Affine-Scaling Methods for Nonlinear Minimization Problems and Nonlinear Systems of Equations with Bound Constraints

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von

Andreas Klug

aus

Dortmund

Würzburg, 2006

Eingereicht bei der Fakultät für Mathematik und Informatik am 12.05.2006

1. Gutachter: Prof. Dr. Christian Kanzow, Universität Würzburg
2. Gutachterin: Prof. Dr. Stefania Bellavia, Universität Florenz, Italien

Tag der mündlichen Prüfung: 28.07.2006

Contents

Chapter 1

Introduction

In this thesis we consider numerical methods for two different types of mathematical problems that both have feasible sets in the form of boxes. The first class of problems are the so called *nonlinear minimization problems with bound constraints*

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \Omega, \tag{P}$$

where the feasible set Ω is given by the box

$$\Omega := \{x \in \mathbb{R}^n : \ l_i \leq x_i \leq u_i \quad \forall i = 1, \ldots, n\}$$

and l_i and u_i denote the lower and upper bounds for x_i and $f : \mathcal{D} \longrightarrow \mathbb{R}$ is the objective function with domain $\mathcal{D} \subseteq \mathbb{R}^n$. Optimization problems of this type appear for example if reasonable solutions can only be expected in a particular area or if f is not defined in every point in $\mathbb{R}^n$ due to singularities for instance. Moreover, if one knows that a solution should exist in a certain area, this kind of information can be used by putting suitable bounds on the variables. In order to document that bound constraints make sense for many optimization problems we cite a part of the introduction of [13] by Conn, Gould and Toint that characterizes the given problem very well.

> Some authors [...] even claim that a vast majority of optimization problems should be considered from the point of view that their variables are indeed restricted to certain meaningful intervals, and should therefore be solved in conjunction with bound constraints. Fortunately, it is the simplest of the inequality constrained problems, because of its structure. On the other hand, in a way it is more complex than many equality type problems ...

As a simple economic example we consider a company that fabricates n products, whose production quantities are denoted by $x_1, \ldots, x_n \in \mathbb{R}$. The company wants

to minimize the total production costs, that are commonly described by

$$K(x) := K^{fix} + \sum_{i=1}^{n} x_i k_i^{var}(x),$$

where K^{fix} denotes the fixed costs and the functions $k_i^{var} : \mathbb{R}^n \to \mathbb{R}$ describe the variable costs for the ith product. Usually the optimal production quantities are determined by unconstrained minimization of K, c.f. [38]. But due to on hand orders that have to be delivered and nonnegativity of the production quantity there exists a minimum fabrication amount $m_i \geq 0$ of each product i. On the other hand limited capacities $c_i > 0$ for the products can lead to a maximal production quantity. Therefore an unconstrained consideration may lead to quantities that are not possible to fabricate. Thus it makes sense to solve problem

$$\text{minimize} \quad K(x) \quad \text{subject to} \quad m_i \leq x_i \leq c_i, \quad i = 1, \ldots, n,$$

taking the given restrictions explicitly into account.

The optimization problem (P) has attracted quite a few researchers during the last years, and a number of different methods for its solution may be found in [7, 8, 13, 14, 26, 27, 34, 47, 48, 49, 70]. The approach we follow in this work is typically called the *affine-scaling* interior-point Newton method. Following an observation by Coleman and Li [11, 12], these methods exploit the fact that the first order optimality conditions of (P) may be rewritten as a (bound constrained) nonlinear system of equations

$$G(x) = 0, \quad x \in \Omega,$$

where $G : \mathbb{R}^n \to \mathbb{R}^n$ is defined by

$$G(x) := D(x)\nabla f(x)$$

for a certain *scaling matrix* $D(x)$, i.e.

$$D(x) = \operatorname{diag}\big(d_1(x), \ldots, d_n(x)\big)$$

is a diagonal matrix with the components

$$d_i(x) := d_i^{CL}(x) := \begin{cases} x_i - l_i, & \text{if } [\nabla f(x)]_i \geq 0, \\ u_i - x_i, & \text{if } [\nabla f(x)]_i < 0 \end{cases}$$

for $i = 1, \ldots, n$. The corresponding method was shown to be locally quadratically convergent in [11, 12] under certain assumptions including *strict complementarity* of the solution x^* of problem (P), i.e., under the assumption that, for

all indices $i \in \{1, \ldots, n\}$, we have

$$x_i^* \in \{l_i, u_i\} \Longrightarrow [\nabla f(x^*)]_i \neq 0.$$

Unfortunately the rate of convergence can slow down, if the strict complementarity condition is violated. This effect was reported by Heinkenschloss et al. in [37] and was overcome by a modification of the scaling matrix, where the scaling can be switched for certain indices.

We follow this idea and introduce a new class of affine-scaling methods for the solution of the box constrained optimization problem (P). This new class differs from the previous works mainly by using a different scaling $D(x)$. To this end, we note that both the Coleman-Li matrix $D(x) = D^{CL}(x)$ and the Heinkenschloss et al. scaling $D(x) = D^{HUU}(x)$ are, in general, discontinuous even at a solution x^*. This makes it relatively difficult to predict the behaviour of Newton's method. Hence we suggest another scaling matrix which is continuous (in fact, Lipschitz continuous) around a solution of problem (P). It turns out that the use of locally Lipschitz continuous scaling matrices simplifies the algorithm to some extent and, in particular, allows a relatively short and straightforward convergence proof. Of central importance for our new scaling matrix, however, is the fact that we are able to identify the degenerate indices correctly, where an index i is called *degenerate* at a solution x^* of problem (P), if both $x_i^* \in \{l_i, u_i\}$ and $[\nabla f(x^*)]_i = 0$.

The second type of problems we consider are *nonlinear systems of equations with bound constraints*

$$F(x) = 0 \quad \text{subject to} \quad x \in \Omega, \tag{NE}$$

where the feasible set Ω is again given by

$$\Omega := \{x \in \mathbb{R}^n : \; l_i \leq x_i \leq u_i \quad \forall i = 1, \ldots, n\}$$

and where $F : \mathcal{D} \longrightarrow \mathbb{R}^n$ denotes a suitable function with domain $\mathcal{D} \subseteq \mathbb{R}^n$ which is at least semismooth. This type of problem is, just like the box constrained optimization problem (P), quite important for several reasons. In fact, in a number of applications, the mapping F is not defined outside the box Ω. In some other situations, the unconstrained problem $F(x) = 0$, $x \in \mathbb{R}^n$ might have solutions outside the box Ω, which have no meaning for the applications. And once more if solutions of the unconstrained problem are expected to be located in a certain area, this additional information can be exploited. Furthermore bound constrained nonlinear systems are often used as a reformulation of other problem types, for example chemical equilibrium problems, boundary value problems or (mixed) complementarity problems.

As an example we consider the boundary value problem (BVP) to find a function

$w \in C^2((0,1)) \cap C([0,1])$ such that

$$w'' = \frac{3}{2}w^2, \quad w(0) = 4, \quad w(1) = 1.$$

This problem has the two symbolic solutions

$$w_1(t) = \frac{4}{(1+t)^2}$$

and

$$w_2(t) = C_1^2 \left(\frac{1 - \mathrm{cn}(C_1 t - C_2, k)}{1 + \mathrm{cn}(C_1 t - C_2, k)} - \frac{1}{\sqrt{3}} \right),$$

where $\mathrm{cn}(\xi, k)$ denotes the Jacobian elliptic function with the modulus k. The constants are given by

$$k = \frac{1}{2}\sqrt{2+\sqrt{3}}, \quad C_1 = 4.30310990, \quad C_2 = 2.33464196,$$

where the last two are results of an iterative process, c.f. [63, p. 504]. Both solutions w_1 and w_2 are plotted in Figure 1.1 below. If one wants to solve the

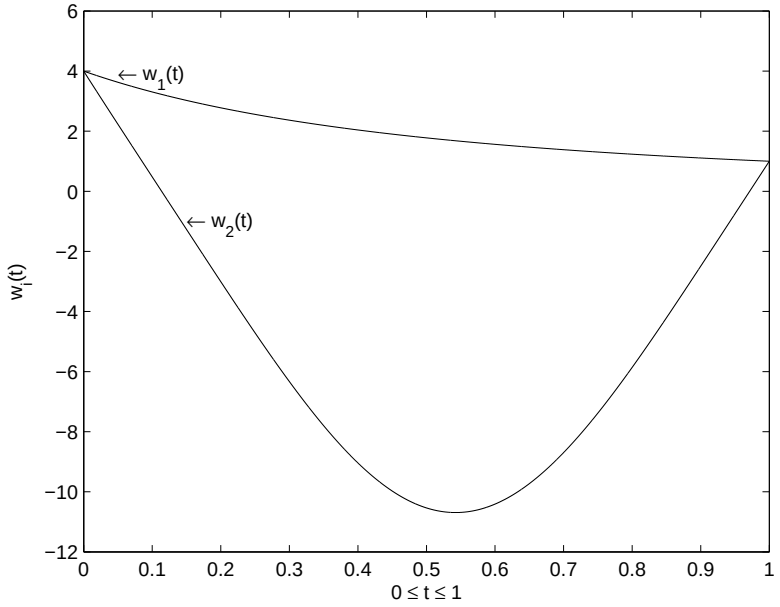

Figure 1.1: Plot of the symbolic BVP solutions

BVP numerically, a vector $x \in \mathbb{R}^n$ satisfying

$$x_k \approx w(t_k) \quad \text{with} \quad t_k := \frac{k-1}{n-1}, \; k = 1, \ldots, n$$

has to be computed. Assuming sufficient smoothness of w the standard $O(h^2)$ discretization arising from

$$w''(t) \approx \frac{w(t+h) - 2w(t) + w(t-h)}{h^2}$$

and the boundary values lead to the nonlinear system $F(x) = 0$ defined by

$$\begin{aligned} F_1(x) &= x_1 - 4, \\ F_k(x) &= 2x_k - x_{k-1} - x_{k+1} + \frac{3}{2}h^2 x_k^2, \qquad k = 2, \ldots, n-1, \\ F_n(x) &= x_n - 1 \end{aligned}$$

with $h = 1/(n-1)$ to determine the points $x_k \approx w(t_k)$. However the given BVP has two solutions and it is therefore useful to consider the bound constrained system

$$F(x) = 0, \quad l_i \le x_i \le u_i, \quad \forall i = 1, \ldots, n$$

to determine one specific solution. By use of the lower bounds $l_i := 0$ and upper bounds $u_i := \infty$, $i = 1, \ldots, n$ one can approximate the positive solution w_1 for instance. We will recall this example later in Section 4.5, where we will apply our affine-scaling method to the reformulation and compute a numerical solution of the BVP.

The unconstrained nonlinear system $F(x) = 0$ with $\Omega = \mathbb{R}^n$ and F continuously differentiable is discussed in several books including [1, 18, 20, 44, 45, 55]. Extensions of the classical Newton method for the solution of the unconstrained problem with F being semismooth may be found in [58, 57, 56, 29]. Despite the popularity of the unconstrained system of nonlinear equations, the number of references dealing with the box constrained problem (NE) (and with F being either smooth or nonsmooth) is still very limited. Currently, we are only aware of the papers [2, 3, 4, 5, 6, 39, 41, 46, 59, 64, 65]. Most of these papers, however, appeared during the last few years, and we believe that the box constrained problem (NE) is of increasing interest.

The main motivation for the method we propose later is a series of papers by Bellavia et al. [2, 3, 4, 5, 6]. These authors consider a class of affine-scaling interior-point methods for the solution of problem (NE) which have very good numerical properties. In particular, in our experience, the practical performance of these methods is better than the behaviour of the active-set type methods discussed in [39, 41, 59]. However, Bellavia et al. consider smooth equations only and assume that the Jacobian $F'(x)$ is nonsingular in order to show that a certain inner iteration is finite. Our aim is therefore to generalize their method to the class of semismooth equations (which is important if we want to apply

our method to complementarity problems, for example) without using the nonsingularity assumption. Moreover, we allow a more general choice of the scaling matrix and try to simplify the method and the corresponding convergence analysis by using a simple rule for the transition from global to local fast convergence.

The remaining thesis is organized as follows: In chapter 2 we prepare theoretical background material including optimality conditions and results from nonsmooth analysis. The bound constrained optimization problem (P) is considered in chapter 3. After a review of some numerical methods we present our affine-scaling approach and its theoretical and numerical properties. The emphasis lies on the local convergence properties. Chapter 4 deals with numerical methods for the problem (NE). In particular a new method for semismooth nonlinear systems with box constraints is proposed. Global and local convergence properties are established and the method is tested on various examples, especially smooth and semismooth reformulations of mixed complementarity problems. In chapter 5 a brief conclusion is carried out.

Closing this section a few words regarding our notation: For a vector $x \in \mathbb{R}^n$ we denote by x_i and sometimes by $[x]_i$ its ith component. If $F : \mathbb{R}^n \to \mathbb{R}^n$ is a vector-valued mapping, F_i is used for its ith component function. In the differentiable case $F'(x)$ denotes the Jacobian of F at a point $x \in \mathbb{R}^n$, whereas $\nabla F(x)$ is the transposed Jacobian. In particular, if $m = 1$, the gradient $\nabla F(x)$ is viewed as a column vector. If F is locally Lipschitz we write $\partial F(x)$ for the generalized Jacobian of F at x. Throughout this text, $\|\cdot\|$ denotes the Euclidean vector norm or the corresponding matrix norm, while $\|\cdot\|_\infty$ denotes the maximum norm. Furthermore, $P_\Omega(x)$ is the (Euclidean) projection of a vector $x \in \mathbb{R}^n$ onto the feasible set Ω. Note that this projection can be calculated quite easily since we are dealing with box constraints only. Given a matrix $A \in \mathbb{R}^{n\times n}$, we write A_i for the ith column of this matrix. If A is positive semidefinite, we write $A^{1/2}$ for its positive semidefinite square root. Moreover we often use the short-hand notation F_k for the mapping F evaluated at a point $x^k \in \mathbb{R}^n$. Finally $B_\varepsilon(x^*)$ stands for the open Euclidean ball of radius $\varepsilon > 0$ around the point $x^* \in \mathbb{R}^n$ and the Landau symbols are denoted by o and O.

Chapter 2

Theoretical Basics

In this chapter we provide the theoretical basics needed for our further considerations. On the one hand we consider optimality conditions for the bound constrained nonlinear program (P) and the already mentioned reformulation of the first order necessary condition in the form of a nonlinear system of equations. Since this nonlinear system is in general not continuously differentiable and since we want to allow semismooth functions later, some results from nonsmooth analysis are also presented.

2.1 Optimality Conditions

We consider the nonlinear minimization problem with bound constraints

$$\text{minimize } f(x) \quad \text{subject to} \quad x \in \Omega \tag{P}$$

with

$$\Omega := \{x \in \mathbb{R}^n : \; l_i \leq x_i \leq u_i \quad \forall i = 1, \ldots, n\},$$

lower bounds l_i, upper bounds u_i and a twice continuously differentiable objective function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. A simple necessary optimality condition for the problem (P) is given in the next Lemma, see also [27].

Lemma 2.1 *Let $x^* \in \Omega$ be a local minimum of the optimization problem* (P). *Then*

$$[\nabla f(x^*)]_i \begin{cases} = 0, & \text{if } \; l_i < x_i^* < u_i, \\ \geq 0, & \text{if } \; x_i^* = l_i, \\ \leq 0, & \text{if } \; x_i^* = u_i \end{cases} \tag{2.1}$$

holds.

As noted in the introduction, the first order necessary optimality condition (2.1) is equivalent to a nonlinear system of equations involving the Coleman-Li scaling matrix. This equivalence can be extended to more general scaling matrices. More precisely, we have the following result, cf. Heinkenschloss et al. [37].

Lemma 2.2 *Let $x^* \in \Omega$. Then x^* satisfies the first order optimality conditions* (2.1) *if and only if it is a solution of the nonlinear system of equations*

$$G(x) := D(x)\nabla f(x) = 0, \tag{2.2}$$

where $D(x) := diag\big(d_1(x), \ldots, d_n(x)\big)$ is any scaling matrix having the following properties on the feasible set Ω:

$$d_i(x) \begin{cases} = 0, & \text{if } x_i = l_i \text{ and } [\nabla f(x)]_i > 0, \\ = 0, & \text{if } x_i = u_i \text{ and } [\nabla f(x)]_i < 0, \\ \geq 0, & \text{if } x_i \in \{l_i, u_i\} \text{ and } [\nabla f(x)]_i = 0, \\ > 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

Motivated by Lemma 2.2, some methods for solving the bound constrained optimization problem (P) apply a Newton-type method to the corresponding nonlinear system (2.2), taking into account explicitly the simple bound constraints $x \in \Omega$.

In order to describe a sufficient optimality condition for the optimization problem (P), we introduce some index sets.

Definition 2.3 Let $x \in \Omega$ and the index set $I := \{1, \ldots, n\}$ be given. Then we call

$$I_0(x) := \big\{i \in I : x_i \in \{l_i, u_i\}\big\}$$

the *set of active indices* and

$$I_{00}(x) := \big\{i \in I_0(x) : [\nabla f(x)]_i = 0\big\}$$

the *set of degenerate indices*.

By use of these index sets strict complementarity of a solution x^* of problem (P) reduces to $I_{00}(x^*) = \emptyset$. Moreover the strong second order sufficiency condition can be described in the following form.

Definition 2.4 A point $x^* \in \Omega$ with (2.1) is said to satisfy the *strong second order sufficiency condition* (SSOSC for short) if

$$d^T \nabla^2 f(x^*) d > 0$$

holds for all nonzero $d \in T(x^*) := \{z \in \mathbb{R}^n : z_i = 0 \; \forall i \in I_0(x^*) \backslash I_{00}(x^*)\}$.

2.2 Nonsmooth Analysis

In this section we present a short introduction to the relevant parts of nonsmooth analysis. We consider two different aspects. The first one deals with generalized

derivatives and allows us to compute a suitable substitute for the not existing Jacobian of a nonsmooth but locally Lipschitz continuous function. This allows us the development of a Newton-type algorithm. The second aspect we consider, are semismooth functions. These are not continuously differentiable, but possess a property of smooth functions that ensures fast local convergence of our Newton-type iteration.

2.2.1 Generalized Jacobians

The reformulation of the optimality conditions presented in Lemma 2.2 is in general not differentiable in a solution x^* of (P) and we are therefore not able to compute Jacobians of the function G defined in that lemma. But under mild assumptions on G, we are able to compute a substitute for the Jacobian, called generalized Jacobian in the sense of Clarke [10]. As a minimum requirement we need the following definition.

Definition 2.5 Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open. A function $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ is called *locally Lipschitz continuous*, if for each $x \in \mathcal{O}$ two constants (depending on x) $\varepsilon = \varepsilon(x) > 0$ and $L = L(x) > 0$ exist, such that

$$\|F(y) - F(z)\| \leq L\|y - z\|, \qquad \forall\, y, z \in B_\varepsilon(x).$$

If F is a locally Lipschitz continuous function, we obtain from a theorem of Rademacher [10, Theorem 2.5.1] that F is differentiable almost everywhere in the following sense. Let $D_F := \{x \in \mathcal{O} : F \text{ is differentiable in } x\}$ denote the set of differentiable points. Then $\mathcal{O} \backslash D_F$ is a set of Lebesgue measure zero and hence for each $x \in \mathcal{O}$ there exists an arbitrary number of sequences $\{x^k\} \subseteq D_F$ with $x^k \longrightarrow x$. This legitimates the following definition.

Definition 2.6 Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ be locally Lipschitz continuous in $x^* \in \mathcal{O}$ and $D_F := \{x \in \mathcal{O} : F \text{ is differentiable in } x\}$ the set of differentiable points of F. Then the set

$$\partial_B F(x^*) := \left\{V \in \mathbb{R}^{m \times n} : \exists\, \{x_k\} \subseteq D_F \text{ with } x_k \to x^* \text{ and } F'(x_k) \to V\right\}$$

is called the *B-subdifferential* of F in x^*, and its convex hull

$$\partial F(x^*) := \operatorname{conv} \partial_B F(x^*)$$

is the *generalized Jacobian* of F in x^*. If $m = 1$, the set $\partial F(x^*)$ is called the *generalized gradient* of F in x^*.

For sake of clarity we give some elementary examples for the B-subdifferential and generalized Jacobians.

Examples 2.7

- The standard example is $F(x) := |x|$, which is locally Lipschitz continuous but not differentiable in $x = 0$. One can easily compute

$$\partial_B F(0) = \{-1, +1\} \quad \text{and} \quad \partial F(0) = [-1, +1].$$

- Another simple example is $F(x) := \max\{0, x\}$. This function is again locally Lipschitz continuous and not differentiable in $x = 0$. Here we get

$$\partial_B F(0) = \{0, +1\} \quad \text{and} \quad \partial F(0) = [0, +1].$$

- A multidimensional standard example is the Euclidian norm $F(x) := \|x\|_2$. This function is Lipschitz continuous but not differentiable in $x = 0$. For all $x \neq 0$ we obtain the Jacobian

$$F'(x) = \frac{x}{\|x\|},$$

  which has a Euclidean norm of one. We therefore get

$$\partial_B F(0) \subseteq \{x \in \mathbb{R}^n : \|x\|_2 = 1\}.$$

  To prove the opposite inclusion we consider an arbitrary $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$ and define a sequence $\{x^k\}$ by $x^k := x/k$. Then $x^k \to x$, F is differentiable in all x^k and $F'(x^k) \to x$. This yields $x \in \partial_B F(0)$ and

$$\partial_B F(0) = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}, \qquad \partial F(0) = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}.$$

- Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ be continuously differentiable. Then we obtain from Definition 2.6

$$\partial F(x) = \partial_B F(x) = \{F'(x)\}$$

  for all $x \in \mathbb{R}^n$. Hence both the generalized Jacobian in the sense of Clarke and the B-subdifferential are only consisting of the Jacobian of F in x.

- If $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is differentiable but not continuously differentiable the last conclusion does not hold. To see this we consider the function

$$F(x) := \begin{cases} x^2 \sin(\frac{1}{x}), & \text{if } x \neq 0, \\ 0, & \text{if } x = 0 \end{cases}$$

  with $F'(0) = 0$, whereas a simple calculation shows $\partial F(0) = [-1, +1]$.

For our further considerations we need some important properties of the B-subdifferential and the generalized Jacobian, which are presented in the next propositions. The first one is an existence result.

Proposition 2.8 *Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ be locally Lipschitz continuous and $x \in \mathcal{O}$ be given. Then the following holds:*

(a) The B-subdifferential $\partial_B F(x)$ is a nonempty and compact set.

(b) The generalized Jacobian $\partial F(x)$ is a nonempty, convex and compact set.

The last proposition ensures that the generalized Jacobian of a function contains at least one element. But the computation can be difficult. Sometimes it is easier to have a look at the generalized gradients of the component functions of F. In the smooth case the rows of the Jacobian are the gradients of the component functions. For the generalized Jacobian the following inclusion holds.

Proposition 2.9 *Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $f_1, \ldots, f_m : \mathcal{O} \longrightarrow \mathbb{R}$ be locally Lipschitz continuous. Then $F : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ with $F(x) = (f_1(x), \ldots, f_m(x))^T$ is also locally Lipschitz continuous and*

$$\partial F(x) \subseteq \left\{ (g_1, \ldots, g_m)^T : \; g_1 \in \partial f_1(x), \ldots, g_m \in \partial f_m(x) \right\}$$

holds.

This proposition, c.f. [10, Theorem 2.6.2.e], shows that the generalized Jacobian of a function is included in a set represented by the generalized gradients of its component functions. In order to apply this later we need some computational rules for generalized gradients which are quite similar to the smooth case. We start with a chain rule.

Proposition 2.10 *Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ locally Lipschitz continuous. Let $\mathcal{P} \subseteq \mathbb{R}^m$ be an open superset of $F(\mathcal{O})$, $g : \mathcal{P} \longrightarrow \mathbb{R}$ be locally Lipschitz continuous and $f := g \circ F$. Then f is locally Lipschitz continuous and*

$$\partial f(x) \subseteq \mathit{conv}\, \partial g(F(x)) \partial F(x) = \mathit{conv}\, \{vH : \; v \in \partial g(F(x)), H \in \partial F(x)\}$$

holds.

By use of the above chain rule, see also [10, Theorem 2.3.9, Theorem 2.6.6], it is possible to derive the subsequent rules for sums, products and quotients of locally Lipschitz continuous functions.

Proposition 2.11 *Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $f_1, \ldots, f_m : \mathcal{O} \longrightarrow \mathbb{R}$ be locally Lipschitz continuous. Then the following holds:*

(a) The weighted sum

$$\sum_{i=1}^{m} \alpha_i f_i$$

is locally Lipschitz continuous for all $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$ *and*

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

holds for all $\alpha_1, \ldots, \alpha_m \in \mathbb{R}$. *Equality of the sets holds, if all* f_i *with at most one exception are continuously differentiable.*

(b) The product $f_1 f_2$ *is locally Lipschitz continuous and*

$$\partial(f_1 f_2)(x) \subseteq f_2(x)\partial f_1(x) + f_1(x)\partial f_2(x).$$

holds.

(c) The quotient $\frac{f_1}{f_2}$ *is locally Lipschitz continuous and*

$$\partial \left(\frac{f_1}{f_2} \right)(x) \subseteq \frac{f_2(x)\partial f_1(x) - f_1(x)\partial f_2(x)}{f_2(x)^2}$$

holds, if $f_2(x) \neq 0$.

The results above can be found in [10, Propositions 2.3.3, 2.3.13, 2.13.14].

In the last part of Examples 2.7 we have seen that for continuously differentiable functions the generalized gradient reduces to a singleton. Conversely the following weaker assertion holds, see [10, Proposition 2.2.4].

Proposition 2.12 *Let* $\mathcal{O} \subseteq \mathbb{R}^n$ *be open and* $f : \mathcal{O} \longrightarrow \mathbb{R}$ *be locally Lipschitz continuous in a neighbourhood of* $x \in \mathcal{O}$. *If the (nonempty) generalized gradient reduces to a singleton, i.e.* $\partial f(x) = \{g\}$ *with a vector* $g \in \mathbb{R}^n$, *then* f *is differentiable in* x *with gradient* $\nabla f(x) = g$.

With this result we close our consideration of B-subdifferential and generalized Jacobians. For more information concerning these topics we refer to [10].

2.2.2 Semismooth Functions

The crucial part of the next two chapters in this work is the application of Newton-type methods to nonlinear systems of equations. The considered functions are not continuously differentiable, which is a very important assumption for local quadratic convergence of the classical Newton-method. Fortunately local quadratic convergence can be extended to a more general class of objective functions, that are called semismooth. In this section we define semismoothness and present some useful properties of semismooth functions. An important assumption for semismoothness is B-differentiability, so we define this first.

Definition 2.13 Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ be given. Then F is called

(a) *directionally differentiable* in $x \in \mathcal{O}$, if the limit

$$F'(x;d) := \lim_{t \to 0+} \frac{F(x+td) - F(x)}{t}$$

exists for all $d \in \mathbb{R}^n$.

(b) *B-differentiable* in $x \in \mathcal{O}$, if G is directionally differentiable in x and locally Lipschitz continuous near x.

Now we are able to define (strong) semismoothness of a function.

Definition 2.14 Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ B-differentiable. Then F is called

(a) *semismooth* in $x \in \mathcal{O}$, if

$$\|H_k d^k - F'(x;d^k)\| = o(\|d^k\|)$$

holds for all $d^k \to 0$ and $H_k \in \partial F(x+d^k)$.

(b) *strongly semismooth* in $x \in \mathcal{O}$, if

$$\|H_k d^k - F'(x;d^k)\| = O(\|d^k\|^2)$$

holds for all $d^k \to 0$ and $H_k \in \partial F(x+d^k)$.

Note that several other, but equivalent definitions of (strong) semismoothness are known, an overview and examples are given in [29, Definition 7.4.2, Theorem 7.4.3]. Standard examples for semismooth functions are:

Examples 2.15

- The minimum function

$$F(x_1, x_2) := \min\{x_1, x_2\}$$

is strongly semismooth on $\mathbb{R}^2$, see also [29, Proposition 7.4.7] for piecewise affine linear functions.

- The Fischer-Burmeister-function

$$F(x_1, x_2) := \sqrt{x_1^2 + x_2^2} - x_1 - x_2$$

is strongly semismooth on $\mathbb{R}^2$, see [29, p. 685].

- The norm function $F(x) := \|x\|_p$ is strongly semismooth on $\mathbb{R}^n$ for every $p \in \mathbb{N}$ and $p = \infty$, c.f. [29, Proposition 7.4.8].

The examples above are quite important for our further considerations, because the minimum and the Fischer-Burmeister functions are so called NCP-functions that are used to transform a nonlinear complementarity problem into a nonlinear system of equations. This will be a topic later.

In view of the defining expressions it seems to be quite natural that a smooth function is also semismooth. This and other sufficient conditions for (strong) semismoothness are expressed in the next proposition.

Proposition 2.16 *Let $\mathcal{O} \subseteq \mathbb{R}^n$ open, $x \in \mathcal{O}$ and $f : \mathcal{O} \longrightarrow \mathbb{R}$ be given.*

(a) If f is continuously differentiable in a neighborhood of x, then f is semismooth in x.

