In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institute shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

7/25/68

VARIATIONS ON NEWTON'S METHOD IN FINITE DIMENSIONAL SPACES

A THESIS

Presented to

the Faculty of the Graduate Division

by

Robert Lind Horton

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Mathematics

Georgia Institute of Technology

September, 1969

VARIATIONS ON NEWTON'S METHOD

IN FINITE DIMENSIONAL SPACES

Approved:

Chairman

Date approved by Chairman: 8/29/69

ACKNOWLEDGMENTS

I would like to express my gratitude to all who assisted in the preparation of this thesis. Special thanks must go to Dr. W. J. Kammerer, my thesis advisor, whose guidance, patience, and criticisms were indispensable. I am also indebted to Dr. R. H. Kasriel and Dr. D. L. Finn for their reading of this work.

In addition, my appreciation goes to Georgia Tech and the National Science Foundation for a Traineeship during the 1968-69 academic year.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS ii

Chapter

I. INTRODUCTION. 1

II. NEWTON'S METHOD 19

III. DIFFICULTIES OF NEWTON'S METHOD 29

Point Substitution Method
Secant, Wolfe's, and Barnes' Method
Broyden's Method
Freudenstein and Roth's Method

BIBLIOGRAPHY 60

CHAPTER I

INTRODUCTION

In this thesis, we will be dealing with functions whose domains are subsets of R^n, the set of n-tuples $\{(x_1,x_2,\ldots,x_n)^T\}$ whose components are real numbers, and whose ranges are also finite dimensional. Although much of what we consider can be extended to more abstract spaces, the algorithms to be discussed are placed in finite dimensional settings for simplicity. A vector in R^n will be denoted with a bar, $\bar{x}$, to differentiate it from a real number. Also the uniform norm will be used throughout, i.e. $\|\bar{x}\| = \max_{1\le i\le n} \{|x_i|\}$. If A is an $m \times n$ matrix with components a_{ij}, then $\|A\| = \max_{1\le i\le m} \{\sum_{j=1}^{n} |a_{ij}|\}$. Chapter I is a review of notation, terminology, and theorems useful in dealing with functions defined on finite dimensional spaces.

Whenever f is a linear function, there are many methods for solving $f(\bar{x}) = \bar{0}$ using computing machinery. However, when f is nonlinear, the number of algorithms dwindles while their complexity generally increases. One such method, Newton's method, has gained a wide following because it is simple, easy to use and understand, and its results are usually quite satisfactory. Newton's method is described and a convergence proof is presented in Chapter II.

It should be pointed out that there are major drawbacks to this method. One is the necessity of calculating an inverse of f' at each iteration, a time-consuming if not impossible task especially for a

large or complicated system. Another difficulty is the necessity of choosing an initial approximation within a suitable neighborhood of the true solution. Variations on Newton's method which deal with these difficulties are discussed in Chapter III.

Let Ω be an open set in R^n, and f a mapping such that $f: \Omega \to R^m$. Then for $\bar{x} \in \Omega$,

$$f(\bar{x}) = \begin{bmatrix} f_1(x_1,x_2,\ldots,x_n) \\ f_2(x_1,x_2,\ldots,x_n) \\ \vdots \\ f_m(x_1,x_2,\ldots,x_n) \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \bar{y}.$$

The functions $y_i = f_i(\bar{x})$, $i=1,2,\ldots,m$, are called coordinate functions associated with f.

Definition 1.1 A function $f: \Omega \to R^m$ is said to be differentiable at the point $\bar{x} \in \Omega$ if and only if there exists a linear mapping $A: R^n \to R^m$ such that

$$\lim_{\bar{h}\to\bar{0}} = \frac{\|f(\bar{x}+\bar{h}) - f(\bar{x}) - A\bar{h}\|}{\|\bar{h}\|} = 0.$$

The linear mapping A is said to be the derivative of f at $\bar{x}$, and one writes $f'(\bar{x}) = A$. If f is differentiable at each $\bar{x} \in \Omega$ then f is said to be differentiable on Ω.

Theorem 1.1 If Ω is an open set in R^n and $f: \Omega \to R^m$ is differentiable at $\bar{x} \in \Omega$, then $f'(\bar{x})$ is unique.

Proof: Suppose $f'(\bar{x}) = A_1$ and $f'(\bar{x}) = A_2$, then we have

$$\lim_{\bar{h}\to\bar{o}} \frac{\|f(\bar{x}+\bar{h}) - f(\bar{x}) - A_1\bar{h}\|}{\|\bar{h}\|} = \lim_{\bar{h}\to\bar{o}} \frac{\|f(\bar{x}+\bar{h}) - f(\bar{x}) - A_2\bar{h}\|}{\|\bar{h}\|} = 0.$$

The inequality

$$\|(A_1-A_2)\bar{h}\| \leq \|f(\bar{x}+\bar{h}) - f(\bar{x}) - A_1\bar{h}\| + \|f(\bar{x}+\bar{h}) - f(\bar{x}) - A_2\bar{h}\|$$

shows that

$$\lim_{\bar{h}\to\bar{o}} \frac{\|(A_1-A_2)\bar{h}\|}{\|\bar{h}\|} = 0.$$

Therefore, for an arbitrary fixed $\bar{h} \neq \bar{o}$, we have

$$\lim_{t\to o} \frac{\|(A_1-A_2)t\bar{h}\|}{\|t\bar{h}\|} = \|(A_1-A_2)\frac{\bar{h}}{\|\bar{h}\|}\| = 0$$

implying that $(A_1-A_2)\bar{h} = \bar{o}$ for all $\bar{h} \in R^n$.

We shall now develop a relationship between the derivative of f at $\bar{x}$ and the partial derivatives of f at $\bar{x}$. Let Ω be an open subset of

R^n and $f: \Omega \to R^m$ be a differentiable function at $\bar{x} \in \Omega$. Then there exists a linear transformation $A: R^n \to R^m$, which has a matrix representation

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

such that

$$\lim_{\bar{h} \to \bar{o}} \frac{\|f(\bar{x}+\bar{h}) - f(\bar{x}) - A\bar{h}\|}{\|\bar{h}\|} = 0.$$

If $\bar{h} = [t,o,\ldots,o]^T$ and $\bar{d}_1 = [1,o,\ldots,o]^T$, then

$$\lim_{t \downarrow o} \frac{\|f(\bar{x}+\bar{h})-f(\bar{x})-A\bar{h}\|}{\|\bar{h}\|} = \lim_{t \downarrow o} \left\| \frac{1}{t} \begin{bmatrix} f_1(x_1+t,x_2,\ldots,x_n)-f_1(\bar{x}) \\ f_2(x_1+t,x_2,\ldots,x_n)-f_2(\bar{x}) \\ \vdots \\ f_m(x_1+t,x_2,\ldots,x_n)-f_m(\bar{x}) \end{bmatrix} - \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} \right\| = 0.$$

Likewise

$$\lim_{t \uparrow o} \left\| \frac{f(\bar{x}+\bar{h}) - f(\bar{x})}{-t} + A\bar{\delta}_1 \right\| = 0,$$

implying that the partial derivatives with respect to the first

coordinate of each of the coordinate functions f_i exist and

$$D_1 f_i(\bar{x}) = a_{i1} \qquad \text{for } i=1,\ldots,m.$$

A similar argument will show that $a_{ij} = D_j f_i(\bar{x})$ for $i=1,\ldots,m$, $j=1,\ldots,n$.

Thus the matrix representation of $f'(\bar{x})$ is

$$f'(\bar{x}) = A = \begin{bmatrix} D_1 f_1(\bar{x}) & D_2 f_1(\bar{x}) & \cdots & D_n f_1(\bar{x}) \\ D_1 f_2(\bar{x}) & D_2 f_2(\bar{x}) & \cdots & D_n f_2(\bar{x}) \\ \vdots & \vdots & & \vdots \\ D_1 f_m(\bar{x}) & D_2 f_m(\bar{x}) & \cdots & D_n f_m(\bar{x}) \end{bmatrix}$$

which will be called the Jacobian matrix of f' at $\bar{x}$.

Definition 1.2 The differential of f at $\bar{x}$, denoted by $df[\bar{x};\bar{h}]$, is defined to be,

$$df[\bar{x};\bar{h}] = \begin{bmatrix} D_1 f_1(\bar{x}) & D_2 f_1(\bar{x}) & \cdots & D_n f_1(\bar{x}) \\ D_1 f_2(\bar{x}) & D_2 f_2(\bar{x}) & \cdots & D_n f_2(\bar{x}) \\ \vdots & \vdots & & \vdots \\ D_1 f_m(\bar{x}) & D_2 f_m(\bar{x}) & \cdots & D_n f_m(\bar{x}) \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix}$$

whenever the indicated partial derivatives exist. If $f'(\bar{x})$ exists, then $df[\bar{x};\bar{h}] = f'(\bar{x})\bar{h}$, and $\lim_{\bar{h}\to\bar{o}} \|f(\bar{x}+\bar{h}) - f(\bar{x}) - df[\bar{x};\bar{h}]\| = 0$.

Theorem 1.2 Let Ω be an open subset of R^n, and let $f: \Omega \to R^m$ be differentiable at $\bar{x} \in \Omega$. Then f is continuous at $\bar{x}$.

Proof. Since $f'(\bar{x})$ is a linear mapping from R^n into R^m, there exists a constant $M \geq 0$ such that

$$\|f'(\bar{x})\bar{h}\| \leq M\|\bar{h}\| \quad \text{for all } \bar{h} \in R^n.$$

Now given $\varepsilon > 0$, there exists a δ, with $\varepsilon > \delta > 0$, such that

$$\|f(\bar{x}+\bar{h}) - f(\bar{x}) - f'(\bar{x})\bar{h}\| < \varepsilon \qquad \text{whenever } \|\bar{h}\| < \delta.$$

Thus

$$\|f(\bar{x}+\bar{h}) - f(\bar{x})\| < \|f'(\bar{x})\bar{h}\| + \varepsilon \leq M\|\bar{h}\| + \varepsilon < (M+1)\,\varepsilon \qquad \text{whenever } \|\bar{h}\| < \delta$$

So

$$\lim_{\bar{h}\to\bar{0}} \|f(\bar{x}+\bar{h}) - f(\bar{x})\| = 0.$$

Theorem 1.3 (Mean Value Theorem.) Let Ω be an open subset of R^n and $f: \Omega \to R^m$. If f is differentiable on the line segment $\lambda\bar{y} + (1-\lambda)\bar{x} \in \Omega$ for $0 \leq \lambda \leq 1$, $\bar{y}, \bar{x} \in \Omega$, then there exists m points $\bar{x}_1, \ldots, \bar{x}_m$ on this line segment such that $f(\bar{y}) - f(\bar{x}) = A(\bar{y}-\bar{x})$, where A is the linear mapping represented by

$$A = \begin{bmatrix} D_1 f_1(\bar{x}_1) & D_2 f_1(\bar{x}_1) & \cdots & D_n f_1(\bar{x}_1) \\ D_1 f_2(\bar{x}_2) & D_2 f_2(\bar{x}_2) & \cdots & D_n f_2(\bar{x}_2) \\ \vdots & \vdots & & \vdots \\ D_1 f_m(\bar{x}_m) & D_2 f_m(\bar{x}_m) & \cdots & D_n f_m(\bar{x}_m) \end{bmatrix}.$$

Proof. If $h_1(\lambda) = f_1(\lambda\bar{y} + (1-\lambda)\bar{x})$ for $o \leq \lambda \leq 1$, then h_1 is a continuous real valued function of one variable for which the standard mean value theorem holds. Thus there exists a λ_1 in $(o,1)$ such that

$$h(1) - h(o) = h'(\lambda_1)$$

or equivalently

$$f_1(\bar{y}) - f_1(\bar{x}) = D_1 f_1(\bar{x}_1)(y_1 - x_1) + D_2 f_1(\bar{x}_1)(y_2 - x_2) + \cdots + D_n f_1(\bar{x}_1)(y_n - x_n)$$

where $\bar{x}_1 = \lambda_1\bar{y} + (1-\lambda_1)\bar{x}$.

A similar argument applied to each of the coordinate functions will show the existence of $\bar{x}_i = \lambda_i\bar{y} + (1-\lambda_i)\bar{x}$, $o < \lambda_i < 1$, such that

$$f_i(\bar{y}) - f_i(\bar{x}) = D_1 f_i(\bar{x}_i)(y_1 - x_1) + D_2 f_i(\bar{x}_i)(y_2 - x_2) + \cdots + D_n f_i(\bar{x}_i)(y_n - x_n)$$

for $i = 1, \ldots, m$.

Theorem 1.4 (Chain Rule Theorem.) Let Ω be an open set in R^n, and g map Ω into R^k. If g is differentiable at $\bar{x}_o \in \Omega$ and f maps an open set containing $g[\Omega]$ into R^m and f is differentiable at $g(\bar{x}_o)$, then the composite mapping $F = f(g)$ of Ω into R^m is differentiable at $\bar{x}_o$ and

$$F'(\bar{x}_o) = f'(g(\bar{x}_o))g'(\bar{x}_o).$$

Proof. The mapping $f'(g(\bar{x}_o))g'(\bar{x}_o)$ from Ω to R^m is linear. It remains to be shown that

$$\lim_{\bar{x}\to\bar{x}_o} \frac{\|F(\bar{x}) - F(\bar{x}_o) - [f'(g(\bar{x}_o))g'(\bar{x}_o)][\bar{x} - \bar{x}_o]\|}{\|\bar{x} - \bar{x}_o\|} = 0.$$

If

$$r(\bar{x}) = F(\bar{x}) - F(\bar{x}_o) - [f'(g(\bar{x}_o))g'(\bar{x}_o)][\bar{x}-\bar{x}_o]$$

$$= f(g(\bar{x})) - f(g(\bar{x}_o)) - f'(g(\bar{x}_o))[g(\bar{x}) - g(\bar{x}_o)]$$

$$+ f'(g(\bar{x}_o))[g(\bar{x}) - g(\bar{x}_o) - g'(\bar{x}_o)[\bar{x}-\bar{x}_o]]$$

and $A = f'(g(\bar{x}_o))$, $B = g'(\bar{x}_o)$, then

$$\|r(\bar{x})\| \le \|f(g(\bar{x})) - f(g(\bar{x}_o)) - A[g(\bar{x}) - g(\bar{x}_o)]\|$$

$$+ \|A[g(\bar{x}) - g(\bar{x}_o) - B[\bar{x}-\bar{x}_o]]\|.$$

Given $\varepsilon > o$, there exists a $\delta_1 > o$ and a $\delta_2 > o$ such that

(i) $$\|f(g(\bar{x})) - f(g(\bar{x}_o)) - A[g(\bar{x}) - g(\bar{x}_o)]\| < \varepsilon\|g(\bar{x}) - g(\bar{x}_o)\|$$

whenever $\|g(\bar{x}) - g(\bar{x}_o)\| < \delta_1$, and

(ii) $$\|g(\bar{x}) - g(\bar{x}_o) - B[\bar{x} - \bar{x}_o]\| < \varepsilon\|\bar{x} - \bar{x}_o\|$$

and

$$\|g(\bar{x}) - g(\bar{x}_o)\| < \delta_1$$

whenever $\|\bar{x} - \bar{x}_o\| < \delta_2$.

