ANALYZING AND IMPROVING QUASI-NEWTON METHODS FOR UNCONSTRAINED OPTIMIZATION

Robert B. Schnabel

TR 77-320

A Thesis
Department of Computer Science
Cornell University
Ithaca, NY 14853

ANALYZING AND IMPROVING QUASI-NEWTON METHODS FOR UNCONSTRAINED OPTIMIZATION

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment for the Degree of

Doctor of Philosophy

by

Robert B. Schnabel

August, 1977

BIOGRAPHICAL SKETCH

Robert B. Schnabel was born on December 18, 1950 in Queens, New York, and attended grammar and high school there. From 1967 to 1971 he attended Dartmouth College, graduating summa cum laude with high distinction in mathematics. He worked for one and a half years as a computer programmer for The Chase Manhattan Bank before enrolling for graduate work in fall 1973 in the department of computer science at Cornell University. He has spent parts of two summers working in the Applied Mathematics Division of Argonne National Laboratory, and spent the year 1975-6 with Professor John Dennis at the National Bureau of Economic Research, Cambridge, Massachusetts and the Computer Science and Systems Division, Harwell Atomic Energy Research Establishment, England.

DEDICATION

To my parents.

Thanks for everything.

ACKNOWLEDGEMENTS

Thanks to Professor John E. Dennis Jr. for being such a tremendous advisor and chairman. Thanks for Professor Juris Hartmanis for his encouragement and help, to Professors David Gries, John E. Hopcroft, George L. Nemhauser, and Christopher Pottle who also served on my special committee, and to Professor Lee W. Schruben who served as a proxy member. Thanks to the National Bureau of Economic Research, Cambridge, Massachusetts, and the Atomic Energy Research Establishment, Harwell, England for providing the facilities to perform some of this research, and to Dr. David M. Gay and Professor M.J.D. Powell for all their help. Thanks to Cornell University for a first year graduate fellowship, and to the National Science Foundation for a three year graduate fellowship. Thanks to Bea McKay for the great typing job.

This research was supported in part by National Science Foundation grant 73-03413, and contract P04361 of the National Bureau of Economic Research, Cambridge, Massachusetts.

ANALYZING AND IMPROVING QUASI-NEWTON METHODS FOR UNCONSTRAINED OPTIMIZATION

Robert B. Schnabel, Ph.D.
Cornell University 1977

This thesis is concerned with analyzing and improving the performance of quasi-Newton methods for finding the minimum of a real valued function of a finite number of real variables. Quasi-Newton methods are a successful way of iteratively solving this problem when the Hessian matrix of second partial derivatives of the function cannot be cheaply computed. Instead they keep an approximation to the Hessian matrix at the current point, and update it at each iteration. A recent algorithm of Davidon's has contributed two important new ideas to the formulation of such updates, while improving upon the computational performance of existing algorithms. In this thesis we analyze the work of Davidon, propose some changes to his algorithm and test some of these, and extend much of the existing analysis of quasi-Newton updates for optimization problems to a broader class of updates which includes Davidon's.

The two innovations in Davidon's algorithm are the use of a new update class intended to preserve past derivative information in the Hessian approximation, and the selection of the actual update from this class by a method called optimal conditioning. We explain and analyze both of these aspects thoroughly. The use of the new update class seems to be very beneficial, although our analysis suggests several alternative

implementations which are equivalent to Davidon's algorithm on quadratic problems. We extend existing convergence techniques to show that two of these methods enjoy local Q-superlinear convergence under normal assumptions. On the other hand, our theoretical and computational analysis of optimal conditioning indicates that it may not be helpful in most cases, and that further computational testing definitely is required. However we introduce and test an update which uses optimal conditioning only on some selected iterations, and which seems to show promise.

To aid our analysis of Davidon's ideas we introduce the concept of a restricted update class, of which his is an example. We then extend much of the existing analysis of rank-two quasi-Newton updates, including questions of hereditary positive definiteness, minimum change updates, and performance using perfect line search, to include this broadened update class. This work helps suggest some of our proposed changes to Davidon's algorithm, and is useful in deriving and analyzing quasi-Newton updates for other problems. A basic equivalence shows that our analysis extends to a very general class of rank-two updates.

Quasi-Newton methods are currently being applied in a variety of problem areas. We hope that our analysis improves upon their use in unconstrained optimization, and helps make the leading new ideas in the field accessible to other application areas.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

This thesis is concerned with analyzing and improving the performance of quasi-Newton methods for solving nonlinear optimization problems. Quasi-Newton methods, introduced by Davidon (1959), are by now a very successful way of solving various types of nonlinear problems. However, refinements are still being made which improve the performance or extend the applicability of these methods. It is at this rather advanced stage of development of quasi-Newton methods that the work of this thesis takes place.

Thus, in order to understand the questions to be addressed in this thesis, it is necessary and appropriate to first summarize briefly the development of quasi-Newton methods and the important analysis which has been done about them. This puts us in a position to understand recent developments, largely those of Davidon (1975). This then gives us the perspective to ask for these new methods the types of questions which have been answered about previous algorithms and updates, questions like: consistent with our conditions on the update, which is the minimum change update in a specific norm?; which updates preserve positive definiteness?; how do the new algorithms behave under perfect line search?; is the "optimal conditioning" in the new methods a good idea?; are the new methods locally convergent?; how can they be improved?

These are the main questions which are answered in the body of this thesis. This introductory chapter is intended to provide a framework in which they will appear reasonable and natural ones to ask. The seven sections of this chapter discuss, briefly: 1) nonlinear optimization problems; 2) Newton's method; 3) quasi-Newton methods; 4) rank one and two update equations; 5) the behavior and analysis of leading quasi-Newton methods; 6) recent developments in quasi-Newton methods and our perspective on these developments; 7) the material covered in and organization of the remainder of this thesis. A reader with a good background in quasi-Newton methods may wish to read only sections 1.6 and 1.7.

Section 1.1 - Nonlinear Problems

The main problem we discuss in this thesis is the unconstrained minimization problem

$$\min_{x \in R^n} \quad f: R^n \to R \qquad (1.1.1)$$

where R denotes the set of real numbers and R^n Euclidean n-space. We will call a local solution to this problem x*. If f is continuously differentiable, we denote the gradient vector of first partial derivatives of f by

$$g(x) = \begin{pmatrix} df/dx_1 \\ \vdots \\ df/dx_n \end{pmatrix} \Bigg|_x$$

The gradient must be zero at a local minimum of f, though it can also be zero at a maximum or a saddle point. If f is twice continuously differentiable, we denote the Hessian matrix of second partial derivatives by

$$G(x) = \begin{bmatrix} d^2f/dx_1dx_1 \ldots d^2f/dx_1dx_n \\ \vdots \qquad\qquad \vdots \\ d^2f/dx_ndx_1 \ldots d^2f/dx_ndx_n \end{bmatrix}\Bigg|_x$$

The Hessian matrix is symmetric under these continuity assumptions. We let $L(R^n)$ denote the set of real nxn matrices. A symmetric matrix M in $L(R^n)$ is called positive definite if all its eigenvalues are positive, or equivalently if $x^TMx>0$ for all non-zero $x \in R^n$; it is called positive semi-definite if all its eigenvalues are non-negative, or equivalently if $x^TMx \geq 0$ for all $x \in R^n$. If f is twice continuously differentiable, then for x^* to be a local minimum of f, it is necessary that $g(x^*) = 0$ and $G(x^*)$ is positive semi-definite; it is sufficient that $g(x^*) = 0$ and $G(x^*)$ is positive definite. Our theory and methods seldom guarantee that a local minimum x^* is also the global solution to (1.1.1), though in practice this is often not a problem.

The methods of this thesis are also applicable to two related but more structured problems. In the first, the nonlinear least squares problem, f in (1.1.1) is defined by

$$f(x) = \frac{1}{2}\sum_{i=1}^{m} f_i(x)^2,$$

where for $i=1,\ldots,m$, $f_i:R^n \to R$. In this case, the gradient and Hessian of f have special structure which one takes advantage of in solving the problem. We will mention in Section 1.6 how this thesis was partly motivated by recent developments in solving the nonlinear least squares problem.

In the second problem, the constrained minimization problem

$$\min_{x \in D \subset R^n} \quad f: R^n \to R$$

our solution is constrained to lie in some subset D of R^n. While we do not deal with this problem in this thesis, quasi-Newton methods have been used successfully in solving it in the past (c.f., Garcia-Palomares (1973), Han (1974)), and we believe our ideas are applicable to it also.

Finally, a related but less structured problem is the solution of a system of nonlinear equations

$$F(x) = 0, \quad F: R^n \to R^n \tag{1.1.2}$$

Much of the development of quasi-Newton methods has proceeded almost concurrently on problems (1.1.1) and (1.1.2), although the actual methods are different. We do not consider the nonlinear equations problem in this thesis, but many of the ideas of Davidon's (1975) paper and this thesis are developed for the nonlinear equations problem in Gay and Schnabel (1977).

Section 1.2 - Newton's Method and Steepest Descent

Methods for solving problem (1.1.1) are iterative, proceeding from an initial guess x_0 to a series of points $x_1, x_2, \ldots, x_k$ and stopping when x_k is judged sufficiently close to a solution x^*. The two basic methods for producing this series of points are steepest descent and Newton's Method.

The method of steepest descent proceeds by the equation

$$x_{i+1} = x_i - \lambda_i g(x_i), \quad \lambda_i > 0$$

Its name derives from the fact that the negative gradient direction $-g(x_i)$

is the direction of maximum initial rate of decrease in the function f from the point x_i. In theory λ_i is chosen to minimize $f(x_i-\lambda_i g(x_i))$ among all $\lambda_i > 0$. In this case the method is guaranteed to converge to a local minimum or saddle point of f, but its rate of convergence is linear (c.f., Ortega and Rheinholdt (1970), Section 14.3), and is often too slow in practice. (Rates of convergence are discussed in Section 2.2.)

Newton's Method proceeds by the formula

$$x_{i+1} = x_i-G(x_i)^{-1}g(x_i). \tag{1.2.1}$$

This comes from solving for the δ which makes the Taylor series linearization of $g(x_i+\delta)$ about $g(x_i)$,

$$g(x_i+\delta) = g(x_i)+G(x_i)\delta$$

equal to zero. Thus Newton's Method finds the minimum of a positive definite quadratic function in one step; and for general f, provided it starts from a point x_0 sufficiently close to an x^* where $g(x^*) = 0$, its rate of convergence to x^* is quadratic (c.f., Ortega and Rheinboldt (1970), Section 10.2), which is fast. However, it may not converge at all if x_0 is too far from x^*.

Thus, steepest descent has one major asset: it is globally convergent; and two major liabilities: it converges slowly, and it must solve a one-dimensional minimization problem at each iteration. Newton's Method has one major asset: very quick local convergence; and three major liabilities: 1) it is not necessarily globally convergent; 2) it requires the Hessian matrix $G(x_i)$, which is often impossible to obtain and very expensive to even approximate; 3) at each iteration it requires the solution of a system of linear equations, as (1.2.1) is computed by solving $G(x_i)\cdot\delta = -g(x_i)$, $x_{i+1} = x_i+\delta$.

Modern algorithms combine features of both methods, attempting to be globally convergent and have a fast rate of convergence. They also find ways to circumvent the need for the Hessian matrix $G(x_i)$ in Newton's Method, for problems for which it isn't available. Methods using this approach have become known as quasi-Newton methods.

Section 1.3 - Quasi-Newton Methods

We explain in Section 1.2 that the unconstrained minimization problem (1.1.1) is usually solved by the iterative process

$$x_{i+1} = x_i+s_i$$

where

$$s_i = -\lambda_i g(x_i) \tag{1.3.1}$$

or

$$s_i = -\lambda_i B_i^{-1} g(x_i) \tag{1.3.2}$$

or s_i equals some combination of (1.3.1) and (1.3.2), and where B_i in (1.3.2) is $G(x_i)$ or an approximation to it. The first choice of s_i gives the steepest descent direction; the second gives or approximates the Newton direction. The two strategies are combined to try and achieve both the global convergence of steepest descent and the fast local convergence of Newton's Method. The pioneering work in this regard was done by Powell (1970a), and has been continued by Chien (1974) and Dennis and Mei (1975). It has been successful in obtaining both global and fast local convergence for very many problems.

The parameter λ_i in (1.3.1-2) is known as the step length parameter. It determines which point on the line $x_i-\lambda_i d_i$ we select, where d_i is the search direction. As long as $g(x_i)^T d_i>0$, we say direction $-d_i$ is a descent direction, and we know that for small λ_i, $f(x_i-\lambda_i d_i)<f(x_i)$. The process for selecting λ_i is known as a line search. If λ_i is chosen so that it is a (local) solution to

$$\min_{\lambda_i>0} \quad f(x_i-\lambda_i d_i)$$

then the process is called a perfect (or exact) line search. This is generally only theoretically possible. If it is chosen to get some prescribed decrease in $f(x_{i+1})$ from $f(x_i)$, then the process is known as an inexact line search. Line searches are generally expressive (what we mean by this is mentioned shortly), so recent algorithms tend to avoid them or keep them simple. When d_i is the Newton direction $-G(x_i)^{-1}g(x_i)$ (or an approximation to it), $\lambda_i = 1$ is often attempted as this is the value of λ_i for which Newton's Method is locally quadratically convergent. The choice $\lambda_i = 1$ is called direct prediction.

The one remaining factor that makes solution of the unconstrained minimization problem (1.1.1) difficult is that the objective function f is often a routine which is costly to evaluate. This means that analytic derivatives of f are not available, and while the gradient $g(x_i)$ can be approximated by finite differences using n additional function values of f, each approximation of $G(x_i)$ would require $\frac{n^2+3n}{2}$ function evaluations which is likely to be prohibitively costly. Therefore when f is expensive we need to use approximations to $G(x_i)$ which are less costly to compute.

We also would like our algorithm to evaluate f as infrequently as possible, as function evaluation is likely to be by far the most expensive part of our minimization algorithm.

Algorithms which use the philosophy of Newton's Method, but substitute some matrix B_i for the Hessian $G(x_i)$, are called quasi-Newton methods. Such algorithms may also combine the steepest descent strategy -- the algorithms of Powell (1974a), Chien (1974), and Dennis and Mei (1975) are all examples. However, the strategy for combining these approaches is fairly well understood, and the main question about such algorithms is how best to approximate $G(x_i)$. This is the question which is studied in this thesis.

In particular, an iteration of a quasi-Newton method will find a new point

$$x_{i+1} = x_i+s_i$$

where s_i probably depends in some way upon an approximation B_i to $G(x_i)$, and then will form an approximation B_{i+1} to $G(x_{i+1})$ by

$$B_{i+1} = U(B_i)$$

where U is some updating function. The updating function will try to incorporate information it knows about $G(x_{i+1})$ from current and previous iterations; as well as properties it knows G(x) should have, like symmetry and perhaps positive definiteness. Alternatively - since Newton's method uses only the inverse of $G(x_i)$ - the method may keep and update approximations H_i to $G(x_i)^{-1}$, and use them in place of B_i^{-1} in equation (1.3.2).

Section 1.4 - Update Equations

Suppose one has a good approximation B_i to $G(x_i)$, and has just taken a step in a quasi-Newton algorithm to point x_{i+1}. Define

$$s_i = x_{i+1}-x_i,$$

$$y_i = g(x_{i+1})-g(x_i).$$

One now wants to create a good approximation B_{i+1} to $G(x_{i+1})$. Since $G(x_{i+1})s_i \approx y_i$ (with equality for quadratic f), it is reasonable to expect

$$B_{i+1}s_i = y_i.$$

This formula is known as the quasi-Newton equation, and is the main way in which information is supplied to B_{i+1}.

Since $G(x)$ is symmetric, B_i and B_{i+1} should be too. We may also want them to be positive definite. If x_i, x_{i+1} are close to x^*, then it is likely that $G(x_i)$, $G(x_{i+1})$ are positive definite. In any case, if B_i is positive definite then the quasi-Newton direction $-B_i^{-1}g(x_i)$ is a descent direction, which is desirable - so that usually quasi-Newton algorithms keep all the B_i's positive definite. (They must be at least non-singular in order to form $-B_i^{-1}g(x_i)$).

Aside from these considerations, we know little about $G(x_{i+1})$. So (expecially when s_i is small) it is reasonable to expect it to be similar to $G(x_i)$; and thus to form B_{i+1} by changing B_i as little as possible consistent with the above requirements. The smallest rank change to B_i is a matrix of rank one; and there is one update which preserves symmetry and obeys the quasi-Newton equation, the symmetric rank one (SR1) update

$$\bar{B} = B + \frac{(y-Bs)(y-Bs)^T}{\langle y-Bs,s\rangle} \qquad (1.4.1)$$

(We denote by $\langle v,w\rangle$ the inner product v^Tw of two vectors $v,w \in R^n$. Also, we have dropped the subscripts i and replaced B_{i+1} with $\bar{B}$). However, $\bar{B}$ may not be positive definite even when B is, and furthermore can be badly conditioned if the denominator is close to zero. So this update is not in general satisfactory, although we will see later that it has come back into favor for occasional use.

Another measure of least change is gotten by asking $\bar{B}$ to satisfy

$$\min \quad \{\|\bar{B}-B\| \;\big|\; \bar{B}s = y, \;\; \bar{B} \text{ symmetric}\} \qquad (1.4.2)$$

where $\|\cdot\|$ is some matrix norm and B is symmetric. Usually B will be positive definite and we will want $\bar{B}$ to be positive definite also. For this to be possible, we must have $s^T\bar{B}s = s^Ty>0$, but this condition turns out to be satisfied by most algorithms (for instance, if perfect line search is used, then it is true as long as s is a descent direction). We will explain in Chapter 5 that an appropriate norm in problem (1.4.2) is then the Frobenius norm weighted by $\bar{B}^{-1/2}$; and the unique solution is then the rank two update

$$\bar{B} = B + \frac{(y-Bs)y^T + y(y-Bs)^T}{\langle y,s\rangle} - \frac{\langle y-Bs,s\rangle yy^T}{\langle y,s\rangle^2} \qquad (1.4.3)$$

(Dennis and Moré (1977)). This update was first proposed by Davidon (1959), and explained by Fletcher and Powell (1963), and is known as the Davidon-Fletcher-Powell (DFP) update. It was the leading quasi-Newton update for a number of years.

The preceding analysis can also be done for updates of an approximation H to $G(x)^{-1}$ into an approximation $\bar{H}$ to $G(\bar{x})^{-1}$. In this case the quasi-Newton equation is

$$\bar{H}y = s,$$

and if we have H symmetric and positive definite and $\langle s,y\rangle > 0$, then the unique solution to

$$\min \{ \|\bar{H}-H\| \mid \bar{H}y = s, \ \bar{H} \text{ symmetric and positive definite}\}$$

is

$$\bar{H} = H + \frac{(s-Hy)s^T + s(s-Hy)^T}{\langle s,y\rangle} - \frac{\langle s-Hy,y\rangle ss^T}{\langle s,y\rangle^2} \tag{1.4.4}$$

where $\|\cdot\|$ is now the Frobenius norm weighted by $\bar{H}^{-1/2}$ (Dennis and Moré (1977)). This update was proposed in 1970 by Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970) and so is known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) update. It has replaced the DFP as the most successful quasi-Newton update.

If $H = B^{-1}$, then the inverse $\bar{B}$ of the BFGS $\bar{H}$ given by (1.4.4) can be calculated by

$$\bar{B} = B + \frac{yy^T}{\langle y,s\rangle} - \frac{Bss^TB}{\langle s,Bs\rangle} . \tag{1.4.5}$$

Thus the DFP and BFGS can also be written as the choices $\phi = 0,1$ in the one-parameter class of updates

$$\bar{B}(\phi) = B + \frac{(y-Bs)y^T + y(y-Bs)^T}{\langle y,s\rangle} - \frac{\langle y-Bs,s\rangle yy^T}{\langle y,s\rangle^2}$$

$$- \phi\langle s,Bs\rangle \left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]\left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]^T . \qquad (1.4.6)$$

This class was first proposed by Broyden (1967). For all updates in it, $\bar{B}s = y$, $\bar{B}$ is symmetric if B is symmetric, and the column and row spaces of $\bar{B}$-B are in the subspace spanned by y and Bs. The choice $\phi = -\langle s,Bs\rangle / \langle y-Bs,s\rangle$ corresponds to the SR1 update (1.4.1). The matrices $\bar{B}(\phi)$, $\phi\epsilon[0,1]$ are known as the convex class. If B is symmetric and positive definite and $\langle s,y\rangle > 0$, then $\bar{B}(\phi)$ is symmetric and positive definite for the entire convex class (Fletcher (1970)).

One can see a duality between equations (1.4.3) and (1.4.4), with $H,\bar{H},y,s$ taking the places of $B,\bar{B},s,y$ respectively. Similarly there is a dual inverse representation of the one parameter class: the BFGS and DFP are the choices $\phi = 0,1$ in

$$\bar{H}(\phi) = H + \frac{(s-Hy)s^T + s(s-Hy)^T}{\langle s,y\rangle} - \frac{\langle s-Hy,y\rangle ss^T}{\langle s,y\rangle^2}$$

$$- \phi\langle y,Hy\rangle \left[\frac{s}{\langle s,y\rangle} - \frac{Hy}{\langle y,Hy\rangle}\right]\left[\frac{s}{\langle s,y\rangle} - \frac{Hy}{\langle y,Hy\rangle}\right]^T \qquad (1.4.7)$$

The choices $\bar{H}(\phi)$, $\phi\epsilon[0,1]$ give the $\bar{H}$ form of the convex class.

Section 1.5 - Behavior of the Leading Methods

Our development of quasi-Newton methods to this point gives no real assurance that they will work - that they will solve unconstrained minimization problems. In practice they do. Various implementations of the DFP and BFGS, and related algorithms such as Powell's hybrid (1970a) have all been successful, meaning that they solve most problems at a reasonable cost. The consensus has lately developed that the BFGS update does a better job than the DFP (c.f., Brodlie (1976)).

Quasi-Newton implementations have moved away from the use of perfect or almost perfect line search, and this is significant because Dixon (1972a) has shown that if perfect line search is used, then the sequence of points x_i generated by the algorithm

$$\begin{aligned} &x_{i+1} = x_i - \lambda_i B_i^{-1} g(x_i) \\ &B_{i+1}(\phi) \text{ given by (1.4.6)} \end{aligned} \tag{1.5.1}$$

is independent of the choices of ϕ. However, without perfect line search there seems to be a fair difference in performance between various methods, for instance the DFP and the BFGS. Also, if perfect line search is used, any method from the one-parameter class will find the exact minimum of a quadratic function in n+1 or fewer steps (Fletcher and Powell (1963) and Dixon (1972)) but without perfect line search, none of these methods is in general exact on quadratics.

In practice, we hope to let $\lambda_i = 1$ in (1.5.1) as often as possible - whenever it provides a reasonable decrease in f(x). (This choice of λ_i avoids the cost in function evaluations of a line search.) Broyden, Dennis, and Moré (1973) have shown that the DFP and BFGS are locally Q-superlinearly convergent in this case, under reasonable conditions. As long as G(x) is Lipschitz continuous near the minimum x^* of f and $G(x^*)$ is positive definte, they show that once $\|B_i-G(x^*)\|$ and $\|x_i-x^*\|$ become sufficiently small, that $\|B_k-G(x^*)\|$ stays small for all $k>i$ and x_k

converges superlinearly to x^* as $k \to \infty$. Q-superlinear convergence is not as fast as the quadratic rate of Newton's method, but it is still a fast rate of convergence on the computer; and in practice, once these methods are near to a solution, they hone in on it quickly. The techniques of Stoer (1975) show also that any update from the convex class ($\phi \epsilon [0,1]$) is locally Q-linearly convergent using $\lambda_i = 1$ in (1.5.1); but only the DFP and BFGS ($\phi = 0,1$) have been shown to have Q-superlinear convergence.

Section 1.6 - Recent Developments

Recently, Davidon (1975) introduced a new quasi-Newton algorithm which has performed better than the DFP and BFGS in practice. It differs from these algorithms mainly in that it uses a new update equation

$$\bar{B}(\phi) = B + \frac{(y-Bs)v^T + v(y-Bs)^T}{\langle v,s\rangle} - \frac{\langle y-Bs,s\rangle vv^T}{\langle v,s\rangle^2}$$

$$- \phi \left[\frac{y-Bs}{\langle y-Bs,s\rangle} - \frac{v}{\langle v,s\rangle}\right]\left[\frac{y-Bs}{\langle y-Bs,s\rangle} - \frac{v}{\langle v,s\rangle}\right]^T \qquad (1.6.1)$$

where v is not (necessarily) in the subspace spanned by y and Bs. Instead v is chosen to be orthogonal to a certain subspace Z (which is also orthogonal to y-Bs). Z changes from iteration to iteration, in such a way that if $x_{i+1} = x_i - \lambda_i B_i^{-1} g(x_i)$ is used, then the algorithm will find the exact minimum of a quadratic function in n+1 or fewer steps, using inexact line search or even $\lambda_i = 1$ at each iteration.

In Section 2.4 we give a derivation of Davidon's algorithm as we understand it. We view Davidon's update as satisfying the additional restriction,

$$(\bar{B}-B)z = 0 \text{ for all } z \in Z, \qquad Z \text{ an m dimensional subspace of } R^n,\ m<n \tag{1.6.2}$$

as well as $\bar{B}s = y$, $\bar{B}$ symmetric and positive definite. Davidon uses this additional restriction to enable him to retain past quasi-Newton equations $B_{i+1}s_j = y_j$, $j<i$, for quadratic problems, and this leads to the finite termination for quadratic problems without using line searches. Therefore we consider restriction (1.6.2) an important new idea in quasi-Newton algorithms.

Davidon shows that there exist vectors $\hat{y}$, $\hat{s}$ such that his update (1.6.1) can be written

$$\bar{B}(\hat{\phi}) = B + \frac{(\hat{y}-B\hat{s})\hat{y}^T + \hat{y}(\hat{y}-B\hat{s})^T}{\langle \hat{y},\hat{s}\rangle} - \frac{\langle \hat{y}-B\hat{s},\hat{s}\rangle \hat{y}\hat{y}^T}{\langle \hat{y},\hat{s}\rangle^2}$$
$$- \hat{\phi}\langle \hat{s},B\hat{s}\rangle \left[\frac{\hat{y}}{\langle \hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle \hat{s},B\hat{s}\rangle}\right]\left[\frac{\hat{y}}{\langle \hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle \hat{s},B\hat{s}\rangle}\right]^T \tag{1.6.3}$$

where $\hat{y}$ and $\hat{s}$ are certain projections of y and s. (This transformation is discussed in Section 2.4 and in Chapter 3.) However, rather than use the DFP or BFGS choices, $\hat{\phi} = 0,1$ respectively, he chooses $\hat{\phi}$ by the "optimal conditioning" criterion also suggested by Oren and Spedicato (1976). This criterion seeks to hold down the condition number of $\bar{B}(\hat{\phi})$ by choosing $\hat{\phi}$ to minimize the condition number of $B^{-1/2}\bar{B}(\hat{\phi})B^{-1/2}$, which is a bound on the condition number of $\bar{B}(\hat{\phi})$ (see Chapter 7). The $\hat{\phi}$ so chosen may be within or outside the convex class $\hat{\phi}\in[0,1]$, and sometimes gives the choice $\hat{\phi} = \frac{-\langle \hat{s},B\hat{s}\rangle}{\langle \hat{y}-B\hat{s},\hat{s}\rangle}$, the symmetric rank one update. This strategy for choosing $\hat{\phi}$ is rather different from any previous quasi-Newton algorithm, and we consider it the other important new idea in Davidon's algorithm.

Actually, Davidon maintains and updates an approximation H to the inverse of the Hessian, using the "dual" formula to (1.6.1).

$$\bar{H}(\phi) = H + \frac{(s-Hy)w^T + w(s-Hy)^T}{\langle w,y\rangle} - \frac{\langle s-Hy,y\rangle ww^T}{\langle w,y\rangle^2}$$

$$- \phi\left[\frac{s-Hy}{\langle s-Hy,y\rangle} - \frac{w}{\langle w,y\rangle}\right]\left[\frac{s-Hy}{\langle s-Hy,y\rangle} - \frac{w}{\langle w,y\rangle}\right]^T \qquad (1.6.4)$$

where w is not (necessarily) in the subspace spanned by s and Hy, and is orthogonal to some subspace Z_H (which is orthogonal to s-Hy). However, the analysis of formulas (1.6.1) and (1.6.4) is virtually identical, and we use mainly the former as we find the concepts associated with it slightly more intuitive.

Our interest in Davidon's (1975) ideas is strengthened by the fact that Dennis and Mei (1975) have used Davidon's updating strategy in a hybrid algorithm like Powell's (1970a), and have gotten better computational results than either. Also, Dennis (1976), has recently introduced a quasi-Newton algorithm for the nonlinear least squares problem which uses an update of form (1.6.1). This contributed to our interest in studying such updates.

Finally, Huang (1970) and Oren (1972) have suggested using two additional parameters in the rank-two quasi-Newton update. They suggest replacing B with γB and y with ρy in (1.4.6), so that the update equation becomes

$$\overline{B}(\phi,\rho,\gamma) = \gamma B + \frac{(\rho y-\gamma Bs)y^T+y(\rho y-\gamma Bs)^T}{\langle y,s\rangle} - \frac{\langle\rho y-\gamma Bs\rangle yy^T}{\langle y,s\rangle^2}$$

$$- \phi\gamma\langle s,Bs\rangle\left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]\left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]^T \quad (1.6.5)$$

The effect of the parameter γ is equivalent to making the change $B = \gamma B$ before making the update. This allows a simple rank n change to the matrix B. The introduction of the parameter ρ changes the quasi-Newton equation to $\overline{B}s = \rho y$. These modifications have been proposed to ensure non-singularity of $\overline{B}$, and to take account of non-quadratic behavior of the objective function. Experimentation with the use of these parameters in algorithms is still going on.

Section 1.7 - Topics to be Covered

In this thesis we are mainly concerned with examining Davidon's (1975) algorithm; and the properties of the type of update he introduces, which requires $\overline{B}$ to be in the set

$$S = \{\overline{B}\in L(R^n) | \overline{B} \text{ symmetric and positive definite,}$$

$$\overline{B}s = y,\ (\overline{B}-B)z = 0 \text{ for all } z\in Z\} \quad (1.7.1)$$

where B is a symmetric and positive definite matrix $\in L(R^n)$, s and y are vectors $\in R^n$, s nonzero, and Z is an m dimensional subspace of R^n, $m<n$. Many of the questions we ask and answer are guided by the past history and analysis of quasi-Newton methods, which have been reviewed in this chapter.

In Chapter 2 we discuss background material which is required by the remainder of this thesis. We review the vector and matrix norms we will ... convergence; and projection matrices and their properties, which we will employ in discussing the new quasi-Newton updates. We then derive and explain Davidon's (1975) update for quadratic problems, and discuss the various ways it may be extended to non-quadratic functions. We indicate what we think are the fundamental new ideas in quasi-Newton methods introduced by Davidon's algorithm.

In Chapter 3 we demonstrate two important algebraic equivalences between one parameter classes of rank-two updates. The first shows that the general new class (1.6.1) can be rewritten in the familiar form (1.6.3), for appropriate values of $\hat{y}$, $\hat{s}$, and $\hat{\phi}$. This allows us to define "restricted DFP" and "restricted BFGS" updates which will be useful to us; and to extend the analysis of updates of form (1.6.3) to the more general updates (1.6.1). The second equivalence shows that for certain values of ϕ, the one-parameter class (1.6.1) can be rewritten as

$$\bar{B}(\sigma) = B + \frac{(y-Bs)u^T+u(y-Bs)^T}{\langle u,s\rangle} - \frac{\langle y-Bs,s\rangle uu^T}{\langle u,s\rangle^2},$$

$$u = v+\sigma(y-Bs).$$

This parameterization is shown to contain all the "important" updates, and in many ways is more convenient to use. It is helpful in our discussion of optimal conditioning in Chapter 7.

In the following two chapters we discuss the important considerations in selecting updates from the set S of restricted updates (equation 1.7.1). In Section 2.4 we will have demonstrated the necessary conditions on Z in order for S to be nonempty when the restriction that $\bar{B}$ be positive definite is omitted. In Chapter 4 we complete the analysis of when S is nonempty by showing the necessary and sufficient conditions such that a positive $\bar{B}\in S$ exists. We find that the larger Z is, the less likely it is for a positive definite update to exist. This result has important implications to the use of Davidon's (1975) algorithm.

Then in Chapter 5 we derive the updates in S which minimize $\|\bar{B}-B\|$ or $\|\bar{H}-H\|$, for the same class of norms used by Dennis and Moré (1977). For the same norms for which Dennis and Moré found the DFP and BFGS to be the least change updates in the case when Z is the null set, we show that the corresponding restricted DFP and BFGS updates are the least change updates when a general set Z is used in defining S. These results also help us comment and advise upon the updates used in Davidon's (1975) algorithm.

In Chapter 6 we examine the behavior of the new general one-parameter class of rank-two updates (1.6.1) using perfect line search. Davidon (1975) has written that the points generated by his algorithm using perfect line search are independent of the choices of ϕ in (1.6.1); but this is because his class of updates reduces to the DFP-BFGS class (1.4.6) in this case. However, if perfect line search is started in Davidon's algorithm at any step after the first, we show that the updates will not (in general) correspond to the DFP-BFGS class, and the subsequent points generated will depend upon the choices of ϕ. In general, the points generated by a quasi-Newton algorithm (1.5.1) using updates from a one-parameter class (1.6.1) are dependent upon the choices of ϕ, unless the class of updates is the DFP-BFGS class. These results indicate that the choice of ϕ may be even more important when using updates from our new one-parameter classes.

Next we turn our attention to the procedure called optimal conditioning used in Davidon's (1975) algorithm to select the parameter ϕ in (1.6.3). First in Chapter 7 we give our derivation of the optimal conditioning selection process, resulting in three graphs which aid in our understanding and subsequent analysis. Using the equivalences of Chapter 3, we see that we have extended the optimal conditioning process to any one-parameter class (1.6.1).

Then we analyze how much we are gaining by making Davidon's "optimally conditioned" choice of ϕ, rather than the traditional BFGS or convex class selections. Davidon does optimal conditioning to minimize a bound $K(\phi)$ on the condition number of $\bar{B}(\phi)$; and in minimizing $K(\phi)$ can choose ϕ within or outside the convex class. However we demonstrate that there is really no reason to go outside the convex class from the standpoint of optimal conditioning, for if the minimum bound K_1 is attained by an update outside the convex class, then among all updates in the convex class the minimum bound K_2 is attained by the DFP or the BFGS, and $K_2 < \frac{32}{27} K_1$ (whereas Davidon was concerned about his choice of ϕ making orders of magnitude difference in the condition number of $\bar{B}(\phi)$). Furthermore, we show that as an algorithm converges to a solution, that the bounds $K(\phi)$ approach one for any member of the convex class as well as for the "optimally conditioned" update, so that optimal conditioning becomes unimportant. These conclusions are independent of the meanings of y and s or $\hat{y}$ and $\hat{s}$ in (1.4.6) or (1.6.3).

There are however cases, away from the solution, when the update within the convex class which minimizes the bound on $K(\phi)$ is much better conditioned that the BFGS. The analysis of Chapter 7 therefore leads us to propose a new strategy for choosing ϕ in Davidon's algorithm: choose the "optimally conditioned" update in the convex class, except choose the BFGS instead if it is almost as well conditioned. In Chapter 8 we present computational results using Davidon's (1975) algorithm with this modification. Our algorithm turns out to choose the BFGS most but not all of the time, and performs better than Davidon's algorithm using the BFGS alone, which in turn usually does much better than Davidon's algorithm using optimal conditioning. Mei (1977) had already found that his and Dennis' hybrid algorithm, which uses Davidon's updating strategy, usually performs better with the BFGS than with optimal conditioning. We also find that his algorithm performs slightly better using our combination strategy.

Finally in Chapter 8, we show that our combination optimal conditioning - BFGS strategy does not destroy the Q-superlinear convergence of the BFGS algorithm, when we are dealing with the standard DFP-BFGS class (1.4.6). This is because our algorithm is shown asymptotically to always select the BFGS. Our analysis and experience encourages us to suggest our new strategy in various applications.

In Chapter 9 we discuss ways to extend Davidon's (1975) quadratic algorithm to non-quadratic problems, and analyze the convergence of some of these methods for general nonlinear problems. It has not yet been shown that Davidon's new choices of $\hat{y}$ and $\hat{s}$ in (1.6.3) lead to a locally convergent algorithm, even using only the BFGS choice $\hat{\phi} = 1$. However, we are able to show that a slightly different choice of $\hat{y}$ and $\hat{s}$, which is equivalent to Davidon's for quadratic problems, yields a locally Q-superlinearly convergent algorithm. Our assumptions and techniques are similar to those of Broyden, Dennis, and Moré (1973). The algorithm for which we prove convergence is asymptotically so similar to Davidon's that this proof strategy strongly suggests that Davidon's algorithm is Q-superlinearly convergent.

Finally, we consider the effect of the introduction of the Huang and Oren parameters γ and ρ (see equation 1.6.5) on the local convergence of the standard DFP or BFGS algorithm. In Chapter 10 we give sufficient conditions upon both parameters such that the algorithms will continue to be locally Q-superlinearly convergent under the assumptions of Broyden, Dennis, and Moré (1973). We also show a somewhat weaker necessary condition on these two parameters for Q-superlinear convergence.

In Chapter 11 we summarize our findings and recommendations. We also indicate directions for further research which we believe are suggested by our work.

CHAPTER II

BACKGROUND

In this chapter we present material which is fundamental to the remainder of the thesis. Sections 2.1-2.3 cover mathematical topics. In Section 2.1 we introduce the vector and matrix norms we will use, and their properties which we will need. Section 2.2 defines the rates of convergence which are used in the later chapters. In Section 2.3 we define and prove the important properties of a generalized projection matrix which will be very useful to us in discussing the new quasi-Newton updates.

Finally, in Section 2.4 we motivate and explain the update used in Davidon's (1975) recent algorithm which has led to much of the work of this thesis. We show that it interests us in quasi-Newton updates with the new property that $\bar{B}$-B is orthogonal to a subspace Z of R^n, and that this is exactly the property that Davidon uses to get an algorithm with finite termination on quadratic problems. We give several alternative forms of Davidon's update which are identical on quadratics and will be important to us later. For non-quadratic functions, we show that the new property that $\bar{B}$-B be orthogonal to a subspace Z is not, as it is used in the quadratic case, in general consistent with the requirements that $\bar{B}$ be symmetric and satisfy the quasi-Newton equation. However, various forms of the quadratic update can be adapted to non-quadratic problems, and we exhibit some of the reasonable possibilities, including the update Davidon (1975) chooses. The properties of and differences between these various updates when applied to general nonlinear problems will be discussed extensively in the remainder of this thesis.

Section 2.1 - Norms

We will have many occasions to use vector and matrix norms - for example when deriving least change updates in Chapter 5, and when studying convergence properties in Chapters 8 - 10. We will use the ℓ_2 (Euclidean) vector norm; occasionally its induced matrix norm; but more often a weighted Frobenius norm. We define these norms below, and list those properties which are important to us in some portion of this thesis. For a statement of the properties required of vector and matrix norms, and a good reference for this material, see Stewart (1973), Chapter 4.

<u>Definition 2.1.1</u> Let $v \in R^n = (v_1, \ldots, v_n)$, $A \in L(R^n)$.

The ℓ_2 (Euclidean) norm of v, denoted $\|v\|_2$, is defined as

$$\|v\|_2 = \left(\sum_{i=1}^{n} v_i^2\right)^{1/2} .$$

The ℓ_2-induced norm of A, denoted $\|A\|_2$, is defined as

$$\|A\|_2 = \max_{\substack{v \in R^n \\ v \neq 0}} \frac{\|Av\|_2}{\|v\|_2} .$$

<u>Definition 2.1.2</u> Let $A \in L(R^n)$. The Frobenius norm of A, denoted $\|A\|_{Fr}$, is defined as the ℓ_2 norm of the elements of A,

$$\|A\|_{Fr} = \left(\sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij}^2\right)^{1/2}.$$

Equivalently, $\|A\|_{Fr}$ can be computed by

$$\|A\|_{Fr} = \left(\sum_{j=1}^{n} \|Au_j\|_2^2\right)^{1/2} = (\text{trace } A^T A)^{1/2}$$

[illegible] onormal [illegible]

The weighted Frobenius norm of A, $\|A\|_{Fr,M}$ is defined by

$$\|A\|_{Fr,M} = \|MAM\|_{Fr}$$

where M is a symmetric non-singular matrix $\epsilon L(R^n)$.

Lemma 2.1.3 Let A, $B \epsilon L(R^n)$; u, $v \epsilon R^n$. Then the following properties hold whether $\|\cdot\|$ is the ℓ_2-induced matrix norm or the Frobenius norm:

i) $\|A+B\| \leq \|A\| + \|B\|$

ii) $\|AB\| \leq \|A\| \cdot \|B\|$

iii) $\|Av\|_2 \leq \|A\| \; \|v\|_2$

iv) $\|uv^T\| = \|u\|_2 \; \|v\|_2$

The following properties also hold:

v) $\|AB\|_{Fr} \leq \min \{\|A\|_{Fr} \|B\|_2, \|A\|_2 \; \|B\|_{Fr}\}$

vi) if $\|\cdot\|_P$ and $\|\cdot\|_Q$ are any two matrix norms, then there exists a constant $\eta \geq 1$ such that for any matrix $M \epsilon L(R^n)$, $\|M\|_P \leq \eta \; \|M\|_Q$.

vii) if matrix $M \epsilon L(R^n)$ is symmetric, $\|M\|_2$ = the largest absolute value of the eigenvalues of M.

Section 2.2 - Rates of Convergence

In this section we define the rates of convergence which we will use in this thesis. A sequence of points $\{x_i\}$ is said to converge to x* if

$\lim_{i \to \infty} \| x_i - x^* \| = 0$. The rate of convergence concerns how quickly $\| x_i - x^* \|$ decreases. The three rates of convergence used in this thesis are Q-linear, Q-superlinear, and Q-quadratic. Q-linear convergence is considered the slowest rate of convergence detectable on a computer, Q-superlinear and Q-quadratic are both fast rates of convergence. Many quasi-Newton methods have Q-superlinear convergence; Newton's Method is Q-quadratic, which is faster.

The prefix "Q" differentiates from R-order convergence, which is differently defined. The definitive reference is Ortega and Rheinboldt (1970), Chapter 9.

We denote the open interval from a to b, $a,b \in R$ and $a<b$, by (a,b), and the closed interval from a to b by $[a,b]$. Symbols $[a,b)$ and $(a,b]$ denote intervals closed at one end and open at the other.

<u>Definition 2.2.1</u> Let $x_1, x_2, \ldots$ be a sequence in R^n, $x \in R^n$, $\| \cdot \|$ a vector norm. Then $\{x_i\}$ is said to converge to x^* if $\lim_{i \to \infty} \| x_i - x^* \| = 0$.

The sequence $\{x_i\}$ is said to converge Q-linearly to x^* if there exists a constant $k_1 \in (0,1)$ and an integer i_0 such that for all $i > i_0$,

$$\| x_{i+1} - x^* \| \leq k_1 \| x_i - x^* \|$$

If in addition

$$\lim_{i \to \infty} \frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} = 0$$

then $\{x_i\}$ is said to converge Q-superlinearly to x^*.

The sequence $\{x_i\}$ is said to converge Q-quadratically to x^* if there exists a constant $k_2 \geq 0$ and an integer i_1 such that for all $i > i_1$

$$\|x_{i+1}-x^*\| \leq k_2\|x_i-x^*\|^2$$

Section 2.3 - Projection Matrices

We will use projection matrices as an algebraic tool when we discuss the new quasi-Newton updates. They fit in naturally with the restriction that these updates be orthogonal to a certain subspace Z. Hopefully they also aid the geometric understanding of the new methods.

In this section we define the general projection matrices we will use, and prove their important properties.

Definition 2.3.1 Let Z be an m dimensional subspace of R^n, $0 \leq m < n$; $M \in L(R^n)$ symmetric and positive definite. Let $Z^{\perp}(M) = \{u \in R^n | \langle u, Mz \rangle = 0$ for all $z \in Z\}$. Then a projection matrix projecting orthogonal to Z in the metric M (or equivalently, onto $Z^{\perp}(M)$ in the metric M) is a matrix P such that: $P^2 = P$; $Pz = 0$ for all $z \in Z$; $Pv = v$ for all $v \in Z^{\perp}(M)$.

Lemma 2.3.2 Let $Z, M, Z^{\perp}(M)$ be defined as in Definition 2.3.1. Let $z_1, \ldots, z_m$ be a basis for Z such that $\langle z_j, Mz_k \rangle = 0$, $1 \leq j < k \leq m$; $u_1, \ldots, u_{n-m}$ a basis for $Z^{\perp}(M)$ such that $\langle u_j, Mu_k \rangle = 0$, $1 \leq j < k \leq n-m$. Then the unique projection matrix projecting orthogonal to Z in the metric M is

$$P = I - \sum_{i=1}^{m} \frac{z_i z_i^T M}{\langle z_i, Mz_i \rangle} \tag{2.3.1a}$$

$$= \sum_{i=1}^{n-m} \frac{u_i u_i^T M}{\langle u_i, Mu_i \rangle} \tag{2.3.1b}$$

and $MP = P^T M$.

Proof: Using either (2.3.1a) or (2.3.1b), straight forward algebra shows $P^2 = P$ and $MP = P^TM$. By expressing any vector $z\varepsilon Z$ as $\sum_{i=1}^{m} \alpha_i z_i$ for some constants α_i, and any vector $u\varepsilon Z^{\perp}(M)$ as $\sum_{j=1}^{n-m} \beta_j u_j$ for some constants β_j, we also get from straightforward algebra that $Pz = 0$ and $Pu = u$, using either (2.3.1a) or (2.3.1b). Both proofs rely on the fact that any two vectors $v, w \varepsilon \{z_1,\ldots,z_m,u_1,\ldots,u_{n-m}\}$ obey $\langle v,Mw\rangle = 0$.

To show that P is unique, observe that since M is positive definite, $z_1,\ldots,z_m,u_1,\ldots,u_{n-m}$ are linearly independent and form a basis for R^n. Thus any vector $v\varepsilon R^n = \sum_{i=1}^{m} \alpha_i z_i + \sum_{j=1}^{n-m} \beta_j u_j$ for some constants α_i, β_j. Now let $\bar{P}$ be any projection matrix projecting orthogonal to Z in the metric M. Then by Definition 2.3.1, $\bar{P}v = \sum_{i-1}^{n-m} \beta_j u_j = Pv$. Hence for all $v\varepsilon R^n$, $(\bar{P}-P)v = 0$; so $\bar{P} = P$.

The following is the main form in which we will use projection matrices. We write MZ to denote the m dimensional subspace of R^n spanned by $Mz_1, Mz_2,\ldots,Mz_m$, where M is a non-singular matrix $\varepsilon L(R^n)$, and Z is an m dimensional subspace of R^n spanned by vectors $z_1,z_2,\ldots,z_m \varepsilon R^n$, $m\leq n$.

Lemma 2.3.3 Let Z be an m dimensional subspace of R^n, $0\leq m<n$; $M\varepsilon L(R^n)$ symmetric and positive definite. Define $Z^{\perp} = \{u\varepsilon R^n | \langle u,z\rangle = 0 \text{ for all } z\varepsilon Z\}$. Let $z_1,\ldots,z_m$ be a basis for Z such that $\langle z_j,Mz_k\rangle = 0$, $1\leq j<k\leq m$, and let $u_1,\ldots,u_{n-m}$ be a basis for $Z^{\perp}$ such that $\langle u_j,M^{-1}u_k\rangle = 0$, $1\leq j<k\leq n-m$. Then the unique matrix projecting onto $Z^{\perp}$, or orthogonal to MZ, in the metric M^{-1} is

$$P = I - \sum_{i=1}^{m} \frac{Mz_i z_i^T}{\langle z_i, Mz_i \rangle} \tag{2.3.2a}$$

$$= \sum_{i=1}^{n-m} \frac{u_i u_i^T M^{-1}}{\langle u_i, M^{-1} u_i \rangle} \tag{2.3.2b}$$

and $PM = MP^T$.