(b) If f is continuously differentiable and ∇f is locally Lipschitz continuous in a neighborhood of x, then f is strongly semismooth in x.

(c) If f is convex in a neighborhood of x, then f is semismooth in x.

This can be found in [28, Theorem 7.4.5]. Moreover a function $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ is (strongly) semismooth if and only if all component functions $F_i : \mathcal{O} \longrightarrow \mathbb{R}$ are (strongly) semismooth. This can be seen by exploiting the equivalence of $\|\cdot\|_2$ and $\|\cdot\|_\infty$ in Definition 2.14. Therefore parts (a) and (b) of the last proposition also hold for vector valued functions.

Similar to the last section we need some computational rules for semismooth functions as well. Again we start with a chain rule from [29, Proposition 7.4.4].

Proposition 2.17 *Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ B-differentiable. Let $\mathcal{P} \subseteq \mathbb{R}^m$ be an open superset of $F(\mathcal{O})$, $G : \mathcal{P} \longrightarrow \mathbb{R}^p$ be B-differentiable and $H := G \circ F$. Then the following holds:*

(a) If F is semismooth in $x \in \mathcal{O}$ and G is semismooth in $F(x)$, then H is semismooth in x.

(b) If F is strongly semismooth in $x \in \mathcal{O}$ and G is strongly semismooth in $F(x)$, then H is strongly semismooth in x.

An application of the chain rule proves (strong) semismoothness of weighted sums, products and quotients of (strongly) semismooth functions.

Proposition 2.18 *Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $f_1, f_2 : \mathcal{O} \longrightarrow \mathbb{R}$ be (strongly) semismooth in $x \in \mathcal{O}$. Then it holds, that:*

(a) The weighted sum $\alpha_1 f_1 + \alpha_2 f_2$ is (strongly) semismooth in x for all constants $\alpha_1, \alpha_2 \in \mathbb{R}$.

(b) The product $f_1 \cdot f_2$ is (strongly) semismooth in x.

(c) The quotient $\frac{f_1}{f_2}$ is (strongly) semismooth in x, if $f_2(x) \neq 0$.

As already mentioned semismooth functions share a property with continuously differentiable functions that is crucial for fast local convergence of our Newton-type methods. This property is described in the next proposition and can also be found in [29, Theorem 7.4.3.c].

Proposition 2.19 *Let $\mathcal{O} \subseteq \mathbb{R}^n$ be open and $F : \mathcal{O} \longrightarrow \mathbb{R}^m$ and $x \in \mathbb{R}^n$ be given.*

(a) If F is semismooth in x, then

$$\|F(x+d) - F(x) - Hd\| = o(\|d\|)$$

holds for all $d \to 0$ and $H \in \partial F(x+d)$.

(b) If F is strongly semismooth in x, then

$$\|F(x+d) - F(x) - Hd\| = O(\|d\|^2)$$

holds for all $d \to 0$ and $H \in \partial F(x+d)$.

With his result we close the collection of theoretical background material and consider now nonlinear programming problems with box constraints and apply a new affine scaling method to them.

Chapter 3

Nonlinear Minimization Problems

In this chapter we consider the box constrained nonlinear optimization problem

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \Omega, \tag{P}$$

where the feasible set Ω is given by

$$\Omega := \{x \in \mathbb{R}^n : \; l_i \leq x_i \leq u_i \quad \forall i = 1, \ldots, n\}$$

and l_i and u_i denote the lower and upper bounds, and $f : \mathcal{O} \longrightarrow \mathbb{R}$ is the objective function defined on an open set $\mathcal{O} \subseteq \mathbb{R}^n$ containing the feasible set Ω. Throughout this chapter, we assume that $l_i < u_i$ for all $i = 1, \ldots, n$ and that f is twice continuously differentiable with a locally Lipschitz continuous Hessian on the set $\mathcal{O}$. Moreover we assume that all bounds l_i and u_i are finite, but this is mainly a notational assumption since it simplifies many formulas in our analysis like the definitions of the scaling matrices that will be introduced later. However it is not difficult to see that all results remain true with a suitable redefinition of these scaling matrices, if either l_i or u_i or both are infinite for some indices $i \in \{1, \ldots, n\}$. Before we propose our affine-scaling method for box constrained nonlinear optimization problems, we consider some other known numerical methods for (P).

3.1 Numerical Methods

In the following subsections we describe several of the most important interior point methods for the bound constrained optimization problem (P). The emphasis lies on affine-scaling Newton-type methods since this will be the basic of our further considerations. For sake of consistency we adopt the notation used in the corresponding papers if necessary.

3.1.1 The Coleman-Li Reflective Newton Method

Affine-scaling methods were used only for linear programming at first. The basic idea is the following: Let an iterate $x^k \in \operatorname{int}\Omega$ be given and consider the steepest descent direction $p^k := -\nabla f(x_k)$. In order to prevent this direction from yielding onto the boundary it is multiplied with a scaling matrix $D(x^k) \in \mathbb{R}^{n\times n}$, such that the scaled steepest descent direction

$$p^k := -D(x^k)\nabla f(x^k)$$

is angled away from the nearest boundary. Coleman and Li extend this basic idea in [11] to develop a method for bound constrained optimization problems. Since it is possible to prove at most linear convergence for methods using the scaled steepest descent direction, Coleman and Li consider Newton-type directions as well. The basis of their considerations is the reformulation of the first order necessary optimality conditions in the form of a bound constrained nonlinear system of equations

$$G(x) := D(x)\nabla f(x) = 0, \quad x \in \Omega \tag{3.1}$$

using the *Coleman-Li scaling*

$$D(x) = \operatorname{diag}\big(d_1(x), \ldots, d_n(x)\big)$$

with diagonal elements

$$d_i(x) := d_i^{CL}(x) := \begin{cases} x_i - l_i, & \text{if } [\nabla f(x)]_i \geq 0, \\ u_i - x_i, & \text{if } [\nabla f(x)]_i < 0, \end{cases} \tag{3.2}$$

for $i = 1, \ldots, n$. This reformulation was later extended by Heinkenschloss et al. in [37] for more general types of scaling matrices, see also Lemma 2.2. In [11] a slightly different notation with squared scaling matrices $D(x)^2$ in (3.1) is used. We prefer the given one, which is also used in [19, 37, 42, 43] and omit the case of unbounded components.

Coleman and Li then obtain a Newton-type search direction from the nonlinear system (3.1). Since G is not differentiable if $d_i(x) = 0$ or $[\nabla f(x)]_i = 0$ holds for an $i \in 1, \ldots, n$ they use a scaled modification of the (non existing) Jacobian $\hat{B}(x) \in \mathbb{R}^{n\times n}$ and solve the linear system

$$\hat{B}(x^k)D(x^k)^{-1/2}p = -D(x^k)^{1/2}G(x^k)$$

to get a Newton-type search direction $p_N^k \in \mathbb{R}^n$. By use of this direction and a stepsize strategy of the form

$$x^{k+1} := x^k + t_k p_N^k$$

with an $t_k > 0$ such that $\{x^k\} \subseteq \operatorname{int}\Omega$ it is possible to prove local quadratic convergence of $\{x^k\}$, if $t_k \to 1$ sufficiently fast.

Moreover Coleman and Li introduce a new unconventional line search strategy. In the classical way a given descent direction is truncated to maintain strict feasibility, which can prevent that a full step is taken even near a solution. To overcome this problem Coleman and Li modify the search direction in the following way. If the direction yields outside the boundary of Ω it is reflected from the boundary back into the interior of the feasible set. This can be done several times. Then a steplength $t_k > 0$ along this piecewise linear reflected path $p_k(t)$ is computed. To get strictly feasible iterates it is necessary to ensure that no breakpoint, which is a point where the path is reflected from the boundary, is accepted as next iterate. A model algorithm can be described the following form.

Algorithm 3.1 (model interior-point reflective method)

(S.0) Choose $x^0 \in \operatorname{int}\Omega$ and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Determine a descent direction $p^k \in \mathbb{R}^n$ for f in x^k.

(S.3) Determine the piecewise linear reflective path $p_k(t)$.

(S.4) Determine a steplength $t_k > 0$ by an approximate piecewise line minimization of $f(x^k + p_k(t))$ with respect to t such that $p_k(t_k)$ does not correspond to a breakpoint.

(S.5) Set $x^{k+1} := x^k + p_k(t_k)$.

(S.6) Set $k \leftarrow k + 1$, and go to (S.1).

The termination criterion in the model algorithm above is left open, but later we will use $\|G(x^k)\| < \varepsilon$ with a small $\varepsilon > 0$. In steps (S.2) and (S.3) of the algorithm are still two open questions: The choice of a descent direction and the computation of a suitable steplength. To compute a steplength Coleman and Li decide to use the modified Armijo and Goldstein conditions

$$f(x^k + t_k p^k) < f(x_k) + \sigma_l \big[t_k \nabla f(x^k)^T p^k + \frac{1}{2} t_k^2 \min\{(p^k)^T \nabla^2 f(x^k) p^k, 0\}\big] \quad (3.3)$$

and

$$f(x^k + t_k p^k) > f(x_k) + \sigma_u \big[t_k \nabla f(x^k)^T p^k + \frac{1}{2} t_k^2 \min\{(p^k)^T \nabla^2 f(x^k) p^k, 0\}\big] \quad (3.4)$$

with constants $0 < \sigma_l < \sigma_u < 1$ and a given descent direction $p^k \in \mathbb{R}^n$ for f in x^k. They prove that in each iteration an interval $(t_l^{(k)}, t_u^{(k)})$ containing only a finite number of breakpoints exists, such that conditions (3.3) and (3.4) are satisfied for all stepsizes $t \in (t_l^{(k)}, t_u^{(k)})$ and recommend to use a simple bisection strategy for numerical computations. In the algorithm presented in [11] the condition (3.4) can be replaced with $t_k > t_{\min} > 0$ with a constant $t_{\min} > 0$ in each iteration, such that global convergence still holds.

The more complicated part is the computation of a descent direction that ensures global and fast local convergence. Two conditions are introduced that are essential for global convergence:

- The sequence of search directions $\{p^k\}$ is called *constrained compatible*, if $\{D(x^k)^{-1}p^k\}$ is bounded.
- The sequence of search directions $\{p^k\}$ satisfies the *consistency condition*, if $\nabla f(x^k)^T p^k \to 0$ implies $\|D(x^k)^{1/2}\nabla f(x^k)\| \to 0$.

Coleman and Li consider several directions including scaled steepest descent, some scaled Newton-type directions and solutions of subspace trust-region approaches that satisfy these two conditions and obtain the following global convergence result [11, Theorem 8]:

Theorem 3.2 *Let $\{x^k\}$ be generated by the interior-reflective path algorithm 3.1 with $\{p^k\}$ satisfying consistency and constrained compatibility conditions. Then*

$$\lim_{k\to\infty} \|D(x^k)\nabla f(x^k)\| = 0.$$

In order to obtain a local quadratic rate of convergence the search direction p^k has to be restricted further. Hence Coleman and Li develop a trust-region-type search direction that has two crucial properties. At first it satisfies the constrained compatibility and consistency conditions such that Theorem 3.2 yields global convergence of Algorithm 3.1. At second the method turns into a local Newton-type method with the direction p_N^k, for which local quadratic convergence can be proved. We want to avoid the rather technical details of the trust-region type search direction and describe only the most relevant parts. They consider another extended approximation $\hat{M}_k \approx \hat{B}(x^k)$ of the scaled Jacobian of G in x^k which is depending on a parameter $\tau_\epsilon > 0$ to handle with possible degenerate indices. If $\hat{M}_k$ is positive definite the Newton-type direction

$$p_N^k := -D(x^k)^{1/2}\hat{M}(x^k)^{-1}D(x^k)^{1/2}\nabla f(x^k) \tag{3.5}$$

is used, otherwise s^k is obtained by solving the trust-region subproblem

$$\begin{aligned} \text{minimize} \quad & m_k(p) := \nabla f(x^k)^T p + \frac{1}{2} p^T D(x^k)^{-1/2}\hat{M}_k D(x^k)^{-1/2} p \\ \text{subject to} \quad & \|D(x^k)^{-1/2}p\| \le \Delta_k,\ p \in \mathcal{S}_k \end{aligned} \tag{3.6}$$

where $\Delta_k > 0$ is the trust-region radius and $\mathcal{S}_k$ a suitable subspace of $\mathbb{R}^n$. A search direction computed in this way satisfies both consistency and constrained compatibility conditions and therefore global convergence holds. If this subspace is $\mathcal{S}_k = \mathbb{R}^n$ at each iteration, the following global and local convergence result holds for the interior point reflective method with search direction as described above, see also [11, Theorem 15, Corollary 16].

Theorem 3.3 *Let (P) be given and let $\{x^k\}$ be generated by Algorithm 3.1 with a search direction p^k given by* (3.5) *and* (3.6) *with $\mathcal{S}_k = \mathbb{R}^n$ and a steplength strategy satisfying* (3.3) *and* (3.4)*. Then the following holds:*

(a) Every limit point of $\{x^k\}$ satisfies the first order necessary condition.

(b) If τ_ϵ is sufficiently small and strict complementarity holds for a limit point x^* satisfying SSOSC then $\{x^k\}$ converges to x^* and the convergence rate is quadratic.

Theorem 3.3 shows the main advantages and a disadvantage of the Coleman-Li reflective affine-scaling method: It possesses strong local and global convergence results, but the strict complementarity assumption is needed for fast local convergence.

In fact Coleman and Li prove additional results concerning other subspaces for the trust-region strategy to compute descent directions. But our main interest lies in the local convergence properties so we do not consider these topics and proceed with another affine-scaling method presented by Coleman and Li as well.

3.1.2 The Coleman-Li Trust-Region Methods

In [12] Coleman and Li exploit again their reformulation of the first order necessary condition to develop an affine-scaling method for bound constrained optimization problems. The main difference to [11] described in the last subsection is the globalization technique. In [11] a reflective line search is established while in [12] two ellipsoidal trust-region globalizations are developed. Both methods use the Coleman-Li scaling from (3.2) as well. For sake of consistency we use the notation of the scaling matrices from [19, 37, 42, 43] again. The first method is called *double-trust-region method* and is of rather theoretical interest. In contrast to standard trust-region methods in unconstrained optimization the trust region size is not only controlled by the quality of the used model function, but also by feasibility requirements. The used quadratic model can be described in the following way. Let an iterate $x^k \in \operatorname{int}\Omega$ and a symmetric approximation $B_k \in \mathbb{R}^{n\times n}$ of $\nabla^2 f(x^k)$ be given. Define

$$M_k := B_k + C_k \quad \text{with} \quad C_k := D(x^k)^{-1/2}\operatorname{diag}(\nabla f(x^k))J_k^D D(x^k)^{-1/2},$$

where $J_k^D \in \mathbb{R}^{n\times n}$ is a modification of the in general not existing Jacobian of $D(\cdot)$ in x^k. Then the quadratic model is given by

$$m_k(p) := \nabla f(x^k)^T p + \frac{1}{2} p^T M_k p \tag{3.7}$$

and the subproblem for the double-trust-region method is

$$\text{minimize} \quad m_k(p) \quad \text{subject to } \|D(x^k)^{-1/2} p\| \leq \Delta_k \tag{3.8}$$

with the trust-region radius $\Delta_k > 0$. The double-trust-region method uses an exact solution $p_{ex}^k \in \mathbb{R}^n$ of (3.8) for the computation of a search direction. To maintain strict feasibility p_{ex}^k has to be truncated with a suitable stepsize $t_k \in (0, 1]$ that additionally minimizes the value of the model function m_k along the direction p_{ex}^k within the trust-region. The search direction is then given by $p^k := t_k p_{ex}^k$. This direction is accepted in dependence of two controlling quantities: The first one measures the quality of the quadratic model as an approximation to the objective function f and is given by

$$r_k^f := \frac{f(x^k + p^k) - f(x^k) + \frac{1}{2}(p^k)^T C_k p^k}{m_k(p^k)}. \tag{3.9}$$

The second quantity is defined to ensure a sufficient decrease of the model function in comparison to a truncated scaled steepest descent direction

$$p_{CP}^k := -\tau_{CP} D(x^k) \nabla f(x^k)$$

with a suitable steplength $\tau_{CP} > 0$ later called *Cauchy point*. If

$$r_k^c := \frac{m_k(p^k)}{m_k(p_{CP}^k)} > \beta \tag{3.10}$$

holds with a constant $\beta > 0$ this decreases is achieved. Moreover a condition later called *fraction of Cauchy decrease condition* that is important for global convergence is satisfied by the search direction when (3.10) holds. If both controlling quantities r_k^f and r_k^c are sufficiently large, the iteration is called successful, the direction p^k is accepted and the next iterate is given by $x^{k+1} := x^k + p^k$. The size of the trust-region Δ_k is also controlled by r_k^f and r_k^c. If one of these is too small, it is assumed that the reason for this is that one has trusted the model function on a too large region around x^k. Consequently the trust-region radius is decreased. Otherwise we can trust the model on a larger region and Δ_k is increased. The double-trust-region method has then the following form.

Algorithm 3.4 (double trust-region affine-scaling method)

(S.0) Choose $x^0 \in \operatorname{int}\Omega$, constants $0 < \rho_1, \beta < \rho_2 < 1$, $0 < \omega_1 < 1 < \omega_2$ and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Compute a solution $p_{ex}^k \in \mathbb{R}^n$ of (3.8).

(S.3) Compute $p^k := t_k p_{ex}^k$, r_k^f from (3.9) and r_k^c from (3.10).

(S.4) If $r_k^f > \rho_1$ and $r_k^c > \beta$ hold, we call the iteration "successful" and set $x^{k+1} := x^k + p^k$, otherwise we set $x^{k+1} := x^k$.

(S.5) If $r_k^f \leq \rho_1$ or $r_k^c \leq \beta$, then choose $\Delta_{k+1} \in (0, \omega_1 \Delta_k]$.
If $r_k^f \in (\rho_1, \rho_2)$, then choose $\Delta_{k+1} \in [\omega_1 \Delta_k, \Delta_k]$.
If $r_k^f \geq \rho_2$, then
if $r_k^c \geq \rho_2$, choose $\Delta_{k+1} \in [\Delta_k, \omega_2 \Delta_k]$,
if $r_k^c \in (\beta, \rho_2)$, choose $\Delta_{k+1} \in [\omega_1 \Delta_k, \Delta_k]$,
if $r_k^c < \beta$, choose $\Delta_{k+1} \in (0, \omega_1 \Delta_k]$.

(S.6) Set $k \leftarrow k + 1$, and go to (S.1).

For this algorithm global and local convergence results are proved in [12], but strict complementarity of a solution has to be assumed always. For a global convergence theorem we need the following additional assumptions.

- The objective function f is twice continuously differentiable and the *level set* $L_0 := \{x \in \Omega : f(x) \leq f(x^0)\}$ is compact.
- The search direction p^k satisfies the fraction of Cauchy decrease condition

$$m_k(p^k) < \beta m_k(p_{CP}^k), \quad x^k + p^k \in \operatorname{int} \Omega, \quad \|D(x^k)^{-1/2} p^k\| \leq \beta_0 \Delta_k \quad (3.11)$$

  with a constant $\beta_0 > 0$ and the Cauchy point p_{CP}^k.
- The quadratic model uses the exact Hessian of f, i.e. $B_k = \nabla^2 f(x^k)$ for all k.
- The search direction p^k satisfies

$$m_k(p^k) < \beta^q m_k(\tau^* p^k), \quad x^k + p^k \in \operatorname{int} \Omega, \quad \|D(x^k)^{-1/2} p^k\| \leq \beta_0^q \Delta_k$$

  with constants $\beta_0^q, \beta^q > 0$, where $m_k(\tau^* p^k)$ denotes the minimum value of the model function on the line τp^k with respect to feasibility and the trust-region.

Under this assumptions the following global convergence theorem holds.

Theorem 3.5 *Let (P) be given and the assumptions above be satisfied. Let $\{x^k\}$ be generated by the double-trust-region affine-scaling Algorithm 3.4 and all solutions to (P) satisfy the strict complementarity assumption. Then*

$$\lim_{k \to \infty} \|D(x^k)^{1/2} \nabla f(x^k)\| = 0.$$

In order to obtain fast local convergence a Newton-type direction similar to the reflective method has to be accepted in step (S.2). Let $p_N^k \in \mathbb{R}^n$ be a solution of the Newton-type linear system

$$\hat{M}_k D(x^k)^{-1/2} p = -D(x^k)^{1/2} \nabla f(x^k)$$

with $\hat{M}_k := D(x^k)^{1/2} M_k D(x^k)^{1/2}$. For fast local convergence we have to assume that a truncated direction $t_k p_N^k$ is accepted in step (S.2) of Algorithm 3.4 whenever $\|D(x^k)^{-1/2} p_N^k\| < \Delta_k$ and $t_k p_N^k$ satisfies (3.11) with a constant $\beta_0 \in (0,1)$. Then the following local convergence Theorem [12, Theorem 3.11] holds.

Theorem 3.6 *Let (P) be given and the assumptions of Theorem 3.5 be satisfied. Let $\{x^k\}$ be generated by the double-trust-region affine-scaling Algorithm 3.4 and let $x^* \in \Omega$ be a limit point of $\{x^k\}$ with regular $\hat{M}(x^*)$ that satisfies the strict complementarity assumption. If p_N^k is accepted as described above, then $\{x^k\}$ converges to x^* quadratically.*

For Algorithm 3.4 global and fast local convergence are proved in [12], but strict complementarity of a solution has to be assumed always. Moreover an exact solution of the subproblem (3.8) has to be computed in each iteration. Hence Coleman and Li propose a more practical trust-region method based on the double-trust-region algorithm. The main difference lies in the choice of the search direction and the update of the trust-region radius. Instead of a truncated exact solution of (3.8) a direction satisfying the fraction of Cauchy decrease condition (3.11), for example a truncated negative scaled gradient, is used. This also leads to a sufficient decrease of the quadratic model and is described as follows.

Algorithm 3.7 (practical trust-region affine-scaling method)

(S.0) Choose $x^0 \in \operatorname{int} \Omega$, constants $0 < \rho_1, \beta < \rho_2 < 1$, $0 < \omega_1 < 1 < \omega_2$, $\Lambda_l > 0$ and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Compute a direction $p^k \in \mathbb{R}^n$ that satisfies (3.11).

(S.3) Compute r_k^f from (3.9).

(S.4) If $r_k^f > \rho_1$ we call the iteration "successful" and set $x^{k+1} := x^k + p^k$, otherwise we set $x^{k+1} := x^k$.

(S.5) If $r_k^f \leq \rho_1$, then choose $\Delta_{k+1} \in (0, \omega_1 \Delta_k]$.
If $r_k^f \in (\rho_1, \rho_2)$, then choose $\Delta_{k+1} \in [\omega_1 \Delta_k, \Delta_k]$.
If $r_k^f \geq \rho_2$, then
if $\Delta_k > \Lambda_l$, choose $\Delta_{k+1} \in [\omega_1 \Delta_k, \Delta_k]$ or $\Delta_{k+1} \in [\Delta_k, \omega_2 \Delta_k]$,
if $\Delta_k \leq \Lambda_l$, choose $\Delta_{k+1} \in [\Delta_k, \omega_2 \Delta_k]$.

(S.6) Set $k \leftarrow k+1$, and go to (S.1).

For the practical affine-scaling Algorithm 3.4, Theorems 3.5 and 3.6 hold without assuming strict complementarity of the solutions, which is a big advantage of this method. However Heinkenschloss et al. observe in [37] that a violation of the strict complementarity assumption severely slows down the convergence speed. A theoretical reason is given later in section 3.2.

3.1.3 The Dennis-Vicente Trust-Region Method

The affine-scaling interior point trust-region method proposed by Dennis and Vicente in [19] is similar to the practical trust-region method of Coleman and Li in [12] and uses the same scaling matrix (3.2). The main difference lies in the computation of the search direction. Dennis and Vicente consider only global convergence properties of their method and therefore do not regard Newton-type directions to obtain fast local convergence. Hence the reformulation of the first order necessary optimality conditions in Lemma 2.2, used by the other affine-scaling methods to obtain a search direction, is here of minor importance. The affine-scaling approach of Dennis and Vicente is thus not motivated by Newton's method applied to the function G in (3.1) but by the property that the scaled steepest descent direction $-D(x^k)\nabla f(x^k)$ is angled away from the boundary. However (3.1) delivers a useful termination criterion. To obtain a direction we consider the following trust-region subproblem

$$\text{minimize} \quad m_k(p) \quad \text{subject to } \|S_k^{-1}p\| \le \Delta_k, \quad \sigma_k(l-x^k) \le p \le \sigma_k(u-x^k) \tag{3.12}$$

with the model function

$$m_k(p) := f(x^k) + \nabla f(x^k)^T p + \frac{1}{2} p^T B_k p,$$

$B_k \in \mathbb{R}^{n\times n}$ with $B_k \approx \nabla^2 f(x^k)$, $\sigma_k \in (0,\sigma]$ with a constant $\sigma \in (0,1)$, a regular matrix $S_k \in \mathbb{R}^{n\times n}$ and the trust-region radius $\Delta_k > 0$. The scaling happens here implicitly by the choice of the matrix S_k. Dennis and Vicente consider the cases

$$S_k := I \quad \text{and} \quad S_k := D(x^k).$$

The latter choice is similar to the trust-region methods of Coleman and Li, which correspond to $S_k := D(x^k)^{1/2}$. An exact computation of a solution of (3.12) can be avoided if the search direction satisfies a fraction of Cauchy decrease condition. Dennis and Vicente define the *Cauchy point* p_{CP}^k as the solution of the one dimensional problem

$$\begin{aligned} \text{minimize} \quad & m_k(p(\tau)) \\ \text{subject to} \quad & p(\tau) = -\tau D(x^k)^{2q}\nabla f(x^k), \quad \tau > 0 \\ & \|S_k^{-1}p(\tau)\| \le \Delta_k, \quad \sigma_k(l-x^k) \le p(\tau) \le \sigma_k(u-x^k), \end{aligned}$$

with an $q \geq \frac{1}{2}$. Their *fraction of Cauchy decrease condition* is then given by

$$m_k(0) - m_k(p^k) \geq \beta(m_k(0) - m_k(p^k_{CP})), \tag{3.13}$$

with a constant $\beta > 0$ and the feasibility requirements

$$\|S_k^{-1} p^k\| \leq \Delta_k, \quad \sigma_k(l - x^k) \leq p^k \leq \sigma_k(u - x^k). \tag{3.14}$$

With a direction satisfying these conditions (for example the Cauchy point p^k_{CP} itself) Dennis and Vicente develop a trust-region algorithm for (P). Acceptance of the search direction and the size of the trust-region radius are controlled by

$$r_k := \frac{f(x^k) - f(x^k + p^k)}{m_k(0) - m_k(p^k)}, \tag{3.15}$$

which again can be seen as a measure of the quality of the model function. The resulting method has the following form:

Algorithm 3.8 (Dennis-Vicente trust-region affine-scaling method)

(S.0) Choose $x^0 \in \operatorname{int} \Omega$, constants $0 < \sigma, \omega, \rho < 1$, $\varepsilon > 0$, $q \geq \frac{1}{2}$ and set $k := 0$.

(S.1) If $\|D(x^k)^q \nabla f(x^k)\| \leq \varepsilon$: STOP.

(S.2) Compute a direction $p^k \in \mathbb{R}^n$ that satisfies (3.13) and (3.14).

(S.3) Compute r_k given by (3.15).

(S.4) If $r_k > \rho$, we call the iteration "successful" and set $x^{k+1} := x^k + p^k$ and $\Delta_{k+1} \geq \Delta_k$,
otherwise we set $x^{k+1} := x^k$ and $\Delta_{k+1} := \omega \|p^k\|$.

(S.5) Set $k \leftarrow k + 1$, and go to (S.1).

The global convergence properties of this algorithm are described in the next Theorem.

Theorem 3.9 *Let (P) be given with f being continuously differentiable and bounded from below on the level set $L_0 := \{x \in \Omega : f(x) \leq f(x^0)\}$. Let $\{x^k\}$ be generated by the trust-region affine-scaling algorithm 3.8. If $\{B_k\}$ and $\{S_k^{-1} D(x^k)^q\}$ are uniformly bounded and L_0 is compact, then*

$$\lim_{k \to \infty} \|D(x^k)^q \nabla f(x^k)\| = 0.$$

If $S_k = I$ or $S_k = D(x^k)^q$ holds, then $\{S_k^{-1} D(x^k)^q\}$ is bounded if L_0 is compact. Hence the assumptions of this result are very mild compared with Theorem 3.5.

The main advantage of the affine-scaling methods of Dennis and Vicente is that global convergence is guaranteed by the fraction of Cauchy decrease condition (3.13) and (3.14) without assuming strict complementarity. Unfortunately steepest descent directions and Cauchy points do not provide fast local convergence, which is the main disadvantage of this method.

