The Hong Kong Polytechnic University

Department of Applied Mathematics

Globalization Techniques for Solving Nonlinear Problems and Applications

Jinhai Chen

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

March 2009

Certification of Originality

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of a university or other institute of higher learning, except where due acknowledgment is made in the text.

_____ (Signed)

Jinhai Chen (Name of student)

Abstract

This thesis is concentrated on the study of global techniques to the numerical solution of nonsmooth problems. These problems are directly relevant to the variational inequality problems, complementarity problems, heat transmission problems in medium, parabolic obstacle problems within financial mathematics, and control-state constrained optimal control problems. These global approaches are discussed deeply on convergence theories and computational results.

A vital point in the implementation of the global approaches to the minimization of a nonsmooth (merit) function, such as Newton methods based on path search, line search, and trust region algorithms, is the calculation of a (generalized) Jacobian matrix equation effectively, especially for large-scale problems. In Chapter 2, we consider a Krylov subspace strategy for the underlying global methods to solving nonsmooth equations. Such strategy has a main advantage of computing a generalized Jacobian matrix equation, especially, in many applications where the (generalized) Jacobian matrix is not practically computable, or is expensive to obtain. Another global strategy is to consider avoiding the minimization of a nonsmooth (merit) function, pseudotransient continuation in Chapter 3 may be a nice choice for this purpose.

Many practical problems have certain structures; if we could find these structures, we may design more suitable algorithms. These algorithms have not only global convergence, but also have special properties, for example: finite termination, monotonicity. Indeed, the Newton-type methods considered to solve piece-wise systems in Chapter 4 have a remarkable monotone convergence. These piece-wise systems arise from the discretizations of heat transmission problems in a medium, parabolic obstacle problems in financial mathematics.

Other strategies, for example: smoothing strategy, nonmonotone strategy, etc., have also rather good effects in most of the practical implementations, even if in the infinite dimensional spaces. To show this, numerical solution of optimal control problems subject to mixed control-state constraints has been investigated in Chapter 5. The necessary conditions of the optimal control problems are stated in terms of a local minimum

principle. By use of the Fischer-Burmeister function, the local minimum principle is transformed into an equivalent nonlinear and nonsmooth equation in appropriate Banach spaces. This nonlinear and nonsmooth equation is solved by inexact nonsmooth and smoothing Newton methods. The globalized methods are developed in a very general setting that allows for non-monotonicity of squared residual norm values at subsequent iterates. Numerical examples are presented to demonstrate the efficiency of these presented approaches.

Acknowledgments

I would like to thank my thesis advisor, Professor Liqun Qi, for providing timely direction and critical evaluation of my thesis research. I am thankful for his guidance as a mentor and for his wisdom as a professor. His style of supervision has encouraged growth, and he has molded an aspiring researcher to follow in his path. I am very fortunate to have had the opportunity to work with him.

I am grateful to Professor C. K. Kelley from North Carolina State University for having technical discussions with me that made me rethink certain concepts.

Special thanks go to Prof. Luigi Brugnano from Università di Firenze for his very helpful notes on piecewise linear system. Many thanks go to Prof. Matthias Gerdts from University of Würzburg and Martin Kunkel at University of Hamburg for providing me crucial suggestions on control-state constrained optimal control problems.

I have been infinitely lucky to have met Miss Jing-Yi Guo when we were both Ph.D. students in the Hong Kong Polytechnic University. As I made further acquaintance with Jing-Yi, I found her to be the most thoughtful, encouraging, and inspiring angel I could wish to have. Without her encouragement, I would not have finished this long journey.

This dissertation is dedicated to my parents. I thank my parents for years of toiling and putting up with my "life-long" studies. I am equally indebted to my elder sister Jinping Chen and brother-in-law Zengmao Cheng for their love and constant support.

Contents

Tables and Figures

Tables

Figures

Notation

$\mathbb{R}^n$	real $n-$dimensional Euclidean space
$x \in \mathbb{R}^n$	an $n-$dimensional vector
$x^\top$	the transpose of a vector x
$\{x_k\}$	a sequence of vectors $x_1, x_2, x_3, \ldots$
$\|x\|$	norm of x ($2-$norm unless otherwise stated)
$x^\top y$	the standard inner product of vector x and $y \in \mathbb{R}$
$f(x) : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$	a mapping from domain Ω onto range $\mathbb{R}^m$
$\nabla f(x)$	gradient of f
$f'(x)$	derivative of f
$\nabla^2 f(x)$	Hessian of f
$f'(x,h)$	directional derivative of f at x in a direction $h \in \mathbb{R}^n$
$O(\delta)$	the same order quantity as the small $\delta \in \mathbb{R}^n$
$o(\delta)$	higher order quantity than the small δ
conv$S$	convex hull of the set $S \subseteq \mathbb{R}^n$

For a locally Lipschitz function $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^m$

$\partial f(x)$	the generalized derivative of f in the sense of Clarke
$\partial_B f(x)$	the $B-$subdifferential of f

Chapter 1

Introduction

Consider the system of nonlinear equations

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

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a locally Lipschitzian function.

It is well known that if F is continuously differentiable (i.e., smooth) and F' is locally Lipschitz and invertible at a solution x^*, then there exists a ball $S(x^*, r),\ r > 0$ such that for any $x_0 \in S(x^*, r)$, the Newton method

$$x^{k+1} = x^k - F'(x^k)^{-1} F(x^k), \quad k \geq 0 \tag{1.2}$$

is quadratically convergent to x^*, see [98, 40].

In the nonsmooth case, $F'(x_k)$ may not exist. The generalized Newton method proposes to use generalized Jacobian matrix of F to play the role of F' in the Newton method (1.2) in the finite dimensional case.

The (generalized) Newton method is the prototype of many local, fast algorithms for solving (non)smooth equations. These algorithms have excellent convergence rates if the starting iterate point belongs to a suitably chosen neighborhood of the desired solution. In addition, the damped Newton and the damped Gauss-Newton methods were presented for improving the global convergence of algorithm [98, 40].

In this chapter, we mainly review the semismooth Newton method and some globalization techniques of the (generalized) Newton methods.

1.1 Generalized Newton-type Methods

In this section, we introduce the generalized Newton-type methods. For convenience, we collect first concepts about nonsmooth analysis.

The notion of B(ouligand)-derivative was proposed by Robinson in [115]. A function $F:\mathbb{R}^n\to\mathbb{R}^n$ is said to be B-differentiable at a point x if F has a one-sided directional derivative $F'(x,d)$ at x (see (1.6)) and

$$\lim_{d\to 0}\frac{F(x+d)-F(x)-F'(x,d)}{\|d\|}=0. \tag{1.3}$$

(1.3) can be written as $F(x+d)=F(x)+F'(x,d)+o(\|d\|)$, see [109]. Shapiro [123] showed that a locally Lipschitzian function F is B-differentiable at x if and only if it is directionally differentiable. Therefore, there is no difference between B-derivatives and directional derivatives in this chapter.

Suppose that $F:\mathbb{R}^n\to\mathbb{R}^n$ is locally Lipschitz but not necessarily smooth. Let

$$D_F:=\{x\in\mathbb{R}^n: F \text{ is differentiable at } x\}.$$

Then the generalized derivative of F at x in the sense of Clarke [30] is defined by

$$\partial F(x)=\text{conv}\partial_B F(x),$$

where $\text{conv}\partial_B F(x)$ denotes the convex hull of the set

$$\partial_B F(x)=\left\{\lim_{\substack{x^j\to x\\ x^j\in D_F}} F'(x^j)\right\},$$

which is called the $B-$subdifferential of F at $x\in\mathbb{R}^n$. The set $\partial F(x)$ is nonempty, convex and compact for any fixed point x. The function $F:\mathbb{R}^n\to\mathbb{R}^n$ is called semismooth [113, 109] at $x\in\mathbb{R}^n$ if F is directionally differentiable at x and for any $V\in\partial F(x+h)$,

$$Vh-f'(x,h)=o(\|h\|), \text{ holds as } h\to 0.$$

The function F is called $p-$order semismooth at x if the term $o(\|h\|)$ in the above

expression is replaced by $O(\|h\|^{1+p})$, and called strongly semismooth at x if the term $o(\|h\|)$ in the above expression is replaced by $O(\|h\|^2)$ [113, 109]. We say that F is BD-regular at $x \in \mathbb{R}^n$ if all the elements in $\partial_B F(x)$ are $n \times n$ nonsingular matrices.

In fact, semismoothness was originally introduced by Mifflin [96] for functionals, which plays an important role in the global convergence theory of nonsmooth optimization, see Facchinei and Pang [46]. Qi and Sun [113, 109] extended the concept of semismoothness to vector-valued functions.

We also need some lemmas for our discussion.

Lemma 1.1 (see [113, Lemma 2.2]). *Suppose that $F : \mathbb{R}^n \to \mathbb{R}^n$ is a locally Lipschitzian function and $F'(x,h)$ exists for any h at x. Then*

(i) *$F'(x,h)$ is Lipschitzian in h,*

(ii) *for any h there exists a $V \in \partial F(x)$ such that $F'(x,h) = Vh$.*

Lemma 1.2 (see [102, Proposition 3]). *If F is BD-regular at $x \in \mathbb{R}^n$, then there is a neighborhood N of x and a positive constant α such that for any $y \in N$ and $V \in \partial_B F(y)$, V is nonsingular and $\|V^{-1}\| \le \alpha$. If, furthermore, $F(x) = 0$ and F is semismooth at x, then there is a neighborhood N' of x and a positive constant β such that*

$$\|F(y)\| \ge \beta \|y - x\|, \quad \forall y \in N'.$$

The Newton method for nonsingular nonsmooth equations using the generalized Jacobian matrix is defined by

$$x^{k+1} = x^k - (V^k)^{-1} F(x^k), \quad V^k \in \partial F(x^k). \tag{1.4}$$

A local superlinear convergence theorem is given in [113], where it is assumed that all $V \in \partial F(x^*)$ are nonsingular.

Qi [109] suggested a modified version of method (1.4) in the form

$$x^{k+1} = x^k - (V^k)^{-1} F(x^k), \quad V^k \in \partial_B F(x^k), \tag{1.5}$$

and gave a local superlinear convergence theorem for method (1.5). His theorem reduced the nonsingularity requirement on all members of $\partial F(x^*)$ to all members of

$\partial_B F(x^*)$.

Another modification is an iteration function method introduced by Han, Pang and Rangaraj [61] using an iteration function $G(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. If F has a one-sided directional derivative

$$F'(x,d) := \lim_{t \downarrow 0} \frac{F(x+td) - F(x)}{t} \tag{1.6}$$

and $G(x,d) = F'(x,d)$, a variant of the iteration function method can be defined by

$$\begin{cases} \text{solve } F(x^k) + G(x^k,d) = 0, \\ \text{set } x^{k+1} = x^k + d. \end{cases} \tag{1.7}$$

This modification is actually a generalization of Pang's B-differentiable Newton method [101, 100].

In practice, we should note that computing the exact solution (1.7) could be expensive if n is large and, for any n, may not be justified when x is far from a solution. This difficulty motivates us to invoke another classical tool for smooth (nonsmooth) equations: the inexact Newton method [122, 37, 101, 94, 137]. Actually, the notion of inexact solution in algorithms for solving nonsmooth equations was suggested in [101] and has been employed in [94, 74, 45].

Algorithm 1.3. INEXACT NEWTON METHOD

Let x^0 be given. For $k = 0$ step 1 until convergence do:

Find some $\eta_k \in [0,1)$ and a vector d^k that satisfy

$$\|F(x^k) + V^k d^k\| \le \eta_k \|F(x^k)\|. \tag{1.8}$$

Set $x^{k+1} = x^k + d^k$.

Here $V^k \in \partial_B F(x^k)$, $\{\eta_k\}$ is a sequence of forcing terms.

1.2 Globalization Techniques

The semismooth Newton method and its convergence results can be applied to some important mathematical programming problems such as nonlinear complementarity problems, variational inequalities and KKT conditions of optimization; see [45, 74, 94] and

the state-of-the-art monograph [46].

Note that (1.5) is only convergent locally under semismoothness assumption. A natural question is that whether (1.5) can be globalized similar to classic Newton's method for solving smooth equations or not. In general, the answer is negative because the function Θ defined by

$$\Theta(x) = \frac{1}{2}\|F(x)\|^2 \tag{1.9}$$

is not smooth. Fortunately, in some especial but important cases, Θ can be smooth though F itself may not smooth. For example, in order to solve a nonlinear complementarity problem (NCP), i.e., find a vector $x \in \mathbb{R}^n$ such that

$$x \geq 0, \quad G(x) \geq 0, \quad x^T G(x) = 0,$$

where $G : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable with $G(x) = (G_1(x), \ldots, G_n(x))^T$, we usually reformulate this NCP as a system of nonlinear equations

$$F_i(x) = \phi(x_i, G_i(x)), \quad i = 1, \ldots, n, \tag{1.10}$$

via a so-called complementarity function $\phi : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\phi(a,b) = 0 \Longleftrightarrow a \geq 0, b \geq 0, ab = 0.$$

The well-known min function ($\phi(a,b) = \min(a,b)$) and Fischer-Burmeister function ($\phi(a,b) = \sqrt{a^2+b^2} - (a+b)$) are both complementarity functions. Furthermore, the (1.10) reformulated by Fischer-Burmeister function [48] is not differentiable at $x = 0$, but $\Theta(x)$ is smooth. See Jiang and Ralph [72], Qi [110]. By assuming that $\Theta(x)$ is smooth, various globalized semismooth Newton methods could be established, see [72, 110, 46] for details.

As the classical Newton method, roughly speaking, there are two seminal classes of globalization techniques

- Line search strategy
- Trust region strategy

Most globalized methods are achieved via the above two strategies. For example,

we illustrate a global Newton method with line search strategy,

Algorithm 1.4. INEXACT SEMISMOOTH NEWTON METHOD WITH LINE SEARCH

Let x^0 be given, $\beta \in (0,1)$, $\eta_{\max} \in [0,1)$ and $0 < \theta_{\min} < \theta_{\max} < 1$ be given.

For $k = 0$ step 1 until convergence do:

Find some $\eta_k \in [0, \eta_{\max}]$ and a vector d^k that satisfy

$$\|F(x^k) + V^k d^k\| \leq \eta_k \|F(x^k)\|, \tag{1.11}$$

and then choose $\theta \in [\theta_{\min}, \theta_{\max}]$, update $\lambda_k \leftarrow \theta\lambda_k$ until the following inequality is satisfied

$$\Theta(x^k + \lambda_k d^k) \leq \Theta(x^k) + \beta\lambda_k \nabla\Theta(x^k)^T d^k \tag{1.12}$$

where $\nabla\Theta(x^k) = (V^k)^\top F(x^k)$ with $V^k \in \partial_B F(x^k)$.

Set $x^{k+1} = x^k + \lambda_k d^k$.

The merits of those methods based on the above globalization techniques are that they are globally and superlinearly (quadratically) convergent, and at each iteration only a system of linear equations needs to be solved [118].

In the practical implementations of these algorithms, some issues should be paid exceptional attention, such as the calculation of a generalized Jacobian matrix (or gradient) for large-scale problems, minimization of a nonsmooth (merit) function, structures of the problems, smoothing strategy, nonmonotone strategy, etc., because these strategies could improve considerably the efficiency of the underlying approaches. In order to show these asserts, a Krylov subspace strategy for solving nonsmooth equations are proposed in Chapter 2. Such strategy has a main advantage of computing a generalized Jacobian matrix, especially, in many applications where the (generalized) Jacobian matrix equation is not directly computable, or is expensive to obtain. Instead of the minimization of a nonsmooth (merit) function, another global strategy, pseudo-transient continuation, is considered in Chapter 3. Chapter 4 designs the Newton-type methods to solve piece-wise systems arise from the discretizations of heat transmission problems in a medium, parabolic obstacle problems within financial mathematics, due to certain structure of these problems. Chapter 5 investigates the numerical solution of optimal control problems subject to mixed control-state constraints. By use of the Fischer-Burmeister function, the necessary conditions of the optimal control problems

stated in terms of a local minimum principle is transformed into an equivalent nonlinear and nonsmooth equation in appropriate Banach spaces. This nonlinear and nonsmooth equation is solved by inexact nonsmooth and smoothing Newton methods. The globalized methods are developed in a very general setting that allows for non-monotonicity of squared residual norm values at subsequent iterates.

Chapter 2

Inexact Newton-Krylov Algorithms

In this chapter[1] we consider the system of nonlinear equations

$$F(x) = 0, \tag{2.1}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is a locally Lipschitzian function. Qi [109] suggested a modified version of method (1.4) in the form

$$x^{k+1} = x^k - (V^k)^{-1} F(x^k), \quad V^k \in \partial_B F(x^k), \tag{2.2}$$

and gave a local superlinear convergence theorem for method (2.2). His theorem reduced the nonsingularity requirement on all members of $\partial F(x^*)$ to all members of $\partial_B F(x^*)$. Further references to the development of approaches based on this idea can be found in [107, 112, 128, 28, 27, 29] and references therein. Among them the first globally and superlinearly (quadratically) convergent smoothing Newton method was proposed by Chen, Qi and Sun in [29].

Modification of an iteration function method was introduced by Han, Pang and Rangaraj [61] using an iteration function $G(\cdot,\cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$. If $F(x)$ has a one-sided directional derivative and $G(x,d) = F'(x,d)$, a variant of the iteration function method can be defined by

Algorithm 2.1**.** Let x^0 be given. For $k = 0$ step 1 until convergence do:

[1] This chapter is taken from [25].

Find a vector d^k that satisfies

$$F(x^k) + G(x^k, d^k) = 0. \tag{2.3}$$

Set $x^{k+1} = x^k + d^k$.

This modification is actually a generalization of Pang's B-differentiable Newton method [101, 100]. Other analysis of Newton methods for Lipschitz equations as well as extensions and applications can be found in the monograph [46].

On the other hand, computing an exact solution of Jacobian linear system in (2.2) or (2.3) can be expensive if n is large and, for any n, may not be justified when x is far from a solution. This difficulty motivates us to invoke another classical tool for smooth (nonsmooth) equations: the inexact Newton method [122, 37, 101, 94, 137]. Actually, the notion of inexact solution in algorithms for solving nonsmooth equations was suggested in [101] and has been employed in [94, 74, 45].

In [94] iteration processes of a general form led to the following algorithm, called inexact Newton method (when $\partial_B F = \{F'(x^k)\}$, i.e. $F(x)$ is strictly differentiable).

Algorithm 2.2**.** Let x^0 be given. For $k = 0$ step 1 until convergence do:

Find some $\eta_k \in [0,1)$ and a vector d^k that satisfy

$$\|F(x^k) + V^k d^k\| \le \eta_k \|F(x^k)\|. \tag{2.4}$$

Set $x^{k+1} = x^k + d^k$.

Here $V^k \in \partial_B F(x^k)$, $\{\eta_k\}$ is a sequence of forcing terms.

If $F(x)$ has Fréchet differential $F'(x)$, the global convergence of the inexact Newton method is obtained by augmenting the inexact Newton condition with a sufficient monotone decrease condition on the merit function $\|F(x)\|$ introduced by Eisenstat and Walker [122].

Algorithm 2.3**.** Let x^0 be given. For $k = 0$ step 1 until convergence do:

Find some $\eta_k \in [0,1)$ and a vector d^k that satisfy

$$\|F(x^k) + F'(x^k) d^k\| \le \eta_k \|F(x^k)\|, \tag{2.5}$$

and

$$\|F(x^k + d^k)\| \leq (1 - \beta(1 - \eta_k))\|F(x^k)\|, \text{ where } \beta \in (0,1). \tag{2.6}$$

In the kth iteration, the acceptability conditions on the trial step d^k are used to the monotone technique. Martínez and Qi [94] established a global convergence of the inexact Newton method for nonsmooth equations using a sufficient monotone decrease condition on the merit function $\frac{1}{2}\|F(x)\|^2$.

Several authors have succeeded using Krylov subspace methods inside a Newton iteration in the context of the systems of smooth equations and unconstrained optimization. See [6, 7, 10, 12, 13, 79, 80, 84] and references therein. One of the main advantages of Krylov subspace methods is that these solvers are often faster than direct methods even if the Jacobian matrix is small and dense. In this chapter we attempt to use variants of Newton's iteration in association with Krylov subspace methods for solving the generalized Jacobian linear systems.

This chapter is organized as follows. We review inexact Newton methods and present a nonmonotone version of inexact Newton-Krylov methods for nonsmooth equations in Section 2.1. In Section 2.3 we analyze the corresponding global and local convergence. In Section 2.4, we give a smoothing variant of inexact Newton-Krylov methods for nonsmooth problems. In Section 2.5, we give some numerical examples. Finally, we make some concluding remarks in Section 2.6.

Notation: Throughout the chapter $\|\cdot\|$ will denote the Euclidean norm. However, it is easy to verify that many results are independent of this choice.

2.1 Inexact Newton Methods

By the use of nonsmooth analysis illustrated in Chapter 1, this section will present the basic ideas of inexact Newton methods and Newton-Krylov methods for nonsmooth equations. The blanket assumptions made throughout this chapter are Assumptions 2.4 and 2.5.

Assumption 2.4**.** The function $F(x)$ is semismooth or stronger, p-order semismooth, $0 < p \leq 1$.

Assumption 2.5**.** Each component function F_i of $F(x)$ is continuously differentiable on $\mathbf{R}^n \setminus F_i^{-1}(0)$.

Ulbrich [132] has shown that Assumptions 2.5 holds, e.g., for complementarity problems based on Fisher-Burmeister function [48], which is used in our numerical tests in Section 2.5. The continuous differentiability of the merit function $h(x) = \frac{1}{2}\|F(x)\|^2$ was also established by Ulbrich (see [132, Lemma 4.2]) under the Assumption 2.5.

Lemma 2.6. *Under the Assumptions 2.4 and 2.5 on $F : \mathbb{R}^n \to \mathbb{R}^n$, the merit function $h(x)$ is continuously differentiable on $\mathbb{R}^n$ with gradient $\nabla h(x) = V^\top F(x)$, where $V \in \partial F(x)$ is arbitrary.*

Generally, global convergence of inexact Newton method for smooth equations can be obtained by many globalization techniques, such as: the inexact-Newton backtracking method, general inexact-Newton trust region methods and dogleg implementations introduced by Eisenstat and Walker in [122], see also [103]. Another globalization technique we referred here is augmenting the inexact generalized Newton condition with some line search methods, such as: the Armijo-Goldstein rule on the merit function $h(x)$ introduced by Brown and Saad [12, 13], wherein the step length λ_k of d^k in (2.4), must satisfy

$$h(x^k + \lambda_k d^k) \le h(x^k) + \beta \lambda_k \nabla h(x^k)^\top d^k \tag{2.7}$$

where $\beta \in (0,1)$.

In our algorithms of this chapter, relaxing the acceptability conditions on the trial step d^k, we suggest to use the nonmonotone technique:

$$h(x^{l(k)}) = \max_{0 \le j \le m(k)} \{h(x^{k-j})\}$$

instead of $h(x^k)$ in (2.7), where $m(0) = 0$ and $0 \le m(k) \le \min\{m(k-1)+1, M\}$, $k \ge 1$. When $F(x)$ is continuously differentiable in a neighborhood of x^k, this nonmonotonic technique can also be found in [39, 59, 137]. A nonmonotonic criterion should bring about speeding up the convergence progress in some ill-conditioned cases. For the case of $F(x)$ being nonsmooth, the nonmonotonic inexact Newton algorithm can be stated as the following general form:

Algorithm 2.7. INEXACT NEWTON METHOD FOR NONSMOOTH EQUATIONS

Let x^0 be given, $\beta \in (0,1)$, $\lambda_0 \in (0,1]$, $\eta_{\max} \in [0,1)$ and $0 < \theta_{\min} < \theta_{\max} < 1$ be given.

For $k = 0$ step 1 until convergence do:

Select an element of $V^k \in \partial_B F(x^k)$. Find some $\eta_k \in [0, \eta_{\max}]$ and a vector d^k that satisfy

$$\|F(x^k) + V^k d^k\| \leq \eta_k \|F(x^k)\|, \tag{2.8}$$

and then choose $\theta \in [\theta_{\min}, \theta_{\max}]$, update $\lambda_k \leftarrow \theta\lambda_k$ until the following inequality is satisfied:

$$h(x^k + \lambda_k d^k) \leq h(x^{l(k)}) + \beta\lambda_k \nabla h(x^k)^\top d^k \tag{2.9}$$

where $\nabla h(x) = (V^k)^\top F(x^k)$, and $h(x^{l(k)}) = \max\limits_{0 \leq j \leq m(k)} \{h(x^{k-j})\}$, with the nonmonotone index function $m(0) = 0, 0 \leq m(k) \leq \min\{m(k-1)+1, M\}$, $k \geq 1$.

Set $x^{k+1} = x^k + \lambda_k d^k$.

Here the framework for reducing λ_k is taken from [40], which allows much flexibility and sophistication in reducing λ_k. Lemma 2.12 in Section 2.3 of this chapter will ensure that λ_k satisfies (2.9). For simplicity, we use the particular sequence of backtracking parameters $\lambda_k = 1, \omega, \omega^2, \ldots$ $(\omega \in (0,1))$.

2.2 Krylov Subspace Strategy

If $F(x)$ is Fréchet differentiable and n is large, a d^k satisfying the residual condition (2.8) is often obtained by using an iterative procedure for linear systems. In [12], Brown and Saad considered using the Arnoldi and GMRES [117, 11] algorithms for nonsymmetric linear systems to obtain d^k's satisfying the residual condition (2.8) and proved the existence of such a d^k.

At each iteration of the inexact Newton method, we must obtain an approximate solution of the nonlinear system (2.8) which we rewrite as

$$Vd = -F(x). \tag{2.10}$$

If d^0 is an initial guess for the true solution of (2.10), then letting $d = d^0 + z$, we have the equivalent system

$$Vz = r^0, \tag{2.11}$$

where $r^0 = -F(x) - Vd^0$ is the initial residual. For a general matrix A and a vector v,

define the Krylov subspace $K(A,v,m)$ by

$$K(A,v,m) = \text{span}\{v, Av, \ldots, A^{m-1}v\}.$$

Let K^m denote

$$K^m \equiv K(V, r^0, m).$$

Both Arnoldi's method and GMRES find an approximate solution

$$d^m = d^0 + z^m, \text{ with } z^m \in K^m,$$

such that either

$$(-F(x) - Vd^m) \perp K^m \left(\text{ equivalently } (r^0 - Vz^m) \perp K^m \right) \tag{2.12}$$

for Arnoldi's method, or

$$\|F(x) + Vd^m\| = \min_{d \in d^0 + K^m} \|F(x) + Vd^m\| \left(= \min_{z \in K^m} \|r^0 - Vz\| \right) \tag{2.13}$$

for GMRES. Note that condition (2.13) is equivalent to demanding that the residual $r^m = -F(x) - F'(x, d^m)$ be orthogonal to $F'(x, K^m)$. Combined with Algorithm 2.7, we get the *inexact Newton-Krylov* methods (such as inexact Newton-GMRES) for nonsmooth equations.

For simplicity, we have omitted details of the practical implementations of the above linear and nonlinear methods, which are discussed at length in [6, 11, 12, 13, 10] for smooth problems. Some other inexact Newton-Krylov methods for nonsmooth equations such as inexact Newton-CGS, inexact Newton-BiCG [10, 79, 80], and so on, can be established similarly.

Due to the semismoothness of F at x^k and Lemma 1.1, the matrix-vector multiplication $V^k d^k$ is not usually approximated well by a difference quotient of the form

$$F'(x^k, d^k) \approx \frac{F(x^k + td^k) - F(x^k)}{t}, \tag{2.14}$$

compared with the smooth case.

2.3 Convergence Theory

In this section, we develop a theoretical foundation for an inexact Newton-Krylov algorithm of nonsmooth problems. Given $x^0 \in \mathbb{R}^n$, we denote a sequence generated in our algorithm by $\{x^k\} \subseteq \mathbb{R}^n$, and the level set of $\|F(x)\|$ by

$$L(x^0) = \{x \in \mathbb{R}^n | \|F(x)\| \leq \|F(x^0)\|\}. \tag{2.15}$$

Throughout this section we make the following assumption.

Assumption 2.8**.** The sequence $\{x^k\}$ generated by Algorithm 2.7 is contained in a compact set $L(x^0)$ on $\mathbb{R}^n$.

In the kth iteration, to guarantee that the current iteration will make progress towards the solution in one step of the Algorithm 2.7, we must know how the generalized inexact Newton step d^k satisfies (2.8).

Proposition 2.9**.** *Assume that there exists d^k satisfying (2.8) when $\|(V^k)^\top F(x^k)\| = 0$ where $V^k \in \partial_B F(x^k)$, then $\|F(x^k)\| = 0$. Further, if $F(x)$ is BD-regular at x^k, then $d^k = 0$.*

Proof. If d^k satisfies the inequality (2.8) when $\|(V^k)^\top F(x^k)\| = 0$, squaring Euclidean norm in both sides of the inequality (2.8), we have

$$\|F(x^k)\|^2 + 2[(V^k)^\top F(x^k)]^\top d^k + \|V^k d^k\|^2 = \|F(x^k) + V^k d^k\|^2 \leq \eta_k^2 \|F(x^k)\|^2, \tag{2.16}$$

which implies that, since $\eta_k \in (0,1)$ and $(V^k)^\top F(x^k) = 0$,

$$0 \leq \|V^k d^k\|^2 \leq -(1-\eta_k^2)\|F(x^k)\|^2 \leq 0.$$

So, $\|F(x^k)\| = 0$. Furthermore, the fact that $V^k \in \partial_B F(x^k)$ are nonsingular means that $d^k = 0$. ■

Proposition 2.10**.** *Suppose that there exists a $\overline{d^k}$ such that for all $V^k \in \partial_B F(x^k)$ we have $\|F(x^k) + V^k \overline{d^k}\| < \|F(x^k)\|$. Then there exists $\eta_{\min} \in [0,1)$ such that, for any $\eta_k \in [\eta_{\min}, 1)$, there is a d^k such that (2.8) holds.*

Proof. Clearly $F(x^k) \neq 0$ and hence $\overline{d^k} \neq 0$. Set

$$\eta_{\min} = \frac{\|F(x^k) + V^k \overline{d^k}\|}{\|F(x^k)\|}. \tag{2.17}$$

For any $\eta_k \in [\eta_{\min}, 1)$, let $d^k = \dfrac{1-\eta_k}{1-\eta_{\min}} \overline{d^k}$. Since the norm function is convex, we have that

$$\begin{aligned} \|F(x^k) + V^k d^k\| &\leq \frac{\eta_k - \eta_{\min}}{1-\eta_{\min}} \|F(x^k)\| + \frac{1-\eta_k}{1-\eta_{\min}} \|F(x^k) + V^k \overline{d^k}\| \\ &= \frac{\eta_k - \eta_{\min}}{1-\eta_{\min}} \|F(x^k)\| + \frac{1-\eta_k}{1-\eta_{\min}} \eta_{\min} \|F(x^k)\| \\ &= \eta_k \|F(x^k)\|. \end{aligned} \tag{2.18}$$

The proof is complete. ■

The following proposition shows the relation between the gradient $\nabla h(x^k) = (V^k)^\top F(x^k)$ of the objective function and the step d^k generated by the proposed algorithm. We can see from Proposition 2.11 that the generalized inexact Newton step d^k is a descent direction for $h(x)$ at the current approximation x^k.

If $V^k \in \partial_B F(x^k)$ is nonsingular,

$$\kappa_k = \operatorname{cond}(V^k) = \|(V^k)^{-1}\| \cdot \|V^k\| \tag{2.19}$$

represents the Euclidean condition number of the matrix $V^k \in \partial_B F(x^k)$.

Proposition 2.11**.** *Suppose that d^k satisfies (2.8) with all $V^k \in \partial_B F(x^k)$ being nonsingular. Then d^k is descent direction for $h(x)$ at x^k, i.e.,*

$$-\nabla h(x^k)^\top d^k \geq (1-\eta_k)\|F(x^k)\|^2 > 0, \tag{2.20}$$

$$\frac{|\nabla h(x^k)^\top d^k|}{\|d^k\|} \geq \frac{1-\eta_k}{(1+\eta_k)\kappa_k} \|\nabla h(x^k)\| \geq 0, \tag{2.21}$$

where κ_k is defined by (2.19), $\eta_k \in [0,1)$ is given in (2.8).

Proof. Let r^k be the residual associated with d^k such that $F(x^k) + V^k d^k = r^k$, where

$V^k \in \partial_B F(x^k)$. From

$$\nabla h(x^k)^\top d^k = F(x^k)^\top V^k d^k = F(x^k)^\top [r^k - F(x^k)], \tag{2.22}$$

and taking the norm in the right-hand side of (2.22), we have that

$$\nabla h(x^k)^\top d^k \leq \|F(x^k)\| \cdot \|r^k\| - \|F(x^k)\|^2 \leq -(1-\eta_k)\|F(x^k)\|^2. \tag{2.23}$$

Clearly, $(V^k)^\top F(x^k) \neq 0$ and hence $d^k \neq 0$. Next, $V^k d^k = r^k - F(x^k)$. Thus, $d^k = (V^k)^{-1}[r^k - F(x^k)]$, taking the norm, we have

$$\|d^k\| = \|(V^k)^{-1}\|(\|r^k\| + \|F(x^k)\|) \leq (1+\eta_k)\|(V^k)^{-1}\|\|F(x^k)\|. \tag{2.24}$$

Combining (2.23) and (2.24), we have that

$$\frac{|\nabla h(x^k)^\top d^k|}{\|d^k\|} \geq \frac{(1-\eta_k)\|F(x^k)\|^2}{(1+\eta_k)\|(V^k)^{-1}\|\|F(x^k)\|} = \frac{(1-\eta_k)\|F(x^k)\|}{(1+\eta_k)\|(V^k)^{-1}\|} \tag{2.25}$$

and hence as a result, using the fact that $\|\nabla h(x^k)\| \leq \|F(x^k)\|\|V^k\|$, we get

$$\frac{|\nabla h(x^k)^\top d^k|}{\|\nabla h(x^k)\|\|d^k\|} \geq \frac{(1-\eta_k)\|F(x^k)\|}{(1+\eta_k)\|F(x^k)\|(\|V^k\|\|(V^k)^{-1}\|)} = \frac{1-\eta_k}{(1+\eta_k)\kappa_k}. \tag{2.26}$$

So, the conclusions of the proposition are true. ■

Lemma 2.12. *Let d^k satisfy (2.8) with all $V^k \in \partial_B F(x^k)$ being nonsingular. Suppose that there exist $\eta_{\max} \in [0,1)$ and $\kappa \geq 0$ such that $\eta_k \leq \eta_{\max}$ in (2.8) and $\kappa_k \leq \kappa$ in (2.19). If $\nabla h(x^k) \neq 0$ and $\beta \in (0,1)$, then the proposed algorithm will produce an iterate $x^{k+1} = x^k + \lambda_k d^k$ satisfying (2.9) in a finite number of backtracking steps.*

Proof. By Lemma 2.6, we know $\nabla h(x)$ is continuous. Since $\|\nabla h(x^k)\| \neq 0$, by continuity there exist $\delta > 0$ and $\varepsilon > 0$ such that $\nabla h(x) \geq \varepsilon$ for all x with $\|x^k - x\| \leq \delta$. Using the mean value theorem, we have that with $0 \leq \upsilon_k \leq 1$, the following inequality holds:

$$\begin{aligned} h(x^k + \lambda_k d^k) &= h(x^k) + \beta\lambda_k \nabla h(x^k)^\top d^k + (1-\beta)\lambda_k \nabla h(x^k)^\top d^k \\ &\quad + \lambda_k[\nabla h(x^k + \upsilon_k \lambda_k d^k)^\top d^k - \nabla h(x^k)^\top d^k] \\ &= h(x^k) + \beta\lambda_k \nabla h(x^k)^\top d^k + \lambda_k[(1-\beta)\nabla h(x^k)^\top d^k + \zeta_k], \end{aligned} \tag{2.27}$$

where for convenience we have set $\zeta_k = [\nabla h(x^k + \upsilon_k \lambda_k d^k)^\top - \nabla h(x^k)^\top] d^k$. Since $\nabla h(x)$ is continuous, there exists sufficiently small λ_k such that

$$\|\nabla h(x^k + \upsilon_k \lambda_k d^k) - \nabla h(x_k)\| \leq (1-\beta)\frac{1-\eta_{\max}}{(1+\eta_{\max})\kappa}\varepsilon.$$

Note that from the assumption we have

$$|\zeta_k| = |[\nabla h(x^k + \upsilon_k \lambda_k d^k)^\top - \nabla h(x^k)^\top] d^k| \leq \frac{(1-\beta)(1-\eta_{\max})\varepsilon}{(1+\eta_{\max})\kappa}\|d^k\|.$$

Since (2.21) means $\nabla h(x^k)^\top d^k \leq -\dfrac{1-\eta_{\max}}{(1+\eta_{\max})\kappa}\varepsilon\|d^k\|$, we have that after a finite number of reductions, the last term in brackets in the right-hand side of (2.27) will become negative and the corresponding λ_k will be acceptable, that is, we have that in a finite number of backtracking steps, λ_k must satisfy

$$h(x^k + \lambda_k d^k) \leq h(x^k) + \beta \lambda_k \nabla h(x^k)^\top d^k.$$

Since $h(x^k) \leq h(x^{l(k)})$, the conclusion of the lemma holds. ■

Now we state a seminal global convergence result for smooth equations in [59].

Theorem 2.13. *Let $\{x^k\}$ be a sequence defined by*

$$x^{k+1} = x^k + \alpha_k d^k, \quad d^k \neq 0.$$

Let $a > 0$, $\gamma \in (0,1)$, $\omega \in (0,1)$ and let M be a nonnegative integer. Assume that

(i) *the level set $\Omega^0 = \{x : h(x) \leq h(x^0)\}$ is compact;*

(ii) *there exist positive numbers c_1, c_2 such that for all k,*

$$\nabla h(x^k)^\top d^k \leq -c_1 \|\nabla h(x^k)\|^2, \tag{2.28}$$

$$\|d^k\| \leq c_2 \|\nabla h(x^k)\|; \tag{2.29}$$

(iii) *$\alpha_k = \omega^{k_i} a$, where k_i is the first nonnegative integer k for which*

$$h(x^k + \lambda_k d^k) \leq h(x^{l(k)}) + \gamma \omega^k a \nabla h(x^k)^\top d^k \tag{2.30}$$

where $h(x^{l(k)}) = \max_{0\le j\le m(k)}\{h(x^{k-j})\}$, *with the nonmonotone index function* $m(0) = 0, 0 \le m(k) \le \min\{m(k-1)+1, M\}$, $k \ge 1$.

