rangle_{W^*, W} + \rho_k(\sigma), \end{aligned}$$

where

$$\rho_k(\sigma) := (1 - \gamma) \sigma \langle f'(w^k), s^k \rangle_{W^*, W} + \sigma \|f'(w^k + \tau_\sigma s^k) - f'(w^k)\|_{W^*} \|s^k\|_W.$$

Now we use the uniform continuity of  $f'$  to choose  $\delta \in (0, \rho)$  so small that

$$\|f'(w^k + \tau_\sigma s^k) - f'(w^k)\|_{W^*} < (1 - \gamma)\varepsilon \quad \forall \sigma \in [0, \delta / \|s^k\|_W].$$

This is possible since

$$\|\tau_\sigma s^k\|_W \leq \sigma \|s^k\|_W \leq \delta.$$

Then

$$\begin{aligned} \rho_k(\sigma) &= (1 - \gamma) \sigma \langle f'(w^k), s^k \rangle_{W^*, W} + \sigma \|f'(w^k + \tau_\sigma s^k) - f'(w^k)\|_{W^*} \|s^k\|_W \\ &\leq -(1 - \gamma)\varepsilon \sigma \|s^k\|_W + (1 - \gamma)\varepsilon \sigma \|s^k\|_W = 0. \end{aligned}$$

Next, we prove the admissibility of Armijo step sizes under mild conditions.

**Lemma 2.3** *Let  $f'$  be uniformly continuous on  $N_0^\rho = \{w + s : f(w) \leq f(w^0), \|s\|_W \leq \rho\}$  for some  $\rho > 0$ . We consider Algorithm 2.1, where  $(\sigma_k)$  is generated by the Armijo rule and the descent directions  $s^k$  are chosen such that they are not too short in the following sense:*

$$\|s^k\|_W \geq \phi \left( -\frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W} \right),$$

where  $\phi : [0, \infty) \rightarrow [0, \infty)$  is monotonically increasing and satisfies  $\phi(t) > 0$  for all  $t > 0$ . Then the step sizes  $(\sigma_k)$  are admissible.

*Proof* Assume that there exist an infinite set  $K$  and  $\varepsilon > 0$  such that

$$\frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W} \leq -\varepsilon \quad \forall k \in K.$$

Then

$$\|s^k\|_W \geq \phi \left( -\frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W} \right) \geq \phi(\varepsilon) =: \eta > 0 \quad \forall k \in K.$$

By Lemma 2.2, for  $k \in K$  we have either  $\sigma_k = 1$  or  $\sigma_k \geq \delta/(2\|s^k\|)$ . Hence,

$$\sigma_k \|s^k\|_W \geq \min\{\delta/2, \eta\} \quad \forall k \in K.$$

This shows

$$\begin{aligned} f(w^k + \sigma_k s^k) - f(w^k) &\leq \gamma \sigma_k \langle f'(w^k), s^k \rangle_{W^*, W} = \gamma \sigma_k \|s^k\|_W \frac{\langle f'(w^k), s^k \rangle_{W^*, W}}{\|s^k\|_W} \\ &\leq -\gamma \min\{\delta/2, \eta\} \varepsilon \quad \forall k \in K. \end{aligned}$$

Therefore

$$f(w^k + \sigma_k s^k) - f(w^k) \not\rightarrow 0.$$

In the Banach space setting, the computation of descent directions is not straightforward. Note that the negative derivative of  $f$  is *not* suitable, since  $W^* \ni f'(w^k) \notin W$ .

In the Hilbert space setting, however, we *can* choose  $W^* = W$  and  $\langle \cdot, \cdot \rangle_{W^*, W} = \langle \cdot, \cdot \rangle_W$  by the Riesz representation theorem. Then we have  $f'(w^k) = \nabla f(w^k) \in W$  and  $-\nabla f(w^k)$  is the direction of steepest descent, as we will show below.

Certainly the most well-known descent method is the steepest descent method. In Banach space, the steepest descent directions of  $f$  at  $w$  are defined by  $s = t d_{sd}$ ,  $t > 0$ , where  $d_{sd}$  solves

$$\min_{\|d\|_W=1} \langle f'(w), d \rangle_{W^*, W}.$$

Now consider the case where  $W = W^*$  is a Hilbert space. Then

$$d_{sd} = -\frac{\nabla f(w)}{\|\nabla f(w)\|_W}.$$

In fact, by the Cauchy-Schwarz inequality,

$$\begin{aligned} \min_{\|d\|_W=1} \langle f'(w), d \rangle_{W^*, W} &= \min_{\|d\|_W=1} (\nabla f(w), d)_W \geq -\|\nabla f(w)\|_W \\ &= \left( \nabla f(w), -\frac{\nabla f(w)}{\|\nabla f(w)\|_W} \right)_W. \end{aligned}$$

Therefore,  $-\nabla f(w)$  is a steepest descent direction. This is the reason why the steepest descent method is also called gradient method.

It should be mentioned that the steepest descent method is usually very inefficient. Therefore, the design of efficient globally convergent methods works as follows: A locally fast convergent method (e.g., Newton's method) is used to generate

trial steps. If the generated step satisfies a (generalized) angle test ensuring admissibility of the step, the step is selected. Otherwise, another search direction is chosen, e.g., the steepest descent direction.

### 2.2.2 Optimization on Closed Convex Sets

We now develop descent methods for simply constrained problems of the form

$$\min f(w) \quad \text{s.t.} \quad w \in S \quad (2.3)$$

with  $W$  a Hilbert space,  $f : W \rightarrow \mathbb{R}$  continuously  $F$ -differentiable, and  $S \subset W$  closed and convex.

*Example 2.1* A scenario frequently found in practice is

$$W = L^2(\Omega), \quad S = \left\{ u \in L^2(\Omega) : a(x) \leq u(x) \leq b(x) \text{ a.e. on } \Omega \right\}$$

with  $L^\infty$ -functions  $a, b$ . It is then very easy to compute the projection  $P_S$  onto  $S$ , which will be needed in the following:

$$P_S(w)(x) = P_{[a(x), b(x)]}(w(x)) = \max(a(x), \min(w(x), b(x))).$$

The presence of the constraint set  $S$  requires to take care that we stay feasible with respect to  $S$ , or—if we think of an infeasible method—that we converge to feasibility. In the following, we consider a feasible algorithm, i.e.,  $w^k \in S$  for all  $k$ .

If  $w^k$  is feasible and we try to apply the unconstrained descent method, we have the difficulty that already very small step sizes  $\sigma > 0$  can result in points  $w^k + \sigma s^k$  that are infeasible. The backtracking idea of considering only those  $\sigma \geq 0$  for which  $w^k + \sigma s^k$  is feasible is not viable, since very small step sizes or even  $\sigma_k = 0$  might be the result.

Therefore, instead of performing a line search along the ray  $\{w^k + \sigma s^k : \sigma \geq 0\}$ , we perform a line search along the projected path

$$\left\{ P_S(w^k + \sigma s^k) : \sigma \geq 0 \right\},$$

where  $P_S$  is the projection onto  $S$ . Of course, we have to ensure that along this path we achieve sufficient descent as long as  $w^k$  is not a stationary point. Unfortunately, not any descent direction is suitable here.

*Example 2.2* Consider

$$S = \left\{ w \in \mathbb{R}^2 : w_1 \geq 0, w_1 + w_2 \geq 3 \right\}, \quad f(w) = 5w_1^2 + w_2^2.$$

Then, at  $w^k = (1, 2)^T$ , we have  $\nabla f(w^k) = (10, 4)^T$ . Since  $f$  is convex quadratic with minimum  $\bar{w} = 0$ , the Newton step is

$$d^k = -w^k = -(1, 2)^T.$$

This is a descent direction, since

$$\nabla f(w^k)^T d^k = -18.$$

But, for  $\sigma \geq 0$ , there holds

$$P_S(w^k - \sigma d^k) = P_S((1 - \sigma)(1, 2)^T) = (1 - \sigma) \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \sigma \begin{pmatrix} 3/2 \\ 3/2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \frac{\sigma}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

From

$$\nabla f(w^k)^T \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 6$$

we see that we are getting ascent, not descent, along the projected path, although  $d^k$  is a descent direction.

The example shows that care must be taken in choosing appropriate search directions for projected methods. Since the projected descent properties of a search direction are more complicated to judge than in the unconstrained case, it is out of the scope of this chapter to give a general presentation of this topic. In the finite dimensional setting, we refer to [84] for a detailed discussion. Here, we only consider the projected gradient method.

### Algorithm 2.3 (Projected gradient method)

0. Choose  $w^0 \in S$ .

For  $k = 0, 1, 2, 3, \dots$ :

1. Set  $s^k = -\nabla f(w^k)$ .
2. Choose  $\sigma_k$  by a projected step size rule such that  $f(P_S(w^k + \sigma_k s^k)) < f(w^k)$ .
3. Set  $w^{k+1} := P_S(w^k + \sigma_k s^k)$ .

For abbreviation, let

$$w_\sigma^k = w^k - \sigma \nabla f(w^k).$$

We will prove global convergence of this method. To do this, we need the facts about the projection operator  $P_S$  collected in Lemma 1.10.

The following result shows that along the projected steepest descent path we achieve a certain amount of descent:

**Lemma 2.4** *Let  $W$  be a Hilbert space and let  $f : W \rightarrow \mathbb{R}$  be continuously  $F$ -differentiable on a neighborhood of the closed convex set  $S$ . Let  $w^k \in S$  and assume*

that  $\nabla f$  is  $\alpha$ -order Hölder-continuous with modulus  $L > 0$  on

$$\left\{ (1-t)w^k + tP_S(w_\sigma^k) : 0 \leq t \leq 1 \right\},$$

for some  $\alpha \in (0, 1]$ . Then there holds

$$f(P_S(w_\sigma^k)) - f(w^k) \leq -\frac{1}{\sigma} \|P_S(w_\sigma^k) - w^k\|_W^2 + L \|P_S(w_\sigma^k) - w^k\|_W^{1+\alpha}.$$

*Proof*

$$\begin{aligned} f(P_S(w_\sigma^k)) - f(w^k) &= (\nabla f(v_\sigma^k), P_S(w_\sigma^k) - w^k)_W \\ &= (\nabla f(w^k), P_S(w_\sigma^k) - w^k)_W \\ &\quad + (\nabla f(v_\sigma^k) - \nabla f(w^k), P_S(w_\sigma^k) - w^k)_W \end{aligned}$$

with appropriate  $v_\sigma^k \in \{(1-t)w^k + tP_S(w_\sigma^k) : 0 \leq t \leq 1\}$ .

Now, since  $w_\sigma^k - w^k = \sigma s^k = -\sigma \nabla f(w^k)$  and  $w^k = P_S(w^k)$ , we obtain

$$\begin{aligned} -\sigma (\nabla f(w^k), P_S(w_\sigma^k) - w^k)_W &= (w_\sigma^k - w^k, P_S(w_\sigma^k) - w^k)_W \\ &= (w_\sigma^k - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\ &= (P_S(w_\sigma^k) - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\ &\quad + \underbrace{(w_\sigma^k - P_S(w_\sigma^k), P_S(w_\sigma^k) - P_S(w^k))_W}_{\geq 0} \\ &\geq (P_S(w_\sigma^k) - P_S(w^k), P_S(w_\sigma^k) - P_S(w^k))_W \\ &= \|P_S(w_\sigma^k) - w^k\|_W^2. \end{aligned}$$

Next, we use

$$\|v_\sigma^k - w^k\|_W \leq \|P_S(w_\sigma^k) - w^k\|_W.$$

Hence,

$$\begin{aligned} (\nabla f(v_\sigma^k) - \nabla f(w^k), P_S(w_\sigma^k) - w^k)_W &\leq \|\nabla f(v_\sigma^k) - \nabla f(w^k)\|_W \|P_S(w_\sigma^k) - w^k\|_W \\ &\leq L \|v_\sigma^k - w^k\|_W^\alpha \|P_S(w_\sigma^k) - w^k\|_W \\ &\leq L \|P_S(w_\sigma^k) - w^k\|_W^{1+\alpha}. \end{aligned}$$

We now consider the following

### 2.2.2.1 Projected Armijo Rule

Choose the maximum  $\sigma_k \in \{1, 1/2, 1/4, \dots\}$  for which

$$f(P_S(w^k + \sigma_k s^k)) - f(w^k) \leq -\frac{\gamma}{\sigma_k} \|P_S(w^k + \sigma_k s^k) - w^k\|_W^2.$$

Here  $\gamma \in (0, 1)$  is a constant.

In the unconstrained case, we recover the classical Armijo rule:

$$\begin{aligned} f(P_S(w^k + \sigma_k s^k)) - f(w^k) &= f(w^k + \sigma_k s^k) - f(w^k), \\ -\frac{\gamma}{\sigma_k} \|P_S(w^k + \sigma_k s^k) - w^k\|_W^2 &= -\frac{\gamma}{\sigma_k} \|\sigma_k s^k\|_W^2 = -\gamma \sigma_k \|s^k\|_W^2 \\ &= \gamma \sigma_k (\nabla f(w^k), s^k)_W. \end{aligned}$$

As a stationarity measure  $\Sigma(w) = \|p(w)\|_W$  we use the norm of the *projected gradient*

$$p(w) \stackrel{\text{def}}{=} w - P_S(w - \nabla f(w)).$$

In fact, the first-order optimality conditions for (2.3) are

$$w \in S, \quad (\nabla f(w), v - w)_W \geq 0 \quad \forall v \in S.$$

By Lemma 1.10, this is equivalent to

$$w - P_S(w - \nabla f(w)) = 0.$$

As a next result we show that projected Armijo step sizes exist.

**Lemma 2.5** *Let  $W$  be a Hilbert space and let  $f : W \rightarrow \mathbb{R}$  be continuously  $F$ -differentiable on a neighborhood of the closed convex set  $S$ . Then, for all  $w^k \in S$  with  $p(w^k) \neq 0$ , the projected Armijo rule terminates successfully.*

*Proof* We proceed as in the proof of Lemma 2.4 and obtain (we have not assumed Hölder continuity of  $\nabla f$  here)

$$f(P_S(w_\sigma^k)) - f(w^k) \leq \frac{-1}{\sigma} \|P_S(w_\sigma^k) - w^k\|_W^2 + o(\|P_S(w_\sigma^k) - w^k\|_W).$$

It remains to show that, for all small  $\sigma > 0$ ,

$$\frac{\gamma - 1}{\sigma} \|P_S(w_\sigma^k) - w^k\|_W^2 + o(\|P_S(w_\sigma^k) - w^k\|_W) \leq 0.$$

But this follows easily from (Lemma 1.10(e)):

$$\frac{\gamma - 1}{\sigma} \|P_S(w_\sigma^k) - w^k\|_W^2 \leq \underbrace{(\gamma - 1) \|p(w^k)\|_W}_{<0} \|P_S(w_\sigma^k) - w^k\|_W.$$

**Theorem 2.4** *Let  $W$  be a Hilbert space,  $f : W \rightarrow \mathbb{R}$  be continuously  $F$ -differentiable, and  $S \subset W$  be nonempty, closed, and convex. Consider Algorithm 2.1 and assume that  $f(w^k)$  is bounded below. Furthermore, let  $\nabla f$  be  $\alpha$ -order Hölder continuous on*

$$N_0^\rho = \left\{ w + s : f(w) \leq f(w^0), \|s\|_W \leq \rho \right\}$$

for some  $\alpha > 0$  and some  $\rho > 0$ . Then

$$\lim_{k \rightarrow \infty} \|p(w^k)\|_W = 0.$$

*Proof* Set  $p^k = p(w^k)$  and assume  $p^k \not\rightarrow 0$ . Then there exist  $\varepsilon > 0$  and an infinite set  $K$  with  $\|p^k\|_W \geq \varepsilon$  for all  $k \in K$ .

By construction we have that  $f(w^k)$  is monotonically decreasing and by assumption the sequence is bounded below. For all  $k \in K$ , we obtain

$$f(w^k) - f(w^{k+1}) \geq \frac{\gamma}{\sigma_k} \|P_S(w^k + \sigma_k s^k) - w^k\|_W^2 \geq \gamma \sigma_k \|p^k\|_W^2 \geq \gamma \sigma_k \varepsilon^2,$$

where we have used the Armijo condition and Lemma 1.10(e). This shows  $(\sigma_k)_K \rightarrow 0$  and  $(\|P_S(w^k + \sigma_k s^k) - w^k\|_W)_K \rightarrow 0$ .

For large  $k \in K$  we have  $\sigma_k \leq 1/2$  and therefore, the Armijo condition did not hold for the step size  $\sigma = 2\sigma_k$ . Hence,

$$\begin{aligned} & -\frac{\gamma}{2\sigma_k} \|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^2 \\ & \leq f(P_S(w^k + 2\sigma_k s^k)) - f(w^k) \\ & \leq -\frac{1}{2\sigma_k} \|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^2 + L \|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{1+\alpha}. \end{aligned}$$

Here, we have applied Lemma 2.4 and the fact that by Lemma 1.10(e)

$$\|P_S(w^k + 2\sigma_k s^k) - w^k\|_W \leq 2 \|P_S(w^k + \sigma_k s^k) - w^k\|_W \xrightarrow{K \ni k \rightarrow \infty} 0.$$

Hence,

$$\frac{1-\gamma}{2\sigma_k} \|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^2 \leq L \|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{1+\alpha}.$$

From this we derive

$$(1-\gamma) \|p^k\|_W \|P_S(w^k + 2\sigma_k s^k) - w^k\|_W \leq L \|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^{1+\alpha}.$$

Hence,

$$(1-\gamma)\varepsilon \leq L \|P_S(w^k + 2\sigma_k s^k) - w^k\|_W^\alpha \leq L 2^\alpha \|P_S(w^k + \sigma_k s^k) - w^k\|_W^\alpha \xrightarrow{K \ni k \rightarrow \infty} 0.$$

This is a contradiction.

A careful choice of search directions will allow to extend the convergence theory to more general classes of projected descent algorithms. For instance, in finite dimensions,  $q$ -superlinearly convergent projected Newton methods and their globalization are investigated in [14, 84]. In an  $L^2$  setting, the superlinear convergence of projected Newton methods was investigated by Kelley and Sachs in [85].

### 2.2.3 General Optimization Problems

For more general optimization problems than we discussed so far, one usually globalizes by choosing step sizes based on an Armijo-type rule that is applied to a suitable merit function. For instance, if we consider problems of the form

$$\min_w f(w) \quad \text{s.t.} \quad e(w) = 0, \quad c(w) \in \mathcal{K},$$

with functions  $f : W \rightarrow \mathbb{R}$ ,  $e : W \rightarrow Z$ , and  $c : W \rightarrow R$ , where  $W$ ,  $Z$ , and  $R$  are Banach spaces and  $\mathcal{K} \subset R$  is a closed convex cone, a possible choice for a merit function is

$$m_\rho(w) = f(w) + \rho \|e(w)\|_Z + \rho \operatorname{dist}(c(w), \mathcal{K})$$

with penalty parameter  $\rho > 0$ . In the case of equality constraints, a global convergence result for reduced SQP methods based on this merit function is presented in [82]. Other merit functions can be constructed by taking the norm of the residual of the KKT system, the latter being reformulated as a nonsmooth operator equation, see Sect. 2.5. This residual-based type of globalization, however, does not take into account second-order information.

## 2.3 Newton-Based Methods—A Preview

To give an impression of modern Newton-based approaches for optimization problems, we first consider all these methods in the finite dimensional setting:  $W = \mathbb{R}^n$ .

### 2.3.1 Unconstrained Problems—Newton's Method

Consider

$$\min_{w \in \mathbb{R}^n} f(w) \tag{2.4}$$

with  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  twice continuously differentiable.

From analysis we know that the first-order optimality conditions are:

$$\nabla f(w) = 0. \tag{2.5}$$

Newton's method for (2.4) is obtained by applying Newton's method to (2.5).