3.1.4 The Heinkenschloss-Ulbrich-Ulbrich Method

Heinkenschloss, Ulbrich and Ulbrich present in [37] a local affine-scaling interior point Newton method, which can be seen as an extension of the Coleman-Li methods from [11, 12]. Origin of their considerations are the local convergence properties of these Newton-type methods. A theoretical and numerical analysis of the bound constrained problem

$$\text{minimize} \quad f(x) := -\frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 - x_1^2 x_2 + x_1 \quad \text{subject to } 0 \le x_1, x_2 \le 1$$

shows that affine-scaling Newton methods using the slightly modified Coleman-Li scaling with

$$d_i(x) := d_i^{CL}(x) := \begin{cases} x_i - l_i, & \text{if } [\nabla f(x)]_i > 0, \\ u_i - x_i, & \text{if } [\nabla f(x)]_i < 0, \\ \min\{x_i - l_i, u_i - x_i\}, & \text{if } [\nabla f(x)]_i = 0, \end{cases} \tag{3.16}$$

for all $i = 1, \ldots, n$ do not converge quadratically to the solution $x^* = (0, 0)^T$ that satisfies the strong second order sufficiency condition but not strict complementarity. To overcome this disadvantage Heinkenschloss et al. introduce two modifications of the local Coleman-Li methods: A change of the scaling matrix and a projection strategy to maintain strict feasibility instead of truncating the direction.

Since a violation of strict complementarity slows down the local convergence speed Heinkenschloss et al. decide to estimate the indices for which strict complementarity does not hold and switch off the scaling for these indices. Their scaling matrix has the following form

$$D(x) = \operatorname{diag}\big(d_1(x), \ldots, d_n(x)\big)$$

with

$$d_i(x) := d_i^{HUU}(x) := \begin{cases} d_i^{CL}(x), & \text{if } |[\nabla f(x)]_i| < \min\big\{x_i - l_i, u_i - x_i\big\}^q \\ & \text{or } \min\big\{x_i - l_i, u_i - x_i\big\} < |[\nabla f(x)]_i|^q, \\ 1, & \text{otherwise,} \end{cases} \tag{3.17}$$

for $i = 1, \ldots, n$ with a constant $q > 1$. The requirements (2.3) are satisfied for this scaling. Hence the reformulation from Lemma 2.2 of the optimality conditions in the form of the nonlinear system

$$G(x) := D(x)\nabla f(x) = 0, \qquad x \in \Omega$$

is possible. Exploiting this reformulation again, a Newton-type search direction is obtained from the linear system

$$M(x^k)p = -G(x^k), \tag{3.18}$$

where $M(x^k)$ is an approximation to the possibly not existing Jacobian of G in x^k that is given by

$$M(x) := D(x)\nabla^2 f(x) + S(x)$$

with a diagonal matrix $S(x) := \operatorname{diag}\big(s_1(x), \ldots, s_n(x)\big)$ defined by

$$s_i(x) := s_i^{HUU}(x) := \begin{cases} |[\nabla f(x)]_i|, & \text{if } |[\nabla f(x)]_i| < \min\{x_i - l_i, u_i - x_i\}^q \\ & \text{or } \min\{x_i - l_i, u_i - x_i\} < |[\nabla f(x)]_i|^q, \\ 0, & \text{else,} \end{cases} \tag{3.19}$$

for $i = 1, \ldots, n$. Since the sequence $\{M^{-1}(x^k)\}$ is not bounded, which is shown by an example in [37], a second scaling of (3.18) with the matrix

$$W(x) := \operatorname{diag}\big(w_1(x), \ldots, w_n(x)\big)$$

with diagonal elements

$$w_i(x) := \frac{1}{d_i(x) + s_i(x)}, \tag{3.20}$$

for $i = 1, \ldots, n$ is necessary. The product $H(x) := W(x)M(x)$ is shown to be nonsingular with bounded inverse in a neighborhood of a solution of (P) satisfying SSOSC. Then the search direction p^k is obtained by solving the linear system

$$H(x^k)p = -W(x^k)G(x^k) \tag{3.21}$$

and a local algorithm based on this direction is given by:

Algorithm 3.10 (Heinkenschloss et al. affine-scaling method)

(S.0) Choose $x^0 \in \operatorname{int}(\Omega), \sigma \in (0,1)$, and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Compute $H(x^k) := W(x^k)D(x^k)\nabla^2 f(x^k) + W(x^k)S(x^k)$ with $D(x^k)$, $S(x^k)$ and $W(x^k)$ being given by (3.17), (3.19), and (3.20) respectively.

(S.3) Let $p^k \in \mathbb{R}^n$ be a solution of the linear system $H(x^k)p = -W(x^k)G(x^k)$.

(S.4) Compute $\sigma_k := \max\big\{\sigma, 1 - \|P_\Omega(x^k + p^k) - x^k\|\big\}$.

(S.5) Set $x^{k+1} := x^k + \sigma_k\big(P_\Omega(x^k + p^k) - x^k\big)$.

(S.6) Set $k \leftarrow k + 1$, and go to (S.1).

Strict feasibility of the iterates is ensured by a projection strategy using the projection mapping

$$P_\Omega(x) = \max\{l, \min\{u, x\}\}$$

from $\mathbb{R}^n$ on the feasible set Ω. The projected direction is then truncated by σ_k, which converges fast enough to 1 that the following local convergence result holds.

Theorem 3.11 *Let (P) be given with f twice continuously differentiable and $\nabla^2 f$ locally Lipschitz continuous. Let $\{x^k\}$ be generated by Algorithm 3.10 and let $x^* \in \Omega$ be a solution of (P) satisfying SSOSC. Then $\{x^k\}$ converges locally with order $\min\{q, 2\}$ to x^*, if $x^0 \in int\Omega$ is sufficiently close to x^*.*

With a choice $q \geq 2$ we obtain local quadratic convergence for Algorithm 3.10. Theorem 3.11 even holds if the linear system in step (S.3) is solved inexactly with some restriction on the error term, see [37] for more details.

The main advantage of this method is fast local convergence without assuming strict complementarity of the solution. Disadvantages are that the set of degenerate indices can be underestimated in (3.17) even close to a solution, which will be shown by some numerical examples later, and that global convergence results are missing.

With this conclusion we stop the consideration of the four affine-scaling methods for bound constrained optimization problems that are of huge importance for our further considerations. Of course other numerical methods with strong theoretical and practical convergence properties are known, but we restrict our detailed review to finite dimensional affine-scaling methods and give only few information about some other interesting methods.

- In [7] Bertsekas proposes globally convergent scaled projected gradient and Newton methods. The Newton-type method converges locally quadratically under a strict complementarity assumption.

- In [8] the limited memory BFGS method from unconstrained optimization is extended by Byrd et al. to the case of bound constraints and is considered from the numerical point of view. The software package LBFGS-B using this method is described in [70].

- Conn, Gould and Toint develop a trust-region method for bound constrained optimization problems in [13] that uses a generalized Cauchy point to obtain global convergence. Under the strict complementarity assumption this method turns into the unconstrained method after a finite number of iterations, which ensure fast local convergence. Numerical experiences are presented in [14].

- Globally convergent extensions of the last method are proposed by Friedlander at al. in [34] and by Lescrenier in [47], where the latter one is additionally locally quadratically convergent without a strict complementarity assumption.

- In [26] Facchinei et al. present a non-monotone active set Newton algorithm with stabilization technique, which converges globally and locally quadratically without assuming strict complementarity.

- A different non-monotone truncated Newton method was later proposed by Facchinei et al. in [27]. Global and local quadratical convergence are proved also without strict complementarity assumption.

- A pattern search method using feasible descent directions is described by Lewis and Torczon in [48] and a global convergence proof is given.

- A Newton-type method that possesses global and local superlinear convergence properties without assuming strict complementarity is developed by Lin and Moré in [49].

- In [62] descent methods from unconstrained optimization are extended by Schwarz and Polak to the bound constrained case by a projection strategy providing global convergence as well.

- Ulbrich and Ulbrich extend the Coleman-Li methods to the case of infinite dimensional problems and obtain a globally convergent affine-scaling method in [67].

- Another globally convergent infinite dimensional extension of the Coleman-Li methods is proposed by Ulbrich et al. in [67]. By use of a weak strict complementarity assumption local superlinear convergence is achieved.

3.2 Singularity Problems of Affine-Scaling Newton Methods

A careful convergence analysis in Heinkenschloss et al. [37] shows that the affine-scaling interior-point Newton method using the Coleman-Li scaling matrices from (3.16) is, in general, not quadratically convergent if strict complementarity does not hold at a local minimum x^*. The aim of this section is to give another reason for the failure of a whole class of affine-scaling methods in the absence of strict complementarity. In subsequent sections, this result will be used in order to motivate our new choice of the scaling matrix $D(x)$. As noted in Lemma 2.2, the first order necessary optimality condition (2.1) is equivalent to the nonlinear system of equations (2.2)

$$G(x) := D(x)\nabla f(x) = 0, \qquad x \in \Omega$$

using a suitable scaling matrix. Motivated by this observation, some methods for solving the bound constrained optimization problem (P) apply a Newton-type method to the corresponding nonlinear system (2.2) (taking into account explicitly the simple bound constraints $x \in \Omega$). Unfortunately, it turns out that the (generalized) Jacobian of the mapping G is singular under fairly mild

assumptions if strict complementarity does not hold. This is the main result we want to show in this section.

To this end, we assume that $D(x)$ is at least locally Lipschitz continuous around a local minimum x^* of problem (P). Then the mapping G is also locally Lipschitz. Hence we can compute its generalized Jacobian in the sense of Clarke [10]. By calculating the generalized Jacobian of the mapping G, we obtain the following negative result, where, in addition to our previous assumptions, we also assume that the scaling matrix $D(x) = \operatorname{diag}\big(d_1(x), \ldots, d_n(x)\big)$ has the property that

$$d_i(x) = 0, \quad \text{if } x_i \in \{l_i, u_i\}. \tag{3.22}$$

This is a rather natural condition since the components $d_i(x)$ usually represent an estimate for the distance of the component x_i to the boundary of the feasible set Ω.

Theorem 3.12 *Let x^* be a local minimum of (P) such that strict complementarity does not hold. Suppose further that $D(x) = \operatorname{diag}\big(d_1(x), \ldots, d_n(x)\big)$ is locally Lipschitz continuous and satisfies* (2.3) *and* (3.22)*. Then:*

(a) The ith component function G_i is differentiable with gradient $\nabla G_i(x^*) = 0$ for every index i where strict complementarity is violated.

(b) All elements of the generalized Jacobian $\partial G(x^*)$ are singular.

Proof. Recall that $G(x) := D(x)\nabla f(x)$ is locally Lipschitz continuous. Using the product rule for generalized gradients from Proposition 2.11 for the ith component $G_i(x) = d_i(x)[\nabla f(x)]_i$, it follows that

$$\partial G_i(x^*) \subseteq [\nabla f(x^*)]_i \partial d_i(x^*) + d_i(x^*)\partial[\nabla f(x^*)]_i.$$

Since strict complementarity does not hold at x^*, there is an index i_0 such that both $x^*_{i_0} \in \{l_{i_0}, u_{i_0}\}$ and $[\nabla f(x^*)]_{i_0} = 0$. For this particular component, we therefore get

$$\partial G_{i_0}(x^*) \subseteq \underbrace{[\nabla f(x^*)]_{i_0}}_{=0} \partial d_{i_0}(x^*) + \underbrace{d_{i_0}(x^*)}_{=0 \text{ by } (3.22)} \partial[\nabla f(x^*)]_{i_0} = \{0\}.$$

Since the generalized gradient $\partial G_{i_0}(x^*)$ is nonempty (see Proposition 2.8), it follows that $\partial G_{i_0}(x^*) = \{0\}$. But then Proposition 2.12 implies that G_{i_0} is differentiable in x^*, and its gradient is given by $\nabla G_{i_0}(x^*) = 0$. However, since we have

$$\partial G(x^*) \subseteq \big\{(g_1, \ldots, g_n)^T : \; g_1 \in \partial G_1(x^*), \ldots, g_n \in \partial G_n(x^*)\big\}$$

due to Proposition 2.9, it follows that each element $V \in \partial G(x^*)$ has a zero row and is therefore singular. This completes the proof of both statements. □

The previous proof shows that Theorem 3.12 actually holds under much weaker conditions. In fact, the local Lipschitz continuity of the scaling matrix $D(x)$ has been exploited only in the degenerate components. The other components, where strict complementarity is satisfied, are not really important. The only difficulty which arises without assuming local Lipschitz continuity of all components $d_i(x)$ is that we have to use an extended definition for a generalized Jacobian for non-Lipschitzian functions. However, whatever this extended definition might be, if we require that the ith row of such a more general Jacobian is equal to the gradient of the ith component function $G_i(x)$ whenever this function is differentiable at the current point (and this is a very natural condition), then it follows that Theorem 3.12 still holds. Moreover, the proof of Theorem 3.12 clearly shows that the statement also holds if property (3.22) is only satisfied at the local minimum x^* of problem (P).

We note that both the Coleman-Li scaling $d_i^{CL}(x)$ as well as the Heinkenschloss et al. scaling $d_i^{HUU}(x)$ satisfy (2.3). Moreover, $d_i^{CL}(x)$ has the property (3.22) which turned out to be quite negative in the discussion by Heinkenschloss et al. [37]. Here we introduce another scaling matrix

$$D^{MIN}(x) = \operatorname{diag}\big(d_1^{MIN}(x), \ldots, d_n^{MIN}(x)\big)$$

defined by

$$d_i^{MIN}(x) := \min\big\{x_i - l_i + \gamma \max\{0, -[\nabla f(x)]_i\}, u_i - x_i + \gamma \max\{0, [\nabla f(x)]_i\}\big\} \tag{3.23}$$

for $i = 1, \ldots, n$ and some constant $\gamma > 0$. This scaling matrix will play an important role in this work, and it has the advantage of being locally Lipschitz continuous. Moreover, it is easy to see that it satisfies (2.3). Furthermore, (3.22) also holds at a local minimum x^* of problem (P). Therefore, we obtain the following result as a direct consequence of Theorem 3.12 (and the previous notes).

Corollary 3.13 *Let x^* be a local minimum of (P) such that strict complementarity does not hold. Suppose further that $D(x) = D^{MIN}(x)$ denotes the scaling matrix with its components defined by* (3.23)*. Then all elements of the generalized Jacobian $\partial G^{MIN}(x^*)$ are singular.*

We note that we cannot apply Theorem 3.12 directly to the Coleman-Li scaling since $D^{CL}(x)$ is, in general, discontinuous (and therefore not locally Lipschitz continuous). Nevertheless, a related singularity problem was also observed for this scaling in Heinkenschloss et al. [37, pp. 621–622]. In fact, this observation

was the main motivation to introduce another scaling matrix. However, the Heinkenschloss et al. scaling is also discontinuous in general, even around a local minimum x^* (namely in those components where $[\nabla f(x^*)]_i = 0$). Since the behaviour of Newton's method is usually less predictable for discontinuous functions than for smooth ones, we prefer to work with scaling matrices which are at least locally Lipschitz continuous around a local minimum x^*. Hence the scaling matrix from (3.23) is a natural candidate, but in view of Corollary 3.13, it has to be modified in order to avoid the strict complementarity assumption.

3.3 Identification of Active and Degenerate Indices

The analysis from our previous section shows that it is quite important for fast local convergence to identify the degenerate indices in a local minimum of problem (P). The aim of this section is therefore to describe a simple and computationally efficient technique for the identification of these indices. To this end we use the notation of Definition 2.3 for the sets of *active indices*

$$I_0(x) := \big\{ i \in \{1, \ldots, n\} : \; x_i \in \{l_i, u_i\} \big\}$$

and *degenerate indices*

$$I_{00}(x) := \big\{ i \in I_0(x) : \; [\nabla f(x)]_i = 0 \big\}$$

in $x \in \Omega$ again. In order to identify the index set $I_{00}(x^*)$ exactly in a neighbourhood of a local minimum x^* of (P), we use an idea from Facchinei et al. [25] and specialize or modify their results to our situation. The fundamental definition from [25] is the following one.

Definition 3.14 A function $\rho : \mathbb{R}^n \longrightarrow \mathbb{R}_+$ is called an *identification function* for (P) if, for an isolated $x^* \in \Omega$ satisfying (2.1), the following properties hold:

(i) ρ is continuous in x^*,

(ii) $\rho(x^*) = 0$,

(iii) $\lim_{x \to x^*,\, x \neq x^*} \frac{\rho(x)}{\|x - x^*\|} = +\infty$.

Note that, in Definition 3.14, we call a vector x^* satisfying (2.1) *isolated* if there is a whole neighbourhood around this point such that x^* is the only vector satisfying the first order optimality conditions (2.1) in this neighbourhood. In our local convergence analysis to be presented in Section 3.5, this local uniqueness

condition is a consequence of another assumption (strong second order sufficiency condition) and, therefore, not as restrictive as it might appear in the beginning. Later in this section, we will give two examples of suitable identification functions. For the moment, however, we assume that we have such an identification function ρ. Using this identification function ρ, we define an estimate of the active indices $I_0(x)$ by

$$\begin{aligned} A_0(x) :&= \big\{ i \in I : \; x_i - l_i \leq \rho(x) \text{ or } u_i - x_i \leq \rho(x) \big\} \\ &= \big\{ i \in I : \; \min\{x_i - l_i, u_i - x_i\} \leq \rho(x) \big\}. \end{aligned} \tag{3.24}$$

Then we have the following result, which shows that $A_0(x)$ is equal to the set $I_0(x^*)$ in a sufficiently small neighbourhood of a solution x^* of (P), i.e., we are able to identify the set of active indices correctly.

Theorem 3.15 *Let ρ be an identification function for problem (P) and $x^* \in \Omega$ be an isolated vector satisfying* (2.1)*. Then there exists an $\epsilon > 0$, such that*

$$A_0(x) = I_0(x^*)$$

holds for all $x \in B_\epsilon(x^*)$.

Proof. The proof is similar to [25, Theorem 2.3] and is presented here for the sake of completeness.

First let $i \in I_0(x^*)$. Then we either have $x_i^* = l_i$ or $x_i^* = u_i$. Consider the case $x_i^* = l_i$ and define $g_i(x) := x_i - l_i$ (the argument is similar if $x_i^* = u_i$). Since $g_i(x^*) = 0$ and g_i is Lipschitz continuous with constant $L = 1$, we get

$$g_i(x) \leq g_i(x^*) + \|x - x^*\| = \|x - x^*\|$$

for all $x \in \mathbb{R}^n$. Using the definition of an identification function, we therefore obtain

$$x_i - l_i = g_i(x) \leq \|x - x^*\| \leq \rho(x) \quad \forall x \in B_{\epsilon_1}(x^*)$$

for some $\epsilon_1 > 0$ sufficiently small. Hence we have $i \in A_0(x)$ for all $x \in B_{\epsilon_1}(x^*)$.

Conversely, take an arbitrary index $i \notin I_0(x^*)$. Then we have $l_i < x_i^* < u_i$. Using $\rho(x^*) = 0$ and a continuity argument, it follows that $\rho(x) < \min\{x_i - l_i, u_i - x_i\}$ for all $x \in B_{\epsilon_2}(x^*)$ for some $\epsilon_2 > 0$ sufficiently small (note that the choice of ϵ_2 depends on the index i, but since there are only finitely many $i \in \{1, \ldots, n\}$, we may choose $\epsilon_2 > 0$ independent of i). Hence $i \notin I_0(x^*)$ implies $i \notin A_0(x)$ for all $x \in B_{\epsilon_2}(x^*)$, and this is equivalent to $A_0(x) \subseteq I_0(x^*)$ for all $x \in B_{\epsilon_2}(x^*)$.
Using $\epsilon := \min\{\epsilon_1, \epsilon_2\}$, we therefore obtain the desired result. □

Now we are able to estimate the active constraints exactly, but since we want to identify the degenerate ones, we also use the set

$$A_+(x) := \{i \in A_0(x) : \ \lambda_i(x) > \rho(x)\} \tag{3.25}$$

where $\lambda(x)$ is a *multiplier function*, i.e., $\lambda(x)$ is continuously differentiable (local Lipschitz continuity would be enough for our purpose) and has the property that

$$\lambda(x^*) = \lambda^*$$

for any vector x^* satisfying (2.1) and the corresponding (unique) Lagrange multiplier λ^* for problem (P). The interested reader is referred to [22, 21] for some suitable examples of multiplier functions. Note that these multiplier functions can be evaluated quite easily in the case of bound constrained optimization problems.

Theorem 3.16 *Let ρ be an identification function for problem (P) and $x^* \in \Omega$ be an isolated vector satisfying* (2.1)*. Then there exists an $\epsilon > 0$, such that*

$$A_+(x) = I_0(x^*) \backslash I_{00}(x^*)$$

holds for all $x \in B_\epsilon(x^*)$.

Proof. The technique of proof is taken from [25, Theorem 2.4] and included here for the sake of clarity.

First consider an index $i \in I_0(x^*) \backslash I_{00}(x^*)$. Since $i \in I_0(x^*)$, Theorem 3.15 shows that $i \in A_0(x)$ for all x sufficiently close to x^*. Furthermore, since $i \notin I_{00}(x^*)$ and $I_{00}(x^*)$ may be rewritten as $I_{00}(x^*) = \{i \in I_0(x^*) : \ \lambda_i^* = 0\}$ in terms of the multipliers λ_i^*, we have $\lambda_i(x^*) = \lambda_i^* > 0$, whereas $\rho(x^*) = 0$ holds. By continuity, this implies $\lambda_i(x) > \rho(x)$ for all $x \in B_{\epsilon_1}(x^*)$ for a suitable constant $\epsilon_1 > 0$, so that $i \in A_+(x)$.

To prove the converse inclusion, suppose that $i \in I_{00}(x^*)$. Then $\lambda_i(x^*) = \lambda_i^* = 0$. Moreover, since the multiplier function is continuously differentiable and, therefore, locally Lipschitz continuous around x^*, there is a constant $c > 0$, such that

$$\lambda_i(x) \leq |\lambda_i(x) - \lambda_i(x^*)| \leq \|\lambda(x) - \lambda(x^*)\| \leq c\|x - x^*\| \leq \rho(x)$$

for all x sufficiently close to x^*. Hence we have $i \notin A_+(x)$ for all these x. Hence $A_+(x) \subseteq I_0(x^*) \setminus I_{00}(x^*)$ for all $x \in B_{\epsilon_2}(x^*)$ and a suitable constant $\epsilon_2 > 0$. Consequently, the statement holds with $\epsilon := \min\{\epsilon_1, \epsilon_2\}$. □

Using Theorems 3.15 and 3.16, it follows that

$$A_{00}(x) := A_0(x) \backslash A_+(x) \tag{3.26}$$

is an exact estimation of the set of degenerate indices in the sense that an $\epsilon > 0$ exists, such that

$$A_{00}(x) = I_{00}(x^*) \tag{3.27}$$

holds for all $x \in B_\epsilon(x^*)$, where x^* is any isolated vector satisfying the optimality conditions (2.1).

Hence we have reached our goal provided that we have an identification function ρ. In the remaining part of this section, we therefore introduce two suitable mappings, which turn out to be identification functions under certain assumptions. The main assumption that will be used here is the strong second order sufficiency condition SSOSC, which is given in Definition 2.4. We note that weaker conditions are possible for the definition of identification functions, however, the strong second order sufficiency condition will also be used in our local convergence analysis of Section 3.5, so this condition is needed in any case.

Note that SSOSC is equivalent to saying that the submatrix $\nabla^2 f(x^*)_{\bar{J}\bar{J}}$ is positive definite, where

$$J := I_0(x^*) \setminus I_{00}(x^*) \quad \text{and} \quad \bar{J} := \{1, \ldots, n\} \setminus J. \tag{3.28}$$

We therefore get the following consequence from the definition of SSOSC.

Lemma 3.17 *Let $x^* \in \Omega$ be a point satisfying* (2.1) *and SSOSC. Then the vectors*

$$e_i \ (i \in J) \quad \textit{and} \quad [\nabla^2 f(x^*)]_i \ (i \in \bar{J})$$

are linearly independent, where the index sets J and $\bar{J}$ are defined in (3.28)*, and e_i denotes the ith unit vector in $\mathbb{R}^n$.*

Proof. Consider an arbitrary linear combination

$$\sum_{i \in J} \alpha_i e_i + \sum_{i \in \bar{J}} \alpha_i [\nabla^2 f(x^*)]_i = 0. \tag{3.29}$$

Without loss of generality, we may assume that $J = \{1, \ldots, r\}$ with $r := |J|$. Then we can rewrite (3.29) as

$$M\alpha = 0 \quad \text{with} \quad M := \begin{pmatrix} I_r & \nabla^2 f(x^*)_{J\bar{J}} \\ 0 & \nabla^2 f(x^*)_{\bar{J}\bar{J}} \end{pmatrix},$$

where $\alpha := (\alpha_1, \ldots, \alpha_n)^T$. In view of our assumption, however, the block matrix $\nabla^2 f(x^*)_{\bar{J}\bar{J}}$ is positive definite and, therefore, nonsingular. This implies that the matrix M is also nonsingular. Consequently, we obtain $\alpha = 0$, thus giving the desired result. □

We now present our first identification function.

Theorem 3.18 *Let $x^* \in \Omega$ be a point satisfying* (2.1) *and SSOSC. Define*

$$\rho_1(x) := \sqrt{\|\phi_1(x)\|}$$

with

$$\phi_1(x) := x - P_\Omega\big(x - \nabla f(x)\big).$$

Then ρ_1 is an identification function for problem (P).

Proof. It is obvious or well-known that ρ_1 satisfies the two conditions (i) and (ii) of Definition 3.14. In order to verify requirement (iii), we note that SSOSC and [28, Proposition 6.2.4] imply that there is a constant $\gamma > 0$ such that

$$\|x - x^*\| \le \gamma \|\phi_1(x)\| \tag{3.30}$$

holds for all x in a sufficiently small neighbourhood of x^*. More precisely, note that our mapping ϕ_1 is identical to what is called the natural residual in [28], and that SSOSC implies (for box constrained optimization problems) strong regularity in the sense of Robinson (see [60]). However, strong regularity implies semistability (see [28, p. 434]), and therefore [28, Proposition 6.2.4] can be applied.
As a consequence of (3.30), we get

$$\frac{\rho_1(x)}{\|x - x^*\|} = \frac{\sqrt{\|\phi_1(x)\|}}{\|x - x^*\|} \ge \frac{\sqrt{\|\phi_1(x)\|}}{\gamma\|\phi_1(x)\|} = \frac{1}{\gamma\sqrt{\|\phi_1(x)\|}} \to +\infty$$

for $x \to x^*, x \ne x^*$. Hence ρ_1 is an identification function. □

Our second identification function is given by

$$\rho_2(x) := \sqrt{\|\phi_2(x)\|}, \tag{3.31}$$

where the components $\phi_i^{(2)}$ of ϕ_2 are defined by

$$\phi_i^{(2)}(x) := 2x_i - l_i - u_i - |x_i - l_i - [\nabla f(x)]_i| + |x_i - u_i - [\nabla f(x)]_i|, \quad i = 1, \ldots, n. \tag{3.32}$$

It turns out, however, that ρ_2 is not much different from ρ_1. In fact, an elementary calculation shows that $\phi_2 = 2\phi_1$. Hence we immediately obtain the following result from Theorem 3.18.

Theorem 3.19 *Let $x^* \in \Omega$ be a point satisfying* (2.1) *and SSOSC. Let ρ_2 be defined as in* (3.31), (3.32). *Then ρ_2 is an identification function for problem (P).*

We close this section by noting that both Theorems 3.18 and 3.19 hold under weaker assumptions. In fact, the central result used in order to prove Theorem 3.18 was Proposition 6.2.4 from [28], and this result holds for any isolated vector x^* satisfying (2.1) and a condition called *semistability* in [28]. We refer the interested reader to [28] for further details on semistability.

3.4 Description of the Method

In this section we present our affine-scaling interior-point Newton method for the solution of the bound constrained optimization problem (P). Basically it is a Newton-type method applied to the reformulation (2.2) of the optimality conditions (2.1) for a suitable scaling matrix $D(x)$. The condition $x \in \Omega$ is guaranteed by generating strictly feasible iterates only. This, in turn, is done by incorporating a projection step as described, for example, in [37], although other choices would also be possible, see, e.g., [40].

In order to state our choice of the scaling matrix, we assume that we have an identification function ρ which allows us to define a set $A_{00}(x)$ via (3.26) which then identifies the set of degenerate indices $I_{00}(x^*)$ in a neighbourhood of a local minimum x^* of problem (P). Examples of suitable functions ρ having this property under the SSOSC assumption were given in Section 3.3.