So

$$\|f(g(\bar{x})) - f(g(\bar{x}_o)) - A[g(\bar{x}) - g(\bar{x}_o)]\| < \varepsilon\|g(\bar{x}) - g(\bar{x}_o)\|$$

$$= \varepsilon\|g(\bar{x}) - g(\bar{x}_o) - B[\bar{x} - \bar{x}_o] + B[\bar{x} - \bar{x}_o]\|$$

$$\leq \varepsilon^2\|\bar{x} - \bar{x}_o\| + \varepsilon\|B\| \, \|\bar{x} - \bar{x}_o\|$$

whenever $\|\bar{x} - \bar{x}_o\| < \delta_2$, and

$$\|A[g(\bar{x}) - g(\bar{x}_o) - B[\bar{x} - \bar{x}_o]]\| \leq \varepsilon \, \|A\| \, \|\bar{x} - \bar{x}_o\|$$

whenever $\|\bar{x} - \bar{x}_o\| < \delta_2$, which imply that

$$\frac{\|r(\bar{x})\|}{\|\bar{x}-\bar{x}_o\|} < \varepsilon^2 + \varepsilon\ (\|B\| + \|A\|)$$

whenever $\|\bar{x}-\bar{x}_o\| < \delta_2$.

Thus it has been shown that

$$\lim_{\bar{x}\to\bar{x}_o} \frac{\|r(\bar{x})\|}{\|\bar{x}-\bar{x}_o\|} = 0.$$

As an example of the chain rule theorem, let $g: \Omega \to R^1$ be given by $g(x_1,x_2) = \ln(x_1+x_2)$ where $\Omega = \{(x_1,x_2)^T: x_1 + x_2 > 0\}$. Let $f: R^1 \to R^3$ be defined by $f(t) = [t,t^2,t^3]^T$. Then the function

$$F(x_1,x_2) = f(g(x_1,x_2)) = [\ln(x_1+x_2),(\ln(x_1+x_2))^2,(\ln(x_1+x_2))^3]^T$$

has a derivative given by

$$F'(x_1,x_2) = \begin{bmatrix} \frac{1}{x_1+x_2} & \frac{1}{x_1+x_2} \\ \frac{2\ln(x_1+x_2)}{(x_1+x_2)} & \frac{2\ln(x_1+x_2)}{(x_1+x_2)} \\ \frac{3(\ln(x_1+x_2))^2}{(x_1+x_2)} & \frac{3(\ln(x_1+x_2))^2}{(x_1+x_2)} \end{bmatrix}.$$

As proven in the theorem, the derivative of F is also given by

$$F'(x_1,x_2) = f'(g(x_1,x_2))g'(x_1,x_2)$$

$$= \begin{bmatrix} 1 \\ 2\ \ln(x_1+x_2) \\ 3(\ln(x_1+x_2))^2 \end{bmatrix} \begin{bmatrix} \frac{1}{x_1+x_2}, & \frac{1}{x_1+x_2} \end{bmatrix}.$$

Definition 1.3 Let Ω be an open subset of R^n and $f: \Omega \to R^m$. Then f is said to be continuously differentiable on Ω if and only if the partial derivatives, $D_j f_i(\bar{x})$ for i=1,...,m and j=1,...,n exist and are continuous on Ω, and we write $f \in C'[\Omega]$. As Rudin [17, p. 192] shows, an equivalent definition would require that f be a continuous mapping of Ω into the space of linear mappings of R^n into R^m.

Theorem 1.5 Let Ω be an open subset of R^n and $f: \Omega \to R^m$. If f is continuously differentiable on Ω, then f is differentiable on Ω.

Proof. Let $\bar{x}_o \in \Omega$ and $\varepsilon > o$ be given. Since $f \in C'[\Omega]$, there exists a $\delta > o$ such that $|D_j f_i(\bar{x}) - D_j f_i(\bar{x}_o)| < \frac{\varepsilon}{n}$ for i=1,...,m and j=1,...,n whenever $\bar{x} \in S$ where $S = \{\bar{x}: \|\bar{x}-\bar{x}_o\| < \delta\} \subset \Omega$. Also by the Mean Value Theorem there exists a sequence $\{\bar{x}_i\}$, i=1,...,n, in Ω such that $f_i(\bar{x}_o+\bar{h}) - f_i(\bar{x}_o) = D_1 f_i(\bar{x}_i)h_1 + D_2 f_i(\bar{x}_i)h_2 + \cdots + D_n f_i(\bar{x}_i)h_n$ where $\bar{x}_o + \bar{h} \in \Omega$. Then we have

$$\|f(\bar{x}_o+\bar{h}) - f(\bar{x}_o) - f'(\bar{x}_o)\bar{h}\| =$$

$$\max_{1\le i\le m} \left[\sum_{j=1}^{n} |(D_j f_i(\bar{x}_i) - D_j f_i(\bar{x}_o))h_i| \right] < \varepsilon \|\bar{h}\|$$

whenever $\|\bar{h}\| < \delta$ and $\bar{x}_o + \bar{h} \in \Omega$, which implies that

$$\lim_{\bar{h}\to\bar{o}} \frac{\|f(\bar{x}_o+\bar{h}) - f(\bar{x}_o) - f'(\bar{x}_o)\bar{h}\|}{\|\bar{h}\|} = 0.$$

Definition 1.4 Let Ω be an open subset of R^n and f map Ω into R^m. Then f is said to be locally one to one on Ω if and only if about any point $\bar{x} \in \Omega$ there exists a neighborhood of $\bar{x}$ in which f is one to one.

As an example, the function $f(x) = x^2$ is not locally one to one on R^1; however, if zero is deleted from its domain, then f is locally one to one on its domain.

Definition 1.5 Let f be a mapping from R^n into R^n. Then f^{-1} is said to be the inverse of f if and only if f^{-1} is a function that maps $f[R^n]$ onto R^n where $f[R^n]$ R^n such that $f^{-1}(f(\bar{x})) = \bar{x}$ for all $\bar{x} \in R^n$.

Theorem 1.6 Let Ω be an open subset of R^n and f map Ω into R^n. If $f \in C'[\Omega]$, $f'(\bar{x}_o)$ is nonsingular at some $\bar{x}_o \in \Omega$ and $\bar{y}_o = f(\bar{x}_o)$, then there exist open sets U R^n and V R^n such that $\bar{x}_o \in U$ and $\bar{y}_o \in V$, f is one to one on U, and f[U] = V. Moreover if f^{-1} is the inverse of f defined on V, then $f^{-1} \in C'[V]$ and $[f^{-1}(f(\bar{x}_o))]' f'(\bar{x}_o) = I$.

Proof. The proof may be found in Rudin [17], p. 193.

As an illustration of Theorem 1.6, let $\Omega = \{\bar{x} \in R^2: x_1 > o\}$ and

$$f(x_1,x_2) = \begin{bmatrix} x_1^2 \\ \dfrac{x_2}{x_1} \end{bmatrix}.$$

Then f is continuously differentiable on Ω, and its derivative is given by

$$f'(x_1,x_2) = \begin{bmatrix} 2x_1 & 0 \\ -\dfrac{x_2}{(x_1)^2} & \dfrac{1}{x_1} \end{bmatrix}.$$

Since the determinant of $f'(x_1,x_2)$ is nonzero for any $(x_1,x_2)^T \varepsilon \Omega$, Theorem 1.6 applies, and we can find

$$f^{-1}(y_1,y_2) = \begin{bmatrix} \sqrt{y_1} \\ y_2\sqrt{y_1} \end{bmatrix}$$

Since f^{-1} is continuously differentiable,

$$[f^{-1}(y)]' = \begin{bmatrix} \dfrac{1}{2\sqrt{y_1}} & 0 \\ \dfrac{y_2}{2\sqrt{y_1}} & \sqrt{y_1} \end{bmatrix}.$$

Now then

$$[f^{-1}(f(x_1,x_2))]' = \begin{bmatrix} \frac{1}{2x_1} & 0 \\ \frac{x_2}{2x_1^2} & x_1 \end{bmatrix}$$

and

$$[f^{-1}(f(x_1,x_2))]'f'(x_1,x_2) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

as the theorem implies. Recall that a bilinear operator C(x,y) that maps $R^n \times R^n$ into R^m is a mapping which is linear in $\bar{x}$ for each $\bar{y} \in R^n$ and in $\bar{y}$ for each $\bar{x} \in R^n$.

Definition 1.5 The second derivative $f''(\bar{x})$ of a mapping $f: R^n \to R^m$ at a point $\bar{x} \in R^n$ is defined to be a bilinear operator from $R^n \times R^n$ into R^m such that

$$\lim_{\bar{k}\to\bar{o}} \left\| \frac{f'(\bar{x}+\bar{k})\bar{h} - f'(\bar{x})\bar{h}}{\|\bar{k}\|} - \frac{f''(\bar{x},\bar{k},\bar{h})}{\|\bar{k}\|} \right\| = 0$$

where $f''(\bar{x})$ is of the form, $f''(\bar{x}) = [B_1(\bar{x}),B_2(\bar{x}),\ldots,B_m(\bar{x})]^T$ and $f''(\bar{x},\bar{k},\bar{h}) = [\bar{k}^T B_1(\bar{x})\bar{h}, \bar{k}^T B_2(\bar{x})\bar{h},\ldots,\bar{k}^T B_m(\bar{x})\bar{h}]^T$ with

$$B_i(\bar{x}) = \begin{bmatrix} b^i_{11}(\bar{x}) & \cdots & b^i_{1n}(x) \\ b^i_{12}(\bar{x}) & \cdots & b^i_{2n}(\bar{x}) \\ \vdots & & \vdots \\ b^i_{1n}(\bar{x}) & \cdots & b^i_{nn}(\bar{x}) \end{bmatrix} \quad \text{for } i=1,\ldots,m .$$

Thus,

$$(1.1)\qquad \lim_{\bar{k}\to\bar{o}} \left\| \frac{f'(\bar{x}+\bar{k})\bar{h}-f'(\bar{x})\bar{h}}{\|\bar{k}\|} - \frac{1}{\|\bar{k}\|}\,[\bar{k}^T B_1\bar{h},\bar{k}^T B_2\bar{h},\ldots,\bar{k}^T B_m\bar{h}]^T \right\| = 0$$

must hold for every $\bar{k},\bar{h} \in R^n$.

Let $\bar{k} = [t,o,\ldots,o]^T \in R^n$. Then (1.1) becomes

$$\lim_{t\to o} \left\| \frac{1}{|t|} \begin{bmatrix} [D_1f_1(\bar{x}+\bar{k}) - D_1f_1(\bar{x})]h_1 + [D_2f_1(\bar{x}+\bar{k}) - D_2f_1(\bar{x})]h_2 + \ldots + [D_nf_1(\bar{x}+\bar{k}) - D_nf_1(\bar{x})]h_n \\ [D_1f_2(\bar{x}+\bar{k}) - D_1f_2(\bar{x})]h_1 + [D_2f_2(\bar{x}+\bar{k}) - D_2f_2(\bar{x})]h_2 + \ldots + [D_nf_2(\bar{x}+\bar{k}) - D_nf_2(\bar{x})]h_n \\ \vdots \\ [D_1f_m(\bar{x}+\bar{k}) - D_1f_m(\bar{x})]h_1 + [D_2f_m(\bar{x}+\bar{k}) - D_2f_m(\bar{x})]h_2 + \ldots + [D_nf_m(\bar{x}+\bar{k}) - D_nf_m(\bar{x})]h_n \end{bmatrix} - \begin{bmatrix} b^1_{11}h_1 + b^1_{12}h_2 + \ldots + b^1_{1n}h_n \\ b^2_{11}h_1 + b^2_{12}h_2 + \ldots + b^2_{1n}h_n \\ \vdots \\ b^m_{11}h_1 + b^m_{12}h_2 + \ldots + b^m_{1n}h_n \end{bmatrix} \right\| = 0$$

By taking the limit as $t\downarrow o$ and as $t\uparrow o$, we see that

$$\begin{aligned} b^1_{11} &= D_{11}f_1(\bar{x}) & b^2_{11} &= D_{11}f_2(\bar{x}) & b^m_{11} &= D_{11}f_m(\bar{x}) \\ b^1_{12} &= D_{21}f_1(\bar{x}) & b^2_{12} &= D_{21}f_2(\bar{x}) & b^m_{12} &= D_{21}f_m(\bar{x}) \\ &\vdots & &\vdots & &\vdots \\ b^1_{1n} &= D_{n1}f_1(\bar{x}) & b^2_{1n} &= D_{n1}f_2(\bar{x}) & b^m_{1n} &= D_{n1}f_m(\bar{x}). \end{aligned}$$

In the same manner it can be shown that

$$B_i = \begin{bmatrix} D_{11}f_i(\bar{x}) & D_{21}f_i(\bar{x}) & \cdots & D_{n1}f_i(\bar{x}) \\ D_{12}f_i(\bar{x}) & D_{22}f_i(\bar{x}) & \cdots & D_{n2}f_i(\bar{x}) \\ \vdots & & & \\ D_{1n}f_i(\bar{x}) & D_{2n}f_i(\bar{x}) & & D_{nn}f_i(\bar{x}) \end{bmatrix} \quad \text{for } i=1,\ldots,m.$$

We shall define a norm on the second derivative in the following manner

$$\|f''(\bar{x})\| = \max_{1\le k\le m} \sum_{i=1}^{n} \sum_{j=1}^{n} |b_{ij}^{k}| \ .$$

Although this norm is not the operator norm, it has the property that

$$\|f''(\bar{x},\bar{h},\bar{k})\| \le \|f''(\bar{x})\| \ \|\bar{h}\| \ \|\bar{k}\| \ .$$

Theorem 1.7 (Taylor's Theorem) Let f be a mapping from R^n into R^m. If f and all of its second partial derivatives exist on the closed region $T = \{\bar{x} \in R^n: a_1 \le x_1 \le c_1, a_2 \le x_2 \le c_2, \ldots, a_n \le x_n \le c_n\}$, then

$$f(\bar{x}+\bar{h}) = f(\bar{x}) + f'(\bar{x})\bar{h} + \frac{1}{2} B(\bar{\xi},\bar{h},\bar{h})$$

whenever $\bar{x}$ and $\bar{x} + \bar{h}$ are in T, and $B(\bar{\xi},\bar{h},\bar{h})$ is given by

$$B(\bar{\xi},\bar{h},\bar{h}) = \begin{bmatrix} \bar{h}^T & B_1 & \bar{h} \\ \bar{h}^T & B_2 & \bar{h} \\ & \vdots & \\ \bar{h}^T & B_m & \bar{h} \end{bmatrix}$$

with

$$Bi = \begin{bmatrix} D_{11}f_i(\bar{x}+\xi_i\bar{h}) & D_{21}f_i(\bar{x}+\xi_i\bar{h}) & \cdots & D_{n1}f_i(\bar{x}+\xi_i\bar{h}) \\ D_{12}f_i(\bar{x}+\xi_i\bar{h}) & D_{22}f_i(\bar{x}+\xi_i\bar{h}) & \cdots & D_{n2}f_i(\bar{x}+\xi_i\bar{h}) \\ \vdots & \vdots & & \vdots \\ D_{1n}f_i(\bar{x}+\xi_i\bar{h}) & D_{2n}f_i(\bar{x}+\xi_i\bar{h}) & \cdots & D_{nn}f_i(\bar{x}+\xi_i\bar{h}) \end{bmatrix}$$

where $o < \xi_i < 1$ for $i=1,\ldots,m$.

