**A Global Convergence Theory
for Arbitrary Norm Trust Region Methods
for Nonlinear Equations¹**

by

Mohammed El Hallabi²

Technical Report 87-5, April 1987

¹A Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Rice University.

²This author was sponsored by AMIDEAST, Washington, DC on behalf of the Moroccan-American Commission for Educational and Cultural Exchange, Rabat, Morocco.

To my parents

In the name of Allah, the Beneficent, the Merciful.

**“And say, Oh my Lord, increase my knowledge”
Q’uran sura 20, ayat 114.**

**“Seek knowledge from the cradle to the grave.”
The Prophet Muhammad (peace be upon him).**

RICE UNIVERSITY

A GLOBAL CONVERGENCE THEORY
FOR ARBITRARY NORM TRUST REGION METHODS
FOR NONLINEAR EQUATIONS

by

MOHAMMEDI EL HALLABI

A Thesis Submitted
In Partial Fulfillment Of The
Requirements For The Degree

DOCTOR OF PHILOSOPHY

Approved, Thesis Committee:

Richard A. Tapia,
Professor of Mathematical Sciences
Chairman

John E. Dennis, Jr.,
Professor of Mathematical Sciences
Co-Chairman

R.E. Bixby,
Professor of Mathematical Sciences

R.J. DeFigueiredo,
Professor of Electrical
and Computer Engineering

Houston, Texas

May, 1987

A GLOBAL CONVERGENCE THEORY
FOR ARBITRARY NORM TRUST REGION METHODS
FOR NONLINEAR EQUATIONS

MOHAMMEDI EL HALLABI

ABSTRACT

In this research we extend the Levenberg-Marquardt algorithm for approximating zeros of the nonlinear system $F(x)=0$, where F is continuously differentiable from $\mathbb{R}^n$ to $\mathbb{R}^n$. Instead of the l_2 -norm, arbitrary norms can be used in the objective function and in the trust region constraint. The algorithm is shown to be globally convergent. This research was motivated by the recent work of Duff, Nocedal and Reid. A key point in our analysis is that the tools from nonsmooth analysis, namely locally Lipschitz analysis, allow us to establish essentially the same properties for our algorithm that have been established for the Levenberg-Marquardt algorithm using the tools from smooth optimization. In our analysis, the sequence generated by the algorithm is the couple (x_k, δ_k) where x_k is the iterate and δ_k the trust region radius. Since the successor (x_{k+1}, δ_{k+1}) of (x_k, δ_k) is not unique we model our algorithm by a point-to-set map and then apply Zangwill's theorem of convergence to our case. It is shown that our algorithm reduces locally to Newton's method.

ACKNOWLEDGEMENTS

It is a pleasure to thank my advisors, Professors Richard A. Tapia and John E. Dennis. I am grateful to Professor Tapia for his generous and frequent support, encouragement and inspiration throughout my years at Rice, and to Professor Dennis for being so generous with his time and ideas. I thank Professors Robert E. Bixby and Rui J. DeFigueiredo for judging the present work.

I would like to thank the "Direction de l'Ecole Hassania des Travaux Publics" and the "Direction de la Formation Professionnelle" of the Ministry of Equipment of the Kingdom of Morocco for allowing me to go on leave to improve my knowledge at Rice University.

I am grateful to the Moroccan-American Commission for Educational and Cultural Exchange (MACECE) for supporting my graduate work in the U.S.A., and to AMIDEAST, especially Mrs. Grace Metz and Judy Bienman, who worked out administrative problems for me so that I could concentrate only on my graduate work.

Mathematical and personal thanks to A.M. Morshedi for the fruitful discussions we shared every time he was at the Department of Mathematical Sciences.

I deeply appreciated the presence, in the Mathematical Sciences Department, of Mahmoud El Alem, in whom I could find a breeze of my culture. For always being very helpful I thank Basalat Ali Raja and Omar (Douglas) Monk.

Mathematical and personal thanks go to Daniel Woods, Maria Rosa Celis, Michael Pearlman, Hector Martinez and Christina Maciel.

I would also like to express my appreciation to Vivian Choi for typing this thesis, and to Judy Forney for the work she has put into keeping the department running and looking after the graduate program.

TABLE OF CONTENTS

Chapter 1 Introduction	1
1.1 The Problem and the Standard Newton's Method	1
1.2 Line-search Globalization Strategy	3
1.3 Trust Region Globalization Strategy	5
1.4 The Duff, Nocedal and Reid Approach	9
1.5 Other Formulations	11
1.6 The Present Work	20
Chapter 2 Background Material	24
2.1 Point-to-Set Maps	24
2.2 Convex Functions	31
2.3 Lipschitz Functions	36
Chapter 3 Existence and Equivalence Properties of the Various Derivatives of $f = \ F\ _g$	42
3.1 Case of General Norm	42
3.2 The Case Where $\ \cdot\ _g$ is the l_1 -norm	56
3.3 The Case Where $\ \cdot\ _g$ is the l_∞ -norm	59
Chapter 4 Optimality Conditions	62
4.1 Optimality Conditions	62
Chapter 5 The Algorithm and Convergence Results	72
5.1 The Trust Region Algorithm	72
5.2 Global Convergence Results	78
5.3 Condition of Convergence to a Solution of the Nonlinear System	93
5.4 Rate of Convergence	95

CHAPTER 1

Introduction

1.1. The Problem and the Standard Newton's Method

In this research, we consider the problem of solving the system of nonlinear equations

$$F(x) = 0, \quad (1.1.1)$$

where $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a continuously differentiable function. We will be concerned with the fact that the Jacobian of F at x , say $F'(x)$, may be sparse. Such systems arise, for example, in the solution of the algebraic equations produced by finite elements or finite difference discretization of boundary value problems for ordinary and partial differential equations.

Problem (1.1.1) is often solved by Newton's method. Namely, given an initial approximation x_0 to the solution x_* , successive approximations $x_1, x_2, \dots$ are generated by solving the linear system

$$F'(x_k)s = -F(x_k), \quad k = 0, 1, 2, \dots \quad (1.1.2)$$

for the Newton step s_k and letting

$$x_{k+1} = x_k + s_k \quad k = 0, 1, 2, \dots \quad (1.1.3)$$

If the initial iterate x_0 is an arbitrary point of some sufficiently small

neighborhood $N(x_*)$ of the solution x_* , $F'(x_*)$ is nonsingular, and F is sufficiently smooth, then the method gives very rapid convergence [Dennis and Schnabel (1983)]. Because x_0 is assumed to be any point of $N(x_*)$ we say that the method is locally convergent.

In the use of Newton's method, difficulties arise when the Newton step lies outside the region where the linear model

$$F(x_k) + F'(x_k)s \quad (1.1.4)$$

is a good approximation to $F(x_k + s)$. One effective remedy when this occurs is to restrict the step s to a region where the linear model (1.1.4) can be trusted. Specifically in the well-known Levenberg-Marquardt trust region approach the step s_k is the solution of the optimization problem

$$\text{minimize} \quad \frac{1}{2} \|F(x_k) + F'(x_k)s\|_2^2 \quad (1.1.5a)$$

$$\text{subject to} \quad \frac{1}{2} \|s\|_2^2 \leq \delta_k. \quad (1.1.5b)$$

For details see Chapter 6 of Dennis and Schnabel (1983).

To better understand the above approach, it should be viewed as a globalization strategy for the locally convergent Newton's method for solving the unconstrained minimization problem

$$\text{minimize}_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} \|F(x)\|_2^2. \quad (1.1.6)$$

In (1.1.6) the square of the l_2 -norm is introduced for the convenience of dealing

with a differentiable objective function.

By a globalization strategy of a locally convergent method for the smooth problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} f(x), \quad (1.1.7)$$

i.e. the function f is differentiable, we mean a modification to the algorithm that attempts to force it to converge to a local (or global) solution of problem (1.1.7), starting from any given x_0 in $\mathbb{R}^n$. Two of the most popular globalization strategies for a locally convergent method are the line-search and the model trust-region strategies.

Because the algorithms proposed in this research will be of the second kind, we will touch briefly on the line-search strategy and present the trust region strategy in more detail.

1.2. Line-search Globalization Strategy

The idea behind a line-search strategy is simple. Namely, given a descent direction d_k of f at x_k , i.e. a d_k such that

$$\nabla f(x_k)^T d_k < 0, \quad (1.2.1)$$

we take a step in that direction that is acceptable according to some rule. Specifically, at the k^{th} iteration we set $x_{k+1} = x_k + \lambda_k d_k$ for some $\lambda_k > 0$ which makes $\lambda_k d_k$ an acceptable step. Different rules may be used to define an

acceptable step. One of the first rules was proposed by Armijo (1966). Then, various rules were studied by Goldstein and Price (1967) and Wolfe (1969). In all these versions an acceptable step has to satisfy at least the so-called α -condition

$$f(x_k + \lambda_k d_k) \leq f(x_k) + \lambda_k \alpha \nabla f(x_k)^T d_k, \quad (1.2.2)$$

where α is a small fixed constant in $(0, 1)$. Additional conditions may be required. The convergence theory for such algorithms shows that fast local convergence will not be lost provided that asymptotically one chooses $\lambda_k = 1$ [see Dennis and Moré (1977)]. This fact motivated Dennis and Schnabel (1983) to consider the backtracking algorithm

Given $\alpha \in (0, \frac{1}{2})$ and $0 < l < u < 1$,

$$\lambda_k = 1,$$

$$\text{while } f(x_k + \lambda_k d_k) > f(x_k) + \alpha \lambda_k \nabla f(x_k)^T d_k$$

do

$$\lambda_k = : \rho \lambda_k \text{ for some } \rho \in [l, \mu];$$

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

For more details concerning the line-search strategy we refer the reader to Dennis and Schnabel (1983).

1.3. Trust Region Globalization Strategy

In the trust region globalization strategy we build a local model of f at x_k , say $m_k(\cdot)$, which satisfies at least the properties

$$m_k(0) = f(x_k) \quad (1.3.1)$$

$$\nabla m_k(0) = \nabla f(x_k). \quad (1.3.2)$$

Given such a model and a trust region radius δ_k , we obtain s_k as the solution of the optimization problem

$$\text{minimize } m_k(s) \quad (1.3.3a)$$

$$\text{subject to } \|s\| \leq \delta_k. \quad (1.3.3b)$$

In (1.3.3.b), $\|\cdot\|$ is usually chosen to be the l_2 -norm. We accept the step s_k and set $x_{k+1} = x_k + s_k$ if the model $m_k(\cdot)$ can be trusted, i.e. if

$$\frac{\text{actred}(k)}{\text{prered}(k)} \geq \alpha, \quad (1.3.4)$$

where α is a (small) fixed constant in $(0, 1)$ and

$$\text{actred}(k) = f(x_k + s_k) - f(x_k) \quad (1.3.5)$$

and

$$\text{prered}(k) = m_k(s_k) - f(x_k). \quad (1.3.6)$$

In most cases the local model $m_k(\cdot)$ is convex, and consequently, we have

$$\nabla m_k(0)^T s_k \leq m_k(s_k) - m_k(0), \quad (1.3.7)$$

or, using the equalities (1.3.1) and (1.3.2),

$$\nabla f(x_k)^T s_k \leq m_k(s_k) - f(x_k). \quad (1.3.8)$$

Using (1.3.5) and (1.3.6) we can rewrite (1.3.4) as

$$f(x_k + s_k) \leq f(x_k) + \alpha \left\{ m_k(s_k) - f(x_k) \right\}, \quad (1.3.9)$$

which, because of (1.3.8), can be viewed as a relaxation of the α -condition

$$f(x_k + s_k) \leq f(x_k) + \alpha \nabla f(x_k)^T s_k. \quad (1.3.10)$$

Either criterion (1.3.4) or (1.3.10) is used for the acceptance or rejection of the step s_k . Moré and Sorenson (1982) use the former; Dennis and Schnabel (1983) prefer the latter. The main issue here is that f is the merit function used to determine when one approximate solution is better than another. If the step s_k is rejected, then the trust region radius δ_k is decreased, and problem (1.3.3) is solved with this new trust region radius. For a discussion regarding the reduction factor, we refer the reader to Dennis and Schnabel (1983). When the step is accepted, the parameter δ_k is updated by comparing the values $\text{actred}(k)$ and $\text{prered}(k)$ given in (1.3.5) and (1.3.6). Namely, if

$$\frac{\text{actred}(k)}{\text{prered}(k)} < \beta, \quad (1.3.11)$$

where $\beta \in (\alpha, 1)$, then the trust region radius is decreased by choosing

$$\delta_{k+1} \in \left[\mu_1 \|s_k\|, \mu_2 \|s_k\| \right], \quad (1.3.12)$$

where μ_1 and μ_2 are constants such that $0 < \mu_1 \leq \mu_2 < 1$. If (1.3.11) does not

hold, then δ_k is increased. The following two ways of increasing the radius of the trust region are commonly used:

- (1) δ_{k+1} is chosen to be any scalar in

$$\left[\|\delta_k\|, \gamma \|\delta_k\| \right], \quad \gamma > 1; \quad (1.3.13)$$

- (2) δ_{k+1} is chosen to be any scalar in

$$\left[\delta_k, \gamma \delta_k \right], \quad \gamma > 1. \quad (1.3.14)$$

In the present work, we consider the following third possibility for increasing the trust region radius:

- (3) δ_{k+1} is chosen to be any scalar in

$$\left[\|\delta_k\|, \max(\delta_k, \gamma \|\delta_k\|) \right], \quad \gamma > 1. \quad (1.3.15)$$

We have seen that the Levenberg-Marquardt algorithm is a trust region globalization strategy for Newton's method for solving problem (1.1.6). The model of the objective function used at each iteration is

$$m_k(s) = \|F(x_k) + F'(x_k)s\|_2. \quad (1.3.16)$$

Because of the special structure of the constraint in the convex program (1.1.5), the Kuhn-Tucker conditions are both necessary and sufficient for s_k to be a solution. Consequently problem (1.1.5) is solved by

$$s(\mu) = -\{F'(x_k)^T F'(x_k) + \mu I\}^{-1} F'(x_k)^T F(x_k) \quad (1.3.17)$$

for the unique nonnegative $\mu_k \in \mathbb{R}$ such that $\|s(\mu_k)\| = \delta_k$, unless $\|s(0)\| \leq \delta_k$, in which case $s(0) = s_k^N$, i.e. the Newton step, is the solution.

In the earlier implementations of the Levenberg-Marquardt method, control is gained by working directly with the parameter μ , while in the later improved implementations, control was gained by varying the parameter δ_k and computing μ by a very efficient iterative method suggested by Hebden (1973) and refined by Moré (1977). This Levenberg-Marquardt algorithm (or the l_2 -trust-region method for nonlinear systems) has proven to be an effective algorithm with good global behavior.

The Hebden-Moré implementation of the Levenberg-Marquardt algorithm is known to be robust. However, for large problems, it has the disadvantage that the system (1.3.17) has to be solved for several values of μ at each iteration. It is not obvious how one utilizes sparsity here since multiplying a matrix by its transpose may destroy the sparsity. Therefore, replacing the problem of solving (1.1.5) by the problem (1.3.17) is not attractive for large problems.

To avoid solving (1.3.17) at each iteration, Powell (1970) devised his now famous dogleg version of the Levenberg-Marquardt trust-region algorithm. Dennis and Mei (1979) extended Powell's dogleg to what they called the double dogleg. The dogleg algorithm approximates the solution of the trust-region problem (1.1.5) by restricting $x_k + s$ to lie on the piecewise linear curve connecting x_k , the Cauchy-point (minimizer of the functional in (1.1.5a) in the

direction of steepest descent) and the Newton point (unconstrained minimizer of the functional in (1.1.5a)). The optimization problem (1.1.5) restricted to this piecewise linear curve (dogleg) can readily be solved. In most cases, the dogleg step will be a good approximation to the locally constrained optimal step, i.e. the solution of problem (1.1.5).

The dogleg and double dogleg methods have proven to be effective for many problems. However, we cannot expect them to be as robust as the Levenberg-Marquardt algorithm. In fact, Reid (1973) adopted the dogleg method to the sparse case, and reported finding examples for which the method did not converge, but the Levenberg-Marquardt method did converge.

1.4. The Duff, Nocedal and Reid Approach

Duff, Nocedal and Reid (1984) suggested replacing the square of the l_2 -norm in (1.1.6) and (1.1.5a) by the l_1 -norm and the square of the l_2 -norm in (1.1.5b) by the l_∞ -norm. They point out that in this way one can use linear programming techniques to solve the model trust-region problem (1.1.5), and therefore take advantage of any sparsity patterns in the Jacobian $F'(x_k)$. The Levenberg-Marquardt approach does not usually allow one to take advantage of sparsity. Duff, Nocedal and Reid (1984) replaced the α -condition (1.3.10) by the α -condition

$$\|F(x+s)\|_1 \leq \|F(x)\|_1 - \alpha \|F'(x)s\|_1. \quad (1.4.1)$$

One way to view (1.4.1) in the context of (1.3.10) is that since the l_1 -norm is not necessarily differentiable, Duff, Nocedal and Reid merely differentiated underneath the norm. This is equivalent to assuming that the l_1 -norm is linear, and we do know that the l_1 -norm is piecewise linear. The main drawback to (1.4.1) is demonstrated by the following inequalities (which we will establish later):

$$-\|F'(x)s\|_1 \leq f'(x,s) \leq \|F(x) + F'(x)s\|_1 - \|F(x)\|_1. \quad (1.4.2)$$

In (1.4.2), $f'(x,s)$ is the one-sided directional derivative of $f(x) = \|F(x)\|_1$ at x in the direction s . It is possible to have $|f'(x,s)|$ so small relative to $\|F'(x)s\|_1$ that $-\alpha\|F'(x)s\|_1$ is less than $f'(x,s)$. This implies, as we will see in Section 5.1, that the α -condition (1.4.1) cannot always be satisfied. The fact that Duff, Nocedal and Reid chose the wrong generalization of the α -condition (1.3.10) is a minor point, and we will presently show that it is not difficult to state a correct α -condition for their formulation. Their algorithm uses (1.3.12) and (1.3.14) to update the trust-region radius.

Duff, Nocedal and Reid do not give any convergence results. They give a very detailed description of their implementation and point out that it is very competitive with the other methods and can succeed when other approaches fail. We feel that (1.4.2) implies that the Duff, Nocedal and Reid α -condition (1.4.1)

is excessively conservative, and it may not be possible to establish global convergence of their algorithm (more will be said about this in Section 5.1). On the other hand, our formulation of the correct α -condition will allow us to obtain a global convergence result.

1.5. Other Formulations

The use of a different norm in (1.1.5) and (1.1.6) instead of the square of the l_2 -norm has been suggested and investigated by many authors. Madsen (1975) uses the l_∞ -norm in both (1.1.5) and (1.1.6) and considers the overdetermined systems where $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$ $n \leq m$. He defines a stationary point of the function f as a point x which satisfies

$$f(x) = \|F(x)\|_\infty = \min \left\{ \|F(x) + F'(x)s\|_\infty \mid s \in \mathbb{R}^n \right\}, \quad (1.5.1)$$

or equivalently

$$\psi(x_k) = 0, \quad (1.5.2)$$

where

$$\psi(x_k) = \|F(x_k)\|_\infty - \min \{ \|F(x_k) + F'(x_k)s\|_\infty \mid s \in \mathbb{R}^n \}, \quad (1.5.3)$$

or simply because the function $m_k(s)$ is convex

$$\psi(x_k) = \|F(x_k)\|_\infty - \min \{ \|F(x_k) + F'(x_k)s\|_\infty / \|s\|_\infty \leq 1 \}. \quad (1.5.4)$$

His algorithm is a trust-region type method where the model $m_k(s)$ is

$\|F(x_k) + F'(x_k)s\|_\infty$. Madsen shows that if F is twice continuously differentiable and the sequence generated by his algorithm, say $\{x_k, k \in \mathbb{N}\}$, is bounded, then

$$\lim_{k \rightarrow +\infty} d(x_k, S) = 0 \quad (1.5.5)$$

where S is the set of stationary points of the function $f(x) = \|F(x)\|_\infty$, and where

$$d(x_k, S) = \min \{ d(x_k, y) \mid y \in S \}. \quad (1.5.6)$$

The limit (1.5.5) does not imply that any accumulation point of the sequence $\{x_k, k \in \mathbb{N}\}$ is in S . It only implies that there exists a subsequence $\{x_{k_j}, j \in \mathbb{N}\}$ whose accumulation points are in S , or equivalently, that

$$\liminf_{k \rightarrow \infty} \psi(x_k) = 0, \quad (1.5.7)$$

as we will establish in the following lemma.