Now, having a suitable identification function and a corresponding set $A_{00}(x)$, we define our scaling matrix by $D(x) = \operatorname{diag}\big(d_1(x), \ldots, d_n(x)\big)$ with

$$d_i(x) := \begin{cases} 1, & \text{if } i \in A_{00}(x), \\ d_i^{MIN}(x), & \text{if } i \notin A_{00}(x), \end{cases} \tag{3.33}$$

where the latter components are given by

$$d_i^{MIN}(x) := \min\big\{x_i - l_i + \gamma \max\{0, -[\nabla f(x)]_i\}, u_i - x_i + \gamma \max\{0, [\nabla f(x)]_i\}\big\}$$

for all $i = 1, \ldots, n$ with a constant $\gamma > 0$. Eventually, this definition differs from the one in (3.23) only in the degenerate indices. Then it is easy to see that $D(x)$ has the property (2.3). Moreover, we will see below that it is locally Lipschitzian around a local minimum x^* of (P) under suitable assumptions. However, it does not have the natural property from (3.22). On the other hand, in view of Corollary 3.13, we know that this property must be violated in order to have a chance to get nonsingular (generalized) Jacobians in the absence of strict complementarity.

We summarize these observations and some related properties in the following result.

Lemma 3.20 *Let x^* be an isolated vector satisfying* (2.1) *and suppose that $A_{00}(x) = I_{00}(x^*)$ holds in a neighbourhood of x^*. Then the scaling matrix $D(x)$ defined in* (3.33) *is locally Lipschitz continuous and strongly semismooth in a neighbourhood of x^*.*

Proof. Note that $D(x)$ is locally Lipschitz and strongly semismooth if and only if each component function $d_i(x)$ is locally Lipschitz and strongly semismooth. Therefore, the local Lipschitz property follows simply from the fact that the set $A_{00}(x)$ does not change locally. Similarly, this also implies the strong semismoothness of each $d_i(x)$ since the min- and max-functions and the composition of strongly semismooth functions are known to be strongly semismooth, see [32]. □

We note that the scaling matrix $D(x)$ is not necessarily Lipschitz continuous if we are far away from a local minimum x^* of (P). In fact, in this case the function might even become discontinuous. However, since we are only interested in the local analysis, Lemma 3.20 states a desirable property of our scaling matrix $D(x)$ around a local minimum x^* of (P). In general, this property holds neither for the Coleman-Li scaling $D^{CL}(x)$ nor for the Heinkenschloss et al. matrix $D^{HUU}(x)$.

Having defined our scaling matrix $D(x)$, we now want to apply a Newton-type method to the corresponding function $G(x) = D(x)\nabla f(x)$. The problem is that this mapping is not differentiable everywhere. As a suitable replacement of the Jacobian, we take

$$M(x) := D(x)\nabla^2 f(x) + S(x) \tag{3.34}$$

where $S(x) := \operatorname{diag}\big(s_1(x), \ldots, s_n(x)\big)$ is a diagonal matrix with

$$s_i(x) \approx d_i'(x)[\nabla f(x)]_i$$

($d_i'(x)$ being the partial derivative of the mapping d_i with respect to the component x_i) being given by

$$s_i(x) := \begin{cases} 0, & \text{if } i \in A_{00}(x), \\ \delta_i[\nabla f(x)]_i \text{ for an arbitrary } \delta_i \in \partial d_i(x), & \text{if } i \notin A_{00}(x). \end{cases} \tag{3.35}$$

Note that the entry $s_i(x)$ of the matrix $S(x)$ corresponds to the exact derivative of the mapping $d_i(x)[\nabla f(x)]_i$ at a continuously differentiable point. In general, we have the following simple but important result.

Theorem 3.21 *Let x^* be an isolated vector satisfying* (2.1) *and suppose that $A_{00}(x) = I_{00}(x^*)$ holds in a neighbourhood of x^*. Then the function $G(x) := D(x)\nabla f(x)$ with $D(x)$ being defined by* (3.33) *is strongly semismooth in a neighbourhood of x^*. Moreover, every element $M(x) \in \partial_B G(x)$ has a representation of the form* (3.34) *with $S(x)$ being the matrix from* (3.35).

Proof. Since the product of two strongly semismooth functions in again strongly semismooth, the first statement follows from Lemma 3.20 together with our general smoothness assumption on the mapping f. The remaining statements follow directly from the definition of the B-subdifferential, see Definition 2.6 (note, however, that, usually, (3.34), (3.35) contain more elements than those belonging to $\partial_B G(x)$). □

We are now in the position to state our Newton-type method for the solution of the bound constrained optimization problem (P).

Algorithm 3.22 (Projected Affine-Scaling Interior-Point Newton Method)

(S.0) Choose $x^0 \in \operatorname{int}(\Omega), \sigma \in (0,1)$, and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Compute $M(x^k) := D(x^k)\nabla^2 f(x^k) + S(x^k)$ with $D(x^k)$ and $S(x^k)$ being given by (3.33) and (3.35), respectively.

(S.3) Let $p^k \in \mathbb{R}^n$ be a solution of the linear system $M(x^k)p = -G(x^k)$.

(S.4) Compute $\sigma_k := \max\{\sigma, 1 - \|P_\Omega(x^k + p^k) - x^k\|\}$.

(S.5) Set $x^{k+1} := x^k + \sigma_k\big(P_\Omega(x^k + p^k) - x^k\big)$.

(S.6) Set $k \leftarrow k + 1$, and go to (S.1).

Note that steps (S.4) and (S.5) obviously guarantee the strict feasibility of all iterates x^k.

There are two differences between Algorithm 3.22 and the corresponding method from Heinkenschloss et al. [37]: First, we use a different way to compute the entries $d_i(x)$ and $s_i(x)$ by using the identification results from Section 3.3. Second, we do not use a further scaling of the matrices $M(x)$ as done in [37]. This further scaling was important in [37] in order to carry out a local convergence analysis. In particular, the exact identification result incorporated in our method turns out to be quite helpful also in order to simplify the local convergence analysis. This will be done in the following section.

3.5 Local Convergence Analysis

The aim of this section is to show that Algorithm 3.22 is locally quadratically convergent under the SSOSC assumption; in particular, we do not need the strict complementarity condition. Hence the local convergence result is identical to the one shown in Heinkenschloss et al. [37] for their algorithm. However,

our method of proof is completely different. Rather than using relatively lengthy and technical calculations, we heavily apply standard results from nonsmooth analysis and, in this way, obtain a relatively simple proof for local quadratic convergence.

Throughout this section, we assume implicitly that we have chosen an identification function such that the corresponding index set $A_{00}(x)$ has the exact identification property (3.27) under the SSOSC assumption. Suitable candidates for such an identification function were given in Section 3.3. We begin with the following result.

Theorem 3.23 *Let $x^* \in \Omega$ be a vector satisfying* (2.1) *and SSOSC. Then the mapping G is differentiable at x^* with $G'(x^*) = M(x^*)$ being nonsingular. Moreover, there is a neighbourhood of x^* and a constant $c > 0$ such that $M(x)$ is nonsingular with*

$$\|M(x)^{-1}\| \leq c$$

for all x in this neighbourhood.

Proof. Taking into account the definition of $M(x^*)$ in (3.34), (3.35), it follows after some elementary calculations that the ith column vector A_i of $A := M(x^*)^T$ is given by

$$A_i = \begin{cases} [\nabla f(x^*)]_i e_i, & \text{if } x_i^* = l_i \text{ and } [\nabla f(x^*)]_i > 0, \\ -[\nabla f(x^*)]_i e_i, & \text{if } x_i^* = u_i \text{ and } [\nabla f(x^*)]_i < 0, \\ [\nabla^2 f(x^*)]_i, & \text{if } i \in I_{00}(x^*), \\ d_i(x^*)[\nabla^2 f(x^*)]_i, & \text{if } i \notin I_0(x^*). \end{cases}$$

In particular, each column A_i is single-valued. Consequently, G is differentiable in x^* with $\partial_B G(x^*) = \{M(x^*)\}$, cf. Proposition 2.12.

Now consider the equation $A\alpha = 0$ for some $\alpha \in \mathbb{R}^n$. In view of the above representation of the columns of A, this may be rewritten as

$$\begin{aligned} 0 \;=\; & \sum_{i:x_i^*=l_i,[\nabla f(x^*)]_i>0} \alpha_i [\nabla f(x^*)]_i e_i - \sum_{i:x_i^*=u_i,[\nabla f(x^*)]_i<0} \alpha_i [\nabla f(x^*)]_i e_i + \ldots \\ & \ldots \sum_{i \in I_{00}(x^*)} \alpha_i [\nabla^2 f(x^*)]_i + \sum_{i \notin I_0(x^*)} \alpha_i d_i(x^*) [\nabla^2 f(x^*)]_i. \end{aligned}$$

Furthermore, using SSOSC and Lemma 3.17, we obtain

$$\begin{aligned} \alpha_i [\nabla f(x^*)]_i &= 0 \quad \forall i : x_i^* = l_i, [\nabla f(x^*)]_i > 0, \\ -\alpha_i [\nabla f(x^*)]_i &= 0 \quad \forall i : x_i^* = u_i, [\nabla f(x^*)]_i < 0, \\ \alpha_i &= 0 \quad \forall i \in I_{00}(x^*), \\ \alpha_i d_i(x^*) &= 0 \quad \forall i \notin I_0(x^*). \end{aligned}$$

Since $d_i(x^*) > 0$ for all $i \notin I_0(x^*)$, we get $\alpha = 0$. Consequently, the matrix A and, therefore, $M(x^*)$ itself is nonsingular.

Using this nonsingularity as well as $\partial_B G(x^*) = \{M(x^*)\}$ and Theorem 3.21, it follows from [57, Lemma 2.6] that there is a constant $c > 0$ and a neighbourhood of x^* such that $M(x)$ is nonsingular with $\|M(x)^{-1}\| \leq c$ for all x in this neighbourhood. □

We are now in the position to prove our main local convergence result.

Theorem 3.24 *Let $x^* \in \Omega$ be a vector satisfying* (2.1) *and SSOSC. Then there is a neighbourhood of x^* such that, for any starting point $x^0 \in int\Omega$ from this neighbourhood, Algorithm 3.22 is well-defined and generates a sequence $\{x^k\}$ which converges to x^* with a quadratic rate of convergence.*

Proof. In view of Theorem 3.23, there are constants $\varepsilon_1 > 0$ and $c > 0$ such that

$$\|M(x)^{-1}\| \leq c \quad \forall x \in B_{\varepsilon_1}(x^*). \tag{3.36}$$

Furthermore, Theorem 3.21 and standard properties of (strongly) semismooth functions (see Proposition 2.19) imply that there is a constant $\varepsilon_2 > 0$ such that

$$\|G(x) - G(x^*) - M(x)(x - x^*)\| \leq \frac{1}{4c}\|x - x^*\| \quad \forall x \in B_{\varepsilon_2}(x^*). \tag{3.37}$$

Using the definition of σ_k in (S.4) of Algorithm 3.22, we also see that there is an $\varepsilon_3 > 0$ such that

$$\sigma_k \geq \frac{3}{4} \quad \forall x^k \in B_{\varepsilon_3}(x^*) \tag{3.38}$$

(to this end, note that $\|P_\Omega(x^k + p^k) - x^k\| = \|P_\Omega(x^k + p^k) - P_\Omega(x^k)\| \leq \|p^k\|$ is very small in a neighbourhood of the solution x^* since then $G(x^k)$ is small and, therefore, the same holds for p^k in view of the nonsingularity of $M(x^k)$).

Now choose $x^0 \in \operatorname{int}\Omega \cap B_\varepsilon(x^*)$ with $\varepsilon := \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$. Then $M(x^0)$ is nonsingular, p^0 from (S.3) of Algorithm 3.22 exists, and we obtain

$$\begin{aligned} \|x^0 + p^0 - x^*\| &\leq \|x^0 - x^* - M(x^0)^{-1}G(x^0)\| \\ &\leq \|M(x^0)^{-1}\|\,\|G(x^0) - G(x^*) - M(x^0)(x^0 - x^*)\| \\ &\leq \frac{1}{4}\|x^0 - x^*\| \end{aligned}$$

from (3.36) and (3.37). The definition of x^1 in Algorithm 3.22 together with (3.38) and the nonexpansiveness of the projection operator then implies

$$\begin{aligned} \|x^1 - x^*\| &= \big\|x^0 + \sigma_0\big(P_\Omega(x^0 + p^0) - x^0\big) - x^*\big\| \\ &= \big\|\sigma_0\big(P_\Omega(x^0 + p^0) - x^*\big) + (1 - \sigma_0)(x^0 - x^*)\big\| \\ &\le \sigma_0\big\|P_\Omega(x^0 + p^0) - P_\Omega(x^*)\big\| + (1 - \sigma_0)\|x^0 - x^*\| \\ &\le \|x^0 + p^0 - x^*\| + \frac{1}{4}\|x^0 - x^*\| \\ &\le \frac{1}{2}\|x^0 - x^*\|. \end{aligned} \tag{3.39}$$

In particular, x^1 is also in the ball with radius ε around x^*. By induction, it follows that $\{x^k\} \subseteq \operatorname{int}\Omega$ is well-defined and satisfies

$$\|x^{k+1} - x^*\| \le \frac{1}{2}\|x^k - x^*\| \quad \forall k \in \mathbb{N}.$$

Hence the sequence $\{x^k\}$ converges (at least linearly) to x^*.

In order to verify the local quadratic rate of convergence, we recall that the strong semismoothness of G (see Theorem 3.21) implies that

$$\|G(x^k) - G(x^*) - M(x^k)(x^k - x^*)\| = O(\|x^k - x^*\|^2),$$

see Proposition 2.19. Using (3.36), we therefore get

$$\begin{aligned} \|x^k + p^k - x^*\| &= \|x^k - x^* - M(x^k)^{-1}G(x^k)\| \\ &\le \|M(x^k)^{-1}\|\,\|G(x^k) - G(x^*) - M(x^k)(x^k - x^*)\| \\ &= O(\|x^k - x^*\|^2). \end{aligned}$$

Following (3.39), this implies

$$\begin{aligned} \|x^{k+1} - x^*\| &\le \sigma_k\|P_\Omega(x^k + p^k) - P_\Omega(x^*)\| + (1 - \sigma_k)\|x^k - x^*\| \\ &\le \|x^k + p^k - x^*\| + (1 - \sigma_k)\|x^k - x^*\| \\ &= O(\|x^k - x^*\|^2) + (1 - \sigma_k)\|x^k - x^*\|. \end{aligned}$$

Exploiting once again the local Lipschitz continuity of the mapping G around x^* (see Theorem 3.21), we get from (S.3) of Algorithm 3.22 together with (3.36)

$$\begin{aligned} 1 - \sigma_k &= \|P_\Omega(x^k + p^k) - x^k\| \\ &\le \|p^k\| \\ &= O(\|G(x^k)\|) \\ &= O(\|G(x^k) - G(x^*)\|) \\ &= O(\|x^k - x^*\|). \end{aligned}$$

Altogether, we therefore have $\|x^{k+1} - x^*\| = O(\|x^k - x^*\|^2)$. □

3.6 Globalization

So far the method presented in Algorithm 3.22 possesses only local convergence properties and a globalization becomes necessary. Unfortunately the trust-region globalization techniques used, for example, in [12, 19, 67] use the Coleman-Li scaling and require at least that $d_i(x) \to 0$ if x converges to a point on the boundary of Ω. This essential property is not satisfied for our scaling (3.33) and the Heinkenschloss et al. scaling (3.17) if the scaling is switched off in a component. Therefore those globalizations cannot be used in the case of a degenerate solution. Moreover it is in the moment not possible for us to prove or disprove that the projected direction

$$p_{PN}^k := \sigma_k\big(P_\Omega(x^k + p^k) - x^k\big)$$

is locally a descent direction for f. Hence a line-search globalization for Algorithm 3.22 is difficult as well.

To overcome these problems we use an idea from methods for nonlinear systems of equations that has been applied in [43] to the semismooth bound constrained case and is described in Chapter 4. More precisely it can be shown that if a current iterate x^k is sufficiently close to a solution x^* satisfying SSOSC, the condition

$$\|G(x^k + p_{PN}^k)\| \le \eta\|G(x^k)\| \tag{3.40}$$

is satisfied for an $\eta \in (0,1)$ and stays satisfied if $x^{k+1} := x^k + p_{PN}^k$ is the next iterate. Thus our local method can be combined with any globally convergent interior-point method for bound constrained optimization problems by use of condition (3.40). For sake of simplicity we assume that the globally convergent method produces strictly feasible iterates of the form

$$x^{k+1} := x^k + p_{gc}^k \in \operatorname{int}\Omega$$

with a suitable direction $p_{gc}^k \in \mathbb{R}^n$ (e.g. a truncated projected steepest descent direction) that provides a global convergence result of the form

$$\liminf_{k\to\infty} \|D(x^k)\nabla f(x^k)\| = 0. \tag{3.41}$$

The globalized method then has the following form.

Algorithm 3.25 (Globalized Affine-Scaling Interior-Point Newton Method)

(S.0) Choose $x^0 \in \operatorname{int}(\Omega), \sigma, \eta \in (0,1)$, and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Compute $M(x^k) := D(x^k)\nabla^2 f(x^k) + S(x^k)$ with $D(x^k)$ and $S(x^k)$ being given by (3.33) and (3.35), respectively.

(S.3) Let $p_N^k \in \mathbb{R}^n$ be a solution of the linear system $M(x^k)p = -G(x^k)$.

(S.4) Compute $\sigma_k := \max\{\sigma, 1 - \|P_\Omega(x^k + p_N^k) - x^k\|\}$.

(S.5) Compute $p_{PN}^k := \sigma_k\big(P_\Omega(x^k + p_N^k) - x^k\big)$.

(S.6) If $\|G(x^k + p_{PN}^k)\| \leq \eta\|G(x^k)\|$ holds, set $x^{k+1} := x^k + p_{PN}^k$,
Otherwise compute p_{gc}^k and set $x^{k+1} := x^k + p_{gc}^k$.

(S.7) Set $k \leftarrow k + 1$, and go to (S.1).

This algorithm inherits the global convergence properties (3.41) of the method using p_{gc}^k.

Theorem 3.26 *Let the global convergence assumptions for the method using p_{gc}^k be satisfied and let Algorithm 3.25 generate an infinite sequence $\{x^k\}$. Then*

$$\liminf_{k\to\infty} \|G(x^k)\| = 0. \tag{3.42}$$

If p_{PN}^k is rejected in step (S.6) only for a finite number of iterations, then we obtain

$$\lim_{k\to\infty} \|G(x^k)\| = 0.$$

Proof. If $\|G(x^k + p_{PN}^k)\| > \eta\|G(x^k)\|$ holds only for a finite number of iterations, the method turns eventually into Algorithm (3.22) and (3.40) leads to $\|G(x^k)\| \to 0$. Otherwise the global convergence property (3.41) of the method using p_{gc}^k ensures the existence of a subsequence of $\{\|G(x^k)\|\}$ converging to zero. □

Depending on the globally convergent method used, (3.41) can use a different scaling matrix, e.g. $D(x) := D^{CL}(x)$ with the Coleman-Li scaling from (3.2). Then (3.42) has to be replaced by (3.41).

To establish fast local convergence we have to prove that Algorithm 3.25 turns into Algorithm 3.22 if the iterates are sufficiently close to a suitable solution of (P). This property can be transferred from the method for bound constrained semismooth systems of equations proposed in Chapter 4. Thus the following global convergence result holds.

Theorem 3.27 *Let $x^* \in \Omega$ be a vector satisfying* (2.1) *and SSOSC. Then there is a neighbourhood of x^*, such that for any starting point $x^0 \in int\Omega$ from this neighbourhood, Algorithm 3.25 is well-defined and generates a sequence $\{x^k\}$, which converges to x^* with a quadratic rate of convergence.*

Proof. Since $M(x^*)$ is nonsingular and G is strongly semismooth the proof Theorem 4.21 (to be presented later) can be easily transferred to our problem by considering the system $G(x) = 0$, $x \in \Omega$. Thus we only sketch the arguments and refer to Theorem 4.21 or [43, Theorem 4.1] for more details.

Just like in the proof of Theorem 3.24 there exist constants $\varepsilon_1, c > 0$ such that (3.36) holds. Since G is locally Lipschitz continuous, constants $L_1, \varepsilon_2 > 0$ exist with

$$\|G(x) - G(y)\| \leq L_1\|x - y\| \quad \forall x, y \in B_{\varepsilon_2}(x^*).$$

Due to the nonsingularity of $M(x^*)$ and [10, Theorem 7.1.1] the locally Lipschitz continuous inverse function G^{-1} exists in a neighbourhood of $G(x^*)$, and we obtain constants $\varepsilon_3 > 0$ and $L_2 > 0$ such that

$$\left\|G^{-1}(G(x)) - G^{-1}(G(y))\right\| \leq L_2\|G(x) - G(y)\| \quad \forall x, y \in B_{\varepsilon_3}(x^*).$$

Similar to (3.37) there exists an $\varepsilon_4 > 0$ with

$$\|G(x) - G(x^*) - M(x)(x - x^*)\| \leq \min\left\{\frac{\eta}{2cL_1L_2}, \frac{1}{4c}\right\}\|x - x^*\|.$$

Continuity of G leads to a constant $\varepsilon_5 > 0$ with

$$\|G(x)\| \leq \min\left\{\frac{\eta}{2cL_1L_2}, \frac{1-\sigma}{c}\right\} \quad \forall x \in B_{\varepsilon_5}(x^*).$$

With $\varepsilon_6 > 0$ defined by (3.38) we set $\varepsilon := \min\{\varepsilon_i : i = 1, \ldots, 6\}$ and assume $x^k \in \operatorname{int}\Omega \cap B_\varepsilon(x^*)$. On this basis one can prove that

$$\|x^k + p_N^k - x^*\| \leq \min\left\{\frac{\eta}{2L_1L_2}, \frac{1}{4}\right\}\|x^k - x^*\|$$

and consequently $x^k + p_{PN}^k \in B_\varepsilon(x^*)$ holds. Moreover $1 - \sigma_k \leq c\|F(x^k)\|$ can be shown. By use of these properties and the nonexpansiveness of the projection mapping we obtain

$$\begin{aligned}
\|G(x^k + p_{PN}^k)\| &= \|G(x^k + p_{PN}^k) - G(x^*)\| \\
&\leq L_1\|x^k + p_{PN}^k - x^*\| \\
&\leq L_1\sigma_k\|P_\Omega(x^k + p_N^k) - P_\Omega(x^*)\| + L_1(1 - \sigma_k)\|x^k - x^*\| \\
&\leq L_1\sigma_k\|x^k + p_N^k - x^*\| + L_1c\|G(x^k)\|\,\|x^k - x^*\| \\
&\leq \frac{\eta}{2L_2}\|x^k - x^*\| + \frac{\eta}{2L_2}\|x^k - x^*\| \\
&= \frac{\eta}{L_2}\left\|G^{-1}(G(x^k)) - G^{-1}(G(x^*))\right\| \\
&\leq \eta\|G(x^k) - G(x^*)\| = \eta\|G(x^k)\|.
\end{aligned}$$

Thus p_{PN}^k is accepted in step (S.6) and $x^{k+1} := x^k + p_{PN}^k$ follows. Like in the proof of Theorem 3.24 one can show

$$\|x^{k+1} - x^*\| \leq \frac{1}{2}\|x^k - x^*\|$$

and hence $x^{k+1} \in B_\varepsilon(x^*)$. By induction (3.40) holds for all sufficiently large k. Hence the transition from the global to the local method is established and the assertion of Theorem 3.24 completes the proof. □

The last two theorems show that transferring the ideas of Chapter 4 or [43] to the reformulated optimality conditions of (P) leads to a globalized method. However this approach is only directed on the reformulated KKT-conditions and not on a descent property for the objective function. We are aware of this severe disadvantage but use it nevertheless in absence of alternatives.

3.7 Numerical Examples

In this section, we want to illustrate the local behaviour of the different scaling strategies using two standard test problems. To this end, we implemented Algorithm 3.22 in MATLAB using $\sigma = 0.9995$ and the termination criterion $\|G(x)\| \leq 10^{-25}$. This is a relatively small tolerance, however, since we compare the pure local behaviour of some methods only, we prefer to have a small tolerance in order to see some interesting effects.

We then consider the following three methods which differ in the choice of the matrix $M(x) = D(x)\nabla^2 f(x) + S(x)$ from (3.34):

- Coleman-Li [11, 12]: Here $d_i(x) = d_i^{CL}(x)$ is defined by (3.16) and

  $$s_i(x) := s_i^{CL}(x) := |[\nabla f(x)]_i|$$

  for all $i = 1, \ldots, n$;

- Heinkenschloss et al. [37]: Here we take $d_i(x) = d_i^{HUU}(x)$ from (3.17) and

  $$s_i(x) := s_i^{HUU}(x) := \begin{cases} |[\nabla f(x)]_i| & \text{if } |[\nabla f(x)]_i| < \min\{x_i - l_i, u_i - x_i\}^q \\ & \text{or } \min\{x_i - l_i, u_i - x_i\} < |[\nabla f(x)]_i|^q \\ 0 & \text{else} \end{cases}$$

  with $q = 2$;

- new method: Here $d_i(x)$ and $s_i(x)$ are defined by (3.33) (using $\gamma = 10^{-3}$) and (3.35), respectively. In order to define the index set $A_{00}(x)$ for these choices of $d_i(x)$ and $s_i(x)$, we use the identification function ρ_2 from (3.31), (3.32) and take a suitable multiplier function $\lambda(x)$ from [21, Proposition 5] (using the parameters $\gamma_1 := \gamma_2 := 0.1$ in that reference).

Note that Heinkenschloss et al. [37] use a further scaling of the matrix $M(x)$ in order to verify theoretically the uniform nonsingularity of certain Jacobian-type matrices, but that this additional scaling is not necessary for the algorithm since it can be cancelled on both sides of the corresponding linear systems to be solved in their method, see Algorithm 3.10. Hence our implementation is equivalent to their method.

