EXPLOITING SPARSITY

IN NEWTON-LIKE METHODS

by

Earl Stanley Marwil

TR78-335

EXPLOITING SPARSITY IN NEWTON-LIKE METHODS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment for the Degree of

Doctor of Philosophy

by

Earl Stanley Marwil

January 1978

Biographical Sketch

The author was born in Bartlesville, Oklahoma on July 31, 1950. He spent most of his youth there, attended the public schools, and graduated from Sooner High School in Bartlesville in May 1968. In May 1971, he graduated from the University of Oklahoma with high honors, having earned a Bachelor of Science degree in mathematics. He then began graduate study in the Center for Applied Mathematics at Cornell University. In May 1974, he was awarded a Master of Science; and he completed his Doctor of Philosophy at Cornell in January 1978.

To my friend, Denise

Acknowledgements

The author is very grateful to Cornell University for the financial support it has given and for the use of its facilities. My most sincere thanks go to my parents for their love and generosity during my graduate study. I would like to thank the members of my special committee for their patience: Larry Payne, minor member, Al Schatz, who served as chairman of the special committee, and John Dennis, who advised me on this thesis. I am very thankful to John for the valuable comments, for the time, and for the encouragement and understanding, which he has given so freely. Also, a special thanks to Jorge Moré, who served on the special committee and advised me on parts of this thesis when he was at Cornell. Finally I wish to thank Denise, to whom this work is dedicated, for "the eggs".

Table of Contents

1. Introduction

1.1 General.

Nonlinear problems occur in physics, engineering, economics, ecology, and many other fields. Let $F: \mathbb{R}^n \to \mathbb{R}^n$ be a nonlinear function on n-dimensional real linear space. The problem of solving a system of nonlinear equations is to find a solution to the equation

$$F(x) = 0. \tag{1.1.1}$$

Numerical analysts seek solutions by iterative techniques.

The classical Newton's method for solving (1.1.1) is a linearization of the Taylor expansion of F. Given a starting point or initial approximation, x_0, to the solution, the iteration can be defined by

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k). \tag{1.1.2}$$

<u>Theorem</u> 1.1.1 [27] If there is a point $x^* \in D_0$ an open, convex set such that $F(x^*) = 0$, F is differentiable in D_0, F' is Lipschitz continuous in D_0, and $F'(x^*)$ is nonsingular, then the Newton iteration, (1.1.2), converges quadratically to x^*, a solution of (1.1.1), for any $x_0 \in D_0$.

Quadratic convergence means for a positive constant α, the sequence $\{x_k\}$ behaves as

$$||x_{k+1} - x^*|| \leq \alpha||x_k - x^*||^2. \tag{1.1.3}$$

A Newton-like method replaces the Jacobian, F', in (1.1.2) by an invertible operator. Given $x_0 \in \mathbb{R}^n$ and a sequence of invertible operators $\{A_k\}$,

$$x_{k+1} = x_k - A_k^{-1}F(x_k) \tag{1.1.4}$$

defines an iterative process which approximates Newton's method. The invertible operator A_k is meant to approximate the Jacobian at x_k, $F'(x_k)$.

There are numerous articles that have been written about iterations of the form (1.1.4) and other generalizations of Newton's method. See [33] for an excellent bibliography including those topics. Recently a subclass of iterations of the form (1.1.4) has received a lot of attention. This was begun by Davidon [7], Fletcher and Powell [18], and Broyden [2]. The idea is to update the approximate Jacobian at each iteration to obtain the next approximate Jacobian.

Given x_0 and B_0 , an initial approximation to the Jacobian at x_0, define $\{x_k, B_k\}$ iteratively by

$$x_{k+1} = x_k - B_k^{-1}F(x_k) \tag{1.1.5}$$

and

(1.1.6) B_{k+1} = a function of B_k, $x_{k+1}-x_k$, and $F(x_{k+1})-F(x_k)$.

This class of methods are frequently called quasi-Newton methods. The reason for this is the requirement that B_{k+1} satisfy the matrix-vector equation:

$$B_{k+1}(x_{k+1} - x_k) = F(x_{k+1}) - F(x_k). \tag{1.1.7}$$

This equation is called the quasi-Newton equation. Newton's method is based on the truncated Taylor expansion:

$$F'(x_{k+1})(x_{k+1} - x_k) = F(x_{k+1}) - F(x_k).$$

For x_{k+1} close to the solution, $F(x_{k+1}) = 0$, and $F'(x_k)$ in place of $F'(x_{k+1})$ defines Newton's method. Equation (1.1.7) uses B_{k+1} in place of $F'(x_{k+1})$ and in one dimension it reduces to the secant equation. The quasi-Newton methods are also called secant methods.

The quasi-Newton methods exhibit good local convergence properties [4], [15] and behave well in practice for a large variety of problems [15]. The type of convergence usually exhibited is superlinear. That is, for $\{x_k\}$

(1.1.8) $$||x_{k+1} - x^*|| \leq \alpha_k ||x_k - x^*||$$

with $\alpha_k \underset{k\to\infty}{\to} 0$. Many authors have discussed various techniques to enlarge the region where an initial guess will lead to convergence.

Problems of the form (1.1.1) which have a special structure will be considered here. Intuitively, a system of nonlinear equations is <u>sparse</u> if each equation depends on only a few variables in the domain space. A linear problem is considered to be sparse when the matrix has fewer than 10% nonzero entries [16].

The quasi-Newton methods are usually implemented replacing (1.1.5) by:

(1.1.9) Solve $B_k s_k = -F(x_k)$ for s_k,

(1.1.10) then set $x_{k+1} = x_k + s_k$.

This is done for numerical stability.

Since there is a linear system to be solved at each iteration, it is very natural to define sparse nonlinear equations as those which have fewer than 10% nonzeroes in the Jacobian matrix. In the methods discussed here linear systems with the same zero-nonzero structure or sparseness structure occur at each iteration.

1.2 Examples.

Specific types of nonlinear problems are abundant. They include power flow systems [3], integrating stiff ordinary differential equations [5], nonlinear electrical networks [5], flow networks in physiological models, finite analog of nonlinear boundary value problems [37], certain econometric models, and fixed point problems [34].

Example 1.2.1 A two point boundary value problem, Rheinboldt [34]. Consider the differential equation

$$-\frac{d}{dt}\left[c\left(\frac{du}{dt}\right)^2\right] = f(t) \quad \text{for } t \in (0,1) \tag{1.2.1}$$

and where

$$u(0) = u(1) = 0, \tag{1.2.2}$$

$$c(s) = 1 + s^2, \quad \text{and}$$

$$f(t) = 600(1 - 2t)^2 + 20.$$

Rheinboldt notes that this problem exhibits properties typical of elliptical boundary value problems in higher dimensions. The exact solution to (1.2.1) is

$$u(t) = 10t(1-t).$$

On a uniform mesh $t_j = jh$, $j = 0,1,\ldots,n+1$, $h = (n+1)^{-1}$, a standard approximation of (1.2.1) is of the form

$$(1.2.3) \qquad A(x) = b, \quad x \in \mathbb{R}^n,$$

where $A(x) = (a_{ij}(x))_{n\times n}$, with

$$(1.2.4) \qquad a_{ij}(x) = \begin{cases} (\alpha_i - \alpha_{i-1})/h, & \text{for } j = i, \\ \alpha_k/h, & \text{for } |j-i| = 1, \ k = \min(i,j), \\ 0, & \text{otherwise,} \end{cases}$$

for $i,j = 1,2,\ldots,n$, and where

$$x_0 = x_{n+1} = 0$$

and $$\alpha_i = c[x_{i-1} - x_i)/h] .$$

Here, x_i is the i^{th} component of the vector x.

<u>Example</u> 1.2.2 A problem of heated networks, Porsching [28], which generates a sparse nonlinear system.

Although in the context that is considered in [28] the end result is in the form of the fixed point iteration, the underlying nonlinear system is sparse. The equations governing the network flow are:

$$(1.2.5) \qquad \begin{aligned} Aw &= b \\ B(w)h &= q \\ w &= DA^T p + r(h). \end{aligned}$$

The following definitions identify the matrices and vectors occurring in (1.2.5).

$A \in \mathbf{R}^n \times \mathbf{R}^m$ is the incidence matrix of positive flow in arcs k at nodes i,

$$a_{ik} = \begin{cases} \pm 1, & \text{depending on direction;} \\ 0, & \text{if not connected.} \end{cases}$$

$w \in \mathbf{R}^n$ is the vector of unknown flow rates.

$b \in \mathbf{R}^n$ is a known vector of the boundary flow rates.

$$B(w) \in \mathbf{R}^{n+1} \times \mathbf{R}^{n+1}, \text{ with}$$

$$b_{ij}(w) = \begin{cases} -|w_k|, & \text{if arc k connects nodes i and j and flow is toward i,} \\ 0, & \text{otherwise.} \end{cases}$$

for $i \neq j$, and $$b_{ii} = -\sum_{j \neq i} b_{ij} .$$

$h \in \mathbf{R}^{n+1}$ is the vector of unknown enthalpy (a measure of thermal energy).

q is a known vector of the thermal energies of the boundary flows.

$p \in \mathbf{R}^n$ is a vector of pressures.

$$r(h) = (\ldots, c_k \rho((h_i - h_j)/2), \ldots)^T,$$

with ρ as the fluid density.

$$D = \mathrm{diag}(d_k).$$

The constants c_k, d_k are determined by the Laminar Flow Law, which is an equation of the form

$$w_k = d_k(p_i - p_j) + c_k \rho((h_i + h_j)/2), \quad k = 1,2,\ldots,m,$$

with the flow in arc k from node i to node j.

The nonlinear system (1.2.5) is sparse if A is sparse.

The preceding examples give an indication of the problems which lead to a system of sparse nonlinear equations. In example 1.2.1, when the mesh size becomes small, the dimension of the nonlinear difference equations is large. In many large scale problems the system is likely to be sparse. The choice of the difference formula and the ordering of the nodes of the mesh in higher order problems can be made so that each equation involves only a few variables.

1.3 Literature.

Tewarson and Stephenson [37] state some considerations for design and selection of methods to solve large systems of nonlinear equations. These considerations include storage requirements, rate of convergence, the domain of attraction and choosing the initial guess, computational instability and large programming costs. Algorithms are developed which require only a few of the equations at a time. These are based on a discretized Newton method and have as a particularly useful application flow network problems.

Porsching [28] presents two source problems, including Example 1.2.2, and discusses some methods for solution of sparse nonlinear equations. For the network flow problem, a continuation method is applied to the system posed as a fixed point problem. He also gives an elliptic boundary value problem arising in nonlinear diffusion and discusses solution by accelerating the Gauss-Seidel method by overrelaxation.

In Kevorkian and Snoek [25] large scale steady state design problems are posed as a system of nonlinear equations. Graph theoretic arguments are used to develop a tearing procedure. Reid [32] gives FORTRAN subroutines, adapting the Levenberg-Marquardt algorithm for least squares, to solve nonlinear equations. Rheinboldt [34] discusses the underlying reasons for some of the difficulties encountered in large dimensional problems. Another aspect of that paper is the discussion of some techniques for reducing the dimension of the nonlinear system.

The idea of adapting quasi-Newton methods to take advantage of sparsity was first published in 1970 by Schubert [36]. He modified Broyden's method [2] to preserve sparsity in the approximate Jacobian. Broyden [3] later gave a local and linear convergence result for Schubert's method.

Powell [29] suggested using a variational method to develop algorithms to minimize functions which have many zero second derivatives. This is reiterated in his survey paper [31]. Toint [38] has applied the variational techniques as in Greenstadt [23]. This is further discussed in Chapter 3.

For general surveys, Rheinboldt [33] contains good background material, an overview of many different types of available methods, and an extensive bibliography. For a discussion of quasi-Newton methods, the survey by Dennis and Moré [15] reviews the development of those methods, gives convergence results and suggests many open problems. The recent paper by Duff [16] has a comprehensive bibliography for sparse linear equations.

1.4 Organization.

In Chapter 2 superlinear convergence is shown for Schubert's method; a Kantorovich analysis is also given. Chapter 3 contains the development of a symmetric version of Schubert's update. An algorithm is given and a linear convergence result is proved. In Chapter 4 a matrix factorization is used for the solution of nonlinear equations. The Schubert update formula is used to update a triangular factor while the other factor is not changed for a finite number of iterations. Continuity of the factorization is shown before proving the local and superlinear convergence of the algorithms. Chapter 5 suggests some further applications of these sparse techniques to other problems. Also, the findings are summarized.

All of the algorithms described here are considered in the context of being used locally. New notation will be introduced as required. Some frequently quoted theorems are stated in the next section for reference.

1.5 Frequently quoted theorems.

Theorem 1.5.1 Bounded Deterioration Theorem, Broyden, Dennis, and Moré [4]. (The reference for the theorem as stated here is [15].)

Let $F: \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in an open convex set D_0. Let $x^* \in D_0$ be such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular. Suppose F' is Lipschitz continuous at x^* : there exists $\kappa > 0$ such that

$$||F'(x) - F'(x^*)|| \leq \kappa||x - x^*|| \ , \quad x \in D_0.$$

If there exist constants α_1 and $\alpha_2 \geq 0$ for which

(1.5.1) $||B_{k+1}-F'(x^*)|| \leq [1+\alpha_1\sigma(x_k,x_{k+1})]||B_k-F'(x^*)|| + \alpha_2\sigma(x_k,x_{k+1})$

with $\sigma(x_k,x_{k+1}) = \max(||x_k-x^*||,||x_{k+1}-x^*||)$, then there exist $\epsilon,\delta > 0$ such that if $x_0 \in D_0$ and $||x_0-x^*|| < \epsilon$ and $||B_0-F'(x^*)|| < \delta$ tben the iteration (1.1.5) and (1.1.6) is well defined and converges linearly to x^*.

<u>Theorem</u> 1.5.2 Banach Perturbation Lemma [27].

Let $A, C \in L(\mathbb{R}^n)$ and assume A is invertible, with $||A^{-1}|| \leq \alpha$. If $||A - C|| \leq \beta$ and $\beta\alpha < 1$, then C is also invertible and

$$||C^{-1}|| \leq \frac{\alpha}{1-\alpha\beta} .$$

<u>Theorem</u> 1.5.3 Superlinear convergence [14].

Let $F\colon \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in D_0, open, convex. Suppose there exists $x^* \in D_0$ such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular. Let $\{B_k\}$ be a sequence of nonsingular matrices in $L(\mathbb{R}^n)$. If there exists an $x_0 \in D_0$ such that $\{x_k\}$, defined by

$$x_{k+1} = x_k - B_k^{-1}F(x_k) \quad k = 0,1,\ldots$$

remains in D_0 and $x_k \neq x^*$ for any k, and $x_k \to x^*$. Then $\{x_k\}$ converges superlinearly to x^* if and only if

$$(1.5.2) \qquad \lim_{k\to\infty} \frac{||[B_k - F'(x^*)](x_{k+1} - x_k)||}{||x_{k+1} - x_k||} = 0.$$

2. Local and Superlinear Convergence of Schubert's Method

2.1 Preliminaries.

For continuously differentiable F, Broyden's method [2] for solving (1.1.1) generates a sequence of matrices which approximate the Jacobian. Given x_0 and B_0 nonsingular, define the iteration:

$$x_{k+1} = x_k - B_k^{-1}F(x_k), \text{ and} \tag{2.1.1}$$

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T}{s_k^T s_k}, \tag{2.1.2}$$

where $y_k = F(x_{k+1}) - F(x_k)$ and $s_k = x_{k+1} - x_k$. The matrix B_k is corrected or updated to obtain the next approximate Jacobian B_{k+1}. The formula (2.1.2) is called the Broyden update; it replaces (1.1.6) in the general formulation of the quasi-Newton methods.

<u>Definition</u> 2.1.1 For vectors u,v $\mathbb{R}^n$, define the subspace $Q_{v,u}$ by $Q_{v,u} = \{A \in L(\mathbb{R}^n) : Au = v\}$. This is the quotient of v by u.

Broyden's update is a projection of B_k onto Q_{y_k,s_k}. This observation is due to the fact [15] that B_{k+1} is the unique solution to

$$\min \{ \|\hat{B} - B_k\|_F : \hat{B} \in Q_{y_k,s_k} \}, \tag{2.1.3}$$

where $\|\cdot\|_F$ is the Frobenius norm. Another way of viewing the Frobenius norm of a matrix in $L(\mathbb{R}^n)$ is as the ℓ_2 norm of a vector in $\mathbb{R}^{n^2}$. The convergence properties of Broyden's method are given in the following theorem. Let $J(x) = J_F(x)$ be the Jacobian of F at the point x.

<u>Theorem</u> 2.1.2 [15] Suppose $F: \mathbb{R}^n \to \mathbb{R}^n$ satisfies (2.1.4) and (2.1.5):

(2.1.4)(a) F is continuously differentiable in an open convex set D_0.

(b) There exists an $x^* \in D_0$ such that $F(x^*) = 0$ and $J(x^*)$ is nonsingular.

(2.1.5) There exists a constant, $\kappa > 0$ such that

$$||J(x) - J(\bar{x})||_F \leq \kappa ||x - \bar{x}||_2 \quad \text{for all } x, \bar{x} \in D_0.$$

Then there exist $\epsilon, \delta > 0$ such that if $x_0 \in D_0$ and B_0 nonsingular satisfy $||x_0 - x^*||_2 < \epsilon$ and $||B_0 - J(x^*)||_F < \delta$ then Broyden's method generates $\{B_k\}$ with B_k nonsingular and $\{x_k\}$ which converges to x^* and convergence is superlinear in the sense of (1.1.8).

The update (2.1.2) does not take advantage of any zero-nonzero structure. The second term on the right in (2.1.2) is a vector outer-product, and is in general a full matrix.

2.2 Schubert's update.

Schubert [36] proposed a modification of Broyden's method to accommodate systems of nonlinear equations with sparse Jacobian. Following the definition for linear problems [16], a Jacobian will be considered "sparse" if it has less than 10% nonzero entries. Recall that the iteration (2.1.1) and (2.1.2) would be implemented as (1.1.9), (1.1.10), and (2.1.2). The solution of a linear system with B_k as the coefficient matrix is required. The pattern of zero and nonzero entries, or sparsity pattern, in the Jacobian is assumed to be fixed for all x.

<u>Definition</u> 2.2.1 For $j = 1,2,\ldots,n$, define $Z_j \subset \mathbb{R}^n$ by

$$(2.2.1)\quad Z_j = \{v \in \mathbb{R}^n : e_i^T v = 0 \text{ for all } i \text{ such that } [J(x)]_{ji} = 0 \text{ for all } x \in \mathbb{R}^n\}.$$

Z_j is the subspace determined by the sparsity pattern of the j^{th} row of $J(x)$.

Also, define the set of matrices Z that preserve the sparseness structure of J as

$$(2.2.2)\qquad Z \equiv \{A \in L(\mathbb{R}^n) : A^T e_j \in Z_j \quad \text{for } j = 1,2,\ldots,n\}.$$

A matrix, B, is said to preserve the sparseness structure of J if $B \in Z$.