Proof: We apply Lemma 2.3.2, with $MZ, Z^{\perp}$ and M^{-1} in Lemma 2.3.3 taking the places of Z, $Z^{\perp}(M)$ and M in Lemma 2.3.2. Then $PM = MP^T$ and (2.3.2b) follow immediately. Finally from equation (2.3.1a),

$$P = I - \sum_{i=1}^{m} \frac{w_i w_i^T M^{-1}}{\langle w_i, M^{-1} w_i \rangle} \tag{2.3.3}$$

where $w_1, \ldots, w_m$ is a basis for MZ such that $\langle w_j, M^{-1} w_k \rangle = 0$, $1 \leq j < k \leq m$. Since $Mz_1, \ldots, Mz_m$ is such a basis for MZ, substituting $w_i = Mz_i$ into (2.3.3) gives (2.3.2a).

In the remainder of this thesis, we will denote a subspace Z spanned by m (not necessarily linearly independent) vectors $z_1, \ldots, z_m \in R^n$ by $[z_1, \ldots, z_m]$.

Section 2.4 - Davidon's 1975 Algorithm

Many people consider the way a minimization routine acts on a (positive definite) quadratic function to be important. This is because a general nonlinear function resembles a quadratic near its minimum. Newton's method finds the exact minimum of a quadratic in one step, without a line search. So it is perhaps surprising that the DFP and BFGS methods, while they are locally Q-superlinearly convergent on general nonlinear functions, do not (necessarily) arrive at the exact minimum of a quadratic in a finite number of steps, unless exact line searches are used. Davidon (1975) set out to modify these quasi-Newton methods so that they would work exactly on quadratics, without requiring line searches at all.

In this section we motivate Davidon's approach to this problem. We see that we would like our update to preserve past quasi-Newton equations, which leads us to the idea of restricting our updates so that at each iteration, $B_{i+1}-B_i$ is orthogonal to a subspace Z_i. This is the basis of Davidon's algorithm for quadratics, which has several equivalent forms.

For non-quadratic problems, we show that the restriction of $B_{i+1}-B_i$ to the desired Z_i is no longer consistent with the quasi-Newton equation $B_{i+1}s_i = y_i$. Therefore we can modify our objectives in several ways. These lead to various algorithms, which are different on non-quadratic problems but all reduce to Davidon's basic algorithm for quadratics. Davidon (1975) proposes one of these algorithms for general problems; we discuss several in this thesis. We think this situation is analagous to that for conjugate gradient methods, where several different algorithms - Hestenes-Stiefel, Fletcher-Reeves, Polak-Ribiere - all are generalizations of the same method for quadratic problems.

Davidon's (1975) paper and algorithm are concerned with approximations H_i to the inverse Hessian $G(x_i)^{-1}$. In this section we will instead consider using approximations B_i to the Hessian $G(x_i)$, because we find this a bit easier conceptually. We are also inclined to take this approach because in some applications, like nonlinear least squares, we are only interested in approximating a derivative matrix, and not its inverse. In any event, the principles used in developing the two algorithms are almost identical, and the updates produced are dual in the sense that the DFP and

BFGS are. It may be the case that the inverse approach produces a better computational method.

For convex quadratic problems $f(x) = \frac{1}{2}x^TAx - b^Tx$, $b,x\epsilon R^n$, $A\epsilon L(R^n)$ symmetric and positive definite, the Hessian $G(x)$ equals A at all points x. Furthermore, if our approximate Hessian B_i equals A at any iteration, then the next direct prediction quasi-Newton step, $x_{i+1} = x_i - B_i^{-1}\, g(x_i)$, yields the exact minimum of f. So we would like our quasi-Newton method to eventually give $B_i = A$.

This seems a reasonable aim, since for quadratics, $y_i = As_i$, $(s_i = x_{i+1} - x_i,\ y_i = g(x_{i+1}) - g(x_i))$, and hence the quasi-Newton equation gives $B_{i+1}s_i = y_i = As_i$. That is, it makes B_{i+1} "look like" A in the direction s_i. If this information were cumulative - if after k iterations $B_ks_j = As_j$ for $j = 0,1,\ldots,k-1$ - then B_n would equal A if $s_0,\ldots,s_{n-1}$ were linearly independent. However, the DFP and BFGS (and other rank-two updates used before Davidon (1975)) do not have this property, because each update destroys old information as it adds new information. For example, after one iteration $B_1s_0 = y_0 = As_0$; after the next, $B_2s_1 = y_1 = As_1$, but $B_2s_0 \neq y_0$. For instance with the DFP

$$B_2s_0 = B_1s_0 + \left(\frac{y_1-B_1s_1}{\langle y_1,s_1\rangle} - \frac{\langle y_1-B_1s_1,s_1\rangle y_1}{\langle y_1,s_1\rangle^2}\right)\langle y_1,s_0\rangle$$

$$+ \frac{y_1}{\langle y_1,s_1\rangle}\ \langle y_1-B_1s_1,s_0\rangle$$

and while $\langle y_1-B_1s_1,s_0\rangle = \langle y_1,s_0\rangle - \langle s_1,B_1s_0\rangle = \langle As_1,s_0\rangle - \langle s_1,As_0\rangle = 0$, in general $\langle y_1,s_0\rangle \neq 0$, and $B_2s_0 \neq B_1s_0(=y_0)$.

What we would like is an algorithm which at each step retains symmetry (we'll concern ourselves with positive definiteness in Chapter 4), yields $B_{i+1}s_i = y_i$, but keeps $B_{i+1}s_j = y_j$ for $j<i$. If we're maintaining these properties at each step, then $B_i s_j$ will equal y_j for $j<i$, and our last requirement becomes $B_{i+1}s_j = B_i s_j$, $j<i$. So we will get $B_n = A$ (assuming $s_0,\ldots,s_{n-1}$ linearly independent) if each B_{i+1} satisfies

$$\{B_{i+1} \text{ symmetric, } B_{i+1}s_i = y_i,\ (B_{i+1}-B_i)z_i = 0 \text{ for all } z_i \in Z_i\},$$

$$Z_i = [s_0,\ldots,s_{i-1}]. \tag{2.4.1}$$

In Chapter 5 we show how to select matrices B_{i+1} which satisfy (2.4.1) and minimize $\|B_{i+1}-B_i\|_{Fr,M}$. For now let us find rank-two updates satisfying (2.4.1) by considering the update

$$B_{i+1} = B_i + \frac{(y_i-B_is_i)v_i^T+v_i(y_i-B_is_i)^T}{\langle v_i,s_i\rangle} - \frac{\langle y_i-B_is_i,s_i\rangle v_iv_i^T}{\langle v_i,s_i\rangle^2}, \tag{2.4.2}$$

a generalization of the DFP which is equal to it for $v_i = y_i$. Equation (2.4.2) retains symmetry and obeys the quasi-Newton equation for any value of v_i such that $\langle v_i,s_i\rangle \neq 0$; and will keep $B_{i+1}s_j = B_is_j$ if $\langle y_i-B_is_i,s_j\rangle = \langle v_i,s_j\rangle = 0$. If f is quadratic, B_i symmetric and $B_is_j = y_j$, then $\langle y_i-B_is_i,s_j\rangle = \langle y_i,s_j\rangle-\langle s_i,B_is_j\rangle = \langle As_i,s_j\rangle-\langle s_i,As_j\rangle = 0$. Therefore if $B_is_j = y_j$ for all $j<i$, update (2.4.2) will satisfy (2.4.1) for quadratic functions if

$$\langle v_i,s_j\rangle = 0, \quad j = 0,1,\ldots,i-1. \tag{2.4.3}$$

By choosing each v_i to satisfy (2.4.3) and $\langle v_i, s_i\rangle \neq 0$, Davidon (1975) obtains an algorithm which is exact on quadratics.

Therefore Davidon (1975) proposes an algorithm of the form

Algorithm 2.4.4

Given x_0, B_0 non-singular

At each iteration

$$x_{i+1} = x_i - B_i^{-1} g(x_i)$$

B_{i+1} given by (2.4.2), $s_i = x_{i+1} - x_i$, $y_i = g(x_{i+1}) - g(x_i)$

v_i chosen to satisfy (2.4.3), $\langle v_i, s_i\rangle \neq 0$, and B_{i+1} non-singular

This choice of v_i is possible as long as $v_i \notin [s_0, \ldots, s_{i-1}]$. If f is a positive definite quadratic function, we show in Chapter 4 that Davidon's (1975) choice of v_i results in each B_i positive definite. Below we show that algorithm 2.4.4 will find the minimum of a quadratic function in n+1 or fewer iterations.

Theorem 2.4.1 (similar to Theorem 2, Davidon (1975)): Let $f: R^n \to R$, $f(x) = \frac{1}{2} x^T Ax - b^T x$, $b \in R^n$, $A \in L(R^n)$ symmetric and positive definite. Consider the sequences $(x_0, \ldots, x_i)$ of points $\in R^n$ and $(B_0, \ldots, B_i)$ of matrices $\in L(R^n)$ generated by Algorithm 2.4.4. Then as long as $s_i \notin [s_0, \ldots, s_{i-1}]$ there exist $v_i \in R^n$ satisfying the conditions of Algorithm 2.4.4; and for the first k such that $s_k \in [s_0, \ldots, s_{k-1}]$, $x_{k+1} = A^{-1}b = x^*$, the minimum of f(x). This $k \leq n+1$, and if n+1 interations are required, then $B_n = A$.

Proof: At each iteration, if $s_i \notin [s_0,\ldots,s_{i-1}]$, it is clear that there exist v_i satisfying $\langle v_i,s_j\rangle = 0$, $j<i$, and $\langle v_i,s_i\rangle \neq 0$. In Chapter 4 we show that at least one such choice maintains non-singularity. The above discussion shows that then $B_{i+1}s_j = y_j = As_j, j=0,\ldots,i$. Now consider the first k such that $s_k \epsilon [s_0,\ldots,s_{k-1}]$. Then $B_k s_k = As_k$, so that $x_{k+1} = x_k+s_k = x_k+A^{-1}B_k s_k = x_k+A^{-1}B_k(-B_k^{-1}g_k) = x_k-A^{-1}(Ax_k-b) = A^{-1}b = x^*$.

Finally, assume that for each m<n, $s_j \notin [s_0,\ldots,s_{j-1}]$. Then $s_0,\ldots,s_{n-1}$ are linearly independent and span R^n. Therefore since $B_n s_j = As_j$ for $j = 0,1,\ldots,n-1$, $B_n = A$. Furthermore, s_n must be $\epsilon[s_0,\ldots,s_{n-1}]$, and so $x_{n+1} = x^*$.

What remains is the actual choice of v_i. Since the value $v_i = y_i$ has been successful, Davidon first suggests (the dual of) $v_i = \hat{y}_i = P_i y_i$, where P_i is a projection matrix projecting orthogonal to $[s_0,\ldots,s_{i-1}]$. He defines

$$P_i = I - \sum_{j=0}^{i-1} \frac{\hat{y}_j\hat{s}_j^T}{\langle\hat{y}_j,\hat{s}_j\rangle} \qquad (2.4.5)$$

where $\hat{y}_j = P_j y_j$, $\hat{s}_j = P_j^T s_j$. Each $\hat{y}_i$ can be thought of as the projection of y_i onto $[y_0,\ldots,y_i]$ and orthogonal to $[s_0,\ldots,s_{i-1}]$; each s_i as the projection of s_i onto $[s_0,\ldots,s_i]$ and orthogonal to $[y_0,\ldots,y_{i-1}]$. Therefore

$$\langle\hat{y}_i,s_j\rangle = \langle\hat{y}_i,\hat{s}_j\rangle = 0,\ j = 0,1,\ldots,i-1 \qquad (2.4.6a)$$

$$\langle\hat{s}_i,y_j\rangle = \langle\hat{s}_i,\hat{y}_j\rangle = 0,\ j = 0,1,\ldots,i-1 \qquad (2.4.6b)$$

$$\langle\hat{s}_i,\hat{y}_i\rangle = \langle\hat{s}_i,y_i\rangle = \langle s_i,\hat{y}_i\rangle \qquad (2.4.6c)$$

If $v_i = \hat{y}_i$ is used in Algorithm 2.4.4, then for quadratic problems

$B_{i+1}s_j = y_j$, $j=0,\ldots,i$ due to (2.4.6a), and $P_i(y_i-B_is_i) = y_i-B_is_i$ from (2.4.5), since each $\hat{s}_j \epsilon [s_0,\ldots,s_j]$ and $\langle y_i-B_is_i,s_j\rangle = 0$ for all $j<i$. Using this we can show by induction that for each i

$$P_iB_i = B_iP_i^T \tag{2.4.7a}$$

$$y_i-B_is_i = \hat{y}_i-B_i\hat{s}_i \tag{2.4.7b}$$

$$B_{i+1}\hat{s}_j = \hat{y}_j,\ j=0,\ldots,i \tag{2.4.7c}$$

Therefore for quadratic problems, an equivalent way to write P_i is

$$P_i' = I - \sum_{j=0}^{i-1} \frac{B_i\hat{s}_j\hat{s}_j^T}{\langle\hat{s}_j,B_i\hat{s}_j\rangle} \tag{2.4.8}$$

or

$$P_i'' = I - \sum_{j=0}^{i-1} \frac{\hat{y}_j\hat{y}_jB_i^{-1}}{\langle\hat{y}_j,B_i^{-1}\hat{y}_j\rangle} \tag{2.4.9}$$

Thus $v_i = \hat{y}_i$ is the projection of y_i orthogonal to $[y_0,\ldots,y_{i-1}]$ in the metric B_i^{-1}; $\hat{s}_i$ is the projection of s_i orthogonal to $[s_0,\ldots,s_{i-1}]$ in the metric B_i.

If $\langle v_i,s_j\rangle = \langle y_i-B_is_i,s_j\rangle = 0$ for $j=0,\ldots,i-1$ and f is quadractic then the vector

$$v_{i+1} = v_i\langle y_i-B_is_i,s_i\rangle - (y_i-B_is_i)\langle v_i,s_i\rangle \tag{2.4.10}$$

is orthogonal to s_j, $j=0,\ldots,i$. Although for quadratic f, v_1 defined by (2.4.10) may not $= P_1y_1$ even if $v_0=y_0$, it still leads to n+1 step termination in Algorithm 2.4.4; and in the dual formulation discussed below for updates to B_i^{-1}, the two strategies are identical for quadratics. Since (2.4.10) is a simpler computation than $v_i = P_i y_i$ using (2.4.5), (2.4.8) or (2.4.9), Davidon (1975) actually uses (the dual of) this formula in his algorithm for quadratics. Powell (1976b) shows that the choice of v_0 does not affect the finite termination property; but it does lead to different intermediate v_i's, B_i's, and x_i's.

Since $y_i - B_i s_i = \hat{y}_i - B_i \hat{s}_i$, $\langle y_i, s_i \rangle = \langle \hat{y}_i, \hat{s}_i \rangle$, and $\langle B_i s_i, s_i \rangle = \langle B_i \hat{s}_i, \hat{s}_i \rangle$ from (2.4.6-7), update (2.4.2) with $v_i = \hat{y}_i$, can be rewritten

$$B_{i+1} = B_i + \frac{(\hat{y}_i - B_i \hat{s}_i)\hat{y}_i^T + \hat{y}_i(\hat{y}_i - B_i \hat{s}_i)^T}{\langle \hat{y}_i, \hat{s}_i \rangle} - \frac{\langle \hat{y}_i - B_i \hat{s}_i, \hat{s}_i \rangle \hat{y}_i \hat{y}_i^T}{\langle \hat{y}_i, \hat{s}_i \rangle^2} \tag{2.4.11}$$

This is the form in which Davidon discusses his update and which was alluded to in Chapter 1. However, Davidon calculates v_i by (2.4.10), and $\hat{y}_i, \hat{s}_i$ by $\hat{y}_i = \overline{P}_i y_i$, $\hat{s}_i = \overline{P}_i^T s_i$, where $\overline{P}_i$ projects onto $[v_i, y_i - B_i s_i]$ in the B_i^{-1} metric, rather than using $\hat{y}_i = P_i y_i$, $\hat{s}_i = P_i^T s_i$ as discussed above. In the dual case, this is another way to obtain the same update for quadratics. Although even in the quadratic case matrix $\overline{P}_i$ is not equal to P_i, for quadratic or non-quadratic functions it will obey $\overline{P}_i B_i = B_i \overline{P}_i^T$, $\overline{P}_i(y_i - B_i s_i) = y_i - B_i s_i$, so that (2.4.2) and (2.4.11) are still equivalent. So $\overline{P}_i$ has many properties of P_i, and is simpler to calculate.

Either formula (2.4.2) or (2.4.11) can be made part of a one-parameter class

$$B_{i+1}(\phi) = B_i + \frac{(y_i-B_is_i)v_i^T+v_i(y_i-B_is_i)^T}{\langle v_i,s_i\rangle} - \frac{\langle y_i-B_is_i,s_i\rangle v_iv_i^T}{\langle v_i,s_i\rangle^2}$$

$$+ \phi\left[\frac{y_i-B_is_i}{\langle y_i-B_is_i,s_i\rangle} - \frac{v_i}{\langle v_i,s_i\rangle}\right]\left[\frac{y_i-B_is_i}{\langle y_i-B_is_i,s_i\rangle} - \frac{v_i}{\langle v_i,s_i\rangle}\right]^T \qquad (2.4.12)$$

or

$$B_{i+1}(\hat{\phi}) = B_i + \frac{(\hat{y}_i-B_i\hat{s}_i)\hat{y}_i^T+\hat{y}_i(\hat{y}_i-B_i\hat{s}_i)^T}{\langle \hat{y}_i,\hat{s}_i\rangle} - \frac{\langle \hat{y}_i-B_i\hat{s}_i,\hat{s}_i\rangle \hat{y}_i\hat{y}_i^T}{\langle \hat{y}_i,\hat{s}_i\rangle^2}$$

$$- \hat{\phi}\langle \hat{s}_i,B_i\hat{s}_i\rangle\left[\frac{\hat{y}_i}{\langle \hat{y}_i,\hat{s}_i\rangle} - \frac{B_i\hat{s}_i}{\langle \hat{s}_i,B_i\hat{s}_i\rangle}\right]\left[\frac{\hat{y}_i}{\langle \hat{y}_i,\hat{s}_i\rangle} - \frac{B_i\hat{s}_i}{\langle \hat{s}_i,B_i\hat{s}_i\rangle}\right]^T \qquad (2.4.13)$$

respectively, $\phi,\hat{\phi}\in R$. Theorem 2.4.1 is still true if any update from (2.4.12), or (2.4.13) with $\hat{s}_i,\hat{y}_i$ defined above, is used in Algorithm 2.4.4. Davidon (1975) in fact uses some update from (2.4.13), rather than (2.4.11), in his algorithm.

In the case when f is not quadratic, there no longer exists a matrix A such that $y_i = As_i$ for all i. However we might still want to preserve recent quasi-Newton equations -- perhaps $B_{i+1}s_j = y_j$, $j = i-\ell,\ldots,i$ for some positive integer $\ell<n$. So we would like B_{i+1} to satisfy (2.4.1), where $Z_i = [s_{i-\ell},\ldots,s_{i-1}]$. Lemma 2.4.2 shows that this is not usually possible.

<u>Lemma 2.4.2</u>: Let $B\in L(R^n)$ be symmetric; $s,y\in R^n$ such that $s\neq 0$, $y\neq Bs$; Z an m dimensional subspace of R^n, $0\leq m<n$. Define $Z^\perp = \{u\in R^n | \langle u,z\rangle = 0$ for all $z\in Z\}$. Then the set $S = \{\bar{B}\in L(R^n) | \bar{B}$ symmetric, $\bar{B}s = y$, $(\bar{B}-B)z = 0$ for all

$z \in Z\}$ is nonempty if and only if $s \notin Z$ and $(y-Bs) \in Z^{\perp}$.

Proof: First assume S is nonempty. Then for any $\bar{B} \in S$ and $z \in Z$, $(\bar{B}-B)z = 0$. Since $(\bar{B}-B)s = y-Bs \neq 0$, $s \notin Z$. Furthermore, since $s^T(\bar{B}-B)z = 0$ for all $z \notin Z$ and $(\bar{B}-B)$ is symmetric, $(y-Bs)^T z = 0$ for all $z \in Z$. Thus $(y-Bs) \in Z^{\perp}$.

Now assume $s \notin Z$ and $(y-Bs) \in Z^{\perp}$. If $\langle y-Bs,s\rangle \neq 0$, then the symmetric rank one update $\bar{B} = B + \frac{(y-Bs)(y-Bs)^T}{\langle y-Bs,s\rangle}$ is in S. If $\langle y-Bs,s\rangle = 0$, then since since $s \notin Z$, there exists $v \in Z^{\perp}$ such that $\langle v,s\rangle \neq 0$. Then $\bar{B} = B + \frac{(y-Bs)v^T + v(y-Bs)^T}{\langle v,s\rangle}$ is in S.

The condition $s_i \notin Z_i$ will be satisfied as long as we select ℓ such that s_i is independent of $s_{i-\ell},\ldots,s_{i-1}$. However for non-quadratic f, $(y_i-B_i s_i)$ will not in general be orthogonal to Z_i $(=[s_{i-\ell},\ldots,s_{i-1}])$ for any $\ell>0$. Therefore in general there is no update which retains symmetry, satisfies the quasi-Newton equation, and preserves any past quasi-Newton equations.

However, the quadratic algorithms using the updates described above can still be applied to non-quadratic problems. It is hoped that these updating formulas will still produce good Hessian approximations. Davidon (1975) chooses to use equation (2.4.12), with v_i defined by (2.4.10). This is computationally convenient. Occasionally he must reset $v_i = y_i$. In Chapter 3 we show that (2.4.12) can be written in the equivalent form (2.4.13), where $\hat{y}_i = \bar{P}_i y_i$, $\hat{s}_i = \bar{P}_i^T s_i$, $\bar{P}_i$ the projection matrix onto $[v_i, y_i-B_i s_i]$ in the metric B_i^{-1}, $\hat{\phi}$ a function of s_i, y_i, v_i and ϕ. Davidon (1975) always discusses his general method in this form, but we believe that the fundamental ideas involved are clearer from (2.4.12).

Another way to adapt the quadratic algorithm to non-quadratic problems would be to use equation (2.4.12), with $v_i = P_i y_i$, P_i given by (2.4.5) except with the summation going from $i-\ell$ to $i-1$ for some $\ell<n$. In this case $\langle v_i,s_j\rangle$ still is 0, $j = i-\ell,\ldots,i-1$. This update requires more housekeeping work but may come closer to preserving information from previous quasi-Newton equations. A third alternative would be to use equation (2.4.13), with $\hat{y}_i = P_i y_i$, $\hat{s}_i = P_i^T s_i$, P_i defined as above. If we do this we get $B_{i+1}\hat{s}_j = \hat{y}_j$, $j = i-\ell,\ldots,i$ but no longer (necessarily) $B_{i+1}s_i = y_i$. All these alternatives are extensions of the same update for quadratics, but they lead to different algorithms for non-quadratic problems. They are discussed more thoroughly in Chapter 9.

Actually, Davidon (1975) updates approximations H_i to the inverse Hessian $G(x_i)^{-1}$. For quadratics he is interested in having $H_{i+1}y_j = s_j$, $j = 0,\ldots,i$, and this leads to the one-parameter class of updates

$$H_{i+1}(\phi) = H_i + \frac{(s_i-H_iy_i)w_i^T+w_i(s_i-H_iy_i)^T}{\langle w_i,y_i\rangle} - \frac{\langle s_i-H_iy_i,y_i\rangle w_iw_i^T}{\langle w_i,y_i\rangle^2}$$

$$+ \phi\left[\frac{s_i-H_iy_i}{\langle s_i-H_iy_i,y_i\rangle} - \frac{w_i}{\langle w_i,y_i\rangle}\right]\left[\frac{s_i-H_iy_i}{\langle s_i-H_iy_i,y_i\rangle} - \frac{w_i}{\langle w_i,y_i\rangle}\right]^T \qquad (2.4.14)$$

$w_i = \hat{s}_i = P_i^T s_i$, P_i given by (2.4.5) with $\hat{s}_j,\hat{y}_j$ defined as before, $\phi\in R$. This update obeys

$$\{H_{i+1}-H_i \text{ symmetric, } H_{i+1}y_i = s_i,\ (H_{i+1}-H_i)z = 0 \text{ for all } z\in Z_H,$$

$$Z_H = [y_0,\ldots,y_{i-1}]\} \qquad (2.4.15)$$

for quadratic functions; and if H_i is substituted for B_i^{-1} in Algorithm 2.4.4, then the algorithm still finds the minimum of a quadratic function in n+1 or fewer steps. Furthermore for quadratics the same w_i can be calculated by $w_0 = s_0$,

$$w_{i+1} = w_i\langle s_i - H_i y_i, y_i\rangle - (s_i - H_i y_i)\langle w_i, y_i\rangle. \tag{2.4.16}$$

For non-quadratics, the algorithm actually proposed in Davidon's (1975) paper uses update (2.4.14), with w_i defined using (2.4.16) and occasionally resetting $w_i = s_i$. Algorithm 2.4.4 is used but a simple line search is employed if the decrease in f(x) produced by the direct prediction step is unsatisfactory. The parameter ϕ in (2.4.14) is chosen by a criterion called optimal conditioning, which seeks to hold down the growth in the condition number of H_i by minimizing a bound on this condition number. The use of optimal conditioning is a second innovation in Davidon's paper, whose effect we have tried to separate from that of his new update. We discuss it extensively in Chapters 7 and B.

Finally, Davidon does not really use formula (2.4.14), but rather makes a rank-one update to a non-symmetric matrix J_i for which $J_i J_i^T = H_i$ to yield a matrix J_{i+1} such that $J_{i+1}J_{i+1}^T = H_{i+1}(\phi)$. In doing this he is helped by considering an equivalent form of formula (2.4.14),

$$H_{i+1}(\phi) = H_i + \frac{(\hat{s}_i - H_i\hat{y}_i)\hat{s}_i^T + \hat{s}_i(\hat{s}_i - H_i\hat{y}_i)^T}{\langle \hat{s}_i, \hat{y}_i\rangle} - \frac{\langle \hat{s}_i - H_i\hat{y}_i, \hat{y}_i\rangle \hat{s}_i\hat{s}_i^T}{\langle \hat{s}_i, \hat{y}_i\rangle^2}$$

$$- \hat{\phi}\langle \hat{y}_i, H_i\hat{y}_i\rangle \left[\frac{\hat{s}_i}{\langle \hat{s}_i, \hat{y}_i\rangle} - \frac{H_i\hat{y}_i}{\langle \hat{y}_i, H_i\hat{y}_i\rangle}\right]\left[\frac{\hat{s}_i}{\langle \hat{s}_i, \hat{y}_i\rangle} - \frac{H_i\hat{y}_i}{\langle \hat{y}_i, H_i\hat{y}_i\rangle}\right]^T \tag{2.4.17}$$

where $\hat{s}_i = Q_i s_i$, $\hat{y}_i = Q_i^T y_i$, Q_i the matrix projecting onto $[w_i, s_i - H_i y_i]$ in the metric H_i^{-1}. This equivalence is discussed in Chapter 3, for it is helpful to later analysis. However, we do not consider the actual implementation of Davidon's update to have much bearing upon the performance of his algorithm, as the update seems to perform just as well when H_i itself is kept and updated in algorithms such as that of Dennis and Mei (1975).

The development of updates to H_i is algebraically equivalent to the updates to B_i with the terms H,w,y,s and Z_H taking the places of B,v,s,y and Z respectively. Therefore our discussion in this section of alternative ways to extend Davidon's quadratic update to non-quadratic problems can also be applied to algorithms which update H_i. Lemma 2.4.2, which shows that in general we can not maintain current and past quasi-Newton equations, has an analagous form for inverse updates, given in Section 3.3. Throughout this thesis we will continue to concentrate on updates to B_i, but will often mention the equivalent results for updates to H_i.

Davidon's (1975) algorithm, as implemented by Davidon and in other places including Argonne National Laboratory, performs somewhat better than previous quasi-Newton methods. It contains two distinct innovations, which seem applicable to other problems. Therefore it has supplied much of the impetus to the work of this thesis.

CHAPTER III

RANK-TWO QUASI-NEWTON UPDATES

Discussion of Davidon's (1975) paper in Section 2.4 has shown us the utility of considering various classes of rank-two quasi-Newton updates. In this chapter we demonstrate two basic equivalences between differently written rank-two quasi-Newton update classes. These equivalences help us understand some properties of our new updates; and aid in our analysis of them in following chapters.

In Section 3.1 we derive a restricted DFP-BFGS class

$$\hat{B}(\hat{\phi}) = B + \frac{(\hat{y}-B\hat{s})\hat{y}^T+\hat{y}(\hat{y}-B\hat{s})^T}{\langle\hat{y},\hat{s}\rangle} - \frac{\langle\hat{y}-B\hat{s},\hat{s}\rangle\hat{y}\hat{y}^T}{\langle\hat{y},\hat{s}\rangle^2}$$

$$- \hat{\phi}\langle\hat{s},B\hat{s}\rangle\left[\frac{\hat{y}}{\langle\hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle\hat{s},B\hat{s}\rangle}\right]\left[\frac{\hat{y}}{\langle\hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle\hat{s},B\hat{s}\rangle}\right]^T \qquad (3.0.1)$$

$\hat{y} = Py$, $\hat{s} = P^Ts$, P a projection matrix in the metric B^{-1} which satisfies our new restricted updating conditions; and show that it is algebraically equivalent to the "v-class"

$$\bar{B}(\phi) = B + \frac{(y-Bs)v^T+v(y-Bs)^T}{\langle v,s\rangle} - \frac{\langle y-Bs,s\rangle vv^T}{\langle v,s\rangle^2}$$

$$+ \phi\left[\frac{y-Bs}{\langle y-Bs,s\rangle} - \frac{v}{\langle v,s\rangle}\right]\left[\frac{y-Bs}{\langle y-Bs,s\rangle} - \frac{v}{\langle v,s\rangle}\right]^T \qquad (3.0.2)$$

for $v\epsilon[\hat{y},B\hat{s}]$. From this equivalence we learn some important properties of our new updates. It is also helpful to our discussion of positive definiteness in Chapter 4 and optimal conditioning in Chapter 7. The following section demonstrates that an important portion of the general one-parameter class (3.0.2) can be represented as

$$\hat{B}(\sigma) = B + \frac{(y-Bs)u^T+u(y-Bs)^T}{\langle u,s\rangle} - \frac{\langle y-Bs,s\rangle uu^T}{\langle u,s\rangle^2}$$

$$u = v+\sigma(y-Bs). \tag{3.0.3}$$

This new parameterization, applicable as well to the standard DFP-BFGS class, is quite helpful to our understanding of quasi-Newton updates. It has uses in implementation of algorithms, and in our analysis of optimal conditioning in Chapter 7. Finally in Section 3.3, we state some of these same types of results for updates to inverse Hessian approximations H. We consider this a useful reference, as some algorithms may wish to use updates in this form.

Section 3.1 - Restricted Updates-A Basic Equivalence

We return in this section to our consideration of quasi-Newton updates to a symmetric matrix B which satisfy

$$\{\bar{B} \text{ symmetric}, \bar{B}s = y, (\bar{B}-B)z = 0 \text{ for all } z\epsilon Z\} \tag{3.1.1}$$

for B, $\bar{B}\epsilon L(R^n)$, s,y nonzero vectors in R^n, Z an m dimensional subspace of R^n, m<n. Lemma 2.4.2 gives the necessary and sufficient conditions for there to exist $\bar{B}$ satisfying (3.1.1). In Theorem 3.1.1 we define and derive a restricted DFP-BFGS class of form (3.0.1) which satisfies (3.1.1) if the conditions of Lemma 2.4.2 are met. This class of updates is important to our discussion of minimum norm restricted updates in Chapter 5.

Theorem 3.1.1 also shows that the restricted DFP-BFGS class is equivalent to a v-class of form (3.0.2). We discuss the importance of this equivalence.

Davidon (1975) in his general algorithm uses updates which instead of (3.1.1) satisfy

$$\{\bar{B} \text{ symmetric}, \ \bar{B}s = y, \ (\bar{B}-B)\epsilon[y-Bs,v]\} \tag{3.1.2}$$

for some $v \epsilon R^n$. (We write $M\epsilon[v_i,\ldots,v_j]$ if the row and column spaces of the symmetric matrix $M\epsilon L(R^n)$ are contained in the subspace spanned by vectors $v_i,\ldots,v_j \epsilon R^n$.) We show that conditions (3.1.2) are a special case of (3.1.1), and in Corollary 3.1.2 show that conditions (3.1.2) are satisfied by a v-class of form (3.0.2), or an equivalent restricted DFP-BFGS class of form (3.0.1). This equivalence helps us understand the properties of updates of form (3.0.2). Also, by showing us how to convert v-class (3.0.2) to the DFP-BFGS form (3.0.1), it enables us to answer questions about v-classes of updates, using techniques developed for the DFP-BFGS class. We use this approach in our examination of positive definiteness of general classes in Chapter 4, and optimal conditioning of general classes in Chapter 7.

<u>Theorem 3.1.1</u>. Let $B\epsilon L(R^n)$ be symmetric and positive definite; $s,y\epsilon R^n$ such that $s\neq 0$, $y\neq Bs$; $\phi, \hat{\phi}$ real parameters. Let Z be an m dimensional subspace of R^n, $0\leq m<n$. Assume $y-Bs\epsilon Z^\perp$ and $s\notin Z$. Let $S = \{\bar{B}\epsilon L(R^n) | \bar{B} \text{ symmetric}, \bar{B}s = y, (\bar{B}-B)z = 0 \text{ for all } z\epsilon Z\}$. Let P be the matrix projecting onto $Z^\perp$ (orthogonal to BZ) in the metric B^{-1}, and define $\hat{y} = Py$, $\hat{s} = P^T s$. Then for any $\bar{B}\epsilon S$, $\bar{B}\hat{s} = \hat{y}$. For any $v\epsilon[\hat{y},B\hat{s}]$ for which $\langle v,s\rangle \neq 0$, the one parameter v-class of updates

$$\bar{B}(\phi,v) = B + \frac{(y-Bs)v^T+v(y-Bs)^T}{\langle v,s\rangle} - \frac{\langle y-Bs,s\rangle vv^T}{\langle v,s\rangle^2}$$

$$+ \phi\left[(y-Bs) - \frac{v\langle y-Bs,s\rangle}{\langle v,s\rangle}\right]\left[(y-Bs) - \frac{v\langle y-Bs,s\rangle}{\langle v,s\rangle}\right]^T \qquad (3.1.3)$$

is in S; and if $\langle\hat{y},\hat{s}\rangle \neq 0$, then for any such v which is also linearly independent of $\hat{y}-B\hat{s}$, (3.1.3) is equivalent to the restricted DFP-BFGS class

$$\hat{B}(\hat{\phi}) = B + \frac{(\hat{y}-B\hat{s})\hat{y}^T+\hat{y}(\hat{y}-B\hat{s})^T}{\langle\hat{y},\hat{s}\rangle} - \frac{\langle\hat{y}-B\hat{s},\hat{s}\rangle\hat{y}\hat{y}^T}{\langle\hat{y},\hat{s}\rangle^2}$$

$$- \hat{\phi}\langle\hat{s},B\hat{s}\rangle\left[\frac{\hat{y}}{\langle\hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle\hat{s},B\hat{s}\rangle}\right]\left[\frac{\hat{y}}{\langle\hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle\hat{s},B\hat{s}\rangle}\right]^T \qquad (3.1.4)$$

<u>Proof</u>: By Lemma 2.4.2, S is nonempty. We first let $\bar{B}$ be any member of S, and show that $\bar{B}s = \hat{y}$. Using the formula (2.3.2a) to define P gives

$$P = I - \sum_{i=1}^{m} \frac{Bz_iz_i^T}{\langle z_i,Bz_i\rangle}$$

where $[z_1,\ldots,z_m]$ is a basis for Z such that $\langle z_i,Bz_j\rangle = 0$, $1\leq i<j\leq m$. Since $(\bar{B}-B)z = 0$ for all $z\in Z$, $P = I - \sum_{i=1}^{m} \frac{\bar{B}z_iz_i^T}{\langle z_i,\bar{B}z_i\rangle}$. Thus $P\bar{B} = \bar{B}P^T$, and since $\bar{B}s = y$, $Py = P\bar{B}s = \bar{B}P^Ts$, or $\hat{y} = \bar{B}\hat{s}$.

We next show that $\bar{B}(\phi,v)\in S$ for any $\phi\in R$ and $v\in[\hat{y},B\hat{s}]$ such that $\langle v,s\rangle \neq 0$. Clearly $\bar{B}(\phi,v)$ is symmetric, and simple algebra shows

$\overline{B}(\phi,v)\cdot s = y$. Finally, since by Lemma 2.3.3, $B\hat{s} = BP^Ts = PBs$, $B\hat{s}$ and $\hat{y}(= Py)$ are in $Z^{\perp}$. Hence $v\in Z^{\perp}$, and since also $y-Bs\in Z^{\perp}$, $(\overline{B}(\phi,v)-B)z = 0$ for all $z\in Z$. Thus $\overline{B}(\phi,v)\in S$.

Next we show that if $\langle\hat{y},\hat{s}\rangle \neq 0$, the restricted DFP-BFGS class (3.1.4) is equivalent to the v-class $\overline{B}(\phi,\hat{y})$. Since $P^2 = P$ by Lemma 2.3.3, $\langle\hat{y},\hat{s}\rangle = \langle Py,P^Ts\rangle = \langle P^2y,s\rangle = \langle Py,s\rangle = \langle\hat{y},s\rangle$. Since $y-Bs\in Z^{\perp}$, $P(y-Bs) = y-Bs$. Also $P(y-Bs) = Py-PBs = Py-BP^Ts = \hat{y}-B\hat{s}$. Hence $y-Bs = \hat{y}-B\hat{s}$, and $\langle\hat{y}-B\hat{s},\hat{s}\rangle = \langle P(y-Bs),P^Ts\rangle = \langle P^2(y-Bs),s\rangle = \langle y-Bs,s\rangle$. These equalities show that $\overline{B}(0,\hat{y}) = \hat{B}(0)$. Secondly, when $v = \hat{y}$ simple algebra shows

$$- \hat{\phi}\langle\hat{s},B\hat{s}\rangle \left[\frac{\hat{y}}{\langle\hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle\hat{s},B\hat{s}\rangle}\right]\left[\frac{\hat{y}}{\langle\hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle\hat{s},B\hat{s}\rangle}\right]^T$$

$$= \phi\left[(y-Bs) - \frac{v\langle y-Bs,s\rangle}{\langle v,s\rangle}\right]\left[(y-Bs) - \frac{v\langle y-Bs,s\rangle}{\langle v,s\rangle}\right]^T$$

for $\phi = -\hat{\phi}/\langle\hat{s},B\hat{s}\rangle$. ($\langle\hat{s},B\hat{s}\rangle$ is nonzero because $\hat{s}\neq 0$ and B is positive definite.) Therefore when $v = \hat{y}$, (3.1.4) is equivalent to (3.1.3), with $\overline{B}(\phi,\hat{y}) = \hat{B}(\hat{\phi})$ for $\phi = -\hat{\phi}/\langle\hat{s},B\hat{s}\rangle$.

Finally, if v in (3.1.3) $= \alpha\hat{y}+\beta B\hat{s}$ for real numbers α, β such that $\langle v,s\rangle \neq 0$, straightforward algebra shows that $\overline{B}(\phi,\hat{y}) = \overline{B}(\overline{\phi},v)$ for

$$\phi = \overline{\phi}\,\frac{(\alpha+\beta)^2b^2}{(\alpha b+\beta c)^2} - \frac{\beta[2\alpha b+\beta(b+c)]}{(\alpha b+\beta c)^2}$$ where $b = \langle\hat{y},\hat{s}\rangle$, $c = \langle\hat{s},B\hat{s}\rangle$. The term $\alpha b+\beta c = \langle v,s\rangle$ is nonzero, as is $(\alpha+\beta)$ because v is linearly independent of $\hat{y}-B\hat{s}$. (If $v = k(\hat{y}-B\hat{s})$ then $\overline{B}(\phi,v)$ is the SR1 update for all $\phi\in R$). Therefore there is a 1-1 and onto mapping from $\overline{B}(\phi,\hat{y})$ to $\overline{B}(\overline{\phi},v)$, and hence v-class (3.1.3) is equivalent to the restricted DFP-BFGS class (3.1.4).

So we see that as long as $\langle\hat{y},\hat{s}\rangle \neq 0$, there will exist a class of updates with the algebraic form of the standard DFP-BFGS class which satisfies the restricted update requirements (3.1.1). These updates satisfy the additional quasi-Newton equation $\overline{B}\hat{s} = \hat{y}$, which we think of as the quasi-Newton equation in the restricted updating space, and which is different from the standard quasi-Newton equation $\overline{B}s = y$ unless $y \varepsilon Z^{\perp}$, in which case $\hat{y} = y$ and $\hat{s} = s$. The restricted DFP-BFGS updates are equivalent to "v-updates" of form (3.1.3), which helps show that we have changed the standard DFP-BFGS class only by changing vector v from its standard value $v = y$.

We have multiplied the vector following ϕ in (3.0.2) by $\langle y-Bs,s\rangle$ in (3.1.3). This avoids having special cases when $\langle y-Bs,s\rangle = 0$. Note also that if $\hat{y}$ and $B\hat{s}$ are linearly dependent, which may happen even if y and Bs are not, that all the updates in (3.1.3) and (3.1.4) reduce to the SR1.

While our analysis in Section 2.4 suggests that Davidon (1975) should use updates which satisfy (3.1.1), he instead requires his updates to satisfy (3.1.2), for some prescribed vector v. This however is just a special case of (3.1.1), where Z is the n-2 dimensional space orthogonal to y-Bs and v, so that $\overline{B}-B$ is restricted to the two dimensional subspace spanned by y-Bs and v. Thus we can use Theorem 3.1.1 to analyze updates satisfying (3.1.2).

<u>Corollary 3.1.2</u> Let $B \varepsilon L(R^n)$ be symmetric and positive definite; $s,y \varepsilon R^n$ such that $s \neq 0$, $y \neq Bs$; $\phi, \hat{\phi}$ real parameters. Let $v \varepsilon R^n$ such that y-Bs and v are linearly independent, and $\langle v,s\rangle \neq 0$. Let $s = \{\overline{B} \varepsilon L(R^n) | \overline{B}$ symmetric, $\overline{B}s = y$, $(\overline{B}-B) \varepsilon [y-Bs,v]\}$. Let P be the matrix projecting onto $[y-Bs,v]$ in the metric B^{-1}; define $\hat{y} = Py$, $\hat{s} = P^T s$. Then for any $\overline{B} \varepsilon S$, $\overline{B}\hat{s} = \hat{y}$, and

each member of the one-parameter v-class (3.1.3) is in S. If $\langle\hat{y},\hat{s}\rangle \neq 0$, then each member of the restricted DFP-BFGS class (3.1.4) is in S; and if $\hat{y}$ and $B\hat{s}$ are linearly independent, then (3.1.3) and (3.1.4) are equivalent, with $\hat{B}(\hat{\phi}) = \overline{B}(\phi,v)$ for

$$\hat{\phi} = -\phi\frac{(\alpha+\beta)^2b^2c}{(\alpha b+\beta c)^2} + \frac{\beta c[2\alpha b+\beta(b+c)]}{(\alpha b+\beta c)^2} \tag{3.1.5}$$

where α, β are constants such that $v = \alpha\hat{y}+\beta B\hat{s}$; $b = \langle\hat{y},\hat{s}\rangle$, $c = \langle\hat{s},B\hat{s}\rangle$.

Proof: We apply Theorem 3.1.1 with $Z^{\perp} = [y-Bs,v]$. Its conditions are satisfied since $y-Bs\in Z^{\perp}$, and $s\notin Z$ due to $\langle v,s\rangle \neq 0$. Thus any $\overline{B}\in S$ satisfies $\overline{B}\hat{s} = \hat{y}$, and if $\langle\hat{y},\hat{s}\rangle \neq 0$ each $B(\hat{\phi})$ given by (3.1.4) is in S. Clearly $\overline{B}(\phi,v)$ given by (3.1.3) is also in S for all $\phi\in R$. If $\hat{y}$ and $B\hat{s}$ are linearly independent, then since both are in $[y-Bs,v]$, they form a basis for this subspace and there exist α, β such that $v = \alpha\hat{y}+\beta B\hat{s}$. (The constants α and β can be determined algebraically, but are not of interest.) Theorem 3.1.1 then shows that the classes $\hat{B}(\hat{\phi})$ and $\overline{B}(\phi,v)$ are equivalent; and by straightforward algebra, using again $\langle v,s\rangle = \langle v,\hat{s}\rangle$ and $\langle y-Bs,s\rangle = \langle\hat{y}-B\hat{s},\hat{s}\rangle$, we find that $\hat{B}(\hat{\phi}) = \overline{B}(\phi,v)$ when $\hat{\phi}$, ϕ are related by (3.1.5).

Davidon emphasizes that his updates satisfy two quasi-Newton equations, $\overline{B}s = y$ and $\overline{B}\hat{s} = \hat{y}$. Corollary 3.1.2 shows that any v-update (3.1.3) will have this property. The two equations are different unless $v\in[y,Bs]$, in which case $\hat{s} = s$ and $\hat{y} = y$. The corollary also shows that we can convert almost any v-update (3.1.3) into the restricted DFP-BFGS from (3.1.4), with which we have much experience. This will be of great help to us in determining in Chapter 4 the necessary and sufficient conditions for there to exist updates which satisfy (3.1.1) or (3.1.2) and which in addition give $\bar{B}$ positive definite if B is.

Section 3.2 - A New Parameterization

In this section we show that a significant portion of the one-parameter classes (3.0.1) or (3.0.2) can be rewritten in the "u-form" (3.0.3). The updates which can be written this way are exactly those for which the rank-two matrix $\bar{B}$-B is the difference ww^T-zz^T of two rank-one matrices. Such updates have the fairly attractive property that at most one eigenvalue of $\bar{B}$ is smaller than the smallest eigenvalue of B, and at most one eigenvalue of $\bar{B}$ is larger than the largest eigenvalue of B. The DFP, BFGS, and all updates in the convex class are of this form; and we will see in Chapter 7 that it is from precisely such updates that Davidon (1975) chooses his "optimally conditioned" update. Therefore a parameterization of only this portion of (3.0.1) or (3.0.2) is important. The new u-form (3.0.3) is also more suitable to some existing analysis techniques, such as the convergence analysis of Broyden, Dennis, and Moré (1973), or the symmetrization development of Powell (1970b) and Dennis (1972).

Theorem 3.2.1 shows exactly which members of the general v-class (3.0.2) are of form ww^T-zz^T, and derives the new u-form class $\hat{B}(\sigma)$ which has two values of σ corresponding to each such $\bar{B}(\phi)$; and a slightly different form $\hat{B}(\omega)$ which corresponds 1-1 with this subset of $\bar{B}(\phi)$. Corollary 3.2.2 then applies this analysis to the DFP-BFGS class. This is the application which we use in Chapter 7.