This, again, is done by linearizing  $G = \nabla f$  about the current iterate  $w^k$  and equating this linearization to zero:

$$G(w^k) + G'(w^k)s^k = 0, \quad w^{k+1} = w^k + s^k.$$

It is well-known—and will be proved later in a much more general context—that Newton's method converges q-superlinearly if  $G$  is  $C^1$  and  $G'(\bar{w})$  is invertible.

### 2.3.2 Simple Constraints

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be  $C^2$  and let  $S \subset \mathbb{R}^n$  be a nonempty closed convex set.

We consider the problem

$$\min_{w \in \mathbb{R}^n} f(w) \quad \text{s.t.} \quad w \in S.$$

The optimality conditions, written in a form that directly generalizes to a Banach space setting, are:  $w = \bar{w}$  solves

$$w \in S, \quad \nabla f(w)^T (v - w) \geq 0 \quad \forall v \in S. \quad (2.6)$$

This is a *Variational Inequality*, which we abbreviate  $\text{VI}(\nabla f, S)$ .

Note that the necessity of  $\text{VI}(\nabla f, S)$  can be derived very easily: For all  $v \in S$ , the line segment  $\{\bar{w} + t(v - \bar{w}) : 0 \leq t \leq 1\}$  connecting  $\bar{w}$  and  $v$  is contained in  $S$  (convexity) and therefore, the function

$$\phi(t) := f(\bar{w} + t(v - \bar{w}))$$

is nondecreasing at  $t = 0$ :

$$0 \leq \phi'(0) = \nabla f(\bar{w})^T (v - \bar{w}).$$

Similarly, in the Banach space setting, we will have that  $w = \bar{w}$  solves

$$w \in S, \quad \langle f'(w), v - w \rangle_{W^*, W} \geq 0 \quad \forall v \in S$$

with  $S \subset W$  closed, convex and  $f' : W \rightarrow W^*$ .

Note that if  $S = \mathbb{R}^n$ , then (2.6) is equivalent to (2.5).

#### 2.3.2.1 Nonsmooth Reformulation Approach and Generalized Newton Methods

In the development of projected descent methods we already used the important fact that the VI (2.6) is equivalent to

$$w - P_S(w - \theta \nabla f(w)) = 0, \quad (2.7)$$

where  $\theta > 0$  is fixed.

*Example 2.3* If  $S$  is a box, i.e.,

$$S = [a_1, b_1] \times \cdots \times [a_n, b_n],$$

then  $P_S(w)$  can be computed very easily as follows:

$$P_S(w)_i = \max(a_i, \min(w_i, b_i)).$$

It is instructive (and not difficult) to check the equivalence of (2.6) and (2.7) by hand.

The function

$$\Phi(w) := w - P_S(w - \theta \nabla f(w))$$

is locally Lipschitz continuous ( $P_S$  is non-expansive and  $\nabla f$  is  $C^1$ ), but cannot be expected to be differentiable. Therefore, *at a first sight*, Newton's method is *not* applicable.

However, a second look shows that  $\Phi$  has nice properties if  $S$  is sufficiently nice. To be more concrete, let

$$S = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

be a box in the following. Then  $\Phi$  is *piecewise* continuously differentiable, i.e., it consists of finitely many  $C^1$ -pieces  $\Phi^j : \mathbb{R}^n \rightarrow \mathbb{R}^n$ ,  $j = 1, \dots, m$ . More precisely, every component  $\Phi_i$  of  $\Phi$  consists of three pieces:

$$w_i - a_i, \quad w_i - b_i, \quad w_i - (w_i - \theta \nabla f(w)_i) = \theta \nabla f(w)_i,$$

hence  $\Phi$  consists of (at most)  $3^n$  pieces  $\Phi^j$ .

Denote by

$$A(w) = \left\{ j : \Phi^j(w) = \Phi(w) \right\}$$

the active indices at  $w$  and by

$$I(w) = \left\{ j : \Phi^j(w) \neq \Phi(w) \right\}$$

the set of inactive indices at  $w$ .

By continuity,  $I(w) \supset I(\bar{w})$  in a neighborhood  $U$  of  $\bar{w}$ . Now consider the following

**Algorithm 2.5** (Generalized Newton's method for piecewise  $C^1$  equations)

0. Chose  $w^0$  (sufficiently close to  $\bar{w}$ ).

For  $k = 0, 1, 2, \dots$ :

1. Choose  $M_k \in \{(\Phi^j)'(w^k) : j \in A(w^k)\}$  and solve

$$M_k s^k = -\Phi(w^k).$$

2. Set  $w^{k+1} = w^k + s^k$ .

For  $w^k$  close to  $\bar{w}$ , we have  $A(w^k) \subset A(\bar{w})$  and thus  $s^k$  is the Newton step for the  $C^1$  equation

$$\Phi^{j_k}(w) = 0,$$

where  $j_k \in A(w^k) \subset A(\bar{w})$  is the active index with  $M_k = (\Phi^{j_k})'(w^k)$ .

Therefore, if all the finitely many Newton processes for

$$\Phi^j(w) = 0, \quad j \in A(\bar{w})$$

converge locally fast, our generalized Newton's method converges locally fast, too. In particular, this is the case if  $f$  is  $C^2$  and all  $(\Phi^j)'(\bar{w})$ ,  $j \in A(\bar{w})$ , are invertible.

### 2.3.2.2 SQP Methods

A further appealing idea is to obtain an iterative method by linearizing  $\nabla f$  in  $\text{VI}(\nabla f, S)$  about the current iterate  $w^k \in S$ :

$$w \in S, \quad (\nabla f(w^k) + \nabla^2 f(w^k)(w - w^k))^T (v - w) \geq 0 \quad \forall v \in S.$$

The solution  $w^{k+1}$  of this VI is then the new iterate. The resulting method, of course, can just as well be formulated for general variational inequalities  $\text{VI}(F, S)$  with  $C^1$ -function  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . We obtain the following method:

**Algorithm 2.6** (Josephy-Newton method for  $\text{VI}(F, S)$ )

0. Choose  $w^0 \in S$  (sufficiently close to the solution  $\bar{w}$  of  $\text{VI}(F, S)$ ).

For  $k = 0, 1, 2, \dots$ :

1. STOP if  $w^k$  solves  $\text{VI}(F, S)$  (holds if  $w^k = w^{k-1}$ ).
2. Compute the solution  $w^{k+1}$  of

$$\text{VI}(F(w^k) + F'(w^k)(\cdot - w^k), S) :$$

$$w \in S, \quad (F(w^k) + F'(w^k)(w - w^k))^T (v - w) \geq 0 \quad \forall v \in S$$

that is closest to  $w^k$ .

In the case  $F = \nabla f$ , it is easily seen that  $\text{VI}(\nabla f(w^k) + \nabla^2 f(w^k)(\cdot - w^k), S)$  is the first-order necessary optimality condition of the problem

$$\min_{w \in \mathbb{R}^n} \nabla f(w^k)^T (w - w^k) + \frac{1}{2} (w - w^k)^T \nabla^2 f(w^k) (w - w^k) \quad \text{s.t.} \quad w \in S.$$

The objective function is quadratic, and in the case of box constraints, we have a box-constrained quadratic program.

This is why this approach is called sequential quadratic programming.

**Algorithm 2.7** (Sequential Quadratic Programming for simple constraints)

0. Chose  $w^0 \in \mathbb{R}^n$  (sufficiently close to  $\bar{w}$ ).

For  $k = 0, 1, 2, \dots$ :

1. Compute the first-order optimal point  $s^k$  of the QP

$$\min_{s \in \mathbb{R}^n} \nabla f(w^k)^T s + \frac{1}{2} s^T \nabla^2 f(w^k) s \quad \text{s.t.} \quad w^k + s \in S$$

that is closest to 0.

2. Set  $w^{k+1} = w^k + s^k$ .

The local convergence analysis of the Josephy-Newton method is intimately connected with the locally unique and Lipschitz-stable solvability of the parameterized VI

$$\text{VI}(F(\bar{w}) + F'(\bar{w})(\cdot - \bar{w}) - p, S) :$$

$$w \in S, \quad (F(\bar{w}) + F'(\bar{w})(w - \bar{w}) - p)^T (v - w) \geq 0 \quad \forall v \in S.$$

In fact, if there exist open neighborhoods  $U_p \subset \mathbb{R}^n$  of 0,  $U_w \subset \mathbb{R}^n$  of  $\bar{w}$ , and a Lipschitz continuous function  $U_p \ni p \mapsto w(p) \in U_w$  such that  $w(p)$  is the unique solution of  $\text{VI}(F(\bar{w}) + F'(\bar{w})(\cdot - \bar{w}) - p, S)$  in  $U_w$ , then  $\text{VI}(F, S)$  is called *strongly regular* at  $\bar{w}$ .

As we will see, strong regularity implies local q-superlinear convergence of the above SQP method if  $f$  is  $C^2$ .

In the case  $S = \mathbb{R}^n$  we have

$$\text{VI}(F, \mathbb{R}^n): \quad F(w) = 0.$$

Hence, the Josephy-Newton method for  $\text{VI}(F, \mathbb{R}^n)$  is Newton's method for  $F(w) = 0$ . Furthermore, from

$$\text{VI}(F(\bar{w}) + F'(\bar{w})(\cdot - \bar{w}) + p, \mathbb{R}^n): \quad F(\bar{w}) + F'(\bar{w})(w - \bar{w}) + p = 0$$

we see that in this case strong regularity is the same as the invertibility of  $F'(\bar{w})$ .

### 2.3.3 General Inequality Constraints

We now consider general nonlinear optimization in  $\mathbb{R}^n$ :

$$\min_{w \in \mathbb{R}^n} f(w) \quad \text{s.t.} \quad e(w) = 0, \quad c(w) \leq 0, \quad (2.8)$$

where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $e : \mathbb{R}^n \rightarrow \mathbb{R}^p$ , and  $c : \mathbb{R}^n \rightarrow \mathbb{R}^m$  are  $C^2$  and  $\leq$  is meant component-wise.

Denote by

$$L(w, \lambda, \mu) = f(w) + \lambda^T c(w) + \mu^T e(w)$$

the Lagrange function of problem (2.8).

Under a constraint qualification (CQ), the first-order optimality conditions (KKT conditions) hold at  $(\bar{w}, \bar{\lambda}, \bar{\mu})$ :

$$\begin{aligned} \nabla_w L(\bar{w}, \bar{\lambda}, \bar{\mu}) &= \nabla f(\bar{w}) + c'(\bar{w})^T \bar{\lambda} + e'(\bar{w})^T \bar{\mu} = 0, \\ \bar{\lambda} &\geq 0, \quad \nabla_{\lambda} L(\bar{w}, \bar{\lambda}, \bar{\mu})^T (z - \bar{\lambda}) = c(\bar{w})^T (z - \bar{\lambda}) \leq 0 \quad \forall z \geq 0, \\ \nabla_{\mu} L(\bar{w}, \bar{\lambda}, \bar{\mu}) &= e(\bar{w}) = 0. \end{aligned} \quad (2.9)$$

*Remark 2.1*

- (a) An easy way to remember these conditions is the following:  $(\bar{w}, \bar{\lambda}, \bar{\mu})$  is a first-order saddle point of  $L$  on  $\mathbb{R}^n \times (\mathbb{R}_+^m \times \mathbb{R}^p)$ .
- (b) The second equation can be equivalently written in the following way:

$$\bar{\lambda} \geq 0, \quad c(\bar{w}) \leq 0, \quad c(\bar{w})^T \bar{\lambda} = 0.$$

The KKT system consists of two equations and the variational inequality  $\text{VI}(-c(\bar{w}), \mathbb{R}_+^m)$ . This is a VI w.r.t.  $\lambda$  that is parameterized by  $\bar{w}$ . Also, since equations are special cases of variational inequalities, we have that (2.9) is in fact the same as  $\text{VI}(-\nabla L, \mathbb{R}^n \times \mathbb{R}_+^m \times \mathbb{R}^p)$ .

We now can use the same techniques as for simple constraints.

### 2.3.3.1 Nonsmooth Reformulation Approach and Generalized Newton Methods

Using the projection, we rewrite the VI in (2.9) as a nonsmooth equation:

$$\Phi(w, \lambda) := \lambda - P_{\mathbb{R}_+^m}(\lambda + \theta c(w)) = 0,$$

where  $\theta > 0$  is fixed. The reformulated KKT system

$$G(w, \lambda, \mu) := \begin{pmatrix} \nabla f(w) + c'(w)^T \lambda + e'(w)^T \mu \\ \Phi(w, \lambda) \\ e(w) \end{pmatrix} = 0$$

is a system of  $n + m + p$  equations in  $n + m + p$  unknowns.

The function on the left is  $C^1$ , except for the second row which is piecewise  $C^1$ . Therefore, the generalized Newton's method for piecewise smooth equations (Algorithm 2.5) can be applied. It is q-superlinearly convergent if  $(G^j)'(\bar{w}, \bar{\lambda}, \bar{\mu})$  is invertible for all active indices  $j \in A(\bar{w}, \bar{\lambda}, \bar{\mu})$ .

### 2.3.3.2 SQP Methods

As already observed, the KKT system is identical to  $\text{VI}(-\nabla L, \mathbb{R}^n \times \mathbb{R}_+^m \times \mathbb{R}^p)$ .

The SQP method for (2.8) can now be derived as in the simply constrained case by linearizing  $-\nabla L$  about the current iterate  $x^k := (w^k, \lambda^k, \mu^k)$ : The resulting subproblem is  $\text{VI}(-\nabla L(x^k) - \nabla L(x^k)(\cdot - x^k), \mathbb{R}^n \times \mathbb{R}_+^m \times \mathbb{R}^p)$ , or, in detail:

$$\begin{aligned} \nabla_w L(x^k) + \nabla_{wx} L(x^k)(x - x^k) &= 0 \\ \lambda &\geq 0, \quad (c(w^k) + c'(w^k)(w - w^k))^T(z - \lambda) \leq 0 \quad \forall z \geq 0, \\ e(w^k) + e'(w^k)(w - w^k) &= 0. \end{aligned} \quad (2.10)$$

As in the simply constrained case, it is straightforward to verify that (2.10) is equivalent to the KKT conditions of the following quadratic program:

$$\begin{aligned} \min_w \quad & \nabla f(w^k)^T(w - w^k) + \frac{1}{2}(w - w^k)^T \nabla_{ww} L(x^k)(w - w^k) \\ \text{s.t.} \quad & e(w^k) + e'(w^k)(w - w^k) = 0, \quad c(w^k) + c'(w^k)(w - w^k) \leq 0. \end{aligned}$$

## 2.4 Generalized Newton Methods

We have seen in the previous section that we can reformulate KKT systems of finite dimensional optimization problems as nonsmooth equations. This also holds true for PDE-constrained optimization with inequality constraints, as we will sketch below. In finite dimensions, we observed that a projection-based reformulation results in a piecewise  $C^1$ -function to which a Newton-type method can be applied. In order to develop similar approaches in a function space framework, it is important to find minimum requirements on the operator  $G : X \rightarrow Y$  that allow us to develop and analyze a Newton-type method for the (possibly nonsmooth) operator equation

$$G(x) = 0. \quad (2.11)$$

### 2.4.1 Motivation: Application to Optimal Control

We will show now that the optimality conditions of constrained optimal control problems can be converted to nonsmooth operator equations.

Consider the following elliptic optimal control problem:

$$\begin{aligned} \min_{y \in H_0^1(\Omega), u \in L^2(\Omega)} \quad & J(y, u) \stackrel{\text{def}}{=} \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 \\ \text{s.t.} \quad & Ay = u, \quad \beta_l \leq u \leq \beta_r. \end{aligned}$$

Here,  $y \in H_0^1(\Omega)$  is the state, which is defined on the open bounded domain  $\Omega \subset \mathbb{R}^n$ , and  $u \in L^2(\Omega)$  is the control. Furthermore,  $A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega) = H_0^1(\Omega)^*$  is a (for simplicity) linear elliptic partial differential operator, e.g.,  $A = -\Delta$ .

The control is subject to pointwise bounds  $\beta_l < \beta_r$ . The objective is to drive the state as close to  $y_d \in L^2(\Omega)$  as possible. The second part of the objective function penalizes excessive control costs; the parameter  $\alpha > 0$  is typically small.

We eliminate the state  $y$  via the state equation, i.e.,  $y = y(u) = A^{-1}u$ , and obtain the reduced problem

$$\begin{aligned} \min_{u \in L^2(\Omega)} \quad & \hat{J}(u) \stackrel{\text{def}}{=} J(y(u), u) \stackrel{\text{def}}{=} \frac{1}{2} \|A^{-1}u - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 \\ \text{s.t.} \quad & \beta_l \leq u \leq \beta_r. \end{aligned}$$

The feasible set is

$$S = \left\{ u \in L^2(\Omega) : \beta_l \leq u \leq \beta_r \right\}$$

and the optimality conditions are given by

$$\text{VI}(\nabla \hat{J}, S) : \quad u \in S, \quad (\nabla \hat{J}(u), v - u)_{L^2(\Omega)} \geq 0 \quad \forall v \in S.$$

Using the projection  $P_S(u) = P_{[\beta_l, \beta_r]}(u(\cdot))$  onto  $S$ , this can be rewritten as

$$\Phi(u) \stackrel{\text{def}}{=} u - P_{[\beta_l, \beta_r]}(u - \theta \nabla \hat{J}(u)) = 0,$$

where  $\theta > 0$  is fixed. This is a nonsmooth operator equation in the space  $L^2(\Omega)$ . Hence, we were able to convert the optimality system into a nonsmooth operator equation.

### 2.4.2 A General Superlinear Convergence Result

Consider the operator equation (2.11) with  $G : X \rightarrow Y$ ,  $X, Y$  Banach spaces.

A general Newton-type method for (2.11) has the form

**Algorithm 2.8** (Generalized Newton's method)

0. Choose  $x^0 \in X$  (sufficiently close to the solution  $\bar{x}$ ).

For  $k = 0, 1, 2, \dots$ :

1. Choose an invertible operator  $M_k \in \mathcal{L}(X, Y)$ .
2. Obtain  $s^k$  by solving

$$M_k s = -G(x^k), \tag{2.12}$$

and set  $x^{k+1} = x^k + s^k$ .

We now investigate the generated sequence  $(x^k)$  in a neighborhood of a solution  $\bar{x} \in X$ , i.e.,  $G(\bar{x}) = 0$ .

For the distance  $d^k := x^k - \bar{x}$  to the solution we have

$$\begin{aligned} M_k d^{k+1} &= M_k(x^{k+1} - \bar{x}) = M_k(x^k + s^k - \bar{x}) = M_k d^k - G(x^k) \\ &= G(\bar{x}) + M_k d^k - G(x^k). \end{aligned}$$

Hence, we obtain:

1.  $(x^k)$  converges q-linearly to  $\bar{x}$  with rate  $\gamma \in (0, 1)$  iff

$$\|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X \leq \gamma \|d^k\|_X \quad \forall k \text{ with } \|d^k\|_X \text{ suff. small.} \quad (2.13)$$

2.  $(x^k)$  converges q-superlinearly to  $\bar{x}$  iff

$$\|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X = o(\|d^k\|_X) \quad \text{for } \|d^k\|_X \rightarrow 0. \quad (2.14)$$

3.  $(x^k)$  converges with q-order  $1 + \alpha > 1$  to  $\bar{x}$  iff

$$\|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X = O(\|d^k\|_X^{1+\alpha}) \quad \text{for } \|d^k\|_X \rightarrow 0. \quad (2.15)$$

In 1., the estimate is meant uniformly in  $k$ , i.e., there exists  $\delta_\gamma > 0$  such that

$$\|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X \leq \gamma \|d^k\|_X \quad \forall k \text{ with } \|d^k\|_X < \delta_\gamma.$$

In 2.,  $o(\|d^k\|_X)$  is meant uniformly in  $k$ , i.e., for all  $\eta \in (0, 1)$ , there exists  $\delta_\eta > 0$  such that

$$\|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X \leq \eta \|d^k\|_X \quad \forall k \text{ with } \|d^k\|_X < \delta_\eta.$$

The condition in 3. and those stated below are meant similarly.