Proof. Let $\phi_i(t) = f_i(x_1+h_1t, x_2+h_2t, \ldots, x_n+h_nt)$ where f_i is a coordinate function associated with f.

Then using Taylor's Theorem for a function of one variable on $\phi_i(t)$ we have

$$\phi_i(1) = \phi_i(0) + \phi_i'(0) + \frac{1}{2}\phi_i''(\xi_i) \quad \text{with} \quad 0 < \xi_i < 1$$

which is the same as

$$f_i(\bar{x}+\bar{h}) = f_i(\bar{x}) + \sum_{k=1}^{n} D_k f_i(\bar{x})h_k + \frac{1}{2}\left[\sum_{k=1}^{n} D_{kk}f_i(\bar{x}+\xi_i\bar{h})h_k^2 + 2\sum_{\substack{2\le k\le n \\ 1\le j<k}} D_{kj}f_i(\bar{x}+\xi_i\bar{h})h_kh_j\right].$$

Thus we have $f(\bar{x}+\bar{h}) = f(\bar{x}) + f'(\bar{x})\bar{h} + \frac{1}{2} B(\bar{\xi},\bar{h},\bar{h})$.

Definition 1.6 If $g(\bar{x})$: $R^n \to R^m$ is given by $g(\bar{x}) = [g_1(\bar{x}), g_2(\bar{x}), \ldots, g_m(\bar{x})]^T$, then $D_j g(\bar{x})$ is defined to be

$$D_j g(\bar{x}) = [D_j g_1(\bar{x}),\ D_j g_2(\bar{x}), \ldots, D_j g_m(\bar{x})]^T \quad \text{for} \quad j=1,2,\ldots,n.$$

Definition 1.7 If

$$J(\bar{x}) = \begin{bmatrix} a_{11}(\bar{x}) & a_{12}(\bar{x}) & \cdots & a_{1n}(\bar{x}) \\ a_{21}(\bar{x}) & a_{22}(\bar{x}) & \cdots & a_{2n}(\bar{x}) \\ \vdots & \vdots & & \vdots \\ a_{m1}(\bar{x}) & a_{m2}(\bar{x}) & \cdots & a_{mn}(\bar{x}) \end{bmatrix}$$

where $\bar{x} \in R^n$ and $a_{ij}(\bar{x})$: $R^n \to R^1$ for $i=1,\ldots,m$, $j=1,\ldots,n$, then $D_j J(\bar{x})$ is defined to be

$$D_j J(\bar{x}) = \begin{bmatrix} D_j a_{11}(\bar{x}) & D_j a_{12}(\bar{x}) & \cdots & D_j a_{1n}(\bar{x}) \\ D_j a_{21}(\bar{x}) & D_j a_{22}(\bar{x}) & \cdots & D_j a_{2n}(\bar{x}) \\ \vdots & \vdots & & \vdots \\ D_j a_{m1}(\bar{x}) & D_j a_{m2}(\bar{x}) & \cdots & D_j a_{mn}(\bar{x}) \end{bmatrix}.$$

CHAPTER II

NEWTON'S METHOD

In this chapter Newton's method is motivated and described. A proof of the convergence of this method is provided, and the convergence is shown to be quadratic. The reader can find most of these results in Isaacson and Keller [9].

Let $\bar{g}(\bar{x}) = [g_1(\bar{x}), g_2(\bar{x}), \ldots, g_n(\bar{x})]^T$ be a mapping of a subset of R^n into R^n.

Definition 2.1 A vector $\bar{x}$ is said to be a fixed point of $\bar{g}$ if and only if $\bar{g}(\bar{x}) = \bar{x}$.

Theorem 2.1 Let $\bar{\alpha}$ be a fixed point of $\bar{g}$, and let $\bar{g}'(\bar{x})$ exist and satisfy $\|\bar{g}'(\bar{x})\| \leq \lambda < 1$ for all $\bar{x}$ in the sphere $S(\bar{\alpha},\rho) = \{\bar{x}: \|\bar{x}-\bar{\alpha}\| \leq \rho\}$. Then

(i) $\bar{\alpha}$ is the unique fixed point of $\bar{g}$ in $S(\bar{\alpha},\rho)$ and

(ii) for any initial estimate $\bar{x}^0 \in S(\bar{\alpha},\rho)$, the iterates $\{\bar{x}^n\}$, where $\bar{x}^{k+1} = \bar{g}(\bar{x}^k)$ for $k=0,1,2,\ldots$, converge to $\bar{\alpha}$.

Proof. Using Theorem 1.3, we have

$$\bar{g}(\bar{x}) = \bar{g}(\bar{y}) + A(\bar{x}-\bar{y}) \quad \text{for} \quad \bar{x},\bar{y} \in S(\bar{\alpha},\rho)$$

where A is the $n \times n$ matrix mentioned in Theorem 1.3. Note that $\|A\| \leq \lambda$. Then

(2.1) $$\|g(\bar{x}) - g(\bar{y})\| \leq \|A\| \, \|\bar{x}-\bar{y}\| \leq \lambda\|\bar{x}-\bar{y}\|$$
for $\bar{x},\bar{y} \in S(\bar{\alpha},\rho)$.

Now let $\bar{y} = \bar{\alpha}$ and we have

(2.2) $$\|g(\bar{x}) - \bar{\alpha}\| \leq \lambda\|\bar{x}-\bar{\alpha}\| \quad \text{for } \bar{x} \in S(\bar{\alpha},\rho).$$

If $\bar{x} = \bar{x}^0$, then $\|g(\bar{x}^0)-\bar{\alpha}\| = \|\bar{x}^1-\bar{\alpha}\| \leq \lambda\|\bar{x}^0-\bar{\alpha}\| \leq \lambda\rho$. If $\bar{x} = \bar{x}^1$, then $\|g(\bar{x}^1)-\bar{\alpha}\| = \|\bar{x}^2-\bar{\alpha}\| \leq \lambda\|\bar{x}^1-\bar{\alpha}\| \leq \lambda^2\|\bar{x}^0-\bar{\alpha}\| \leq \lambda^2\rho$, and by the obvious induction argument we have

$$\|\bar{x}^k-\bar{\alpha}\| \leq \lambda^k\rho \quad \text{for } k=0,1,2,\ldots$$

Notice that $\|\bar{x}^k-\bar{\alpha}\| \leq \|\bar{x}^{k-1}-\bar{\alpha}\| \leq \cdots \leq \|\bar{x}^0-\bar{\alpha}\| \leq \rho$. So $\bar{x}^k \in S(\bar{\alpha},\rho)$ for $k=0,1,2,\ldots$ Also since $\|\bar{x}^k-\bar{\alpha}\| \leq \lambda^k\rho$ and $\lambda < 1$, we have $\lim_{k\to\infty} \bar{x}^k = \bar{\alpha}$. So (ii) holds.

For uniqueness, let $\bar{\beta}$ be another fixed point of g in $S(\bar{\alpha},\rho)$. Then $\|\bar{\alpha}-\bar{\beta}\| = \|g(\bar{\alpha})-g(\bar{\beta})\| \leq \lambda\|\bar{\alpha}-\bar{\beta}\| < \|\bar{\alpha}-\bar{\beta}\|$ which is a contradiction.

Definition 2.2 Let $\{\bar{x}^k\}$ be a sequence in R^n that converges to $\bar{\alpha} \in R^n$. Then the sequence is said to be rth order convergent if $\|\bar{x}^{k+1}-\bar{\alpha}\| \leq M\|\bar{x}^k-\bar{\alpha}\|^r$ for $k=0,1,2,\ldots$ where M and r are positive real numbers. An iterative scheme $\{\bar{x}^{k+1} = g(\bar{x}^k), k=0,1,2,\ldots\}$ is said to have rth order convergence at a fixed point $\bar{\alpha} = g(\bar{\alpha})$ if there exists a neighborhood S of $\bar{\alpha}$ such that for each initial value $\bar{x}^0 \in S$, the iteration $\bar{x}^{k+1} = g(\bar{x}^k)$, $k=0,1,2,\ldots$ is rth order convergent to $\bar{\alpha}$.

By (2.2), the iteration examined in Theorem 2.1 is first order convergent. An alternate terminology is that it converges linearly. Let us now determine conditions on g which will provide a faster rate of convergence.

Theorem 2.2 Let $\bar{\alpha}$ be a fixed point of g, and let $g''(\bar{x})$ exist and be bounded on $S(\bar{\alpha},\rho)$. If $g'(\bar{\alpha}) = 0$ and $\bar{x}^0 \in S(\bar{\alpha},\rho)$, then the iterates $\{\bar{x}^k\}$ given by $\bar{x}^{k+1} = g(\bar{x}^k)$ for k=0,1,2,... are 2nd order convergent. (Second order convergence is often called quadratic convergence.)

Proof. Expanding $g(\bar{x})$ about $\bar{\alpha}$, using Taylor's theorem, gives

$$g(\bar{x}) = g(\bar{a}) + B(\bar{\xi},\bar{x}-\bar{\alpha},\bar{x}-\bar{\alpha}).$$

Note that $\|B(\bar{x})\|$ is bounded by say M for $\bar{x} \in S(\bar{\alpha},\rho)$. If g is evaluated at $\bar{x}^k$, the result is $g(\bar{x}^k) - g(\bar{\alpha}) = B(\bar{\xi}^k,\bar{x}^k-\bar{\alpha},\bar{x}^k-\bar{\alpha})$. So

$$\|\bar{x}^{k+1}-\bar{\alpha}\| \leq \|B(\bar{\xi}^k)\|\|\bar{x}^k-\bar{\alpha}\|^2 \leq M\|\bar{x}^k-\bar{\alpha}\|^2, \quad \text{for } k=0,1,2,...$$

Now let us concentrate on solving $f(\bar{x}) = \bar{o}$, where $f(\bar{x})$ is a mapping from a subset of R^n into R^n. Rewrite this system as a fixed point problem with $g(\bar{x}) = \bar{x} - A(\bar{x})f(\bar{x})$, where $A(\bar{x})$ is an n×n matrix with components $a_{ij}(\bar{x})$. In addition, $A(\bar{x})$ must be nonsingular so that the solutions of the two systems will coincide. The simplest choice for $A(\bar{x})$ is A, a constant nonsingular nth order matrix.

However, from Theorems 2.1 and 2.2, we know that the algorithm $\bar{x}^{k+1} = \bar{x}^k - Af(\bar{x}^k)$ will converge quadratically whenever $\bar{x}^o$ is "close enough" to $\bar{\alpha}, g'(\bar{\alpha}) = 0$, and g satisfies the hypotheses of Theorem 2.2. Since differentiating $g(\bar{x}) = \bar{x} - Af(\bar{x})$ gives $g'(\bar{x}) = I - Af'(\bar{x})$, $g'(\bar{\alpha})$ will equal zero if $f'(\bar{\alpha})$ is nonsingular and $A = [f'(\bar{\alpha})]^{-1}$. In practice however the solution $\bar{\alpha}$ is generally not available, so we let $A(\bar{x}) = [f'(\bar{x})]^{-1}$ for each iteration. Thus A now depends on $\bar{x}$. This is Newton's method. A recap of the procedure follows.

Newton's Method: To solve $f(\bar{x}) = \bar{o}$, choose an initial approximation $\bar{x}^o$ to the root. Compute successive iterations using the formula $\bar{x}^{k+1} = g(\bar{x}^k) = \bar{x}^k - [f'(\bar{x}^k)]^{-1} f(\bar{x}^k)$ for k=0,1,2,... Normally, instead of computing $[f'(\bar{x}^k)]^{-1}$, the equation is rearranged to read $f'(\bar{x}^k)(\bar{x}^{k+1}-\bar{x}^k) = -f(\bar{x}^k)$. This linear system is then solved for the correction vector $\bar{\rho}^k = \bar{x}^{k+1} - \bar{x}^k$ so that $\bar{x}^{k+1} = \bar{\rho}^k + \bar{x}^k$.

Consider the following 2×2 illustration of Newton's method. Let $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and $f(\bar{x}) = \begin{bmatrix} f_1(x_1,x_2) \\ f_2(x_1,x_2) \end{bmatrix}$. Then $f'(\bar{x}) = \begin{bmatrix} D_1f_1(\bar{x}) & D_2f_1(\bar{x}) \\ D_1f_2(\bar{x}) & D_2f_2(\bar{x}) \end{bmatrix}$. The iterates are given by $f'(\bar{x}^k)(\bar{x}^{k+1}-\bar{x}^k) = -f(\bar{x}^k)$, which when written out in component form becomes,

$$\begin{bmatrix} D_1f_1(x_1^k,x_2^k) & D_2f_1(x_1^k,x_2^k) \\ D_1f_2(x_1^k,x_2^k) & D_2f_2(x_1^k,x_2^k) \end{bmatrix} \begin{bmatrix} x_1^{k+1} - x_1^k \\ x_2^{k+1} - x_2^k \end{bmatrix} = - \begin{bmatrix} f_1(x_1^k,x_2^k) \\ f_2(x_1^k,x_2^k) \end{bmatrix}.$$

Thus we have

$$(2.3)\quad \begin{cases} (x_1^{k+1}-x_1^k)D_1f_1(x_1^k,x_2^k) + (x_2^{k+1}-x_2^k)D_2f_1(x_1^k,x_2^k) + f_1(x_1^k,x_2^k) = 0 \\ (x_1^{k+1}-x_1^k)D_1f_2(x_1^k,x_2^k) + (x_2^{k+1}-x_2^k)D_2f_2(x_1^k,x_2^k) + f_2(x_1^k,x_2^k) = 0 \end{cases}$$

for $k=0,1,2,\ldots$

Geometrically speaking in the one dimensional case, Newton's method approximates the graph of the function f with the tangent line to f at $\bar{x}^k$. The next iterate is then the zero of this tangent line. In two dimensions, as in the previous illustration, the equation $z = (x_1-x_1^k)D_1f_1(x_1^k,x_2^k) + (x_2-x_2^k)D_2f_1(x_1^k,x_2^k) + f_1(x_1^k,x_2^k)$ denotes the tangent plane to the surface $z = f_1(x_1,x_2)$ at the point $(x_1^k,x_2^k,f_1(x_1^k,x_2^k))$. So the solution to system (2.3) represents the intersection of the tangent planes to $z = f_1(x_1,x_2)$ at $(x_1^k,x_2^k,f_1(x_1^k,x_2^k))$ and to $z = f_2(x_1,x_2)$ at $(x_1^k,x_2^k,f_2(x_1^k,x_2^k))$ in the $x_1x_2(z=o)$ plane. Thus in two dimensions, tangent planes instead of tangent lines are used.