Corresponding to each Z_j, there is a projection operator that maps $\mathbb{R}^n$ onto Z_j.

<u>Definition</u> 2.2.2 For $j = 1,2,\ldots,n$, define $D_j \in L(\mathbb{R}^n)$ by

$$(2.2.3)\qquad D_j = \operatorname{diag}(d_{j1}, d_{j2}, \ldots, d_{jn}), \text{ where}$$

$$d_{ij} = \begin{cases} 0, & \text{if } [J(x)]_{ji} = 0 \text{ for all } x \in \mathbb{R}^n, \\ 1, & \text{otherwise.} \end{cases}$$

It is clear that D_j is a projection.

<u>Lemma</u> 2.2.3 Let $J(x)$ be the Jacobian of $F \in C^1(\mathbb{R}^n)$. Then for $j = 1,2,\ldots,n$,

$$(2.2.4)\qquad e_j^T J(x) = e_j^T J(x) D_j \quad \text{for all } x \in \mathbb{R}^n.$$

(The vector e_j is the j^{th} column of the n×n identity matrix.)

<u>Proof</u>: By Definition 2.2.1, $[e_j^T J(x)]^T \in Z_j$. For any $v \in Z$, $D_j v = v$.

Equation (2.2.4) follows.

To simplify notation for a given iteration, no subscript will denote the current iterate, and the subscript "plus", $_+$, will indicate the next iterate. For example, (2.1.2) would be

$$B_+ = B + \frac{(y - Bs)s^T}{s^T s} \tag{2.2.5}$$

The idea Schubert had was that the update should be made row-wise to preserve the sparsity pattern. The Schubert update can be summarized as

$$B_+ = B + \sum_{\substack{j=1 \\ s_j \neq 0}}^{n} e_j e_j^T \frac{(y - Bs)}{s_j^T s_j} s_j^T , \tag{2.2.6}$$

where $s_j = D_j s$. Note that for $s_j = 0$, the zero correction is added to the j^{th} row of B. For a scalar $\alpha \in \mathbb{R}$, the notation of the pseudo-inverse is a superscript "plus":

$$\alpha^+ = \begin{cases} \alpha^{-1}, & \text{if } \alpha \neq 0. \\ 0, & \text{if } \alpha = 0. \end{cases}$$

The update (2.2.6) can be rewritten as

$$B_+ = B + \sum_{j=1}^{n} (s_j^T s_j)^+ e_j^T (y - Bs)\, e_j s_j^T. \tag{2.2.7}$$

Lemma 2.2.4 Let $F \in C^1(\mathbb{R}^n)$. Let $B \in Z$; $x_+, x \in \mathbb{R}^n$. Let $s = x_+ - x$ and $y = F(x_+) - F(x)$. If B_+ is defined by (2.2.7), then $B_+ \in Q_{y,s} \cap Z$.

Proof: For $i = 1,2,\ldots,n$, multiply both sides of (2.2.7) by e_i^T to obtain

$$e_i^T B_+ = e_i^T B + (s_i^T s_i)^+ e_i^T (y - Bs)\, s_i^T . \tag{2.2.8}$$

Since $B \in Z$, it follows that $B^T e_i \in Z_i$. By definition, $s_i \in Z_i$. Linearity implies $B_+^T e_i \in Z_i$, and thus $B_+ \in Z$.

To prove $B_+ \in Q_{y,s}$, examine $B_+ s$ and y component by component. If $s_i \neq 0$, then

$$e_i^T B_+ s = e_i^T Bs + (s_i^T s_i)^+ \, e_i^T (y - Bs) \, s_i^T s \; . \tag{2.2.9}$$

Since s_i is a projection of s, $s_i^T s = s_i^T s_i$. For $s_i \neq 0$, $(s_i^T s_i)^+ = (s_i^T s_i)^{-1}$ by definition, and equation (2.2.9) reduces to

$$e_i^T B_+ s = e_i^T Bs + e_i^T (y - Bs) = e_i^T y. \tag{2.2.10}$$

If $s_i = 0$, then by the Mean Value Theorem [27], there exists $\bar{t}_i \in (0,1)$ such that

$$\begin{aligned} e_i^T y &= e_i^T [F(x_+) - F(x)] \\ &= e_i^T J(x + \bar{t}_i s) s \\ &= e_i^T J(x + \bar{t}_i s) s_i, \quad \text{because } J \in Z, \\ &= 0, \quad \text{since } s_i = 0. \end{aligned} \tag{2.2.11}$$

Thus, for $i = 1,2,\ldots,n$, $e_i^T B_+ s = e_i^T y$, which implies that $B_+ s = y$. Therefore $B_+ \in Q_{y,s}$.

The Schubert update is the unique solution to a minimization problem. The update B_+ can be viewed as the projection of B onto $Q_{y,s} \cap Z$.

<u>Theorem</u> 2.2.5 Let $B \in Z$; $y,s \in \mathbb{R}^n$ with $s \neq 0$. Define B_+ by (2.2.7). Then B_+ is the unique solution to $\min\{\|\hat{B} - B\|_F : \hat{B} \in Q_{y,s} \cap Z\}$.

<u>Proof</u>: Let $\hat{B} \in Q_{y,s} \cap Z$. Then $\hat{B}s = y$ and

$$(2.2.12) \qquad ||B_+ - B||_F = ||\sum_{i=1}^{n} (s_i^T s_i)^+ e_i^T (\hat{B}-B) s_i \; e_i s_i^T ||_F$$

$$= (\sum_{i=1}^{n} |(s_i^T s_i)^+ e_i^T (\hat{B}-B) s_i|^2 \; ||s_i||_2^2)^{\frac{1}{2}}$$

$$\leq (\sum_{i=1}^{n} ||e_i^T (\hat{B}-B)||_2^2)^{\frac{1}{2}} = ||\hat{B} - B||_F .$$

Uniqueness follows since $f: L(\mathbb{R}^n) \to \mathbb{R}^n$ defined by $f(A) = ||A-B||_F$ is strictly convex and $Q_{y,s} \cap Z$ is convex.

2.3 Local and superlinear convergence.

In 1971, Broyden gave a local convergence proof for Schubert's method [3]. He conjectured that the method was superlinearly convergent based on numerical results. A different, simple proof of local convergence is by an application of the Bounded Deterioration Theorem of Broyden, Dennis and Moré [4]. Although superlinear convergence was first proved for Broyden's method in [4], the techniques for the proof in [13] are extended to Schubert's method here.

The Bounded Deterioration Theorem gets its name from the hypothesis on the behavior of the approximate Jacobians (1.5.1). Define $E = B - J(x^*)$ and $E_+ = B_+ - J(x^*)$ for $B, B_+ \in L(\mathbb{R}^n)$. Then (1.5.1) may be written as

$$(2.3.1) \qquad ||E_+|| \leq [1 + \alpha_1 \sigma(x,x_+)]||E|| + \alpha_2 \sigma(x,x_+).$$

The norm $||\cdot||$ is unspecified. The error $||E_+||$ may be worse than $||E||$, but only in a controlled way. The difficulty in applying the

Bounded Deterioration Theorem is to show for a particular norm that the estimate (2.3.1) is satisfied for the update method generating the sequence of approximate Jacobians.

For Broyden's method, the Frobenius norm was one for which (2.3.1) was proved in [13]. That is the same norm in which Broyden's update was a minimum norm correction. See (2.1.3). Theorem 2.2.5 shows that Schubert's update is also a minimum norm correction in the Frobenius norm. The following theorem provides an estimate for the Schubert update in the Frobenius norm. This is the crux of the subsequent convergence results.

Theorem 2.3.1 Let $B, J \in Z$ and $y, s \in \mathbb{R}^n$ with $s \neq 0$. If B_+ is defined by (2.2.7), then

$$(2.3.2) \quad ||B_+-J||_F^2 \le ||B-J||_F^2 - \frac{||(B-J)s||_2^2}{||s||_2^2} + \sum_{i=1}^{n} (s_i^T s_i)^+ [e_i^T (y-Js)]^2.$$

Proof: By definition of the Frobenius norm,

$$||B_+-J||_F^2 = \sum_{i,j=1}^{n} [e_i^T (B_+-J) e_j]^2.$$

Substituting for B_+ using (2.2.7) gives

$$(2.3.3) \quad ||B_+-J||_F^2 = \sum_{i,j=1}^{n} [e_i^T (B-J) e_j + (s_i^T s_i)^+ e_i^T (y-Bs)\, s_i^T e_j]^2$$

$$= \sum_{i,j=1}^{n} \{[e_i^T (B-J) e_j]^2 + 2 e_i^T (B-J) e_j (s_i^T s_i)^+ e_i^T (y-Bs) e_j^T s_i$$
$$+ [(s_i^T s_i)^+]^2 [e_i^T (y-Bs)]^2 (e_j^T s_i)^2\}$$

$$= ||B-J||_F^2 + 2\sum_{i=1}^{n}(s_i^Ts_i)^+e_i^T(B-J)s_i\; e_i^T(y-Bs) + \sum_{i=1}^{n}(s_i^Ts_i)^+[e_i^T(y-Bs)]^2.$$

The second equality follows by expanding the quadratic; the third is by summing each term over $j = 1,2,\ldots,n$. Now, add and subtract Js in both the second and third terms. Then,

$$(2.3.4)\quad ||B_+-J||_F^2 = ||B-J||_F^2 + 2\sum_{i=1}^{n}(s_i^Ts_i)^+e_i^T(B-J)s_i\;[e_i^T(y-Js) + e_i^T(B-J)s]$$

$$+ \sum_{i=1}^{n}(s_i^Ts_i)^+[e_i^T(y-Js) + e_i^T(J-B)s]^2$$

$$= ||B-J||_F^2 + \sum_{i=1}^{n}(s_i^Ts_i)^+\{2e_i^T(B-J)s_i[e_i^T(y-Js)] + 2e_i^T(B-J)s_i[e_i^T(B-J)s]$$

$$+ [e_i^T(y-Js)]^2 + 2e_i^T(y-Js)[e_i^T(J-B)s] + [e_i^T(J-B)s]^2\}.$$

Since $B,J \in Z$, $(B-J)^Te_i \in Z_i$, and $e_i^T(B-J)s_i = e_i^T(B-J)s$. The first and fourth terms inside the summation cancel; the second and fifth combine to yield

$$(2.3.5)\quad ||B_+-J||_F^2 = ||B-J||_F^2 + \sum_{i=1}^{n}(s_i^Ts_i)^+\{[e_i^T(y-Js)]^2 - [e_i^T(B-J)s]^2\}.$$

Finally observe that $||s_i||_2 \le ||s||_2$ implies that $-(s_i^Ts_i)^+ \le -||s||_2^{-2}$. Therefore

$$(2.3.6)\quad ||B_+-J||_F^2 \le ||B-J||_F^2 - \frac{||(B-J)s||_2^2}{||s||_2^2} + \sum_{i=1}^{n}(s_i^Ts_i)^+[e_i^T(y-Js)]^2.$$

<u>Theorem</u> 2.3.2 Suppose $F\colon \mathbb{R}^n \to \mathbb{R}^n$ satisfies (2.1.4), and (2.3.7).

(2.3.7) There exists $K = (k_1,k_2,\ldots,k_n) \in \mathbb{R}^n$ with $k_i \ge 0$ for $i = 1,$

$2,\ldots,n,$ such that

$$||e_i^T[J(x) - J(\bar{x})]||_2 \le k_i||x - \bar{x}||_2 \quad \text{for all} \quad x,\bar{x} \in D_0.$$

Then there exist $\epsilon, \delta > 0$ such that if $x_0 \in D_0$ and B_0 nonsingular satisfy $||x_0 - x^*||_2 < \epsilon$ and $||B_0 - J(x^*)||_F < \delta$ then Schubert's method generates $\{B_k\}$ with B_k nonsingular and $\{x_k\}$ converges locally and superlinearly to x^*.

<u>Proof</u>: Let $\kappa = ||K||_2$; κ is a Lipschitz constant for J in the Frobenius norm. From Theorem 2.3.1 with J replaced by $J^* = J(x^*)$, $E = B - J^*$, and $E_+ = B_+ - J^*$,

$$(2.3.8) \qquad ||E_+||_F^2 \le ||E||_F^2 - \frac{||Es||_2^2}{||s||_2^2} + \sum_{i=1}^{n} (s_i^T s_i)^+[e_i^T(y-J^*s)]^2.$$

For each $s_i \neq 0$, estimate the term in the sum.

For $j = 1,2,\ldots,n$, define $\phi_j\colon \mathbb{R} \to \mathbb{R}$ by

$$(2.3.9) \qquad \phi_j(t) = e_j^T[-F(x) + F(x+ts) - tJ^*s].$$

Then $\phi_j(0) = 0$, $\phi_j(1) = e_j^T[-F(x) + F(x_+) - J^*s] = e_j^T(y - J^*s)$, and $\phi_j'(t) = e_j^T[J(x+ts)s - J^*s] = e_j^T[J(x+ts) - J^*]s_j$, since $J \in Z$. By the Mean Value Theorem [27], there exists $\bar{t}_j \in (0,1)$ such that $\phi_j(1) = \phi_j'(\bar{t}_j)$. Thus,

$$(2.3.10) \qquad \frac{|e_i^T(y-J^*s)|^2}{||s_i||_2^2} = \frac{|e_i^T[J(x+\bar{t}_i s) - J^*]s_i|^2}{||s_i||_2^2}$$

$$\le ||e_i^T[J(x+\bar{t}_i s) - J^*]||_2^2 \text{ , by the Cauchy-Schwartz inequality,}$$

$$\le k_i^2||x+\bar{t}_i s - x^*||_2^2 \text{ , by hypothesis,}$$

$\leq k_i^2 \sigma(x,x_+)^2$, since $\bar{t}_i \in (0,1)$,

and where $\sigma(x,x_+) = \max(||x - x^*||_2, ||x_+ - x^*||_2)$.

If any $s_i = 0$, then $0 = (s_i^T s_i)^+[e_i^T(y-J^*s)]^2 \leq k_i^2\sigma(x,x_+)^2$.

Therefore,

$$||E_+||_F^2 \leq ||E||_F^2 - \frac{||Es||_2^2}{||s||_2^2} + \kappa\sigma(x,x_+)^2 . \tag{2.3.11}$$

Disregarding the negative term, it follows that

$$||E_+||_F \leq ||E||_F + \kappa\,\sigma(x,x_+), \tag{2.3.12}$$

since if $a,b,c \geq 0$ and $a^2 \leq (b^2 + c^2) \leq (b + c)^2$ then $a \leq b + c$.

Inequality (2.3.1) is satisfied in the Frobenius norm with $\alpha_1 = 0$ and $\alpha_2 = \kappa$. The hypotheses of the Bounded Deterioration Theorem 1.5.1 [15] are satisfied and this proves local and linear convergence of Schubert's method.

Theorem 1.5.3 gives a necessary and sufficient condition for superlinear convergence. To show that (1.5.2) is satisfied for Schubert's method requires a more careful estimate of the right hand side of inequality (2.3.10). Observe that

$$||E(I - \frac{ss^T}{s^Ts})||_F^2 = ||E||_F^2 - \left(\frac{||Es||_2}{||s||_2}\right)^2 .$$

This and (2.3.11) give

$$||E_+||_F \leq [\,||E||_F^2 - \frac{||Es||_2^2}{||s||_2^2}\,]^{\frac{1}{2}} + \kappa\sigma(x,x_+). \tag{2.3.13}$$

The proof now proceeds as in Dennis and More [p. 58,15]. For $a,b > 0$ $(a^2 - b^2)^{\frac{1}{2}} \leq a - \frac{b^2}{2a}$ so that

$$(2.3.14) \quad \|E_+\|_F \leq \|E\|_F - (2\|E\|_F^{-1}) \frac{\|Es\|_2}{\|s\|_2}^2 + \kappa\sigma(x,x_+).$$

Let $\eta_k = \|E_k\|_F = \|B_k - J^*\|_F$, $\psi_k = \frac{\|E_k s_k\|_2}{\|s_k\|_2}$, $\sigma_k = \sigma(x_k,x_{k+1})$.

Inequality (2.3.14) is now

$$(2.3.15) \qquad \eta_{k+1} \leq [1 - (2\eta_k^2)^{-1} \psi_k^2]\eta_k + \kappa\sigma_k.$$

This gives $\eta_{k+1} \leq \eta_k + \kappa\sigma_k$. The sequence $\{x_k\}$ is linearly convergent which implies $\{\eta_k\}$ is bounded. Let η be an upper bound for $\{\eta_k\}$. Then

$$(2.3.16) \qquad (2\eta^{-1})\psi_k^2 \leq \eta_k - \eta_{k+1} + \kappa\sigma_k, \text{ and}$$

$$(2\eta^{-1}) \sum_{k=0}^{\infty} \psi_k^2 \leq \eta_0 + \kappa\sum_{k=0}^{\infty} \sigma_k$$

forcing $\lim_{k\to\infty} \psi_k = 0$. Condition (1.5.2) is satisfied and superlinear convergence of Schubert's method follows.

2.4 A Kantorovich analysis.

A general Newton-like method is of the form

$$(2.4.1) \qquad x_{k+1} = x_k - A_k^{-1}F(x_k)$$

with x_0 given and $\{A_k\}$ a sequence of invertible operators. Dennis [8] gave a Kantorovich analysis for the iteration (2.4.1). In [9] he studied

Broyden's method with a Kantorovich analysis. The Kantorovich analysis consists of constructing a scalar iteration to majorize the vector iteration. The scalar sequence is used to study the properties of the vector sequence.

The first lemma is a corollary to a theorem in [27]. The bound will be used in proving the Kantorovich theorem.

Lemma 2.4.1 Let $F \in C^1(D_0)$ for a convex $D_0 \subset \mathbb{R}^n$. Let F satisfy (2.3.7); let $\kappa = ||K||_2$. Then for any $x, y \in D_0$,

(2.4.2) $$||F(y) - F(x) - F'(x)(y-x)||_2 \leq \tfrac{1}{2}\kappa||y-x||_2^2.$$

Proof: In vector notation, with the integration occurring component-wise,

(2.4.3) $$F(y)-F(x)-F'(x)(y-x) = \int_0^1[F'((1-t)x+ty)-F'(x)](y-x)dt.$$

Thus,

(2.4.4) $$||F(y)-F(x)-F'(x)(y-x)||_2 \leq ||\int_0^1[F'((1-t)x+ty)-F'(x)](y-x)dt||_2$$

$$\leq \int_0^1||F'((1-t)x+ty)-F'(x)||_2||y-x||_2dt.$$

Now $||\cdot||_2 \leq ||\cdot||_F$. By definition of $||\cdot||_F$ and (2.3.7),

$$||F'(x)-F'(y)||_F \leq \kappa||x-y||_2.$$

Thus

(2.4.5) $$||F(y) - F(x) - F'(x)(y-x)||_2 \leq \int_0^1 \kappa t||y-x||_2^2\, dt$$

$$= \tfrac{1}{2}\kappa||y-x||_2^2.$$

The following is another technical lemma that is essential to the Kantorovich analysis. The proof depends on the estimate obtained in Theorem 2.3.1. The estimate in this lemma shows the relationship of the error of B_{k+1} as an approximation to $J(x_{k+1})$ to the error of the initial approximation and the total length of the steps taken.