<u>Theorem 3.2.1</u> Let $B\varepsilon R^n$ be symmetric; $s,y\varepsilon R^n$ such that $\langle v,s\rangle \neq 0$, $y \neq Bs$, $v \neq y-Bs$; ϕ, σ real parameters. Consider the one-parameter classes of up-

dates

$$\bar{B}(\phi) = \frac{(y-Bs)v^T+v(y-Bs)^T}{<v,s>} - \frac{<y-Bs,s>vv^T}{<v,s>^2}$$

$$+ \phi\left[(y-Bs) - \frac{v<y-Bs,s>}{<v,s>}\right]\left[(y-Bs) - \frac{v<y-Bs,s>}{<v,s>}\right]^T \qquad (3.2.1)$$

and

$$\hat{B}(\sigma) = \frac{(y-Bs)u^T+u(y-Bs)^T}{<u,s>} - \frac{<y-Bs,s>uu^T}{<u,s>^2}, \; u = v+\sigma(y-Bs). \qquad (3.2.2)$$

Let $\alpha = <y-Bs,s>$. If $\alpha = 0$, then there is a 1-1 and onto mapping from (3.2.1) to (3.2.2), with $\bar{B}(\phi) \doteq \hat{B}(\sigma)$ for $\sigma = \phi<v,s>/2$. If $\alpha \neq 0$, let $\bar{\phi} = \phi/\alpha^2$, and let S_1 be the set of $\bar{\phi}$ for which $\bar{B}(\bar{\phi})$ has the form $\bar{B}(\bar{\phi}) = B+ww^T-zz^T$, w,z nonzero vectors in R^n. Then $S_1 = \{\bar{\phi}|\alpha(\alpha-\bar{\phi})>0\}$ and there exists a 1-2 and onto mapping from $\bar{B}(\bar{\phi})$, $\bar{\phi}\varepsilon S_1$ to $\hat{B}(\sigma)$, $\sigma\varepsilon R\sim\{-<v,s>/\alpha\}$, with $\bar{B}(\bar{\phi}) = \hat{B}(\sigma)$ for $\sigma = -<v,s>\left[\frac{1}{\alpha} \pm \frac{1}{(\alpha(\alpha-\bar{\phi}))^{1/2}}\right]$. If (3.2.2) is reparameterized as $\hat{B}(\omega)$, $\sigma = -<v,s>[1\pm\omega^{1/2}]/\alpha$, then there exists a 1-1 and onto mapping from $\bar{B}(\bar{\phi})$, $\bar{\phi}\varepsilon S_1$ to $\hat{B}(\omega)$, $\omega\varepsilon(0,\infty)$ with $\bar{B}(\bar{\phi}) = \hat{B}(\omega)$ for $\omega = \alpha/(\alpha-\bar{\phi})$.

<u>Proof</u>: If $<y-Bs,s> = 0$ then $\bar{B}(\phi) = \frac{(y-Bs)v^T+v(y-Bs)^T}{<v,s>} + \phi(y-Bs)(y-Bs)^T$

$$= \frac{(y-Bs)(v+[\phi<v,s>/2](y-Bs))^T+(v+[\phi<v,s>/2](y-Bs))(y-Bs)^T}{<v,s>}$$

$$= \hat{B}(\phi<v,s>/2).$$

If $<y-Bs,s> \neq 0$, then (3.2.1) can be rewritten

$$\bar{B}(\bar{\phi}) = B + \frac{(y-Bs)(y-Bs)^T}{\alpha} - (\alpha-\bar{\phi})\left[\frac{y-Bs}{\alpha} - \frac{v}{\langle v,s\rangle}\right]\left[\frac{y-Bs}{\alpha} - \frac{v}{\langle v,s\rangle}\right]^T, \tag{3.2.3}$$

$\alpha = \langle y-Bs,s\rangle$, $\bar{\phi} = \phi/\alpha^2$. Brodlie, Gourlay, and Greenstadt (1973) have shown that for any member of (3.2.1), $\bar{B}(\bar{\phi})-B$ is expressible in exactly one of the forms ww^T, $-ww^T$, ww^T+zz^T, ww^T-zz^T, $-ww^T-zz^T$, w,z nonzero vectors in R^n. Therefore we see that $\bar{B}(\bar{\phi})-B$ has the form ww^T-zz^T if and only if α and $\alpha-\bar{\phi}$ have the same sign; that is, if and only if $\alpha(\alpha-\bar{\phi}) > 0$.

If $\sigma \neq -\langle v,s\rangle/\alpha$, then for u defined in (3.2.2), $\langle u,s\rangle \neq 0$ and (3.2.2) can be rewritten

$$\hat{B}(\sigma) = B + \frac{(y-Bs)(y-Bs)^T}{\alpha} - \alpha\left[\frac{y-Bs}{\langle y-Bs,s\rangle} - \frac{u}{\langle u,s\rangle}\right]\left[\frac{y-Bs}{\langle y-Bs,s\rangle} - \frac{u}{\langle u,s\rangle}\right]^T$$

$$= B + \frac{(y-Bs)(y-Bs)^T}{\alpha} - \frac{\alpha\langle v,s\rangle^2}{[\langle v,s\rangle+\alpha\sigma]^2}$$

$$\left[\frac{y-Bs}{\langle y-Bs,s\rangle} - \frac{v}{\langle v,s\rangle}\right]\left[\frac{y-Bs}{\langle y-Bs,s\rangle} - \frac{v}{\langle v,s\rangle}\right]^T \tag{3.2.4}$$

Therefore for all $\sigma\epsilon R\sim\{-\langle v,s\rangle/\alpha\}$, $\hat{B}(\sigma)-B$ is of the form ww^T-zz^T, w,z nonzero. Thus $\bar{B}(\bar{\phi})$ can be equivalent to some $\hat{B}(\sigma)$ only if $\bar{\phi}\epsilon S_1$. In this case, we see by matching the coefficients of the last terms of (3.2.3) and (3.2.4) that

$\bar{B}(\bar{\phi}) = \hat{B}(\sigma)$ when $\alpha-\bar{\phi} = \dfrac{\alpha\langle v,s\rangle^2}{[\langle v,s\rangle+\alpha\sigma]^2}$, or $\sigma = -\langle v,s\rangle\left[\dfrac{1}{\alpha} \pm \dfrac{1}{(\alpha(\alpha-\bar{\phi}))^{1/2}}\right]$.

If $\sigma = -\langle v,s\rangle[1 \pm \omega^{1/2}]/\alpha$ (a 1-2 and onto map from $\omega\epsilon(0,\infty)$ to $\sigma\epsilon R\sim\{-\langle v,s\rangle/\alpha\}$), then

$$\hat{B}(\omega) = B + \frac{(y-Bs)(y-Bs)}{\alpha} - \frac{\alpha}{\omega}\left[\frac{y-Bs}{\alpha} - \frac{v}{\langle v,s\rangle}\right]\left[\frac{y-Bs}{\alpha} - \frac{v}{\langle v,s\rangle}\right]^T$$

$\omega\varepsilon(0,\infty)$, and $\hat{B}(\omega) = \overline{B}(\bar{\phi})$ for $\bar{\phi}\varepsilon S_1$, $\omega\varepsilon(0,\infty)$, and $\alpha-\bar{\phi} = \alpha/\omega$; $\omega = \alpha/(\alpha-\bar{\phi})$.

One application of Theorem 3.2.1 is in the computer implementation of quasi-Newton updates. Gay (1976) discusses this problem, and applying his suggestions is easier if updates are in u-form (3.2.2). This form also makes it clear that we are using our free parameter to vary the vector u within the subspace spanned by v and y-Bs.

When we discuss optimal conditioning in Chapter 7, we will only be concerned with update classes of the DFP-BFGS form (3.1.4), because we have shown in Section 3.1 that the other update classes we are interested in are algebraically equivalent to this one. Furthermore, we will be interested in exactly those updates of form ww^T-zz^T. Therefore in Corollary 3.2.2 we derive the new parameterization $B(\omega)$ for these members of the DFP-BFGS class.

<u>Corollary 3.2.2</u> Let $B\varepsilon L(R^n)$ be symmetric; $s,y\varepsilon R^n$ such that $\langle y,s\rangle$, $\langle s,Bs\rangle$, $\langle y-Bs,s\rangle \neq 0$; ϕ,σ,ω real parameters. Consider the DFP-BFGS class

$$\overline{B}(\phi) = B + \frac{(y-Bs)y^T+y(y-Bs)^T}{\langle y,s\rangle} - \frac{\langle y-Bs,s\rangle yy^T}{\langle y,s\rangle^2}$$

$$- \phi\langle s,Bs\rangle\left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]\left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]^T \qquad (3.2.5)$$

and the one parameter class

$$\hat{B}(\omega) = B + \frac{(y-Bs)u^T+u(y-Bs)^T}{\langle u,s\rangle} - \frac{\langle y-Bs,s\rangle uu^T}{\langle u,s\rangle^2} ,$$

$$u = y+\sigma(y-Bs), \ \sigma = \frac{-\langle y,s\rangle[1\pm\omega^{1/2}]}{\langle y-Bs,s\rangle} \qquad (3.2.6)$$

Let S_2 be the set of ϕ such that $\bar{B}(\phi)-B$ is of the form ww^T-zz^T, w,z nonzero vectors in R^n. Then $S_2 = \{\phi | 1 + \phi\frac{\langle y-Bs,s\rangle}{\langle s,Bs\rangle} > 0\}$, and there exists a 1-1 and onto mapping from $\bar{B}(\phi)$, $\phi\epsilon S_2$ to $\hat{B}(\omega)$, $\omega\epsilon(0,\infty)$ with $\bar{B}(\phi) = \hat{B}(\omega)$ for $\omega = \frac{\langle s,Bs\rangle}{\langle s,Bs\rangle+\phi\langle y-Bs,s\rangle}$.

In most applications, B is positive definite, so that $\langle s,Bs\rangle > 0$; and the iteration is performed so that $\langle y,s\rangle > 0$, as this is a necessary condition for there to exist positive definite $\bar{B}$ satisfying $\bar{B}s = y$. In this case the set S_2 in Corollary 3.2.2 contains $\phi\epsilon[0,1]$, so that the DFP, BFGS, and the entire convex class have the form ww^T-zz^T. Note that the SR1 update, $\phi = \frac{-\langle s,Bs\rangle}{\langle y-Bs,s\rangle}$ in (3.2.5), $u = y-Bs$ in (3.2.6) is the boundary of the set S_2, and also the limit of $\hat{B}(\sigma)$ as $\sigma\to\pm\infty$, or $\hat{B}(\omega)$ as $\omega\to\infty$.

Corollary 3.2.2 shows that each update $\bar{B}(\phi)$ in (3.2.5) with $\phi\epsilon S_2$ can be written in u-form (3.2.6) for two different values of u. For example, the DFP can have $u = y$ or $y - \frac{2b}{b-c}(y-Bs)$, where $b = \langle y,s\rangle$, $c = \langle s,Bs\rangle$; the BFGS, $u = y + \frac{-b\pm\sqrt{bc}}{b-c}(y-Bs)$. However the analysis of Dennis and Moré (1974) shows that we can expect the values $\frac{b-c}{b}$ at successive iterations to approach zero; so that asymptotically only one of each of these vectors u is numerically stable. For the DFP it is $u = y$; for the BFGS, $u = y + \frac{-b+\sqrt{bc}}{b-c}(y-Bs) =$

$y - \frac{\sqrt{b}}{\sqrt{b}+\sqrt{c}}(y-Bs)$, which approaches $\frac{1}{2}(y+Bs)$ as $\frac{b-c}{b} \to 0$. This shows the interesting fact that the BFGS is in some sense the average of the DFP choice, $u = y$, and the Greenstadt update, $u = Bs$ ($\sigma = -1$ in (3.2.6)).

Section 3.3 - Inverse Updates

We conclude this chapter by discussing restricted quasi-Newton updates to matrices H which approximate the inverse of the Hessian. Algebraically, our main theorems are analagous to the equivalences of Section 3.1 and 3.2: we replace B,v,P,S and Z by H,w,Q,S_H and Z_H, and exchange the roles of s and y. This is an extension of the duality between the B and H forms of the standard DFP-BFGS classes which was first noted by Fletcher (1970). These results are useful because they are likely to be of interest to some algorithms in this inverse update form.

We first give the lemma, analagous to Lemma 2.4.2, describing the necessary and sufficient conditions for there to exist an update $\bar{H}$ to a symmetric matrix H satisfying the restricted inverse updating conditions

$$\{\bar{H} \text{ symmetric}, \bar{H}y = s, (\bar{H}-H)z = 0 \text{ for all } z \in Z_H\} \tag{3.3.1}$$

where s,y are vectors in R^n, y nonzero, $H,\bar{H}$ matrices in $L(R^n)$, Z_H is an m dimensional subspace of R^n, $m<n$. In Theorem 3.3.2 we derive the restricted inverse DFP-BFGS class for updates satisfying (3.3.1), and show that it is equivalent to a w-class of updates (see equation 3.3.2). This theorem is the dual to Theorem 3.1.1. We then show, in Lemma 3.3.3, that if $H = B^{-1}$, $Z_H = BZ$, and $\hat{s},\hat{y} \neq 0$, that the restricted updating conditions (3.1.1) and the restricted inverse updating conditions (3.3.1) are equivalent. This equivalence will be of help in discussing the relation between restricted

and inverse restricted updates in future chapters. Finally in Theorem 3.3.4, we show that a significant portion of the general one-parameter class of rank-two updates to H can be reparameterized in a u-form in a manner analagous to that of Section 3.2. This is the dual of Theorem 3.2.1. We do not state the dual versions of Corollaries 3.1.2 and 3.2.2 - their form should be clear from the results of this section.

Lemma 3.3.1 Let $H \in L(R^n)$ be symmetric; $s, y \in R^n$ such that $y \neq 0$, $s \neq Hy$; Z_H an m dimensional subspace of R^n, $0 \leq m < n$. Then the set $S_H = \{\bar{H} \in L(R^n) |$ $\bar{H}$ symmetric, $\bar{H}y = s$, $(\bar{H}-H)z = 0$ for all $z \in Z_H\}$ is nonempty if and only if $y \notin Z_H$ and $(s-Hy) \in Z_H^{\perp}$.

Note that in Davidon's (1975) algorithm for quadratics, $(Z_H)_i = [y_0,\ldots,y_{i-1}]$, $y_j = g(x_{j+1}) - g(x_j)$. Therefore since $Z_i = [s_0,\ldots,s_{i-1}]$, $s_j = x_{j+1}-x_j$ and $B_i s_j = y_j$, $j = 0,\ldots,i-1$, $(Z_H)_i = BZ_i$. For non-quadratic problems, there will probably be no $\ell < i$ such that H_{i+1} exists satisfying (3.3.1) with $(Z_H)_i = [y_\ell, y_{\ell+1},\ldots,y_{i-1}]$, because $(s_i-H_i y_i)$ will not be contained in $(Z_H)_i^{\perp}$. This situation is analagous to the one we encountered in Section 2.4 in extending Davidon's quadratic algorithm to non-quadratic problems.

Theorem 3.3.2 Let $H \in L(R^n)$ be symmetric and positive definite; $s, y \in R^n$ such that $y \neq 0$, $s \neq Hy$; $\phi, \hat{\phi}$ real parameters. Let Z_H be an m dimensional subspace of R^n, $0 \leq m < n$. Assume $s-Hy \in Z_H^{\perp}$ and $y \notin Z_H$. Let $S_H = \{\bar{H} \in L(R^n) | \bar{H}$ symmetric, $\bar{H}y = s$, $(\bar{H}-H)z = 0$ for all $z \in Z_H\}$. Let Q be the matrix projecting onto $Z_H^{\perp}$ (orthogonal to HZ_H) in the metric H^{-1}; define $\hat{s} = Qs$, $\hat{y} = Q^T y$. Then for any $\bar{H} \in S_H$, $\bar{H}\hat{y} = \hat{s}$. For any $w \in [\hat{s}, H\hat{y}]$ for which $\langle w,y \rangle \neq 0$, the one parameter w-class of updates

$$\bar{H}(\phi,w) = H + \frac{(s-Hy)w^T + w(s-Hy)^T}{\langle w,y\rangle} - \frac{\langle s-Hy,y\rangle ww^T}{\langle w,y\rangle^2}$$

$$+ \phi\left[(s-Hy) - \frac{w\langle s-Hy,y\rangle}{\langle w,y\rangle}\right]\left[(s-Hy) - \frac{w\langle s-Hy,y\rangle}{\langle w,y\rangle}\right]^T \qquad (3.3.2)$$

is in S_H, and if $\langle\hat{s},\hat{y}\rangle \neq 0$, then for any such w which is also linearly independent of $\hat{s}-H\hat{y}$, (3.3.2) is equivalent to the restricted DFP-BFGS class

$$\hat{H}(\phi) = H + \frac{(\hat{s}-H\hat{y})\hat{s}^T + \hat{s}(\hat{s}-H\hat{y})^T}{\langle\hat{s},\hat{y}\rangle} - \frac{\langle\hat{s}-H\hat{y},\hat{y}\rangle\hat{s}\hat{s}^T}{\langle\hat{s},\hat{y}\rangle^2}$$

$$-\hat{\phi}\langle\hat{y},H\hat{y}\rangle\left[\frac{\hat{s}}{\langle\hat{s},\hat{y}\rangle} - \frac{H\hat{y}}{\langle\hat{y},H\hat{y}\rangle}\right]\left[\frac{\hat{s}}{\langle\hat{s},\hat{y}\rangle} - \frac{H\hat{y}}{\langle\hat{y},H\hat{y}\rangle}\right]^T \qquad (3.3.3)$$

From the formulas for projection matrices in Section 2.3, we see that if the subspaces Z,Z_H and the matrices B,H of (3.1.1) and (3.3.1) are related by $H = B^{-1}$, $Z_H = BZ$, then the matrices P and Q of Theorems 3.1.1 and 3.3.2 are related by $Q = P^T$. In this case, if s and y are identical in the two theorems, so are $\hat{s}$ and $\hat{y}$. Then the restricted DFP-BFGS class satisfies (3.1.1) if and only if the restricted inverse DFP-BFGS class satisfies (3.3.1), because $y-Bs\in Z^{\perp}$ if and only if $s-Hy\in Z_H^{\perp}$, and $\langle\hat{s},\hat{y}\rangle \neq 0$ implies $\hat{s},\hat{y} \neq 0$, which imply $s\notin Z$, $y\notin Z_H$. Also, the updates produced by the restricted inverse DFP-BFGS formula (3.3.3) are the inverses of the updates produced by the restricted DFP-BFGS formula (3.1.4), because these two formulas are algebraically equivalent to the B and H forms of the standard DFP-BFGS class, which are known to have this relationship. For instance, the inverse of the restricted DFP $\hat{B}(0)$ is $\hat{H}(1)$; the inverse of the restricted BFGS $\hat{B}(1)$ is $\hat{H}(0)$; and in general $\hat{H}(\hat{\phi}) = \hat{B}(\phi)^{-1}$ for

$$\hat{\phi} = \frac{1-\phi}{1+\phi\left[\frac{\langle\hat{s},B\hat{s}\rangle\langle\hat{y},B^{-1}\hat{y}\rangle}{\langle\hat{s},\hat{y}\rangle^2} - 1\right]}$$ (c.f. Stoer (1975)), which is a 1-1 and onto

mapping from $\phi\varepsilon R$ to $\hat{\phi}\varepsilon R$.

We formally state the relation between the restricted updating conditions (3.1.1) and the inverse restricted updating conditions (3.3.1) in Lemma 3.3.3.

<u>Lemma 3.3.3</u> Let $B,H\varepsilon L(R^n)$ be symmetric and non-singular matrices such that $B = H^{-1}$, $s,y\varepsilon R^n$ non zero vectors such that $y\neq Bs$. Let Z be an m dimensional subspace of R^n, $o\leq m<n$; and define $Z_H = BZ$, $Z_H^{\perp} = B^{-1}Z^{\perp}$. Let P be the matrix projecting onto $Z^{\perp}$(orthogonal to BZ) in the metric B^{-1}, Q the matrix projecting onto $Z_H^{\perp}$(orthogonal to HZ_H) in the metric H^{-1}; define $\hat{s} = Qs$, $\hat{y} = Py$. Then $P = Q^T$, and a non-singular matrix $\bar{B}$ is in the set $S = \{\bar{B}\varepsilon L(R^n) | \bar{B} \text{ symmetric}, \bar{B}s = y, (\bar{B}-B)z = 0 \forall z\varepsilon Z\}$ if and only if the non-singular matrix $\bar{H} = \bar{B}^{-1}$ is in the set $S_H = \{\bar{H}\varepsilon L(R^n) | \bar{H} \text{ symmetric}, \bar{H}y = s, (\bar{H}-H)z = 0 \forall z\varepsilon Z_H\}$. Furthermore, as long as $\hat{s}$ and $\hat{y}$ are both nonzero, then the conditions for S and S_H to be nonempty are identical.

<u>Proof</u>: The relation $P = Q^T$ comes directly from using formula (2.3.2a) in Lemma 2.3.3 for P and Q. This gives $P = I - \sum_{i=1}^{m} \frac{Bz_i z_i^T}{\langle z_i,Bz_i\rangle}$, where $z_1,\ldots,z_m$ is a basis for Z such that $\langle z_j,Bz_k\rangle = 0$, $1\leq j<k\leq m$. If we then define $w_i = Bz_i$, $i=1,\ldots,m$, then $w_1,\ldots,w_m$ is a basis for $Z_H = BZ$ such that

$\langle w_j, Hw_k\rangle = 0$, $1\le j<k\le m$; and $P = I - \sum_{i=1}^{m} \frac{w_i w_i^T H}{\langle w_i, Hw_i\rangle} = Q^T$ by Lemma 2.3.3.

Next, let $\bar{B}$ be a non-singular matrix in S, and $\bar{H} = \bar{B}^{-1}$. Then $\bar{H}$ is symmetric, and $\bar{B}s = y$ implies $\bar{H}y = s$. Also, if w is any vector in Z_H, then $w = Bz$ for some $z\varepsilon Z$, and $(\bar{H}-H)w = (\bar{H}-H)Bz = \bar{H}Bz-z = \bar{H}(B-\bar{B})z = 0$. Thus $\bar{H}\varepsilon S_H$. Similarly if $\bar{H}\varepsilon S_H$ then $\bar{B} = \bar{H}^{-1}\varepsilon S$.

Finally, is $\hat{s}, \hat{y} \neq 0$, then $s \notin Z$ and $y \notin Z_H$. Thus by Lemmas 2.4.2 and 3.3.1, S and S_H are nonempty if and only if $y-Bs\varepsilon Z^{\perp}$, $s-Hy\varepsilon Z_H^{\perp}$ respectively. But since $Z_H^{\perp} = HZ^{\perp}$, $y-Bs\varepsilon Z^{\perp}$ if and only if $-H(y-Bs) = s-Hy\varepsilon Z_H^{\perp}$.

Note that if under the conditions of Lemma 3.3.3, $y-Bs\varepsilon Z^{\perp}$ but $\hat{y} = 0$, then S_H is empty because $\hat{y} = 0$ implies $y\varepsilon\hat{Z}$; but S is nonempty, because $\hat{y}-B\hat{s} \neq 0$ and $\hat{y} = 0$ imply $\hat{s} \neq 0$, $s \notin Z$. However in this case each $\bar{B}\varepsilon S$ is singular, as by Theorem 3.1.1 it satisfies $\bar{B}\hat{s} = \hat{y} = 0$. This, or the dual situation when $\hat{s} = 0$, $\hat{y} \neq 0$, is the only case in which conditions (3.1.1) and (3.3.1) are not equivalent when $Z_H = BZ$ and $B = H^{-1}$.

Lemma 3.3.3 shows that if we want our update to satisfy the restricted updating conditions (3.1.1), but want instead to update an inverse Hessian approximation H, then we should use updates $\bar{H}$ satisfying the restricted inverse updating conditions (3.3.1), where $Z_H = BZ$. This information will be helpful to us in later chapters. It might also be important to algorithms, where it may be easiest to think of updating conditions in terms of updates to B, but most efficient to actually keep and update inverse approximations H.

We conclude Chapter 3 by giving the H form of the new parameterization which was introduced in Section 3.2.

Theorem 3.3.4 Let $H \epsilon L(R^n)$ be symmetric; $s,w,y \epsilon R^n$ such that $\langle w,y\rangle \neq 0$, $s \neq Hy$, $w \neq s-Hy$; ϕ,σ real parameters. Consider the one-parameter w-class of updates (3.3.2), and the class

$$\hat{H}(\sigma) = \frac{(s-Hy)u^T+u(s-Hy)^T}{\langle u,y\rangle} - \frac{\langle s-Hy,y\rangle uu^T}{\langle u,y\rangle^2},$$

$$u = w+\sigma(s-Hy). \tag{3.3.4}$$

Let $\alpha = \langle s-Hy,y\rangle$. If $\alpha=0$, then there is a 1-1 and onto mapping from (3.3.2) to (3.3.4), with $\bar{H}(\phi,w) = \hat{H}(\sigma)$ for $\sigma = \phi\langle w,y\rangle/2$. If $\alpha \neq 0$, let $\bar{\phi} = \phi/\alpha^2$, and let S_1 be the set of $\bar{\phi}$ for which $\bar{H}(\bar{\phi},w)$ has the form $\bar{H} = H+vv^T-zz^T$, v,z nonzero vectors ϵR^n. Then $S_1 = \{\bar{\phi} | \alpha(\alpha-\bar{\phi}) > 0\}$, and there exists a 1-2 and onto mapping from $\bar{H}(\bar{\phi},w)$, $\bar{\phi}\epsilon S_1$, to $\hat{H}(\sigma)$, $\sigma\epsilon R-\{-\langle w,y\rangle/\alpha\}$, with $\bar{H}(\bar{\phi},w) = \hat{H}(\sigma)$ for $\sigma = -\langle w,y\rangle\left[\frac{1}{\alpha} \pm \frac{1}{(\alpha(\alpha-\bar{\phi}))^{1/2}}\right]$. If (3.3.4) is reparameterized as $\hat{H}(\omega)$, $\sigma = -\langle w,y\rangle\left[1\pm\omega^{1/2}\right]/\alpha$, then there exists a 1-1 and onto mapping from $\bar{H}(\bar{\phi},w)$, $\bar{\phi}\epsilon S_1$ to $\hat{H}(\omega)$, $\omega\epsilon(0,\infty)$, with $\bar{H}(\bar{\phi},w) = \hat{H}(\omega)$ for $\omega = \alpha/(\alpha-\bar{\phi})$.

CHAPTER IV

HEREDITARY POSITIVE DEFINITENESS

In Section 2.4 we showed that Davidon's (1975) algorithm leads us to consider updates which retain symmetry, obey the quasi-Newton equation, and leave $\bar{B}$-B orthogonal to some subspace Z. In Lemma 2.4.2 we gave the necessary and sufficient conditions for such updates to exist; and in Chapter 3 we showed that if these conditions are met, then there almost always exists a restricted DFP-BFGS class which satisfies them. We now want to add one more property to our updates: that they satisfy the hereditary positive definiteness condition,

$$B \text{ positive definite implies } \bar{B} \text{ positive definite.}$$

That is, we will be interested in updates $\bar{B}\epsilon L(R^n)$ to symmetric and positive definite matrices $B\epsilon L(R^n)$ which satisfy

$$\{\bar{B} \text{ symmetric and positive definite, } \bar{B}s = y,\ (\bar{B}-B)z = 0 \text{ for all } z\epsilon Z\} \tag{4.0.1}$$

for $s,y\epsilon R^n$, $s\neq 0$, Z an m dimensional subspace of R^n, $m<n$. In this chapter we discuss the conditions under which such updates exist.

There are two main reasons why many quasi-Newton optimization algorithms require successive Hessian approximations B_i to be positive definite. First, if a unique local minimum exists, then at and near this minimum the actual Hessian $G(x_i)$ will be positive definite, and so this is a good property to require of the Hessian approximation B_i (or the inverse Hessian approximation H_i). Secondly, if we use the usual search direction $-B_i^{-1}$

$g(x_i)$, then the directional derivative in the search direction is $-\langle g(x_i), B_i^{-1} g(x_i)\rangle$. Therefore if B_i (and hence B_i^{-1}) is positive definite, we are assured that our search direction is a descent direction.

In Section 4.1 we derive the necessary and sufficient conditions for there to exist updates satisfying (4.0.1). We first state Fletcher's (1970) result that an important portion of the DFP-BFGS class retains positive definiteness if $\langle s,y\rangle > 0$. We then show in Theorem 4.1.2 that a portion of the restricted DFP-BFGS class satisfies (4.0.1) if $\langle \hat{s},\hat{y}\rangle > 0$, and that no symmetric and positive definite restricted quasi-Newton updates exist if $\langle \hat{s},\hat{y}\rangle \leq 0$ ($\hat{s},\hat{y}$ as defined in Chapters 2 and 3, and again in Theorem 4.1.2). In Corollary 4.1.3 we prove the reasonable property that as the dimension of the subspace Z in (4.0.1) increases, the chance that we can preserve positive definiteness decreases. In Section 4.2 we apply our results to Davidon's (1975) algorithm, and show that they encourage us to adapt his algorithm to non-quadratic problems differently than he has. We propose an update which we think is more in keeping with Davidon's objectives. Finally in Section 4.3 we re-state our main result in terms of updates to inverse Hessian approximations H.

Section 4.1 - Hereditary Positive Definiteness for Restricted Updates

In this section we consider the conditions under which there exist restricted quasi-Newton updates which inherit symmetry and positive definiteness (condition (4.0.1)). This differs from our work of Sections 2.4 and 3.1 by the addition of the positive definiteness condition, which is our main concern here. Theorem 4.1.1 repeats Fletcher's (1970) theorem that if $\langle s,y\rangle > 0$, then the standard DFP-BFGS class $\bar{B}(\phi)$ (see 4.1.1)) retains symmetry and positive definiteness and obeys the quasi-Newton equation for all ϕ less than some constant which is greater than one. Theorem 4.1.2 generalizes this result to restricted updates, showing that to satisfy (4.0.1) we must have $<\hat{s},\hat{y}> > 0$ in addition to the restrictions on Z given by Lemma 2.4.2; and that in this case the restricted DFP-BFGS class $\hat{B}(\hat{\phi})$ (see (4.1.3)) also satisfies (4.0.1) for all $\hat{\phi}$ less than some constant in $(1,\infty)$. This is one reason why the restricted DFP-BFGS class was introduced in Chapter 3; we will see another reason in Chapter 5. Finally in Corollary 4.1.3 we show that as we enlarge the restricting subspace Z in (4.0.1) - which would probably be done in an attempt to preserve more old information - the chance that we can satisfy (4.0.1) decreases. It may be possible to preserve positive definiteness if we use restricting subspace Z_1, but not if we use a larger subspace Z_2 which contains Z_1. This result is important to the formulation of new algorithms.

<u>Theorem 4.1.1</u> (Fletcher (1970)) Let $B \in L(R^n)$ be symmetric and positive definite; $s,y \in R^n$ such that $<s,y> \neq 0$; ϕ a real parameter. Consider the DFP-BFGS class of updates

$$\bar{B}(\phi) = B + \frac{(y-Bs)y^T + y(y-Bs)^T}{<y,s>} - \frac{<y-Bs,s>yy^T}{<y,s>^2}$$

$$- \phi<s,Bs>\left[\frac{y}{<y,s>} - \frac{Bs}{<s,Bs>}\right]\left[\frac{y}{<y,s>} - \frac{Bs}{<s,Bs>}\right]^T \tag{4.1.1}$$

Then $\bar{B}(\phi)$ is positive definite if and only if $<s,y> > 0$ and $\phi < \alpha$,

$$\alpha = \frac{<y,B^{-1}y><s,Bs>}{<y,B^{-1}y><s,Bs>-<s,y>^2} \, . \tag{4.1.2}$$

One way to prove this result is shown in Chapter 7. Note that the theorem is independent of any meaning of s and y. Note also that since B is positive definite, the numerator of α in (4.1.2) is positive; the denominator is smaller than the numerator; and by setting $B = B^{1/2}\, B^{1/2}$, $B^{1/2}$ the symmetric and positive definite square root of B, and $v = B^{-1/2}y$, $w = B^{1/2}s$, the denominator becomes $\langle v,v\rangle\langle w,w\rangle - \langle v,w\rangle^2$ which is nonnegative by the Cauchy-Schwartz inequality. Thus α in (4.1.2) is greater than 1, which shows that if $\langle s,y\rangle > 0$, then the DFP ($\phi = 0$ in (4.1.1)), BFGS ($\phi = 1$), and the entire convex class ($\phi\in[0,1]$) are positive definite. The SR1 update, $\phi = -\langle s,Bs\rangle/\langle y-Bs,s\rangle$, may or may not preserve positive definiteness.

We now consider hereditary positive definiteness for restricted updates.

<u>Theorem 4.1.2</u> Let $B\in L(R^n)$ be symmetric and positive definite; $s,y\in R^n$ such that $s\neq 0$, $y\neq Bs$; ϕ a real parameter. Let Z be an m dimensional subspace of R^n, $0\leq m<n$. Let P be the matrix projecting onto $Z^\perp$ (orthogonal to BZ) in the metric B^{-1}; define $\hat{y} = Py$, $\hat{s} = P^Ts$. Then the set $S = \{\bar{B}\in L(R^n)\,|\,\bar{B}$ symmetric and positive definite, $\bar{B}s = y$, $(\bar{B}-B)z = 0$ for all $z\in Z\}$ is nonempty if and only if $(y-Bs)\in Z^\perp$, $s\notin Z$, and $\langle\hat{s},\hat{y}\rangle > 0$; and if these conditions are met, then the restricted DFP-BFGS class

$$\hat{B}(\phi) = B + \frac{(\hat{y}-B\hat{s})\hat{y}^T+\hat{y}(\hat{y}-B\hat{s})^T}{\langle\hat{y},\hat{s}\rangle} - \frac{\langle\hat{y}-B\hat{s},\hat{s}\rangle\hat{y}\hat{y}^T}{\langle\hat{y},\hat{s}\rangle^2}$$

$$- \hat{\phi}\langle\hat{s},B\hat{s}\rangle\left[\frac{\hat{y}}{\langle\hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle\hat{s},B\hat{s}\rangle}\right]\left[\frac{\hat{y}}{\langle\hat{y},\hat{s}\rangle} - \frac{B\hat{s}}{\langle\hat{s},B\hat{s}\rangle}\right]^T \qquad (4.1.3)$$

is in S for all $\hat{\phi}<\hat{\alpha}$, $\hat{\alpha} = \dfrac{\langle\hat{y},B^{-1}\hat{y}\rangle\langle\hat{s},B\hat{s}\rangle}{\langle\hat{y},B^{-1}\hat{y}\rangle\langle\hat{s},B\hat{s}\rangle-\langle\hat{s},\hat{y}\rangle^2}$. Furthermore, these updates then represent the smallest rank change $\bar{B}-B$ for which $\bar{B}$ is necessarily in S.

Proof: Let $S' = \{\bar{B}\epsilon L(R^n) | \bar{B}$ symmetric, $\bar{B}s = y$, $(\bar{B}-B)z = 0$ for all $z\epsilon Z\}$. Then S' contains the set S. From Lemma 2.4.2 we know that S' is nonempty if and only if $y-Bs\epsilon Z^{\perp}$ and $s\notin Z$. Let us assume these conditions are met. Then from Theorem 3.1.1, we know that the restricted DFP-BFGS class (4.1.3) is in S' as long as $<\hat{s},\hat{y}> \neq 0$. We must now determine whether S' has any positive definite members.

From Theorem 3.1.1 we also know that $\bar{B}\hat{s} = \hat{y}$ for any $\bar{B}\epsilon S'$; and due to $s\notin Z$ and formula (2.3.2b) for P, $\hat{s} = P^T s \neq 0$. Therefore for any $\bar{B}\epsilon S'$ to be positive definite, $<\hat{s},\bar{B}\hat{s}> = <\hat{s},\hat{y}>$ must be positive. On the other hand, if $<\hat{s},\hat{y}> > 0$, then the restricted DFP-BFGS class (4.1.3) is well-defined and in S'; and then by Theorem 4.1.1, its members with $\hat{\phi}<\hat{\alpha}$ are positive definite, and so are in S.

Finally, we see that if S is nonempty, then there exists $\bar{B}\epsilon S$ such that $\bar{B}-B$ is a matrix of rank two. The only rank one update in S' is the SR1 update $\bar{B} = B + \frac{(y-Bs)(y-Bs)^T}{<y-Bs,s>}$, which may have a negative eigenvalue or be undefined even if $<\hat{s},\hat{y}>$ or $<s,y> > 0$. Therefore if S is nonempty, the smallest rank update necessarily in S has rank two.

Thus if any updates exist satisfying (4.0.1), an important portion of the restricted DFP-BFGS class (4.1.3) will be among them. These will include the restricted DFP, BFGS, and the entire restricted convex class, $\hat{\phi}\epsilon[0,1]$.

The condition for hereditary positive definiteness to be possible, $0 < <\hat{s},\hat{y}>$ ($= <P^T s,Py> = <s,Py>$), is different than the condition $<s,y> > 0$ for unrestricted updates. If Davidon's (1975) algorithm is applied to a quadratic function with positive definite Hessian A, and $y_i = g(x_{i+1})$ -

$g(x_i)$, $s_i = x_{i+1} - x_i$, $Z_i = [s_0,\ldots,s_{i-1}]$, then both conditions are satisfied at each iteration. This is because for quadratic functions $y_i = As_i$, so that $<s_i,y_i> = <s_i,As_i> > 0$ as long as $s_i \neq 0$; and also $\hat{y}_i = A\hat{s}_i$ using Davidon's algorithm, so that $<\hat{s}_i,\hat{y}_i> = <\hat{s}_i,A\hat{s}_i> > 0$ as long as $\hat{s}_i \neq 0$. Corollary 4.1.3 shows how the conditions $<s,y> > 0$ and $<\hat{s},\hat{y}> > 0$ are related in the non-quadratic case.

Corollary 4.1.3 Let $B\varepsilon L(R^n)$ be symmetric and positive definite; $s,y\varepsilon R^n$ such that $s\neq 0$, $y\neq Bs$. Let Z_1,Z_2 be subspaces of R^n of dimensions m_1 and m_2 respectively, $0\leq m_1<m_2<n$, such that $Z_1 C Z_2$. Assume $y-Bs\varepsilon Z_2^{\perp}$, $s\notin Z_2$. For $i = 1,2$, let P_i be the matrix projecting onto $Z_i^{\perp}$ (orthogonal to BZ_i) in the metric B^{-1}; define $\hat{y}_i = P_i y$, $\hat{s}_i = P_i^T s$. Let $S_i = \{\bar{B}\varepsilon L(R^n)|\bar{B}$ symmetric and positive definite, $\bar{B}s = y$, $(\bar{B}-B)z = 0$ for all $z\varepsilon Z_i\}$, $i = 1,2$. Then for $i = 1,2$, S_i is nonempty if and only if $<s,P_i y> > 0$; and $<s,P_1 y> \geq <s,P_2 y>$, with equality if and only if $\hat{y}_2 = \hat{y}_1$, $\hat{s}_2 = \hat{s}_1$. Furthermore, $<s,P_1 y> > 0 > <s,P_2 y>$ is possible.

Proof: Since $Z_1 C Z_2$, $s\notin Z_2$ implies $s\notin Z_1$; and $Z_2^{\perp} C Z_1^{\perp}$, so that $y-Bs\varepsilon Z_2^{\perp}$ implies $y-Bs\varepsilon Z_1^{\perp}$. Thus we know from Theorem 4.1.2 that for $i = 1,2$, S_i is nonempty if and only if $<\hat{s}_i,\hat{y}_i> = <P_i^T s,P_i y> = <s,P_i^2 y> = <s,P_i y> > 0$.

Using formula (2.3.2a) for projection matrices, we have for $i = 1,2$

$$P_i = I - \sum_{j=1}^{m_i} \frac{Bz_j z_j^T}{<z_j,Bz_j>}$$

where $[z_1,\ldots,z_{m_i}]$ is a basis for Z_i such that $<z_j,Bz_k> = 0$, $1\leq j<k\leq m_i$. Therefore if we let $[z_1,\ldots,z_{m_1}]$ be such a basis for Z_1, and extend it

to such a basis $[z_1,\ldots,z_{m_1}, z_{m_1+1},\ldots,z_{m_2}]$ for Z_2, we have

$$P_2 = P_1 - \sum_{j=m_1+1}^{m_2} \frac{Bz_j z_j^T}{\langle z_j, Bz_j\rangle} \tag{4.1.4}$$

Thus

$$\langle s,P_2y\rangle = \langle s,P_1y\rangle - \sum_{j=m_1+1}^{m_2} \frac{\langle s,Bz_j\rangle\langle z_j,y\rangle}{\langle z_j,Bz_j\rangle},$$

$$= \langle s,P_1y\rangle - \sum_{j=m_1+1}^{m_2} \frac{\langle z_j,y\rangle^2}{\langle z_j,Bz_j\rangle} \tag{4.1.5}$$

because $\langle y\text{-}Bs\rangle \in Z_2^{\perp}$ implies that $\langle y\text{-}Bs, z_j\rangle = 0$, $\langle y,z_j\rangle = \langle Bs,z_j\rangle$ for $j = 1,\ldots,m_2$. Since $\langle z_j,Bz_j\rangle > 0$ due to B positive definite, (4.1.5) gives $\langle s,P_2y\rangle \geq \langle s,P_1y\rangle$, with equality if and only if

$$\langle y,z_j\rangle = 0, \quad j = m_1+1,\ldots,m_2. \tag{4.1.6}$$

From (4.1.4) and $\hat{y}_i = P_iy$, $\hat{s}_i = P_i^Ts$, (4.1.6) is precisely the necessary and sufficient condition for $\hat{y}_2 = \hat{y}_1$, $\hat{s}_2 = \hat{s}_1$.

As an example that $\langle s,P_1y\rangle > 0 > \langle s,P_2y\rangle$ is possible, let $B = I$, $s = (1,\ldots,1,1)$, $y = (1,\ldots,1,\alpha)$, $\alpha = \frac{3}{2} + m_1 - n$, $z_i = (\underbrace{0,\ldots,0}_{i-1},1,0,\ldots,0)$, $i = 1,\ldots,m_2$. (Recall that $0\leq m_1<m_2<n$.) Then $y\text{-}Bs = (0,\ldots,0,\alpha-1)$ so that $\langle y\text{-}Bs,z_i\rangle = 0$, $i=1,\ldots,m_2$, which implies $y\text{-}Bs\in Z_2^{\perp}$. Also $s\notin Z_2$, as any vector in Z_2 must have zeros in its last $n\text{-}m_2$ positions. Thus the conditions of Corollary 4.1.3 are met; and $\hat{s}_i = (\underbrace{0,\ldots,0}_{m_i},1,\ldots,1,1)$.

$\hat{y}_i = (\underbrace{0,\ldots,0}_{m_i},1,\ldots,1,\alpha)$, $i = 1,2$. Therefore $\langle s,P_1y\rangle = \langle\hat{s}_1,\hat{y}_1\rangle = \frac{1}{2}$, $\langle s,P_2y\rangle = \langle\hat{s}_2,\hat{y}_2\rangle = \frac{1}{2} - (m_2-m_1)$; and $\langle s,P_1y\rangle > 0 > \langle s,P_2y\rangle$.

So the more we restrict our updating space, the less likely it becomes that updates exist which preserve positive definiteness. Corollary 4.1.3 can be applied with $m_1 = 0$, in which case Z_1 is the empty set, and S_1 is the set of unrestricted symmetric and positive definite quasi-Newton updates. In this case the corollary tells us that for any restricting space Z, $\langle s,y\rangle \geq \langle\hat{s},\hat{y}\rangle = \langle s,Py\rangle$ (P the matrix projecting onto $Z^{\perp}$ in the metric B^{-1}) with equality only if $\hat{s} = s$ and $\hat{y} = y$. So $\langle s,y\rangle > 0$ is a necessary condition for restricted positive definite updates to exist, but it is not in general sufficient, as Corollary 4.1.3 tells us that $\langle s,y\rangle > 0 > \langle\hat{s},\hat{y}\rangle$ is possible.

In a real algorithm, we would construct Z such that $y-Bs\in Z^{\perp}$ and $s\notin Z$; and we would likely perform our line search so that $\langle s,y\rangle > 0$, but not necessarily so that $\langle s,Py\rangle > 0$. Whether we can use a restricted update or have to use an unrestricted one would likely depend upon whether $\langle s,Py\rangle > 0$. In Section 4.2 we apply Corollary 4.1.3 to show that Davidon (1975) chooses Z in a way which makes $\langle s,Py\rangle > 0$ less likely than some other reasonable choices of Z.

Section 4.2 - Hereditary Positive Definiteness in Davidon's Recent Algorithm

In Davidon's (1975) recent algorithm, if the restricted update cannot preserve positive definiteness because $\langle\hat{s},\hat{y}\rangle \leq 0$, he goes back to using the standard DFP-BFGS class, which will have positive definite members because $\langle s,y\rangle > 0$ will have been guaranteed by the line search. However this strategy has the disadvantage that by using the standard DFP-BFGS class we abandon our attempt to preserve old derivative information. Therefore we are interested in adapting Davidon's (1975) quadratic update to non-quadratic problems in a way which has a good chance of preserving positive definiteness. In this section we first show, in Corollary 4.2.1, how the analysis of hereditary positive definiteness in Section 4.1 applies to Davidon's algorithm. This analysis indicates that Davidon may not be making a very good choice of Z from among the possibilities which reduce to his basic algorithm in the quadratic case. We then discuss alternative choices of Z which seem to have at least as good a chance of preserving old information, and a better chance of being consistent with hereditary positive definiteness.

<u>Corollary 4.2.1</u> Let $B\varepsilon L(R^n)$ be symmetric and positive definite; $s,v,y\varepsilon R^n$ such that v and $y-Bs$ are linearly independent, and either $\langle v,s\rangle \neq 0$ or $\langle y-Bs,s\rangle \neq 0$. Let P be the matrix projecting onto $[y-Bs,v]$ in the metric B^{-1}; define $\hat{y} = Py$, $\hat{s} = P^Ts$. Let $S_1 = \{\bar{B}\varepsilon L(R^n)|\bar{B}$ symmetric and positive definite, $\bar{B}s = y\}$ and $S_2 = \{\bar{B}\varepsilon S_1|(\bar{B}-B)z = 0$ for all $z\varepsilon Z\}$. Then S_1 is nonempty if and only if $\langle s,y\rangle > 0$; S_2 is nonempty if and only if $\langle\hat{s},\hat{y}\rangle > 0$; $\langle s,y\rangle \geq \langle\hat{s},\hat{y}\rangle$ with equality if and only if $\hat{s} = s$ and $\hat{y} = y$; and $\langle s,y\rangle > 0 > \langle\hat{s},\hat{y}\rangle$ is possible.

<u>Proof</u>: We apply Corollary 4.1.3 with $Z_1 = \{\}$, $Z_2 = [y-Bs,v]$. Thus $Z_1^{\perp} = R^n$; $m_1 = 0$, $m_2 = n-2$; $P_1 = I$, $P_2 = P$; $\hat{s}_1 = s$, $\hat{y}_1 = y$, $\hat{s}_2 = \hat{s}$, $\hat{y}_2 = \hat{y}$; and S_1, S_2 are correctly defined. Therefore since $y-Bs\varepsilon[y-Bs,v]$, and $\langle y-Bs,s\rangle \neq 0$ or $\langle v,s\rangle \neq 0$ implies $s\not\varepsilon Z_2$, the conclusions of Corollary 4.2.1 follow directly from Corollary 4.1.3.

From Corollary 4.1.3 we know that the larger the subspace Z the smaller the chance that $\langle\hat{s},\hat{y}\rangle$ is positive. Davidon (1975) has chosen Z with dimension n-2, which is the largest dimension Z can have if the set of restricted quasi-Newton updates is to have more than one member. By doing this he is reducing the chance that a positive definite restricted update will exist. Furthermore, while in the quadratic case $y_i - B_i s_i$ and v_i are orthogonal to all the past values s_j, $j<i$, in the general case all we can really say is that v_i ($= (y_{i-1}-B_{i-1}s_{i-1}) \ \langle v_{i-1},s_{i-1}\rangle -$ $v_{i-1} \ \langle y_{i-1}-B_{i-1}s_{i-1},s_{i-1}\rangle$) is orthogonal to s_{i-1}, and $B_{i+1}-B_i$ is orthogonal to any projection of s_{i-1} which is orthogonal to $y_i-B_is_i$ and v_i, for example

$$(s_{i-1})_p = s_{i-1} - \frac{u_i\langle s_{i-1},y_i-B_is_i\rangle}{\langle u_i,y_i-B_is_i\rangle} \tag{4.2.1}$$

for any u_i orthogonal to v_i and not to $y_i-B_is_i$.