Our first test example is the famous Rosenbrock-function

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

This function has a unique global minimum at $x^* := (1, 1)^T$. Therefore we use $l := (0, 0)^T$ and $u := (1, 1)^T$ to obtain the degenerate set $I_{00}(x^*) = \{1, 2\}$. Since we are interested in the local convergence properties, we change the standard starting point to $x^0 := (0.999, 0.999)^T$. Table 3.1 contains the corresponding numerical results for the Coleman-Li scaling. For each iteration k, we report the size of the stopping criterion $\|G(x^k)\|$ as well as the distance of the current iterate (and its components) to the known solution.

k	$\|G(x^k)\|$	$\|x_1^k - x_1^*\|$	$\|x_2^k - x_2^*\|$	$\|x_k - x^*\|$
0	3.994004e-01	1.000000e-03	1.000000e-03	1.414214e-03
1	3.967247e-04	9.945479e-04	1.987114e-03	2.222104e-03
2	5.463162e-07	5.425928e-04	1.085095e-03	1.213194e-03
3	1.296976e-07	2.714348e-04	5.429716e-04	6.070379e-04
4	3.162359e-08	1.357598e-04	2.716074e-04	3.036467e-04
5	7.809617e-09	6.789056e-05	1.358342e-04	1.518554e-04
6	1.940620e-09	3.394796e-05	6.792478e-05	7.593576e-05
7	4.836979e-10	1.697465e-05	3.396431e-05	3.796989e-05
8	1.207434e-10	8.487491e-06	1.698263e-05	1.898545e-05
9	3.016327e-11	4.243788e-06	8.491436e-06	9.492851e-06
⋮	⋮	⋮	⋮	⋮
31	1.659987e-24	1.011746e-12	2.024603e-12	2.263326e-12
32	4.003273e-25	5.058176e-13	1.012301e-12	1.131638e-12
33	1.000818e-25	2.529088e-13	5.061507e-13	5.658192e-13
34	2.359936e-26	1.264544e-13	2.531308e-13	2.829593e-13

Table 3.1: Numerical results for the Rosenbrock function using the Coleman-Li scaling

The results in Table 3.1 indicate a relatively slow (linear) rate of convergence. The situation is significantly better for the Heinkenschloss et al. method, and the corresponding results are given in Table 3.2, where we include one further

column which gives the index set

$$\begin{aligned}\tilde{A}_{00}(x) :=& \big\{i : \ |[\nabla f(x)]_i| < \min\{x_i - l_i, u_i - x_i\}^p\big\} \\ & \cup \big\{i : \ \min\{x_i - l_i, u_i - x_i\} < |[\nabla f(x)]_i|^p\big\}\end{aligned}$$

which, in view of the definition of their scaling matrices in (3.17), may be viewed as the counterpart of our index set $A_{00}(x)$.

k	$\|G(x^k)\|$	$\|(x^k)_1 - x_1^*\|$	$\|(x^k)_2 - x_2^*\|$	$\|x^k - x^*\|$	$\tilde{A}_{00}(x^k)$
0	3.994004e-01	1.000000e-03	1.000000e-03	1.414214e-03	$\emptyset$
1	2.394061e-03	9.945479e-04	1.987114e-03	2.222104e-03	$\{1,2\}$
2	2.644000e-04	4.972739e-07	1.587753e-06	1.663803e-06	$\{1,2\}$
3	4.481673e-10	5.981726e-11	1.208688e-10	1.348605e-10	$\{1,2\}$
4	0	0	0	0	$\{1,2\}$

Table 3.2: Numerical results for the Rosenbrock function using the Heinkenschloss et al. scaling

Table 3.2 clearly shows the local quadratic convergence of the Heinkenschloss et al. method. The same is true for our scaling technique, and the corresponding numerical results are given in Table 3.3. In fact, we need one iteration less than the Heinkenschloss et al. method. According to our experience, this is mainly due to the fact that our identification technique for the degenerate indices $I_{00}(x^*)$ is more effective. In fact, comparing the results in Tables 3.2 and 3.3, we see that we are able to identify the correct set from the very beginning, whereas this is not true for the Heinkenschloss et al. scaling.

k	$\|G(x^k)\|$	$\|(x^k)_1 - x_1^*\|$	$\|(x^k)_2 - x_2^*\|$	$\|x^k - x^*\|$	$A_{00}(x^k)$
0	4.481984e-01	1.000000e-03	1.000000e-03	1.414214e-03	$\{1,2\}$
1	2.245015e-04	5.000000e-07	5.000000e-07	7.071068e-07	$\{1,2\}$
2	1.587703e-10	3.536060e-13	3.536060e-13	5.000744e-13	$\{1,2\}$
3	0	0	0	0	$\{1,2\}$

Table 3.3: Numerical results for the Rosenbrock function using the new scaling

To illustrate this point further, we take the Wood-function

$$\begin{aligned}f(x) :=& \ 100(x_2 - x_1^2) + (1 - x_1)^2 + 90(x_4 - x_3)^2 + (1 - x_3)^2 + 10(x_2 + x_4 - 2)^2 \\ & + 0.1(x_2 - x_4)^2\end{aligned}$$

as our second test problem. This function admits an unconstrained minimum in $x^* := (1, 1, 1, 1)^T$. We use the bounds $l := (1, 1, 1, 0.99)^T$ and $u := (3, 3, 3, 3)^T$ and obtain the degenerate set $I_{00}(x^*) = \{1, 2, 3\}$. Note, however, that also the

fourth component is almost degenerate. As a very good local starting point, we take $x^0 := 1.001 \cdot (1,1,1,1)^T$. The corresponding numerical results using the Coleman-Li scaling are given in Table 3.4.

k	$\|G(x^k)\|$	$\|x^k - x^*\|$
0	4.255314e-01	2.000000e-03
1	3.088548e-04	2.021504e-03
2	7.816318e-05	1.012668e-03
3	1.964852e-05	5.069441e-04
4	4.925803e-06	2.536260e-04
5	1.233169e-06	1.268516e-04
6	3.085074e-07	6.343543e-05
7	7.715390e-08	3.172013e-05
8	1.929178e-08	1.586067e-05
9	4.823369e-09	7.930485e-06
⋮	⋮	⋮
34	5.684342e-14	2.366777e-13
35	5.684342e-14	1.182266e-13
36	5.684342e-14	5.933163e-14
37	2.327831e-26	2.994722e-14

Table 3.4: Numerical results for the Wood function using the Coleman-Li scaling

Again, we see that the convergence of the Coleman-Li method is rather slow. On the other hand, we get much faster convergence for the Heinkenschloss et al. scaling, as documented in Table 3.5. However, we also see that the estimation $\tilde{A}_{00}(x^k)$ of the degenerate index set $I_{00}(x^*)$ is sometimes incorrect even very close to the solution.

k	$\|G(x^k)\|$	$\|x^k - x^*\|$	$\tilde{A}_{00}(x^k)$
0	4.255314e-01	2.000000e-03	$\{\}$
1	2.428585e-04	2.021504e-03	$\{2\}$
2	6.143252e-05	1.012668e-03	$\{2\}$
3	1.543420e-05	5.069441e-04	$\{2\}$
4	3.868204e-06	2.536260e-04	$\{1,2,3\}$
5	8.814318e-03	1.268516e-04	$\{1,2,3\}$
6	2.824142e-06	2.268045e-08	$\{1,2,3\}$
7	3.067384e-13	5.874748e-16	$\{1,2,3\}$
8	5.684342e-14	1.110223e-16	$\{1,2\}$
9	0	0	$\{1,2,3\}$

Table 3.5: Numerical results for the Wood function using the Heinkenschloss et al. scaling

The situation is significantly better when using our new scaling technique. Table 3.6 gives the results obtained with our method. Although the fourth index is initially viewed as being degenerate (as is to be expected), our technique eventually finds the correct index set and converges much faster than the Heinkenschloss et al. method.

k	$\|G(x^k)\|$	$\|x^k - x^*\|$	$A_{00}(x^k)$
0	5.823476e-01	2.000000e-03	$\{1,2,3,4\}$
1	1.391926e-03	9.067386e-06	$\{1,2,3,4\}$
2	2.574718e-08	1.609195e-10	$\{1,2,3\}$
3	0	0	$\{1,2,3\}$

Table 3.6: Numerical results for the Wood function using the new scaling

The preceeding numerical examples confirm the observation of Heinkenschloss et al. [37] that a violation of the strict complementarity assumption slows down the convergence rate of the affine-scaling Newton method using the Coleman-Li scaling. This problem can be overcome by the Heinkenschloss et al. scaling or the new scaling introduced in (3.33) and [43]. The method using the new scaling seems to perform slightly better due to the exact estimation technique for the degenerate indices. With this conclusion we abandon considering the optimization problem (P) and proceed with the second problem class of nonlinear systems subject to bound constraints.

Chapter 4

Nonlinear Systems of Equations

This chapter deals with affine-scaling methods for the bound constrained nonlinear system of equations

$$F(x) = 0 \quad \text{subject to} \quad x \in \Omega, \tag{NE}$$

where the feasible set Ω is given by

$$\Omega := \{x \in \mathbb{R}^n : \; l_i \leq x_i \leq u_i \quad \forall i = 1, \ldots, n\}$$

and where $F : \mathcal{O} \longrightarrow \mathbb{R}^n$ is at least semismooth on an open set $\mathcal{O} \subseteq \mathbb{R}^n$ containing the box Ω. Moreover we assume that the lower and upper bounds satisfy $-\infty \leq l_i < u_i \leq +\infty$ for all $i = 1, \ldots, n$.

Similar to the last chapter we consider some recently proposed numerical methods for (NE) before we describe our affine-scaling method and its theoretical and numerical properties.

4.1 Numerical Methods

In subsequent subsections several numerical methods for bound constrained nonlinear systems are shortly described. Our main interest lies again on affine-scaling methods, which are motivated by corresponding methods for bound constrained optimization problems. The notation of the described papers is partially adapted in order to obtain a consistent representation. This mainly concerns the used scaling matrices. If possible we call the methods with the names and abbreviations introduced by their authors.

4.1.1 The STRN Method

Bellavia, Macconi and Morini propose in [2] a scaled trust-region interior point Newton method (STRN for short) for bound constrained nonlinear systems with

continuously differentiable objective functions. Origin of their method is the unconstrained local Newton method that converges locally quadratically under suitable assumptions. The used direction $p_N^k \in \mathbb{R}^n$ is the solution of the Newton equation

$$F'(x^k)p = -F(x^k). \tag{4.1}$$

Because of the bound constraints this direction has to be truncated to obtain an interior point method. Let $x^k \in \operatorname{int}\Omega$ be given an let $\lambda(p^k)$ denote the steplength from x^k to the nearest boundary in the direction $p^k \in \mathbb{R}^n$, then the stepsize is given by

$$t(p^k) := \begin{cases} 1, & \text{if} \quad x^k + p^k \in \operatorname{int}\Omega, \\ \max\{\theta, 1 - \|p^k\|\}\lambda(p^k), & \text{otherwise}, \end{cases} \tag{4.2}$$

with a constant $\theta \in (0,1)$. If the next iterate is defined by use of this stepsize and the Newton direction from (4.1) with

$$x^{k+1} := x^k + t(p_N^k)p_N^k,$$

one obtains a strictly feasible truncated Newton method that possesses fast local convergence properties if additional assumptions on p_N^k and the solution of (NE) are satisfied.

In order to obtain a globally convergent method the merit function

$$f(x) := \frac{1}{2}\|F(x)\|^2$$

is considered and an affine-scaling trust-region method similar to the Coleman-Li trust-region methods in [12] and Section 3.1.2 is applied to the problem

$$\text{minimize } f(x) \quad \text{subject to} \quad x \in \Omega.$$

With a given iterate $x^k \in \operatorname{int}\Omega$ the model function

$$m_k(p) := \frac{1}{2}\|F(x^k) + F'(x^k)p\|^2 \tag{4.3}$$

arises from a linearization of F (not f as in the case of optimization problems). This function is trusted and minimized on an ellipsoidal trust-region and a search direction $p^k \in \mathbb{R}^n$ is obtained by solving

$$\text{minimize} \quad m_k(p) \quad \text{subject to} \quad \|D(x^k)^{-1/2}p\| \le \Delta_k \tag{4.4}$$

with the trust-region radius $\Delta_k > 0$. The used scaling matrix in [2] is (in adapted notation) given by

$$D(x) = \operatorname{diag}\big(d_1(x), \ldots, d_n(x)\big),$$

where

$$d_i(x) := \begin{cases} x_i - l_i, & \text{if } [\nabla f(x)]_i \geq 0 \text{ and } l_i > -\infty, \\ u_i - x_i, & \text{if } [\nabla f(x)]_i < 0 \text{ and } u_i < \infty, \\ 1, & \text{otherwise.} \end{cases} \tag{4.5}$$

This scaling matrix is again the Coleman-Li scaling from [11, 12], similar to (3.2) and (3.16), but the case of unbounded components is included this time. As we have seen in the trust-region methods by Coleman-Li and Dennis-Vicente in Sections 3.1.2 and 3.1.3 the Cauchy point plays an important role for the global convergence properties of trust-region methods. Therefore Bellavia et al. define the Cauchy point by

$$p_{CP}^k := -\tau_{CP} D(x^k) \nabla f(x^k) \tag{4.6}$$

where the scalar τ_{CP} is the solution of

$$\begin{aligned} \text{minimize} \quad & m_k(p(\tau)) \quad \text{subject to} \quad && p(\tau) = -\tau D(x^k)\nabla f(x^k),\ \tau > 0, \\ & && \|D(x^k)^{-1/2} p(\tau)\| \leq \Delta_k. \end{aligned} \tag{4.7}$$

Different from, for example, the Dennis-Vicente method this Cauchy point and the subproblem (4.4) does not take into account the box constraints, because the stepsize strategy from (4.2) is applied later in the method to maintain strict feasibility. For global convergence proofs it is important again that the search direction p^k satisfies a fraction of Cauchy decrease condition

$$r_k^c := \frac{m_k(0) - m_k(t(p^k)p^k)}{m_k(0) - m_k(t(p_{CP}^k)p_{CP}^k)} \geq \beta_1 \tag{4.8}$$

with a scalar $\beta_1 \in (0, 1]$. If the current candidate for a search direction does not satisfy (4.8) the Cauchy point itself can be accepted. Moreover the acceptance of the search direction and the size of the trust-region is controlled by

$$r_k^f := \frac{f(x^k) - f(x^k + t(p^k)p^k)}{m_k(0) - m_k(t(p^k)p^k)} \geq \beta_2 \tag{4.9}$$

with $\beta_2 \in (0, 1)$ and can be seen as a measure for the agreement of m_k and f. The STRN algorithm can now be described in the following form:

Algorithm 4.1 (Scaled Trust-Region Newton (STRN) method)

(S.0) Choose $x^0 \in \operatorname{int} \Omega$, $\Delta_0 > 0$, constants $\beta_1 \in (0, 1]$, $0 < \beta_2 < \beta_3 < 1$, $0 < \omega_1 < 1 < \omega_2$, $\theta \in (0, 1)$ and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Repeat:

(R.1) Compute a solution p^k of (4.4).
(R.2) Compute the Cauchy point p^k_{CP} from (4.6).
(R.3) Compute $t(p^k)$ and $t(p^k_{CP})$ by using (4.2) and r^c_k from (4.8).
(R.4) If $r^c_k < \beta_1$, then set $p^k := p^k_{CP}$.
(R.5) Set $\Delta^*_k := \Delta_k$.
(R.6) Set $\Delta_k := \omega_1 \Delta_k$.
(R.7) Compute r^f_k given by (4.9).

Until $r^f_k \geq \beta_2$.

(S.3) Set $x^{k+1} := x^k + t(p^k)p^k$ and $\Delta_k = \Delta^*_k$.

(S.4) If $r^f_k \geq \beta_3$, set $\Delta_{k+1} := \omega_2 \Delta_k$.
Otherwise set $\Delta_{k+1} := \Delta_k$.

(S.5) Set $k \leftarrow k+1$, and go to (S.1).

Global and local convergence Theorems for this method are carried out in [2]. The following general assumptions are used:

(A1) F' is Lipschitz continuous on $\mathcal{L} := \bigcup_{k=0}^{\infty} \{x \in \Omega : \ \|x - x^k\| \leq r\}$ with a constant $r > 0$.

(A2) $\|F'(x)\|$ is bounded from above on $\mathcal{L}$.

At first it is important to ensure that the repeat loop in step (S.2) of the STRN-Algorithm 4.1 terminates in each iteration to obtain a well defined method. This is the assertion of the following Lemma, see also [2, Lemma 3.4].

Lemma 4.2 *Let (A1) be satisfied and $F'(x^k)$ be nonsingular and $F(x^k) \neq 0$ for a $k \in \mathbb{N}_0$. Then the repeat loop in step (S.2) of the kth iteration of the STRN-Algorithm 4.1 terminates after a finite number of inner iterations.*

Hence the method is well defined if all Jacobians $F'(x^k)$ are nonsingular and under this additional assumption global convergence [2, Theorems 3.1, 3.2] can be shown.

Theorem 4.3 *Let (A1) and (A2) be satisfied and $\{x^k\}$ be generated by the STRN-Algorithm 4.1 and bounded. Then*

$$\lim_{k \to \infty} \|D(x^k)^{1/2} \nabla f(x^k)\| = 0.$$

If x^* is an isolated limit point of $\{x^k\}$ such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular, then $\{x^k\}$ converges to x^*.

The convergence rate of the STRN method is described in the next Theorem.

Theorem 4.4 *Let (A1) and (A2) be satisfied and* $\{x^k\}$ *be generated by the STRN-Algorithm 4.1. Let* x^* *be a solution of* (NE) *with nonsingular* $F'(x^*)$*. If* $x^k \to x^*$*, then* $\{\Delta_k\}$ *is bounded away from zero. Moreover:*

(a) If $\|D(x^k)p_N^k\| \to 0$ *and* $t(p_N^k) \geq 1 - \sqrt{1-\beta_1}$ *for all sufficiently large* k*, then* $\|F(x^k)\| \to 0$ *linearly.*

(b) If $\|D(x^k)p_N^k\| \to 0$ *and* $t(p_N^k) \to 1$*, then* $x^k \to x^*$ *superlinearly.*

(c) If $\|D(x^k)p_N^k\| \to 0$ *and* $t(p_N^k) = 1$ *for all sufficiently large* k*, then* $x^k \to x^*$ *quadratically.*

As a further consequence of this Theorem, see also [2, Theorem 3.2, Corollary 3.2], $\{x^k\}$ converges quadratically to strictly feasible solutions of (NE) satisfying the assumptions of Theorem 4.4.

The practical properties of the STRN method are examined by Bellavia et al. in [2] very detailed. Their considerations include 32 different test examples and a comparison to the ASTN method from [41] and the IGTN method from [46]. Conclusively the STRN method is robust and less costly then the other methods taking into account the number of needed function evaluations. On basis of this method the STRSCNE code was developed and described in [3]. This code inherits the robustness and fast convergence properties of the STRN method and is therefore taken as a reference code for the numerical properties of the later presented method.

Recapitulating the main advantages of the STRN method are very fast theoretical and numerical convergence properties. But the assumptions for the theoretical properties are strict, i.e. $F'(x^k)$ has to be nonsingular in each iteration and an interior solution is needed for a quadratic rate of convergence theorem. This is a disadvantage, which is overcome by the later proposed IATR method in [5].

4.1.2 The IATR Method

In [5] Bellavia and Morini develop an interior point affine-scaling trust-region (IATR) method for (NE) that can be seen as a successor of the STRN method from the last section or [2]. This method shares many properties of the STRN method, but is locally quadratically convergent even to solutions on the boundary of the feasible set. To achieve this aim two main modifications of the STRN method are incorporated. Both are concerning the choice of the search direction. At first the stepsize strategy (4.2) to maintain strict feasibility is (in large parts) replaced by a modified projection strategy. At second the direction candidates satisfying a fraction of Cauchy decrease condition are computed in a different

way.

Basis of the IATR algorithm is again an affine-scaling trust-region optimization approach applied to the problem

$$\text{minimize } f(x) := \frac{1}{2}\|F(x)\|^2 \quad \text{subject to} \quad x \in \Omega$$

with the merit function f. To obtain a search direction the ellipsoidal trust-region subproblem

$$\text{minimize} \quad m_k(p) \quad \text{subject to} \quad \|D(x^k)^{-1/2}p\| \leq \Delta_k \tag{4.10}$$

with the model function m_k defined in (4.3) is considered. The scaling matrix D is a slightly modified Coleman-Li scaling with diagonal elements

$$d_i(x) := \begin{cases} x_i - l_i, & \text{if } [\nabla f(x)]_i > 0 \text{ and } l_i > -\infty, \\ u_i - x_i, & \text{if } [\nabla f(x)]_i < 0 \text{ and } u_i < \infty, \\ \min\{x_i - l_i, u_i - x_i\}, & \text{if } [\nabla f(x)]_i = 0 \text{ and } (l_i > -\infty \text{ or } u_i < \infty), \\ 1, & \text{otherwise,} \end{cases} \tag{4.11}$$

that differs from (4.5) only in the third case. To compute the search direction two vectors are involved: The exact solution of (4.10) denoted by p_{ex}^k and the Cauchy point p_{CP}^k defined in (4.6). Since the subproblem (4.10) does not take into account the box constraints both directions can lead to infeasible iterates. Therefore the projection-type direction $\bar{p}_{ex}^k$ with

$$[\bar{p}_{ex}^k]_i := \begin{cases} \min\big\{(1-\alpha)(l_i - x_i^k), 2(l_i - x_i^k) - [p_{ex}^k]_i\big\}, & \text{if } x_i^k + [p_{ex}^k]_i \leq l_i, \\ [p_{ex}^k]_i, & \text{if } x_i^k + [p_{ex}^k]_i \in (l_i, u_i), \\ \max\big\{(1-\alpha)(u_i - x_i^k), 2(u_i - x_i^k) - [p_{ex}^k]_i\big\}, & \text{if } x_i^k + [p_{ex}^k]_i \geq u_i, \end{cases} \tag{4.12}$$

for all $i = 1, \ldots, n$ with a constant $\alpha \in (0, 1)$ is computed. Then $x^k + \bar{p}_{ex}^k$ is strictly feasible and can for small α be seen as the result of a projection on Ω followed by a small step into the interior of Ω. Moreover $\bar{p}_{ex}^k$ shares some important properties of projected directions, that will also be used later in this chapter. In case of the Cauchy point p_{CP}^k the stepsize rule from (4.2) is still used to obtain interior iterates when the direction

$$\bar{p}_{CP}^k := t(p_{CP}^k)p_{CP}^k \tag{4.13}$$

is used. The search direction p^k is then located on the connecting line between $\bar{p}_{ex}^k$ and $\bar{p}_{CP}^k$

$$p^k := \mu\bar{p}_{CP}^k + (1-\mu)\bar{p}_{ex}^k, \tag{4.14}$$

where $\mu \in [0, 1)$ is computed such that the fraction of Cauchy decrease condition

$$r_k^c := \frac{m_k(0) - m_k(p^k)}{m_k(0) - m_k(\bar{p}_{CP}^k)} \geq \beta_1 \tag{4.15}$$

with a constant $\beta_1 \in (0,1)$ and

$$m_k(p^k) \leq m_k(\bar{p}_{ex}^k) \tag{4.16}$$

are satisfied. Existence and computation of such a μ are shown in [5]. The remaining algorithm is very similar to the STRN method but the update of the trust-region size is different.

Algorithm 4.5 (Interior point Affine-scaling Trust-Region (IATR) method)

(S.0) Choose $x^0 \in \operatorname{int}\Omega$, $\Delta_0, \Delta_{\min} > 0$, constants $\alpha, \beta_1, \beta_2, \omega_1, \theta \in (0,1)$ and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Set $\bar{\Delta}_k := \max\{\Delta_{\min}, \Delta_k\}$, set $\Delta_k^* := \bar{\Delta}_k$.

(S.3) Repeat:

- (R.1) Set $\Delta_k^* := \Delta_k$.
- (R.2) Compute a solution p_{ex}^k of (4.10).
- (R.3) Compute the Cauchy point p_{CP}^k from (4.6).
- (R.4) Compute $\bar{p}_{ex}^k$ and $\bar{p}_{CP}^k$ by using (4.12) and (4.13).
- (R.5) Compute $\mu \in [0,1)$ such that p^k satisfies (4.15) and (4.16).
- (R.6) Set $\Delta_k^* := \omega_1 \Delta_k$.
- (R.7) Compute r_k^f given by (4.9).

Until $r_k^f \geq \beta_2$.

(S.4) Set $x^{k+1} := x^k + p^k$.

(S.5) Choose Δ_{k+1} (e.g. like in Algotrihm 4.1).

(S.6) Set $k \leftarrow k+1$, and go to (S.1).

By use of the assumptions

(A1) F' is Lipschitz continuous on $\mathcal{L} := \bigcup_{k=0}^{\infty} \{x \in \Omega : \ \|x - x^k\| \leq r\}$ with a constant $r > 0$,

(A2) $\|F'(x)\|$ is bounded from above on $\mathcal{L}$,

global and local convergence results for the IATR method from Algorithm 4.5 can be established. By assuming that all Jacobians $F'(x^k)$ are nonsingular the algorithm is well defined, see also the next Lemma or [5, Lemma 3.2].

Lemma 4.6 *Let (A1) be satisfied and $F'(x^k)$ be nonsingular and $F(x^k) \neq 0$ for a $k \in \mathbb{N}_0$. Then the repeat loop in step (S.3) of the kth iteration of the IATR-Algorithm 4.5 terminates after a finite number of inner iterations.*

On this basis Bellavia and Morini prove a global convergence Theorem [5, Theorem 3.1] similar to Theorem 4.3.

Theorem 4.7 *Let (A1) and (A2) be satisfied and $\{x^k\}$ be generated by the IATR-Algorithm 4.5 and bounded. Then*

$$\lim_{k\to\infty} \|D(x^k)^{1/2}\nabla f(x^k)\| = 0.$$

If $x^* \in int\Omega$ is a limit point of $\{x^k\}$ such that $F'(x^*)$ is nonsingular, then $\{F(x^k)\}$ converges to 0 and all accumulation points of $\{x^k\}$ solve (NE).

The global convergence properties of the STRN and the IATR method do not differ much because both methods use a fraction of Cauchy decrease condition to establish global convergence for the trust-region approach to the bound constrained optimization problem with the merit function. But with the modifications of the search direction it is possible to prove a stronger local convergence Theorem [5, Theorem 3.2], which leads to quadratic convergence for solutions on the boundary as well.

Theorem 4.8 *Let (A1) and (A2) be satisfied and $\{x^k\}$ be generated by the IATR-Algorithm 4.5. Let x^* be a limit point of $\{x^k\}$ with $F(x^*) = 0$ and nonsingular $F'(x^*)$. Then $\{x^k\}$ converges to x^* and the rate of convergence is quadratic.*

In [5] also numerical experiments with a huge amount of test problems are carried out. The IATR method is able to solve nearly all of the given problems and seems to be very efficient since in most iterations only one function evaluation is necessary.

Conclusively it can be said, that the IATR method inherits the advantages of the STRN method and possesses a stronger local convergence theory, which eliminates a disadvantage of the STRN method. The smaller disadvantage that all Jacobians $F'(x^k)$ have to be nonsingular to get a well defined method still exists.

4.1.3 The SIATR Method

The IATR method from [5] and the last section possesses good theoretical and numerical properties. In [6] Bellavia and Morini extend this method to a subspace interior point affine-scaling trust-region method (SIATR) for large scale

nonlinear equations with bound constraints. Most of the numerical effort of the IATR method has to be done solving the trust-region subproblem (4.10). Hence the restriction of this subproblem to suitable subspaces can be an advantage in the large scale case.

The general framework of the SIATR method is equal to the IATR method. Basis of the subspace strategy is the trust-region subproblem

$$\text{minimize} \quad m_k(p) \quad \text{subject to} \quad \|S(x^k)p\| \leq \Delta_k \tag{4.17}$$

with the trust-region radius $\Delta_k > 0$ and the model function m_k from (4.3). Similar to Chapter 3.1.3 the matrix $S(x^k)$ is defined by

$$S(x^k) := I \quad \text{or} \quad S(x^k) := D(x^k)^{-1/2}$$

where $D(x^k)$ is the Coleman-Li scaling matrix from (4.11). Since the computation of an exact solution of (4.17) can be be very expensive in the large scale case, the subproblem is replaced by the *subspace trust-region subproblem*

$$\text{minimize} \quad m_k(p) \quad \text{subject to} \quad p \in \mathcal{S}_k,\ \|S(x^k)p\| \leq \Delta_k \tag{4.18}$$

with a subset $\mathcal{S}_k$ of $\mathbb{R}^n$. This subset has to satisfy the following condition:

$$m_k(p_k^*) \leq \eta_k^2 m_k(0), \tag{4.19}$$

where p_k^* denotes a solution of

$$\text{minimize} \quad m_k(p) \quad \text{subject to} \quad p \in \mathcal{S}_k \tag{4.20}$$

and $\eta_k \in [0, 1)$. Two different subspace approaches and ways to compute an exact solution p_{ex}^k of (4.18) are carried out in [6], but for sake of brevity we pass on the details. If a solution p_{ex}^k is computed the search direction is again computed on the connecting line between the truncated Cauchy point $\bar{p}_{CP}^k$ from (4.13) and the projection-type direction $\bar{p}_{ex}^k$ defined in (4.12) with the solution p_{ex}^k of (4.18)

$$p^k := \mu \bar{p}_{CP}^k + (1 - \mu)\bar{p}_{ex}^k \tag{4.21}$$

with a $\mu \in [0, 1)$ such that the fraction of Cauchy decrease condition (4.15) is satisfied. Then the SIATR method can be described as follows.

Algorithm 4.9 (Subspace Interior point Affine-scaling Trust-Region (SIATR) method)

(S.0) Choose $x^0 \in \operatorname{int} \Omega$, $\Delta_0, \Delta_{\min} > 0$, constants $\alpha, \beta_1, \beta_2, \omega_1, \theta \in (0, 1)$ and set $k := 0$.

(S.1) If x^k satisfies a suitable termination criterion: STOP.

(S.2) Set $\bar{\Delta}_k := \max\{\Delta_{\min}, \Delta_k\}$, set $\Delta_k := \bar{\Delta}_k/\omega_1$ and choose $\eta_k \in [0, 1)$

(S.3) Repeat:

(R.1) Set $\Delta_k := \omega_1 \Delta_k$.
(R.2) Find $\mathcal{S}_k \subseteq \mathbb{R}^n$ such that (4.19) holds.
(R.3) Compute a solution p_{ex}^k of (4.18).
(R.4) Compute the Cauchy point p_{CP}^k from (4.6).
(R.5) Compute $\bar{p}_{ex}^k$ and $\bar{p}_{CP}^k$ by using (4.12) and (4.13).
(R.6) Compute $\mu \in [0, 1)$ such that p^k satisfies (4.15).
(R.7) Compute r_k^f given by (4.9).

Until $r_k^f \geq \beta_2$.

(S.4) Set $x^{k+1} := x^k + p^k$.

(S.5) Choose Δ_{k+1} (e.g. like in Algotrihm 4.1).

(S.6) Set $k \leftarrow k + 1$, and go to (S.1).

Under the assumption (A1) from the last two sections the assertion of Lemma 4.6 holds for Algorithm 4.9 as well. Therefore it is well defined and global convergence can be established, see [6, Theorems 4.1, 4.2].

Theorem 4.10 *Let (A1) and (A2) be satisfied and $\{x^k\}$ be generated by the SIATR-Algorithm 4.9 and bounded. Then*

$$\lim_{k\to\infty} \|D(x^k)^{1/2} \nabla f(x^k)\| = 0.$$

If $x^* \in int\Omega$ is a limit point of $\{x^k\}$ such that $F'(x^*)$ is nonsingular, then $\{F(x^k)\}$ converges to 0 and all accumulation points of $\{x^k\}$ solve (NE). *If x^* is an isolated limit point of $\{x^k\}$ such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular, then $\{x^k\}$ converges to x^*.*

The global convergence properties are very similar to those of the IATR and the STRN method. The local properties differ stronger due to the choice of the matrix S_k and the reduction of the subproblem to subspaces.