Lemma 2.4.2 Let $F: \mathbb{R}^n \to \mathbb{R}^n$ satisfy (2.1.4)(a) and (2.3.7). Set $\kappa = ||K||_2$. Let $\{x_k\}$ and $\{B_k\}$ be generated by Schubert's method. If $\{x_i\}_{i=1}^{m+1} \subset D_0$, then

$$(2.4.6) \quad ||B_{m+1}-J(x_{m+1})||_F \le ||B_0-J(x_0)||_F + \frac{3}{2}\kappa \sum_{k=0}^{m} ||x_{k+1}-x_k||_2.$$

Proof: An application of Theorem 2.3.1 yields

$$(2.4.7) \quad ||B_{m+1}-J(x_{m+1})||_F^2 \le ||B_m-J(x_{m+1})||_F^2 - \frac{||(B_m-J(x_{m+1}))s_m||_2^2}{||s_m||_2^2} + \sum_{i=1}^{n} (s_i^T s_i)^+[e_i^T(y_m-J(x_{m+1})s_m)]^2$$

$$\le ||B_m-J(x_{m+1})||_F^2 + \sum_{i=1}^{n} (s_i^T s_i)^+[e_i^T(y_m-J(x_{m+1})s_m)]^2.$$

Here $s_i = D_i s_m$. The techniques in Lemma 2.4.1 can be applied to obtain $|e_i^T y_m - J(x_{m+1})s_m| \le \frac{1}{2}k_i||s_i||_2||s_m||_2$. Thus,

$$(2.4.8) \quad ||B_{m+1}-J(x_{m+1})||_F^2 \le ||B_m-J(x_{m+1})||_F^2 + \frac{1}{4}\sum_{i=1}^{n} k_i^2||s_m||_2^2$$

$$\le ||B_m-J(x_{m+1})||_F^2 + \frac{1}{4}||K||_2^2||x_{m+1}-x_m||_2^2, \text{ and}$$

$$(2.4.9) \quad ||B_{m+1}-J(x_{m+1})||_F \le ||B_m-J(x_{m+1})||_F + \frac{1}{2}\kappa||x_{m+1}-x_m||_2$$

$$\le ||B_m-J(x_m)||_F + ||J(x_{m+1})-J(x_m)||_F + \frac{1}{2}\kappa||x_{m+1}-x_m||_2$$

$$\leq ||B_m - J(x_m)||_F + \frac{3}{2}\kappa||x_{m+1} - x_m||_2,$$

since $||J(x_{m+1}) - J(x_m)||_F \leq \kappa||x_{m+1} - x_m||_2$. Therefore

$$||B_{m+1} - J(x_{m+1})||_F \leq ||B_0 - J(x_0)||_F + \frac{3}{2}\kappa\sum_{k=0}^{n}||x_{k+1} - x_k||_2.$$

The following theorem asserts the existence of a zero of F. The technique for proving the result is the Kantorovich analysis. The details are almost identical to the analysis of Broyden's method in [9].

__Theorem__ 2.4.3 Let $F: \mathbb{R}^n \to \mathbb{R}^n$ satisfy (2.1.4)(a) and (2.3.7). Let $\kappa = ||K||_2$ and let $B_0 \in L(\mathbb{R}^n)$ be nonsingular and suppose there exist constants $\beta, \delta, \eta > 0$ such that $||B_0^{-1}||_2 \leq \beta$, $||J(x_0) - B_0||_F \leq \delta$ and $||B_0^{-1}F(x_0)||_2 \leq \eta$.

(a) Define $h' = \dfrac{\beta\kappa\eta}{(1-\beta\delta)^2}$ and $r'_\pm = \dfrac{1 \pm (1-2h')^{\frac{1}{2}}}{\beta\kappa}(1-\beta\delta)$. If $\beta\delta < 1$ and $h' \leq \frac{1}{2}$ and $B(x_0, r'_-) \subset D_0$, then F has a root $x^* \in \bar{B}(x_0, r'_-)$ and if $h' < \frac{1}{2}$, x^* is unique in $D_0 \cap B(x_0, r'_+)$. The sequence $\{x'_k\}$ generated by

$$x'_{k+1} = x'_k - B_0^{-1}F(x'_k) \qquad (2.4.10)$$

converges to x^* from any $x'_0 \in D_0 \cap B(x_0, r'_+)$.

(b) Define $h = \dfrac{\beta\kappa\eta}{(1-3\beta\delta)^2}$ and $r = \dfrac{1-(1-8h)^{\frac{1}{2}}}{4\beta\kappa}(1-3\beta\delta)$. If $3\beta\delta < 1$, $h \leq \frac{1}{8}$ and $B(x_0, r) \subset D_0$, then Schubert's method converges to x^*.

__Proof__: (a) Let $x'_0 \in D_0$ and suppose $||x'_0 - x_0||_2 < r_0 \leq r'_+$. Consider the iteration (2.4.10), and note that the approximate Jacobian B_0 is used throughout the iteration. In the first part of the proof, (2.4.10) will be used to establish the existence and uniqueness of the root x^* of F. The convergence of the sequence $\{x_k\}$ is established at the same time.

Define a scalar function $f_1: \mathbb{R} \to \mathbb{R}$ by

$$(2.4.11) \qquad f_1(t) = \tfrac{1}{2}\kappa t^2 - (b_0 - \delta)t + b_0||B_0^{-1}F(x_0)||_2$$

where $b_0 = \beta^{-1}$. Define the scalar iteration

$$(2.4.12) \qquad t'_{k+1} = t'_k + \beta f_1(t'_k), \quad \text{for} \quad t'_0 \in [||x_0 - x'_0||_2, r_0).$$

The next step in the proof is to show that $\{t'_k\}$ majorizes the vector sequence $\{x'_k\}$. Let $t' \in [r'_-, r'_+)$. The proof proceeds by induction:

$$||x'_1 - x'_0||_2 = ||B_0^{-1}F(x_0)||_2 \le t'_1 - t'_0.$$

Now let $x'_k \in \overline{B}(x_0, t') \cap D_0$ and suppose that $t'_k \in (0, t')$ for $k = 0,1,\ldots,$ m, and

$$(2.4.13) \qquad ||x'_{k+1} - x'_k||_2 \le t'_{k+1} - t'_k \quad \text{for} \quad k = 0,1,\ldots,m-1.$$

Then $\sum_{k=0}^{m-1} ||x'_{k+1} - x'_k||_2 \le t'_m - t'_0$. From the vector iteration,

$$x'_{k+1} - x'_k = -B_0^{-1}F(x'_k), \quad \text{and}$$

$$0 = -F(x'_{k-1}) - B_0(x'_k - x'_{k-1}), \quad \text{so that}$$

$$x'_{m+1} - x'_m = -B_0^{-1}[F(x'_m) - F(x'_{m-1}) - B_0(x'_m - x'_{m-1})]. \quad \text{Thus,}$$

$$(2.4.14) \qquad ||x'_{m+1} - x'_m||_2 = ||B_0^{-1}[F(x'_m) - F(x'_{m-1}) - B_0(x'_m - x'_{m-1})]||_2$$

$$\leq \beta[\|F(x_m')-F(x_{m-1}')-J(x_{m-1}')(x_m'-x_{m-1}')\|_2 + (\|J(x_{m-1}')-J(x_0')\|_2$$

$$+ \|J(x_0')-J(x_0)\|_2 + \|J(x_0)-B_0\|_2)\|x_m'-x_{m-1}'\|_2]$$

$$\leq \beta[\tfrac{1}{2}\kappa\|x_m'-x_{m-1}'\|_2^2 + (\delta + \kappa\|x_0-x_0'\|_2 + \kappa\sum_{k=0}^{m-2}\|x_{k+1}'-x_k'\|_2)$$

$$\|x_m'-x_{m-1}'\|_2]$$

$$\leq \beta[\tfrac{1}{2}\kappa(t_m'-t_{m-1}')^2 + (\delta + \kappa t_{m-1}')(t_m'-t_{m-1}')],$$

by (2.4.13), and $\|x_0-x_0'\|_2 \leq t_0'$. Now, from (2.4.12) and (2.4.11),

$$\text{(2.4.15)} \qquad \frac{f_1(t_{m-1}')}{t_m'-t_{m-1}'} + f_1'(t_{m-1}') = b_0 + \kappa t_{m-1}' - (b_0 - \delta)$$

$$= \delta + \kappa t_{m-1}'.$$

Substituting (2.4.15) in (2.4.14) yields

$$\text{(2.4.16)} \quad \|x_{m+1}'-x_m'\|_2 \leq \beta[\tfrac{1}{2}\kappa(t_m'-t_{m-1}')^2 + f_1(t_{m-1}') + f_1'(t_{m-1}')(t_m'-t_{m-1}')]$$

$$\leq \beta f_1(t_m') = t_{m+1}' - t_m' .$$

Therefore $\sum_{k=0}^{m}\|x_{k+1}'-x_k'\|_2 \leq t_{m+1}'-t_0' < t'$ for all m. Thus the sequence $\{x_k'\}$ is Cauchy and must converge to some $x^* \in \bar{B}(x_0',t')$. By definition of $r_\pm$, r_- is the only solution of $f_1(t) = 0$ in the interval $[0,t']$; $t' + \beta f_1(t') \leq t'$ since $f_1(t) \leq 0$ for $t \in [r_-',r_+')$, with equality only for $t' = r_-'$. By a theorem in Kantorovich and Akilov [24], and quoted in [10], x^* is unique in $\bar{B}(x_0,r_-')$ and x_k' converges to x^* from any $x_0' \in \bar{B}(x_0,r_-')$.

If $h' < \frac{1}{2}$, then r'_+ and r'_- are distinct, and the same theorem implies x^* is unique in $D_0 \cap B(x_0, r'_+)$ and $\{x'_k\}$ converges to x^* from any $x'_0 \in D_0 \cap B(x_0, r'_+)$.

(b) The first step is to show that $\{B_k\}$ are invertible and to produce bounds on $||B_k^{-1}||_2$. Assume for $m \geq 0$ that

$$(2.4.17) \qquad \sum_{k=0}^{m} ||x_{k+1} - x_k||_2 < r.$$

By hypothesis, $||B_0^{-1}||_2 \leq \beta$ so that

$$(2.4.18) \qquad ||B_0^{-1}B_{m+1} - I||_F \leq \beta ||B_{m+1} - B_0||_F$$

$$\leq \beta[||B_{m+1} - J(x_{m+1})||_F + ||J(x_{m+1}) - J(x_0)||_F + ||J(x_0) - B_0||_F],$$

using the triangle inequality after adding and subtracting both $J(x_{m+1})$ and $J(x_0)$. Now Lemma 2.4.2 can be applied to the term $||B_{m+1} - J(x_{m+1})||_F$ yielding

$$(2.4.19) \qquad ||B_0^{-1}B_{m+1} - I||_F \leq \beta[2||J(x_0) - B_0||_F + \frac{3}{2}\kappa \sum_{k=0}^{m} ||s_k||_2$$

$$+ ||J(x_{m+1}) - J(x_0)||_F]$$

$$\leq \beta[2\delta + \frac{3}{2}\kappa \sum_{k=0}^{m} ||s_k||_2 + \gamma ||x_{m+1} - x_0||_2],$$

since $||J(x_0) - B_0||_F \leq \delta$ by hypothesis, and

$$||J(x_{m+1}) - J(x_0)||_F \leq (\sum_{i=1}^{n} ||e_i^T(J(x_{m+1}) - J(x_m))||_2^2)^{\frac{1}{2}}$$

$$\leq (\sum_{i=1}^{n} k_i^2 ||x_{m+1} - x_0||_2^2)^{\frac{1}{2}}$$

$$= \kappa ||x_{m+1}-x_0||_2 .$$

Applying the triangle inequality to $||x_{m+1}-x_0||_2$ gives

$$(2.4.20) \qquad ||B_0^{-1}B_{m+1}-I||_F \le 2\beta\delta + \frac{5}{2}\beta\kappa \sum_{k=0}^{m} ||s_k||_2$$

$$< 2\beta\delta + \frac{5}{2}\beta\kappa r , \quad \text{by (2.4.17)},$$

$$= 2\beta\delta + \frac{5}{8}[1-(1-8h)^{\frac{1}{2}}](1-3\beta\delta) ,$$

by definition of r. Now, $0 \le h \le \frac{1}{8}$ implies that $0 \le (1-8h) \le 1$ and $[1-(1-8h)^{\frac{1}{2}}] \le 1$. Then $c \equiv \frac{5}{8}[1-(1-8h)^{\frac{1}{2}}] < 1$. Therefore,

$$(2.4.21) \qquad ||B_0^{-1}B_{m+1}-I||_F < 3\beta\delta + c(1-3\beta\delta)$$

$$= (1-c)3\beta\delta + c$$

$$< (1-c) + c = 1,$$

since $3\beta\delta < 1$ by hypothesis. Since $||\cdot||_2 \le ||\cdot||_F$, (2.4.21) implies $||B_0^{-1}B_{m+1}-I||_2 < 1$ and by the Banach Lemma 1.5.2,

$$(B_0^{-1}B_{m+1})^{-1}B_0^{-1} = B_{m+1}^{-1} \quad \text{exists and}$$

$$(2.4.22) \qquad ||B_{m+1}^{-1}||_2 < \frac{\beta}{(1-2\beta\delta-\frac{5}{2}\beta\kappa\sum_{k=0}^{m}||s_k||_2)} .$$

The next step in the proof is to define the scalar iteration which majorizes the sequence $\{x_k\}$ produced by Schubert's method. Let $b_0 = \beta^{-1}$, $h_0 = \beta$ and set $t_0 = 0$. Define the scalar function $f: \mathbb{R} \to \mathbb{R}$ by

$$(2.4.23) \qquad f(t) = 2\kappa t^2 - (b_0 - 3\delta)t + b_0\eta.$$

Define $h_k = \dfrac{\beta}{(1-2\beta\delta-\frac{5}{2}\beta\kappa t_k)}$ and $b_k = h_k^{-1}$. Define the scalar iteration

$$(2.4.24) \qquad t_{k+1} = t_k + h_k f(t_k).$$

Next, the sequence $\{t_k\}$ is shown to be strictly increasing and bounded above, hence convergent. An induction is used to prove $\{t_k\}$ remains bounded. Observe that

$$t_1 = t_0 + h_0 f(t_0) = h_0 b_0 \eta = \eta,$$

and $\eta \geq ||B_0^{-1}F(x_0)||_2 = ||s_1||_2$. Since r is the smallest root of the quadratic (2.4.23) or equivalently $2\beta\kappa t^2 - (1-3\beta\delta)t + \eta$, elementary analysis shows that $r > \eta$. Thus

$$(2.4.25) \qquad r > ||s_1||_2.$$

Assume that $t_k \in (0,r)$ for $k = 1,2,\ldots,m$. Then by (2.4.24)

$$(2.4.26) \qquad r - t_{m+1} = r - t_m + h_m[f(r) - f(t_m)]$$

$$= (r-t_m)[1+h_m f'(\xi)], \quad \text{for} \quad \xi \in (t_m,r),$$

since $f(r) = 0$ and by an application of the Mean Value Theorem [27]. Next, the derivative is shown to be negative. Differentiating (2.4.23), and evaluating f' at ξ gives

$$(2.4.27) \qquad f'(\xi) = 4\ \xi - (b_0 - 3\delta)$$

$$< 4\kappa r - (b_0 - 3\delta)$$

$$= \left[1-(1-8h)^{\frac{1}{2}}\right](b_0-3\delta) - (b_0-3\delta), \quad \text{by definition of } r,$$

$$= -(1-8h)^{\frac{1}{2}}(b_0-3\delta) \leq 0.$$

This implies $1 + h_m f'(\xi) < 1$ and with (2.4.26) yields $r-t_{m+1} < r-t_m$. Also, $h_m \geq \beta = h_0$, so that

$$h_m f'(\xi) \geq -h_m(b_0-3\delta) \geq -h_0(b_0-3\delta) = -1+3\beta\delta > -1. \tag{2.4.28}$$

Thus $1 + h_m f'(\xi) > 0$, and $0 < r - t_{m+1} < r_0 - t_m$. Therefore $\{t_k\}$ is strictly increasing in $[0,r]$. Hence $\{t_k\}$ converges to $t^* \leq r$. If $t^* < r$, then $t^* = t^* + (1 - 2\beta\delta - \frac{5}{2}\beta\kappa t^*)^{-1} f(t^*)$ and $f(t^*)$ must be zero. Therefore $t^* = r$ since r is the smallest root of f in $[0,r]$.

It remains to show that $\sum_{k=0}^{m} ||s_k||_2 < r$ for all m. The first step of the induction is satisfied by (2.4.25). Suppose that $x_0, x_1, \ldots, x_m$ satisfy

$$||x_{k+1} - x_k||_2 \leq t_{k+1} - t_k \quad k = 0,1,\ldots,m-1. \tag{2.4.29}$$

Then $\sum_{k=0}^{m-1} ||s_k||_2 \leq t_m < r$ and the first part of the proof of (b) shows B_m^{-1} exists with $||B_m^{-1}||_2 \leq h_m$. To complete the induction, it remains to show that $||x_{m+1} - x_m||_2 \leq t_{m+1} - t_m$. For Schubert's method,

$$x_{m+1}-x_m = -B_m^{-1}F(x_m) \quad \text{and} \quad 0 = -F(x_{m-1})-B_{m-1}s_{m-1}.$$

Now,
$$x_{m+1}-x_m = -B_m^{-1}\left[F(x_m) - F(x_{m-1}) - B_{m-1}s_{m-1}\right],$$

and taking norms yields

$$(2.4.30)\qquad \|x_{m+1}-x_m\|_2 = \|B_m^{-1}[F(x_m)-F(x_{m-1})-B_{m-1}s_{m-1}]\|_2$$

$$\leq h_m[\|F(x_m)-F(x_{m-1})-J(x_{m-1})s_{m-1}\|_2 + \|J(x_{m-1})-B_{m-1}\|_2\|s_{m-1}\|_2]$$

$$\leq h_m[\tfrac{1}{2}\kappa\|x_m-x_{m-1}\|_2^2 + (\delta + \tfrac{3}{2}\kappa\sum_{k=0}^{m-2}\|s_k\|_2)\|x_m-x_{m-1}\|_2]$$

by an application of Lemmas 2.4.1 and 2.4.2. The induction hypothesis (2.4.29) with (2.4.30) gives

$$(2.4.31)\qquad \|x_{m+1}-x_m\|_2 \leq h_m[\tfrac{1}{2}\kappa(t_m-t_{m-1})^2 + (\delta + \tfrac{3}{2}\kappa t_{m-1})(t_m-t_{m-1})].$$

The scalar function f satisfies

$$(2.4.32)\qquad \frac{f(t_{m-1})}{(t_m-t_{m-1})} + f'(t_{m-1}) = b_{m-1} + f'(t_{m-1}), \quad \text{by (2.4.24)}$$

$$= b_0(1-2h_0\delta-\tfrac{5}{2}h_0\kappa t_{m-1}) + 4\kappa t_{m-1} - (b_0-3\delta),$$

by definition of h_{m-1}, b_{m-1}, h_0 and b_0, and evaluating f' at t_{m-1}. This simplifies to

$$(2.4.33)\qquad \frac{f(t_{m-1})}{(t_m-t_{m-1})} + f'(t_{m-1}) = b_0 + 2\delta - \tfrac{5}{2}\kappa t_{m-1} + 4\kappa t_{m-1} - b_0 + 3\delta$$

$$= \delta + \tfrac{3}{2}\kappa t_{m-1}.$$

Substituting this expression in (2.4.31) yields

$$(2.4.34)\qquad \|x_{m+1}-x_m\|_2 \leq h_m[\tfrac{1}{2}\kappa(t_m-t_{m-1})^2 + f(t_{m-1}) + f'(t_{m-1})(t_m-t_{m-1})]$$

$$\leq h_m[2\kappa(t_m-t_{m-1})^2 + f'(t_{m-1})(t_m-t_{m-1}) + f(t_{m-1})]$$

$= h_m f(t_m)$, by the Taylor expansion of f at t_m,

$= t_{m+1} - t_m$, by (2.4.18).