Theorem 2.3 Let f map a subset of R^n into R^n and $\bar{\alpha}$ be a fixed point of g where $g(\bar{x}) = \bar{x} - [f'(\bar{x})]^{-1}f(\bar{x})$. Now if $\|g'(\bar{x})\| \leq \lambda < 1$ and $g''(\bar{x})$ exists and is bounded for $\bar{x}$ in $S(\bar{\alpha},\rho)$ where $\rho \geq \|\bar{\alpha}-\bar{x}^o\|$, then Newton's method converges quadratically.

Proof. Theorem 2.1 implies that the iterates converge. If it can be shown that $g'(\bar{\alpha}) = \bar{o}$, then Theorem 2.2 will give second order convergence. The jth column of $g'(\bar{x})$ is given by

$$D_j g(\bar{x}) = \bar{\delta}_j - f'(\bar{x})^{-1} D_j f(\bar{x}) - [D_j f'(\bar{x})^{-1}] f(\bar{x})$$

where

$$\bar{\delta}_j = [o,\ldots,o,1_j,o,\ldots,o]^T \quad \text{and} \quad D_j f'(\bar{x})^{-1} = -f'(\bar{x})^{-1}[D_j f'(\bar{x})] f'(\bar{x})^{-1}.$$

Then we have

$$D_j g(\bar{\alpha}) = \bar{\delta}_j - f'(\bar{\alpha})^{-1} D_j f(\bar{\alpha}) - [D_j f'(\bar{\alpha})^{-1}] f(\bar{\alpha})$$

$$= \bar{\delta}_j - \bar{\delta}_j = \bar{o} \quad \text{for } j=1,2,\ldots,n.$$

It has been shown that if $\bar{x}^o$ is "close enough" to $\bar{\alpha}$, so that $\|g'(\bar{x})\| \leq \lambda < 1$ for $\bar{x} \in S(\bar{\alpha},\rho)$ where $\rho = \|\bar{\alpha}-\bar{x}^o\|$, then Newton's method converges. If in addition $f'(\bar{x})$ is nonsingular at $\bar{\alpha}$ and differentiable, the convergence is of second order. However we need a sufficient condition for the convergence of Newton's method when the root $\bar{\alpha}$ is unknown.

Theorem 2.4 Let f be a mapping from an open subset of R^n into R^n, and the initial approximation $\bar{x}^o$ be such that $f'(\bar{x}^o)$ has an inverse with norm bounded by

$$\|[f'(\bar{x}^o)]^{-1}\| \leq a. \tag{2.4}$$

If the difference of the first two iterates is bounded by

(2.5) $$\|\bar{x}^1-\bar{x}^0\| = \|[f'(\bar{x}^0)]^{-1}f(\bar{x}^0)\| \leq b,$$

and the coordinate functions of f have continuous second partials so that

(2.6) $$\|f''(\bar{x})\| \leq c \quad \text{for all } \bar{x} \in S(\bar{x}^0,2b),$$

and if, in addition,

(2.7) $$a,b,c \text{ are such that } a \cdot b \cdot c \leq \frac{1}{2},$$

then (i) the iterates are uniquely defined and lie in $S(\bar{x}^0,2b)$, and

(ii) the iterates converge to some element $\bar{\alpha}$ such that $f(\bar{\alpha}) = \bar{0}$ and $\|\bar{x}^k-\bar{\alpha}\| \leq \frac{2b}{2^k}$.

Proof. It is convenient to introduce the following notation:

$$J^k = f'(\bar{x}^k); \quad H^k = [f'(\bar{x}^k)]^{-1}; \quad A^{k+1} = I - H^kJ^{k+1}.$$

Now by a lengthy induction argument we wish to establish the following for k=0,1,2,...

(1) $$\|\bar{x}^{k+1} - \bar{x}^k\| \leq \frac{b}{2^k}$$

(2) $$\|\bar{x}^{k+1} - \bar{x}^0\| \leq 2b$$

(3) $$\|A^{k+1}\| = \|H^k(J^k-J^{k+1})\| \leq \frac{1}{2}$$

$$(4)\quad \|H^{k+1}\| = \|(I-A^{k+1})^{-1}H^k\| \le 2^{k+1}a.$$

Since (1) and (2) are true for k = o by hypothesis, $\bar{x}^1$ and $\bar{x}^o$ are elements of $S(\bar{x}^o,2b)$ where the second partials of the coordinate functions are continuous. This enables us to use Theorem 1.3 to show $J^1 = J^o + B(\xi^1,\bar{x}^1-\bar{x}^o,\cdot)$ which implies that $\|J^1-J^o\| \le c\|\bar{x}^1-\bar{x}^o\|$. This bound on the norm of the difference of the first two Jacobians along with (2.4), (2.5), and (2.7) establishes the following inequality $\|A^1\| \le \|H^o\|\ \|J^o-J^1\| \le ac\|\bar{x}^1-\bar{x}^o\| \le abc \le \frac{1}{2}$, which proves (3) for k=o. Now since $\|A^1\| < 1$, it is a well-known result that $I-A^1$ is nonsingular and $\|(I-A^1)^{-1}\| \le \dfrac{1}{1 - \|A^1\|}$. So we have

$$\|H^1\| \le \|H^o\|\ \|(I-A^1)^{-1}\| \le \frac{\|H^o\|}{1 - \|A^1\|} \le 2a,$$

which establishes (4) for k = o. Observe that $J^1 = J^o(I-A^1)$ since J^o is nonsingular by hypothesis, and thus J^1 also has an inverse.

Let us now prove (1), (2), (3), and (4) for k = n by assuming these same inequalities hold for $o \le k \le n-1$. Since (3) is valid for k equal to n-1, $\|A^n\| < 1$ which implies that J^n is nonsingular by an argument in the preceding paragraph. Thus $\bar{x}^{n+1} = \bar{x}^n - H^n f(\bar{x}^n)$ is uniquely defined with $\|\bar{x}^{n+1} - \bar{x}^n\| \le \|H^n\|\ \|f(\bar{x}^n)\|$. Now to get a bound on $\|f(\bar{x}^n)\|$, recall that (2) is valid for $o \le k \le n-1$ which implies $\bar{x}^n \in S(\bar{x}^o,2b)$ so that Taylor's theorem can be used to produce

$$f(\bar{x}^n) = f(\bar{x}^{n-1}) + J^{n-1}[\bar{x}^n - \bar{x}^{n-1}] + \frac{1}{2} B(\bar{\xi}^n, \bar{x}^n - \bar{x}^{n-1}, \bar{x}^n - \bar{x}^{n-1})$$

$$= \frac{1}{2} B(\bar{\xi}^n, \bar{x}^n - \bar{x}^{n-1}, \bar{x}^n - \bar{x}^{n-1}).$$

Taking the norm of both sides of the above equation and using (2.6), we see that $\|f(\bar{x}^n)\| \leq \frac{c}{2} \|\bar{x}^n - \bar{x}^{n-1}\|^2$. Thus from (1) and (4) at k = n-1 and from (2.7) it is possible to conclude that

$$\|\bar{x}^n - \bar{x}^{n-1}\| \leq \|H^n\| \, \|f(\bar{x}^n)\| \leq \frac{c}{2} (2^n a) \|\bar{x}^n - \bar{x}^{n-1}\|^2 \leq \frac{c}{2} (2^n a) \left\{ \frac{b}{2^{n-1}} \right\}^2 =$$

$$\frac{ab^2c}{2^{n-1}} \leq \frac{b}{2^n},$$

which establishes (1).

The repeated use of (1) will establish (2) in the following way,

$$\|\bar{x}^{n+1} - \bar{x}^o\| \leq \sum_{i=o}^{k} \|\bar{x}^{i+1} - \bar{x}^i\| \leq b \sum_{i=o}^{k} \frac{1}{2^i} \leq 2b .$$

Now since $\bar{x}^{n+1}$ is in $S(\bar{x}^o, 2b)$, the Mean Value Theorem can be used to show $J^{n+1} = J^n + B(\bar{\xi}^{n+1}, \bar{x}^{n+1} - \bar{x}^n, \cdot)$ so that $\|J^{n+1} - J^n\| \leq c\|\bar{x}^{n+1} - \bar{x}^n\|$. This result together with (1), (4), and (2.7) implies that $\|A^{n+1}\| \leq \|H^n\| \, \|J^n - J^{n+1}\| \leq 2^n ac \frac{b}{2^n} = abc \leq \frac{1}{2}$ which proves (3) for k = n. Recall that $\|A^{n+1}\| < 1$ implies that J^{n+1} is non-singular and $\|I - A^{n+1}\| \leq \frac{1}{1 - \|A^{n+1}\|}$. So (4) can be established by noting that

$$\|H^{n+1}\| \leq \frac{\|H^n\|}{1 - \|A^{n+1}\|} \leq 2^{n+1}a.$$

Now (i) follows from (4) which implies that the H^k exist for all k and thus the sequence of iterates $\{\bar{x}^k\}$ is well defined, and (2) shows that $\bar{x}^{k+1} \in S(\bar{x}^o, 2b)$ for all k.

For the convergence argument we use (1) to show that

$$\|\bar{x}^{k+m} - \bar{x}^k\| \leq \sum_{i=k}^{k+m-1} \|\bar{x}^{i+1} - \bar{x}^i\| \leq b \sum_{i=k}^{k+m-1} \frac{1}{2^i} \leq \frac{b}{2^{k-1}}.$$

So the iterates $\{\bar{x}^k\}$ form a Cauchy sequence in R^n and converge to some $\bar{\alpha}$ in $S(\bar{x}^o, 2b)$. Recall that $\|f(\bar{x}^k)\| \leq \frac{c}{2}\|\bar{x}^k - \bar{x}^{k-1}\|^2 \leq \frac{c}{2}\left(\frac{b}{2^{k-1}}\right)^2 = \frac{2b^2c}{4^k}$. So $\lim_{k\to\infty}\|f(\bar{x}^k)\| = 0$. But we also have $\lim_{k\to\infty}\|f(\bar{x}^k)\| = \|f(\bar{\alpha})\|$. Thus $f(\bar{\alpha}) = \bar{0}$. Also $\|\bar{x}^{k+m} - \bar{x}^k\| \leq \frac{b}{2^{k-1}}$ implies that $\lim_{m\to\infty}\|\bar{x}^{k+m} - \bar{x}^k\| = \|\bar{\alpha} - \bar{x}^k\| \leq \frac{2b}{2^k}$.

CHAPTER III

DIFFICULTIES OF NEWTON'S METHOD

Because of its simplicity and rapid convergence, Newton's method is often used to solve nonlinear systems of the form $f(\bar{x}) = \bar{o}$ where $\bar{x} \in R^n$ and $f(\bar{x}) \in R^n$. Nevertheless there are serious difficulties with this method. The excessive number of calculations required to compute the inverse of the Jacobian matrix at each iteration is one of the principal complaints. The calculation of the derivative alone requires $n(n+1)$ functional evaluations, not to mention the calculations required for the inverse. Not only is this process inefficient from a programming point of view, but the derivative may be ill-conditioned or may not even exist. Another major drawback to the method is the necessity of guessing an initial approximation within a suitable neighborhood of the solution so that convergence is assured.

Point Substitution Method

Let us turn our attention to the problem of calculating a new Jacobian matrix for each iteration. Our purpose is to show that instead of evaluating the Jacobian at each iterate, as Newton's method requires, the first derivative term in the algorithm can be calculated at any arbitrary point in a particular region of R^n and the iterates will still converge. Thus those points at which f' is easily calculated can be chosen, while those points where f' is not defined or

ill-conditioned can be avoided. The following results can be found in Bartle [2].

Lemma 3.1 Let Ω be an open set in R^n and $f'(\bar{x})$ exist and be bounded for all $\bar{x} \in \Omega$. Furthermore for $\bar{x}^o \in \Omega$, let $\delta(\bar{x}^o,\varepsilon)$ be a real number such that $\|f'(\bar{x}) - f'(\bar{x}^o)\| \leq \varepsilon$ whenever $\bar{x} \in S(\bar{x}^o,\delta(\bar{x}^o,\varepsilon)) \subset \Omega$. Then if $\bar{x}^1$ and $\bar{x}^2$ are in $S(\bar{x}^o,\delta(\bar{x}^o,\varepsilon))$, we have

$$\|f(\bar{x}^1) - f(\bar{x}^2) - f'(\bar{x}^o)[\bar{x}^1 - \bar{x}^2]\| \leq \varepsilon\|\bar{x}^1 - \bar{x}^2\|.$$

Proof. Using the Mean Value Theorem in Chapter I we conclude that $f(\bar{x}^1) - f(\bar{x}^2) = A(\bar{x}^1 - \bar{x}^2)$ where A is an nth order matrix with $\|A - f'(\bar{x}^o)\| \leq \varepsilon$. So

$$\|f(\bar{x}^1) - f(\bar{x}^2) - f'(\bar{x}^o)[\bar{x}^1 - \bar{x}^2]\| \leq \|A - f'(\bar{x}^o)\| \, \|\bar{x}^1 - \bar{x}^2\|$$

$$\leq \varepsilon \, \|\bar{x}^1 - \bar{x}^2\|.$$

Lemma 3.2 In addition to the hypotheses of Lemma 3.1, if $f'(\bar{x})^{-1}$ is also defined and bounded on a neighborhood N of $\bar{x}^o$, then for any $a > \|f'(\bar{x}^o)^{-1}\|$, there exists a β such that

(i) if $\|\bar{x} - \bar{x}^o\| \leq \beta$, then $\|f'(\bar{x})^{-1}\| < a$ and

(ii) if $\|\bar{x}^k - \bar{x}^o\| \leq \beta$ for k=1,2,3, then

$$\|f(\bar{x}^1) - f(\bar{x}^2) - f'(\bar{x}^3)[\bar{x}^1 - \bar{x}^2]\| \leq \frac{1}{2a}\|\bar{x}^1 - \bar{x}^2\|.$$

Proof. If $\varepsilon = a - \|f'(\bar{x}^o)^{-1}\|$, then there exists a $\delta_1 > o$ such that if $\|\bar{x} - \bar{x}^o\| < \delta_1$, then $\|f'(\bar{x})^{-1} - f'(\bar{x}^o)^{-1}\| < \varepsilon$. Thus if $\|\bar{x} - \bar{x}^o\| < \delta_1$, we have

$$\|f'(\bar{x})^{-1}\| - \|f'(\bar{x}^o)^{-1}\| \leq \|f'(\bar{x})^{-1} - f'(\bar{x}^o)^{-1}\| < \varepsilon = a - \|f'(\bar{x}^o)^{-1}\|$$

and so $\|f'(\bar{x})^{-1}\| < a$. Also by the continuity of f', there exists a $\delta_2 > o$ such that if $\|\bar{x} - \bar{x}^o\| < \delta_2$, then $\|f'(\bar{x}) - f'(\bar{x}^o)\| < \frac{1}{4a}$. Now if $\bar{x}^1$, $\bar{x}^2$, and $\bar{x}^3$ are in $S(\bar{x}^o,\beta)$ where $\beta = \min\{1,\frac{\delta_1}{2},\frac{\delta_2}{2}\}$, then Lemma 3.1 can be used to show

$$\|f(\bar{x}^1) - f(\bar{x}^2) - f'(\bar{x}^o)[\bar{x}^1 - \bar{x}^2]\| \leq \frac{1}{4a}\|\bar{x}^1 - \bar{x}^2\|.$$