The relation of $B_{i+1}-B_i$ to s_{i-2} in Davidon's scheme is already quite vague. Therefore it seems it would be just as reasonable to select a u_i in (4.2.1) and define $Z' = [(s_{i-1})_p]$, as $(s_{i-1})_p$ seems to be the only vector of importance to which $B_{i+1}-B_i$ is orthogonal in Davidon's scheme. Since Davidon's choice of $Z^{\perp} = [y_i-B_is_i,v_i]$ is equivalent to defining $Z = [(s_{i-1})_p,w_1,\ldots,w_{n-3}]$ for some vectors $w_1,\ldots,w_{n-3}\varepsilon R^n$ which are orthogonal to $y_i-B_is_i$ and v_i, Corollary 4.1.3 shows that the chance of preserving positive definiteness is greater using Z'. However the new choice Z' does not lead to an algorithm with finite termination for quadratics.

Since we would like to preserve more old derivative information, a second alternative is to define $\overline{s}_j$ as the projection of s_j orthogonal to $B_i^{-1}(y_i-B_is_i)$ in B_i metric,

$$\bar{s}_j = s_j - \frac{B_i^{-1}(y_i-B_is_i)\langle s_j,y_i-B_is_i\rangle}{\langle y_i-B_is_i,B_i^{-1}(y_i-B_is_i)\rangle}$$

$j = i-\ell,\ldots,i-1$; and $Z = [\bar{s}_{i-\ell},\ldots,\bar{s}_{i-1}]$ for some $\ell<n$. For quadratic problems $\bar{s}_j = s_j$ (because $\langle s_j,y_i-B_is_i\rangle = 0$), and if $\ell=i$ then using the restricted DFP-BFGS class with this Z yields Davidon's (1975) update for quadratics. However this method is also attractive for non-quadratics. Each $\bar{s}_j$ is orthogonal to $y_i-B_is_i$, so $y_i-B_is_i\epsilon Z^{\perp}$. The larger the choice of ℓ, the less likely it is that positive definite updates will exist, but the more $\bar{s}_j$'s $(B_{i+1}-B_i)$ will be orthogonal to. Therefore we can balance the choice of ℓ, perhaps selecting the largest $\ell \leq$ some constant $\lambda \leq n-2$ for which $\langle\hat{s},\hat{y}\rangle$ is positive. This strategy probably has a better chance of preserving positive definiteness than Davidon's (1975) algorithm, where Z is fixed, includes $(s_{i-1})_p$, (which is similar to $\bar{s}_{i-1}$), and has dimension n-2. It may also be better at preserving old information, as $(B_{i+1}-B_i)$ would be orthogonal to $\bar{s}_{i-1},\ldots,\bar{s}_{i-\ell}$, whereas in Davidon's scheme it is only orthogonal to some projections of s_{i-1}. Since this strategy also reduces to Davidon's basic algorithm in the quadratic case, we think that in practice it might be at least as successful as his general algorithm. It does require more overhead, but this is not very important for problems where evaluating f is the major expense. We have not yet tested this suggestion, but we discuss it and related algorithms further in Chapter 9.

Section 4.3 - Hereditary Positive Definiteness for Restricted Inverse Updates

We conclude this chapter by re-stating our main theorem, Theorem for updates to approximations H to the inverse Hessian $G(x)^{-1}$. We include this theorem because it is likely that some algorithms will want to use our result in this form.

Theorem 4.3.1 Let $H \in L(R^n)$ be symmetric and positive definite; $s, y \in R^n$ such that $y \neq 0$, $s \neq Hy$; $\hat{\phi}$ a real parameter. Let Z_H be an m dimensional subspace of R^n, $0 \leq m < n$. Let Q be the matrix projecting onto $Z_H^{\perp}$ (orthogonal to HZ_H) in the metric H^{-1}, define $\hat{s} = Qs$, $\hat{y} = Q^T y$. Then the set $S_H = \{\bar{H} \in L(R^n) | \bar{H}$ symmetric and positive definite, $\bar{H}y = s$, $(\bar{H}-H)z = 0$ for all $z \in Z_H\}$ is nonempty if and only if $s - Hy \in Z_H^{\perp}$, $y \notin Z_H$, and $\langle \hat{s}, \hat{y} \rangle > 0$; and if these conditions are met, then the inverse restricted DFP-BFGS class

$$\hat{H}(\hat{\phi}) = H + \frac{(\hat{s}-H\hat{y})\hat{s}^T + \hat{s}(\hat{s}-H\hat{y})^T}{\langle \hat{s}, \hat{y} \rangle} - \frac{\langle \hat{s}-H\hat{y}, \hat{y} \rangle \hat{s}\hat{s}^T}{\langle \hat{s}, \hat{y} \rangle^2}$$

$$- \hat{\phi}(\hat{y}, H\hat{y}) \left[\frac{\hat{s}}{\langle \hat{s}, \hat{y} \rangle} - \frac{H\hat{y}}{\langle \hat{y}, H\hat{y} \rangle} \right] \left[\frac{\hat{s}}{\langle \hat{s}, \hat{y} \rangle} - \frac{H\hat{y}}{\langle \hat{y}, H\hat{y} \rangle} \right]^T \qquad (4.3.1)$$

is in S_H for all $\hat{\phi} < \dfrac{\langle \hat{s}, H^{-1}\hat{s} \rangle \langle \hat{y}, H\hat{y} \rangle}{\langle \hat{s}, H^{-1}\hat{s} \rangle \langle \hat{y}, H\hat{y} \rangle - \langle \hat{s}, \hat{y} \rangle^2}$. Furthermore, these updates represent the smallest rank change $(\bar{H}-H)$ for which $\bar{H}$ is necessarily in S_H.

As we showed in Lemma 3.3.3, if Z_H in Theorem 4.3.1 and Z in Theorem 4.1.1 are related by $Z_H = BZ$, then $Q = P^T$; and then if s and y are the same in these two places, so are $\hat{s}$ and $\hat{y}$. In this case the restricted inverse DFP-BFGS class (4.3.1) gives the inverses of the restricted DFP-BFGS class (4.1.6), and $\langle \hat{s}, \hat{y} \rangle > 0$ is the necessary and sufficient condition for either class to have positive definite members. If $\langle \hat{s}, \hat{y} \rangle > 0$, then the restricted inverse DFP ($\hat{\phi} = 1$ in (4.3.1)), the BFGS ($\hat{\phi} = 0$) and the entire convex class ($\hat{\phi} \in [0,1]$) retain positive definiteness.

CHAPTER V

MINIMUM CHANGE UPDATES

In Chapter 3 we discussed updates $\bar{B}$ to a symmetric matrix B which retain symmetry, satisfy the quasi-Newton equation $\bar{B}s = y$, and leave $\bar{B}-B$ orthogonal to some subspace Z. In Chapter 4 we considered those $\bar{B}$ which in addition preserve positive definiteness if B is positive definite. We have seen that the restricted DFP-BFGS class obeys these conditions if it is possible to satisfy them. However, there are other updates which obey them too. In this chapter we give a good reason to select the restricted DFP or BFGS: we show that the restricted DFP minimizes $\|\bar{B}-B\|_{Fr,\hat{B}^{-1/2}}$, and the restricted BFGS minimizes $\|\bar{B}^{-1}-B^{-1}\|_{Fr,\hat{B}^{1/2}}$, among all updates $\bar{B}$ satisfying our conditions, where (throughout this chapter) $\hat{B}$ is the updated matrix we actually choose. These results are an extension of the work of Dennis and Moré (1977) deriving minimum norm unrestricted updates.

We start Section 5.1 by explaining why we are interested in updates which minimize $\|\bar{B}-B\|$, or $\|\bar{B}^{-1}-B^{-1}\|$. Next we state Dennis and Moré's (1977) result that if any symmetric and positive definite updates exist satisfying $\bar{B}s=y$, then $\|\bar{B}-B\|_{Fr,\hat{B}^{-1/2}}$ is minimized by the standard DFP update; and we comment on why this norm is probably the best one to use is measuring the change $\bar{B}-B$. Then we show that among restricted updates, $\|\bar{B}-B\|_{Fr,\hat{B}^{-1/2}}$ is minimized by the restricted DFP. This lends some support to using the restricted DFP in algorithms making restricted quasi-Newton updates.

However, most optimization algorithms actually use B^{-1}, and not B, during their iteration. Therefore we explain that we are probably more interested in minimizing $\|\bar{B}^{-1}-B^{-1}\|$, or $\|\bar{H}-H\|$ if we are updating approximations H to the inverse Hessian. In Section 5.2 we repeat Dennis and Moré's (1977) result that among unrestricted symmetric and positive definite updates satisfying $\bar{B}s=y$ (or $\bar{H}y=s$), $\|\bar{B}^{-1}-B^{-1}\|_{Fr,\hat{B}^{1/2}}$ (or $\|\bar{H}-H\|_{Fr,\hat{H}^{-1/2}}$) is minimized by the BFGS update. We discuss why this is an appropriate norm for measuring change to inverse Hessian approximations. Since the BFGS has been the most successful unrestricted quasi-Newton update (c.f. Brodlie (1976)), we are encouraged that minimizing $\|\bar{B}^{-1}-B^{-1}\|_{Fr,\hat{B}^{1/2}}$ may be a good way to select updates. We then show that in the case where we restrict the subspace of $\bar{B}-B$ (or $\bar{H}-H$), that $\|\bar{B}^{-1}-B^{-1}\|_{Fr,\hat{B}^{1/2}}$ (or $\|\bar{H}-H\|_{Fr,\hat{H}^{-1/2}}$) is minimized by the restricted BFGS update. This analysis supports the restricted BFGS as the most desirable formula to use in algorithms making restricted quasi-Newton updates. In Chapter 8 we will see that this choice is supported in practice.

Most of the material in this chapter is also contained in Schnabel (1977).

Section 5.1 - Minimum Change Restricted Updates

We have shown in Section 2.4 and Chapter 4 that our optimization algorithms are likely to be interested in updating a symmetric and positive definite matrix B into a symmetric and positive definite matrix $\bar{B}$, so that $\bar{B}s=y$ and $\bar{B}-B$ is orthogonal to some subspace Z, where B and $\bar{B}$ approximate $G(x)$ and $G(\bar{x})$ respectively, $s = \bar{x}-x$ and $y = g(\bar{x})-g(x)$. We have also shown the necessary and sufficient conditions upon s, y, and Z so that any such $\bar{B}$ exist. If they do, then among them are the restricted DFP-BFGS class (4.1.3) as well as many other updates. Therefore we need a way to choose between these possibilities. Since we have no other information to put into $\bar{B}$, and since we hope B is a fairly good approximation to G(x) and hence, especially when $\bar{x}$-x is small, also to $G(\bar{x})$, it seems reasonable to want to change B as little as possible subject to the aforementioned conditions. Therefore in this section we determine which symmetric and positive definite restricted quasi-Newton update minimizes $\|\bar{B}-B\|$ for some appropriate norm $\|\cdot\|$.

Actually, though, our optimization algorithms use the approximate Hessian $\bar{B}$ only to form the next search direction $-\bar{B}^{-1}g(\bar{x})$; and since they only use the inverse of $\bar{B}$, it may be even more reasonable to minimize $\|\bar{B}^{-1}-B^{-1}\|$ subject to our other conditions on the update. This is equivalent to minimizing $\|\bar{H}-H\|$ for algorithms which maintain approximations H to the inverse of the Hessian, and update them subject to $\bar{H}$ symmetric and positive definite, $\bar{H}y = s$ and $\bar{H}-H$ orthogonal to $H^{-1}Z$. This is the approach we take in Section 5.2. It encourages us to use the restricted BFGS update, whereas the analysis of this section supports the use of the restricted DFP.

We start by stating Dennis and Moré's (1977) theorem giving the $\bar{B}$ which minimizes $\|\bar{B}-B\|_{Fr,M}$ subject to $\bar{B}$ symmetric and $\bar{B}s = y$. The DFP is the minimum norm update when $M = \hat{B}^{-1/2}$, where $\hat{B}$ denotes the updated matrix we actually chose (here, the DFP), and we explain why this norm is probably the most natural one to use. Then in Theorem 5.1.2 we derive the minimum change restricted update in the same general Frobenius norm. The proof is an extension of the techniques of Dennis and Moré (1977). Corollary 5.1.3 shows that among restricted updates, $\|\bar{B}-B\|_{Fr,\hat{B}^{-1/2}}$ is minimized by the restricted DFP.

<u>Theorem 5.1.1</u> (Dennis and Moré (1977)) Let $B, M \in L(R^n)$ be symmetric, M non-singular; $s, y \in R^n$ such that $s \neq 0$ and $y \neq Bs$. Then the unique solution to

$$\min \{ \| \bar{B}-B \|_{Fr,M} \mid \bar{B} \in L(R^n) \text{ symmetric}, \bar{B}s = y \} \qquad (5.1.1)$$

is

$$\hat{B} = B + \frac{(y-Bs)v^T + v(y-Bs)^T}{\langle v,s \rangle} - \frac{\langle y-Bs, s \rangle vv^T}{\langle v,s \rangle^2}, \quad v = M^{-2}s \qquad (5.1.2)$$

If $M=I$, (5.1.1) is solved by the Powell-symmetric-Broyden update, $v = s$ in (5.1.2). If B is positive definite and $M = B^{-1/2}$, (5.1.1) is solved by the Greenstadt update, $v = Bs$ in (5.1.2). If B is positive definite and $\langle s,y \rangle > 0$, then the DFP update, $v = \bar{B}s = y$ in (5.1.2), is positive definite and is the unique solution to (5.1.1) with $M = \hat{B}^{-1/2}$, or with $M = B_+^{-1/2}$ for any other matrix B_+ satisfying the update conditions.

<u>Proof</u>: A simpler version of the proof of Theorem 5.1.2.

If we are dealing with minimization problems where $\bar{B}$ approximates $G(\bar{x})$, then especially as $\bar{x}$ approaches the solution x^*, the natural scaling of the variables is by $G(x^*)^{-1/2}$ (assuming $G(x^*)$ is positive definite). This scaling, which makes the contour lines near the solution circular, corresponds to using norms weighted by $G(x^*)^{-1/2}$. Therefore we would really like to solve (5.1.1) with $M = G(x^*)^{-1/2}$. Since our next Hessian approximation $\hat{B}$ is likely to be our best approximation to $G(x^*)$ thus far, we think it is best to solve (5.1.1) with a norm weighted by $\hat{B}^{-1/2}$. The Frobenius norm is natural because it is the obvious measure of the size of the change in a matrix, the ℓ_2 norm of the changes in its elements. Therefore we think the Frobenius norm weighted by $\hat{B}^{-1/2}$ is the best norm we can use in measuring $\bar{B}-B$. Thus Theorem 5.1.1 supports the use of the DFP update.

We now derive the minimum change restricted update in a general weighted Frobenius norm.

Theorem 5.1.2 Let $B, M \in L(R^n)$ be symmetric, M non-singular; $s, y \in R^n$ such that $s \neq 0$, $y \neq Bs$. Let Z be an m dimensional subspace of R^n, $0 \leq m < n$. Assume $y-Bs \in Z^{\perp}$ and $s \notin Z$. Let $P[M]$ be the matrix projecting onto $Z^{\perp}$ (orthogonal to $M^{-2}Z$) in the metric M^2. Then the unique solution to

$$\min \{\|\bar{B}-B\|_{Fr,M} \mid \bar{B} \in L(R^n) \text{ symmetric, } \bar{B}s = y,\ (\bar{B}-B)z = 0 \text{ for all } z \in Z\} \tag{5.1.3}$$

is

$$\hat{B} = B + \frac{(y-Bs)v^T + v(y-Bs)^T}{\langle v,s\rangle} - \frac{\langle y-Bs,s\rangle vv^T}{\langle v,s\rangle^2}, \quad v = P[M]\, M^{-2}s \tag{5.1.4}$$

Proof: Let $S = \{\bar{B} \in L(R^n) \mid \bar{B}$ symmetric, $\bar{B}s = y$, $(\bar{B}-B)z = 0$ for all $z \in Z\}$. Then S is non-empty by Lemma 2.4.2. $\hat{B}$ is well-defined because $s \notin Z$ implies $P[M]^Ts \neq 0$, and $P[M]^2 = P[M]$, $P[M]M^{-2} = M^{-2}P[M]^T$, and M symmetric and non-singular imply $\langle v,s\rangle = \langle P[M]M^{-2}s, P[M]^Ts\rangle = \langle M^{-2}P[M]^Ts, P[M]^Ts\rangle = \langle M^{-1}P[M]^Ts, M^{-1}P[M]^Ts\rangle \neq 0$. Also $\hat{B} \in S$ due to $\hat{B}$ symmetric, $\hat{B}s = y$, and $y-Bs$, $v \in Z^{\perp}$.

Now consider any $\bar{B} \in S$. Define $\bar{E} = M(\bar{B}-B)M$, $\hat{E} = M(\hat{B}-B)M$. Since $y-Bs = (\bar{B}-B)s$ and $\bar{B}-B$ is symmetric, subtracting B from both sides of (5.1.4) and pre and post-multiplying by M gives

$$\hat{E} = \frac{\bar{E}(M^{-1}s)(Mv)^T + (Mv)(M^{-1}s)^T\bar{E}}{\langle Mv, M^{-1}s\rangle} - \frac{\langle M^{-1}s, \bar{E}(M^{-1}s)\rangle (Mv)(Mv)^T}{\langle Mv, M^{-1}s\rangle^2} \quad (5.1.5)$$

Using formula (2.3.2a) for P[M],

$$v = M^{-2}s - \sum_{i=1}^{m} \frac{\langle M^{-2}s, z_i\rangle}{\langle M^{-2}z_i, z_i\rangle} M^{-2}z_i \quad (5.1.6)$$

where $z_1,\ldots,z_m$ is a basis for Z such that $\langle z_j, M^{-2}z_k\rangle = 0$, $1 \leq j < k \leq m$. Define $\delta = v - M^{-2}s$. By (5.1.6), $\delta \epsilon M^{-2}Z$, so since $(\bar{B}-B)z = 0$ for any $z \epsilon Z$, $(\bar{B}-B)M^2\delta = 0$. Therefore $\bar{E}(Mv) = M(\bar{B}-B)M^2v = M(\bar{B}-B)(s+M^2\delta) = M(\bar{B}-B)s = M(y-Bs)$. Similarly $\hat{E}(Mv) = M(y-Bs)$. Thus $\bar{E}(Mv) = \hat{E}(Mv)$.

Now consider any $w \epsilon R^n$ such that $\langle w, Mv\rangle = 0$. Then (5.1.5) gives

$$\hat{E}w = \frac{(Mv)(M^{-1}s)^T\bar{E}w}{\langle Mv, M^{-1}s\rangle} \quad (5.1.7)$$

From above, $(M^{-1}s)^T\bar{E} = (Mv)^T\bar{E}$, and $\langle Mv, M^{-1}s\rangle = \langle Mv, Mv - M\delta\rangle = \langle Mv, Mv\rangle$ as $v \epsilon Z^{\perp}$, $M^2\delta \epsilon Z$. Therefore from (5.1.7)

$$\hat{E}w = \frac{(Mv)(Mv)^T\bar{E}w}{\langle Mv, Mv\rangle}$$

and using Lemma 2.1.3

$$\|\hat{E}w\|_2 \leq \left\| \frac{(Mv)(Mv)^T}{\langle Mv, Mv\rangle} \right\|_2 \|\bar{E}w\|_2 = \|\bar{E}w\|_2.$$

Thus $\|\hat{E}u_i\|_2 \leq \|\bar{E}u_i\|_2$ for the set of orthonormal vectors $\{u_i\} = (\frac{Mv}{\|Mv\|_2}, w_1, \ldots, w_{n-1})$. Therefore by Definition 2.1.2 of the Frobenius norm, $\|\hat{E}\|_{Fr} \leq \|\bar{E}\|_{Fr}$, or equivalently, $\|\hat{B}-B\|_{Fr,M} \leq \|\bar{B}-B\|_{Fr,M}$. $\hat{B}$ is the unique solution to (5.1.3) because $f:L(R^n) \to R$, $f(\bar{B}) = \|\bar{B}-B\|_{Fr,M}$ is strictly convex on S.

Note that Theorem 5.1.1 is the special case of Theorem 5.1.2 when $Z = \{\}$, $P[M] = I$ for all M. Again, we are most interested in the solution to the restricted minimum change problem (5.1.3) when $\bar{B}$ positive definite is possible and $M = \hat{B}^{-1/2}$, $\hat{B}$ our next Hessian approximation. At first glance, this solution (5.1.4) would seem to require a fore-knowledge of $\hat{B}$. However Corollary 5.1.3 shows that this is not so, and that if any symmetric and positive definite restricted quasi-Newton updates $\bar{B}$ exist, then $\|\bar{B}-B\|_{Fr,\hat{B}^{-1/2}}$ is minimized by the restricted DFP.

<u>Corollary 5.1.3</u> Let $B\in L(R^n)$ be symmetric and positive definite; $s,y\in R^n$ such that $s\neq 0$, $y\neq Bs$. Let Z be an m dimensional subspace of R^n, $0\leq m<n$. Assume $y-Bs\in Z^{\perp}$, $s\notin Z$. Let P = the matrix projecting onto $Z^{\perp}$ (orthogonal to BZ) in the metric B^{-1}; define $\hat{y} = Py$, $\hat{s} = P^Ts$, and assume $\langle\hat{s},\hat{y}\rangle > 0$. Then the restricted DFP update

$$B^* = B + \frac{(\hat{y}-B\hat{s})\hat{y}^T+\hat{y}(\hat{y}-B\hat{s})^T}{\langle\hat{y},\hat{s}\rangle} - \frac{\langle\hat{y}-B\hat{s},\hat{s}\rangle\hat{y}\hat{y}^T}{\langle\hat{y},\hat{s}\rangle^2} \qquad (5.1.8)$$

is positive definite and is the unique solution to (5.1.3) with $M = B^{*-1/2}$, or with $M = B_+^{-1/2}$ for any other matrix B_+ satisfying the update conditions.

<u>Proof:</u> Let $S = \{\bar{B}\in L(R^n) | \bar{B}$ symmetric and positive definite, $\bar{B}s = y$, $(\bar{B}-B)z = 0$ for all $z\in Z\}$. Then by Theorem 4.1.2, S is nonempty and $B^*\in S$. It only remains to show that B^* is equivalent to the solution given by Theorem 5.1.2 to problem (5.1.3) with $M = B^{*-1/2}$,

$$\bar{B} = B + \frac{(y-Bs)v^T+v(y-Bs)}{<v,s>} - \frac{<y-Bs,s>vv^T}{<v,s>^2},$$

$$v = P[B*^{-1/2}]B*s \qquad (5.1.9)$$

By Lemma 2.3.3, $P[B*^{-1/2}] = I - \sum_{i=1}^{m} \frac{B*z_i z_i^T}{<z_i,B*z_i>}$, where $z_1,\ldots,z_m$ are a basis for Z such that $<z_j,B*z_k> = 0$, $1\le j<k\le m$. Since $(\bar{B}-B)z = 0$ for any $z\varepsilon Z$ and any $\bar{B}\varepsilon S$, $B*z_i = Bz_i$, $i=1,\ldots,m$. Therefore the formula (2.3.2a) for P gives $P[B*^{-1/2}] = P$ ($=P[B^{-1/2}]$ in the notation of Theorem 5.1.2). Thus v in (5.1.9) = PB*s = Py = $\hat{y}$; and as was shown in the proof of Theorem 3.1.1, $y-Bs = \hat{y}-B\hat{s}$, $<y-Bs,s> = <\hat{y}-B\hat{s},\hat{s}>$, and $<\hat{y},s> = <\hat{y},\hat{s}>$. Therefore the restricted DFP (5.1.8) is equivalent to (5.1.9), and is the unique solution to (5.1.3) with $M = B*^{-1/2}$. The proof with M^{-2} equal to any other $B_+\varepsilon S$ is identical.

Corollary 5.1.3 supports the selection of the restricted DFP among all symmetric and positive definite restricted quasi-Newton updates. However, as we have said, it seems more important to minimize the change $\bar{B}^{-1}-B^{-1}$, since it is the inverse of the Hessian approximation which we really use. We address this problem in Section 5.2.

Section 5.2 - Minimum Change Restricted Inverse Updates

In this section we show which update $\bar{B}$ to a symmetric matrix B minimizes $\|\bar{B}^{-1}-B^{-1}\|_{Fr,M}$, subject to $\bar{B}$ symmetric, $\bar{B}s = y$, and $\bar{B}-B$ orthogonal to some subspace Z. From Lemma 3.3.3 we know that this is equivalent to asking which update $\bar{H}$ to $H = B^{-1}$ minimizes $\|\bar{H}-H\|_{Fr,M}$, subject to $\bar{H}$ symmetric, $\bar{H}y = s$, and $\bar{H}-H$ orthogonal to $Z_H = BZ$. This is the form in which we solve our problem - it is simply the dual of the work of Section 5.1, and no new proofs are required. In Theorem 5.2.1 we repeat the result of Dennis and Moré (1977) which shows that for unrestricted updates, $\|\bar{H}-H\|_{Fr,\hat{H}^{-1/2}}$ is minimized by the BFGS update, where $\hat{H}$ is our next inverse Hessian approximation. We comment on why the Frobenius norm weighted by $\hat{H}^{-1/2}$ is the appropriate one to use in measuring $\bar{H}-H$. Since the BFGS has been the most successful unrestricted update (c.f. Brodlie (1976)), we are encouraged in the use of this minimum norm criterion. In Theorem 5.2.2 we show that for restricted updates, $\|\bar{H}-H\|_{Fr,\hat{H}^{-1/2}}$ is minimized by the restricted BFGS. This lends support to the use of the restricted BFGS in algorithms like Davidon's (1975). Davidon makes a different choice, based on his optimal conditioning strategy, which we discuss in Chapters 7 and 8. The computer results of Chapter 8 support the use of the restricted BFGS rather than Davidon's optimally conditioned update, and hence give further support to the use of the minimum norm criterion for selecting updates.

<u>Theorem 5.2.1</u> (Dennis and Moré (1977)) Let $H, M \in L(R^n)$ be symmetric, M non-singular; $s, y \in R^n$ such that $y \neq 0$ and $s \neq Hy$. Then the unique solution to

$$\min \{\|\bar{H}-H\|_{Fr,M} \mid \bar{H} \in L(R^n) \text{ symmetric}, \bar{H}y = s\} \tag{5.2.1}$$

is

$$\hat{H} = H + \frac{(s-Hy)w^T + w(s-Hy)^T}{\langle w,y\rangle} - \frac{\langle s-Hy,y\rangle ww^T}{\langle w,y\rangle^2}, \quad w = M^{-2}y \tag{5.2.2}$$

If H is positive definite and $\langle s,y\rangle > 0$, then the BFGS update, $w = \bar{H}y = s$

in (5.2.2), is positive definite and is the unique solution to (5.2.1) with $M = \hat{H}^{-1/2}$, or with $M = H_+^{-1/2}$ for any other matrix H_+ satisfying the update conditions.

The natural scaling of the variables in any optimization problem which was discussed in Section 5.1 corresponds to scaling the inverse Hessian by $G(x^*)^{1/2}$. Therefore we would really like to solve (5.2.1) with $M = G(x^*)^{1/2}$. Since the best approximation we have to $G(x^*)$ is likely to be $\hat{H}^{-1}$, we are most interested in the solution to (5.2.1) with $M = \hat{H}^{-1/2}$. Thus Theorem 5.1.1 supports the use of the BFGS update. Note that if B is non-singular, then (5.2.1) is equivalent to

$$\min \{ \| \bar{B}^{-1} - B^{-1} \|_{Fr,M} | \bar{B} \in L(R^n) \text{ symmetric and non-singular, } \bar{B}s = y \} \qquad (5.2.3)$$

so that the B form of the BFGS solves (5.2.3) when B is positive definite, $\langle s,y \rangle > 0$ and $M = \hat{B}^{1/2}$.

We now state the analog of Theorem 5.1.2 and Corollary 5.1.3 for restricted inverse updates.

<u>Theorem 5.2.2</u> Let H, $M \in L(R^n)$ be symmetric, M non-singular; $s, y \in R^n$ such that $y \neq 0$, $s \neq Hy$. Let Z_H be an m dimensional subspace of R^n, $0 \leq m < n$. Assume $s - Hy \in Z_H^{\perp}$ and $s \notin Z_H$. Let $Q[M]$ be the matrix projecting onto $Z_H^{\perp}$ (orthogonal to $M^{-2} Z_H$) in the metric M^2. Then the unique solution to

$$\min \{ \| \bar{H} - H \|_{Fr,M} | \bar{H} \in L(R^n) \text{ symmetric, } \bar{H}y = s, (\bar{H} - H)z = 0 \text{ for all } z \in Z_H \} \qquad (5.2.4)$$

is

$$\hat{H} = H + \frac{(s-Hy)w^T + w(s-Hy)^T}{\langle w,y\rangle} - \frac{\langle s-Hy,y\rangle ww^T}{\langle w,y\rangle^2}, \quad w = Q[M]\cdot M^{-2}y.$$

If H is positive definite, define $Q = Q[H^{-1/2}]$, $\hat{s} = Qs$, $\hat{y} = Q^Ty$. Then if $\langle \hat{s},\hat{y}\rangle > 0$, the restricted BFGS update

$$H^* = H + \frac{(\hat{s}-H\hat{y})\hat{s}^T + \hat{s}(\hat{s}-H\hat{y})^T}{\langle \hat{s},\hat{y}\rangle} - \frac{\langle \hat{s}-H\hat{y},\hat{y}\rangle \hat{s}\hat{s}^T}{\langle \hat{s},\hat{y}\rangle^2} \qquad (5.2.5)$$

is positive definite and is the unique solution to (5.2.4) with $M = H^{*-1/2}$, or with $M = H_+^{-1/2}$ for any other matrix H_+ satisfying the update conditions.

Theorem 5.2.2 supports the use of the restricted BFGS (5.2.6) when using restricted updates. From Lemma 3.3.3, we know that if $H = B^{-1}$ and $Z_H = BZ$, then (5.2.4) is equivalent to

$$\min \{ \| \bar{B}^{-1} - B^{-1} \|_{Fr,M} \mid \bar{B} \in L(R^n) \text{ symmetric and non-singular},$$

$$\bar{B}s = y,\ (\bar{B}-B)z = 0 \text{ for all } z \in Z\} \qquad (5.2.6)$$

Therefore the solution to (5.2.6) when B is positive definite and $\langle \hat{s},\hat{y}\rangle > 0$ is the B form of the restricted BFGS, (4.1.3) with $\hat{\phi} = 1$. Since we feel that $\bar{B}^{-1} - B^{-1}$ is the quantity we really want to minimize, we think Theorem 5.2.2 supports the use of the restricted BFGS update even if we are keeping approximations to $G(x)$ instead of $G(x)^{-1}$.

Therefore the analysis of this chapter would encourage us to use the restricted BFGS in an algorithm like Davidon's (1975) - or at least a member of the restricted convex class, a convex combination of the two minimum change updates, the restricted DFP and BFGS. Davidon (1975) instead selects the positive definite member of the restricted DFP-BFGS class which minimizes a certain bound on the condition number of $\bar{B}$ - and in doing so may select an update within or outside the convex class. Our computer results in Chapter B show that Davidon's algorithm using this update is inferior to the same algorithm using the restricted BFGS in most of our tests. This seems to further support the minimum norm criterion for selecting updates. Furthermore, our analysis in Chapter 7 shows that the restricted BFGS usually, and some member of the restricted convex class always, keeps Davidon's bound on the condition number almost as low as his optimally conditioned update. This leads us to propose in Chapters 7 and B an algorithm which uses mostly the restricted BFGS, and occasionally other members of the restricted convex class. It is a bit superior to the pure restricted BFGS algorithm in our tests.

CHAPTER VI

PERFECT LINE SEARCH

We now begin to examine the behavior of optimization algorithms using our new quasi-Newton updates. In this chapter, we analyze the performance of various updates in a perfect line search algorithm, which has the form

Perfect Line Search Algorithm 6.0.1

given $f:R^n \to R$, f continuously differentiable, $g(x) = \nabla f(x)$

$x_0 \epsilon R^n$, $H_0 \epsilon L(R^n)$, H_0 symmetric

at each iteration

$d_i = H_i g_i$

$$x_{i+1} = x_i - \lambda_i d_i,\ \lambda_i \neq 0 \text{ chosen so that } g(x_{i+1})^T d_i = 0 \tag{6.0.2}$$

$H_{i+1} = U(H_i)$, U an updating function with range and domain the symmetric matrices $\epsilon L(R^n)$.

The distinguishing feature of such an algorithm is that λ_i is chosen by a perfect line search (6.0.2). This means that $f(x_{i+1})$ is a local minimum of $f(x)$ along the line $x(\lambda_i) = x_i - \lambda_i d_i$, so that the directional derivative of $f(x)$ along this line, $-g(x_{i+1})^T d_i$, = 0. We refer to a specific perfect line search strategy as one for which the next point x_{i+1} depends only on the last point x_i and the search direction d_i.

Examples are $x_{i+1} = \{x_i-\lambda_i d_i | \lambda_i$ solves $\min_{\lambda \in R} f(x_i-\lambda d_i)\}$ or $x_{i+1} = \{x_i-\lambda_i d_i | \lambda_i$ solves $\min_{\lambda \in R} |\lambda|$ subject to $g(x_i-\lambda d_i)^T d_i = 0\}$.

While it was once thought that perfect or near-perfect line searches were important to optimization algorithms, they are now considered a waste of function evaluations, and instead most algorithms either set $\lambda = 1$ or do a simple line search at each iteration. However, the performance of algorithms using perfect line search remains of theoretical interest, especially as we expect that once x_i is close to the minimum of f and $H_i = G(x_i)$, that the choice $\lambda_i = 1$ will be close to the value of λ_i we would obtain with a perfect line search. Interest in perfect line search algorithms was especially enhanced by Dixon's (1972a) remarkable result that if the standard DFP-BFGS class of updates (equation 1.4.7) is used in Algorithm 6.0.1, then the sequence $\{x_1,x_2,...\}$ of points produced is independent of the choices of update within the class.

Our purpose in this chapter is two-fold. First, we want to examine how algorithms using new quasi-Newton updates perform with perfect line search, including Davidon's (1975) recent algorithm, algorithms using the w-class of updates (equation 3.3.2), and algorithms using the updating parameters of Huang (1970) and Oren (1972) (equation 1.6.5). We are interested whether the use of such updates in Algorithm 6.0.1 produces different sequences $\{x_i\}$ than the DFP-BFGS updates, and whether within each class of updates the sequence $\{x_i\}$ is dependent upon the choices of the free-parameter ϕ. Secondly, we are assembling in one place and consistent notation various results of Dixon (1972a,b), Powell (1972), and Spedicato (1975) which all pertain to some aspect of this problem. Together with our results they provide a unified treatment of perfect line search quasi-Newton algorithms.

Throughout this chapter we consider algorithms which keep approximations H_i to the inverse of the Hessian $G(x_i)^{-1}$, rather than algorithms keeping approximations B_i to $G(x_i)$, because H_i is the matrix we need in analyzing Algorithm 6.0.1. In Section 6.1 we analyze how the leading one-parameter classes perform using perfect line search. First we give a version of Dixon's (1972a) result which is similar to a proof by Powell (1972), and shows that the sequence of points produced by algorithm 6.0.1 using the standard DFP-BFGS class of updates is independent of the choices of the free parameter ϕ. Then we show that if we use instead a w-class (3.3.2) where $w \notin [s,Hy]$, that the sequence of points generated by Algorithm 6.0.1 does depend on the choices of ϕ. This is a stronger version of a result of M.J.D. Powell's in Dixon (1972a). Since we have shown in Chpater 3 that our restricted DFP-BFGS update classes (3.3.3) correspond to w-classes where $w \notin [s,Hy]$, this result indicates that the choice of ϕ may be more important to such classes than to the standard DFP-BFGS.

In Section 6.2 we examine the performance of Davidon's (1975) algorithm using perfect line search. Davidon has said that his new algorithm also has the property that when perfect line searches are used, the sequence of points generated is independent of the free parameter ϕ. We show that this is because Davidon's class of updates reduces to the DFP-BFGS class in this case. Furthermore, if perfect line search is started at any iteration after the first, then the subsequent sequence of points will depend upon the subsequent choices of ϕ. This indicates that Davidon's algorithm may also be more sensitive to the choices of ϕ than an algorithm using the standard DFP-BFGS class.

Finally, in Section 6.3 we consider updates using the additional parameters of Huang (1970) and Oren (1972). Here we consolidate existing work and add new results. We extend a result of Spedicato's (1975) to show that we need only consider algorithms using one or the other of these additional parameters; and give a version of Dixon's (1972b) result that shows that the points generated by Algorithm 6.0.1 using the DFP-BFGS with these parameters will depend upon these two parameter values, but will not depend upon the choices of the parameter ϕ. However, the sequence of points produced by an algorithm using a restricted DFP-BFGS class with these parameters is dependent upon the choices of ϕ.

Section 6.1 - One-Parameter Classes with Perfect Line Search

In this section we examine the behavior of perfect line search algorithm 6.0.1 using updates from the standard DFP-BFGS class

$$\bar{H}(\phi) = H + \frac{(s-Hy)s^T+s(s-Hy)^T}{<s,y>} - \frac{<s-Hy,y>ss^T}{<s,y>^2}$$

$$+ \phi\left[Hy - \frac{s<y,Hy>}{<s,y>}\right]\left[Hy - \frac{s<y,Hy>}{<s,y>}\right]^T \tag{6.1.1}$$

and the more general w-class

$$\bar{H}(\phi) = H + \frac{(s-Hy)w^T+w(s-Hy)^T}{<w,y>} - \frac{<s-Hy,y>ww^T}{<w,y>^2}$$

$$+ \phi\left[s-Hy - \frac{w<s-Hy,y>}{<w,y>}\right]\left[s-Hy - \frac{w<s-Hy,y>}{<w,y>}\right]^T \tag{6.1.2}$$

where $s,w,y \in R^n$, H a symmetric matrix in $L(R^n)$, ϕ a real number. If H is positive definite, we know from Chapter 3 that (6.1.2) is equivalent to the restricted DFP-BFGS class (3.3.3), in which we are particularly interested in this thesis. In Theorems 6.1.1 and 6.1.2 we give a version of Dixon's (1972a) result which is similar to that in Powell (1972), and shows that when Algorithm 6.0.1 is used with the DFP-BFGS class (6.1.1) and the standard values $s = x_{i+1} - x_i$, $y = g(x_{i+1}) - g(x_i)$, then the sequence $\{x_1, x_2, \ldots\}$ of points generated is independent of the choice of ϕ_i at each iteration; and the value of H_{i+1} is dependent upon ϕ_i but independent of $\{\phi_0, \ldots, \phi_{i-1}\}$. This last fact is similar to the result of Powell (1972), and he has observed that it points to a central role of the BFGS update. Then we show that if update (6.1.2) is used instead in the same algorithm with the same definitions of s and y, (and $w \notin [s, Hy]$ so that (6.1.2) is not equivalent to (6.1.1)) that already the second search direction d_1 is in general dependent upon ϕ_0, so that x_2 will vary for different ϕ_0's. This not only shows that any restricted DFP-BFGS update produces different points than the standard DFP-BFGS class when used in Algorithm 6.0.1, but also indicates that the choices of ϕ are probably more important to this class.

We have written the ϕ-term in (6.1.1) differently than the standard (1.4.7) so that we can allow $\langle y, Hy \rangle = 0$.

<u>Theorem 6.1.1</u> (Dixon (1972a)) Let $f: R^n \to R$ be continuously differentiable; $H \in L(R^n)$ symmetric, $x \in R^n$. Define $g = g(x)$ and assume $\langle g, Hg \rangle \neq 0$. Let $\bar{x} = x - \lambda Hg$ be chosen by a perfect line search. Define $\bar{g} = g(\bar{x})$, let $s = \bar{x} - x$, $y = \bar{g} - g$, and assume $H\bar{g} \neq 0$. Let $\bar{H}(\phi) \in L(R^n)$ be chosen from the DFP-BFGS class (6.1.1), and let $\bar{d} = \bar{H}(0)\bar{g}$. Then the direction of $\bar{H}(\phi)\bar{g}$ is independent of ϕ as long as $\phi \neq -1/\langle \bar{g}, H\bar{g} \rangle$; and $\bar{H}(\phi) = \bar{H}(0) + \phi \bar{d}\bar{d}^T$.

<u>Proof</u>: Since $\bar{x}$ is chosen by a perfect line search, $\langle \bar{g}, Hg \rangle = 0$. Since

also $\langle g,Hg\rangle \neq 0$, $\bar{g} \neq g$, so that $\bar{x} \neq x$. Therefore $\lambda \neq 0$, and $\langle s,y\rangle = \lambda\langle g,Hg\rangle \neq 0$, so that (6.1.1) is well-defined.

Substituting $s = -\lambda Hg$, $y = \bar{g}-g$ into (6.1.1) and using $\langle\bar{g},Hg\rangle = 0$ gives

$$\bar{d} = \bar{H}(0)\bar{g} = H\bar{g} + Hg\,\frac{\langle\bar{g},H\bar{g}\rangle}{\langle g,Hg\rangle} \tag{6.1.3}$$

$$Hy - \frac{s\langle y,Hy\rangle}{\langle s,y\rangle} = \bar{d} \tag{6.1.4}$$

$$\bar{H}(\phi)\bar{g} = \bar{d} + \phi\bar{d}\langle\bar{g},H\bar{g}\rangle = (1+\phi\langle\bar{g},H\bar{g}\rangle)\bar{H}(0)\bar{g} \tag{6.1.5}$$

From (6.1.4), $\bar{H}(\phi) = \bar{H}(0) + \phi\bar{d}\bar{d}^T$. From (6.1.3), the new search direction $\bar{d}$ can be zero only if the nonzero vectors Hg, $H\bar{g}$ obey $Hg = \alpha H\bar{g}$, $\alpha \neq 0$. However, this would imply $0 = \langle g,H\bar{g}\rangle = \alpha\langle g,Hg\rangle$, which is impossible. Thus $\bar{d} \neq 0$, and from (6.1.5) the vector $\bar{H}(\phi)\bar{g}$ has direction $\bar{d}$ unless $\phi = -1/\langle\bar{g},\bar{H}\bar{g}\rangle$, in which case $\bar{H}(\phi)\bar{g} = 0$.

<u>Theorem 6.1.2</u> (modification of Dixon (1972a)) Let $f:R^n \to R$ be continuously differentiable; $H_0 \in L(R^n)$ symmetric, $x_0 \in R^n$. Let the sequences $\{x_i\}$, $\{H_i\}$, $i = 1,2,\ldots$ be generated from x_0, H_0 by Algorithm 6.0.1 using a specific perfect line search strategy and the BFGS update, (6.1.1) with $\phi = 0$, using at each iteration $s = x_{i+1}-x_i$, $y = g(x_{i+1})-g(x_i)$, $H = H_i$, $H_{i+1} = \bar{H}(0)$. Define $g_i = g(x_i)$. Let $\{\hat{x}_i\},\{\hat{H}_i\}$, $i = 1,2,\ldots$ also be generated from x_0, H_0 by Algorithm 6.0.1 using the same perfect line search strategy and DFP-BFGS update (6.1.1), where $s = \hat{x}_{i+1}-\hat{x}_i$, $y = g(\hat{x}_{i+1})-g(\hat{x}_i)$, $H = \hat{H}_i$ and $\hat{H}_{i+1} = \bar{H}(\phi_i)$ for any $\phi_i \neq -1/\langle g_{i+1}, H_i g_{i+1}\rangle$. Assume each $\langle g_i, H_i g_i\rangle \neq 0$ and define $d_i = H_i g_i$. Then for $i = 0,1,\ldots$, $\hat{x}_{i+1} = x_{i+1}$,

$d_{i+1} \neq 0$, $\hat{H}_{i+1}g(\hat{x}_{i+1})$ has the same direction as d_{i+1}, and $\hat{H}_{i+1} = H_{i+1}+$ $\phi_i d_{i+1}d_{i+1}^T$.

Proof: The proof is by induction. For $i = 0$, the conclusions are true by Theorem 6.1.1. Assume they are true for $i = 0,\ldots,k-1$. Then $\hat{x}_k = x_k$, $d_k = H_k g_k \neq 0$, $\hat{H}_k = H_k+\phi_{k-1}d_k d_k^T$, and $\hat{H}_k g(\hat{x}_k)$ has direction d_k. Therefore since we are using a specific perfect line search strategy, $\hat{x}_{k+1} = x_{k+1}$, and $\langle g_{k+1},H_k g_k\rangle = 0$. Since $\langle g_k,H_k g_k\rangle \neq 0$, $g_{k+1} \neq g_k$ so that $x_{k+1} \neq x_k$.

Thus H_{k+1} and $\hat{H}_{k+1}$ are both formed with $s = x_{k+1}-x_k$, $y = g_{k+1}-g_k$. Since $s = -\lambda_k H_k g_k \neq 0$, $\langle s,y\rangle = \lambda_k\langle g_k,H_k g_k\rangle$ so that both updates are well-defined. Now substituting $s = -\lambda_k H_k g_k$ and $y = g_{k+1}-g_k$ into (6.1.1) and using $\langle g_{k+1},H_k g_k\rangle = 0$ yields

$$d_{k+1} = H_{k+1}g_{k+1} = H_k g_{k+1} + H_k g_k \frac{\langle g_{k+1},H_k g_{k+1}\rangle}{\langle g_k,H_k g_k\rangle}$$

which is nonzero by the same reasoning as was used in the proof of Theorem 6.1.1. Substituting also $\hat{H}_k = H_k+\phi_{k-1}d_k d_k^T$ and $d_k = -s_k/\lambda_k$ yields, after some cancellation

$$\hat{H}_{k+1} = H_{k+1}+\phi_k d_{k+1}d_{k+1}^T .$$

Therefore $\hat{H}_{k+1}g_{k+1} = H_{k+1}g_{k+1}+\phi_k d_{k+1}\langle d_{k+1},g_{k+1}\rangle = d_{k+1}+\phi_k d_{k+1}\langle g_{k+1},H_k g_{k+1}\rangle$ which has direction d_{k+1} since $\phi_k \neq -1/\langle g_{k+1},H_k g_{k+1}\rangle$.

So, as long as we exclude the choice of ϕ_i for which $H_{i+1}(\phi_i)g(x_{i+1}) = 0$, the choices of ϕ_i do not affect the performance of perfect line search algorithm 6.0.1 using the DFP-BFGS updates (6.1.1). They do affect the matrices H_{i+1}, and the conditioning of these matrices may influence the performance of Algorithm 6.0.1 in finite precision arithmetic. In our parameterization, the matrix $H_{i+1}(\phi_i)$ is also independent of the values of ϕ_j, $j<i$. Note however that in the standard parameterization $H_{i+1}(\overline{\phi}_i)$ (1.4.7) (where $<y,Hy> \neq 0$ is required), $\overline{\phi}_i = -\phi_i<y_i,H_iy_i>$, and H_i depends upon previous $\overline{\phi}_j$'s. Therefore the only $\overline{\phi}_i$ for which $H_{i+1}(\overline{\phi}_i)$ will be independent of $\overline{\phi}_j$, $j<i$, is the BFGS update, $\overline{\phi}_i = 0$. This indicates that the BFGS has a unique property among those updates from the standard DFP-BFGS class (1.4.7) of erasing the effects of past updates, and so may be a stabilizing influence on a quasi-Newton algorithm. Powell (1972) notes this property, and suggests that it may contribute to the success of the BFGS algorithm.

We now consider perfect line search algorithms using w-class update (6.1.2).