Then

(a) *the sequence* $\{x^k\}$ *remains in* Ω^0 *and every limit point* $\bar{x}$ *satisfies* $\nabla h(\bar{x}) = 0$;

(b) *no limit point of* $\{x^k\}$ *is a local maximum of* $h(x)$;

(c) *if the number of the stationary points of* $h(x)$ *in* Ω^0 *is finite, then the sequence* $\{x^k\}$ *converges.*

From Theorem 2.13, we can get the global convergence result of Algorithm 2.7.

Theorem 2.14. *Let* $\{x_k\} \subset \mathbf{R}^n$ *be a sequence generated by Algorithm 2.7. Let* d^k *satisfy (2.8) with all* $V^k \in \partial_B F(x^k)$ *being nonsingular. Suppose that there exist* $\eta_{\max} \in [0,1)$ *and* $\kappa \ge 0$ *such that* $\eta_k \le \eta_{\max}$ *in (2.8) and* $\kappa_k \le \kappa$ *in (2.19). Assume that* $\nabla h(x^k) \ne 0$ *and* $\beta \in (0,1)$. *Then*

(a) *the sequence* $\{x^k\}$ *remains in* $L(x^0)$ *and every limit point* x^* *satisfies* $h(x^*) = 0$,

(b) *if the number of the stationary points of* $h(x)$ *in* $L(x^0)$ *is finite, then the sequence* $\{x^k\}$ *converges.*

Proof. We will show the three conditions of the Theorem 2.13 hold. The level set $L(x^0)$ is compact by Assumption 2.8; thus the first condition holds. Since $L(x^0)$ is compact and $\{x^k\} \subset L(x^0)$, the assumption that $F(x)$ is locally Lipschitzian implies that it is uniformly Lipschitzian in $L(x^0)$ and, therefore, the $V^k \in \partial_B F(x^k)$ are uniformly bounded in norm. Since the condition numbers $\kappa_k = \mathrm{cond}(V^k)$ are assumed to be uniformly bounded in the Lemma 2.12, it follows that the inverses $(V^k)^{-1}$ are also uniformly bounded in norm. Then, as in (2.23), we have

$$\nabla h(x^k)^\top d^k \le -(1-\eta_k)\|F(x^k)\|^2 \le -(1-\eta_{\max})\left(\frac{C}{\kappa}\right)^2 \|\nabla h(x^k)\|^2, \tag{2.31}$$

where $\eta_k \le \eta_{\max}$, $\kappa_k \le \kappa$, and $\|(V^k)^{-1}\| \le C$ for all k. Moreover, as in (2.24),

$$\|d^k\| \le (1+\eta_k)\|(V^k)^{-1}\|\|F(x^k)\| \le (1+\eta_{\max})C^2\|\nabla h(x^k)\|. \tag{2.32}$$

With (2.31) and (2.32), the second condition of the Theorem 2.13 holds. Finally, the third condition holds just by (2.9) with the sequence of backtracking parameters $\lambda_k = 1, \omega, \omega^2, \ldots$. Thus, by Theorem 2.13, the conclusion of the theorem holds. ■

Theorem 2.15. *Under the assumptions of Theorem 2.14, let x^* be any limit point of the sequence $\{x^k\}$ generated by Algorithm 2.7 and a BD-regular point of F. If $\beta < \frac{1}{2}$ and $\eta_k \to 0$, then the whole sequence $\{x^k\}$ converges to x^* and the stepsize satisfies $\lambda_k = 1$ for large enough k.*

Proof. Note that

$$\begin{aligned} h(x^k + d^k) - h(x^k) - \nabla h(x^k)^\top d^k &= \frac{1}{2}\|F(x^k) + V^k d^k + o(\|d^k\|)\|^2 \\ &\quad - \frac{1}{2}\|F(x^k)\|^2 - \nabla h(x^k)^\top d^k \\ &= \frac{1}{2}\|V^k d^k\|^2 + o(\|d^k\|^2). \end{aligned}$$

This gives

$$\begin{aligned} h(x^k + d^k) \;\leq\; & h(x^{l(k)}) + \beta \nabla h(x^k)^\top d^k + \left(\frac{1}{2} - \beta\right) \nabla h(x^k)^\top d^k \\ & + \frac{1}{2}\left(\nabla h(x^k)^\top d^k + \|V^k d^k\|^2\right) + o(\|d^k\|^2). \end{aligned} \tag{2.33}$$

Next, by (2.23) and (2.24), we get

$$\nabla h(x^k)^\top d^k \leq -(1 - \eta_k)\|F(x^k)\|^2,$$

$$\|d^k\| \leq (1 + \eta_k)\|(V^k)^{-1}\|\|F(x^k)\|, \tag{2.34}$$

and

$$\|V^k d^k\| \leq \|r^k\| + \|F(x^k)\| \leq (1 + \eta_k)\|F(x^k)\|.$$

So, (2.33) can be rewritten as follows

$$h(x^k + d^k) \;\leq\; h(x^{l(k)}) + \beta \nabla h(x^k)^\top d^k + \left(\frac{1}{2} - \beta\right) \nabla h(x^k)^\top d_k$$

$$
\begin{aligned}
& + \frac{1}{2}\left(\nabla h(x^k)^\top d^k + \|V^k d^k\|^2\right) + o(\|d^k\|^2) \\
\leq\ & h(x^{l(k)}) + \beta \nabla h(x^k)^\top d^k \\
& - \left[\left(\frac{1}{2} - \beta\right)(1-\eta_k) + \frac{1}{2}(1-\eta_k) - \frac{1}{2}(1+\eta_k)^2\right] \|F(x^k)\|^2 + o(\|d^k\|^2) \\
\leq\ & h(x^{l(k)}) + \beta \nabla h(x^k)^\top d^k
\end{aligned} \tag{2.35}
$$

for all large enough k, the last inequality is deduced because the third term in brackets in the right-hand side of (2.35) will become negative by

$$
\left(\frac{1}{2} - \beta\right)(1-\eta_k) + \frac{1}{2}(1-\eta_k) - \frac{1}{2}(1+\eta_k)^2 \to \left(\frac{1}{2} - \beta\right), \text{ as } \eta_k \to 0,
$$

and by (2.34), $o(\|d^k\|^2) = o(\|F(x^k)\|^2)$ for the last term. By the above inequality and since λ_k given in (2.9) is bounded away from 1 as $\eta_k \to 0$, we know that the acceptance rule (2.9) means that, for large k,

$$
x^{k+1} = x^k + d^k, \tag{2.36}
$$

which implies that for large enough k, the stepsize $\lambda_k = 1$.

Since x^* is a limit point of $\{x^k\}$, Theorem 2.14 gives $F(x^*) = 0$ which means that x^* is a BD-regular zero solution of $F(x)$. By (2.36) and (2.38) of the following Theorem 2.16, the whole sequence $\{x^k\}$ converges to x^*. Hence the theorem is proved. ■

In what follows, we will analyze the local convergence of Algorithm 2.7 which can be taken as a supplement for those conclusions of the inexact Newton method for nonsmooth problems by Martinez and Qi [94], Facchinei, Fischer, and Kanzow [44, 45].

Theorem 2.16 (see [94, 44]). *Assume that $F(x)$ is semismooth in a neighborhood of x^* and that x^* is a BD-regular zero solution of F. There are numbers $\overline{\eta} > 0$ and $\varepsilon > 0$ such that, if $\|x^0 - x^*\| \leq \varepsilon$ and $\eta_k < \overline{\eta}$ for all k, then the sequence $\{x^k\}$ generated by*

$$
\|F(x^k) + V^k d^k\| \leq \eta_k \|F(x^k)\|, \quad x^{k+1} = x^k + d^k \tag{2.37}
$$

converges to x^*, and the convergence is linear in the sense that there exists $\tau \in (0,1)$

such that for all k,

$$\|x^{k+1}-x^*\| \le \tau\|x^k-x^*\|. \tag{2.38}$$

Moreover, $x^k \to x^*$ superlinearly if and only if

$$\|r^k\| = o(\|F(x^k)\|) \text{ as } k \to \infty, \tag{2.39}$$

where $r^k = F(x^k)+V^k d^k$. If $F(x)$ is p-order semismooth at x^*, then $x^k \to x^*$ with order at least $1+p$ if and only if

$$\|r^k\| = O(\|F(x^k)\|^{1+p}) \text{ as } k \to \infty. \tag{2.40}$$

Since $F(x)$ is locally Lipschitzian, by Lemma 1.2, there exist $\delta > 0$ and $\zeta > \zeta' > 0$ such that

$$\zeta'\|x^k-x^*\| \le \|F(x^k)\| \le \zeta\|x^k-x^*\|$$

for all $\|x^k-x^*\| \le \delta$. Therefore,

$$\begin{aligned} \frac{\zeta}{1-\rho_k} &= \frac{\zeta\|x^k-x^*\|}{\|x^k-x^*\|-\|x^{k+1}-x^*\|} \\ &\ge \frac{\|F(x^k)\|}{\|x^{k+1}-x^k\|} \\ &\ge \frac{\zeta'\|x^k-x^*\|}{\|x^k-x^*\|+\|x^{k+1}-x^*\|} = \frac{\zeta'}{1+\rho_k}, \end{aligned} \tag{2.41}$$

where $\rho_k = \dfrac{\|x^{k+1}-x^*\|}{\|x^k-x^*\|}$. Then an equivalent result in terms of the steps $\{d^k\}$ for (2.39) and (2.40) is expressed that $x^k \to x^*$ superlinearly if and only if

$$\|r^k\| = o(\|d^k\|) \text{ as } k \to \infty,$$

where $r^k = F(x^k)+V^k d^k$. Furthermore, if $F(x)$ is p-order semismooth at x^*, $x^k \to x^*$ with order at least $1+p$ if and only if

$$\|r^k\| = O(\|d^k\|^{1+p}) \text{ as } k \to \infty.$$

On the other hand, if the condition

$$\|F(x^k)+V^k d^k\| \le \eta_k \|F(x^k)\|$$

is replaced by the stronger condition with the forcing sequence $\{\eta_k\}$ of p-order for

$$\|F(x^k)+V^k d^k\| \le \eta_k \|F(x^k)\|^{1+p}, \quad \forall k = 0,1,\ldots, \tag{2.42}$$

that is, $\eta_k = \eta\|F(x^k)\|^p$, where η is any nonnegative constant, one can show $1+p$ order superlinear convergence of the iterative sequence $\{x^k\}$.

Theorem 2.17**.** *Suppose that F is BD-regular at x^*, which is a solution of (2.1). Assume also that in a neighborhood N of x^*, for any $y \in N$ and $V \in \partial_B F(y)$, the following inequality holds*

$$\|F(y)-F(x^*)-V(y-x^*)\| \le \gamma\|y-x^*\|^{1+p}, \tag{2.43}$$

where γ is called p-order semismooth constant at x^*. Then there exists an $\varepsilon \in (0,1]$ such that if $\|x^0-x^*\| \le \varepsilon$, $\beta = \max\{\|V^{-1}\| | V \in \partial_B F(x)\}$ for all x with $\|x-x^*\| \le \varepsilon$, then the sequence $\{x^k\}$ generated by (2.42) superlinearly converges to x^* with order at least $1+p$ in the sense that

$$\|x^{k+1}-x^*\| \le \beta(\gamma+\eta\zeta^{1+p})\|x^k-x^*\|^{1+p}, \quad \forall k = 0,1,\ldots, \tag{2.44}$$

where ζ is the locally Lipschitizan constant of $F(x)$ at x^*.

Proof. The proof of theorem is similar to that of Theorem 2.16. We will omit it here. ■

2.4 A Smoothing Variant

In this section, suppose that a smoothing function [27, 29] of problem (2.1) is available, i.e., a function $G(t,x) : [0,\infty) \times \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable for any $t > 0$ and $x \in \mathbb{R}^n$,

$$\lim_{t \downarrow 0, y \to x} G(t,y) = F(x). \tag{2.45}$$

A smoothing variant of the inexact Newton-Krylov methods for nonsmooth problems can be obtained via replacing $F(x)$ by $(t, G(t,x))^\top$ in Algorithm 2.7. However, it is not easy to get a smoothing function for a general problem (2.1) in practice, thus the main aim of this section is algorithmic. We focus on the theoretical study of the smoothing variant of the inexact Newton-Krylov methods for nonsmooth problems.

We treat the smoothing parameter t as a variable [112], denote

$$F(z) = F(t,x) = \begin{pmatrix} t \\ G(t,x) \end{pmatrix} = 0, \quad h(z) = \frac{1}{2}\|F(z)\|^2. \tag{2.46}$$

From (2.46), for any $t \neq 0$ a straightforward calculation yields

$$F'(z) = \begin{pmatrix} 1 & 0 \\ G'_t(t,x) & G'_x(t,x) \end{pmatrix}. \tag{2.47}$$

Choose $\bar{t} \in (0,\infty)$ and $\gamma \in (0,1)$ such that $\gamma\bar{t} < 1$. Let $\bar{z} = (\bar{t}, 0) \in (0,\infty) \times \mathbb{R}^n$. Define $\zeta(z) = \gamma \min\{1, h(t,x)\}$. Let $z^k = (t^k, x^k) \in (0,\infty) \times \mathbb{R}^n$ denote the iterate at iteration k. Then the algorithm framework of the smoothing variant of the inexact Newton-Krylov methods is stated below.

Algorithm 2.18**.** SMOOTHING VARIANT OF INEXACT NEWTON-KRYLOV METHODS

Let x^0 be given, $\beta \in (0,1)$, $\lambda_0 \in (0,1]$, $\eta_{\max} \in [0,1)$, $t^0 = \bar{t}$, and $0 < \theta_{\min} < \theta_{\max} < 1$ be given.

For $k = 0$ step 1 until convergence do:

Find some $\eta_k \in [0, \eta_{\max}]$ and a vector $d^k = (\zeta(z^k)\bar{t} - t^k, d_x^k)$ that satisfy

$$\|F(z^k) + F'(z^k)d^k - \zeta(z^k)\bar{z}\| \leq \eta_k \|F(x^k)\|, \tag{2.48}$$

and then choose $\theta \in [\theta_{\min}, \theta_{\max}]$, update $\lambda_k \leftarrow \theta\lambda_k$ until the following inequality is satisfied:

$$h(z^k + \lambda_k d^k) \leq h(z^{l(k)}) + \beta\lambda_k \nabla h(z^k)^\top d^k \tag{2.49}$$

where $h(z^{l(k)}) = \max\limits_{0 \leq j \leq m(k)} \{h(z^{k-j})\}$, with the nonmonotone index function $m(0) = 0, 0 \leq m(k) \leq \min\{m(k-1)+1, M\}$, $k \geq 1$.

Set $z^{k+1} = z^k + \lambda_k d^k$.

The following assertion shows that Algorithm 2.18 is well-defined.

Proposition 2.19. *Suppose z^k is not a solution of $h(z)$ in (2.46). Then*

(i) $\zeta(z^k)\bar{t} \le t^{k+1} \le t^k$,

(ii) d^k *is a descent direction of* $h(z)$ *at* z^k, *i.e.,* $\nabla h(z^k)^\top d^k < 0$.

Proof. Let the residual $r^k = F(z^k) + F'(z^k)d^k$. Then

$$\begin{aligned} \nabla h(z^k)^\top d^k &= F(z^k)^\top F'(z^k)d^k \\ &= F(z^k)^\top (r^k - F(z^k)) \\ &\le \|F(z^k)\|\|r^k\| - \|F(z^k)\|^2 \\ &\le -(1-\eta_k)\|F(z^k)\|^2 < 0, \end{aligned} \tag{2.50}$$

which implies (ii).

We show (i) by induction. For $k = 1$, the result is trivial. Suppose (i) holds for some $k \ge 1$, then

$$\begin{aligned} t^k = \max\left\{t^k, \zeta(z^k)\bar{t}\right\} \ge t^{k+1} &= t^k + \lambda_k d_t^k \\ &= t^k + \lambda_k(\zeta(z^k)\bar{t} - t^k) \\ &= (1-\lambda_k)t^k + \lambda_k\zeta(z^k)\bar{t} \\ &\ge \min\left\{t^k, \zeta(z^k)\bar{t}\right\} = \zeta(z^k)\bar{t} > 0. \end{aligned} \tag{2.51}$$

This completes the induction. ■

Convergence theories of smoothing variant 2.18 of the inexact Newton-Krylov methods can be established following an analysis analogous to the analysis for Algorithm 2.7 in Section 2.3 with slight and technical modifications. At the end of this section, we present some convergence conclusions for smoothing variant 2.18. These assertions can be proved exactly as in the proofs of Theorems 2.14 and 2.15, we therefore omit the detailed proofs here.

Theorem 2.20. *Let $\{z_k\} \subset \mathbb{R}^{n+1}$ be a sequence generated by Algorithm 2.18. Let d^k satisfy (2.48) with $F'(z^k)$ being nonsingular. Suppose that there exist $\eta_{\max} \in [0,1)$ and*

$\kappa \geq 0$ such that $\eta_k \leq \eta_{\max}$ in (2.48) and $\kappa_k = \|(F'(z^k))^{-1}\|\|F'(z^k)\| \leq \kappa$. Assume that $\nabla h(z^k) \neq 0$ and $\beta \in (0,1)$. Then

(a) *the sequence $\{z^k\}$ remains in $L(z^0)$ (see (2.15)) and every limit point z^* satisfies $h(z^*) = 0$,*

(b) *if the number of the stationary points of $h(z)$ in $L(z^0)$ is finite, then the sequence $\{z^k\}$ converges.*

Theorem 2.21**.** *Under the assumptions of Theorem 2.20, let $z^* = (t^*, x^*)$ be any limit point of the sequence $\{z^k\}$ generated by Algorithm 2.18 and a BD-regular point of F. If $\beta < \frac{1}{2}$ and $\eta_k \to 0$, then the entire sequence $\{z^k\}$ converges to z^* and the stepsize satisfies $\lambda_k = 1$ for large enough k.*

2.5 Numerical Experiments

In our numerical experiments, a set of problems was defined using classical smooth systems of the form $f(x) = 0$ with $f : \mathbb{R}^n \to \mathbb{R}^n$, $f = (f_1, \ldots, f_n)^\top$. Associated to each smooth system, we generated the following nonlinear complementarity problem:

$$x_i \geq 0, \quad f_i \geq 0, \quad x_i f_i(x) = 0, \quad \forall i = 1, 2, \ldots, n.$$

By the Fisher-Burmeister function [48] $\phi_{FB}(a,b) = \sqrt{a^2+b^2} - (a+b)$, $(a,b) \in \mathbb{R}^2$, solving nonlinear complementarity problem is equivalent to solving the semismooth system of nonlinear equations [72, 110] below :

$$F(x) = \begin{pmatrix} \sqrt{x_1^2 + f_1^2(x)} - (x_1 + f_1(x)) \\ \vdots \\ \sqrt{x_n^2 + f_n^2(x)} - (x_n + f_n(x)) \end{pmatrix} = 0.$$

Herein, we consider two classes of complementarity problems:

Example 2.22**.** (Linear complementarity)

$f(x) = Mx + q$, where $n = 800, 1600, 2400, 3200$,

$q = (-1,\ldots,-1)^\top$,
$M_{ii} = 4(i-1)+1, \quad i = 1,\ldots,n,$
$M_{ij} = M_{ji} = M_{ii}+1, \quad i = 1,\ldots,n-1, j = i+1,\ldots,n$

Example 2.23. (Nonlinear complementarity)
$f(x) = Mx+q(x)$, where $n = 4800, 5600,$
$q(x) = (q_1(x_1), q_2(x_2),\ldots,q_n(x_n))^\top$, $q_i(x) = 4\exp(x_i)$, $i = 1,2,\ldots,n,$
$M_{ii} = 4, \quad i = 1,\ldots,n,$
$M_{ij} = M_{ji} = -1.5, \quad i = 1,\ldots,n-1, j = i+1,\ldots,n.$

In this section, all numerical experiments are achieved in MATLAB 7.3. The Krylov subspace strategies: GMRES, GMRES(m), CGS and TFQMR [79, 80] are employed in Algorithm 2.7. The corresponding approaches derived from these Krylov subspace strategies are denoted by Newton-GMRES, Newton-GMRES(m), Newton-CGS and Newton-TFQMR, respectively. Here GMRES(m) employs restarted strategy [118, 58]. The forcing terms in inexact methods has been chosen as in [43, 79, 80], i.e, *given* $\gamma \in [0,1]$, $\alpha \in (1,2]$, *and* $\eta_0 \in [0,1)$, *choose*

$$\eta_k = \gamma \left(\frac{\|F(x^k)\|}{\|F(x^{k-1})\|} \right)^{\alpha}, \quad k = 1,2,\ldots.$$

In the actual computations, the initial value x^0 is taken randomly. The initial iterative values of Krylov subspace methods used in Algorithm 2.7 are all zero vectors. The approaches derived from different Krylov subspace strategies in Algorithm 2.7 terminate once the current iteration attains a prescribed stopping tolerance ε

$$\text{ERROR} = \frac{\|F(x)\|}{\|F(x^0)\|} \leq \varepsilon$$

or the admissible Newton largest iteration step counter reaches I_{max}. We take $\varepsilon = 10^{-6}$, the largest iteration steps number $I_{max} = 20$. Inner iterative steps number, namely, the maximal dimension of Krylov subspace $k_{max} = 100$.

With dimensions of nonlinear equations $n = 800, 1600, 2400, 3200, 4800, 5600$, the performance results of Newton-GMRES, Newton-GMRES(m), Newton-CGS and Newton-TFQMR methods are shown in Table 2.1. To compare the convergence speed of algo-

TABLE 2.1
Numerical results of algorithms

Example	Method	Dimension	Time (Second)	IT	ERROR
2.22	Newton-CGS	800	2.7853	15	2.3649e-008
		1600	11.951	16	2.1422e-007
		2400	25.339	16	1.1245e-008
		3200	56.959	18	6.2692e-008
	Newton-GMRES	800	2.2164	15	2.5261e-010
		1600	9.2131	16	7.7245e-010
		2400	19.201	16	4.7092e-008
		3200	44.961	18	4.6324e-009
	Newton-GMRES(20)	800	2.2107	15	3.5883e-009
		1600	9.1390	16	8.2396e-010
		2400	20.047	16	4.7092e-008
		3200	44.672	18	2.7380e-009
	Newton-TFQMR	800	2.2340	15	3.7041e-008
		1600	9.5967	17	2.1892e-009
		2400	19.543	16	4.8883e-009
		3200	45.577	18	4.7232e-007
2.23	Newton-CGS	4800	25.702	4	2.0013e-012
		5600	46.682	5	2.1610e-012
	Newton-GMRES	4800	12.494	4	4.2682e-012
		5600	16.733	4	2.1588e-012
	Newton-GMRES(20)	4800	12.492	4	7.8759e-013
		5600	16.978	4	3.1400e-013
	Newton-TFQMR	4800	53.180	8	8.8062e-014
		5600	36.264	4	3.9334e-013

rithms, we also draw curves of relative error-iterative step number $(\log_{10}(\text{ERROR}) - \text{IT})$ for two examples in the dimension $n = 3200, 5600$, respectively.

Table 2.1 gives the numerical results of Examples 2.22 and 2.23 by using Newton-Krylov subspace methods. From Table 2.1, we can see that Newton-Krylov subspace methods perform very well for solving Examples 2.22 and 2.23 with different dimensions, though the Example 2.22 is dense. There is some difference between different dimensions. It seems that Newton-Krylov methods have better numerical behavior for nonlinear complementarity problems, than for linear complementarity problems.

Fig. 2.1 and Fig. 2.2 give the curves of the relative residuals versus iterative numbers of Examples 2.22 and 2.23 when $n = 3200, 5600$, respectively. These figures, combined with Table 2.1, enhances the feasibility of Newton-Krylov methods for solving nonsmooth equations. Especially, it seems that Newton-GMRES and Newton-GMRES(20) has better numerical behaviors than Newton-CGS, Newton-TFQMR, for

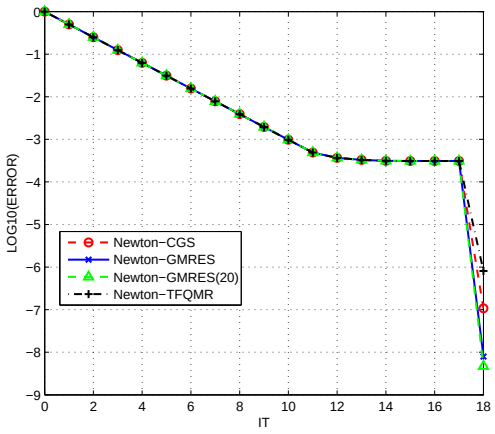

FIGURE 2.1 *Curves of the relative residuals versus iterative numbers of Example 2.22 when* $n = 3200$

Examples 2.22 and 2.23.

2.6 Contributions and Future Research

This chapter has introduced some variants of the inexact Newton method for solving systems of nonlinear equations defined by locally Lipschitzian functions. These methods combine inexact Newton iteration with Krylov subspace methods for solving the generalized Jacobian linear systems. Convergence theorems are proved under the condition of semismoothness. The preliminary numerical tests arising from the complementarity problems show that the proposed algorithms are feasible for solving large-scale nonlinear systems for which the functions are locally Lipschitz continuous.

It should be pointed out that our implementation in Section 2.5 is still in an early stage. Four directions in future research can be pursued to improve the current implementation:

(1) more global convergence strategies such as model trust region techniques;

(2) differences in numerical performances among the proposed inexact Newton-Krylov methods;

(3) how to precondition the proposed inexact Newton-Krylov methods for nonsmooth problems;

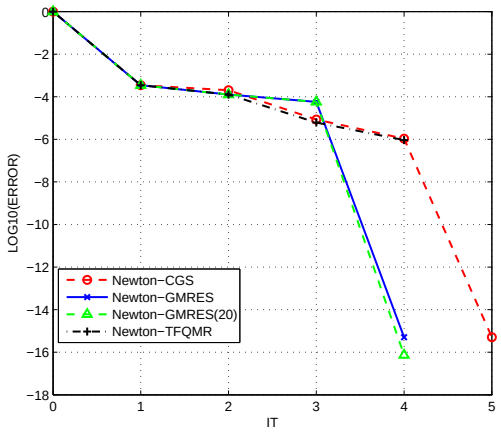

FIGURE 2.2 *Curves of the relative residuals versus computation time of Example 2.23 when* $n = 5600$

(4) how to choose more effective forcing terms of inexact Newton method for non-smooth problems [94].

We would like continue our testing of the proposed inexact Newton-Krylov methods on more practical problems. One interesting task is to test these methods on safe operations of electrical power systems, such as: the security region problems of power systems [126, 135].

Chapter 3

Pseudotransient Continuation

In this chapter[1] we are concerned with finding a solution to the following system of nonlinear equations with inequality constraints

$$\begin{cases} f(x) = 0 \\ g(x) \leq 0, \end{cases} \tag{3.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are at least semismooth functions. The feasible set of (3.1) is defined as

$$\Omega = \{x \in \mathbb{R}^n \mid g(x) \leq 0\}. \tag{3.2}$$

Here we suppose that the problem (3.1) is well-defined, i.e., $f(x)$ has at least a solution on Ω. When $\Omega = \{x \in \mathbb{R}^n \mid l \leq x \leq u\}$, $l_i \in \mathbb{R} \cup \{-\infty\}$ and $u_i \in \mathbb{R} \cup \{-\infty\}$, $l_i < u_i$, $i = 1, \ldots, n$, the problem (3.1) reduces to a bound-constrained system of nonlinear equations or unconstrained equations,

$$\begin{cases} f(x) = 0 \\ x \in \Omega. \end{cases} \tag{3.3}$$

Recently, concerning the solution of the reduced problem (3.3), Bellavia et al. [4, 5] presented a class of affine scaling trust-region interior point methods for this problem that combined ideas from the classical trust-region Newton method solving the unconstrained system of equations and the recent affine scaling approach [67, 133] for the solution of constrained optimization problems given by Coleman and Li in [33]. These

[1] This chapter is taken from [26].

affine scaling trust region approaches have been extended to a class of semismooth equations by Kanzow and Klug [75, 76]. Kanzow et al. [77] also proposed global projected Levenberg-Marquardt methods for the numerical solution of general nonsquare systems of bound-constrained nonlinear equations. The method proposed in [132] by Ulbrich is based on a Newton-like method with projection wrapped into a trust-region technique [111, 129]. It requires the minimization of quadratic problems on a box and achieves quadratic convergence. Sellami and Robinson [119, 120] developed the homotopy method [1], which is fairly robust and can find a solution of (3.3).

There are also some methods concerning the solution of the general problem (3.1). We recall extensions of the Newton method to systems of mixed equalities and inequalities [35], global quadratic algorithms based on backtracking line search [18, 52], and the recent trust-region methods [38, 50, 89], which are all based on suitable transformations of the problem (3.1) and vary widely from a computational point of view.

We note that all of the above methods are established via to solve the underlying optimization problem

$$\min \Theta(x) = \frac{1}{2}\|F(x)\|^2, \quad x \in \Omega. \tag{3.4}$$

To be more precise, these methods have been mainly concerned with finding a stationary point or a local minimizer of (3.4), which is not necessarily a solution of (3.3). This means that there is a "hump" that can block progress of the above approaches applied to the constrained equations. In other words, standard globalization strategies such as line search or trust region methods employed in these methods often stagnate at local minima. How to jump over this "hump" or avoid the standard globalization strategies to get a solution of (3.3) is crucial and very difficult in practice.

Other methods often used are to find a solution u^* of

$$F(u) = 0 \tag{3.5}$$

by ordinary differential equation (ODE) dynamics. More precisely, suppose $F : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous and

$$u^* = \lim_{t\to\infty} u(t) \tag{3.6}$$

where $u(t)$ is the solution of the initial value problem (IVP) [3, 122]

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -F(u), \qquad u(0) = u_0. \tag{3.7}$$

(3.6) is usually called as a *stability condition*. Solving (3.5) is equivalent to finding the steady-state solution u^* of the IVP (3.7). The pseudotransient continuation (Ψtc) approach was usually used for solving the dynamic system (3.7) in the recent literature [31, 51, 68, 82, 81]. As shown in [81], Ψtc could be taken as a predictor-corrector method for efficient integration of the time-dependent differential equation (3.7) to find a steady-state solution, which differs from the traditional continuation methods, pseudo-arclength continuation method, and homotopy methods [114].

The most common form of Ψtc for the case that F is continuously differentiable is the iteration

$$u^+ = u^c - (\delta_c^{-1} I + F'(u^c))^{-1} F(u^c), \tag{3.8}$$

where I is an identity matrix, u^c is the current iteration and u^+ is the new iteration. One common way to control δ is "Switched Evolution Relaxation" (SER) [97]

$$\delta_+ = \min\left(\delta_c \frac{\|F(u^c)\|}{\|F(u^+)\|},\ \delta_{\max}\right). \tag{3.9}$$

Using $\delta_{\max} = \infty$ is common. As indicated by (3.9), through searching important transients in the early iteration, and Ψtc grows near u^*, (3.8) becomes Newton's method.

Ψtc has succeeded in avoiding the standard globalization strategies such as line search or trust region methods to get a solution of (3.5), via taking advantage of the underlying structure of the problems [82]. Ψtc has also been successfully applied to problems in differential algebraic equation dynamics [31, 51], computational fluid dynamics [32, 97], plasma dynamics [85], hydrology [47] and optimal control [57, 63, 64, 65, 66]. In [81], Kelley et al. investigated the Ψtc method for a class of constrained problems in which projections onto the tangent space of the constraints are easy to compute. These problems included the dynamic formulation of bound-constrained optimization problems and inverse eigenvalue and singular value problems.

In this chapter we focus on how the results in [81] can be applied to a class (BD-regular) of constrained semismooth nonlinear equations of the form (3.1). The semis-

moothness of the problem (3.1) makes Ψtc differ from the methods in which smooth problems are handled, such as the ODE methods of Bogges [9], Keller [78], Klopfenstein [83], Incerti et al. [69], Kubíček [87], Smale [124], and the continuation methods mentioned by Allgower and Georg [1].

The chapter is organized as follows. In Section 3.1 and 3.2, by using slack variables, we transform (3.1) to a BD-regular nonlinear equation subject to simple bound constraints on the variables, a situation to which the theory in [81] applies. Some illustrative examples are presented in Section 3.3. Some final comments are made in the last section.

Some words about the notation used in this chapter. For a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}^n$, we denote the Jacobian matrix of f at $x \in \mathbb{R}^n$ by $f'(x)$, and the gradient of f by ∇f at $x \in \mathbb{R}^n$. Throughout our chapter, $\|\cdot\|$ denotes the Euclidean norm.

3.1 Reformulation

In this section, we will reformulate problem (3.1) by adding a slack variable $\gamma \in \mathbb{R}^m$ whose nonnegativity is imposed by means of simple bounds. The problem (3.1) then becomes a square problem where $u = (x, \gamma)$ and $F : \mathbb{R}^{m+n} \to \mathbb{R}^{m+n}$ is given by

$$F(u) = \begin{pmatrix} f(x) \\ g(x) + \gamma \end{pmatrix} = 0, \qquad \gamma \geq 0. \tag{3.10}$$

It follows from the semismoothness of $f(x)$ and $g(x)$ that $F(u)$ is also semismooth, and a generalized derivative $V \in \partial_B F(u)$ has the form below

$$V = \begin{pmatrix} W_{n\times n} & 0 \\ U_{m\times n} & I \end{pmatrix}, \tag{3.11}$$

where $\begin{pmatrix} W_{n\times n} \\ U_{m\times n} \end{pmatrix} \in \partial_B \begin{pmatrix} f(x) \\ g(x) \end{pmatrix}$.

On the other hand, we may take $g(x) + \gamma^2 = 0$, $\gamma^2 = (\gamma_1^2, \ldots, \gamma_m^2)^\top$ and transfer (3.1)

as an unconstrained problem

$$F(u) = \begin{pmatrix} f(x) \\ g(x) + \gamma^2 \end{pmatrix} = 0. \tag{3.12}$$

Then a generalized derivative $V \in \partial_B F(u)$ has the form

$$V = \begin{pmatrix} W_{n\times n} & 0 \\ U_{m\times n} & 2\mathrm{diag}(\gamma_1, \ldots, \gamma_m) \end{pmatrix}. \tag{3.13}$$

We note that the nonsingularity of the generalized derivative of (3.12) depends on $W_{n\times n}$ and vector γ, and the nonsingularity of (3.10) only relies on $W_{n\times n}$. In other words, the nonsingularity of the generalized derivative of (3.10) is easier to control. Therefore, to get a solution of (3.1), one would like to design some algorithms by solving (3.10) rather than (3.12). Based on the above, we only consider the problem (3.10).

3.2 Pseudotransient Continuation

Let u^* denote a solution of $F(u) = 0$, and

$$e = u - u^*$$

denote the error. We will consider a Ψtc iteration of the form

$$u^+ = \mathscr{P}(u^c - (\delta_c^{-1} I + V(u^c))^{-1} F(u^c)), \tag{3.14}$$

where $\mathscr{P}$ is the projection in the 2-norm onto $\Omega = \{u \mid \gamma \geq 0\}$, i.e., $\mathscr{P}(u) = (x,\ \max\{0, \gamma\})$ with max meant componentwise.

The theory we will develop applies equally well to the inexact formulation

$$u^+ = \mathscr{P}(u^c + s), \tag{3.15}$$

where

$$\|(\delta_c^{-1} I + V(u^c))s + F(u^c))\| \leq \eta_c \|F(u^c)\|. \tag{3.16}$$

Define a neighborhood of the trajectory from u_0 as

$$S(\varepsilon) = \bigcup_{t \geq 0} B(u(t), \varepsilon), \tag{3.17}$$

where $B(u, \varepsilon)$ is the open ball of radius ε centered at u.

Kelley et al. [81] gave a convergence conclusion of the projected pseudotransient continuation method under the following Assumptions 3.1 and 3.2.

Assumption 3.1**.** $\mathscr{P}$ is called a Lipschitz projection onto Ω if

(i) $\mathscr{P}(u) = u$ for all $u \in \Omega$.

(ii) There are $M_{\mathscr{P}}$, $\varepsilon_{\mathscr{P}}$ such that for all $u \in \Omega$ and v such that $\|u - v\| \leq \varepsilon_{\mathscr{P}}$

$$\|\mathscr{P}(v) - u\| \leq \|v - u\| + M_{\mathscr{P}}\|v - u\|^2. \tag{3.18}$$

Assumption 3.2**.** Assume that H is a sufficiently good approximation to the generalized derivative of F and satisfies the conditions below:

(i) There are M_H, ε_H such that

$$\|H(u)\| \leq M_H, \ \forall u \in S(\varepsilon_H). \tag{3.19}$$

For all $\varepsilon > 0$ there is $\overline{\varepsilon}$ such that if $u \in S(\varepsilon_H)$ and $\|u - u^*\| > \varepsilon$ then

$$\|F(u)\| > \overline{\varepsilon}. \tag{3.20}$$

(ii) There are ε_L so that if $\|u^c - u^*\| \leq \varepsilon_L$, then $H(u^c)$ is nonsingular,

$$\|(I + \delta H(u^c))^{-1}\| \leq (1 + \beta\delta)^{-1}, \textit{ for some } \beta > 0 \textit{ and all } \delta > 0, \tag{3.21}$$

and the Newton iteration

$$u_N^+ = u^c - H(u^c)^{-1}F(u^c) \tag{3.22}$$

reduces the error by a (small) factor $r \in [0,1)$ for all $u^c \in \{u \mid \|u-u^*\| \le \varepsilon_L\}$, i.e.,

$$\|e_N^+\| \le r\|e^c\|. \tag{3.23}$$

Theorem 3.3 (see [81]). *Let F be locally Lipschitz continuous, and assume that the stability condition (3.6) and Assumption 3.2 hold. Let the sequence $\{\delta_k\}$ be updated via (3.9) and let $\delta^* > 0$ such that*

$$\frac{M_{\mathscr{P}}\varepsilon_L}{\beta} < \delta^* \le \delta_n \tag{3.24}$$

for all k. Assume that the Q-factor r in (3.23) satisfies

$$r < \frac{(1+M_{\mathscr{P}}\varepsilon_L)^{-1} - (1+\beta\delta^*)^{-1}}{2}, \tag{3.25}$$

where β is the constant in (3.21). Then if δ_0 and the sequence $\{\eta_k\}$ are sufficiently small, the inexact Ψtc iteration

$$u^{k+1} = \mathscr{P}(u^k + s^k),$$

where

$$\|(\delta_k^{-1}I + H(u^k))s^k + F(u^k)\| \le \eta_k\|F(u^k)\|$$

converges to u^*. Moreover, there is $K > 0$ such that for n sufficiently large

$$\|e^{k+1}\| \le \|e_N^{k+1}\| + K\|e^k\|(\eta_k + \delta_k^{-1}), \tag{3.26}$$

where

$$\|e_N^{k+1}\| = \|(u^k - H(u^k)^{-1}F(u^k)) - u^*\|.$$

In this chapter, $\mathscr{P}$ is the projection in the 2-norm onto $\Omega = \{u \mid \gamma \ge 0\}$. Thus, by Assumption 3.1, $\mathscr{P}$ is a Lipschitz projection. Based on the semismoothness assumption on F and Proposition 1.2, we can rewrite Assumption 3.2 as follows:

Assumption 3.4. Let F be semismooth with the conditions below:

(i) There are M_0, ε_0 such that $V(u)$ is nonsingular for all $u \in \{u \mid \|u-u^*\| \le \varepsilon_0\}$ (i.e.,

$F(u)$ is BD-regular at u^*), and

$$\|(\delta^{-1}I + V(u))^{-1}\| \le M_0, \ \forall \delta > 0. \tag{3.27}$$

(ii) There are M_1, ε_1 such that

$$\|V(u)\| \le M_1, \forall u \in S(\varepsilon_1). \tag{3.28}$$

Remark 3.5**.** Note that

$$\det(V(u^*)) = \det \begin{pmatrix} W^*_{n\times n} & 0 \\ U^*_{m\times n} & I \end{pmatrix} = \det(W^*_{n\times n}), \tag{3.29}$$

where $\begin{pmatrix} W^*_{n\times n} \\ U^*_{m\times n} \end{pmatrix} \in \partial_B \begin{pmatrix} f(x^*) \\ g(x^*) \end{pmatrix}$. Then $F(u)$ is BD-regular at $u^* = (x^*, \gamma^*)$ if and only if $W^*_{n\times n}$ is nonsingular.