So

$$\|f(\bar{x}^1) - f(\bar{x}^2) - f'(\bar{x}^3)[\bar{x}^1-\bar{x}^2]\| \leq \|f(\bar{x}^1) - f(\bar{x}^2) - f'(\bar{x}^o)[\bar{x}^1-\bar{x}^2]\|$$

$$+ \|[f'(\bar{x}^3)-f'(\bar{x}^o)][\bar{x}^1-\bar{x}^2]\| \leq \frac{1}{4a}\|\bar{x}^1-\bar{x}^2\|$$

$$+ \frac{1}{4a}\|\bar{x}^1-\bar{x}^2\| = \frac{1}{2a}\|\bar{x}^1-\bar{x}^2\|.$$

Theorem 3.1 Let Ω be an open subset of R^n, and $f: \Omega \to R^n$ have a bounded first derivative on Ω such that $f'(\bar{x}^o)$ has an inverse that satisfies $\|f'(\bar{x}^o)^{-1}\| < a < \infty$. Then if $\|f(\bar{x}^o)\| < \frac{\beta}{2a}$ where β is as in Lemma 3.2, and each $\bar{z}^k$ for k=1,2,3,... is an arbitrary element of $S(\bar{x}^o,\beta)$ except for $\bar{z}^o$ which equals $\bar{x}^o$, the iterative process $\bar{x}^{k+1} = \bar{x}^k - [f'(\bar{z}^k)]^{-1}f(\bar{x}^k)$ for k=0,1,2,... converges to a unique solution $\tilde{\alpha}$ of

$f(\bar{x}) = \bar{o}$ in $S(\bar{x}^o,\beta)$. Furthermore $\|\bar{x}^k - \bar{\alpha}\| \leq \frac{\beta}{2^k}$.

Proof. Using an inductive approach, we will establish the following:

(1) $\|\bar{x}^k - \bar{x}^o\| < \beta$

(2) $\|\bar{x}^k - \bar{x}^{k-1}\| \leq a\|f(\bar{x}^{k-1})\|$, and

(3) $\|f(\bar{x}^k)\| \leq \frac{1}{2a}\|\bar{x}^k - \bar{x}^{k-1}\|$ for $k=1,2,\ldots$

By hypothesis, $\|\bar{x}^1 - \bar{x}^o\| = \|f'(\bar{x}^o)^{-1}f(\bar{x}^o)\| \leq a\|f(\bar{x}^o)\| < \frac{\beta}{2} < \beta$ which proves (1) and (2) for $k=1$. Since $f(\bar{x}^1) = f(\bar{x}^1) - f(\bar{x}^o) - f'(\bar{x}^o)[\bar{x}^1-\bar{x}^o]$, Lemma 3.1 shows that $\|f(\bar{x}^1)\| \leq \frac{1}{2a}\|\bar{x}^1 - \bar{x}^o\|$. Thus (1), (2), and (3) are true for $k=1$. Now suppose they are true for $1 \leq k \leq n$. Using Lemma 3.2, $\bar{x}^{n+1} - \bar{x}^n = -f'(\bar{z}^n)^{-1}f(\bar{x}^n)$ implies that $\|\bar{x}^{n+1} - \bar{x}^n\| \leq a\|f(\bar{x}^n)\|$ which proves (2) for $k = n+1$.

Since $\|\bar{x}^{n+1} - \bar{x}^n\| \leq a\|f(\bar{x}^n)\| \leq \frac{1}{2}\|\bar{x}^n - \bar{x}^{n-1}\|$, we can see inductively that

$$\|\bar{x}^{n+1} - \bar{x}^o\| \leq \sum_{i=o}^{n} 2^{-i}\|\bar{x}^1 - \bar{x}^o\| < \frac{\beta}{2}\sum_{i=o}^{n} 2^{-i} < \beta.$$

This establishes (1) for $k = n+1$.

Now $\bar{x}^{n+1} \in S(\bar{x}^o,\beta)$ and $f(\bar{x}^{n+1}) = f(\bar{x}^{n+1}) - f(\bar{x}^n) - f'(\bar{z}^n)(\bar{x}^{n+1}-\bar{x}^n)$ imply by Lemma 3.2 that $\|f(\bar{x}^{n+1})\| \leq \frac{1}{2a}\|\bar{x}^{n+1} - \bar{x}^n\|$. Thus (3) is proven.

Notice that $\|\bar{x}^{k+p} - \bar{x}^k\| \leq \sum_{i=1}^{p} \|\bar{x}^{k+i} - \bar{x}^{k+i-1}\| \leq a\|f(\bar{x}^o)\|2^{-k}\{\sum_{i=o}^{p-1} 2^{-i}\} < \frac{\beta}{2^k}$ for all $p \geq 1$.

So the $\{\bar{x}^k\}$'s from a Cauchy sequence in $S(\bar{x}^o,\beta) \subset R^n$, and therefore converge to some $\bar{\alpha} \in S(\bar{x}^o,\beta)$ with $\lim_{p\to\infty}\|\bar{x}^{k+p} - \bar{x}^k\| = \|\bar{\alpha} - x^k\| \leq \frac{\beta}{2^k}$ for $k=0,1,2,\ldots$ Statement (3) shows that $\lim_{k\to\infty}\|f(\bar{x}^k)\| = \|f(\bar{\alpha})\| = o$ proving that $\bar{\alpha}$ is a solution of $f(\bar{x}) = \bar{o}$.

Regarding uniqueness, suppose $\bar{\gamma}$ is another solution of $f(\bar{x}) = \bar{o}$ in $S(\bar{x}^o,\beta)$.

Then

$$\|\bar{\gamma}-\bar{\alpha}\| = \|f'(\bar{x}^o)^{-1}f'(\bar{x}^o)[\bar{\gamma}-\bar{\alpha}]\| \leq a\|f'(\bar{x}^o)[\bar{\gamma}-\bar{\alpha}]\| =$$

$$a\|f(\bar{\gamma}) - f(\bar{\alpha}) - f'(\bar{x}^o)[\bar{\gamma}-\bar{\alpha}]\| < \frac{1}{2}\|\bar{\gamma}-\bar{\alpha}\|$$

by Lemma 3.2. So $\|\bar{\gamma} - \bar{\alpha}\| = o$ which contradicts the hypothesis that $\bar{\gamma} \neq \bar{\alpha}$.

Observe that Lemma 3.2 is not needed in the proof if $\bar{z}^k = \bar{x}^k$ for $k=1,2,\ldots$ Note also from the convergence factor that $\|\bar{x}^{k+1} - \bar{\alpha}\| \leq \frac{1}{2}\|\bar{x}^k - \bar{\alpha}\|$ for $k=0,1,2,\ldots$ Since the convergence rate has not been shown to be any higher than one, ease in handling the Jacobian matrix has apparently cost us the quadratic convergence rate. In fact, with $\bar{z}^k = \bar{x}^o$ for $k=1,2,3,\ldots$, this scheme has only linear convergence when used in R^1 on the function $f(x) = x^2$. However when the $\bar{z}^k$'s are all chosen equal to $\bar{x}^o$, the linear operator remains constant throughout the algorithm, and the general recursion formula becomes $\bar{x}^{k+1} = \bar{x}^k - f'(\bar{x}^o)^{-1}f(\bar{x}^k)$ for $k=0,1,2,\ldots$

Notice that each iteration requires only n coordinate function evaluations. Thus with respect to time, the reduction in the number of calculations per iteration tends to offset any reduction in the convergence rate.

Given $f(\bar{x})$ and $\bar{x}^o$, let $\{\bar{x}^k\}$ be the iterates computed by Newton's method, and let $\{\tilde{x}^k\}$ be the iterates computed by the point substitution method with $\bar{z}^k = \bar{x}^o$ for k=1,2,... Then if $L = \|\bar{x}^{-1} - \bar{x}^o\|$, $M = \|f'(\bar{x}^o)^{-1}\|$, $N = \|f''(\bar{x}^o)\|$, and $h = LMN < \frac{1}{2}$, error bounds for the two methods are given by the following formulas (which can be found in Lohr and Rall [13]):

$$\|\bar{x}^k - \bar{\alpha}\| \leq \frac{L(2h)^{2k-1}}{2^{k-1}} = r_k$$

$$\|\tilde{x}^k - \bar{\alpha}\| \leq 2hL(1 - \sqrt{1-2h})^k = r_{\tilde{k}}$$

where $\bar{\alpha}$ is the solution to $f(\bar{x}) = \bar{o}$ approached by both $\{\bar{x}^k\}$ and $\{\tilde{x}^k\}$. Now let t_1 and t_2 signify the times required to calculate $f(\bar{x}^k)$ and $f'(\bar{x}^k)^{-1}$, respectively, and let t_3 denote the time required to compute $\bar{x}^{k+1} = \bar{x}^k - f'(\bar{x}^k)^{-1}f(\bar{x}^k)$ given $f(\bar{x}^k)$ and $f'(\bar{x}^k)^{-1}$. If ε is the desired accuracy, $\eta_\varepsilon = \min\{k: r_k \leq \varepsilon\}$, and $\eta_{\tilde{\varepsilon}} = \min\{k: r_{\tilde{k}} \leq \varepsilon\}$, then the point substitution process will give the desired accuracy in the least amount of time whenever $\eta_{\tilde{\varepsilon}}(t_1+t_3) < \eta_\varepsilon(t_1+t_2+t_3)$. Now, Lohr and Rall [13] suggest a procedure which combines the two methods to reduce computing time. In this procedure the error and time analyses are repeated after each iteration. At the kth iteration let t_1^k, t_2^k, and

t_3^k be the times required to compute $f(\bar{x}^{k-1})$, $f'(\bar{x}^{k-1})^{-1}$ and $\bar{x}^k = \bar{x}^{k-1} - f'(\bar{x}^{k-1})^{-1}f(\bar{x}^{k-1})$ given $f(\bar{x}^{k-1})$ and $f'(\bar{x}^{k-1})^{-1}$, respectively. Then substitute $\bar{x}^k$ for $\bar{x}^1$ and $\bar{x}^{k-1}$ for $\bar{x}^o$ in the formulas for L, M and N (call these new values L^k, M^k and N^k), and using L^k, M^k and N^k in the error bound expressions compute η_ε^k and $\eta_{\tilde{\varepsilon}}^k$, the number of additional iterates needed for each method to converge. In addition let t_4^k be the time required to compute η_ε^k and $\eta_{\tilde{\varepsilon}}^k$. Then the procedure is to use Newton's method for k iterations until

$$\eta_{\tilde{\varepsilon}}^k(t_1^k + t_3^k) < \eta_\varepsilon^k(t_1^k + t_2^k + t_3^k + t_4^k)$$

and then switch to the point substitution algorithm until the convergence criterion is satisfied.

Another special case of this modified method is to use the same Jacobian for r iterations where r is some positive integer chosen beforehand. Thus we let

$$\bar{z}^k = \bar{x}^o \quad \text{for } k=1,2,\ldots,r-1$$

$$\bar{z}^k = \bar{x}^r \quad \text{for } k=r,r+1,\ldots,2r-1$$

$$\bar{z}^k = \bar{x}^{2r} \quad \text{for } k=2r,2r+1,\ldots,3r-1$$

$$\vdots$$

This scheme also reduces the number of calculations required per iteration when compared to Newton's Method. Again, in order to reduce

the time required for each iteration, it may have been necessary to forfeit the quadratic convergence.

Corollary (Bartle [2]). Let $f: \Omega \to R^n$ be as in Theorem 3.1. Consider a set of bounded linear operators $\{T_k\}$ where $T_k: R^n \to R^n$ for $k=0,1,2,\ldots$ such that the following conditions are satisfied,

(i) $\|T_n(\bar{x}) - f'(\bar{x}^o)\| < \frac{1}{4a}$ for all $\bar{x} \in S(\bar{x}^o,\beta)$, $n=0,1,2,\ldots$

and

(ii) $\|T_n^{-1}(\bar{x})\| < a$ for all $\bar{x} \in S(\bar{x}^o,\beta)$, $n=0,1,2,\ldots$

Then the sequence of iterates defined by $\bar{x}^{k+1} = \bar{x}^k - T_k^{-1}(\bar{x}^k)f(\bar{x}^k)$ for $k=0,1,2,\ldots$ will converge to $\bar{\alpha} \in S(\bar{x}^o,\beta)$ where $\bar{\alpha}$ is a solution of the equation $f(\bar{x}) = \bar{o}$.

Proof. The proof is the same as the proof of Theorem 3.1. Condition (i) implies that Lemma 3.2 holds, and condition (ii) takes the place of requiring that $\{\bar{z}^k\} \subset S(\bar{x}^o,\beta)$.

Thus instead of choosing new points at which to evaluate the Jacobian matrix, it may be possible to switch to a different, more easily managed set of bounded linear operators.

Secant, Wolfe's, and Barnes' Methods

Another method commonly used to circumvent the calculation of the derivative is the secant method. If $f: R^1 \to R^1$, let x^1 and x^o be two initial approximations to a root α of $f(x) = o$. Then the secant iterates are given by the formula,

$$x^{k+1} = x^k - \left(\frac{x^k - x^{k-1}}{f(x^k) - f(x^{k-1})}\right) f(x^k)$$

or equivalently

$$x^{k+1} = \left(\frac{f(x^k)}{f(x^k) - f(x^{k-1})}\right)x^{k-1} - \left(\frac{f(x^{k-1})}{f(x^k) - f(x^{k-1})}\right)x^k \quad \text{for } k=1,2,\ldots$$

Notice that this algorithm is similar to Newton's method except that $[f'(x^k)]^{-1}$ has been approximated by $\frac{x^k - x^{k-1}}{f(x^k) - f(x^{k-1})}$. However the secant method requires only one functional evaluation per iteration while Newton's method in the same setting uses two. The following theorem can be found in Isaacson and Keller [9].