Theorem 4.11 *Let (A1) and (A2) be satisfied and $\{x^k\}$ be generated by the SIATR-Algorithm 4.9. Let $\{x^k\}$ converge to x^* with $F(x^*) = 0$ and nonsingular $F'(x^*)$ and assume that either $S_k := I$ for all k or $S_k := D(x^k)^{-1/2}$ for all k with $\|D(x^k)^{-1/2} p^k\| \to 0$. Then p^k satisfies* (4.15) *with $\bar{\Delta}_k$ for all sufficiently large k. If $\eta_k \to 0$, then $\{x^k\}$ converges to x^* superlinearly. If $\eta_k = O(\|F(x^k)\|)$, then $\{x^k\}$ converges to x^* quadratically.*

Proof of the last theorem can be found in [6, Theorem 4.3]. The convergence rate depends on η_k since this term is closely related to the error term of the inexactly solved Newton equation, see [6, 17] for more details.

As a special choice for the subspace $\mathcal{S}_k$ in the SIATR method Bellavia, Macconi and Morini propose in [4] a two-dimensional subspace trust-region method for bound constrained nonlinear systems. The subspace is given by

$$\mathcal{S}_k := \operatorname{span}\{p_I^k, \nabla f(x^k)\}$$

with an inexact solution p_I^k of the Newton equation. The subspace subproblem (4.18) with $S_k := I$ for all k is approximately solved by a dogleg strategy and the assertions of Theorems 4.10 and 4.11 hold.

Advantages and disadvantages of the SIATR method are mostly inherited from the IATR method. The SIATR method possesses strong global and local convergence results. The local results can be weaker depending on choice of the matrix S_k and the subspace $\mathcal{S}_k$. Numerical experiments are not described in [6], but especially for large scale problems the SIATR method could perform better than other methods. This is also indicated by the successful high dimensional tests for the two-dimensional subspace method in [4].

4.1.4 The Ulbrich method

The non-monotone trust-region method for semismooth equations with bound constraints proposed by Ulbrich in [65] is unlike to the other considered methods neither an interior point nor an affine-scaling method. Nevertheless it shares some important properties with those methods and the method we will propose later in this chapter. The function F is assumed to be at least semismooth on an open set containing Ω. Basis of the algorithm is the Newton-type direction p_N^k from the linear system

$$M_k p = -F(x^k) \tag{4.22}$$

where $x^k \in \Omega$ is the current iterate and $M_k \in \mathbb{R}^{n\times n}$ is a nonsingular approximation to an $H \in \partial F(x^k)$. Due to the bound constraints this direction is projected back on the feasible set and one obtains the projected direction

$$p_{PN}^k := P_\Omega(x^k + p_N^k) - x^k.$$

Since the aim is not an interior point approach, no further truncation of this direction is necessary. For a local method using the direction p_{PN}^k local quadratic convergence can be shown under assumptions including BD-regularity of the solution x^*, strong semismoothness of F and the following restriction on M_k

$$\mu_k := \min_{H\in\partial_B F(x^k)} \|(M_k - H)p_N^k\| \le \delta\|p_N^k\| \tag{4.23}$$

for all k with a sufficiently small $\delta > 0$. In order to embed this local direction into a globally convergent method the merit function

$$f(x) := \frac{1}{2}\|F(x)\|^2$$

is considered again. Under the assumption that f is continuously differentiable on an open set containing Ω a trust-region algorithm is applied to the problem

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \Omega. \tag{4.24}$$

With the model function

$$m_k(p) := \nabla f(x^k)^T p + \frac{1}{2}\|M_k p\|^2$$

the trust-region subproblem

$$\text{minimize} \quad m_k(p) \quad \text{subject to} \quad x^k + p \in \Omega,\ \|p\|_\infty \leq \Delta_k \tag{4.25}$$

with the trust-region radius $\Delta_k > 0$ is considered. Since the second restriction is chosen in terms of the maximum norm, subproblem (4.25) is a box constrained convex quadratic program with the feasible set

$$X_k := [l - x^k, u - x^k] \cap [-\Delta_k, \Delta_k]^n.$$

A feasible projected Newton direction is then given by

$$p_P^k := P_{X_k}(p_N^k). \tag{4.26}$$

The used search direction p^k has to satisfy the *feasibility condition*

$$x^k + p^k \in \Omega \quad \text{and} \quad \|p^k\|_\infty \leq \beta_1 \Delta_k \tag{4.27}$$

with a constant $\beta_1 \geq 1$ and the *reduction condition*

$$\text{pred}_k(p^k) := -m_k(p^k) \geq \beta_2 \chi(x^k) \min\{1, \Delta_k, \chi(x^k)\} \tag{4.28}$$

with $\beta_2 > 0$ and a *criticality measure* χ. A criticality measure is a continuous mapping $\chi : \Omega \longrightarrow \mathbb{R}_+$ with $\chi(x^*) = 0$ iff x^* satisfies the first order necessary optimality condition for (4.24). For example

$$\chi(x) := \|x - P_\Omega(x - \nabla f(x))\|$$

can be used. Moreover the acceptance of a search direction is controlled by an unconventional reduction ratio r_k, because non monotonicity of $\{f(x^k)\}$ should be allowed, see [65] for more details. In adapted notation the method has the following form.

Algorithm 4.12 (Non-Monotone Trust-Region Method)

(S.0) Choose $x^0 \in \Omega$, $\Delta_0 > \Delta_{\min} \geq 0$, $\eta, \sigma, \theta \in (0,1)$, $0 < \rho_1 < \rho_2 < 1$, and $0 \leq \omega_1, \omega_2 < 1 < \omega_3$, set $k := 0$.

(S.1) If $\chi(x^k) = 0$: STOP.

(S.2) Compute the Newton-direction p_N^k from (4.22).

(S.3) Compute p_P^k given by (4.26).

(S.4) If p_P^k satisfies (4.28), then set $p^k := p_P^k$. Otherwise compute p^k satisfying (4.27) and (4.28).

(S.5) Compute r_k.

(S.6) Update the trust-region radius according to the following rules:
If $r_k \geq \rho_1$, choose $\Delta_{k+1} \in (\omega_1 \Delta_k, \omega_2 \Delta_k]$.
If $r_k \in (\rho_1, \rho_2)$, choose $\Delta_{k+1} \in [\omega_2 \Delta_k, \max\{\Delta_{\min}, \Delta_k\}] \cap [\Delta_{\min}, \infty)$.
If $r_k \geq \rho_2$, choose $\Delta_{k+1} \in (\Delta_k, \max\{\Delta_{\min}, \omega_3 \Delta_k\}] \cap [\Delta_{\min}, \infty)$.

(S.7) If $r_k > \rho_1$, we call iteration k "successful" and set $x^{k+1} := x^k + p^k$.
Otherwise set $x^{k+1} := x^k$.

(S.8) Set $k \leftarrow k + 1$, and go to (S.1).

Global and local convergence results for this algorithm are carried out in [65]. As usual we restate the global ones [65, Theorems 4.8, 4.9] first.

Theorem 4.13 *Let Algorithm 4.12 generate an infinite sequence* $\{x^k\}$ *and let* $\omega_1 > 0$ *or* $\Delta_{\min} = 0$. *Then*

$$\liminf_{k \to \infty} \chi(x^k) = 0.$$

If there exists a set containing all x^k *on which* χ *is uniformly continuous and bounded, then*

$$\lim_{k \to \infty} \chi(x^k) = 0.$$

To establish fast local convergence it is important to ensure that the projected Newton direction p_{PN}^k is accepted close to a BD-regular solution x^*. Therefore the following implication is required

$$\|x^k - x^*\| < \beta_3 \text{ and } \text{pred}_k(p_{PN}^k) \geq \beta_4 f(x^k) \Longrightarrow p_{PN}^k \text{ satisfies (4.28)} \qquad (4.29)$$

with constants $\beta_3 > 0$, $\beta_4 \in (0,1)$. A concrete implementation of the rather abstract conditions (4.29) and (4.28) by means of a fraction of Cauchy decrease condition and including affine-scaling matrices is described in [65]. The local convergence properties of Algorithm 4.12 can now be described as follows.

Theorem 4.14 *Let Algorithm 4.12 with $\Delta_{\min} > 0$ generate an infinite sequence $\{x^k\}$ and let x^* be a BD-regular solution of* (NE)*. Then there exist constants $\delta, \varepsilon > 0$ such that the following holds: If the index k' satisfies $\|x^{k'} - x^*\| \leq \varepsilon$, $k' - 1$ was successful and if* (4.23) *holds for all $k \geq k'$, then:*

(a) The direction p_P^k is accepted, $p^k = p_P^k = p_{PN}^k$ holds and the kth iteration is successful for all $k \geq k'$.

(b) The sequence $\{x^k\}$ converges to x^*.

(c) If $\mu_k / \|p_N^k\| \to 0$ holds, the convergence rate is superlinear.

(d) If F is strongly semismooth in x^* and

$$\limsup_{k \to \infty} \frac{\mu_k}{\|p_N^k\|^2}$$

holds, the rate of convergence is quadratic.

In fact Ulbrich's convergence result considers also convergence rates between 1 and 2 for so called p-order semismooth functions, but we restrict this to the strongest possible rate.

In the numerical section of [65] the presented method is applied to a semismooth reformulation of Mixed Complementarity Problems (MCPs) from the MCPLIB [23]. The results are very impressive. A reason for this could be that the subproblems (4.25) are solved with a QP solver if the used Cauchy decrease condition is not satisfied instead of taking the Cauchy point itself as it is done in other methods. This seems to be very effective, but is also costly.

Conclusive it can be said that the non-monotone trust-region method by Ulbrich possesses strong global and local convergence results and is able to handle semismooth functions. The numerical properties are very good, but an efficient QP solver is needed, which is a disadvantage of this method. Depending on the given problem it can be a disadvantage that this method is not an interior point method. This can be the case if F is not defined on the boundary of Ω.

Beside the methods already described in this section there are only few methods for nonlinear systems taking explicitly into account box constraints. These other methods have strong global and local convergence properties as well, but since our main interest lies on affine-scaling methods, we describe them only shortly.

- In [41] Kanzow proposes an active-set type Newton method for (NE). The objective function is assumed to be smooth and the generated iterates are feasible. Global and local convergence results are carried out. The rate of convergence is quadratic under standard assumptions. A numerical comparison of this method with the STRN method is presented in [2].

- Kozakevich, Martínez and Santos describe in [46] a feasible inexact Newton-type approach for bound constrained smooth equations that has to solve a convex QP with bound constraints in each iteration. Global and local superlinear convergence theorems and numerical tests are established.
- A method for nonsmooth nonlinear systems with bound constraints is developed by Qi, Tong and Li in [59]. This active set projected trust-region algorithm is shown to be globally and locally quadratically convergent and numerical experiments are described.

After this summary over existing methods for bound constrained nonlinear systems we describe our affine-scaling approach to semismooth systems of equations with bound constraints.

4.2 Description of the Method

This section gives a detailed description of our trust-region-type method for the solution of problem (NE). To this end, we first recall that (NE) is closely related to the box constrained optimization problem

$$\text{minimize } f(x) := \frac{1}{2}\|F(x)\|^2 \quad \text{subject to} \quad x \in \Omega. \tag{4.30}$$

In fact, every solution x^* of (NE) is a global minimum of (4.30). Conversely, if x^* is a minimum of (4.30) such that $f(x^*) = 0$, then x^* is also a solution of (NE).

Regarding the mapping f defined in (4.30), we make the following assumption, which we assume to hold throughout the remaining part of this chapter.

(A) The mapping f from (4.30) is continuously differentiable.

Assumption (A) obviously holds if F itself is continuously differentiable. However, there are also some interesting situations where f is continuously differentiable although F is not differentiable (but semismooth), see e.g. [30, 16, 9] for some examples in the context of complementarity problems.

We exploit the relation between the two problems (NE) and (4.30) and apply again the observation of Coleman and Li [11, 12] that the first order optimality conditions of (4.30) are equivalent to the nonlinear system of equations

$$D(x)\nabla f(x) = 0, \quad x \in \Omega, \tag{4.31}$$

with a suitable scaling matrix

$$D(x) = \operatorname{diag}\big(d_1(x), \ldots, d_n(x)\big).$$

Originally, Coleman and Li [11, 12] consider only one particular choice of the scaling matrix $D(x)$. However Heinkenschloss et al. [37] extended this equivalence to a rather general class of scaling matrices satisfying the conditions (2.3), i.e.

$$d_i(x) \begin{cases} = 0, & \text{if } x_i = l_i \text{ and } [\nabla f(x)]_i > 0, \\ = 0, & \text{if } x_i = u_i \text{ and } [\nabla f(x)]_i < 0, \\ \geq 0, & \text{if } x_i \in \{l_i, u_i\} \text{ and } [\nabla f(x)]_i = 0, \\ > 0, & \text{else} \end{cases}$$

for all $i = 1, \ldots, n$ and all $x \in \Omega$, see also Lemma 2.2. In fact, the reader may find some other scaling matrices (satisfying these conditions and sometimes having better convergence properties than the original Coleman-Li-scaling) in [37, 42, 67].

In this work, we allow the scaling matrix satisfying (2.3) to be from a rather general class, see Assumption (C) below. Several existing scaling matrices from the literature satisfy our assumptions, for example, we may take the Coleman-Li-scaling [11, 12], defined by

$$d_i^{CL}(x) := \begin{cases} x_i - l_i, & \text{if } [\nabla f(x)]_i > 0 \text{ and } l_i > -\infty, \\ u_i - x_i, & \text{if } [\nabla f(x)]_i < 0 \text{ and } u_i < \infty, \\ \min\{x_i - l_i, u_i - x_i\}, & \text{if } [\nabla f(x)]_i = 0 \text{ and } (l_i > -\infty \text{ or } u_i < \infty), \\ 1, & \text{else,} \end{cases} \tag{4.32}$$

for $x \in \Omega$ (more precisely, this is the modified Coleman-Li-scaling suggested by Heinkenschloss et al. [37] or (3.16) with respect to infinite bounds, see also [5, 6] or (4.11)), or the *minimum-scaling*

$$d_i^{MIN}(x) := \begin{cases} 1, & \text{if } l_i = -\infty \text{ and } u_i = +\infty, \\ d_i(x), & \text{otherwise,} \end{cases} \tag{4.33}$$

with

$$d_i(x) := \min\left\{x_i - l_i + \gamma \max\{0, -[\nabla f(x)]_i\}, u_i - x_i + \gamma \max\{0, [\nabla f(x)]_i\}\right\},$$

where $\gamma > 0$ is a given constant, cf. (4.33) or [42]. Both scaling matrices may be used in order to prove suitable global and local convergence results. Nevertheless, we stress that the minimum-scaling has some additional properties (see Assumption (D) below) that allows us to prove stronger global convergence results than for the Coleman-Li-scaling (see Theorem 4.20 below).

In order to construct a suitable method for the solution of problem (NE), we follow an interior-point trust-region approach for (4.30) similar to those in [12, 19, 67] for box constrained optimization, although it should be pointed out that our quadratic model is different due to the fact that we deal with nonlinear

systems of equations. Given an iterate $x^k \in \operatorname{int}\Omega$, we consider the quadratic model

$$m_k(p) := \frac{1}{2}\|F(x^k) + H_k p\|^2 \approx \frac{1}{2}\|F(x^k + p)\|^2$$

on the scaled trust-region

$$\big\{p \in \mathbb{R}^n : \|D(x^k)^{-1/2} p\| \le \Delta_k\big\},$$

where $\Delta_k > 0$ denotes the trust-region radius, and where $H_k \in \partial F(x^k)$ is an element of the generalized Jacobian of F at x^k. In order to get the next iterate $x^{k+1} \in \operatorname{int}\Omega$, we first compute an approximate solution $p^k \in \mathbb{R}^n$ of the subproblem

$$\text{minimize} \quad m_k(p) \quad \text{s.t.} \quad x^k + p \in \operatorname{int}\Omega,\ \|D(x^k)^{-1/2} p\| \le \Delta_k. \tag{4.34}$$

Following the standard trust-region philosophy, we then define the predicted and actual reductions by

$$\begin{aligned} \operatorname{ared}_k(p^k) &:= f(x^k) - f(x^k + p^k) \qquad \text{and} \\ \operatorname{pred}_k(p^k) &:= m_k(0) - m_k(p^k) = f(x^k) - m_k(p^k), \end{aligned}$$

respectively. If the quotient

$$r_k := \frac{\operatorname{ared}_k(p^k)}{\operatorname{pred}_k(p^k)} \tag{4.35}$$

is sufficiently large, we accept the quadratic model, compute $x^{k+1} := x^k + p^k$, and possibly increase the trust-region radius Δ_k. Otherwise we reject the step, set $x^{k+1} := x^k$ again and decrease the radius Δ_k.

Hence it remains to specify the computation of our approximate solution p^k. To this end, we first define the modified Cauchy-step

$$p^k_{CP} := -\tau_{CP} D_k \nabla f(x^k),$$

where $D_k := D(x^k)$ and $\tau_{CP} = \tau^k_{CP} \in \mathbb{R}$ is a solution of the one-dimensional problem

$$\begin{aligned} \text{minimize} \quad m_k(p(\tau)) \quad \text{s.t.} \qquad & p(\tau) = -\tau D_k \nabla f(x^k),\ \|D_k^{-1/2} p(\tau)\| \le \Delta_k \\ & \theta(l - x^k) \le p(\tau) \le \theta(u - x^k), \tau \ge 0, \end{aligned} \tag{4.36}$$

where $\theta \in (0,1)$ is a given constant which guarantees that $x^k + p^k_{CP} \in \operatorname{int}\Omega$, see also Dennis and Vicente [19] or Section 3.1.3. We then compute a vector $p^k \in \mathbb{R}^n$ satisfying the fraction of Cauchy-decrease condition

$$m_k(p^k) \le m_k(p^k_{CP}), \quad x^k + p^k \in \operatorname{int}\Omega, \quad \|D_k^{-1/2} p^k\| \le \Delta_k; \tag{4.37}$$

in particular, we may take $p^k = p^k_{CP}$ (although, in practice, we try other steps first, see the corresponding comments in Section 4.5). With such a choice of p^k, it is possible to prove a global convergence result. In order to get fast local convergence, we also use a projected interior-point Newton-type step. Since the (generalized) Newton direction

$$p^k_N := -H_k^{-1}F(x^k) \qquad (\text{where } H_k \in \partial F(x^k)) \tag{4.38}$$

for the unconstrained problem $F(x) = 0$ does, in general, not satisfy the condition $x^k + p^k_N \in \operatorname{int}\Omega$, we follow [37, 65] and use the projected and truncated Newton direction

$$p^k_{PN} := \sigma_k\big(P_\Omega(x^k + p^k_N) - x^k\big), \tag{4.39}$$

with

$$\sigma_k := \max\big\{\sigma, 1 - \|P_\Omega(x^k + p^k_N) - x^k\|\big\} \tag{4.40}$$

for some constant $\sigma \in (0,1)$. We then have $x^{k+1} := x^k + p^k_{PN} \in \operatorname{int}\Omega$, and we will see in our convergence analysis that this choice guarantees local fast convergence under suitable conditions.

In order to get a simple transition from the global method with a direction p^k satisfying (4.37) to the local method with the direction p^k_{PN} from (4.39), we also incorporate the test

$$\|F(x^k + p^k_{PN})\| \le \eta\|F(x^k)\| \tag{4.41}$$

in our method, where $\eta \in (0,1)$ denotes another constant. We will see that (4.41) holds automatically in a neighbourhood of a solution of (NE) under suitable assumptions. We are now in a position to give a precise statement of the overall method.

Algorithm 4.15 (Interior-Point Trust-Region Method)

(S.0) Choose $x^0 \in \operatorname{int}\Omega$, $\Delta_0 > 0$, $\varepsilon > 0$, $\eta, \sigma, \theta \in (0,1)$, $0 < \rho_1 < \rho_2 < 1$, and $0 < \omega_1 < 1 < \omega_2$, set $k := 0$.

(S.1) If $\|D_k^{1/2}\nabla f(x^k)\| \le \varepsilon$: STOP.

(S.2) Choose a matrix $H_k \in \partial F(x^k)$ and compute (if possible) p^k_{PN} using (4.39). If (4.41) holds, set $x^{k+1} := x^k + p^k_{PN}, \Delta_{k+1} := \omega_2\Delta_k$, and go to (S.5); otherwise go to (S.3).

(S.3) Compute $p^k \in \mathbb{R}^n$ satisfying (4.37), and define r_k by (4.35).
If $r_k \ge \rho_1$, we call iteration k "successful" and set $x^{k+1} := x^k + p^k$;
otherwise we set $x^{k+1} := x^k$.

(S.4) Update the trust-region radius according to the following rules:
If $r_k < \rho_1$, set $\Delta_{k+1} := \omega_1 \Delta_k$.
If $r_k \in [\rho_1, \rho_2)$, set $\Delta_{k+1} := \Delta_k$.
If $r_k \geq \rho_2$, set $\Delta_{k+1} := \omega_2 \Delta_k$.

(S.5) Set $k \leftarrow k + 1$, and go to (S.1).

We give a number of comments with some simple properties of Algorithm 4.15.

Remark 4.16 (a) The termination criterion in step (S.1) checks whether the current iterate x^k is an approximate stationary point of the box constrained optimization problem (4.30).

(b) All iterates x^k belong to the interior of the box Ω. Hence the inverse diagonal matrices $D_k^{-1/2}$ occurring, e.g., in (4.36), (4.37), always exist since the elements $d_i(x^k)$ are positive according to (2.3).

(c) The computation of p_{PN}^k in step (S.2) requires the (generalized) Jacobian H_k from (4.38) to be nonsingular. If this matrix turns out to be singular, we immediately switch to step (S.3).

(d) Taking into account the previous comments, Algorithm 4.15 is well-defined in the sense that all steps can actually be carried out without any additional assumptions on problem (NE).

(e) The entire sequence $\{f(x^k)\}$ is monotonically decreasing. Equivalently, this means that we have $\|F(x^{k+1})\| \leq \|F(x^k)\|$ for all $k \in \mathbb{N}$. In fact, this is obvious if the test (4.41) gets accepted in step (S.2) of Algorithm 4.15. Otherwise, we compute p^k satisfying (4.37) in step (S.3). If the iteration is not successful, we have $\|F(x^{k+1})\| = \|F(x^k)\|$, otherwise we have

$$r_k \geq \rho_1 \Longleftrightarrow f(x^k) - f(x^k + p^k) \geq \rho_1 \big(f(x^k) - m_k(p^k)\big).$$

Here, the expression on the right-hand side is nonnegative because we have

$$m_k(p^k) \leq m_k(p_{CP}^k) \leq m_k(0) = f(x^k)$$

in view of (4.37) and the definition of the Cauchy step p_{CP}^k. Hence, the inequality $f(x^k) \geq f(x^k + p^k) = f(x^{k+1})$ also holds for all successful iterations in step (S.3).

In our subsequent convergence analysis, we assume throughout that $\varepsilon = 0$ and that Algorithm 4.15 generates an infinite sequence, i.e., it does not terminate after a finite number of iterations satisfying the first order optimality conditions of problem (4.30).

4.3 Global Convergence

The aim of this section is to prove a global convergence result for Algorithm 4.15. To this end, we need two more assumptions. The first is a boundedness assumption, which is rather standard in trust-region methods (see e.g. [15]), and the second one is a condition regarding the choice of the diagonal scaling matrix $D(x)$.

(B) The sequence $\{H_k\}$ generated by Algorithm 4.15 is bounded.

(C) The scaling matrix $D(x)$ satisfies (2.3) and is bounded on Ω. Furthermore, there exists a constant $\alpha > 0$ such that

$$\alpha d_i(x) \leq \begin{cases} x_i - l_i, & \text{if } [\nabla f(x)]_i > 0 \text{ and } l_i > -\infty, \\ u_i - x_i, & \text{if } [\nabla f(x)]_i < 0 \text{ and } u_i < +\infty \end{cases}$$

for all $i = 1, \ldots, n$ and all $x \in \operatorname{int} \Omega$.

Note that the last part of (C) is satisfied both by the modified Coleman-Li-scaling (4.32) and the minimum-scaling (4.33) with $\alpha = 1$. Furthermore, all remaining conditions hold automatically if Ω itself is bounded, i.e., if all lower and upper bounds l_i and u_i are finite (this follows, e.g., from the upper semicontinuity of the generalized Jacobian, see [10]). This assumption is quite realistic in many cases since otherwise one may replace infinite bounds by sufficiently large bounds.

There is a simple consequence of Assumption (B) that will play a crucial role in our subsequent analysis and that is therefore stated explicitly in the following remark.

Remark 4.17 Suppose that Assumptions (A) and (B) hold. Then the sequence $\{\nabla f(x^k)\}$ is bounded. To see this, note that Assumption (A) and [10, Proposition 2.2.4 and Theorem 2.6.6] together imply that we can write the gradient as $\nabla f(x^k) = H_k^T F(x^k)$ with H_k being the matrix from step (S.2) of Algorithm 4.15. Now $\{H_k\}$ is bounded in view of Assumption (B). Moreover, $\{\|F(x^k)\|\}$ is also bounded because of Remark 4.16 (e), so that $\{\nabla f(x^k)\}$ must indeed be bounded.

We now state a technical lemma that leads to a lower bound for the predicted reduction. Results of this kind are standard for trust-region methods, see, in particular, [67, Lemma 6.1] and [19, Lemma 4.1].

Lemma 4.18 *Suppose that Assumptions (A) and (C) hold, and let $p^k \in \mathbb{R}^n$ satisfy the fraction of Cauchy-decrease condition* (4.37)*. Then*

$$\mathit{pred}_k(p^k) \geq \frac{1}{2}\|D_k^{1/2} g^k\| \min\Big\{\Delta_k, \frac{\|D_k^{1/2} g^k\|}{\|D_k^{1/2} H_k^T H_k D_k^{1/2}\|}, \theta\alpha \frac{\|D_k^{1/2} g^k\|}{\|g^k\|_\infty}\Big\} \tag{4.42}$$

where $g^k := \nabla f(x^k)$ denotes the gradient of f at x^k. If, in addition, Assumption (B) holds, then there exists a constant $C > 0$ such that

$$pred_k(p^k) \geq C\|D_k^{1/2} g^k\|^2 \min\{\Delta_k, 1\}. \tag{4.43}$$

Proof. The proof is essentially the same as in [67, 19], except that we have a different quadratic model since we deal with box constrained nonlinear equations instead of bound constrained optimization problems. For the sake of completeness, however, we include the full proof here.

Consider a fixed iterate $x^k \in \operatorname{int}\Omega$, and recall that the stepsize $\tau \geq 0$ in (4.36) has to satisfy the two feasibility requirements

$$\|D_k^{-1/2} p(\tau)\| \leq \Delta_k \tag{4.44}$$

and

$$\theta(l - x^k) \leq p(\tau) \leq \theta(u - x^k). \tag{4.45}$$

Let τ_Δ and τ_Ω denote the two maximum stepsizes such that (4.44) and (4.45) hold, respectively. Since $p(\tau) = -\tau D_k g^k$, an elementary calculation shows that

$$\begin{aligned} \tau_\Delta &= \frac{\Delta_k}{\|D_k^{1/2} g^k\|} \qquad \text{and} \\ \tau_\Omega &= \theta \min\Big\{ \min_{i:[D_k g^k]_i < 0} \Big\{ \frac{u_i - x_i^k}{-[D_k g^k]_i} \Big\}, \min_{i:[D_k g^k]_i > 0} \Big\{ \frac{l_i - x_i^k}{-[D_k g^k]_i} \Big\} \Big\}. \end{aligned}$$

Hence the solution $\tau_{CP} = \tau_{CP}^k$ of the one-dimensional problem (4.36) has to belong to the interval $[0, \tau_+]$, where

$$\tau_+ := \min\{\tau_\Delta, \tau_\Omega\}.$$

(Note that τ_Δ, τ_Ω and, therefore, τ_+ are well-defined because $D_k g^k \neq 0$ since otherwise we would have stopped in step (S.1) of Algorithm 4.15.) From the Cauchy-decrease condition (4.37), we therefore obtain

$$m_k(p^k) \leq m_k(p_{CP}^k) = \min_{\tau \in [0, \tau_+]} \phi(\tau)$$

with

$$\phi(\tau) := m_k(-\tau D_k g^k).$$

Let $\tau^* (= \tau_{CP})$ be a solution of

$$\text{minimize } \phi(\tau) \quad \text{s.t.} \quad \tau \in [0, \tau_+].$$

Using the notation

$$F_k := F(x^k), \quad \hat{g}^k := D_k^{1/2} g^k, \quad \text{and} \quad \hat{M}_k := D_k^{1/2} H_k^T H_k D_k^{1/2},$$

and recalling that $g^k = H_k^T F_k$ (see Remark 4.17), the function ϕ may be rewritten as

$$\begin{aligned}\phi(\tau) &= \frac{1}{2}\|F_k\|^2 - \tau F_k^T H_k D_k g^k + \frac{1}{2}\tau^2 (g^k)^T D_k H_k^T H_k D_k g^k \\ &= \frac{1}{2}\|F_k\|^2 - \tau \|\hat{g}^k\|^2 + \frac{1}{2}\tau^2 (\hat{g}^k)^T \hat{M}_k \hat{g}^k.\end{aligned}$$

Note that $\hat{M}_k$ is always positive semidefinite. We now distinguish four cases.