It is convenient, and often done, to split the smallness assumption on

$$\|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X$$

in two parts:

1. *Regularity condition:*

$$\|M_k^{-1}\|_{Y \rightarrow X} \leq C \quad \forall k \geq 0. \quad (2.16)$$

2. *Approximation condition:*

$$\|G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k\|_Y = o(\|d^k\|_X) \quad \text{for } \|d^k\|_X \rightarrow 0 \quad (2.17)$$

or

$$\|G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k\|_Y = O(\|d^k\|_X^{1+\alpha}) \quad \text{for } \|d^k\|_X \rightarrow 0. \quad (2.18)$$

We obtain

**Theorem 2.9** Consider the operator equation (2.11) with  $G : X \rightarrow Y$ , where  $X$  and  $Y$  are Banach spaces. Let  $(x^k)$  be generated by the generalized Newton method (Algorithm 2.8). Then:

1. If  $x^0$  is sufficiently close to  $\bar{x}$  and (2.13) holds then  $x^k \rightarrow \bar{x}$   $q$ -linearly with rate  $\gamma$ .
2. If  $x^0$  is sufficiently close to  $\bar{x}$  and (2.14) (or (2.16) and (2.17)) holds then  $x^k \rightarrow \bar{x}$   $q$ -superlinearly.
3. If  $x^0$  is sufficiently close to  $\bar{x}$  and (2.15) holds (or (2.16) and (2.18)) then  $x^k \rightarrow \bar{x}$   $q$ -superlinearly with order  $1 + \alpha$ .

*Proof* 1. Let  $\delta > 0$  be so small that (2.13) holds for all  $x^k$  with  $\|d^k\|_X < \delta$ . Then, for  $x^0$  satisfying  $\|x^0 - \bar{x}\|_X < \delta$ , we have

$$\begin{aligned} \|x^1 - \bar{x}\|_X &= \|d^1\|_X = \|M_0^{-1}(G(\bar{x} + d^0) - G(\bar{x}) - M_0 d^0)\|_X \leq \gamma \|d^0\|_X \\ &= \gamma \|x^0 - \bar{x}\|_X < \delta. \end{aligned}$$

Inductively, let  $\|x^k - \bar{x}\|_X < \delta$ . Then

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_X &= \|d^{k+1}\|_X = \|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X \\ &\leq \gamma \|d^k\|_X = \gamma \|x^k - \bar{x}\|_X < \delta. \end{aligned}$$

Hence, we have

$$\|x^{k+1} - \bar{x}\|_X \leq \gamma \|x^k - \bar{x}\|_X \quad \forall k \geq 0.$$

2. Fix  $\gamma \in (0, 1)$  and let  $\delta > 0$  be so small that (2.13) holds for all  $x^k$  with  $\|d^k\|_X < \delta$ . Then, for  $x^0$  satisfying  $\|x^0 - \bar{x}\|_X < \delta$ , we can apply 1. to conclude  $x^k \rightarrow \bar{x}$  with rate  $\gamma$ .

Now, (2.14) immediately yields

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_X &= \|d^{k+1}\|_X = \|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X = o(\|d^k\|_X) \\ &= o(\|x^k - \bar{x}\|_X) \quad (k \rightarrow \infty). \end{aligned}$$

3. As in 2, but now

$$\begin{aligned} \|x^{k+1} - \bar{x}\|_X &= \|d^{k+1}\|_X = \|M_k^{-1}(G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k)\|_X = O(\|d^k\|_X^{1+\alpha}) \\ &= O(\|x^k - \bar{x}\|_X^{1+\alpha}) \quad (k \rightarrow \infty). \end{aligned}$$

We emphasize that an inexact solution of the Newton system (2.12) can be interpreted as a solution of the same system, but with  $M_k$  replaced by a perturbed operator  $\tilde{M}_k$ . Since the condition (2.14) (or the conditions (2.16) and (2.17)) remain valid if  $M_k$  is replaced by a perturbed operator  $\tilde{M}_k$  and the perturbation is sufficiently small, we see that the fast convergence of the generalized Newton's method is not affected if the system is solved inexactly and the accuracy of the solution

is controlled suitably. The Dennis-Moré condition [36] characterizes perturbations that are possible without destroying q-superlinear convergence.

We will now specialize on particular instances of generalized Newton methods. The first one, of course, is Newton's method itself.

### 2.4.3 The Classical Newton's Method

In the classical Newton's method, we assume that  $G$  is continuously F-differentiable and choose  $M_k = G'(x^k)$ .

The regularity condition then reads

$$\|G'(x^k)^{-1}\|_{Y \rightarrow X} \leq C \quad \forall k \geq 0.$$

By Banach's Lemma (asserting continuity of  $M \mapsto M^{-1}$ ), this holds true if  $G'$  is continuous at  $\bar{x}$  and

$$G'(\bar{x}) \in \mathcal{L}(X, Y) \quad \text{is continuously invertible.}$$

This condition is the textbook regularity requirement in the analysis of Newton's method.

Fréchet differentiability at  $\bar{x}$  means

$$\|G(\bar{x} + d^k) - G(\bar{x}) - G'(\bar{x})d^k\|_Y = o(\|d^k\|_X).$$

Now, due to the continuity of  $G'$ ,

$$\begin{aligned} & \|G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k\|_Y \\ &= \|G(\bar{x} + d^k) - G(\bar{x}) - G'(\bar{x} + d^k)d^k\|_Y \\ &\leq \|G(\bar{x} + d^k) - G(\bar{x}) - G'(\bar{x})d^k\|_Y + \|(G'(\bar{x}) - G'(\bar{x} + d^k))d^k\|_Y \\ &\leq o(\|d^k\|_X) + \|G'(\bar{x}) - G'(\bar{x} + d^k)\|_{X \rightarrow Y} \|d^k\|_X \\ &= o(\|d^k\|_X) \quad \text{for } \|d^k\|_X \rightarrow 0. \end{aligned}$$

Therefore, we have proved the superlinear approximation condition.

If  $G'$  is  $\alpha$ -order Hölder continuous near  $\bar{x}$ , we even obtain the approximation condition of order  $1 + \alpha$ . In fact, let  $L > 0$  be the modulus of Hölder continuity. Then

$$\begin{aligned} & \|G(\bar{x} + d^k) - G(\bar{x}) - M_k d^k\|_Y \\ &= \|G(\bar{x} + d^k) - G(\bar{x}) - G'(\bar{x} + d^k)d^k\|_Y \\ &= \left\| \int_0^1 (G'(\bar{x} + td^k) - G'(\bar{x} + d^k))d^k dt \right\|_Y \end{aligned}$$

$$\begin{aligned}
&\leq \int_0^1 \|G'(\bar{x} + td^k) - G'(\bar{x} + d^k)\|_{X \rightarrow Y} dt \|d^k\|_X \\
&\leq L \int_0^1 (1-t)^\alpha \|d^k\|_X^\alpha dt \|d^k\|_X = \frac{L}{1+\alpha} \|d^k\|_X^{1+\alpha} = O(\|d^k\|_X^{1+\alpha}).
\end{aligned}$$

Summarizing, we have proved the following

**Corollary 2.1** *Let  $G : X \rightarrow Y$  be a continuously  $F$ -differentiable operator between Banach spaces and assume that  $G'(\bar{x})$  is continuously invertible at the solution  $\bar{x}$ . Then Newton's method (i.e., Algorithm 2.8 with  $M_k = G'(x^k)$  for all  $k$ ) converges locally  $q$ -superlinearly. If, in addition,  $G'$  is  $\alpha$ -order Hölder continuous near  $\bar{x}$ , the order of convergence is  $1 + \alpha$ .*

*Remark 2.2* The choice of  $M_k$  in the classical Newton's method,  $M_k = G'(x^k)$ , is *point-based*, since it depends on the point  $x^k$ .

#### 2.4.4 Generalized Differential and Semismoothness

If  $G$  is nonsmooth, the question arises if a suitable substitute for  $G'$  can be found. We follow [134, 136] here; a related approach can be found in [87] and [69]. Thinking at subgradients of convex functions, which are set-valued, we consider set-valued generalized differentials  $\partial G : X \rightrightarrows \mathcal{L}(X, Y)$ . Then we will choose  $M_k$  point-based, i.e.,

$$M_k \in \partial G(x^k).$$

If we want every such choice  $M_k$  to satisfy the superlinear approximation condition, then we have to require

$$\sup_{M \in \partial G(\bar{x}+d)} \|G(\bar{x}+d) - G(\bar{x}) - Md\|_Y = o(\|d\|_X) \quad \text{for } \|d\|_X \rightarrow 0.$$

This approximation property is called semismoothness [134, 136]:

**Definition 2.1** (Semismoothness) Let  $G : X \rightarrow Y$  be a continuous operator between Banach spaces. Furthermore, let be given the set-valued mapping  $\partial G : X \rightrightarrows Y$  with nonempty images (which we will call generalized differential in the sequel). Then

(a)  $G$  is called  $\partial G$ -semismooth at  $x \in X$  if

$$\sup_{M \in \partial G(x+d)} \|G(x+d) - G(x) - Md\|_Y = o(\|d\|_X) \quad \text{for } \|d\|_X \rightarrow 0.$$

(b)  $G$  is called  $\partial G$ -semismooth of order  $\alpha > 0$  at  $x \in X$  if

$$\sup_{M \in \partial G(x+d)} \|G(x+d) - G(x) - Md\|_Y = O(\|d\|_X^{1+\alpha}) \quad \text{for } \|d\|_X \rightarrow 0.$$

**Lemma 2.6** *If  $G : X \rightarrow Y$  is continuously  $F$ -differentiable near  $x$ , then  $G$  is  $\{G'\}$ -semismooth at  $x$ . Furthermore, if  $G'$  is  $\alpha$ -order Hölder continuous near  $x$ , then  $G$  is  $\{G'\}$ -semismooth at  $x$  of order  $\alpha$ . Here,  $\{G'\}$  denotes the setvalued operator  $\{G'\} : X \rightrightarrows \mathcal{L}(X, Y)$ ,  $\{G'\}(x) = \{G'(x)\}$ .*

*Proof*

$$\begin{aligned} & \|G(x+d) - G(x) - G'(x+d)d\|_Y \\ & \leq \|G(x+d) - G(x) - G'(x)d\|_Y + \|G'(x)d - G'(x+d)d\|_Y \\ & \leq o(\|d\|_X) + \|G'(x) - G'(x+d)\|_{X \rightarrow Y} \|d\|_X = o(\|d\|_X). \end{aligned}$$

Here, we have used the definition of  $F$ -differentiability and the continuity of  $G'$ .

In the case of  $\alpha$ -order Hölder continuity we have to work a little bit more:

$$\begin{aligned} & \|G(x+d) - G(x) - G'(x+d)d\|_Y \\ & = \left\| \int_0^1 (G'(x+td) - G'(x+d))d \, dt \right\|_Y \\ & \leq \int_0^1 \|G'(x+td) - G'(x+d)\|_{X \rightarrow Y} \|d\|_X \, dt \leq \int_0^1 L(1-t)^\alpha \|d\|_X^\alpha \, dt \|d\|_X \\ & = \frac{L}{1+\alpha} \|d\|_X^{1+\alpha} = O(\|d\|_X^{1+\alpha}). \end{aligned}$$

**Example 2.4** For locally Lipschitz-continuous functions  $G : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , the standard choice for  $\partial G$  is Clarke's generalized Jacobian:

$$\partial^{cl} G(x) = \text{conv} \left\{ M : x^k \rightarrow x, \, G'(x^k) \rightarrow M, \, G \text{ differentiable at } x^k \right\}. \quad (2.19)$$

This definition is justified since  $G'$  exists almost everywhere on  $\mathbb{R}^n$  by Rademacher's theorem (which is a deep result).

**Remark 2.3** The classical definition of semismoothness for functions  $G : \mathbb{R}^n \rightarrow \mathbb{R}^m$  [105, 113] is equivalent to  $\partial^{cl} G$ -semismoothness, where  $\partial^{cl} G$  is Clarke's generalized Jacobian defined in (2.19), in connection with directional differentiability of  $G$ .

Next, we give a concrete example of a semismooth function:

**Example 2.5** Consider  $\psi : \mathbb{R} \rightarrow \mathbb{R}$ ,  $\psi(x) = P_{[a,b]}(x)$ ,  $a < b$ , then Clarke's generalized derivative is

$$\partial^{cl} \psi(x) = \begin{cases} \{0\} & x < a \text{ or } x > b, \\ \{1\} & a < x < b, \\ \text{conv}\{0, 1\} = [0, 1] & x = a \text{ or } x = b. \end{cases}$$

The  $\partial^{cl}\psi$ -semismoothness of  $\psi$  can be shown easily:

For all  $x \notin \{a, b\}$  we have that  $\psi$  is continuously differentiable in a neighborhood of  $x$  with  $\partial^{cl}\psi \equiv \{\psi'\}$ . Hence, by Lemma 2.6,  $\psi$  is  $\partial^{cl}\psi$ -semismooth at  $x$ .

For  $x = a$ , we estimate explicitly: For small  $d > 0$ , we have  $\partial^{cl}\psi(x) = \{\psi'(a + d)\} = \{1\}$  and thus

$$\sup_{M \in \partial^{cl}\psi(x+d)} |\psi(x+d) - \psi(x) - Md| = a + d - a - 1 \cdot d = 0.$$

For small  $d < 0$ , we have  $\partial^{cl}\psi(x) = \{\psi'(a + d)\} = \{0\}$  and thus

$$\sup_{M \in \partial^{cl}\psi(x+d)} |\psi(x+d) - \psi(x) - Md| = a - a - 0 \cdot d = 0.$$

Hence, the semismoothness of  $\psi$  at  $x = a$  is proved.

For  $x = b$  we can do exactly the same.

The class of semismooth operators is closed with respect to a wide class of operations, see [134]:

**Theorem 2.10** *Let  $X, Y, Z, X_i, Y_i$  be Banach spaces.*

- (a) *If the operators  $G_i : X \rightarrow Y_i$  are  $\partial G_i$ -semismooth at  $x$  then  $(G_1, G_2)$  is  $(\partial G_1, \partial G_2)$ -semismooth at  $x$ .*
- (b) *If  $G_i : X \rightarrow Y, i = 1, 2$ , are  $\partial G_i$ -semismooth at  $x$  then  $G_1 + G_2$  is  $(\partial G_1 + \partial G_2)$ -semismooth at  $x$ .*
- (c) *Let  $G_1 : Y \rightarrow Z$  and  $G_2 : X \rightarrow Y$  be  $\partial G_i$ -semismooth at  $G_2(x)$  and  $x$ , respectively. Assume that  $\partial G_1$  is bounded near  $y = G_2(x)$  and that  $G_2$  is Lipschitz continuous near  $x$ . Then  $G = G_1 \circ G_2$  is  $\partial G$ -semismooth with*

$$\partial G(x) = \{M_1 M_2 : M_1 \in \partial G_1(G_2(x)), M_2 \in \partial G_2(x)\}.$$

*Proof* Parts (a) and (b) are straightforward to prove.

Part (c):

Let  $y = G_2(x)$  and consider  $d \in X$ . Let  $h(d) = G_2(x + d) - y$ . Then, for  $\|d\|_X$  sufficiently small,

$$\|h(d)\|_Y = \|G_2(x + d) - G_2(x)\|_Y \leq L_2 \|d\|_X.$$

Hence, for  $M_1 \in \partial G_1(G_2(x + d))$  and  $M_2 \in \partial G_2(x + d)$ , we obtain

$$\begin{aligned} & \|G_1(G_2(x + d)) - G_1(G_2(x)) - M_1 M_2 d\|_Z \\ &= \|G_1(y + h(d)) - G_1(y) - M_1 h(d) + M_1(G_2(x + d) - G_2(x) - M_2 d)\|_Z \\ &\leq \|G_1(y + h(d)) - G_1(y) - M_1 h(d)\|_Z \\ &\quad + \|M_1\|_{Y \rightarrow Z} \|G_2(x + d) - G_2(x) - M_2 d\|_Y. \end{aligned}$$

By assumption, there exists  $C$  with  $\|M_1\|_{Y \rightarrow Z} \leq C$  if  $\|d\|_X$  is sufficiently small. Taking the supremum with respect to  $M_1, M_2$  and using the semismoothness of  $G_1$  and  $G_2$  gives

$$\begin{aligned}
 & \sup_{M \in \partial G(x+d)} \|G(x+d) - G(x) - Md\|_Z \\
 & \leq \sup_{M_1 \in \partial G_1(y+h(d))} \|G_1(y+h(d)) - G_1(y) - M_1 h(d)\|_Z \\
 & \quad + C \sup_{M_2 \in \partial G_2(x+d)} \|G_2(x+d) - G_2(x) - M_2 d\|_Y \\
 & = o(\|h(d)\|_Y) + o(\|d\|_X) = o(\|d\|_X).
 \end{aligned}$$

### 2.4.5 Semismooth Newton Methods

The semismoothness concept ensures the approximation property required for generalized Newton methods. In addition, we need a regularity condition, which can be formulated as follows:

There exist constants  $C > 0$  and  $\delta > 0$  such that

$$\|M^{-1}\|_{Y \rightarrow X} \leq C \quad \forall M \in \partial G(x) \quad \forall x \in X, \quad \|x - \bar{x}\|_X < \delta. \quad (2.20)$$

Under these two assumptions, the following generalized Newton method for semismooth operator equations is q-superlinearly convergent:

**Algorithm 2.11** (Semismooth Newton's method)

0. Choose  $x^0 \in X$  (sufficiently close to the solution  $\bar{x}$ ).

For  $k = 0, 1, 2, \dots$ :

1. Choose  $M_k \in \partial G(x^k)$ .
2. Obtain  $s^k$  by solving

$$M_k s^k = -G(x^k),$$

and set  $x^{k+1} = x^k + s^k$ .

The local convergence result is a simple corollary of Theorem 2.9:

**Theorem 2.12** *Let  $G : X \rightarrow Y$  be continuous and  $\partial G$ -semismooth at a solution  $\bar{x}$  of (2.11). Furthermore, assume that the regularity condition (2.20) holds. Then there exists  $\delta > 0$  such that for all  $x^0 \in X$ ,  $\|x^0 - \bar{x}\|_X < \delta$ , the semismooth Newton method (Algorithm 2.11) converges q-superlinearly to  $\bar{x}$ .*

*If  $G$  is  $\partial G$ -semismooth of order  $\alpha > 0$  at  $\bar{x}$ , then the convergence is of order  $1 + \alpha$ .*

*Proof* The regularity condition (2.20) implies (2.16) as long as  $x^k$  is close enough to  $\bar{x}$ . Furthermore, the semismoothness of  $G$  at  $\bar{x}$  ensures the q-superlinear approximation condition (2.17).

In the case of  $\alpha$ -order semismoothness, the approximation condition (2.18) with order  $1 + \alpha$  holds.

Therefore, Theorem 2.9 yields the assertions.

### 2.4.5.1 Semismooth Newton Method for Finite Dimensional KKT Systems

At the beginning of this chapter we have seen that we can rewrite the KKT conditions of the NLP

$$\min f(w) \quad \text{s.t.} \quad e(w) = 0, \quad c(w) \leq 0$$

in the following form:

$$G(x) \stackrel{\text{def}}{=} \begin{pmatrix} \nabla_w L(w, \lambda, \mu) \\ \lambda - P_{\mathbb{R}_+^p}(\lambda + c(w)) \\ e(w) \end{pmatrix} = 0,$$

where we have set  $x = (w, \lambda, \mu)$ . With the developed results, we now can show that the function  $G$  on the left is semismooth. In fact,  $\nabla_w L$  is  $\{\nabla_{wx} L\}$ -semismooth and  $e$  is  $\{e'\}$ -semismooth.