Theorem 3.2 Let x^0 and x^1 be two initial approximations of α, a root of $f: R^1 \to R^1$, and suppose all the iterates $\{x^k\}$ of the secant method for this system lie in $S(\alpha,\rho)$ for $\rho > 0$ such that $0 \le \left|\frac{f''(x)}{2f'(y)}\right| \le M$ for all $x,y \in S(\alpha,\rho)$. Now if $|x^0 - \alpha| \le \frac{\beta}{M}$ and $|x^1 - \alpha| \le \frac{\beta}{M}$ where $\beta < 1$, then $\lim_{k\to\infty} x^k = \alpha$ and the rate of convergence is given by $|x^{k+1} - \alpha| \le \beta^{g_k} |x^0 - \alpha|$ where g_k is a sequence of real numbers such that $g_1 = 1$, $g_2 = 2$, and $g_k = g_{k-1} + g_{k-2} + 1$ for $k=3,4,\ldots$

Proof. Let us introduce the divided difference notation

$$f[x^{k-1},x^k] = \frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}},$$

and

$$f[\alpha,x^k,x^{k-1}] = \frac{f[x^{k-1},x^k] - f[x^k,\alpha]}{x^{k-1} - \alpha}.$$

Now Theorem 1.3 enables us to write

$$f[x^{k-1},x^k] = f'(\xi^k) \tag{3.1}$$

where ξ^k lies between x^{k-1} and x^k, and the identity,

$$f[\alpha,x^k,x^{k-1}] = \frac{f''(\eta^k)}{2} \tag{3.2}$$

with η^k lying between x^{k-1}, x^k and α, can be found in Milne-Thomson [14]. Use the secant formula and the divided difference notation to see that

$$x^{k+1} - \alpha = x^k - \alpha - f(x^k)\left[\frac{x^k - x^{k-1}}{f(x^k) - f(x^{k-1})}\right]$$

$$= (x^k - \alpha)\left[\frac{f[x^{k-1},x^k] - f[x^k,\alpha]}{f[x^k,x^{k-1}]}\right]$$

$$= (x^k - \alpha)(x^{k-1} - \alpha)\left[\frac{f[\alpha,x^k,x^{k-1}]}{f[x^k,x^{k-1}]}\right].$$

Then

$$|x^{k+1} - \alpha| = |x^k - \alpha|\,|x^{k-1} - \alpha|\left|\frac{f''(\eta^k)}{2f'(\xi^k)}\right| \le M|x^k - \alpha|\,|x^{k-1} - \alpha|$$

for $k = 1,2,\ldots$

Thus for k = 1,

$$|x^2 - \alpha| \leq M|x^1 - \alpha|\ |x^o - \alpha| \leq \beta^{g_1}\ |x^o - \alpha|.$$

Now suppose $|x^k - \alpha| \leq \beta^{g_k}\ |x^o - \alpha|$ for k=2,3,4,...,n, then

$$|x^{n+1} - \alpha| \leq M|x^n - \alpha|\ |x^{n-1} - \alpha| \leq M\ \beta^{g_n}|x^o - \alpha|\ \beta^{g_{n-1}}\ |x^o - \alpha|$$

$$\leq \beta^{g_n+g_{n-1}+1}\ |x^o - \alpha| \leq \beta^{g_{n+1}}\ |x^o - \alpha|.$$

Jeaves [10] computes the convergence rate of the secant method as $\frac{1 + \sqrt{5}}{2}$, which places it at a disadvantage when compared to Newton's method. However because it requires fewer calculations per iteration, the secant method often converges to a solution more quickly than does Newton's method.

It is important to notice that the secant algorithm can also be characterized in the following way: At the kth step, solve the linear system

$$a_1^{(k)} + a_2^{(k)} = 1$$

$$a_1^{(k)} f(x^{k-1}) + a_2^{(k)} f(x^k) = 0 \qquad \text{for } a_1^{(k)} \text{ and } a_2^{(k)}.$$

Then let $x^{k+1} = a_1^{(k)} x^{k-1} + a_2^{(k)} x^k$, and the iterates will be the same as those produced by the earlier version of the secant method. With

this in mind, Wolfe [23] generalizes the secant method in the one dimensional case to higher dimensional spaces as follows.

Let $f(\bar{x})$ be a transformation from R^n into R^n, and $\bar{x}^o, \bar{x}^1, \ldots, \bar{x}^n$ be n+1 initial approximations in R^n to $\bar{\alpha}$ where $\bar{\alpha} \in R^n$ is a solution of $f(\bar{x}) = \bar{o}$.

Now let

$$A = \begin{bmatrix} f_1(\bar{x}^o) & f_1(\bar{x}^1) & \cdots & f_1(\bar{x}^n) \\ f_2(\bar{x}^o) & f_2(\bar{x}^1) & \cdots & f_2(\bar{x}^n) \\ \vdots & \vdots & & \vdots \\ f_n(\bar{x}^o) & f_n(\bar{x}^1) & \cdots & f_n(\bar{x}^n) \\ 1 & 1 & \cdots & 1 \end{bmatrix}$$

$\bar{a} = [a_o, a_1, \ldots, a_n]^T$ and, $\bar{b} = [o, \ldots, o, 1]^T \in R^n$.

Solve

$A\bar{a} = \bar{b}$ to get

$\bar{a} = \bar{A}^{-1}\bar{b}$ if A is non-singular, which shows that $\bar{a}$ is the n+1 st column of A^{-1}.

Notice that $\bar{a}$ satisfies $\sum_{i=o}^{n} a_i f_j(\bar{x}^i) = o$ for $j=1,2,\ldots,n$ and

$$\sum_{i=o}^{n} a_i = 1.$$

Now the n+1 st iterate is defined to be

$$\bar{x}^{n+1} = \sum_{i=o}^{n} a_i \bar{x}^i.$$

Replace $\bar{x}^j$ with $\bar{x}^{k+1}$ where $\|f(\bar{x}^j)\| \geq \|f(\bar{x}^i)\|$ for i=o,...,n and repeat the process. Now if Wolfe's method is slightly altered, a type of quadratic convergence can readily be shown. The alteration only changes the vector to be replaced by the new iterate. Instead of letting the new iterate $\bar{x}^{n+1}$ replace whichever of the previous n+1 iterates had the largest functional value in norm, let $\bar{x}^k$ replace $\bar{x}^{k-(n+1)}$. This is used to establish a definite pattern in the computations.

Theorem 3.3 Let f: $R^n \to R^n$. Then if

(i) $f''(\bar{x})$ and $[f'(\bar{x})]^{-1}$ exist for $\bar{x} \in S(\bar{\alpha},\rho)$ where $\bar{\alpha}$ is a solution of $f(\bar{x}) = \bar{o}$ and $\rho > o$,

(ii) the a_i's described above remain bounded so that $\frac{\|a_i f''(\bar{x})\|}{2} \leq b$ for i=o,...,n in all iterations for $x \in S(\bar{\alpha},\rho)$,

(iii) $\bar{x}^k \in S(\bar{\alpha},\rho)$ for k=o,1,... and

(iv) $(n+1)\, b\rho = \theta < 1$,

the iterates computed by the modified Wolfe's method converge to $\bar{\alpha}$.

Proof. First note that there is no loss of generality in assuming that $f'(\bar{\alpha}) = I$ providing that $f'(\bar{\alpha})$ is nonsingular. If $f'(\bar{\alpha}) \neq I$, let $g(\bar{x}) = f'(\bar{\alpha})^{-1} f(\bar{x})$. Then $g'(\bar{\alpha}) = I$, and the algorithm can be applied to g since any root of g will also be a root of f.

Using Taylor's theorem on f gives

$$f(\bar{x}^{n+1}) = f'(\bar{\alpha})[\bar{x}^{n+1} - \bar{\alpha}] + \frac{1}{2} B(\bar{\xi}^{n+1}, \bar{x}^{n+1}-\bar{\alpha}, \bar{x}^{n+1}-\bar{\alpha}).$$

If we observe that

$$f'(\bar{\alpha})[\bar{x}^{n+1} - \bar{\alpha}] = \sum_{i=o}^{n} a_i f'(\bar{\alpha})[\bar{x}^{i} - \bar{\alpha}]$$

and

$$f'(\bar{\alpha})[\bar{x}^{i} - \bar{\alpha}] = f(\bar{x}^{i}) - \frac{1}{2} B(\bar{\xi}^{i}, \bar{x}^{i}-\bar{\alpha}, \bar{x}^{i}-\bar{\alpha})$$

for i=o,1,...,n,

then

$$f(\bar{x}^{n+1}) = \sum_{i=o}^{n} a_i (f(\bar{x}^{i}) - \frac{1}{2} B(\bar{\xi}^{i}, \bar{x}^{i}-\bar{\alpha}, \bar{x}^{i}-\bar{\alpha}))$$

$$+ \frac{1}{2} B(\bar{\xi}^{n+1}, \bar{x}^{n+1}-\bar{\alpha}, \bar{x}^{n+1}-\bar{\alpha})$$

$$= \sum_{i=o}^{n} a_i (- \frac{1}{2} B(\bar{\xi}^{i}, \bar{x}^{i}-\bar{\alpha}, \bar{x}^{i}-\bar{\alpha}))$$

$$+ \frac{1}{2} B(\bar{\xi}^{n+1}, \bar{x}^{n+1}-\bar{\alpha}, \bar{x}^{n+1}-\bar{\alpha}) .$$

However $f'(\bar{\alpha}) = I$, shows that $f(\bar{x}^{n+1}) = [\bar{x}^{n+1} - \bar{\alpha}] + \frac{1}{2} B(\bar{\xi}^{n+1}, \bar{x}^{n+1}-\bar{\alpha}, \bar{x}^{n+1}-\bar{\alpha})$. Thus since $[\bar{x}^{n+1}-\bar{\alpha}] = - \sum_{i=o}^{n} a_i [\frac{1}{2} B(\bar{\xi}^{i}, \bar{x}^{i}-\bar{\alpha}, \bar{x}^{i}-\bar{\alpha})]$, we have

$$\|\bar{x}^{n+1} - \bar{\alpha}\| \leq \sum_{i=o}^{n} \frac{\|a_i f''(\bar{\xi}^i)\|}{2} \|\bar{x}^i - \bar{\alpha}\|^2 \leq b \sum_{i=o}^{n} \|\bar{x}^i - \bar{\alpha}\|^2$$

$$\leq (n+1)b\rho^2 = \theta\rho$$

Continuing in this manner gives the following results:

$$\|\bar{x}^{n+2} - \bar{\alpha}\| \leq b \sum_{i=1}^{n+1} \|\bar{x}^i - \bar{\alpha}\|^2 = bn\rho^2 + b\theta^2\rho^2 = b\rho^2(n+\theta^2) < \theta\rho$$

$$\|\bar{x}^{n+3} - \bar{\alpha}\| \leq b\rho^2(n-1 + 2\theta^2) < \theta\rho$$

$$\vdots$$

$$\|\bar{x}^{2n+1} - \bar{\alpha}\| \leq b\rho^2(n+1\ \theta^2) = \theta^3\rho$$

$$\|\bar{x}^{2n+2} - \bar{\alpha}\| \leq b\rho^2(n\theta^2 + \theta^6) < \theta^3\rho$$

$$\vdots$$

$$\|\bar{x}^{3n+1} - \bar{\alpha}\| \leq b\rho^2((n+1)\theta^6) = \theta^7\rho\ .$$

So the bound on the error is $\theta^k\rho$ where $\theta < 1$ and k is initially equal to one but then more than doubles after each cycle of n+1 iterations.

We will now describe another algorithm, based on the secant method, which is also designed to avoid the time consuming job of computing the derivative at each iteration. Due to Barnes [1], this method is claimed to be of particular value when a good approximation to the root and to the Jacobian matrix can be found. Indeed the

process first requires an initial guess of the solution and first derivative. It then proceeds to use as few functional evaluations as possible to correct the Jacobian matrix at each iteration. Barnes comments that in practice his method seems to be more reliable than Wolfe's secant method. To describe this method let $\bar{x}^o$ be the initial approximation to the root, J^o be the approximation to the Jacobian matrix at $\bar{x}^o$, and f^k the value of f at $\bar{x}^k$ for k=0,1,... Compute $\bar{\rho}^o$ from $J^o\bar{\rho}^o = -f^o$, and then let $\bar{x}^1 = \bar{x}^o + \bar{\rho}^o$. Notice that if J^o is the Jacobian matrix at $\bar{x}^o$, this first iterate is identical to the first iterate of Newton's method.

Let us choose the next approximate Jacobian matrix, J^1, so that

$$f^1 = f^o + J^1\bar{\rho}^o. \tag{3.3}$$

Then if $J^1 = J^o + D^o$ where D^o is the correction matrix, we want $f^1 = f^o + (J^o+D^o)\bar{\rho}^o$ to be satisfied, or equivalently,

$$f^1 = D^o\bar{\rho}^o. \tag{3.4}$$

A solution to (3.4) is $D^o = \dfrac{f^1\bar{z}^{oT}}{\bar{z}^{oT}\bar{\rho}^o}$ where $\bar{z}^o$ is an element of R^n to be chosen later.

The general iterative scheme for Barnes' method is

$$\bar{\rho}^k = -(J^k)^{-1}f^k$$

$$\underline{x}^{k+1} = \underline{x}^{k} + \underline{\rho}^{k}$$

$$D^{k} = \frac{f^{k+1}\underline{z}^{k^T}}{\underline{z}^{k^T}\underline{\rho}^{k}}$$

$$J^{k+1} = J^{k} + D^{k} \qquad \text{for } k=0,1,2,\ldots$$

Note from (3.4) that $D^{k}\underline{\rho}^{k} = f^{k+1}$. Since $J^{k+1} = J^{k} + D^{k}$, we see that $J^{k+1}\underline{\rho}^{k} = J^{k}\underline{\rho}^{k} + D^{k}\underline{\rho}^{k} = f^{k+1} - f^{k}$.

Now the $\underline{z}^{k}$'s are chosen to be orthogonal to the correction vectors to the solution. Thus if $k \geq n$, $\underline{z}^{k}$ is chosen orthogonal to $\{\underline{\rho}^{k-n+1},\ldots,\underline{\rho}^{k-1}\}$ (previous n-1 steps), or if $k < n$, then $\underline{z}^{k}$ is chosen orthogonal to $\{\underline{\rho}^{o},\ldots,\underline{\rho}^{k-1}\}$. For simplicity, choose $\underline{z}^{k}$ to be the linear combination of the previous k-1 or n-1 correction vectors which is orthogonal to the appropriate elements mentioned above. If the $\underline{z}^{k}$'s are also taken to be unit vectors, they can be formed using the Gram Schmidt process.