<u>Theorem 6.1.3</u> Let $f:R^n \to R$ be continuously differentiable; $H \in L(R^n)$ symmetric, $x \in R^n$. Define $g = g(x)$, and assume $<g,Hg> \neq 0$. Let $\overline{x} = x-\lambda Hg$ be chosen by a perfect line search. Define $\overline{g} = g(\overline{x})$, let $s = \overline{x}-x$, $y = \overline{g}-g$, and assume $<\overline{g},H\overline{g}> \neq 0$, $s \neq Hy$. Let $\overline{H} \in L(R^n)$ be chosen from the one-parameter w-class (6.1.2), where ϕ is a real parameter, $w \in R^n$ such that $<w,y> \neq 0$ and $w \notin [s,Hy]$. Let $u = s-Hy - \frac{w<s-Hy,y>}{<w,y>}$. Then for any two different values ϕ_1, ϕ_2 of ϕ in (6.1.2), $\overline{H}(\phi_1)\overline{g}$ and $\overline{H}(\phi_2)\overline{g}$ are linearly independent unless either $\lambda = 1$ or $<u,\overline{g}> = 0$. If H is positive definite, $<u,\overline{g}> = 0$ if and only if $\hat{s}$ and $H\hat{y}$ are linearly dependent, where $\hat{s} = Qs$, $\hat{y} = Q^Ty$, Q is the matrix projecting onto $[s-Hy,w]$ in the metric H^{-1}.

<u>Proof</u>: Since $\overline{x}$ was chosen by perfect line search, $<\overline{g},Hg> = 0$, and as was shown in the proof of Theorem 6.1.1, $\overline{x} \neq x$. Thus λ, $s \neq 0$.

We treat separately the cases $<s-Hy,y> = 0$ and $\neq 0$. In the former case, $\lambda \neq 1$ because $<s-Hy,y> = -<\bar{g},H\bar{g}>$ when $\lambda = 1$, and $<\bar{g},H\bar{g}> \neq 0$; and $<u,\bar{g}> \neq 0$ because $<u,\bar{g}> = -<\bar{g},H\bar{g}>$ when $<s-Hy,y> = 0$. Also by substituting $s = -\lambda Hg$ and $y = \bar{g}-g$ into (6.1.2) and using $<\bar{g},Hg> = 0$ we get

$$\bar{H}(\phi)\bar{g} = v_1 - \phi(s-Hy)<\bar{g},H\bar{g}>$$

$$v_1 = H\bar{g} + (s-Hy)\frac{<w,\bar{g}>}{<w,y>} - w\frac{<\bar{g},H\bar{g}>}{<w,y>}$$

Since $<\bar{g},H\bar{g}> \neq 0$ and $H\bar{g} = Hy-s/\lambda$, $v_1 \notin [s,Hy]$ and v_1, $s-Hy$ are linearly independent. Thus when $<s-Hy,y> = 0$, $\bar{H}(\phi)\bar{g} \neq 0$, and has a different direction for each value of ϕ.

Now assume $<s-Hy,y> \neq 0$. Then straightforward algebra shows that (6.1.2) can be rewritten as

$$\bar{H}(\phi) = H + \frac{(s-Hy)(s-Hy)^T}{<s-Hy,y>} + \left\{\phi - \frac{1}{<s-Hy,y>}\right\}uu^T .$$

Therefore

$$\bar{H}(\phi)\bar{g} = v_2 + \left\{\phi - \frac{1}{s-Hy,y}\right\} u<u,\bar{g}>$$

$$v_2 = \frac{(\lambda-1)}{<s-Hy,y>}[H\bar{g}<g,Hg> + Hg<\bar{g},H\bar{g}>] \qquad (6.1.6)$$

Since $v_2 \epsilon [s,Hy]$ (as $H\bar{g} = Hy-s/\lambda$, $Hg = -s/\lambda$) and $u \neq 0$, $u \notin [s-Hy]$, $\bar{H}(\phi)\bar{g}$ given by (6.1.6) will have a different direction for each ϕ unless $v_2 = 0$ or $<u,\bar{g}> = 0$. The former can happen only if $\lambda = 1$, for in the proof of Theorem 6.1.1 we showed that $Hg<\bar{g},H\bar{g}> + H\bar{g}<g,Hg> \neq 0$ when $<\bar{g},Hg> = 0$ $\neq <g,Hg>$ and $H\bar{g} \neq 0$.

Finally, if H is positive definite, we show that $\langle u,\bar{g}\rangle = 0$ if and only if $\hat{s}$ and $H\hat{y}$ are linearly dependent. By Lemma 2.3.2, Q is well-defined and $Q^2 = Q$, $HQ^T = QH$, $Qu = u$. Since $\langle u,y\rangle = 0$, $0 = \langle Qu,y\rangle = \langle u,Q^Ty\rangle = \langle u,\hat{y}\rangle$. If $\langle u,\bar{g}\rangle = 0$, then $\langle u,g\rangle = \langle u,\bar{g}\rangle - \langle u,y\rangle = 0$, and $0 = -\lambda\langle u,g\rangle = \langle u,H^{-1}s\rangle = \langle Qu,H^{-1}s\rangle = \langle u,Q^TH^{-1}s\rangle = \langle u,H^{-1}Qs\rangle = \langle u,H^{-1}\hat{s}\rangle$. Then since u, $H\hat{y}$ and $\hat{s}$ are all in $[s-Hy,w]$ and $u \neq 0$, $H\hat{y}$ and $\hat{s}$ must be linearly dependent. ($\hat{s}$ can be 0; $H\hat{y}$ cannot since $\langle w,y\rangle \neq 0$ implies $\hat{y} \neq 0$.) Conversely if $\hat{s}$ and $H\hat{y}$ are linearly dependent, then $\langle u,y\rangle = \langle u,\hat{y}\rangle = 0$ implies $\langle u,H^{-1}\hat{s}\rangle = 0 = -\lambda\langle u,g\rangle$. Thus $\langle u,g\rangle = 0$ since $\lambda \neq 0$, and $\langle u,\bar{g}\rangle = \langle u,g\rangle + \langle u,y\rangle = 0$.

So the sequences of points produced by a perfect line search algorithm using a w-class of updates will in general depend upon the choices of the free parameter ϕ. One thing this shows us is that such an algorithm with $w \notin [s,Hy]$ will in general generate different points than a DFP-BFGS perfect line search algorithm - because any DFP-BFGS update produces the same points as the SR1 update, choice $\phi = 1/\langle s-Hy,y\rangle$ in w-class (6.1.2) (or (6.1.1)), which by Theorem 6.1.3 generates different points than any other choice of ϕ in (6.1.2). (It can easily be shown that the DFP-BFGS class and a w-class with $w \notin [s,Hy]$ also produce different points if $\langle s-Hy,y\rangle = 0$.) This result also follows from a theorem in Dixon (1972a) which is attributed to M.J.D. Powell.

From Theorem 3.3.2, we know that any restricted DFP-BFGS class (3.3.3) for which $\hat{s} \neq s$ and $\hat{y} \neq y$ corresponds to a w-class (6.1.2) with $w \notin [s,Hy]$. Therefore unless $\hat{s}$ and $H\hat{y}$ are linearly dependent - in which case all of the restricted DFP-BFGS class reduces to the SR1 update - Theorem 6.1.3 shows that the points generated by a perfect line search algorithm using a restricted DFP-BFGS class will depend upon the choices of ϕ unless $\lambda = 1$. This suggests that the choice of ϕ may be more important in actual algorithms using a restricted DFP-BFGS class than in the same algorithm using the standard DFP-BFGS.

In Section 6.2 we use the theorems of this section to analyze the behavior of Davidon's (1975) algorithm using perfect line search.

Section 6.2 - Davidon's Recent Algorithm with Perfect Line Search

In this section we examine the behavior of Davidon's (1975) algorithm if perfect line searches are used. From our discussion in Chapter 2, we know that Davidon's algorithm uses a w-class of updates. Therefore from our work in Section 6.1 we would expect that the points generated by the algorithm using perfect line search would depend upon the choices of ϕ - but Davidon (1975) says this is not so. We show in Lemma 6.2.1 that this is because with perfect line search, Davidon's initial choice $w_0 = s_0$, and his updating formula for w_i,

$$w_{i+1} = w_i\langle s_i - H_i y_i, y_i\rangle - (s_i - H_i y_i)\langle w_i, y_i\rangle \tag{6.2.1}$$

each w_i is (some multiple of) s_i, so that Davidon's updating class becomes the standard DFP-BFGS. However Theorem 6.2.2 shows that for any choice $w_0 \notin [s_0, H_0 y_0]$, the points produced by Davidon's algorithm will in general depend upon the choices of ϕ. This shows that if perfect line search is started at any iteration after the first, that the subsequent points generated will depend on the subsequent choices of ϕ. It also suggests for actual implementations, that even near the solution where λ is likely to approximate the perfect line search value, that the sequence of points will depend upon the values of ϕ.

Lemma 6.2.1 Let $f:R^n \to R$ be continuously differentiable; $H_0 \in L(R^n)$ symmetric, $x_0 \in R^n$, $w_0 \in R^n$ nonzero. Let the sequences $\{x_i\}$, $\{H_i\}$, $\{w_i\}$, $i = 1,2,\ldots$ be generated from x_0, H_0, w_0 by perfect line search algorithm 6.0.1, using updates from the w-class

$$H_{i+1}(\phi_i) = H_i + \frac{(s_i - H_i y_i)w_i^T + w_i(s_i - H_i y_i)^T}{\langle w_i, y_i\rangle} - \frac{\langle s_i - H_i y_i, y_i\rangle w_i w_i^T}{\langle w_i, y_i\rangle^2}$$

$$+ \phi_i \left[s_i - H_i y_i - \frac{w_i\langle s_i - H_i y_i, y_i\rangle}{\langle w_i, y_i\rangle}\right]\left[s_i - H_i y_i - \frac{w_i\langle s_i - H_i y_i, y_i\rangle}{\langle w_i, y_i\rangle}\right]^T \tag{6.2.2}$$

where $s_i = x_{i+1} - x_i$, $y_i = g(x_{i+1}) - g(x_i)$, ϕ_i is a real parameter, and w_i is generated by formula (6.2.1). Define $g_i = g(x_i)$, and assume $\langle g_i, H_i g_i\rangle \neq 0$, $i = 0,1,\ldots$, and $\langle w_0, y_0\rangle \neq 0$. Let $V = \{[s_0, H_0 y_0] \sim [s_0 - H_0 y_0]\}$. If $w_0 \in V$, then for $i = 0,1,\ldots$, the w-class (6.2.2) is equivalent to the DFP-BFGS class (6.1.1), and $w_{i+1} = \alpha_{i+1} s_{i+1}$, α_{i+1} a nonzero real number.

Proof: The proof is by induction. For $i = 1$, update (6.2.2) is well-defined because $\langle w_0, y_0\rangle \neq 0$; and since also $w_0 \in V$, from Theorem 3.3.2, (6.2.2) is equivalent to the DFP-BFGS class (6.1.1). Since $w_0 \in V$, $w_0 = \beta s_0 + \gamma H_0 y_0$, for some real numbers β, γ such that $\beta + \gamma \neq 0$; and substituting this into (6.2.1) along with $s_0 = -\lambda_0 H_0 g_0$, $y_0 = g_1 - g_0$, and using $\langle g_1, H_0 g_0\rangle = 0$ yields $w_1 = (\beta+\gamma)\lambda_0 v$, $v = H_0 g_1 \langle g_0, H_0 g_0\rangle + H_0 g_0 \langle g_1, H_0 g_1\rangle$. From the proof of Theorem 6.1.1, $\lambda_0 \neq 0$ due to $\langle g_0, H_0 g_0\rangle \neq 0$, v is nonzero and has the same direction as the next search direction $H_1 g_1 = -s_1/\lambda_1$, and $\lambda_1 \neq 0$ due to $\langle g_1, H_1 g_1\rangle \neq 0$. Therefore $w_1 = \alpha_1 s_1$ for some nonzero number α_1.

If we assume the theorem is true for $i = 0,1,\ldots,k-1$, then $w_k = \alpha_k s_k$, $\alpha_k \neq 0$, and since λ_k is also nonzero, $\langle w_k, y_k\rangle = \alpha_k \langle s_k, y_k\rangle = \alpha_k \lambda_k \langle g_k, H_k g_k\rangle \neq 0$. Thus for $i = k$, update (6.2.2) is well-defined, and clearly is equivalent to the DFP-BFGS class (6.1.1). The remainder of the induction step proceeds exactly like the case $i = 0$.

Using the terminology we used in discussing Davidon's (1975) algorithm in Chapter 2, Lemma 6.2.1 can be interpreted as saying that when perfect line searches are used, s_i and $H_i y_i$ are orthogonal to the restricting subspace $(Z_H)_i$ at each iteration. This is reasonable, because $(Z_H)_i$ is chosen so that the algorithm will have finite termination on quadratic problems; and it has long been known (Fletcher and Powell (1963), Broyden (1970)) that the standard DFP-BFGS class has this property with perfect line search.

Theorem 6.2.2 fully characterizes the behavior of Davidon's algorithm using perfect line search.

<u>Theorem 6.2.2</u> Let $f:R^n \to R$ be continuously differentiable, $H_0 \in L(R^n)$ symmetric, $x_0 \in R^n$, $w_0 \in R^n$ nonzero, $\{\phi_0, \phi_1, \ldots\}$ real numbers. Let the sequences $\{x_i\}$, $\{H_i\}$, $\{w_i\}$, $i = 1,2,\ldots$ be generated from x_0, H_0, w_0 by Algorithm 6.0.1 using a specific perfect line search strategy, forming $H_{i+1}(\phi_i)$ by update (6.2.2) where $s_i = x_{i+1} - x_i$ and $y_i = g(x_{i+1}) - g(x_i)$, and forming w_{i+1} by equation (6.2.1). Define $g_i = g(x_i)$, and assume $\langle g_i, H_i g_i\rangle \neq 0$, $\phi_i \neq -1/\langle g_{i+1}, H_i g_{i+1}\rangle$, $\langle w_i, y_i\rangle \neq 0$, $i = 0,1,\ldots$ Then if $w_0 \in \{[s_0, H_0 y_0] \sim [s_0 - H_0 y_0]\}$, the sequence of points $\{x_i\}$ is independent of the sequence of parameters $\{\phi_i\}$. If $w_0 \notin [s_0, H_0 y_0]$, then $\{x_i\}$ is independent of $\{\phi_i\}$ if and only if either $\lambda_0 = 1$, or $\langle w_1, g_1\rangle = 0$ and either

$H_0g_2 = 0$ or $w_0 \epsilon S$, where $S = \{[s_0-H_0y_0, H_0g_2] \sim \{[s_0-H_0y_0] \wedge [H_0g_2]\}\}$.

Proof: If $w_0 \epsilon \{[s_0, H_0y_0] \sim [s_0-H_0y_0]\}$, then by Lemma 6.2.1, the updating class (6.2.2) is equivalent to the DFP-BFGS class (6.1.1) at each iteration. Thus by Lemma 6.1.2, the sequence $\{x_i\}$ is independent of the sequence $\{\phi_i\}$.

If $w_0 \notin [s_0, H_0y_0]$, we know from Theorem 6.1.3 that H_1g_1 is independent of ϕ_1 if and only if $\lambda_0 = 1$ or $<w_1, g_1> = 0$ (since u as defined in Theorem 6.1.3 $= -w_1/<w_0, y_0>$. In these cases let $\hat{H}_1$, $\hat{H}_2$ be the matrices produced by update (6.2.2) using $\phi = 0$ both times. If $\lambda_0 = 1$, then the proof of Theorem 6.1.3 shows that $w_1 \epsilon [s_1]$; so since $H_1 = \hat{H}_1 + \phi_0 w_1 w_1^T / <w_0, y_0>^2$, we are in the same position as if we had started with $w_0 = s_0$, and hence the remaining x_i's are independent of the remaining ϕ_i's.

Now suppose $\lambda_0 \neq 1$ and $<w_1, g_1> = 0$. From (6.2.2) (after some cancellation), $H_2 = \hat{H}_2 + \phi_1 u_1 u_1^T$, $u_1 = s_1 - \hat{H}_1 y_1 - w_1(<s_1 - \hat{H}_1 y_1, y_1>/<w_1, y_1>) = -w_2/<w_1, y_1>$. This is the same situation as after one iteration. Therefore if $w_1 \epsilon \{[s_1, H_1y_1] \sim [s_1 - H_1y_1]\}$ it is proven above that $\{x_i\}$ is independent of $\{\phi_i\}$. If $w_1 \notin [s_1, H_1y_1]$, Theorem 6.1.3 shows that H_2g_2 is independent of ϕ_1 if and only if $\lambda_1 = 1$ or $<u_1, g_2> = 0$. We show that since $<w_1, g_1> = 0$, $w_1 \epsilon \{[s_1, H_1y_1] \sim [s_1 - H_1y_1]\}$ if and only if $H_0g_2 = 0$ or $w_0 \epsilon S$, while $\lambda_1 = 1$ and $<u_1, g_2> = 0$ are both impossible. This final part of the proof is quite long and not illuminating.

If $w_1 \epsilon [s_1, H_1y_1] = [H_1g_1, H_1g_2]$, then since $<w_1, g_1> = <H_1g_2, g_1> = 0$ and $<H_1g_1, g_1> \neq 0$, $w_1 \epsilon [H_1g_2]$ ($[H_1g_2] = [s_1 - H_1y_1]$ only if $\lambda_1 = 1$ which we show below is impossible). From (6.2.2), $H_1g_2 = v + \alpha w_1$, where

$$v = H_0g_2 + z(<z, g_2>/<z, y_0>), \tag{6.2.3}$$

$z = s_0 - H_0y_0$, and $\alpha = (\phi_0 - (1/<s_0 - H_0y_0, y_0>))<w_1, g_2>/<w_0, y_0>^2$. (It was shown in the proof of Theorem (6.1.3) that $\langle z,y_o\rangle \neq 0$ if $\langle w_1,g_1\rangle = 0$.) Thus $w_1 \epsilon [H_1 g_2]$ if and only if $v = 0$ or $v \epsilon [w_1]$. For v to be 0, $H_o g_2$ must equal kz for some number k, and substituting this into (6.2.3) gives $v = 0$ if and only if $\langle g_o,H_o g_2\rangle = 0$. If $k = 0$, $H_o g_2 = 0$ and $v = 0$, $w_1 \epsilon [H_1 g_2]$. If $k \neq 0$, $\langle g_o,H_o g_2\rangle = (1-\lambda_o)\langle g_o,H_o g_o\rangle/k \neq 0$ and $v \neq 0$. If this case we require $v \epsilon [w_1]$, and since $w_1 = z\langle w_o,y_o\rangle - w_o\langle z,y_o\rangle$, (6.2.2) shows that $w_o \epsilon [z,H_o g_2]$ is necessary for $v \epsilon [w_1]$. Since $w_o \neq z$, assume without loss of generality $w_o = z+\alpha H_o g_2$, $\alpha \neq 0$ because $w_1 \neq 0$. Substituting this into $\langle w_1,g_1\rangle = 0$ gives

$$\langle z,g_1\rangle\langle H_o g_2,g_o\rangle = \langle z,g_o\rangle\langle H_o g_2,g_1\rangle \tag{6.2.4}$$

while from $\langle H_1 g_2,g_1\rangle = 0 = \langle v,g_1\rangle$ we have

$$(1-\lambda_o)\langle z,g_1\rangle\langle H_o g_2,g_o\rangle = \langle z,g_o\rangle\langle H_o g_2,g_1\rangle \tag{6.2.5}$$

Subtracting (6.2.5) from (6.2.4) yields $\lambda_o\langle z,g_1\rangle\langle H_o g_2,g_o\rangle = 0$, which implies $\langle H_o g_2,g_o\rangle = 0$ since $\langle g_1,H_o g_1\rangle \neq 0$ due to $\langle g_1,H_1 g_1\rangle \neq 0$. Thus $w_1 = -\alpha H_o g_2\langle z,y_o\rangle$, and from (6.2.4), $\langle H_o g_2,g_1\rangle = 0$ so that $\langle z,g_2\rangle = 0$, $v = H_o g_2$, $v \epsilon [w_1]$. Therefore $w_1 \epsilon \{[s_1,H_1 y_1] \sim [s_1-H_1 y_1]\}$ if and only if $w_o \epsilon S$.

To show that $\lambda_1 = 1$ or $\langle u_1,g_2\rangle = 0$ are impossible, we recall that the techniques of proof in Theorem 6.1.3 show that if $\lambda_1 = 1$, then $H_2 g_2 = u_1\langle u_1,g_2\rangle(\phi_1-(1/\langle s_1-\hat{H}_1 y_1,y_1\rangle))$. Since $\langle u_1,y_1\rangle = 0$ and $\langle w_1,g_1\rangle = 0$, $\langle u_1,g_2\rangle = \langle u_1,g_1\rangle = \langle s_1-\hat{H}_1 y_1,y_1\rangle = (1-\lambda_1)\langle g_1,H_1 g_1\rangle$. Thus if $\lambda_1 = 1$, $\langle u_1,g_2\rangle = 0$ and $H_2 g_2 = 0$ for all ϕ_1, which contradicts our assumptions. If $\lambda_1 \neq 1$, then $\langle u_1,g_2\rangle \neq 0$ since $\langle g_1,H_1 g_1\rangle \neq 0$. Therefore $\langle w_1,g_1\rangle = 0$ implies $\lambda_1 \neq 1$ and $\langle u_1,g_2\rangle \neq 0$.

Powell (1976b) has shown that the finite termination of Davidon's

(1975) algorithm on quadratics is independent of the choice of w_0. Theorem 6.2.2 shows however that the property that the points produced by the algorithm are independent of the choices of ϕ_i when perfect line search is used is true only with Davidon's choice $w_0 = s_0$ (or equivalently, $w_0 \epsilon \{[s_0, H_0 y_0] \sim [s_0 - H_0 y_0]\}$. This means that if perfect line search were started in Davidon's algorithm at any iteration after the first, that since then w_i would not in general be in $[s_i, H_i y_i]$ even if $w_0 = s_0$, the subsequent points generated by the algorithm would depend upon the subsequent choices of ϕ. This differs from an algorithm using the DFP-BFGS class, where if perfect line search is started at any iteration, then the remaining points will depend on the previous but not the remaining choices of ϕ.

In particular, we think the real applicability of perfect line search theorems to present-day algorithms is as a guide to the behavior of these algorithms near the solution, where it is likely that the λ's produced by direct prediction or a simple line search will approximate the perfect line search values of λ. Thus the results of this section indicate that near the solution, since the values of w_i have little chance of being in $[s_i, H_i y_i]$, that the performance of Davidon's (1975) algorithm is likely to be more influenced by the choices of ϕ than is an algorithm using the standard DFP-BFGS class. Therefore we think that Davidon's (1975) statement, that his algorithm shares with the DFP-BFGS the property that the points generated by perfect line search are independent of ϕ, does not completely characterize the performance of his algorithm using perfect line search -- for the analysis of this section shows that only the DFP-BFGS class has the property that if perfect line search is started after the first iteration, then the points generated are independent of subsequent choices of ϕ, and this leads to differing conclusions about the performances of the algorithms as actually implemented.

Section 6.3 - The Huang and Oren Updates with Perfect Line Search

We complete our analysis of quasi-Newton algorithms using perfect line search by considering the w-class of updates using the parameters of Huang (1970) and Oren (1972),

$$\bar{H}(\phi,\rho,\gamma) = \gamma H + \frac{(\rho s-\gamma Hy)w^T + w(\rho s-\gamma Hy)^T}{\langle w,y\rangle} - \frac{\langle \rho s-\gamma Hy, y\rangle ww^T}{\langle w,y\rangle^2}$$

$$+ \phi\left[\rho s-\gamma Hy - \frac{w\langle \rho s-\gamma Hy,y\rangle}{\langle w,y\rangle}\right]\left[\rho s-\gamma Hy - \frac{w\langle \rho s-\gamma Hy,y\rangle}{\langle w,y\rangle}\right]^T \qquad (6.3.1)$$

where $s,y\in R^n$ are $\bar{x} - x$ and $g(\bar{x})-g(x)$ respectively, $w\in R^n$ is such that $\langle w,y\rangle \neq 0$, $H\in L(R^n)$ is symmetric, ϕ is a real parameter, and ρ,γ are non-zero real numbers. The effect of the parameter ρ is to change the quasi-Newton equation to $\bar{H}y = \rho s$, which may be done to keep $\bar{H}$ non-singular or to incorporate additional derivative information. The parameter γ is used to make a simple rank n change to H, $\hat{H} = \gamma H$, in addition to the rank-two update. This may be especially useful in the initial iterations of an algorithm if we don't have a good initial approximation H_0. Since we would like $\gamma Hy = s$, a possible heuristic is $\gamma = \langle s,y\rangle/\langle y,Hy\rangle$, so that $\langle \gamma Hy,y\rangle = \langle s,y\rangle$.

In Theorem 6.3.1, we show that with perfect line searches, an algorithm using update (6.3.1) with parameter values (ϕ_i,ρ_i,γ_i) will produce the same sequence of points as an algorithm using parameter values $(\phi_i\gamma_i\alpha_i,\ \rho_i/\gamma_i\alpha_i,\ 1)$, $\alpha_i = \prod_{j=0}^{i-1}\gamma_j$ or an algorithm using $(\phi_i\rho_i,\ 1,\ \gamma_i\rho_{i-1}/\rho_i)$. This is an extension of a theorem of Spedicato (1975), and shows that we can analyze the behavior of algorithms using Huang-Oren update (6.3.1) and perfect line search by considering updates which use only the Huang, and not the Oren, parameter. We do this for the remainder of the section. In Theorem 6.3.2 we show that the points produced by an algorithm using update (6.3.1) with $w = s$ are dependent upon the values of ρ; but in Theorem 6.3.3 we show they are independent of the values of ϕ. Both these theorems are modifications of a result in Dixon (1972b). Finally in Theorem 6.3.4 we show that the points produced by a perfect line search algorithm using update (6.3.1) with $w \notin [s,Hy]$ depend in general upon the values of ϕ.

<u>Theorem 6.3.1</u> (modification of Spedicato (1975)). Let $f:R^n \to R$ be continuously differentiable; $H_0 \varepsilon L(R^n)$ symmetric; $x_0 \varepsilon R^n$, $\{w_0, w_1, \ldots\}$ a sequence of nonzero n-vectors, $\{\phi_i\}$, $\{\rho_i\}$, $\{\gamma_i\}$, $i = 0,1,\ldots$ sequences of real numbers, $\rho_i, \gamma_i \neq 0$ for all i. Let the sequences $\{x_i\}$, $\{H_i\}$, $i = 1,2,\ldots$ be generated from x_0, H_0 by Algorithm 6.0.1 using a specific perfect line search strategy and update (6.3.1), using at each iteration $s = x_{i+1}-x_i$, $y = g(x_{i+1})-g(x_i)$, $w = w_i$, $H = H_i$, $H_{i+1} = \overline{H}(\phi_i, \rho_i, \gamma_i)$. Let the sequences $\{\overline{x}_i\}$, $\{\overline{H}_i\}$ and $\{\hat{x}_i\}$, $\{\hat{H}_i\}$ also be generated from x_0, H_0 by Algorithm 6.0.1 using the same perfect line search strategy, and update (6.3.1) using at each iteration $s = \overline{x}_{i+1}-\overline{x}_i$, $y = g(\overline{x}_{i+1})-g(\overline{x}_i)$, $w = w_i$, $H = \overline{H}_i$, $\overline{H}_{i+1} = \overline{H}(\phi_i\gamma_i\alpha_i, \rho_i/\gamma_i\alpha_i, 1)$, and $s = \hat{x}_{i+1}-\hat{x}_i$, $y = g(\hat{x}_{i+1})-g(\hat{x}_i)$, $w = w_i$, $H = \hat{H}_i$, $\hat{H}_{i+1} = \overline{H}(\phi_i\rho_i, 1, \gamma_i\beta_i/\rho_i)$ respectively where $\alpha_i = \prod_{j=0}^{i-1} \gamma_i$ for $i \geq 1$, $\alpha_0 = 1$ and $\beta_i = \rho_{i-1}$ for $i \geq 1$, $\beta_0 = 1$. Assume $\langle w_i, y_i \rangle \neq 0$, $i = 0,1,\ldots$. Then for $i = 0,1,\ldots$, $x_i = \overline{x}_i = \hat{x}_i$ and $H_i = \alpha_i\overline{H}_i = \beta_i\hat{H}_i$.

<u>Proof:</u> The proof is by induction on i. For $i = 0$ it is trivially true,

since all three sequences start from x_0, H_0. Assume it is true for $i = k$. Then $x_k = \bar{x}_k = \hat{x}_k$, and $H_k = \alpha_k \bar{H}_k = \beta_k \hat{H}_k$, $\alpha_k, \beta_k \neq 0$. Thus the next search direction $H_k g(x_k)$ is the same in all three cases, and since we are using a specific line search strategy, $x_{k+1} = \bar{x}_{k+1} = \hat{x}_{k+1}$. Therefore s and y are the same in all three updates. Now consider making H a parameter in the left hand side of update (6.3.1), and we see that $\bar{H}(\phi,\rho,\gamma,\sigma H) = \bar{H}(\phi,\rho,\sigma\gamma,H)$ for any constant σ. Also by inspection, $\bar{H}(\phi,\rho,\gamma,H) = \rho\bar{H}(\phi\rho, 1, \gamma/\rho, H)$, and $\bar{H}(\phi,\rho,\gamma,H) = \gamma\bar{H}(\phi\gamma, \rho/\gamma, 1, H)$. Therefore $H_{k+1} = \bar{H}(\phi_k,\rho_k,\gamma_k,H_k) = \bar{H}(\phi_k,\rho_k,\gamma_k,\alpha_k\bar{H}_k) = \bar{H}(\phi_k,\rho_k,\gamma_k\alpha_k,\bar{H}_k) = \gamma_k\alpha_k\bar{H}(\phi_k\gamma_k\alpha_k, \rho_k/\gamma_k\alpha_k, 1, \bar{H}_k) = \gamma_k\alpha_k\bar{H}_{k+1} = \alpha_{k+1}\bar{H}_{k+1}$; and $H_{k+1} = \bar{H}(\phi_k,\rho_k,\gamma_k,\beta_k\hat{H}_k) = \bar{H}(\phi_k,\rho_k,\gamma_k\beta_k,\hat{H}_k) = \rho_k\bar{H}(\phi_k\rho_k, 1, \gamma_k\beta_k/\rho_k, \hat{H}_k) = \rho_k\hat{H}_{k+1} = \beta_{k+1}\hat{H}_{k+1}$, which completes the induction.

Thus we see that we can characterize the behavior of perfect line search algorithms using the Huang and Oren parameters by considering updates with only the Huang (or only the Oren) parameter. Therefore in the remainder of this section we consider for simplicity the special case of update (6.3.1) where $\gamma = 1$. Next we analyze the behavior of a perfect line search algorithm using (6.3.1) with $w = s$, the standard DFP-BFGS class modified to include the Huang and Oren paremters. We see that the sequence of points generated is still independent of the values of ϕ, but dependent upon the values of ρ (and γ).

<u>Theorem 6.3.2</u> (modification of Dixon (1972b) Let $f:R^n \to R$ be continuously differentiable; $H_0 \in L(R^n)$ symmetric, $x_0 \in R^n$. Let x_1, H_1, x_2, H_2 be generated from x_0, H_0 by Algorithm 6.0.1 using a specific perfect line search strategy and update

$$H_{i+1}(\rho_i) = H_i + \frac{(\rho_i s_i - H_i y_i)s_i^T + s_i(\rho_i s_i - H_i y_i)^T}{\langle s_i, y_i\rangle}$$

$$- \frac{\langle \rho_i s_i - H_i y_i, y_i\rangle s_i s_i^T}{\langle s_i, y_i\rangle^2} \tag{6.3.2}$$

where $s_i = x_{i+1} - x_i$, $y_i = g(x_{i+1}) - g(x_i)$, ρ_0, ρ_1 nonzero real numbers. Let $\hat{x}_1$, $\hat{H}_1$, $\hat{x}_2$, $\hat{H}_2$ be generated from x_0, H_0 by Algorithm 6.0.1 using the same perfect line search strategy and update (6.3.2) with $s_i = \hat{x}_{i+1} - \hat{x}_i$, $y_i = g(\hat{x}_{i+1}) - g(\hat{x}_i)$ and $\hat{\rho}_0$, $\hat{\rho}_1$ any nonzero real numbers such that $\hat{\rho}_0 \neq \rho_0$. Define $g_i = g(x_i)$ and assume $\langle g_i, H_i g_i\rangle \neq 0$, $i = 0,1$. Then $H_1 g_1$ and $\hat{H}_1 g_1$ have the same direction, but $H_2 g_2$ and $\hat{H}_2 g_2$ do not unless $\langle g_2, H_0 g_0\rangle = 0$ or $H_1 g_0$ and $H_1 g_2$ are linearly dependent.

<u>Proof</u>: Since the first search direction is $H_0 g_0$ in both cases, $\hat{x}_1 = x_1$ and by perfect line search $\langle s_0, g_1\rangle = 0$. Since $\langle g_0, H_0 g_0\rangle \neq 0$, $s_0 = -\lambda_0 H_0 g_0 \neq 0$ and $\langle s_0, y_0\rangle = \lambda_0 \langle g_0, H_0 g_0\rangle \neq 0$ so that H_1, $\hat{H}_1$ are well defined. By (6.3.2), $\hat{H}_1 = H_1 + ((\hat{\rho}_0 - \rho_0)/\langle s_0, y_0\rangle) s_0 s_0^T$, so $\hat{H}_1 g_1 = H_1 g_1$. Thus $\hat{x}_2 = x_2$ and by perfect line search $\langle s_1, g_2\rangle = 0$. Let $\alpha = (\hat{\rho}_0 - \rho_0)/\langle s_0, y_0\rangle$; since $\hat{\rho}_0 \neq \rho_0$, $\alpha \neq 0$. By $\langle g_1, H_1 g_1\rangle \neq 0$, $s_1 = -\lambda_1 H_1 g_1 \neq 0$ and $\langle s_1, y_1\rangle \neq 0$, so H_2, $\hat{H}_2$ are well-defined. By (6.3.2), $\hat{H}_2 = H_2 + ((\hat{\rho}_1 - \rho_1)/\langle s_1, y_1\rangle) s_1 s_1^T + \alpha v v^T$, $v = s_0 - (\langle s_0, y_1\rangle / \langle s_1, y_1\rangle) s_1$, so $\hat{H}_2 g_2 = H_2 g_2 + \alpha v \langle s_0, g_2\rangle$. Therefore $\hat{H}_2 g_2$ and $H_2 g_2$ have the same direction if and only if either $\langle s_0, g_2\rangle = 0$, which is equivalent to $\langle H_0 g_0, g_2\rangle = 0$, or $H_2 g_2$ and v have the same direction. Since $H_2 g_2 = H_1 g_2 + H_1 g_1 (\langle g_2, H_1 g_2\rangle / \langle g_1, H_1 g_1\rangle)$, $v = -\lambda_0 (H_0 g_0 + H_1 g_1 (\langle g_2, H_0 g_0\rangle / \langle g_1, H_1 g_1\rangle))$, and $H_1 g_1$, $H_1 g_2$ are linearly independent, $v \in [H_2 g_2]$ if and only if $H_0 g_0 \in [H_1 g_2]$. Since also $H_0 g_0 = (H_1 g_0 - H_1 g_1)/\lambda_0$,

$v\epsilon[H_2 g_2]$ if and only if $H_1 g_0 \epsilon [H_1 g_2]$.

Theorem 6.3.3 (modification of Dixon (1972b)) Let $f:R^n \to R$ be continuously differentiable; $H_0 \epsilon L(R^n)$ symmetric, $x_0 \epsilon R^n$, $\{\rho_i\}$, $i = 0,1,\ldots$ a sequence of nonzero numbers ϵR. Let the sequences $\{x_i\}$, $\{H_i\}$, $i = 1,2,\ldots$ be generated from x_0, H_0 by Algorithm 6.0.1 using a specific perfect line search strategy and update (6.3.2) with $s_i = x_{i+1} - x_i$, $y_i = g(x_{i+1}) - g(x_i)$. Define $g_i = g(x_i)$. Let $\{\hat{x}_i\}$, $\{\hat{H}_i\}$, $i = 1,2,\ldots$ also be generated from x_0, H_0 by Algorithm 6.0.1 using the same perfect line search strategy and update $\hat{H}_{i+1}(\phi_i, \rho_i) = H_{i+1}(\rho_i) + \phi_i u_i u_i^T$, H_{i+1} given by (6.3.2) with $\hat{H}_i$ replacing H_i, $u_i = \hat{H}_i y_i - s_i(\langle y_i, \hat{H}_i y_i\rangle / \langle s_i, y_i\rangle)$, $s_i = \hat{x}_{i+1} - \hat{x}_i$, $y_i = g(\hat{x}_{i+1}) - g(\hat{x}_i)$, ϕ_i a real number $\neq -1/\langle g_{i+1}, H_i g_{i+1}\rangle$. Assume each $\langle g_i, H_i g_i\rangle \neq 0$, and define $d_i = H_i g_i$. Then for $i = 0,1,\ldots$, $\hat{x}_{i+1} = x_{i+1}$, $d_{i+1} \neq 0$, $\hat{H}_{i+1} g(x_{i+1})$ has the same direction as d_{i+1}, and $\hat{H}_{i+1} = H_{i+1} + \phi_i d_{i+1} d_{i+1}^T$.

Proof: Virtually the same as the proofs of Theorem 6.1.1 and 6.1.2.

Using Theorem 6.3.1, Theorem 6.3.2 shows that the sequence of points generated by the Huang-Oren DFP-BFGS class ((6.3.1) with $w = s$) is in general different from those generated by the plain DFP-BFGS class unless $\rho = \gamma = 1$. The condition $\langle g_2, H_0 g_0\rangle$ will be satisfied if the objective function f is quadratic (Huang (1970)), but not for general non-quadratic f. The condition that $H_1 g_0$ and $H_1 g_2$ are linearly dependent is equivalent to $H_0 g_0$, $H_1 g_2$ linearly dependent by the proof of Theorem 6.3.2. Since $\langle g_1, H_0 g_0\rangle = \langle g_1, H_1 g_2\rangle = 0$, this will occur if $n = 2$ but not in general if $n > 2$. This fact is contained in Dixon (1972a) and is attributed to M.J.D. Powell.

Theorem 6.3.3 shows that the sequence of points generated by any Huang-Oren DFP-BFGS class in a perfect line search algorithm is independent of the choices of ϕ. Theorem 6.3.4 shows that this is not true for the general Huang-Oren w-class (6.3.1).

<u>Theorem 6.3.4</u> Let $f:R^n \to R$ be continuously differentiable; $H \in L(R^n)$ symmetric, $x \in R^n$. Define $g = g(x)$, and assume $\langle g,Hg\rangle \neq 0$. Let $\bar{x} = x-\lambda Hg$ be chosen by a perfect line search. Define $\bar{g} = g(\bar{x})$, let $s = \bar{x}-x$, $y = \bar{g}-g$, and assume $\langle \bar{g},H\bar{g}\rangle \neq 0$. Let $\bar{H} \in L(R^n)$ be chosen from the one parameter class $\bar{H}(\phi,\rho,\gamma) = \bar{H}(\phi)$ given by (6.3.1) with ϕ a free real parameter, ρ a fixed nonzero number such that $\rho s \neq Hy$, $\gamma = 1$, and $w \in R^n$ such that $\langle w,y\rangle \neq 0$ and $w \notin [s,Hy]$. Let $u = s-Hy - w(\langle s-Hy,y\rangle/\langle w,y\rangle)$. Then for any two different real numbers ϕ_1, $\phi_2 \in R^n$, $\bar{H}(\phi_1)g_1$ and $\bar{H}(\phi_2)g_1$ are linearly independent unless either $\lambda = 1/\rho$ or $\langle u,\bar{g}\rangle = 0$. If H is positive definite, then the condition $\langle u,\bar{g}\rangle = 0$ is equivalent to $\hat{s}$, $H\hat{y}$ linearly dependent, where $\hat{s} = Qs$, $\hat{y} = Q^Ty$, Q the matrix projecting onto $[s-Hy,w]$ in the metric H^{-1}.

<u>Proof</u>: Virtually identical to the proof of Theorem 6.1.3.

Again, by Theorem 6.3.1 we know that the conclusions of Theorem 6.3.4 also apply if the Oren parameter $\gamma \neq 1$ in (6.3.1). Also, identically to the proof of Theorem 3.3.2 we can show that unless $\hat{s}$ and $H\hat{y}$ are linearly dependent, update class (6.3.1) is equivalent to a restricted DFP-BFGS class with Huang and Oren parameters. Therefore Theorem 6.3.4 shows that the points generated by a restricted Huang-Oren DFP-BFGS class will in general depend upon the choices of ϕ.

CHAPTER VII

OPTIMAL CONDITIONING - THEORY

In this thesis we have studied updates $\bar{B} \epsilon L(R^n)$ to a symmetric and positive definite matrix $B \epsilon L(R^n)$ which satisfy: 1) $\bar{B}s' = y'$ for s', y' nonzero vectors ϵR^n; 2) $\bar{B}$ is symmetric and positive definite; 3) $(\bar{B}-B)z = 0$ for all $z \epsilon Z$, Z an m dimensional subspace of R^n, $0 \leq m \leq n-2$. From Chapters 2-4 we learned that if any $\bar{B}$ exists satisfying these conditions, then among them will be the updates

$$\bar{B}(\phi) = B + \frac{(y-Bs)y^T + y(y-Bs)^T}{\langle y,s\rangle} - \frac{\langle y-Bs,s\rangle yy^T}{\langle y,s\rangle^2}$$

$$- \phi\langle s,Bs\rangle \left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]\left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]^T \qquad (7.0.1)$$

where s and y equal s', y' if y' and Bs' are orthogonal to Z, and are projections of them otherwise (see Theorem 4.1.2), and ϕ is any number less than $\langle s,Bs\rangle\langle y,B^{-1}y\rangle / (\langle s,Bs\rangle\langle y,B^{-1}y\rangle - \langle s,y\rangle^2)$.

One way to choose an update satisfying the above three conditions was shown in Chapter 5: choose the $\bar{B}$ which minimizes $\|\bar{B}-B\|_{Fr,M}$, or $\|\bar{B}^{-1}-B^{-1}\|_{Fr,M}$ subject to these conditions, for some symmetric non-singular matrix M. For example, if $\hat{B}$ is any matrix which satisfies the above conditions, then $\|\bar{B}-B\|_{Fr,\hat{B}^{-1/2}}$ is minimized by the DFP update, $\phi = 0$ in (7.0.1), while $\|\bar{B}^{-1}-B^{-1}\|_{Fr,\hat{B}^{1/2}}$ is minimized by $\phi = 1$ in (7.0.1), the BFGS. (In this chapter, we will call the updates $\bar{B}(0)$ and $\bar{B}(1)$ in (7.0.1) the DFP [illegible] BFGS respect[illegible] independent of the meanings of [illegible] and y.) These two [illegible] the BFGS have [illegible] when $Z = \{\}$.

Davidon's (1975) recent algorithm, where Z is not empty, chooses its update from a class of form (7.0.1). However, rather than choosing the BFGS or the DFP, Davidon chooses the "optimally conditioned" $\overline{B}(\phi)$: the update in (7.0.1) which minimizes a certain bound on the condition number of $\overline{B}(\phi)$. In Chapters 7 and 8 we examine whether this is a good idea. In Chapter 7 we analyze optimal conditioning theoretically, and in Chapter 8 we present computer results comparing the BFGS and the "optimally conditioned" updates. Our basic conclusion is that optimal conditioning is probably not helpful, at least in most cases. However, the analysis of Chapter 7 leads us to suggest in Chapter 8 an updating strategy which uses optimal conditioning in a very restricted manner along with the BFGS; and which seems slightly superior to the BFGS in our tests. Further testing is definitely required.

The condition number of a nonsingular matrix A, denoted $K(A)$, is defined as $\| A\|_2 \; \| A^{-1}\|_2$. If A is symmetric and positive definite, $K(A)$ is the ratio of the largest to the smallest eigenvalue of A. It is well-known that the difficulty in solving a linear system $Ax = b$ is proportional to $K(A)$ (c.f. Wilkinson and Reinsch (1971)). Since in a quasi-Newton method, $\overline{B}(\phi)$ is used in solving a linear system $\overline{B}(\phi)d = -\overline{g}$, (or if $\overline{H}(\phi) = \overline{B}(\phi)^{-1}$ is approximated, $\overline{H}(\phi)$ is used in a matrix-vector multiplication $d = -\overline{H}(\phi)\overline{g})$ it is reasonable to try to hold down the condition number of $\overline{B}(\phi)$ at each iteration. Both Davidon (1975) and Oren and Spedicato (1976) seem to have had this idea independently, although our justification follows more closely that of Oren and Spedicato. We show in Lemma 7.0.1 that $K(B^{-1/2}\overline{B}(\phi)B^{-1/2})$ is an upper bound on $K(\overline{B}(\phi))/K(B)$, where $B^{-1/2}$ is the symmetric and positive definite square root of B^{-1}. Both Davidon (1975) and Oren and Spedicato (1976) then derive and use the positive definite update in (7.0.1) which minimizes $K(B^{-1/2}\bar{B}(\phi)B^{-1/2})$, and call this the optimally conditioned (OC) update from (7.0.1).

<u>Lemma 7.0.1</u> Let $B, \bar{B} \in L(R^n)$ be symmetric and non-singular, B positive definite. Then $K(\bar{B}) \leq K(B) \cdot K(B^{-1/2}\bar{B}\, B^{-1/2})$.

<u>Proof</u>: Since $B^{1/2}$ is symmetric, $\| B^{1/2}\|_2{}^2 = \| B\|_2$ and $\| B^{-1/2}\|_2{}^2 = \| B^{-1}\|$. Thus $\| \bar{B}\|_2 = \| B^{1/2}(B^{-1/2}\, \bar{B}\, B^{-1/2})B^{1/2}\|_2 \leq \| B\|_2 \; \| B^{-1/2}\, \bar{B}\, B^{-1/2}\|_2$ and similarly $\| \bar{B}^{-1}\| \leq \| B^{-1}\| \; \| B^{1/2}\, \bar{B}^{-1}\, B^{1/2}\|$. Thus $K(\bar{B}) = \| \bar{B}\|_2 \; \| \bar{B}^{-1}\|_2 \leq (\| B\| \; \| B^{-1}\|)(\| B^{-1/2}\, \bar{B}\, B^{-1/2}\|)(\| B^{1/2}\, \bar{B}^{-1}\, B^{1/2}\|) = K(B)\; K(B^{-1/2}\, \bar{B}\, B^{-1/2})$.

This bound upon $K(\bar{B}(\phi))$ is not likely to be tight. Hopefully though it will be reasonable, and it has the advantage that since $B^{-1/2}\, \bar{B}(\phi)\, B^{-1/2}$ differs from the identity by a matrix of rank 2, $K(B^{-1/2}\, \bar{B}(\phi)\, B^{-1/2})$ - henceforth denoted $K(\phi)$ - is analytically obtainable, which allows us to give a simple formula in terms of B, s, and y for the ϕ_{OC} which minimizes $K(B^{-1/2}\, \bar{B}(\phi)\, B^{-1/2})$ among positive definite $\bar{B}(\phi)$. This was done by both Davidon (1975) and Oren and Spedicato (1976). We give a more straightforward derivation of the optimally conditioned update in Section 7.1, resulting also in additional information, contained in Figures 7.1.1a-c, which is of much help to our understanding and subsequent analysis.

The positive definite update which minimizes the bound on the condition number of $\bar{B}(\phi)$ may be within or outside the convex class, whereas almost all previous algorithms had used $\phi \in [0,1]$. In Section 7.2 we determine whether we have gained much, from the standpoint of conditioning, by considering updates outside the convex class. We define the OCX update as the (positive definite) $\bar{B}(\phi_{OCX})$ which minimizes $K(\phi)$ over $\phi\epsilon[0,1]$, and K(OC), K(OCX), K(DFP), and K(BFGS) as $K(\phi)$ for $\phi = \phi_{OC}, \phi_{OCX}$, 0 and 1 respectively; and show that when the OC update is outside the convex class, the OCX update is the DFP or the BFGS, and $K(OCX) < \frac{32}{27} K(OC)$. This is a tiny difference - we are concerned with orders of magnitude variations in $K(\phi)$. Thus since the convex class is recommended by minimum norm considerations and past experience, it seems unwarranted to go outside the convex class for reasons of conditioning. We do show however that K(BFGS) (or K(DFP)) may be far greater than K(OCX), so there does seem to be some argument for choosing members of the convex class other than the BFGS.