Furthermore, as shown above,  $\psi(t) = P_{\mathbb{R}_+}(t)$  is  $\partial^{cl}\psi$ -semismooth with

$$\partial^{cl}\psi(t) = \{0\} \quad (t < 0), \quad \partial^{cl}\psi(t) = \{1\} \quad (t > 0), \quad \partial^{cl}\psi(0) = [0, 1].$$

Hence, by the sum and chain rules from Theorem 2.10

$$\phi_i(w, \lambda_i) \stackrel{\text{def}}{=} \lambda_i - P_{\mathbb{R}_+}(\lambda_i + c_i(w)),$$

is semismooth with respect to

$$\partial\phi_i(w, \lambda_i) := \left\{ (-g_i c'_i(w), 1 - g_i) : g_i \in \partial^{cl}\psi(\lambda_i + c_i(w)) \right\}.$$

Therefore, the operator  $\Phi(w, \lambda) = \lambda - P_{\mathbb{R}_+^p}(\lambda + c(w))$  is semismooth with respect to

$$\partial\Phi(w, \lambda) := \left\{ (-D_g c'_i(w), I - D_g) : D_g = \text{diag}(g_i), g_i \in \partial^{cl}\psi(\lambda_i + c_i(w)) \right\}.$$

This shows that  $G$  is semismooth with respect to

$$\partial G(x) \stackrel{\text{def}}{=} \left\{ \begin{pmatrix} \nabla_{ww} L(x) & c'(w)^T & e'(w)^T \\ -D_g c'(w) & I - D_g & 0 \\ e'(w) & 0 & 0 \end{pmatrix}; \right. \\ \left. D_g = \text{diag}(g_i), g_i \in \partial^{cl}\psi(\lambda_i + c_i(w)) \right\}.$$

Under the regularity condition

$$\|M^{-1}\| \leq C \quad \forall M \in \partial G(x) \quad \forall x, \quad \|x - \bar{x}\| < \delta,$$

where  $\bar{x} = (\bar{w}, \bar{\lambda}, \bar{\mu})$  is a KKT triple, Theorem 2.12 is applicable and yields the q-superlinear convergence of the semismooth Newton method.

*Remark 2.4* The compact-valuedness and the upper semicontinuity of Clarke's generalized differential [34] even allows to reduce the regularity condition to

$$\|M^{-1}\| \leq C \quad \forall M \in \partial G(\bar{x}).$$

*Remark 2.5* We also can view  $G$  as a piecewise smooth equation and apply Algorithm 2.5. In fact, it can be shown that Clarke's generalized Jacobian is the convex hull of the Jacobians of all essentially active pieces [123, 134]. We are not going into details here.

### 2.4.5.2 Discussion

So far, we have looked at semismooth Newton methods from an abstract point of view. The main point, however, is to prove semismoothness for concrete instances of nonsmooth operators. In particular, we aim at reformulating KKT systems arising in PDE-constrained optimization in the same way as we did this in finite dimensions in the above section. We will investigate this in detail in Sect. 2.5.

It should be mentioned that the class of semismooth Newton method includes as a special case the *primal dual active set strategy*, see [13, 69].

## 2.5 Semismooth Newton Methods in Function Spaces

In the finite dimensional setting we have shown that variational inequalities and complementarity conditions can be reformulated as nonsmooth equations. We also described how generalized Newton methods can be developed that solve these nonsmooth equations.

In Sect. 2.4.5 we introduced the concept of semismoothness for nonsmooth operators and developed superlinearly convergent generalized Newton methods for semismooth operator equations. We now will show that, similar to the finite dimensional case, it is possible to reformulate variational inequalities and complementarity conditions in function space.

### 2.5.1 Pointwise Bound Constraints in $L^2$

Let  $\Omega \subset \mathbb{R}^n$  be measurable with measure  $0 < |\Omega| < \infty$ . If boundary spaces are considered,  $\Omega$  can also be a measurable surface, e.g., the boundary of an open Lipschitz domain, on which  $L^p$ -spaces can be defined.

We consider the problem

$$\min_{u \in L^2(\Omega)} f(u) \quad a \leq u \leq b \quad \text{a.e. on } \Omega$$

with  $f : L^2(\Omega) \rightarrow \mathbb{R}$  twice continuously F-differentiable. We can admit unilateral constraints ( $a \leq u$  or  $u \leq b$ ) just as well. To avoid distinguishing cases, we will focus on the bilateral case  $a, b \in L^\infty(\Omega)$ ,  $b - a \geq \nu > 0$  on  $\Omega$ . We also could consider problems in  $L^p(\Omega)$ ,  $p \neq 2$ . However, for the sake of compact presentation, we focus on the case  $p = 2$ , which is the most important situation.

It is convenient to transform the bounds to constant bounds, e.g., via

$$u \mapsto \frac{u - a}{b - a}.$$

Hence, we will consider the problem

$$\min_{u \in L^2(\Omega)} f(u), \quad \beta_l \leq u \leq \beta_r \quad \text{a.e. on } \Omega \quad (2.21)$$

with constants  $\beta_l < \beta_r$ . Let  $U = L^2(\Omega)$  and  $S = \{u \in L^2(\Omega) : \beta_l \leq u \leq \beta_r\}$ . We choose the standard dual pairing  $\langle \cdot, \cdot \rangle_{U^*, U} = (\cdot, \cdot)_{L^2(\Omega)}$  and then have  $U^* = U = L^2(\Omega)$ . The optimality conditions are

$$u \in S, \quad (\nabla f(u), v - u)_{L^2(\Omega)} \geq 0 \quad \forall v \in S.$$

We now use the projection  $P_S$  onto  $S$ , which is given by

$$P_S(v)(x) = P_{[\beta_l, \beta_r]}(v(x)), \quad x \in \Omega.$$

Then the optimality conditions can be written as

$$\Phi(u) := u - P_S(u - \theta \nabla f(u)) = 0, \quad (2.22)$$

where  $\theta > 0$  is arbitrary, but fixed. Note that, since  $P_S$  coincides with the pointwise projection onto  $[\beta_l, \beta_r]$ , we have

$$\Phi(u)(x) = u(x) - P_{[\beta_l, \beta_r]}(u(x) - \theta \nabla f(u)(x)).$$

Our aim now is to define a generalized differential  $\partial\Phi$  for  $\Phi$  in such a way that  $\Phi$  is semismooth.

By the chain rule and sum rule that we developed, this reduces to the question how a suitable differential for the superposition  $P_{[\beta_l, \beta_r]}(v(\cdot))$  can be defined.

### 2.5.2 Semismoothness of Superposition Operators

More generally than the superposition operator in the previous subsection, we look at the superposition operator

$$\Psi : L^p(\Omega)^m \rightarrow L^q(\Omega), \quad \Psi(w)(x) = \psi(w_1(x), \dots, w_m(x))$$

with  $1 \leq q \leq p \leq \infty$ .

Here,  $\psi : \mathbb{R}^m \rightarrow \mathbb{R}$  is assumed to be Lipschitz continuous. Since we aim at semismoothness of  $\Psi$ , it is more than natural to assume semismoothness of  $\psi$ . As differential we choose Clarke's generalized differential  $\partial^{cl}\psi$ . Now it is reasonable to define  $\partial\Psi$  in such a way that, for all  $M \in \partial\Psi(w + d)$ , the remainder

$$|(\Psi(u + d) - \Psi - Md)(x)| = |\psi(w(x) + d(x)) - \psi(w(x)) - (Md)(x)|$$

becomes pointwise small if  $|d(x)|$  is small. By semismoothness of  $\psi$ , this, again, holds true if  $(Md)(x) \in \partial^{cl}\psi(w(x) + d(x))d(x)$  is satisfied.

Hence, we define:

**Definition 2.2** Let  $\psi : \mathbb{R}^m \rightarrow \mathbb{R}$  be Lipschitz continuous and  $(\partial^{cl}\psi\cdot)$  semismooth. For  $1 \leq q \leq p \leq \infty$ , consider

$$\Psi : L^p(\Omega)^m \rightarrow L^q(\Omega), \quad \Psi(w)(x) = \psi(w_1(x), \dots, w_m(x)).$$

We define the differential

$$\partial\Psi : L^p(\Omega)^m \rightrightarrows \mathcal{L}(L^p(\Omega)^m, L^q(\Omega)),$$

$$\partial\Psi(w) = \left\{ M : Mv = g^T v, \ g \in L^\infty(\Omega)^m, \ g(x) \in \partial^{cl}\psi(w(x)) \text{ for a.a. } x \in \Omega \right\}.$$

The operator  $\Phi$  in (2.22) is naturally defined as a mapping from  $L^2(\Omega)$  to  $L^2(\Omega)$ . Therefore, since  $\nabla f$  maps to  $L^2(\Omega)$ , we would like the superposition  $v \mapsto P_{[\beta_l, \beta_r]}(v(\cdot))$  to be semismooth from  $L^2(\Omega)$  to  $L^2(\Omega)$ . But this is not true, as the following Lemma shows in great generality.

**Lemma 2.7** Let  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  be any Lipschitz continuous function that is not affine linear. Furthermore, let  $\Omega \subset \mathbb{R}^n$  be nonempty, open and bounded. Then, for all  $q \in [1, \infty)$ , the operator

$$\Psi : L^q(\Omega) \ni u \mapsto \psi(u(\cdot)) \in L^q(\Omega)$$

is not  $\partial\Psi$ -semismooth.

*Proof* Fix  $b \in \mathbb{R}$  and choose  $g_b \in \partial\psi(b)$ . Since  $\psi$  is not affine linear, there exists  $a \in \mathbb{R}$  with

$$\psi(a) \neq \psi(b) + g_b(a - b).$$

Hence,

$$\rho := |\psi(b) - \psi(a) - g_b(b - a)| > 0.$$

Let  $x_0 \in \Omega$  and  $U_\varepsilon = (x_0 - h_\varepsilon, x_0 + h_\varepsilon)^n$ ,  $h_\varepsilon = \varepsilon^{1/n}/2$ . Define

$$u(x) = a, \quad x \in \Omega, \quad d_\varepsilon(x) = \begin{cases} b - a & x \in U_\varepsilon, \\ 0 & x \notin U_\varepsilon. \end{cases}$$

Then

$$\|d_\varepsilon\|_{L^q} = \left( \int_{\Omega} |d_\varepsilon(x)|^q dx \right)^{1/q} = \left( \int_{U_\varepsilon} |b-a|^q dx \right)^{1/q} = \varepsilon^{1/q} |b-a|.$$

Choose some  $g_a \in \partial\psi(a)$  and define

$$g_\varepsilon(x) = \begin{cases} g_b & x \in U_\varepsilon, \\ g_a & x \notin U_\varepsilon. \end{cases}$$

Then  $M : L^q(\Omega) \ni v \mapsto g_\varepsilon \cdot v \in L^q(\Omega)$  is an element of  $\partial\Psi(u + d_\varepsilon)$ . Now, for all  $x \in \Omega$ ,

$$\begin{aligned} & |\psi(u(x) + d_\varepsilon(x)) - \psi(u(x)) - g_\varepsilon(x)d_\varepsilon(x)| \\ &= \begin{cases} |\psi(b) - \psi(a) - g_b(b-a)| = \rho > 0, & x \in U_\varepsilon, \\ |\psi(a) - \psi(a) - g_a(a-a)| = 0, & x \notin U_\varepsilon. \end{cases} \end{aligned}$$

Therefore,

$$\begin{aligned} & \|\Psi(u + d_\varepsilon) - \Psi(u) - Md_\varepsilon\|_{L^q} \\ &= \left( \int_{\Omega} |\psi(u(x) + d_\varepsilon(x)) - \psi(u(x)) - g_\varepsilon(x)d_\varepsilon(x)|^q dx \right)^{1/q} \\ &= \left( \int_{U_\varepsilon} \rho^q dx \right)^{1/q} = \varepsilon^{1/q} \rho = \frac{\rho}{|b-a|} \|d_\varepsilon\|_{L^q}. \end{aligned}$$

Note that the trouble is not caused by the nonsmoothness of  $\psi$ , but by the nonlinearity of  $\psi$ .

Fortunately, Ulbrich [134, 136] proved a result that helps us. See also [69]. To formulate the result in its full generality, we extend our definition of generalized differentials to superposition operators of the form  $\psi(G(\cdot))$ , where  $G$  is a continuously F-differentiable operator.

**Definition 2.3** Let  $\psi : \mathbb{R}^m \rightarrow \mathbb{R}$  be Lipschitz continuous and  $(\partial^{cl}\psi)$  semismooth. Furthermore, let  $1 \leq q \leq p \leq \infty$  be given, consider

$$\Psi_G : Y \rightarrow L^q(\Omega), \quad \Psi_G(y)(x) = \psi(G(y)(x)),$$

where  $G : Y \rightarrow L^p(\Omega)^m$  is continuously F-differentiable and  $Y$  is a Banach space. We define the differential

$$\begin{aligned} & \partial\Psi_G : Y \rightrightarrows \mathcal{L}(Y, L^q(\Omega)), \\ & \partial\Psi_G(y) = \left\{ M : Mv = g^T(G'(y)v), \quad g \in L^\infty(\Omega)^m, \right. \\ & \quad \left. g(x) \in \partial^{cl}\psi(G(y)(x)) \text{ for a.a. } x \in \Omega \right\}. \end{aligned} \tag{2.23}$$

Note that this is just the differential that we would obtain by the construction in part (c) of Theorem 2.10.

Now we can state the following semismoothness result.

**Theorem 2.13** *Let  $\Omega \subset \mathbb{R}^n$  be measurable with  $0 < |\Omega| < \infty$ . Furthermore, let  $\psi : \mathbb{R}^m \rightarrow \mathbb{R}$  be Lipschitz continuous and semismooth. Let  $Y$  be a Banach space,  $1 \leq q < p \leq \infty$ , and assume that the operator  $G : Y \rightarrow L^q(\Omega)^m$  is continuously  $F$ -differentiable and that  $G$  maps  $Y$  locally Lipschitz continuously to  $L^p(\Omega)$ . Then, the operator*

$$\Psi_G : Y \rightarrow L^q(\Omega), \quad \Psi_G(y)(x) = \psi(G(y)(x)),$$

*is  $\partial\Psi_G$ -semismooth, where  $\partial\Psi_G$  is defined in (2.23).*

*Addition: Under additional assumptions, the operator  $\Psi_G$  is  $\partial\Psi_G$ -semismooth of order  $\alpha > 0$  with  $\alpha$  appropriate.*

A proof can be found in [134, 136].

### 2.5.3 Pointwise Bound Constraints in $L^2$ Revisited

We return to the operator  $\Phi$  defined in (2.22). To be able to prove the semismoothness of  $\Phi : L^2(\Omega) \rightarrow L^2(\Omega)$  defined in (2.22), we thus need some kind of smoothing property of the mapping

$$u \mapsto u - \theta \nabla f(u).$$

Therefore, we assume that  $\nabla f$  has the following structure:

There exist  $\alpha > 0$  and  $p > 2$  such that

$$\nabla f(u) = \alpha u + H(u), \tag{2.24}$$

$H : L^2(\Omega) \rightarrow L^2(\Omega)$  continuously  $F$ -differentiable,

$H : L^2(\Omega) \rightarrow L^p(\Omega)$  locally Lipschitz continuous.

This structure is met by many optimal control problems, as illustrated in Sect. 2.5.4.

If we now choose  $\theta = 1/\alpha$ , then we have

$$\Phi(u) = u - P_{[\beta_l, \beta_r]}(u - (1/\alpha)(\alpha u + B(u))) = u - P_{[\beta_l, \beta_r]}(-(1/\alpha)B(u)).$$

Therefore, we have achieved that the operator inside the projection satisfies the requirements of Theorem 2.13. We obtain:

**Theorem 2.14** *Consider the problem (2.21) with  $\beta_l < \beta_r$  and let the continuously  $F$ -differentiable function  $f : L^2(\Omega) \rightarrow \mathbb{R}$  satisfy condition (2.24). Then, for*

$\theta = 1/\alpha$ , the operator  $\Phi$  in the reformulated optimality conditions (2.22) is  $\partial\Phi$ -semismooth with

$$\begin{aligned}\partial\Phi : L^2(\Omega) &\rightrightarrows \mathcal{L}(L^2(\Omega), L^2(\Omega)), \\ \partial\Phi(u) &= \left\{ M ; M = I + \frac{g}{\alpha} \cdot H'(u), \ g \in L^\infty(\Omega), \right. \\ &\quad \left. g(x) \in \partial^{cl} P_{[\beta_l, \beta_r]}(-(1/\alpha)H(u)(x)) \text{ for a.a. } x \in \Omega \right\}.\end{aligned}$$

Here,

$$\partial^{cl} P_{[\beta_l, \beta_r]}(t) = \begin{cases} \{0\} & t < \beta_l \text{ or } t > \beta_r, \\ \{1\} & \beta_l < t < \beta_r, \\ [0, 1] & t = \beta_l \text{ or } t = \beta_r. \end{cases}$$

*Proof* Setting  $q = 2$ ,  $\psi = P_{[\beta_l, \beta_r]}$  and  $G = -(1/\alpha)H$ , we can apply Theorem 2.13 and obtain that the operator  $\Psi_G : L^2(\Omega) \rightarrow L^2(\Omega)$  is  $\partial\Psi_G$ -semismooth. Therefore,  $\Phi = I - \Psi_G$  is  $(I - \partial\Psi_G)$ -semismooth by Theorem 2.10. Since  $\partial\Phi = I - \partial\Psi_G$ , the proof is complete.

For the applicability of the semismooth Newton method (Algorithm 2.11) we need, in addition, the following regularity condition:

$$\|M^{-1}\|_{L^2(\Omega) \rightarrow L^2(\Omega)} \leq C \quad \forall M \in \partial\Phi(u) \quad \forall u \in L^2(\Omega), \quad \|u - \bar{u}\|_{L^2(\Omega)} < \delta.$$

Sufficient conditions for this regularity assumption in the flavor of second order sufficient optimality conditions can be found in [134, 135].

## 2.5.4 Application to Optimal Control

Consider the following elliptic optimal control problem:

$$\begin{aligned}\min_{y \in H_0^1(\Omega), u \in L^2(\Omega)} J(y, u) &\stackrel{\text{def}}{=} \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 \\ \text{s.t.} \quad & Ay = r + Bu, \quad \beta_l \leq u \leq \beta_r.\end{aligned}\tag{2.25}$$

Here,  $y \in H_0^1(\Omega)$  is the state, which is defined on the open bounded domain  $\Omega \subset \mathbb{R}^n$ , and  $u \in L^2(\Omega_c)$  is the control, which is defined on the open bounded domain  $\Omega_c \subset \mathbb{R}^m$ . Furthermore,  $A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega) = H_0^1(\Omega)^*$  is a (for simplicity) linear elliptic partial differential operator, e.g.,  $A = -\Delta$ , and  $r \in H^{-1}(\Omega)$  is given.

The control operator  $B : L^{p'}(\Omega_c) \rightarrow H^{-1}(\Omega)$  is continuous and linear, with  $p' \in [1, 2)$  (the reason why we do not choose  $p' = 2$  here will become clear later; note however, that  $L^2(\Omega_c)$  is continuously embedded in  $L^{p'}(\Omega_c)$ ). For instance, distributed control on the whole domain  $\Omega$  would correspond to the choice  $\Omega_c = \Omega$

and  $B : u \in L^{p'}(\Omega) \mapsto u \in H^{-1}(\Omega)$ , where  $p'$  is chosen in such a way that  $H_0^1(\Omega)$  is continuously embedded in the dual space  $L^p(\Omega)$ ,  $p = p'/(p' - 1)$ , of  $L^{p'}(\Omega)$ .