The induction is complete.

Therefore $\sum_{k=0}^{m} ||x_{k+1}-x_k||_2 \leq t_m < r$ for all m. The sequence $\{x_k\}$ is Cauchy and must converge to some $x^{**} \in \bar{B}(x_0,r)$. From Lemma 2.4.2,

(2.4.35) $$||B_m - J(x_m)||_2 \leq \delta + \frac{3}{2}\kappa \sum_{k=0}^{m} ||x_{k+1}-x_k||_2$$

$$\leq \delta + \frac{3}{2}\kappa t_m$$

and $$||B_m||_2 \leq ||B_m - J(x_m)||_2 + ||J(x_m)||_2$$

$$\leq ||J(x_m)||_2 + \delta + \frac{3}{2}\kappa t_m$$

$$\leq ||J(x_0)||_2 + \delta + \frac{5}{2}\kappa t_m$$

$$< ||J(x_0)||_2 + \delta + \frac{5}{2}\kappa r \equiv c, \text{ a constant.}$$

Hence,

(2.4.36) $$||F(x^{**})||_2 = \lim_{k\to\infty} ||F(x_k)||_2 = \lim_{k\to\infty} ||B_k||_2 ||x_{k+1}-x_k||_2$$

$$\leq c \lim_{k\to\infty} ||x_{k+1}-x_k||_2 = 0$$

Now, $r < r'_+$ and the uniqueness of x^* implies $x^{**} = x^*$.

3. Symmetric Schubert Update

3.1 Preliminaries.

Let $f: \mathbb{R}^n \to \mathbb{R}$ be given. Consider the problem of finding a local minimum of f on some open set $D_0 \subset \mathbb{R}^n$. Let $x^* \in D_0$ be the minimum. so that

$$f(x^*) = \min \{f(x): x \in D_0\}.$$

In the following, assume that f is twice continuously differentiable.

When solving this problem using a Newton-like method to find a zero of $g(x) = \nabla f(x)$, it is usually the case that an approximation to $J_g(x) = \nabla^2 f(x)$, the Jacobian of g or the Hessian of f, looks as much like J_g as possible. Since $\nabla^2 f$ is symmetric, it is approximated by a symmetric matrix.

<u>Definition</u> 3.1.1 Define the set of symmetric matrices S by
$S = \{A \in L(\mathbb{R}^n) : A = A^T\}$.

A Newton-like method, (1.1.4), for the solution of the minimization problem can be written as: given $x \in D_0$ and $B \in S$, nonsingular,

(3.1.1)
$$x_+ = x - B^{-1} g(x)$$
$$B_+ \in S \cap Q_{y,s},$$

where $y = g(x_+) - g(x)$ and $s = x_+ - x$. Specific methods for choosing a B_+ in $S \cap Q_{y,s}$ have the property that near the solution, B_+ is nonsingular.

Suppose that $\nabla^2 f$ is sparse. The quasi-Newton methods, (3.1.1) require the solution of a linear system, namely

$$\text{(3.1.2)} \qquad Bs = -g(x).$$

The objective is to define a method of updating B to preserve the sparseness structure of the Hessian. Some quasi-Newton methods for solving the minimization problem are designed to update $B^{-1} = H$ [15]. However, there is no guarantee that H will be sparse if B is. Here, only methods for updating the Hessian are considered. As in section 2.2, $\nabla^2 f$ will be considered sparse if it has less than 10% nonzeroes. The sparsity pattern is again assumed to be fixed for all x.

Assume further that

$$\text{(3.1.3)} \qquad [\nabla^2 f]_{ii} \neq 0 \quad \text{for} \quad i = 1,2,\ldots,n.$$

Otherwise, the i^{th} component of ∇f is linear in x_i, so that x_i can be written as a function of the remaining x_j, $j = 1,\ldots,i-1, i+1,\ldots,n$, and the dimension of the problem can be reduced.

3.2 Notation and technical lemmas.

Definition 3.2.1 Let $s \in \mathbb{R}^n$ and define the components of s by

$$\text{(3.2.1)} \qquad \rho_i = e_i^T s \quad \text{for} \quad i = 1,2,\ldots,n.$$

The subspaces Z_j, $j = 1,2,\ldots,n$, have the same meaning as in Definition 2.2.2 with J(x) now being replaced by $J_g(x)$. Instead of using the projection operators defined in Chapter 2, it will be more convenient, because of the complexity, to work with the individual components.

Definition 3.2.2 For $j = 1,2,\ldots,n$ define the ℓ_2 projection of $s \in \mathbb{R}^n$ onto Z_j by s_j with s_j defined component-wise by

(3.2.2)
$$e_i^T s_j = \begin{cases} \rho_i, & \text{if } [\nabla^2 f]_{ji} \neq 0, \\ 0, & \text{if } [\nabla^2 f]_{ji} = 0. \end{cases}$$

Lemma 3.2.3 Let $s \in \mathbb{R}^n$ and suppose $f: \mathbb{R}^n \to \mathbb{R}$ is C^2 and $[\nabla^2 f]_{ii} \neq 0$ for $i = 1,2,\ldots,n$. Then

(3.2.3)(a)
$$e_j^T s = e_j^T s_j,$$

(b)
$$e_j^T s \; e_i^T s_j = e_j^T s_i \; e_i^T s_j, \quad \text{and}$$

(c)
$$w = \sum_{j=1}^{n} e_i^T s_j \alpha_j e_j \in Z_i \text{ for } \alpha_j \in \mathbb{R},\ j = 1,2,\ldots,n.$$

Proof: By hypothesis, $[\nabla^2 f]_{jj} \neq 0$, which implies $e_j^T s_j = \rho_j = e_j^T s$.

(b): If $[\nabla^2 f]_{ji} \neq 0$, then by symmetry $[\nabla^2 f]_{ij} \neq 0$ so that $e_j^T s_i = \rho_j = e_j^T s$. Hence (3.2.3)(b) holds. If $[\nabla^2 f]_{ji} = 0$ then $e_i^T s_j = 0$ and both sides of (3.2.3)(b) are zero.

(c): $e_j^T w = e_i^T s_j \alpha_j = 0$ if $[\nabla^2 f]_{ij} = 0$. Therefore $w \in Z_i$.

Definition 3.2.4 Let $s \in \mathbb{R}^n$. Define $P \in L(\mathbb{R}^n)$ by

(3.2.4)
$$P = \tfrac{1}{2}[I - \sum_{j=1}^{n} (s_j^T s_j)^{+} e_j^T s \; s_j e_j^T].$$

Define $A, D \in L(\mathbb{R}^n)$ by

(3.2.5)
$$A = \sum_{j=1}^{n} \rho_j s_j e_j^T$$
$$= [\rho_1 s_1 \mid \rho_2 s_2 \mid \cdots \mid \rho_n s_n],$$

and

(3.2.6) $$D = \mathrm{diag}\,(s_1^Ts_1, s_2^Ts_2, \ldots, s_n^Ts_n)$$

Observe that A is the Frobenius projection of ss^T onto Z. Also, P can be written as

(3.2.7) $$P + \tfrac{1}{2}[I - AD^+].$$

<u>Definition</u> 3.2.5 For $v \in \mathbb{R}^n$ and $M \in L(\mathbb{R}^n)$ and an integer $0 < m \leq n$, define $M_m \in L(\mathbb{R}^m)$ to be the lower right $m \times m$ submatrix of $M = M_n$, and $v_m \in \mathbb{R}^m$ to be the last m components of v.

The matrix P is a projection operator. This is a consequence of the following lemma on the eigenvalues of P.

<u>Lemma</u> 3.2.6 The eigenvalues of P are in $\left[0, \max\left(\tfrac{1}{2}, 1 - \min\limits_{\rho_i \neq 0} \frac{\rho_i^2}{s_i^Ts_i}\right)\right]$

and the eigenvalues of $(I-P)$ are in $\left[\min\left(\tfrac{1}{2}, \min\limits_{\rho_i \neq 0} \frac{\rho_i^2}{s_i^Ts_i}\right), 1\right]$.

<u>Proof</u>: Without loss of generality, assume all the zero components of s are ordered first. That is,

$$\rho_i = 0, \quad i = 1,2,\ldots,\ell, \quad \text{and}$$

$$\rho_i \neq 0, \quad i = \ell+1,\ldots,n.$$

Then $A_{n-\ell}$ is the nonzero part of A, and $D_{n-\ell}$ is invertible by (3.2.3)(a) and Definition 3.2.5. Let $m = n-\ell$.

From (3.2.7), partitioning the matrix,

$$(3.2.8) \qquad P = \tfrac{1}{2}\left[I - \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & A_m D_m^{-1} \end{array}\right]\right]$$

$$= \left[\begin{array}{c|c} \tfrac{1}{2} I & 0 \\ \hline 0 & \tfrac{1}{2}[I_m - A_m D_m^{-1}] \end{array}\right]$$

It is clear that P has ℓ eigenvalues equal to $\frac{1}{2}$, and the remaining $n-\ell$ are the eigenvalues of $\frac{1}{2}[I_m - A_m D_m^{-1}]$. Similarly, $I-P$ has ℓ eigenvalues equal to $\frac{1}{2}$, and its other $n-\ell$ are the eigenvalues of $\frac{1}{2}[I_m + A_m D_m^{-1}]$.

It is now sufficient to consider P and $(I-P)$ of dimension m, so the subscripts are omitted temporarily. Observe that

$$P = \tfrac{1}{2}[I - AD^{-1}] \text{ and}$$

$$(I-P) = \tfrac{1}{2}[I + AD^{-1}]$$

are similar to

$$\tfrac{1}{2} D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}} \text{ and } \tfrac{1}{2} D^{-\frac{1}{2}}(D + A)D^{-\frac{1}{2}}$$

respectively.

Schnabel [35] observed that $D + A$ is the sum of a diagonal positive definite matrix and positive semi-definite rank one matrices. Thus,

$$(3.2.9)\quad D+A = 2\begin{bmatrix} \rho_1^2 & & & \\ & \rho_2^2 & & \\ & & \ddots & \\ & & & \rho_n^2 \end{bmatrix} + \sum_{\substack{i<j \\ [\nabla^2 f]_{ij} \neq 0}} \begin{array}{c} \begin{bmatrix} \rho_j^2 & \cdots & \rho_i\rho_j \\ \vdots & & \vdots \\ \rho_j\rho_i & \cdots & \rho_i^2 \end{bmatrix} \end{array}\begin{array}{c} \cdots i \\ \\ \cdots j \end{array} .$$

By the Interleaving Eigenvalue Lemma, see Wilkinson [pp. 95-98,41] for example,

$$(3.2.10)\qquad \min \text{ e.v. } (D + A) \geq \min_i 2\rho_i^2 > 0.^{(\dagger)}$$

Similarly $D - A$ can be written as a sum of positive semi-definite rank-one matrices

$$(3.2.11)\qquad D - A = \sum_{\substack{i<j \\ [\nabla^2 f]_{ij} = 0}} \begin{bmatrix} \rho_j^2 & \cdots & -\rho_i\rho_j \\ \vdots & & \vdots \\ -\rho_j\rho_i & \cdots & \rho_i^2 \end{bmatrix}\begin{array}{c} \cdots i \\ \\ \cdots j \end{array} .$$

So $\min \text{ e.v. } (D - A) \geq 0$.

Therefore P_m is positive semi-definite and

$$\min \text{ e.v. } (I - P)_m \geq \min_i \frac{\rho_i^2}{s_i^T s_i} > 0.$$

In general, $\min \text{ e.v. } P > 0$ and

$$\min \text{ e.v. } (I - P)_m \geq \min(\tfrac{1}{2}, \min_{i \neq 0} \frac{\rho_i^2}{s_i^T s_i}).$$

This gives the lower bounds.

The bounds on the largest eigenvalues follow from

(†) Toint [38] also showed that $D + A$ is positive definite.

$$x^Tx = x^T(I-P)x + x^TPx.$$

This completes the proof of Lemma 3.2.6.

Lemma 3.2.7 Let $u,v,s \in \mathbb{R}^n$ and P be given by (3.2.7). Then

(3.2.12) $$\left\langle D^+(I-P)^{-1}u,v\right\rangle = \left\langle D^+(I-P)^{-1}v,u\right\rangle.$$

Proof: Suppose that the zero components of s are ordered first. Then

$$I - P = \left[\begin{array}{c|c} \frac{1}{2}I_\ell & 0 \\ \hline 0 & \frac{1}{2}(I_m + A_mD_m^{-1}) \end{array}\right]$$

Now,

(3.2.13) $$\left\langle D^+(I-P)^{-1}u,v\right\rangle = \sum_{i=1}^{\ell}[2(s_i^Ts_i)^+e_i^Tu]e_i^Tv + \left\langle D_m^{-1}(I-P)_m^{-1}u_m,v_m\right\rangle .$$

Note that $(I-P)_mD_m = \frac{1}{2}(D_m + A_m)$ which is symmetric and positive definite by (3.2.10). Its inverse must be symmetric and positive definite also, so that

(3.2.14) $$\left\langle D_m^{-1}(I-P)_m^{-1}u_m,v_m\right\rangle = \left\langle u_m,D_m^{-1}(I-P)_m^{-1}v_m\right\rangle .$$

From (3.2.13) and (3.2.14),

(3.2.15) $$\left\langle D^+(I-P)^{-1}u,v\right\rangle = \sum_{i=1}^{\ell} e_i^Tu[2(s_i^Ts_i)^+e_i^Tv] + \left\langle u_m,D_m^{-1}(I-P)_m^{-1}v_m\right\rangle$$

$$= \left\langle u,D^+(I-P)^{-1}v\right\rangle ,$$

and the result follows.

Lemma 3.2.8 Let $s \in \mathbb{R}^n$, and P be defined by (3.2.7).

(a) If $v \in \mathbb{R}^n$ is such that $e_i^T v = 0$ for all i such that $s_i = 0$, then

$$(3.2.16) \qquad -\left\langle D^+(I-P)^{-1}v,v\right\rangle \le -\frac{v^Tv}{s^Ts}.$$

(b) If $u \in \mathbb{R}^n$, then

$$(3.2.17) \qquad \left\langle D^+(I-P)^{-1}u,u\right\rangle \le \max\,(2, \max_{\rho_j \neq 0} \frac{s_j^Ts_j}{\rho_j^2}) \sum_{i=1}^{n} (s_i^Ts_i)^+(e_i^Tu)^2 .$$

<u>Proof</u>: Again assume that the zero components of s are ordered first.

(a): Then,

$$(3.2.18) \qquad -\left\langle D^+(I-P)^{-1}v,v\right\rangle = -\sum_{i=1}^{\ell} 2(s_i^Ts_i)^+(e_i^Tv)^2 - \left\langle D_m^{-1}(I-P)_m^{-1}v_m,v_m\right\rangle$$

$$\le -\frac{2}{s^Ts} \sum_{\substack{i=1 \\ s_i \neq 0}}^{\ell} (e_i^Tv)^2 - ||(I-P)_mD_m||_2^{-1}v_m^Tv_m ,$$

since $s_i^Ts_i \le s^Ts$ and $\langle Mx,x\rangle \ge \frac{||x||_2^2}{||M||_2}$ for nonsingular M.

By Lemma 3.2.6, $||(I-P)_m||_2 \le 1$, so

$$(3.2.19) \qquad ||(I-P)_mD_m||_2 \le ||(I-P)_m||_2||D_m||_2$$

$$\le ||D_m||_2$$

$$\le \max_{\ell < j \le n} s_j^Ts_j$$

$$\le s^Ts.$$

Therefore, (3.2.18) and (3.2.19) yield

$$(3.2.20)\qquad -\left\langle D^{+}(I-P)^{-1}v,v\right\rangle \le -\frac{1}{s^Ts}\Big[2\sum_{\substack{i=0\\ s_i\ne 0}}^{\ell}(e_i^Tv)^2 + v_m^Tv_m\Big]$$

$$\le -\frac{1}{s^Ts}\Big[\sum_{\substack{i=1\\ s_i\ne 0}}^{\ell}(e_i^Tv)^2 + v_m^Tv_m\Big]$$

$$= -\frac{v^Tv}{s^Ts}\ ,$$

since for $s_i = 0$, $e_i^Tv = 0$.

(b): Direct evaluation gives

$$(3.2.21)\quad \left\langle D^{+}(I-P)^{-1}u,u\right\rangle = \sum_{i=1}^{\ell} 2(s_i^Ts_i)^{+}(e_i^Tu)^2 + \left\langle D_m^{-1}(I-P)_m^{-1}u_m,u_m\right\rangle$$

$$= \sum_{i=1}^{\ell} 2(s_i^Ts_i)^{+}(e_i^Tu)^2 + \left\langle 2(D_m+A_m)^{-1}u_m,u_m\right\rangle$$

$$\le 2\sum_{i=1}^{\ell}(s_i^Ts_i)^{+}(e_i^Tu)^2 + ||2D_m^{\frac12}(D_m+A_m)^{-1}D_m^{\frac12}||_2\,||D_m^{-\frac12}u_m||_2^2$$

$$= 2\sum_{i=1}^{\ell}(s_i^Ts_i)^{+}(e_i^Tu)^2 + ||(I-P)_m^{-1}||_2\left\langle D_m^{-1}u_m,u_m\right\rangle$$

$$\le 2\sum_{i=1}^{\ell}(s_i^Ts_i)^{+}(e_i^Tu)^2 + \max_{\rho_j\ne 0}\frac{s_j^Ts_j}{\rho_j^2}\left\langle D_m^{-1}u_m,u_m\right\rangle$$

by Lemma (3.2.6) and

$$||(I-P)_m^{-1}||_2 = \frac{1}{\min\ e.v.(I-P)_m}$$

for positive definite $(I-P)_m$. Therefore,

$$\left\langle D^+(I-P)^{-1}u,u\right\rangle \le \max\left(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\right) \sum_{i=1}^{n} (s_i^T s_i)^+ (e_i^T u)^2$$

completing the proof.