LEMMA 1.5.1 *If the sequence $\{x_k, k \in \mathbb{N}\}$ generated by the algorithm of Madsen (1975) is bounded, then the limit (1.5.5) is equivalent to the limit inferior in (1.5.7).*

Proof. Suppose that $\liminf_{k \rightarrow +\infty} \psi(x_k) = 0$. Then there exists a subsequence $\{x_k, k \in \mathbb{N}'\}$ such that $\{\psi(x_k), k \in \mathbb{N}'\}$ converges to 0 in $\mathbb{R}$. The sequence $\{x_k, k \in \mathbb{N}\}$ is bounded and the function ψ is continuous (see Powell (1983));

therefore any accumulation point of $\{x_k, k \in N'\}$, say x_* , verifies

$$\psi(x_*) = 0$$

and consequently it is a stationary point of $f = \|F\|_\infty$.

Now suppose that $\lim_{k \rightarrow +\infty} d(x_k, S) = 0$. Then there exists a subsequence $\{x_{k_j}, j \in \mathbb{N}\}$ and a sequence $\{y_j \in S \mid j \in \mathbb{N}\}$ such that

$$d(x_{k_j}, y_j) < \frac{1}{j+1} \quad \text{for all } j \in \mathbb{N}. \quad (1.5.8)$$

The subsequence $\{x_{k_j}, j \in \mathbb{N}\}$ is bounded; consequently, the sequence $\{y_j \in S, j \in \mathbb{N}\}$ is also bounded. If x and $y \in S$ are accumulation points of $\{x_{k_j}, j \in \mathbb{N}\}$ and $\{y_j \in S, j \in \mathbb{N}\}$ respectively then, by (1.5.8), we have that $x = y$, and consequently x is a stationary point. •

Madsen uses the criterion (1.3.4) to accept or reject the step. But he uses different criteria to update the trust region radius. The trust-region radius is decreased if (1.3.11) holds. It is increased whenever the inequality

$$\|F(x_k + s_k) - (F(x_k) + F'(x_k)s_k)\|_\infty \leq \mu [f(x_k) - f(x_k + s_k)] \quad (1.5.9)$$

is satisfied, where μ is a fixed constant in $(0, 1)$. The radius δ_k is set equal to $\|s_k\|$ if neither (1.3.11) nor (1.5.9) holds. No relationship is mentioned between the constant β in (1.3.11) and μ in (1.5.9). This lack of dependence

may lead to a situation where one cannot decide whether to increase or to decrease the trust-region radius. For example, suppose that the test for increasing δ_k , i.e. (1.5.9), is satisfied. Then we have

$$f(x_k + s_k) - \|F(x_k) + F'(x_k)s_k\|_\infty \leq \mu [f(x_k) - f(x_k + s_k)]$$

or

$$- [f(x_k) - f(x_k + s_k)] + f(x_k) - \|F(x_k) + F'(x_k)s_k\|_\infty \leq \mu [f(x_k) - f(x_k + s_k)],$$

and

$$(1 + \mu)^{-1} [f(x_k) - \|F(x_k) + F'(x_k)s_k\|_\infty] \leq [f(x_k) - f(x_k + s_k)], \quad (1.5.10)$$

which does not contradict the test for decreasing the trust-region radius, i.e. we may also have

$$f(x_k) - f(x_k + s_k) \leq \beta [f(x_k) - \|F(x_k) + F'(x_k)s_k\|_\infty]$$

as long as $(1 + \mu)^{-1} \leq \beta$. Therefore in the algorithm of Madsen (1975) the constants μ and β should satisfy a restriction of the form

$$(1 + \mu)^{-1} > \beta. \quad (1.5.11)$$

In fact in his numerical experiments Madsen used $\mu = \beta = 0.25$, which satisfies (1.5.11).

Powell (1983) also considered a trust-region algorithm for solving problem (1.1.6), where $F: \mathbb{R}^n \rightarrow \mathbb{R}^m$, $n \leq m$ (overdetermined system of nonlinear

equations) is continuously differentiable and the square of the l_2 -norm was replaced by any continuous convex function h . In order to provide theory applicable to the often used cases where h is a norm or a power of a norm on $\mathbb{R}^m$, he assumed that h is bounded below and $h(F) \rightarrow +\infty$ whenever $\|F\| \rightarrow +\infty$. He also assumed that the function $f(x) = h(F(x))$ is restricted to a compact subset $X \subset \mathbb{R}^n$. Powell defined a stationary point of the function $f(x) = h(F(x))$ to be a point x such that $\psi(x) = 0$ where ψ is defined by

$$\psi(x) = h(F(x)) - \min \{h(F(x) + F'(x)s) \mid \|s\| \leq 1\} \quad (1.5.12)$$

(more will be said about this definition in Chapter 5). Observe that this is Madsen's definition (1.5.3) of a stationary point where $\|\cdot\|_\infty$ is replaced by h . Powell used the following local model:

$$m_k(s) = h(F(x_k) + F'(x_k)s) + \frac{1}{2} s^T B_k s, \quad (1.5.13)$$

where $\{B_k, k \in \mathbb{N}\}$ is a bounded sequence of symmetric matrices. He proves that if the sequence generated by his algorithm, say $\{x_k, k \in \mathbb{N}\}$, satisfies the property that for any μ there exists $\epsilon > 0$ such that whenever $\psi(x_k) \geq \mu$ the inequality

$$m_k(s_k) \leq h(F(x_k)) - \epsilon \min[1, \delta_k] \quad (1.5.14)$$

holds, then $\psi(x_k) \rightarrow 0$ as $k \rightarrow +\infty$. To show that this property is a practical one Powell (1983) points out that the solution to (1.1.5) satisfies this property.

Suppose that we choose h as one-half of the square of the l_2 -norm and that the overdetermined system

$$F(x) = 0 \quad (1.5.15)$$

or equivalently

$$\frac{1}{2} \|F(x)\|_2^2 = 0 \quad (1.5.16)$$

has a solution. When applied to the unconstrained problem of minimizing $f(x) = \frac{1}{2} \|F(x)\|_2^2$, Newton's method is

$$x_{k+1} = x_k - [F'(x_k)^T F'(x_k) + S(x_k)]^{-1} F'(x_k)^T F(x_k), \quad (1.5.17)$$

where $S(x_k) = \sum_{i=1}^m F_i(x_k) \nabla^2 F_i(x_k)$. We know that this method is locally q -quadratically convergent under mild assumptions. For the zero residual problems (1.5.15), instead of (1.5.17) one usually considers the modification where $S(x_k)$ is removed. This gives us the Gauss-Newton method, i.e.

$$x_{k+1} = x_k - [F'(x_k)^T F'(x_k)]^{-1} F'(x_k)^T F(x_k), \quad (1.5.18)$$

which is still q -quadratically convergent (for details see Chapter 10 of Dennis and Schnabel (1983)).

In a trust-region globalization strategy for solving (1.5.15), one expects the trust-region constraint not to be binding near a solution, and consequently the problem (1.3.3) to be unconstrained. Therefore, any solution of problem (1.3.3)

would satisfy

$$\nabla_s m_k(s_k) = 0. \quad (1.5.19)$$

In Powell's model trust-region algorithm (1983) we would have

$$F'(x_k)^T(F'(x_k)s_k + F(x_k)) + B_k s_k = 0. \quad (1.5.20)$$

From (1.5.18), s_k cannot be the Gauss-Newton step unless $B_k s_k^{GN} = 0$, and it is not obvious when this will be satisfied with a nonzero B_k . We can rewrite (1.5.20) as

$$[F'(x_k)^T F'(x_k) + B_k] s_k = -F'(x_k)^T F(x_k), \quad (1.5.21)$$

and conclude from (1.5.17) that s_k will be the Newton step if B_k is equal to $S(x_k)$, a quantity which we wish not to calculate.

Powell's (1983) trust-region algorithm with a nonzero B_k is probably not suitable for solving zero (or small) residual nonlinear systems, and it should be considered an algorithm for nonzero residual problems, which are, after all, minimization problems rather than nonlinear systems. Indeed, it is for the case of nonzero residual problems that Powell and Yuan (1983) derive conditions on $B_k, k \in \mathbb{N}$, which imply superlinear convergence in the case where h is either the l_1 -norm or the l_∞ -norm.

In a similar approach to Powell (1983a, b), Yuan (1983) replaced condition (1.5.14) by a very simple condition, namely $h(F(x_{k+1})) < h(F(x_k))$. He obtained a global convergence result of Madsen's type, i.e.

$$\liminf_{k \rightarrow +\infty} \psi(x_k) = 0$$

where ψ is defined by (1.5.12).

The matrices $B_k, k \in \mathbb{N}$ have their origin in the approach suggested by Han (1978, 1981) for solving the min-max problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \gamma(x) = \max(f_1(x), \dots, f_m(x)), \quad (1.5.22)$$

where $f_i, i = 1, \dots, m$ are twice continuously differentiable. Problem (1.5.22) is equivalent to

$$\begin{aligned} &\text{minimize} \quad \mu \\ &\text{subject to} \quad f_i(x) \leq \mu \quad i = 1, \dots, m. \end{aligned}$$

Han's approach contains a line-search globalization strategy where the search direction d_k is the solution of

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2} s^T B_k s + \mu \\ &\text{subject to} \quad f_i(x_k) + \nabla f_i(x_k)^T s \leq \mu \quad i = 1, \dots, m, \end{aligned} \quad (1.5.23)$$

where $\{B_k, k \in \mathbb{N}\}$ is a sequence of matrices such that

$$M_1 \|s\|_2^2 \leq s^T B_k s \leq M_2 \|s\|_2^2 \quad \forall s \in \mathbb{R}^n, \quad (1.5.24)$$

where $0 < M_1 \leq M_2$. When this approach is applied to the problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \|F(x)\|_\infty,$$

problem (1.5.23) becomes

$$\begin{aligned} & \text{minimize} \quad \frac{1}{2} s^T B_k s + \mu \\ & \text{subject to} \quad |f_i(x_k) + \nabla f_i(x_k)^T s| \leq \mu \quad i=1, \dots, m, \end{aligned}$$

or

$$\text{minimize}_{s \in \mathbb{R}^n} \quad \frac{1}{2} s^T B_k s + \|F(x_k) + F'(x_k)s\|_\infty. \quad (1.5.25)$$

Adding a trust-region constraint in Problem (1.5.25), we relax the positive definiteness condition (1.5.24) on the matrices B_k and obtain Powell's approach with $h(F) = \|F\|_\infty$.

The solution of problem (1.5.23), say s_k , satisfies

$$\gamma'(x_k, s_k) \leq -s_k^T B_k s_k, \quad (1.5.26)$$

where $\gamma'(x_k, s_k)$ is the one-sided directional derivative of γ at x_k in the direction s_k . Han uses a backtracking line-search to obtain the subsequent iterate $x_{k+1} = x_k + \lambda_k s_k$ and requires that the α -condition

$$\gamma(x_k + \lambda_k s_k) \leq \gamma(x_k) + \lambda_k \alpha (-s_k^T B_k s_k)$$

holds. We remark that, in the context of (1.3.10), i.e. $f(x+s) \leq f(x) + \lambda_k \alpha \nabla f(x)^T s$, Han used an approximate directional derivative that majorizes the one-sided directional derivative of γ at x_k , unlike Duff, Nocedal and Reid (1984) who used a quantity which minorized the one-sided directional derivative (see (1.4.1) and (1.4.2)). Han gives a global convergence result.

Solving problem (1.1.6) when the square of the l_2 -norm is replaced by a different norm has generated considerable interest, mainly for overdetermined systems. For overdetermined systems ($n < m$) Osborne and Watson (1969), Anderson and Osborne (1977), Osborne and Watson (1978) and Watson (1979) present algorithms which use line-search globalization strategies. A Newton-type algorithm for overdetermined system was considered in Dennis, Gay and Welch (1981a, b). Completely different approaches were considered by El Attar, Vidya and Dutta (1979) and Murray and Overton (1981). In the former, the authors replaced the square of the l_2 -norm by the l_1 -norm and then solved a sequence of parameterized differentiable unconstrained problems. Theoretically, the sequence of parameterized solutions converges to the solution of problem (1.1.6) as the parameter goes to zero. Numerically, this makes the unconstrained problems ill-conditioned. In the later paper, the authors used the l_1 -norm and proposed a method based on projected Lagrangian methods for constrained optimization which requires successively solving quadratic programs in the same number of variables n .

1.6. The Present Work

In the present work, we propose a globally convergent algorithm for approximating zeros of the square nonlinear system $F(x)=0$, where F is continuously differentiable. Our approach consists of approximating the solution

of problem (1.1.6) where the square of the l_2 -norm can be replaced by any norm on $\mathbb{R}^n$.

Our general class of algorithms has been motivated by the algorithm proposed by Duff, Nocedal and Reid (1984). At each iteration we solve the following problem

$$\begin{aligned} \text{minimize} \quad & m_k(s) = \|F(x_k) + F'(x_k)s\|_a \\ \text{subject to} \quad & \|s\|_b \leq \delta_k, \end{aligned} \quad (1.6.1)$$

where $\|\cdot\|_a$ and $\|\cdot\|_b$ are two arbitrary but fixed norms on $\mathbb{R}^n$. Although our intention is that both $\|\cdot\|_a$ and $\|\cdot\|_b$ should be either the l_1 -norm or the l_∞ -norm so that the subproblem can be solved as a linear program, we feel that the formulation of the algorithm in terms of arbitrary norms has not only simplified the presentation and analysis but has given considerable credibility to the philosophy and structure of the algorithm.

In our analysis, we consider the sequence $\{(x_k, \delta_k), k \in \mathbb{N}\}$ generated by the algorithm. Since the iterate (x_{k+1}, δ_{k+1}) is not unique, i.e. the subproblem may have many solutions and the choice of δ_{k+1} is far from unique, we model our algorithm with a point-to-set map. Therefore in the first part of Chapter 2, we review some basic results concerning point-to-set maps and the convergence theorem of Zangwill (1969) for such models. The objective function of the subproblem (1.6.1) is convex. In order to exploit this fact fully, in part 2 of Chapter 2 we review results from the theory of convex analysis. For this review

we will appeal to Rockafellar (1970) for results regarding the existence and other properties of the one-sided directional derivative of the convex function $m_k(\cdot)$ and its subdifferential at zero. To end the second chapter we review results concerning locally Lipschitz functions, mainly the existence of the generalized directional derivative and the generalized gradient at x of a function that is Lipschitz near x , i.e. Lipschitz on some neighborhood of x .

In Chapter 3 we apply the results reviewed in parts 2 and 3 of Chapter 2 to the function $m_k(\cdot)$ and $f(x) = \|F(x)\|_a$. We show that f is regular, i.e. the one-sided directional derivative at x is equal to the generalized directional derivative. Also, we show that the one-sided directional derivative and the subdifferential at zero of the convex model $m_k(\cdot)$ at x_k are respectively equal to the generalized directional derivative and the generalized gradient of f at x_k .

In the fourth chapter we give the well-known necessary conditions for x_* to solve problem

$$\text{minimize}_{x \in \mathbb{R}^n} f(x) = \|F(x)\|_a. \quad (1.6.2)$$

that is, zero is in the generalized gradient of f at x_* , i.e. $0 \in \partial f(x_*)$. The results from the second chapter allow us to prove that this condition is equivalent to saying that the one-sided directional derivative of f at x_* is nonnegative in all directions in $\mathbb{R}^n$, i.e.

$$f'(x_*, s) \geq 0 \quad \forall s \in \mathbb{R}^n. \quad (1.6.3)$$

Consequently, we will define a stationary point x_* as a point for which (1.6.3) holds, and then show that although our definition seems more natural, it is equivalent to the one used by Madsen (1975), and by Powell (1983) and Yuan (1983). We also show that a stationary point x_* , at which $F'(x_*)$ is nonsingular, is actually a solution of the nonlinear system of equation $F(x)=0$. We establish the inequalities

$$- \|F'(x)s\|_\alpha \leq f'(x,s) \leq \|F(x) + F'(x)s\|_\alpha - \|F(x)\|_\alpha,$$

which demonstrate the conservatism of the α -condition used by Duff, Nocedal and Reid (1984).

In Chapter 5, we define our model trust-region algorithm. We model it with a point-to-set map and show that the theorem of Zangwill (1969) reviewed in part one of the second chapter applies to our algorithm. Consequently, our algorithm is globally convergent to a stationary point of $f = \|F\|_\alpha$. We end the chapter by proving that if the sequence $\{x_k, k \in \mathbb{N}\}$ converges to a stationary point x_* such that $F'(x_*)$ is nonsingular and F' is Lipschitz near x_* , then our algorithm reduces to Newton's method for large k and therefore is q -quadratically convergent to a zero of F .

CHAPTER 2

Background Material

This chapter consists of three parts. In the first part we review some basic results from the theory of point-to-set maps including the Zangwill convergence theorem. The second part deals with results from convex analysis and includes properties of the one-sided directional derivative of a proper convex function. Finally, in the third part, we review some results related to locally Lipschitz functions, including properties of the generalized directional derivative.

2.1. Point-to-Set Maps

DEFINITION 2.1.1 *Let A be a map defined from $X \subset \mathbb{R}^n$ into 2^Y where $Y \subset \mathbb{R}^m$. We say that A is a point-to-set map (multivalued function or multifunction) if $A(x)$ is not a singleton set for at least one x in X . If $A(x)$ is a singleton for every x in X , then we call A a point-to-point map or a point-to-point function. Even though (in a set theoretic sense) the range spaces are not the same, no distinction will be made between the point-to-point function f and f considered as a function in the usual sense.*

In numerical optimization, most algorithms are iterative in nature. Namely, given a starting point z_0 , a sequence $\{z_k, k \in \mathbb{N}\}$ is generated according to prescribed rules. Each iterate is chosen from a well-defined subset of $X \subset \mathbb{R}^n$ that is rarely a singleton (for example in a model trust-region globalization strategy, problem (1.1.5) may have more than one solution and the iterate is defined to be $x_{k+1} = x_k + s_k$, where s_k is any solution of problem (1.1.5)). Therefore the iterate z_{k+1} is defined via a point-to-set map which characterizes algorithm.

Model Algorithm

Given a point $z_0 \in \mathbb{R}^n$, a sequence of points $\{z_k, k \in \mathbb{N}\}$ is generated recursively according to the defining relation

$$z_{k+1} \in A(z_k)$$

where A is a point-to-set map and any point in the set $A(z_k)$ is an acceptable successor point of z_k .

Notice that the model does not specify the type of problem we are solving. We refer to the set of solutions as the solution set P . For a specific application, A and P must be defined.

We now recall a well-known definition.

DEFINITION 2.1.2 Consider a function f defined on $X \subset \mathbb{R}^n$ into $\mathbb{R}^p$. We say that f is continuous at $x \in X$ if given $W(f(x))$, an open neighborhood of $f(x)$, there exists $V(x)$, an open neighborhood of x , such that

$$x' \in V(x) \cap X \Rightarrow f(x') \in W(f(x)). \quad (2.1.1)$$

holds. If f is continuous at every x in X , then it is said to be continuous on X .