The control is subject to pointwise bounds  $\beta_l < \beta_r$ . The objective is to drive the state as close to  $y_d \in L^2(\Omega)$  as possible. The second part penalizes excessive control costs; the parameter  $\alpha > 0$  is typically small.

We eliminate the state  $y$  via the state equation, i.e.,  $y = y(u) = A^{-1}(r + Bu)$ , and obtain the reduced problem

$$\begin{aligned} \min_{u \in L^2(\Omega)} \quad & \hat{J}(u) \stackrel{\text{def}}{=} J(y(u), u) \stackrel{\text{def}}{=} \frac{1}{2} \|y(u) - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 \\ \text{s.t.} \quad & \beta_l \leq u \leq \beta_r. \end{aligned}$$

This problem is of the form (2.21).

For the gradient we obtain

$$\begin{aligned} (\nabla \hat{J}(u), d)_{L^2(\Omega)} &= (y(u) - y_d, y'(u)d)_{L^2(\Omega)} + \alpha(u, d)_{L^2(\Omega_c)} \\ &= (y'(u)^*(y(u) - y_d) + \alpha u, d)_{L^2(\Omega_c)}. \end{aligned}$$

Therefore,

$$\begin{aligned} \nabla \hat{J}(u) &= y'(u)^*(y(u) - y_d) + \alpha u = B^*(A^{-1})^*(A^{-1}(r + Bu) - y_d) + \alpha u \\ &= \alpha u + B^*(A^{-1})^*(A^{-1}(r + Bu) - y_d) \stackrel{\text{def}}{=} \alpha u + H(u). \end{aligned}$$

Since  $B \in \mathcal{L}(L^{p'}(\Omega_c), H^{-1}(\Omega))$ , we have  $B^* \in \mathcal{L}(H_0^1(\Omega), L^p(\Omega_c))$  with  $p = p'/(p' - 1) > 2$ . Hence, the affine linear operator

$$H(u) = B^*(A^{-1})^*(A^{-1}(r + Bu) - y_d)$$

is a continuous affine linear mapping  $L^2(\Omega_c) \rightarrow L^p(\Omega)$ .

Therefore, we can apply Theorem 2.13 to rewrite the optimality conditions as a semismooth operator equation

$$\Phi(u) \stackrel{\text{def}}{=} u - P_{[\beta_l, \beta_r]}(-(1/\alpha)H(u)) = 0.$$

The Newton system reads

$$\left( I + \frac{1}{\alpha} g^k \cdot H'(u^k) \right) s^k = -\Phi(u^k), \quad (2.26)$$

where  $g \cdot H'(u)$  stands for  $v \mapsto g \cdot (H'(u)v)$  and  $g^k \in L^\infty(\Omega_c)$  is chosen such that

$$g^k(x) \begin{cases} = 0 & -(1/\alpha)H(u^k)(x) \notin [\beta_l, \beta_r], \\ = 1 & -(1/\alpha)H(u^k)(x) \in (\beta_l, \beta_r), \\ \in [0, 1] & -(1/\alpha)H(u^k)(x) \in \{\beta_l, \beta_r\}. \end{cases}$$

The linear operator on the left has the form

$$M_k \stackrel{\text{def}}{=} I + \frac{1}{\alpha} g^k \cdot H'(u^k) = I + \frac{1}{\alpha} g^k \cdot B^*(A^{-1})^* A^{-1} B.$$

For solving (2.26), it can be advantageous to note that  $s^k$  solves (2.26) if and only if  $s^k = d_u^k$  and  $(d_y^k, d_u^k, d_\mu^k)^T$  solves

$$\begin{pmatrix} I & 0 & A^* \\ 0 & I & -\frac{1}{\alpha} g^k \cdot B^* \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} d_y^k \\ d_u^k \\ d_\mu^k \end{pmatrix} = \begin{pmatrix} 0 \\ -\Phi(u^k) \\ 0 \end{pmatrix}. \quad (2.27)$$

As we will see later in Sect. 2.8.2, this system is amenable to multigrid methods.

### 2.5.5 General Optimization Problems with Inequality Constraints in $L^2$

We now consider problems of the form

$$\min_{w \in W} f(w) \quad \text{s.t.} \quad e(w) = 0, \quad c_j(w) \leq 0 \quad \text{a.e. on } \Omega_j, \quad j = 1, \dots, m.$$

Here  $W$  and  $Z$  are Banach spaces,  $f : W \rightarrow \mathbb{R}$ ,  $e : W \rightarrow Z$ , and  $c_j : W \rightarrow L^2(\Omega_j)$  are twice continuously F-differentiable. The sets  $\Omega_j \subset \mathbb{R}^{n_j}$  are assumed to be measurable and bounded.

This, in particular, includes control-constrained optimal control problems with  $L^2$ -control  $u$  and state  $y \in Y$ :

$$\min_{y \in Y, u \in L^2(\Omega)} J(y, u) \quad \text{s.t.} \quad e(y, u) = 0, \quad a_i \leq u_i \leq b_i, \quad i = 1, \dots, l,$$

with  $y \in Y$  denoting the state,  $u \in L^2(\Omega_1) \times \dots \times L^2(\Omega_l)$  denoting the controls, and  $a_i, b_i \in L^\infty(\Omega_i)$ .

In this case, we have

$$\begin{aligned} w &= (y, u), & m &= 2l, & c_{2i-1}(y, u) &= a_i - u_i, \\ c_{2i}(y, u) &= u_i - b_i, & i &= 1, \dots, l. \end{aligned}$$

To simplify the presentation, consider the case  $m = 1$ , i.e.,

$$\min_{w \in W} f(w) \quad \text{s.t.} \quad e(w) = 0, \quad c(w) \leq 0 \quad \text{a.e. on } \Omega. \quad (2.28)$$

The Lagrange function is given by

$$L : W \times L^2(\Omega) \times Z^* \rightarrow \mathbb{R},$$

$$L(w, \lambda, \mu) = f(w) + (\lambda, c(w))_{L^2(\Omega)} + \langle \mu, e(w) \rangle_{Z^*, Z}.$$

Assuming that a CQ holds at the solution  $\bar{w} \in W$ , the KKT conditions hold:

There exist  $\bar{\lambda} \in L^2(\Omega)$  and  $\bar{\mu} \in Z^*$  such that  $(\bar{w}, \bar{\lambda}, \bar{\mu})$  satisfies

$$L_w(\bar{w}, \bar{\lambda}, \bar{\mu}) = 0, \quad (2.29)$$

$$e(\bar{w}) = 0, \quad (2.30)$$

$$c(\bar{w}) \leq 0, \quad \bar{\lambda} \geq 0, \quad (\bar{\lambda}, c(\bar{w}))_{L^2(\Omega)} = 0. \quad (2.31)$$

The last line can equivalently be written as  $\text{VI}(-c(\bar{w}), \mathcal{K})$  with  $\mathcal{K} = \{u \in L^2(\Omega) : u \geq 0\}$  and this VI can again be rewritten using the projection onto  $\mathcal{K}$ :

$$\bar{\lambda} - P_{\mathcal{K}}(\bar{\lambda} + \theta c(\bar{w})) = 0$$

with fixed  $\theta > 0$ . Since  $P_{\mathcal{K}}(u) = P_{[0, \infty)}(u(\cdot))$ , we again have to deal with a superposition operator.

To make the whole KKT system a semismooth equation, we need to get a smoothing operator inside of the projection.

We need additional structure to achieve this. Since it is not very enlightening to define this structure in full generality without giving a motivation, we look at an example first.

## 2.5.6 Application to Elliptic Optimal Control Problems

### 2.5.6.1 Distributed Control

Very similar as in Sect. 2.5.4, we consider the following control-constrained elliptic optimal control problem

$$\min_{y \in H_0^1(\Omega), u \in L^2(\Omega)} J(y, u) \stackrel{\text{def}}{=} \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 \quad (2.32)$$

$$\text{s.t.} \quad Ay = r + Bu, \quad u \leq b.$$

Here  $\Omega \subset \mathbb{R}^n$  is an open bounded domain and  $A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega)$  is a second order linear elliptic operator, e.g.,  $A = -\Delta$ . Furthermore,  $b \in L^\infty(\Omega)$  is an upper bound on the control,  $r \in H^{-1}(\Omega)$  is a source term, and  $B \in \mathcal{L}(L^{p'}(\Omega_c), H^{-1}(\Omega))$ ,  $p' \in [1, 2)$  is the control operator. For a more detailed explanation of the problem setting, see Sect. 2.5.4.

We convert this control problem into the form (2.28) by setting

$$\begin{aligned} w &= (y, u), & W &= Y \times U, & Y &= H_0^1(\Omega), & U &= L^2(\Omega), \\ Z &= H^{-1}(\Omega), & e(y, u) &= Ay - Bu - r, & c(y, u) &= u - b. \end{aligned}$$

Note that  $e$  and  $c$  are continuous affine linear operators. Hence,

$$e_y(y, u) = A, \quad e_u(y, u) = -B, \quad c_y(y, u) = 0, \quad c_u(y, u) = I.$$

The Lagrange function is

$$L(y, u, \lambda, \mu) = J(y, u) + (\lambda, c(y, u)_{L^2(\Omega)}) + \langle \mu, e(y, u) \rangle_{H_0^1(\Omega), H^{-1}(\Omega)}.$$

We write down the optimality conditions:

$$\begin{aligned} L_y(y, u, \lambda, \mu) &= J_y(y, u) + c_y(y, u)^* \lambda + e_y(y, u)^* \mu = y - y_d + A^* \mu = 0, \\ L_u(y, u, \lambda, \mu) &= J_u(y, u) + c_u(y, u)^* \lambda + e_u(y, u)^* \mu = \alpha u + \lambda - B^* \mu = 0, \\ \lambda &\geq 0, \quad c(y, u) = u - b \leq 0, \quad (\lambda, c(y, u))_{L^2(\Omega)} = (\lambda, u - b)_{L^2(\Omega)} = 0, \\ e(y, u) &= Ay - Bu - r = 0. \end{aligned}$$

The second equation yields  $\lambda = B^* \mu - \alpha u$  and inserting this, we arrive at

$$\begin{aligned} A^* \mu &= -(y - y_d), & (\text{adjoint equation}) \\ B^* \mu - \alpha u &\geq 0, \quad u \leq b, \quad (B^* \mu - \alpha u, u - b)_{L^2(\Omega)} = 0, \\ Ay &= r + Bu. & (\text{state equation}) \end{aligned}$$

We can reformulate the complementarity condition by using the projection  $P_{[0, \infty)}$  as follows:

$$b - u - P_{[0, \infty)}(b - u - \theta(B^* \mu - \alpha u)) = 0.$$

If we choose  $\theta = 1/\alpha$ , this simplifies to

$$\Phi(u, \mu) := u - b + P_{[0, \infty)}(b - (1/\alpha)B^* \mu) = 0.$$

Since  $B^* \in \mathcal{L}(H_0^1(\Omega), L^p(\Omega))$  with  $p = p'/(p' - 1) > 2$ , we see that

$$(u, \mu) \in L^2(\Omega) \times H_0^1(\Omega) \mapsto b - (1/\alpha)B^* \mu \in L^p(\Omega)$$

is continuous and affine linear, and thus  $\Phi$  is  $\partial\Phi$ -semismooth w.r.t.

$$\partial\Phi : L^2(\Omega) \times H_0^1(\Omega) \rightrightarrows \mathcal{L}(L^2(\Omega) \times H_0^1(\Omega), L^2(\Omega)),$$

$$\partial\Phi(u, \mu) = \{M; M = (I, -(g/\alpha) \cdot B^*), g \in L^\infty(\Omega),$$

$$g(x) \in \partial^{cl} P_{[0, \infty)}(b(x) - (1/\alpha)(B^* \mu)(x)) \text{ for a.a. } x \in \Omega\}.$$

Here,

$$\partial^{cl} P_{[0,\infty)}(t) = \begin{cases} \{0\} & t < 0, \\ \{1\} & t > 0, \\ [0, 1] & t = 0. \end{cases} \quad (2.33)$$

The semismooth Newton system looks as follows

$$\begin{pmatrix} I & 0 & A^* \\ 0 & I & -(g^k/\alpha) \cdot B^* \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} s_y \\ s_u \\ s_\mu \end{pmatrix} = - \begin{pmatrix} y^k - y_d + A^* \mu^k \\ u^k - b + P_{[0,\infty)}(b - (1/\alpha)B^* \mu^k) \\ Ay^k - Bu^k - r \end{pmatrix}. \quad (2.34)$$

It is important to note that this equation has exactly the same linear operator on the left as the extended system in (2.27). In particular, the regularity condition for the Newton system (2.34) is closely connected to the regularity condition for (2.26).

### 2.5.6.2 Neumann Boundary Control

We now consider a similar problem as before, but with Neumann boundary control:

$$\begin{aligned} \min_{y \in H^1(\Omega), u \in L^2(\partial\Omega)} \quad & J(y, u) \stackrel{\text{def}}{=} \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\partial\Omega)}^2 \\ \text{s.t.} \quad & -\Delta y + cy = r \quad \text{in } \Omega, \\ & \frac{\partial y}{\partial \nu} = u \quad \text{in } \partial\Omega, \\ & u \leq b \quad \text{in } \partial\Omega. \end{aligned} \quad (2.35)$$

Here  $\Omega \subset \mathbb{R}^n$  is an open bounded Lipschitz domain and  $c \in L^\infty(\Omega)$ ,  $c > 0$ . Furthermore,  $b \in L^\infty(\partial\Omega)$  is an upper bound on the control and  $r \in H^1(\Omega)^*$  is a source term.

The weak formulation of the state equation including boundary condition is

$$\int_{\Omega} (\nabla y \cdot \nabla v + cyv) dx = \int_{\Omega} rv dx + \int_{\partial\Omega} uv dS(x) \quad \forall v \in H^1(\Omega),$$

which in operator form can be written as

$$Ay = r + Bu,$$

where

$$B \in \mathcal{L}(L^2(\partial\Omega), H^1(\Omega)^*), \quad \langle Bu, v \rangle_{H^1(\Omega)^*, H^1(\Omega)} = \int_{\partial\Omega} uv dS(x),$$

$$A \in \mathcal{L}(H^1(\Omega), H^1(\Omega)^*),$$

$$\langle Ay, v \rangle_{H^1(\Omega)^*, H^1(\Omega)} = \int_{\Omega} (\nabla y \cdot \nabla v + cyv) dx \quad \forall v \in H^1(\Omega).$$

The adjoint  $B^* \in \mathcal{L}(H^1(\Omega), L^2(\partial\Omega))$  of  $B$  is given by  $B^*v = v|_{\partial\Omega}$ . In fact,

$$(B^*v, w)_{L^2(\partial\Omega)} = \langle Bw, v \rangle_{H^1(\Omega)^*, H^1(\Omega)} = \int_{\partial\Omega} wv dS(x) = (v, w)_{L^2(\partial\Omega)}.$$

This control problem assumes the form (2.28) by setting

$$\begin{aligned} w &= (y, u), & W &= Y \times U, & Y &= H^1(\Omega), & U &= L^2(\partial\Omega), \\ Z &= H^1(\Omega)^*, & e(y, u) &= Ay - Bu - r, & c(y, u) &= u - b. \end{aligned}$$

The operators  $e$  and  $c$  are continuous and affine linear with derivatives

$$e_y(y, u) = A, \quad e_u(y, u) = -B, \quad c_y(y, u) = 0, \quad c_u(y, u) = I.$$

The Lagrange function reads

$$L(y, u, \lambda, \mu) = J(y, u) + (\lambda, c(y, u))_{L^2(\partial\Omega)} + \langle \mu, e(y, u) \rangle_{H^1(\Omega), H^1(\Omega)^*}.$$

We write down the optimality conditions:

$$L_y(y, u, \lambda, \mu) = J_y(y, u) + c_y(y, u)^* \lambda + e_y(y, u)^* \mu = y - y_d + A^* \mu = 0,$$

$$L_u(y, u, \lambda, \mu) = J_u(y, u) + c_u(y, u)^* \lambda + e_u(y, u)^* \mu = \alpha u + \lambda - B^* \mu = 0,$$

$$\lambda \geq 0, \quad c(y, u) = u - b \leq 0, \quad (\lambda, c(y, u))_{L^2(\partial\Omega)} = (\lambda, u - b)_{L^2(\partial\Omega)} = 0,$$

$$e(y, u) = Ay - Bu - r = 0.$$

The second equation yields  $\lambda = B^* \mu - \alpha u$  and using this to eliminate  $\lambda$ , we arrive at

$$A^* \mu = -(y - y_d), \quad (\text{adjoint equation})$$

$$B^* \mu - \alpha u \geq 0, \quad u \leq b, \quad (B^* \mu - \alpha u, u - b)_{L^2(\partial\Omega)} = 0, \quad (2.36)$$

$$Ay = r + Bu. \quad (\text{state equation})$$

Inserting  $Av = A^*v = -\Delta v + cv$ ,  $B^*v = v|_{\partial\Omega}$ , and the definition of  $B$ , we can express this system as a coupled system of elliptic partial differential equations:

$$-\Delta \mu + c\mu = -(y - y_d) \quad \text{in } \Omega,$$

$$\frac{\partial \mu}{\partial \nu} = 0 \quad \text{in } \partial\Omega,$$

$$\mu|_{\partial\Omega} - \alpha u \geq 0, \quad u \leq b, \quad (\mu|_{\partial\Omega} - \alpha u)(u - b) = 0 \quad \text{in } \partial\Omega,$$

$$\begin{aligned} -\Delta y + cy &= r \quad \text{in } \Omega, \\ \frac{\partial y}{\partial \nu} &= u \quad \text{in } \partial\Omega. \end{aligned}$$

Here, we have written the complementarity condition pointwise. Note that in the adjoint equation we have homogeneous Neumann boundary conditions since a Neumann boundary condition  $\frac{\partial y}{\partial \nu} = h$  would result in the term  $Bh$  on the right hand side of the differential equation. Since no such term is present in the adjoint equation, we must have  $h = 0$ .