Now we can give the convergence of the Ψtc approach for the nonsmooth system of nonlinear equations with inequality constraints (3.1).

Theorem 3.6**.** *Assume that the stability condition* (3.6) *and Assumption* 3.4 *hold. Let the sequence* $\{\delta_k\}$ *be updated via* (3.9) *and satisfy* $\delta_k \ge \delta^* > 0$*. Then if* δ_0 *and the sequence* $\{\eta_k\}$ *are sufficiently small, the inexact* Ψtc *iteration*

$$u^{k+1} = \mathscr{P}(u^k + s^k),$$

where

$$\|(\delta_k^{-1}I + V(u^k))s^k + F(u^k)\| \le \eta_k \|F(u^k)\| \tag{3.30}$$

converges to u^**. Moreover, for k sufficiently large*

$$\|e^{k+1}\| \le M_0(\delta_k^{-1} + L\eta_k)\|e^k\| + o(\|e^k\|), \tag{3.31}$$

or if F is semismooth of order p,

$$\|e^{k+1}\| \le M_0(\delta_k^{-1} + L\eta_k)\|e^k\| + O(\|e^k\|^{1+p}), \tag{3.32}$$

where L is the Lipschiz constant of F at u^*.

Proof. The outline of the proof follows those in [31, 51, 82, 81]. It is a direct application of the convergence results [81] (i.e., Theorem 3.3), we do not repeat it here. For the *local phase*, from the (i) of Assumption 3.4, we have

$$\begin{aligned} e^{k+1} &= u^{k+1} - u^* \\ &= \mathscr{P}(u^k + s^k) - u^* \\ &= \mathscr{P}\Big(u^k - (\delta_k^{-1} I + V(u^k))^{-1} F(u^k) \\ &\quad + (\delta_k^{-1} I + V(u^k))^{-1}[(\delta_k^{-1} I + V(u^k))s^k + F(u^k)]\Big) - u^*. \end{aligned}$$

Thus

$$\begin{aligned} \|e^{k+1}\| &= \Big\|\mathscr{P}\Big(u^k - (\delta_k^{-1} I + V(u^k))^{-1} F(u^k) \\ &\quad + (\delta_k^{-1} I + V(u^k))^{-1}[(\delta_k^{-1} I + V(u^k))s^k + F(u^k)]\Big) - u^*\Big\| \\ &\leq \Big\|(\delta_k^{-1} I + V(u^k))^{-1}((\delta_k^{-1} I + V(u^k))e^k - F(u^k) \\ &\quad + (\delta_k^{-1} I + V(u^k))s^k + F(u^k)\Big\| \\ &= M_0\Big[\|F(u^k) - F(u^*) - (\delta_k^{-1} I + V(u^k))e^k\| \\ &\quad + \|(\delta_k^{-1} I + V(u^k))s^k + F(u^k))\|\Big] \\ &\leq M_0(\delta_k^{-1}\|e^k\| + \|V(u^k)e^k - F(u^k)\| + \eta_k\|F(u^k)\|) \\ &\leq M_0(\delta_k^{-1} + L\eta_k)\|e^k\| + o(\|e^k\|), \end{aligned}$$

where L is the Lipschiz constant of F at u^*. The case that F is p-order semismooth at u^* is similar. Hence error bounds (3.31), (3.32) hold, i.e., the Ψtc iteration converges at least locally Q-linearly. ■

Remark 3.7**.** The Newton-like iterations (3.22) and (3.23) in Assumption 3.2 of Theorem 3.3 are not expressed explicitly in Theorem 3.6, which are involved in the high-order error terms of (3.31) and (3.32) in Theorem 3.6.

3.3 Numerical Tests

In this section, we discuss the feasibility of the Ψtc approach for solving nonlinear equations with inequality constraints.

3.3.1 Application I

We consider the Karush-Kuhn-Tucker (KKT for short) system of the following nonlinear programming (denoted by NP) problem:

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0 \end{aligned} \tag{3.33}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable.

Suppose a suitable constraint qualification holds (for example the Mangasarian-Fromovitz or the Slater constraint qualification [46]) for the constraints of NP problem (3.33). Then we can reformulate the KKT system for the NP problem (3.33) as follows:

$$\begin{cases} \nabla f(x) + \sum_{i=1}^{m} \mu_i \nabla g_i(x) = 0 \\ \mu_i \geq 0, \;\; g_i(x) \leq 0, \;\; \mu_i g_i(x) = 0 \;\; i = 1, \ldots, m. \end{cases} \tag{3.34}$$

Let

$$L(x,\mu) = \begin{pmatrix} \nabla f(x) + \mu \nabla g(x) \\ \mu_1 g_1(x) \\ \vdots \\ \mu_m g_m(x) \end{pmatrix}_{(m+n)\times(m+n)}, \quad G(x,\mu) = \begin{pmatrix} g(x) \\ -\mu \end{pmatrix}_{(m+n)\times 2m},$$

where $\mu = (\mu_1, \ldots, \mu_m)^\top$. Then (3.34) is equivalent to

$$\begin{cases} L(x,\mu) & = 0 \\ G(x,\mu) & \leq 0, \end{cases} \tag{3.35}$$

which is the exact expression of (3.1). Hence, we can use the Ψtc we mention to solve

this class of NP problems.

Example 3.8**.** Consider the KKT system of the following NP problem:

$$\begin{aligned} \min \quad & f(x) = \frac{1}{2}x^\top Ax - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_{-1}^{x_i}\exp(\max\{0,v\})\,\mathrm{d}v \\ \text{s.t.} \quad & g(x) = \sum_{i=1}^{n} x_i^2 \le 1, \quad n = 100, 1000, \end{aligned}$$

where

$$A = \begin{pmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}_{n\times n}.$$

We will test the KKT system of the above example by Ψtc with (3.30). Especially, we determine a solution of (3.30) by using a direct method (such as: Gaussian elimination method) [58, 79, 80]. So the forcing term in (3.30) $\eta_c = 0$. We also test this problem by STRSCNE solver for constrained system of nonlinear equations in [4, 5], which combines Newton method and an elliptical trust-region procedure.

In the implementation of algorithms, we set the initial iterative value $x^0 = (1,\ldots,1)^\top$, $\delta_0 = 0.49$ for $n = 100$, $\delta_0 = 2.35$ for $n = 1000$. The algorithm terminates once the current iteration attains a prescribed stopping tolerance ε (i.e., $\|F(x)\| \le \varepsilon$) or the admissible largest iteration step counter reaches I_{max}. We take $\varepsilon = 10^{-6}$, the largest iteration steps number $I_{max} = 100$.

In Figure 3.1 we plot the residuals versus iterative step numbers of Ψtc and STRSCNE solver for Example 3.8, respectively. Figure 3.1 shows that both Ψtc and STRSCNE solver can solve this problem well. Also it seems that Ψtc has better numerical behaviors than STRSCNE solver for this problem.

3.3.2 Application II

Consider the implicit complementarity problems with the following form:

Find $y \in \mathbb{R}^n$ such that

$$y - m(y) \ge 0, \quad F(y) \ge 0, \quad F(y)^\top (y - m(y)) = 0, \tag{3.36}$$

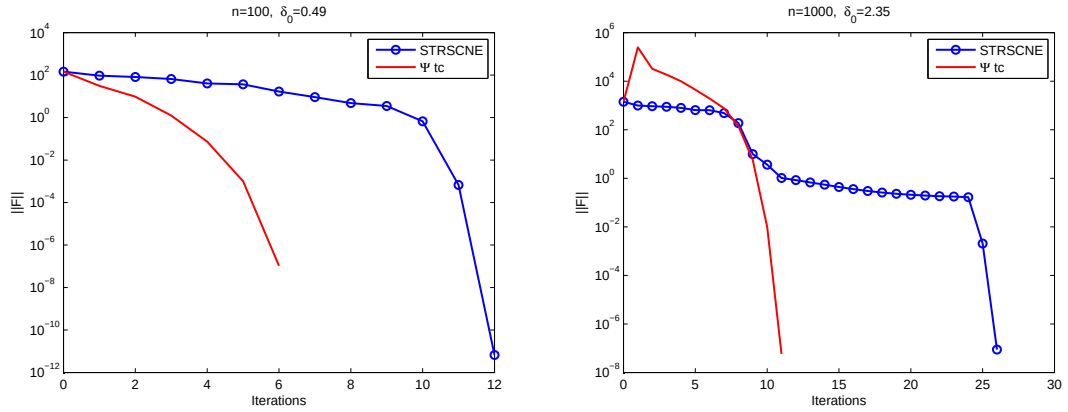

FIGURE 3.1 *Curves of the residuals versus iterative step numbers of Example 3.8.*

where $F(y)$ and $m(y) : \mathbb{R}^n \to \mathbb{R}^n$ are twice continuously differentiable.

It is clear to see that problem (3.36) can be rewritten as (3.1).

Example 3.9**.** Consider an implicit complementarity problem [2, 71, 99] under the conditions:

$$F(y) = Ay + b = \begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}_{n\times n} y + \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

$m(y) = \varphi(Ay+b)$ with $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ being twice continuously differentiable. Function φ is defined by the choices below:

(i) $\varphi_i(x) = -\dfrac{1}{2} - x_i, \quad i = 1,2,\ldots,n,$

(ii) $\varphi_i(x) = -\dfrac{1}{2} - \dfrac{3}{2}x_i + \dfrac{1}{4}x_i^2, \quad i = 1,2,\ldots,n,$

where $n = 4, 40, 400$.

For each choice of φ, three starting vectors were used, namely,

(a) $(0,0,\ldots,0)_{n\times 1}$,

(b) $(-0.5,-0.5,\ldots,-0.5)_{n\times 1}$,

(c) $(-1,-1,\ldots,-1)_{n\times 1}$.

In [99], the first approach adopted to Example 3.9 is an iterative scheme to compute fixed points of an operator S:

$$y^{k+1} = y^k - (E - V^k)^{-1}(y^k - S(y^k)), \tag{3.37}$$

where $V^k \in \partial S(y^k)$. The second approach is a Newton variant applied to the semismooth operator

$$H(y) = \min\{y - m(y), F(y)\} = 0. \tag{3.38}$$

Jiang et al., [71] proposed a trust-region approach to solve Example 3.9. In [2], a software BOX-QUACAN was used to solve the problem 3.9, this software was also based on trust-region strategy.

In the implementation of our Ψtc for Example 3.9, we use the similar termination parameters of Example 3.8. Tables 3.1 and 3.2 show the comparison of our data and the results of [2, 71, 99] when $n = 4, 40, 400$. "—" represents that the approach is not convergent. The results in Table 3.1 are quite promising. The numbers of iterations and function evaluations of Ψtc seem comparable to the results in [2, 71, 99], and less than STRSCNE solver in [4, 5].

In Figure 3.2 we plot the residuals versus iterative step numbers for Example 3.9 when $n = 400$. Figure 3.2 also shows us the feasibility of pseudotransient continuation approach for solving implicit complementarity problem, which can be taken as a system of nonlinear equations with inequality constraints.

3.4 Verification of the Assumptions

We will now explore verification of Assumption 3.4 for Examples 3.8 and 3.9.

TABLE 3.1
The numerical comparison: the number of iterations (IT), the number of function evaluations (FE) and the residual norms.

Choice	n	Initial Vector	Ψtc				STRSCNE in [4, 5]			Results in [2]	
			δ_0	IT	FE	Residual	IT	FE	Residual	IT	FE
(i)	4	(a)	5	6	7	5.9850e-009	14	15	1.84107e-009	4	5
		(b)	5	5	6	1.2124e-009	5	6	2.12138e-007	4	5
		(c)	0.53	14	15	8.5834e-008			—	4	5
	40	(a)	5	6	7	4.5193e-009	14	15	1.73937e-006	7	10
		(b)	5	5	6	7.1720e-010	16	17	3.76983e-006	6	8
		(c)	0.23	23	24	6.3782e-011			—	5	7
	400	(a)	5	6	7	4.2351e-009	21	22	1.14767e-007	8	12
		(b)	5	5	6	6.6979e-010	13	14	4.38376e-007	7	10
		(c)	0.07	38	39	7.1056e-011			—	6	8
(ii)	4	(a)	5	5	6	2.0655e-007	9	10	1.64424e-008	5	6
		(b)	5	4	5	7.3408e-008	4	5	2.79303e-010	6	8
		(c)	0.60	14	15	3.6181e-007			—	3	4
	40	(a)	5	5	6	1.3856e-007	7	8	1.41530e-006	7	11
		(b)	5	4	5	3.9909e-008	13	14	1.09028e-006	6	9
		(c)	0.37	16	17	1.2930e-011			—	6	8
	400	(a)	5	5	6	1.2657e-007	7	8	9.55374e-006	9	14
		(b)	5	4	5	3.4260e-008	9	10	1.46470e-007	7	11
		(c)	0.14	23	24	1.2202e-008			—	7	10

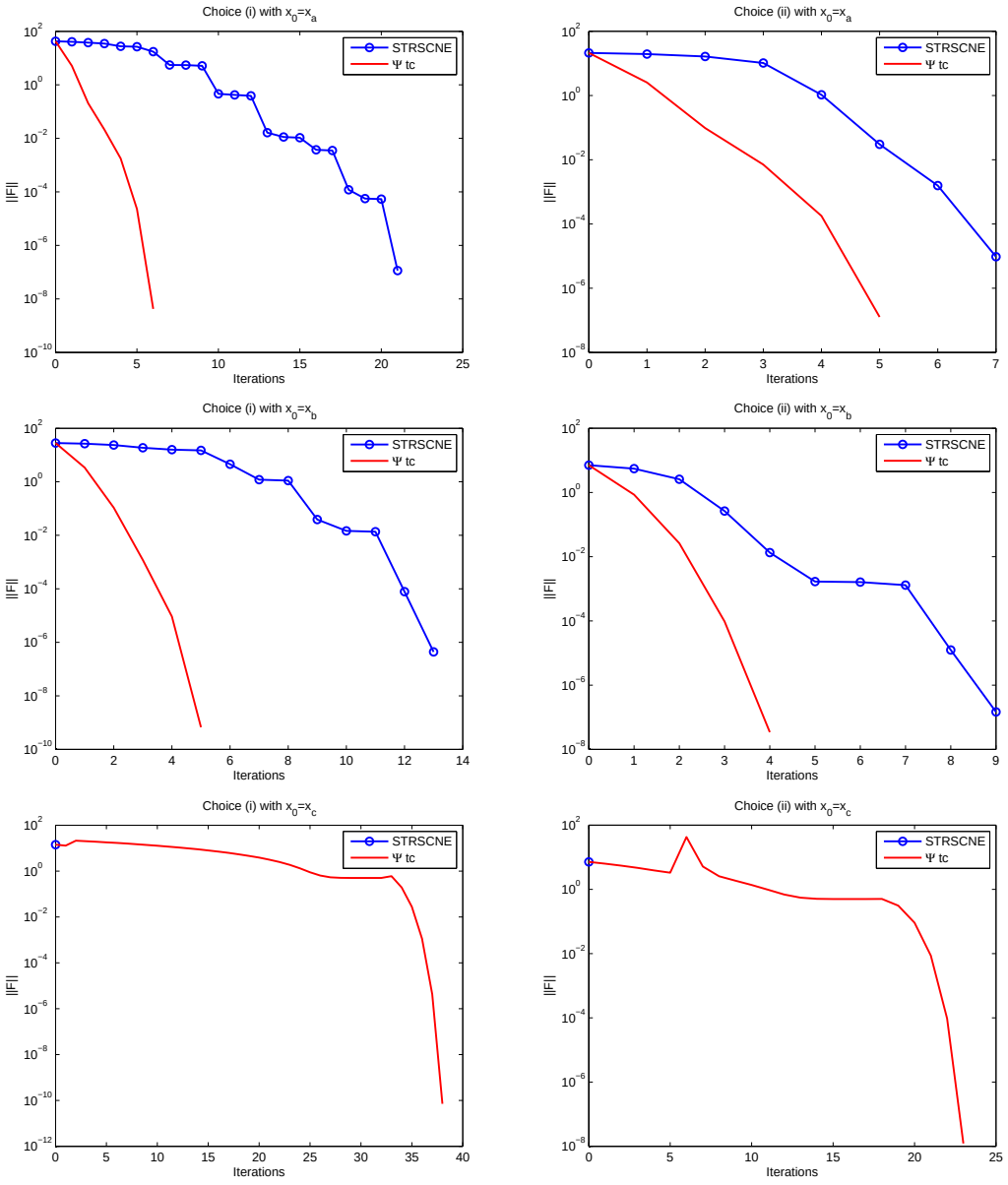

FIGURE 3.2 *Curves of the residuals versus iterative step numbers of Example 3.9.*

TABLE 3.2
The known results: the number of iterations (IT), the number of function evaluations (FE).

Choice	n	Initial Vector	(3.37) in [99]	(3.38) in [99]	Results in [71]	
			IT	IT	IT	FE
(i)	4	(a)	2	14	5	17
		(b)	2	41	4	15
		(c)	—	—	5	11
(ii)	4	(a)	3	15	5	17
		(b)	—	15	4	15
		(c)	—	56	5	11

3.4.1 Case of Example 3.8

We can rewrite the KKT system of the Example 3.8 as a constrained system of nonlinear equations,

$$
\begin{aligned}
F(u) &= \begin{pmatrix} f(x,\mu) \\ g(x,\mu)+s \end{pmatrix} \\
&= \begin{pmatrix} Ax - \dfrac{1}{\sqrt{n}}\exp(\max\{0,x\}) + 2\mu(x_1,\ldots,x_n)^\top \\ \mu(\sum_{i=1}^{n} x_i^2 - 1) \\ \sum_{i=1}^{n} x_i^2 - 1 + s_1 \\ -\mu + s_2 \end{pmatrix} = 0, \quad s = (s_1, s_2) \geq 0.
\end{aligned}
$$

Here function of vectors, $\exp(\max\{0,x\})$ for example, are understood to mean componentwise evaluation in the discussion. From the known result for scalar function $\max\{0,x\}$,

$$
\partial_B \max\{0,x\} = \begin{cases} \{0\}, & x < 0, \\ \{0,1\}, & x = 0, \\ \{1\}, & x > 0, \end{cases} \tag{3.39}
$$

we know

$$\partial_B F(u) = \begin{pmatrix} \mathscr{A}_{n\times n} & \beta_{n\times 1} & 0 & 0 \\ \mu\beta_{n\times 1}^\top & \alpha & 0 & 0 \\ \beta_{n\times 1}^\top & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}_{(n+3)\times(n+3)}, \tag{3.40}$$

where

$$\mathscr{A}_{n\times n} = A - \frac{1}{\sqrt{n}}\exp(\max\{0,x\})W_{n\times n} + 2\mu\,\mathrm{diag}(1,\ldots,1)^\top,$$

$$W_{n\times n} \in \partial_B \max\{0,x\},$$

$$\beta_{n\times 1} = 2(x_1,\ldots,x_n)^\top,$$

$$\alpha = \sum_{i=1}^{n} x_i^2 - 1.$$

Since A is positive definite, the determinant $\det(W_{n\times n}) \le 1$ for all $W_{n\times n} \in \partial_B \max\{0,x\}$, and $\mu \ge 0$, it follows that the minimum eigenvalue of $\mathscr{A}$ denoted by $\lambda_{\min}(\mathscr{A}) > 0$.

F is clearly semismooth and we use $V(u^k) \in \partial_B F(u^k)$. Then by taking determinant of $V(u^k)$, we have

$$\det(V(u^k)) = \det(\mathscr{A}_{n\times n}(u^k))\det((\alpha - \mu\beta^\top \mathscr{A}_{n\times n}^{-1}\beta)(u^k)). \tag{3.41}$$

It follows from the compactness of $S(\varepsilon_1)$ that (ii) in Assumption 3.4 holds for Example 3.8. From the positive definiteness of $\mathscr{A}_{n\times n}(u^k)$, $\alpha \le 0$ and the strict complementarity at u^* (deduced from the computational results), we know the validity of the first part of (i) in Assumption 3.4 (i.e., F is BD-regular at u^*).

From BD-regularity of F at u^* and the known matrix perturbation lemma in [98, p. 45], for all sufficiently large $\delta_k > 0$, we have

$$\|(\delta_k^{-1}I + V(u^k))^{-1}\| \le \frac{\|(V(u^k))^{-1}\|}{1 - \delta_k^{-1}\|(V(u^k))^{-1}\|} \tag{3.42}$$

which implies the (3.27) of Assumption 3.4.

3.4.2 Case of Example 3.9

The implicit complementarity problem (3.36) is equivalent to a constrained system of nonlinear equations,

$$G(u) = \begin{pmatrix} F_1(y)(y_1 - m_1(y)) \\ \vdots \\ F_n(y)(y_n - m_n(y)) \\ -F(y) + s_1 \\ m(y) - y + s_2 \end{pmatrix} = 0, \quad s = (s_1, s_2) \geq 0. \tag{3.43}$$

Since $F(y)$ and $m(y)$ are twice continuously differentiable, $G(u)$ is clearly smooth and its Jacobian matrix has the following form

$$JG(u) = \begin{pmatrix} \mathscr{A}_{n\times n} & 0 & 0 \\ -F'(y) & I_{n\times n} & 0 \\ m'(y) - I & 0 & I_{n\times n} \end{pmatrix}_{3n\times 3n}, \tag{3.44}$$

where

$$\mathscr{A}_{n\times n} = \begin{pmatrix} 2\alpha_1 + 3\beta_1 & -\alpha_1 - \beta_1 & & \\ -\alpha_2 - \beta_2 & 2\alpha_2 + 3\beta_2 & \ddots & \\ & \ddots & \ddots & -\alpha_{n-1} - \beta_{n-1} \\ & & -\alpha_n - \beta_n & 2\alpha_n + 3\beta_n \end{pmatrix}_{n\times n},$$

$$(\alpha_1, \ldots, \alpha_n)^\top = Ay + b \geq 0,$$

$$(\beta_1, \ldots, \beta_n)^\top = (A + I)y + b + \operatorname{diag}\left(\frac{1}{2}, \ldots, \frac{1}{2}\right) \geq 0.$$

Here for simplify, we only consider the choice (i) of $m(y)$. The other choice for $m(y)$ can be discussed similarly.

From the computational results in [2, 71, 99] and this chapter, it follows that there exists a neighborhood of u^* such that $2\alpha_i + 3\beta_i > 0$ and $\alpha_i + \beta_i > 0$ for $i = 1, \ldots, n$. By the known results about tridiagonal matrix [98, p. 51], we claim that $\mathscr{A}_{n\times n}$ is irreducibly

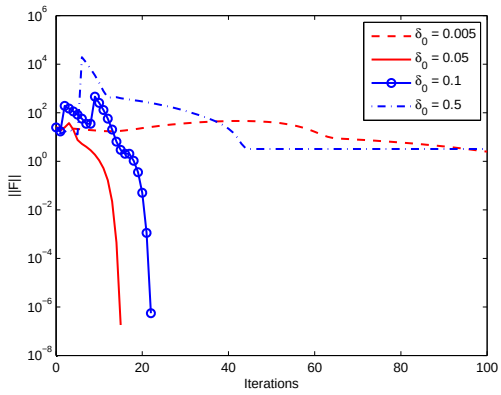

FIGURE 3.3 *Curves of the residuals versus iterative step numbers of problem* (3.45) *with different* δ_0.

diagonally dominant and similar to a positive definite matrix, which implies the validity of (i) of Assumption 3.4. Meanwhile, the compactness of $S(\varepsilon_1)$ guarantees that (ii) in Assumption 3.4 holds for Example 3.9.

3.4.3 Choices of δ_0

In Subsections 3.3.1 and 3.3.2, we tested the numerical examples without the choices of δ_0 involved. We remark in this subsection that δ_0 should be chosen optimally according to the choices of initial iterative values of Ψtc for these problems. To see this, let us consider a simple numerical example [17]

$$\begin{cases} -13+x_1-x_2^3+5x_2^2-2x_2=0 \\ -29+x_1-x_2^3+x_2^2-14x_2=0, \end{cases} \tag{3.45}$$

where $x=(x_1,x_2)\in\Omega$ with $\Omega=\{x=(x_1,x_2)\mid 2.5\le x_1\le 10,\ 2\le x_2\le 8\}$.

Taking the similar termination parameters of Ψtc for Example 3.8 and $x^0=(0.5,-2)$ as an initial iterative value, we test Ψtc for problem (3.45) with $\delta_0=0.005,0.05,0.1,0.5$, respectively.

In Figure 3.3 we plot the residuals versus iterative step numbers for problem (3.45). Figure 3.3 shows us that $\delta_0=0.05$ should be optimal for problem (3.45) with initial iterative value $x^0=(0.5,-2)$, compared with the other three choices of δ_0.

3.5 Summary

The main objective of this research was to show how pseudotransient continuation can be applied to nonlinear equations with inequality constraints. We have reported on numerical results to illustrate the ideas.

Chapter 4

Piece-wise System

This chapter will provide two classes of piece-wise systems, both of which have special structures. By exploring these structures, several effective Newton-type methods will be designed. These approaches have a remarkable finite termination property.

4.1 Piece-wise System I

Brugnano and Casulli [15] considered the numerical solution of a piecewise linear system,

$$\max\{0,x\} + Tx = b, \tag{4.1}$$

where

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \qquad \max\{0,x\} = \begin{pmatrix} \max\{0,x_1\} \\ \vdots \\ \max\{0,x_n\} \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix},$$

$b \in \mathbb{R}^n$ is known, $T \in \mathbb{R}^{n\times n}$ is an irreducible, symmetric, and (at least) positive semidefinite matrix satisfying either of the following properties:

(A1) T is a Stieltjes matrix, *i.e.*, a symmetric M-matrix (see, *e.g.*, [98]), or

(A2) $\texttt{null}(T) \equiv \texttt{span}(v)$ with $v > 0$ (componentwise), and $T + D$ is a Stieltjes matrix for all diagonal matrices $D = \text{diag}(d_1,\ldots,d_n)$ with $\sum_{i=1}^{n} d_i > 0$, $d_i \geq 0$, $i = 1,2,\ldots,n$.

The system (4.1) arises from the semi-implicit methods for the numerical simulation of free-surface hydrodynamics (see, *e.g.*, [20, 127]) and the numerical solutions of large-scale complementarity problems (see, *e.g.*, [34, 46]). Under the assumption (A1) or (A2), Brugnano and Casulli [15] proposed an efficient Newton-type approach with a finite termination property for solving system (4.1).

In this section[1], we first relax the assumptions (A1) and (A2), and then prove the finite termination of the Newton-type approach under our relaxed conditions. To this end, let us consider an extended linear system,

$$Tx + S\max\{0, x\} = b, \tag{4.2}$$

where $S \in \mathbb{R}^{n\times n}$ is a nonnegative matrix, *i.e.*, $S \geq 0$ (see, *e.g.*, [98]), and matrices $T, S \in \mathbb{R}^{n\times n}$ satisfy one of the following properties:

(i) T and $T+S$ are monotone matrices, *i.e.*, $T^{-1} \geq 0$, $(T+S)^{-1} \geq 0$,

(ii) T is singular, and for every $x \in \mathbb{R}^n$ there exists an entry of $b - Tx$ is positive. $T + SD$ is a monotone matrix for all diagonal matrices $D = \mathrm{diag}(d_1, \ldots, d_n)$ with $\sum_{i=1}^{n} d_i > 0$, $d_i \in [0,1]$, $i = 1, \ldots, n$.

System (4.1) is actually a special expression of system (4.2) with $S = I_{n\times n}$. As will be shown in Section 4.1.2, assumptions (A1) and (A2) are much stronger than assumptions (i) and (ii), respectively. Due to these observations, we call system (4.2) as an *extended piecewise linear system*.

The organization of this section is as follows. In Subsection 4.1.1, we discuss a Newton-type method for solving system (4.2), and prove its finite termination property. In Subsection 4.1.2, we establish some results on the existence of solution for system (4.2). Subsection 4.1.3 illustrate monotonicity of iterative sequence generated by our Newton-type method. Finally, we give our conclusion in Subsection 4.1.4.

4.1.1 Newton-type Iteration

In order to derive the Newton-type iteration for solving system (4.2), we propose some results on matrix splitting.

[1] This section is taken from [23].

Definition 4.1. *Let $A, B \in \mathbb{R}^{n\times n}$. Then $A = B - C$ is a regular splitting of A if B is invertible, $B^{-1} \geq 0$, and $C \geq 0$; it is a weak regular splitting if the condition $C \geq 0$ is replaced by $B^{-1}C \geq 0$ and $CB^{-1} \geq 0$.*

Clearly, a regular splitting is a weak regular splitting, but the converse is not true, see [98, p. 56]. The following lemma shows that there is a close connection between weak regular splitting and nonnegative inverse, see [98, Theorem 2.4.17].

Lemma 4.2. *Let $A \in \mathbb{R}^{n\times n}$ and suppose that $A = B - C$ is a weak regular splitting. Then the spectral radius $\rho(B^{-1}C) < 1$ if and only if $A^{-1} \geq 0$.*

From the definition of M-matrix (see [98, Definition 2.4.7]), we also have that an M-matrix is a monotone matrix, but the converse is not true.

Example 4.3. Let $A = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}$, then $A^{-1} = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$. Thus A is monotone, but not M-matrix.

The next result gives a characterization of monotone matrices.

Proposition 4.4. *$A \in \mathbb{R}^{n\times n}$ is monotone (i.e., $A^{-1} \geq 0$) if and only if there exist two monotone matrices B^1 and B^2 such that $B^1 \leq A \leq B^2$.*

Proof. Let $B^1, B^2 \in \mathbb{R}^{n\times n}$ be monotone and satisfy $B^1 \leq A \leq B^2$. Since $C^2 = B^2 - B^1 \geq 0$, it follows from Definition 4.1 that $B^1 = B^2 - C^2$ is a regular splitting. By Lemma 4.2 and $(B^1)^{-1} \geq 0$, the spectral radius $\rho((B^2)^{-1}C^2) < 1$.

Moreover, let $C = B^2 - A$, then $A = B^2 - C$ and $(B^2)^{-1} \geq 0$, $C \geq 0$. From Definition 4.1, it follows that $A = B^2 - C$ is a regular splitting. Since $C = B^2 - A \leq B^2 - B^1 = C^2$ and $(B^2)^{-1} \geq 0$, it suffices to show $(B^2)^{-1}C \leq (B^2)^{-1}C^2$. An application of the general comparison theorem [98, Theorem 2.4.9] yields $\rho((B^2)^{-1}C) \leq \rho((B^2)^{-1}C^2) < 1$, which establishes the monotonicity of A.

Conversely, suppose that A is monotone, then the conclusion follows readily if $B^1 = B^2 = A$. ■

Proposition 4.5. *System (4.2) is equivalent to the following system*

$$[T + SP(x)]x = b, \tag{4.3}$$

where $P(x) = \text{diag}(p(x_1), \ldots, p(x_n))$, $p(x_i)$, $i = 1, 2, \ldots, n$, *are piecewise constant functions defined as*

$$p(x_i) = \begin{cases} 1 & \text{if } x_i > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.4}$$

Proof. The equality $P(x)x = \max\{0, x\}$ implies the validity of conclusion. ■

The left-hand side of system (4.3) is not everywhere differentiable but semismooth. Therefore, Qi's generalized Newton method [109, 113] can be used to solve system (4.3)

$$x^{k+1} = x^k - \left(T + SV^k\right)^{-1} \left[\left(T + SP^k\right) x^k - b\right],$$

where

$$V^k \in \partial_B \max\{0, x^k\} = \{\text{diag}(v_1^k, \ldots, v_n^k)\}$$

with v_i^k, $i = 1, \ldots, n$ are given by:

$$v_i = \begin{cases} 0, & x_i^k < 0; \\ 0 \text{ or } 1, & x_i^k = 0; \\ 1, & x_i^k > 0. \end{cases}$$

Here $\partial_B \max\{0, x^k\}$ is called the B-subdifferential of $\max\{0, x\}$ at $x^k \in \mathbb{R}^n$[109, 113]. By the convergence theory of Qi's generalized Newton method, the above method enjoys locally quadratic convergence if an initial vector x^0 is chosen suitably. However, if further observation is given to the expression of V^k, we may design some Newton-type methods with remarkable convergence properties, such as finite termination, global monotonicity. More precisely, by taking different approximations of the B-subdifferential of $\max\{0, x^k\}$, two Newton-type methods for solving system (4.3) are established. One is

$$x^{k+1} = x^k - \left(T + SP^k\right)^{-1} \left[\left(T + SP^k\right) x^k - b\right],$$

which simplifies to the following Picard iteration,

$$\left(T + SP^k\right) x^{k+1} = b, \qquad k = 0, 1, 2, \ldots, \tag{4.5}$$

where the upper index k denotes the iteration step and $P^k = P(x^k)$. The other is

$$y^{k+1} = x^k - (T+S)^{-1}\left[\left(T + SP(x^k)\right)x^k - b\right], \tag{4.6}$$

Remark 4.6. Taking $S = I_{n\times n}$, the Picard iteration (4.5) is the iterative approach mentioned in [15] for solving piecewise linear system (4.2).

In the sequel we will establish the finite termination of the Picard iteration (4.5), and global monotone convergence of iteration (4.6) under the assumption (i) or (ii). We first show the iteration (4.5) is well-defined for solving system (4.3).

Theorem 4.7. *Let matrices T,S in system* (4.3) *satisfy either* (i) *or* (ii). *If T,S satisfy* (ii) *assume also that $P^0 \neq 0$. Then $T+SP^k$ is a monotone matrix and the iteration* (4.5) *is well defined for all $k \geq 0$.*

Proof. By assumption (i), we claim that $T \leq T + SP^k \leq T + S$. From Proposition 4.4, we have that $T + SP^k$ is a monotone matrix, and thus the iteration (4.5) is well defined.

On the other hand, if T,S satisfy (ii) and $P^0 \neq 0$, then $T + SP^0$ is a monotone matrix. Next, by induction, one assumes that for $k \geq 1$ one has $P^{k-1} \neq 0$. Therefore, the vector x^k, satisfying

$$\left(T + SP^{k-1}\right)x^k = b,$$

is well defined. Then, one has

$$SP^{k-1}x^k = b - Tx^k.$$

By assumption (ii), there exists at least an entry of $b - Tx^k$ is positive. Consequently, $P^k \neq 0$, $T + SP^k$ is a monotone matrix, and x^{k+1} is well defined. ■

The iteration (4.5) allows a very simple stopping criterion as provided by the following lemma.

Lemma 4.8. *Let matrices T,S in system* (4.3) *satisfy either* (i) *or* (ii). *If T,S satisfy* (ii) *assume also that $P^0 \neq 0$. If for some $k \geq 0$ one gets $P^{k+1} = P^k$, then $x^* = x^{k+1}$ is an exact solution of problem* (4.3), (4.4).

Proof. Since $P^{k+1} = P^k$, one has

$$\left(T + SP^k\right) x^{k+1} = \left(T + SP^{k+1}\right) x^{k+1} = b.$$

Then the assertion follows from Proposition 4.5. ■

The following theorem shows that the iteration (4.5) has a remarkable finite termination property.

Theorem 4.9**.** *Let matrices T, S in system* (4.3) *satisfy either* (i) *or* (ii). *If T, S satisfy* (ii) *assume also that $P^0 \neq 0$. Then the iteration* (4.5) *is monotonically decreasing and converges to an exact solution of problem* (4.3), (4.4) *in at most $n+1$ iterations.*

Proof. The iterative scheme (4.5) implies the following equality

$$\left(T + SP^k\right) x^{k+1} = \left(T + SP^{k-1}\right) x^k = b, \quad k = 1, 2, \ldots,$$

which implies

$$\left(T + SP^k\right) x^{k+1} = \left(T + SP^k\right) x^k - \xi^k, \tag{4.7}$$

where $\xi^k \equiv S\left(P^k - P^{k-1}\right) x^k$.

By denoting hereafter by p_i^k the ith diagonal entry of P^k, one has

$$p_i^k - p_i^{k-1} \neq 0 \quad \Rightarrow \quad \begin{cases} p_i^k = 1 \quad \text{and} \quad p_i^{k-1} = 0 \quad \Rightarrow \quad x_i^k > 0, \quad \Rightarrow \quad \xi_i^k > 0, \\ \qquad\qquad\qquad \text{or} \\ p_i^k = 0 \quad \text{and} \quad p_i^{k-1} = 1 \quad \Rightarrow \quad x_i^k \leq 0 \quad \Rightarrow \quad \xi_i^k \geq 0. \end{cases}$$

This implies $\xi^k \geq 0$. By Theorem 4.7, it follows that $\left(T + SP^k\right)^{-1} \geq 0$, and consequently, equation (4.7) implies $x^{k+1} \leq x^k$. Hence, $P^{k+1} \leq P^k$ for all $k = 1, 2, \ldots$.