3.3 Symmetric Schubert update.

Powell [30] derived the Powell symmetric Broyden update from the Broyden update (2.1.2) by iteratively projecting B_+ into S and then projecting back to $Q_{y,s}$. Using the same technique, Dennis [11] derived most of the symmetric updates satisfying the linear equation $B_+ s = y$ from the rank-one updates

$$B_+ = B + \frac{(y-Bs)c^T}{c^T s} \tag{3.3.1}$$

with various choices of $c \in \mathbb{R}^n$.

Here, the iterative double projection technique is applied to Schubert's update, alternately projecting into S and back to $Z \cap Q_{y,s}$ to derive the symmetric Schubert update.

Let $B \in S$ be given. Define C_1 by Schubert's update (2.2.5),

$$C_1 = B + \sum_{j=1}^{n} (s_j^T s_j)^+ e_j^T (y-Bs)\, e_j s_j^T .$$

Set $$C_2 = \tfrac{1}{2}(C_1 + C_1^T),$$

$$\vdots$$

$$C_{2k+1} = C_{2k} + \sum_{j=1}^{n} (s_j^T s_j)^+ e_j^T (y-C_{2k}s)\, e_j s_j^T , \tag{3.3.2}$$

and $$C_{2k+2} = \tfrac{1}{2}(C_{2k+1} + C_{2k+1}^T)$$

$$= C_{2k} + \tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+e_j^T(y-C_{2k}s)(e_js_j^T + s_je_j^T) .$$

Let $G_k = C_{2k}$; then

(3.3.3) $$G_{k+1} = G_k + \tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+e_j^T(y-G_ks)(e_js_j^T + s_je_j^T),$$

and $$G_{k+1}s = G_ks + \tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+e_j^T(y-G_ks)(e_js_j^Ts_j + s_je_j^Ts),$$

and $$y-G_{k+1}s = y-G_ks - \tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+e_j^T(y-G_ks)(e_js_j^Ts_j + s_je_j^Ts).$$

Set $w_k = y-G_ks$. Then

$$w_{k+1} = w_k - \tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+[e_js_j^Ts_j + s_je_j^Ts]e_j^Tw_k$$

$$= [I - \tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+(e_js_j^Ts_j + s_je_j^Ts)e_j^T]w_k$$

$= Pw_k$, with P defined by (3.2.4)

$$= P^kw_0.$$

Now, $\lim_{k\to\infty} G_k = G_0 + \sum_{k=0}^{\infty}(G_{k+1} - G_k)$, and from (3.3.3),

$$\sum_{k=0}^{\infty}(G_{k+1} - G_k) = \sum_{k=0}^{\infty}\tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+(e_js_j^T + s_je_j^T)e_j^Tw_k$$

$$= \tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+(e_js_j^T + s_je_j^T)\, e_j^T\sum_{k=0}^{\infty}w_k$$

$$= \tfrac{1}{2}\sum_{j=1}^{n} (s_j^T s_j)^+ (e_j s_j^T + s_j e_j^T)\, e_j^T \sum_{k=0}^{\infty} P^k w_0 .$$

Hence,

$$\lim_{k\to\infty} G_k = B + \tfrac{1}{2}\sum_{j=1}^{n} (s_j^T s_j)^+ (e_j s_j^T + s_j e_j^T)\, e_j^T \left(\sum_{k=0}^{\infty} P^k\right)(y-Bs).$$

The update is given by

(3.3.4) $$B_+ = B + \tfrac{1}{2}\sum_{j=1}^{n} (s_j^T s_j)^+ e_j^T \lambda\, (e_j s_j^T + s_j e_j^T) ,$$

where $$\lambda = \left(\sum_{k=0}^{\infty} P^k\right)(y-Bs)$$

and P is defined by (3.2.4).

Lemma 3.2.6 implies that the maximum eigenvalue of P is less than 1, since $\min\limits_{\rho_i \neq 0} \dfrac{\rho_i^2}{(s_i^T s_i)} > 0$. Therefore, by the Neumann Lemma [27],

$$\sum_{k=0}^{\infty} P^k = (I-P)^{-1}$$

and λ is the solution of

(3.3.5) $$(I - P)\lambda = (y-Bs).$$

Observe that the matrix $(I-P)$ has the same sparseness structure as $\nabla^2 f$. Although $(I-P)$ is not symmetric we can rewrite the update so that the linear system is symmetric and positive definite. Again assume that the zero components of s are ordered first, then

(3.3.6) $$B_+ = B + \sum_{\substack{j=1\\ s_j\neq 0}}^{n} (e_j s_j^T + s_j e_j^T)\, e_j^T \hat{\lambda}\ ,$$

where $\hat{\lambda}$ is the solution to

(3.3.7) $$\left[\begin{array}{c|c} G_\ell & 0 \\ \hline 0 & D_m + A_m \end{array}\right] \hat{\lambda}_\ell = (y-Bs)\quad ,$$

and $G = \operatorname{diag}(\gamma_1,\gamma_2,\dots,\gamma_\ell)$ with γ_i defined by

$$\gamma_i = \begin{cases} s_i^T s_i, & \text{if } s_i\neq 0, \\ 1, & \text{if } s_i=0. \end{cases}$$

3.4 Properties of the update.

The first lemma of this section will verify that the update has the desired structure.

Lemma 3.4.1 Let $B\in S\cap Z$ and $y,s\in \mathbb{R}^n$. Then B_+ defined by (3.3.4) satisfies $B_+ \in S\cap Z\cap Q_{y,s}$.

Proof: Obviously $B_+\in S$. To show $B_+\in Z$, it is sufficient to check that $B^T e_i \in Z_i$ for each i. For $i = 1,2,\dots,n$,

(3.4.1) $$e_i^T B_+ = e_i^T B + \tfrac{1}{2}(s_i^T s_i)^+ e_i^T \lambda s_i^T + \tfrac{1}{2}\sum_{j=1}^{n} e_i^T s_j (s_j^T s_j)^+ e_j^T \lambda e_j^T\ .$$

The transposes of the first and second terms on the right are in Z_i; by Lemma 3.2.3(c), the transpose of the third term is too. Therefore, $B_+\in Z$.

To see that $B_+\in Q_{y,s}$, form the vector $B_+ s$ and check component-wise that $e_i^T B_+ s = e_i^T y$, $i = 1,2,\dots,n$. If $s_i \neq 0$,

$$(3.4.2) \qquad e_i^T B_+ s = e_i^T Bs + \tfrac{1}{2} e_i^T \lambda + \tfrac{1}{2} \sum_{j=1}^{n} e_i^T s_j (s_j^T s_j)^+ e_j^T \lambda e_j^T s$$

$$= e_i^T Bs + e_i^T (I-P)\lambda, \quad \text{from (3.2.4)},$$

$$= e_i^T y.$$

If $s_i = 0$, then

$$(3.4.3) \qquad e_i^T B_+ s = e_i^T B_+ s_i = 0.$$

By the Mean Value Theorem [27], there exist $\bar{t}_i \in (0,1)$ such that

$$e_i^T (g_+ - g) = e_i^T J_g (x + \bar{t}_i s) s .$$

Then $e_i^T y = e_i^T J_g (x + \bar{t}_i s) s_i = 0$ because $J_g \in Z$. Therefore $B_+ \in Q_{y,s}$ and the proof is complete.

The following estimate will be used to prove the convergence properties of the algorithm given in section 3.5.

<u>Theorem</u> 3.4.2 Let $B, J \in S \cap Z$ and $y, s \in \mathbb{R}^n$ with $s \neq 0$, and let B_+ be defined by (3.3.4). Then

$$(3.4.4) \qquad ||B_+ - J||_F^2 \le ||B-J||_F^2 - \frac{||(B-J)s||_2^2}{||s||_2^2}$$

$$+ \max\left(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\right) \sum_{i=1}^{n} (s_i^T s_i)^+ [e_i^T (y - Js]^2 .$$

<u>Proof</u>: Expand the expression for the Frobenius norm of $B_+ - J$.

(3.4.5) $$||B_+ - J||_F^2 = \sum_{i,j=1}^{n} [e_i^T(B_+ - J)e_j$$

$$= \sum_{i,j=1}^{n} [e_i^T(B-J)e_j + \tfrac{1}{2}(s_i^Ts_i)^+ e_i^T\lambda s_i^T e_j + \tfrac{1}{2}e_i^T s_j (s_j^T s_j)^+ e_j^T\lambda]^2$$

$$= \sum_{i,j=1}^{n} \{[e_i^T(B-J)e_j]^2 + 2e_i^T(B-J)e_j(s_i^Ts_i)^+ e_i^T\lambda e_j^T s_i$$

$$+ \tfrac{1}{2}[(s_i^Ts_i)^+(e_i^T\lambda)(e_j^Ts_i)]^2 + \tfrac{1}{2}(s_i^Ts_i)^+ e_i^T\lambda e_j^T s_i (s_j^Ts_j)^+ e_j^T\lambda e_i^T s_j\}$$

$$= ||B-J||_F^2 + \sum_{i=1}^{n} (s_i^Ts_i)^+ e_i^T\lambda\{2e_i^T(B-J)s + \tfrac{1}{2}e_i^T\lambda + \tfrac{1}{2}\sum_{j=1}^{n} (s_j^Ts_j)^+ e_j^T s_i e_i^T s_j e_j^T\lambda\}.$$

Now observe that

$$y-Bs = (I-P)\lambda = \tfrac{1}{2}\lambda + \tfrac{1}{2}[\sum_{j=1}^{n} (s_j^Ts_j)^+ e_j^T s\, s_j e_j^T]\lambda ,$$

and

(3.4.6) $$e_i^T(y-Bs) = \tfrac{1}{2}e_i^T\lambda + \tfrac{1}{2}\sum_{j=1}^{n}(s_j^Ts_j)^+ e_j^T s\, e_i^T s_j e_j^T\lambda$$

$$= \tfrac{1}{2}e_i^T\lambda + \tfrac{1}{2}\sum_{j=1}^{n} (s_j^Ts_j)^+ e_j^T s_i e_i^T s_j e_j^T\lambda , \quad \text{by (3.2.3)(b).}$$

Therefore, (3.4.6) applied to (3.4.5) yields

(3.4.7) $$||B_+-J||_F^2 = ||B-J||_F^2 + \sum_{i=1}^{n} (s_i^Ts_i)^+ e_i^T\lambda[2e_i^T(B-J)s + e_i^T(y-Bs)]$$

$$= ||B-J||_F^2 + \sum_{i=1}^{n} (s_i^Ts_i)^+ e_i^T\lambda[e_i^T(B-J)s + e_i^T(y-Bs)] .$$

To complete the proof, examine the sum on the right hand side of

(3.4.7). Let $u = y - Js$ and $v = (B - J)s$. Then

$$\lambda = (I-P)^{-1}(y-Bs) = (I-P)^{-1}(u-v),$$

and

$$(3.4.8) \qquad \sum_{i=1}^{n} (s_i^T s_i)^+ e_i^T \lambda [e_i^T(B-J)s + e_i^T(y-Js)]$$

$$= \sum_{i=1}^{n} (s_i^T s_i)^+ e_i^T (I-P)^{-1}(u-v)[e_i^T v + e_i^T u]$$

$$= \sum_{i=1}^{n} e_i^T D^+ (I-P)^{-1}(u-v)\; e_i^T(v + u)$$

$$= \left\langle D^+(I-P)^{-1}u, v\right\rangle - \left\langle D^+(I-P)^{-1}v, v\right\rangle + \left\langle D^+(I-P)^{-1}u, u\right\rangle - \left\langle D^+(I-P)^{-1}v, u\right\rangle$$

Now, substituting (3.4.8) into (3.4.7) gives

$$(3.4.9) \quad ||B_+ - J||_F^2 = ||B-J||_F^2 + \left\langle D^+(I-P)^{-1}u, v\right\rangle - \left\langle D^+(I-P)^{-1}v, v\right\rangle + \left\langle D^+(I-P)^{-1}u, u\right\rangle - \left\langle D^+(I-P)^{-1}v, u\right\rangle$$

$$= ||B-J||_F^2 - \left\langle D^+(I-P)^{-1}v, v\right\rangle + \left\langle D^+(I-P)^{-1}u, u\right\rangle,$$

by Lemma 3.2.7,

$$\leq ||B-J||_F^2 - \frac{v^T v}{s^T s} + \max\Big(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\Big) \sum_{i=1}^{n} (s_i^T s_i)^+ (e_i^T u)^2$$

$$= ||B-J||_F^2 - \frac{||(B-J)s||_2^2}{||s||_2^2} + \max\Big(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}\Big) \sum_{i=1}^{n} (s_i^T s_i)^+ [e_i^T(y-Js)]^2.$$

The inequality in (3.4.9) follows from Lemma 3.2.8 and since $s_i = 0$

implies $e_i^T v = e_i^T(B - J)s = e_i^T(B - J)s_i = 0$. The last equality is by substituting for u and v, and the proof is complete.

Finally, B_+ defined by (3.3.4) is the solution to

$$\min \{||\hat{B} - B||_F : \hat{B} \in S \cap Z \cap Q_{y,s}\}.$$

It is straightforward to apply the variational techniques [23], [22] to this constrained problem. The solution by Toint [38] to the problem posed in the variational form [31] is the same as the B_+ in (3.3.6).

3.5 An algorithm.

Let $x \in \mathbb{R}^n$, $B \in S \cap Z$, positive definite be given.

1. Solve $B\Delta x = -g$ for Δx, the step to be taken; $g = g(x)$.
2. Set $x_+ = x + \Delta x$.
3. Evaluate $g_+ = g(x_+)$; test for convergence.
4. Set $y = g_+ - g$.
5. For each i such that $|e_i^T \Delta x| < \dfrac{||\Delta x_i||_2}{M^{\frac{1}{2}}}$ with $M \geq 2$ fixed,

   set $\rho_i = 0$; otherwise $\rho_i = e_i^T \Delta x$. This defines s, the step used to update B.
6. Compute B_+ from $B, s,$ and y using (3.3.6) and (3.3.7).

The convergence theorem in the next section will show that B_+ is positive definite in the region of local convergence. An implementation of the algorithm might avoid singularity of the Hessian approximation by re-evaluating the Hessian $B_k = J_g(x_k)$ or by adding a positive diagonal matrix to a singular or nearly singular B_k. One possible choice for starting the iteration would be to take $B_0 = I$.

The test in step 5 means that no component of Δx should tend to zero faster than its corresponding projection. Also, this ensures that the eigenvalues of P, defined by (3.2.4), are not tending to 1 as $k\to\infty$, or those of the matrix associated with the linear system (3.3.7) are not tending to 0 as $k\to\infty$. This is critical for the conditioning of the linear systems involving λ or $\hat{\lambda}$, and it is also important in the convergence proof.

The test is about what would happen anyway on the machine. When the size of any component is less than machine precision times the size of the corresponding projection, that component is insignificant. In other words, for $M^{\frac{1}{2}} = \frac{1}{n\times\text{mach.eps.}}$, the algorithm is roughly the one carried out on the computer.

On the other hand, for small M, say $M = 2$, it is likely that many ρ_i are set to zero. In that case, the correction to B doesn't take much work. This is close to the idea of keeping the same approximation to the Hessian for several iterations [15].

In step 6, the form of the update (3.3.6) is preferable to (3.3.4) since the linear system (3.3.7) is symmetric. This, of course, requires some bookkeeping to do the permutation on the vector s and the corresponding permutations of B and y. However, the important block in that symmetric system has the same sparseness structure as the corresponding block of the permuted Hessian. Also, this is a possibly smaller system to solve.

3.6 Convergence.

<u>Theorem</u> 3.6.1 Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be $C^2(D_0)$ for D_0 open and convex;

assume there exists $x^* \in D_0$ such that $\nabla f(x^*) = 0$, $\nabla^2 f(x^*)$ is positive definite and $[\nabla^2 f(x)]_{ii} \neq 0$ for all $x \in D_0$. If there exists $K = (k_1, k_2, \ldots, k_n) \in \mathbb{R}^n$ with $k_i \geq 0$ such that

$$(3.6.1) \qquad \|e_i^T(\nabla^2 f(x) - \nabla^2 f(x^*))\|_2 \leq k_i \|x - x^*\|,$$

for $i = 1,2,\ldots,n$ and for all $x \in D_0$, then there exist $\epsilon, \delta > 0$ such that for (x_0, B_0), $B_0 \in Z \cap S$ which satisfy

$$\|B_0 - \nabla^2 f(x^*)\|_2 < \delta$$

and

$$\|x_0 - x^*\|_2 < \epsilon ,$$

then the symmetric Schubert method generates $\{B_k\}$ with B_k well-defined and $\{x_k\}$ which converges linearly to x^*.

<u>Proof</u>: Choose $M \geq 2$, and set $\kappa = \|K\|_2$. From Theorem 3.4.2 with $J = J^* = \nabla^2 f(x^*)$,

$$(3.6.2) \quad \|B_+ - J^*\|_F^2 \leq \|B - J^*\|_F^2 + \max(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}) \sum_{\substack{i=1 \\ s_i \neq 0}}^{n} \frac{[e_i^T(y - J^* s)]^2}{s_i^T s_i} .$$

For $s_i \neq 0$, an application of the Mean Value Theorem [27] gives

$$(3.6.3) \qquad \frac{|e_i^T(y - J^* s)|^2}{\|s\|_2^2} \leq k_i^2 \sigma(x, x+s)^2 ,$$

where $\sigma(x, x+s) = \max(\|x - x^*\|_2, \|x+s - x^*\|_2)$. Therefore,

$$(3.6.4) \qquad \max(2, \max_{\rho_j \neq 0} \frac{s_j^T s_j}{\rho_j^2}) \sum_{\substack{i=1 \\ s_i \neq 0}}^{n} \frac{[e_i^T(y - J^* s)]^2}{s_i^T s_i} \leq M \kappa^2 \sigma(x, x+s)^2$$

and this with (3.6.2) implies

(3.6.5) $$||B_+ - J^*||_F^2 \leq ||B - J^*||_F^2 + M\kappa^2 \sigma(x,x+s)^2$$

and

(3.6.6) $$||B_+ - J^*||_F \leq ||B - J^*||_F + \alpha\sigma(x,x+s)$$

where $\alpha = \kappa M^{\frac{1}{2}}$.