Property (2.1.1) may be written in two forms:

- (1) For any open set W in $\mathbb{R}^p$ such that $\{f(x)\} \subset W$ there exists an open neighborhood of $x, V(x)$, such that $x' \in X \cap V(x)$ implies that $\{f(x')\} \subset W$, or
- (2) For any open set W in $\mathbb{R}^p$ such that $W \cap \{f(x)\} \neq \emptyset$, there exists an open neighborhood $V(x)$ of x such that $x' \in X \cap V(x)$ implies $\{f(x')\} \cap W \neq \emptyset$.

Each of these two forms motivates one of the following generalizations of continuity to point-to-set maps.

Let $A: X \subset \mathbb{R}^n \rightarrow 2^{\mathbb{R}^p}$ be a point-to-set map. Also let x be a point in X .

DEFINITION 2.1.3 (upper semi-continuity). The point-to-set map A is said to be upper semi-continuous (u.s.c.) at x if for any open set W in $\mathbb{R}^p$ containing $A(x)$, there exists an open neighborhood V of x such that $A(x')$ is contained in W whenever x' is in V . If A is u.s.c. at every x in X , then it is said to be u.s.c. on X .

The last definition of the upper semi-continuity should not be confused with the common definition of the upper semi-continuity of a point-to-point function; that is, the function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is u.s.c. at x (equivalently, $-f$ is l.s.c. at x) if and only if for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\|x' - x\| < \delta \Rightarrow f(x') \leq f(x) + \epsilon.$$

DEFINITION 2.1.4 (lower-semi-continuity). *The point-to-set map A is said to be lower-semi-continuous (l.s.c.) at x if for any open set W in $\mathbb{R}^p$ intersecting $A(x)$, there exists an open neighborhood V of x such that W intersects $A(x')$ whenever x' is in V . If A is l.s.c. at every x in X , then it is said to be l.s.c. on X .*

We can consider two additional properties of the point-to-set map A by looking at equivalent notions of continuity of a point-to-point function f . Specifically, recall that Definition 2.1.2 is equivalent to the statement that f is continuous at x if, given a sequence $\{x_k, k \in \mathbb{N}\}$ which converges to x , then the sequence $\{f(x_k), k \in \mathbb{N}\}$ converges to $f(x)$.

DEFINITION 2.1.5 (closed point-to-set map). *The point-to-set map A is said to be closed at x if $\{x_k, k \in \mathbb{N}\}$ converges to x and $\{y_k \in A(x_k), k \in \mathbb{N}\}$ converges to y implies that $y \in A(x)$. If A is closed at every x in $X \subset \mathbb{R}^m$, then it is said to be closed on X .*

As we shall see, the property of being closed will be very important for the convergence to a solution of the model algorithm. In parts of the literature, a closed point-to-set map sometimes is called an upper semi-continuous map. Therefore it is important to specify which definition is being used in a particular application.

DEFINITION 2.1.6 (open point-to-set map). *The point-to-set A is said to be open at x , if $\{x_k, k \in \mathbb{N}\}$ converges to x , implies that for each y in $A(x)$ there exist $k_0 \in \mathbb{N}$ and $y_k \in A(x_k)$ for each $k \geq k_0$ such that $\{y_k, k \geq k_0\}$ converges to y . If A is open at every x in X , then it is said to be open on X .*

DEFINITION 2.1.7 (continuous point-to-set map). *If A is both closed and open at x , then A is said to be continuous at x . It is said to be continuous on X if it is continuous at every x in X .*

In the following three propositions we will review some special properties of point-to-set maps.

PROPOSITION 2.1.8 (composition of a point-to-set map and a function). *Let $X \subset \mathbb{R}^n, Y \subset \mathbb{R}^m, Z \subset \mathbb{R}^p$ and $x \in X$. Also let $f: X \rightarrow Y$ be a continuous point-to-point function and $A_1: Y \rightarrow 2^Z$ be a point-to-set map. Consider the point-to-set map $A: X \rightarrow 2^Z$ defined by $A(x) = A_1(f(x))$. If A_1 is closed (open) at x , then A is closed (open) at x .*

Proof. For a proof of this proposition see Proposition 5.6.4 of Huard (1972a) or Corollary 4.2.1 of Zangwill (1969). •

Our second proposition deals with point-to-set maps defined by inequalities.

PROPOSITION 2.1.9. Let X and Y be subsets of $\mathbb{R}^m$ and $\mathbb{R}^p$ respectively. Also let $f: X \times Y \rightarrow \mathbb{R}$ be a point-to-point function and finally let $A_1: D \subset X \rightarrow 2^Y$ and $A_2: D \subset X \rightarrow 2^Y$ be the point-to-set maps defined by

$$A_1(x) = \{y \in Y \mid f(x, y) \geq 0\}$$

$$A_2(x) = \{y \in Y \mid f(x, y) > 0\}$$

where D is the subset of X consisting of points x for which $A_1(x) \neq \emptyset$ and $A_2(x) \neq \emptyset$. The following implications hold:

- (i) If Y is closed in $\mathbb{R}^p$ and f is u.s.c., then A_1 is closed on D ;
- (ii) if f is l.s.c., then A_2 is open on D ;
- (iii) if f is l.s.c. at $x \in D$ and

$$y \in A_1(x) \implies y \in \overline{A_2(x)},$$

then A_1 is open at x .

Proof. This proposition can be found as Proposition 5.8.6 of Huard (1972a). •

Problem (1.1.5) may be viewed as a parameterized optimization problem: the parameter is (x_k, δ_k) and we solve (1.1.5) for s_k . For this reason we wish to

include the following proposition from parametric optimization.

PROPOSITION 2.1.9. Let Y and Z be subsets of $\mathbb{R}^m$ and $\mathbb{R}^p$ respectively. Also let $A: Y \rightarrow 2^Z$ be a point-to-set map, $f: Z \times Y \rightarrow \mathbb{R}$ be a point-to-point function and consider the following point-to-set map

$$\begin{aligned} \psi(y) &= \left\{ z \in A(y) \mid f(z, y) \leq f(z', y), \forall z' \in A(y) \right\} \\ &= \operatorname{argmin} \left\{ f(z, y) \mid z \in A(y) \right\}. \end{aligned}$$

Finally, let y be any point in Y . If A is continuous at y and if f is continuous on $A(y) \times \{y\}$, then ψ is closed at y .

Proof. This proposition can be found as Proposition 6.10.11 of Huard (1972a) or as Theorem 8 of Hogan (1973). •

Our motivation for using point-to-set maps to model our algorithm stems from the following theorem due to Zangwill (1969). We present the theorem as stated in Huard (1979).

THEOREM 2.1.11. Let E , a compact set of $\mathbb{R}^m$, $P \subset E$ (a solution set), $A: E \rightarrow 2^E$, a point-to-set map and $h: E \rightarrow \mathbb{R}$ a continuous function be such that for any $z \in E$ and $z \notin P$ we have

- (i) $A(z) \neq \emptyset$
- (ii) $h(z') < h(z)$ for any z' in $A(z)$
- (iii) A is closed at z .

Suppose that a sequence $\{z_k, k \in \mathbb{N}\}$ has been obtained by the following recursion relation: let z_0 be any point in E , if $z_k \notin P$ then $z_{k+1} \in A(z_k)$, otherwise $z_{k+1} = z_k$. Then any accumulation point z_* of $\{z_k, k \in \mathbb{N}\}$ is contained in P .

Proof. This theorem can be found as Convergence Theorem A of Zangwill (1969) or as a consequence of Corollary 3 and Remark 6 of Huard (1975). •

More details regarding point-to-set maps can be found in Berge (1966), Delahaye and Denel (1978), Denel (1977), Huard (1975), Meyer (1970) and Hogan (1973).

2.2. Convex Functions

We recall some basic definitions and properties from convex analysis that will be used later on. The main reference used here is Rockafellar (1970).

DEFINITION 2.2.1 (epigraph of a real function). Let C be a subset of $\mathbb{R}^n$ and let h be a real valued function defined on C . The epigraph of h is the set

$$\text{epi}(h) = \{(x, r) \in C \times \mathbb{R} \mid h(x) \leq r\}. \quad (1.2.1)$$

DEFINITION 2.2.2 (convex function and convex set). Let C be a subset of $\mathbb{R}^n$ and let h be a real function defined on C . Then

- (i) The subset of C is said to be convex if $\lambda x + (1-\lambda)y \in C$ for all λ in $[0, 1]$ and for all x, y in C .
- (ii) The function h is said to be convex on C if

$$h(\lambda x + (1-\lambda)y) \leq \lambda h(x) + (1-\lambda)h(y) \quad (2.2.3)$$

for all λ in $[0, 1]$ and for all x, y in C .

DEFINITION 2.2.3 (support function of a convex set). Let C be a convex set in $\mathbb{R}^n$. The support function $S(\cdot \mid C)$ is defined by

$$S(x \mid C) = \sup \{x^T y \mid y \in C\}. \quad (2.2.4)$$

DEFINITION 2.2.4 (effective domain of a convex function). Let h be a convex function defined on $\mathbb{R}^n$. The effective domain of h , denoted $\text{dom}(h)$, is defined to be the subset of $\mathbb{R}^n$ consisting of all x for which there exists a scalar r such that $(x, r) \in \text{epi}(h)$.

DEFINITION 2.2.5 (relative interior of a convex set). Let C be a convex subset of $\mathbb{R}^n$. Also let $\text{aff}(C)$ denote the smallest affine subspace of $\mathbb{R}^n$ containing C . Then the relative interior of C , denoted $r_i(C)$, is the subset of $\text{aff}(C)$ consisting of the elements x such that there exists $\epsilon > 0$ satisfying

$(x + \epsilon B) \cap \text{aff}(C) \subset C$, where B is the unit ball of $\mathbb{R}^n$.

DEFINITION 2.2.6 (proper convex function). A convex function h on $\mathbb{R}^n$ is said to be proper if

- (i) $\text{epi}(h)$ is not empty,
- (ii) there exists at least one x in $\mathbb{R}^n$ such that $h(x) < +\infty$, and
- (iii) $h(x) > -\infty$ for all x in $\mathbb{R}^n$.

For a convex function h defined on $\mathbb{R}^n$ and differentiable at x , the gradient $\nabla h(x)$ satisfies the inequality

$$h(y) - h(x) \geq \nabla h(x)^T (y - x), \quad (2.2.5)$$

for all y in $\mathbb{R}^n$. For a proof of this result we refer the reader to page 84 of Ortega and Rheinboldt (1970). When h is not differentiable, the gradient is no longer defined; however we may be able to obtain vectors in $\mathbb{R}^n$ which play the same role as the gradient in inequality (2.2.5).

DEFINITION 2.2.7 (the subdifferential of a convex function). Let h be a convex function defined on $\mathbb{R}^n$, and let x be a point in $\mathbb{R}^n$ such that $h(x)$ is finite. Then a vector g in $\mathbb{R}^n$ is said to be a subgradient of h at x if

$$h(y) - h(x) \geq g^T (y - x) \quad (2.2.6)$$

for all y in $\mathbb{R}^n$. The set of all subgradients of h at x is called the subdifferential

of h at x and is denoted by $\partial h(x)$.

It is well known that a convex function h has a one-sided directional derivative at any point x such that $h(x)$ is finite. This fact is given by the following theorems from Rockafellar.

THEOREM 2.2.8. Let h be a convex function, and let x be a point in $\mathbb{R}^n$ such that $h(x)$ is finite. Then for each $d \in \mathbb{R}^n$, the limit

$$\lim_{t \downarrow 0} \frac{h(x + d) - h(x)}{t} \quad (2.2.7)$$

exists and is called the one-sided directional derivative of h at x in the direction d and is denoted by $h'(x; d)$.

When h is differentiable at x , its subdifferential consists of the single element $\nabla h(x)$ and the one-sided directional derivative at x in the direction d is given by $\nabla h(x)^T d$. If h is not differentiable, then the relationship between the one-sided directional derivative of h at x , in the direction d and its subdifferential at x , $\partial h(x)$, is given by the following theorem.

THEOREM 2.2.9. Let h be a proper convex function on $\mathbb{R}^n$. Then for any x in $r_i(\text{dom}(h))$, the subdifferential of h at x , $\partial h(x)$, is a nonempty, compact and convex set. Furthermore, we have

$$h'(x; d) = \max \{g^T d \mid g \in \partial h(x)\}, \quad (1.2.9)$$

or equivalently $h'(x; d)$ is the support function of $\partial h(x)$.

The previous theorem leads us to the following useful result.

THEOREM 2.2.10. *Let h be a proper convex function on $\mathbb{R}^n$. Then for any x in $r_i(\text{dom}(h))$ we have*

$$h'(x; d) \leq h(x + d) - h(x) \quad (2.2.10)$$

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

Proof. The proof follows directly from Definition 2.2.7 and Theorem 2.2.9. •

THEOREM 2.2.11. *Let h be a proper convex function on $\mathbb{R}^n$. Then for any x in $r_i(\text{dom}(h))$,*

$$\partial h(x) = \left\{ g \in \mathbb{R}^n \mid g^T d \leq h'(x; d) \quad \forall \quad d \in \mathbb{R}^n \right\}. \quad (2.2.11)$$

Proof. Theorem 2.2.9 implies that the set defined by the right hand side of (2.2.11) contains the subdifferential of h at x , i.e. $\partial h(x)$. On the other hand let g in $\mathbb{R}^n$ be such that

$$g^T d \leq h'(x; d) \quad \forall \quad d \in \mathbb{R}^n. \quad (2.2.12)$$

The inequalities (2.2.10) and (2.2.12) give

$$g^T d \leq h(x + d) - h(x), \quad \forall \quad d \in \mathbb{R}^n$$

which implies that g is in $\partial h(x)$ by (2.2.7). •

2.3. Lipschitz Functions

Let X be an open set in $\mathbb{R}^n$. A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be Lipschitz on X if there exists a positive constant K such that

$$|f(y) - f(y')| \leq K \|y - y'\|,$$

for all y and y' in X , where $\| \cdot \|$ is a given norm on $\mathbb{R}^n$.

DEFINITION 2.3.1 [Clarke (1983), p. 25]. *Let x be any point in $\mathbb{R}^n$ and f a real function defined on $\mathbb{R}^n$. If f is Lipschitz on some neighborhood $N(x)$ of x , then f is said to be Lipschitz near x .*

A real function f that is Lipschitz near some x in $\mathbb{R}^n$ may not have a one-sided directional derivative at x ; however we do have the existence of a more general notion.

DEFINITION 2.3.2 [Clarke (1983), p. 25]. *Let x be any point in $\mathbb{R}^n$, and let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a function that is Lipschitz near x . The generalized directional derivative of f at x in the direction d in $\mathbb{R}^n$ is defined to be*

$$\limsup_{\substack{y \rightarrow x \\ t \downarrow 0}} \frac{1}{t} [f(y + td) - f(y)]$$

and is denoted by $f^\circ(x; d)$.

The utility of the generalized directional derivative f° stems from the properties given in the following proposition and theorem.

PROPOSITION 2.3.3 [Clarke (1983)]. *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be Lipschitz near some point x in $\mathbb{R}^n$. Then*

- (i) *the function $d \mapsto f^\circ(x; d)$ defined on $\mathbb{R}^n$ is finite, positively homogeneous, subadditive and satisfies*

$$|f^\circ(x; d)| \leq K \|d\|$$

for all d in $\mathbb{R}^n$, where K is the local Lipschitz constant for f and

- (ii) $f^\circ(x, -d) = (-f)^\circ(x; d)$.

Proof. The proof of this proposition can be found in Clarke (1983) in pp. 25-26. •

One of the most useful properties of the functional f° on $\mathbb{R}^n \times \mathbb{R}^n$ is given in the following theorem.

THEOREM 2.3.4 [Clarke (1983) pp. 25-26]. *Suppose that $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is Lipschitz near x for all x in $\mathbb{R}^n$. Then the generalized directional derivative*

functional $f^\circ: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ defined by $f^\circ(x; d)$ is upper semi-continuous.

Proof. The proof of this theorem can be found in pages 25-26 of Clarke (1983). •

DEFINITION 2.3.5 [Clarke (1983) page 27]. *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be Lipschitz near some x in $\mathbb{R}^n$. Then the generalized gradient of f at x is the set*

$$\partial f(x) = \left\{ g \in \mathbb{R}^n \mid g^T d \leq f^\circ(x; d), \quad \forall d \in \mathbb{R}^n \right\}.$$

Theorem 2.2.9 states that the one-sided directional derivative of a proper convex f at x in $r_i(\text{dom}(f))$ is indeed the support function of the subdifferential of f at x . In the following we give a similar property for the generalized directional derivative and the generalized gradient of a function f that is Lipschitz near some x in $\mathbb{R}^n$.

THEOREM 2.3.6. *Let x be a point in $\mathbb{R}^n$, and let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be Lipschitz near x . Then*

- (i) *the generalized gradient $\partial f(x)$ is a nonempty compact set of $\mathbb{R}^n$, and*
(ii) *the function $d \mapsto f^\circ(x, d)$ is the support function of the generalized gradient $\partial f(x)$, i.e.*

$$f^\circ(x; d) = \max \{ g^T d \mid g \in \partial f(x) \}.$$

Proof. The proof of this theorem can be found in page 27 of Clarke (1981). •

Convex functions are locally Lipschitz in general as is demonstrated by the following theorem due to Roberts and Varberg (1974).

THEOREM 2.3.7. *Let U be an open convex subset of $\mathbb{R}^n$. Also let $f: U \rightarrow \mathbb{R}$ be a convex function. If there exists a point $\bar{x}$ and a neighborhood $N(\bar{x})$ of $\bar{x}$ such that f is bounded on $N(\bar{x})$, then f is Lipschitz near any point x in U .*

Proof. This theorem can be found as Proposition 2.2.6 in Clarke (1983) or Roberts and Vanberg (1974). •

A consequence of the last theorem is that a convex function may have both a one-sided and a generalized directional derivative. In fact we have the following proposition.

PROPOSITION 2.3.8 *Let $U \subset \mathbb{R}^n$ be an open convex set and let $f: U \rightarrow \mathbb{R}$ be a convex function. Assume that f is Lipschitz near some x in U . Then $f^\circ(x; d)$ coincides with $f'(x, d)$ for every d in $\mathbb{R}^n$ and the generalized gradient $\partial f(x)$ coincides with the subdifferential of f at x in the sense of convex analysis.*

Proof. This proposition can be found as Proposition 2.2.7 of Clarke (1983) or as Proposition 3 of Clarke (1976). •

The following example shows that we can have $f'(x; d) \neq f^\circ(x; d)$ even though f is Lipschitz near x .

Example 2.3.9. Consider $f: \mathbb{R} \rightarrow \mathbb{R}$ defined by $f(x) = -|x|$. This function has at $x = 0$ in the direction $d = 1$:

(1) a one-sided directional derivative

$$f'(0, 1) = \lim_{t \downarrow 0} \frac{-|t|}{t} = -1$$

and

(2) a generalized directional derivative because it is Lipschitz near zero.

Moreover for $t > 0$ and y such that $y + t < 0$, we have that

$$\frac{f(y + td) - f(y)}{t} = \frac{-|y + t| + |y|}{t} = 1,$$

which implies that

$$\limsup_{\substack{y \rightarrow 0 \\ t \downarrow 0}} \frac{f(y + td) - f(y)}{t} \geq 1$$

and therefore $f^\circ(0; 1) > f'(0; 1)$.

A function f such that the one-sided directional derivative $f'(x; d)$ is equal to the generalized directional derivative $f^\circ(x; d)$ is said to be *regular*. We saw that a convex function is regular at x , but Example 2.3.9 shows that a function that is Lipschitz near x is not necessarily regular at x .

To end the background material on Lipschitz functions we give the following important property of the point-to-set map $\partial f: x \rightarrow \partial f(x)$.

THEOREM 2.3.8. *Let x be any point in $\mathbb{R}^n$, and let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be Lipschitz near x . Then the point-to-set map $x \rightarrow \partial f(x)$ is closed.*

Proof. The proof of this theorem can be found in Clarke (1976) or Clarke (1983). •

The reader interested in more detail concerning these notions is referred to Rockafellar (1980), Clarke (1975), Clarke (1976), Clarke (1981) and Clarke (1983).