Finally, from Lemma 4.8, it follows that if $P^{k+1} = P^k$ then x^{k+1} is an exact solution of system (4.3). Conversely, one obtains $P^{k+1} \neq P^k$ and, since $0 \leq P^{k+1} \leq P^k$ for all $k = 1, 2, \ldots$, this may occur at most $n - m + 1$ times where $m = \sum_{i=1}^{n} p(x_i^0)$. ■

4.1.2 Existence Results

We now present some conclusions on the existence of the solution for problem (4.3), (4.4). These results complete the framework under assumptions (i) and (ii).

Theorem 4.10**.** *Let matrices T,S in system* (4.3) *satisfy either* (i) *or* (ii)*. Then the solution of problem* (4.3)*,* (4.4) *exists and is unique.*

Proof. The existence of a solution has been established constructively by Theorem 4.9. It remains to show the uniqueness. For any two vectors x and y, it follows from (4.4) that

$$P(x)x - P(y)y = Q(x,y)\cdot(x-y)\,, \tag{4.8}$$

where $Q = \mathrm{diag}(q_1,\ldots,q_n)$, the diagonal entries q_i satisfy the inequalities $0 \le q_i \le 1$, $i = 1,2,\ldots,n$. In fact, one of the following four cases occurs:

(a) $x_i, y_i > 0 \quad \Rightarrow \quad p(x_i) = p(y_i) = 1 \quad \Rightarrow \quad q_i = 1;$

(b) $x_i, y_i \le 0 \quad \Rightarrow \quad p(x_i) = p(y_i) = 0 \quad \Rightarrow \quad q_i = 0;$

(c) $x_i > 0 \ge y_i \quad \Rightarrow \quad p(x_i) = 1,\ p(y_i) = 0 \quad \Rightarrow \quad 0 < q_i = \dfrac{x_i}{x_i - y_i} \le 1;$

(d) $x_i \le 0 < y_i \quad \Rightarrow \quad p(x_i) = 0,\ p(y_i) = 1 \quad \Rightarrow \quad 0 < q_i = \dfrac{y_i}{x_i - y_i} \le 1.$

Assume now that x and y are both solutions of system (4.3) such that

$$[T + SP(x)]x = b, \qquad [T + SP(y)]y = b.$$

Thus,

$$[T + SP(x)]x - [T + SP(y)]y = (T + SQ)(x-y) = 0. \tag{4.9}$$

By Proposition 4.4, if T,S in system (4.3) satisfy (i), it follows that $T + SQ$ is certainly a monotone matrix, and thus $x = y$.

On the other hand, if T,S in system (4.3) satisfy (ii), one has

$$SP(x)x = b - Tx, \quad SP(y)y = b - Ty.$$

Consequently, $P(x) \neq 0$, $P(y) \neq 0$ and hence at least one of the diagonal entries of Q is strictly positive. Thus, $T+SQ$ is a monotone matrix and the uniqueness $(x=y)$ follows readily from (4.9). This establishes the statement. ■

Note that assumption (ii) can be rewritten as three conditions stated below:

(iii) T is singular;

(iv) $T+SD$ is a monotone matrix for all diagonal matrices $D=\mathrm{diag}(d_1,\ldots,d_n)$ with $\sum_{i=1}^{n} d_i > 0,\ d_i \in [0,1],\ i=1,2,\ldots,n$;

(v) for every $x \in \mathbb{R}^n$ there exists an entry of $b-Tx$ is positive.

Consequently, some existence results for solution of the problem (4.3), (4.4) under the partial assertions of assumption (ii) could also be established. To be more precise, we will present an existence result for the solution of problem (4.3), (4.4) without the assumptions (iv) and (v). To this end, we suppose that

(vi) a vector $u \in \mathbb{R}^n$ exists such that $Tu=b$,

(vii) a vector $v>0$ exists such that $v \in \mathtt{null}(T)$.

Then the following result is deduced.

Proposition 4.11**.** *Let T,S of problem* (4.3), (4.4) *satisfy assumptions* (iii) *and* (vii). *If $v^\top b>0$, then T,S satisfy assumption* (v).

Proof. By assumptions (iii) and (vii), we have

$$v^\top(b-Tx)=v^\top b>0,$$

which implies the validity of conclusion. ■

Remark 4.12**.** Proposition 4.11 actually shows that assumption (ii) is weaker than the condition (A2) used in [15].

Furthermore, a conclusion can be deduced if matrices T,S in system (4.3) satisfy assumptions (iii), (vi) and (vii).

Theorem 4.13. *Let T, S of problem* (4.3), (4.4) *satisfy assumptions* (iii), (vi) *and* (vii). *Assume that*

(a) $v^\top b = 0$, *then a solution exists but is not unique,*

(b) $v^\top b < 0$, *then the problem has no solution.*

Proof. Let u_i and v_i denote the ith entries of the vectors u and v in assumptions (vi) and (vii), respectively, then for all $\alpha \geq \max\limits_{1\leq i\leq n} \dfrac{u_i}{v_i}$ the vector

$$x(\alpha) = u - \alpha v$$

satisfies

$$x(\alpha) \leq 0, \qquad Tx(\alpha) = b.$$

Consequently, $x(\alpha)$ is a solution of problem (4.3), (4.4). The assertion (a) thus follows from Proposition 4.14.

To prove the assertion (b), assume that a solution x exists, then from (4.3) one has

$$v^\top [T + SP(x)]x = v^\top SP(x)x = v^\top b < 0,$$

which is a contradiction with the assertions that $v > 0$ and $SP(x)x \geq 0$. ■

Finally, if matrices T, S in system (4.3) satisfy neither (i) nor (ii), we have the following result.

Proposition 4.14. *Suppose that*

(a) *the set* $\{u \leq 0 \mid Tu = b\} \neq \emptyset$, *then the solution of problem* (4.3), (4.4) *exists and all elements in this set are the solutions of problem* (4.3), (4.4),

(b) *for every* $x \in \mathbb{R}^n$, *there exists an entry of* $b - Tx$ *is negative, then the problem* (4.3), (4.4) *has no solution.*

Proof. The statements follow readily from the expression of problem (4.3), (4.4). ■

4.1.3 Monotonicity of Iterative Sequence

Theorem 4.7 claims that there is a finite monotonically decreasing sequence $\{x^k\}$ converges to the solution x^* of problem (4.3), (4.4). Thus $\{x^k\}$ constitutes the upper bounds for x^*. In this section we will consider the construction of an additional, monotonically increasing sequence, which provides the lower bounds for the solution x^*.

Theorem 4.15. *Let matrices T,S in system* (4.3) *satisfy either* (i) *or* (ii). *If T,S satisfy* (ii) *assume also that $P^0 \neq 0$. Assume further that there exists an initial value y^0 such that $\left(T+SP(y^0)\right)y^0 - b \leq 0$. Then the iteration* (4.6), i.e.,

$$y^{k+1} = y^k - (T+S)^{-1}\left[\left(T+SP(y^k)\right)y^k - b\right], \tag{4.10}$$

is well-defined. Moreover, the iterative sequence $\{y^k\}$ is monotonically increasing and converges to an exact solution of problem (4.3), (4.4).

Proof. The matrices T,S in problem (4.3), (4.4) satisfy either (i) or (ii), then $(T+S)^{-1}$ exists, and thus the iteration (4.10) is well defined. Next, we show by induction that

$$y^0 \leq y^{k-1} \leq y^k, \quad \left(T+SP(y^k)\right)y^k - b \leq 0.$$

Suppose this holds for some $k \geq 0$, then $(T+S)^{-1} \geq 0$ and $\left(T+SP(y^k)\right)y^k - b \leq 0$ together imply that $y^k \leq y^{k+1}$.

Taking $x = y^{k+1}$ and $y = y^k$ in (4.8), together with (4.10), we obtain

$$\begin{aligned}\left(T+SP(y^{k+1})\right)y^{k+1} - b &= \left(T+SP(y^k)\right)y^k - b + \left(T+SQ(y^{k+1},y^k)\right)\left(y^{k+1}-y^k\right) \\ &\leq \left(T+SP(y^k)\right)y^k - b + (T+S)\left(y^{k+1}-y^k\right) \\ &= 0.\end{aligned} \tag{4.11}$$

This completes the induction. Suppose that x^* is a solution of $(T+SP(x^*))x^* = b$. Taking $x = y^k$ and $y = x^*$ in (4.8), together with (4.11), we have

$$\left(T+SP(y^k)\right)y^k - (T+SP(x^*))x^* = \left(T+SQ(y^k,x^*)\right)\left(y^k - x^*\right) \leq 0. \tag{4.12}$$

If assumption (i) holds for problem (4.3), (4.4), then by assumption (i), we claim that $T \leq T + SQ(y^k, x^*) \leq T + S$. From Proposition 4.4, it follows that $T + SQ(y^k, x^*)$ in (4.12) is a monotone matrix, which implies $y^k \leq x^*$.

Alternatively, if assumption (ii) holds for problem (4.3), (4.4) and $P^0 \neq 0$, by the increasing monotonicity of $\{y^k\}$, we have that $P^{k+1} \geq P^k \geq P^0 \neq 0$ for all $k = 1, 2, \ldots$. It follows from (4.8) and (4.12) that at least one of the diagonal entries of $Q(y^k, x^*)$ in (4.12) is strictly positive. Thus, $T + SQ(y^k, x^*)$ in (4.12) is a monotone matrix, then $y^k \leq x^*$ holds.

Now $\{y^k\}$ is an upper bounded, monotonically increasing sequence thus has a limit y^*. By (4.10), we have

$$y^{k+1} - y^k = -(T+S)^{-1}\left[\left(T + SP(y^k)\right)y^k - b\right] \geq 0. \tag{4.13}$$

But $\lim_{k\to\infty}\left(y^{k+1} - y^k\right) = 0$, so that

$$\lim_{k\to\infty}\left(\left(T + SP(y^k)\right)y^k - b\right) = 0$$

which implies $\left(T + SP(y^*)\right)y^* - b = 0$. Theorem 4.10 about the unique existence for the solution of problem (4.3), (4.4) then implies $y^* = x^*$. This establishes the assertions. ■

In practice, the determination of x^{k+1} from iteration (4.5) can be accomplished quite efficiently by using a preconditioned Krylov subspace method (see, *e.g.*, [116]). This is particularly the case in applications where T is a sparse and very large matrix. For the choice of a starting point for the used preconditioned Krylov subspace method in each iteration (4.5), the following result provide a criterion.

Proposition 4.16. *Let matrices T, S in system* (4.3) *satisfy either* (i) *or* (ii). *If T, S satisfy* (ii) *assume also that $P^0 \neq 0$. Assume further that at the kth iteration* (4.27)*, there exist two vectors h^{k+1} and g^{k+1} such that*

$$\left(T + SP^k\right)h^{k+1} \leq b, \qquad \left(T + SP^k\right)g^{k+1} \geq b. \tag{4.14}$$

Then the exact solution x^{k+1} of the kth iteration (4.5) satisfies

$$h^{k+1} \leq x^{k+1} \leq g^{k+1}. \tag{4.15}$$

Proof. By Theorem 4.7, it follows that $T + SP^k$ is a monotone matrix and the iteration (4.5) is well defined for all $k \geq 0$. From the monotonicity of $T + SP^k$, *i.e.* $\left(T + SP^k\right)^{-1} \geq 0$, we have

$$\begin{aligned}\left(T + SP^k\right)^{-1}\left(T + SP^k\right) h^{k+1} &\leq \left(T + SP^k\right)^{-1} b \\ &= x^{k+1} \\ &\leq \left(T + SP^k\right)^{-1}\left(T + SP^k\right) g^{k+1},\end{aligned}$$

which implies (4.15) holds. ■

Remark 4.17**.** By Theorem 4.9, we obtain $x^{k+1} \leq x^k$. Hence using x^k as a starting point is reasonable and convenient for the preconditioned Krylov subspace method employed in each iteration (4.27).

4.1.4 Contributions and Future Research

This section shows that under more relaxed assumption (i) or (ii) (compared with condition (A1) or (A2) used in [15]), the Newton-type method converges to an exact solution of the given system in a finite number of steps. The existence results of solution for the piecewise linear system are established.

Finally, we point out that under condition (A1) or (A2), two constructive iterative methods to solve a piecewise linear system of the form

$$\max\{l, \min\{u, x\}\} + Tx = b, \quad l, u, b \in \mathbb{R}^n, l = (l_i) \leq u = (u_i)$$

are analyzed with a finite termination property in [14]. Under more relaxed assumption (i) or (ii), the finite termination property for these algorithms in [14] can also be established following an analogous analysis to this chapter with slight and technical modifications.

4.2 Piece-wise system II

In this section[2] we consider the numerical solution of another special linear systems whose coefficient matrix is a piecewise constant function of the solution itself, *i.e.*,

$$Sx + T\max\{0,x\} = b, \tag{4.16}$$

where the operators max is to be intended componentwise, $b \in \mathbb{R}^n$ is known, $S \in \mathbb{R}^{n\times n}$ is a positive diagonal matrix. $T \in \mathbb{R}^{n\times n}$ is an irreducible, symmetric, and (at least) positive semidefinite matrix, satisfying either one of the following properties:

T1: T is a Stieltjes matrix, i.e., a symmetric M-matrix (see, *e.g.*, [98]), or

T2: $\mathtt{null}(T) \equiv \mathtt{span}(v)$ with $v > 0$ (componentwise), and $T + D$ is a Stieltjes matrix for all diagonal matrices $D = \mathrm{diag}(d_1,\ldots,d_n)$ with $\sum_{i=1}^{n} d_i > 0$, $d_i \geq 0$, $i = 1,2,\ldots,n$.

Note that upon a suitable variable transformation, the following problems can be taken back to problem (4.16):

$$Sx + T\max\{\xi,x\} = b, \tag{4.17}$$

$$Sx + T\min\{\xi,x\} = b, \tag{4.18}$$

where ξ is a given vector.

The efficient solution of system (4.16) is of interest in numerical optimization because this system can be cast as a *linear complementarity problem* (see, *e.g.*, [34]). In fact, by setting $y = \max\{0,x\}$ and $z = y - x$, system (4.16) can be formulated either as a horizontal linear complementarity problem,

$$(S+T)y = z + b, \qquad y^\top z = 0, \qquad y,z \geq 0, \tag{4.19}$$

or, equivalently, as a standard linear complementarity problem,

$$y = Mz + q, \qquad y^\top z = 0, \qquad y,z \geq 0, \tag{4.20}$$

[2] This section is taken from [16].

where $q = (S+T)^{-1}b$ and $M = (S+T)^{-1}$ is a positive definite matrix (see [41, 8]).

When the size of a linear complementarity problem is reasonably small, it can be solved by means of a (direct) pivoting method (see, *e.g.*, [34, 41, 88]). For large and sparse problems, however, these methods suffer from unacceptable roundoff error accumulation and excessive storage requirement. Iterative (indirect) methods such as, interior-point type methods (see, *e.g.*, [105, 106, 108, 136]), nonsmooth Newton methods (see, *e.g.*, [36, 109, 113, 74]), are therefore employed to solve large-scale complementarity problems. Facchinei and Pang in their monograph [46] presented a comprehensive, state-of-the-art treatment of the iterative solution for complementarity problem. However, these iterative methods are characterized by having a convergence without a global monotonicity and, moreover, generally occurring only in the limit of an infinite number of iterations.

Another application of system (4.16) is the absolute value programming introduced by Mangasarian and Meyer (see, *e.g.*, [93, 91, 92]). Since $|x| = 2\max\{0,x\} - x$, then system (4.16) could be reformulated as

$$(2I+T)x + T|x| = 2b, \tag{4.21}$$

which is an absolute value equation. It should be note that the main difference between (4.21) and the absolute value equations mentioned in [93, 91, 92] is that the matrix T in (4.21) satisfies assumptions **T1** or **T2**. This crucial difference causes that the problem (4.21) is not NP-hard. For details of the NP-hardness about a general absolute value programming, see the above mentioned references [93, 91, 92].

More recently, Brugnano and Casulli [15] also considered the solution of large systems in the form

$$\max\{0,x\} + Tx = b, \tag{4.22}$$

where $T \in \mathbb{R}^{n\times n}$ satisfies either **T1** or **T2**. System (4.22) arises from the use of semi-implicit methods for the numerical simulation of free-surface hydrodynamics (see, *e.g.*, [20, 21]). More precisely, a correct formulation of numerical methods for free-surface hydrodynamics, that guarantees nonnegative water depths for any time step, requires the solution of a large and sparse system in the form (4.22). Due to the piecewise characterization of $\max\{0,x\}$, Brugnano and Casulli [15] then called system (4.22) as

a *piecewise linear system*. An efficient semi-iterative Newton-type approach for solving system (4.22) was derived and its convergence within a finite number of iterations was also established in [15]. Because of this, we call system (4.16) an *extended piecewise linear system*.

The main contribution of this section is to generalize the semi-iterative Newton-type approach employed in [15] to solve the extended piecewise linear system (4.16). In the next subsection the semi-iterative Newton-type procedures for solving system (4.16) will be derived. A remarkable monotone convergence of the semi-iterative Newton-type procedures will also be established. In Subsection 4.2.2, some numerical tests are provided to confirm the excellent convergence properties of the proposed algorithms, also presenting the application of systems in the form (4.16) for solving real-life problems. Finally, in Subsection 4.2.3, a few concluding remarks are given.

4.2.1 The Newton-type Iteration

Some preliminary results are stated at first in order to derive the Newton-type iterative procedure for solving system (4.16) and prove its convergence. Their proof is straightforward and is, therefore, omitted.

Lemma 4.18**.** *Let T satisfy either* **T1** *or* **T2**. *Then, for any diagonal matrix P with non-negative diagonal entries, matrix $S+TP$ is an M-matrix and, therefore, $(S+TP)^{-1} \geq 0$.*

Lemma 4.19**.** *System* (4.16) *is equivalent to the following system*

$$[S+TP(x)]\,x = b, \tag{4.23}$$

where $P(x) = \mathrm{diag}(p(x_1),\ldots,p(x_n))$, where $p(x_i)$, $i = 1,2,\ldots,n$, are piecewise constant functions defined as

$$p(x_i) = \begin{cases} 1 & \text{if } x_i \geq 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.24}$$

It is to be noted that the left-hand side of system (4.23) is not everywhere differentiable. Nevertheless, a *Newton-type method* for solving system (4.23) can be deduced,

$$x^{k+1} = x^k - \left(S+TP^k\right)^{-1}\left[\left(S+TP^k\right)x^k - b\right], \qquad k = 0,1,\ldots, \tag{4.25}$$

where the upper index k denotes the iteration step and

$$P^0 = 0, \qquad P^k = P(x^k), \quad k = 1,2,\ldots. \tag{4.26}$$

This simplifies to the following Picard iteration,

$$P^0 = 0, \qquad \left(S + TP^k\right) x^{k+1} = b, \qquad k = 0,1,2,\ldots. \tag{4.27}$$

In the sequel we will establish the convergence of the Picard iteration (4.27). We first show that the iteration (4.27) is well-defined.

Theorem 4.20**.** *Let matrix T in system* (4.23) *satisfy either* **T1** *or* **T2**. *Then $S+TP^k$ is an M-matrix and the iteration* (4.27) *is well defined for all $k \geq 0$.*

Proof. Trivial, from Lemma 4.18. ■

The iteration (4.27) allows a very simple stopping criterion, as provided by the following lemma.

Lemma 4.21**.** *Let matrix T in system* (4.23) *satisfy either* **T1** *or* **T2**. *If, for some $k \geq 0$, one gets $(P^{k+1} - P^k)x^{k+1} = 0$, then $x^* = x^{k+1}$ is an exact solution of problem* (4.23)–(4.24).

Proof. Since $(P^{k+1} - P^k)x^{k+1} = 0$, one has

$$\left(S + TP^k\right) x^{k+1} = \left(S + TP^{k+1}\right) x^{k+1} = b.$$

Then the assertion follows from Lemma 4.19. ■

Next result provides a monotonicity property for the iteration (4.27).

Theorem 4.22**.** *Let matrix T in system* (4.23) *satisfy either* **T1** *or* **T2**. *Then $P^k x^{k+1} \geq P^{k-1}x^k$.*

Proof. From (4.27) one obtains that

$$(S + TP^k)x^{k+1} = b = (S + TP^{k-1})x^k, \qquad k \geq 1.$$

Consequently, from left multiplication by P^k, one obtains that

$$\begin{aligned}(S+P^kT)P^kx^{k+1} &= SP^kx^k+P^kTP^{k-1}x^k\\ &= (S+P^kT)P^{k-1}x^k+S(P^k-P^{k-1})x^k\\ &\geq (S+P^kT)P^{k-1}x^k,\end{aligned}$$

Since one readily verifies that, because of the definition (4.26), $S(P^k-P^{k-1})x^k \geq 0$. Because of the result of Theorem 4.20, matrix $(S+P^kT)$ is an M-matrix and, therefore, $(S+P^kT)^{-1} \geq 0$. The thesis then follows immediately. ■

As a consequence of the previous monotonicity property, convergence in a finite number of steps is established.

Corollary 4.23**.** *Let matrix T in system* (4.23) *satisfy either* **T1** *or* **T2**. *Then iteration* (4.27) *converges in at most n steps.*

Proof. Indeed, from the result of Theorem 4.22, one obtains that

$$P^kx^{k+1} \geq P^{k-1}x^k \geq P^0x^1 = 0, \qquad k = 1,2,\ldots.$$

Consequently, from the definitions (4.24) and (4.26), one obtains, by setting, as usual $x^k = (x_i^k)$ and $x^{k+1} = (x_i^{k+1})$,

$$x_i^k \geq 0 \Rightarrow x_i^{k+1} \geq 0, \qquad i.e. \qquad P^{k+1} \geq P^k \geq 0.$$

By considering the stopping criterion provided by Theorem 4.21, one has that at each step one either has $P^k = P^{k-1}$, and then the solution has been reached, or $P^k \neq P^{k-1}$. The latter case can obviously happen at most n times, where n is the dimension of the problem. ■

Remark 4.24**.** Even though Corollary 4.23 establishes the finite convergence of iteration (4.27), nevertheless the corresponding upper bound may be large, when the dimension of the system is large. However, convergence practically occurs in just a few iterates, as it is also confirmed by the numerical tests in Section 4.2.2.

Next, we present a conclusion on the existence of the solution for problem (4.23). We need the following preliminary result.

Lemma 4.25**.** *With reference to matrix P defined in* (4.24)*, for any two vectors x and y, one has*

$$P(x)x - P(y)y = W \cdot (x - y)\,, \tag{4.28}$$

where

$$W = \mathrm{diag}(w_1, \ldots, w_n), \qquad \textit{with} \qquad 0 \le w_i \le 1, \quad i = 1, 2, \ldots, n.$$

Proof. For each $i = 1, 2, \ldots, n$, it follows from (4.24) that either one of the following four cases occurs, for any two vectors $x = (x_i)$ and $y = (y_i)$:

(a) $x_i, y_i \ge 0 \quad \Rightarrow \quad p(x_i) = p(y_i) = 1 \quad \Rightarrow \quad w_i = 1;$

(b) $x_i, y_i < 0 \quad \Rightarrow \quad p(x_i) = p(y_i) = 0 \quad \Rightarrow \quad w_i = 0;$

(c) $x_i \ge 0 > y_i \quad \Rightarrow \quad p(x_i) = 1,\ p(y_i) = 0 \quad \Rightarrow \quad 0 \le w_i = \dfrac{x_i}{x_i - y_i} < 1;$

(d) $x_i < 0 \le y_i \quad \Rightarrow \quad p(z_i) = 0,\ p(y_i) = 1 \quad \Rightarrow \quad 0 \le w_i = \dfrac{y_i}{y_i - x_i} < 1.$

This shows the validity of (4.28). ■

Theorem 4.26**.** *Let matrix T in system* (4.23) *satisfy either* **T1** *or* **T2***. Then, the solution of problem* (4.23)–(4.24) *exists and is unique.*

Proof. The existence of a solution has been established constructively by Corollary 4.23. It remains to prove its uniqueness. Assume that x and y are both solutions of system (4.23), i.e.,

$$[S + TP(x)]\,x = b, \qquad [S + TP(y)]\,y = b.$$

From Lemma 4.25, it follows that

$$[S + TP(x)]\,x - [S + TP(y)]\,y = (S + TW)\,(x - y) = 0, \tag{4.29}$$

where W is a diagonal matrix with nonnegative diagonal entries. By Lemma 4.18, if matrix T in system (4.23) satisfies either **T1** or **T2**, it follows that $I + TW$ is an M-matrix and, thus, $x = y$. ■

For completeness, we mention that, for Problem (4.17), the corresponding iteration is:

$$P^0 = 0, \qquad Sx^{k+1} + TP^k(x^{k+1} - \xi) = b - T\xi, \qquad k = 0,1,2,\ldots, \tag{4.30}$$

where $\xi = (\xi_i)$ and

$$P^k = P(x^k) = \mathrm{diag}(p(x_i^k)), \qquad p(x_i^k) = \begin{cases} 1 & \text{if } x_i \geq \xi_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.31}$$

Similarly, for Problem (4.18), the corresponding iteration is given by

$$Q^0 = I, \qquad Sx^{k+1} + TQ^k(x^{k+1} - \xi) = b - T\xi, \qquad k = 0,1,2,\ldots, \tag{4.32}$$

where

$$Q^k = Q(x^k) = \mathrm{diag}(q(x_i^k)), \qquad q(x_i^k) = \begin{cases} 1 & \text{if } x_i < \xi_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.33}$$

4.2.2 Numerical Tests

We here consider a test problem, concerning the heat transmission in a medium which, only for sake of simplicity, is assumed to be homogeneous.

Consider, at first, a tiny wire of length L which is heated at $x = 0$, at temperature u_0, and is insulated at $x = L$. If the wire is initially at temperature $u = 0$, and its *thermal diffusivity* is κ, then the governing equation for $u(x,t)$ is the well-known *heat equation*,

$$\frac{\partial}{\partial t}u(x,t) = \kappa \frac{\partial^2}{\partial x^2}u(x,t), \qquad 0 < x < L, \quad t > 0, \tag{4.34}$$

with boundary conditions

$$u(x,0) \equiv 0, \quad 0 < x < L, \qquad \text{and} \qquad u(0,t) = u_0, \quad \frac{\partial}{\partial x}u(L,t) = 0, \quad t > 0. \tag{4.35}$$

Let then consider a discretization for the space variable with stepsize

$$\Delta x = \frac{L}{N+1}, \tag{4.36}$$

and of the time variable with stepsize Δt. By setting, as usual, by u_i^k the numerical approximation to $u(i\Delta x, k\Delta t)$, one then obtains the discrete equation

$$\frac{u_i^{k+1}-u_i^k}{\Delta t} = \frac{\kappa}{\Delta x^2}\left(u_{i-1}^{k+1} - 2u_i^{k+1} + u_{i+1}^{k+1}\right), \qquad i = 1,\ldots,N, \quad k \geq 0, \tag{4.37}$$

with boundary conditions

$$u_i^0 = 0, \quad i = 1,\ldots,N, \qquad u_0^k = u_0, \quad u_{N+1}^k = u_N^k, \quad k \geq 1. \tag{4.38}$$

Upon multiplication of (4.37) by Δt, one may cast all the equations as a linear system in the form

$$(I+T)u^{k+1} = u^k + \eta, \tag{4.39}$$

where

$$u^k = \begin{pmatrix} u_1^k \\ \vdots \\ u_N^k \end{pmatrix}, \qquad T = \frac{\kappa\Delta t}{\Delta x^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & 2 & -1 \\ & & -1 & 2 \end{pmatrix}, \qquad \eta = \frac{\kappa u_0 \Delta t}{\Delta x^2}\begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{4.40}$$

We observe that matrix T is **T1**: in general this will be the case if the temperature is prescribed in at least one point of the boundary (as in the present case); on the other hand, matrix T will satisfy property **T2** when only the heat flux is prescribed (i.e., only Neumann conditions are specified for the problem). The solution of Problem (4.38)-(4.35) can then be easily approximated by solving the discrete problem (4.37)-(4.38). However, let now consider the following modification to the original problem, i.e., assume that the wire is plugged into a thermostat, which cools the wire as soon as its temperature reaches a specifield theshold $u_{\max}$ (which we assume, for simplicity, to be constant, even though it may vary both in space and time). In such a case, the discrete equation (4.37) is no more a correct formalization of the new problem. In order to obtain a new, more appropriate discrete problem, let us rewrite Equation (4.37) as

$$\left(u_i^{k+1} - u_i^k\right)\Delta x = \kappa\Delta t\left(\frac{u_{i+1}^{k+1}-u_i^{k+1}}{\Delta x} - \frac{u_i^{k+1}-u_{i-1}^{k+1}}{\Delta x}\right), \qquad i = 1,\ldots,N, \quad k \geq 0,$$

which can be read as a *conservation law*: namely, by considering that the heat flux is directed in the direction of the negative gradient, the difference of heat at x_i, at the next time step, is obtained as the difference between the heat flux "entering" from the left,

$$-\kappa\Delta t\frac{u_i^{k+1}-u_{i-1}^{k+1}}{\Delta x},$$

and that "exiting" from the right,

$$-\kappa\Delta t\frac{u_{i+1}^{k+1}-u_i^{k+1}}{\Delta x}.$$

Consequently, by defining

$$\phi_i^k=\min\{u_i^k,u_{\max}\},\qquad i=1,\ldots,N,\quad k\geq 0,$$

the generalization of the model for the new problem becomes:

$$\left(u_i^{k+1}-\phi_i^k\right)\Delta x=\kappa\Delta t\left(\frac{\phi_{i+1}^{k+1}-\phi_i^{k+1}}{\Delta x}-\frac{\phi_i^{k+1}-\phi_{i-1}^{k+1}}{\Delta x}\right),\qquad i=1,\ldots,N,\quad k\geq 0.$$

Evidently, the quantity

$$\psi_i^{k+1}\equiv u_i^{k+1}-\phi_i^{k+1},$$

will be the temperature fall due to the action of the thermostat. Consequently, the problem to be solved, at the k-th time step, will be

$$u^{k+1}+T\min\{u^{k+1},u_{\max}\}=b\equiv\min\{u^k,u_{\max}\}+\eta,\tag{4.41}$$

with the vector $u_{\max}$ containing the given upper bound due to the introduction of the thermostat (i.e., a constant vector, in the present example), and matrix T satisfying either **T1** or **T2**, depending on the specified boundary conditions (i.e., **T1**, for the current example). Clearly, Problem (4.41) is in the form (4.18) and, then, the corresponding iteration (4.32)-(4.33) has to be used. In Table 4.1 we list the number of iterations required for convergence, when

$$\kappa=10^{-3}\,m^2/s,\qquad L=1\,m,\qquad u_0=5\,C\qquad u_{\max}\equiv 2,\qquad \Delta t=10^2\,s.\tag{4.42}$$

TABLE 4.1
Number of iterations required for solving problem (4.36) and (4.40)–(4.42) for various values of N.

$k \setminus N$	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
1	3	3	3	3	3	3	3	3	3	3
2	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3
4	3	3	3	3	3	3	3	3	3	3
5	3	3	3	3	3	3	3	3	3	3
6	3	3	3	3	3	3	3	3	3	3
7	3	3	3	3	3	3	3	3	3	3
8	3	3	3	3	3	3	3	3	3	3
9	3	3	3	3	3	3	3	3	3	3
10	3	3	3	3	3	3	3	3	3	3

As one can see, the number of the required iterations is remarkably small and insensitive of grid resolution. For completeness, in Figure 4.1 there is the plot of the computed solution at

$$t = 100, 200, \ldots, 1000\, s.$$

The second problem is derived from the *parabolic obstacle problem* within financial mathematics [104, 121]. This problem is outlined as follows. Let Q be parabolic cylinder in $\mathbb{R}^n \times \mathbb{R}$, and let $\phi(x,t)$ be parabolically $C^{0,\alpha}$ in Q. Set

$$H(u) = F(D^2u, Du, u, x, t) - D_t u \tag{4.43}$$

where F is a full nonlinear uniformly elliptic operator within certain homogeneity properties for which the regularity theory of viscosity solutions applies. Let u solve the parabolic obstacle problem

$$\begin{cases} (u-\phi)H(u) & = 0, \\ u & \geq \phi, \text{ in } Q, \\ H(u) & \leq 0, \end{cases} \tag{4.44}$$

with boundary datum

$$u(x,t) = g(x,t) \geq \phi(x,t) \text{ on } \partial_p Q, \tag{4.45}$$

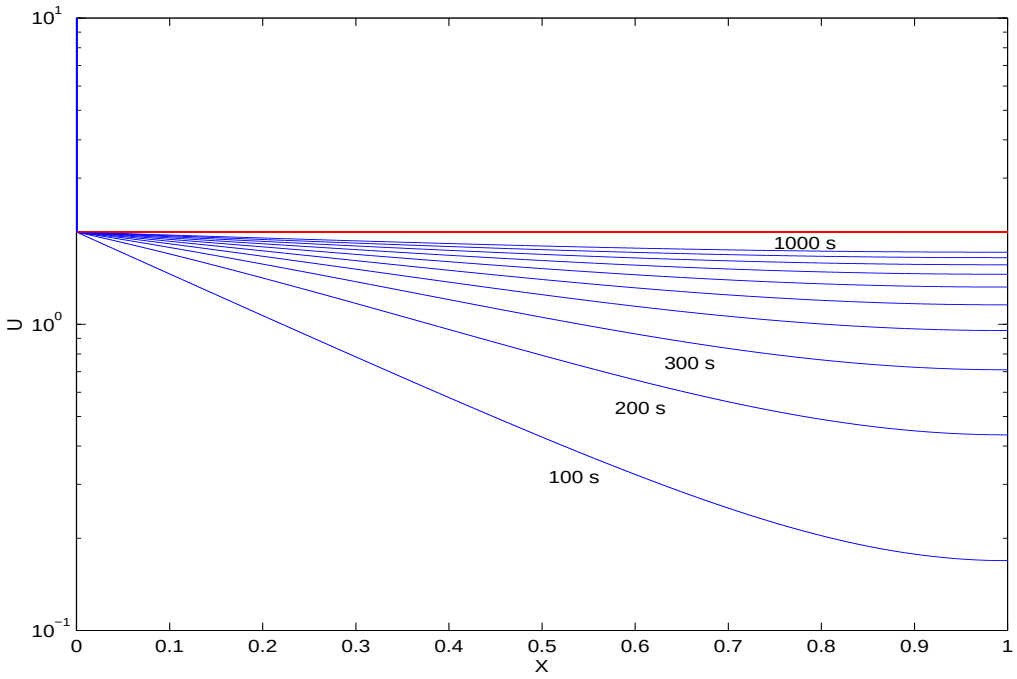

FIGURE 4.1 *Computed solution of problem* (4.36) *and* (4.40)–(4.42).

TABLE 4.2
Obstacles in the applications.

Obstacle ϕ	Applications
$\max\{0, E - x_1\}$	1-dimension contract, American put
$\min\{\max\{0, E - x_1\}, \max\{0, E - x_2\}\}$	min option, American put

$$u(x,0) = g(x,0) = \phi(x,0). \tag{4.46}$$

Here ∂_p denote the parabolic boundary. In fact, equations (4.44) is equivalent to a nonsmooth system

$$\min\{u - \phi, -H(u)\} = 0. \tag{4.47}$$

The obstacle ϕ in general has singularities, usually representing a change in the nature of the contract in applications in finance. Examples of obstacles that appear in finance are given in Table 4.2. (here E is a constant and it denotes the exercise price). See [104] for more details.

In the present test $H(u)$ in parabolic obstacle problem (4.44) is described by a parabolic equation

$$H(u) = -\frac{\partial}{\partial t}u(x,t) + \frac{\partial^2 u(x,t)}{\partial x_1^2} + \frac{\partial^2 u(x,t)}{\partial x_2^2}, \quad Q = [-l,m]\times[-l,m]\times[0,T). \quad (4.48)$$

The obstacle is given by

$$\phi(x) = \max\{0, \min\{x_1, x_2\}\}. \quad (4.49)$$

Moreover, we only consider the numerical solution of (4.44) with a fixed boundary (4.45), that is, $g(x,t) = \phi(x,t)$ in (4.45). More complicated cases involved a free boundary will not be discussed herein.

Let then consider a consistent implicit discretization for the space variables with stepsize

$$\Delta x_1 = \Delta x_2 = \frac{l+m}{N+1}, \quad (4.50)$$

and of the time variable with stepsize Δt. By setting, as usual, by $u_{i,j}^k$ the numerical approximation to $u(i\Delta x_1, j\Delta x_2, k\Delta t)$, one then obtains the discrete equation

$$\min\left\{H_{i,j}^{k+1}, u_{i,j}^{k+1} - \phi_{i,j}^{k+1}\right\} = 0, \qquad i,j = 1,\ldots,N, \quad k \geq 0, \quad (4.51)$$

where

$$H_{i,j}^{k+1} = \frac{u_{i,j}^{k+1} - u_{i,j}^k}{\Delta t} - \frac{u_{i-1,j}^{k+1} - 2u_{i,j}^{k+1} + u_{i+1,j}^{k+1}}{(\Delta x_1)^2} - \frac{u_{i,j-1}^{k+1} - 2u_{i,j}^{k+1} + u_{i,j+1}^{k+1}}{(\Delta x_2)^2}, \quad (4.52)$$

$$\phi_{i,j}^{k+1} = \max\left\{0, \min\left\{-l + \frac{i}{N+1}(l+m), -l + \frac{j}{N+1}(l+m)\right\}\right\}. \quad (4.53)$$

The corresponding boundary conditions are

$$u_{i,j}^0 = \phi_{i,j}^0, \quad i,j = 0,\ldots,N+1. \quad (4.54)$$

Then we may cast all the equations as a linear system in the form

$$u^{k+1} + \max\left\{Tu^{k+1} + \left(\frac{1}{\Delta t} - 1\right)u^{k+1}, \frac{1}{\Delta t}u^k - \phi^{k+1}\right\} = \frac{1}{\Delta t}u^k, \quad (4.55)$$

where

$$u^k = \begin{pmatrix} u^k_{1,1} \\ \vdots \\ u^k_{N,1} \\ \vdots \\ u^k_{1N} \\ \vdots \\ u^k_{NN} \end{pmatrix}, \quad T = \frac{1}{(\Delta x_1)^2} \begin{pmatrix} A & -I & & \\ -I & \ddots & \ddots & \\ & \ddots & A & -I \\ & & -I & A \end{pmatrix}, \quad \phi^{k+1} = \begin{pmatrix} \phi^{k+1}_{1,1} \\ \vdots \\ \phi^{k+1}_{N,1} \\ \vdots \\ \phi^{k+1}_{1N} \\ \vdots \\ \phi^{k+1}_{NN} \end{pmatrix}. \tag{4.56}$$

Herein I is $N \times N$ identity matrix, and

$$A = \begin{pmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & 4 & -1 \\ & & -1 & 4 \end{pmatrix}_{N \times N} \tag{4.57}$$

We observe that matrix T is **T1** if the time stepsize and space stepsize satisfy

$$\frac{1}{\Delta t} > 1 - \frac{8 \sin^2\left(\frac{\pi}{2}\Delta x_1\right)}{\Delta x_1^2}. \tag{4.58}$$

Generally this will be the case if the temperature is prescribed in at least one point of the boundary (as in the present case); on the other hand, matrix T will satisfy property **T2** when only the heat flux is prescribed (i.e., Neumann conditions are specified for the problem. However, under the assumption $D_t u \geq D_t \phi$ in Q, which mathematically corresponds to the Stefan problem case. It is beyond the problem considered herein). Clearly, parabolic obstacle problem (4.44) is in the form (4.17) and, then, the corresponding iteration (4.30)-(4.31) has to be used. In Table 4.3 we list the number of iterations required for convergence, when

$$l = 10, \qquad m = 4, \qquad T = 5\,s, \qquad \Delta t = 0.05\,s. \tag{4.59}$$

As one can see, the number of the required iterations is remarkably small and insensitive

TABLE 4.3
Number of iterations required for solving problem (4.44) *and* (4.45)–(4.46) *for various values of N.*

$k \setminus N$	16	24	32	48	56	64
1	3	3	3	2	2	2
2	3	3	3	2	2	2
3	3	3	3	2	2	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
98	3	3	3	2	2	2
99	3	3	3	2	2	2
100	3	3	3	2	2	2

of grid resolution. For completeness, in Figure 4.2 there is the plot of the computed solution at $t = 5\,s$.