However in Section 7.3 we show that if a direct prediction BFGS algorithm is Q-superlinearly convergent and some reasonable conditions are satisfied, then $K(BFGS) \rightarrow 1$ as the algorithm converges to the solution. We know that the standard BFGS algorithm ($s = \bar{x}-x$, $y = g(\bar{x})-g(x)$, $\phi = 1$ in (7.0.1)) is Q-superlinearly convergent for a wide class of problems, and the work of Chapter 9 makes it seem likely that the restricted BFGS is too. Therefore our result indicates that optimal conditioning is likely to become unimportant as an algorithm nears the solution. The analysis of Sections 7.2 and 7.3 leads us to propose and test in Chapter 8 an algorithm which selects the OCX if $K(BFGS) > \frac{3}{2} K(OCX)$ and the BFGS otherwise. In practice this results in almost sole use of the BFGS.

Finally in Section 7.4 we examine the relation of restricting the updating space to the conditioning of $\bar{B}$. We obtain the reasonable result that K(OC) is larger if we restrict our updating space than if we don't. This indicates that these may be a trade-off between preserving old information and holding down a bound on the condition number of $\bar{B}$.

Except as specifically stated in Section 7.3, the results of this chapter are independent of the meanings of s and y in equation (7.0.1). Thus our analysis applies to standard and restricted DFP-BFGS classes, and by the equivalences of Chapter 3, to general one parameter v-classes as well. Much of the material in this chapter is also contained in Schnabel (1976).

Section 7.1 - Derivation of the Optimally Conditioned Update

In this section we derive the "optimally conditioned" update, the $\bar{B}(\phi_{OC})$ which minimizes $K(B^{-1/2}\bar{B}(\phi)B^{-1/2})$, a bound on the condition number of $\bar{B}(\phi)$, among all positive definite $\bar{B}(\phi)$ in (7.0.1). This has been previously done by Davidon (1975) and Oren and Spedicato (1976), so our work is largely a review. However we take a more straightforward approach, analytically calculating the eigenvalues of $B^{-1/2}\bar{B}(\phi)B^{-1/2}$, and this gives us information about the curves $K(\phi)$ which is reflected in the graphs at the end of this section, and is of help in analyzing the optimal conditioning process in Section 7.2.

Lemma 7.1.1 gives the eigenvalues of $B^{-1/2}\bar{B}(\phi)B^{-1/2}$ as a function of ϕ. Since $B^{-1/2}\bar{B}(\phi)B^{-1/2} = I$ + a matrix of rank two, it has n-2 unit eigenvalues and two other eigenvalues denoted $\lambda_{\pm}(\phi)$. Lemma 7.1.2 uses these eigenvalues to determine when $\bar{B}(\phi)$ is positive definite. In Lemma 7.1.3 we show that $\lambda_{\pm}(\phi)$ are both monotonically decreasing functions of ϕ. We explain that if $n>2$ this means that the OC update must have the form $\bar{B}(\phi_{OC}) = B + ww^T - zz^T$, so we need consider only updates $\bar{B}(\omega)$ in the parameterization of Section 3.2. Theorem 7.1.4 then derives the "optimally conditioned" update by finding the update $\bar{B}(\omega)$ which minimizes $K(B^{-1/2}\bar{B}(\omega)B^{-1/2})$. Its proof leads to Figures 7.1.1a-c.

Note that if we are instead updating approximations H to $G(x)^{-1}$ with updates $\bar{H}(\bar{\phi})$ from the inverse DFP-BFGS class (1.4.7), then $K(\bar{H}(\bar{\phi}))/K(H) \le K(H^{-1/2}\bar{H}(\bar{\phi})H^{-1/2})$. Therefore if $H = B^{-1}$, the $\bar{H}(\bar{\phi}_{oc})$ which minimizes $K(H^{-1/2}\bar{H}(\bar{\phi})H^{-1/2})$ will equal $\bar{B}(\phi_{oc})^{-1}$, since $K(H^{-1/2}\bar{H}(\bar{\phi})H^{-1/2}) = K(B^{-1/2}\bar{B}(\phi)B^{-1/2})$ for $\bar{H}(\bar{\phi}) = \bar{B}(\phi)^{-1}$. Thus all our analysis of optimal conditioning is independent of whether we are updating B or H.

We start by analytically determining the eigenvalues of $B^{-1/2}\bar{B}(\phi)B^{-1/2}$. We only consider the case when y and Bs are linearly independent, as otherwise the entire class (7.0.1) reduces to the SR1 update.

<u>Lemma 7.1.1</u> Let $n \ge 2$, $B \in L(R^n)$ symmetric and positive definite; $s, y \in R^n$ such that $\langle s,y \rangle \ne 0$ and y, Bs are linearly independent; ϕ a real parameter, $\bar{B}(\phi)$ given by (7.0.1). Define

$$a = \langle y, B^{-1}y \rangle,\ b = \langle y,s \rangle,\ c = \langle s,Bs \rangle. \tag{7.1.1}$$

Then $B^{-1/2}\bar{B}(\phi)B^{-1/2}$ has n-2 unit eigenvalues, and two remaining eigenvalues

$$\lambda_{\pm}(\phi) = f_1(\phi) \pm (f_1(\phi)^2 - f_2(\phi))^{1/2} \tag{7.1.2a}$$

where

$$f_1(\phi) = \frac{a(b+c) - \phi(ac-b^2)}{2b^2},\ f_2(\phi) = \frac{a}{b} - \frac{\phi(ac-b^2)}{bc} \tag{7.1.2b}$$

<u>Proof</u>: Define $A(\phi) = B^{-1/2}\bar{B}(\phi)B^{-1/2}$. Then $A(\phi) = I + A_1(\phi)$, where $A_1(\phi)$ is a matrix of rank two $\epsilon[B^{-1/2}y, B^{1/2}s]$. Thus $A(\phi)$ has n-2 unit eigenvalues corresponding to the n-2 dimensional eigenspace orthogonal to $[B^{-1/2}y, B^{1/2}s]$, and two eigenvalues given by (7.1.2) which are found by solving the system $A(\phi) \cdot (B^{-1/2}y + \alpha B^{1/2}s) = \lambda(B^{-1/2}y + \alpha B^{1/2}s)$.

Lemma 7.1.2 Let $n \geq 2$, $B \in L(R^n)$ symmetric and positive definite; $s, y \in R^n$ such that $\langle s,y\rangle \neq 0$ and y, Bs are linearly independent; ϕ a real parameter; $\bar{B}(\phi)$ given by (7.0.1); a,b,c defined by (7.1.1). Then $\bar{B}(\phi)$ is positive definite if and only if

$$b>0 \text{ and } \phi < \frac{ac}{ac-b^2} \tag{7.1.3}$$

Proof: Let $B^{-1/2}$ be the symmetric and positive definite square root of B^{-1}. Then $\bar{B}(\phi)$ is positive definite if and only if $B^{-1/2}\bar{B}(\phi)B^{-1/2}$ is. By Lemma 7.1.1, $B^{-1/2}\bar{B}(\phi)B^{-1/2}$ is positive definite if and only if $\lambda_-(\phi) > 0$. This clearly requires $f_1(\phi) > 0$ and $f_2(\phi) > 0$. Since $f_1(\phi) = \frac{c}{2b} f_2(\phi) + \frac{a}{2b}$, then if $b > 0$ and $f_2(\phi) > 0$ then $f_1(\phi) > 0$, while if $b < 0$ and $f_2(\phi) > 0$ then $f_1(\phi) < 0$. Thus $\lambda_-(\phi) > 0$ if and only if $b > 0$ and $f_2(\phi) > 0$. By writing $a = \langle B^{-1/2}y, B^{-1/2}y\rangle$, $b = \langle B^{-1/2}y, B^{1/2}s\rangle$ and $c = \langle B^{1/2}s, B^{1/2}s\rangle$, the Cauchy -Schwartz inequality gives

$$ac-b^2 > 0 \tag{7.1.4}$$

and therefore $f_2(\phi) > 0$ is equivalent to $\phi < ac/(ac-b^2)$ if $b > 0$.

Since $B^{-1/2}\bar{B}(\phi)B^{-1/2}$ has n-2 unit eigenvalues and remaining eigenvalues $\lambda_\pm(\phi)$, minimizing $K(\phi)$ among ϕ such that $\bar{B}(\phi)$ is positive definite is equivalent to

$$\min_{\phi \in R} \frac{\max(\lambda_+(\phi),1)}{\min(\lambda_-(\phi),1)} \text{ subject to } \lambda_-(\phi) > 0 \tag{7.1.5}$$

In Lemma 7.1.3 we show that $\lambda_\pm(\phi)$ are both monotomically decreasing functions of ϕ for all $\phi \in R$. This shows that if $n>2$, problem (7.1.5) is equivalent to

$$\min_{\phi\in R} \frac{\lambda_+(\phi)}{\lambda_-(\phi)} \text{ subject to } 0 < \lambda_-(\phi) \le 1 \le \lambda_+(\phi) \tag{7.1.6}$$

which will allow us to switch to the parameterization $\bar{B}(\omega)$. The case n=2 is discussed briefly at the end of this section.

Lemma 7.1.3 Let $\lambda_\pm(\phi)$ be defined by (7.1.2), where a,b,c are nonzero constants $\in R$ such that $ac-b^2 > 0$. Then $\lambda_\pm(\phi)$ are monotonically decreasing functions of ϕ.

Proof: Straightforward differentiation shows that

$$\frac{d\lambda_\pm(\phi)}{d\phi} = \frac{-(ac-b^2)}{2b^2}\left[1 \pm \frac{(f_1(\phi)-\frac{b}{c})}{\left[(f_1(\phi)-\frac{b}{c})^2+\frac{ac-b^2}{c^2}\right]^{1/2}}\right]$$

Thus $d\lambda_\pm(\phi)/d(\phi) < 0$ for all ϕ.

Thus the positive definite update which minimizes $K(B^{-1/2}\bar{B}(\phi)B^{-1/2})$ is one for which $0 < \lambda_-(\phi)\le 1\le\lambda_+(\phi)$, or equivalently, $-1\le\hat{\lambda}_-(\phi)\le 0\le\hat{\lambda}_+(\phi)$, where $\hat{\lambda}_\pm(\phi)$ are the eigenvalues of $B^{-1/2}(\bar{B}(\phi)-B)B^{-1/2}$. It is well known (c.f. Brodlie, Gourlay, and Greenstadt (1973)) that a rank two matrix in $L(R^n)$ has one non-negative and one non-positive eigenvalue if and only if it is of the form ww^T-zz^T, $w,z\in R^n$. Therefore to solve the optimal conditioning problem we only need consider updates $\bar{B}(\phi)$ for which $B^{-1/2}(\bar{B}(\phi)-B)B^{-1/2}$, and thus $\bar{B}(\phi)-B$, has this form. In Section 3.2 it was shown that such updates are precisely the ones which can be represented in the $\hat{B}(\omega)$ parameterization (3.2.6), if we include the SR1 update which is

$\lim_{\omega\to\infty} \hat{B}(\omega)$. Therefore the optimal conditioning problem (7.1.6) is equivalent to

$$\min_{\omega>0} \frac{\lambda_+(\omega)}{\lambda_-(\omega)} \text{ subject to } \lambda_-(\omega) > 0 \tag{7.1.7}$$

where $\lambda_\pm(\omega)$ are found by substituting

$$\phi = \frac{b^2c}{\omega(b-c)} - \frac{c}{b-c} \tag{7.1.8}$$

into (7.1.2) (Formula (7.1.8), for which $\bar{B}(\phi) = \hat{B}(\omega)$, is given in Corollary 3.2.2). We derive the "optimally conditioned" update in Theorem 7.1.4 by solving (7.1.7). The case $b = c$ turns out to require no additional work.

<u>Theorem 7.1.4</u> (Davidon (1975)) Let $n>2$, $B\epsilon L(R^n)$ be symmetric and positive definite; $s,y\epsilon R^n$ such that $\langle s,y\rangle > 0$ and y, Bs are linearly independent, ϕ a real parameter; $\bar{B}(\phi)$ given by (7.0.1), a,b,c defined by (7.1.1). Then among all positive definite $\bar{B}(\phi)$, $K(B^{-1/2}\bar{B}(\phi)B^{-1/2})$ is minimized by $\bar{B}(\phi_{oc})$ where

$$\phi_{oc} = \begin{cases} \dfrac{a(c-b)}{ac-b^2} & 2ac-b(a+c) > 0 \\[2ex] \dfrac{c}{c-b} & 2ac-b(a+c) \le 0 \end{cases} \tag{7.1.9}$$

<u>Proof:</u> From the above discussion, we know that unless $b = c$, ϕ_{oc} is found by solving (7.1.7), with $\phi(\omega)$ given by (7.1.8). Therefore we solve $d(\lambda_+(\omega)/\lambda_-(\omega))/d\omega = 0$. Using the chain rule

$$\frac{d}{d\omega}\left[\frac{\lambda_+(\omega)}{\lambda_-(\omega)}\right] = \frac{-(ac-b^2)f_3(\omega)}{\lambda_-(\phi)^2[f_1(\phi)^2-f_2(\phi)]^{1/2}\ 2\omega^2(b-c)^2},$$

$$f_3(\omega) = \frac{bc(ac-b^2)}{\omega} - [2ac-b(a+c)] \qquad (7.1.10)$$

Let $\alpha = 2ac-b(a+c)$. Then since $ac-b^2 > 0$ by (7.1.4), if $\alpha > 0$

$$\frac{d}{d\omega}\left[\frac{\lambda_+(\omega)}{\lambda_-(\omega)}\right] \begin{cases} = 0 & \omega = \dfrac{bc(ac-b^2)}{2ac-b(a+c)} \equiv \omega_{oc} \\ < 0 & 0 < \omega < \omega_{oc} \\ > 0 & \omega > \omega_{oc} \end{cases} \qquad (7.1.11)$$

while if $\alpha \leq 0$, $d(\lambda_+(\omega)/\lambda_-(\omega))/d\omega < 0$ for all $\omega > 0$. Therefore if $\alpha > 0$, ω_{oc} solves (7.1.7) and corresponds to $\phi_{oc} = a(c-b)/(ac-b^2)$, and $\overline{B}(\phi_{oc})$ is positive definite by (7.1.3). If $\alpha \leq 0$, the solution to (7.1.7) = $\lim_{\omega\to\infty} \hat{B}(\omega)$ = the SR1 update, $\phi = c/(c-b)$, which is positive definite in this case. If b=c, arguements of continuity still show that $\phi_{oc} = a(c-b)/(ac-b^2)$.

The solution to the optimal conditioning problem is summarized in Figures 7.1.1a-c, which graph the condition number of $B^{-1/2}\hat{B}(\omega)B^{-1/2}$ for all positive definite $\hat{B}(\omega)$. We have divided the problem into 3 cases which will be useful to our analysis in Section 7.2, and have distorted the ω scale so that the SR1 update, $\lim_{\omega\to\infty} \hat{B}(\omega)$, is shown. The positive definite updates are all those to the left of ω = PD in Case 1, all those to the right of ω = PD in Case 2, and all $\omega > 0$ in Case 3.

The important information about the optimal conditioning problem is contained in equations (7.1.10) and (7.1.11), and these graphs. They show that in Case 1, the OC update is in the convex class, while in Cases 2 and 3, it is outside the convex class and may be the SR1. In all 3 cases, the curve $K(\omega) = \lambda_+(\omega)/\lambda_-(\omega)$ is monotonically decreasing for $\omega < \omega_{OC}$, and monotonically increasing for $\omega > \omega_{OC}$ if ω_{OC} is finite.

The solution to the optimal conditioning problem is independent of the meanings of s and y in (7.0.1). Also, in Chapter 3 we have shown that the general v-class (3.1.3) is equivalent to a class of form (3.1.4) unless the projections $\hat{y}, B\hat{s}$ of y and Bs onto [y-Bs,v] in the metric B^{-1} are linearly dependent. In this case $\hat{y} = kB\hat{s}$, so that k is an eigenvalue of all $B^{-1/2}\overline{B}(\phi)B^{-1/2}$ corresponding to eigenvector $B^{1/2}\hat{s}$. Therefore there exist positive definite $\overline{B}(\phi)$ only if $k > 0$, and since only one eigenvalue of $B^{-1/2}\overline{B}(\phi)B^{-1/2}$ is a function of ϕ, the "optimally conditioned" $\overline{B}(\phi)$ occurs when it is 1, and $\overline{B}(\phi)$ is the SR1.

If n=2, Davidon (1975) shows that $\phi_{OC} = a(c-b)/(ac-b^2)$ always. However the case n=2 is not considered important because in practice we are seldom concerned with such small problems.

Case 1

$b \leq a$, $b < c$

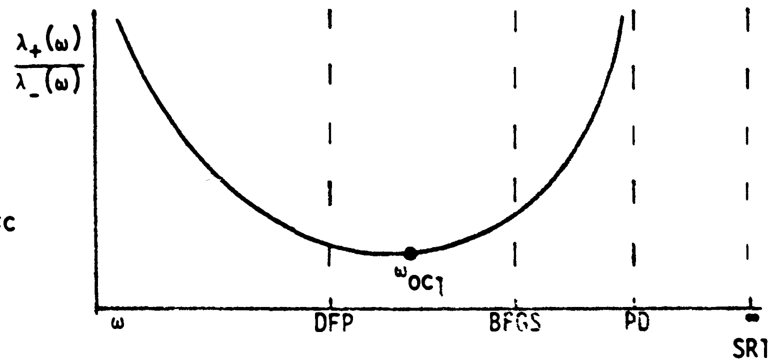

Figure 7.1.1a

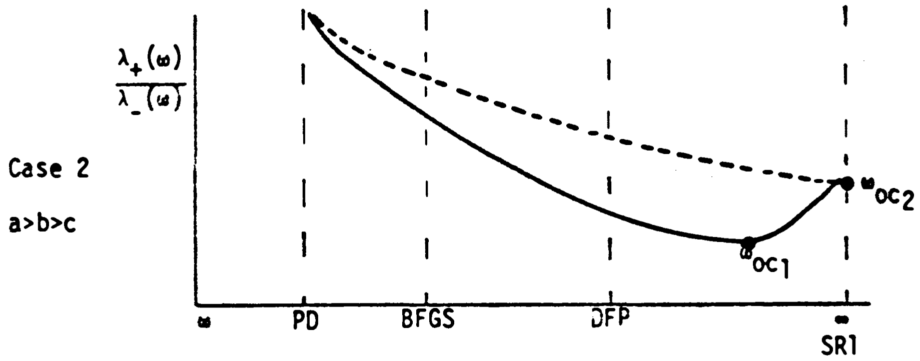

Figure 7.1.1b

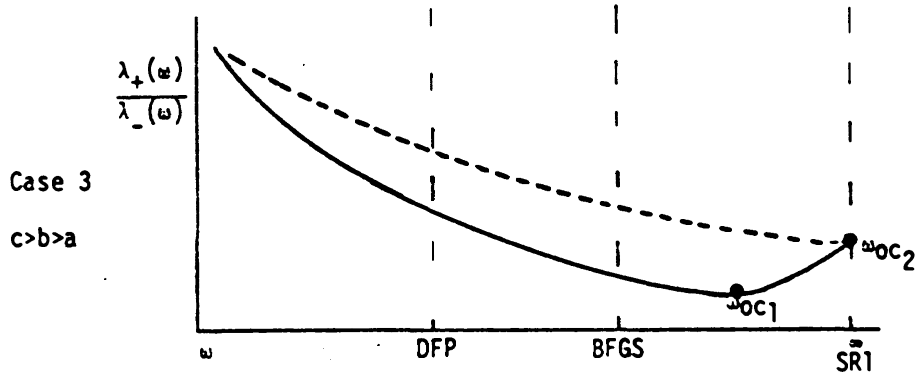

Figure 7.1.1c

solid line: $2ac-b(a+c) > 0$, $\omega_{oc} = \omega_{oc_1}$

dotted line: $2ac-b(a+c) \leq 0$, $\omega_{oc} = \omega_{oc_2}$

DFP $= b^2$; BFGS $= bc$; SR1 $= \omega_{oc_2} = +\infty$;

$$PD = \frac{b(ac-b^2)}{a-b}\ ; \quad \omega_{oc_1} = \frac{bc(ac-b^2)}{2ac-b(a+c)}$$

Section 7.2 - Optimal Conditioning Restricted to the Convex Class

The optimal conditioning strategy of Davidon (1975) and Oren and Spedicato (1976) can select updates from (7.0.1) either within or outside the convex class $\phi\epsilon[0,1]$, whereas previous algorithms have mainly used the DFP or BFGS (ϕ = 0,1 respectively) and almost always ϕ in the convex class. In this section we examine whether, from the standpoint of conditioning, it is worthwhile to consider updates outside the convex class.

We define the OCX update $\bar{B}(\phi_{OCX})$ as the "optimally conditioned" update in the convex class, the $\bar{B}(\phi)$ in (7.0.1) for which $K(B^{-1/2}\bar{B}(\phi)B^{-1/2})$ is minimized among $\phi\epsilon[0,1]$. From Figures 7.1.1a-c, the OCX update is the OC if $b\le a,c$, the DFP if $a>b>c$, and the BFGS if $a<b<c$. We let $K(\phi) = K(B^{-1/2}\bar{B}(\phi)B^{-1/2})$, and write K(OC), K(OCX), K(DFP), and K(BFGS) for $K(\phi)$ with $\phi = \phi_{OC}, \phi_{OCX}$, 0 and 1 respectively. In Lemma 7.2.1 we give formulas for these quantities. Then with the help of Lemmas 7.2.2 and 7.2.3, Theorem 7.2.4 contains the rather surprising result that for any class (7.0.1) with $n>2$, $K(OCX) < \frac{32}{27} K(OC)$. This means that we have made virtually no improvement in the bound on $\bar{B}(\phi)$ by considering updates outside the convex class. On the other hand, Lemmas 7.2.5 and 7.2.6 show that either K(DFP), K(BFGS) or both may be much larger than K(OCX). Therefore the results of this section indicate that from the viewpoint of conditioning, there is very little reason to consider updates outside the convex class, but there may be reason to consider updates in the convex class other than the BFGS (or DFP). We examine the relation between K(BFGS) and K(OCX) more closely in Section 7.3.

<u>Lemma 7.2.1</u> Let $n>2$, $B\epsilon L(R^n)$ symmetric and positive definite; $s,y\epsilon R^n$ such that $\langle s,y\rangle > 0$ and y,Bs are linearly independent; ϕ a real parameter; $\bar{B}(\phi)$ given by (7.0.1); a,b,c defined by (7.1.1). Then

$$K(DFP) = [a(b+c)^2-2b^3+(b+c)\sqrt{a^2(b+c)^2-4ab^3}]/2b^3 \qquad (7.2.1)$$

$$K(BFGS) = [(a+b)^2c-2b^3+(a+b)\sqrt{(a+b)^2c^2-4b^3c}]/2b^3 \tag{7.2.2}$$

$$K(OC) = \begin{cases} [2ac-b^2+2\sqrt{a^2c^2-ab^2c}]/b^2 & 2ac>b(a+c) \quad (7.2.3a) \\ (a-b)/(b-c) & 2ac\leq b(a+c)\ \&\ a>b \quad (7.2.3b) \\ (c-b)/(b-a) & 2ac\leq b(a+c)\ \&\ b>a \quad (7.2.3c) \end{cases}$$

$$K(OCX) = \begin{cases} K(OC) & b\leq a,c \\ K(DFP) & a>b>c \\ K(BFGS) & c>b>a \end{cases} \tag{7.2.4}$$

Proof: From Section 7.1, we know that in all these cases, $K(\phi) = \lambda_+(\phi)/\lambda_-(\phi)$. Equations (7.2.1) and (7.2.2) come from substituting $\phi = 0,1$ into (7.1.2), while (7.2.3) comes from using ϕ_{OC} given by (7.1.9). Equation (7.2.4) is clear from Figures 7.1.1a-c.

To compare K(OC) and K(OCX) we consider Cases 2 and 3 of Figure 7.1.1, where the OC and OCX updates differ. In Case 2 the OCX update is the DFP. To compare K(OC) to K(DFP) we then consider the cases when $\bar{B}(\phi_{OC})$ is and is not the SR1. Lemma 7.2.2 shows that when $\bar{B}(\phi_{OC})$ is not the SR1 in Case 2, $K(OFP) < \frac{9}{8} K(OC)$, and Lemma 7.2.3 shows that when it is, $K(DFP) < \frac{32}{27} K(SR1)$. In Case 3 the OCX update is the BFGS, and the comparison of K(BFGS) to K(OC) turns out to be algebraically equivalent to the comparison of K(DFP) to K(OC) in Case 2.

Lemma 7.2.2 Let $n>2$, $B\in L(R^n)$ symmetric and positive definite; $s,y\in R^n$ such that $\langle s,y\rangle > 0$ and y,Bs are linearly independent; ϕ a real parameter, $\bar{B}(\phi)$

given by (7.0.1), a,b,c defined by (7.1.1). Then if $a>b>c$ and $2ac \geq b(a+c)$, the OCX update is the DFP, the OC update has $\phi = a(c-b)/(ac-b^2)$, and $K(DFP) < \frac{9}{8}K(OC)$.

Proof: The fact that the OCX update is the DFP and the OC has $\phi = a(c-b)/(ac-b^2)$ is represented by the solid line in Figure 7.1.1b. Their condition numbers are given by (7.2.1) and (7.2.3a).

Let

$$a = \alpha c, \quad b = \beta c. \tag{7.2.5}$$

Then $a > b > c$ implies $\alpha > \beta > 1$, and $ac - b^2 > 0$ implies $\alpha > \beta^2$. From α, $\beta > 1$ and $b \leq \frac{2ac}{a+c}$ we find $\beta \leq \frac{2\alpha}{\alpha+1} < 2$, $\alpha \geq \frac{\beta}{2-\beta}$. Therefore we require the maximum value of

$$f_1(\alpha,\beta) \triangleq \frac{K(DFP)}{K(OC)} = \frac{\alpha(\beta+1)^2-2\beta^3+(\beta+1)\sqrt{\alpha^2(\beta+1)^2-4\alpha\beta^3}}{2\beta[2\alpha-\beta^2+2\sqrt{\alpha^2-\alpha\beta^2}]} \tag{7.2.6}$$

$$\text{subject to } \beta\epsilon(1,2),\ \alpha\epsilon\left[\frac{\beta}{2-\beta},\infty\right). \tag{7.2.7}$$

We show that this maximum has a least upper bound of $\frac{9}{8}$, which is $\lim_{\beta\to 2} f(\frac{\beta}{2-\beta},\beta)$.

We show first that for any fixed $\beta\epsilon(1,2)$, $f_1(\alpha,\beta)$ is a monotonically decreasing function of α for $\alpha \geq \frac{\beta}{2-\beta}$:

$$\frac{df_1(\alpha,\beta)}{d\alpha} = \frac{-[f_2(\alpha,\beta)+f_3(\alpha,\beta)+f_4(\alpha,\beta)]}{f_5(\alpha,\beta)}, \tag{7.2.8}$$

$$f_2(\alpha,\beta) \triangleq (\beta-1)^2[\alpha^2(\beta+1)+\sqrt{\alpha^2-\alpha\beta^2}\cdot\sqrt{\alpha^2(\beta+1)^2-4\alpha\beta^3}],$$

$$f_3(\alpha,\beta) \triangleq [\alpha(\beta-1)^2+2\beta^3]\sqrt{\alpha^2(\beta+1)^2-4\alpha\beta^3},$$

$$f_4(\alpha,\beta) \triangleq (\beta+1)[\alpha(\beta-1)^2-2\beta^3]\sqrt{\alpha^2-\alpha\beta^2},$$

$$f_5(\alpha,\beta) \triangleq 2\beta^{-1}[2\alpha-\beta^2+2\sqrt{\alpha^2-\alpha\beta^2}]^2\sqrt{\alpha^2-\alpha\beta^2}\sqrt{\alpha^2(\beta+1)^2-4\alpha\beta^3}.$$

With (α,β) subject to (7.2.7), the two expressions appearing under square root signs are positive, and

$$f_2(\alpha,\beta),\quad f_3(\alpha,\beta),\quad f_5(\alpha,\beta) > 0 \tag{7.2.9}$$

Also

$$f_3(\alpha,\beta)^2-f_4(\alpha,\beta)^2=\alpha\beta^2(\beta-1)^2[8\alpha\beta[\alpha(\beta+1)^2-4\beta^3]+[\alpha(\beta-1)-2\beta^3]^2]. \tag{7.2.10}$$

The right hand side of (7.2.10) is positive if (α,β) is subject to (7.2.7) so that

$$f_3(\alpha,\beta) > |f_4(\alpha,\beta)|. \tag{7.2.11}$$

Substituting (7.2.9), (7.2.11) in (7.2.8) gives $\frac{df_1(\alpha,\beta)}{d\alpha} < 0$, so that for any fixed $\beta\epsilon(1,2)$, $f_1(\alpha,\beta)$ is monotonically decreasing for $\alpha \geq \frac{\beta}{2-\beta}$.

Therefore the maximum of (7.2.6), (7.2.7) is the maximum of $f_1(\frac{\beta}{2-\beta},\beta)$, $\beta\epsilon(1,2)$. We show that $f_1(\frac{\beta}{2-\beta},\beta)$ is an increasing function of β in this region. Define

$$g_1(\beta) = f_1(\tfrac{\beta}{2-\beta},\beta) = [(2\beta^3-3\beta^2+2\beta+1)+(\beta^2-1)\sqrt{4\beta+1}]/2\beta^3$$

Then

$$\frac{dg_1(\beta)}{d\beta} = \frac{g_2(\beta)+g_3(\beta)}{2\beta^4\sqrt{4\beta+1}}$$

$$g_2(\beta) \triangleq (-2\beta^3-\beta^2+10\beta+3),$$

$$g_3(\beta) \triangleq (3\beta^2-4\beta-3)\sqrt{4\beta+1}.$$

For $\beta\varepsilon(1,2)$, $(4\beta+1) > 0$, $g_2(\beta) > 0$, and $g_2(\beta) > |g_3(\beta)|$ as

$$g_2(\beta)^2-g_3(\beta)^2 = 4\beta^4(2-\beta)(6-\beta) > 0.$$

Therefore $\frac{dg_1(\beta)}{d\beta} > 0$ for $\beta\varepsilon(1,2)$. Therefore the maximum of (7.2.6), (7.2.7) has a least upper bound of $g_1(2) = \frac{9}{8}$.

Lemma 7.2.3 Let $n>2$, $B\varepsilon L(R^n)$ symmetric and positive definite; $s,y\varepsilon R^n$ such that $<s,y> > 0$ and y,Bs are linearly independent; ϕ a real parameter; $\bar{B}(\phi)$ given by (7.0.1); a,b,c defined by (7.1.1). Then if $a>b>c$ and $2ac \leq b(a+c)$, the OCX update is the DFP, the OC update is the SR1, and $K(DFP) < \frac{32}{27} K(OC)$ $(= \frac{32}{27} K(SR1))$.

Proof: The fact that the OCX update is the DFP and the OC is the SR1 is represented by the dotted line in Figure (7.1.1b). Their condition numbers are given by (7.2.1) and (7.2.3b).

We again define α, β by (7.2.5). Therefore $\alpha > \beta > 1$, $\alpha > \beta^2$, and from also $b \geq \frac{2ac}{a+c}$, we find either $\beta\varepsilon(1,2)$ and $\alpha\varepsilon(\beta^2,\frac{\beta}{2-\beta}]$ or $\beta\varepsilon[2,\infty)$ and $\alpha\varepsilon(\beta^2,\infty)$. Therefore we require the maximum value of

$$f_6(\alpha,\beta) = \frac{K(DFP)}{K(OC)} = \frac{\alpha(\beta+1)^2-2\beta^3+(\beta+1)\sqrt{\alpha^2(\beta+1)^2-4\alpha\beta^3}}{2\beta^3} \cdot \frac{\beta-1}{\alpha-\beta} \tag{7.2.12}$$

subject to

$$\beta\varepsilon(1,2),\ \alpha\varepsilon(\beta^2,\tfrac{\beta}{2-\beta}] \tag{7.2.13}$$

or to

$$\beta\varepsilon[2,\infty),\ \alpha\varepsilon(\beta^2,\infty). \tag{7.2.14}$$

We show that this maximum has a least upper bound of $\frac{32}{27}$, which is $\lim\limits_{\alpha\to\infty} f(\alpha,3)$.

We consider first $f_6(\alpha,\beta)$ as a function of α with $\beta\varepsilon(1,\infty)$ fixed. Then

$$\frac{df_6(\alpha,\beta)}{d\alpha} = \frac{(\beta-1)}{2\beta^3}\cdot\frac{1}{(\alpha-\beta)^2}\cdot\frac{1}{\sqrt{\alpha^2(\beta+1)^2-4\alpha\beta^3}}\cdot f_7(\alpha,\beta),$$

$$f_7(\alpha,\beta) \triangleq (\beta+1)[\alpha(\beta^2-2\beta-1)+2\beta^3] + (\beta^2-2\beta-1)\sqrt{\alpha^2(\beta+1)^2-4\alpha\beta^3}.$$

The term under the square root is positive for (7.2.13) or (7.2.14). If $\beta\varepsilon(1,1+\sqrt{2})$, then $(\beta^2-2\beta-1) < 0$, and $f_7(\alpha,\beta) = 0$ for $\alpha = -\beta(\beta+1)^2/2(\beta^2-2\beta-1)$, which is the maximum point of (7.2.12), (7.2.13), (7.2.14) for fixed $\beta\varepsilon(1,1+\sqrt{2})$. For this value of α, $f_6(\alpha,\beta) = \frac{4\beta}{3\beta+1}$, which for $\beta\varepsilon(1,1+\sqrt{2})$ has a strict upper bound of $\lim\limits_{\beta\to1+\sqrt{2}} \frac{4\beta}{3\beta+1} = 2(2-\sqrt{2}) = 1.172$.

If $\beta \geq 1+\sqrt{2}$, $(\beta^2-2\beta-1) \geq 0$ and $f_7(\alpha,\beta) > 0$ for (α,β) subject to (3.11). Therefore in this case $\frac{df_6(\alpha,\beta)}{d\alpha} > 0$ for $\alpha > \beta^2$, so that the maximum of (7.2.12), (7.2.14) for fixed $\beta \geq 1+\sqrt{2}$ has upper bound

$$\lim_{\alpha\to\infty} f_6(\alpha,\beta) = (\beta-1)(\beta+1)^2/\beta^3.$$

This function of β has maximum value $\frac{32}{27} = 1.\overline{185}$ at $\beta = 3$. Therefore the maximum value of (7.2.12) subject to (7.2.13) or (7.2.14) has a strict upper bound of $\lim_{\alpha\to\infty} f_6(\alpha,3) = \frac{32}{27}$.

In Case 3, c>b>a, we prove the analogs of Lemmas 7.2.2 and 7.2.3 where the BFGS replaces the DFP. The proofs proceed identically if we simply replace (7.2.5) by

$$c = \alpha a, \quad b = \beta a$$

Therefore we have proven the main result of this section.

Theorem 7.2.4 Let $n>2$, $B \in L(R^n)$ symmetric and positive definite; $s,y \in R^n$ such that $\langle s,y\rangle > 0$ and y,Bs are linearly independent; ϕ a real parameter; $\bar{B}(\phi)$ given by (7.0.1); a,b,c defined by (7.1.1). Let $\bar{B}(\phi_{OC})$ and $\bar{B}(\phi_{OCX})$ be Davidon's optimally conditioned update and the optimally conditioned update in the convex class, respectively -- the updates which minimize the condition number of $B^{-1/2}\bar{B}(\phi)B^{-1/2}$ among all positive definite $\bar{B}(\phi)$, and among all $\bar{B}(\phi)$ with $\phi\in[0,1]$, respectively. Let K(OC), K(OCX) be the condition numbers of $B^{-1/2}\bar{B}(\phi)B^{-1/2}$ for $\phi = \phi_{OC}, \phi_{OCX}$ respectively. Then $K(OCX) < \frac{32}{27} K(OC)$.

Furthermore, the OC update is in the convex class, and therefore is the OCX, if $b \le a$ and $b \le c$. Otherwise, the OC update is not in the convex class, and the OCX is the BFGS if $a < c$, the DFP if $a > c$.

The 32/27 maximum ratio of K(OCX) to K(OC) is a very small difference, and not by itself reason to go outside the convex class in choosing $\bar{B}$. Davidon was concerned about differentiating between updates for which $K(\bar{B}(\phi))$ differ by orders of magnitude, and mentions that when $b \ll a,c$, K(OC) is much smaller than K(DFP) or K(BFGS). We show that this is true in Lemma 7.2.5. Note that $b \ll a,c$ can only occur in Case 1. However in Lemma 7.2.6 we show that in Case 2, while $K(OCX) = K(DFP) < \frac{32}{27} K(OC)$, $K(DFP) \ll K(BFGS)$ is possible. The same is true in Case 3 if the roles of the DFP and BFGS are interchanged. These lemmas show that from the viewpoint of conditioning, it may be important which update in the convex class we select.

<u>Lemma 7.2.5</u> Let a,b,c be positive numbers such that $b<a,c$, and let K(DFP), K(BFGS), K(OC) be given by (7.2.1), (7.2.2), (7.2.3a) respectively. Then for any $0<\varepsilon<1$, $K(OC) < \varepsilon \cdot \min\{K(DFP), K(BFGS)\}$ is possible.

<u>Proof</u>: Let $\frac{b}{a} = \frac{b}{c} = \frac{\varepsilon}{4}$, $0<\varepsilon<1$. Then $K(DFP) = K(BFGS) > 64\varepsilon^{-3}$; $K(OC) < 64\varepsilon^{-2}$. Therefore $K(OC) < \varepsilon \cdot \min$ K(DFP), K(BFGS) .

<u>Lemma 7.2.6</u> Let a,b,c be positive numbers such that $a>b>c$, and let K(DFP), K(BFGS) be given by (7.2.1) and (7.2.2). Then for any $0<\varepsilon<1$, $K(DFP) < \varepsilon \cdot K(BFGS)$ is possible.

<u>Proof</u>: Rewrite K(DFP), K(BFGS) as

$$K(DFP) = \left[(1+\tfrac{b}{c})^2 - 2(\tfrac{b}{a})(\tfrac{b}{c})^2 + (1+\tfrac{b}{c})\sqrt{(1+\tfrac{b}{c})^2 - 4(\tfrac{b}{a})(\tfrac{b}{c})^2}\right] / 2(\tfrac{b}{a})(\tfrac{b}{c})^2 \tag{7.2.15}$$

$$K(BFGS) = \left[(1+\tfrac{b}{a})^2 - 2(\tfrac{b}{a})^2(\tfrac{b}{c}) + (1+\tfrac{b}{a})\sqrt{(1+\tfrac{b}{a})^2 - 4(\tfrac{b}{a})^2(\tfrac{b}{c})}\right] / 2(\tfrac{b}{a})^2(\tfrac{b}{c}) \tag{7.2.16}$$

Let $\frac{b}{a} = \frac{\varepsilon}{4}$, $\frac{b}{c} = 1 + \frac{\varepsilon}{4}$, $0<\varepsilon<1$. Then from (7.2.15), (7.2.16), $K(DFP) < 16\varepsilon^{-1}$, $K(BFGS) > 16\varepsilon^{-2}$. Therefore $K(DFP) < \varepsilon \cdot K(BFGS)$.

Taken together with Theorem 7.2.4, Lemmas 7.2.5 and 7.2.6 show that while restricting our update to the convex class is probably not harmful from the viewpoint of conditioning, always selecting the BFGS (or DFP) may be. From the perspective of Figures 7.1.1a-c, we have shown that the curves $K(\phi)$ hardly rise between ω_{OC} and the nearest ω in the convex class, but that they may rise sharply within the convex class. However in Section 7.3 we show that they are also likely to be flat within the convex class in many cases, and our computational results in Chapter 8 indicate that in practice this occurs most of the time.

When $n=2$, Lemmas 7.2.3 and Theorem 7.2.4 are no longer true because the OC update is not the SR1 when $2ac < b(a+c)$, and then K(OC) may be much less than K(OCX). This is not considered important because we are not very concerned with two-dimensional problems.

Section 7.3 - Optimal Conditioning in Q-Superlinearly Convergent Algorithms

In Section 7.2 we showed that for any update class of form (7.0.1), $K(OCX) < \frac{32}{27} K(OC)$, but K(BFGS) (or K(DFP)) may be much greater than K(OCX). Therefore from the standpoint of conditioning it looks like we can restrict our updates to the convex class, but not necessarily to the BFGS. However, in this section we show that for the update classes we actually use, it is likely that K(BFGS) (and $K(\phi)$ for any $\phi\in[0,1]$) approaches 1 as we near the solution, so that Davidon's (1975) and Oren and Spedicato's (1976) optimal conditioning becomes unimportant.

Powell (1976a) and Broyden, Dennis, and Moré (1973) have shown that when $s = \bar{x}-x$ and $y = g(\bar{x})-g(x)$ in 7.0.1, the DFP and BFGS updates achieve local Q-superlinear convergence on a wide class of problems. In Theorem

7.3.1 we show that under the assumptions of Broyden, Dennis, and Moré, if the BFGS algorithm converges Q-superlinearly, then K(BFGS) approaches 1 at successive iterations. This result makes us suspect that optional conditioning will usually be unimportant in the later iterations of algorithms using the standard DFP-BFGS class.

Davidon's (1975) recent algorithm uses different values of s and y in (7.0.1). His algorithm has been more successful than the standard BFGS in practice, so we expect it also achieves local Q-superlinear convergence on many problems. In Theorem 7.3.2 we show that if Davidon's algorithm using the restricted BFGS converges Q-superlinearly and several reasonable assumptions are met, then K(BFGS) will also approach 1 in this case. This leads us to suspect that optimal conditioning is also often unimportant at later iterations in algorithms using Davidon's restricted updates. Our tests in Chapter B confirm this suspicion.

We use the H form of the update in Theorems 7.3.1 and 7.3.2 as it is easier to analyze the BFGS in this form. All norms are ℓ_2 or its induced matrix norm.

Theorem 7.3.1 Let $f:R^n \to R$ be twice differentiable in the open convex set D, and suppose for some $x^* \epsilon D$ and $p>0$, $K \geq 0$

$$\| G(x)-G(x^*) \| \leq K \| x-x^* \|^p$$

where $g(x^*) = 0$ and $G(x^*)$ is symmetric and positive definite. Consider the sequence $(x_1, H_1, x_2, H_2, \ldots)$ generated from $x_0 \epsilon R^n$, $H_0 \epsilon L(R^n)$ symmetric and positive definite, by the BFGS algorithm

$$x_{i+1} = x_i - H_i g(x_i)$$

$$H_{i+1} = H_i + \frac{(s_i-H_iy_i)s_i^T+s_i(s_i-H_iy_i)^T}{<s_i,y_i>} - \frac{<s_i-H_iy_i,y_i>s_is_i^T}{<s_i,y_i>^2}$$

where $s_i = x_{i+1}-x_i$, $y_i = g(x_{i+1})-g(x_i)$, and $<s_i,y_i> > 0$ is assumed, $i = 0,1,\ldots$. If $\{x_i\}$ converges Q-superlinearly to x^*, then $\lim_{i\to\infty} \| H_{i+1}-H_i \| = 0$ and $\lim_{i\to\infty} K(H_i^{-1/2}H_{i+1}H_i^{-1/2}) = 1$.

Proof: Define $A = G(x^*)^{-1}$. Broyden, Dennis, and Moré (1973) show that under the assumptions of this theorem, there exists a constant $\rho>1$ and a positive integer i_o such that for all $i > i_o$

$$<s_i,y_i> \geq \frac{2}{3} \| s_i \|^2/\| A\| \qquad (7.3.1)$$

$$\| s_i \|/\rho \leq \| y_i \| \leq \rho \| s_i \| \qquad (7.3.2)$$

$$\frac{\| s_i-Ay_i \|}{\| s_i\|} \leq K\| A\| \max \{\| x_{i+1}-x^* \|^p, \| x_i-x^* \|^p\} \qquad (7.3.3)$$

They also show that under these assumptions

$$\lim_{i\to\infty} \frac{\| (A-H_i)y_i \|}{\| y_i \|} = 0 \qquad (7.3.4)$$

Thus since

$$\| H_{i+1}-H_i \| \leq \frac{2\| s_i-H_iy_i \| \|s_i\|}{<s_i,y_i>} + \frac{\| s_i-H_iy_i\| \|y_i\| \|s_i\|^2}{<s_i,y_i>^2}$$

using (7.3.1), (7.3.2) and the triangle inequality we find

$$\| H_{i+1}-H_i \| \leq \frac{3\| s_i-H_iy_i\| \|A\|}{\| s_i \|} + \frac{\frac{9}{4}\rho \| s_i-H_iy_i\| \|A\|^2}{\| s_i \|}$$

$$\leq (3+\frac{9}{4}\rho \| A\|) \| A\| \left[\frac{\| s_i - Ay_i \|}{\| s_i \|} + \frac{\rho \| (A-H_i)y_i \|}{\| y_i \|}\right]$$

Therefore from (7.3.3), (7.3.4), and $\{x_i\} \to x^*$, $\lim_{i\to\infty} \| H_{i+1} - H_i \| = 0$.

Dennis and Moré (1974) show that under the conditions of this theorem, H_i is positive definite for all i and $\{\| H_i^{-1} \|\}$ is uniformly bounded. Therefore $H_i^{-1/2}$ exists, $\{\| H_i^{-1/2} \|\}$ is uniformly bounded, and $\lim_{i\to\infty} H_i^{-1/2} H_{i+1} H_i^{-1/2} = \lim_{i\to\infty} (I + H_i^{-1/2}(H_{i+1} - H_i) H_i^{-1/2}) = I$. Thus $\lim_{i\to\infty} K(H_i^{-1/2} H_{i+1} H_i^{-1/2}) = 1$.

Under the assumptions of Theorem 7.3.1, we can also show that K(DFP) approaches 1 (using the B form of the update). This means that at a near final iteration, the curves in Figures 7.1.1a-c, while rising at their ends, will virtually equal the line $K(\phi) = 1$ for a region including the OC update and the entire convex class. The OC update may still be outside the convex class, but there is little point to the optimal conditioning strategy in this case, as $K(\phi) = 1$ for established and successful updates.

Davidon's (1975) recent algorithm uses updates from the restricted DFP-BFGS class, which has the same algebraic form as (7.0.1) but different meanings of s and y. In Chapter 3 we showed that this class is equivalent to a v-class (3.1.3), or in H form to a w-class (3.3.2). For example the restricted BFGS can be written

$$H_{i+1} = H_i + \frac{(s_i - H_i y_i) w_i^T + w_i (s_i - H_i y_i)^T}{\langle w_i, y_i \rangle} - \frac{\langle s_i - H_i y_i, y_i \rangle w_i w_i^T}{\langle w_i, y_i \rangle^2}$$

(7.3.5)

where s and y still are $\bar{x}-x$ and $g(\bar{x})-g(x)$, respectively, and $w_i = Q_i s_i$ is the projection in the metric H_i^{-1} of s_i onto some subspace including $s_i - H_i y_i$. In Theorem 7.3.2 we show that if a direct prediction algorithm using update (7.3.5) (with any value of w_i) converges Q-superlinearly and in addition

$$\langle w_i, y_i \rangle > \alpha \| w_i \| \, \| y_i \| \qquad (7.3.6)$$

for some $\alpha>0$ and $\{\| H_i \|\}$, $\{\| H_i^{-1} \|\}$ are uniformly bounded for all $i >$ some constant i_1, then $\lim_{i\to\infty} K(H_i^{-1/2} H_{i+1} H_i^{-1/2}) = 1$. We think these assumptions are reasonable ones in the case of the restricted BFGS. In particular, $\langle w_i, y_i \rangle = \langle Q_i s_i, y_i \rangle$ must then be positive for any positive definite updates to exist, and most implementations will use a condition like (7.3.6) to switch to the standard DFP-BFGS class when w_i and y_i get too close to being perpendicular.