CHAPTER 3

Existence and Equivalence Properties

of the Various Derivatives of $f = \|F\|$

In this chapter we establish the existence and other properties of the generalized and the one-sided directional derivatives of the function $f(x) = \|F(x)\|$, where F is a continuously differentiable function from $\mathbb{R}^n$ into $\mathbb{R}^n$ and where $\|\cdot\|$ denotes any norm on $\mathbb{R}^n$. We show that the generalized gradient of f at x in $\mathbb{R}^n$ is equal to the generalized gradient at $s=0$ (commonly called subdifferential in the convex case) of the convex local model $m_x(s) = \|F(x) + F'(x)s\|$, i.e.

$$\partial f(x) = \partial m_x(0).$$

These results will allow us to establish later that our notion of stationary points of $f = \|F\|$ is equivalent to the standard one used in nonsmooth optimization and to the one used by Madsen (1975), Anderson and Osborne (1977), Powell (1983) and Yuan (1983a, b) and others. We also show that this result implies that the local model $m_x(s)$ at $s=0$ and the merit function $f(x) = \|F(x)\|$ at x have the same descent directions. This fact is known for the l_2 -norm (Levenberg-Marquardt) formulation and is important in the convergence

analysis. Finally we look closely at the special case where the norm is either the l_1 or l_∞ norm, since, in practice these choices are most likely.

3.1. Case of General Norm

In this section we establish that the function defined by

$$f(x) = ||F(x)|| \quad (3.1.1)$$

and its local model at x defined by

$$m_x(s) = ||F(x) + F'(x)s||, \quad (3.1.2)$$

where $|| \cdot ||$ is an arbitrary norm on $\mathbb{R}^n$, have the same one-sided directional derivatives at x and $s=0$ respectively. We also establish that f is a regular function and that the generalized gradient of f at every x in $\mathbb{R}^n$ is equal to the generalized gradient at $s=0$, of the convex local model m_x . For a convex function, the generalized gradient is equal to the subdifferential.

THEOREM 3.1.1. *Let x and d be points in $\mathbb{R}^n$. Also let $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be differentiable at x . Then*

(i)

$$f'(x; d) = m_x'(0; d). \quad (3.1.3)$$

Moreover, if in addition F is continuously differentiable in a neighborhood of x , then

(ii) *the generalized directional derivative of f exists,*

(iii) *the function f is regular at x , i.e.*

$$f'(x; s) = f^\circ(x; s) \quad (3.1.4)$$

for all s in $\mathbb{R}^n$, and

(iv) *the generalized gradient of f at x coincides with the subdifferential at $s=0$ of the convex local model m_x , i.e.*

$$\partial f(x) = \partial m_x(0). \quad (3.1.5)$$

The proof of Theorem 3.1.1 will require the following lemma.

LEMMA 3.1.1 *Let X be any open set in $\mathbb{R}^n$. If $F: X \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable on X , then the point-to-set map $M: X \rightarrow 2^{\mathbb{R}^n}$ defined by*

$$M(x) = \partial m_x(0) \quad (3.1.6)$$

is closed and upper semi-continuous at every x in X .

Proof of the lemma: First, we prove that M is closed at every x in X . Let x be any point in X . We have, from (3.1.2) and Definition (2.2.7), that $g \in M(x)$ if and only if

$$||F(x) + F'(x)s|| - ||F(x)|| \geq g^T s \quad (3.1.7)$$

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

Let $\{x_k, k \in \mathbb{N}\}$ be a sequence that converges to x ; also, because $M(x_k)$ is not empty, let $\{g_k, k \in \mathbb{N}\}$ be a sequence that converges to g and such that $g_k \in M(x_k)$ for all $k \in \mathbb{N}$. Then, for any fixed s in $\mathbb{R}^n$, we have

$$\|F(x_k) + F'(x_k)s\| - \|F(x_k)\| \geq g_k^T s,$$

which implies that

$$\|F(x) + F'(x)s\| - \|F(x)\| \geq g^T s. \quad (3.1.8)$$

Therefore g belongs to $\partial m_x(0) = M(x)$ and consequently M is closed at $x \in X$.

Since the point $x \in X$ was arbitrary, the functional M is closed on X .

Secondly, we prove that M is upper semi-continuous on X . Suppose that M is not u.s.c. at some x_0 in X . Then, there exists some open neighborhood W of $M(x_0)$ such that for any neighborhood V_0 of x_0 there exists x' in V_0 satisfying $M(x') \not\subset W$. This implies that there exist two sequences $\{x_k, k \in \mathbb{N}\}$ and $\{g_k, k \in \mathbb{N}\}$ such that

$$\begin{aligned} x_k &\rightarrow x_0, \\ g_k &\in M(x_k) \text{ and } g_k \notin W. \end{aligned} \quad (3.1.9)$$

Let us show that the sequence $\{g_k, k \in \mathbb{N}\}$ is bounded. The fact that g_k belongs to $M(x_k) = \partial m_{x_k}(0)$ implies that

$$\|F(x_k) + F'(x_k)s\| - \|F(x_k)\| \geq g_k^T s \quad (3.1.10)$$

for all s in $\mathbb{R}^n$, consequently

$$\|F'(x_k)s\| \geq g_k^T s \quad (3.1.11)$$

for all s in $\mathbb{R}^n$. In particular for $s = g_k$, this gives

$$\|F'(x_k)\| \geq c \|g_k\|_2 \quad (3.1.12)$$

for some constant c independent of k . On the other hand, the sequence $\{x_k, k \in \mathbb{N}\}$ converges to x and F is continuously differentiable in a neighborhood $N(x)$ of x . Consequently

$$\|F'(x_k)\| \leq \beta \quad (3.1.13)$$

for all $k \in \mathbb{N}$, where β is independent of k . The inequalities (3.1.12) and (3.1.13) imply that the sequence $\{g_k, k \in \mathbb{N}\}$ is bounded. Therefore there exists a subsequence $\{g_k, k \in \mathbb{N}'\}$ that converges to some g in $\mathbb{R}^n$. Because of (3.1.9), W is open, and $M(x_0) \subset W$, the limit point g does not belong to $M(x_0)$. In summary then, by assuming that M is not u.s.c. at x_0 we have proved that there exist a sequence $\{x_k, k \in \mathbb{N}'\}$ that converges to x_0 and a sequence $\{g_k \in M(x_k), k \in \mathbb{N}'\}$ that converge to a g that does not belong to $M(x_0)$. This contradicts that the point-to-set map M defined by $M(x) = \partial m_x(0)$ is closed on X . Therefore M is upper semi-continuous on X . •

Now we are prepared to prove Theorem 3.1.1.

Proof of the theorem.

Part (i). the existence of the one-sided directional derivative at $s=0$ in the direction d of the convex local model m_x , i.e., $m_x'(0; d)$, is the result of a straightforward application of Theorem 2.2.8. Now let us establish that the one-sided directional derivative of the function f defined by (3.1.1) at x in the direction d , $f'(x; d)$, exists and is equal to $m_x'(0; d)$. The one-sided direction derivative of f at x in the direction d is defined to be the limit

$$\lim_{t \downarrow 0} \frac{1}{t} [f(x + td) - f(x)] \quad (3.1.14)$$

whenever this limit exists. The hypothesis of differentiability of F at x implies that for any positive scalar t

$$F(x + td) = F(x) + tF'(x)d + o(t) \quad (3.1.15)$$

where

$$\lim_{t \rightarrow 0} \frac{o(t)}{t} = 0. \quad (3.1.16)$$

Therefore we have

$$f(x + td) - f(x) = \|F(x) + tF'(x)d + o(t)\| - \|F(x)\|$$

which implies (by the triangle inequality) that for any positive scalar t

$$\frac{m_x(d) - m_x(0)}{t} - \frac{\|o(t)\|}{t} \leq \frac{f(x + td) - f(x)}{t} \quad (3.1.17a)$$

and

$$\frac{f(x + td) - f(x)}{t} \leq \frac{m_x(td) - m_x(0)}{t} + \frac{\|o(t)\|}{t}. \quad (3.1.17b)$$

The above inequalities (3.1.17a) and (3.1.17b) and the limit (3.1.16) give

$$\lim_{t \downarrow 0} \frac{m_x(td) - m_x(0)}{t} = \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t}$$

or

$$m_x'(0; d) = f'(x; d). \quad (3.1.18)$$

Part (ii). We now establish the existence of the generalized directional derivative of f at x in the direction d . Because of Proposition 2.3.3, it is sufficient to show that the function f is Lipschitz near x . Moreover because of the inequality

$$|f(y) - f(y')| \leq \|F(y) - F(y')\| \quad (3.1.19)$$

it is sufficient to show that F is Lipschitz near x . However, since F is continuously differentiable in a neighborhood of x , F is Lipschitz near x (see the corollary of Proposition 2.21 in Clarke (1983)). Consequently, the generalized directional derivative of f at x in the direction d , $f'(x; d)$, exists.

Part (iii). We now show that the function $f = \|F\|$ is regular at x , i.e. $f^o(x; s) = f'(x; s)$ for all s in $\mathbb{R}^n$.

The proof of this result is motivated by a similar result in Clarke (1981) for a function which, instead of satisfying the hypothesis of Theorem 3.1.1, is

assumed to have a property called quasi-differentiability (we refer the reader to Pschenichney (1974) for the theory of quasi-differentiability). We do not have the quasi-differentiability property in our case. Let x be a point in X such that F is continuously differentiable on some neighborhood, say $N(x)$, of x . Since

$$f^o(x;0) = f'(x;0) \quad (3.1.20)$$

and both $f'(x;\cdot)$ and $f^o(x;\cdot)$ are positively homogeneous, it is sufficient to prove (3.1.4) for s of l_2 -norm one. Let s be any direction in $\mathbb{R}^n$ such that

$$\|s\|_2 = 1. \quad (3.1.21)$$

The generalized directional derivative of the function f at x in the direction s is given by

$$\limsup_{\substack{y \rightarrow x \\ t \downarrow 0}} \frac{1}{t} [f(y+ts) - f(y)]. \quad (3.1.22)$$

It is straightforward, from (3.1.14) and (3.1.22), that

$$f'(x;s) \leq f^o(x;s). \quad (3.1.23)$$

Consequently, it suffices to establish that

$$f^o(x;s) \leq f'(x;s). \quad (3.1.24)$$

Let ϵ be any positive scalar and let W be any open set such that

$$M(x) \subset W \subset M(x) + \epsilon B \quad (3.1.25)$$

where B is the l_2 -unit ball in $\mathbb{R}^n$ (W exists because $M(x)$ is a bounded set).

Since the point-to-set map M is u.s.c. at x , there exists a neighborhood V of x such that

$$x' \in V \Rightarrow M(x') \in W. \quad (3.1.26)$$

Given the neighborhood V of x , there exists a neighborhood V' of x and a positive scalar τ such that

$$y \in V' \text{ and } \theta \in [0, \tau] \Rightarrow y + \theta s \in V. \quad (3.1.27)$$

Consequently, the relations (3.1.25), (3.1.26) and (3.1.27) imply that

$$M(y + \theta s) \subset M(x) + \epsilon B, \quad (3.1.28)$$

for all y in V' and all θ in $[0, \tau]$. Theorem 2.2.8, applied to the local model of f at the point $y + \theta s$, and part (i) of the present theorem give

$$f'(y + \theta s; s) = \max \{g^T s \mid g \in M(y + \theta s)\}.$$

Now by (3.1.28), this implies that

$$f'(y + \theta s; s) \leq \max \{g^T s \mid g \in M(x)\} + \max \{g^T s \mid g \in \epsilon B\}.$$

By Theorem 2.2.8 and part (i) above we have

$$f'(y + \theta s; s) \leq f'(x; s) + \epsilon \max \{g^T s \mid g \in B\}.$$

and since $\|s\|_2 = 1$

$$f'(y + \theta s; s) \leq f'(x; s) + \epsilon. \quad (3.1.29)$$

Let us fix y in V' and consider the function h defined on $[0, \tau]$ by

$$h(\theta) = f(y + \theta s). \quad (3.1.30)$$

To establish that $f''(x; s) \leq f'(x; s)$, we will use the facts that h is Lipschitz near $\theta = 0$ and that h is absolutely continuous on $[0, \tau]$. Therefore we first establish these two properties.

Let us first demonstrate that h is Lipschitz near zero. We have

$$|h(\theta) - h(\theta')| = |f(y + \theta s) - f(y + \theta' s)|$$

and we know that f is Lipschitz near x , i.e. f is Lipschitz in a neighborhood of x , and without loss of generality we can suppose that V is such a neighborhood. Consequently, we have

$$|h(\theta) - h(\theta')| \leq K(x) |\theta - \theta'|, \quad (3.1.31)$$

for all θ and θ' in $[0, \tau]$ where $K(x) \neq 0$ is the local Lipschitz constant relative to f near x . This establishes the first desired property of the function h . Now we will demonstrate the absolute continuity of h on $[0, \tau]$. Let $\{I_i = (\theta_i, \theta_i'), i = 1, \dots, p\}$ be any finite family of subintervals of $[0, \tau]$ such that $I_i \cap I_j = \emptyset$ for $i \neq j$. Also let δ and γ be any positive scalars such that

$$\delta < \frac{\gamma}{K(x)} \quad \text{and} \quad \sum_{i=1}^p (\theta_i' - \theta_i) < \delta. \quad (3.1.32)$$

Then using (3.1.31) and (3.1.32) we get

$$\sum_{i=1}^p |h(\theta_i') - h(\theta_i)| < \gamma \quad (3.1.33)$$

which establishes the absolute continuity of the function h on $[0, \tau]$.

The absolute continuity of h on $[0, \tau]$ allows us to write

$$h(t) - h(0) = \int_0^t h'(\theta) d\theta \quad (3.1.34)$$

for all t in $[0, \tau]$. Now from (3.1.30) we obtain

$$h'(\theta) = f'(y + \theta s; s). \quad (3.1.35)$$

The equalities (3.1.34) and (3.1.35) imply

$$h(t) - h(0) = \int_0^t f'(y + \theta s; s) d\theta$$

for all t in $[0, \tau]$, and

$$\frac{h(t) - h(0)}{t} = \frac{1}{t} \int_0^t f'(y + \theta s; s) d\theta$$

for all t in $(0, \tau]$. Using the inequality (3.1.29) and the definition of h given in (3.1.30) we obtain

$$\frac{f(y + ts) - f(y)}{t} \leq f'(x; s) + \epsilon, \quad (3.1.36)$$

for all t in $(0, \tau]$. Finally, since in (3.1.36) y is an arbitrary point in the neighborhood V' of x , we have

$$\limsup_{\substack{y \rightarrow x \\ t \downarrow 0}} \frac{f(y + ts) - f(y)}{t} \leq f'(x; s) + \epsilon, \quad (3.1.37)$$

and since (3.1.37) holds for any positive scalar ϵ , we have

$$f^0(x; s) \leq f'(x; s). \quad (3.1.38)$$

The two inequalities (3.1.23) and (3.1.38) imply that

$$f^0(x; s) = f'(x; s) \quad (3.1.39)$$

Therefore the function f is regular at x .

Part (iv). Finally we establish that $\partial f(x)$, i.e. the generalized gradient of the function f at x , coincides with $\partial m_x(0)$, i.e. the subdifferential of the convex local m_x at zero. From Definition 2.3.5, we have

$$\partial f(x) = \{g \in \mathbb{R}^n \mid f^0(x; d) \geq g^T d, \forall d \in \mathbb{R}^n\} \quad (3.1.40)$$

and from Theorem 2.2.11 and equality (3.1.3)

$$\partial m_x(0) = \{g \in \mathbb{R}^n \mid f'(x; d) \geq g^T d, \forall d \in \mathbb{R}^n\}. \quad (3.1.41)$$

The properties (3.1.40) and (3.1.41) together with (3.1.4) imply (3.1.5), which ends the proof of the theorem. •

The function $m_x: \mathbb{R}^n \rightarrow \mathbb{R}$ defined by $m_x(s) = \|F(x) + F'(x)s\|$ is convex, therefore by Theorem 2.2.10 we have

$$m_x'(0; s) \leq \|F(x) + F'(x)s\| - \|F(x)\|$$

for all s in $\mathbb{R}^n$, which by (3.1.3), i.e. $m_x'(0; s) = f'(x; s)$ implies

$$f'(x; s) \leq \|F(x) + F'(x)s\| - \|F(x)\| \quad (3.1.42)$$

for all s in $\mathbb{R}^n$. When s is the Newton step, say s^N , (3.1.42) becomes

$$f'(x; s^N) \leq -\|F(x)\| \quad (3.1.43)$$

or

$$f'(x; s^N) \leq -\|F'(x)s^N\|. \quad (3.1.44)$$

We now show that in fact the quantity $-\|F'(x)s\|$ is a lower bound for $f'(x; s)$ for all s in $\mathbb{R}^n$, and therefore (3.1.44) is an equality.

THEOREM 3.1.3. *Let x be any point in $\mathbb{R}^n$ and also let $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be differentiable at x . Then*

$$-\|F'(x)s\| \leq f'(x; s) \leq \|F(x) + F'(x)s\| - \|F(x)\| \quad (3.1.45)$$

for all s in $\mathbb{R}^n$. Moreover if $F'(x)$ is nonsingular, then for $s = s^N$, the Newton step, the inequalities in (3.1.45) become equalities, i.e.

$$-\|F'(x)s^N\| = f'(x; s^N) \quad (3.1.46)$$

and

$$f'(x; s^N) = \|F(x) + F'(x)s^N\| - \|F(x)\|; \quad (3.1.47a)$$

equivalently, because $F(x) + F'(x)s^N = 0$

$$f'(x; s^N) = -\|F(x)\|. \quad (3.1.47b)$$

Proof. By (3.1.3) and Theorem (2.2.10) it is sufficient to establish the first inequality in (3.1.45). We have for any positive scalar t and any s in $\mathbb{R}^n$

$$\left| \|F(x) + F'(x)ts\| - \|F(x)\| \right| \leq t \|F'(x)s\|$$

which in turn implies

$$-\|F'(x)s\| \leq \frac{1}{t} \left[\|F(x) + F'(x)ts\| - \|F(x)\| \right] \quad (3.1.48)$$

and by passing to the limit in (3.1.48) as t decreases to zero, we obtain

$$-\|F'(x)s\| \leq \lim_{t \downarrow 0} \frac{1}{t} \left[\|F(x) + F'(x)ts\| - \|F(x)\| \right] .$$

From Theorem 3.1.3 we have

$$-\|F'(x)s\| \leq f'(x; s) . \quad (3.1.49)$$

Now suppose that $F'(x)$ is nonsingular and let s^N denote the Newton step.

Then (3.1.45) becomes

$$-\|F'(x)s^N\| \leq f'(x; s^N) \leq -\|F(x)\| ,$$

but, since $F(x) + F'(x)s^N = 0$, we have

$$-\|F'(x)s^N\| = f'(x; s^N) \quad (3.1.49a)$$

$$\text{or} \quad -\|F(x)\| = f'(x; s^N) \quad (3.1.49b)$$

and

$$f'(x; s^N) = \|F(x) + F'(x)s^N\| - \|F(x)\| . \quad (3.1.50)$$

The equalities (3.1.49) and (3.1.50) imply that the inequalities in (3.1.45) are

equalities in this case.

In an algorithm that uses an α -condition criterion to accept or reject a given step s , the quantity of decrease should be, for a fixed α , at least $\alpha f'(x; s)$. When the one-sided directional derivative is not available or expensive to obtain, one usually tries to use some approximation to it. Theorem 3.1.3 means, on the one hand, that the quantity $-\|F'(x)s\|$ is not a practical approximation whenever s is far away from being the Newton step, because $\alpha(-\|F'(x)s\|)$ might be smaller than $f'(x; s)$, and on the other hand that the quantity $\|F(x) + F'(x)s\| - \|F(x)\|$ is a practical one. We believe that any acceptable approximation to $f'(x; s)$ should come from the interval $[f'(x; s), 0]$ as long as $f'(x; s)$ is negative (i.e. as long as x is not stationary as defined in Chapter 4).