4.2.3 Summary

A simple semi-iterative Newton-type procedure for solving certain extended piecewise linear systems has been derived and investigated. It is shown that under rather general assumptions, the iterates are well defined and monotonically converge to the exact solution of the given system in a finite number of steps. Simple, non trivial, numerical examples prove the effectiveness of the proposed method for solving free-surface problems derived from real-life applications.

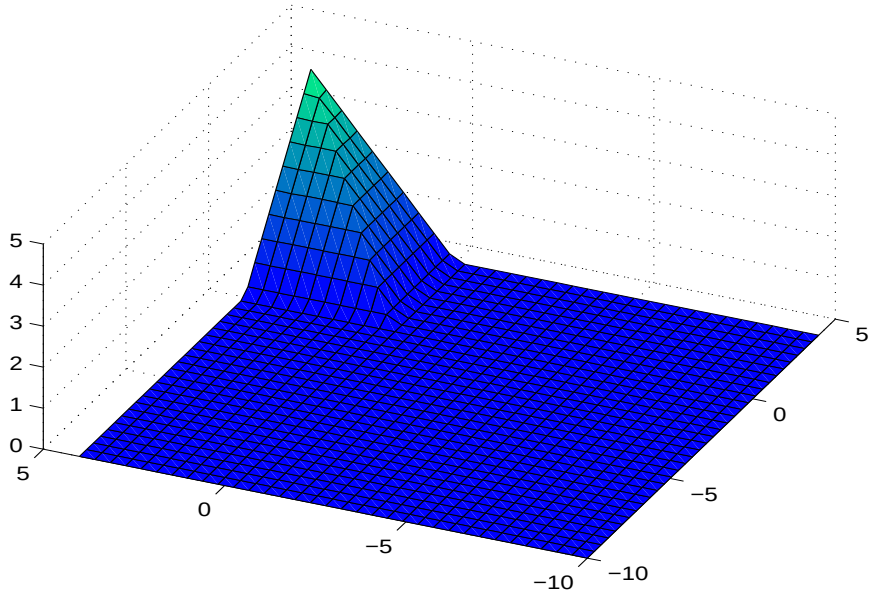

FIGURE 4.2 *Computed solution of problem* (4.44) *and* (4.45)–(4.46) *at* $T = 5\,s$.

Chapter 5

Numerical Solution for Optimal Control Problem

This chapter [1] considers the following optimal control problem (OCP for simplicity) subject to mixed control-state constraints:

$$
\text{(OCP)} \qquad \begin{aligned}
&\text{Minimize} && \int_0^1 f_0(x(t),u(t))\,\mathrm{d}t \\
&\text{w.r.t.} && x \in W^{1,\infty}([0,1],\mathbb{R}^{n_x}), u \in L^{\infty}([0,1],\mathbb{R}^{n_u}), \\
&\text{s.t.} && x'(t) = f(x(t),u(t)) \text{ a.e. in } [0,1], \\
& && \psi(x(0),x(1)) = 0, \\
& && c(x(t),u(t)) \leq 0 \text{ a.e. in } [0,1].
\end{aligned}
$$

Without loss of generality, the discussion is restricted to autonomous problems on the fixed time interval $[0,1]$. The functions $f_0 : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$, $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}$, $\psi : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_\psi}$, $c : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_c}$, are supposed to be at least twice continuously differentiable (i.e., C^2) w.r.t. all arguments. As usual, the Banach space $L^{\infty}([0,1],\mathbb{R}^n)$ consists of all measurable functions $h : [0,1] \to \mathbb{R}^n$ with

$$\|h\|_{\infty} = \operatorname*{ess\,sup}_{0 \leq t \leq 1} \|h(t)\| < \infty,$$

[1] This chapter is taken from [24].

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$. For $1 \le q < \infty$ the Banach space $L^q([0,1],\mathbb{R}^n)$ consists of all measurable functions $h : [0,1] \to \mathbb{R}^n$ with

$$\|h\|_q = \left(\int_0^1 \|h(t)\|^q \, \mathrm{d}t \right)^{\frac{1}{q}} < \infty.$$

For $1 \le q \le \infty$ the Banach space $W^{1,q}([0,1],\mathbb{R}^n)$ consists of all absolutely continuous functions $h : [0,1] \to \mathbb{R}^n$ with

$$\|h\|_{1,q} = \max\{\|h\|_q, \|h'\|_q\} < \infty.$$

Several approaches towards the numerical solution of OCP have been investigated in the literature. The so-called direct discretization method is based on a discretization of the infinite dimensional optimal control problem and leads to a finite dimensional nonlinear program, cf., e.g., Gerdts [53]. The direct discretization method turns out to be very robust in practice. Nevertheless, the computational effort grows at a nonlinear rate with the number of grid points used for discretization. Another numerical method for optimal control problems is the so-called indirect method, this approach attempts to satisfy the necessary conditions that are provided by the well-known minimum principle [62] numerically. Although the indirect method usually leads to the most accurate solutions, it suffers from the drawback that it requires a sufficiently good initial guess of the solution in order to converge. One crucial task is to estimate the sequence of active and inactive intervals of the control-state constraint. For more details about the direct discretization methods and indirect methods for OCP, we refer to Büskens [19], Gerdts [55], Grötschel et al. [60], Ioffe and Tihomirov [70] and the references therein.

Most recently, Gerdts [56] analyzed the local and global convergence properties of a nonsmooth Newton method for the numerical solution of OCP. This method was based on a nonsmooth reformulation of the necessary optimality conditions for the OCP, see Gerdts [54]. More precisely, the reformulation of the necessary conditions leads to the nonsmooth equation

$$F(z) = 0, \qquad F : Z \to Y, \tag{5.1}$$

where Z and Y are appropriate Banach spaces. Application of the globalized nonsmooth

Newton method generates sequences $\{z^k\}$, $\{d^k\}$ and $\{\alpha_k\}$ related by the iteration

$$z^{k+1} = z^k + \alpha_k d^k, \qquad k = 0, 1, 2, \ldots.$$

Herein, the search direction d^k is the solution of the linear operator equation

$$V^k(d^k) = -F(z^k) \tag{5.2}$$

and the step length $\alpha_k > 0$ is determined by a line-search procedure of Armijo's type for a suitably defined merit function. The linear operator V^k is chosen from an appropriately defined generalized Jacobian matrix $\partial_* F(z^k)$.

However, computing the exact solution via (5.2) could be expensive for large dimensional problems (e.g. problems originating from discretized partial differential equations) and may not be justified when x^k is far from a solution. These difficulties motivate us to invoke another classical tool for nonsmooth equations (5.1): the inexact Newton method. Actually, the notion of inexact solution in algorithms for solving nonsmooth equations in finite dimensions was suggested by Pang [101], and has been employed by Martínez and Qi [94], Kanzow [74], Facchinei and Kanzow [45].

Following the general framework of Ulbrich [131, 130] which was used to solve certain optimal control problems subject to partial differential equations, one of the contributions of this chapter is to extend the inexact nonsmooth and smoothing Newton method to infinite spaces. The other contribution is the application to the numerical solution of the OCP. The application of the inexact nonsmooth and smoothing Newton method to this problem class has not been investigated in detail by now.

The chapter is organized as follows. Section 5.1 reformulates OCP to (5.1) by exploitation of the minimum principle. Section 5.2 introduces the inexact nonsmooth Newton method and establishes the locally superlinear convergence under comparatively mild assumptions. Section 5.3 analyzes the global convergence properties of the inexact nonsmooth Newton method based on a non-monotonic backtracking strategy. Section 5.4 proposes a smoothing approach for the numerical solution of OCP, and illustrates its convergence. Numerical experiments are presented in Section 5.5. Finally, we make some conclusions and comment on possible further work in Section 5.7.

5.1 Reformulation

The (augmented) Hamilton function $H : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_c} \to \mathbb{R}$ is defined by

$$H(x,u,\lambda,\eta) = f_0(x,u) + \lambda^\top f(x,u) + \eta^\top c(x,u).$$

We summarize the well-known minimum principle for OCP. Throughout the rest of the chapter we will use the abbreviation $f[t]$ for $f(x(t),u(t))$ and likewise for other functions with time dependent arguments. Moreover, for an index set I and a vector c with components c_i we define $c_I = (c_i)_{i\in I}$.

Let (x_*,u_*) be a (weak) local minimum of OCP and, in addition to the smoothness assumptions made above, let the following assumptions be satisfied at (x_*,u_*):

(i) (Linear independence) There exist $\alpha > 0$ and $\beta > 0$ such that

$$\|c'_{I_\alpha(t),u}[t]^\top \zeta\| \geq \beta\|\zeta\|$$

for all ζ of appropriate dimension. Herein, the index set I_α is defined by

$$I_\alpha(t) = \{i \in \{1,\dots,n_c\} \mid c_i[t] \geq -\alpha\}.$$

(ii) (Controllability) For every $q \in \mathbb{R}^{n_\psi}$ there exists a solution (x,u,ρ) of the linear system

$$\begin{aligned} x'(t) - f'_x[t]x(t) - f'_u[t]u(t) &= 0, \\ \psi'_{x_0}x(0) + \psi'_{x_1}x(1) &= q, \\ c'_x[t]x(t) + c'_u[t]u(t) + S_\alpha(t)\rho(t) &= 0, \end{aligned}$$

where $S_\alpha(t) = \operatorname{diag}(c_{i,\alpha}(t))$ and $c_{i,\alpha}(t) = \min\{c_i[t]+\alpha, 0\}$.

Under these assumptions, Malanowski [90, Thm. 4.3, p. 86] shows the regularity of the Lagrange multipliers associated with OCP. In particular, the multiplier l_0 associated with the objective function can be normalized to one and the linear operator defined by the linear system in (ii) is surjective under the assumptions (i) and (ii) in

Malanowski [90, Lem. 4.1]. Under assumptions (i) and (ii) there exist Lagrange multipliers $\lambda_* \in W^{1,\infty}([0,1],\mathbb{R}^{n_x})$, $\eta_* \in L^\infty([0,1],\mathbb{R}^{n_c})$, and $\sigma_* \in \mathbb{R}^{n_\psi}$ with

$$\begin{cases} x_*'(t) - f(x_*(t),u_*(t)) & = 0 \\ \lambda_*'(t) + H_x'(x_*(t),u_*(t),\lambda_*(t),\eta_*(t))^\top & = 0 \\ \psi(x_*(0),x_*(1)) & = 0 \\ \lambda_*(0) + \psi_{x_0}'(x_*(0),x_*(1))^\top \sigma_* & = 0 \\ \lambda_*(1) - \psi_{x_1}'(x_*(0),x_*(1))^\top \sigma_* & = 0 \\ H_u'(x_*(t),u_*(t),\lambda_*(t),\eta_*(t))^\top & = 0. \end{cases} \tag{5.3}$$

Furthermore, the complementarity conditions hold a.e. in $[0,1]$:

$$\eta_*(t) \geq 0, \qquad c(x_*(t),u_*(t)) \leq 0, \qquad \eta_*(t)^\top c(x_*(t),u_*(t)) = 0. \tag{5.4}$$

The convex and locally Lipschitz continuous Fischer-Burmeister function [48, 49] $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$\varphi(a,b) = \sqrt{a^2+b^2} - a - b. \tag{5.5}$$

The Fischer-Burmeister function has the nice property that $\varphi(a,b) = 0$ holds if and only if $a,b \geq 0$ and $ab = 0$. Hence, the complementarity conditions (5.4) are equivalent with the equality

$$\varphi(-c_i(x_*(t),u_*(t)),\eta_{i*}(t)) = 0, \qquad i = 1,\ldots,n_c,$$

that has to hold almost everywhere in $[0,1]$. Rather than working with the derivative of φ, which does not exist at the origin, we will work with Clarke's generalized Jacobian matrix [30] of φ:

$$\partial\varphi(a,b) \quad = \quad \begin{cases} \left\{ \left(\dfrac{a}{\sqrt{a^2+b^2}} - 1, \dfrac{b}{\sqrt{a^2+b^2}} - 1 \right) \right\}, & \text{if } (a,b) \neq (0,0), \\ \left\{ (s,r) \in \mathbb{R}^2 \;\middle|\; (s+1)^2 + (r+1)^2 \leq 1 \right\}, & \text{if } (a,b) = (0,0). \end{cases}$$

Notice, that $\partial\varphi(a,b)$ is a non-empty, convex and compact set. For $1 \leq q \leq \infty$ let the

Banach spaces

$$\begin{aligned} Z_q &= W^{1,q}([0,1],\mathbb{R}^{n_x}) \times L^q([0,1],\mathbb{R}^{n_u}) \times W^{1,q}([0,1],\mathbb{R}^{n_x}) \times L^q([0,1],\mathbb{R}^{n_c}) \times \mathbb{R}^{n_\psi}, \\ Y_{1,q} &= L^q([0,1],\mathbb{R}^{n_x}) \times L^q([0,1],\mathbb{R}^{n_x}) \times \mathbb{R}^{n_\psi} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times L^q([0,1],\mathbb{R}^{n_u}), \\ Y_{2,q} &= L^q([0,1],\mathbb{R}^{n_c}) \end{aligned}$$

be equipped with the maximum norm for product spaces and $z_* = (x_*, u_*, \lambda_*, \eta_*, \sigma_*)$. Then, the necessary conditions (5.3)-(5.4) are equivalent with the nonlinear equation

$$F(z_*) = \begin{pmatrix} F_1(z_*) \\ F_2(z_*) \end{pmatrix} = 0, \tag{5.6}$$

where $F_1 : Z_\infty \to Y_{1,q}$ and $F_2 : Z_\infty \to Y_{2,q}$ denote the smooth and the nonsmooth part of $F : Z_\infty \to Y_q = Y_{1,q} \times Y_{2,q}$ with $1 \le q \le \infty$, respectively:

$$F_1(z)(\cdot) = \begin{pmatrix} x'(\cdot) - f(x(\cdot),u(\cdot)) \\ \lambda'(\cdot) + H'_x(x(\cdot),u(\cdot),\lambda(\cdot),\eta(\cdot))^\top \\ \psi(x(0),x(1)) \\ \lambda(0) + \psi'_{x_0}(x(0),x(1))^\top \sigma \\ \lambda(1) - \psi'_{x_1}(x(0),x(1))^\top \sigma \\ H'_u(x(\cdot),u(\cdot),\lambda(\cdot),\eta(\cdot))^\top \end{pmatrix}, \quad F_2(z)(\cdot) = \omega(z(\cdot)), \tag{5.7}$$

where $\omega = (\omega_1, \ldots, \omega_{n_c})^\top : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_c} \times \mathbb{R}^{n_\psi} \to \mathbb{R}^{n_c}$ and

$$\omega_i(\bar{x},\bar{u},\bar{\lambda},\bar{\eta},\bar{\sigma}) := \varphi(-c_i(\bar{x},\bar{u}),\bar{\eta}_i), \qquad i = 1,\ldots,n_c. \tag{5.8}$$

For technical reasons, which become apparent later, we consider F as a mapping from Z_∞ into Y_q. However, we note that

$$\mathrm{im}(F) \subseteq Y_\infty \subset Y_q \quad \text{for every } 1 \le q < \infty.$$

5.2 Inexact Nonsmooth Newton Method

By (5.8), the derivative $F'(z^k)$ does not exist since the component F_2 in (5.6) is not differentiable. Hence, we need to find a substitute for the derivative F'. In finite dimensional spaces, such a substitute for locally Lipschitz continuous functions may be chosen from the generalized Jacobian matrix of F defined by

$$\partial F(z) = \operatorname{co}\left\{ V \;\middle|\; V = \lim_{\substack{z^i \in D_F \\ z^i \to z}} F'(z^i) \right\},$$

where "co" is the shorthand for "convex hull of", D_F denotes the set of points at which F is differentiable. However, in infinite dimensional spaces it is more difficult to define an appropriate generalized Jacobian matrix since locally Lipschitz continuous functions in general are not differentiable almost everywhere. Motivated by the chain rule in finite dimensions we define the point to set mapping $\partial_* F : Z_\infty \rightrightarrows \mathscr{L}(Z_q, Y_q)$ for some $1 \le q \le \infty$ according to

$$\partial_* F(z^k)(z) = \left\{ \begin{pmatrix} F_1'(z^k)(z) \\ -S(\cdot)\left(c_x'[\cdot]x + c_u'[\cdot]u\right) + R(\cdot)\eta \end{pmatrix} \;\middle|\; \begin{array}{l} S = \operatorname{diag}(s_1, \ldots, s_{n_c}), \\ R = \operatorname{diag}(r_1, \ldots, r_{n_c}), \\ (s_i, r_i) \in \partial\varphi[\cdot] \text{ a.e.}, \\ s_i(\cdot), r_i(\cdot) \text{ measurable} \end{array} \right\}$$

and use this set as a generalized Jacobian matrix. The same idea was introduced earlier in Ulbrich [131, Def. 3.35, p. 47]. Notice that the first component F_1 of F in (5.7) is continuously Fréchet-differentiable as a mapping from Z_∞ to Y_∞ with

$$F_1'(z^k)(z) = \begin{pmatrix} x'(\cdot) - f_x'[\cdot]x(\cdot) - f_u'[\cdot]u(\cdot) \\ \lambda'(\cdot) + H_{xx}''[\cdot]x(\cdot) + H_{xu}''[\cdot]u(\cdot) + H_{x\lambda}''[\cdot]\lambda(\cdot) + H_{x\eta}''[\cdot]\eta(\cdot) \\ \psi_{x_0}'x(0) + \psi_{x_1}'x(1) \\ \lambda(0) + \psi_{x_0x_0}''(\sigma^k, x(0)) + \psi_{x_0x_1}''(\sigma^k, x(1)) + \left(\psi_{x_0}'\right)^\top \sigma \\ \lambda(1) - \psi_{x_1x_0}''(\sigma^k, x(0)) - \psi_{x_1x_1}''(\sigma^k, x(1)) - \left(\psi_{x_1}'\right)^\top \sigma \\ H_{ux}''[\cdot]x(\cdot) + H_{uu}''[\cdot]u(\cdot) + H_{u\lambda}''[\cdot]\lambda(\cdot) + H_{u\eta}''[\cdot]\eta(\cdot) \end{pmatrix}, \tag{5.9}$$

provided that the functions f_0, f, c, ψ are C^2 w.r.t. all arguments. Herein, all functions are evaluated at $z^k = (x^k, u^k, \lambda^k, \eta^k, \sigma^k) \in Z_\infty$. The definition of Fréchet differentiability then implies that F_1 is as well Fréchet differentiable as a mapping from Z_∞ to Y_q for every $1 \le q \le \infty$.

Ulbrich [131, Def. 3.1, p. 34.] gave a definition about semismoothness and p-order semismoothness of F in Banach space.

Definition 5.1. *Let $F : \tilde{Z} \subset Z \to Y$ be defined on an open subset $\tilde{Z}$ of the Banach space Z with images in the Banach space Y. Further, let be given a set-valued mapping $\partial_* F : \tilde{Z} \Rightarrow L(Z,Y)$, and let $z \in \tilde{Z}$.*

(a) *We say that F is $\partial_* F$-semismooth at z if F is continuous near z and*

$$\|F(z) - F(z_*) - V(z - z_*)\|_Y = o(\|z - z_*\|_Z), \qquad \forall V \in \partial_* F(z),$$

as $\|z - z_*\|_Z \to 0$,

(b) *We say that F is p-order $\partial_* F$-semismooth at z, $0 < p \le 1$, if F is continuous near z and*

$$\|F(z) - F(z_*) - V(z - z_*)\|_Y = O(\|z - z_*\|_Z^{1+p}), \qquad \forall V \in \partial_* F(z), \tag{5.10}$$

as $\|z - z_*\|_Z \to 0$.

By the above analysis on the generalized Jacobian matrix $\partial_* F(z^k)$, we can present the following algorithm.

Algorithm 5.2. LOCAL INEXACT NONSMOOTH NEWTON METHOD

(0) Choose $z^0 \in Z_\infty$.

(1) If some stopping criterion is satisfied, stop.

(2) Choose an arbitrary $V^k \in \partial_* F(z^k)$ and compute the search direction d^k from the linear equation

$$V^k(d^k) + F(z^k) = r^k, \tag{5.11}$$

where $\max\{\|r^k\|_{Y_\infty}, \|r^k\|_{Y_q}\} \le \rho^k \|F(z^k)\|_{Y_q}$, $0 \le \rho^k \le \bar{\rho} < 1$.

(3) Set $z^{k+1} = S^k(z^k + d^k)$, $k = k+1$, and goto (1).

To globalize Algorithm 5.2 conveniently in the next section, we will use $q = 2$. The smoothing operator $S^k : Z_q \to Z_\infty$, see [131], maps $z^k + d^k \in Z_q$ back to Z_∞. The smoothing operator S^k in Step (3) is necessary if r^k is in Y_q, but not in Y_∞. As we shall argue later, the smoothing step can be omitted in certain situations.

The assumptions needed to prove local convergence of the method are similar to those in Martínez and Qi [94], Kanzow [74], Facchinei and Kanzow [45], and Ulbrich [131]. $\partial_* F(z)$ is called non-singular if for every $V \in \partial_* F(z)$ the inverse operator V^{-1} exists and if it is linear and bounded, i.e. $V^{-1} \in \mathscr{L}(Y_q, Z_q)$. In fact, it suffices if the non-singularity assumptions are satisfied for certain elements of $\partial_* F$ provided that only these elements are used in the algorithm. For the upcoming computations we used the element corresponding to the choices

$$s_i(t) = \begin{cases} -1, & \text{if } c_i[t] = 0,\ \eta_i(t) = 0, \\ \dfrac{-c_i[t]}{\sqrt{c_i[t]^2 + \eta_i(t)^2}} - 1, & \text{otherwise,} \end{cases}$$

$$r_i(t) = \begin{cases} 0, & \text{if } c_i[t] = 0,\ \eta_i(t) = 0, \\ \dfrac{\eta_i(t)}{\sqrt{c_i[t]^2 + \eta_i(t)^2}} - 1, & \text{otherwise.} \end{cases}$$

Theorem 5.3. *Let z_* be a zero of F. Suppose that there exist constants $\Delta > 0$ and $C > 0$ such that for every $\|z - z_*\|_{Z_\infty} < \Delta$ the generalized Jacobian matrix $\partial_* F(z)$ is non-singular and $\|V^{-1}\|_{\mathscr{L}(Y_q, Z_q)} \le C$ for every $V \in \partial_* F(z)$. Moreover, let there exist a constant $C_S > 0$ such that*

$$\|S^k(z^k + d^k) - z_*\|_{Z_\infty} \le C_S \|z^k + d^k - z_*\|_{Z_q}$$

for all k. Let $\rho^k = O(\|F(z^k)\|_{Y_q}^{\tilde{q}})$ for some $\tilde{q} > 0$. Then the following assertions hold.

(i) *If F for some $1 \le q \le \infty$ is $\partial_* F$-semismooth at z_*, then for z^0 sufficiently close to z_* the inexact nonsmooth Newton method converges superlinearly to z_*.*

(ii) *If F for some $1 \le q \le \infty$ is p-order $\partial_* F$-semismooth at z, then for z^0 sufficiently close to z_* the inexact nonsmooth Newton method converges at order $1 + \min\{p, \tilde{q}\}$ to z_*.*

Furthermore, if $F(z^k) \neq 0$ for all k and if there is a constant $\widetilde{C}_S$ with $\|S^k(z^k+d^k) - z^k\|_{Z_\infty} \leq \widetilde{C}_S \|d^k\|_{Z_q}$ then the residual values converge superlinearly:

$$\lim_{k\to\infty} \frac{\|F(z^{k+1})\|_{Y_q}}{\|F(z^k)\|_{Y_q}} = 0. \tag{5.12}$$

Proof. By the assumptions in theorem, the algorithm is well-defined in some neighborhood of z_*. It holds

$$\begin{aligned} V^k(z^k+d^k-z_*) &= V^k(z^k-z_*) + V^k d^k \\ &= V^k(z^k-z_*) - F(z^k) + F(z_*) + V^k d^k + F(z^k). \end{aligned}$$

Since $F : Z_\infty \to Y_q$ is locally Lipschitzian, there exist constants L and $\delta > 0$ such that if $\|z - z^*\|_{Z_\infty} \leq \delta$, then

$$\|F(z)\|_{Y_q} = \|F(z) - F(z_*)\|_{Y_q} \leq L\|z - z_*\|_{Z_\infty}. \tag{5.13}$$

The assertions in (i) and (ii) follow from

$$\begin{aligned} \|z^{k+1} - z_*\|_{Z_\infty} &= \|S^k(z^k+d^k) - z_*\|_{Z_\infty} \\ &\leq C_S \cdot \|z^k + d^k - z_*\|_{Z_q} \\ &= C_S \cdot \|(V^k)^{-1}\left(V^k(z^k-z_*) - F(z^k) + F(z_*) + V^k d^k + F(z^k)\right)\|_{Z_q} \\ &\leq C_S \cdot \|(V^k)^{-1}\|_{\mathscr{L}(Y_q,Z_q)} \cdot \left(\|F(z^k) - F(z_*) - V^k(z^k-z_*)\|_{Y_q} + \|r^k\|_{Y_q}\right) \\ &\leq C_S \cdot C \cdot \left(\|F(z^k) - F(z_*) - V^k(z^k-z_*)\|_{Y_q} + \|F(z^k)\|_{Y_q} \cdot O\left(\|F(z^k)\|_{Y_q}^{\tilde{q}}\right)\right) \\ &= C_S \cdot C \cdot \left(\|F(z^k) - F(z_*) - V^k(z^k-z_*)\|_{Y_q} + O\left(\|F(z^k)\|_{Y_q}^{1+\tilde{q}}\right)\right) \\ &= \begin{cases} o(\|z^k - z_*\|_{Z_\infty}), & \text{in case (i)}, \\ O\left(\|z^k - z_*\|_{Z_\infty}^{1+\min\{p,\tilde{q}\}}\right), & \text{in case (ii)}. \end{cases} \end{aligned} \tag{5.14}$$

Herein, we exploited the local Lipschitz continuity of F in (5.13).

Let $\varepsilon > 0$ be arbitrary. According to Equation (5.14) there exists $\delta > 0$ with

$$\|z^{k+1} - z_*\|_{Z_\infty} \leq \varepsilon \|z^k - z_*\|_{Z_\infty} \quad \text{whenever} \quad \|z^k - z_*\|_{Z_\infty} \leq \delta. \tag{5.15}$$

Notice that for any $\delta > 0$ there exists some $k_0(\delta)$ such that $\|z^k - z_*\| \leq \delta$ for every $k \geq k_0(\delta)$ since z^k converges to z_*.

By the local Lipschitz continuity of F we get

$$\|F(z^{k+1})\|_{Y_q} = \|F(z^{k+1}) - F(z_*)\|_{Y_q} \leq L\|z^{k+1} - z_*\|_{Z_\infty} \leq L\varepsilon\|z^k - z_*\|_{Z_\infty}$$

locally around z_* and the inexact Newton iteration implies

$$\begin{aligned} \|z^{k+1} - z^k\|_{Z_\infty} &\leq \widetilde{C}_S \cdot \|(V^k)^{-1}\|_{\mathscr{L}(Y_q,Z_q)} \cdot (\|F(z^k)\|_{Y_q} + \|r^k\|_{Y_q}) \\ &\leq \widetilde{C}_S \cdot C \cdot \left(\|F(z^k)\|_{Y_q} + O\left(\|F(z^k)\|_{Y_q}^{1+\tilde{q}}\right)\right). \end{aligned}$$

Thus,

$$\begin{aligned} \|z^k - z_*\|_{Z_\infty} &\leq \|z^{k+1} - z^k\|_{Z_\infty} + \|z^{k+1} - z_*\|_{Z_\infty} \\ &\leq \widetilde{C}_S \cdot C \cdot \left(\|F(z^k)\|_{Y_q} + O\left(\|F(z^k)\|_{Y_q}^{1+\tilde{q}}\right)\right) + \|z^{k+1} - z_*\|_{Z_\infty} \\ &\leq \widetilde{C}_S \cdot C \cdot \left(\|F(z^k)\|_{Y_q} + O\left(\|F(z^k)\|_{Y_q}^{1+\tilde{q}}\right)\right) + \varepsilon\|z^k - z_*\|_{Z_\infty} \end{aligned}$$

and

$$\|z^k - z_*\|_{Z_\infty} \leq \frac{\widetilde{C}_S C}{1-\varepsilon}\left(\|F(z^k)\|_{Y_q} + O\left(\|F(z^k)\|_{Y_q}^{1+\tilde{q}}\right)\right). \tag{5.16}$$

Finally,

$$\|F(z^{k+1})\|_{Y_q} \leq L\varepsilon\|z^k - z_*\|_{Z_\infty} \leq \frac{L\varepsilon\widetilde{C}_S C}{1-\varepsilon}\left(\|F(z^k)\|_{Y_q} + O\left(\|F(z^k)\|_{Y_q}^{1+\tilde{q}}\right)\right). \tag{5.17}$$

Since $F(z^k) \neq 0$ and ε may be arbitrarily small, (5.12) holds. ■

It is straightforward to show that the first component F_1 is continuously Fréchet-differentiable and that (5.10) with $p = 1$ holds for F_1, see Gerdts [56].

The second component $F_2(z)(t) = \omega(z(t))$ of F in (5.7) is a superposition operator as in Ulbrich [131, Sec.3.3] that maps L^∞-functions to L^q-functions. It was shown in Ulbrich [131, Thms. 3.44,3.48] that the superposition operator F_2 is semismooth as a mapping from Z_∞ to $Y_{2,q}$ for every $1 \leq q < \infty$, if the following assumptions are satisfied:

- The operator $G : Z_\infty \to Y_{2,q}$, $1 \leq q < \infty$, defined by $G(z)(\cdot) = (c(x(\cdot), u(\cdot)), \eta(\cdot))$ is continuously Fréchet differentiable.

- The mapping $z \in Z_\infty \mapsto G(z) \in Y_{2,\infty}$ is locally Lipschitz continuous.
- φ is Lipschitz continuous and semismooth.

Please note that $q = \infty$ is excluded. Note that the three conditions above are satisfied owing to the following reasons. The Fischer-Burmeister function $\varphi : \mathbb{R}^2 \to \mathbb{R}$ is Lipschitz continuous and semismooth, see Fischer [49, Lem. 20]. The mapping $z \in Z_\infty \mapsto G(z) \in Y_{2,\infty}$ is continuously Fréchet differentiable (and thus locally Lipschitz continuous), if c is continuously differentiable. This implies that the operator G as a mapping from Z_∞ to $Y_{2,q}$ for every $1 \le q < \infty$ is continuously Fréchet differentiable. Hence, the operator F_2 is semismooth as an operator from Z_∞ to $Y_{2,q}$ with $1 \le q < \infty$. Summarizing, we obtain the following local convergence result.

Theorem 5.4**.** *Let z_* be a zero of F and let $1 \le q < \infty$. Suppose that there exist constants $\Delta > 0$ and $C > 0$ such that for every $\|z - z_*\|_{Z_\infty} < \Delta$ the generalized Jacobian $\partial_* F(z)$ is non-singular and $\|V^{-1}\|_{\mathscr{L}(Y_q, Z_q)} \le C$ for every $V \in \partial_* F(z)$. Moreover, let there exist a constant $C_S > 0$ such that*

$$\|S^k(z^k + d^k) - z_*\|_{Z_\infty} \le C_S \|z^k + d^k - z_*\|_{Z_q}$$

for all k. Let $\rho^k = O(\|F(z^k)\|_{Y_q}^{\tilde{q}})$ for some $\tilde{q} > 0$. Then the inexact nonsmooth Newton method converges locally at a superlinear rate, if f_0, f, c, ψ are C^2.

Furthermore, if $F(z^k) \neq 0$ for all k and if there is a constant $\widetilde{C}_S$ with $\|S^k(z^k + d^k) - z^k\|_{Z_\infty} \le \widetilde{C}_S \|d^k\|_{Z_q}$ then the residual values converge superlinearly:

$$\lim_{k \to \infty} \frac{\|F(z^{k+1})\|_{Y_q}}{\|F(z^k)\|_{Y_q}} = 0.$$

In the sequel, we give some preliminary discussions on $V \in \partial_* F(z^k)$ in equation (5.11) of Algorithm 5.2. In fact, the linear operator equation (5.11) in step (2) of Algo-

rithm 5.2 can be stated as

$$\begin{aligned}\begin{pmatrix} x' \\ \lambda' \end{pmatrix} - \begin{pmatrix} f'_x & 0 \\ -H''_{xx} & -H''_{x\lambda} \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} - \begin{pmatrix} f'_u & 0 \\ -H''_{xu} & -H''_{x\eta} \end{pmatrix} \begin{pmatrix} u \\ \eta \end{pmatrix} \\ = - \begin{pmatrix} (x^k)' - f \\ (\lambda^k)' + \left(H'_x\right)^\top \end{pmatrix} + r_1^k;\end{aligned} \tag{5.18}$$

$$\begin{aligned}&\begin{pmatrix} \psi'_{x_0} & 0 & 0 \\ (\psi'_{x_0}{}^\top \sigma^k)'_{x_0} & I & \psi'_{x_0}{}^\top \\ -(\psi'_{x_1}{}^\top \sigma^k)'_{x_0} & 0 & -\psi'_{x_1}{}^\top \end{pmatrix} \begin{pmatrix} x(0) \\ \lambda(0) \\ \sigma \end{pmatrix} + \begin{pmatrix} \psi'_{x_1} & 0 & 0 \\ (\psi'_{x_0}{}^\top \sigma^k)'_{x_1} & 0 & 0 \\ -(\psi'_{x_1}{}^\top \sigma^k)'_{x_1} & I & 0 \end{pmatrix} \begin{pmatrix} x(1) \\ \lambda(1) \\ \sigma \end{pmatrix} \\ &\qquad = - \begin{pmatrix} \psi(x^k(0), x^k(1)) \\ \lambda^k(0) + \psi'_{x_0}{}^\top \sigma^k \\ \lambda^k(1) - \psi'_{x_1}{}^\top \sigma^k \end{pmatrix} + r_2^k;\end{aligned} \tag{5.19}$$

and

$$\mathscr{A} \begin{pmatrix} u \\ \eta \end{pmatrix} + \begin{pmatrix} H''_{ux} & H''_{u\lambda} \\ -Sc'_x & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} = - \begin{pmatrix} (H'_u)^\top \\ \omega(z^k(\cdot)) \end{pmatrix} + r_3^k, \tag{5.20}$$

where

$$\mathscr{A} = \begin{pmatrix} H''_{uu} & \left(c'_u\right)^\top \\ -Sc'_u & R \end{pmatrix}. \tag{5.21}$$

Herein, every function is evaluated at the current iterate z^k. If the inverse operator $\mathscr{A}^{-1}$ exists, equation (5.20) can be solved for u and η according to

$$\begin{pmatrix} u \\ \eta \end{pmatrix} = -\mathscr{A}^{-1} \left[\begin{pmatrix} H''_{ux} & H''_{u\lambda} \\ -Sc'_x & 0 \end{pmatrix} \begin{pmatrix} x \\ \lambda \end{pmatrix} + \begin{pmatrix} (H'_u)^\top \\ \omega(z^k(\cdot)) \end{pmatrix} - r_3^k \right]. \tag{5.22}$$

A sufficient condition for the non-singularity of $\mathscr{A}$ is given below in Theorem 5.5. The constant σ in (5.19) can be viewed as a solution of the differential equation $\sigma' = 0$.