<u>Theorem 7.3.2</u> Let $f: R^n \to R$ satisfy the hypothesis of Theorem 7.3.1. Consider the sequence $(x_1, H_1, x_2, H_2, \ldots)$ generated from $x_0 \epsilon R^n$, $H_0 \epsilon L(R^n)$ symmetric by the iteration

$$x_{i+1} = x_i - H_i g(x_i)$$

$$H_{i+1} \text{ given by (7.3.5)}$$

where $s_i = x_{i+1} - x_i$, $y_i = g(x_{i+1}) - g(x_i)$, and $w_i \epsilon R^n$ obeys $\langle w_i, y_i \rangle \neq 0$, $i = 0,1,\ldots$. Then if $\{x_i\}$ converges Q-superlinearly to x^*, and if there exists a constant $\alpha>0$ and a positive integer i_1 such that for all $i > i_1$, (7.3.6) holds and $\{\| H_i \|\}$, $\{\| H_i^{-1} \|\}$ are uniformly bounded, then

[illegible]

Proof: The proof is similar to that of Theorem 7.3.1, using (7.3.6) in place of (7.3.1). Again define $A = G(x^*)^{-1}$. Equations (7.3.2) and (7.3.3) still follow from our assumptions on f. Equation (7.3.4) is again true because Dennis and Moré (1974) show that under the assumptions of this theorem,

$$\lim_{i\to\infty} \frac{\| (A^{-1}-H_i^{-1})s_i \|}{\| s_i \|} = 0$$

and this is easily shown to imply (7.3.4) if $\{\| H_i \|\}$ is uniformly bounded and (7.3.2), (7.3.3) hold. Therefore since

$$\| H_{i+1}-H_i \| \le \frac{2\| s_i-H_iy_i \| \|w_i\|}{\langle w_i, y_i\rangle} + \frac{\| s_i-H_iy_i \| \|y_i\| \|w_i\|^2}{\langle w_i, y_i\rangle^2},$$

(7.3.6) gives

$$\| H_{i+1}-H_i \| \le \frac{2\| s_i-H_iy_i \|}{\alpha\| y_i \|} + \frac{\| s_i-H_iy_i \|}{\alpha^2\| y_i \|}$$

The remainder of the proof is virtually identical to that of Theorem 7.3.1.

The results of Section 7.2 and 7.3 lead us to propose in Chapter 8 a quasi-Newton algorithm which uses the (restricted) BFGS update unless $K(BFGS) > \psi\cdot K(OCX)$ for some constant ψ ($\psi = 1.5$ seems good) in which case it uses the (restricted) OCX. We find that in our tests this updating strategy usually works considerably better than Davidon's optimal conditioning in his algorithm, and slightly better than the restricted BFGS.

Section 7.4 - Optimal Conditioning and Restricted Updates

Finally in Chapter 7, we examine the effect of restricting the subspace of our updates on the condition number of $\bar{B}$. One restricts the updating space so that the update will preserve derivative information from previous iterations. Therefore it seems reasonable that this additional property will be offset by a rise in the condition number of $\bar{B}$. Theorem 7.4.1 shows that this is likely to be true, by showing that the minimum bound K(OC) for the standard DFP-BFGS class is less than or equal to the minimum bound K(OCR) for any restricted DFP-BFGS class.

<u>Theorem 7.4.1</u> Let $B \in L(R^n)$ be symmetric and positive definite; $s,y \in R^n$ such that $\langle s,y \rangle > 0$ and y, Bs are linearly independent; ϕ a real parameter; $\bar{B}(\phi)$ given by (7.0.1). Let Z be an m dimensional subspace of R^n including $y-Bs$, $m \leq n-2$; let P be the matrix projecting onto Z in the B^{-1} metric, define $\hat{y} = Py$, $\hat{s} = P^T s$, and assume $\hat{s} \neq s$, $\hat{y} \neq y$, $\langle \hat{s},\hat{y} \rangle > 0$, and $\hat{y}, B\hat{s}$ linearly independent. Let $\hat{B}(\hat{\phi})$ be the class of updates given by (7.0.1) if s, y, and ϕ are replaced by $\hat{s}, \hat{y}$, and $\hat{\phi}$ respectively. Let $\bar{B}(\phi_{OC})$ and $\hat{B}(\hat{\phi}_{OCR})$ be the optimally conditioned updates from the classes $\bar{B}(\phi)$ and $\hat{B}(\hat{\phi})$, respectively -- the positive definite updates which minimize the condition number of $B^{-1/2}\bar{B}(\phi)B^{-1/2}$ and $B^{-1/2}\hat{B}(\hat{\phi})B^{-1/2}$, respectively. Let $\bar{B}(SR1)$ be the SR1 update. Let K(OC), K(OCR) and K(SR1) = the condition numbers of $B^{-1/2}\bar{B}(\phi_{OC})B^{-1/2}$, $B^{-1/2}\hat{B}(\hat{\phi}_{OCR})B^{-1/2}$, and $B^{-1/2}\bar{B}(SR1)B^{-1/2}$ respectively. Then $K(OC) \leq K(OCR)$, with equality if and only if $\bar{B}(OC) = \hat{B}(OCR) = \bar{B}(SR1)$.

<u>Proof</u>: Let a,b,c be defined by (7.1.1). Then by Theorem 7.2.4, [illegible] $(c-b)/(ac-b^2)$ if $2ac > b(a+c)$, $\bar{B}(\phi_{OC}) = \bar{B}(SR1)$ if $2ac \leq b(a+c)$.

Similarly, define $\hat{a} = \langle\hat{y},B^{-1}\hat{y}\rangle$, $\hat{b} = \langle\hat{s},\hat{y}\rangle$, $\hat{c} = \langle\hat{s},B\hat{s}\rangle$. Then again by Theorem 7.2.4, $\hat{\phi}_{OCR} = \hat{a}(\hat{c}-\hat{b})/(\hat{a}\hat{c}-\hat{b}^2)$ if $2\hat{a}\hat{c} > \hat{b}(\hat{a}+\hat{c})$, $\hat{B}(\hat{\phi}_{OCR}) = \overline{B}(SR1)$ if $2\hat{a}\hat{c} \leq \hat{b}(\hat{a}+\hat{c})$. We show that if $\overline{B}(\phi_{OC}) = \overline{B}(SR1)$ then $\hat{B}(\hat{\phi}_{OCR}) = \overline{B}(SR1)$; while if $\overline{B}(\phi_{OC}) \neq \overline{B}(SR1)$, then $\hat{B}(\hat{\phi}_{OCR})$ may or may not be $\overline{B}(SR1)$, and $K(OC) < K(OCR)$.

By Corollary 4.2.1, $b = \hat{b}+\delta$ for some $\delta>0$. From the proof of Theorem 3.1.1, $\langle\hat{y}-B\hat{s},\hat{s}\rangle = \langle y-Bs,s\rangle$; thus $c = \hat{c}+\delta$. Similarly $\langle\hat{s}-B^{-1}\hat{y},\hat{y}\rangle = \langle s-B^{-1}y,y\rangle$, so that $a = \hat{a}+\delta$. Therefore $2ac-b(a+c) = 2\hat{a}\hat{c}-\hat{b}(\hat{a}+\hat{c})+\delta[\hat{a}-2\hat{b}+\hat{c}]$. The term $\hat{a}-2\hat{b}+\hat{c}$ is positive since it equals $(\hat{y}-B\hat{s})^TB^{-1}(\hat{y}-B\hat{s})$. Hence when $\overline{B}(\phi_{OC}) = \overline{B}(SR1)$, $2ac-b(a+c) \leq 0$ implies $2\hat{a}\hat{c}-\hat{b}(\hat{a}+\hat{c}) < 0$, so that $\hat{B}(\hat{\phi}_{OCR}) = \overline{B}(SR1)$ also, and $K(OC) = K(OCR)$.

If $2ac > b(a+c)$, $\overline{B}(\phi_{OC}) \neq \overline{B}(SR1)$, and $2\hat{a}\hat{c}-\hat{b}(\hat{a}+\hat{c})$ may be positive or negative. If it is not positive, $\hat{B}(\phi_{OCR}) = \overline{B}(SR1)$. Therefore $\overline{B}(SR1)$ is positive definite, and since it is not the optimally conditioned update from the class $\overline{B}(\phi)$, $K(OC) < K(SR1)$. Thus $K(OC) < K(OCR)$.

Finally if $2ac > b(a+c)$ and $2\hat{a}\hat{c} > \hat{b}(\hat{a}+\hat{c})$; $K(OC) = \lambda_+/\lambda_-$, where $\lambda_\pm = \frac{a}{b} \pm \sqrt{(\frac{a}{b})^2-(\frac{a}{c})}$ by (7.1.2) with $\phi = a(c-b)/(ac-b^2)$. Similarly $K(OCR) = \hat{\lambda}_+/\hat{\lambda}_-$, $\hat{\lambda}_\pm = \frac{\hat{a}}{\hat{b}} \pm \sqrt{(\frac{\hat{a}}{\hat{b}})^2 - \frac{\hat{a}}{\hat{c}}}$. Therefore $\lambda_+/\lambda_- = (1+\sqrt{1-\alpha})/(1-\sqrt{1-\alpha})$, $\alpha = b^2/ac$, $\hat{\lambda}_+/\hat{\lambda}_- = (1+\sqrt{1-\hat{\alpha}})/(1-\sqrt{1-\hat{\alpha}})$, $\hat{\alpha} = \hat{b}^2/\hat{a}\hat{c}$, and simple algebra shows that $\lambda_+/\lambda_- < \hat{\lambda}_+/\hat{\lambda}_-$ if and only if $\hat{\alpha}<\alpha$. Substituting $a,b,c = \hat{a}+\delta$, $\hat{b}+\delta$, $\hat{c}+\delta$ respectively, $\alpha-\hat{\alpha} = [\delta^2(\hat{a}\hat{c}-\hat{b}^2) + \delta\hat{b}(2\hat{a}\hat{c}-\hat{b}(\hat{a}+\hat{c}))]/ac\hat{a}\hat{c}$. Since $\hat{a}\hat{c}-\hat{b}^2 > 0$ by 7.1.4, $\hat{\alpha}<\alpha$ and $K(OC) < K(OCR)$.

The techniques of Theorem 7.4.1 apply equally well if we compare two restricted updating classes projecting orthogonal to subspaces Z_1 and Z_2 respectively, where $Z_1 \subset Z_2$. In this case they show that the minimum bound

$K(OCR_1)$ for the less restricted class is less than or equal to the minimum bound $K(OCR_2)$ for the more restricted class. Finally, it is interesting to note that, if K(BFGS) is the bound for the standard BFGS and K(BFGS-R) the bound for the restricted BFGS, then K(BFGS) is not necessarily less than or equal to K(BFGS-R), although it is whenever $2ac \geq b(a+c)$.

CHAPTER VIII

OPTIMAL CONDITIONING - PRACTICE

In Chapter 7 we discussed the use of optimal conditioning to choose the parameter ϕ in the DFP-BFGS update

$$\overline{B}(\phi) = B + \frac{(y-Bs)y^T + y(y-Bs)^T}{\langle y,s\rangle} - \frac{\langle y-Bs,s\rangle yy^T}{\langle y,s\rangle^2}$$

$$- \phi\langle s,Bs\rangle \left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]\left[\frac{y}{\langle y,s\rangle} - \frac{Bs}{\langle s,Bs\rangle}\right]^T \qquad \text{(B.0.1)}$$

where $s,y \in R^n$, B, $\overline{B} \in L(R^n)$, $\phi \in R$. Our theoretical results lead us to wonder whether it is better to use the "optimally conditioned" (OC) update (the positive definite update $\overline{B}(\phi_{OC})$, ϕ_{OC} given by (7.1.9), which minimizes a bound on the condition number of $\overline{B}(\phi)$), or the optimally conditioned update restricted to the convex class (OCX, the update $\overline{B}(\phi_{OCX})$ which minimizes the same bound among $\phi \in [0,1]$), or whether either will be preferable to the BFGS ($\phi=1$).

Davidon (1975) did not explicitly test the optimal conditioning idea when he introduced it, for he introduced optimal conditioning and projected updates simultaneously, and only compared an algorithm using both to an algorithm using neither. Mei (1977) was the first to compare the BFGS and OC updates, in his MINOP algorithm which uses Davidon's projected updates and the dogleg strategy for selecting search directions. He found the BFGS to give clearly superior results in his tests. In this chapter we report on tests of various strategies for choosing ϕ in both the MINOP and Davidon's (1975) OCOPTR algorithm. We find that the BFGS performs better than either the OC or the OCX most of the time. Furthermore, our data confirms our theoretical analysis by showing that the BFGS is as well conditioned as the OCX (as measured by Davidon's bounds) most but not all of the time. This suggests the use of a hybrid BFGS-OCX update which uses the OCX only when it is significantly better conditioned than the BFGS. We test such an update and find that it is often but not always an improvement over the BFGS. We think the use of this update requires further consideration and testing.

In Section B.1 we define our BFGS-OCX update, as well as a hybrid BFGS-OC update which we also test. In Section B.2 we present and analyze our computational test results. Finally in Section B.3 we prove that when the BFGS-OCX update is used in the standard DFP-BFGS class, it does not change the local Q-superlinear convergence of the BFGS algorithm, because asymptotically only the BFGS update is used. This result, which is confirmed in practice, strengthens our confidence in the BFGS-OCX hybrid update.

Section B.1 - The BFGS-OCX Update

Our analysis in Chapter 7 showed that the bound on the condition number of the BFGS update, K(BFGS), may be much larger than the bound on the condition number of the OCX update, K(OCX), although near the solution they are likely to be almost equivalent. In our tests we find that the ratio K(BFGS)/K(OCX) is hardly greater than 1 at most iterations, but on a few iterations per run it may be larger, occasionally greater than 10 and very rarely greater than 100. Therefore it seems natural to propose an algorithm which selects the OCX update on those few iterations when K(BFGS)/K(OCX) is significantly greater than 1, and the BFGS otherwise. Thus we define the BFGS-OCX update as

$$\bar{B}_{BFGS\text{-}OCX} = \begin{cases} \bar{B}(\phi_{OCX}) & K(BFGS) > \psi \cdot K(OCX) \\ \bar{B}(1) & \text{otherwise} \end{cases}$$

where $\psi > 1$ is a constant, $\bar{B}(\phi)$ given by (B.0.1), ϕ_{OCX}, K(OCX), K(BFGS) defined in Chapter 7. We see in Section B.2 that the value we currently recommend for ψ is 1.5. This still results in selecting the OCX update only a very small fraction of the time.

We first used the OCX rather than the OC update in combination with the BFGS because our analysis of Chapter 7 supported the OCX. However the performance of the OC and OCX algorithms in our tests reveal no clear cut favorite. Therefore (since the BFGS-OCX algorithm does appear promising) we have also tested a BFGS-OC hybrid

$$\bar{B}_{BFGS\text{-}OC} = \begin{cases} \bar{B}(\phi_{OC}) & K(BFGS) > \psi \cdot K(OC) \\ \bar{B}(1) & \text{otherwise} \end{cases}$$

where ϕ_{OC} and K(OC) are also given in Chapter 7, and again $\psi > 1$ is a constant. The results of all these tests are given in Section B.2. The comparison between the two hybrid updates is inconclusive.

Section B.2 - Computational Results

We have tested the OC, OCX, BFGS, BFGS-OCX and BFGS-OC updating strategies in what currently seem to be two of the leading unconstrained minimization algorithms: Davidon's (1975) OCOPTR and Dennis and Mei's (1975) MINOP. Both algorithms use Davidon's (1975) projections in making the update, but differ in the techniques used to obtain the search direction and the new point in that direction. In Dennis and Mei's tests their algorithm performs somewhat better than Davidon's. We have made no changes to either algorithm except to vary the process for selecting the update parameter ϕ. Dennis and Mei update an approximation H to the inverse of the Hessian by

$$\bar{H}(\bar{\phi}) = H + \frac{(s-Hy)s^T+s(s-Hy)^T}{<s,y>} - \frac{<s-Hy,y>ss^T}{<s,y>^2}$$

$$- \bar{\phi}<y,Hy>\left[\frac{s}{<s,y>}-\frac{Hy}{<y,Hy>}\right]\left[\frac{s}{<s,y>}-\frac{Hy}{<y,Hy>}\right]^T \qquad \text{(B.2.1)}$$

so that $\bar{\phi}_{BFGS} = 0$,

$$\bar{\phi}_{OC} = \begin{cases} c(a-b)/(ac-b^2) & 2ac > b(a+c) \\ a/(a-b) & \text{otherwise} \end{cases} \qquad \text{(B.2.2)}$$

$$\bar{\phi}_{OCX} = \begin{cases} \bar{\phi}_{OC} & b \leq a,c \\ 0 & a<b<c \\ 1 & a>b>c \end{cases} \qquad \text{(B.2.3)}$$

where $a = <y,Hy>$, $b = <s,y>$, $c = <s,H^{-1}s>$. (Actually Dennis and Mei implement B.2.1 as $\bar{H}(\hat{\phi})$, $\hat{\phi} = 1-\bar{\phi}$). Davidon (1975) also updates $\bar{H}$, but maintains it in a factored form, and performs his entire iteration in the metric weighted by $H^{-1/2}$.

Initially we used the four problems tested by Dennis and Mei (1975),

1) Box's 2-Variable Function (n=2)

$$f(x) = \sum_{i=1}^{10} \left[(e^{-ix_1/10}-e^{-ix_2/10})-(e^{-i/10}-e^{-i})\right]^2$$

with 5 different starts (0,0), (0,20), (5,0), (5,20), (2.5,10)

$x^* = (1,10)$, $f(x^*) = 0$

2) Rosenbrocks Function (n=2)

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

with 5 different starts (-1.2,1), (2,2), (-3.635, 5.621), (6.39, -0.221), (1.489, -2.547)

$x^* = (1,1)$, $f(x^*) = 0$

3) Powell's 4-Variable Function (n=4)

$$f(x) = (x_1+10x_2)^2+5(x_3-x_4)^2+(x_2-2x_3)^4+10(x_1-x_4)^4$$

with 3 different starts (3,-1,0,1), (-0.1,1,-0.1,1), (-0.6,1,-0.6,1)

$x^* = (0,0,0,0)$, $f(x^*) = 0$

4) Woods-Colville's Function (n=4)

$$f(x) = 100(x_2-x_1{}^2)^2+(x_1-1)^2+90(x_4-x_3{}^2)^2+(x_3-1)^2$$
$$+ 10.1[(x_2-1)^2+(x_4-1)^2]+19.8(x_2-1)(x_4-1)$$

with 4 different starts (-3,1,-3,1), (-3,-1,-3,-1),

(-1.2,1,1.2,1), (-1.2,1,-1.2,1)

$x^* = (1,1,1,1)$, $f(x^*) = 0$.

Of these, only Powell's problem has a singular Hessian at the solution. Since this turned out to be a very interesting problem in comparing the OC and BFGS updates, we added a fifth problem, a slight modification of Rosenbrocks function so that $G(x^*)$ is singular:

5) Singular Rosenbrock's Function (n=2)

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

with same 5 starts as for Rosenbrock's Function

$x^* = (1,1)$, $f(x^*) = 0$

We ran each of the five problems from each of their starting points in the OCOPTR and MINOP program, using seven different updates: OC, OCX, BFGS, OCX-BFGS with $\psi = 1.1$, 1.3, 1.5 and OC-BFGS with $\psi = 1.5$. Our stopping criterion in all cases was when the norm of the gradient is less than or equal to 10^{-5}, although Davidon uses a weighted and Dennis and Mei an unweighted norm. All our computation was done in double precision on the IBM 370-168 at Cornell University.

Our results are summarized in Tables B.2.1 and B.2.2. Table B.2.1 lists for each problem and method the number of function and gradient evaluations required to solve the problem from all of the starting points. The totals for MINOP are roughly double those for OCOPTR because each problem was run twice from each starting point in the MINOP algorithm, with two different values of the initial confidence region Δ which the [illegible] requires. (We used the second largest and the second smallest of the five values of Δ tested by Dennis and Mei.) The equivalent function evaluations equal the number of function evaluations plus the number of gradient evaluations times the dimension of the problem. This represents the total number of function evaluations required to solve the problems if the gradients are computed by finite differences, and is considered the most accurate measure of the cost of many practical hard minimization problems, where f(x) is an expensive routine whose derivatives are not analytically available.

Table B.2.2 compares the performance of the updates on each problem by showing, for the OCOPTR and MINOP routines, what per cent fewer or more equivalent function evaluations each method required than the BFGS. This table allows for much easier comparison of the methods than the raw data of Table B.2.1.

It was not our intent to further compare the OCOPTR and MINOP routines, although our data in general confirm the superiority of MINOP. Our most striking conclusion is the superiority of the BFGS update to the CC (or OCX) in either routine, except perhaps on problems with singular Hessians at the solution. It seems reasonable that the OC method will perform better than the BFGS on such problems, as it is likely to keep the Hessian approximations further from singularity, and thus avoid numerical difficulty in computation of the Newton direction. Thus we expected our singular Rosenbrock problem to give results similar to Powell's function. However, while it is a very hard problem and definitely displays the linear convergence characteristic of singular problems, it has just the opposite results as for Powell's problem: the OC update fares considerably more poorly than the BFGS. Thus it seems that the BFGS is clearly superior to the OC update in our tests on non-singular problems, but that more testing is called for with singular problems. Testing on problems of higher dimension is also still necessary. The results of such tests will influence the ultimate use of optimal conditioning.

The comparison between the OC and OCX updates is inconclusive, and probably unimportant as neither strategy is likely to be recommended in general. However, the hybrid updates appear more interesting. It is clear that at a value of ψ like 1.1 they are inferior to the BFGS in our tests, but by the time $\psi = 1.5$ our output shows that they select the BFGS at least 90% of the time, and Table B.2.2 indicates that they may be slightly preferable to the BFGS, especially on singular problems. Thus they may represent a very good compromise between wanting to use the BFGS on nonsingular problems and maybe wanting optimal conditioning on singular problems. It is clear that more testing is called for in this regard, especially on problems of higher dimension. The OC-BFGS looks slightly inferior to the OCX-BFGS (with $\psi = 1.5$) but the data is not conclusive.

The values of ψ tested were somewhat arbitrary, but stemmed partly from the observation (from K(OC), K(OCX), K(BFGS), K(OCX)/K(OC) and K(BFGS)/K(OCX), which we printed at each step of each problem and method) that K(BFGS)/K(OCX) was greater than 1.5 only a few per cent of the time. However it is possible that a higher value of ψ may produce better results. It also needs to be seen what effect an increase in dimension has upon these ratios and the performance of our algorithms, and it may be that ψ need vary with the dimension of the problem.

Our data on the condition number bounds confirmed nicely our theoretical analysis of Chapter 7. The ratio K(OCX)/K(OC) never exceded 32/27, and exceeded 9/8 only when the OC update was the SR1. It did approach these bounds fairly closely on occasion. The ratio K(BFGS/K(OCX) got as large as 100 occasionally and 1000 very rarely, but as previously stated was below 1.5 over 90% of the time. Finally, on problems where there seemed to be numerical Q-superlinear convergence (Box's, Rosenbrock's, and Woods-Colville's) the three condition number bounds seemed indeed to converge to 1, while on problems with strict linear convergence (Powell's and Singular Rosenbrock) they did not.

TABLE B.2.1: Function, Gradient, and Equivalent Function Evaluations Required To Solve Test Problems by Various Methods

a) OCOPTR Algorithm

Problem	Update:	OC	OCX	BFGS	BFGS-OCX $\psi=1.1$	BFGS-OCX $\psi=1.3$	BFGS-OCX $\psi=1.5$	BFGS-OC $\psi=1.5$
Box	F	94	97	97	98	97	97	97
n=2	G	86	91	88	89	88	88	88
5 starts	EFA	266	279	273	276	273	273	273
Powell		96	118	127	110	120	120	111
n=4		77	99	106	91	101	101	92
3 starts		404	514	551	474	524	524	479
Rosenbrock		244	232	224	222	231	224	217
n=2		213	200	187	191	193	187	184
5 starts		670	632	598	604	617	598	585
Woods-Colville		372	365	350	353	326	340	338
n=4		266	261	247	247	239	246	242
4 starts		1436	1409	1338	1341	1282	1324	1306
Singular Rosenbrock		454	458	397	473	392	379	403
n=2		418	448	354	459	347	337	359
5 starts		1290	1354	1105	1576	1086	1053	1121

b) MINOP Algorithm

Problem	Update: OC	OCX	BFGS	BFGS-OCX ψ=1.1	BFGS-OCX ψ=1.3	BFGS-OCX ψ=1.5	BFGS-OC ψ=1.5
Box	204	209	191	203	200	197	202
n=2	170	175	157	170	168	165	170
5 starts; Δ = 2,4	544	559	505	543	536	527	542
Powell	177	186	205	194	192	186	203
n=4	161	170	187	176	174	169	183
3 starts; Δ = 2,4	821	866	953	898	888	862	935
Rosenbrock	443	434	413	436	423	419	403
n=2	421	410	382	412	389	383	377
5 starts; Δ = 2,4	1285	1254	1177	1260	1201	1185	1157
Woods-Colville	621	647	579	631	592	571	569
n=4	560	570	479	559	527	505	508
4 starts; Δ = 8,12	2861	2927	2495	2867	2700	2591	2601
Singular Rosenbrock	939	865	817	849	820	779	808
n=2	884	831	751	800	767	718	763
5 starts; Δ = 2,4	2707	2527	2319	2453	2354	2215	2334

TABLE 8.2.2: Percentage More (+) or Fewer (-) Equivalent Function Evaluations Required In Comparison to BFGS

a) OCOPTR Algorithm

Problem	Update OC	OCX	BFGS-OCX ψ=1.1	BFGS-OCX ψ=1.3	BFGS-OCX ψ=1.5	BFGS-OC ψ=1.5
Box	-2.6	+2.2	+1.1	0	0	0
Powell	-26.7	-6.7	-14.0	-4.9	-4.9	-13.1
Rosenbrock	+12.0	+5.7	+1.0	+3.2	0	-2.2
Woods-Colville	+7.3	+5.3	+0.2	-4.2	-1.0	-2.4
Singular Rosenbrock	+16.7	+22.5	+25.9	-1.7	-4.7	+[illegible].5

b) MINOP Algorithm

Update:	OC	OCX	BFGS-OCX $\psi=1.1$	BFGS-OCX $\psi=1.3$	BFGS-OCX $\psi=1.5$	BFGS-OC $\psi=1.5$
Problem						
Box	+7.7	+10.7	+7.5	+6.1	+4.4	+7.3
Powell	-13.9	-9.1	-5.8	-6.8	-9.6	-1.9
Rosenbrock	+9.2	+6.5	+7.1	+2.0	+0.7	-1.7
Woods-Colville	+14.7	+17.3	+14.9	+8.2	+3.9	+4.3
Singular Rosenbrock	+16.7	+9.0	+5.8	+1.5	-4.5	+0.7

Section B.3 - Q-Superlinear Convergence of the BFGS-OCX Algorithm

The computational results of Section B.2 show that the BFGS-OCX and BFGS-OC may be promising updating strategies. Both of them are intended to use the BFGS update most of the time. In this section we show that when the updating class is the standard DFP-BFGS and some very reasonable assumptions on the function f are satisfied, then asymptotically these strategies choose only the BFGS, and hence are Q-superlinearly convergent by the work of Broyden, Dennis and Moré (1973).

It is easier to prove our results by considering updates to inverse Hessian approximations H. Hence we are discussing an algorithm of the form

Algorithm B.3.1

Given $x_0 \epsilon R^n$, $H_0 \epsilon L(R^n)$ symmetric and positive definite,

$\psi > 1$ constant

at each iteration i:

$$s_i = -H_i g(x_i)$$

$$x_{i+1} = x_i + s_i$$

$$y_i = g(x_{i+1}) - g(x_i)$$

$$H_{i+1}(\bar{\phi}_i) = H_i + \frac{(s_i - H_i y_i)s_i^T + s_i(s_i - H_i y_i)^T}{\langle s_i, y_i \rangle} - \frac{\langle s_i - H_i y_i, y_i \rangle s_i s_i^T}{\langle s_i y_i \rangle^2}$$

$$- \bar{\phi}_i \langle y_i, H_i y_i \rangle \left[\frac{s_i}{\langle s_i, y_i \rangle} - \frac{H_i y_i}{\langle y_i, H_i y_i \rangle}\right]\left[\frac{s_i}{\langle s_i, y_i \rangle} - \frac{H_i y_i}{\langle y_i, H_i y_i \rangle}\right]^T$$

$$\bar{\phi}_i = \begin{cases} 0 & K(\text{BFGS}) \leq \psi \cdot K(\text{OCX}) \\ (\bar{\phi}_i)_{OCX} & K(\text{BFGS}) > \psi \cdot K(\text{OCX}) \end{cases} \tag{B.3.2}$$

where $(\bar{\phi}_i)_{OCX}$ is given by (B.2.3), and K(BFGS), K(OCX) are the bounds on the condition numbers of the BFGS and OCX updates given in Chapter 7. In Theorem B.3.1 we show that Algorithm B.3.1 is locally Q-superlinearly convergent to the solution of many optimization problems for any value of $\psi > 1$. From the techniques of proof, it seems most likely that if an algorithm using Davidon's restricted BFGS update were shown to be linearly or superlinearly convergent, then the corresponding restricted BFGS-OCX algorithm would be also.

<u>Theorem B.3.1</u> Let $f: R^n \to R$ be twice differentiable in the open convex set D, and suppose for some $x^* \epsilon D$ and $p > 0$, $K \geq 0$,

$$\| G(x) - G(x^*) \| \leq K \| x - x^* \|^p \tag{B.3.3}$$

for all $x \epsilon D$, where $g(x^*) = 0$ and $G(x^*)$ is symmetric and positive definite.

Let $x_0 \varepsilon R^n$, $H_0 \varepsilon L(R^n)$ be symmetric. Let the sequences $\{x_i\}$, $\{H_i\}$, $i=0,1,\ldots$ be generated from x_0, H_0 by the BFGS-OCX Algorithm B.3.1. Then for any $\psi > 1$ in Algorithm B.3.1, there exist $\bar{\varepsilon}(\psi)$, $\bar{\delta}(\psi) > 0$ such that if $\| x_0 - x^* \| \leq \bar{\varepsilon}(\psi)$ and $\| H_0 - G(x^*)^{-1} \|_{Fr,G(x^*)^{1/2}} \leq \bar{\delta}(\psi)$, then the BFGS-OCX algorithm selects only the BFGS update, and the sequence $\{x_i\}$ converges Q-superlinearly to x^*.

<u>Proof</u>: Define $A = G(x^*)^{-1}$, $\|\cdot\|_M = \|\cdot\|_{Fr,A^{-1/2}}$, $\|\cdot\| = \|\cdot\|_2$, and let γ, σ, η be constants such that $\gamma \geq \|A\|$, $\sigma \geq \|A^{-1}\|$, $\|\cdot\|_2 \leq \eta\|\cdot\|_M$. Broyden, Dennis and Moré (1973) have shown that for any $r\varepsilon(0,1)$, there exist positive constants $\varepsilon(r)$, $\delta(r)$ such that if $\| x_0 - x^* \| \leq \varepsilon(r)$ and $\| H_0 - A \|_M \leq \delta(r)$, and if $\{x_i\}$, $\{H_i\}$ are generated by the direct prediction BFGS algorithm, then $\| H_i - A \|_M \leq 2\delta(r)$, H_i is positive definite, $\| H_i^{-1} \| \leq \sigma/(1-r)$, $\| x_{i+1} - x^* \| \leq r \| x_i - x^* \|$ for $i = 0,1,\ldots$, and $\{x_i\}$ converges Q-superlinearly to x^*. We show that for any $\psi > 1$, there exist constants $\varepsilon(\psi)$, $\delta(\psi)$ such that if $\| x_0 - x^* \| \leq \min \{\varepsilon(r), \varepsilon(\psi)\}$ and $\| H_0 - A \|_M \leq \min \{\delta(r), \delta(\psi)\}$, then the BFGS-OCX algorithm selects the BFGS update at each iteration. Hence by the result of Broyden, Dennis and Moré, the sequence of points it produces converges Q-superlinearly to x^*.

For any $r\varepsilon(0,1)$ and $\psi > 1$, we let $\varepsilon(\psi)$, $\delta(\psi)$ be positive constants such that

$$\varepsilon(\psi) \leq \frac{1-r}{1+r} \left[\frac{\alpha_1}{2K(2\eta\delta(r)+\gamma)}\right]^{1/P} \tag{B.3.4}$$

$$\delta(\psi) \leq \alpha_1/4\eta\sigma \tag{B.3.5}$$

where

$$\alpha_1 = \alpha^{1/2} \left(\frac{1-r}{\sigma}\right)^{1/2} \frac{1}{4\gamma} \tag{B.3.6}$$

$$\alpha = \frac{\bar{\psi}-1}{3} \left(\frac{1-r}{\sigma}\right)^{3/2} (2\eta\delta(r)+\gamma)^{-1/2} \tag{B.3.7}$$

$$\bar{\psi} = \min(\psi, 2)$$

and we define $\varepsilon = \min\{\varepsilon(\phi), \varepsilon(r)\}$, $\delta = \min\{\delta(\psi), \delta(r)\}$. We show by induction that if $\| x_0 - x^* \| \le \varepsilon$ and $\| H_0 - A \|_M \le \delta$, then at each step the BFGS-OCX algorithm selects the BFGS. However the first step is identical to the induction step, so we omit it.

Assume by induction that at any step i of the BFGS-OCX algorithm, all previous steps have selected the BFGS update. Let $H = H_i$, $x = x_i$, $\bar{x} = x_{i+1}$, $s = s_i$, $y = y_i$, $\bar{H}(\phi) = H_{i+1}(\bar{\phi}_i)$, and $\bar{H}_B = H_{i+1}(0)$. Then by the results of Broyden, Dennis and Moré,

$$\| H^{-1} \|, \ \| \bar{H}_B^{-1} \| \le \sigma/(1-r) \tag{B.3.8}$$

$$\| \bar{x} - x^* \| \le r \| x - x^* \| \tag{B.3.9}$$

$$\| x - x^* \| \le \| x_0 - x^* \| \tag{B.3.10}$$

$$\| H - A \|_M \le 2\delta$$

Since $\| H \| \le \| H - A \| + \| A \| \le \eta \| H - A \|_M + \| A \|$,

$$\| H \| \le 2\eta\delta(r)+\gamma. \tag{B.3.11}$$

We now show that $K(H^{-1/2}\bar{H}_B H^{-1/2}) / K(H^{-1/2}\bar{H}(\phi_{OCX}) H^{-1/2}) < \psi$.

Since for any two matrices M_1 and $M_2 \in L(R^n)$, $K(M_1 M_2) \le K(M_1) \cdot K(M_2)$ and [illegible] $= K(M_1)$

$$K(H^{-1/2}\overline{H}_B H^{-1/2})/K(H^{-1/2}\overline{H}(\phi_{OCX})H^{-1/2}) \leq K(H^{1/2}\overline{H}(\phi_{OCX})^{-1}\overline{H}_B H^{-1/2})$$

$$= K(H^{1/2}\overline{H}_B{}^{-1}\overline{H}(\phi_{OCX})H^{-1/2}) = K(H^{1/2}\overline{H}_B{}^{-1}(\overline{H}_B+\phi_{OCX}T)H^{-1/2})$$

where

$$T = -\langle y,Hy\rangle \left[\frac{s}{\langle s,y\rangle} - \frac{Hy}{\langle y,Hy\rangle}\right]\left[\frac{s}{\langle s,y\rangle} - \frac{Hy}{\langle y,Hy\rangle}\right]^T \qquad \text{(B.3.12)}$$

$= \overline{H}(1) - \overline{H}(0)$. Thus

$$K(H^{-1/2}\overline{H}_B H^{-1/2})/H^{-1/2}\overline{H}(\phi_{OCX})H^{-1/2}) \leq K(I+\phi_{OCX}H^{1/2}\overline{H}_B{}^{-1}TH^{-1/2}) \qquad \text{(B.3.13)}$$

$$= K(I+\overline{T})$$

where $\overline{T} = \phi_{OCX}H^{1/2}\overline{H}_B{}^{-1}T\,H^{-1/2}$. We first derive an upper bound on $\|T\|$.

Straightforward algebra shows that

$$\frac{s}{\langle s,y\rangle} - \frac{Hy}{\langle y,Hy\rangle} = \frac{s\langle y,Hy-s\rangle-(Hy-s)\langle s,y\rangle}{\langle s,y\rangle\langle y,Hy\rangle}$$

Therefore

$$\|T\| \leq \frac{4\|s\|^2\|y\|^2\|Hy-s\|^2}{\langle s,y\rangle^2\langle y,Hy\rangle} \qquad \text{(B.3.14)}$$

To bound $Hy-s$, we write it as $Hy-s = H(y-A^{-1}s)+(H-A)A^{-1}s$. Broyden, Dennis and Moré (1973) show that if (B.3.3) holds, then

$$\|y-A^{-1}s\| \leq K\|s\| \max\{\|x-x^*\|^p, \|\overline{x}-x^*\|^p\}.$$

Also by adding (B.3.9) to $\|x-x^*\| \leq \|x-\overline{x}\| + \|\overline{x}-x^*\|$ we obtain $(1-r)\|x-x^*\| \leq \|s\|$, so that

$$\| y-A^{-1}s \| \leq K \| s \|^{1+P}/(1-r)^P.$$

Thus

$$\| Hy-s \| \leq (\omega_1+\omega_2 \| s \|^P) \| s \|,$$

$$\omega_1 = 2\eta\sigma\delta, \quad \omega_2 = (2\eta\delta(r)+\gamma)K/(1-r)^P \tag{B.3.15}$$

We must find lower bounds for <y,Hy> and <s,y>. Since H is positive definite,

$$\langle y,Hy \rangle \geq \| y \|^2/\| H^{-1} \| \geq \| y \|^2(1-r)/\sigma. \tag{B.3.16}$$

Also

$$\langle s,y \rangle = \langle s,A^{-1}s \rangle + \langle s,y-A^{-1}s \rangle \geq \| s \|^2/\| A \| - \| s \| \, \| y-A^{-1}s \| \geq$$

$$\| s \|^2 \left[\frac{1}{\gamma} - \frac{K \| s \|^P}{(1-r)^P} \right] \tag{B.3.17}$$

We now bound the term in brackets. From (B.3.9), (B.3.10) and $\| s \| \leq \| x-x^* \| + \| \bar{x}-x^* \|$,

$$\| s \| \leq (1+r) \| x-x^* \| \leq (1+r) \| x_0-x^* \| \leq (1+r)\varepsilon(\psi) \tag{B.3.18}$$

Equation (B.3.4) implies

$$\varepsilon(\psi) \leq \frac{1-r}{1+r} \left(\frac{1}{2K\gamma} \right)^{1/P} \tag{B.3.19}$$

because

$$\frac{\alpha_1}{2\eta\delta(r)+\gamma} \leq \left(\frac{\psi-1}{3} \right)^{1/2} \frac{1}{4\gamma} \left[\left(\frac{\sigma}{1-r} \right) (2\eta\delta(r)+\gamma) \right]^{-5/4} \leq \frac{1}{4\sqrt{3}\gamma} < \frac{1}{\gamma}$$

due to $1 < \bar{\psi} \leq 2$ and $(\sigma/(1-r))(2\eta\delta(r)+\gamma) \geq \|H\| \|H^{-1}\| \geq 1$. Thus from (B.3.18) and (B.3.19),

$$\frac{1}{\gamma} - \frac{K\|s\|^p}{(1-r)^p} \geq \frac{1}{\gamma} - K\left(\frac{1+r}{1-r}\right)^p \varepsilon(\psi)^p \geq \frac{1}{\gamma} - \frac{1}{2\gamma} = \frac{1}{2\gamma}$$

and then from (B.3.17)

$$\langle s,y\rangle \geq \|s\|^2/2\gamma \qquad \text{(B.3.20)}$$

Now substituting (B.3.15,16,20) into (B.3.14) yields

$$\|T\| \leq 16\gamma^2(\sigma/(1-r))(\omega_1+\omega_2\|s\|^p)^2$$

From (B.3.5), $\omega_1 \leq 2\eta\sigma\delta(\psi) \leq \alpha_1/2$; and from (B.3.4) and (B.3.18), $\omega_2\|s\|^p \leq \omega_2(1+r)^p\ \varepsilon(\psi)^p \leq \alpha_1/2$. Thus

$$\|T\| \leq 16\gamma^2(\sigma/(1-r))\alpha_1^2 = \alpha \qquad \text{(B.3.21)}$$

Finally, since $\phi_{OCX}\varepsilon[0,1]$, $\|\bar{T}\| \leq \|H^{1/2}\| \|\bar{H}_B^{-1}\| \|T\| \|H^{-1/2}\| \leq (\bar{\psi}-1)/3$ by (B.3.8,11,21) and (B.3.7). Since $(\bar{\psi}-1)/3 \leq 1/3$,

$$K(I+\bar{T}) \leq \frac{1+\|\bar{T}\|}{1-\|\bar{T}\|} \leq 1+3\|\bar{T}\| \leq \bar{\psi} \leq \psi.$$

Thus by (B.3.13), the BFGS-OCX update selects the BFGS, which completes the proof.

The identical theorem to Theorem 8.3.1 can be proven about the BFGS-OC algorithm using the standard DFP-BFGS class, which is the algorithm obtained if (B.3.2) is replaced by

$$\bar{\phi}_i = \begin{cases} 0 & K(BFGS) \le \psi \cdot K(OC) \\ (\bar{\phi}_i)_{OC} & K(BFGS) > \psi \cdot K(OC) \end{cases}$$

where $(\bar{\phi}_i)_{OC}$ is given by (8.2.2), K(OC) given in Chapter 7. The technique of proof is very similar, but seems to require using the Frobenius norm weighted by $H^{-1/2}$ of the matrix ϕ_{OC} T, T given by (B.3.12). Since we show by induction that H_i stays bounded close to $G(x^*)^{-1}$, there is still a μ such that $\|\cdot\|_{Fr,H_i^{-1/2}} \le \mu\|\cdot\|_2$, $i = 0,1,\ldots$.

The algorithms tested in Section B.2 use a restricted BFGS algorithm which has not yet been proven Q-superlinearly convergent. Thus we can not yet prove any convergence results for the corresponding restricted BFGS-OCX algorithm. However, the techniques of proof in Chapter 9 for similar restricted BFGS updates, and the techniques of this section, make it seem most likely that if the restricted BFGS algorithm used in Section B.2 is Q-superlinearly convergent, then the corresponding restricted BFGS-OCX and BFGS-OC algorithms are also.

CHAPTER IX

DAVIDON-LIKE METHODS AND THEIR CONVERGENCE

In Chapter 2 we introduced Davidon's (1975) quasi-Newton algorithm, which has the new feature that it tries to preserve derivative information from past steps. We saw that the algorithm has a basic form for quadratics, and that there are then several different ways to implement it for non-quadratics. Davidon chose what seems the most simple implementation - although its equivalence to his basic algorithm for quadratics depends on all steps being combinations of present and past quasi-Newton directions, and on the update being to inverse Hessian approximations - and with it has gotten good computer results. However, our discussions in Chapters 2 and 4 have suggested that, for reasons of preserving past derivative information for non-quadratics and preserving positive definiteness, we may be interested in other extensions of Davidon's basic algorithm to non-quadratic problems. Also, his implementation seems especially difficult to analyze for convergence. Therefore in this chapter we discuss alternative extensions of Davidon's algorithm to non-quadratic problems, and the convergence of some of these methods.

In Section 9.1 we discuss briefly the considerations in implementing Davidon's ideas for general problems, and introduce the new updates which we will analyze. In Section 9.2 we give a general convergence theorem for the local Q-linear convergence of quasi-Newton methods. This theorem is an extension of the general convergence theorem of Broyden, Dennis, and Moré (1973), applicable to methods which preserve past update information, and also to the updates discussed in Chapter 10. In Section 9.3 we [illegible] local Q-superlinear convergence of two of the methods proposed in Section 9.1. Both are equivalent to Davidon's basic method for quadratics, and we hope that the second especially will be competitive with Davidon's implementation in practice. Our proofs seem to show some basic techniques which are important to analyzing methods which preserve past derivative information.

The theorems and lemmas of Chapters 9 and 10 are written for the general problem of finding a zero of $F(x)$, $F:R^n \to R^n$. This is partly to be consistent with Broyden, Dennis and Moré (1973), but mainly to show that our results apply to this more general problem. The unconstrained minimization problem can be thought of as the special case $F(x) = \nabla f(x)$, $f:R^n \to R$. The work of Sections 9.2 and 9.3 extends the proofs in Gay and Schnabel (1977) of Q-superlinear convergence for a modified Broyden's method for solving $F(x) = 0$ which preserves past derivative information.

Throughout Chapters 9 and 10, we use $\|\cdot\|$ to stand for $\|\cdot\|_2$, and $\|\cdot\|_M$ to stand for $\|\cdot\|_{Fr,M}$. We denote the Jacobian matrix of F at x by $F'(x)$, where $F'(x)_{ij}$ = the derivative of the i^{th} component function of F with respect to the j^{th} component of x. The Jacobian matrix of the gradient function $g(x)$ is the Hessian $G(x)$.

Section 9.1 - Alternative Implementations of Davidon's Basic Algorithm

Davidon (1975) has introduced an unconstrained minimization algorithm which has the property that for quadratic functions f,

$$B_{i+1} \text{ is symmetric and } B_{i+1}s_j = y_j, \quad j = 0,1,\ldots,i \tag{9.1.1}$$

where $s_j = x_{j+1}-x_j$, $y_j = g(x_{j+1})-g(x_j)$, and $B_i \epsilon L(R^n)$ is an approximation to the Hessian $G(x_i)$. Therefore if the first n steps $s_0,\ldots,s_{n-1}$ are linearly independent then B_n is the actual Hessian, and if f is positive definite and direct prediction quasi-Newton steps are taken the method locates the minimum of f in n+1 or fewer iterations.

Equation (9.1.1) can be accomplished at each iteration for quadratic f by an update of the form

$$B_{i+1} = B_i + \frac{(y_i - B_i s_i)v_i^T + v_i(y_i - B_i s_i)^T}{\langle v_i, s_i \rangle} - \frac{\langle y_i - B_i s_i, s_i \rangle v_i v_i^T}{\langle v_i, s_i \rangle^2} \tag{9.1.2}$$

where each v_i obeys

$$\langle v_i, s_j \rangle = 0, \quad j = 0,\ldots,i-1, \quad \langle v_i, s_i \rangle \neq 0 \tag{9.1.3}$$

because for quadratic f we also have

$$\langle y_i - B_i s_i, s_j \rangle = 0, \quad j = 0,\ldots,i-1 \tag{9.1.4}$$

if $B_i s_j = y_j$. A nice way to accomplish (9.1.3) is to define at successive steps

$$v_i = \hat{y}_i = \left(I - \sum_{j=0}^{i-1} \frac{\hat{y}_j \, \hat{s}_j^T}{\langle \hat{y}_j, \hat{s}_j \rangle}\right) y_i \triangleq P_i y_i$$

$$\hat{s}_i = P_i^T s_i \tag{9.1.5}$$

Then each $\hat{y}_i$ is a linear combination of $y_0,\ldots,y_i$ which is orthogonal to s_j, $j = 0,\ldots,i-1$; and each $\hat{s}_i$ is a linear combination of $s_0,\ldots,s_i$ which is orthogonal to y_j, $j = 0,\ldots,i-1$. Also $\hat{y}_i - B_i\hat{s}_i = y_i - B_i s_i$ and $\langle v_i, \hat{s}_i \rangle = \langle v_i, s_i \rangle$, $\langle y_i - B_i s_i, s_i \rangle = \langle \hat{y}_i - B_i \hat{s}_i, \hat{s}_i \rangle$ so that (9.1.2) can be rewritten

$$B_{i+1} = B_i + \frac{(\hat{y}_i - B_i\hat{s}_i)\hat{y}_i{}^T + \hat{y}_i(\hat{y}_i - B_i\hat{s}_i)^T}{\langle \hat{y}_i, \hat{s}_i \rangle} - \frac{\langle \hat{y}_i - B_i\hat{s}_i, \hat{s}_i \rangle \hat{y}_i \hat{y}_i{}^T}{\langle \hat{y}_i, \hat{s}_i \rangle^2} \tag{9.1.6}$$

This is the basic algorithm introduced by Davidon (1975).