If we choose the norms $\|\cdot\|_a$ and $\|\cdot\|_b$ in (1.6.1) to be either the l_1 -norm or l_∞ -norm, then our subproblem (1.6.1) can be reformulated as a linear program. For this reason we take a closer look at the cases where $\|\cdot\|_a$ is either the l_1 -norm or the l_∞ -norm.

3.2. The case where $\|\cdot\|_a$ is the l_1 -norm

In this section, we consider $f(x) = \|F(x)\|_a$ where $\|\cdot\|_a$ is the l_1 -norm. We derive an explicit expression for the one-sided directional derivative $f'(x; s)$

which, by Theorem 3.1.1, is equal to the generalized directional derivative $f^\circ(x; s)$. Using this expression we derive an important property of that derivative for small s .

PROPOSITION 3.2.1. *Let x be any point in $\mathbb{R}^n$. If $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable at x and $f(x) = \|F(x)\|_1$ where $\|\cdot\|_1$ is the l_1 -norm, then for all s in $\mathbb{R}^n$ we have*

$$f'(x; s) = \sum_{i \notin A(x)} \text{sign}(F_i(x)) s^T \nabla F_i(x) + \sum_{i \in A(x)} |s^T \nabla F_i(x)|, \quad (3.2.1)$$

where $A(x) = \{i / F_i(x) = 0\}$. Moreover, for sufficiently small s , we have

$$f'(x; s) = \|F(x) + F'(x)s\|_1 - \|F(x)\|_1. \quad (3.2.2)$$

Proof. By Theorem 3.1.1, establishing (3.2.1) is equivalent to showing that

$$m_x'(0; s) = \sum_{i \notin A(x)} \text{sign}(F_i(x)) s^T \nabla F_i(x) + \sum_{i \in A(x)} |s^T \nabla F_i(x)|. \quad (3.2.3)$$

where $m_x(s) = \|F(x) + F'(x)s\|_1$. Actually, we will show that, for sufficiently small positive t , we have

$$\frac{m_x(ts) - m_x(0)}{t} = \sum_{i \notin A(x)} \text{sign}(F_i(x)) s^T \nabla F_i(x) + \sum_{i \in A(x)} |s^T \nabla F_i(x)|. \quad (3.2.4)$$

Indeed, for t positive we have

$$m_x(ts) - m_x(0) = \sum_{i=1}^n \left[|F_i(x) + ts^T \nabla F_i(x)| - |F_i(x)| \right] \quad (3.2.5)$$

$$= \sum_{i \notin A(x)} \left[|F_i(x) + ts^T \nabla F_i(x)| - |F_i(x)| \right] + \sum_{i \in A(x)} t |s^T \nabla F_i(x)|.$$

If $A(x) = \{1, \dots, n\}$, then the result is obvious. Also, if for all $i \notin A(x)$ we have $s^T \nabla F_i(x) = 0$ or $\text{sign}(F_i(x)) = \text{sign}(s^T \nabla F_i(x))$, then the result is also obvious. Therefore, let us suppose that there exists at least one $j \notin A(x)$ such that $F_j(x) s^T \nabla F_j(x) < 0$. Then, for

$$0 < t < t(x, s) \quad (3.2.6a)$$

where

$$t(x, s) = \min \left\{ \frac{|F_j(x)|}{|s^T \nabla F_j(x)|}, \quad j \notin A(x), \right. \\ \left. \text{and } F_j(x) \cdot s^T \nabla F_j(x) < 0 \right\} \quad (3.2.6b)$$

we have

$$|F_i(x) + ts^T \nabla F_i(x)| = \text{sign}(F_i(x)) (F_i(x) + ts^T \nabla F_i(x)) \quad (3.2.7)$$

for all $i \notin A(x)$. The inequalities (3.2.6) and the equalities (3.2.5) and (3.2.7) give (3.2.4) and therefore (3.2.1). Now, let us establish (3.2.2). Writing s instead of ts in (3.2.5), we obtain

$$m_x(s) - m_x(0) = \sum_{i \notin A(x)} \left[|F_i(x) + s^T \nabla F_i(x)| - |F_i(x)| \right] + \sum_{i \in A(x)} |s^T \nabla F_i(x)|. \quad (3.2.8)$$

On the other hand, if

$$\|s\|_2 < \min \left\{ \frac{|F_j(x)|}{\|\nabla F_j(x)\|_2}, j \notin A(x) \text{ and } \nabla F_j(x) \neq 0 \right\}, \quad (3.2.9)$$

then

$$|s^T \nabla F_i(x)| < |F_i(x)|, \text{ for all } i \notin A(x),$$

and consequently, for $i \notin A(x)$ we have

$$|F_i(x) + s^T \nabla F_i(x)| = \text{sign}(F_i(x))(F_i(x) + s^T \nabla F_i(x)). \quad (3.2.30)$$

If s satisfies (3.2.9), then using (3.2.30) in (3.2.8) we have

$$m_x(s) - m_x(0) = \sum_{i \notin A(x)} \text{sign}(F_i(x)) s^T \nabla F_i(x) + \sum_{i \in A(x)} |s^T \nabla F_i(x)|;$$

which gives (3.2.2) by (3.2.1). •

3.3. The case where $\|\cdot\|_s$ is the l_∞ -norm

In this section we consider $f(x) = \|F(x)\|_s$ where $\|\cdot\|_s$ is the l_∞ -norm.

We derive an explicit expression for the one-sided directional derivative $f'(x; s)$

which, by Theorem 3.1.1, is also the generalized directional derivative $f^c(x; s)$.

Using this expression, we derive an important property of that derivative for small s .

PROPOSITION 3.2.2. *Let f be a real function on $\mathbb{R}^n$ defined by $f(x) = \|F(x)\|_\infty$ where $\|\cdot\|_\infty$ is the l_∞ -norm and $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable at some x such that $f(x) \neq 0$. Then for all s in $\mathbb{R}^n$ we have*

$$f'(x; s) = \max_{i \in M(x)} \left\{ \text{sign}(F_i(x)) s^T \nabla F_i(x) \right\} \quad (3.3.1)$$

where $M(x) = \{i / |F_i(x)| = \|F(x)\|_\infty\}$. Moreover, for sufficiently small s we have

$$f'(x; s) = \|F(x) + F'(x)s\|_\infty - \|F(x)\|_\infty. \quad (3.3.2)$$

Proof. By Theorem 3.1.1, establishing (3.3.1) is equivalent to showing that

$$m_x'(0; s) = \max_{i \in M(x)} \left\{ \text{sign}(F_i(x)) \cdot s^T \nabla F_i(x) \right\}. \quad (3.3.3)$$

Actually, we show that, for sufficiently small positive t , we have

$$\frac{m_x(ts) - m_x(0)}{t} = \max_{i \in M(x)} \left\{ \text{sign}(F_i(x)) s^T \nabla F_i(x) \right\}. \quad (3.3.4)$$

Indeed, for t positive we have

$$m_x(ts) - m_x(0) = \max_{1 \leq i \leq n} |F_i(x) + ts^T \nabla F_i(x)| - \max_{1 \leq i \leq n} |F_i(x)|. \quad (3.3.5)$$

Since t can be arbitrarily small, we can assume without loss of generality that $F_i(x) \neq 0$ for all $i \in \{1, \dots, n\}$. Therefore, let us suppose that

$$|f_i(x)| = f(x) \quad \text{for } i = 1, \dots, p$$

$$\text{and} \quad |f_i(x)| < f(x) \quad \text{for } i = p+1, \dots, n.$$

Therefore, for positive scalar t sufficiently small, the index giving the $\max_{1 \leq i \leq n} |F_i(x) + ts^T \nabla F_i(x)|$ belongs to $\{1, \dots, p\}$. This index depends on t , but this will not cause any problem. We can rewrite (3.3.5), for small positive t , as

$$\begin{aligned} m_x(ts) - m_x(0) &= \max_{i \in M(x)} |F_i(x) + ts^T \nabla F_i(x)| - \max_{i \in M(x)} |F_i(x)| \\ &= \max_{i \in M(x)} |F_i(x) + ts^T \nabla F_i(x)| - f(x) \\ &= \max_{i \in M(x)} \left[|F_i(x) + ts^T \nabla F_i(x)| - |F_i(x)| \right], \end{aligned}$$

which gives

$$m_x(ts) - m_x(0) = t \max_{i \in M} \left[\text{sign}(F_i(x)) s^T \nabla F_i(x) \right] \quad (3.3.6)$$

for sufficiently small positive t . Clearly, (3.3.6) is equivalent to (3.3.4). From (3.3.4), the equality (3.3.3) is obvious.

Now, let us show that for small s we have

$$f'(x; s) = \|F(x) + F'(x)s\|_\infty - \|F(x)\|_\infty.$$

Writing s instead of ts in (3.3.6), we obtain

$$m_x(s) - m_x(0) = \max_{i \in M(x)} \left[\text{sign}(F_i(x)) s^T \nabla F_i(x) \right],$$

which is exactly (3.3.2) by (3.3.1).

CHAPTER 4

Optimality Conditions

In this chapter, we derive optimality conditions for the nondifferentiable optimization problem

$$\text{minimize } f(x) = \|F(x)\|$$

where $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable and $\|\cdot\|$ denotes an arbitrary norm on $\mathbb{R}^n$. These conditions will allow us to define stationary points of f . We show that if the Jacobian of F at a stationary point x_* is nonsingular, then x_* is a solution of the nonlinear system $F(x) = 0$.

4.1. Optimality Conditions

Let us consider the following unconstrained optimization problem

$$\text{minimize } f(x) \quad (4.1.1)$$

where f is any continuous real function defined on $\mathbb{R}^n$. In the smooth case, i.e. f is differentiable, the point x_* is said to be a stationary point of f if the gradient of f vanishes at x_* , i.e.

$$\nabla f(x_*) = 0. \quad (4.1.2)$$

The standard generalization of stationary point to the nonsmooth case is given by the following definition.

DEFINITION 4.1.1. Let x_* be any point in $\mathbb{R}^n$. Also let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be Lipschitz near x_* . Then, we say that x_* is a stationary point of f if zero belongs to the generalized gradient of f at x_* , i.e.

$$0 \in \partial f(x_*). \quad (4.1.3)$$

In our application, the objective function is defined by

$$f(x) = \|F(x)\| \quad (4.1.4)$$

where $\| \cdot \|$ denotes an arbitrary norm on $\mathbb{R}^n$ and $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable. In Chapter 3 we demonstrated that f is a regular function, i.e. at each point x in $\mathbb{R}^n$ and in any direction d in $\mathbb{R}^n$ the usual one-sided directional derivative $f'(x; d)$ exists and is equal to the generalized directional derivative $f^\circ(x; d)$. This enables us to establish the following useful characterization of stationarity

LEMMA 4.1.2. Let x_* be any point in $\mathbb{R}^n$, and also let f be defined by (4.1.4) where F is continuously differentiable and where $\| \cdot \|$ denotes an arbitrary norm on $\mathbb{R}^n$. Then, a necessary and sufficient condition for x_* to be a

stationary point of f is

$$f'(x_*; s) \geq 0 \quad (4.1.5)$$

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

Proof. The proof follows directly from the regularity of f (see Theorem 3.1.1) and Definition 2.3.5. •

Madsen (1975) uses the criterion (1.1.26), i.e.

$$\|F(x_*)\| = \min \{ \|F(x_*) + F'(x_*)s\|, \quad s \in \mathbb{R}^n \} \quad (4.1.6)$$

to define a stationary point in his application where $\| \cdot \|$ is taken as the l_∞ -norm. In the following lemma we show that his definition is also equivalent to the standard definition.

LEMMA 4.1.3. Let x_* be any point in $\mathbb{R}^n$, and also let f be defined by (4.1.4) where F is continuously differentiable and where $\| \cdot \|$ denotes any norm on $\mathbb{R}^n$. Then, x_* is a stationary point of f if and only if zero is a minimizer of the local model m_x of f at x_* , i.e.

$$m_{x_*}(0) \leq m_{x_*}(s) \quad (4.1.7)$$

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

Proof. Suppose that x_* is a stationary point of f . Then, by Lemma 4.1.2, we have

$$f'(x_*; s) \geq 0 \quad (4.1.8)$$

for all s in $\mathbb{R}^n$. But by Theorem 2.2.10

$$f'(x_*; s) \leq m_{x_*}(s) - m_{x_*}(0),$$

which, together with (4.1.8), gives (4.1.7). Now, suppose that (4.1.7) holds and let s be any direction in $\mathbb{R}^n$. Then, by (4.1.7), we have

$$\frac{1}{t} [m_{x_*}(ts) - m_{x_*}(0)] \geq 0 \quad \forall t > 0,$$

which implies that

$$m_{x_*}'(0; s) = \lim_{t \downarrow 0} \frac{1}{t} [m_{x_*}(ts) - m_{x_*}(0)] \geq 0.$$

But by Theorem 3.1.1, we have that $m_{x_*}'(0; s)$ is equal to $f'(x_*; s)$, and since s is an arbitrary direction in $\mathbb{R}^n$, x_* is a stationary point of f . •

Observe that the criterion (4.1.7) is equivalent to the statement

$$\|F(x_*)\| \leq \|F(x_*) + F'(x_*)s\|$$

for all s in $\mathbb{R}^n$, which is, for the special case when $\|\cdot\|$ is the l_∞ -norm, Madsen's criterion (1.1.26). The criterion of stationarity (4.1.7) is equivalent to asking $s=0$ to be an unconstrained global minimizer of the local model $m_{x_*}(\cdot)$. But, since this model is convex, we have that x_* is a stationary point of f if and only if

$$m_{x_*}(0) = \min \{m_{x_*}(s) / \|s\|_b \leq 1\}$$

or equivalently

$$\|F(x_*)\| = \min \{ \|F(x_*) + F'(x_*)s\|_b / \|s\|_b \leq 1 \}.$$

where $\|\cdot\|_b$ is an arbitrary norm on $\mathbb{R}^n$.

In their approach, Powell (1983) and Yuan (1983a, b) use the criterion (1.2.9), i.e.

$$h(F(x)) = \min \{ h(F(x) + F'(x)s) / \|s\| \leq 1 \} \quad (4.1.9)$$

to define the notion of stationarity. Namely, x_* is a stationary point of the function

$$f(x) = h(F(x)) \quad (4.1.10)$$

if (4.1.9) holds for $x = x_*$. Observe that the objective function in (4.1.9) is different from the local model

$$\phi_x(s) = h(F(x) + F'(x)s) + s^T B_k s \quad (4.1.11)$$

used by Powell (1983) and Yuan (1983a, b) in the subproblem.

In the following theorem, we establish that for any x in $\mathbb{R}^n$, the one-sided directional derivative of ϕ_x at $s=0$ is equal to the one-sided directional derivative at x of the function f defined by (4.1.10).

THEOREM 4.1.4. Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be continuously differentiable and $h: \mathbb{R}^n \rightarrow \mathbb{R}$ be a continuous convex function. Also let x be any point in $\mathbb{R}^n$. Then the function f defined by (4.1.10) is Lipschitz near x . Moreover if ϕ_x is defined by (4.1.11), then

$$f'(x; s) = \phi_x'(0; s) \quad (4.1.12)$$

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

Actually, we establish that $f'(x, \cdot)$ is equal to $\theta_x'(0; \cdot)$, where θ is the convex function defined by

$$\theta_x(s) = h(F(x) + F'(x)s), \quad (4.1.13)$$

because it is obvious that

$$\theta_x'(0; s) = \phi_x'(0; s). \quad (4.1.14)$$

Proof of the theorem. By Theorem 2.3.9, we have that the convex function h is Lipschitz near $F(x)$. Consequently in a similar manner to the proof of Theorem 3.1.1 (ii), we establish that f is Lipschitz near x . Now, let us demonstrate that $f'(x; \cdot)$ is equal to $\theta_x'(0; \cdot)$ where θ is defined by (4.1.13). We have that

$$f'(x; s) = \lim_{t \rightarrow 0} \frac{1}{t} [h(F(x + ts)) - h(F(x))], \quad (4.1.15)$$

where s is an arbitrary direction in $\mathbb{R}^n$. Since F is continuously differentiable, we have

$$\begin{aligned} h(F(x + ts)) - h(F(x)) &= [h(F(x) + tF'(x)s) - h(F(x))] \\ &+ [h(F(x) + tF'(x)s + o(t)) - h(F(x) + tF'(x)s)], \end{aligned} \quad (4.1.16)$$

which implies, together with (4.1.14)

$$\begin{aligned} f'(x; s) &= \theta_x'(0; s) \\ \lim_{t \rightarrow 0} \frac{1}{t} [h(F(x) + tF'(x)s + o(t)) - h(F(x) + tF'(x)s)] &= 0. \end{aligned} \quad (4.1.17)$$

Let us calculate the limit in (4.1.17). Since the function h is Lipschitz near $F(x)$, there exists a neighborhood $N(F(x))$ of $F(x)$ and a positive constant $K(x) \neq 0$ such that

$$|h(y) - h(y')| \leq K(x) \|y - y'\|$$

for all y and y' in $N(F(x))$. It is clear that there exists $\epsilon > 0$ such that, for $t \in (0, \epsilon)$, both quantities $F(x) + tF'(x)s + o(t)$ and $F(x) + tF'(x)s$ are in $N(F(x))$. Then, we have

$$|h(F(x) + tF'(x)s + o(t)) - h(F(x) + tF'(x)s)| \leq K(x) \frac{o(t)}{t}.$$

Consequently, the limit in (4.1.17) is zero; which implies that

$$f'(x; s) = \theta_x'(0; s).$$

The fact that $\theta_x'(0; s) = \phi_x'(0; s)$ ends the proof. •

The following corollary shows that the definition of stationarity in Powell (1983) and Yuan (1983a, b) is equivalent to the notion of stationarity stated in Lemma 4.1.2, i.e. x is a stationary point of $f = h(F)$ if and only if $f'(x; s) \geq 0$ for all s in $\mathbb{R}^n$.

COROLLARY 4.1.5. Assume the hypothesis of Theorem 4.1.4. Then

$$h(F(x)) = \min \{h(F(x) + F'(x)s) / \|s\| \leq 1\} \quad (4.1.18)$$

holds if and only if

$$f'(x; s) \geq 0 \quad (4.1.19)$$

for all s in $\mathbb{R}^n$, where f is defined by (4.1.10).

Proof. Because of (4.1.12), the proof is similar to the proof of Lemma 4.1.3. •

Now, because the function h is convex, one can establish that the function f defined by (4.1.10) is regular in a similar manner to the proof of regularity of the function $\|F(x)\|$, where $\|\cdot\|$ is any norm on $\mathbb{R}^n$. This allows us to establish the equivalence of the notion of stationarity used in Powell (1983) and Yuan (1983a, b), that is x is a stationary point of $f = h(F)$ if and only if (4.1.18) holds, and the standard characterization of a stationary point x in nonsmooth optimization, that is $0 \in \partial f(x)$.

From Lemma (4.1.3) it is obvious that a solution of the nonlinear system

$$F(x) = 0 \quad (4.1.20)$$

is a stationary point of $f = \|F\|$. In the following theorem we show that if the Jacobian of F at some stationary point x_* is nonsingular then x_* is indeed a solution of the nonlinear system (4.1.20).

THEOREM 4.1.6. Let $\|\cdot\|$ denote an arbitrary norm on $\mathbb{R}^n$. Also let x_* be a stationary point of $f = \|F\|$. If $F'(x_*)$ is nonsingular, then x_* is a solution of the nonlinear system $F(x) = 0$.

Proof. Suppose that $F(x_*) \neq 0$ and consider the following linear system

$$F(x_*) + F'(x_*)s = 0. \quad (4.1.21)$$

This system has a nonzero solution, s_*^N , which is the Newton step. But by Theorem 3.1.3, we have that

$$f'(x_*; s_*^N) = -\|F(x_*)\|.$$

It follows that $f'(x_*; s_*^N)$ is negative, since $F(x_*) \neq 0$. However, this contradicts the hypothesis that x_* is a stationary point of f . •

Remark. Observe that this result still holds if instead of requiring nonsingularity of $F'(x_*)$ we merely require (4.1.21) to have a solution.