Now,

(3.6.7) $$\begin{aligned} ||x+s - x^*||_2 &\leq ||s||_2 + ||x - x^*||_2 \\ &\leq \sqrt{\tfrac{n}{M}}||\Delta x||_2 + ||x-x^*||_2 \\ &\leq \sqrt{\tfrac{n}{M}}[||x-x^*||_2 + ||x_+-x^*||_2] + ||x-x^*||_2 \\ &\leq (1+\sqrt{\tfrac{n}{M}})||x-x^*||_2 + \sqrt{\tfrac{n}{M}}||x_+-x^*||_2 , \end{aligned}$$

so that

(3.6.8) $$\begin{aligned} \sigma(x,x+s) &\leq (1+\sqrt{\tfrac{n}{M}})||x-x^*||_2 + \sqrt{\tfrac{n}{M}}||x_+-x^*||_2 \\ &\leq (1+\sqrt{\tfrac{n}{M}})\ \sigma(x,x_+). \end{aligned}$$

Thus,

(3.6.9) $$||B_+ - J^*||_F \leq ||B - J^*||_F + \hat{\alpha}\sigma(x,x_+) ,$$

where $\hat{\alpha} = (1+\sqrt{\frac{n}{M}})\alpha$. An application of Theorem 1.5.1 completes the proof.

In the preceding proof observe that since $||s|| \leq ||\Delta x||$, $\sqrt{\frac{n}{M}}$ could be replaced by 1, regardless of M. In general, though, by appropriate choice of M, $\sqrt{\frac{n}{M}} << 1$.

4. Matrix Factorizations

4.1 Basic considerations.

The linear system,

$$Ax = b \tag{4.1.1}$$

with nonsingular $A \in L(\mathbb{R}^n)$ and $b \in \mathbb{R}^n$, is solved by Gaussian elimination. The algorithms most frequently used are the Crout and Dolittle decompositions with partial pivoting. The matrix A is factored into a product of lower and upper triangular matrices:

$$PA = LU, \tag{4.1.2}$$

where P is the nonsingular pivot matrix. In the case where no pivoting is required, the linear system becomes

$$LUx = b. \tag{4.1.3}$$

Then the solution of (4.1.3) is accomplished by solving two triangular linear systems. First $Lw = b$ is solved for w by forward substitution, then $Ux = w$ is solved by backward substitution for x.

For rank-one and rank-two quasi-Newton methods, techniques are available [1], [20] for updating factors of the approximate Jacobian. Brodlie, Gourlay, and Greenstadt [1] show how to express rank-one and rank-two corrections to symmetric positive definite matrices which result in positive definite matrices in product form. This is applied to quasi-Newton methods for unconstrained minimization. The important fact is that the numerical stability of algorithms for those problems is increased. This is because the approximate Jacobian is guaranteed to be positive definite, provided the iteration began with a positive definite matrix; the effect of rounding errors on updating formulae which theoretically maintain positive definiteness is minimized. Gill and Murray [20] propose application of their techniques to the solution of nonlinear equations in addition to modifying the factorization of the Hessian to preserve positive definiteness. In the former case, they recommend using an orthogonal factorization of the approximate Jacobian on the basis of numerical stability.

Gill et al. [19] describe in general, modifying Cholesky factors and QR factors of a matrix following a rank-one change. In a later paper, Gill et al. [21] give methods to compute the LDV factoriazation of a matrix B and the modified factors after the deletion or addition of a column to B.

Broyden [3] proposed dealing with just the factors of a sparse Jacobian. However, the Schubert update formula adds n rank-one corrections and the symmetric version adds n rank-two corrections. There seems to be little hope for a simple method to update factors of the Schubert or the symmetric Schubert methods.

Gill et al. [19] note that "the factors of a sparse matrix are more likely to be sparse than its inverse." Keeping in mind the idea that the Schubert method involves solving sparse linear systems at each iteration, and these systems have the same structure, new algorithms are suggested. These algorithms are based on an LU factorization of the approximate Jacobian. The Schubert update is applied to one of the factors, and preserves the triangular structure along with any additional sparseness structure. Several algorithms are proposed using the Crout and Dolittle factorizations. While we have done no numerical testing on these algorithms, Moré [26] has implemented a version based on updating R of the QR decomposition, and has obtained encouraging preliminary numerical results.

Let $A = (a_{ij})$. The following algorithm gives the factorization (4.1.2) where L is lower triangular and U is unit upper triangular. The pivots are stored in the nonsingular matrix P.

<u>Algorithm</u> 4.1.1 Crout factorization [42].

For $r = 1,2,\ldots,n$, do steps 1 through 4.

1. Compute $\ell_{ir} = a_{ir} - \sum_{k=1}^{r-1} \ell_{ik}u_{kr}$, $i = r,r+1,\ldots,n$.

2. Set $int_r \geq r$ equal to the smallest k for which
$$|\ell_{kr}| = \max_{r \leq j \leq n} |\ell_{jr}| .$$

3. Interchange the r^{th} and int_r^{th} rows of the array. Refer to these rows by their new positions.

4. Compute $u_{rj} = (a_{rj} - \sum_{k=1}^{r-1} \ell_{rk}u_{kj})/\ell_{rr}$, for $j = r+1,\ldots,n$.

The convention $\sum_{1}^{0} = 0$ is used.

The algorithm without steps 2 and 3 is called Crout factorization without pivoting. The pivoting strategy of steps 2 and 3 is the classical one and considered closed, in that there is no flexibility in the choice of pivots.

Instead of the closed pivoting of the standard Crout or Doolittle factorizations, an open pivoting strategy should be used. Duff [16] calls this "threshold pivoting". As Duff notes, the classical selection of pivots was designed to maintain numerical stability. However that selection may destroy the sparsity in the factorization. Threshold pivoting allows selection of nonzero pivots which are at least 1/u times the largest element in the row or column. Erisman and Reid [17] give a formula to monitor the growth of the matrix elements in the factorization. This can be done working only with the sparsity pattern and the factor u. If the growth is too large, a smaller value of u can be chosen. Duff also note that $u = 10$ yields good retention of sparsity and generally good numerical accuracy. He also recommends iterative refinement to improve the solution since solutions for systems with sparse triangular factorizations are fairly cheap.

4.2 Some algorithms.

Consider again problem (1.1.1). The Jacobian, or approximate Jacobian, is decomposed into triangular factors by either the Crout or Dolittle factorizations. The unit triangular factor is held fixed for a finite number of iterations; the other factor is updated using a Schubert correction to preserve the triangular structure and any other zero-nonzero structure of that factor. (If that unit triangular factor were fixed for the entire iterative process, a term involving $||x_0 - x^*||_2$ would occur as a linear growth, denying the bounded deterioration estimate, and the convergence proof cannot be obtained using this type of analysis.)

The algorithms are presented without pivoting for simplicity of the exposition. The implementations would employ an open pivoting strategy as discussed in the previous section. The convergence results in section 4.4 are local and will be based on using a fixed pivoting strategy throughout the iteration. The initial choice of pivots is further discussed at the end of section 4.4.

The first algorithm uses the Dolittle factorization to decompose the Jacobian into a unit lower triangular matrix, L, and an upper triangular matrix, U.

Algorithm 4.2.1

1. Given x_0, the initial guess to the solution of $F(x) = 0$,

   m, a fixed positive integer, and

   $\beta > 0$, a fixed constant.

2. Evaluate $F_0 = F(x_0)$ and $J_0 = J_F(x_0)$.

3. Perform a Dolittle factorization of J_0 to obtain $J_0 = L_0U_0$ where L_0 is unit lower triangular and U_0 is upper triangular. Set $k = 0$.

4. Solve $L_0w_k = -F_k$ and $U_ks_k = w_k$ for w_k and s_k respectively.

5. If $k = m-1$, set $x_0 = x_k + s_k$ and go to 2;

   else set $x_{k+1} = x_k + s_k$,

   $F_{k+1} = F(x_{k+1})$,

   and $y_k = F_{k+1} - F_k$.

6. Solve $L_0v_k = y_k$.

7. Update the j^{th} row of U_k if $||s_k||_2 \leq \beta||(s_k)_j||_2$.

   Let $\chi_\beta(s_k) \equiv \{j: ||s_k||_2 \leq \beta||(s_k)_j||_2\}$.

$$(4.2.1)\quad U_{k+1} = U_k + \sum_{\chi_\beta(s_k)} [(s_k)_j^T(s_k)_j]^+ e_j^T(v_k-U_ks_k)\, e_j(s_k)_j^T .$$

   Replace k by k+1 and go to step 4.

Another technique for getting started in step 2 would be to make a finite difference approximation [6] to $J_F(x_0)$ instead of evaluating all of the partial derivatives.

The other basic algorithm is to use the Crout factorization in place of the Dolittle factorization.

<u>Algorithm</u> 4.2.2

1. Given $x_0 \in \mathbb{R}^n$, m a fixed positive integer, and $\beta > 0$ fixed.
2. Evaluate $F_0 = F(x_0)$ and $J_0 = J_F(x_0)$.

   Set $k = 0$.
3. Factor $J_0 = L_0U_0$ using the Crout decomposition.
4. Solve $L_kw_k = -F_k$ and $U_0s_k = w_k$.
5. If $k = m-1$, set $x_0 = x_k + s_k$ and go to step 2;

   else $x_{k+1} = x_k + s_k$,

   $F_{k+1} = F(x_{k+1})$,

   and $y_k = F_{k+1} - F_k$.
6. Update the j^{th} row of L_k if $||w_k||_2 \leq \beta||(w_k)_j||_2$.

   Let $\chi_\beta(w_k) \equiv \{j: ||w_k||_2 \leq \beta||(w_k)_j||_2\}$.

$$L_{k+1} = L_k + \sum_{\chi_\beta(w_k)} [(w_k)_j^T(w_k)_j]^+ e_j^T(y_k - L_kU_0s_k)\, e_j(w_k)_j^T. \tag{4.2.2}$$

   Replace k by k+1 and go to step 4.

There is one less triangular linear system to be solved at each iteration in this algorithm. Compare with algorithm 4.2.1 at step 6. The trade-off is an additional matrix-vector multiplication in updating the triangular factor. Implicit in these algorithms is the definition of the projections $(\cdot)_j$. In algorithm 4.2.1, the projection is onto the zero-nonzero structure of the j^{th} row of the upper triangular factor. In algorithm 4.2.2, the projection is onto the zero-nonzero structure of the j^{th} row of the lower triangular factor.

The constant β is used to prevent a correction to a row when the projected step along that row is too small relative to the full step. Compare with M used in the algorithm in section 3.5. Also note that this technique would be used in implementing the ordinary Schubert update (2.2.6) or (2.2.7).

Other algorithms are readily suggested. After the initialization of J_0 by either direct evaluation or forming a finite difference approximation, a Schubert (or Broyden) update can be performed every m iterations. That is, form the product

$$B_{m-1} = L_0 U_{m-1} ,$$

and obtain B_m from B_{m-1}, s_m, and y_m by the Schubert update, then factor

$$B_m = L_m U_m$$

by the Dolittle decomposition. Similarly, algorithm 4.2.2 could be modified.

Another modification in the basic algorithms would be to monitor the growth of the elements in the updated triangular factor and to re-evaluate the Jacobian if it becomes poorly conditioned. This would supplement the criterion to re-evaluate every m iterations.

4.3 Continuity of the factorization.

The convergence properties for the algorithms will depend on the continuity of the factorization. The following theorem establishes when the Crout factorization without pivoting exists.

Theorem 4.3.1 [13] Let $A \in L(\mathbb{R}^n)$ be nonsingular. Then the following are equivalent:

(a) Crout reduction without pivoting can be carried out;

(b) There is a unique nonsingular lower triangular matrix, L, and a unique unit upper triangular matrix, U, such that $A = LU$.

(c) All the leading principal submatrices of A are nonsingular.

A similar theorem holds for the Dolittle factorization.

The next theorem gives sufficient conditions for the LU decomposition to be continuous in a neighborhood of a point where the Crout factorization without pivoting is known to exist.

Theorem 4.3.2 Let $A: \mathbb{R}^n \to \mathbb{R}^n$ and suppose for some $x_0 \in \mathbb{R}^n$ that $A(x_0)$ is nonsingular and that there exists $\epsilon_0 > 0$ and $\gamma_{ij} \geq 0$ such that

$$(4.3.1) \qquad |e_i[A(x) - A(y)]e_j| \leq \gamma_{ij}||x - y||_2 ,$$

for $i,j = 1,2,\ldots,n$, and for all $x,y \in B(x_0,\epsilon_0)$. If the Crout factorization without pivoting exists at x_0,

$$A(x_0) = L(x_0)U(x_0) ,$$

then there exists $\epsilon > 0$ such that the Crout factorization without pivoting exists for all $x \in B(x_0,\epsilon)$.

Furthermore, there exist constants c_0 and $d_0 > 0$ such that

$$(4.3.2)(a) \qquad ||L(x) - L(x_0)||_F \leq c_0||x - x_0||_2 ,$$

and

$$(4.3.2)(b) \qquad ||U(x) - U(x_0)||_F \leq d_0||x - x_0||_2 ,$$

for all $x \in B(x_0,\epsilon)$.

Proof: Since $A(x_0)$ is nonsingular and the Crout factorization without pivoting exists at x_0, it follows from Theorem 4.3.1 that the leading principal submatrices of $A(x_0) \equiv {}^kA(x_0)$ $k = 1,2,\ldots,n$, must be nonsingular. By (4.3.1), $A(x)$ and all of its submatrices are Lipschitz continuous in $B(x_0,\epsilon_0)$. Therefore by the Inverse Function Theorem [27] there exist $\epsilon_1,\epsilon_2,\ldots,\epsilon_n$, such that ${}^kA(x)$ is nonsingular in $B(x_0,\epsilon_k)$. Let $\epsilon = \min_{1\leq k\leq n} \epsilon_k$. Then all the leading principal submatrices of $A(x)$ are nonsingular in $B(x_0,\epsilon)$. Another application of Theorem 4.3.1 establishes the existence of the Crout factorization without pivoting for all $x \in B(x_0,\epsilon)$:

$$A(x) = L(x)U(x).$$

Restrict ϵ so that the preceding factorization holds on $\bar{B}(x_0,\epsilon)$. The proof of (4.3.2) is by induction on the dimension [40]. Let $x \in B(x_0,\epsilon)$ and $A(x) = L(x)U(x)$.

For $\ell = 1$,

$${}^{\ell}A(x) = {}^1A(x) = a_{11}(x)$$

and $$a_{11}(x) = a_{11}(x)1 = l_{11}(x)u_{11}(x) = {}^1L(x){}^1U(x) .$$

Thus, $$||{}^1L(x) - {}^1L(x_0)||_F = |a_{11}(x) - a_{11}(x_0)|$$

$$\leq \gamma_{11}||x - x_0||_2$$

$$\equiv c_1||x - x_0||_2 \,,$$

by (4.3.1), and

$$||{}^1U(x) - {}^1U(x_0)||_F = |1 - 1| = 0||x - x_0||_2 \,.$$

By the Inverse Function Theorem, ${}^1L^{-1}$ and ${}^1U^{-1}$ are Lipschitz continuous in $\bar{B}(x_0,\epsilon)$. Let $\hat{c}_1$ and $\hat{d}_1$ be the respective Lipschitz constants.

Assume for $\ell = 2,3,\ldots,k$, $k \leq n-1$, that

(4.3.3) $$||{}^\ell L(x) - {}^\ell L(x_0)||_F \leq c_\ell||x - x_0||_2 \,,$$

and $$||{}^\ell U(x) - {}^\ell U(x_0)||_F \leq d_\ell||x - x_0||_2 \,,$$

for all $x \in B(x_0,\epsilon)$. Furthermore, the inverses, ${}^\ell L^{-1}(x)$ and ${}^\ell U^{-1}(x)$ are Lipschitz continuous on $\bar{B}(x_0,\epsilon)$ with Lipschitz constants $\hat{c}_\ell$ and $\hat{d}_\ell$ respectively.

For $\ell = k+1$, let ${}^{k+1}A(x)$ be written as

(4.3.4) $${}^{k+1}A(x) = \left[\begin{array}{c|c} {}^kA(x) & {}^kv(x) \\ \hline {}^kw^T(x) & a_{k+1,k+1}(x) \end{array}\right] .$$

Since ${}^kA(x) = {}^kL(x){}^kU(x)$, ${}^{k+1}A(x)$ can be factored

$${}^{k+1}A(x) = {}^{k+1}L(x){}^{k+1}U(x)$$

$$= \left[\begin{array}{c|c} {}^{k}L(x) & 0 \\ \hline {}^{k}w^{T}(x)\,{}^{k}U^{-1}(x) & \alpha_{k+1}(x) \end{array}\right] \left[\begin{array}{c|c} {}^{k}U(x) & {}^{k}L^{-1}(x)\,{}^{k}v(x) \\ \hline 0 & 1 \end{array}\right]$$

where $\alpha_{k+1} = a_{k+1,k+1}(x) - {}^{k}w^{T}(x)\,{}^{k}U^{-1}(x)\,{}^{k}L^{-1}(x)\,{}^{k}v(x)$.