We return to the more compact notation of (2.36) and reformulate the complementarity condition by using the projection  $P_{[0,\infty)}$  as follows:

$$b - u - P_{[0,\infty)}(b - u - \theta(B^*\mu - \alpha u)) = 0 \quad \text{in } L(\partial\Omega).$$

If we choose  $\theta = 1/\alpha$ , this simplifies to

$$\Phi(u, \mu) := u - b + P_{[0,\infty)}(b - (1/\alpha)B^*\mu) = 0 \quad \text{in } L(\partial\Omega).$$

From  $B^*v = v|_{\partial\Omega}$  we see that  $B^*$  is a bounded linear operator from  $H^1(\Omega)$  not only to  $L^2(\partial\Omega)$ , but even to  $H^{1/2}(\partial\Omega)$ . By the Sobolev embedding theorem, we can find  $p > 2$  with  $H^{1/2}(\partial\Omega) \hookrightarrow L^p(\partial\Omega)$ . We then have  $B^* \in \mathcal{L}(H^1(\Omega), L^p(\partial\Omega))$  with  $p > 2$ . Hence,

$$(u, \mu) \in L^2(\partial\Omega) \times H^1(\Omega) \mapsto b - (1/\alpha)B^*\mu \in L^p(\partial\Omega)$$

is continuous and affine linear, and thus  $\Phi$  is  $\partial\Phi$ -semismooth w.r.t.

$$\partial\Phi : L^2(\partial\Omega) \times H^1(\Omega) \rightrightarrows \mathcal{L}(L^2(\partial\Omega) \times H^1(\Omega), L^2(\partial\Omega)),$$

$$\partial\Phi(u, \mu) = \{M; M = (I, -(g/\alpha) \cdot B^*), g \in L^\infty(\partial\Omega),$$

$$g(x) \in \partial^{cl} P_{[0,\infty)}(b(x) - (1/\alpha)(B^*\mu)(x)) \text{ for a.a. } x \in \partial\Omega\}.$$

Here,  $\partial^{cl} P_{[0,\infty)}(t)$  is as in (2.33). The semismooth Newton system then is

$$\begin{pmatrix} I & 0 & A^* \\ 0 & I & -(g^k/\alpha) \cdot B^* \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} s_y \\ s_u \\ s_\mu \end{pmatrix} = - \begin{pmatrix} y^k - y_d + A^*\mu^k \\ u^k - b + P_{[0,\infty)}(b - (1/\alpha)B^*\mu^k) \\ Ay^k - Bu^k - r \end{pmatrix}. \quad (2.37)$$

### 2.5.7 Optimal Control of the Incompressible Navier-Stokes Equations

We now discuss how an optimal control problem governed by the 2d incompressible instationary Navier-Stokes equations can be solved by a semismooth Newton method. We use exactly the notation of Sect. 1.8. In particular,  $\Omega \subset \mathbb{R}^2$  is the open

bounded flow domain and  $I = [0, T]$  is the time horizon. By  $V$  we denote the closure of  $\{y \in C_0^\infty(\Omega)^2 : \nabla \cdot y = 0\}$  in  $H_0^1(\Omega)^2$  and by  $H$  its closure in  $L^2(\Omega)^2$ . Given the resulting Gelfand triple  $V \hookrightarrow H \hookrightarrow V^*$  we can write the state equation of the flow control problem as follows: The velocity field  $y \in W(I)$  satisfies

$$\begin{aligned} y_t - \nu \Delta y + (y \cdot \nabla)y &= Bu \quad \text{in } L^2(I; V^*), \\ y|_{t=0} &= y_0 \quad \text{in } H. \end{aligned} \quad (2.38)$$

Here,  $B \in \mathcal{L}(U, L^2(I; V^*))$  is the control operator and  $U$  is a Hilbert space of controls. To be more concrete, we will consider time-dependent control on the right hand side on a subdomain  $\Omega_c$  of the flow domain  $\Omega$ . We achieve this by choosing  $B \in \mathcal{L}(L^2(I \times \Omega_c)^2, L^2(I; V^*))$ ,

$$\langle Bu, w \rangle_{L^2(I; V^*), L^2(I; V)} = (u, w)_{L^2(I \times \Omega_c)^2}.$$

This is well defined, since  $L^2(I; L^2(\Omega)) = L^2(I \times \Omega)$ .

We consider an objective function of the form

$$J(y, u) = \frac{1}{2} \int_0^T \|Ny - q_d\|_{L^2(\Omega_d)^2}^2 dt + \frac{\alpha}{2} \|u\|_{L^2(I \times \Omega_c)^2}^2.$$

Here,  $N \in \mathcal{L}(V, L^2(\Omega_d)^2)$  is an operator that maps the velocity field to the corresponding observation on the set  $\Omega_d \subset \Omega$ . For instance,  $N = I$  or  $N = \text{curl}$  are possible choices. On the control we will pose a pointwise constraint

$$u \in C \quad \text{on } I \times \Omega_c,$$

where  $C \subset \mathbb{R}^2$  is a closed convex set such that the projection  $P_C$  onto  $C$  is semi-smooth.

We thus consider the problem

$$\min_{y, u} J(y, u) \quad \text{s.t.} \quad (y, u) \text{ satisfy (2.38)} \quad \text{and} \quad u \in C \quad \text{on } I \times \Omega_c.$$

The analysis of this problem was discussed in Sect. 1.8. In particular, for any  $u \in U$  the state equation possesses a unique solution  $y(u) \in W(I)$  and the operator  $u \mapsto y(u)$  is infinitely F-differentiable. Since the objective function  $J(y, u)$  is continuous and quadratic, it is infinitely F-differentiable. Therefore, the reduced objective function  $\hat{J}(u) = J(y(u), u)$  is infinitely F-differentiable. The gradient of  $\hat{J}(u)$  can be represented using the adjoint state in the form

$$\nabla \hat{J}(u) = \alpha u - B^* p_1,$$

where  $p_1 = p_1(u) \in L^2(I; V)$  is the adjoint state corresponding to  $(y, u) = (y(u), u)$  given by the weak solution of the adjoint equation

$$\begin{aligned} -(p_1)_t - (y \cdot \nabla)p_1 + (\nabla y)^T p_1 - \nu \Delta p_1 &= -N^*(Ny - q_d) \quad \text{in } I \times \Omega, \\ p_1|_{t=T} &= 0 \quad \text{in } \Omega. \end{aligned}$$

Due to the structure of  $B$  we see that

$$\langle Bu, w \rangle_{L^2(I; V^*), L^2(I; V)} = (u, w)_{L^2(I \times \Omega_c)^2} = (u, B^*w)_{L^2(I \times \Omega_c)^2}.$$

Therefore,  $B^*w = w|_{I \times \Omega_c}$ .

Since  $N \in \mathcal{L}(V, L^2(\Omega_d)^2)$ , we have  $N^* \in \mathcal{L}(L^2(\Omega_d)^2, V^*)$  and thus the right hand side  $-N^*(Ny(u) - q_d)$  maps  $u \in U = L^2(I \times \Omega_c)^2 = L^2(I, L^2(\Omega_c)^2)$  infinitely F-differentiable to  $L^2(I; V^*)$ . From the imbedding  $L^2(I; V^*) \hookrightarrow W(I)^* \cap L^{4/3}(I; V^*)$  and Theorem 1.58 we conclude that the operator

$$u \in U \mapsto p_1(u) \in W^{4/3}(I)$$

is well-defined and Lipschitz continuous on bounded sets.

Furthermore, it can be shown, see [134, 137], that

$$W^{4/3}(I) \hookrightarrow L^q(I \times \Omega)^2, \quad \forall 1 \leq q < \frac{7}{2}.$$

Thus fixing  $q \in (2, 7/2)$  we obtain that

$$u \in U \mapsto p_1(u) \in L^q(I \times \Omega)$$

is well-defined and Lipschitz continuous on bounded sets.

We collect what we have found so far

- $\hat{J} : U \rightarrow \mathbb{R}$  is infinitely F-differentiable.
- The reduced gradient has the following structure:

$$\nabla \hat{J}(u) = \alpha u + H(u)$$

with

$$H(u) = -B^* p_1(u) = -p_1(u)|_{I \times \Omega_c},$$

where  $p_1(u) \in L^2(I; V)$  is the adjoint state.

- The operator  $u \in U \mapsto p_1(u) \in L^2(I; V)$  is infinitely F-differentiable. Furthermore, the operator

$$u \in U \mapsto p_1(u) \in W^{4/3}(I) \hookrightarrow L^q(I \times \Omega)$$

is Lipschitz continuous on bounded sets for  $q \in (2, 7/2)$ . From this, it follows that  $H : U \rightarrow U$  is infinitely F-differentiable and that the operator

$$u \in U \mapsto H(u) \in L^q(I \times \Omega_c)$$

is Lipschitz continuous on bounded sets.

We can write the first order optimality conditions in the form

$$u - P_C(u - \theta \nabla \hat{J}(u)) = 0$$

with  $\theta > 0$  fixed. Choosing  $\theta = 1/\alpha$  and inserting the adjoint representation of  $\nabla \hat{J}(u)$ , we obtain

$$u - P_C(-(1/\alpha)H(u)) = 0. \quad (2.39)$$

We made the assumption that  $P_C$  is semismooth. Due to the properties of the operator  $H$  it now follows from Theorem 2.13 that the operator in equation (2.39) is semismooth from  $U$  to  $U$ . Hence, a semismooth Newton's method can be applied to this optimal control problem. For further details, we refer to [134, 137].

## 2.6 Sequential Quadratic Programming

### 2.6.1 Lagrange-Newton Methods for Equality Constrained Problems

We consider

$$\min_{w \in W} f(w) \quad \text{s.t.} \quad e(w) = 0 \quad (2.40)$$

with  $f : W \rightarrow \mathbb{R}$  and  $e : W \rightarrow Z$  twice continuously F-differentiable.

If  $\bar{w}$  is a local solution and a CQ holds (e.g.,  $e'(\bar{w})$  is surjective), then the KKT conditions hold:

There exists a Lagrange multiplier  $\bar{\mu} \in Z^*$  such that  $(\bar{w}, \bar{\mu})$  satisfies

$$\begin{aligned} L_w(\bar{w}, \bar{\mu}) &= f'(\bar{w}) + e'(\bar{w})^* \bar{\mu} = 0, \\ L_\mu(\bar{w}, \bar{\mu}) &= e(\bar{w}) = 0. \end{aligned}$$

Setting

$$x = (w, \mu), \quad G(w, \mu) = \begin{pmatrix} L_w(w, \mu) \\ e(w) \end{pmatrix},$$

the KKT conditions form a nonlinear equation

$$G(x) = 0.$$

To this equation we can apply Newton's method:

$$G'(x^k)s^k = -G(x^k).$$

Written in detail,

$$\begin{pmatrix} L_{ww}(w^k, \mu^k) & e'(w^k)^* \\ e'(w^k) & 0 \end{pmatrix} \begin{pmatrix} s_w^k \\ s_\mu^k \end{pmatrix} = - \begin{pmatrix} L_w(w^k, \mu^k) \\ e(w^k) \end{pmatrix}. \quad (2.41)$$

The resulting method is called *Lagrange-Newton method*. We need a regularity condition:

$$\begin{pmatrix} L_{ww}(\bar{w}, \bar{\mu}) & e'(\bar{w})^* \\ e'(\bar{w}) & 0 \end{pmatrix} \quad \text{is boundedly invertible.} \quad (2.42)$$

**Theorem 2.15** *Let  $f$  and  $e$  be twice continuously  $F$ -differentiable. Let  $(\bar{w}, \bar{\mu})$  be a KKT pair of (2.40) at which the regularity condition (2.42) holds. Then there exists  $\delta > 0$  such that, for all  $(w^0, \mu^0) \in W \times Z^*$  with  $\|(w^0, \mu^0) - (\bar{w}, \bar{\mu})\|_{W \times Z^*} < \delta$ , the Lagrange-Newton iteration converges  $q$ -superlinearly to  $(\bar{w}, \bar{\mu})$ .*

*If the second derivatives of  $f$  and  $e$  are locally Lipschitz continuous, then the rate of convergence is  $q$ -quadratic.*

*Proof* We just have to apply the convergence theory of Newton's method.

If the second derivatives of  $f$  and  $e$  are locally Lipschitz continuous, then  $G'$  is locally Lipschitz continuous, and thus we have  $q$ -quadratic convergence.

So far, it is not clear what the connection is between the Lagrange-Newton method and sequential quadratic programming.

However, the connection is very close. Consider the following quadratic program:

SQP subproblem:

$$\begin{aligned} \min_{d \in W} \quad & \langle f'(w^k), d \rangle_{W^*, W} + \frac{1}{2} \langle L_{ww}(w^k, \mu^k) d, d \rangle_{W^*, W} \\ \text{s.t.} \quad & e(w^k) + e'(w^k) d = 0. \end{aligned} \quad (2.43)$$

The constraint is linear with derivative  $e'(w^k)$ . As we will show below,  $e'(w^k)$  is surjective for  $w^k$  close to  $\bar{w}$  if  $e'(\bar{w})$  is surjective.

Therefore, at a solution  $d^k$  of (2.43), the KKT conditions hold:

There exists  $\mu_{qp}^k \in Z^*$  such that  $(d^k, \mu_{qp}^k)$  solves

$$\begin{aligned} f'(w^k) + L_{ww}(w^k, \mu^k) d^k + e'(w^k)^* \mu_{qp}^k &= 0 \\ e(w^k) + e'(w^k) d^k &= 0. \end{aligned} \quad (2.44)$$

It is now easily seen that  $(d^k, \mu_{qp}^k)$  solves (2.44) if and only if  $(s_w^k, s_\mu^k) = (d^k, \mu_{qp}^k - \mu^k)$  solves (2.41).

Hence, locally, the Lagrange-Newton method is equivalent to the following method:

**Algorithm 2.16** (SQP method for equality constrained problems)

0. Choose  $(w^0, \mu^0)$  (sufficiently close to  $(\bar{w}, \bar{\mu})$ ).

For  $k = 0, 1, 2, \dots$ :

1. If  $(w^k, \mu^k)$  is a KKT pair of (2.40), STOP.
2. Compute the KKT pair  $(d^k, \mu^{k+1})$  of

$$\begin{aligned} \min_{d \in W} \quad & \langle f'(w^k), d \rangle_{W^*, W} + \frac{1}{2} \langle L_{ww}(w^k, \mu^k) d, d \rangle_{W^*, W} \\ \text{s.t.} \quad & e(w^k) + e'(w^k) d = 0, \end{aligned}$$

that is closest to  $(0, \mu^k)$ .

3. Set  $w^{k+1} = w^k + d^k$ .

For solving the SQP subproblems in step 2, it is important to know if for  $w^k$  close to  $\bar{w}$ , the operator  $e'(w^k)$  is indeed surjective and if there exists a unique solution to the QP.

**Lemma 2.8** *Let  $W$  be a Hilbert space and  $Z$  be a Banach space. Furthermore, let  $e : W \rightarrow Z$  be continuously  $F$ -differentiable and let  $e'(\bar{w})$  be surjective. Then  $e'(w)$  is surjective for all  $w$  close to  $\bar{w}$ .*

*Proof* We set  $B = e'(\bar{w})$ , and  $B(w) = e'(w)$ , and do the splitting  $W = W_0 \perp W_1$  with  $W_0 = \text{Kern}(B)$ . We then see that  $B|_{W_1} \in \mathcal{L}(W_1, Z)$  is bijective and thus continuously invertible (open mapping theorem). Now, by continuity, for  $w \rightarrow \bar{w}$  we have  $B(w) \rightarrow B$  in  $\mathcal{L}(W, Z)$  and thus also  $B(w)|_{W_1} \rightarrow B|_{W_1}$  in  $\mathcal{L}(W_1, Z)$ . Therefore, by the Lemma of Banach,  $B(w)|_{W_1}$  is continuously invertible for  $w$  close to  $\bar{w}$  and thus  $B(w)$  is onto.

Next, we show a second-order sufficient condition for the QP.

**Lemma 2.9** *Let  $W$  be a Hilbert space and  $Z$  be a Banach space. Furthermore, let  $f : W \rightarrow \mathbb{R}$  and  $e : W \rightarrow Z$  be twice continuously  $F$ -differentiable. Let  $e(\bar{w}) = 0$  and assume that  $e'(\bar{w})$  is surjective. In addition, let the following second-order sufficient condition hold at  $(\bar{w}, \bar{\mu})$ :*

$$\langle d, L_{ww}(\bar{w}, \bar{\mu})d \rangle_{W, W^*} \geq \alpha \|d\|_W^2 \quad \forall d \in W \text{ with } e'(\bar{w})d = 0,$$

where  $\alpha > 0$  is a constant. Then, there exists  $\delta > 0$  such that for all  $(w, \mu) \in W \times Z^*$  with  $\|(w, \mu) - (\bar{w}, \bar{\mu})\|_{W \times Z^*} < \delta$  the following holds:

$$\langle d, L_{ww}(w, \mu)d \rangle_{W, W^*} \geq \frac{\alpha}{2} \|d\|_W^2 \quad \forall d \in W \text{ with } e'(w)d = 0.$$

*Proof* Set  $B = e'(\bar{w})$ ,  $B(w) = e'(w)$ ,  $W_0 = \text{Kern}(B)$  and split  $W = W_0 \perp W_1$ . Remember that  $B|_{W_1} \in \mathcal{L}(W_1, Z)$  is continuously invertible.

For any  $d \in \text{Kern}(B(w))$  there exist unique  $d_0 \in W_0$  and  $d_1 \in W_1$  with  $d = d_0 + d_1$ . Our first aim is to show that  $d_1$  is small. In fact,

$$\|Bd_1\|_Z = \|Bd\|_Z = \|(B - B(w))d\|_Z \leq \|B - B(w)\|_{W \rightarrow Z} \|d\|_W.$$

Hence,

$$\begin{aligned} \|d_1\|_W &= \|(B|_{W_1})^{-1} Bd_1\|_W \leq \|(B|_{W_1})^{-1}\|_{Z \rightarrow W_1} \|B - B(w)\|_{W \rightarrow Z} \|d\|_W \\ &\stackrel{\text{def}}{=} \xi(w) \|d\|_W. \end{aligned}$$

Therefore, setting  $x = (w, \mu)$ ,

$$\begin{aligned}
& \langle L_{ww}(x)d, d \rangle_{W^*, W} \\
&= \langle L_{ww}(\bar{x})d, d \rangle_{W^*, W} + \langle (L_{ww}(x) - L_{ww}(\bar{x}))d, d \rangle_{W^*, W} \\
&= \langle L_{ww}(\bar{x})d_0, d_0 \rangle_{W^*, W} + \langle L_{ww}(\bar{x})(d + d_0), d_1 \rangle_{W^*, W} \\
&\quad + \langle (L_{ww}(x) - L_{ww}(\bar{x}))d, d \rangle_{W^*, W} \\
&\geq \alpha \|d_0\|_W^2 - \|L_{ww}(\bar{x})\|_{W \rightarrow W^*} (\|d\|_W + \|d_0\|_W) \|d_1\|_W \\
&\quad - \|L_{ww}(x) - L_{ww}(\bar{x})\|_{W \rightarrow W^*} \|d\|_W^2 \\
&\geq (\alpha(1 - \xi^2(w)) - 2\|L_{ww}(\bar{x})\|_{W \rightarrow W^*} \xi(w) \\
&\quad - \|L_{ww}(x) - L_{ww}(\bar{x})\|_{W \rightarrow W^*}) \|d\|_W^2 \\
&=: \alpha(x) \|d\|_W^2.
\end{aligned}$$

By continuity,  $\alpha(x) \rightarrow \alpha$  for  $x \rightarrow \bar{x}$ .

A sufficient condition for the regularity condition (2.42) is the following:

**Lemma 2.10** *Let  $W$  be a Hilbert space, let  $e'(\bar{w})$  be surjective (this is a CQ), and assume that the following second order sufficient condition holds:*

$$\langle d, L_{ww}(\bar{w}, \bar{\mu})d \rangle_{W, W^*} \geq \alpha \|d\|_W^2 \quad \forall d \in W \text{ with } e'(\bar{w})d = 0,$$

where  $\alpha > 0$  is a constant. Then the regularity condition (2.42) holds.

*Proof* For brevity, set  $A = L_{ww}(\bar{w}, \bar{\mu})$  and  $B = e'(\bar{w})$ . We consider the unique solvability of

$$\begin{pmatrix} A & B^* \\ B & 0 \end{pmatrix} \begin{pmatrix} w \\ \mu \end{pmatrix} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}.$$

Denote by  $W_0$  the null space of  $B$  and by  $W_1$  its orthogonal complement. Then  $W = W_0 \perp W_1$  and  $W_0, W_1$  are Hilbert spaces.

Since  $B$  is surjective, the equation  $Bw = r_2$  is solvable and the set of all solutions is  $w_1(r_2) + W_0$ , where  $w_1(r_2) \in W_1$  is uniquely determined.