Lemma 4.1.3 says that zero is a minimizer of the convex function $m_{x_*}(\cdot)$ whenever x_* is a stationary point of f . But this minimizer might not be unique. Usually in optimization theory we demonstrate the uniqueness of a minimizer by establishing strict convexity of the objective function. Unfortunately the norm is never strictly convex. However, in our case, the following corollary allows us to establish the uniqueness of the minimizer of $m_{x_*}(\cdot)$ over $\mathbb{R}^n$.

COROLLARY 4.1.7. *Assume the hypothesis of Theorem 4.1.6. If $F(x_*)$ is nonsingular, then $s = 0$ is the unique minimizer of the local model $m_{x_*}(\cdot)$ over $\mathbb{R}^n$.*

Proof. Theorem (4.1.6) implies that x_* solves the nonlinear system (4.1.10), i.e.

$$F(x_*) = 0. \quad (4.1.22)$$

Suppose that there exists s in $\mathbb{R}^n$ such that $m_{x_*}(s) = \|F(x_*)\|$. Then (4.1.22) implies that

$$\|F(x_*) + F'(x_*)s\| = 0. \quad (4.1.23)$$

The two equalities (4.1.22) and (4.1.23), together with the hypothesis of nonsingularity of $F'(x_*)$, imply that s is zero.

CHAPTER 5

The Algorithm and Convergence Results

In this chapter, we define our algorithm for approximating a solution of the nondifferentiable optimization problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} f(x) = \|(F)\| \quad (5.0.1)$$

where $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable and where $\|\cdot\|$ denotes an arbitrary norm on $\mathbb{R}^n$. We also demonstrate that our algorithm is globally convergent in the following sense. Starting from a point x_0 in $\mathbb{R}^n$, the algorithm generates a sequence $\{x_k, k \in \mathbb{N}\}$ which has the property that all of its accumulation points are stationary points of f in problem (5.0.1). Furthermore, if any one of the accumulation points of the sequences $\{x_k, k \in \mathbb{N}\}$, say x_* , is such that $F'(x_*)$ is nonsingular, then all accumulation points are necessarily solutions of the nonlinear system

$$F(x) = 0. \quad (5.0.2)$$

The algorithm also has the desirable property that it asymptotically reduces to Newton's method and is therefore q -quadratically convergent.

5.1. The Trust Region Algorithm

Let X and S be subsets of $\mathbb{R}^n$ such that (x, s) is in $X \times S$ if and only if $f'(x; s) < 0$. Let $\gamma: X \times S \rightarrow \mathbb{R}$ be an upper semi-continuous function satisfying

$$\gamma(x, s) \in [f'(x; s), 0], \quad (5.1.1)$$

let $c_i, i = 1, \dots, 5$ be positive scalars such that

$$0 < c_1 < c_2 < 1 \leq c_3, \quad (5.1.2)$$

$$0 < c_4 < c_5 < 1, \quad (5.1.3)$$

and finally let x_0 be any point in $\mathbb{R}^n$ and δ_0 any positive scalar.

Suppose that x_k and δ_k are the iterate and the trust region radius determined by the algorithm at the k^{th} iteration. The algorithm determines x_{k+1} and δ_{k+1} in the following manner:

STEP 1. Obtain s_k as the solution of the model trust region-problem

$$\text{minimize } m_x(s) = \|F(x_k) + F'(x_k)s\| \quad (5.1.4a)$$

$$\text{subject to } \|s\| \leq \delta_k. \quad (5.1.4b)$$

STEP 2. If $f(x_k + s_k) \leq f(x_k) + c_1\gamma(x_k, s_k)$

set $x_{k+1} = x_k + s_k$ and go to STEP 3,

Else, choose δ_k so that

$$c_4 \|s_k\| \leq \delta_k \leq c_5 \|s_k\|,$$

and go to STEP 1,

STEP 3. If $f(x_{k+1}) < f(x_k) + c_2[m_k(s_k) - f(x_k)]$

choose δ_{k+1} so that

$$\|s_k\| \leq \delta_{k+1} \leq \max(\delta_k, c_3 \|s_k\|)$$

and go to STEP 1,

Else, if $f(x_{k+1}) > f(x_k) + c_2[m_k(s_k) - f(x_k)]$,

choose δ_{k+1} so that

$$c_4 \|s_k\| \leq \delta_{k+1} \leq c_5 \|s_k\|,$$

Else, choose δ_{k+1} so that

$$c_4 \|s_k\| \leq \delta_{k+1} \leq \max(\delta_k, c_3 \|s_k\|).$$

Several straightforward choices for the function γ used in STEP 2 of the algorithm are

$$\gamma(x, s) = f'(x; s) \quad (5.1.5)$$

and

$$\gamma(x, s) = \|(F(x) + F'(x)s) - F(x)\|. \quad (5.1.6)$$

For the choice (5.1.5) γ is upper semi-continuous by Theorem 2.3.4, and for the choice (5.1.6) γ is not only upper-semi-continuous but is actually continuous.

In their algorithm, Duff, Nocedal and Reid (1984) used

$$\gamma(x, s) = - \|F'(x)s\|_1. \quad (5.1.7)$$

Our convergence analysis does not allow the Duff, Nocedal and Reid choice (5.1.7). It is not acceptable because of the first inequality established in Theorem 3.1.3, i.e.

$$- \|F'(x)s\| \leq f'(x; s) \quad (5.1.8)$$

for all s in $\mathbb{R}^n$. The choice (5.1.7) of γ does not satisfy condition (5.1.1) of our algorithm except in the very special case where we have

$$- \|F'(x_k)s_k\|_1 = f'(x_k; s_k). \quad (5.1.9)$$

Let us take a look at the case where equality (5.1.9) holds. Equality (5.1.9) and Proposition 3.2.1 imply that

$$\sum_{i \in A(x_k)} \left[|s_k^T \nabla F_i(x_k)| + \text{sign } F_i(x_k) s_k^T \nabla F_i(x_k) \right] = - 2 \sum_{i \in A(x_k)} |s_k^T \nabla F_i(x_k)|, \quad (5.1.10)$$

where

$$A(x_k) = \{i / F_i(x_k) = 0\} \quad (5.1.11)$$

and

$$\sum_{i \in A(x_k)} \equiv 0 \quad \text{if } A(x_k) = \emptyset. \quad (5.1.12)$$

One possible way to obtain (5.1.10) is to ask s_k to satisfy

$$s_k^T \nabla F_i(x_k) = 0 \quad (5.1.13a)$$

for all i in $A(x_k)$, which implies that

$$\sum_{i \in A(x_k)} \left[|s_k^T \nabla F_i(x_k)| + \text{sign } F_i(x_k) s_k^T \nabla F_i(x_k) \right] = 0,$$

or equivalently

$$s_k^T \nabla F_i(x_k) \leq 0 \quad \text{if } F_i(x_k) > 0 \quad (5.1.13b)$$

and

$$s_k^T \nabla F_i(x_k) \geq 0 \quad \text{if } F_i(x_k) < 0. \quad (5.1.13c)$$

Observe that if (5.1.10) holds with $A(x_k) = \emptyset$, then (5.1.13b) and (5.1.13c) are not only a possible way to obtain (5.1.10) but, indeed, they are necessary conditions. Therefore, a possible way of trying to include the choice (5.1.7) of Duff, Nocedal and Reid in our convergence analysis is to ask for s_k to satisfy condition (5.1.13), or equivalently to ask for s_k to solve, instead of the subproblem (5.1.4), the following one:

$$\text{minimize} \quad \|F(x_k) + F'(x_k)s\|_1 \quad (5.1.14a)$$

subject to

$$s^T \nabla F_i(x_k) = 0 \quad i \in A(x_k) \quad (5.1.14b)$$

$$s^T \nabla F_i(x_k) \geq 0 \quad i \in \bar{A}_-(x_k) \quad (5.1.14c)$$

$$s^T \nabla F_i(x_k) \leq 0 \quad i \in \bar{A}_+(x_k) \quad (5.1.14d)$$

$$\|s\| \leq \delta,$$

where

$$A(x_k) = \{i / F_i(x_k) = 0\}$$

$$\bar{A}_+(x_k) = \{i / F_i(x_k) > 0\}$$

$$\bar{A}_-(x_k) = \{i / F_i(x_k) < 0\}.$$

Observe that if the Newton step lies inside the trust region, then it solves problem (5.1.14).

Inequality (5.1.8) implies that the Duff, Nocedal and Reid choice of γ requires too much descent and could be impossible to satisfy. Suppose that there exists some x in $\mathbb{R}^n$ such that s , the solution of problem (5.1.4), does not satisfy (5.1.14b), (5.1.14c) or (5.1.14d) even when the trust region is reduced. Also suppose that for a given α in $(0,1)$, the constant of the α -condition test, there exists a positive scalar μ , which may depend on s , such that

$$\alpha \mu > 1 \quad (5.1.15a)$$

and

$$-||F'(x)s||_1 \leq \mu f'(x, s). \quad (5.1.15b)$$

Then, we have that

$$-\alpha ||F'(x)s||_1 \leq \alpha \mu f'(x, s). \quad (5.1.16)$$

Since $f'(x, s)$ is negative, (5.1.16) and (5.1.15a) imply that

$$-\alpha ||F'(x)s||_1 < f'(x, s)$$

and consequently the α -condition test of Duff, Nocedal and Reid

$$||F(x+s)||_1 \leq ||F(x)||_1 - \alpha ||F'(x)s||_1$$

cannot be satisfied.

Near the solution of Problem (5.0.1) we expect s_k to be the Newton step and in this case the three choices (5.1.5), (5.1.6) and (5.1.7) coincide; consequently the asymptotic properties of the respective algorithms are the same (see Theorem (3.1.3)). For the l_1 -norm or the l_∞ -norm case, when s_k is sufficiently small, the functions defined by (5.1.5) and (5.1.6) are equal (see Proposition 3.2.1 or Proposition 3.2.2), and consequently the respective algorithms will be the same.

5.2. Global Convergence Results

In order to establish the global convergence of our algorithm we will model it by a point-to-set map, and then we show that the hypothesis of Zangwill's Theorem 2.1.1 holds.

We make the assumption that

$$X_0 = \left\{ s \in \mathbb{R}^n / f(x) \leq f(x_0) \right\} \quad (5.2.1)$$

is bounded (and therefore compact). Let Δ be a positive constant such that for all x in X_0

$$\|x\| \leq \frac{1}{2} \Delta \quad \text{and let } \Delta' = c_3 \Delta. \quad (5.2.2)$$

The algorithm generates sequences $\{s_k, k \in \mathbb{N}\}$ and $\{x_k, k \in \mathbb{N}\}$ such that

$$s_k = x_{k+1} - x_k, \quad (5.2.3)$$

for all $k \in \mathbb{N}$. Since $\{x_k, k \in \mathbb{N}\}$ is contained in X_0 , we have

$$\|s_k\| \leq 2 \sup \{ \|x\| / x \in X_0 \}$$

or

$$\|s_k\| \leq \Delta \quad (5.2.4)$$

for all k in $\mathbb{N}$. The algorithm also generates a sequence of positive scalars $\{\delta_k, k \in \mathbb{N}\}$ satisfying

$$0 < \delta_{k+1} \leq \max(\delta_j, c_3 \|s_j\|) \quad (5.2.5)$$

for some $j \leq k$. The inequalities (5.2.4), (5.2.5) and Definition (5.2.2) imply that

$$0 < \delta_k \leq \Delta' \quad (5.2.6)$$

for all k in $\mathbb{N}$.

THEOREM 5.2.1. *Let $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be continuously differentiable, and let f be defined by $f(x) = \|F(x)\|$, where $\|\cdot\|$ denotes an arbitrary norm on $\mathbb{R}^n$. Also let x_0 be any point in $\mathbb{R}^n$. Assume that the level set $\{x \in \mathbb{R}^n / f(x) \leq f(x_0)\}$ is bounded. Then any accumulation point of the*

sequence $\{x_k, k \in \mathbb{N}\}$ generated by the Algorithm 5.1 is a stationary point of the function f .

The proof of the theorem will require the use of the following lemma.

LEMMA 5.2.2. *Let the function $F: X_0 \rightarrow \mathbb{R}$ be continuously differentiable. Suppose that for some fixed integer k , the algorithm loops indefinitely between Step 2 and Step 1. Then x_k is necessarily a stationary point of f .*

Proof of the lemma. Let $\{s_j, j \in \mathbb{N}\}$ be the sequence generated by the algorithm by solving

$$\text{minimize } m_k(s) = \|F(x_k) + F'(x_k)s\| \quad (5.2.7a)$$

$$\text{subject to } \|s\| \leq \delta_k, \quad (5.2.7b)$$

where $\delta_{k(i+1)} \leq c_5 \|s_{k_i}\| \leq c_5 \delta_{k_i}$ for all j in $\mathbb{N}$. Observe that $\delta_{k_0} = \delta_k$ and $0 < c_5 < 1$ so the sequence $\{\|s_{k_j}\|, j \in \mathbb{N}\}$ is decreasing to zero. The test in step 2 fails for all j in $\mathbb{N}$. Thus we have

$$f(x_k + s_{k_j}) > f(x_k) + c_1 \gamma(x_k, s_{k_j}), \quad (5.2.8)$$

and by (5.1.1), i.e. $\gamma(x_k, s_{k_j}) \geq f'(x_k, s_{k_j})$

$$f(x_k + s_{k_j}) > f(x_k) + c_1 f'(x_k, s_{k_j}). \quad (5.2.9)$$

On the other hand, the hypothesis of continuous differentiability of F gives the equality

$$F(x_k + \varepsilon_k) = F(x_k) + F'(x_k)\varepsilon_k + o(\|\varepsilon_k\|), \quad (5.2.10)$$

so we have

$$f(x_k + \varepsilon_k) \leq m_k(\varepsilon_k) + o(\|\varepsilon_k\|). \quad (5.2.11)$$

Inequalities (5.2.9) and (5.2.11) imply that we can write

$$m_k(\varepsilon_k) + o(\|\varepsilon_k\|) > m_k(0) + c_1 f'(x_k, \varepsilon_k). \quad (5.2.12)$$

The sequence $\{d_j = \frac{\varepsilon_k}{\|\varepsilon_k\|}, j \in N\}$ is bounded; therefore there exists a

subsequence $\{d_j, j \in N' \subset N\}$ convergent to some d_* with $\|d_*\| = 1$.

Inequality (5.2.12) implies

$$\frac{m_k(\varepsilon_k) - m_k(0)}{\|\varepsilon_k\|} + \frac{o(\|\varepsilon_k\|)}{\|\varepsilon_k\|} > c_1 f'(x_k, d_j). \quad (5.2.13)$$

Let us set $\|\varepsilon_k\| = t_j$; then $\{t_j, j \in N'\}$ is decreasing to zero. We can rewrite

(5.2.13) as

$$\frac{m_k(t_j d_j) - m_k(0)}{t_j} + \frac{o(t_j)}{t_j} > c_1 f'(x_k, d_j) \quad (5.2.14)$$

or

$$\begin{aligned} \frac{m_k(t_j d_*) - m_k(0)}{t_j} + \frac{m_k(t_j d_j) - m_k(t_j d_*)}{t_j} \\ + \frac{o(t_j)}{t_j} > c_1 f'(x_k, d_j). \end{aligned} \quad (5.2.15)$$

We know, since t_j is decreasing to zero, that

$$\lim_{j \rightarrow +\infty} \frac{m_k(t_j d_*) - m_k(0)}{t_j} = m_k'(0; d_*) \quad (5.2.16)$$

$$\lim_{j \rightarrow +\infty} \frac{o(t_j)}{t_j} = 0. \quad (5.2.17)$$

$$\begin{aligned} \frac{|m_k(t_j d_j) - m_k(t_j d_*)|}{t_j} &= \frac{\|F_k + F_k' d_j\| - \|F_k + t_j F_k' d_*\|}{t_j} \\ &\leq \|F_k'\| \|d_j - d_*\|; \end{aligned}$$

where $F_k = F(x_k)$ and $F_k' = F'(x_k)$; consequently

$$\lim_{j \rightarrow +\infty} \frac{m_k(t_j d_j) - m_k(t_j d_*)}{t_j} = 0. \quad (5.2.18)$$

Inequality (5.2.15), the limits (5.2.16), (5.2.17) and (5.2.18), and the fact that $m_k'(0; d_*) = f'(x_k, d_*)$ (see Theorem 3.1.1) imply that

$$f'(x_k, d_*) \geq c_1 f'(x_k, d_*) \quad (5.2.19)$$

or

$$(1 - c_1) f'(x_k, d_*) \geq 0. \quad (5.2.20)$$

Because $1 - c_1 > 0$ we have that

$$f'(x_k, d_*) \geq 0. \quad (5.2.21)$$

But, we also have $f'(x_k, d_j) < 0$ for all j in N' , which implies, since $f'(x_k, \cdot)$ is continuous, that

$$f'(x_k, d_*) \leq 0. \quad (5.2.22)$$

From inequalities (5.2.21) and (5.2.22), we obtain

$$f'(x_k, d_*) = 0. \quad (5.2.23)$$

Now, we prove that $f'(x_k, s) \geq 0$ for all s in $\mathbb{R}^n$. Let s be any vector of norm one (this is sufficient since $f'(x_k, \cdot)$ is positively homogeneous). Also let μ_j be a positive scalar such that

$$\|\mu_j s\| = \|s_{k_j}\|, \quad j \in \mathbb{N}, \quad (5.2.24)$$

and let

$$y_j = \mu_j s. \quad (5.2.25)$$

Because $\|y_j\| = \|s_{k_j}\|$ and $s_{k_j} \in \operatorname{argmin} \{m_k(s) / \|s\| \leq \delta_{k_j}\}$ we have

$$m_k(s_{k_j}) \leq m_k(y_j). \quad (5.2.26)$$

By Theorem 2.2.20, we have

$$f'(x_k, s_{k_j}) \leq m_k(s_{k_j}) - m_k(0)$$

which, with (5.2.26), implies

$$f'(x_k, s_{k_j}) \leq m_k(y_j) - m_k(0).$$

Moreover, because $\|s_{k_j}\| = \|y_j\|$, we have

$$f'(x_k; \frac{s_{k_j}}{\|s_{k_j}\|}) \leq \frac{m_k(y_j) - m_k(0)}{\|y_j\|}. \quad (5.2.27)$$

Let us rewrite inequality (5.2.27) in the following form:

$$f'(x_k; d_j) \leq \frac{m_k(t_j u_j) - m_k(0)}{t_j}, \quad (5.2.28)$$

where

$$d_j = \frac{s_{k_j}}{\|s_{k_j}\|}, \quad u_j = \frac{y_j}{\|y_j\|}, \quad t_j = \|y_j\|. \quad (5.2.29)$$

But $u_j = \frac{y_j}{\|y_j\|} = \frac{\mu_j s}{\|\mu_j s\|} = \frac{s}{\|s\|} = s$; so (5.2.28) becomes

$$f'(x_k, d_j) \leq \frac{m_k(t_j s) - m_k(0)}{t_j}. \quad (5.2.30)$$

The sequence $\{t_j = \|y_j\|, j \in \mathbb{N}\}$ is decreasing to zero since $\|y_j\| = \|s_{k_j}\|$ and $\{\|s_{k_j}\|, j \in \mathbb{N}\}$ is decreasing to zero. Inequality (5.2.30) implies that

$$\lim_{j \rightarrow +\infty} f'(x_k, d_j) \leq \lim_{j \rightarrow +\infty} \frac{m_k(t_j s) - m_k(0)}{t_j},$$

and because the function $d \mapsto f'(x_k; d)$ is continuous we have

$$f'(x_k; d_*) \leq f'(x_k; s). \quad (5.2.31)$$

Equality (5.2.23) and inequality (5.2.31) imply that

$$f'(x_k, s) \geq 0, \quad (5.2.32)$$

where s is any unitary vector. Since $f'(x_k, \cdot)$ is positively homogeneous, we have that x_k is a stationary point of the function f in problem

(5.0.1).