Thus

$$D^{i}\underline{\rho}^{j} = \frac{f^{i+1}\,\underline{z}^{i^T}}{\underline{z}^{i^T}\underline{\rho}^{i}}\,\underline{\rho}^{j} = 0 \quad \text{for } 1\leq i-j<n.$$

Hence $J^{i+1}\,\underline{\rho}^{j} = [J^{j+1} + D^{j+1} + \ldots + D^{i}]\underline{\rho}^{j} = J^{j+1}\,\underline{\rho}^{j}$ for $1\leq i-j<n$, and so

$$J^{k}\underline{\rho}^{k-i} = J^{(k-i+1)}\underline{\rho}^{k-i}$$

$$= f^{k-i+1} - f^{k-i} = \delta f^{k-i} \quad \text{for } i<k, o<i\leq n.$$

Therefore for $k \geq n$, we have

$$J^{k}[\underline{\rho}^{k-n}, \underline{\rho}^{k-n+1}, \ldots, \underline{\rho}^{k-1}] = [\delta f^{k-n}, \delta f^{k-n+1}, \ldots, \delta f^{k-1}].$$

Now it may be of benefit to show how $\underline{x}^{k+1}$ where $k \geq n$ can be expressed in terms of the previous $n + 1$ iterates and their function values. Since $\delta f^{k-i-1} = J^{k}\underline{\rho}^{k-i-1}$ for $o\leq i\leq n$, we see that $f^{k-i} - J^{k}\underline{x}^{k-i} = f^{k-i-1} - J^{k}\underline{x}^{k-i-1} = L$ for $o\leq i\leq n$, where L is an $n\times 1$ matrix. Thus the equation $f^{k-i} - f^{k-i-1} = J^{k}[\underline{x}^{k-i} - \underline{x}^{k-i-1}]$ for $o\leq i\leq n$ can be written as

$$f^{k-i} = J^{k}\underline{x}^{k-i} + L \quad \text{for } o\leq i\leq n \tag{3.5}$$

Now if I is the $1 \times (n+1)$ matrix with each element one, then (3.5) can be rewritten as

$$F = J^{k}X + LI$$

where

$$F = \begin{bmatrix} f_1^{k-n} & f_1^{k-n+1} & \cdots & f_1^{k} \\ \vdots & \vdots & & \vdots \\ f_n^{k-n} & f_n^{k-n+1} & \cdots & f_n^{k} \end{bmatrix}$$

$$X = \begin{bmatrix} x_1^{k-n} & x_1^{k-n+1} & \cdots & x_1^{k} \\ \vdots & \vdots & & \vdots \\ x_n^{k-n} & x_n^{k-n+1} & \cdots & x_n^{k} \end{bmatrix}$$

Then $F_i = J_i^k X + L_i I$ for $i=1,\ldots,n+1$ when F_i, J_i^k, L_i denote the ith rows of the respective matrices, or equivalently

$$F_i = [J^k | L]_i \begin{bmatrix} X \\ \hline I \end{bmatrix} .$$

The transpose gives

$$F_i^T = [X^T | I^T] \begin{bmatrix} J^{k^T} \\ \hline L^T \end{bmatrix};$$

So by Cramer's Rule

$$L_i = \frac{\det[X^T | F_i^T]}{\det[X^T | I^T]} = \frac{\det[\frac{X}{F_i}]}{\det[\frac{X}{I}]}$$

This gives L_i in terms of the n + 1 previous vectors and their function values. Now $\bar{x}^{k+1} - \bar{x}^k = -[J^k]^{-1} f^k$. So

$$\begin{aligned} \bar{x}^{k+1} &= \bar{x}^k - [J^k]^{-1} f^k \\ &= [J^k]^{-1}[J^k \bar{x}^k - f^k] \\ &= -[J^k]^{-1} L, \end{aligned}$$

which shows how $\bar{x}^{k+1}$ can be expressed as a function of the previous

iterates and their functional values. A convergence proof for this method can be found in Barnes [1], page 69.

Broyden's Method

The idea of approximating the inverse of the Jacobian matrix and then "improving" it at each iteration is also incorporated by Broyden [3,4] in his papers on quasi-Newton methods. The approximate Jacobian matrices of this method, $\{B^k\}$, satisfy a particular property to be described later of the true Jacobian matrix. In addition, the improvement of B^k at each iteration is developed in a manner designed to minimize the number of calculations required.

Broyden's method is used, as is Newton's method, to solve $f(\bar{x}) = \bar{o}$ where f is a differentiable function from R^n into R^n, and the algorithms of the two methods are very similar. For this new method let B^o and $\bar{x}^o$ be initial approximations to the Jacobian matrix and the root of f, respectively. Then calculate the improvement vector, $\bar{\rho}^k$, at the kth step by $\bar{\rho}^k = -H^k f(\bar{x}^k)$ where H^k is the inverse of B^k, and compute the k+1st iterate using the formula $\bar{x}^{k+1} = \bar{x}^k + t^k\bar{\rho}^k$ for $k=1,2,\ldots$ where $t^k \in R^1$ is designed to assist the convergence of the iterates.

The t^k's are to be chosen in the following manner. Once a direction of improvement, $\bar{\rho}^k$, has been obtained from $-H^k f^k$, the function at the new iterate $f^{k+1} = f(\bar{x}^k + t\bar{\rho}^k)$ can be treated as a function of t. Now t^k should be chosen to either minimize the norm of f^{k+1} or to simply reduce the norm of f^{k+1} so that $\|f(\bar{x}^k+t^k\bar{\rho}^k)\| \leq \|f^k\|$. The minimization procedure gives the greatest improvement in the solution but

may require an excessive number of calculations. On the other hand, the norm reduction strategy normally requires fewer calculations but may not provide as good an immediate improvement in the solution. Broyden states that his experiments suggest that the norm reduction method consumes less computer time and is therefore more desirable. It is also of interest to note that this technique can be used occasionally with the regular Newton's method to induce convergence in an ordinarily divergent system.

Unfortunately the convergence of the iterates is not guaranteed in this algorithm. However the norms of the functional values of the iterates form a non-increasing sequence, and in practice this is often enough for convergence.

Let us now discuss the appearance of the B^k's. Recall that at the kth step

$$\bar{x}^{k+1} = \bar{x}^k + t\bar{\rho}^k \quad \text{for } t = t^k. \tag{3.6}$$

Let $g(t) = \bar{x}^k + t\bar{\rho}^k$ so that $F(t) = f(g(t)) = f(\bar{x}^k + t\bar{\rho}^k)$. Then taking the derivative according to Theorem 1.4, we have

$$F'(t) = f'(g(t))g'(t) = f'(g(t))\bar{\rho}^k . \tag{3.7}$$

Now to approximate $F'(t)$, expand F about t^k using Taylor's theorem to show

$$F(t^k - s^k) = F(t^k) - s^k F'(t^k) + (s^k)^2 F''(\xi^k),$$

where $s^k \in R^1$ is to be chosen later and ξ^k lies between $t^k - s^k$ and s^k. Ignoring the error term in the Taylor expansion gives $s^k F'(t^k) \approx F(t^k) - F(t^k - s^k)$, and from (3.7) we have

$$s^k[f'(g(t))\rho^{-k}] \approx F(t^k) - F(t^k - s^k) = f(\bar{x}^{k+1}) - f(\bar{x}^k + (t^k - s^k)\rho^{-k}) .$$

This is the property that the approximate Jacobian matrix will be required to satisfy. Thus B^{k+1} will be chosen so that

$$s^k[B^{k+1}\rho^{-k}] = f(\bar{x}^{k+1}) - f(\bar{x}^k + (t^k - s^k)\rho^{-k}). \tag{3.8}$$

In choosing s^k, several considerations should be kept in mind. In order for the above approximations to be valid, the second and higher order terms of Taylor's expansion must be negligible, and so $|s^k|$ should be as small as possible. On the other hand, $|s^k|$ must be large enough so that no appreciable round off error creeps into the computation of the right-hand side of (3.8). Now in order to avoid extra calculations s^k will be related to the t^k already chosen. If the first trial value of t was used as t^k, then let $s^k = t^k$ in which case $f(\bar{x}^k + (t^k - s^k)\rho^{-k}) = f(\bar{x}^k)$. Thus B^{k+1} is required to satisfy $t^k[B^{k+1}\rho^{-k}] = f(\bar{x}^{k+1}) - f(\bar{x}^k)$, and no extra functional evaluations are needed. However if $f(\bar{x}^k + t\rho^{-k})$ was evaluated at more than one value of t, let t^* be the closest such value to t^k. Now if $|t^k| \leq |t^k - t^*|$, again let $s^k = t^k$. Otherwise let $s^k = t^k - t^*$ so that $f(\bar{x}^k + (t^k - s^k)\rho^{-k}) = f(\bar{x}^k + t^*\rho^{-k})$. Now B^{k+1} must satisfy

$(t^k - t^*)[B^{k+1}\bar{\rho}^k] = f(\bar{x}^{k+1}) - f(\bar{x}^k + t^*\bar{\rho}^k)$, and again no new functional evaluations are required.

For convenience, let $[B^k]^{-1} = H^k$ and $y^k = f(\bar{x}^{k+1}) - f(\bar{x}^k + (t^k - s^k)\bar{\rho}^k)$, then (3.8) can be written as

$$H^{k+1}\bar{y}^k = s^k\bar{\rho}^k. \tag{3.9}$$

Definition 3.2 An algorithm for finding a root of a differentiable function $f: R^n \to R^n$ which requires an initial estimate of both the root and the Jacobian matrix of f, and which produces its iterates by the formula $\bar{x}^{k+1} = \bar{x}^k + t^k\bar{\rho}^k$ with $\bar{\rho}^k = -H^k f^k$ and H^{k+1} satisfying $H^{k+1}\bar{y}^k = s^k\bar{\rho}^k$ as described above is defined to be a quasi-Newton method.

In order for such an algorithm to be well defined, each B^k must be nonsingular so that H^k can be computed, and $|s^k|$ must be small enough so that the Taylor approximation is valid. Broyden comments that as $\bar{x}^k$ approaches the root of f, Taylor's approximation does improve, and in his numerical experiments, quadratic convergence is often approached.

Theorem 3.4 If the iterates $\{\bar{x}^k\}$ of a quasi-Newton method lie within a neighborhood N of $\bar{\alpha}$, a solution of $f(\bar{x}) = \bar{o}$, where $f''(\bar{x})$ is bounded in N then $\lim_{k\to\infty} \|[H^{k+1} - f'(\bar{x}^{k+1})^{-1}]\bar{y}^k\| = o$.

Proof. By Taylor's theorem, we have

$$f(\bar{x}^{k+1} - s^k\bar{\rho}^k) = f(\bar{x}^{k+1}) - f'(\bar{x}^{k+1})[s^k\bar{\rho}^k] + \frac{1}{2}B[\bar{\xi}^k, s^k\bar{\rho}^k, s^k\bar{\rho}^k].$$

Since $\bar{y}^k = f(\bar{x}^{k+1}) - f(\bar{x}^k + (t^k - s^k)\bar{\rho}^k) = f(\bar{x}^{k+1}) - f(\bar{x}^{k+1} - s^k\bar{\rho}^k)$, we see that $\bar{y}^k = f'(\bar{x}^{k+1})[s^k\bar{\rho}^k] - \frac{1}{2} B[\bar{\xi}^k, s^k\bar{\rho}^k, s^k\bar{\rho}^k]$. Then we can solve for $s^k\bar{\rho}^k$,

$$s^k\bar{\rho}^k = f'(\bar{x}^{k+1})^{-1}[\bar{y}^k + \frac{1}{2} B[\bar{\xi}^k, s^k\bar{\rho}^k, s^k\bar{\rho}^k]].$$

But $s^k\bar{\rho}^k = H^{k+1}\bar{y}^k$ so that

$$H^{k+1}\bar{y}^k = f'(\bar{x}^{k+1})^{-1}[\bar{y}^k + \frac{1}{2} B[\bar{\xi}^k, s^k\bar{\rho}^k, s^k\bar{\rho}^k]]. \tag{3.10}$$

Now because the iterates $\{\bar{x}^k\}$ converge, $\lim_{k\to\infty} s^k\bar{\rho}^k = 0$, and applying this to (3.10) we have

$$\lim_{k\to\infty} \| [H^{k+1} - f'(\bar{x}^{k+1})^{-1}]\bar{y}^k \| = 0.$$

Let us now show that the improvement in the inverse of the approximate Jacobian matrix can actually be calculated from the requirement that $H^{k+1}\bar{y}^k = s^k\bar{\rho}^k$ where $\bar{y}^k = f(\bar{x}^{k+1}) - f(\bar{x}^k + (t^k - s^k)\bar{\rho}^k)$. If D^k is the improvement matrix at the kth step, then we have $H^{k+1} = H^k + D^k$.

Thus $H^{k+1}\bar{y}^k = H^k\bar{y}^k + D^k\bar{y}^k$ where $H^{k+1}\bar{y}^k = s^k\bar{\rho}^k$, so that

$$D^k\bar{y}^k = s^k\bar{\rho}^k - H^k\bar{y}^k.$$

One such solution for D^k is $D^k = [s^k\bar{\rho}^k - H^k\bar{y}^k]\bar{z}^{kT}$ where $\bar{z}^{kT}$ must

satisfy $\bar{z}^{k^T}\bar{y}^k = 1$. A more general solution would be

$$D^k = [s^k\bar{\rho}^k]\bar{q}^{k^T} - H^k\bar{y}^k\bar{z}^{k^T} \tag{3.11}$$

where $\bar{q}^{k^T}\bar{y}^k = \bar{z}^{k^T}\bar{y}^k = 1$. Thus we would calculate H^{k+1} using the formula

$$H^{k+1} = H^k - H^k\bar{y}^k\bar{z}^{k^T} + [s^k\bar{\rho}^k]\bar{q}^{k^T}. \tag{3.12}$$

As a particular case of the general solution, let

$$\bar{z}^{k^T} = -\bar{q}^{k^T} = \frac{\bar{\rho}^{k^T}H^k}{\bar{\rho}^{k^T}H^k\bar{y}^k}$$

which gives

$$H^{k+1} = H^k - \frac{[H^k\bar{y}^k + s^k\bar{\rho}^k]\bar{\rho}^{k^T}H^k}{\bar{\rho}^{k^T}H^k\bar{y}^k}.$$

This particular solution also arises in another quite natural way. Recall that the condition B^{k+1} must satisfy is

$$\bar{y}^k = f(\bar{x}^k + t^k\bar{\rho}^k) - f(\bar{x}^k + (t^k - s^k)\bar{\rho}^k) = s^k[B^{k+1}\bar{\rho}^k].$$

This relates the change in $f(\bar{x})$ to the change in $\bar{x}$ in the direction $\bar{\rho}^k$. Since there is no information available about changes in any

direction other than $\bar{\rho}^k$, let $B^{k+1}\bar{q}^k = B^k\bar{q}^k$ whenever $\bar{q}^{kT}\bar{\rho}^k = 0$. With this in mind, solve (3.8) for B^{k+1}

$$B^{k+1} = B^k + \frac{\bar{y}^k\bar{\rho}^{kT}}{\bar{\rho}^{kT}[s^k\bar{\rho}^k]} - \frac{B^k[s^k\bar{\rho}^k]\bar{\rho}^{kT}}{\bar{\rho}^{kT}[s^k\bar{\rho}^k]} \tag{3.13}$$

$$= B^k + \frac{[\bar{y}^k - B^k[s^k\bar{\rho}^k]]\bar{\rho}^{kT}}{\bar{\rho}^{kT}[s^k\bar{\rho}^{kT}]} \quad .$$

Now to solve for H^{k+1}, use Householder's formula which states that if A and $A + \bar{x}\bar{y}^T$ are nonsingular where $\bar{x},\bar{y} \in R^n$, then

$$[A + \bar{x}\bar{y}^T]^{-1} = A^{-1} - \frac{A^{-1}[\bar{x}\bar{y}^T]A^{-1}}{1 + \bar{y}^TA^{-1}\bar{x}} \quad .$$