Introducing (5.22) into the differential equation (5.18), augmenting this system by $\sigma' = 0$, and taking into account the boundary conditions (5.19), yields the linear boundary value problem for $\xi = (x, \lambda, \sigma)^\top$:

$$\begin{cases} \xi' = B\xi + b, \\ E_0\xi(0) + E_1\xi(1) = \widehat{q}, \end{cases} \tag{5.23}$$

where

$$B = \begin{pmatrix} f'_x & 0 & 0 \\ -H''_{xx} & -H''_{x\lambda} & 0 \\ 0 & 0 & 0 \end{pmatrix} - \begin{pmatrix} f'_u & 0 \\ -H''_{xu} & -H''_{x\eta} \\ 0 & 0 \end{pmatrix} \mathscr{A}^{-1} \begin{pmatrix} H''_{ux} & H''_{u\lambda} & 0 \\ -Sc'_x & 0 & 0 \end{pmatrix},$$

$$b = -\left[\begin{pmatrix} (x^k)' - f \\ (\lambda^k)' + H_x'^\top \\ 0 \end{pmatrix} - r_1^k + \begin{pmatrix} f'_u & 0 \\ -H''_{xu} & -H''_{x\eta} \\ 0 & 0 \end{pmatrix} \mathscr{A}^{-1} \left(\begin{pmatrix} (H'_u)^\top \\ \omega(z^k(\cdot)) \end{pmatrix} - r_3^k \right) \right],$$

$$E_0 = \begin{pmatrix} \psi'_{x_0} & 0 & 0 \\ (\psi'^\top_{x_0}\sigma^k)'_{x_0} & I & \psi'^\top_{x_0} \\ -(\psi'^\top_{x_1}\sigma^k)'_{x_0} & 0 & -\psi'^\top_{x_1} \end{pmatrix},$$

$$E_1 = \begin{pmatrix} \psi'_{x_1} & 0 & 0 \\ (\psi'^\top_{x_0}\sigma^k)'_{x_1} & 0 & 0 \\ -(\psi'^\top_{x_1}\sigma^k)'_{x_1} & I & 0 \end{pmatrix},$$

$$\widehat{q} = -\begin{pmatrix} \psi(x^k(0), x^k(1)) \\ \lambda^k(0) + \psi'^\top_{x_0}\sigma^k \\ \lambda^k(1) - \psi'^\top_{x_1}\sigma^k \end{pmatrix} + r_2^k.$$

Hence, in each iteration of Algorithm 5.2 we have to solve the linear boundary value problem (5.23). Gerdts [56, Thm. 3.2] gives a sufficient condition for the existence and boundedness of the inverse operator of $\mathscr{A}$ in (5.21).

Proposition 5.5. *Let $z = (x,u,\lambda,\eta,\sigma) \in Z_\infty$ be given. Define the index sets*

$$\begin{aligned} I_>(t) &= \{i \in \{1,\dots,n_c\} \mid c_i[t] = 0,\ \eta_i(t) > 0\}, \\ J_\gamma(t) &= \{i \in \{1,\dots,n_c\} \mid |c_i[t]| \le \gamma\eta_i(t),\ \eta_i(t) \ge 0\}, \qquad \gamma > 0. \end{aligned}$$

Let the following assumptions hold at z:

(i) *Let there exist constants C_1, C_2, C_3 such that a.e. in $[0,1]$ it holds*

$$\|H''_{uu}[t]\| \le C_1, \quad \|c'_u[t]^\top\| \le C_2, \quad \|c'_u[t]\| \le C_3.$$

(ii) *(Coercivity) Let there exist a constant $\alpha > 0$ such that a.e. in $[0,1]$ it holds*

$$d^\top H''_{uu}[t]d \ge \alpha\|d\|^2 \quad \text{for all} \quad d \in \{d \in \mathbb{R}^{n_u} \mid c'_{I_>(t),u}[t]d = 0\}.$$

(iii) *(Linear independence) Let there exist constants $\gamma > 0$ and $\beta > 0$ such that a.e. in $[0,1]$ it holds*

$$\|c'_{J_\gamma(t),u}[t]^\top \zeta\| \ge \beta\|\zeta\| \quad \text{for all } \zeta \text{ of appropriate dimension.}$$

Then, a.e. in $[0,1]$ the inverse operator $\mathscr{A}^{-1}(t)$ exists and it holds $\|\mathscr{A}^{-1}(t)\| \le C$ for some constant C.

For $1 \le q \le \infty$ define from the boundary value problem (5.23) the linear operator $G : W^{1,q}([0,1],\mathbb{R}^{2n_x+n_\psi}) \to L^q([0,1],\mathbb{R}^{2n_x+n_\psi}) \times \mathbb{R}^{2n_x+n_\psi} = \Omega$ by

$$G(\xi)(t) = \begin{pmatrix} \xi'(t) - B(t)\xi(t) \\ E_0\xi(0) + E_1\xi(1) \end{pmatrix},$$

where $\|(\omega_1,\omega_2)\|_\Omega = \max\{\|\omega_1\|_q, \|\omega_2\|\}$. Similar as in Gerdts [56, Thm. 3.3] one can obtain the following non-singularity and boundedness result for the inverse of the operator G.

Proposition 5.6. *Let the following assumptions be satisfied.*

(i) *Let there exist a constant C such that a.e. in $[0,1]$ it holds $\|B(t)\| \leq C$.*

(ii) *Let there exist $\kappa > 0$ such that for all $\zeta \in \mathbb{R}^{2n_x+n_\psi}$ it holds*

$$\| \left(E_0\Phi(0) + E_1\Phi(1)\right) \zeta \| \geq \kappa \|\zeta\|,$$

where Φ is a fundamental solution with $\Phi'(t) = B(t)\Phi(t)$, $\Phi(0) = I$.

Then, for $1 \leq q \leq \infty$ the inverse operator G^{-1} exists and it holds $\|G^{-1}\| \leq K$ for some constant K.

A combination of Propositions 5.5 and 5.6 leads to the following result.

Theorem 5.7**.** *Let z_* be a zero of F. Suppose that there exists a constant $\Delta > 0$ such that for every $\|z - z_*\|_{Z_\infty} < \Delta$ the assumptions of Propositions 5.5 and 5.6 hold with uniform constants. Then, for every $1 \leq q \leq \infty$ the generalized Jacobian $\partial_* F(z)$ is non-singular and there exists a constant $C > 0$ such that $\|V^{-1}\|_{\mathscr{L}(Y_q, Z_q)} \leq C$ for every $V \in \partial_* F(z)$.*

Remark 5.8**.** *Theorem 5.7 holds for every $1 \leq q \leq \infty$. In particular, this implies that every element $V \in \partial_* F(z)$ maps a function in Y_q to a function in Z_q. In particular, if $F(z^k) \in Y_\infty$ and $r^k \in Y_\infty$ then $d^k = V_k^{-1}\left(r^k - F(z^k)\right) \in Z_\infty$. $F(z^k) \in Y_\infty$ holds, if $z^k \in Z_\infty$. Hence, the smoothing operator S^k in step (3) of Algorithm 5.2 can be chosen to be the identity if the initial z^0 is chosen to be in Z_∞ and if every r^k is in Y_∞.*

5.3 Globalization Strategy

In this section, we globalize the local inexact nonsmooth Newton method using the squared L^2-norm (5.24) of F as a merit function. A favorable property of the merit

function (5.24) is that it is Fréchet-differentiable in Z_∞ if f_0, f, c, ψ are C^2.

$$\begin{aligned}\Theta(z) &= \frac{1}{2}\|F(z)\|_{Y_2}^2 \\ &= \frac{1}{2}\int_0^1 \left\|x'(t) - f(x(t),u(t))\right\|^2 \,\mathrm{d}t \\ &+ \frac{1}{2}\int_0^1 \left\|\lambda'(t) + H_x'(x(t),u(t),\lambda(t),\eta(t))^\top\right\|^2 \,\mathrm{d}t \\ &+ \frac{1}{2}\int_0^1 \left\|H_u'(x(t),u(t),\lambda(t),\eta(t))^\top\right\|^2 \,\mathrm{d}t + \frac{1}{2}\sum_{i=1}^{n_c}\int_0^1 \varphi(-c_i(x(t),u(t)),\eta_i(t))^2 \,\mathrm{d}t \\ &+ \frac{1}{2}\|\psi(x(0),x(1))\|^2 + \frac{1}{2}\|\lambda(0) + \psi_{x_0}'(x(0),x(1))^\top\sigma\|^2 \\ &+ \frac{1}{2}\|\lambda(1) - \psi_{x_1}'(x(0),x(1))^\top\sigma\|^2.\end{aligned} \tag{5.24}$$

From [59, 131, 130], we note that the performance of nonlinear programming algorithms can be significantly improved by using non-monotone linear search or trust-region techniques. Thus, in contrast to the traditional approach, relaxing the acceptability conditions on the trial step d^k in our algorithm, we suggest to use the non-monotone technique:

$$\Theta(z^{l(k)}) = \max_{0\le j\le m(k)}\{\Theta(z^{k-j})\} \tag{5.25}$$

instead of $\Theta(z^k)$, where $m(0) = 0$ and $0 \le m(k) \le \min\{m(k-1)+1, M\}$, $k \ge 1$.

In the following algorithm and thereafter we will make use of the norm $\|\cdot\|_{\widehat{Z}}$ on the space $\widehat{Z} := Z_2$, which is defined in Section 5.6.

Algorithm 5.9. GLOBAL INEXACT NONSMOOTH NEWTON METHOD

(0) Choose $z^0 \in Z_\infty$, $\beta \in (0,1)$, $\kappa > 0$, $\rho > 1$, $\sigma \in (0, 1/4)$, m and M.

(1) If some stopping criterion is satisfied, stop.

(2) Choose an arbitrary $V^k \in \partial_* F(z^k)$ and compute the search direction d^k from (5.11). If (5.11) is not solvable or if the condition

$$\Theta'(z^k)(d^k) \le -\kappa\|d^k\|_{\widehat{Z}}^\rho \tag{5.26}$$

is not satisfied, set $d^k = -W^k F(z^k)$, where the linear operator W^k and the norm $\|\cdot\|_{\widehat{Z}}$ are defined in Section 5.6.

(3) Find smallest $i_k \in \mathbb{N}_0$ with

$$\Theta(S^k(z^k + \beta^{i_k} d^k)) \leq \Theta(z^{l(k)}) + \sigma \beta^{i_k} \Theta'(z^k)(d^k) \tag{5.27}$$

and set $\alpha_k = \beta^{i_k}$, where $l(k)$ is chosen by (5.25).

(4) Set $z^{k+1} = S^k(z^k + \alpha_k d^k)$, $k = k+1$, and goto (1).

It holds

Lemma 5.10. *Suppose that z^k is not a stationary point of (5.24). Then d^k is a descent direction of Θ at z^k and*

$$\Theta'(z^k)(d^k) \leq -\min\{\kappa \|d^k\|_{\widehat{Z}}^{\rho}, \|d^k\|_{\widehat{Z}}^2\} < 0. \tag{5.28}$$

Proof. An analysis of the derivative of Θ reveals that for d^k from (5.11) it holds

$$\begin{aligned}
\Theta'(z^k)(d^k) &= \int_0^1 F(z^k)(t)^\top V^k(d^k)(t)\, \mathrm{d}t \\
&= \int_0^1 F(z^k)(t)^\top (r^k(t) - F(z^k)(t))\, \mathrm{d}t \\
&= \int_0^1 F(z^k)(t)^\top r^k(t)\, \mathrm{d}t - \|F(z^k)\|_{Y_2}^2 \\
&\leq \|F(z^k)\|_{Y_2} \cdot \|r^k\|_{Y_2} - \|F(z^k)\|_{Y_2}^2 \\
&\leq (\rho^k - 1) \cdot \|F(z^k)\|_{Y_2}^2 \\
&\leq (\bar{\rho} - 1) \cdot \|F(z^k)\|_{Y_2}^2 < 0.
\end{aligned} \tag{5.29}$$

As a consequence, d^k is a direction of descent of Θ at z^k.

Alternatively, for the direction $d^k = -W^k F(z^k)$ we find $\Theta'(z^k)(d^k) = -\|d^k\|_{\widehat{Z}}^2 < 0$. This, together with (5.26), implies (5.28), i.e., the line-search in the Algorithm 5.9 is well-defined unless z^k is a stationary point of $\Theta(z)$. ∎

As a consequence, for some $\hat{\sigma} \in (0,1)$ there exists $\alpha > 0$ such that

$$\Theta(z^k + \alpha d^k) \leq \Theta(z^k) + \hat{\sigma} \alpha \Theta'(z^k)(d^k). \tag{5.30}$$

Instead for $z^k + \alpha d^k$ we intend to perform a line-search using the smoothing operator $S^k(z^k + \alpha d^k)$. The following growth condition is sufficient to prove the well-posedness of the line-search procedure.

Assumption 5.11**.** For z^k and $d^k = V_k^{-1}\left(r^k - F(z^k)\right)$ let there be a smoothing operator S^k and constants $0 \leq L < 1$, $\gamma > 0$ with

$$\left|\Theta(S^k(z^k + \alpha d^k)) - \Theta(z^k + \alpha d^k)\right| \leq 2L\alpha^{1+\gamma}\Theta(z^k) \tag{5.31}$$

for all $0 \leq \alpha \leq 1$ and all k.

Remark 5.12**.** If in Algorithm 5.9 the gradient direction $d^k = -W^k F(z^k)$ is chosen, then the smoothing operator S^k is obsolete and can be chosen to be the identity. The smoothing operator only becomes relevant if the Newton direction is applied. For notational simplicity the subsequent analysis will be performed with a smooting step.

Lemma 5.13**.** *Let Assumption 5.11 be satisfied. Then there exists $\alpha > 0$ such that*

$$\Theta(S^k(z^k + \alpha d^k)) \leq \Theta(z^k) + \sigma\alpha\Theta'(z^k)(d^k) = \Theta(z^k)\,(1 - 2\sigma\alpha)$$

for some $\sigma \in (0, 1-L)$. Moreover, it holds $0 < 1 - 2\sigma\alpha < 1$ whenever $\sigma \in (0, \min\{1/2, 1-L\})$.

Proof. Inequality (5.30) together with Assumption 5.11 and exploiting $\alpha^{1+\gamma} \leq \alpha$ for $0 \leq \alpha \leq 1$ implies

$$\Theta(S^k(z^k + \alpha d^k)) \leq \Theta(z^k)\,(1 - 2(\hat{\sigma} - L)\alpha)\,.$$

Define $\hat{\sigma} := \sigma + L \in (0,1)$, i.e. $\sigma \in (-L, 1-L)$. Then

$$\Theta(S^k(z^k + \alpha d^k)) \leq \Theta(z^k)\,(1 - 2\sigma\alpha) = \Theta(z^k) + \sigma\alpha\Theta'(z^k)(d^k).$$

Armijo's rule requires $0 < \sigma < 1$. Since $0 \leq L < 1$ and together with $\sigma \in (-L, 1-L)$ this implies $\sigma \in (0, 1-L)$. Because $0 < \alpha \leq 1$ it holds $0 < 1 - 2\sigma\alpha < 1$ whenever $\sigma \in (0, \min\{1/2, 1-L\})$.

Lemma 5.13 guarantees that the line-search in Algorithm 5.9 is well defined.

Now we give a global convergence conclusion deduced from Gerdts [56, Thm. 4.2], which extends the proof presented in [59] for finite dimensions into infinite dimensions.

Theorem 5.14. *Let z_* be an accumulation point of the sequence $\{z^k\}$ generated by Algorithm 5.9. Let all first and second derivatives of the functions f_0, f, c, ψ be uniformly bounded. Let there be a constant C_F such that $\|F(z^k)\|_{Y_\infty} \leq C_F$ for every k. Let Assumption 5.11 hold. Moreover, let a constant $\widetilde{C}_S$ exist with*

$$\|S^k(z^k + \alpha_k d^k) - z^k\|_{\widehat{Z}} \leq \widetilde{C}_S \alpha_k \|d^k\|_{\widehat{Z}} \tag{5.32}$$

for all k.

Then, z_* is a stationary point of $\Theta(z)$, i.e., $\Theta'(z_*) = 0$ (zero operator). Moreover, if the inverse operators $(V^k)^{-1}$ exist for all k, $C > 0$ is a constant such that $\|(V^k)^{-1}\|_{\mathscr{L}(Y_\infty, Z_\infty)} \leq C$ holds for all k, and (5.26) is satisfied by the Newton direction for all but finitely many k, then z_* is a zero of F.

Proof. Let $\{z^k\}_{k \in K \subset \mathbb{N}}$ be a subsequence with $z^k \to z_*$ and $\Theta'(z^k)(d^k) \neq 0$. Then, it follows from Lemma 5.13 that the line-search of Algorithm 5.9 is well-defined.

Suppose $k \in K \subset \mathbb{N}$, by (5.25) and (5.27), we have

$$\begin{aligned}
\Theta(z^{l(k+1)}) &= \max_{0 \leq j \leq m(k+1)} \{\Theta(z^{k+1-j})\} \\
&\leq \max_{0 \leq j \leq m(k)+1} \{\Theta(z^{k+1-j})\} \\
&= \max\{\Theta(z^{k+1}), \Theta(z^{l(k)})\} = \Theta(z^{l(k)}),
\end{aligned}$$

which implies that the sequence $\{\Theta(z^{l(k)})\}$ is non-increasing and together with the non-negativity of Θ the sequence $\{\Theta(z^{l(k)})\}$ converges. From (5.27) it follows that for $k \geq M$,

$$\Theta(z^{l(k)}) = \Theta(S^k(z^{l(k)-1} + \alpha_{l(k)-1} d^{l(k)-1})) \leq \Theta(z^{l(l(k)-1)}) + \sigma \alpha_{l(k)-1} \Theta'(z^{l(k)-1})(d^{l(k)-1}). \tag{5.33}$$

This, together with convergence of $\{\Theta(z^{l(k)})\}$, yields

$$\lim_{k(\in K) \to \infty} \alpha_{l(k)-1} \Theta'(z^{l(k)-1})(d^{l(k)-1}) = 0. \tag{5.34}$$

By Lemma 5.10, we have

$$\lim_{k(\in K)\to\infty} \alpha_{l(k)-1}\|d^{l(k)-1}\|_{\widehat{Z}} = 0. \tag{5.35}$$

We now prove that

$$\lim_{k(\in K)\to\infty} \alpha_k \Theta'(z^k)(d^k) = 0. \tag{5.36}$$

Let $\widehat{l}(k) = l(k+M+2)$. We first show by induction that for any given $j \geq 1$,

$$\lim_{k(\in K)\to\infty} \alpha_{\widehat{l}(k)-j}\|d^{\widehat{l}(k)-j}\|_{\widehat{Z}} = 0, \tag{5.37}$$

and

$$\lim_{k(\in K)\to\infty} \Theta(z^{\widehat{l}(k)-j}) = \lim_{k(\in K)\to\infty} \Theta(z^{l(k)}). \tag{5.38}$$

If $j = 1$, since $\{\widehat{l}(k)\} \subset \{l(k)\}$, (5.37) follows from (5.34). This in turn implies

$$\begin{aligned} \|z^{\widehat{l}(k)} - z^{\widehat{l}(k)-1}\|_{\widehat{Z}} &= \|S^{\widehat{l}(k)-1}(z^{\widehat{l}(k)-1} + \alpha_{\widehat{l}(k)-1}d^{\widehat{l}(k)-1}) - z^{\widehat{l}(k)-1}\|_{\widehat{Z}} \\ &\leq \widetilde{C}_S\alpha_{\widehat{l}(k)-1}\|d^{\widehat{l}(k)-1}\|_{\widehat{Z}} \to 0. \end{aligned} \tag{5.39}$$

From this we intend to deduce the convergence of the function values in (5.38). Notice that Θ is continuous w.r.t. the norm in Z_∞ but not necessarily w.r.t. the norm $\|\cdot\|_{\widehat{Z}}$ and that (5.39) does not necessarily hold w.r.t. to the norm in Z_∞. Nevertheless, (5.39) implies that for almost every $t \in [0,1]$ $z^{\widehat{l}(k)}(t)$ converges to $z^{\widehat{l}(k)-1}(t)$. As $\|F(z^k)(t)\|_{Y_2} \leq C_Y\|F(z^k)\|_{Y_\infty} \leq C_Y C_F$ for every k this implies that (5.38) holds for $j = 1$ by Kolmogorov and Fomin [86, Thm. 1, p. 56] [2].

Assume now that (5.37) and (5.38) hold for a given j. Then by (5.27) it follows for $k \geq M$,

$$\Theta(z^{\widehat{l}(k)-j}) \leq \Theta(z^{l(\widehat{l}(k)-j-1)}) + \sigma\alpha_{\widehat{l}(k)-j-1}\Theta'(z^{\widehat{l}(k)-j-1})(d^{\widehat{l}(k)-j-1}). \tag{5.40}$$

[2] Alternatively, Fatou's Lemma can be used to avoid the assumption $\|F(z^k)\|_{Y_\infty} \leq C_F$. Then, instead of working with the limit we have to work with the limes inferior.

Taking limits for $k(\in K) \to \infty$, we have, by (5.38)

$$\lim_{k(\in K)\to\infty} \alpha_{\widehat{l}(k)-j-1}\Theta'(z^{\widehat{l}(k)-j-1})(d^{\widehat{l}(k)-j-1}) = 0, \tag{5.41}$$

This, together with Lemma 5.10, yields

$$\lim_{k(\in K)\to\infty} \alpha_{\widehat{l}(k)-j-1}\|d^{\widehat{l}(k)-j-1}\|_{\widehat{Z}} = 0. \tag{5.42}$$

Moreover this implies $\|z^{\widehat{l}(k)-j} - z^{\widehat{l}(k)-j-1}\|_{\widehat{Z}} \to 0$ by exploitation of (5.32). Thus by (5.38) and the same reasoning as above, we have

$$\lim_{k(\in K)\to\infty} \Theta(z^{\widehat{l}(k)-j-1}) = \lim_{k(\in K)\to\infty} \Theta(z^{\widehat{l}(k)-j}) = \lim_{k(\in K)\to\infty} \Theta(z^{l(k)}). \tag{5.43}$$

This completes the induction.

Now for any $k \in K \subset \mathbb{N}$, it holds

$$\|z^{\widehat{l}(k)} - z^{k+1}\|_{\widehat{Z}} \le \sum_{j=0}^{\widehat{l}(k)-k-2} \|z^{\widehat{l}(k)-j} - z^{\widehat{l}(k)-j-1}\|_{\widehat{Z}}. \tag{5.44}$$

By (5.27), we have $\widehat{l}(k) - k - 1 = l(k+M+2) - k - 1 \le M+1$. This, together with (5.44) and (5.37), yields

$$\lim_{k(\in K)\to\infty} \|z^{k+1} - z^{\widehat{l}(k)}\|_{\widehat{Z}} = 0. \tag{5.45}$$

By (5.38) and the above reasoning, this implies

$$\lim_{k(\in K)\to\infty} \Theta(z^{k+1}) = \lim_{k(\in K)\to\infty} \Theta(z^{\widehat{l}(k)}). \tag{5.46}$$

Taking limits in (5.25) for $k(\in K) \to \infty$, it follows from (5.46) that

$$\lim_{k(\in K)\to\infty} \alpha_k \Theta'(z^k)(d^k) = 0. \tag{5.47}$$

By Lemma 5.10 again, we also have

$$\lim_{k(\in K)\to\infty} \alpha_k \|d^k\|_{\widehat{Z}} = 0. \tag{5.48}$$

Next, we consider two cases for (5.47).

Case 1: Assume

$$\alpha = \liminf_{k(\in K)\to\infty} \alpha_k > 0.$$

We need to consider two subcases. Suppose first that $d^k = -W^k F(z^k)$ holds for infinitely many $k \in K \subset \mathbb{N}$. Then it follows from (5.27) that for some infinite subset $K' \subseteq K$

$$\lim_{k(\in K')\to\infty} -\Theta'(z^k)(W^k F(z^k)) = 0. \tag{5.49}$$

Hence z_* is a stationary point of Θ by Lemma 5.21.

On the other hand, if (5.26) and the condition that (5.11) is solvable hold for all but finitely many $k \in K$. Then from (5.29), it follows that

$$\Theta'(z^k)(d^k) \le (\bar{\rho} - 1) \cdot \|F(z^k)\|_{Y_2}^2 \le 0$$

and thus $\lim_{k(\in K)\to\infty} \|F(z^k)\|_{Y_2} = 0$. By the continuity of $\|\cdot\|_{Y_2}$ and F (in Z_∞), z_* is a zero of F.

Case 2: Assume that there is a subsequence $\{z^k\}_{k\in J}$, $J \subseteq K$, with $\lim_{k(\in J)\to\infty} \alpha_k = 0$.

The sequence $\{d^k\}_{k\in J}$ is bounded since

$$\begin{aligned} 0 \le \|d^k\|_{Z_\infty} &\le \max\left\{ \|W^k F(z^k)\|_{Z_\infty}, \|(V^k)^{-1}(F(z^k) - r^k)\|_{Z_\infty} \right\} \\ &\le \max\left\{ \widetilde{C}\|F(z^k)\|_{Y_\infty}, C(1 + C_Y \rho^k)\|F(z^k)\|_{Y_\infty} \right\} \\ &\le \max\left\{ \widetilde{C}\|F(z^0)\|_{Y_\infty}, C(1 + C_Y \bar{\rho})\|F(z^0)\|_{Y_\infty} \right\}. \end{aligned} \tag{5.50}$$

where C_Y is a constant satisfying $\|\cdot\|_{Y_2} \le C_Y \|\cdot\|_{Y_\infty}$ and $\tilde{C}$ is a constant with $\|W^k\|_{\mathcal{L}(Y_\infty, Z_\infty)} \le \tilde{C}$. Note that the linear operator W^k is uniformly bounded as the first and second derivatives of f_0, f, c, ψ are assumed to be uniformly bounded, see also Section 5.6 for a detailed description of W^k.

Since d^k is bounded in Z_∞, it is also bounded in the space $\widehat{Z} = Z_2$, which is a Hilbert space and thus reflexive. From Werner [134, Thm. III.3.7], it follows that there exists a weakly convergent subsequence $\{d^k\}_{k\in\widehat{J}}$, $\widehat{J} \subseteq J$. Hence, by [56, Thm. 4.2], we know

that $\Theta'(z_*)(\cdot)$ can be viewed as elements of $\widehat{Z}^*$ and

$$\Theta'(z_*)(d^k) \to \Theta'(z_*)(d_*).$$

Furthermore, due to the continuity of $\Theta'(\cdot)$ (in Z_∞) for every $\varepsilon > 0$ there exists $\delta > 0$ such that for every $\|z^k - z_*\|_{Z_\infty} \leq \delta$ it holds

$$\begin{aligned}
|\Theta'(z^k)(d^k) - \Theta'(z_*)(d^k)| &= \|d^k\|_{Z_\infty} \left| \Theta'(z^k)\left(\frac{d^k}{\|d^k\|_{Z_\infty}}\right) - \Theta'(z_*)\left(\frac{d^k}{\|d^k\|_{Z_\infty}}\right) \right| \\
&\leq \|d^k\|_{Z_\infty} \cdot \sup_{\|d\|_{Z_\infty}=1} |\Theta'(z^k)(d) - \Theta'(z_*)(d)| \\
&= \|d^k\|_{Z_\infty} \cdot \|\Theta'(z^k) - \Theta'(z_*)\|_{\mathscr{L}(Z_\infty,\mathbb{R})} \leq \varepsilon \|d^k\|_{Z_\infty}.
\end{aligned}$$

For arbitrary $\varepsilon > 0$ we find

$$\begin{aligned}
|\Theta'(z^k)(d^k) - \Theta'(z_*)(d_*)| &\leq |\Theta'(z^k)(d^k) - \Theta'(z_*)(d^k)| \\
&\quad + |\Theta'(z_*)(d^k) - \Theta'(z_*)(d_*)| \\
&\leq \varepsilon \|d^k\|_{Z_\infty} + |\Theta'(z_*)(d^k) - \Theta'(z_*)(d_*)|.
\end{aligned}$$

Since $\varepsilon > 0$ was arbitrary and since d^k is weakly convergent it holds

$$\Theta'(z^k)(d^k) \to \Theta'(z_*)(d_*) \qquad \text{as } k(\in \widehat{J}) \to \infty.$$

Moreover, it holds $\Theta'(z_*)(d_*) \leq 0$ because $\Theta'(z^k)(d^k) < 0$ for every k.

In a similar way the Fréchet differentiability of Θ yields

$$\begin{aligned}
&\left| \frac{1}{\alpha_k} \left(\Theta(S^k(z^k + \alpha_k d^k)) - \Theta(z^k) \right) - \Theta'(z_*)(d_*) \right| \\
&\leq \left| \frac{1}{\alpha_k} \left(\Theta(S^k(z^k + \alpha_k d^k)) - \Theta(z^k) \right) - \Theta'(z^k)(d^k) \right| \\
&\qquad + \left| \Theta'(z^k)(d^k) - \Theta'(z_*)(d_*) \right| \\
&\leq \left| \frac{1}{\alpha_k} \left(\Theta(S^k(z^k + \alpha_k d^k)) - \Theta(z^k + \alpha_k d^k) \right) \right| \\
&\qquad + \left| \frac{1}{\alpha_k} \left(\Theta(z^k + \alpha_k d^k) - \Theta(z^k) \right) - \Theta'(z^k)(d^k) \right|
\end{aligned}$$

$$
\begin{aligned}
&+\left|\Theta'(z^k)(d^k)-\Theta'(z_*)(d_*)\right| \\
&\le 2L\alpha_k^{\gamma}\Theta(z^k)+\frac{1}{\alpha_k}o(\|\alpha_k d^k\|_{Z_\infty})+\left|\Theta'(z^k)(d^k)-\Theta'(z_*)(d^k)\right| \\
&\quad+\left|\Theta'(z_*)(d^k)-\Theta'(z_*)(d_*)\right| \\
&\le 2L\alpha_k^{\gamma}\Theta(z^k)+\|d^k\|_{Z_\infty}\frac{o(\alpha_k\|d^k\|_{Z_\infty})}{\alpha_k\|d^k\|_{Z_\infty}}+\varepsilon\|d^k\|_{Z_\infty} \\
&\quad+\left|\Theta'(z_*)(d^k)-\Theta'(z_*)(d_*)\right|.
\end{aligned}
$$

Since d^k is weakly convergent it holds

$$
\frac{1}{\alpha_k}\left(\Theta(S^k(z^k+\alpha_k d^k))-\Theta(z^k)\right)\to\Theta'(z_*)(d_*)\qquad\text{as } k(\in\widehat{J})\to\infty.
$$

The line search in step (3) of the algorithm and Assumption 5.11 yield

$$
\begin{aligned}
\Theta'(z^k)(d^k)+o(\frac{\alpha_k}{\beta})+2L\left(\frac{\alpha_k}{\beta}\right)^{\gamma}\Theta(z^k) &\ge \frac{\Theta(z^k+\frac{\alpha_k}{\beta}d^k)-\Theta(z^k)+2L\left(\frac{\alpha_k}{\beta}\right)^{1+\gamma}\Theta(z^k)}{\frac{\alpha_k}{\beta}} \\
&\ge \frac{\Theta(S^k(z^k+\frac{\alpha_k}{\beta}d^k))-\Theta(z^k)}{\frac{\alpha_k}{\beta}} \\
&> \frac{\Theta(z^{l(k)})+\sigma\frac{\alpha_k}{\beta}\Theta'(z^k)(d^k)-\Theta(z^k)}{\frac{\alpha_k}{\beta}} \\
&\ge \frac{\Theta(z^k)+\sigma\frac{\alpha_k}{\beta}\Theta'(z^k)(d^k)-\Theta(z^k)}{\frac{\alpha_k}{\beta}} \\
&= \sigma\Theta'(z^k)(d^k).
\end{aligned}
$$

Passing to the limit and exploiting the previous considerations yields

$$
\Theta'(z_*)(d_*)\ge\sigma\Theta'(z_*)(d_*).
$$

Since $\sigma\in(0,1)$ and $\Theta'(z_*)(d_*)\le 0$ the above inequality only holds for $\Theta'(z_*)(d_*)=0$.

Repeating the process in previous case, it follows that either (5.49) for some infinite subset $K'\subset\mathbb{N}$, i.e., z_* is a stationary point of Θ, or

$$
0\ge-(1-\bar{\rho})\cdot\|F(z^k)\|_{Y_2}^2\ge\Theta'(z^k)(d^k)\to\Theta'(z_*)(d_*)=0.
$$

In the latter case, z_* is a zero of F by the continuity of F. Therefore, in either case, we established the conclusion. ■

The above result only shows that each accumulation point is a zero of F, provided that eventually only Newton steps are accepted by the algorithm. Two important questions arise:

- Does the algorithm eventually accept Newton steps (close to a zero of F)?
- Does the global method allow a final locally superlinear convergence?

The locally superlinear convergence would follow from the local convergence Theorem 5.3 if we were able to show that $\alpha_k = 1$ satisfies non-monotonic Armijo's rule for all sufficiently large k.

Theorem 5.15. *Let the assumptions of Theorems 5.4 and 5.14 be valid with $q = 2$. In Algorithm 5.9 let $\sigma \in (0, 1/4)$ and either $\kappa > 0$ sufficiently small and $\rho = 2$ or $\kappa > 0$ and $\rho > 2$ sufficiently large (the magnitude will be given in the proof). Then, for sufficiently large k the step length $\alpha_k = 1$ is accepted and the global method turns into the local one.*

Proof. Owing to $\|(V^k)^{-1}\|_{\mathscr{L}(Y_2,Z_2)} \le C$ and $\|r^k\|_{Y_2} \le \bar{\rho}\|F(z^k)\|_{Y_2}$, the Newton direction d^k in (5.11) satisfies

$$\|d^k\|_{Z_2} \le C\left(\|F(z^k)\|_{Y_2} + \|r^k\|_{Y_2}\right) \le C(1+\bar{\rho})\|F(z^k)\|_{Y_2}.$$

This, together with (5.29), yields that

$$\Theta'(z^k)(d^k) \le (\bar{\rho}-1)\cdot\|F(z^k)\|_{Y_2}^2 \le \frac{\bar{\rho}-1}{C^2(1+\bar{\rho})^2}\|d^k\|_{Z_2}^2, \tag{5.51}$$

which implies that for sufficiently large k, the search direction d^k from (5.11) in Algorithm 5.9 satisfies the condition (5.26) with either $\kappa = \dfrac{1-\bar{\rho}}{C^2(1+\bar{\rho})^2}$ and $\rho = 2$, or with $\kappa > 0$ and $\rho > 2$ sufficiently large.

The superlinear convergence of the residual norms $\|F(z^k)\|_{Y_2}$ was shown in Theorem 5.4, that is, for any $\varepsilon > 0$ and sufficiently large k it holds

$$\|F(S^k(z^k + d^k))\|_{Y_2} \le \varepsilon\|F(z^k)\|_{Y_2}. \tag{5.52}$$

In addition,

$$
\begin{aligned}
|\Theta'(z^k)(d^k)| &= \left| \int_0^1 F(z^k)(t)^\top V^k(d^k)(t)\,\mathrm{d}t \right| \\
&= \left| \int_0^1 F(z^k)(t)^\top (r^k(t) - F(z^k)(t))\,\mathrm{d}t \right| \\
&= \left| \int_0^1 F(z^k)(t)^\top r^k(t)\,\mathrm{d}t - \|F(z^k)\|_{Y_2}^2 \right| \\
&\le \|F(z^k)\|_{Y_2} \cdot \|r^k\|_{Y_2} + \|F(z^k)\|_{Y_2}^2 \\
&\le (\rho^k + 1) \cdot \|F(z^k)\|_{Y_2}^2 \\
&\le (\bar{\rho} + 1) \cdot \|F(z^k)\|_{Y_2}^2.
\end{aligned} \tag{5.53}
$$

This, together with (5.52) and

$$
\varepsilon := \sqrt{1 - 2\sigma(\bar{\rho} + 1)} > 0,
$$

implies that for sufficiently large k

$$
\begin{aligned}
\Theta(z^{l(k)}) + \sigma\Theta'(z^k)(d^k) &\ge \Theta(z^k) + \sigma\Theta'(z^k)(d^k) \\
&= \frac{1}{2}\|F(z^k)\|_{Y_2}^2 + \sigma\Theta'(z^k)(d^k) \\
&\ge \frac{1}{2}\|F(z^k)\|_{Y_2}^2 - \sigma(\bar{\rho} + 1) \cdot \|F(z^k)\|_{Y_2}^2 \\
&= (1 - 2\sigma(\bar{\rho} + 1)) \cdot \frac{1}{2} \cdot \|F(z^k)\|_{Y_2}^2 \\
&\ge \frac{1 - 2\sigma(\bar{\rho} + 1)}{\varepsilon^2} \cdot \frac{1}{2} \cdot \|F(S^k(z^k + d^k))\|_{Y_2}^2 \\
&= \Theta(S^k(z^k + d^k)),
\end{aligned}
$$

i.e. non-monotonic Armijo's line-search accepts $\alpha_k = 1$ and $z^{k+1} = S^k(z^k + d^k)$. ■

5.4 A Smoothing Newton Approach

In this section, we first present a smoothing reformulation for complementarity condition (5.4), and then design a smoothing Newton approach for the reformulations of the OCP. To this end, there exist several smoothing functions available, for example, the

smoothing Fischer-Burmeister function [73] defined by $\varphi : \mathbb{R}_+ \times \mathbb{R}^2 \to \mathbb{R}$:

$$\varphi(\mu, a, b) = \sqrt{a^2 + b^2 + 2\mu} - a - b, \tag{5.54}$$

and the Chen-Hanker-Kanzow-Smale (CHKS) [22, 73, 125] smoothing function defined by $\varphi : \mathbb{R}^3 \to \mathbb{R}$:

$$\varphi(\mu, a, b) = \sqrt{(a-b)^2 + 4\mu^2} - a - b. \tag{5.55}$$

In what follows, without loss of generality, we choose the smoothing Fischer-Burmeister function (5.54) to illustrate our approach. The cases of choosing other smoothing functions can be discussed similarly.