Case 1: Suppose that $\tau^* \in (0, \tau_+)$. Then τ^* is an unconstrained minimum of ϕ, and we therefore get

$$0 = \phi'(\tau^*) = -\|\hat{g}^k\|^2 + \tau^* (\hat{g}^k)^T \hat{M}_k \hat{g}^k. \tag{4.46}$$

This gives the explicit formula

$$\tau^* = \frac{\|\hat{g}^k\|^2}{(\hat{g}^k)^T \hat{M}_k \hat{g}^k}.$$

(Note that the denominator is nonzero, because otherwise (4.46) would lead to $D_k^{1/2} g^k = 0$ and we would have stopped in step (S.1) of Algorithm 4.15.) This implies

$$\phi(\tau^*) = \frac{1}{2}\|F_k\|^2 - \frac{1}{2}\frac{\|\hat{g}^k\|^4}{(\hat{g}^k)^T \hat{M}_k \hat{g}^k} \leq \frac{1}{2}\|F_k\|^2 - \frac{1}{2}\frac{\|\hat{g}^k\|^2}{\|\hat{M}_k\|}.$$

Case 2: Assume that $\tau^* = \tau_+$ and $\tau_+ = \tau_\Delta$. If, in addition, we have $(\hat{g}^k)^T \hat{M}_k \hat{g}^k > 0$, the necessary optimality condition $\phi'(\tau^*) \leq 0$ implies

$$\tau^* \leq \frac{\|\hat{g}^k\|^2}{(\hat{g}^k)^T \hat{M}_k \hat{g}^k}. \tag{4.47}$$

We therefore get

$$\begin{aligned}\phi(\tau^*) &\leq \frac{1}{2}\|F_k\|^2 - \tau_\Delta \|\hat{g}^k\|^2 + \frac{1}{2}\tau_\Delta \left(\frac{\|\hat{g}^k\|^2}{(\hat{g}^k)^T \hat{M}_k \hat{g}^k}\right) (\hat{g}^k)^T \hat{M}_k \hat{g}^k \\ &= \frac{1}{2}\|F_k\|^2 - \frac{1}{2}\tau_\Delta \|\hat{g}^k\|^2 \\ &= \frac{1}{2}\|F_k\|^2 - \frac{1}{2}\Delta_k \|\hat{g}^k\|\end{aligned}$$

from the definition of τ_Δ. On the other hand, if we have $(\hat{g}^k)^T \hat{M}_k \hat{g}^k = 0$, we also obtain

$$\phi(\tau^*) = \frac{1}{2}\|F_k\|^2 - \tau_\Delta \|\hat{g}^k\|^2 = \frac{1}{2}\|F_k\|^2 - \frac{1}{2}\Delta_k \|\hat{g}^k\|.$$

Case 3: Suppose that $\tau^* = \tau_+$ and $\tau_+ = \tau_\Omega$. Here we first take a closer look at τ_Ω. Using Assumption (C), we get the following lower bound for the maximum stepsize τ_Ω:

$$\begin{aligned}
\tau_\Omega &= \theta \min\Big\{ \min_{i:[D_k g^k]_i<0} \Big\{ \frac{u_i - x_i^k}{-[D_k g^k]_i} \Big\}, \min_{i:[D_k g^k]_i>0} \Big\{ \frac{l_i - x_i^k}{-[D_k g^k]_i} \Big\} \Big\} \\
&= \theta \min\Big\{ \min_{i:[D_k g^k]_i<0} \Big\{ \frac{u_i - x_i^k}{d_i(x^k)|g_i^k|} \Big\}, \min_{i:[D_k g^k]_i>0} \Big\{ \frac{x_i^k - l_i}{d_i(x^k)|g_i^k|} \Big\} \Big\} \\
&\geq \theta \min\Big\{ \min_{i:[D_k g^k]_i<0} \Big\{ \frac{u_i - x_i^k}{d_i(x^k)\|g^k\|_\infty} \Big\}, \min_{i:[D_k g^k]_i>0} \Big\{ \frac{x_i^k - l_i}{d_i(x^k)\|g^k\|_\infty} \Big\} \Big\} \\
&\geq \frac{\theta\alpha}{\|g^k\|_\infty}.
\end{aligned}$$

Therefore, if we have $(\hat{g}^k)^T \hat{M}_k \hat{g}^k > 0$, then $\phi'(\tau^*) \leq 0$, hence (4.47) holds, and we get

$$\phi(\tau^*) \leq \frac{1}{2}\|F_k\|^2 - \frac{1}{2}\tau_\Omega \|\hat{g}^k\|^2 \leq \frac{1}{2}\|F_k\|^2 - \frac{\theta\alpha}{2}\frac{\|\hat{g}^k\|^2}{\|g^k\|_\infty}.$$

On the other hand, if $(\hat{g}^k)^T \hat{M}_k \hat{g}^k = 0$, we also obtain

$$\phi(\tau^*) = \frac{1}{2}\|F_k\|^2 - \tau^* \|\hat{g}^k\|^2 = \frac{1}{2}\|F_k\|^2 - \tau_\Omega \|\hat{g}^k\|^2 \leq \frac{1}{2}\|F_k\|^2 - \frac{\theta\alpha}{2}\frac{\|\hat{g}^k\|^2}{\|g^k\|_\infty}.$$

Case 4: Suppose that $\tau^* = 0$. Then the necessary optimality condition $-\|\hat{g}^k\|^2 = \phi'(\tau^*) \geq 0$ implies $\hat{g}^k = 0$, a contradiction to $D_k^{1/2} g^k \neq 0$. Hence this case does not occur.

Taking all cases together, we get

$$m_k(p^k) \leq \phi(\tau^*) \leq \frac{1}{2}\|F_k\|^2 - \frac{1}{2}\min\Big\{ \Delta_k \|\hat{g}^k\|, \frac{\|\hat{g}^k\|^2}{\|\hat{M}_k\|}, \theta\alpha \frac{\|\hat{g}^k\|^2}{\|g^k\|_\infty} \Big\}.$$

Consequently, we obtain the lower bound

$$\mathrm{pred}_k(p^k) = m_k(0) - m_k(p^k) \geq \frac{1}{2}\|\hat{g}^k\| \min\Big\{ \Delta_k, \frac{\|\hat{g}^k\|}{\|\hat{M}_k\|}, \theta\alpha \frac{\|\hat{g}^k\|}{\|g^k\|_\infty} \Big\} \tag{4.48}$$

for the predicted reduction, which is precisely the statement from (4.42).

Now suppose that Assumptions (A)–(C) hold. Then the sequences $\{\|H_k\|\}$, $\{\|g^k\|_\infty\}$ and $\{\|D_k^{1/2}\|\}$ are bounded, cf. Remark 4.17. Hence $\{\|\hat{M}_k\|\}$ and $\{\|\hat{g}^k\|\}$ are bounded as well. Therefore, (4.48) yields the existence of a constant $C > 0$

such that

$$\begin{aligned}\mathrm{pred}_k\left(p^k\right) &\geq \frac{1}{2}\|\hat{g}^k\| \min\left\{\frac{\Delta_k\|\hat{g}^k\|}{\|\hat{g}^k\|}, \frac{\|\hat{g}^k\|}{\|\hat{M}_k\|}, \theta\alpha\frac{\|\hat{g}^k\|}{\|g^k\|_\infty}\right\} \\ &= \frac{1}{2}\|\hat{g}^k\|^2 \min\left\{\frac{\Delta_k}{\|\hat{g}^k\|}, \frac{1}{\|\hat{M}_k\|}, \frac{\theta\alpha}{\|g^k\|_\infty}\right\} \\ &\geq C\|\hat{g}^k\|^2 \min\{\Delta_k, 1\}\end{aligned}$$

for all $k \in \mathbb{N}$, and this proves the second statement. □

Note that the proof of Lemma 4.18 shows, in particular, how the Cauchy step p_{CP}^k can be computed in practice.

We are now in the position to state the first main global convergence result for Algorithm 4.15. To this end, note that we are dealing with two different search directions p_{PN}^k (the projected Newton-like step) and p^k (the Cauchy-like step). While the former will play a central role for the local rate of convergence, the Cauchy-like step is the main tool for showing global convergence. This is similar to some existing results stated in [12, 19, 67], for example. Note that the local direction p_{PN}^k does not destroy the global convergence of the overall method.

Theorem 4.19 *Suppose that Assumptions (A)–(C) hold. Then*

$$\liminf_{k\to\infty} \|D_k^{1/2}\nabla f(x^k)\| = 0. \tag{4.49}$$

Moreover, if the direction p_{PN}^k is accepted an infinite number of times in step (S.2) of Algorithm 4.15, we have

$$\lim_{k\to\infty} \|F(x^k)\| = 0. \tag{4.50}$$

Proof. First recall from Remark 4.16 (e) that the entire sequence $\{\|F(x^k)\|\}$ is monotonically decreasing. Hence, if the test (4.41) in step (S.2) of Algorithm 4.15 is satisfied an infinite number of times, we immediately obtain (4.50). In particular, this implies $\|\nabla f(x^k)\| \to 0$ and therefore (4.49) since the sequence $\{D_k\}$ stays bounded in view of Assumption (C).

It remains to consider the case where the direction p_{PN}^k is accepted only a finite number of times. Without loss of generality, we may assume that this never happens, so we always compute the direction p^k. Suppose that (4.49) does not hold. Then there is a constant $\delta > 0$ such that

$$\|\hat{g}^k\| \geq \delta \quad \forall k \in \mathbb{N}, \tag{4.51}$$

where, again, we write $\hat{g}^k := D_k^{1/2}\nabla f(x^k)$.

In the first part of this proof, we show that this implies

$$\sum_{k=0}^{\infty} \Delta_k < \infty. \tag{4.52}$$

In fact, if there is only a finite number of successful iterations, we have $\Delta_{k+1} = \omega_1 \Delta_k$ for all $k \in \mathbb{N}$ sufficiently large, and (4.52) follows from $\omega_1 \in (0,1)$ and the convergence of the geometric series. Otherwise, there is an infinite number of successful iterations. Let k_i denote the indices of the successful iterations. Since $\{f(x^k)\}$ is monotonically decreasing and bounded from below, the entire sequence $\{f(x^k)\}$ converges. In particular, we have

$$\sum_{k=0}^{\infty} \left(f(x^k) - f(x^{k+1}) \right) < \infty. \tag{4.53}$$

From (4.43) and (4.51), we obtain

$$f(x^{k_i}) - f(x^{k_i+1}) = \text{ared}_{k_i}(p^{k_i}) \geq \rho_1 \text{pred}_{k_i}(p^{k_i}) \geq \rho_1 C \delta^2 \min\{\Delta_{k_i}, 1\} > 0 \tag{4.54}$$

for all successful iterations. Since the expression on the left-hand side of (4.54) converges to zero, it follows that $\min\{\Delta_{k_i}, 1\} = \Delta_{k_i}$ for all sufficiently large k_i. Consequently, (4.54) implies

$$\Delta_{k_i} \leq \frac{1}{\rho_1 C \delta^2} \left(f(x^{k_i}) - f(x^{k_i+1}) \right).$$

Since $\{f(x^k)\}$ is monotonically decreasing, it therefore follows from (4.53) that

$$\sum_{i=0}^{\infty} \Delta_{k_i} \leq \frac{1}{\rho_1 C \delta^2} \sum_{i=0}^{\infty} \left(f(x^{k_i}) - f(x^{k_i+1}) \right) \leq \frac{1}{\rho_1 C \delta^2} \sum_{k=0}^{\infty} \left(f(x^k) - f(x^{k+1}) \right) < \infty. \tag{4.55}$$

If there are unsuccessful iterations between two successful ones, say k_i and k_{i+1}, we have

$$\Delta_{k_i+1} \leq \omega_2 \Delta_{k_i} \quad \text{and} \quad \Delta_{l+1} = \omega_1 \Delta_l \quad \forall l \in \{k_i + 1, \ldots, k_{i+1} - 1\}.$$

This implies

$$\sum_{l=k_i+1}^{k_{i+1}-1} \Delta_l \leq \Delta_{k_i+1} \sum_{j=0}^{\infty} \omega_1^j = \frac{1}{1-\omega_1} \Delta_{k_i+1} \leq \frac{\omega_2}{1-\omega_1} \Delta_{k_i}.$$

Together, we obtain from (4.55) that

$$\begin{aligned}\sum_{k=0}^{\infty}\Delta_k &= \sum_{k\in\{k_i\}}\Delta_k + \sum_{k\notin\{k_i\}}\Delta_k \\ &\le \sum_{i=0}^{\infty}\Delta_{k_i} + \frac{\omega_2}{1-\omega_1}\sum_{i=0}^{\infty}\Delta_{k_i} \\ &= \left(1+\frac{\omega_2}{1-\omega_1}\right)\sum_{i=0}^{\infty}\Delta_{k_i} < \infty,\end{aligned}$$

and the proof of (4.52) is complete. As a consequence of (4.52), it follows that

$$\min\{\Delta_k, 1\} = \Delta_k \quad \text{for all } k \in \mathbb{N} \text{ sufficiently large,} \tag{4.56}$$

and

$$\|p^k\| \le \|D_k^{1/2}\|\,\|D_k^{-1/2}p^k\| \le \|D_k^{1/2}\|\Delta_k \le C_1\Delta_k \longrightarrow 0 \tag{4.57}$$

since there is a constant $C_1 > 0$ such that $\|D_k^{1/2}\| \le C_1$ for all $k \in \mathbb{N}$ in view of Assumption (C). Moreover, we obtain from (4.57) that

$$\|x^{k+p} - x^k\| \le \sum_{j=0}^{p-1}\|x^{k+j+1} - x^{k+j}\| \le \sum_{j=0}^{p-1}\|p^{k+j}\| \le C_1\sum_{j=0}^{p-1}\Delta_{k+j}.$$

Consequently, (4.52) implies that $\{x^k\}$ is a Cauchy sequence and therefore convergent.

In the next part of the proof, we show that $\lim_{k\to\infty} r_k = 1$. To this end, first note that

$$\begin{aligned}\left|\text{pred}_k\left(p^k\right)\right|\,|r_k - 1| &= \left|\text{pred}_k\left(p^k\right)\right|\left|\frac{\text{ared}_k\left(p^k\right)}{\text{pred}_k\left(p^k\right)} - 1\right| \\ &= \left|\text{ared}_k\left(p^k\right) - \text{pred}_k\left(p^k\right)\right| \\ &= \left|f(x^k+p^k) - f(x^k) + m_k(0) - m_k(p^k)\right|.\end{aligned}$$

From the mean-value theorem, we therefore get the existence of a vector ξ^k between x^k and $x^k + p^k$ such that

$$\begin{aligned}\left|\text{pred}_k\left(p^k\right)\right|\,|r_k - 1| &= \left|\nabla f(\xi^k)^T p^k - \nabla f(x^k)^T p^k - \frac{1}{2}(p^k)^T H_k^T H_k p^k\right| \\ &\le \|\nabla f(\xi^k) - \nabla f(x^k)\|\|p^k\| + \frac{1}{2}\|H_k p^k\|^2 \\ &\le \|\nabla f(\xi^k) - \nabla f(x^k)\|\,C_1\Delta_k + \frac{1}{2}\|H_k\|^2 C_1^2\Delta_k^2,\end{aligned}$$

where the last inequality follows from (4.57). Dividing this expression by $\Delta_k > 0$, using Assumption (B), noting that $\Delta_k \to 0$ and $\|\nabla f(\xi^k) - \nabla f(x^k)\| \to 0$ since

∇f is continuous and both sequences $\{x^k\}, \{\xi^k\}$ converge to the same point (see (4.57)), we obtain

$$\frac{\left|\text{pred}_k\left(p^k\right)\right|}{\Delta_k}|r_k - 1| \to 0. \tag{4.58}$$

However, using (4.43), (4.51), and (4.56), we have

$$\frac{\text{pred}_k\left(p^k\right)}{\Delta_k} \geq C\|D_k^{1/2} g^k\|^2 \geq C\delta^2$$

for all $k \in \mathbb{N}$ sufficiently large. This implies $|r_k - 1| \to 0$ because of (4.58). This, in turn, gives $\Delta_{k+1} \geq \Delta_k$ for all these k, a contradiction to (4.52). □

Note that, if the entire sequence $\{x^k\}$ remains bounded, then (4.49) guarantees that at least one accumulation point of this sequence is a stationary point of the optimization problem (4.30), whereas (4.50) guarantees that every accumulation point is a solution of the box constrained system of equations (NE).

In order to prove a stronger convergence result than Theorem 4.19 with the limit inferior in (4.49) being replaced by the limit, we need to introduce another assumption, see also [67, Assumption (A.6)].

(D) The scaled gradient $D(x)^{1/2}\nabla f(x)$ is uniformly continuous.

Note that Assumption (D) is satisfied automatically on compact sets if $D(x)$ denotes the minimum-scaling from (4.33). This follows from the fact that both ∇f and $D(x)^{1/2}$ are continuous and therefore uniformly continuous on compact sets. This is in contrast to the scaling from (4.32) which is not continuous.

Theorem 4.20 *Suppose that Assumptions (A)–(D) hold. Then*

$$\lim_{k\to\infty} \|D_k^{1/2}\nabla f(x^k)\| = 0.$$

Proof. Similar to the proof of Theorem 4.19, we may assume that the test (4.41) in step (S.2) of Algorithm 4.15 is never accepted, so we always compute the direction p^k in step (S.3).

Suppose our statement is not true. Then there exists a constant $\delta > 0$ and a subsequence $\{x^k\}_K$ such that

$$\|D_k^{1/2}\nabla f(x^k)\| \geq 2\delta \quad \forall k \in K. \tag{4.59}$$

In view of Theorem 4.19, we have $\liminf_{k\to\infty} \|D_k^{1/2}\nabla f(x^k)\| = 0$. Therefore, we can find for each $k \in K$ an iteration index $\ell(k) > k$ such that

$$\|D_\ell^{1/2}\nabla f(x^\ell)\| \geq \delta \quad \forall k \leq \ell < \ell(k)$$

and

$$\|D_{\ell(k)}^{1/2}\nabla f(x^{\ell(k)})\| < \delta, \quad k \in K. \tag{4.60}$$

Using Assumptions (B), (C), there exist constants $C_1 > 0$ and $C_2 > 0$ such that $\|D_k^{1/2}\| \le C_1$ and $\|H_k\| \le C_2$ for all $k \in \mathbb{N}$.

Now, let $k \in K$ be fixed for the moment, take an arbitrary ℓ with $k \le \ell < \ell(k)$, and assume that the ℓ-th iteration is successful. Then we obtain from Lemma 4.18 that

$$f(x^\ell) - f(x^{\ell+1}) \ge \rho_1\big(f(x^\ell) - m_\ell(p^\ell)\big) \ge \rho_1 C\|D_\ell^{1/2}\nabla f(x^\ell)\|^2 \min\{\Delta_\ell, 1\}.$$

Since the left-hand side converges to zero and since

$$\|x^{\ell+1} - x^\ell\| = \|p^\ell\| \le \|D_\ell^{1/2}\|\,\|D_\ell^{-1/2}p^\ell\| \le \|D_\ell^{1/2}\|\Delta_\ell \le C_1\Delta_\ell,$$

we obtain

$$f(x^\ell) - f(x^{\ell+1}) \ge \rho_1 C\|D_\ell^{1/2}\nabla f(x^\ell)\|^2\Delta_\ell \ge \frac{\delta^2\rho_1 C}{C_1}\|x^{\ell+1} - x^\ell\|.$$

Trivially, this inequality also holds if the ℓ-th iteration is not successful. Consequently, we get

$$\begin{aligned}\frac{\delta^2\rho_1 C}{C_1}\|x^{\ell(k)} - x^k\| &\le \frac{\delta^2\rho_1 C}{C_1}\sum_{\ell=k}^{\ell(k)-1}\|x^{\ell+1} - x^\ell\| \\ &\le \sum_{\ell=k}^{\ell(k)-1}\big(f(x^\ell) - f(x^{\ell+1})\big) \\ &= f(x^k) - f(x^{\ell(k)})\end{aligned}$$

for all $k \in K$. Since $\{f(x^k)\}$ converges, this implies

$$\big\{\|x^{\ell(k)} - x^k\|\big\}_{k\in K} \to 0.$$

In view of Assumption (D), we therefore have

$$\big\{\|D_{\ell(k)}^{1/2}\nabla f(x^{\ell(k)}) - D_k^{1/2}\nabla f(x^k)\|\big\}_{k\in K} \to 0.$$

On the other hand, it follows from (4.59) and (4.60) that

$$\begin{aligned}\big\|D_{\ell(k)}^{1/2}\nabla f(x^{\ell(k)}) - D_k^{1/2}\nabla f(x^k)\big\| &\ge \big\|D_k^{1/2}\nabla f(x^k)\big\| - \big\|D_{\ell(k)}^{1/2}\nabla f(x^{\ell(k)})\big\| \\ &\ge 2\delta - \delta = \delta.\end{aligned}$$

This contradiction completes the proof. □

4.4 Local Convergence

In this section, we consider the local convergence properties of Algorithm 4.15. More precisely, we show in the following result that, locally, the projected and truncated Newton-direction is always accepted in step (S.2), and that this direction guarantees superlinear and even quadratic convergence under suitable assumptions.

Theorem 4.21 *Let $\{x^k\}$ be a sequence generated by Algorithm 4.15, and let x^* be an accumulation point of this sequence such that $F(x^*) = 0$ and all elements $H_* \in \partial F(x^*)$ are nonsingular. Then the following statements hold:*

(a) The entire sequence $\{x^k\}$ converges to x^*.

(b) The rate of convergence is Q-superlinear.

(c) If F is strongly semismooth, the rate of convergence is Q-quadratic.

Proof. Since all elements $H_* \in \partial F(x^*)$ are nonsingular, it follows from [58, Proposition 3.1] that there exist constants $\varepsilon_1 > 0$ and $c > 0$ such that

$$\|H(x)^{-1}\| \leq c \quad \forall x \in B_{\varepsilon_1}(x^*), \ \forall H(x) \in \partial F(x). \tag{4.61}$$

Moreover, being semismooth, F is locally Lipschitz continuous. Hence there exist constants $\varepsilon_2 > 0$ and $L_1 > 0$ with

$$\|F(x) - F(y)\| \leq L_1 \|x - y\| \quad \forall x, y \in B_{\varepsilon_2}(x^*). \tag{4.62}$$

Furthermore, the nonsingularity assumption and [10, Theorem 7.1.1] implies that the inverse function F^{-1} exists in a sufficiently small neighbourhood of $F(x^*)$, and this function is also locally Lipschitz. Consequently, we get the existence of two constants $\varepsilon_3 > 0$ and $L_2 > 0$ such that

$$\left\|F^{-1}(F(x)) - F^{-1}(F(y))\right\| \leq L_2 \|F(x) - F(y)\| \quad \forall x, y \in B_{\varepsilon_3}(x^*). \tag{4.63}$$

Using Proposition 2.19 and the semismoothness of F, we see that there is another constant $\varepsilon_4 > 0$ such that

$$\|F(x) - F(x^*) - H(x)(x - x^*)\| \leq \min\left\{\frac{\eta}{2cL_1L_2}, \frac{1}{4c}\right\} \|x - x^*\| \tag{4.64}$$

for all $x \in B_{\varepsilon_4}(x^*)$ and all $H(x) \in \partial F(x)$, where η denotes the constant from (4.41). Moreover, by continuity, there is a constant $\varepsilon_5 > 0$ with

$$\|F(x)\| \leq \min\left\{\frac{\eta}{2cL_1L_2}, \frac{1-\sigma}{c}\right\} \quad \forall x \in B_{\varepsilon_5}(x^*), \tag{4.65}$$

where $\sigma \in (0,1)$ denotes the constant from Algorithm 4.15. Finally, the nonsingularity assumption implies that there is another constant $\varepsilon_6 > 0$ such that σ_k from (4.40) satisfies

$$\sigma_k = 1 - \big\|P_\Omega(x^k + p_N^k) - x^k\big\| \geq \frac{3}{4} \quad \forall x^k \in B_{\varepsilon_6}(x^*). \tag{4.66}$$

Now define $\varepsilon := \min\{\varepsilon_i : i = 1, \ldots, 6\}$. Since x^* is an accumulation point of the sequence $\{x^k\}$, we can choose an iterate x^k (the index k is fixed for the moment) such that $x^k \in B_\varepsilon(x^*) \cap \operatorname{int}\Omega$. We will show that the next iterate also belongs to this neighbourhood and, in fact, is actually much closer to x^* than x^k is. The proof of statement (a) then follows by a simple induction argument.

To this end, we first note that H_k is nonsingular in view of (4.61), and we therefore obtain from (4.64) that

$$\begin{aligned} \|x^k + p_N^k - x^*\| &= \|x^k - H_k^{-1}F(x^k) - x^*\| \\ &\leq \|H_k^{-1}\|\|F(x^k) - F(x^*) - H_k(x^k - x^*)\| \\ &\leq \min\Big\{\frac{\eta}{2L_1L_2}, \frac{1}{4}\Big\}\|x^k - x^*\|, \end{aligned} \tag{4.67}$$

in particular, $x^k + p_N^k$ also belongs to the neighbourhood $B_\varepsilon(x^*)$ of the solution x^*. Since we can write

$$x^k + p_{PN}^k - x^* = \sigma_k\big(P_\Omega(x^k + p_N^k) - x^*\big) + (1 - \sigma_k)(x^k - x^*), \tag{4.68}$$

it is easy to see that this also implies that the vector $x^k + p_{PN}^k$ is in the neighbourhood $B_\varepsilon(x^*)$ of x^*. Using $x^k \in \operatorname{int}\Omega$, (4.61), (4.65), and the nonexpansiveness of the projection operator, we get

$$\|P_\Omega(x^k + p_N^k) - x^k\| = \|P_\Omega(x^k + p_N^k) - P_\Omega(x^k)\| \leq \|p_N^k\| \leq \|H_k^{-1}\|\|F(x^k)\| \leq 1 - \sigma.$$

In view of (4.40), (4.66), this yields

$$1 - \sigma_k = \|P_\Omega(x^k + p_N^k) - x^k\| \leq \|H_k^{-1}\|\|F(x^k)\| \leq c\|F(x^k)\|. \tag{4.69}$$

Using (4.62), (4.68), (4.67), (4.69), (4.65), (4.63), $\sigma_k \leq 1$, and the nonexpansiveness of the projection operator, we get

$$\begin{aligned} \|F(x^k + p_{PN}^k)\| &= \|F(x^k + p_{PN}^k) - F(x^*)\| \\ &\leq L_1\|x^k + p_{PN}^k - x^*\| \\ &\leq L_1\sigma_k\|x^k + p_N^k - x^*\| + L_1(1 - \sigma_k)\|x^k - x^*\| \\ &\leq \frac{\eta}{2L_2}\|x^k - x^*\| + L_1c\|F(x^k)\|\,\|x^k - x^*\| \\ &\leq \frac{\eta}{2L_2}\|x^k - x^*\| + \frac{\eta}{2L_2}\|x^k - x^*\| \end{aligned}$$

$$\begin{aligned} &= \frac{\eta}{L_2}\big\|F^{-1}(F(x^k)) - F^{-1}(F(x^*))\big\| \\ &\leq \eta\|F(x^k) - F(x^*)\| \\ &= \eta\|F(x^k)\|. \end{aligned}$$

Hence the projected Newton direction is accepted in step (S.2), and the next iterate is given by $x^{k+1} = x^k + p^k_{PN}$. Together with (4.66), (4.67), and (4.68), this implies

$$\begin{aligned} \|x^{k+1} - x^*\| &= \|x^k + p^k_{PN} - x^*\| \\ &\leq \sigma_k\|x^k + p^k_N - x^*\| + (1-\sigma_k)\|x^k - x^*\| \\ &\leq \|x^k + p^k_N - x^*\| + \frac{1}{4}\|x^k - x^*\| \\ &\leq \frac{1}{2}\|x^k - x^*\|. \end{aligned}$$

Therefore, we also have $x^{k+1} \in B_\varepsilon(x^*)$. Using an induction argument, it follows that the test $\|F(x^k + p^k_{PN})\| \leq \eta\|F(x^k)\|$ is satisfied for all sufficiently large $k \in \mathbb{N}$, so that we have $x^{k+1} = x^k + p^k_{PN}$ and

$$\|x^{k+1} - x^*\| \leq \frac{1}{2}\|x^k - x^*\| \tag{4.70}$$

for all $k \in \mathbb{N}$ large enough. In particular, the sequence $\{x^k\}$ is well-defined and converges (at least) linearly to x^*.

To prove superlinear convergence, we consider the inequality

$$\|x^{k+1} - x^*\| \leq \sigma_k\|x^k + p^k_N - x^*\| + (1-\sigma_k)\|x^k - x^*\| \tag{4.71}$$

again, cf. (4.68). For sufficiently large $k \in \mathbb{N}$, it follows from (4.69) and (4.62) that

$$1 - \sigma_k \leq c\|F(x^k)\| = c\|F(x^k) - F(x^*)\| = O(\|x^k - x^*\|).$$

Using Proposition 2.19 and the semismoothness of F, we get

$$\|x^k + p^k_N - x^*\| \leq \|H_k^{-1}\|\|F(x^k) - F(x^*) - H_k(x^k - x^*)\| = o(\|x^k - x^*\|).$$

Since $\sigma_k \to 1$, we therefore obtain from (4.71) that $\|x^{k+1} - x^*\| = o(\|x^k - x^*\|)$. Hence the local rate of convergence is superlinear.