We have

$$\langle d, Ad \rangle_{W, W^*} \geq \alpha \|d\|_W^2 \quad \forall d \in W_0.$$

Hence, by the Lax-Milgram Lemma 1.8, there exists a unique solution  $w_0(r_1, r_2) \in W_0$  to the problem

$$w_0 \in W_0, \quad \langle Aw_0, d \rangle_{W^*, W} = \langle r_1 - Aw_1(r_2), d \rangle_{W^*, W} \quad \forall d \in W_0.$$

Since  $B$  is surjective, we have for all  $z^* \in \text{Kern}(B^*)$ :

$$\langle z^*, Z \rangle_{Z^*, Z} = \langle z^*, BW \rangle_{Z^*, Z} = \langle B^* z^*, W \rangle_{W^*, W} = \langle \{0\}, W \rangle_{W^*, W} = \{0\}.$$

Hence,  $\text{Kern}(B^*) = \{0\}$  and thus  $B^*$  is injective. Also, since  $BW = Z$  is closed, the closed range theorem yields

$$B^*Z^* = \text{Kern}(B)^\perp = W_0^\perp.$$

Here, for  $S \subset X$

$$S^\perp = \{x' \in X^* : \langle x', s \rangle_{X^*, X} = 0 \ \forall s \in S\}.$$

By construction,  $r_1 - Aw_0(r_1, r_2) - Aw_1(r_2) \in W_0^\perp$ . Hence, there exists a unique  $\mu(r_1, r_2) \in Z^*$  such that

$$B^*\mu(r_1, r_2) = r_1 - Aw_0(r_1, r_2) - Aw_1(r_2).$$

Therefore, we have found the unique solution

$$\begin{pmatrix} w \\ \mu \end{pmatrix} = \begin{pmatrix} w_0(r_1, r_2) + w_1(r_2) \\ \mu(r_1, r_2) \end{pmatrix}.$$

## 2.6.2 The Josephy-Newton Method

In the previous section, we were able to derive the SQP method for equality-constrained problems by applying Newton's method to the KKT system.

For inequality constrained problems this is not directly possible since the KKT system consists of operator equations and a variational inequality. As we will see, such a combination can be most elegantly written as a

### 2.6.2.1 Generalized Equation

$$\text{GE}(G, N): \quad 0 \in G(x) + N(x).$$

Here,  $G : X \rightarrow Y$  is assumed to be continuously F-differentiable and  $N : X \rightrightarrows Y$  is a set-valued mapping with closed graph.

For instance, the variational inequality  $\text{VI}(F, S)$ , with  $F : W \rightarrow W^*$  and  $S \subset W$  closed and convex, can be written as

$$0 \in F(w) + N_S(w),$$

where  $N_S$  is the normal cone mapping of  $S$ :

**Definition 2.4** Let  $S \subset W$  be a nonempty closed convex subset of the Banach space  $W$ . The *normal cone*  $N_S(w)$  of  $S$  at  $w \in W$  is defined by

$$N_S(w) = \begin{cases} \{y \in W^* : \langle y, z - w \rangle_{W^*, W} \leq 0 \ \forall z \in S\}, & w \in S, \\ \emptyset, & w \notin S. \end{cases}$$

This defines a set-valued mapping  $N_S : W \rightrightarrows W^*$ .

The Josephy-Newton method for generalized equations looks as follows:

**Algorithm 2.17** (Josephy-Newton method for  $\text{GE}(G, N)$ )

0. Choose  $x^0 \in X$  (sufficiently close to the solution  $\bar{x}$  of  $\text{GE}(G, N)$ ).

For  $k = 0, 1, 2, \dots$ :

1. STOP if  $x^k$  solves  $\text{GE}(G, N)$  (holds if  $x^k = x^{k-1}$ ).
2. Compute the solution  $x^{k+1}$  of

$$\begin{aligned} &\text{GE}(G(x^k) + G'(x^k)(\cdot - x^k), N) : \\ &0 \in G(x^k) + G'(x^k)(x - x^k) + N(x) \end{aligned}$$

that is closest to  $x^k$ .

In the classical Newton's method, which corresponds to  $N(x) = \{0\}$  for all  $x$ , an essential ingredient is the regularity condition that  $G'(\bar{x})$  is continuously invertible.

This means that the linearized equation

$$p = G(\bar{x}) + G'(\bar{x})(x - \bar{x})$$

possesses the unique solution  $x(p) = \bar{x} + G'(\bar{x})^{-1}p$ , which of course depends linearly and thus Lipschitz continuously on  $p \in Y$ .

The appropriate generalization of this regularity condition is the following:

**Definition 2.5** (Strong regularity) The generalized equation  $\text{GE}(G, N)$  is called *strongly regular* at a solution  $\bar{x}$  if there exist  $\delta > 0$ ,  $\varepsilon > 0$  and  $L > 0$  such that, for all  $p \in Y$ ,  $\|p\|_Y < \delta$ , there exists a unique  $x = x(p) \in X$  with  $\|x(p) - \bar{x}\|_X < \varepsilon$  such that

$$p \in G(\bar{x}) + G'(\bar{x})(x - \bar{x}) + N(x)$$

and  $x(p)$  is Lipschitz continuous:

$$\|x(p_1) - x(p_2)\|_X \leq L\|p_1 - p_2\|_Y \quad \forall p_1, p_2 \in Y, \quad \|p_i\|_X < \delta, \quad i = 1, 2.$$

It is a milestone result of Robinson [117] that then the following holds:

**Theorem 2.18** Let  $X$ ,  $Y$ , and  $Z$  be Banach spaces. Furthermore, let  $\bar{z} \in Z$  be fixed and assume that  $\bar{x}$  is a solution of

$$\text{GE}(G(\bar{z}, \cdot), N): \quad 0 \in G(\bar{z}, x) + N(x)$$

at which the GE is strongly regular with Lipschitz modulus  $L$ . Assume that  $G$  is  $F$ -differentiable with respect to  $x$  near  $(\bar{z}, \bar{x})$  and that  $G$  and  $G_x$  are continuous at  $(\bar{z}, \bar{x})$ .

Then, for every  $\varepsilon > 0$ , there exist neighborhoods  $Z_\varepsilon(\bar{z})$  of  $\bar{z}$ ,  $X_\varepsilon(\bar{x})$  of  $\bar{x}$ , and a mapping  $x : Z_\varepsilon(\bar{z}) \rightarrow X_\varepsilon(\bar{x})$  such that, for all  $z \in Z_\varepsilon(\bar{z})$ ,  $x(z)$  is the (locally) unique solution of the generalized equation

$$0 \in G(z, x) + N(x), \quad x \in X_\varepsilon(\bar{x}).$$

In addition,

$$\|x(z_1) - x(z_2)\|_X \leq (L + \varepsilon) \|G(z_1, x(z_2)) - G(z_2, x(z_2))\|_Y \quad \forall z_1, z_2 \in Z_\varepsilon(\bar{z}).$$

From this, it is not difficult to derive fast local convergence of the Josephy-Newton method:

**Theorem 2.19** *Let  $X, Y$  be Banach spaces,  $G : X \rightarrow Y$  continuously  $F$ -differentiable, and let  $N : X \rightrightarrows Y$  be set-valued with closed graph. If  $\bar{x}$  is a strongly regular solution of  $\text{GE}(G, N)$ , then the Josephy-Newton method (Algorithm 2.17) is locally  $q$ -superlinearly convergent in a neighborhood of  $\bar{x}$ . If, in addition,  $G'$  is  $\alpha$ -Hölder continuous near  $\bar{x}$ , then the order of convergence is  $1 + \alpha$ .*

*Proof* For compact notation, we set  $B_\delta(x) = \{y \in X : \|y - x\|_X < \delta\}$ .

Let  $L$  be the Lipschitz modulus of strong regularity. We set  $Z = X$ ,  $\bar{z} = \bar{x}$  and consider

$$\bar{G}(z, x) \stackrel{\text{def}}{=} G(z) + G'(z)(x - z).$$

Since  $\bar{G}(\bar{z}, \cdot)$  is affine linear, we have

$$\bar{G}(\bar{z}, \bar{x}) + \bar{G}_x(\bar{z}, \bar{x})(x - \bar{x}) = \bar{G}(\bar{z}, x) = G(\bar{z}) + G'(\bar{z})(x - \bar{z}) = G(\bar{x}) + G'(\bar{x})(x - \bar{x}).$$

Therefore,  $\text{GE}(\bar{G}(\bar{z}, \cdot), N)$  is strongly regular at  $\bar{x}$  with Lipschitz constant  $L$ . Theorem 2.18 is applicable and thus, for  $\varepsilon > 0$ , there exist neighborhoods  $Z_\varepsilon(\bar{z})$  of  $\bar{z} = \bar{x}$  and  $X_\varepsilon(\bar{x})$  of  $\bar{x}$  such that, for all  $z \in Z_\varepsilon(\bar{x})$ ,

$$0 \in \bar{G}(z, x) + N(x) = G(z) + G'(z)(x - z) + N(x), \quad x \in X_\varepsilon(\bar{x})$$

has a unique solution  $x(z)$  that satisfies

$$\forall z_1, z_2 \in Z_\varepsilon(\bar{z}) = Z_\varepsilon(\bar{x}) :$$

$$\begin{aligned} \|x(z_1) - x(z_2)\|_X &\leq (L + \varepsilon) \|\bar{G}(z_1, x(z_2)) - \bar{G}(z_2, x(z_2))\|_Y \\ &= (L + \varepsilon) \|G(z_1) - G(z_2) + G'(z_1)(x(z_2) - z_1) - G'(z_2)(x(z_2) - z_2)\|_Y. \end{aligned}$$

If we choose  $z_1 = z \in Z_\varepsilon(\bar{x})$  and  $z_2 = \bar{x}$ , we obtain  $x(z_2) = \bar{x}$  and thus for all  $z \in Z_\varepsilon(\bar{x})$ :

$$\begin{aligned} \|x(z) - \bar{x}\|_X &\leq (L + \varepsilon) \|G(z) - G(\bar{x}) + G'(z)(\bar{x} - z) - G'(\bar{x})(\bar{x} - \bar{x})\|_Y \\ &= (L + \varepsilon) \|G(z) - G(\bar{x}) - G'(z)(z - \bar{x})\|_Y \end{aligned}$$

$$\begin{aligned}
&\leq (L + \varepsilon) \|G(z) - G(\bar{x}) - G'(\bar{x})(z - \bar{x})\|_Y \\
&\quad + (L + \varepsilon) \|(G'(\bar{x}) - G'(z))(z - \bar{x})\|_Y \\
&\leq (L + \varepsilon) \|G(z) - G(\bar{x}) - G'(\bar{x})(z - \bar{x})\|_Y \\
&\quad + (L + \varepsilon) \|G'(\bar{x}) - G'(z)\|_{X \rightarrow Y} \|z - \bar{x}\|_X \\
&= o(\|z - \bar{x}\|_X) \quad (z \rightarrow \bar{x}).
\end{aligned} \tag{2.45}$$

In the last estimate, we have used the F-differentiability of  $G$  and the continuity of  $G'$ .

Now choose  $\delta > 0$  such that  $B_\delta(\bar{x}) \subset X_\varepsilon(\bar{x})$  and  $B_{5\delta/2}(\bar{x}) \subset Z_\varepsilon(\bar{x})$ . By possibly reducing  $\delta$ , we achieve, using (2.45),

$$\|x(z) - \bar{x}\|_X \leq \frac{1}{2} \|z - \bar{x}\|_X \quad \forall z \in B_\delta(\bar{x}).$$

In particular, this implies

$$x(z) \in B_{\delta/2}(\bar{x}) \subset B_\delta(\bar{x}) \quad \forall z \in B_\delta(\bar{x}).$$

Now observe that, for  $x^k \in B_\delta(\bar{x})$ , the unique solution of  $\text{GE}(G(x^k) + G'(x^k)(\cdot - x^k), N)$  in  $X_\varepsilon(\bar{x})$  is given by  $x(x^k) \in B_{\delta/2}(\bar{x})$ .

From

$$\|x(x^k) - x^k\| \leq \|x(x^k) - \bar{x}\|_X + \|\bar{x} - x^k\|_X < \frac{\delta}{2} + \delta = \frac{3}{2}\delta$$

and  $B_{5\delta/2}(\bar{x}) \subset X_\varepsilon(\bar{x})$  we conclude that  $x(x^k)$  is the solution of  $\text{GE}(G(x^k) + G'(x^k)(\cdot - x^k), N)$  that is closest to  $x^k$ . Hence, for  $x^k \in B_\delta(\bar{x})$ , we have

$$x^{k+1} = x(x^k) \in B_{\delta/2}(\bar{x}) \subset B_\delta(\bar{x}), \quad \|x^{k+1} - \bar{x}\|_X \leq \frac{1}{2} \|x^k - \bar{x}\|_X.$$

Thus, if we choose  $x^0 \in B_\delta(\bar{x})$ , we obtain by induction  $x^k \rightarrow \bar{x}$ .

Furthermore, from (2.45) it follows that

$$\|x^{k+1} - \bar{x}\|_X = \|x(x^k) - \bar{x}\|_X = o(\|x^k - \bar{x}\|_X) \quad (k \rightarrow \infty).$$

This proves the q-superlinear convergence.

If  $G'$  is  $\alpha$ -order Hölder continuous near  $\bar{x}$  with modulus  $L_\alpha > 0$ , then we can improve the estimate (2.45):

$$\begin{aligned}
\|x(z) - \bar{x}\|_X &\leq (L + \varepsilon) \|G(z) - G(\bar{x}) - G'(z)(z - \bar{x})\|_Y \\
&= (L + \varepsilon) \left\| \int_0^1 (G'(\bar{x} + t(z - \bar{x})) - G'(z))(z - \bar{x}) dt \right\|_Y \\
&\leq (L + \varepsilon) \int_0^1 \|G'(\bar{x} + t(z - \bar{x})) - G'(z)\|_{X \rightarrow Y} dt \|z - \bar{x}\|_X
\end{aligned}$$

$$\begin{aligned}
&\leq (L + \varepsilon) \int_0^1 L_\alpha (1-t)^\alpha \|z - \bar{x}\|_X^\alpha dt \|z - \bar{x}\|_X \\
&= \frac{L + \varepsilon}{1 + \alpha} L_\alpha \|z - \bar{x}\|_X^{1+\alpha} \\
&= O(\|z - \bar{x}\|_X^{1+\alpha}) \quad (z \rightarrow \bar{x}).
\end{aligned}$$

Hence,

$$\|x^{k+1} - \bar{x}\|_X = \|x(x^k) - \bar{x}\|_X = O(\|x^k - \bar{x}\|_X^{1+\alpha}) \quad (k \rightarrow \infty).$$

### 2.6.3 SQP Methods for Inequality Constrained Problems

We consider the problem

$$\min_{w \in W} f(w) \quad \text{s.t.} \quad e(w) = 0, \quad c(w) \in \mathcal{K}, \quad (2.46)$$

with  $f : W \rightarrow \mathbb{R}$ ,  $e : W \rightarrow Z$ , and  $c : W \rightarrow R$  twice continuously F-differentiable. Furthermore,  $W, Z, R$  are Banach spaces,  $R$  is reflexive (i.e.,  $R^{**} = R$ ), and  $\mathcal{K} \subset R$  is a nonempty closed convex cone.

For this problem, we define the Lagrange function

$$L(w, \lambda, \mu) = f(w) + \langle \lambda, c(w) \rangle_{R^*, R} + \langle \mu, e(w) \rangle_{Z^*, Z}.$$

We will need the notion of the polar cone.

**Definition 2.6** Let  $X$  be a Banach space and let  $\mathcal{K} \subset X$  be a nonempty closed convex cone. Then the *polar cone* of  $\mathcal{K}$  is defined by

$$\mathcal{K}^\circ = \{y \in X^* : \langle y, x \rangle_{X^*, X} \leq 0 \quad \forall x \in \mathcal{K}\}.$$

Obviously,  $\mathcal{K}^\circ$  is a closed convex cone.

Recall also the definition of the normal cone mapping (Def. 2.4).

Under a constraint qualification, see Sect. 1.7.3.2, the following KKT conditions hold:

There exist Lagrange multipliers  $\bar{\lambda} \in \mathcal{K}^\circ$  and  $\bar{\mu} \in Z^*$  such that  $(\bar{w}, \bar{\lambda}, \bar{\mu})$  satisfies

$$\begin{aligned}
L_w(\bar{w}, \bar{\lambda}, \bar{\mu}) &= 0, \\
c(\bar{w}) &\in \mathcal{K}, \quad \bar{\lambda} \in \mathcal{K}^\circ, \quad \langle \bar{\lambda}, c(\bar{w}) \rangle_{R^*, R} = 0, \\
e(\bar{w}) &= 0.
\end{aligned}$$

The second condition can be shown to be equivalent to  $\text{VI}(-c(\bar{w}), \mathcal{K}^\circ)$ . This is a VI w.r.t.  $\lambda$  with a constant operator parameterized by  $\bar{w}$ .

Now comes the trick, see, e.g., [5]:

By means of the normal cone  $N_{\mathcal{K}^\circ}$ , it is easily seen that  $\text{VI}(-c(w), \mathcal{K}^\circ)$  is equivalent to the generalized equation

$$0 \in -c(w) + N_{\mathcal{K}^\circ}(\lambda).$$

Therefore, we can write the KKT system as a generalized equation:

$$0 \in \begin{pmatrix} L_w(w, \lambda, \mu) \\ -c(w) \\ e(w) \end{pmatrix} + \begin{pmatrix} \{0\} \\ N_{\mathcal{K}^\circ}(\lambda) \\ \{0\} \end{pmatrix}. \quad (2.47)$$

Setting

$$N(w, \lambda, \mu) = \begin{pmatrix} \{0\} \\ N_{\mathcal{K}^\circ}(\lambda) \\ \{0\} \end{pmatrix},$$

and noting  $L_\lambda(w, \lambda, \mu) = c(w)$ ,  $L_\mu(w, \lambda, \mu) = e(w)$ , we can write (2.47) very compactly as  $\text{GE}(-L', N)$ .

The closed graph of the normal cone mapping is proved in the next lemma.

**Lemma 2.11** *Let  $X$  be a Banach spaces and  $S \subset X$  be nonempty, closed, and convex. Then the normal cone mapping  $N_S$  has closed graph.*

*Proof* Let  $\text{graph}(N_S) \ni (x^k, y^k) \rightarrow (\bar{x}, \bar{y})$ . Then  $y^k \in N_S(x^k)$  and thus  $x^k \in S$ , since otherwise  $N_S(x^k)$  would be empty. Since  $S$  is closed,  $\bar{x} \in S$  follows. Now, for all  $z \in S$ , by continuity

$$\langle \bar{y}, z - \bar{x} \rangle_{X^*, X} = \lim_{k \rightarrow \infty} \underbrace{\langle y^k, z - x^k \rangle_{X^*, X}}_{\leq 0} \leq 0,$$

hence  $\bar{y} \in N_S(\bar{x})$ . Therefore,  $(\bar{x}, \bar{y}) \in \text{graph}(N_S)$ .