Applying this to (3.13) with $A = B^k$, $\bar{x} = \frac{\bar{y}^k - B^k[s^k\bar{\rho}^k]}{\bar{\rho}^{kT}[s^k\bar{\rho}^k]}$, and $\bar{y}^T = \bar{p}^{kT}$ shows that

$$H^{k+1} = H^k - \frac{H^k\left[\frac{\bar{y}^k + [H^k]^{-1}[s^k\bar{\rho}^k]}{\bar{\rho}^{kT}[s^k\bar{\rho}^k]}\right]\bar{\rho}^{kT}H^k}{1 + \bar{\rho}^{kT}H^k\left[\frac{\bar{y}^k - [H^k]^{-1}[s^k\bar{\rho}^k]}{\bar{\rho}^{kT}[s^k\bar{\rho}^k]}\right]}$$

$$= H^k - \frac{[H^k \underset{-}{y}^k + s^k \underset{-}{\rho}^k]\underset{-}{\rho}^{k^T} H^k}{\underset{-}{\rho}^{k^T}[s^k \underset{-}{\rho}^k] + \underset{-}{\rho}^{k^T} H^k \underset{-}{y}^k - \underset{-}{\rho}^{k^T}[s^k \underset{-}{\rho}^k]}$$

$$= H^k - \frac{[H^k \underset{-}{y}^k + s^k \underset{-}{\rho}^k]\underset{-}{\rho}^{k^T} H^k}{\underset{-}{\rho}^{k^T} H^k \underset{-}{y}^k}.$$

Another particular method is given by $\underset{-}{z}^{k^T} = -\underset{-}{q}^{k^T} = \frac{\underset{-}{y}^{k^T}}{\underset{-}{y}^{k^T}\underset{-}{y}^k}$, so that

$$H^{k+1} = H^k - \frac{[s^k \underset{-}{\rho}^k + H^k \underset{-}{y}^k]\underset{-}{y}^{k^T}}{\underset{-}{y}^{k^T}\underset{-}{y}^k}.$$

Again this method can be derived another way. Since nothing is given about $H^{k+1}\underset{-}{v}^k$ when $\underset{-}{v}^{k^T}\underset{-}{y}^k = 0$, let $H^{k+1}\underset{-}{v}^k = H^k \underset{-}{v}^k$ in such a case. With this in mind solve (3.9) for H^{k+1}

$$H^{k+1} = H^k - \frac{[s^k \underset{-}{\rho}^k + H^k \underset{-}{y}^k]\underset{-}{y}^{k^T}}{\underset{-}{y}^{k^T}\underset{-}{y}^k}.$$

However Broyden remarks that this method has proven to be unsatisfactory in practice.

A third method suggested is to let $\underset{-}{z}^{k^T} = \frac{\underset{-}{y}^{k^T} H^k}{\underset{-}{y}^{k^T} H^k \underset{-}{y}^k}$ and let $-\underset{-}{q}^{k^T} = \frac{\underset{-}{\rho}^{k^T}}{\underset{-}{\rho}^{k^T}\underset{-}{y}^k}$. Thus the formula for H^{k+1} is

$$H^{k+1} = H^k - \frac{H^k \bar{y}^k \bar{y}^{kT} H^k}{\bar{y}^{kT} H^k \bar{y}^k} - \frac{[s^k - \rho^k]\rho^{kT}}{\rho^{kT} \bar{y}^k} .$$

Clearly H^{k+1} is symmetric whenever H^k is symmetric.

Note that in each of the above three methods we were able to define H^{k+1} in terms of quantities which had already been calculated.

Freudenstein and Roth's Method

Another major drawback to Newton's method is the necessity of guessing an initial approximation near enough to the true solution. A poor initial estimate of the solution may create a divergent sequence of iterates. In actual practice, the complexity of the function often precludes the choosing of a "good" initial estimate. One procedure which may prevent divergence was discussed in the description of Broyden's method. Let $\bar{\rho}^k = -H^k f(\bar{x}^k)$ where $H^k = J(\bar{x}^k)^{-1}$ and then calculate the next iterate, $\bar{x}^{k+1}$, from the formula $\bar{x}^{k+1} = \bar{x}^k + t^k \bar{\rho}^k$ where $\bar{t}$ is chosen so that the norm of $f(\bar{x}^{k+1})$ is either minimized or less than the norm of $f(\bar{x}^k)$. Although the convergence of such a sequence has not been mathematically guaranteed, in practice this norm reduction procedure will occasionally provide convergence when the ordinary Newton's method does not.

Freudenstein and Roth [7] have another method which they have experimented with chiefly using second and third degree polynomials and have found to be quite useful. This method, which they call the parameter-perturbation procedure, makes excellent use of the technique

of permitting the root of one equation to be the first approximation to the root of a similar equation.

If $f(\bar{x}) = \bar{0}$ is the equation to solve, suppose that each f_i can be written as $f_i(\bar{x}) = \sum_{k=o}^{m_i} a_{ik}\phi_{ik}(\bar{x})$ for i=1,...,n where the a_{ik} are real numbers and the ϕ_{ik} are functions of $\bar{x}$. Now consider a second set of equations of the form $g_i^{(o)}(\bar{x}) = \sum_{k=1}^{m_i} q_{ik}^{(o)}\phi_{ik}(\bar{x})$ for i=1,2,...,n where a solution to $g^{(o)}(\bar{x}) = [g_1^{(o)}(\bar{x}),...,g_n^{(o)}(\bar{x})]^T = \bar{0}$ is known. We will now change $g^{(o)}$ into f by a finite number of successive small changes in the coefficients of the ϕ_{ik}'s.

If $g_i^{(j)}(\bar{x}) = \sum_{k=o}^{m_i} q_{ik}^{(j)}\phi_{ik}(\bar{x})$ where $q_{ik}^{(j)} = q_{ik}^{(o)} + (a_{ik}-q_{ik}^{(o)})\frac{j}{N}$ for i=1,...,n and j=1,...,N, then $g^{(N)}(\bar{x}) = [g_1^{(N)}(\bar{x}),...,g_n^{(N)}(\bar{x})]^T = f(\bar{x})$. Now use the known root of $g^{(o)}(\bar{x})$ as the initial approximation for Newton's method to a solution of $g^{(1)}(\bar{x}) = \bar{0}$. Compute the solution to $g^{(1)}(\bar{x}) = \bar{0}$ and use it as the initial approximation in Newton's method to solve $g^{(2)}(\bar{x}) = \bar{0}$. Continue in this way until a solution of $g^{(N)}(\bar{x}) = f(\bar{x})$ is obtained. If the changes in the coefficients of the ϕ_{ik}'s are "small enough", the initial approximations from one step to the next will be good enough for convergence. Actually a more general algorithm than this can be obtained. Let $f(\bar{x})$ and $g^{(o)}(\bar{x})$ be functions of any form where a solution of $g^{(o)}(\bar{x}) = \bar{0}$ is known, then define a sequence of functions by

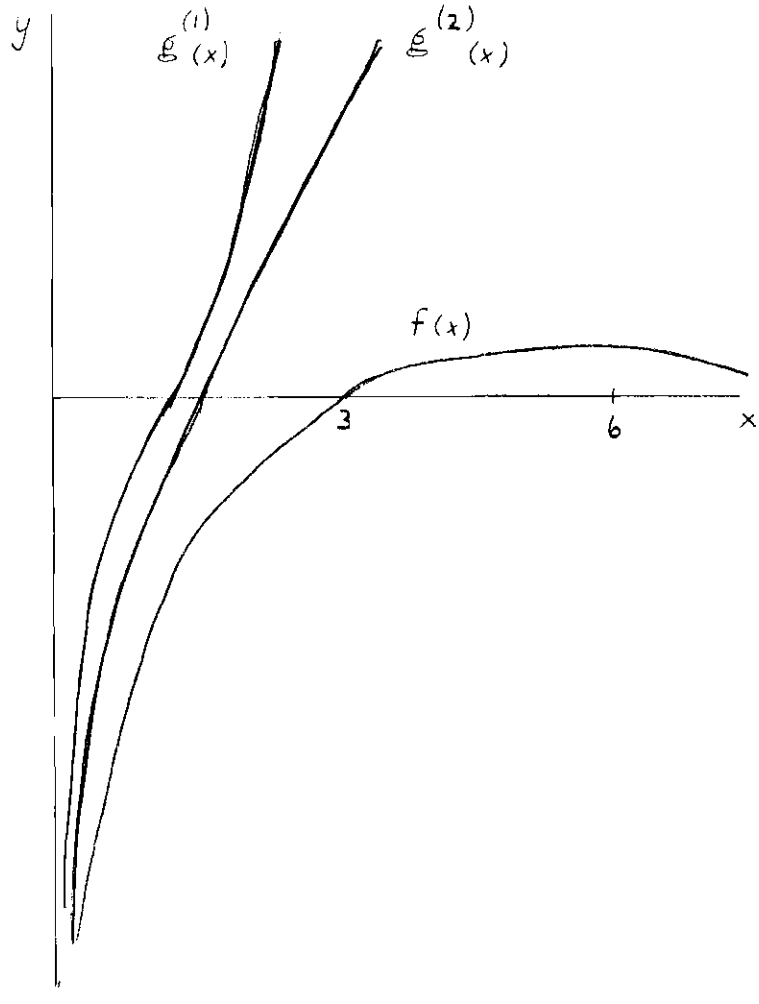
$$g^{(j)}(\bar{x}) = g^{(o)}(\bar{x}) + (f(\bar{x}) - g^{(o)}(\bar{x}))\frac{j}{N} \quad \text{for } j=1,...,N.$$

Now proceed as before.

Of course, $q_{ik}^{(j)} - q_{ik}^{(j-1)}$ need not be equal for all j=1,...,N as described, but rather these step sizes may be varied in any desired manner. For instance if a step size is too large for convergence, it may be halved or quartered or whatever is necessary for the previous root to lie within the radius of convergence of Newton's method for the next equation. Thus each step may require several smaller steps to implement it. In fact the optimal step size may not be known in advance, and so the method may have to be continually modified as it proceeds. Another problem that may arise with this procedure is that the Jacobian matrix may vanish. At the jth step, if the determinant of the Jacobian matrix falls below some predetermined level, change the increments of each $q_{ik}^{(j)}$ individually so that the value of the determinant is increased above the set value and then reintroduce the regular variation.

The usefulness of the parameter-perturbation procedure can be seen in the following example. Let $f(x) = \frac{x-3}{x^2}$ for $x>o$. This function has a root at $x = 3$ and takes on its maximum value of $\frac{1}{12}$ at $x = 6$ On the interval $(6,\infty)$, the function decreases and approaches zero asymptotically. Thus for $x^o = 7$, Newton's method produces a sequence of iterates which increases without bound and thus diverges. However if we employ the parameter-perturbation procedure with $g^{(o)} = x^2$ and $n = 3$, we will arrive at the solution $x = 3$. The data are summarized in the following table and graph.

$g^{(1)}(x) = \frac{2x^2}{3} + \frac{x-3}{3x^2}$	$g^{(2)}(x) = \frac{x^2}{3} + \frac{2x-6}{3x^2}$	$g^{(3)}(x) = f(x)$
$x^0 = 7$	$x^0 = 1$	$x^0 = 1.34$
$x^1 = 3.5$	$x^1 = 1.25$	$x^1 = 1.82$
$x^2 = 1.76$	$x^2 = 1.34$	$x^2 = 2.33$
$x^3 = 1.05$		$x^3 = 2.76$
$x^4 = 1$		$x^4 = 2.97$
		$x^5 = 3.00$



BIBLIOGRAPHY

1. Barnes, J. G. P., "An Algorithm for Solving Non-Linear Equations Based on the Secant Method," *Computer Journal*, Vol. 8, 1965, pp. 66-72.

2. Bartle, R. G., "Newton's Method in Banach Spaces," *Proceedings of the American Mathematical Society*, Vol. 6, 1955, pp. 827-831.

3. Broyden, C. G., "A Class of Methods for Solving Nonlinear Simultaneous Equations," *Mathematics of Computation*, Vol. 19, 1965, pp. 577-593.

4. Broyden, C. G., "Quasi-Newton Methods and Their Application to Function Minimization," *Mathematics of Computation*, Vol. 21, 1967, pp. 368-381.

5. Davidon, William C., "Viable Metric Methods for Mimization," *A.E.C. Research and Development Report*, ANL-5990, 1959.

6. Dennis, J. E., Jr., "On Newton-Like Methods," *Numerische Mathematik*, Vol. 11, 1968, pp. 324-330.

7. Freudenstein, Ferdinand and Bernard Roth, "Numerical Solution of Systems of Nonlinear Equations," *Journal of the Association of Computing Machinery*, Vol. 10, 1963, pp. 550-556.

8. Henrici, Peter, *Elements of Numerical Analysis*, John Wiley and Sons, Inc., New York, 1964.

9. Isaacson, Eugene and Herbert B. Keller, *Analysis of Numerical Methods*, John Wiley and Sons, Inc., New York, 1966.

10. Jeeves, T. A., "Secant Modification of Newton's Method," *Communications of the Association of Computing Machinery*, Vol. 1, 1958, pp. 9-10.

11. Kantorovich, L. U., and G. P. Akilov, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.

12. Kizner, William, "A Numerical Method for Finding Solutions of Nonlinear Equations," *Journal of the Society for Industrial and Applied Mathematics*, Vol. 12, 1964, pp. 424-428.

13. Lohr, L., and L. B. Rall, "Efficient Use of Newton's Method," *I.C.C. Bulletin*, Vol. 6, 1967, pp. 99-103.

14. Milne-Thomson, L. M., *Calculus of Finite Differences,* Macmillan and Co., Limited, London, 1951.

15. Nashed, M. Z., "Some Remarks on Variations and Differentials," *American Mathematical Monthly, Slought Memorial Papers,* Vol. 73, Number 4, April, 1966.

16. Rall, L. B., "Convergence of the Newton Process to Multiple Solutions," *Numerische Mathematik,* Vol. 9, 1966, pp. 23-37.

17. Rudin, Walter, *Principles of Mathematical Analysis,* McGraw Hill, New York, 1964.

18. Traub, J. F., *Iterative Methods for the Solution of Equations,* Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1964.

19. Turner, L. R., "Solution of Nonlinear Systems," *Annals of the New York Academy of Sciences,* Vol. 86, 1960, pp. 817-827.

20. Vandergraft, James S., "Newton's Method for Convex Operators in Partially Ordered Spaces," *SIAM Journal of Numerical Analysis,* Vol. 4, 1967, pp. 406-432.

21. Weeg, Gerard P., and Georgia B. Reed, *Introduction to Numerical Analysis,* Blaisdell Publishing Company, Waltham, Massachusetts, 1966.

22. Wendroff, Burton, *Theoretical Numerical Analysis,* Academic Press, New York, 1966.

23. Wolfe, Phillip, "The Secant Method for Simultaneous Nonlinear Equations," *Communications of the Association of Computing Machinery,* Vol. 2, 1959, pp. 12-13.