Now,

$$(4.3.5) \qquad \|{}^{k+1}L(x) - {}^{k+1}L(x_0)\|_F$$

$$= \left\| \left[\begin{array}{c|c} {}^{k}L(x) & 0 \\ \hline {}^{k}w^{T}(x)\,{}^{k}U^{-1}(x) & \alpha_{k+1}(x) \end{array}\right] - \left[\begin{array}{c|c} {}^{k}L(x_0) & 0 \\ \hline {}^{k}w^{T}(x_0)\,{}^{k}U^{-1}(x_0) & \alpha_{k+1}(x_0) \end{array}\right] \right\|_F$$

$$\leq \|{}^{k}L(x) - {}^{k}L(x_0)\|_F + \|{}^{k}w^{T}(x)\,{}^{k}U^{-1}(x) - {}^{k}w^{T}(x_0)\,{}^{k}U^{-1}(x_0)\|_2$$

$$+ |\alpha_{k+1}(x) - \alpha_{k+1}(x_0)| .$$

An application of the triangle inequality to the second term yields

$$(4.3.6) \qquad \|{}^{k}w^{T}(x)\,{}^{k}U^{-1}(x) - {}^{k}w^{T}(x_0)\,{}^{k}U^{-1}(x_0)\|_2$$

$$\leq \|[{}^{k}w^{T}(x) - {}^{k}w^{T}(x_0)]\,{}^{k}U^{-1}(x)\|_2 + \|{}^{k}w^{T}(x_0)[{}^{k}U^{-1}(x) - {}^{k}U^{-1}(x_0)]\|_2$$

$$\leq \|{}^{k}w(x) - {}^{k}w(x_0)\|_2\|{}^{k}U^{-1}(x)\|_2 + \|{}^{k}w(x_0)\|_2\|{}^{k}U^{-1}(x) - {}^{k}U^{-1}(x_0)\|_F$$

$$\leq (\sum_{i=1}^{k} \gamma_{k,i}^2)^{\frac{1}{2}}\|x-x_0\|_2\|{}^{k}U^{-1}(x)\|_2 + \|{}^{k}w(x_0)\|_2\hat{\delta}_k\|x-x_0\|_2$$

$$\leq \{(\sum_{i=1}^{k} \gamma_{k,i}^2)^{\frac{1}{2}} \sup_{x \in B(x_0,\epsilon)} \|{}^{k}U^{-1}(x)\|_2 + \|{}^{k}w(x_0)\|_2\hat{\delta}_k\}\|x-x_0\|_2$$

$$\equiv \beta_k \|x - x_0\|_2 .$$

Similarly, the third term on the right hand side of (4.3.5) can be estimated:

(4.3.7)
$$|\alpha_{k+1}(x) - \alpha_{k+1}(x_0)|$$

$$\le |a_{k+1,k+1}(x) - a_{k+1,k+1}(x_0)|$$
$$+ |{}^k w^T(x)\,{}^k U^{-1}(x)\,{}^k L^{-1}(x)\,{}^k v(x) - {}^k w^T(x_0)\,{}^k U^{-1}(x_0)\,{}^k L^{-1}(x_0)\,{}^k v(x_0)|$$

$$\le \gamma_{k+1,k+1}||x-x_0||_2 + |({}^k w^T(x) - {}^k w^T(x_0))\,{}^k U^{-1}(x)\,{}^k L^{-1}(x)\,{}^k v(x)|$$
$$+ |{}^k w^T(x_0)[{}^k U^{-1}(x) - {}^k U^{-1}(x_0)]\,{}^k L^{-1}(x)\,{}^k v(x)|$$
$$+ |{}^k w^T(x_0)\,{}^k U^{-1}(x_0)[{}^k L^{-1}(x) - {}^k L^{-1}(x_0)]\,{}^k v(x)|$$
$$+ |{}^k w^T(x_0)\,{}^k U^{-1}(x_0)\,{}^k L^{-1}(x_0)({}^k v(x) - {}^k v(x_0))|$$

$$\le \{\gamma_{k+1,k+1} + (\sum_{i=1}^{k} \gamma_{k+1,i}^2)^{\frac{1}{2}} \sup_{x \in B(x_0,\epsilon)} ||{}^k U^{-1}(x)||_2 ||{}^k L^{-1}(x)||_2 ||{}^k v(x)||_2$$
$$+ ||{}^k w(x_0)||_2\, \partial_k \sup_{x \in B(x_0,\epsilon)} ||{}^k L^{-1}(x)||_2 ||{}^k v(x)||_2$$
$$+ ||{}^k w(x_0)||_2 ||{}^k U^{-1}(x_0)||_2\, \hat{\epsilon}_k \sup_{x \in B(x_0,\epsilon)} ||{}^k v(x)||_2$$
$$+ ||{}^k w(x_0)||_2 ||{}^k U^{-1}(x_0)||_2 ||{}^k L^{-1}(x_0)||_2 (\sum_{i=1}^{k} \gamma_{i,k+1}^2)^{\frac{1}{2}}\}||x-x_0||_2$$

$$\equiv \delta_k \; ||x-x_0||_2 \,.$$

Set $c_{k+1} = c_k + \beta_k + \delta_k$.

To complete the proof, observe that

(4.3.8)
$$||{}^k \ddot{U}(x) - {}^k U(x_0)||_F$$

$$= \left\| \begin{bmatrix} {}^kU(x) & {}^kL^{-1}(x)\,{}^kv(x) \\ \hline 0 & 1 \end{bmatrix} - \begin{bmatrix} {}^kU(x_0) & {}^kL^{-1}(x_0)\,{}^kv(x_0) \\ \hline 0 & 1 \end{bmatrix} \right\|_F$$

$$\le \|{}^kU(x) - {}^kU(x_0)\|_F + \|{}^kL^{-1}(x)\,{}^kv(x) - {}^kL^{-1}(x_0)\,{}^kv(x_0)\|_F$$

$$\le d_k\ \|x-x_0\|_2 + \|{}^kL^{-1}(x)[{}^kv(x) - {}^kv(x_0)]\| + \|[{}^kL^{-1}(x) - {}^kL^{-1}(x_0)]\ {}^kv(x_0)\|_2$$

$$\le \{d_k + \sup_{x \in B(x_0,\epsilon)} \|{}^kL^{-1}(x)\|_2 \ (\sum_{i=1}^{k} \gamma_{i,k+1}^2)^{\frac{1}{2}} + \hat{c}_k \|{}^kv(x_0)\|_2\} \|x-x_0\|_2$$

$$\equiv d_{k+1}\ \|x-x_0\|_2 \ .$$

By the Inverse Function Theorem, ${}^{k+1}L^{-1}(x)$ and ${}^{k+1}U^{-1}(x)$ are Lipschitz continuous on $\bar{B}(x_0,\epsilon)$ with Lipschitz constants $\hat{c}_{k+1}$ and $\hat{d}_{k+1}$ respectively. The induction is complete. Define $c_0 = c_n$ and $d_0 = d_n$ to prove (4.3.2).

A similar theorem holds for the Dolittle decomposition.

<u>Theorem</u> 4.3.3 Let $A: \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and suppose for some $x_0 \in \mathbb{R}^n$ that $A(x_0)$ is nonsingular and that there exists $\epsilon_0 > 0$ and $\gamma_{ij} \ge 0$ such that

$$(4.3.9) \qquad |e_i^T[A(x) - A(y)]e_j| \le \gamma_{ij}\|x - y\|_2$$

for $i,j = 1,2,\ldots,n$, and for all $x,y \in B(x_0,\epsilon_0)$. If the Dolittle factorization without pivoting exists at x_0, $A(x_0) = L(x_0)U(x_0)$, then there exists $\epsilon > 0$ such that the Dolittle factorization without pivoting exists for all $x \in B(x_0,\epsilon)$.

Furthermore, there exist constants c_0 and $d_0 > 0$ such that

(4.3.10)(a) $$||L(x) - L(x_0)||_F \le c_0 \, ||x - x_0||_2 ,$$

and

(4.3.10)(b) $$||U(x) - U(x_0)||_F \le d_0 \, ||x - x_0||_2$$

for all $x \in B(x_0,\epsilon)$.

Proof: The proof is very similar to the proof of Theorem 4.3.2.

4.4 Convergence.

Before stating the convergence theorems, several technical lemmas will be given. The first two lemmas relate the updated triangular matrix factors to the preceding ones.

Lemma 4.4.1 Let $F: \mathbb{R}^n \to \mathbb{R}^n$ and suppose there exists x^* such that $F(x^*) = 0$ and $J^* = J(x^*)$ is nonsingular. Suppose there exists $\epsilon > 0$ and $\gamma_{ij} \ge 0$ such that

(4.4.1) $$|e_i^T[J(x) - J^*]e_j| \le \gamma_{ij} \, ||x - x^*||_2$$

for $i,j = 1,2,\dots,n$, and for all $x \in B(x^*,\epsilon)$.

Then there exists $\epsilon_0 > 0$ and a pivoting strategy, P_0, such that the Crout factorization without pivoting of $P_0 J$ exists at $x_0 \in B(x^*,\epsilon_0)$:

(4.4.2) $$P_0 J(x_0) = L(x_0)U(x_0) .$$

Furthermore, $\{L_\ell\}_{\ell=0}^{m-1}$ defined by (4.2.2) satisfy

(4.4.3) $$||L_{\ell+1}-L^*||_F^2 \le ||L_\ell-L^*||_F^2 + \beta^2 n(\kappa\sigma_\ell + ||L^*||_2 d_0 ||U_0^{-1}||_2 ||x_0-x^*||_2)^2$$

where $\kappa = ||\Gamma||_F$ and $\Gamma = (\gamma_{ij})$, and $\sigma_\ell = \max(||x_{\ell+1}-x^*||_2, ||x_\ell-x^*||_2)$.

<u>Proof</u>: The proof is very similar to that of the following lemma.

<u>Lemma</u> 4.4.2 Let $F:\mathbb{R}^n \to \mathbb{R}^n$ and suppose there exists x^* such that $F(x^*) = 0$ and $J^* = J(x^*)$ is nonsingular. Suppose there exists $\epsilon > 0$ and $\gamma_{ij} \geq 0$ such that

$$|e_i^T [J(x) - J^*]e_j| \leq \gamma_{ij}||x - x^*||_2 \,, \tag{4.4.4}$$

for $i,j = 1,2,\ldots,n$, and for all $x \in B(x^*,\epsilon)$.

Then there exists $\epsilon_0 > 0$ and a pivoting strategy, P_0, such that the Dolittle factorization without pivoting of P_0J exists at $x_0 \in B(x^*,\epsilon_0)$:

$$P_0J(x_0) = L(x_0)U(x_0)\,. \tag{4.4.5}$$

Furthermore, $\{U_\ell\}_{\ell=0}^{m-1}$ defined by (4.2.1) satisfy

$$||U_{\ell+1} - U^*||_F^2 \leq ||U_\ell - U^*||_F^2 + ||L_0^{-1}||_2^2 \, n\beta^2[\kappa\sigma_\ell + ||U^*||_2 c_0 ||x_0-x^*||_2]^2, \tag{4.4.6}$$

where $\beta > 0$ is set in step 1 of the algorithm 4.2.1, $\kappa \equiv ||\Gamma||_F$, $\Gamma = (\gamma_{ij})$, and $\sigma_\ell = \max(||x_{\ell+1} - x^*||_2, ||x_\ell - x^*||_2)$.

<u>Proof</u>: For the first assertion, let P_0 be any pivoting strategy such that the Dolittle factorization of

$$P_0J(x^*) = L^*U^*$$

exists. Apply Theorem 4.3.3 to $A(x) \equiv P_0J(x)$ to obtain ϵ_0 such that the Dolittle factorization without pivoting of

$$A(x) = L(x)U(x)\,.$$

exists for all $x \in B(x_0, \epsilon_0)$. That theorem also gives a $c_0 > 0$ such

$$||L(x) - L^*||_F \le c_0 ||x - x^*||_2$$

for all $x \in B(x^*, \epsilon_0)$.

In particular, if $x_0 \in B(x^*, \epsilon_0)$, then the Dolittle factorization $P_0 J(x_0) = L_0 U_0$ exists. To prove the second assertion, evaluate $||U_{\ell+1} - U^*||_F^2$ using the update (4.2.1). Set $\chi_\beta(s_\ell) = \{i : ||s_\ell||_2 \le \beta ||(s_\ell)_i||_2\}$, then

(4.4.7)
$$||U_{\ell+1} - U^*||_F^2$$

$$= ||U_\ell - U^* + \sum_{\chi_\beta(s_\ell)} [(s_\ell)_i^T (s_\ell)_i]^+ e_i^T (L_0^{-1} y_\ell - U_\ell s_\ell)\ e_i (s_\ell)_i^T||_F$$

$$= \sum_{\chi_\beta(s_\ell)} \{||e_i^T (U_\ell - U^*)(I - [(s_\ell)_i^T (s_\ell)_i]^+ (s_\ell)(s_\ell)_i^T||_2^2$$

$$+ ||e_i^T (L_0^{-1} y_\ell - U^* s_\ell)[(s_\ell)_i^T (s_\ell)_i]^+ (s_\ell)_i^T||_2^2\}$$

$$+ \sum_{\chi_\beta^c(s_\ell)} ||e_i^T (U_\ell - U^*)||_2^2$$

$$\le ||U_\ell - U^*||_F^2 + \sum_{\chi_\beta(s_\ell)} ||[(s_\ell)_i^T (s_\ell)_i]^+ e_i^T L_0^{-1} (y_\ell - L^* U^* s_\ell$$

$$+ L^* U^* s_\ell - L_0 U^* s_\ell)(s_\ell)_i^T||_2^2$$

$$\le ||U_\ell - U^*||_F^2 + \sum_{\chi_\beta(s_\ell)} [||(s_\ell)_i^T (s_\ell)_i]^+ e_i^T L_0^{-1} (y_\ell - J^* s_\ell)(s_\ell)_i^T||_2$$

$$+ ||[(s_\ell^T)_i (s_\ell)_i]^+ e_i^T L_0^{-1} (L^* - L_0) U^* s_\ell (s_\ell)_i^T||_2]^2$$

$$\leq ||U_\ell - U^*||_F^2 + \sum_{X_\beta(s_\ell)} [||L_0^{-1}||_2(||(s_\ell)_i||_2^+||y_\ell - J^*s_\ell||_2$$

$$+ ||(s_\ell)_i||_2^+||L^* - L_0||_2||U^*||_2||s_\ell||_2)]^2$$

$$\leq ||U_\ell - U^*||_F^2 + ||L_0^{-1}||_2^2\beta^2\sum_{i=1}^{n}(\kappa\sigma_\ell + ||U^*||_2 c_0||x_0 - x^*||_2)^2$$

$$\leq ||U_\ell - U^*||_F^2 + ||L_0^{-1}||_2\beta^2 n(\kappa\sigma_\ell + ||U^*||_2 c_0||x_0 - x^*||_2)^2,$$

giving (4.4.6).

There are several norms which can be used for the error estimates. The most convenient weighted norm is the one established at the start of the iteration.

<u>Definition</u> 4.4.3 Define the weighted norm $||\cdot||_{L_0^{-1},F}: L(\mathbb{R}^n) \to \mathbb{R}$ by

(4.4.8) $$||\cdot||_{L_0^{-1},F} \equiv ||L_0^{-1}\cdot||_F \quad ,$$

for L_0 nonsingular, lower triangular. This is just the left weighted Frobenius norm.

<u>Lemma</u> 4.4.4 Let $L: \mathbb{R}^n \to \mathbb{R}^n$ be a continuous lower triangular operator Suppose L^{-1} exists and is continuous on $\bar{B}(x^*,\epsilon)$ for some $x^* \in \mathbb{R}^n$ and $\epsilon > 0$. Then there exist constants $\bar{\eta}$ and $\hat{\eta} > 0$ such that

(4.4.9) $$||\cdot||_{L^{-1}(x),F} \leq \bar{\eta}||\cdot||_F \quad \text{for all} \quad x \in \bar{B}(x^*,\epsilon), \text{ and}$$

(4.4.10) $$||\cdot||_F \leq \hat{\eta}||\cdot||_{L^{-1}(x),F} \quad \text{for all} \quad x \in \bar{B}(x^*,\epsilon).$$

<u>Proof</u>: Let $\bar{\eta} = \sup_{x \in \bar{B}(x^*,\epsilon)} ||L^{-1}(x)||_2$ and $\hat{\eta} = \sup_{x \in \bar{B}(x^*,\epsilon)} ||L(x)||_2$.

Then,

$$||\cdot||_{L^{-1}(x),F} = ||L^{-1}(x)\ \cdot||_F \leq ||L^{-1}(x)||_2||\cdot||_F \leq \bar{\eta}||\cdot||_F ,$$

and

$$||\cdot||_F = ||L(x)L^{-1}(x)\ \cdot||_F \leq ||L(x)||_2||\cdot||_{L^{-1}(x),F} \leq \hat{\eta}||\cdot||_{L^{-1}(x),F} .$$

Let $\eta > 0$ be such that the matrix norms satisfy $||\cdot||_F \leq \eta||\cdot||_2$.

<u>Theorem</u> 4.4.5 Let $F: \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in an open convex set D_0. Suppose there exists $x^* \varepsilon D_0$ such that $F(x^*) = 0$ and $J(x^*)$ is nonsingular. Suppose there exists $\Gamma = (\gamma_{ij})$ with $\gamma_{ij} \geq 0$ such that

(4.4.11) $$|e_i^T[J(x) - J(y)]e_j| \leq \gamma_{ij}||x - y||_2$$

for $i,j = 1,2,\ldots,n$, and for all $x,y \varepsilon D_0$.

Then there exist $\epsilon,\delta > 0$ and a pivoting strategy, P_0, such that if $||x_0-x^*||_2 < \epsilon$ and $||J_0-J^*||_F < \delta$ then the algorithm 4.2.1 generates $\{x_k\}$ which converges locally and superlinearly to x^*.

<u>Proof</u>: Let $\epsilon_1 > 0$ such that $B(x^*,\epsilon_1) \subset D_0$. Set $\gamma \geq ||J^{*-1}||_2$ and define $\kappa = ||\Gamma||_F$.

By Lemma 4.4.2, there exists $\epsilon_0 > 0$ and a pivoting strategy, P_0, such that the Dolittle factorization of $P_0J(x)$ exists for all $x \varepsilon B(x^*,\epsilon_0)$. Furthermore, (4.3.11)(a) can now be written as

(4.4.12) $$||L(x) - L(x^*)||_F \leq c_0||x - x^*||_2$$

for all $x \varepsilon B(x^*,\epsilon_0)$.

Now choose ϵ in $(0,\epsilon_0)$, $\delta > 0$ and $r \in (0,$ [illegible] that

(4.4.13)
$$\alpha_2 \epsilon \left(\frac{1-r^{m-1}}{1-r} \right) < \delta$$
$$\gamma(1+r)[\kappa\epsilon + 2\eta\bar{\eta}\hat{\eta}\delta] \leq r$$

where $\alpha_2 = \alpha_1 + (m-1)\alpha_0$ and $\alpha_0 = \bar{\eta}||U^*||_2 c_0($ [illegible] $\alpha_1 = \sqrt{n}\,\eta\bar{\eta}\beta\kappa$,

and $\hat{\eta} = \sup_{x \in \bar{B}(x^*,\epsilon_0)} ||L(x)||_2$, $\bar{\eta} = \sup_{x \in \bar{B}(x^*,\epsilon_0)}$ [illegible].

Further restrict ϵ so that $||J(x) - J^*||_F < \delta$ whenever $||x - x^*||_2 < \epsilon$. Let $x_0 \in B(x^*,\epsilon)$, then

(4.4.14)
$$||L(x_0) - L(x^*)||_F \leq c_0||x_0 - x^*||_2 .$$

by (4.4.12).

If $||J_0 - J^*||_F < \delta$ then $||J_0 - J^*||_2 <$ [illegible] and (4.4.13) implies that $\gamma(1+r)2\eta\delta \leq r$. Therefore the Banach Lemma [illegible] gives

(4.4.15)
$$||J_0^{-1}||_2 \leq (1+r)\gamma .$$

The iteration of Algorithm 4.2.1 can be viewed as follows: Let $p = im$, $i = 0,1,\ldots,$ and $k \in \{p,p+1,\ldots,p+(m-1)\}$, and

(4.4.16)
$$x_{k+1} = x_k - U_k^{-1}L_p^{-1}F_k$$

with $\{L_p\}$ and $\{U_p\}$ determined in step 5, and $\{U_k\}_{k=p+1}^{p+(m-1)}$ determined in step 7. Then,

$$x_{k+1}-x^* = x_k-x^*-U_k^{-1}L_p^{-1}(F_k-F^*) + U_k^{-1}L_p^{-1}J^*(x_k-x^*) - U_k^{-1}L_p^{-1}J^*(x_k-x^*) ,$$

and

$$(4.4.17)\quad ||x_{k+1}-x^*||_2 \le ||U_k^{-1}L_p^{-1}||_2[\,||F_k-F^*-J^*(x_k-x^*)||_2 + ||L_pU_k-J^*||_2||x_k-x^*||_2] \, .$$

For $p = 0,\ k = 0,$ (4.4.17) becomes

$$(4.4.18)\quad \begin{aligned} ||x_1-x^*||_2 &\le ||J_0^{-1}||_2[(\kappa\epsilon + 2\eta\delta)||x_0-x^*||_2] \\ &\le (1+r)\gamma(\kappa\epsilon + 2\eta\delta)||x_0-x^*||_2 \\ &\le r||x_0-x^*||_2 \, , \end{aligned}$$

applying (4.4.15) and (4.4.13).