If we now apply the Josephy-Newton method to (2.47), we obtain the following subproblem (we set  $x^k = (w^k, \lambda^k, \mu^k)$ ):

$$0 \in \begin{pmatrix} L_w(x^k) \\ -c(w^k) \\ e(w^k) \end{pmatrix} + \begin{pmatrix} L_{ww}(x^k) & c'(w^k)^* & e'(w^k)^* \\ -c'(w^k) & 0 & 0 \\ e'(w^k) & 0 & 0 \end{pmatrix} \begin{pmatrix} w - w^k \\ \lambda - \lambda^k \\ \mu - \mu^k \end{pmatrix} + \begin{pmatrix} \{0\} \\ N_{\mathcal{K}^\circ}(\lambda) \\ \{0\} \end{pmatrix}. \quad (2.48)$$

It is not difficult to see that (2.48) are exactly the KKT conditions of the following quadratic optimization problem:

### 2.6.3.1 SQP Subproblem

$$\begin{aligned} \min_{w \in W} \quad & \langle f'(w^k), w - w^k \rangle_{W^*, W} + \frac{1}{2} \langle L_{ww}(x^k)(w - w^k), w - w^k \rangle_{W^*, W} \\ \text{s.t.} \quad & e(w^k) + e'(w^k)(w - w^k) = 0, \quad c(w^k) + c'(w^k)(w - w^k) \in \mathcal{K}. \end{aligned}$$

In fact, the Lagrange function of the QP is

$$\begin{aligned} L^{qp}(x) = & \langle f'(w^k), w - w^k \rangle_{W^*, W} + \frac{1}{2} \langle L_{ww}(x^k)(w - w^k), w - w^k \rangle_{W^*, W} \\ & + \langle \lambda, c(w^k) + c'(w^k)(w - w^k) \rangle_{W^*, W} \\ & + \langle \mu, e(w^k) + e'(w^k)(w - w^k) \rangle_{Z^*, Z}. \end{aligned}$$

Since

$$\begin{aligned} L_w^{qp}(x) = & f'(w^k) + L_{ww}(x^k)(w - w^k) + c'(w^k)^* \lambda + e'(w^k)^* \mu \\ = & L_w(x^k) + L_{ww}(x^k)(w - w^k) + c'(w^k)^*(\lambda - \lambda^k) + e'(w^k)^*(\mu - \mu^k), \end{aligned}$$

we see that writing down the KKT conditions for the QP in the form (2.47) gives exactly the generalized equation (2.48).

We obtain:

**Algorithm 2.20** (SQP method for inequality constrained problems)

0. Choose  $(w^0, \lambda^0, \mu^0)$  (sufficiently close to  $(\bar{w}, \bar{\lambda}, \bar{\mu})$ ).

For  $k = 0, 1, 2, \dots$ :

1. If  $(w^k, \lambda^k, \mu^k)$  is a KKT triple of (2.46), STOP.
2. Compute the KKT triple  $(d^k, \lambda^{k+1}, \mu^{k+1})$  of

$$\begin{aligned} \min_{d \in W} \quad & \langle f'(w^k), d \rangle_{W^*, W} + \frac{1}{2} \langle L_{ww}(w^k, \lambda^k, \mu^k)d, d \rangle_{W^*, W} \\ \text{s.t.} \quad & e(w^k) + e'(w^k)d = 0, \quad c(w^k) + c'(w^k)d \in \mathcal{K}, \end{aligned}$$

that is closest to  $(0, \lambda^k, \mu^k)$ .

3. Set  $w^{k+1} = w^k + d^k$ .

Since this method is the Josephy-Newton algorithm applied to (2.47), we can derive local convergence results immediately if Robinson's strong regularity condition is satisfied. This condition has to be verified from case to case and is connected to second order sufficient optimality conditions. As an example where strong regularity is verified for an optimal control problem, we refer to [56].

### 2.6.3.2 Application to Optimal Control

For illustration, we consider the nonlinear elliptic optimal control problem

$$\begin{aligned} \min_{y \in H_0^1(\Omega), u \in L^2(\Omega)} J(y, u) &\stackrel{\text{def}}{=} \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 \\ \text{s.t.} \quad &Ay + y^3 + y = u, \quad u \leq b. \end{aligned} \quad (2.49)$$

Here,  $y \in H_0^1(\Omega)$  is the state, which is defined on the open bounded domain  $\Omega \subset \mathbb{R}^n$ ,  $n \leq 3$ , and  $u \in L^2(\Omega)$  is the control. Furthermore,  $A : H_0^1(\Omega) \rightarrow H^{-1}(\Omega) = H_0^1(\Omega)^*$  is a linear elliptic partial differential operator, e.g.,  $A = -\Delta$ . Finally  $b \in L^\infty(\Omega)$  is an upper bound on the control. We convert this control problem into the form (2.46) by setting

$$\begin{aligned} Y &= H_0^1(\Omega), \quad U = L^2(\Omega), \quad Z = H^{-1}(\Omega), \\ e(y, u) &= Ay + y^3 + y - u, \quad c(y, u) = u - b, \\ \mathcal{K} &= \left\{ u \in L^2(\Omega) : u \leq 0 \text{ a.e. on } \Omega \right\}. \end{aligned}$$

One can show (note  $n \leq 3$ ) that the operator  $e$  is twice continuously F-differentiable with

$$e_y(y, u) = A + 3y^2 \cdot I + I, \quad e_{yy}(y, u)(h_1, h_2) = 6yh_1h_2$$

(the other derivatives are obvious due to linearity). Therefore, given  $x^k = (y^k, u^k, \lambda^k, \mu^k)$ , the SQP subproblem reads

$$\begin{aligned} \min_{d_y, d_u} & (y^k - y_d, d_y)_{L^2(\Omega)} + \alpha(u^k, d_u)_{L^2(\Omega)} + \frac{1}{2} \|d_y\|_{L^2(\Omega)}^2 \\ & + \frac{1}{2} \langle \mu^k, 6y^k d_y^2 \rangle_{H_0^1(\Omega), H^{-1}(\Omega)} + \frac{\alpha}{2} \|d_u\|_{L^2(\Omega)}^2 \\ \text{s.t.} \quad & Ay^k + (y^k)^3 + y^k - u^k + Ad_y + 3(y^k)^2 d_y + d_y - d_u = 0, \\ & u_k + d_u \leq b. \end{aligned}$$

## 2.7 State-Constrained Problems

So far, we focused on optimization problems with control constraints. Only very recently, advances in the analysis of Newton based algorithms for state constrained problems have been made and much is to be done yet. We cannot go into a detailed discussion of this topic here. Rather, we just briefly sketch a couple of promising approaches that are suitable for state constrained optimization problems.

### 2.7.1 SQP Methods

In the case of SQP methods, state constraints do not pose direct conceptual difficulties, at least not at a first glance. In fact, sequential quadratic programming is applicable to very general problem settings. The constraints are linearized to generate the QP subproblems, i.e., state constraints arise as linearized state constraints in the subproblems and the difficulties of dealing with state constraints is thus transported to the subproblems. However, the efficient solution of the QP subproblems is not the only challenge. In fact, it is important to emphasize that second order optimality theory is challenging in the case of state constraints. Second order sufficient optimality conditions are closely linked to strong regularity of the generalized equation corresponding to the KKT conditions. Therefore, proving fast local convergence of SQP methods for problems with state constraints is challenging. Recent progress in second order optimality theory, e.g., [31, 57] may help paving the ground for future progress in this field.

### 2.7.2 Semismooth Newton Methods

The application of semismooth reformulation techniques for state constraints poses principal difficulties. In fact, consider for illustration the following model problem:

$$\begin{aligned} \min_{y,u} \quad & J(y, u) := \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 \\ \text{s.t.} \quad & -\Delta y = u \quad \text{on } \Omega, \\ & y = 0 \quad \text{on } \partial\Omega, \\ & y \leq b \quad \text{on } \Omega. \end{aligned} \tag{2.50}$$

Here,  $n \leq 3$  and  $\Omega \subset \mathbb{R}^n$  is open and bounded with  $C^2$  boundary. Furthermore,  $b \in H^2(\Omega)$ ,  $b > 0$ ,  $\alpha > 0$ , and  $y_d \in L^2(\Omega)$ . From regularity results, see Theorem 1.28, we know that for  $u \in U := L^2(\Omega)$  there exists a unique weak solution  $y \in Y := H_0^1(\Omega) \cap H^2(\Omega) \hookrightarrow C(\bar{\Omega})$  of the state equation. The existence and uniqueness of a solution  $(\bar{y}, \bar{u}) \in Y \times U$  is easy to show by standard arguments.

Similar to the analysis of problem (1.144) it can be shown, see, e.g., [12], that the following optimality conditions hold at the solution: There exists a regular Borel measure  $\bar{\mu} \in \mathcal{M}(\Omega)$  and an adjoint state  $\bar{p} \in L^2(\Omega)$  such that

$$-\Delta \bar{y} = \bar{u} \quad \text{on } \Omega, \tag{2.51}$$

$$\bar{y} = 0 \quad \text{on } \partial\Omega, \tag{2.52}$$

$$(\bar{p}, -\Delta v)_{L^2(\Omega)} + \langle \bar{\mu}, v \rangle_{\mathcal{M}(\Omega), C(\bar{\Omega})} = (y_d - \bar{y}, v)_{L^2(\Omega)} \quad \forall v \in Y, \tag{2.53}$$

$$\bar{y} \leq b, \quad \langle \bar{\mu}, v - \bar{y} \rangle_{\mathcal{M}(\Omega), C(\bar{\Omega})} \leq 0 \quad \forall v \in C(\bar{\Omega}), \quad v \leq b, \tag{2.54}$$

$$\alpha \bar{u} - \bar{p} = 0 \quad \text{in } \Omega. \quad (2.55)$$

The difficulty now is that the complementarity condition (2.54) between the function  $\bar{y}$  and the measure  $\bar{\mu}$  cannot be written in a pointwise fashion. Hence, nonsmooth pointwise reformulations as needed for semismooth Newton methods are not possible.

To avoid this difficulty, several approaches were presented recently.

### 2.7.2.1 Moreau-Yosida Regularization

One possibility is to treat the state constraint by a Moreau-Yosida regularization. The state constraint is converted to a penalty term, resulting in the following Moreau-Yosida regularized problem:

$$\begin{aligned} \min \quad & \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Omega)}^2 + \frac{1}{2\gamma} \|\max(0, \hat{\mu} + \gamma(y - b))\|_{L^2(\Omega)}^2 \\ \text{s.t.} \quad & -\Delta y = u \quad \text{on } \Omega, \\ & y = 0 \quad \text{on } \partial\Omega. \end{aligned} \quad (2.56)$$

Here  $\gamma > 0$  is a penalty parameter and  $\hat{\mu} \geq 0$ ,  $\hat{\mu} \in L^2(\Omega)$  is a shift parameter. For this problem without inequality constraints, the optimality conditions are

$$-\Delta \bar{y}_\gamma = \bar{u}_\gamma \quad \text{on } \Omega, \quad (2.57)$$

$$\bar{y}_\gamma = 0 \quad \text{on } \partial\Omega, \quad (2.58)$$

$$-\Delta \bar{p}_\gamma = y_d - \bar{y}_\gamma - \max(0, \hat{\mu} + \gamma(\bar{y}_\gamma - b)) \quad \text{on } \Omega \quad (2.59)$$

$$\bar{p}_\gamma = 0 \quad \text{on } \partial\Omega, \quad (2.60)$$

$$\alpha \bar{u}_\gamma - \bar{p}_\gamma = 0 \quad \text{on } \Omega. \quad (2.61)$$

To make this system more similar to the optimality conditions (2.51)–(2.55), we introduce

$$\bar{\mu}_\gamma = \max(0, \hat{\mu} + \gamma(\bar{y}_\gamma - b)).$$

We then can write the KKT conditions (2.57)–(2.61) in the form

$$-\Delta \bar{y}_\gamma = \bar{u}_\gamma \quad \text{on } \Omega, \quad (2.62)$$

$$\bar{y}_\gamma = 0 \quad \text{on } \partial\Omega, \quad (2.63)$$

$$-\Delta \bar{p}_\gamma + \bar{\mu}_\gamma = y_d - \bar{y}_\gamma \quad \text{on } \Omega, \quad (2.64)$$

$$\bar{p}_\gamma = 0 \quad \text{on } \partial\Omega, \quad (2.65)$$

$$\bar{\mu}_\gamma = \max(0, \hat{\mu} + \gamma(\bar{y}_\gamma - b)) \quad \text{on } \Omega, \quad (2.66)$$

$$\alpha \bar{u}_\gamma - \bar{p}_\gamma = 0 \quad \text{on } \Omega. \quad (2.67)$$

For further discussion, we rewrite (2.66) as follows

$$\begin{aligned} 0 &= \bar{\mu}_\gamma - \max(0, \hat{\mu} + \gamma(\bar{y}_\gamma - b)) \\ &= \bar{\mu}_\gamma - \max\left(0, \bar{\mu}_\gamma + \gamma\left(\bar{y}_\gamma - b + \frac{1}{\gamma}(\hat{\mu} - \bar{\mu}_\gamma)\right)\right). \end{aligned} \quad (2.68)$$

If, just for an informal motivation, we suppose for a moment that  $(\hat{\mu} - \bar{\mu}_\gamma)/\gamma$  becomes small for large  $\gamma$ , then we can interpret (2.68) as an approximation of

$$\bar{\mu}_\gamma = \max(0, \bar{\mu}_\gamma + \gamma(\bar{y}_\gamma - b)). \quad (2.69)$$

From earlier considerations we know that (2.69) is equivalent to

$$\bar{\mu}_\gamma \geq 0, \quad \bar{y}_\gamma - b \leq 0, \quad \bar{\mu}_\gamma (\bar{y}_\gamma - b) = 0,$$

which can be interpreted as a strong formulation of (2.54). This demonstrates the role of (2.66) as an approximation of (2.54).

We collect some results concerning the regularized solution tuple  $(\bar{y}_\gamma, \bar{u}_\gamma, \bar{p}_\gamma, \bar{\mu}_\gamma)$ , which we call primal dual path. The details can be found in [66, 67]:

For any  $\gamma_0 > 0$ , the primal dual path  $\gamma \in [\gamma_0, \infty) \mapsto (\bar{y}_\gamma, \bar{u}_\gamma, \bar{p}_\gamma, \bar{\mu}_\gamma)$  can be shown to be bounded in  $Y \times U \times L^2(\Omega) \times Y^*$  and Lipschitz continuous. In addition,  $\gamma \in (0, \infty) \rightarrow \bar{\mu}_\gamma \in L^2(\Omega)$  is locally Lipschitz continuous. Moreover  $(\bar{y}_\gamma, \bar{u}_\gamma, \bar{p}_\gamma, \bar{\mu}_\gamma)$  converges weakly to  $(\bar{y}, \bar{u}, \bar{p}, \bar{\mu})$  as  $\gamma \rightarrow \infty$  and the convergence  $(\bar{y}_\gamma, \bar{u}_\gamma) \rightarrow (\bar{y}, \bar{u})$  is even strong in  $Y \times U$ .

The idea is now to apply a semismooth Newton method to (2.62)–(2.67) for solving (2.56) approximately and to drive  $\gamma$  to infinity in an outer iteration. The analysis of this approach was carried out in, e.g., [66, 67]. The adaption of the parameter  $\gamma$  can be controlled by models of the optimal value function along the path.

### 2.7.2.2 Lavrentiev Regularization

A second approach to state constrained problems is Lavrentiev regularization [103, 104]. We again consider the problem (2.50). The idea is to replace the constraint

$$y \leq b$$

by

$$y + \varepsilon u \leq b$$

with a parameter  $\varepsilon > 0$ . If we then introduce the new artificial control  $w = y + \varepsilon u$ , we have  $u = (w - y)/\varepsilon$  and thus can express  $u$  in terms of  $w$ . The Lavrentiev

regularized problem, transformed to  $w$ , then is given by

$$\begin{aligned}
 \min J(y, w) &:= \frac{1}{2} \|y - y_d\|_{L^2(\Omega)}^2 + \frac{\alpha}{2\varepsilon^2} \|w - y\|_{L^2(\Omega)}^2 \\
 \text{s.t. } & -\varepsilon \Delta y + y = w \quad \text{on } \Omega, \\
 & y = 0 \quad \text{on } \partial\Omega, \\
 & w \leq b \quad \text{on } \Omega.
 \end{aligned} \tag{2.70}$$

Except for the modified  $L^2$ -regularization, this problem has the form of a control-constrained elliptic optimal control problem. It is not difficult to see that it is uniquely solvable and can be handled by semismooth Newton techniques.

Under suitable assumptions, it can be shown, see [104], that the regularized solution  $(\bar{y}_\varepsilon, \bar{u}_\varepsilon)$  converges strongly to the solution  $(\bar{y}, \bar{u})$  of (2.50) as  $\varepsilon \rightarrow 0^+$ .

## 2.8 Further Aspects

### 2.8.1 Mesh Independence

For numerical computations, we have to discretize the problem (Finite elements, finite differences, ...) and to apply the developed optimization methods to the discretized, finite dimensional problem. One such situation would be, for instance, to apply an SQP method to the discretization  $(P_h)$  of the infinite dimensional problem  $(P)$ . If this is properly done, we can interpret the discrete SQP method as an inexact (i.e. perturbed) version of the SQP method applied to  $(P)$ .

Abstractly speaking, we have an infinite dimensional problem  $(P)$  and an algorithm  $A$  for its solution. Furthermore, we have a family of finite dimensional approximations  $(P_h)$  of  $(P)$ , and discrete versions  $A_h$  of algorithm  $A$ . Here  $h > 0$  denotes the accuracy of discretization (with increasing accuracy as  $h \rightarrow 0$ ). Starting from  $x^0$  and the corresponding discrete point  $x_h^0$ , respectively, the algorithms  $A$  and  $A_h$  will generate sequences  $(x^k)$  and  $(x_h^k)$ , respectively. Mesh independence means that the convergence behavior of  $(x^k)$  and  $(x_h^k)$  become more and more alike as the discretization becomes more and more accurate, i.e., as  $h \rightarrow 0$ . This means, for instance, that  $q$ -superlinear convergence of Alg.  $A$  on a  $\delta$ -neighborhood of the solution implies the same rate of convergence for Alg.  $A_h$  on a  $\delta$ -neighborhood of the corresponding discrete solution as soon as  $h$  is sufficiently small.

Mesh independence results for Newton's method were established in, e.g., [3, 44]. The mesh independence of SQP methods and Josephy-Newton methods was shown, e.g., in [6, 45]. Furthermore, the mesh independence of semismooth Newton methods was established in [68].

### 2.8.2 Application of Fast Solvers

An important ingredient in PDE constrained optimization is the combination of optimization methods with efficient solvers (sparse linear solvers, multigrid, preconditioned Krylov subspace methods, etc.). It is by far out of the scope of this chapter to give details. Instead, we focus on just two simple examples.

For both semismooth reformulations of the elliptic control problems (2.25) and (2.32), we showed that the semismooth Newton system is equivalent to

$$\begin{pmatrix} I & 0 & A^* \\ 0 & I & -\frac{1}{\alpha} g^k \cdot B^* \\ A & -B & 0 \end{pmatrix} \begin{pmatrix} s_y^k \\ s_u^k \\ s_\mu^k \end{pmatrix} = \begin{pmatrix} r_1^k \\ r_2^k \\ r_3^k \end{pmatrix} \quad (2.71)$$

with appropriate right hand side. Here  $A \in \mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$  is an elliptic operator,  $B \in \mathcal{L}(L^{p'}(\Omega_c), H^{-1}(\Omega))$  with  $p' \in [1, 2)$ , and  $g^k \in L^\infty(\Omega_c)$  with  $g^k \in [0, 1]$  almost everywhere. We can do block elimination to obtain

$$\begin{pmatrix} I & A^* & 0 \\ A & -\frac{1}{\alpha} B(g^k \cdot B^*) & 0 \\ 0 & -\frac{g^k}{\alpha} \cdot B^* & I \end{pmatrix} \begin{pmatrix} s_y^k \\ s_\mu^k \\ s_u^k \end{pmatrix} = \begin{pmatrix} r_1^k \\ Br_2^k + r_3^k \\ r_2^k \end{pmatrix}.$$

The first two rows form a  $2 \times 2$  elliptic system for which very efficient fast solvers (e.g., multigrid [62]) exist.

Similar techniques can successfully be used, e.g., for elastic contact problems [139].

### 2.8.3 Other Methods

Our treatment of Newton-type methods is not at all complete. There exist, for instance, interior point methods that are very well suited for optimization problems in function spaces, see, e.g., [121, 122, 138, 140, 145].

<http://www.springer.com/978-1-4020-8838-4>

Optimization with PDE Constraints

Hinze, M.; Pinnau, R.; Ulbrich, M.; Ulbrich, S.

2009, XII, 270 p., Hardcover

ISBN: 978-1-4020-8838-4