If F is strongly semismooth, it follows from Proposition 2.19 that

$$\|x^k + p^k_N - x^*\| \leq \|H_k^{-1}\|\|F(x^k) - F(x^*) - H_k(x^k - x^*)\| = O(\|x^k - x^*\|^2).$$

Since $1 - \sigma_k = O(\|x^k - x^*\|)$, we therefore get $\|x^{k+1} - x^*\| = O(\|x^k - x^*\|^2)$ from (4.71). Hence the rate of convergence is locally quadratic in the strongly semismooth case. □

The assumptions for local quadratic convergence in Theorem 4.21 are satisfied, e.g., if F is continuously differentiable with F' being locally Lipschitz continuous and $F'(x^*)$ being nonsingular. However, in some applications the assumptions of Theorem 4.21 also hold for nonsmooth F, especially in the context of complementarity problems and variational inequalities, see, for example, [16, 29].

Remark 4.22 We close this section by noting that our (global and) local convergence theory would remain true if we would solve the linear system of equations (4.38) only inexactly. The classical reference for smooth (unconstrained) equations is [17], and an extension to semismooth equations may be found in [51, 24]. From a practical point of view, it is our experience, however, that such an extension is much less obvious since it requires an iterative linear system solver (like a Krylov subspace method, see [61]) for nonsymmetric linear systems of equations. Typically, we then need a very good preconditioner in order for such a method to be effective, and the choice of a suitable preconditioner depends very much on the particular problem that we want to solve.

4.5 Numerical Experiments

In this section, we apply Algorithm 4.15 to several test problems of different types. Some of these problems are originally not given in the form of a nonlinear system of equations with box constraints, but can be reformulated in this way. We implemented Algorithm 4.15 in MATLAB using the scaling matrix from (4.33) and the following constants:

$$\begin{aligned} &\sigma = 0.995, \quad \theta = 0.95, \quad \eta = 0.1, \quad \gamma = 1, \quad \omega_1 = 0.25, \quad \omega_2 = 2, \\ &\rho_1 = 0.1, \quad \rho_2 = 0.75, \quad \Delta_0 = 1. \end{aligned}$$

We terminate the iteration, if one of the following criteria is satisfied:

$$\|D_k^{1/2} \nabla f(x^k)\| \leq 10^{-6} \quad \text{or} \quad \|F(x^k)\|_\infty \leq 10^{-6}.$$

As a safeguard we stop the iteration, if

$$k \geq k_{\max} := 500 \quad \text{or} \quad \Delta_k \leq \Delta_{\min} := 10^{-8}.$$

The search direction p^k satisfying (4.37) in step (S.3) is computed in such a way that we try to avoid using the Cauchy point whenever this is possible. To this end, we make use of the following three directions:

- the projected and truncated Newton direction p^k_{PN} from (4.39)
- the truncated Newton direction tp^k_N from (4.38), where $t > 0$ is a stepsize such that $x^k + tp^k_N$ belongs to the interior of Ω
- the Cauchy step p^k_{CP} itself.

At each iteration, we then choose the search direction in the following way:

- If p^k_{PN} satisfies (4.41), we take $x^{k+1} := x^k + p^k_{PN}$.
- Otherwise, if $p^k := tp^k_N$ satisfies (4.37) for some not too small $t > 0$ (we use a backtracking and require $t > 10^{-6}$ in our current implementation), we accept $x^{k+1} := x^k + p^k$ as our new iterate.
- If none of these two strategies work, we use a dogleg-type step from the Cauchy point $x^k + p^k_{CP}$ to the Newton point $x^k + p^k_N$. To this end, we compute a suitable vector on the connecting line between these two points such that this vector is strictly feasible for our box constraints and belongs to the trust-region. If this point satisfies (4.37), it becomes our vector of choice, otherwise we take the Cauchy point itself.

All these choices require that we are able to solve the linear system (4.38). If this is not possible, we take the Cauchy point as our new iterate. Note that all global and local convergence properties remain true for this modification of Algorithm 4.15.

We next give a summary of the test problems that are used in our numerical experiments:

1. **The Chandrasekhar H-equation**. A discretization of Chandrasekhar's H-equation leads to a nonlinear system of equations that depends on a parameter $c \in [0, 1]$, see [44, p. 87] for more details. Since this system has two solutions and only one has a physical meaning, we use the bounds $l_i := 0$ and $u_i := \infty$ for all $i = 1, \ldots, n$. We choose $x^0 \in \mathbb{R}^n$ with $x^0_i := 1$ for all i and consider the cases $c = 0.99$, $c = 0.9999$, and $c = 1$ with $n = 1000$.

2. **The seven-diagonal problem**. This is a nonlinear system of equations of variable dimension that can be found in [50]. The Jacobians have the structure of band matrices, which allows us to consider high dimensional cases. The system has several solutions, so we use the bounds $l_i := 0$ and $u_i := \infty$, $i = 1, \ldots, n$, to avoid negative ones. We choose $n = 100000$ and $x^0_i := 1$ for all $i = 1, \ldots, n$.

3. **A countercurrent reactor problem**. This problem can be found in [50] as well. Again the problem has variable dimension, several solutions, and the Jacobians are band matrices. We consider $n = 10000$, $n = 100000$ and set $l_i := -1$, $u_i := \infty$ and $x_i^0 := 1$ for $i = 1, \ldots, n$.

4. **A chemical equilibrium problem**. In [53, system 1], a nonlinear system of 11 equations and variables is described. A solution of this system is only physically meaningful if all components of the solution are real and positive. Following [4] we augment the system given in [53] to the size of $n = 11000$ and use $l_i := 0$, $u_i := \infty$ and $x_i^0 := 1$ for $i = 1, \ldots, n$.

5. **Boundary value problems**. Discretizing a boundary value problem leads to a nonlinear system of equations. If this problem has several solutions, the use of bound constraints is quite helpful to avoid, for example, negative solutions. We use three different boundary value problems for our numerical test runs, called BVP1, BVP2, and BVP3 in the following. BVP1 is a two-point boundary value problem from [45, Example 2.7.4] which has at least two solutions. We use the discretization given in [45] ($n = 800$) to approximate the function and the first derivative. In order to get positive function values, we set $l_i := 0$ for all odd i, $l_i := -\infty$ for all even i and $u_i = \infty$ for $i = 1, \ldots, n$. BVP2 is taken from [54]. The discretization given in [54] leads to a nonlinear system that has a unique solution in the box defined by $l_i = -0.5$ and $u_i = 0$ for $i = 1, \ldots, n$. We set $n = 500$, use the given bounds and start with $x_i^0 := -0.25$ for all $i = 1, \ldots, n$. Finally, BVP3 is the boundary value problem described in the introduction and in [63, p. 504]. The problem has two solutions, but only one is positive. We use the discretization from the introduction and approximate the positive solution by setting $l_i := 0$, $u_i := \infty$. The dimension of this problem is $n = 500$, and we take $x_i^0 := 1$ for all $i = 1, \ldots, n$.

6. **The Floudas et al. collection**. In [33, Section 14.1], Floudas et al. present a collection of box constrained nonlinear systems of equations. This collection contains nine examples. The dimension of these examples is small, nevertheless, some of these problems are challenging. All examples have finite lower and upper bounds. We choose $x^0 := l + 0.25(u - l)$ as initial iterate for all test problems.

7. **Complementarity problems**. Here one tries to find a solution of the system

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

where $G : \mathbb{R}_+^n \to \mathbb{R}^n$ is a given function that is sometimes not defined outside the nonnegative orthant. This complementarity problem can be

reformulated as a square system of equations with box constraints in the following way:

$$x \geq 0,\ y \geq 0,\ G(x) - y = 0,\ x_i y_i = 0\ \ \forall i = 1, \ldots, n.$$

Similar reformulations are possible for the slightly more general class of mixed complementarity problems, and a large number of (often very difficult) test problems of this class is given in the MCPLIB collection, see [23]. Since the standard starting points for these examples are sometimes on the boundary of the feasible region, we use the strategy from [65] and project the standard starting vector x^0 of the MCPLIB on the smaller box $[\hat{l}, \hat{u}]$ with $\hat{l}_i = l_i + 0.01$ and $\hat{u}_i = u_i - 0.01$ for $i = 1, \ldots, n$ for all finite lower and upper bounds, whereas we use $y_i^0 := 1$ for the slack variables. (Note that all lower and upper bounds from the MCPLIB with absolute value 10^{20} are treated as infinite bounds.)

In the following tables, we present our numerical results. For each test problem, the size of the problem (n), the number of iterations (iter), the evaluations of the function F (eval), the norm of this function in the last iterate ($\|F(x)\|$), and the norm of the stopping criterion in the last iterate ($\|D^{1/2}(x)g(x)\|$) are given. Moreover, the number of iterations needed by the STRSCNE code described in [3] is presented. If a method fails to solve a problem, this is denoted by "–". Table 4.1 contains the results obtained for all examples not taken from the MCPLIB collection.

problem	n	iter	eval	$\|F(x)\|$	$\|D^{1/2}(x)g(x)\|$	strscne
H-equation, $c = 0.99$	1000	8	15	9.253655e-07	1.618650e-05	9
H-equation, $c = 0.9999$	1000	11	21	4.805774e-08	9.663247e-07	11
H-equation, $c = 1$	1000	14	29	3.564824e-07	7.302551e-06	16
7-diagonal	100000	6	8	1.024474e-07	1.201056e-06	6
reactor	10000	20	38	9.408478e-09	2.762200e-08	17
reactor	100000	33	63	1.760088e-08	5.488085e-08	24
chemical-eq.	11000	13	23	3.596600e-08	2.187478e-06	27
BVP1	800	7	13	4.219342e-10	6.250180e-10	6
BVP2	500	2	3	6.250000e-06	2.210701e-08	4
BVP3	500	3	4	3.750000e-07	7.526758e-07	6
Floudas no. 1	2	5	7	9.663381e-13	1.249890e-10	5
Floudas no. 2	5	12	22	9.044287e-08	4.664314e-07	12
Floudas no. 3	2	30	57	3.793600e-09	4.301148e-06	20
Floudas no. 4	2	4	6	1.040490e-07	1.158006e-07	5
Floudas no. 5	5	9	16	1.677537e-09	3.751097e-09	9
Floudas no. 6	8	6	10	1.466922e-07	1.214811e-07	9
Floudas no. 7	9	–	–	–	–	–
Floudas no. 8	2	7	13	1.300791e-07	1.028366e-07	5
Floudas no. 9	1	3	5	2.089180e-05	8.711504e-07	4

Table 4.1: Results for problem classes 1–8

Both our method and the related algorithm from [3] are able to solve all test examples with the exception of one example from the Floudas et al. collection. However, this problem is regarded as very challenging, even the solution given in [33] seems to be wrong. The behaviour of the two methods on the other examples is more or less similar, and there is no clear winner on this set of test problems.

We next present our numerical results for all test examples of dimension $n \geq 1000$ from the MCPLIB. The corresponding numerical results are contained in Table 4.2. The columns have the same meaning as in Table 4.1, in particular, we also compare our results with those obtained by using the STRSCNE code from [3].

							STRSCNE	
problem	n	iter	eval	cpu	$\|\Phi(x)\|$	$\|D^{1/2}(x)g(x)\|$	iter	cpu
bert_oc	5000	14	26	52.21	1.726599e-05	7.703553e-07	17	37.21
bishop	1645	–	–	–	–	–	–	–
bratu	5625	14	28	195.01	2.675836e-06	9.231569e-07	21	158.12
obstacle	2500	13	26	36.68	3.030147e-07	3.258603e-07	20	31.25
opt_cont31	1024	16	30	4.32	5.846047e-07	9.324531e-08	–	–
opt_cont127	4096	14	25	45.99	1.265241e-10	7.121115e-10	–	–
opt_cont255	8192	13	24	173.49	6.015165e-07	4.677748e-06	–	–
opt_cont511	16384	18	33	863.47	1.419340e-07	6.629523e-08	–	–
trafelas	2904	68	136	337.21	9.407251e-07	7.632607e-06	79	292.28

Table 4.2: Results for large-scale mixed complementarity problems (smooth reformulation)

Algorithm 4.15 is able to solve all examples with the exception of the `bishop` problem. In this case, our method compares very favourably with the STRSCNE code from [3] which produces five error messages and is able to solve only four test examples. In this respect, however, it should be noted that the above problems have solutions on the boundary of the box Ω, and that the STRSCNE code is guaranteed to converge locally quadratically only to interior solutions (a modification of that method, which overcomes this disadvantage, has been presented in Section 4.1.2 or [5], but is currently not implemented in the STRSCNE package).

Table 4.2 also gives the CPU times for both methods, and there STRSCNE is somewhat better than our code for those problems which are solved by both methods. On the other hand, the CPU times differ (sometimes significantly) between several test runs on the same problem, hence a serious comparison based on CPU times seems to be difficult. In principle, both methods should have a similar amount of work per iteration since the most time-consuming part is the solution of a linear system of equations.

We stress, however, that the reformulation of the complementarity problems

used for our test runs is not necessarily the best formulation. For example, if a solution x^* of the complementarity problem is degenerate, i.e., if there is at least one index such that both $x_i^* = 0$ and $G_i(x^*) = 0$, then it is easy to see that the Jacobian F' of our reformulated system is singular at the solution, hence we cannot expect quadratic convergence.

We illustrate this point using the Kojima-Shindo example. This is a complementarity problem with four variables which has two solutions, one is nondegenerate and one is degenerate. Using the standard starting point, our method converges to the degenerate solution. The iteration history is given in Table 4.3.

k	$\|\Phi(x^k)\|$	$\|D_k^{1/2} g(x^k)\|$	eval	Δ_k	direction
0	2.475100e+00	1.952834e+01	2	2	proj. Newton
1	5.716714e-02	7.597381e-01	4	4	trunc. Newton
2	9.601738e-03	1.297287e-01	5	8	proj. Newton
3	1.898870e-04	1.429594e-03	7	16	trunc. Newton
4	2.637017e-05	1.983471e-04	8	32	proj. Newton
5	1.328548e-07	1.369230e-06	9	64	proj. Newton
6	2.797033e-09	2.862149e-08	10	128	proj. Newton

Table 4.3: Iteration history for the smooth reformulation of the Kojima-Shindo problem

Clearly, Table 4.3 shows that we do not have quadratic convergence, although the rate of convergence is still relatively fast. If we would require higher accuracy, however, we would run into singularity problems. In fact, if we iterate a bit further, we see that we get very slow convergence using Cauchy points all the time from iteration 8 on.

There exist other reformulations of the complementarity problem as a semismooth system of equations such that the corresponding merit function is continuously differentiable and such that quadratic convergence can still be expected even in the case of degenerate solutions. We refer to [16, 9] for the corresponding background. In particular, using the semismooth reformulation from [9] and applying our code to this reformulation using the Kojima-Shindo example once again, we get the iteration history from Table 4.4.

k	$\|\Phi(x^k)\|$	$\|D_k^{1/2} g(x^k)\|$	eval	Δ_k	direction
0	2.067848e-01	1.722355e+00	3	2	trunc. Newton
1	3.911732e-02	1.252072e+00	4	4	proj. Newton
2	1.118381e-04	8.304539e-04	5	8	proj. Newton
3	7.703886e-09	5.770644e-08	6	16	proj. Newton

Table 4.4: Iteration history for the semismooth reformulation of the Kojima-Shindo problem

Obviously, the method is quadratically convergent in this case. Note that we cannot apply the STRSCNE code from [3] to this semismooth reformulation since this method requires smooth functions F. On the other hand, if we apply our method to the semismooth reformulation of the large-scale mixed complementarity problems from Table 4.2, we get similar results. More precisely, we can solve once again all problems with the exception of `bishop`, and the number of iterations are $9, -, 19, 11, 8, 9, 14, 25, 76, 262$ for the test examples from Table 4.2, respectively, where the $-$ indicates the failure on the `bishop` example.

Chapter 5

Conclusion

In the preceding chapters we have introduced affine-scaling methods for two different types of mathematical problems with box constraints.

The first class of problems are bound constrained optimization problems. We have introduced a new scaling technique for the solution of these problems by affine-scaling interior-point Newton methods. Using this scaling technique, the strict complementarity condition is not needed in order to prove local quadratic convergence. Moreover, this new scaling allows a much simpler local convergence proof by using standard results from nonsmooth analysis. The analysis carried out is essentially local. In absence of a suitable descent property for the objective function a possible globalization based on the first order optimality conditions is presented. The local convergence properties of the new affine-scaling method are illustrated on two well-known numerical examples that show the advantage of our scaling if strict complementarity does not hold for the solution.

The second class of problems considered here are semismooth systems of equations subject to bound constraints. Here we have introduced an interior-point trust-region method. The method also follows the affine-scaling approach and generates strictly feasible iterates. It differs from other methods of this type in the choice of the scaling matrix and the transition from the global to the local method. Moreover, the method can be applied to both continuously differentiable and semismooth systems of equations. Hence the method is applicable to a wider class of problems than other affine-scaling methods, and this was illustrated for the class of complementarity problems. Compared to other methods, we also avoid a nonsingularity assumption that is used in order to get a well-defined method. Global and fast local convergence results are established for this method and its numerical properties are tested on various smooth and semismooth examples.

The two problem types we consider are closely related. We stress, however,

that there is a significant difference in applying affine-scaling methods to either bound constrained optimization problems or to nonlinear equations with box constraints. In the first case the affine-scaling approach is used in order to get good local convergence properties without assuming strict complementarity. Whereas in the second case fast local convergence is guaranteed by a suitable modification of the standard Newton step for the unconstrained problem $F(x) = 0$. The affine-scaling approach is used in order to get suitable global convergence properties.

Bibliography

[1] E.L. Allgower and K. Georg: *Introduction to Numerical Continuation Methods*. John Wiley & Sons, New York, NY, 1979.

[2] S. Bellavia, M. Macconi, and B. Morini: *An affine scaling trust-region approach to bound-constrained nonlinear systems*. Applied Numerical Mathematics 44, 2003, pp. 257–280.

[3] S. Bellavia, M. Macconi, and B. Morini: *STRSCNE: A scaled trust-region solver for constrained nonlinear systems*. Computational Optimization and Applications 28, 2004, pp. 31–50.

[4] S. Bellavia, M. Macconi, and B. Morini: *A two-dimensional trust-region method for large scale bound-constrained nonlinear systems*. Technical Report, submitted for publication.

[5] S. Bellavia and B. Morini: *An interior global method for nonlinear systems with simple bounds*. Optimization Methods and Software 20, 2005, pp. 453–474.

[6] S. Bellavia and B. Morini: *Subspace Trust-Region methods for large bound constrained nonlinear equations*. SIAM Journal on Numerical Analysis, to appear.

[7] D.P. Bertsekas: *Projected Newton methods for optimization problems with simple constraints*. SIAM Journal on Control and Optimization 20, 1982, pp. 221–246.

[8] R.H. Byrd, P. Lu, and J. Nocedal: *A limited memory algorithm for bound constrained optimization*. SIAM Journal on Scientific and Statistical Computing 16, 1995, pp. 1190–1208.

[9] B. Chen, X. Chen, and C. Kanzow: *A penalized Fischer-Burmeister NCP-function*. Mathematical Programming 88, 2000, pp. 211–216.

[10] F.H. Clarke: *Optimization and Nonsmooth Analysis*. Wiley, New York, 1983.

[11] T.F. Coleman and Y. Li: *On the convergence of interior reflective Newton methods for nonlinear minimization subject to bounds*. Mathematical Programming 67, 1994, pp. 189–224.

[12] 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.

[13] A.R. Conn, N.I.M. Gould, and Ph.L. Toint: *Global convergence of a class of trust region algorithms for optimization with simple bounds*. SIAM Journal on Numerical Analysis 25, 1988, pp. 433–460 (Correction in: SIAM Journal on Numerical Analysis 26, 1989, pp. 764–767).

[14] A.R. Conn, N.I.M. Gould, and Ph.L. Toint: *Testing a class of methods for solving minimization problems with simple bounds on the variables*. Mathematics of Computation 50, 1988, pp. 399–430.

[15] A.R. Conn, N.I.M. Gould, and Ph.L. Toint: *Trust-Region Methods*. MPS-SIAM Series on Optimization, SIAM, Philadelphia, PA, 2000.

[16] T. De Luca, F. Facchinei, and C. Kanzow: *A semismooth equation approach to the solution of nonlinear complementarity problems*. Mathematical Programming 75, 1996, pp. 407–439.

[17] R.S. Dembo, S.C. Eisenstat, and T. Steihaug: *Inexact Newton methods*. SIAM Journal on Numerical Analysis 19, 1982, pp. 400–408.

[18] J.E. Dennis and R.B. Schnabel: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[19] J.E. Dennis and L.N. Vicente: *Trust-region interior-point algorithms for minimization problems with simple bounds*. In: H. Fischer, B. Riedmüller and S. Schäffler (eds.): *Applied Mathematics and Parallel Computing. Festschrift for Klaus Ritter*. Physica, Heidelberg, 1996, pp. 97–107.

[20] P. Deuflhard: *Newton Methods for Nonlinear Problems*. Springer, Berlin, Heidelberg, 2004.

[21] G. Di Pillo: *Exact penalty methods*. In: E. Spedicato (ed.): *Algorithms for Continuous Optimization, The State of the Art*. Kluwer, 1994, pp. 209–254.

[22] G. Di Pillo and L. Grippo: *A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints*. SIAM Journal on Control and Optimization 23, 1985, pp. 72–84.

[23] S.P. Dirkse and M.C. Ferris: *MCPLIB: A collection of nonlinear mixed complementarity problems.* Optimization Methods and Software 5, 1995, pp. 319–345.

[24] 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.

[25] F. Facchinei, A. Fischer, and C. Kanzow: *On the accurate identification of active constraints.* SIAM Journal on Optimization 9, 1999, pp. 14–32.

[26] F. Facchinei, J. Júdice, and J. Soares: *An active set Newton algorithm for large-scale nonlinear programs with box constraints.* SIAM Journal on Optimization 8, 1998, pp. 158–186.

[27] F. Facchinei, S. Lucidi, and L. Palagi: *A truncated Newton algorithm for large scale box constrained optimization.* SIAM Journal on Optimization 12, 2002, pp. 1100–1125.

[28] F. Facchinei and J.-S. Pang: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I.* Springer, New York-Berlin-Heidelberg, 2003.

[29] F. Facchinei and J.-S. Pang: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume II.* Springer, New York, Berlin, Heidelberg, 2003.

[30] F. Facchinei and J. Soares: *A new merit function for nonlinear complementarity problems and a related algorithm.* SIAM Journal on Optimization 7, 1997, pp. 225–247.

[31] M.C. Ferris and J.-S. Pang: *Engineering and economic applications of complementarity problems.* SIAM Review 39, 1997, pp. 669–713.

[32] A. Fischer: *Solution of monotone complementarity problems with locally Lipschitzian functions.* Mathematical Programming 76, 1997, pp. 513–532.

[33] C.A. Floudas et al.: *Handbook of test problems in local and global optimization.* Kluwer Academic Publishers, Dordrecht, 1999.

[34] A. Friedlander, J.M. Martínez, and S.A. Santos: *A new trust region algorithm for bound constrained minimization.* Applied Mathematics and Optimization 30, 1994, pp. 235–266.

[35] C. Geiger and C. Kanzow: *Numerische Verfahren zur Lösung unrestringierter Optimierungsaufgaben*. Springer, Berlin-Heidelberg-New York, 1999.

[36] C. Geiger and C. Kanzow: *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin-Heidelberg-New York, 2002.

[37] 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.

[38] E. Jehle, K. Müller and H. Michael: *Produktionswirtschaft*. Verlag Recht und Wirtschaft, Heidelberg, Germany, 4th edition, 1994.

[39] C. Kanzow: *Strictly feasible equation-based methods for mixed complementarity problems*. Numerische Mathematik 89, 2001, pp. 135–160.

[40] C. Kanzow: *Strictly feasible equation-based methods for mixed complementarity problems*. Numerische Mathematik 89, 2001, pp. 135–160.

[41] C. Kanzow: *An active set-type Newton method for constrained nonlinear systems*. In: M.C. Ferris, O.L. Mangasarian, and J.-S. Pang (eds.): *Complementarity: Applications, Algorithms and Extensions*. Kluwer Academic Publishers, Dordrecht, 2001, pp. 179–200.

[42] C. Kanzow and A. Klug: *On affine-scaling interior-point Newton methods for nonlinear minimization with bound constraints*. Computational Optimization and Applications, to appear.

[43] C. Kanzow and A. Klug: *An interior-point affine-scaling trust-region method for semismooth equations with box constraints*. Computational Optimization and Applications, to appear.

[44] C.T. Kelley: *Iterative Methods for Linear and Nonlinear Equations*. SIAM, Philadelphia, PA, 1995.

[45] C.T. Kelley: *Solving Nonlinear Equations with Newton's Method*. SIAM, Philadelphia, PA, 2003.

[46] D.N. Kozakevich , J.M. Martínez, and S.A. Santos: *Solving nonlinear systems of equations with simple constraints*. Computational and Applied Mathematics 16, 1997, pp. 215–235.

[47] M. Lescrenier: *Convergence of trust region algorithms for optimization with bounds when strict complementarity does not hold*. SIAM Journal on Numerical Analysis 28, 1991, pp. 476–495.

[48] R.M. Lewis and V. Torczon: *Pattern search algorithms for bound constrained minimization*. SIAM Journal on Optimization 9, 1999, pp. 1082–1099.

[49] C.-J. Lin and J.J. Moré: *Newton's method for large bound-constrained optimization problems*. SIAM Journal on Optimization 9, 1999, pp. 1100–1127.

[50] L. Lukšan: *Inexact Trust Region Method for Large Sparse Systems of Nonlinear Equations*. Journal of Optimization Theory and Applications 81, 1994, pp. 569–590.

[51] J.M Martínez and L. Qi: *Inexact Newton methods for solving nonsmooth equations*. Journal of Computation and Applied Mathematics 60, 1995, pp. 127–145.

[52] Matlab: Software, The Mathworks Inc., Version 7.1.

[53] K. Meintjes and A.P. Morgan: *Chemical Equilibrium Systems as Numerical Test Problems*, ACM Transactions on Mathematical Software 16, 1990, pp. 143–151.

[54] J.J. Moré and M.Y. Cosnard: *Numerical solution of nonlinear equations*. ACM Transactions on Mathematical Software 5, 1979, pp. 64–85.

[55] J.M. Ortega and W.C. Rheinboldt: *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, London, 1970.

[56] J.-S. Pang and L. Qi: *Nonsmooth equations: motivation and algorithms*. SIAM Journal on Optimization 3, 1993, pp. 443–465.

[57] L. Qi: *Convergence analysis of some algorithms for solving nonsmooth equations*. Mathematics of Operations Research 18, 1993, pp. 227–244.

[58] L. Qi and J. Sun: *A nonsmooth version of Newton's method*. Mathematical Programming 58, 1993, pp. 353–367.

[59] L. Qi, X. Tong, and D. Li: *Active-set projected trust region algorithm for box constrained nonsmooth equations*. Journal of Optimization Theory and Applications 120, 2004, pp. 601–625.

[60] S.M. Robinson: *Strongly regular generalized equations*. Mathematics of Operations Research 5, 1980, pp. 43–62.

[61] Y. Saad: *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, second edition 2003.

[62] A. Schwartz and E. Polak: *Family of projected descent methods for optimization problems with simple bounds.* Journal of Optimization Theory and Applications 92, 1997, pp. 1–31.

[63] J. Stoer and R. Bulirsch: *Introduction to Numerical Analysis.* Springer, New-York, NY, 2nd ed., 1993.

[64] X.J. 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. 187–211.

[65] M. Ulbrich: *Non-monotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems.* SIAM Journal on Optimization 11, 2001, pp. 889–917.

[66] 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.

[67] M. Ulbrich, S. Ulbrich, and M. Heinkenschloss: *Global convergence of trust-region interior-point algorithms for infinite-dimensional nonconvex minimization subject to pointwise bounds.* SIAM Journal on Control and Optimization 37, 1999, pp. 731–764.

[68] J. Werner: *Numerische Mathematik I.* Vieweg-Verlag, Braunschweig-Wiesbaden, 1992.

[69] J. Werner: *Numerische Mathematik II.* Vieweg-Verlag, Braunschweig-Wiesbaden, 1992.

[70] C. Zhu, R.H. Byrd, and J. Nocedal: *Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization.* ACM Transactions on Mathematical Software 23, 1997, pp. 550–560.

Acknowledgment

Closing this thesis I would like to thank all those persons who supported me mathematically and morally during my time in Würzburg. In particular I'm obliged very much to my advisor Prof. Dr. Christian Kanzow for suggesting the interesting topic and for his patient and friendly support that led to our joint publications and consecutively to this thesis. Moreover I would like to thank Prof. Dr. Stefania Bellavia for writing the expertise and pointing out several of the large scaled test examples.