Assume for $p = im,\ i = 0,1,\ldots,j$ that

$$(4.4.19)\quad ||x_{jm} - x^*||_2 \le r||x_{jm-1} - x^*||_2 \, .$$

Then $x_{jm} \in B(x^*,\epsilon)$ and it follows that $||J_{jm} - J^*||_F < \delta$, and $||J_{jm} - J^*||_2 < 2\eta\delta$. From (4.4.13), $2\gamma(1+r)\eta\delta \le r$, and the Banach Lemma [27] gives $||J_{jm}^{-1}||_2 \le (1+r)\gamma$.

For $p = jm$, let $k = p$, and (4.4.17) is

$$(4.4.20)\quad \begin{aligned} ||x_{jm+1} - x^*||_2 &\le ||J_{jm}^{-1}||_2[(\kappa\epsilon + 2\eta\delta)||x_{jm} - x^*||_2] \\ &\le (1+r)\gamma(\kappa\epsilon + 2\eta\delta)||x_{jm} - x^*||_2 \\ &\le r||x_{jm} - x^*||_2 \, . \end{aligned}$$

Assume for $\ell = p,p+1,\ldots,k-1,\ k \le p+(m-1)$ that

$$||x_{\ell+1} - x^*||_2 \le r||x_\ell - x^*||_2 \, .$$

Now,

(4.4.21) $$||L_pU_{\ell+1}-J^*||_{L_p^{-1},F} = ||L_pU_{\ell+1}-L_pU^*+L_pU^*-L^*U^*||_{L_p^{-1},F}$$

$$\leq ||U_{\ell+1}-U^*||_F + ||L_p^{-1}||_2||L_p-L^*||_F||U^*||_2$$

$$\leq ||U_\ell-U^*||_F + ||L_p^{-1}||_2 \sqrt{n}\beta[\kappa\sigma_\ell$$

$$+ ||U^*||_2c_0||x_p-x^*||_2]$$

$$+ ||L_p^{-1}||_2||U^*||_2||L_p-L^*||_F ,$$

by (4.4.6) and the fact that for $a,b,c > 0$, $c^2 \leq a^2 + b^2 \leq (a + b)^2$ implies $c \leq a + b$. Since $||L_p^{-1}||_2 \leq \bar{n}$,

(4.4.22) $$||L_pU_{\ell+1}-J^*||_{L_p^{-1},F} \leq ||L_pU_\ell-L^*U^*||_{L_p^{-1},F}$$

$$+ 2\bar{n}||U^*||_2||L_p-L^*||_F + \alpha_1\sigma_\ell + \bar{n}\sqrt{n}\beta||U^*||_2c_0||x_\ell-x^*||_2$$

$$\leq ||L_pU_\ell- L^*U^*||_{L_p^{-1},F} + \alpha_1\sigma_\ell + \alpha_0||x_\ell - x^*||_2 .$$

Therefore,

(4.4.23) $$||L_pU_{\ell+1} - L^*U^*||_{L_p^{-1},F} \leq ||L_pU_p - L^*U^*||_{L_p^{-1},F} + \alpha_1 \sum_{i=p}^{\ell} \sigma_i$$

$$+ \alpha_0(\ell+1)||x_p - x^*||_2$$

$$\leq \alpha_2 \sum_{i=p}^{\ell} \sigma_i + ||L_pU_p - L^*U^*||_{L_p^{-1},F} ,$$

and for $\ell = k-1$,

(4.4.24) $$||L_pU_k-L^*U^*||_{L_p^{-1},F} \leq ||L_pU_p-L^*U^*||_{L_p^{-1},F} + \alpha_2 \sum_{i=p}^{k-1} r^i\epsilon$$

$$\leq ||L_pU_p - L^*U^*||_{L_p^{-1},F} + \alpha_2\epsilon\left(\frac{1-r^{m-1}}{1-r}\right)r^p$$

$$\leq \bar{\eta}\delta + \delta \leq 2\bar{\eta}\delta .$$

Thus $||L_pU_{\ell+1}-J^*||_2 \leq \eta||L_pU_{\ell+1}-J^*||_F \leq \eta\hat{\eta}||L_pU_{\ell+1}-J^*||_{L_p^{-1},F}$ and this gives $||U_k^{-1}L_p^{-1}||_2 \leq (1+r)\gamma$ which with (4.4.17) yields

$$||x_{k+1}-x^*||_2 \leq (1+r)\gamma[\kappa\epsilon + 2\eta\bar{\eta}\hat{\eta}\delta]||x_k - x^*||_2$$

$$\leq r||x_k-x^*||_2 , \text{ by } (4.4.13).$$

For $\ell = p+(m-1)$, (4.4.17) is

$$||x_{p+m} - x^*||_2 \leq r||x_{p+(m-1)} - x^*||_2 ,$$

so $x_{p+m} \in B(x^*,\epsilon)$ and $||J_{p+m} - J^*||_F < \delta$. This establishes the local and linear convergence of Algorithm 4.2.1. Superlinear convergence follows from Theorem 1.5.3 since for $p = im$, $i = 0,1,\ldots$

$$\frac{||(J_p - J^*)(x_{p+1} - x_p)||_2}{||x_{p+1} - x_p||_2} \leq ||J_p - J^*||_2 \xrightarrow[p\to\infty]{} 0 .$$

The convergence theorem for algorithm 4.2.2 is proved in a similar fashion.

<u>Theorem</u> 4.4.6 Let $F:\mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable in an open convex set D_0. Suppose there exists $x^* \in D_0$ such that $F(x^*) = 0$ and $J(x^*)$ is nonsingular. Suppose there exists $\Gamma = (\gamma_{ij})$ with $\gamma_{ij} \geq 0$ such that

(4.4.25) $$|e_i^T[J(x) - J(y)]e_j| \le \gamma_{ij}||x - y||_2$$

for $i,j = 1,2,\ldots,n$, for all x,y D_0.

Then there exist $\epsilon,\delta > 0$ and a pivoting strategy, P_0, such that if $||x_0 - x^*||_2 < \epsilon$ and $||J_0 - J^*||_F < \delta$ then algorithm 4.2.2 generates $\{x_k\}$ which converges locally and superlinearly to x^*.

For the variations discussed following algorithms 4.2.1 and 4.2.2, convergence theorems can also be proved. In the case where a Schubert update is made every m iterations, the bounded deterioration estimate for the Schubert update would be used every m steps in the inductive proof of local convergence. Superlinear convergence is conjectured.

In the convergence proofs, the pivoting strategy was fixed throughout the iterations. Due to the nature of the local analysis, it was chosen in such a way that the factorization $P_0J(x^*)$ was well-defined. In practice, this is not feasible since x^* and $J(x^*)$ are not known. The pivoting strategy, P_0 , would be chosen so that $P_0J(x_0)$ had the appropriate LU factorization. For x_0 sufficiently close to the solution, $P_0J(x^*)$ will have all principal leading submatrices nonsingular.

5. Further applications and conclusion

5.1 Further applications.

Curtis, Powell and Reid [6] examine the problem of estimating a Jacobian, J, by finite differences when J is sparse with a known sparsity pattern. Let $F:\mathbb{R}^n \to \mathbb{R}^n$ have component functions $f_i:\mathbb{R}^n \to \mathbb{R}$, so that

$$F(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix}$$

Then $J_F = \left(\frac{\partial f_i}{\partial x_j}\right)$.

The finite difference Newton method uses the approximation

$$\frac{\partial f_i}{\partial x_j} \approx \frac{f_i(x + h_j e_j) - f_i(x)}{h_j} \tag{5.1.1}$$

with the step length h_j. This can also be written as a vector approximation to the j^{th} column of J:

$$Je_j \approx \frac{1}{h_j}[F(x + h_j e_j) - F(x)] . \tag{5.1.2}$$

The disadvantage here is that n additional function evaluations are required at each iteration.

Curtis et al. suggested grouping columns that are mutually orthogonal to allow approximating several columns at the same time by one difference. The groupings are accomplished by inspecting the columns and grouping together those which have no unknowns in common with any other column in the group. Sparsity is what makes this feasible. This technique of approximating the Jacobian reduces the number of function evaluations compared to the finite difference Newton method. The choice of step length is discussed in Curtis and Reid [5].

The above problem is considered sparse in [6] when F can be written in the form

$$F(x) = \Phi(x) + Mx$$

where M is a matrix of known derivatives and J_Φ is sparse.

Consider now a Jacobian of the form

$$J(x) = C(x) + A(x) \tag{5.1.3}$$

where $C(x)$ is easily computed and $A(x)$ is the part of the Jacobian to be approximated. Dennis [12] had suggested approximating $A(x)$ using its structure as follows:

Given x_k, x_{k+1}, and B_k approximating $J(x_k)$,

set $s_k = x_{k+1} - x_k$

$y_k = F(x_{k+1}) - F(x_k)$.

Compute $C_{k+1} = C(x_{k+1})$.

Define A_{k+1} approximating $A(x_{k+1})$ to satisfy

$$A_{k+1}s_k = y_k - C_{k+1}s_k \; . \tag{5.1.4}$$

This is a quasi-Newton equation for A_{k+1}. Now the structure of A can be preserved by an appropriate updating formula.

In particular, if $A(x)$ is known to be sparse, Schubert's method can be used to update A_k to obtain a sparse A_{k+1} satisfying (5.1.4).

Based on the convergence properties of Schubert's method, local and superlinear convergence for such an algorithm is conjectured.

5.5 Conclusions.

Gill and Murray [20] comment that advantage should be taken of special structure. The structure of a problem can be important in developing efficient methods for solution.

Schubert's method takes advantage of sparsity in solving nonlinear equations. An approximate Jacobian can be updated row-wise to preserve the sparseness. The advantage here is the reduced storage requirement over a non-sparse algorithm. There is less work; the work to compute the new approximate Jacobian is the order of the nonzeros of the Jacobian. At each iteration in the algorithm, a sparse linear system must be solved. These systems all have the same structure. The preprocessing can be done just once [16] and then applied to each linear system. Schubert's method exhibits the local and superlinear convergence properties that Broyden's method has. The proofs of these properties depended on finding the bounded deterioration estimate (2.3.2). Also, a minimum norm property is satisfied by Schubert's update.

The symmetric Schubert update can be derived in two ways. One is by an iterative projection technique [11]. The other is by variational principles [23] which also establish a minimum norm property for this update. As with the Schubert method, the advantage of reduced storage is present. At each iteration of the symmetric Schubert algorithm, two sparse symmetric linear systems must be solved. All of these systems have the same structure. Once a pivoting strategy is chosen [16], the preprocessing can be applied to all subsequent linear systems.

The algorithm is shown to be locally and linearly convergent to a solution. Again the proof depended on obtaining a bounded deterioration estimate (3.6.2). Superlinear convergence of the symmetric Schubert method is conjectured.

Matrix factorizations are suggested as a means of solving sparse nonlinear equations. The underlying update is Schubert's update which is used to preserve the structure in a triangular factor of an approximate Jacobian. In the factored form, solution of the linear systems are easily performed. The convergence of the algorithms is contingent on the continuity of the factorization. Local and suparlinear convergence are proved for factorizations based on the Crout and Dolittle decompositions.

A further application of the sparse techniques is suggested for problems in which the Jacobian can be written as $J(x) = C(x) + A(x)$. Part of the Jacobian is easily computed, $C(x)$; part of the Jacobian is to be approximated. When $A(x)$ is sparse, an algorithm is suggested and superlinear convergence is conjectured.

The convergence results depend on the Bounded Deterioration Theorem [4]. No numerical results are presented here. Schubert's method has received some testing with good results [3]. Some testing has also been done recently on the sparse symmetric updating formula [38],[39]. Toint claims a superlinear convergence property for that algorithm, however it is not presented. Moré has an indication from his testing of a QR factorization that the factored update methods perform well [26].

Several questions have been raised. The next step is to implement and test the algorithms and verify the theoretical convergence properties discussed here. Sparse nonlinear algorithms have widespread applicability, since most large scale nonlinear problems are likely to be sparse. With the interest in sparse linear problems, codes are available which will make it easier to handle the nonlinear problems using the algorithms presented and discussed here.

Bibliography

[1] K.W. Brodlie, A.R. Gourlay, and J. Greenstadt, Rank one and rank-two corrections to positive definite matrices expressed in product form, J. Inst. Math. Appl. 11(1973), pp. 73-82.

[2] C.G. Broyden, A class of methods for solving nonlinear simultaneous equations, Math. Compl 19(1965), pp. 577-593.

[3] C.G. Broyden, The convergence of an algorithm for solving sparse nonlinear systems, Math. Comp. 25(1971), pp. 285-294.

[4] C.G. Broyden, J.E. Dennis, Jr., and J.J. Moré, On the local and super-linear convergence of quasi-Newton methods, J. Inst. Math. Appl. 12(1973), pp. 223-245.

[5] A.R. Curtis and J.K. Reid, The choice of step lengths when using differences to approximate Jacobian matrices, J. Inst. Math. Appl. 13 (1974), pp. 121-126.

[6] A.R. Curtis, M.J.D. Powell, and J.K. Reid, On the estimation of sparse Jacobian matrices, J. Inst. Math. Appl. 13(1974), pp. 117-119.

[7] W.C. Davidon, Variable metric method for minimization, AEC Res. and Dev. Report ANL-5990, 1959 (revised 1966).

[8] J.E. Dennis, Jr., On convergence of Newton-like methods, in Numerical Methods for Non-linear Algebraic Equations, ed. P. Rabinowitz, Gordon and Breach, London, 1970, pp. 163-181.

[9] J.E. Dennis, Jr., On the convergence of Broyden's method for nonlinear systems of equations, Math. Comp. 25(1971), pp. 559-567.

[10] J.E. Dennis, Jr., Toward a unified convergence theory for Newton-like methods, in Nonlinear Functional Analysis and Applications, ed. L.B. Rall, Academic Press, New York, 1972, pp. 425-472.

[11] J.E. Dennis, Jr., On some methods based on Broyden's secant approximation to the Hessian, in Numerical Methods for Nonlinear Optimization, ed. F.A. Lootsma, Academic Press, London, 1972, pp. 19-34.

[12] J.E. Dennis, Jr., A brief survey of convergence results for quasi-Newton methods, in Nonlinear Programming, ed. R. Cottle and C. Lempke, SIAM-AMS Proceedings 9,1976, pp. 185-199.

[13] J.E. Dennis, Jr., and J.J. Moré, Lecture notes for Numerical Analysis, Computer Science 421-422, Cornell University, 1972-1973.

[14] J.E. Dennis, Jr. and J.J. Moré, A characterization of superlinear convergence and its application to quasi-Newton methods, Math. Comp. 28 (1974), pp. 549-560.

[15] J.E. Dennis, Jr. and J.J. Moré, Quasi-Newton methods, motivation and theory, SIAM Review 19(1977), pp. 46-89.

[16] I.S. Duff, A survey of sparse matrix research, Proceedings of the IEEE 65(1977), pp. 500-535.

[17] A.M. Erisman and J.K. Reid, Monitoring the stability of the triangular factorization of a sparse matrix, Numer. Math. 22(1974), pp. 183-186.

[18] R. Fletcher and M.J.D. Powell, A rapidly convergent descent method for minimization, Computer J. 6(1963), pp. 163-168.

[19] P.E. Gill, G. Golub, W. Murray, and M.A. Saunders, Methods for modifying matrix factorizations, Math. Comp. 28(1974), pp. 505-535.

[20] P.E. Gill and W. Murray, Quasi-Newton methods for unconstrained minimization, J. Inst. Math. Appl. 9(1972), pp. 91-108.

[21] P.E. Gill, W. Murray and M.A. Saunders, Methods for computing the LDV factors of a matrix, Math. Comp. 29(1975), pp. 1051-1077.

[22] D. Goldfarb, A family of variable-metric methods derived by variational means, Math. Comp. 24(1970), pp. 23-26.

[23] J. Greenstadt, Variations on variable-metric methods, Math. Comp. 24(1970), pp. 1-18.

[24] L.V. Kantorovich and G.P. Akilov, Functional Analysis in Normed Spaces, Fitzmatgiz, Moscow, 1959; Engl. transl., Pegamon Press, Oxford, 1964.

[25] A. Kevorkian and J. Snoek, Decomposition in large scale systems theory and applications in solving large sets of nonlinear simultaneous equations, in Decompositions of Large Scale Problems, ed. D. Himmelblau, North Holland, New York, 1973, pp. 491-516.

[26] J.J. Moré, private communication, 1977.

[27] J.M. Ortega and W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[28] T.A. Porsching, On the origins and numerical solution of some sparse nonlinear systems, in Sparse Matrix Computations, Academic Press, New York, 1976, pp. 439-453.

[29] M.J.D. Powell, A survey of numerical methods for unconstrained optimization, SIAM Review 12(1970), pp. 79-97.

[30] M.J.D. Powell, A new algorithm for unconstrained optimization, in Nonlinear Programming, ed. J.B. Rosen, O.L. Mangasarian, K. Ritter, Academic Press, New York, 1970, pp. 31-65.

[31] M.J.D. Powell, A view of unconstrained optimization, AERE Harwell Tech. Rep. C.S.S. 14, 1975.

[32] J.K. Reid, Fortran subroutines for the solution of sparse systems of nonlinear equations, AERE Harwell Tech. Rep. AERE-R7293, 1972.

[33] W.C. Rheinboldt, Methods for Solving Systems of Nonlinear Equations, SIAM, Philadelphia, 1974.

[34] W.C. Rheinboldt, On the solution of large sparse sets of nonlinear equations, in Computational Mechanics, ed. J.T. Oden, Springer-Verlag, Berlin, 1975, pp. 169-194.

[35] R. Schnabel, private communication, 1977.

[36] L.K. Schubert, Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian, Math. Comp. 24(1970), pp. 27-30.

[37] R.P. Tewarson and J.L. Stephenson, Solution of large sparse systems of nonlinear equations, College of Engr. Rep. No. 265, SUNY at Stony Brook, 1975.

[38] Ph.L. Toint, On sparse and symmetric matrix updating subject to a linear equation, Dept. of Appl. Math. and Theor. Phys. Tech. Rep. DAMTP 77/NA1, University of Cambridge, 1977.

[39] Ph. L. Toint, Some numerical results using a sparse matrix updating formula in unconstrained optimization, manuscript, Dept. of Mathematics, FNDP, Namur, Belgium, 1977.

[40] C. Van Loan, private communication, 1977.

[41] J.H. Wilkinson, The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.

[42] J.H. Wilkinson and C. Reinsch, Linear Algebra, Springer-Verlag, New York, 1971.