The necessary conditions (5.3)-(5.4) are equivalent to the nonlinear equations

$$F(\mu, z) = \begin{pmatrix} \mu \\ G(\mu, z) \end{pmatrix} = \begin{pmatrix} \mu \\ F_1(z) \\ F_2(\mu, z) \end{pmatrix} = 0, \tag{5.56}$$

where $F_1 : Z_\infty \to Y_{1,q}$ and $F_2 : \mathbb{R} \times Z_\infty \to Y_{2,q}$ denote the smooth and the smoothing part of $F : \mathbb{R} \times Z_\infty \to Y = Y_{1,q} \times Y_{2,q}$, respectively, where

$$F_2(\mu, z)(\cdot) = \omega(z(\cdot)), \tag{5.57}$$

$\omega = (\omega_1, \ldots, \omega_{n_c})^\top : \mathbb{R} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_c} \times \mathbb{R}^{n_\psi} \to \mathbb{R}^{n_c}$ and

$$\omega_i(\mu, \bar{x}, \bar{u}, \bar{\lambda}, \bar{\eta}, \bar{\sigma}) := \varphi(\mu, -c_i(\bar{x}, \bar{u}), \bar{\eta}_i), \qquad i = 1, \ldots, n_c. \tag{5.58}$$

From (5.56), for any $\mu \neq 0$ a straightforward calculation yields

$$F'(\mu, z) = \begin{pmatrix} 1 & 0 \\ G'_\mu(\mu, z) & G'_z(\mu, z) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & F'_1(z) \\ F'_{2\mu}(\mu, z) & F'_{2z}(\mu, z) \end{pmatrix}, \tag{5.59}$$

where $F_1'(z)$ is defined by (5.9), and

$$F'_{2\mu}(\mu,z)(\cdot) = \begin{pmatrix} \dfrac{1}{\sqrt{c_1[\cdot]^2+\eta_1(\cdot)^2+2\mu}} \\ \vdots \\ \dfrac{1}{\sqrt{c_{n_c}[\cdot]^2+\eta_{n_c}(\cdot)^2+2\mu}} \end{pmatrix}, \tag{5.60}$$

$$F'_{2z}(\mu^k,z^k)(\mu,z) = -S(\cdot)\left(c'_x[\cdot]x(\cdot)+c'_u[\cdot]u(\cdot)\right)+R(\cdot)\eta(\cdot), \tag{5.61}$$

where

$$\begin{aligned} S(\cdot) &= \operatorname{diag}\left(\frac{c_i[\cdot]}{\sqrt{c_i[\cdot]^2+\eta_i(\cdot)^2+2\mu}}-1 \;\middle|\; i=1,\ldots,n_c\right), \\ R(\cdot) &= \operatorname{diag}\left(\frac{\eta_i[\cdot]}{\sqrt{c_i[\cdot]^2+\eta_i(\cdot)^2+2\mu}}-1 \;\middle|\; i=1,\ldots,n_c\right). \end{aligned}$$

As shown in previous section, we employ the squared L^2-norm (5.24) of F as a merit function.

$$\begin{aligned} \Xi(\nu) &= \Xi(\mu,z) \qquad (5.62)\\ &= \frac{1}{2}\|F(\mu,z)\|_{Y_2}^2 \\ &= \frac{1}{2}\mu^2+\frac{1}{2}\int_0^1 \left\|x'(t)-f(x(t),u(t))\right\|^2 \,\mathrm{d}t \\ &+\frac{1}{2}\int_0^1 \left\|\lambda'(t)+H'_x(x(t),u(t),\lambda(t),\eta(t))^\top\right\|^2 \,\mathrm{d}t \\ &+\frac{1}{2}\int_0^1 \left\|H'_u(x(t),u(t),\lambda(t),\eta(t))^\top\right\|^2 \,\mathrm{d}t+\frac{1}{2}\sum_{i=1}^{n_c}\int_0^1 \varphi(\mu,-c_i(x(t),u(t)),\eta_i(t))^2 \,\mathrm{d}t \\ &+\frac{1}{2}\|\psi(x(0),x(1))\|^2+\frac{1}{2}\|\lambda(0)+\psi'_{x_0}(x(0),x(1))^\top\sigma\|^2 \\ &+\frac{1}{2}\|\lambda(1)-\psi'_{x_1}(x(0),x(1))^\top\sigma\|^2. \end{aligned}$$

Similar to the previous section, we can find a direction

$$d^k = -\begin{pmatrix} \mathscr{M}^k \\ \mathscr{N}^k \end{pmatrix} = -\begin{pmatrix} \mu^k + W_\mu^k G(\mu^k, z^k) \\ W_z^k G(\mu^k, z^k) \end{pmatrix} = -W^k F(\mu^k, z^k), \tag{5.63}$$

such that $\Xi'(\nu^k)(d^k) = -\|d^k\|_{\hat{Z}}^2$, where the linear operators W^k, W_μ^k, W_z^k are defined via replacing $\Theta(z^k)$ by $\Xi(\nu^k)$ in Section 5.6. We are not going to use this direction in the following algorithm. Instead we will use a modified direction which guarantees that the component μ^k remains positive throughout the iteration, which cannot be guaranteed for the above gradient-like direction without additional safeguards.

Let the signum function $\operatorname{sgn}(\mathscr{M}^k)$ defined by

$$\operatorname{sgn}(\mathscr{M}^k) = \begin{cases} 1, & \text{if } \mathscr{M}^k > 0; \\ 0, & \text{if } \mathscr{M}^k = 0; \\ -1, & \text{if } \mathscr{M}^k < 0. \end{cases} \tag{5.64}$$

In the sequel let $\nu^k = (\mu^k, z^k) \in \mathbb{R} \times Z_\infty$ denote the iterate at iteration k. Moreover, for a constant γ define

$$\zeta(\nu) = \gamma \min\{1, \Xi(\mu, z)\}, \qquad \nu = (\mu, z).$$

Algorithm 5.16. GLOBAL INEXACT SMOOTHING NEWTON METHOD

(0) Choose $z^0 \in Z_\infty$, $\beta \in (0,1)$, $\kappa > 0$, $\rho > 1$, $\sigma \in (0, 1/4)$, m and M. Choose $\bar{\mu} > 0$ and $\gamma \in (0,1)$ such that $\gamma\bar{\mu} < 1$. Let $\bar{\nu} = (\bar{\mu}, 0)$.

(1) If some stopping criterion is satisfied, stop.

(2) Compute the search direction $d^k = (d_\mu^k, d_z^k)$ by

$$F(\nu^k) + F'(\nu^k)d^k - \zeta(\nu^k)\bar{\nu} = r^k, \tag{5.65}$$

where $r^k = (r_\mu^k, r_z^k)$ satisfies $\max\{\|r^k\|_{Y_2}, \|r^k\|_{Y_\infty}\} \le \rho^k \|F(\nu^k)\|_{Y_2}$, $0 \le \rho^k \le \bar{\rho} <$

1, and $r_\mu^k + \zeta(\nu^k)\bar{\mu} > 0$. If (5.65) is not solvable or if the condition

$$\Xi'(\nu^k)(d^k) \le -\kappa \|d^k\|_{\widehat{Z}}^{\rho} \tag{5.66}$$

is not satisfied, set

$$d^k = -\begin{pmatrix} \dfrac{\operatorname{sgn}(\mathscr{M}^k)\min\{1, |\mathscr{M}^k|\}\mu^k}{1+|\mathscr{M}^k|} \\ W_z^k G(\mu^k, z^k) \end{pmatrix}, \tag{5.67}$$

herein $\mathscr{M}^k$ and W_z^k are defined by (5.63), $\operatorname{sgn}(\mathscr{M}^k)$ is defined by (5.64).

(3) Find smallest $i_k \in \mathbb{N}_0$ with

$$\Xi(S^k(\nu^k + \beta^{i_k} d^k)) \le \Xi(\nu^{l(k)}) + \sigma\beta^{i_k}\Xi'(\nu^k)(d^k) \tag{5.68}$$

and set $\alpha_k = \beta^{i_k}$, where

$$\Xi(\nu^{l(k)}) = \max_{0\le j\le m(k)} \{\Xi(\nu^{k-j})\} \tag{5.69}$$

instead of $\Xi(\nu^k)$, herein $m(0) = 0$ and $0 \le m(k) \le \min\{m(k-1)+1, M\}$, $k \ge 1$.

(4) Set $z^{k+1} = S^k(z^k + \alpha_k d^k)$, $k = k+1$, and goto (1).

Before giving the theoretical analysis to Algorithm 5.16, we first introduce a concept of P(artial)-stationary point.

Definition 5.17. *$\nu_* = (\mu_*, z_*) = (0, z_*)$ is called a P-stationary point of Ξ in (5.62), if $W_z G(\nu_*)(\cdot) \equiv 0$.*

If $(0, z_*)$ is P-stationary point of Ξ in (5.62) then $\Theta'(z_*)(\cdot) = 0$ in (5.24). Nevertheless, it follows from (5.63) that a stationary point of Ξ may have no relationship with a solution (or a stationary point) of Θ in (5.24).

The following Lemma shows that Algorithm 5.16 is well-defined.

Lemma 5.18. *Suppose ν^k is neither a P-stationary point, nor a stationary point of Ξ in (5.62) and $\mu^k > 0$. Then $\mu^{k+1} > 0$ and d^k is a descent direction of Ξ at ν^k, i.e., either*

$$\Xi'(\nu^k)(d^k) \leq -\kappa \|d^k\|_{\hat{Z}}^{\rho} < 0, \tag{5.70}$$

or

$$\Xi'(\nu^k)(d^k) = -\frac{\min\{1, |\mathscr{M}^k|\}|\mathscr{M}^k|}{1+|\mathscr{M}^k|}\mu^k - \|W_z^k G(\mu^k, z^k)\|_{\hat{Z}}^2 < 0. \tag{5.71}$$

Proof. If (5.65) is solvable, then

$$\begin{aligned}
\Xi'(\nu^k)(d^k) &= \int_0^1 F(\nu^k)(t)^\top F'(\nu^k)(d^k)(t)\,\mathrm{d}t \\
&= \int_0^1 F(\nu^k)(t)^\top (r^k(t) - F(\nu^k)(t))\,\mathrm{d}t \\
&= \int_0^1 F(\nu^k)(t)^\top r^k(t)\,\mathrm{d}t - \|F(\nu^k)\|_{Y_2}^2 \\
&\leq \|F(\nu^k)\|_{Y_2} \cdot \|r^k\|_{Y_2} - \|F(\nu^k)\|_{Y_2}^2 \\
&\leq (\rho^k - 1) \cdot \|F(\nu^k)\|_{Y_2}^2 \\
&\leq (\bar{\rho} - 1) \cdot \|F(\nu^k)\|_{Y_2}^2 < 0.
\end{aligned} \tag{5.72}$$

and

$$\begin{aligned}
\mu^{k+1} = \mu^k + \alpha d_\mu^k &= \mu^k + \alpha(r_\mu^k + \zeta(\nu^k)\bar{\mu} - \mu^k) \\
&= (1-\alpha)\mu^k + \alpha(r_\mu^k + \zeta(\nu^k)\bar{\mu}) \\
&\geq \min\{\mu^k, r_\mu^k + \zeta(\nu^k)\bar{\mu}\} > 0.
\end{aligned} \tag{5.73}$$

Notice that the first equation in (5.65) always can be solved exactly and thus it is not necessary to consider the case $r_\mu^k \neq 0$. However, for the sake of completeness we leave this case included.

In addition, (5.70) follows readily from (5.66) in Algorithm 5.16.

On the other hand, if (5.67) holds, then

$$\Xi'(\nu^k)(d^k) = -\int_0^1 F(\nu^k)(t)^\top F'(\nu^k) \begin{pmatrix} \dfrac{\operatorname{sgn}(\mathscr{M}^k)\min\{1, |\mathscr{M}^k|\}\mu^k}{1+|\mathscr{M}^k|} \\ W_z^k G(\mu^k, z^k) \end{pmatrix}(t)\,\mathrm{d}t$$

$$
\begin{aligned}
&= -\int_0^1 \begin{pmatrix} \mu^k \\ G(\mu^k, z^k)(t) \end{pmatrix}^\top \begin{pmatrix} 1 & 0 \\ G'_\mu(\mu^k, z^k) & G'_z(\mu^k, z^k) \end{pmatrix} \\
&\qquad \begin{pmatrix} \dfrac{\operatorname{sgn}(\mathscr{M}^k)\min\{1, |\mathscr{M}^k|\}\mu^k}{1+|\mathscr{M}^k|} \\ W_z^k G(\mu^k, z^k) \end{pmatrix}(t)\, \mathrm{d}t \\
&= -\frac{\min\{1, |\mathscr{M}^k|\}|\mathscr{M}^k|}{1+|\mathscr{M}^k|}\mu^k - \int_0^1 G(\nu^k)(t)^\top G'_z(\nu^k)(W_z^k G(\nu^k))(t)\, \mathrm{d}t \\
&= -\frac{\min\{1, |\mathscr{M}^k|\}|\mathscr{M}^k|}{1+|\mathscr{M}^k|}\mu^k - \|W_z^k G(\mu^k, z^k)\|_Z^2 < 0,
\end{aligned} \tag{5.74}
$$

which implies (5.71). Moreover,

$$
\begin{aligned}
\mu^{k+1} = \mu^k + \alpha d_\mu^k &= \mu^k - \alpha \frac{\operatorname{sgn}(\mathscr{M}^k)\min\{1, |\mathscr{M}^k|\}\mu^k}{1+|\mathscr{M}^k|} \\
&= \left(1 - \frac{\operatorname{sgn}(\mathscr{M}^k)\min\{1, |\mathscr{M}^k|\}\alpha}{1+|\mathscr{M}^k|}\right)\mu^k > 0.
\end{aligned} \tag{5.75}
$$

Therefore, d^k is a direction of descent of Ξ at ν^k and the line-search in the Algorithm 5.16 is well-defined unless ν^k is a stationary (or P-stationary) point of Ξ. ■

The following global convergence result extends the proof presented in Kanzow [73], Qi, Sun, and Zhou [112] for finite dimensions into infinite dimensions.

Theorem 5.19. *Let $\nu_* = (\mu_*, z_*)$ be an accumulation point of the sequence $\{\nu^k\}$ generated by Algorithm 5.16. Let all first and second derivatives of the functions f_0, f, c, ψ be uniformly bounded. Let there be a constant C_F such that $\|F(z^k)\|_{Y_\infty} \le C_F$ for every k. Let Assumption 5.11 hold for Ξ. Moreover, let a constant $\widetilde{C}_S$ exist with (5.32) for all k.*

Then, ν_* is either a P-stationary point, or a stationary point of Ξ. Moreover, if the inverse operators $(V^k)^{-1}$ exist for all k, $C > 0$ is a constant such that $\|(V^k)^{-1}\|_{\mathscr{L}(Y_\infty, Z_\infty)} \le C$ holds for all k, and (5.66) is satisfied by the Newton direction for all but finitely many k, then z_* is a zero of F.

Proof. Let $\{\nu^k\}_{k\in K}$ be a subsequence with $\nu^k \to \nu_*$ and $\Xi'(\nu^k)(d^k) \ne 0$. Replacing

$\Theta(z^k)$ by $\Xi(v^k)$ in Theorem 5.14, together with Lemma 5.18, we have

$$\lim_{k(\in K)\to\infty} \alpha_k \Xi'(v^k)(d^k) = 0, \tag{5.76}$$

and

$$\lim_{k(\in K)\to\infty} \alpha_k \|d^k\|_{\widehat{Z}} = 0.$$

Case 1: Assume

$$\alpha = \liminf_{k(\in K)\to\infty} \alpha_k > 0. \tag{5.77}$$

We need to consider two subcases. Suppose first that

$$d^k = -\begin{pmatrix} \dfrac{\operatorname{sgn}(\mathscr{M}^k)\min\{1,|\mathscr{M}^k|\}\mu^k}{1+|\mathscr{M}^k|} \\ W_z^k G(\mu^k, z^k) \end{pmatrix}$$

holds for infinitely many $k \in K \subset \mathbb{N}$. Then it follows from (5.74) in Lemma 5.18 that for some infinite subset $K' \subseteq K$

$$\begin{aligned} \lim_{k(\in K')\to\infty} \Xi'(v^k)(d^k) &= -\lim_{k(\in K')\to\infty} \frac{\min\{1,|\mathscr{M}^k|\}|\mathscr{M}^k|}{1+|\mathscr{M}^k|}\mu^k \\ &\quad - \lim_{k(\in K')\to\infty} \|W_z^k G(\mu^k, z^k)\|_{\widehat{Z}} \\ &= 0. \end{aligned}$$

Hence v_* is a stationary (or P-stationary) point of Ξ.

On the other hand, if the condition that (5.65) is solvable holds for all but finitely many $k \in K$, then replacing $\Theta(z^k)$ by $\Xi(v^k)$ and repeating the corresponding process used in the Case 1 of Theorem 5.14 shows that v_* is a zero of F.

Case 2: Assume that there is a subsequence $\{z^k\}_{k\in J}$, $J \subseteq K$, with $\lim_{k(\in J)\to\infty} \alpha_k = 0$. By the similar process employed in the Case 2 of Theorem 5.14, the conclusion of theorem is valid. Therefore, in either case, we establish the conclusion. ■

Finally, we discuss the locally superlinear convergence of Algorithm 5.16.

Theorem 5.20**.** *Let the assumptions of Theorems 5.4 and 5.19 be valid. In Algorithm 5.16 let* $\sigma \in (0, 1/4)$ *and either* $\kappa > 0$ *sufficiently and* $\rho = 2$ *or* $\kappa > 0$ *and* $\rho > 2$

sufficiently large. Then, for sufficiently large k the step length $\alpha_k = 1$ *is accepted and the global smoothing method turns into the local one.*

5.5 Numerical Results

We used the smoothing Newton method with non-monotone line-search described in Algorithm 5.16 for the following computations. We did not make use of explicit inexactness (apart from numerical inaccuracies owing to rounding errors), that is we used $r^k = 0$ for our computations. A typical example, for which $r^k \neq 0$ occurs, is if iterative solvers are used for the occurring linear equations in the Newton step. In this case the accuracy of the iteratively obtained solutions has to be adapted as outlined in the above theory.

In each step of the smoothing Newton method a linear boundary value problem defining the search direction has to be solved numerically. For the following computations, a single shooting method was used to solve these boundary value problems. Herein, the differential equations are discretized on the time interval using the explicit Euler method with N equidistant subintervals. The occurring derivatives $(x^k)'$ and $(\lambda^k)'$ are approximated by finite forward differences on the grid.

All computations were performed on a PC with 3 GHz processing speed and 1 GB of memory. The following parameters were used throughout the computations: $\bar{\mu} = 1$, $\gamma = 0.5$, $\beta = 0.9$, $\sigma = 0.1$, $M = 5$.

5.5.1 Rayleigh Example

We consider the Rayleigh problem, which was investigated earlier in Maurer and Augustin [95, p. 39], and in Gerdts [56, Section 5.2]:

Minimize

$$\int_0^{4.5} u(t)^2 + x_1(t)^2 \, \mathrm{d}t \tag{5.78}$$

subject to

$$\begin{cases} x_1' &= x_2, & x_1(0) = -5,\ x_1(4.5) = 0, \\ x_2' &= -x_1 + x_2\left(1.4 - 0.14x_2^2\right) + 4u, & x_2(0) = -5,\ x_2(4.5) = 0, \end{cases} \tag{5.79}$$

TABLE 5.1
Output of the smoothing Newton method for Rayleigh's problem for $N = 1000$ subintervals and Euler discretization: local superlinear convergence.

ITER	ALPHA	||F||	||d||	MU
0	0.729000E+00	0.432539E+02	0.150713E+05	0.635500E+00
1	0.590490E+00	0.665081E+02	0.621723E+04	0.555489E+00
2	0.100000E+01	0.147693E+02	0.690325E+04	0.500000E+00
3	0.100000E+01	0.233098E+01	0.236290E+04	0.500000E+00
4	0.100000E+01	0.342511E+00	0.173333E+04	0.500000E+00
5	0.100000E+01	0.164712E-01	0.509764E+03	0.171255E+00
6	0.100000E+01	0.960063E-02	0.835709E+03	0.823562E-02
7	0.100000E+01	0.205645E-03	0.146192E+03	0.480031E-02
8	0.100000E+01	0.120216E-04	0.248320E+02	0.102823E-03
9	0.100000E+01	0.115521E-05	0.581244E+01	0.601078E-05
10	0.100000E+01	0.193262E-07	0.238858E+00	0.577607E-06
11	0.100000E+01	0.486511E-10	0.347779E-01	0.966310E-08
12	0.100000E+01	0.113434E-14	0.185980E-02	0.243256E-10
13	0.100000E+01	0.100863E-23	0.859600E-05	0.567170E-15

and

$$-1 \leq u(t) \leq 1.$$

It can be checked (see Gerdts [56, Section 5.2]) that the regularity assumptions are satisfied for this problem. We leave the details to the reader.

Table 5.1 shows details of the iterations, i.e. step size α, residual norm $\|F\|_2$, search direction $\|d_z^k\|$, and smoothing parameter μ^k.

Figure 5.1 shows the iterates of the smoothing Newton method for $N = 1000$.

The number of iterations remains nearly constant, which indicates — at least numerically — the mesh independence of the method. Furthermore, the CPU time grows at a linear rate with N. For this example the smoothing method requires 2-5 iterations less compared to the results presented in Gerdts [56, page 347].

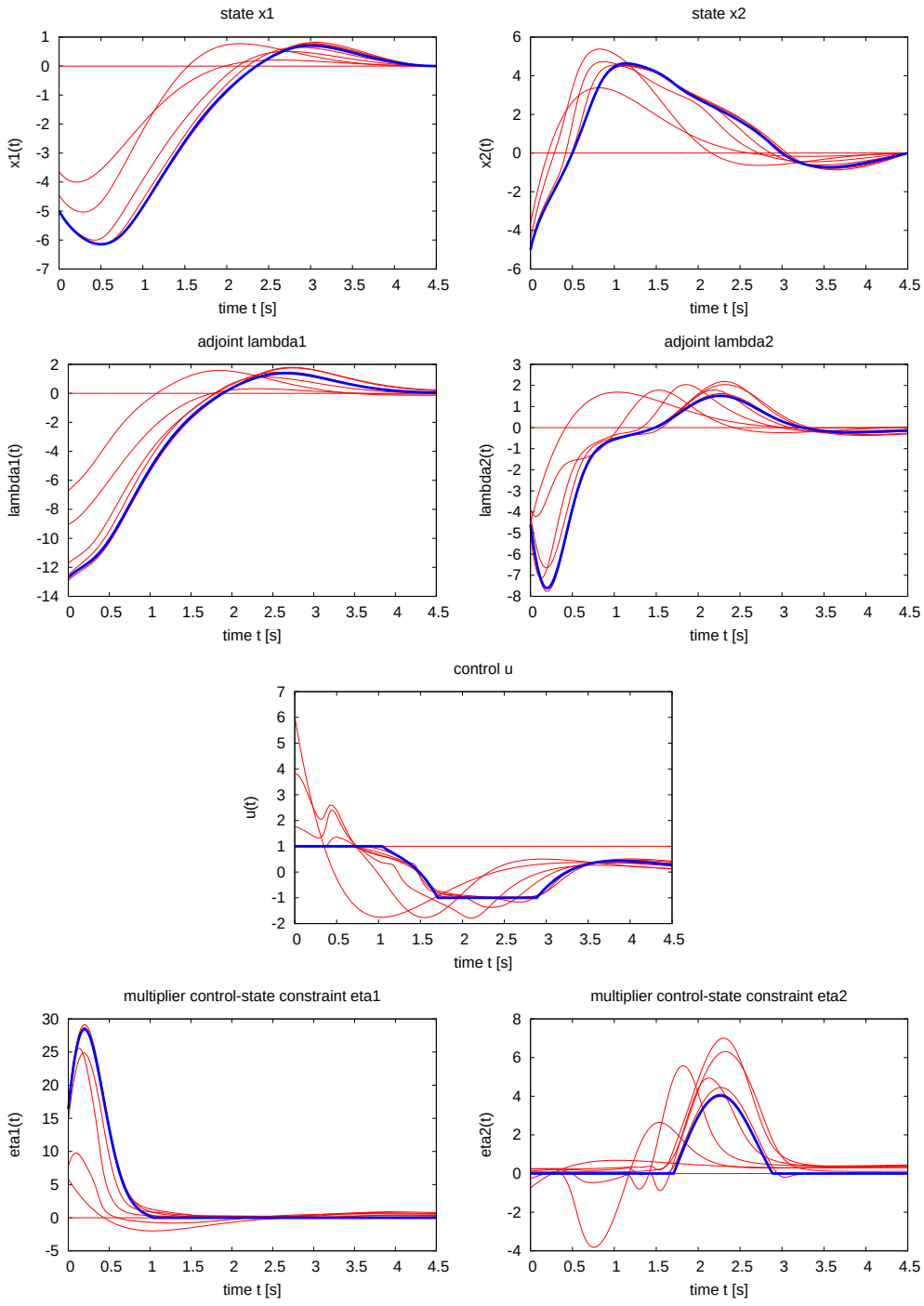

FIGURE 5.1 *Numerical solution of Rayleigh's problem for* $N = 1000$ *Euler steps: Intermediate iterates (thin lines) and converged solution (thick lines).*

N	CPU time [s]	Iterations
100	0.112	12
200	0.236	12
400	0.400	13
800	0.640	13
1600	1.332	14
3200	2.672	15
6400	6.692	16
12800	12.789	16
25600	23.217	16

In addition, reference solutions were computed using a direct discretization method as in Gerdts [53] using an Euler discretization and feasibility tolerance 10^{-10} and optimality tolerance $\sqrt{\text{eps}}$, where eps denotes the machine precision. This direct discretization method needed 2.724 seconds for $N = 100$, 21.565 seconds for $N = 200$ and 311.911 seconds for $N = 400$. This indicates that the smoothing Newton method is extremely efficient, provided that all regularity assumptions are satisfied.

5.5.2 Trolley Example

We consider an optimal control problem for a trolley of mass m_1 moving in a high rack storage area. A load of mass m_2 is attached to the trolley by a rigid cable of length ℓ, cf. Figure 5.2. Herein, x_1 and x_3 denote the x-coordinate of the trolley and its velocity, respectively, and x_2 and x_4 refer to the angle between vertical axis and cable and its velocity, respectively. The acceleration of the trolley can be controlled by the control u which is subject to

$$-0.5 \leq u(t) \leq 0.5. \tag{5.80}$$

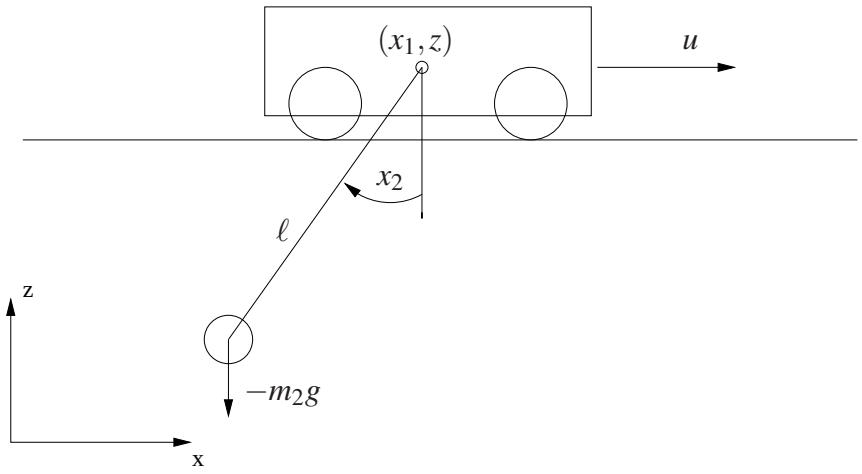

FIGURE 5.2 *Configuration of the trolley and the load.*

The equations of motion of the trolley are given by the following differential equations for the state $x = (x_1, x_2, x_3, x_4)^\top$:

$$\begin{cases} x_1' &= x_3 \\ x_2' &= x_4 \\ x_3' &= \dfrac{m_2^2\ell^3 \sin(x_2)x_4^2 - m_2\ell^2 u + m_2 I_y \ell x_4^2 \sin(x_2) - I_y u + m_2^2\ell^2 g\cos(x_2)\sin(x_2)}{-m_1 m_2\ell^2 - m_1 I_y - m_2^2\ell^2 - m_2 I_y + m_2^2\ell^2\cos(x_2)^2} \\ x_4' &= \dfrac{m_2\ell\left(m_2\ell\cos(x_2)x_4^2\sin(x_2) - \cos(x_2)u + g\sin(x_2)(m_1+m_2)\right)}{-m_1 m_2\ell^2 - m_1 I_y - m_2^2\ell^2 - m_2 I_y + m_2^2\ell^2\cos(x_2)^2} \end{cases} \tag{5.81}$$

The optimal control problem is defined by the task to control the trolley from the given initial position

$$x_1(0) = x_2(0) = x_3(0) = x_4(0) = 0$$

to the terminal position

$$x_1(t_f) = 1,\ x_2(t_f) = x_3(t_f) = x_4(t_f) = 0.$$

within the fixed time $t_f = 2.7$ such that the objective function

$$\frac{1}{2}\int_0^{t_f} u(t)^2 + 5x_4(t)^2 \,\mathrm{d}t$$

TABLE 5.2
Output of globalized non-monotone smoothing Newton method for the trolley example for $N = 1000$ subintervals and Euler discretization: local superlinear convergence.

ITER	ALPHA	||F||	||dx||	MU
0	0.100000E+01	0.317199E+00	0.566262E+04	0.500000E+00
1	0.100000E+01	0.227788E-01	0.208544E+04	0.158600E+00
2	0.100000E+01	0.477624E-02	0.994667E+03	0.113894E-01
3	0.100000E+01	0.136652E+00	0.230600E+04	0.238812E-02
4	0.100000E+01	0.660310E-01	0.561017E+04	0.683261E-01
5	0.100000E+01	0.374706E-02	0.356179E+04	0.330155E-01
6	0.100000E+01	0.261953E-03	0.114805E+04	0.187353E-02
7	0.100000E+01	0.121403E-03	0.763222E+03	0.130977E-03
8	0.100000E+01	0.263449E-05	0.185290E+02	0.607015E-04
9	0.100000E+01	0.450633E-07	0.387605E+01	0.131725E-05
10	0.100000E+01	0.321005E-09	0.940170E+00	0.225316E-07
11	0.100000E+01	0.650199E-12	0.280432E-01	0.160502E-09
12	0.100000E+01	0.161108E-15	0.114973E-02	0.325100E-12
13	0.100000E+01	0.252786E-21	0.203142E-04	0.805541E-16
14	0.100000E+01	0.464054E-26	0.243029E-07	0.126393E-21

is minimized subject to (5.80) and (5.81). The objective function aims at minimizing the steering effort and the angle velocity of the cable to avoid extensive swinging of the load.

Note that the boundary conditions define an entirely symmetric configuration, so the resulting solution should be symmetric, too.

The following parameters were used for the numerical computations:

$$g = 9.81,\ m_1 = 0.3,\ m_2 = 0.5,\ \ell = 0.75,\ r = 0.1,\ I_y = 0.002.$$

Table 5.2 summarizes CPU times for the smoothing Newton method depending on the number N of equidistant intervals used in the linear boundary value problems. Table 5.2 shows the output of the smoothing Newton method with non-monotone line-search, i.e. step size α, residual norm $\|F\|_2$, search direction $\|d_z^k\|$, and smoothing parameter μ^k during iterations. The iterations show the rapid superlinear convergence at the end of the iteration sequence as predicted by Theorem 5.20.

The following table summarizes results for different step sizes. The number of itera-

tions is nearly constant, which indicates – at least numerically – the mesh independence of the method. Furthermore, the CPU time grows at a linear rate with N.

N	CPU time [s]	Iterations
101	0.244	11
201	0.392	13
401	0.752	13
801	1.408	13
1600	3.060	14
3200	6.176	13
6400	11.673	13
12800	22.981	14
25600	45.951	14

As before reference solutions with a direct discretization method were computed and led to the following CPU times: 6.708 seconds for $N = 200$, 62.576 seconds for $N = 400$ and 797.122 seconds for $N = 800$. Again, the smoothing Newton method turns out to be very efficient.

Finally, Figures 5.3 and 5.4 illustrate the iterates of the smoothing Newton method. Notice the symmetry in the solution which is due to the symmetry in the boundary conditions.

Figure 5.4 shows that the smoothing Newton method is able to find the switching structure of the optimal solution without any a priori assumptions.

5.6 Gradient Operator

Our aim is to compute the operator W^k in $d^k = -W^k F(z^k)$ in Algorithm 5.9 given the functional

$$\begin{aligned}\Theta(z) &= \frac{1}{2}\|F(z)\|_{Y_2}^2 \\ &= \frac{1}{2}\int_0^1 \left\|x'(t) - f(x(t),u(t))\right\|^2 \,\mathrm{d}t \\ &\quad + \frac{1}{2}\int_0^1 \left\|\lambda'(t) + H_x'(x(t),u(t),\lambda(t),\eta(t))^\top\right\|^2 \,\mathrm{d}t\end{aligned}$$

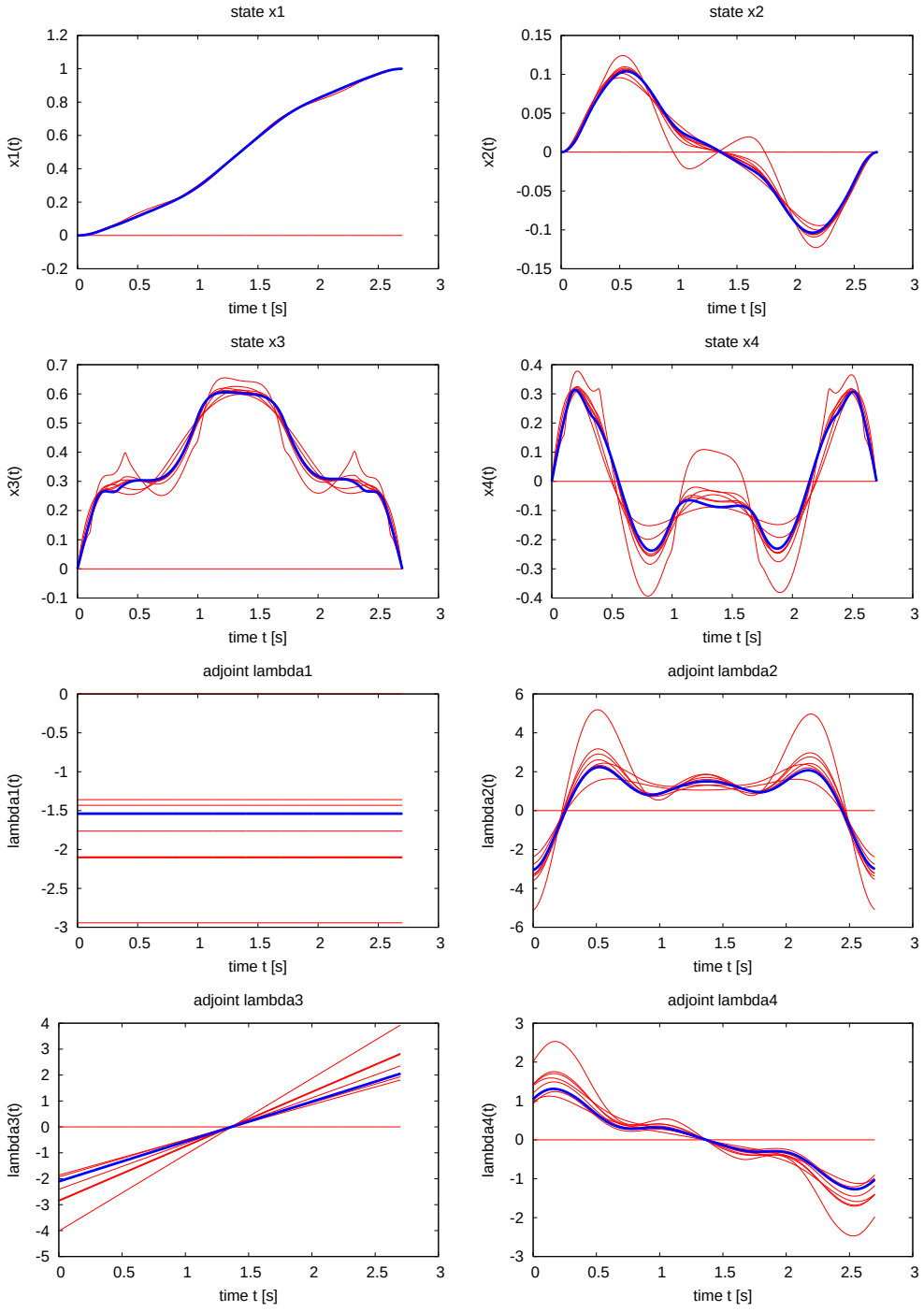

FIGURE 5.3 *Numerical solution of the trolley example for $N = 1000$ Euler steps: States and adjoints at intermediate iterates (thin lines) and converged solution (thick lines).*