The idea behind Lemma 5.2.2 is that if for some fixed integer k the algorithm loops indefinitely between Step 2 and Step 1, then algorithm generates a sequence $\{\delta_k, j \in \mathbb{N}\}$ decreasing to zero. Moreover, if $s_k, j \in \mathbb{N}$, are the respective solutions of problem (5.1.4), then any accumulation point of the sequence $\{\frac{s_k}{\|s_k\|}, j \in \mathbb{N}\}$ is equal to the steepest descent direction, say d_* . Since the one-sided directional derivative at x_k in the direction d_* is nonnegative, there are no descent directions or equivalently x_k is a stationary point of f in problem (5.0.1).

Now, we return to our goal of proving Theorem 5.2.1.

Let us define the compact set E , the solution set P , the point-to-set A and the function h used in Zangwill's Theorem 2.1.11. The set E is defined by

$$E = \left\{ z = (x, \delta) \mid x \in X_0 \text{ and } \delta \in [0, \Delta] \right\},$$

which is compact. The solution set P consists essentially of points (x, δ) such that x is a stationary point of $f \|F\|$. We will say that $(x, \delta) \in P$ if and only if for all $\delta \in [0, \Delta]$, if $\|s\| \leq \delta$ then $f'(x; s) \geq 0$, i.e.

$$P = \left\{ z = (x, \delta) \in E \mid \forall \delta \in [0, \Delta], \|s\| \leq \delta \Rightarrow f'(x; s) \geq 0 \right\}.$$

We define the merit function h by

$$h(x, \delta) = f(x).$$

which is continuous. Finally, we define the point-to-set map A on E . For z in P we set $A(z) = z$ and for $z = (x, \delta)$ in $E - P$ we say that $z' = (x', \delta')$ is in $A(z)$ if there exists a positive scalar μ and a vector s in $\mathbb{R}^n$ such that:

$$(1) \quad \alpha < \mu \leq \delta$$

and

$$(2) \quad x' = x + s$$

where

$$(3) \quad s \in \operatorname{argmin} \{m_x(s) \mid \|s\| \leq \mu\}$$

$$(4) \quad f(x + s) \leq f(x) + c_1 \gamma(x, s),$$

and

$$(5) \quad \text{if } f(x + s) < f(x) + c_2[m_x(s) - m_x(0)],$$

$$\text{then } \|s\| \leq \delta' \leq \max(\delta, c_3 \|s\|),$$

$$\text{else, if } f(x + s) > f(x) + c_2[m_x(s) - m_x(0)],$$

$$\text{then } c_4 \|s\| \leq c_5 \|s\|,$$

$$\text{else, } c_4 \|s\| \leq \delta' \leq \max(\delta, c_3 \|s\|).$$

Now, let us establish that the point-to-set A that model our algorithm is closed on $E - P$.

First, we establish the condition (i) of Zangwill's Theorem 2.1.11. Let (x, δ) be any point in $E-P$. Then x is not stationary. By Lemma 5.2.2 we know that there exists a positive scalar μ such that $0 < \mu \leq \delta$ and a vector $s \neq 0$ such that

$$s \in \operatorname{argmin} \{m_x(s) \mid \|s\| \leq \mu\} \quad (5.2.33a)$$

and

$$f(x+s) \leq f(x) + c_1 \gamma(x; s), \quad (5.2.33b)$$

where $\gamma(x, s) \in [f(x; s), 0]$. If we set $x' = x + s$, the existence of δ' such that $(x', \delta') \in A(x, \delta)$ is obvious and this establishes that condition (i) holds.

Secondly, we establish condition (ii) of Theorem 2.1.11. Let (x, δ) be in $E-P$. It is obvious that for any $(x+s, \delta')$ in $A(x, \delta)$. We have $\gamma(x, s) < 0$ and consequently $f(x') < f(x)$ where $x' = x + s$, so condition (ii) holds. Finally, we establish the third condition of Theorem 2.1.11, i.e. the point-to-set map A is closed on $E-P$. Let (x, δ) be any point in $E-P$. Also let $\{(x_k, \delta_k), k \in \mathbb{N}\}$ be a sequence that converges to some (x, δ) , and let $\{(x'_k, \delta'_k) \in A(x_k, \delta_k), k \in \mathbb{N}\}$ be a sequence that converges to some (x', δ') . We want to establish that

$$(x', \delta') \in A(x, \delta). \quad (5.2.34)$$

By definition, $(x'_k, \delta'_k) \in A(x_k, \delta_k)$ implies that there exists a positive scalar μ_k and a vector s_k such that

$$(i) \quad 0 < \mu_k \leq \delta_k \quad (5.2.35a)$$

$$(ii) \quad s_k \in \operatorname{argmin} \{m_{x_k}(s) \mid \|s\| \leq \mu_k\} \quad (5.2.35b)$$

$$(iii) \quad x'_k = x_k + s_k \quad (5.2.35c)$$

$$(iv) \quad f(x_k + s_k) \leq f(x_k) + c_1 \gamma(x_k; s_k). \quad (5.2.35d)$$

Since the sequences $\{x'_k, k \in \mathbb{N}\}$ and $\{x_k, k \in \mathbb{N}\}$ converge to x' and x respectively, the sequence $\{s_k, k \in \mathbb{N}\}$, where s_k is defined by (5.2.35b) and (5.2.35c), converges to some s_* such that

$$s_* = x' - x, \quad (5.2.36)$$

and consequently it is a bounded sequence, say $\|s_k\| \leq \Delta$ for all $k \in \mathbb{N}$. Because of (5.2.35a), and the fact that the sequence $\{\delta_k, k \in \mathbb{N}\}$ converges, we have that the sequence $\{\mu_k, k \in \mathbb{N}\}$ is bounded (without loss of generality we can assume that $\mu_k \leq \Delta, k \in \mathbb{N}$), and therefore it has a subsequence, say $\{\mu_{k'}, k' \in \mathbb{N}'\}$, that converges to some μ_* . We establish that

$$s_* \in \operatorname{argmin} \{m_x(s) \mid \|s\| \leq \mu_*\}. \quad (5.2.37)$$

For that we will use Proposition 2.1.10. Let us set

$$\phi(s; x, \mu) = m_x(s) \quad (5.2.38a)$$

$$\{s \in \mathbb{R}^n \mid \|s\| \leq \mu\} = S(x, \mu) \quad (5.2.38b)$$

and rewrite (5.2.35b) as follows:

$$s_k \in \operatorname{argmin} \left\{ \phi(s; x_k, \mu_k) \mid s \in S(x_k, \mu_k) \right\}. \quad (5.2.39)$$

To apply Proposition 2.1.10, since ϕ is continuous, it is sufficient to establish that S is a continuous point-to-set map. Observe that the map S is the composition of the projection $pr_2(z, r) = r$ and the point-to-set map $B(r) = \{s \in \mathbb{R}^n \mid r - \|s\| \geq 0\}$, which by Proposition 2.1.9 is continuous. Proposition 2.1.8 allows us to conclude that the point-to-set map $S = B \circ pr_2$ is continuous. Therefore, the point-to-set map

$$\psi: (x, r) \rightarrow \operatorname{argmin} \left\{ \phi(s; z, \alpha) \mid s \in S(x, r) \right\}$$

is closed. We have that $\{(x_k, \mu_k), k \in N'\}$ converges to (x, μ_*) and $\{s_k \in \psi(x_k, \mu_k), k \in N'\}$ converges to s_* and because ψ is closed we obtain that s_* belongs to $\psi(x, \mu_*)$ or equivalently

$$s_* \in \operatorname{argmin} \left\{ m_x(s) \mid \|s\| \leq \mu_* \right\} \quad (5.2.40)$$

which is what we wanted to establish i.e. (5.2.37). Now since x is not stationary, s_* is not zero and consequently $\mu_* > 0$. Because of (5.2.35a) we have

$$0 < \mu_* \leq \delta. \quad (5.2.41)$$

The upper semi-continuity of the real function γ implies that the function g defined by

$$g(x, s) = f(x + s) - f(x) - c_1 \gamma(x, s) \quad (5.2.42)$$

is lower semi-continuous, and consequently its epigraph is closed. by (5.2.35d) we have that

$$(x_k, s_k; 0) \in \operatorname{epi}(g). \quad (5.2.43)$$

On the other hand, the sequence $\{(x_k, s_k; 0), k \in N\}$ converges to $(x, s_*; 0)$, and because of the closedness of $\operatorname{epi}(g)$ and (5.2.43), we conclude that

$$(x, s_*; 0) \in \operatorname{epi}(g), \quad (5.2.44a)$$

or equivalently

$$f(x + s_*) \leq f(x) + c_1 \gamma(x, s_*). \quad (5.2.44.b)$$

Properties (5.2.36), (5.2.37), (5.2.41) and (5.2.44b) establish the first four properties (out of five) needed to conclude that (x', δ') belongs to $A(x, \delta)$. Let us establish the fifth property. Suppose that

$$f(x + s_*) - f(x) - c_2 \left\{ m_x(s_*) - m_x(0) \right\} < 0. \quad (5.2.45)$$

The sequence $\{(x_k, s_k), k \in N\}$ converges to (x, s_*) , so for large k in N we have

$$f(x_k + s_k) - f(x_k) - c_2 \left\{ m_x(s_k) - m_k(0) \right\} < 0, \quad (5.2.46)$$

which gives

$$\|s_k\| \leq \delta_k' \leq c_3 \max(\delta_k, c_3 \|s_k\|). \quad (5.2.47)$$

When δ is not equal to $c_3||s_*||$ it is obvious that (5.2.47) implies that

$$||s_*|| \leq \delta' \leq c_3 \max(\delta, c_3||s_*||). \quad (5.2.48)$$

Suppose that δ is equal to $c_3||s_*||$, and let

$$N_1 = \left\{ k \in \mathbb{N} / \delta_k < c_3||s_k|| \right\}.$$

If N_1 is finite then there exists an integer k_0 such that for all $k \geq k_0$ we have

$\delta_k \geq c_3||s_k||$. Consequently, we obtain that

$$||s_k|| \leq \delta_k' \leq \delta_k \quad (5.2.49)$$

for all $k \geq k_0$, and therefore

$$||s_*|| \leq \delta' \leq \max(\delta, c_3||s_*||). \quad (5.2.50)$$

Suppose that N_1 is infinite, i.e. for all $j \in \mathbb{N}$, there exists $k_j \in \mathbb{N}$, such that $\delta_{k_j} < c_3||s_{k_j}||$. Then we have for $j \in \mathbb{N}$

$$||s_{k_j}|| \leq \delta_{k_j}' \leq c_3||s_{k_j}|| \quad (5.2.51)$$

and

$$\lim_{j \rightarrow +\infty} ||s_{k_j}|| \leq \lim_{j \rightarrow +\infty} \delta_{k_j}' \leq c_3 \lim_{j \rightarrow +\infty} ||s_{k_j}||$$

or

$$||s_*|| \leq \delta' \leq \max(\delta, c_3||s_*||). \quad (5.2.52)$$

Now suppose that

$$f(x + s_*) - f(x) - c_2 \left\{ m_x(s_*) m_x(0) \right\} > 0. \quad (5.2.53)$$

Because $\{(x_k, s_k), k \in \mathbb{N}\}$ converges to (x, s_*) and the function on the left in the inequality (5.2.53) is continuous, we have for sufficiently large k in $\mathbb{N}$ that

$$f(x_k + s_k) - f(x_k) - c_2 \left\{ m_k(s_k) - m_k(0) \right\} > 0, \quad (5.2.54)$$

and consequently, by the algorithm,

$$c_4||s_k|| \leq \delta_k' \leq c_5||s_k|| \quad (5.2.55)$$

for sufficiently large k , which implies that

$$(5.2.55)$$

$$c_4||s_*|| \leq \delta' \leq c_5||s_*||. \quad (5.2.56)$$

Finally if neither (5.2.45) nor (5.2.53) hold, then necessarily we have

$$f(x + s_*) - f(x) - c_2 \left\{ m_x(s_*) - m_x(0) \right\} = 0. \quad (5.2.57)$$

Let us denote by N' the set of integers k such that (5.2.46) holds for k . If N' is finite, then there is an integer k' such that for all $k \geq k'$, (5.2.54) and consequently (5.2.55) holds. Therefore by passing to the limit when $k \rightarrow +\infty$ we obtain

$$c_4||s_*|| \leq \delta' \leq c_5||s_*|| \quad (5.2.58)$$

and consequently

$$c_4 \|s_*\| \leq \delta' \leq \max(\delta, c_3 \|s_*\|). \quad (5.2.59)$$

Suppose that N' is infinite. Then for all j in N' , there exists $k_j \geq j, k_j \in N'$. By a similar argument used in the case where (5.2.46) holds for all k in N , we obtain that

$$\|s_*\| \leq \delta' \leq \max(\delta, c_3 \|s_*\|), \quad (5.2.59)$$

and consequently

$$c_4 \|s_*\| \leq \delta' \leq \max(\delta, c_3 \|s_*\|). \quad (5.2.60)$$

This establishes the fifth condition for (x', δ') to belong to $A(x, \delta)$. Consequently the point-to-set map A is closed on $E-P$, which ends the proof of the conditions of Zangwill's Theorem 2.1.11, and demonstrates Theorem 5.2.1. •

5.3. Condition of Convergence to a Solution of the Nonlinear System

In this section, we establish a mild condition which guarantees that any accumulation point of the sequence $\{x_k, k \in N\}$ generated by Algorithm 5.1 is actually a solution of the nonlinear system $F(x)=0$. We call x a nonsingular point if $F'(x)$ is nonsingular.

THEOREM 5.3.1. *Let $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be continuously differentiable. If the sequence $\{x_k, k \in N\}$ generated by Algorithm 5.1 has at least one nonsingular accumulation point x_* , then any accumulation point of the sequence $\{x_k, k \in N\}$*

is necessarily a solution of the nonlinear system $F(x)=0$.

To prove this theorem we will need the following lemma.

LEMMA 5.3.2. *Let $h: \mathbb{R}^n \rightarrow \mathbb{R}$ be continuous on $\mathbb{R}^n$. Also let $\{z_k, k \in N\}$ be a bounded sequence such that the sequence $\{h(z_k), k \in N\}$ is decreasing. Then the function h is constant on the set of accumulation points of $\{z_k, k \in N\}$.*

Proof of Lemma 5.3.2. Let z_* and z_*' be two accumulation points of the sequence $\{z_k, k \in N\}$. Then, there exist two subsequences $\{z_k, k \in N\}$ and $\{z_k, k \in N'\}$ that converge respectively to z_* and z_*' . We have that for every j in N , there exists k_j in N' such that

$$h(z_{k_j}) \leq h(z_j) \quad (5.3.1a)$$

$$k_j \geq j. \quad (5.3.1b)$$

Because h is continuous, the inequalities (5.3.1) imply that

$$h(z_*') \leq h(z_*). \quad (5.3.2)$$

Since the roles of z_*' and z_* in establishing (5.3.2) are symmetric, we may conclude that

$$h(z_*) \leq h(z_*'). \quad (5.3.3)$$

The two inequalities (5.3.2) and (5.3.3) imply that

$$h(z_*) = h(z_*'),$$

which establishes the lemma. •

Proof of Theorem 5.3.1. By Theorem 5.2.1 we have that x_* is a stationary point of $f = \|F\|$, and by Theorem 4.1.4 we have that x_* is a solution of the nonlinear system $F(x) = 0$, i.e.

$$F(x_*) = 0. \quad (5.3.5)$$

On the other hand the function $f = \|F\|$ is continuous, the sequence $\{x_k, k \in \mathbb{N}\}$ is bounded and the sequence $\{f(x_k), k \in \mathbb{N}\}$ is decreasing. Therefore f is constant on the set of accumulation points of $\{x_k, k \in \mathbb{N}\}$. Hence by (5.3.5) we have

$$F(\bar{x}) = 0 \quad (5.3.6)$$

for any accumulation point $\bar{x}$ of the sequence $\{x_k, k \in \mathbb{N}\}$. •

5.4. Rate of Convergence

In this section, we show that if the sequence $\{x_k, k \in \mathbb{N}\}$ generated by Algorithm 5.1 converges to x_* which is nonsingular, then the method, for large k , reduces to Newton's method and consequently the convergence of $\{x_k, k \in \mathbb{N}\}$ to x_* is q -quadratic.

THEOREM 5.4.1. *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be continuously differentiable. Suppose that the sequence $\{x_k, k \in \mathbb{N}\}$ generated by Algorithm 5.1 converges to a nonsingular point x_* . Also suppose that F' is Lipschitz near x_* . Then, for sufficiently large k , x_k is the Newton iterate for the nonlinear equation $F(x) = 0$, and consequently $\{x_k\}$ converges to x_* q -quadratically.*

Proof. To prove that the algorithm, for large k , is equivalent to Newton's method, it is sufficient to establish that the test

$$f(x_{k+1}) < f(x_k) + c_2 [m_k(s_k) - m_k(0)] \quad (5.4.1)$$

is satisfied for large k . Let us establish (5.4.1). Since F is continuously differentiable we have

$$f(x_k + s_k) = \|F(x_k) + F'(x_k)s_k + o(\|s_k\|)\|, \quad (5.4.2)$$

which implies

$$f(x_k) - f(x_k + s_k) \geq f(x_k) - m_k(s_k) - o(\|s_k\|). \quad (5.4.3)$$

On the other hand, since x_k is not a stationary point for all k in $\mathbb{N}$, Theorem 5.2.1 implies that

$$f(x_k) - m_k(s_k) > 0. \quad (5.4.4)$$

The two last inequalities give

$$\frac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)} \geq 1 - \frac{o(\|s_k\|)}{f(x_k) - m_k(s_k)}, \quad (5.4.5)$$

or equivalently

$$\frac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)} \geq 1 - \frac{o(\|s_k\|)}{\|s_k\|} \frac{\|s_k\|}{f(x_k) - m_k(s_k)}. \quad (5.4.6)$$

Let us show that the ratio

$$\frac{f(x_k) - m_k(s_k)}{\|s_k\|}$$

is bounded away from zero. Since $\{x_k, k \in \mathbb{N}\}$ converges to x_* , $F'(x_*)$ is nonsingular and F is continuously differentiable, there exists k_* in $\mathbb{N}$ such that $F'(x_k)$ is nonsingular for all $k \geq k_*$. We have two cases for $k \geq k_*$, either the Newton step s_k^N is inside the trust region or not. First suppose that $\|s_k^N\| \leq \delta_k$. Then we have $s_k = s_k^N$ and

$$f(x_k) - m_k(s_k) = \|F'(x_k)s_k^N\|. \quad (5.4.7)$$

Since the matrix $F'(\cdot)$ is nonsingular, there exists a positive constant λ_* such that

$$\|F'(x_*)d\|_a \geq 2\lambda_* \|d\|_b \quad (5.4.8)$$

for all d in $\mathbb{R}^n$ (λ_* depends on the norms being used in both sides of inequalities (5.4.8)). The subscripts a and b have been added to remind the reader that we are using different norms. Because $\{F'(x_k), k \in \mathbb{N}\}$ converges to $F'(x_*)$, there exists an integer (we take it equal to k_* for simplicity) such that

for all $k \geq k_*$ we have

$$\|F'(x_k)d\| \geq \lambda_* \|d\| \quad (5.4.9)$$

for all d in $\mathbb{R}^n$. Consequently (5.4.7) implies that

$$f(x_k) - m_k(s_k) \geq \lambda_* \|s_k\|$$

or equivalently

$$\frac{f(x_k) - m_k(s_k)}{\|s_k\|} \geq \lambda_*. \quad (5.4.10)$$

Now suppose that $\|s_k^N\| > \delta_k$. Let us set

$$\mu_k = \frac{\delta_k}{\|s_k^N\|} \in (0, 1) \quad (5.4.11a)$$

and

$$\bar{s}_k = \mu_k s_k^N. \quad (5.4.11b)$$

We have, since $\|\bar{s}_k\| = \delta_k$,

$$f(x_k) - m_k(\bar{s}_k) \leq f(x_k) - m_k(s_k) \quad (5.4.12)$$

and from Definition (5.4.11b)

$$\mu_k f(x_k) - m_k(\bar{s}_k) = \mu_k f(x_k). \quad (5.4.13)$$

Inequality (5.4.12) and equality (5.4.13) imply

$$\mu_k f(x_k) \leq f(x_k) - m_k(s_k). \quad (5.4.14)$$

which, together with $\|s_k\| \leq \|\bar{s}_k\|$, give

$$\frac{\mu_k f(x_k)}{\|s_k\|} \leq \frac{f(x_k) - m_k(s_k)}{\|s_k\|} \quad (5.4.15)$$

By (5.4.11b), (5.4.15) we have

$$\frac{f(x_k)}{\|s_k^N\|} \leq \frac{f(x_k) - m_k(s_k)}{\|s_k\|}; \quad (5.4.16)$$

or because $F(x_k) + F'(x_k)s_k^N = 0$

$$\frac{\|F'(x_k)s_k^N\|}{\|s_k^N\|} \leq \frac{f(x_k) - m_k(s_k)}{\|s_k\|} \quad (5.4.17)$$

Using inequality (5.4.9) we get

$$\lambda_* \leq \frac{f(x_k) - m_k(s_k)}{\|s_k\|} \quad (5.4.18)$$

Inequalities (5.4.10) and (5.4.18) can be used to show that

$$\lambda_* \leq \frac{f(x_k) - m_k(s_k)}{\|s_k\|} \quad (5.4.19)$$

for all $k \geq k_*$.

Property (5.4.19) and inequality (5.4.6) imply that for $k \geq k_*$ we have

$$\frac{f(x_k) - f(x_k + s_k)}{f(x_k) - m_k(s_k)} \geq 1 - \frac{1}{\lambda_*} \frac{o(\|s_k\|)}{\|s_k\|}$$

We have that there exists an integer k_* such that

$$1 - \frac{1}{\lambda_*} \frac{o(\|s_k\|)}{\|s_k\|} > c_2$$

for all $k \geq k_*$. Consequently, for $k \geq k_*$ we have

$$f(x_k) - f(x_k + s_k) > c_2 [f(x_k) - m_k(s_k)] \quad (5.4.20)$$

which is effectively (5.4.1). Observe that, because of (5.1.2), (5.4.1) or (5.4.20) above implies that for $k \geq k_*$ the solution of the problem

$$\begin{aligned} & \text{minimize} \quad \|F(x_k) + F'(x_k)s\|_a \\ & \text{subject to} \quad \|s\|_b \leq \delta_k \end{aligned}$$

is always an acceptable step in terms of Algorithm 5.1. Inequality (5.4.1) also implies that the trust region radius δ_k , for $k \geq k_*$, is updated according to the rule

$$\|s_k\| \leq \delta_{k+1} \leq \max(\delta_k, c_3 \|s_k\|). \quad (5.4.21)$$

Suppose that there exists an integer that we choose equal to k_* for simplicity such that

$$s_k \neq s_k^N, \quad (5.4.22)$$

for all $k \geq k_*$ where s_k^N denotes the Newton step. This implies that

$$\|s_k\| = \delta_k \quad (5.4.23)$$

for all $k \geq k_*$. Consequently the criterion (5.4.21) is equivalent to

$$\delta_k \leq \delta_{k+1} \leq \max(\delta_k, c_3 \delta_k) \quad (5.4.24)$$

which implies, because of (5.4.23), that

$$\|\delta_k\| \leq \|\delta_{k+1}\| \text{ for } k \geq k_*. \quad (5.4.25)$$

Inequality (5.4.25) contradicts the fact that $\{\delta_k = x_{k+1} - x_k, k \in \mathbb{N}\}$ converges to zero. Therefore for all j in $\mathbb{N}$ there exists an integer $k_j \geq j$ such that

$$\delta_{k_j} = \delta_{k_j}^N. \quad (5.4.26)$$

and consequently x_{k_j+1} is the Newton's iterate from x_{k_j} .

Let k' be the smallest integer, greater than k_* , such that (5.4.26) holds for $k' = k_j$, i.e.

$$\delta_{k'} = \delta_{k'}^N. \quad (5.4.27)$$

and such that $x_{k'}$ is in a sufficiently small neighborhood of x_* , say $N(x_*)$, where Newton's method generates steps such that

$$\|\delta_{k+1}^N\| \leq M \|\delta_k^N\|^2, \quad k \geq k' \quad (5.4.28)$$

for some positive constant M independent of k . We want to show that, starting from $x_{k'}$, our algorithm and Newton's method generate the same sequences $\{x_k, k \geq k'\}$. Because of (5.4.26), we can assume that $N(x_*)$ is such that

$$M \|\delta_{k'}^N\| \leq 1 \quad (5.4.29)$$

for all $k \geq k'$. Consequently, inequalities (5.4.29) and (5.4.28) imply that

$$\|\delta_{k+1}^N\| \leq \|\delta_k^N\|, \quad (5.4.30)$$

which, together with (5.4.27), gives

$$\|\delta_{k+1}^N\| \leq \|\delta_{k'}\|. \quad (5.4.31)$$

But, by (5.4.24) we have that $\|\delta_{k'}\| \leq \delta_{k+1}$ which, together with (5.4.31), implies that

$$\|\delta_{k+1}^N\| \leq \delta_{k+1}; \quad (5.4.32)$$

and consequently

$$\delta_{k+1} = \delta_{k+1}^N. \quad (5.4.33)$$

By induction, we establish that

$$\delta_k = \delta_k^N$$

for all $k \geq k'$. Consequently the sequence $\{x_k, k \geq k'\}$ generated by Algorithm 5.1 is the sequence of the Newton's iterates from $x_{k'}$, and hence, Algorithm 5.1 is asymptotically q -quadratically convergent.

REFERENCES

- ANDERSON, M.R. and M.R. OSBORNE [1977a]. Discrete nonlinear approximation problems in polyhedral norms, *Numer. Math.*, 28, pp. 143-156.
- ANDERSON, M.R. and M.R. OSBORNE [1977b]. Discrete nonlinear approximation problems in polyhedral norms: A Levenberg algorithm, *Numer. Math.*, 28, pp. 157-170.
- ARMJO, L. [1966]. Minimization of functions having Lipschitz-continuous first partial derivatives, *Pacific J. Math.*, 16, pp. 1-3.
- AUSLANDER, A. [1977]. Minimisation sans contraintes de fonctions localement Lipschitziennes, *Comptes Rend. Acad. Sc.*, Paris, 284, pp. 959-961.
- AUSLANDER, A. [1978]. Minimization de fonctions localement Lipschitziennes: Applications à la programmation mi-convex, mi-differentiable, *Nonlinear Programming 3*, O. Mangasarian, R. Meyer, S. Robinson, eds., Academic Press, New York, N.Y., pp. 429-459.
- BERGE, C. [1963]. *Topological Spaces*, translated by E.M. Patterson, Oliver & Boyd.
- BERGE, C. [1966]. *Espaces topologiques. Fonctions multivoques*, Dunod, Paris.
- CHARALAMBOUS, C. [1979]. On conditions for optimality of the nonlinear l_1 problem, *Math. Programming*, 17, pp. 123-135.
- CLARKE, F.H. [1973]. *Necessary Conditions for Nonsmooth Problems in Optimal Control and Calculus of Variations*, Ph.D. thesis, University of Washington.
- CLARKE, F.H. [1975]. Generalized gradients and applications, *Trans. Amer. Math. Soc.*, 205, pp. 247-262.
- CLARKE, F.H. [1976]. A new approach to Lagrange multipliers, *Math. Op. Res.*, 1, pp. 165-174.
- CLARKE, F.H. [1981]. Generalized gradients of Lipschitz functionals, *Adv. Math.*, 40, pp. 52-67.
- CLARKE, F.H. [1983]. *Optimization and Nonsmooth Analysis*, Canadian Math. Soc. Series of Monographs and Advanced Texts, John Wiley and Sons Publications.
- DANSKIN, J.M. [1967]. *The Theory of Max Min*, Springer-Verlag, New York.
- DANTZIG, G. FOLKMAN and SHAPIRO [1967]. On the continuity of the minimum set of a continuous function. *J. Math. Anal. Appl.*, 17, pp. 519-548.
- DELAHAYE, J.P. and J. DENEL [1978]. Equivalences des continuités des applications multivoques dans les espaces topologiques. *Laboratoire de Calcul, publication no. 111*, University of Lille I.
- DEM'YANOV, V.F. and V.N. MALOZEMOV [1971]. On the theory of nonlinear minimax problems, *Russian Math. Soc.*, 26, pp. 57-115.
- DEM'YANOV, V.F. and A.M. RUBINOV [1980]. On quasidifferentiable functionals, *Soviet Math. Dokl.*, 21, pp. 14-17.
- DENNEL, J. [1977]. On the continuity of point-to-set map with applications to optimization, *Proceedings of the 2nd Symposium on Operations Research*, Aachen.
- EL-ATTAR, R.A., M. VIDYASAGAR and S.R.K. DUTTA [1979]. An algorithm for l_1 -norm minimization with applications to nonlinear l_1 -approximation. *SIAM J. Num. Anal.*, 6, pp. 70-86.
- DENNIS, J.E. Jr. [1977]. Nonlinear least squares and equations, in *The State of the Art in Numerical Analysis*, D. Jacobs, ed., Academic Press, London, pp. 269-312.
- DENNIS, J.E. Jr., D.M. GAY and R.E. WELSCH [1981a]. An adaptive nonlinear least-squares algorithm, *TOMS*, 7, pp. 348-368.
- DENNIS, J.E. Jr., D.M. GAY and R.E. WELSCH [1981b]. Algorithm 573 NL2SOL - An adaptive nonlinear least-squares algorithm [E4], *TOMS*, 7, pp. 369-383.
- DENNIS, J.E. Jr. and R.B. SCHNABEL [1983]. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, New Jersey.

- DENNIS, J.E. Jr. and R.A. TAPIA [1976]. Supplementary terminology for nonlinear iterative methods, *SIGNUM Newsletter*, 11, no.4, pp. 4-6.
- DUFF, I.S., J. NOCEDAL and J.K. REID [1984]. The use of linear programming for the solution of sparse sets of nonlinear equations, Department of Electrical Engineering and Computer Science Report CSS 145, Northwestern University, Evanston, Illinois 60201.
- FLETCHER, R. [1971]. A modified Marquardt subroutine for nonlinear least squares, A.E.R.E. Harwell Report TP 476.
- FLETCHER, R. [1980]. *Practical Methods of Optimization, vol. 1, Unconstrained Optimization*, John Wiley and Sons, New York.
- FLETCHER, R. [1981]. *Practical Methods of Optimization, vol. 2, Constrained Optimization*, John Wiley and Sons, Ltd., Chichester.
- HALD, J. and K. MADSEN [1985]. Combined LP and quasi-Newton methods for nonlinear t_1 optimization, *SIAM J. Numer. Anal.*, 22, pp. 68-80.
- HAN, S.P. [1978]. Superlinear convergence of a minimax method. Report TR78-336, Department of Computer Science, Cornell University.
- HAN, S.P. [1981]. Variable metric methods for minimizing a class of nondifferentiable functions, *Math. Programming*, 20, pp. 1-13.
- HEBDEN, M.D. [1973]. An algorithm for minimization using exact second derivatives, Technical report TP515, Atomic Energy Research Establishment, Harwell, England.
- HIRIART-URRUTY, J.B. [1979a]. Tangent cones, generalized gradients and mathematical programming in Banach spaces, *Math. Op. Research*, 4, pp. 79-97.
- HIRIART-URRUTY, J.B. [1979b]. Refinements of necessary optimality conditions in nondifferentiable programming I. *Appl. Math. Optim.*, 5, pp. 63-82.
- HIRIART-URRUTY, J.B. [1981]. A better insight into the generalized gradient of the absolute value of a function. *Appl. Anal.*, 12, pp. 239-249.
- HIRIART-URRUTY, J.B. [1982]. Refinements of necessary optimality conditions in nondifferentiable programming II. *Math. Programming Study*, 19, pp.

120-139.

- HOGAN, W.W. [1973a]. Point-to-set maps in mathematical programming. *SIAM Review*, 3, pp. 591-603.
- HOGAN, W.W. [1973b]. The continuity of the perturbation function of a convex program. *Operat. Research*, 21, pp. 351-352.
- HUARD, P. [1972a]. *Optimisation dans $\mathbb{R}^n$, 1^{ère} partie: Eléments Théoriques*. Laboratoire de Calcul, University of Lille I.
- HUARD, P. [1972b]. *Optimisation dans $\mathbb{R}^n$, 2^{ème} partie: Algorithmes Généraux*. Laboratoire de Calcul, University of Lille I.
- HUARD, P. [1974]. Tentative de synthèse dans les méthodes de programmation non-linéaire. *Cahiers du Centre d'études de R.O.*, no. 16, pp. 347-367.
- HUARD, P. [1975]. Optimization and point-to-set maps. *Math. Programming*, 8, pp. 308-331.
- HUARD, P. [1979]. *Point-to set maps and mathematical programming*, *Math. Programming Study* 10, North Holland Publishing, Amsterdam.
- IOFFE, A.D., and V.L. LEVIN [1972]. Subdifferentials of convex functions, *Trans. Moscow Math. Soc.*, 26, pp. 1-72.
- LEVENBERG, K. [1944]. A method for the solution of certain problems in least squares, *Quart. Appl. Math.*, 2, pp. 164-168.
- LEMARECHAL, C. [1978]. Nonlinear programming and nonsmooth optimization. Report de Recherche No. 332, INRIA, Le Chesnay, France.
- MADSEN, K. [1975]. An algorithm for minimax solution of overdetermined systems of non-linear equations, *IMA J. Numer. Anal.*, 16, pp. 321-328.
- MADSEN, K. and H. SCHJAER-JACOBSEN [1978]. Linearly constrained minimax optimization, *Math. Programming*, 14, pp. 208-223.
- MARQUARDT, D. [1963]. An algorithm for least-squares estimation of nonlinear parameters, *SIAM J. Appl. Math.*, 11, pp. 431-441.

- MEYER, R.R. [1970]. The validity of a family of optimization methods, *SIAM J. Control*, 1, pp. 41-54.
- MIFLIN, R. [1977a]. An algorithm for constrained optimization with semi-smooth functions. *Math. Op. Res.*, vol. 2, no. 2, pp. 191-207.
- MIFLIN, R. [1977b]. Semismooth and semi-convex functions in constrained optimization. *SIAM J. Cont. Optim.*, 15, pp. 959-972.
- MORE, J.J. [1977]. The Levenberg-Marquardt algorithm: implementation and theory, *Numerical Analysis*, G.A. Watson, ed., pp. 105-116, Lecture Notes in Mathematics 630, Springer-Verlag, Berlin, Heidelberg and New York.
- MORE, J.J. and D.C. SORENSEN [1982]. Newton's method, Argonne National Labs Report ANL-82-8, Argonne, Illinois.
- MORE, J.J. and D.C. SORENSEN [1983]. Computing a trust region step, *SIAM J. Sci. Statist. Comput.*, 4, pp. 553-572.
- ORTEGA, J.M. and W.C. RHEINBOLDT [1970]. *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.
- OSBORNE, M.R. [1976]. Nonlinear least squares - the Levenberg algorithm revisited, *J. Austral. Math. Soc. Ser. B*, 19, pp. 343-357.
- OSBORNE, M.R. and G.A. WATSON [1969]. An algorithm for minimax approximation in the non-linear case, *Comput. J.*, 12, pp. 64-69.
- OSBORNE, M.R. and G.A. WATSON [1978]. Nonlinear approximation problems in vector norms, *Numerical Analysis*, Dundee, *Lectures Notes in Mathematics*, no. 630, ed. G.A. Watson, Springer-Verlag, Berlin.
- POWELL, M.J.D. [1970]. A hybrid method for nonlinear equations, in *Numerical Methods for Nonlinear Algebraic Equations*, P. Rabinowitz, ed., Gordon and Breach, London, pp. 87-114.
- POWELL, M.J.D. [1983]. General algorithms for discrete nonlinear approximation calculations, Department of Applied Mathematics and Theoretical Physics Report DAMTP 1983/NA2, University of Cambridge, Cambridge, England.

- POWELL, M.J.D. and Y. YUAN [1983]. Conditions for superlinear convergence in l_1 and l_∞ solutions of overdetermined nonlinear equations, Department of Applied Mathematics and Theoretical Physics DAMTP 1983/NA8, University of Cambridge, Cambridge, England.
- PSHENICHNYI, B.N. [1971]. *Necessary Conditions for an Extremum*, Dekker, New York.
- REID, J.K. [1973]. Least squares solution of sparse systems of non-linear equations by a modified Marquardt algorithm, in *Proceedings of the NATO Conf. at Cambridge, July 1972*, North Holland, Amsterdam, pp. 437-445.
- ROBERTS, A.W. and D.E. VANBERG [1974]. Another proof that convex functions are locally Lipschitz, *Am. Math. Mon.*, 81, pp. 1014-1016.
- ROBINSON, S.M. and R.R. MEYER [1973]. Lower semi-continuity of multivalued linearization mappings, *SIAM J. Control*, 3, pp. 525-533.
- ROCKAFELLAR, R.T. [1966]. Characterization of the subdifferentials of convex functions, *Pacific J. Math.*, 17, pp. 497-510.
- ROCKAFELLAR, R.T. [1970]. *Convex Analysis*, Princeton Mathematics Services, No. 28, Princeton University.
- ROCKAFELLAR, R.T. [1976]. Lagrange multipliers in optimization, *SIAM-AMS Proceedings*, 9, Cottle, R.W. and Lemke, C.E., eds., pp. 145-168.
- ROCKAFELLAR, R.T. [1979]. *La Théorie des Sous-Gradients et Ses Applications à l'optimisation: Fonctions Convexes et Non Convexes*, Collection Chaire Aisenstadt, Presses de l'université de Montréal, 130 pp.
- ROCKAFELLAR, R.T. [1980]. Generalized directional derivatives and subgradients of nonconvex functions, *Canadian J. Math.*, 32, pp. 157-180.
- ROCKAFELLAR, R.T. [1981]. *The Theory of Subgradients and Its Applications to Problems of Optimization: Convex and Nonconvex Functions*, Holderman Verlag, Berlin.
- SORENSEN, D.C. [1982]. Newton's method with a model trust region modification, *SIAM J. Numer. Anal.*, 19, pp. 409-426.

- SPRINGARN, J.E. [1981]. Submonotone subdifferentials of Lipschitz functions. *Trans. Amer. Math. Soc.*, 264, pp. 77-89.
- THIBAUT, L. [1982]. Sudifferentials of nonconvex vector-valued functions, *J. Math. Anal. Appl.*, 86, pp. 319-344.
- WATSON, G.A. [1979]. The minimax solution of an overdetermined system of nonlinear equations, *Journal Instit. Math. Appl.*, 23, pp. 167-180.
- WOLFE, P. [1969]. Convergence conditions for ascent methods, *SIAM Review*, 11, pp. 226-235.
- WOLFE, P. [1971]. Convergence conditions for ascent methods II: some corrections, *SIAM Review*, 13, pp. 185-188.
- YUAN, Y. [1983a]. Global convergence of trust region algorithms for nonsmooth optimization, Department of Applied Mathematics and Theoretical Physics Report DAMTP 1983/NA13, University of Cambridge, Cambridge, England.
- YUAN, Y. [1983b]. Some properties of trust region algorithms for nonsmooth optimization, Department of Applied Mathematics and Theoretical Physics Report DAMTP 1983/NA4, University of Cambridge, Cambridge, England.
- ZANGWILL, W.I. [1969]. *Nonlinear Programming. A Unified Approach*. Prentice Hall.