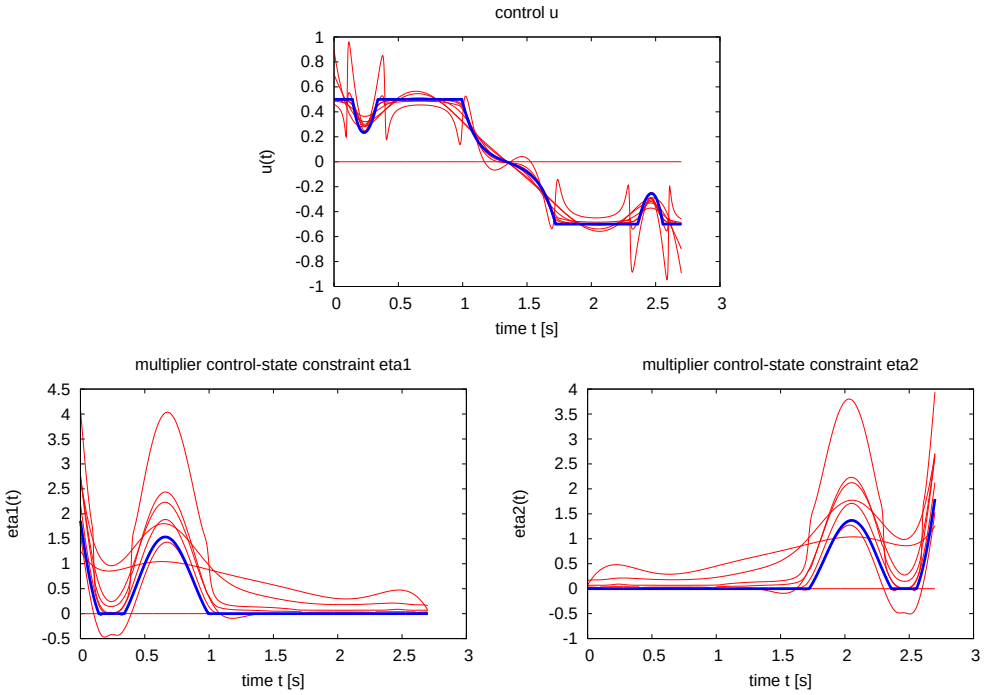

FIGURE 5.4 *Numerical solution of the trolley example for $N = 1000$ Euler steps: Control and multipliers at intermediate iterates (thin lines) and converged solution (thick lines).*

$$
\begin{aligned}
&+\frac{1}{2}\int_0^1 \left\| H_u'(x(t),u(t),\lambda(t),\eta(t))^\top \right\|^2 \,\mathrm{d}t \\
&+\frac{1}{2}\int_0^1 \|\omega(x(t),u(t),\lambda(t),\eta(t),\sigma)\|^2 \,\mathrm{d}t \\
&+\frac{1}{2}\|\psi(x(0),x(1))\|^2+\frac{1}{2}\|\lambda(0)+\psi_{x_0}'(x(0),x(1))^\top\sigma\|^2 \\
&+\frac{1}{2}\|\lambda(1)-\psi_{x_1}'(x(0),x(1))^\top\sigma\|^2,
\end{aligned}
$$

where ω is given by (5.8). Differentiating Θ at z^k and partial integration yields

$$
\begin{aligned}
\Theta'(z^k)(z) &= \int_0^1 \left((x^k)'(t)-f[t]\right)^\top \left(x'(t)-f_x'[t]x(t)-f_u'[t]u(t)\right)\,\mathrm{d}t \\
&\quad+\int_0^1 \left((\lambda^k)'(t)+H_x'[t]^\top\right)^\top \Big(\lambda'(t)+H_{xx}''[t]x(t) \\
&\quad+H_{xu}''[t]u(t)+H_{x\lambda}''[t]\lambda(t)+H_{x\eta}''[t]\eta(t)\Big)\,\mathrm{d}t \\
&\quad+\int_0^1 H_u'[t]\left(H_{ux}''[t]x(t)+H_{uu}''[t]u(t)+H_{u\lambda}''[t]\lambda(t)+H_{u\eta}''[t]\eta(t)\right)\,\mathrm{d}t \\
&\quad+\int_0^1 \omega[t]^\top\left(-S(t)\left(c_x'[t]x(t)+c_u'[t]u(t)\right)+R(t)\eta(t)\right)\,\mathrm{d}t \\
&\quad+(\psi(x^k(0),x^k(1)))^\top\left(\psi_{x_0}'(x^k(0),x^k(1))x(0)+\psi_{x_1}'(x^k(0),x^k(1))x(1)\right) \\
&\quad+\left(\lambda^k(0)+(\psi_{x_0}'(x^k(0),x^k(1)))^\top\sigma^k\right)^\top\Big(\lambda(0)+((\psi_{x_0}')^\top\sigma^k)_{x_0}'x(0) \\
&\quad+((\psi_{x_0}')^\top\sigma^k)_{x_1}'x(1)+(\psi_{x_0}')^\top\sigma\Big) \\
&\quad+\left(\lambda^k(1)-(\psi_{x_1}'(x^k(0),x^k(1)))^\top\sigma^k\right)^\top\Big(\lambda(1)-((\psi_{x_1}')^\top\sigma^k)_{x_0}'x(0) \\
&\quad-((\psi_{x_1}')^\top\sigma^k)_{x_1}'x(1)-(\psi_{x_1}')^\top\sigma\Big) \\
&= \int_0^1 g_1(t)^\top x'(t)\,\mathrm{d}t+\int_0^1 g_2(t)^\top u(t)\,\mathrm{d}t+\int_0^1 g_3(t)^\top\lambda'(t)\,\mathrm{d}t \\
&\quad+\int_0^1 g_4(t)^\top\eta(t)\,\mathrm{d}t+\Delta_1^\top x(0)+\Delta_2^\top x(1)+\Delta_3^\top\lambda(0)+\Delta_4^\top\lambda(1)+\Delta_5^\top\sigma,
\end{aligned}
$$

where

$$
\begin{aligned}
g_1(t)^\top &= \left((x^k)'(t)-f[t]\right)^\top+\int_0^t\Big(\left((x^k)'(\tau)-f[\tau]\right)^\top f_x'[\tau] \\
&\quad-\left((\lambda^k)'(\tau)+H_x'[\tau]^\top\right)^\top H_{xx}''[\tau]-H_u'[\tau]H_{ux}''[\tau]-\omega[\tau]^\top S(\tau)c_x'[\tau]\Big)\,\mathrm{d}\tau,
\end{aligned}
$$

$$
\begin{aligned}
g_2(t)^\top &= -\Big((x^k)'(t) - f[t]\Big)^\top f'_u[t] + \Big((\lambda^k)'(t) + H'_x[t]^\top\Big)^\top H''_{xu}[t] \\
&\quad + H'_u[t]H''_{uu}[t] - \omega[t]^\top S(t)c'_u[t], \\
g_3(t)^\top &= \Big((\lambda^k)'(t) + H'_x[t]^\top\Big)^\top \\
&\quad - \int_0^t \left(\Big((\lambda^k)'(\tau) + H'_x[\tau]^\top\Big)^\top H''_{x\lambda}[\tau] + H'_u[\tau]H''_{u\lambda}[\tau]\right) \mathrm{d}\tau, \\
g_4(t)^\top &= \Big((\lambda^k)'(t) + H'_x[t]^\top\Big)^\top H''_{x\eta}[t] + H'_u[t]H''_{u\eta}[t] + \omega[t]^\top R(t), \\
\Delta_1^\top &= \psi^\top \psi'_{x_0} + \Big(\lambda^k(0) + (\psi'_{x_0})^\top \sigma^k\Big)^\top ((\psi'_{x_0})^\top \sigma^k)'_{x_0} \\
&\quad - \Big(\lambda^k(1) - (\psi'_{x_1})^\top \sigma^k\Big)^\top ((\psi'_{x_1})^\top \sigma^k)'_{x_0}, \\
\Delta_2^\top &= \psi^\top \psi'_{x_1} + \Big(\lambda^k(0) + (\psi'_{x_0})^\top \sigma^k\Big)^\top ((\psi'_{x_0})^\top \sigma^k)'_{x_1} \\
&\quad - \Big(\lambda^k(1) - (\psi'_{x_1})^\top \sigma^k\Big)^\top ((\psi'_{x_1})^\top \sigma^k)'_{x_1} + \int_0^1 \Big(\Big((x^k)'(\tau) - f[\tau]\Big)^\top f'_x[\tau] \\
&\quad - \Big((\lambda^k)'(\tau) + H'_x[\tau]^\top\Big)^\top H''_{xx}[\tau] - H'_u[\tau]H''_{ux}[\tau] - \omega[\tau]^\top S(\tau)c'_x[\tau]\Big) \mathrm{d}\tau, \\
\Delta_3^\top &= \Big(\lambda^k(0) + (\psi'_{x_0})^\top \sigma^k\Big)^\top, \\
\Delta_4^\top &= \Big(\lambda^k(1) - (\psi'_{x_1})^\top \sigma^k\Big)^\top \\
&\quad + \int_0^1 \left(\Big((\lambda^k)'(\tau) + H'_x[\tau]^\top\Big)^\top H''_{x\lambda}[\tau] + H'_u[\tau]H''_{u\lambda}[\tau]\right) \mathrm{d}\tau, \\
\Delta_5^\top &= \Big(\lambda^k(0) + (\psi'_{x_0})^\top \sigma^k\Big)^\top (\psi'_{x_0})^\top - \Big(\lambda^k(1) - (\psi'_{x_1})^\top \sigma^k\Big)^\top (\psi'_{x_1})^\top.
\end{aligned}
$$

Moreover, it holds

$$
x(1) = x(0) + \int_0^1 x'(t)\,\mathrm{d}t, \qquad \lambda(1) = \lambda(0) + \int_0^1 \lambda'(t)\,\mathrm{d}t.
$$

Introducing these relations into the above formula yields

$$
\begin{aligned}
\Theta'(z^k)(z) &= \int_0^1 (\Delta_2 + g_1(t))^\top x'(t)\,\mathrm{d}t + \int_0^1 g_2(t)^\top u(t)\,\mathrm{d}t \\
&\quad + \int_0^1 (\Delta_4 + g_3(t))^\top \lambda'(t)\,\mathrm{d}t + \int_0^1 g_4(t)^\top \eta(t)\,\mathrm{d}t \\
&\quad + (\Delta_1 + \Delta_2)^\top x(0) + (\Delta_3 + \Delta_4)^\top \lambda(0) + \Delta_5^\top \sigma.
\end{aligned}
$$

Define $d^k = (x, \lambda, u, \eta, \sigma) \in Z_\infty$ by

$$\begin{aligned} x'(t) &= -(\Delta_2 + g_1(t)), \qquad x(0) = -(\Delta_1 + \Delta_2), \\ \lambda'(t) &= -(\Delta_4 + g_3(t)), \qquad \lambda(0) = -(\Delta_3 + \Delta_4), \\ u(t) &= -g_2(t), \\ \eta(t) &= -g_4(t), \\ \sigma &= -\Delta_5. \end{aligned}$$

Then, d^k can be expressed as $d^k = -W^k F(z^k)$ with a continuous (and thus bounded) linear operator $W^k : Y_\infty \to Z_\infty$. Notice that d^k is actually an element of the space Z_∞. We omit the technical details of defining W^k explicitly although this is straightforward by exploiting the linear structure of $g_1, \ldots, g_4$ and $\Delta_1, \ldots, \Delta_5$.

We note, that W^k is uniformly bounded, i.e. there exists a constant C independent of k with

$$\|W^k h\|_{Z_\infty} \le C \|h\|_{Y_\infty} \qquad \text{for every } k,$$

if all first and second derivatives of the functions f_0, f, c, ψ are uniformly bounded.

Actually, d^k can be interpreted as the negative gradient of Θ at z^k in the Hilbert space $\widehat{Z} = Z_2$ equipped with the norm $\|d\|_{\widehat{Z}} = \sqrt{\langle d, d \rangle_{\widehat{Z} \times \widehat{Z}}}$ with the inner product

$$\langle v, w \rangle_{\widehat{Z} \times \widehat{Z}} := \langle v_x, w_x \rangle_{1,2} + \langle v_u, w_u \rangle_2 + \langle v_\lambda, w_\lambda \rangle_{1,2} + \langle v_\eta, w_\eta \rangle_2 + v_\sigma^\top w_\sigma,$$

where $v = (v_x, v_u, v_\lambda, v_\eta, v_\sigma) \in Z_\infty$, $w = (w_x, w_u, w_\lambda, w_\eta, w_\sigma) \in Z_\infty$. The inner products $\langle \cdot, \cdot \rangle_{1,2}$ in the space $W^{1,2}$ and $\langle \cdot, \cdot \rangle_2$ in the space L^2 for $v, w \in W^{1,2}$ and $v, w \in L^2$, respectively, are defined by

$$\langle v, w \rangle_{1,2} = v(0)^\top w(0) + \langle v', w' \rangle_2, \quad \langle v, w \rangle_2 = \int_0^1 v(t)^\top w(t) \, \mathrm{d}t.$$

Using this norm, the search direction $d^k = -W^k F(z^k)$ satisfies

$$\Theta'(z^k)(d^k) = -\|W^k F(z^k)\|_{\widehat{Z}}^2 = -\|d^k\|_{\widehat{Z}}^2$$

and hence d^k is a direction of descent unless $d^k = 0$.

In addition, we should note that by replacing $\Theta(z^k)$ with $\Xi(\nu^k)$ of Section 5.4 in the above discussions, we can find a direction

$$d^k = -\begin{pmatrix} \mathscr{M}^k \\ \mathscr{N}^k \end{pmatrix} = -\begin{pmatrix} \mu^k + W_\mu^k G(\mu^k, z^k) \\ W_z^k G(\mu^k, z^k) \end{pmatrix} = -W^k F(\mu^k, z^k),$$

in Algorithm 5.16 such that $\Xi'(\nu^k)(d^k) = -\|d^k\|_{\widehat{Z}}^2$ with linear operators W^k, W_μ^k, W_z^k. More specifically, $d^k = (\mu, x, \lambda, u, \eta, \sigma) \in \mathbb{R} \times Z_\infty$ is given by

$$\begin{aligned}
\mu &= -\left(\mu^k + \int_0^1 \omega[t]^\top F_{2\mu}'(\nu^k)(t)\,\mathrm{d}t\right), \\
x'(t) &= -(\Delta_2 + g_1(t)), \qquad x(0) = -(\Delta_1 + \Delta_2), \\
\lambda'(t) &= -(\Delta_4 + g_3(t)), \qquad \lambda(0) = -(\Delta_3 + \Delta_4), \\
u(t) &= -g_2(t), \\
\eta(t) &= -g_4(t), \\
\sigma &= -\Delta_5,
\end{aligned}$$

where $F_{2\mu}'$ is defined in (5.60) and S and R in g_1, g_2, and g_4 are defined in (5.61). The details are just a verbatim repetition of the above procedures. We omit these here.

Finally, we establish an auxiliary result on stationary points.

Lemma 5.21**.** *Let $\Theta : Z_\infty \to \mathbb{R}$ be Fréchet differentiable. Let $z_* \in Z_\infty$ and $d_* \in Z_\infty$ be given with $\Theta'(z_*)(d_*) = 0$. Furthermore, let $\{z^k\}_{k\in\mathbb{N}}$ and $d^k = -W^k F(z^k)$ be sequences with $z^k \to z_*$ in Z_∞ and $\Theta'(z^k)(d^k) \to 0$.*

Then z_* is a stationary point of Θ, i.e. $\Theta'(z_*)(d) = 0$ for every $d \in Z_\infty$.

Proof. Assume that there exists $d \in Z_\infty$ with $\|d\|_{\widehat{Z}} = 1$ and $\Theta'(z_*)(d) \neq 0$. It holds

$$0 = \lim_{k\to\infty} \Theta'(z^k)(d^k) = -\lim_{k\to\infty} \|d^k\|_{\widehat{Z}}^2$$

and thus for some constant C

$$0 \leq |\Theta'(z^k)(d)| = |\langle d^k, d\rangle_{\widehat{Z}\times\widehat{Z}}| \leq C\|d^k\|_{\widehat{Z}}\|d\|_{\widehat{Z}} = C\|d^k\|_{\widehat{Z}} \to 0.$$

The continuity of $\Theta'(\cdot)$ implies

$$0 = \lim_{k\to\infty} \Theta'(z^k)(d) = \Theta'(z_*)(d) \neq 0,$$

which is a contradiction. ■

5.7 Contributions and Future Research

This chapter studied the numerical solutions for the optimal control problems subject to mixed control-state constraints via the inexact nonsmooth and smoothing Newton methods. Global convergence of the proposed algorithms is established under a non-monotonic backtracking strategy. The locally superlinear convergence under certain regularity conditions is analyzed. Numerical examples show that our approaches are very promising.

It would be interesting to establish the locally quadratic convergence for the inexact nonsmooth Newton method we mentioned. Also the feasibility of weakening the regularity conditions of the generalized Jacobian matrix at a zero of the OCP presents further challenges. This would be particularly beneficial for problems with pure state constraints or singular controls as in these cases the presented regularity assumptions for uniform non-singularity do not hold. We are currently starting to investigate this semismoothness of the reformulation of OCP in more details.

Bibliography

[1] E. L. ALLGOWER AND K. GEORG, *Numerical Continuation Methods: An Introduction*, Springer, Berlin, 1990.

[2] R. ANDREANI, A. FRIEDLANDER, AND S. A. SANTOS, *On the resolution of the generalized nonlinear complementarity problem,* SIAM Journal on Optimization, 12 (2001), pp. 303–321.

[3] U. M. ASCHER AND L. R. PETZOLD, *Computer Methods for Ordinary Differential Equations and Differential Algebraic Equations*, SIAM, Philadelphia, 1998.

[4] S. BELLAVIA, M. MACCONI, AND B. MORINI, *An affine scaling trust-region approach to bounded-constrained nonlinear systems*, Applied Numerical Mathematics, 44 (2003), pp. 257–280.

[5] S. BELLAVIA, M. MACCONI, AND B. MORINI, *STRSCNE: A scaled trust-region solver for constrained nonlinear equations*, Computational Optimization and Applications, 28 (2004), pp. 31–50.

[6] S. BELLAVIA AND B. MORINI, *A globally convergent Newton-GMRES subspace method for systems of nonlinear equations*, SIAM Journal on Scientific Computing, 23 (2001), pp. 940–960.

[7] S. BELLAVIA AND B. MORINI, *Subspace trust-region methods for large bound constrained nonlinear equations*, SIAM Journal on Numerical Analysis, 44 (2006), pp. 1535–1555.

[8] S. C. BILLUPS AND K. G. MURTY, *Complementarity problems*, Journal of Computational and Applied Mathematics, 124 (2000), pp. 303–318.

[9] P. T. BOGGS, *The solution of nonlinear systems of equations by A-stable integration techniques*, SIAM Journal on Numerical Analysis, 8 (1971), pp. 767–785.

[10] P. N. BROWN, *A local convergence theory for combined inexact-Newton/finite-difference projection methods*, SIAM Journal on Numerical Analysis, 24 (1987), pp. 407–434.

[11] P. N. BROWN, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM Journal on Scientific and Statistical Computing, 12 (1991), pp. 58–78.

[12] P. N. BROWN AND Y. SAAD, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM Journal on Scientific and Statistical Computing, 11 (1990), pp. 450–481.

[13] P. N. BROWN AND Y. SAAD, *Convergence theory of nonlinear Newton-Krylov algorithms*, SIAM Journal on Optimization, 4 (1994), pp. 297–330.

[14] L. BRUGNANO AND V. CASULLI, *Iterative solution of piecewise linear systems and applications to flows in porous media*, SIAM Journal on Scientific Computing, to appear.

[15] L. BRUGNANO AND V. CASULLI, *Iterative solution of piecewise linear systems*, SIAM Journal on Scientific Computing, 30 (2008), pp. 463–472.

[16] L. BRUGNANO AND J. H. CHEN, *On Newton-type methods for solving some extended piecewise linear systems*, (2008), submitted for publication.

[17] L. G. BULLARD AND L. T. BIEGLER, *Iterative linear programming strategies for constrained simulation*, Computers and Chemial Engineering, 15 (1991), pp. 239–254.

[18] J. V. BURKE AND M. C. FERRIS, *A Gauss-Newton method for convex composite optimization*, Mathematical Programming, 71 (1995), pp. 179–194.

[19] C. BÜSKENS, *Optimierungsmethoden und Sensitivitätsanalyse für optimale Steuerprozesse mitSteuer- und Zustandsbeschränkungen*, PhD thesis, Fachbereich Mathematik, Westfälische Wilhems-Universität Münster, 1998.

[20] V. CASULLI, *Semi-implicit finite difference methods for the two-dimensional shallow water equations*, Journal of Computational Physics, 86 (1990), pp. 56–74.

[21] V. CASULLI AND P. ZANOLLI, *Semi-implicit modeling of nonhydrostatic free-surface flows for environmental problems*, Mathematical and Computer Modelling, 36 (2002), pp. 1131–1149.

[22] B. CHEN AND P. T. HARKER, *A non-interior-point continuation method for linear complementarity problem*, SIAM Journal on Matrix Analysis and Applications, 14 (1993), pp. 1168-1190.

[23] J. H. CHEN AND R. P. AGARWAL, *On the finite termination of a Newton-type approach for solving piecewise linear systems*, (2008), submitted for publication.

[24] J. H. CHEN AND M. GERDTS, *Numerical solution for control-state constrained optimal control problems with inexact nonsmooth and smoothing Newton methods*, (2008), submitted for publication.

[25] J. H. CHEN AND L. QI, *Globally and superlinearly convergent inexact Newton-Krylov algorithms for solving nonsmooth equations*, (2008), submitted for publication.

[26] J. H. CHEN AND L. QI, *Pseudotransient continuation for solving systems of nonsmooth equations with inequality constraints*, (2008), submitted for publication.

[27] X. J. CHEN, Z. NASHED, AND L. QI, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM Journal on Numerical Analysis, 38 (2000), pp. 1200–1216.

[28] X. J. CHEN, Z. NASHED, AND L. QI, *Covergence of Newton's method for singular smooth and nonsmooth equations using adaptive outer inverses*, SIAM Journal on Optimization, 7 (1997), pp. 445–462.

[29] X. J. CHEN, L. QI, AND D. SUN, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities*, Mathematics of Computation, 67 (1998), pp. 519–540.

[30] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley & Sons, New York, 1983.

[31] T. S. COFFEY, C. T. KELLEY, AND D. E. KEYES, *Pseudo-transient continuation and differential-algebraic equations*, SIAM Journal on Scientific Computing, 25 (2003), pp. 553–569.

[32] T. S. COFFEY, R. J. MCMULLAN, C. T. KELLEY, AND D. S. MCRAE, *Globally convergent algorithms for nonsmooth nonlinear equations in computational fluid dynamics*, Journal of Computational and Applied Mathematics, 152 (2003), pp. 69–81.

[33] T. F. COLEMAN AND Y. LI, *An interior trust region approach for nonlinear minimization subject to bounds*, SIAM Journal on Optimization, 6 (1996), pp. 418–445.

[34] R. W. COTTLE, J.-S. PANG, AND R. E. STONE, *The Linear Complementarity Problem*, Academic Press, San Diego, CA, 1992.

[35] J. W. DANIEL, *Newton's method for nonlinear inequalities*, Numerische Mathematik, 6 (1973), pp. 381–387.

[36] T. DE LUCA, F. FACCHINEI, AND C. KANZOW, *A semismooth equation approach to the solution of nonlinear complementary problems*, Mathematical Programming, 75 (1996), pp. 407–439.

[37] R. S. DEMBO, S.C. EISENSTAT, AND T. STEIHAUG, *Inexact Newton methods*, SIAM Journal on Numerical Analysis, 19 (1982), pp. 400–408.

[38] J. E. DENNIS, M. EL-ALEM, AND K. WILLIAMSON, *A trust-region approach to nonlinear systems of equalities and inequalities*, SIAM Journal on Optimization, 9 (1999), pp. 291–315.

[39] N. Y. DENG, Y. XIAO, AND F. J. ZHOU, *A nonmonotonic trust region algorithm*, Journal of Optimization Theory and Applications, 76 (1993), pp. 259–285.

[40] J. E. DENNIS JR. AND R. B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall Series in Computation; Mathematics, Englewood Cliffs, NJ, 1983.

[41] B. C. EAVES, *The linear complementarity problem*, Management Science, 17 (1971), pp. 612–634.

[42] S. C. EISENSTAT AND H. F. WALKER, *Globally convergent inexact Newton methods*, SIAM Journal on Optimization, 4 (1994), pp. 393–422.

[43] S. C. EISENSTAT AND H. F. WALKER, *Choosing the forcing terms in an inexact Newton method*, SIAM Journal on Scientific Computing, 17 (1996), pp. 16–32.

[44] F. FACCHINEI, A. FISCHER, AND C. KANZOW, *Inexact Newton methods for semismooth equations with applications to variational inequality problems*, in G. Di Pillo and F. Giannessi (eds.): Nonlinear Optimization and Applications, Plenum Press, New York, (1996), pp. 125–139.

[45] F. FACCHINEI AND C. KANZOW, *A nonsmooth inexact Newton method for the solution of large-scale nonlinear complementarity problems*, Mathematical Programming, 76 (1997), pp. 493–512.

[46] F. FACCHINEI AND J.-S. PANG, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.

[47] M. W. FARTHING, C. E. KEES, T. COFFEY, C. T. KELLEY, AND C. T. MILLER, *Efficient steady-state solution techniques for variably saturated groundwater flow*, Advances in Water Resources, 26 (2003), pp. 833–849.

[48] A. FISCHER, *A special Newton-type optimization method*, Optimization, 24 (1992), pp. 269–284.

[49] A. FISCHER, *Solution of monotone complementarity problems with locally Lipschitzian functions*, Mathematical Programming, 76 (1997), pp. 513–532.

[50] R. FLETCHER AND S. LEYFFER, *Filter-type algorithms for solving systems of algebraic equations and inequalities*, High Performance Algorithms and Software for Nonlinear Optimization, G. Di Pillo and A. Murli, editors, Kluwer Academic Publishers, 2003, pp. 259–278.

[51] K. R. FOWLER AND C. T. KELLEY, *Pseudo-transient continuation for nonsmooth nonlinear equations*, SIAM Journal on Numerical Analysis, 43 (2005), pp. 1385–1406.

[52] U. M. GARCIA-PALOMARES AND A. RESTUCCIA, *A global quadratic algorithm for solving a system of mixed equalities and inequalities*, Mathematical Programming, 21 (1981), pp. 290–300.

[53] M. GERDTS, *Direct shooting method for the numerical solution of higher index DAE optimal control problems*, Journal of Optimization Theory and Applications, 117 (2003), pp. 267–294.

[54] M. GERDTS, *A nonsmooth Newton's method for control-state constrained optimal control problems*, Mathematics and Computers in Simulation, 79 (2008), pp. 925–936.

[55] M. GERDTS, *Optimal Control of Ordinary Differential Equations and Differential-algebraic Equations*, Habilitation, Universität Bayreuth, Bayreuth, 2006, http://www.mathematik.uni-wuerzburg.de/ gerdts/habilitation.pdf.

[56] M. GERDTS, *Global convergence of a nonsmooth newton method for control-state constrained optimal control problems*, SIAM Journal on Optimization, 19 (2008), pp. 326–350.

[57] I. GHERMAN AND V. SCHULZ, *Preconditioning of one-shot pseudo-timestepping methods for shape optimization*, Proceedings in Applied Mathematics and Mechanics, 5 (2005), pp. 741–742.

[58] A. GREENBAUM, *Iterative Methods for Solving Linear Systems*, Frontiers in Applied Mathematics, 17, SIAM, Philadelphia, 1997.

[59] L. GRIPPO, F. LAMPARIELLO, AND S. LUCIDI, *A nonmonotone line search technique for Newton's methods*, SIAM Journal on Numerical Analysis, 23 (1986), pp. 707–716.

[60] M. GRÖTSCHEL, S. O. KRUMKE, AND J. RAMBAU, *Online Optimization of Large Scale Systems*, Springer, Berlin, 2001.

[61] S. P. HAN, J.-S. PANG, AND N. RANGRAJ, *Globally convergent Newton methods for nonsmooth equations*, Mathematics of Operations Research, 17 (1992), pp. 586–607.

[62] R. F. HARTL, S. P. SETHI, AND G. VICKSON, *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Review, 37 (1995), pp. 181–218.

[63] S. B. HAZRA AND V. SCHULZ, *Simultaneous pseudo-timestepping for PDE-model based optimization problems*, BIT Numerical Mathematics, 44 (2004), pp. 457–472.

[64] S. B. HAZRA, V. SCHULZ, AND J. BREZILLON, *Simultaneous pseudo-time stepping for 3D aerodynamic shape optimization*, (2005), preprint.

[65] S. B. HAZRA, V. SCHULZ, J. BREZILLON, AND N. GAUGER, *Aerodynamic shape optimization using simultaneous pseudo-timestepping*, Journal of Computational Physics, 204 (2005), pp. 46–64.

[66] S. B. HAZRA AND V. SCHULZ, *Simultaneous pseudo-timestepping for aerodynamic shape optimization problems with constraints*, SIAM Journal on Scientific Computing, 28 (2006), pp. 1078–1099.

[67] M. HEINKENSCHLOSS, M. ULBRICH, AND S. ULBRICH, *Superlinear and quadratic convergence of affine-scaling interior-point Newton methods for problems with simple bounds without strict complementarity assumption*, Mathematical Programming, 86 (1999), pp. 615–635.

[68] D. J. HIGHAM, *Trust region algorithms and time step selection*, SIAM Journal on Numerical Analysis, 37 (1999), pp. 194–210.

[69] S. INCERTI, V. PARISI, AND F. ZIRILLI, *A new method for solving nonlinear simultaneous equations*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 779–789.

[70] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of Extremal Problems*, Vol. 6 of Studies in Mathematics and its Applications, Amsterdam, New York, Oxford, 1979, North-Holland Publishing Company.

[71] H. JIANG, M. FUKUSHIMA, L. QI, AND D. SUN, *A trust region method for solving generalized complementarity problems*, SIAM Journal on Optimization, 8 (1998), pp. 140–157.

[72] H. JIANG AND D. RALPH, *Global and local superlinear convergence analysis of Newton-type methods for semismooth equations with smooth least squares*, in Reformulation: Nonsmooth, Piecewise Smooth, Semismooth and Smoothing Methods, M. Fukushima and L. Qi, eds., Kluwer Academic Publishers, Dordrecht, 1999, pp. 181–209.

[73] C. KANZOW, *Some noninterior continuation methods for linear complementarity problems*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 851–868.

[74] C. KANZOW, *Inexact semismooth Newton methods for large-scale complementarity problems*, Optimization Methods and Software, 19 (2004), pp. 309–325.

[75] C. KANZOW AND A. KLUG, *On affine-scaling interior-point Newton methods for nonlinear minimization with bound constraints*, Computational Optimization and Applications, 35 (2006), pp. 177–197.

[76] C. KANZOW AND A. KLUG, *An interior-point affine-scaling trust-region method for semismooth equations with box constraints*, Computational Optimization and Applications, 37 (2007), pp. 329–353.

[77] C. KANZOW, N. YAMASHITA, AND M. FUKUSHIMA, *Levenberg-Marquardt methods for constrained non-linear equations with strong local convergence properties*, Journal of Computational and Applied Mathematics, 172 (2004), pp. 375–397.

[78] H. B. KELLER, *Global homotopies and Newton methods*, in: Recent Advances in Numerical Analysis (C. de Boor and G. H. Golub, eds.), Academic Press, New York, 1979, pp. 73–94.

[79] C. T. KELLEY, *Iterative Methods for Linear and Nonlinear Equations*, Frontiers in Applied Mathematics, Vol.16, SIAM, Philadelphia, 1995.

[80] C. T. KELLEY, *Solving Nonlinear Equations with Newton's Method*, Fundamentals of Algorithms, SIAM, Philadelphia, 2003.

[81] C. T. KELLEY, L.-Z. LIAO, L. QI, M. T. CHU, J. P. REESE, AND C. WINTON, *Projected pseudotransient continuation*, SIAM Journal on Numerical Analysis, 46 (2008), pp. 3071–3083.

[82] C. T. KELLEY AND D. E. KEYES, *Convergence analysis of pseudo-transient continuation*, SIAM Journal on Numerical Analysis, 35 (1998), pp. 508–523.

[83] R. W. KLOPFENSTEIN, *Zeros of nonlinear functions*, Journal of the Association for Computing Machinery, 8 (1961), pp. 366–373.

[84] D. A. KNOLL AND D. E. KEYES, *Jacobian-free Newton-Krylov methods: A survey of approaches and applications*, Journal of Computational Physics, 193 (2004), pp. 357–397.

[85] D. A. KNOLL AND P. MCHUGH, *Enhanced nonlinear iterative techniques applied to a nonequilibrium plasma flow*, SIAM Journal on Scientific Computing, 19 (1998), pp. 291–301.

[86] A. N. KOLMOGOROV AND S. V. FOMIN, *Elements of the theory of functions and functional analysis*, Dover Publications Inc., New York, 1999, Vol. 2, originally published by Graylock Press, Rochester, New York, 1957, 1961.

[87] M. KUBÍČEK, *Algorithm 502: Dependence of solution of nonlinear systems on a parameter*, ACM Transactions on Mathematical Software, 2 (1976), pp. 98–107.

[88] C. E. LEMKE AND J. T. HOWSON, *Equilibrium points of bimatrix games*, SIAM Journal on Applied Mathematics, 12 (1964), pp. 413–423.

[89] M. MACCONI, B. MORINI, AND M. PORCELLI, *Trust-region quadratic methods for nonlinear systems of mixed equalities equalities*, (2007), preprint.

[90] K. MALANOWSKI, *On normality of Lagrange multipliers for state constrained optimal control problems*, Optimization, 52 (2003), pp. 75–91.

[91] O. L. MANGASARIAN, *Absolute value programming*, Computational Optimization and Applications, 36 (2007), pp. 43–53.

[92] O. L. MANGASARIAN, *Absolute value equation solution via concave minimization*. Optimization Letters, 1 (2007), pp. 3–8.

[93] O. L. MANGASARIAN AND R. R. MEYER, *Absolute value equations*, Linear Algebra and its Applications, 419 (2006), pp. 359–367.

[94] J. M. MARTÍNEZ AND L. QI, *Inexact Newton's method for solving nonsmooth equations*, Journal of Computational and Applied Mathematics, 60 (1995), pp. 127–145.

[95] H. MAURER AND D. AUGUSTIN, *Sensitivity analysis and real-time control of parametric optimal control problems using boundary value methods*, in Online Optimization of Large Scale Systems, M. Grötschel, S. O. Krumke, and J. Rambau, eds., Springer, 2001, pp. 17–55.

[96] R. MIFFLIN, *Semismooth and semiconvex functions in constrained optimization*. SIAM Journal on Control and Optimization, 15 (1977), pp. 957-972.

[97] W. MULDER AND B. V. LEER, *Experiments with implicit upwind methods for the Euler equations*, Journal of Computational Physics, 59 (1985), pp. 232–246.

[98] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[99] J. V. OUTRATA AND J. ZOWE, *A Newton method for a class of quasi-variational inequalities*, Computational Optimization and Applications, 4 (1995), pp. 5–21.

[100] J.-S. PANG, *Newton's method for B-differentiable equations*, Mathematics of Operations Research, 15 (1990), pp. 311–341.

[101] J.-S. PANG, *A B-differentiable equation-based, globally and locally quadratically convergent algorithm for nonlinear programs, complementarity and variational inequality problems*, Mathematical Programming, 51 (1991), pp. 101–131.

[102] J.-S. PANG AND L. QI, *Nonsmooth equations: motivation and algorithms*, SIAM Journal on Optimization, 3 (1993), pp. 443–465.

[103] R. P. PAWLOWSKI, J. N. SHADID, J. P. SIMONIS, AND H. F. WALKER, *Globalization techniques for Newton-Krylov methods and applications to the fully-coupled solution of the Navier-Stokes equations*, SIAM Review, 46 (2006), pp. 700–721.

[104] A. PETROSYAN AND H. SHAHGHOLIAN, *Parabolic obstacle problems applied to finance*, in Recent Developments in Nonlinear Partial Differential Equations, Contemp. Math., 439 (2007), pp. 117–133.

[105] F. A. POTRA, *A superlinearly convergent predictor-corrector method for degenerate LCP in a wide neighborhood of the central path with $O(\sqrt{n}L)$-iteration complexity*, Mathematical Programming, Ser. A, 100 (2004), pp. 317–337.

[106] F. A. POTRA AND X. LIU, *Corrector-predictor methods for sufficient linear complementarity problems in a wide neighborhood of the central path*, SIAM Journal on Optimization, 17 (2006), pp. 871–890.

[107] F. A. POTRA, L. QI, AND D. SUN, *Secant methods for semismooth equations*, Numerische Mathematik, 80 (1995), pp. 305–324.

[108] F. A. POTRA AND S. J. WRIGHT, *Interior point methods*, Journal of Computational and Applied Mathematics, 124, (2000), pp. 281–302.

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

[110] L. QI, *Regular pseudo-smooth NCP and BVIP functions and globally and quadratically convergent generalized Newton methods for complementarity and variational inequality problems*, Mathematics of Operations Research, 24 (1999), pp. 440–471.

[111] L. QI, X. TONG, AND D. LI, *An active-set projected trust region algorithm for box constrained nonsmooth equations*, Journal of Optimization Theory and Applications, 120 (2004), pp. 601–625.

[112] L. QI, D. SUN, AND G. ZHOU, *A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities*, Mathematical Programming, 87 (2000), pp. 1–35.

[113] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Mathematical Programming, 58 (1993), pp. 353–367.

[114] W. C. RHEINBOLDT, *Numerical Analysis of Parametrized Nonlinear Equations*, John Wiley and Sons, New York, 1986.

[115] S. M. ROBINSON, *Local structure of feasible sets in nonlinear programming. III. Stability and sensitivity*, Mathematical Programming Study, 30 (1987), pp. 45–66.

[116] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, 2nd Edition, SIAM, Philadelphia, PA, 2003.

[117] Y. SAAD AND M.H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 856–869.

[118] Y. SAAD AND H. A. VAN DER VORST, *Iterative solution of linear systems in the 20th century*, Journal of Computational and Applied Mathematics, 123 (2000), pp. 1–33.

[119] H. SELLAMI AND S. M. ROBINSON, *Homotopies based on nonsmooth equations for solving nonlinear variational inequalities*, in: Nonlinear Optimization and Applications (G. Di Pillo and F. Giannessi, eds.), Plenum Press, New York, 1996.

[120] H. SELLAMI AND S. M. ROBINSON, *Implementation of a continuation method for normal maps*, Mathematical Programming, 76 (1997), pp. 563–578.

[121] H. SHAHGHOLIAN, *Free boundary regularity close to initial state for parabolic obstacle problem*, Trans. Amer. Math. Soc. 360 (2008), pp. 2077–2087.

[122] L. F. SHAMPINE, *Numerical Solution of Ordinary Differential Equations*, Chapman and Hall, New York, 1994.

[123] A. SHAPIRO, *On concepts of directional differentiability*, Journal of Optimization Theory and Applications, 66 (1990), pp. 477–487.

[124] S. SMALE, *A convergent process of price adjustment and global Newton methods*, Journal of Mathematical Economics, 3 (1976), pp. 1–14.

[125] S. SMALE, *Algorithms for solving equations*, in Proceeding of International Congress of Mathematicians, A. M. Gleason, ed., American Mathematics Society, Providence, Rhode Island, 1987, pp. 172–195.

[126] Y.-H. SONG, *Modern Optimization Techniques in Power Systems*, Kluwer Academic Publishers, Boston, USA, 1999.

[127] G. S. STELLING AND S. P. A. DUYNMEYER, *A staggered conservative scheme for every Froude number in rapidly varied shallow water flows*, International Journal for Numerical Methods in Fluids, 43 (2003), pp. 1329–1354.

[128] D. SUN, R. S. WOMERSLEY, AND H. QI, *A feasible semismooth asymptotically Newton method for mixed complementarity problems*, Mathematical Programming, 94 (2002), pp. 167–187.

[129] X. TONG AND L. QI, *On the convergence of a trust region method for solving constrained nonlinear equations with degenerate solutions*, Journal of Optimization Theory and Applications, 123 (2004), pp. 187–212.

[130] M. ULBRICH, *Semismooth newton methods for operator equations in function spaces*, SIAM Journal on Optimization, 13 (2003), pp. 805–841.

[131] M. ULBRICH, *Nonsmooth Newton-like Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, Habilitation, Technical University of Munich, Munich, 2002.

[132] M. ULBRICH, *Nonmonotone trust-region methods for bound-constrained semismooth systems of equations with applications to nonlinear mixed complementarity problems*, SIAM Journal on Optimization, 11 (2001), pp. 889–917.

[133] M. ULBRICH AND S. ULBRICH, *Superlinear convergence of affine-scaling interior-point newton methods for infinite-dimensional nonlinear problems with*

pointwise bounds, SIAM Journal on Control and Optimization, 38 (2000), pp. 1938–1984.

[134] D. WERNER, *Funktionalanalysis*, Springer, Berlin-Heidelberg-New York, 1995.

[135] A. J. WOOD AND B. F. WOLLENBERG, *Power Generation, Operation, and Control*, John Wiley and Sons, New York, USA, 1996.

[136] S. J. WRIGHT, *Primal-Dual Interior Point Methods*, SIAM, Philadelphia, PA, 1997.

[137] D. T. ZHU, *Affine scaling inexact generalized Newton algorithm with interior backtracking technique for solving bound-constrained semismooth equations*, Journal of Computational and Applied Mathematics, 187 (2006), pp. 227–252.