When we attempt to extend Davidon's basic algorithm to general functions f we encounter two fundamental problems. The first is that once we proceed more than n iterations, we can at most hope to preserve update information from the past n-1 iterations. Indeed, if $s_0, \ldots, s_{n-1}$ are linearly independent then $\hat{y}_i$ given by (9.1.5) will be 0 for $i \geq n$. Thus we must adjust our goals to concern only recent past derivative information, and adjust formula (9.1.5) accordingly. The second and more serious problem is that we have shown in Lemma 2.4.2 that even if $B_i s_j = y_j$, then a new matrix B_{i+1} which preserves symmetry and the quasi-Newton equation $B_{i+1}s_i = y_i$ can also have $B_{i+1}s_j = y_j$ only if $\langle y_i - B_i s_i, s_j \rangle = 0$. For general functions f this will not hold.

An obvious thing we can do is continue to use update (9.1.2) with v_i orthogonal to recent past s_j's, so that if f is close to quadratic and $\langle y_i - B_i s_i, s_j \rangle = 0$, then $B_{i+1}s_j$ will $= B_i s_j$. To do this we need to redefine v_i to be orthogonal to "recent past s_j's." We do this by adjusting equation (9.1.5) so that we not only ask that $\hat{y}_i$ be orthogonal to at most n-1 vectors, but also that if $\hat{y}_i$ is orthogonal to $s_\ell, \ldots s_{i-1}$ that s_i be numerically linearly independent of $s_\ell, \ldots, s_{i-1}$, so that the goals $(B_{i+1} - B_i)s_i = y_i - B_i s_i \neq 0$ and $(B_{i+1} - B_i)s_j = 0$, $j = \ell, \ldots, i-1$ are numerically consistent. This is accomplished in Algorithm 9.1.7.

Algorithm 9.1.7

Given $F: R^n \to R^n$ differentiable; $\tau > 1$ constant; $\ell_{-1} = 0$.

At each iteration i, $i = 0,1,\ldots$

$$s_i = x_{i+1} - x_i \neq 0$$

$$y_i = F(x_{i+1}) - F(x_i)$$

$$Q_i = \sum_{j=\ell_{i-1}}^{i-1} \frac{\hat{y}_j \hat{s}_j^T}{\langle \hat{y}_j, \hat{s}_j \rangle}$$

If $\| s_i \| > \tau \| s_i - Q_i^T s_i \|$

then $P_i = I$, $\ell_i = i$

else $P_i = I - Q_i$, $\ell_i = \ell_{i-1}$

$$\hat{y}_i = P_i y_i$$

$$\hat{s}_i = P_i^T s_i$$

In a comparable algorithm for the nonlinear equations problem, Gay and Schnabel (1977) find that 10 is a reasonable value for τ. We now prove some basic properties of the vectors produced in Algorithm 9.1.7.

Lemma 9.1.1 Let $F:R^n \to R^n$ be continuously differentiable, $(x_0, x_1, \ldots)$ be a sequence $\in R^n$ such that $x_{i+1} \neq x_i$, $i = 0,1,\ldots$; $\tau > 1$ constant; and let the sequences $(s_0, s_1, \ldots)$, $(y_0, y_1, \ldots)$, $(\hat{s}_0, \hat{s}_1, \ldots)$, $(\hat{y}_0, \hat{y}_1, \ldots)$ $\in R^n$ be generated from $(x_0, x_1, \ldots)$ by Algorithm 9.1.7. Then for $i = 0,1,\ldots$ and ℓ_i defined in Algorithm 9.1.7, $(s_{\ell_i}, \ldots, s_i)$ are linearly independent and

$$\| s_i \| \leq \tau \| \hat{s}_i \| \tag{9.1.8}$$

$$i - \ell_i < n \tag{9.1.9}$$

$$\langle\hat{y}_i,\hat{s}_j\rangle = \langle\hat{y}_i,s_j\rangle = 0 \qquad j = \ell_i,\ldots,i-1 \tag{9.1.10}$$

$$\langle\hat{s}_i,\hat{y}_j\rangle = \langle\hat{s}_i,y_j\rangle = 0 \qquad j = \ell_i,\ldots,i-1 \tag{9.1.11}$$

$$\langle\hat{y}_i,\hat{s}_i\rangle = \langle\hat{y}_i,s_i\rangle = \langle y_i,\hat{s}_i\rangle \tag{9.1.12}$$

Proof: The proof is by induction. For $i = 0$, $Q_0 = 0$, $\ell_0 = 0$, $\hat{y}_0 = y_0$, $\hat{s}_0 = s_0$ so that the conclusions of the lemma are trivially true.

Now assume Lemma 9.1.1 is true for $i = 0,\ldots,k-1$. Then for $i = k$, if $\|s_k\| > \tau\|s_k-Q_k^Ts_k\|$ then $\ell_k = k$, $\hat{y}_k = y_k$, $\hat{s}_k = s_k$ and the conclusions of the lemma are again trivially true. If $\|s_k\| \le \tau\|s_k-Q_k^Ts_k\|$, then $\ell_k < k$ and $\hat{s}_k = s_k-Q_k^Ts_k$, so that (9.1.8) holds. Thus since $s_k \neq 0$, $\hat{s}_k = s_k - \sum_{j=\ell_k}^{k-1} \hat{s}_j(\langle\hat{y}_j,s_k\rangle/\langle\hat{y}_j,\hat{s}_j\rangle) \neq 0$, which by induction hypothesis (9.1.10) implies that s_k is not a linear combination of $\hat{s}_{\ell_k},\ldots,\hat{s}_{k-1}$, and hence not a linear combination of $s_{\ell_k},\ldots,s_{k-1}$. Therefore by the induction hypothesis $s_{\ell_k},\ldots,s_k$ are linearly independent, which implies that (9.1.9) holds. To prove (9.1.10), we see by the definition of $\hat{y}_i$ and induction hypotheses (9.1.10) and (9.1.11) that for any $m\epsilon[\ell_k,k-1]$

$$\langle\hat{y}_k,\hat{s}_m\rangle = \langle y_k,\hat{s}_m\rangle - \sum_{j=\ell_k}^{k-1} \frac{\langle\hat{y}_j,\hat{s}_m\rangle\langle\hat{s}_j,y_k\rangle}{\langle\hat{y}_j,\hat{s}_j\rangle} =$$

$$\langle y_k,\hat{s}_m\rangle - \frac{\langle\hat{y}_m,\hat{s}_m\rangle\langle\hat{s}_m,y_k\rangle}{\langle\hat{y}_m,\hat{s}_m\rangle} = 0.$$

Since each $s_m\epsilon[\hat{s}_{\ell_k},\ldots,\hat{s}_m]$, also $\langle\hat{y}_k,s_m\rangle = 0$. Thus (9.1.10) is true. The proof of (9.1.11) is the same with the roles of s and y reversed, and then (9.1.12) follows from (9.1.10) and (9.1.11).

Thus we see that we can still pick v_i in (9.1.2) such that if $\langle y_i-B_is_i,s_j\rangle = 0$ and $\ell_i \leq j < i$ then $B_{i+1}s_j = B_is_j$. One obvious strategy which follows is to use $\hat{y}_i,\hat{s}_i$ as defined in Algorithm 9.1.7 in restricted DFP update (9.1.6). This update has the disadvantage that, since $\hat{y}_i-B_i\hat{s}_i$ no longer necessarily equals $y_i-B_is_i$, $B_{i+1}s_i$ will not necessarily be y_i. This would probably impair its computational performance. However it has the interesting property that $B_{i+1}\hat{s}_j = \hat{y}_j$, $j = \ell_i,\ldots,i$, and it can be shown to be locally Q-superlinearly convergent (when used in a direct prediction quasi-Newton algorithm) by techniques which can then be expanded to more promising updates. Therefore we formally define this update as Update 9.1.13.

Update 9.1.13

B_{i+1} given by (9.1.6)

where $\hat{s}_i,\hat{y}_i$ are given by Algorithm 9.1.7.

A second idea would be to use $\hat{y}_i$ given by Algorithm 9.1.7 in v-update (9.1.2). Then we would have $B_{i+1}s_i = y_i$ at each iteration, and $B_{i+1}s_j = B_is_j$ for $j\varepsilon[\ell_i,i-1]$ if $\langle y_i-B_is_i,s_j\rangle = 0$. We think this update might lead to as successful an algorithm as Davidon's, and we intend to test it computationally. Furthermore, we also show in Section 9.3 that the direct prediction quasi-Newton algorithm using this update is locally Q-superlinearly convergent. Hence we define Update 9.1.14 as:

Update 9.1.14

B_{i+1} given by (9.1.2)

where $v_i = \hat{y}_i$ given by Algorithm 9.1.7

Both Updates 9.1.13 and 9.1.14 are equivalent to Davidon's basic update for quadratics, as long as $\| s_i \| \leq \tau \cdot \| s_i - Q_i^T s_i \|$ for all $i < n$. One disadvantage Update 9.1.14 may have is that it is messy to monitor the theoretical positive definiteness of B_{i+1}, although from Chapter 3 we know how to do it, and computationally we must monitor it separately anyhow. Davidon however uses a restricted DFP update (or the associated one parameter class) which in form (9.1.2) has v_i = the projection $\overline{P}_i y_i$ of y_i onto $[y_i - B_i s_i, u_i]$, u_i some vector orthogonal to s_{i-1}. (Actually Davidon implements the dual of this idea for updates to H_i, as discussed in Section 2.4.) We know from Chapter 3 that such an update can also be written in restricted DFP form (9.1.5) with $\hat{y}_i = \overline{P}_i y_i$ and $\hat{s}_i = \overline{P}_i^T s_i$, and so its theoretical positive definiteness is easy to monitor. If the update is not positive definite, Davidon goes back to the standard DFP-BFGS class. However such a strategy has the disadvantage that even if $\langle y_i - B_i s_i, s_j \rangle = 0$, $B_{i+1} s_i$ will only necessarily be $B_i s_i$ if $j = i-1$. Secondly, it was shown in Chapter 4 that we expect such a strategy to be less likely to produce positive definite restricted updates than another which we mention again below. Finally, in the B form this algorithm is not even equivalent to Davidon's basic algorithm, and while it is in the H form, for some applications an inverse update has no clear meaning or use.

Thus the final update we propose is very similar to (9.1.14), but embodies the restricted DFP-BFGS philosophy. Realizing that $B_{i+1}s_j = B_i s_j$ is impossible (along with $B_{i+1}-B_i$ symmetric and $B_{i+1}s_i = y_i$) unless $\langle y_i - B_i s_i, s_j \rangle = 0$, we define $\bar{s}_j$ as the projection of s_j orthogonal to $B_i^{-1}(y_i - B_i s_i)$ in metric B_i,

$$\bar{s}_j = s_j - \frac{(B_i^{-1}y_i - s_i)\langle y_i - B_i s_i, s_j \rangle}{\langle y_i - B_i s_i, B_i^{-1}(y_i - B_i s_i)\rangle} .$$

Thus $\langle \bar{s}_j, y_i - B_i s_i \rangle = 0$, and $\bar{s}_j = s_j$ if $\langle y_i - B_i s_i, s_j \rangle = 0$. We then let B_{i+1} be the restricted DFP update as defined in Theorem 3.1.1 for $Z_i = [\bar{s}_\ell, \ldots, \bar{s}_{i-1}]$, $\ell(> i-n)$ the smallest number such that the resultant $\langle \hat{s}_i, \hat{y}_i \rangle > 0$, $\hat{s}_i, \hat{y}_i$ defined by Theorem 3.1.1. This update has already been discussed in Section 4.2. As was explained there, it will be positive definite, and preserve information in directions $\bar{s}_\ell, \ldots, \bar{s}_{i-1}$, while Davidon's algorithm preserves information only in some directions which are projections of s_{i-1}. It is also equivalent to Davidon's algorithm for quadratics. We have not proven the Q-superlinear convergence of an algorithm using this update, as it would appear to be even messier than the proofs in Section 9.2 (which are already quite long). However, it seems very likely that the same techniques would be successful. We would first like to implement and test this update, which we consider the most promising one mentioned in this thesis. For unconstrained minimization we would actually implement the dual (restricted BFGS) form.

Section 9.2 - A General Convergence Theorem for Quasi-Newton Methods

In this section we prove a general Q-linear convergence theorem for quasi-Newton methods. This theorem is an extension of a similar theorem in Broyden, Dennis, and Moré (1973), which was a key to their Q-superlinear convergence proofs for the DFP and BFGS. We extend it to apply to updates which depend on previous steps in addition to the current step; and also to include a slightly more general term which we use in Chapter 10. The proof is hardly different from that of Broyden, Dennis and Moré. The theorem is used in our proofs of Q-superlinear convergence of two Davidon-like methods in Section 9.3.

In Lemma 9.2.1 we state a fundamental property of Lipschitz continuous functions. Theorem 9.2.2 is our convergence theorem. In Theorem 9.2.3 we state the dual form of Theorem 9.2.2 for updates to inverse Hessian approximations H. It is apparent from the work of Broyden, Dennis and Moré (1973) that the proof of such a dual result is almost identical to the proof of Theorem 9.2.2.

In Theorems 9.2.2 and 9.2.3, we denote by $2^{L(R^n)}\sim\{\}$ the collection of all nonempty subsets of $L(R^n)$.

__Lemma 9.2.1__ Let $F:R^n \to R^n$ be differentiable in the open convex set D, and suppose for some $x^* \epsilon D$ and $p > 0$, $K \geq 0$ that

$$\| F'(x)-F'(x^*)\| \leq K\| x-x^*\|^p \tag{9.2.1}$$

for all $x \epsilon D$. Then for $u,v \epsilon D$

$$\| F(v)-F(u)-F'(x^*)(v-u)\| \leq K\| v-u\| \max \{\| v-x^*\|^p, \| u-x^*\|^p\} \tag{9.2.2}$$

__Proof:__ Follows immediately from Section 3.2.5 of Ortega and Rheinboldt (1970).

Theorem 9.2.2 Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume that for some $x^* \epsilon D$ and $p > 0$, $K \geq 0$, (9.2.1) holds where $F(x^*) = 0$ and $F'(x^*)$ is nonsingular. Define $A = F'(x^*)$. Let q be an integer ≥ 0, and let $U:(R^n)^{q+1} \times L(R^n) \to 2^{L(R^n)} \sim \{\}$ be defined in a neighborhood $N = N_1^{q+1} \times N_2$ of $(x^*,\ldots,x^*,A)$ where $N_1 \epsilon D$ and N_2 contains only nonsingular matrices. Consider the sequences $(x_0,x_1,\ldots)\epsilon R^n$ and $(B_0,B_1,\ldots)\epsilon L(R^n)$ generated from x_0,B_0 by the iteration

$$x_{i+1} = x_i - B_i^{-1}F(x_i) \tag{9.2.3a}$$

$$B_{i+1} = U(x_i,\ldots,x_{i-q},B_i) \tag{9.2.3b}$$

where we define $x_{i-q} = x_0$ for $i \leq q$. Suppose there exist non-negative constants α_1,α_2 and a non-singular symmetric $M \epsilon L(R^n)$ such that for any $(x_i,\ldots,x_{i-q},B_i)\epsilon N$ generated by (9.2.3), the function U satisfies

$$\| B_{i+1}-A\|_M \leq (1+\alpha_1 m_i+\delta_i)\| B_i-A\|_M + (\alpha_2 m_i+\omega_i) \tag{9.2.4}$$

for each $B_{i+1}\epsilon U(x_i,\ldots,x_{i-q},B_i)$, where

$$m_i = \max \{\| x_{i+1}-x^*\|^p, \| x_i-x^*\|^p,\ldots,\| x_{i-q}-x^*\|^p \tag{9.2.5}$$

and δ_i, $\omega_i \geq 0$. Suppose there also exists a non-negative constant $\delta < 1$ such that for all $j \geq 0$, $\sum_{i=0}^{j} \delta_i < \delta$. Then for each $r\epsilon(0,1)$ and any non-negative constant ω, there exist positive constants $\epsilon(r)$, $\delta(r)$ such that if $\| x_0-x^*\| \leq \epsilon(r)$, $\| B_0-A\|_M \leq \delta(r)$, and $\sum_{i=0}^{j} \omega_i \leq \omega\delta(r)$ for all $j \geq 0$, then the sequence $\{x_i\}$ is well-defined and converges to x^*. Furthermore

$$\| x_{i+1}-x^*\| \leq r\| x_i-x^*\| \tag{9.2.6}$$

for each $i \geq 0$, and $\{\| B_i \|\}$, $\{\| B_i^{-1} \|\}$ are uniformly bounded.

Proof: The proof is only a slight modification of the proof in Broyden, Dennis and Moré (1973) of a slightly less general theorem, where $\varsigma_i, \omega_i = 0$ in (9.2.4) and $q = 0$ in (9.2.3b) and (9.2.5). It consists of a double induction argument which shows that (9.2.6) holds and $\| B_i - A \|_M$ is uniformly bounded at each step.

Let $r \in (0,1)$ be given, also γ such that $\| A^{-1} \| \leq \gamma$, and η such that $\| \cdot \| \leq \eta \| \cdot \|_M$. Define $\beta = \frac{\omega+2}{1-\varsigma}$. Choose $\delta(r) = \delta$ and $\varepsilon(r) = \varepsilon$ such that

$$\gamma(1+r)(K\varepsilon^P + \beta\eta\delta) \leq r, \tag{9.2.7}$$

$$(\alpha_1\beta\delta + \alpha_2)(q + \frac{1}{1-r^P})\, \varepsilon^P \leq \delta. \tag{9.2.8}$$

This can be done by selecting any $t \in (0,1)$ and

$$\delta \leq t \frac{r}{\gamma(1+r)\beta\eta}$$

$$\varepsilon^P \leq \min\left\{ \frac{\delta}{(\alpha_1\beta\delta + \alpha_2)(q + \frac{1}{1-r^P})},\ \frac{(1-t)r}{\gamma(1+r)K} \right\}.$$

If necessary further restrict ε and δ so that $x \in N_1$ whenever $\| x - x^* \| \leq \varepsilon$ and $B \in N_2$ whenever $\| B - A \|_M \leq \beta\delta$. We show by induction that

$$\| B_k - A \|_M \leq \beta\delta \tag{9.2.9}$$

$$\| x_{k+1} - x^* \| \leq r \| x_k - x^* \|, \tag{9.2.10}$$

For $k = 0$, (9.2.9) is true from $\|B_0-A\|_M \leq \delta$, since $\beta \geq 2$. Inequality (9.2.10) is proven by the exact same techniques as in the induction step, so we omit it here.

Now assume (9.2.9), (9.2.10) hold for $k = 0,1,\ldots,\ell-1$. Thus for any $k\varepsilon[0,\ell-1]$ (9.2.4) can be written

$$\|B_{k+1}-A\|_M \leq (1+\alpha_1\|x_{k-q}-x^*\|^p+\varsigma_i)\|B_k-A\|_M+(\alpha_2\|x_{k-q}-x^*\|^p+\omega_i). \quad (9.2.11)$$

Summing (9.2.11) for $k = 0$ through $\ell-1$ (and recalling $x_{k-q} \triangleq x_0$ for $k \leq q$) gives

$$\|B_\ell-A\|_M \leq \|B_0-A\|_M + \sum_{k=0}^{\ell-1}\left\{(\alpha_1\|x_{k-q}-x^*\|^p+\varsigma_i)\|B_k-A\|_M + (\alpha_2\|x_{k-q}-x^*\|^p+\omega_i)\right\}$$

$$\leq \|B_0-A\|_M+(\alpha_1\beta\delta+\alpha_2)[(q+1)\varepsilon^p+r^p\varepsilon^p+r^{2p}\varepsilon^p+\ldots+r^{(\ell-1-q)p}\varepsilon^p] + \sum_{k=0}^{\ell-1}(\varsigma_i\beta\delta+\omega_i)$$

$$\leq \|B_0-A\|_M+(\alpha_1\beta\delta+\alpha_2)(q+\frac{1}{1-r^p})\varepsilon^p+\varsigma\beta\delta+\omega\delta. \quad (9.2.12)$$

Using $\|B_0-A\|_M \leq \delta$, (9.2.8), and $\beta = \frac{\omega+2}{1-\varsigma}$, (9.2.12) yields

$$\|B_\ell-A\|_M \leq (2+\varsigma\beta+\omega)\delta = \beta\delta$$

which proves (9.2.9).

To prove (9.2.10), we first show that B_ℓ^{-1} is bounded in norm. Using the Banach Lemma (Ortega and Rheinboldt (1970),p45)

$$\| B_\ell^{-1} \| \leq \frac{\| A^{-1} \|}{1-\| A^{-1} \| \|B_\ell-A\|} \leq \frac{\gamma}{1-\gamma\beta\eta\delta}$$

as long as $\| A^{-1} \| \|B_\ell-A\| < 1$; and since from (9.2.7) $\gamma\beta\eta\delta \leq r/(1+r)$, $\| B_\ell^{-1} \| \leq \gamma(1+r)$. We now write (9.2.3a) as

$$x_{\ell+1}-x^* = -B_\ell^{-1}\{[F(x_\ell)-F(x^*)-A(x_\ell-x^*)]+[(A-B_\ell)(x_\ell-x^*)]\}$$

(recalling $F(x^*) = 0$). Therefore using Lemma 9.2.1 and the induction hypothesis,

$$\begin{aligned} \| x_{\ell+1}-x^* \| &\leq \| B_\ell^{-1} \| [\| F(x_\ell)-F(x^*)-A(x_\ell-x^*) \| + \| A-B_\ell \| \, \| x_\ell-x^* \|] \\ &\leq \gamma(1+r)[K\| x_\ell-x^* \|^{1+p}+\beta\eta\delta \| x_\ell-x^* \|] \\ &\leq \gamma(1+r)(K\varepsilon^p+\beta\eta\delta) \| x_\ell-x^* \| \end{aligned}$$

so by (9.2.7)

$$\| x_{\ell+1}-x^* \| \leq r \| x_\ell-x^* \|$$

which proves (9.2.10) and completes the induction.

<u>Theorem 9.2.3</u> Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume that for some $x^* \varepsilon D$ and $p > 0$, $K \geq 0$, (9.2.1) holds where $F(x^*) = 0$ and $F'(x^*)$ is non-singular. Define $A = F'(x^*)$. Let q be an integer ≥ 0 and let $U:(R^n)^{q+1} \times L(R^n) \to 2^{L(R^n)}\sim\{\}$ be defined in a neighborhood $N = N_1^{q+1} \times N_2$ of $(x^*,\ldots,x^*, A^{-1})$ where $N_1 \varepsilon D$. Consider the sequences $(x_0,x_1,\ldots)\varepsilon R^n$ and $(H_0,H_1,\ldots)\varepsilon L(R^n)$ generated from x_0,H_0 by the iteration

$$x_{i+1} = x_i-H_iF(x_i)$$

$$H_{i+1} = U(x_i,\ldots,x_{i-q},H_i) \tag{9.2.13}$$

where we define $x_{i-q} = x_0$ for $i \le q$. Suppose there exist non-negative constants α_1, α_2 and a non-singular symmetric $M \in L(R^n)$ such that for any $(x_i,\ldots,x_{i-q},H_i) \in N$ generated by (9.2.13), the function U satisfies

$$\| H_{i+1}-A^{-1}\|_M \le (1+\alpha_1 m_i+\xi_i)\| H_i-A^{-1}\|_M+(\alpha_2 m_i+\omega_i)$$

for each $H_{i+1} \in U(x_i,\ldots,x_{i-q},H_i)$, where

$$m_i = \max \{\|x_{i+1}-x^*\|^p, \| x_i-x^*\|^p,\ldots,\| x_{i-q}-x^*\|^p\}$$

and ξ_i, $\omega_i \ge 0$. Suppose there also exists a non-negative constant $\xi < 1$ such that $\sum_{i=1}^{j} \xi_i < \xi$ for all $j > 0$. Then for each $r \in (0,1)$ and any non-negative constant ω, there exist positive constants $\varepsilon(r)$, $\delta(r)$ such that if $\| x_0-x^*\| \le \varepsilon(r)$, $\| H_0-A^{-1}\|_M \le \delta(r)$ and $\sum_{i=0}^{j} \omega_i \le \omega\delta(r)$ for all $j \ge 0$, then the sequence $\{x_i\}$ is well-defined and converges to x^*. Furthermore

$$\| x_{i+1}-x^*\| \le r\| x_i-x^*\|$$

for each $i \ge 0$, and $\{\| H_i\|\}$, $\{\| H_i^{-1}\|\}$ are uniformly bounded.

Section 9.3 - Q-Superlinear Convergence of Some Davidon-Like Methods

In this section we prove the local Q-superlinear convergence, using direct quasi-Newton steps, of the first two Davidon-like algorithms proposed in Section 9.1. Both algorithms reduce to Davidon's (1975) basic algorithm on quadratic problems. Our techniques are extension of techniques Broyden, Dennis, and Moré (1973) and Dennis and Moré (1974) have used to prove the Q-superlinear convergence of the DFP, BFGS and other algorithms.

The two proofs seem to point out some basic techniques required, or at least helpful, in proving convergence of quasi-Newton methods which incorporate past derivative information. In these methods the vectors $s_i = x_{i+1}-x_i$ and $y_i = g(x_{i+1})-g(x_i)$ are replaced by projections $\hat{s}_i$ and $\hat{y}_i$, so that quasi-Newton equations will hold or almost hold for the past ℓ_i steps. This results in $B_{i+1}\hat{s}_i = \hat{y}_i$, or $B_{i+1}\hat{s}_i \doteq \hat{y}_i$. So first, it always seems necessary to prove that this is a "good" quasi-Newton equation, by showing that $F'(x^*)\hat{s}_i \doteq \hat{y}_i$ as x_i approaches x^*, just as the Lipschitz condition (9.2.1) implies $F'(x^*)s_i \doteq y_i$. In our experience, the equation

$$\| y_i-F'(x^*)s_i\| \leq k\| s_i\| \max \{\| x_{i+1}-x^*\|^p, \| x_i-x^*\|^p\}$$

is replaced by

$$\| \hat{y}_i-F'(x^*)\hat{s}_i\| \leq \hat{k}\| \hat{s}_i\| \max \{\| x_{i+1}-x^*\|^p,\ldots,\| x_{i-\ell_i}-x^*\|^p$$

Secondly, it seems necessary to prove that we are "successful" in not changing B_i much in the ℓ_i directions in which we try to keep it "almost unchanged" -- that is to show that either $B_{i+1}\hat{s}_j = \hat{y}_j$ for $i-\ell_i \leq j \leq i$ or at least that $\| \hat{y}_j-B_{i+1}\hat{s}_j\|$ is bounded by a term similar to that bounding $\| \hat{y}_i-F'(x^*)\hat{s}_i\|$ above. Finally, we seem to require constants $\tau,\mu \geq 1$ such that $\| s_i\|/\tau \leq \| \hat{s}_i\| \leq \mu\| s_i\|$ for all i.

These are the types of relations that are proven in the three lemmas of this section. In Lemma 9.3.1 we derive an upper bound on $\| \hat{y}_i-F'(x^*)\hat{s}_i\|$, for the $\hat{s}_i$, $\hat{y}_i$, defined in Algorithm 9.1.7, which we use in both our Davidon-like algorithms. In Lemma 9.1.2 we derive an upper bound on $\| \hat{s}_i\|/\| s_i\|$. Using these two lemmas we are able to prove in Theorem 9.3.3 the Q-superlinear convergence of the first Davidon-like algorithm, which uses Update 9.1.13, and for which $B_{i+1}\hat{s}_j = \hat{y}_j$, j =

$i-\ell_i,\ldots,i$. Then in Lemma 9.3.4 we derive an upper bound on $\|\hat{y}_j-B_{i+1}\hat{s}_j\|$, $i-\ell_i \le j \le i$ for the second algorithm, which uses Update 9.1.14, and using this we prove the Q-superlinear convergence of this algorithm in Theorem 9.3.5. In Theorem 9.3.6 we state the same results for the dual algorithms which update H_i. It is clear from our work and that of Broyden, Dennis and Moré that the same techniques of proof apply to these algorithms.

The proofs of the three lemmas consist mainly of some fairly tedious but straightforward algebraic calculations, often showing that terms are small by expressing them as (sum of) terms which are zero for quadratics, and hence small near the solution of non-quadratics.

<u>Lemma 9.3.1</u> Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^* \epsilon D$ and $p > 0$, $K \ge 0$ that (9.2.1) holds where $F(x^*) = 0$ and $F'(x^*)$ is symmetric and positive definite. Define $A = F'(x^*)$. Let $(x_0,x_1,\ldots)$ be a sequence ϵR^n, and let τ, ℓ_i, s_i, y_i, $\hat{s}_i$, $\hat{y}_i$ be defined by Algorithm 9.1.7. Define

$$m_i = \max\{\|x_{i+1}-x^*\|^p, \|x_i-x^*\|^p, \ldots \|x_{i-\ell_i}-x^*\|^p\}$$

$$\gamma = (\|A\| \, \|A^{-1}\|)^{1/2}$$

$$\phi_j = K(3\gamma)^j \frac{(3^{j+1}-1)}{2} \tau^j$$

$$\sigma_i = \phi_{i-\ell_i}$$

$$\sigma = \phi_{n-1}$$

where i and j are non-negative integers. Then for any non-negative integer i, if $m_i \leq (2\sigma\| A^{-1} \|)^{-1}$ then

$$\| \hat{y}_i - A\hat{s}_i \| \leq \| \hat{s}_i \| \sigma_i m_i \tag{9.3.1}$$

$$\langle \hat{y}_i, \hat{s}_i \rangle \geq \| \hat{y}_i \| \, \| \hat{s}_i \| / 3\gamma \tag{9.3.2}$$

Proof: The proof is by induction, and requires mainly some rather tedious algebraic calculations. For $i = 0$, $\hat{s}_0 = s_0$, $\hat{y}_0 = y_0$, so $\| \hat{y}_0 - A\hat{s}_0 \| = \| y_0 - As_0 \| \leq \| s_0 \| Km_0$ by Lemma 9.2.1 with $v = x_1$, $u = x_0$. Thus since $\ell_0 = 0$ by Algorithm 9.1.7 and $\sigma_0 = K$, (9.3.1) is true for $i = 0$. To show (9.3.2),

$$\begin{aligned} \langle \hat{y}_0, \hat{s}_0 \rangle &= \langle A^{1/2}\hat{s}_0, A^{1/2}\hat{s}_0 \rangle - \langle \hat{y}_0 - A\hat{s}_0, \hat{s}_0 \rangle \\ &\geq \| A^{1/2}\hat{s}_0 \|^2 - \sigma_0 m_0 \| \hat{s}_0 \|^2 \\ &\geq \| A^{1/2}\hat{s}_0 \|^2 (1 - \sigma_0 m_0 \| A^{-1} \|). \end{aligned} \tag{9.3.3}$$

Also since $A^{-1/2}\hat{y}_0 = A^{1/2}\hat{s}_0 + A^{-1/2}(\hat{y}_0 - A\hat{s}_0)$,

$$\begin{aligned} \| A^{-1/2}\hat{y}_0 \| &\leq \| A^{1/2}\hat{s}_0 \| + \| A^{-1/2} \| \, \| \hat{s}_0 \| \sigma_0 m_0 \\ &\leq \| A^{1/2}\hat{s}_0 \| (1 + \sigma_0 m_0 \| A^{-1} \|). \end{aligned} \tag{9.3.4}$$

Substituting (9.3.4) for one of the terms $\| A^{1/2}\hat{s}_0 \|$ in (9.3.3),

$$\begin{aligned} \langle \hat{y}_0, \hat{s}_0 \rangle &\geq \frac{\| A^{1/2}\hat{s}_0 \| \, \| A^{-1/2}\hat{y}_0 \| (1 - \sigma_0 m_0 \| A^{-1} \|)}{(1 + \sigma_0 m_0 \| A^{-1} \|)} \\ &\geq \frac{\| \hat{s}_0 \| \, \| \hat{y}_0 \|}{\gamma} \frac{(1 - \sigma_0 m_0 \| A^{-1} \|)}{(1 + \sigma_0 m_0 \| A^{-1} \|)} \end{aligned}$$

$$\geq \frac{\|\hat{s}_0\|\ \|\hat{y}_0\|}{\gamma}\frac{(1-m_0\sigma\|A^{-1}\|)}{(1+m_0\sigma\|A^{-1}\|)} \geq \frac{\|\hat{s}_0\|\ \|\hat{y}_0\|}{3\gamma}$$

since $\sigma_0 < \sigma$ and $\sigma m_0\|A^{-1}\| \leq \frac{1}{2}$. This proves (9.3.2) for $i = 0$.

Now assume (9.3.1) and (9.3.2) are true for $i = 0,\ldots,k-1$. For $i = k$, if $\ell_k = k$, then $\hat{s}_k = s_k$, $\hat{y}_k = y_k$ and the proof is identical to that for $i = 0$. If $\ell_k < k$ then by the definitions of $\hat{s}_k$ and $\hat{y}_k$,

$$\hat{y}_k - A\hat{s}_k = y_k - As_k + \sum_{j=\ell_k}^{k-1} \frac{\hat{y}_j\langle\hat{s}_j,y_k\rangle - A\hat{s}_j\langle\hat{y}_j,s_k\rangle}{\langle\hat{s}_j,\hat{y}_j\rangle} \tag{9.3.5}$$

The numerator of the fraction in (9.3.5) can be rewritten

$$\hat{y}_j\langle\hat{s}_j,y_k\rangle - A\hat{s}_j\langle\hat{y}_j,s_k\rangle = \hat{y}_j[\langle\hat{s}_j,y_k\rangle - \langle\hat{y}_j,s_k\rangle] + (\hat{y}_j - A\hat{s}_j)\langle\hat{y}_j,s_k\rangle$$

$$= \hat{y}_j[\langle\hat{s}_j,y_k - As_k\rangle - \langle\hat{y}_j - A\hat{s}_j,s_k\rangle] + (\hat{y}_j - A\hat{s}_j)\langle\hat{y}_j,s_k\rangle \tag{9.3.6}$$

Thus from (9.3.5) and (9.3.6).

$$\|\hat{y}_k - A\hat{s}_k\| \leq \|y_k - As_k\|$$

$$+ \sum_{j=\ell_k}^{k-1} \frac{\|\hat{y}_j\|\ \|\hat{s}_j\|\ \|y_k - As_k\| + 2\|\hat{y}_j\|\ \|s_k\|\ \|\hat{y}_j - A\hat{s}_j\|}{|\langle\hat{y}_j,\hat{s}_j\rangle|}$$

and using induction relations (9.3.1) and (9.3.2),

$$\|\hat{y}_k - A\hat{s}_k\| \leq \|y_k - As_k\| + 3\gamma \sum_{j=\ell_k}^{k-1} [\|y_k - As_k\| + 2\|s_k\|\sigma_j m_j].$$

By Lemma 9.2.1, $\|y_k - As_k\| \leq \|s_k\| Km_k$; and using also $m_j \leq m_k$ for

$\ell_k \le j < k$,

$$\| \hat{y}_k - A\hat{s}_k \| \le [1+(k-\ell_k)3\gamma] \| s_k \| m_k$$

$$+ (3\gamma)2\| s_k \| \sum_{j=\ell_k}^{k-1} K(3\gamma)^{j-\ell_k} \frac{(3^{j-\ell_k+1}-1)}{2} \tau^{j-\ell_k} m_j$$

$$\le \| s_k \| K(3\gamma)^{k-\ell_k} \tau^{k-\ell_k-1} m_k[(k-\ell_k+1)+2\sum_{j=1}^{k-\ell_k} (3^j-1)/2]$$

$$= \| s_k \| K(3\gamma)^{k-\ell_k} \tau^{k-\ell_k-1} m_k[(3^{k-\ell_k+1}-1)/2]$$

$$\le \| \hat{s}_k \| \sigma_k m_k \ .$$

This proves (9.3.1). The proof of (9.3.2) then proceeds identically to the case $i = 0$.

Lemma 9.3.2 Let F satisfy the conditions of Lemma 9.3.1. Let $(x_0, x_1, \ldots)$ be a sequence ϵR^n, let τ, ℓ_i, s_i, y_i, $\hat{s}_i$, $\hat{y}_i$, P_i be defined by Algorithm 9.1.7, and let A, m_i, γ, σ_i, σ be defined as in Lemma 9.3.1. Then for any non-negative integer i, if $m_i \le (2\sigma \| A^{-1} \|)^{-1}$, then $\| P_i \| \le 2\gamma$.

Proof: The proof basically shows that $A^{-1/2}P_iA^{1/2}$ is close to being a Euclidean projection matrix, which has norm 1.

Let $v_1, \ldots, v_{n+\ell_i-i}$ be a collection of linearly independent vectors ϵR^n such that

$$\langle v_j, \hat{y}_k \rangle = 0 \quad 1 \le j \le n+\ell_i-i \text{ and } \ell_i \le k \le i-1 \tag{9.3.7}$$

We first show that $\hat{s}_{\ell_i},\ldots,\hat{s}_{i-1},\ v_1,\ldots,v_{n+\ell_i-i}$ are linearly independent, and thus form a basis for R^n. Assume there exist constants $\alpha_{\ell_i},\ldots,\alpha_{i-1}$, $\beta_1,\ldots,\beta_{n+\ell_i-i}$ such that

$$u = \sum_{k=\ell_i}^{i-1} \alpha_k \hat{s}_k + \sum_{j=1}^{n+\ell_i-i} \beta_j v_j = 0$$

Since for each $k\epsilon[\ell_i,i-1]$, Lemma 9.1.1 and equation (9.3.7) yield $\langle u,\hat{y}_k\rangle = \alpha_k\langle\hat{s}_k,\hat{y}_k\rangle$, and $\langle\hat{s}_k,\hat{y}_k\rangle > 0$ by Lemma 9.3.1, $u = 0$ implies $\alpha_k = 0$, $k = \ell_i,\ldots,i-1$. Then since the v_i's are linearly independent, $u = 0$ implies that $\beta_j = 0$, $j = 1,\ldots,n+\ell_i-i$. Thus $\hat{s}_{\ell_i},\ldots,\hat{s}_{i-1},\ v_1,\ldots,v_{n+\ell_i-i}$ are linearly independent and form a basis for R^n.

Now consider any vector $w\epsilon R^n$, and write

$$w = s+v, \quad s = \sum_{k=\ell_i}^{i-1} \alpha_k \hat{s}_k, \quad v = \sum_{j=1}^{n+\ell_i-i} \beta_j v_j \ .$$

From Lemma 9.1.1 and equation (9.3.7), $P_i\hat{s}_k = 0$, $k = \ell_i,\ldots,i-1$, and $P_i v_j = v_j$, $j = 1,\ldots,n+\ell_i-i$. Thus

$$\|P_i\| = \max_{w\neq 0} \frac{\|P_i w\|}{\|w\|} = \max_{w\neq 0} \frac{\|v\|}{\|s+v\|}$$

$$\leq \max_{w\neq 0} \frac{\|A^{-1/2}\|\ \|A^{1/2}v\|}{\|A^{1/2}(s+v)\| / \|A^{1/2}\|} =$$

$$= \gamma \max_{w\neq 0} \frac{\|A^{1/2}v\|}{\|A^{1/2}(s+v)\|} \tag{9.3.8}$$

Now

$$\| A^{1/2}(s+v)\|^2 \geq \| A^{1/2}s\|^2 + \| A^{1/2}v\|^2 - 2\langle As,v\rangle \qquad (9.3.9)$$

and using (9.3.7) and Lemma 9.3.1,

$$\langle As,v\rangle = \sum_{k=\ell_i}^{i-1} \alpha_k \langle A\hat{s}_k,v\rangle = \sum_{k=\ell_i}^{i-1} \alpha_k[\langle\hat{y}_k,v\rangle+\langle A\hat{s}_k-\hat{y}_k,v\rangle]$$

$$\leq \| v\| \sum_{k=\ell_i}^{i-1} \alpha_k \| \hat{s}_k\| \sigma_k m_k \leq \| v\| \, \| s\| \, \sigma_{i-1} \, m_{i-1}$$

$$\leq \| A^{1/2}v\| \, \| A^{1/2}s\| \, \| A^{-1}\| \, \sigma m_{i-1} \qquad (9.3.10)$$

Substituting (9.3.10) into (9.3.9),

$$\| A^{1/2}(s+v)\|^2 \geq \frac{1}{2}\{(1-\sigma m_{i-1}\| A^{-1}\|)[\| A^{1/2}s\| +\| A^{1/2}v\|]^2$$

$$+ (1+\sigma m_{i-1}\| A^{-1}\|)[\| A^{1/2}s\| -\| A^{1/2}v\|]^2\}$$

$$\geq \frac{1}{4} \| A^{1/2}v\|^2 \qquad (9.3.11)$$

since $1-\sigma m_i \| A^{-1}\| \geq \frac{1}{2}$. Substituting (9.3.11) into (9.3.8) completes the proof.

Lemma 9.3.2 is used in the Q-superlinear convergence proofs in Theorems 9.3.3 and 9.3.5 to show that $\| \hat{s}_i\| \leq \| P_i^T\| \, \| s_i\| = \| P_i\| \, \| s_i\| \leq 2\gamma \| s_i\|$.

Theorem 9.3.3 Let $F: R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^* \epsilon D$ and $p > 0$, $K \geq 0$ that (9.2.1) holds, where $F(x^*) = 0$ and $F'(x^*)$ is symmetric and positive definite. Define $A = F'(x^*)$. Let the sequence $(x_0,x_1,\ldots)\epsilon R^n$ be generated from $x_0 \epsilon R^n$ and a symmetric

$B_0 \epsilon L(R^n)$ by the quasi-Newton algorithm

$$x_{i+1} = x_i - B_i^{-1} F(x_i)$$

$$B_{i+1} = U(x_i, \ldots, x_{i-\ell_i}, B_i) \text{ given by Update (9.1.13)}$$

Then there exist constants ϵ, $\delta > 0$ such that if $\| x_0 - x^* \| \leq \epsilon$ and $\| B_0 - A \|_{A^{-1/2}} \leq \delta$, the sequence $\{x_i\}$ is well-defined and converges Q-superlinearly to x^*.

Proof: We use Lemma 9.3.1 and the work of Broyden, Dennis and Moré (1973) to show that Update (9.1.13) satisfies (9.2.4). Using Theorem 9.2.2, this proves the linear convergence of our algorithm. We then use Lemmas 9.3.1 and 9.3.2 and work of Dennis and Moré (1974) to prove Q-superlinear convergence.

Let τ, ℓ_i, s_i, y_i, $\hat{s}_i$, $\hat{y}_i$, P_i be defined by Algorithm 9.1.7, m_i, γ, σ_i, σ be defined as in Lemma 9.3.1, and let $M = A^{-1/2}$. Broyden, Dennis and Moré (1973) (Lemma 5.2) show that if $\| A^{-1/2}\hat{y}_i - A^{1/2}\hat{s}_i \| \leq \beta \| A^{1/2}\hat{s}_i \|$ for some $\beta \epsilon [0,1/3]$ then update (9.1.13) satisfies

$$\| B_{i+1} - A \|_M \leq \left[\sqrt{1-\alpha\theta_i^2} + \frac{15}{4} \frac{\| A^{-1/2}\hat{y}_i - A^{1/2}\hat{s}_i \|}{\| A^{1/2}\hat{s}_i \|} \right] \| B_i - A \|_M$$

$$+ 2(1+2\sqrt{n}) \frac{\| A^{-1/2}\hat{y}_i - A^{1/2}\hat{s}_i \|}{\| A^{1/2}\hat{s}_i \|}, \qquad (9.3.12)$$

$$\alpha \epsilon [3/8, 1], \quad \theta_i = \frac{\| A^{-1/2}(B_i - A)\hat{s}_i \|}{\| B_i - A \|_M \; \| A^{1/2}\hat{s}_i \|}$$

From Lemma 9.3.1, we know that if $m_i \leq (3\sigma\|A^{-1}\|)^{-1}$, then

$$\frac{\|A^{-1/2}\hat{y}_i - A^{1/2}\hat{s}_i\|}{\|A^{1/2}\hat{s}_i\|} \leq \frac{\|A^{-1/2}\|\;\|\hat{y}_i - A\hat{s}_i\|}{\|\hat{s}_i\| / \|A^{-1/2}\|}$$

$$\leq \sigma_i m_i \|A^{-1}\| \leq \sigma m_i \|A^{-1}\| \leq \frac{1}{3} \tag{9.3.13}$$

Thus (9.3.12) holds and substituting (9.3.13) into it,

$$\|B_{i+1}-A\|_M \leq [\sqrt{1-\alpha\theta_i^2} + (15/4)\sigma\|A^{-1}\|m_i]\|B_i-A\|_M$$

$$+ 2(1+2\sqrt{n})\sigma\|A^{-1}\|\, m_i \tag{9.3.14}$$

Therefore there exists an open neighborhood around x^* - any neighborhood for which $m_i \leq (3\sigma\|A^{-1}\|)^{-1}$ - for which Update 9.1.13 satisfies equation (9.2.4), with $\alpha_1 = (15/4)\sigma\|A^{-1}\|$, $\alpha_2 = 2(1+2\sqrt{n})\sigma\|A^{-1}\|$, $\xi_i = \omega_i = 0$, $q = n-1$; and so by Theorem 9.2.2 there exist positive constants ε, δ such that for $\|x_0-x^*\| \leq \varepsilon$ and $\|B_0-A\|_M \leq \delta$, $\{x_i\}$ is well-defined and converges Q-linearly to x^*.

To demonstrate Q-superlinear convergence, we use Theorem 2.2 of Dennis and Moré (1974), which shows that under the conditions of this theorem and given the linear convergence just proven, $\{x_i\}$ is Q-superlinearly convergent to x^* if and only if

$$\limsup_{i\to\infty} \frac{\|(B_i-A)s_i\|}{\|s_i\|} = 0 \tag{9.3.15}$$

We use equation (9.3.14) to show that

$$\limsup_{i\to\infty} \frac{\| (B_i-A)\hat{s}_i \|}{\| \hat{s}_i \|} = 0 \tag{9.3.16}$$

and then with the help of Lemmas 9.3.1 and 9.3.2 show that (9.3.16) implies (9.3.15).

If $\liminf_{i\to\infty} \| B_i-A \|_M = 0$ then (9.3.16) clearly holds. If $\liminf_{i\to\infty} \| B_i-A \|_M$ is greater than or equal to some positive constant, then since $\lim_{i\to\infty} m_i = 0$ and $\alpha \geq 3/B$, it is clear from (9.3.14) that $\limsup_{i\to\infty} \theta_i = 0$, which implies (9.3.16) since $\| B_i-A \|_M \leq 2\delta$ from the proof of Theorem 9.2.1. Now using the definition of $\hat{s}_i$, Lemmas 9.3.1 and 9.3.2, and the fact that $B_i\hat{s}_j = \hat{y}_j$, $j = \ell_i,\ldots,i-1$ (whose very simple proof we omit)

$$\frac{\| (B_i-A)s_i \|}{\| s_i \|} \leq \frac{\| (B_i-A)\hat{s}_i \|}{\| s_i \|} + \sum_{j=\ell_i}^{i-1} \frac{\| (B_i-A)\hat{s}_j \| \, |<\hat{y}_j, s_i>|}{\| s_i \| <\hat{s}_j, \hat{y}_j>}$$

$$\leq \frac{\| (B_i-A)\hat{s}_i \|}{\| \hat{s}_i \| / \| P_i \|} + \sum_{j=\ell_i}^{i-1} \frac{\| \hat{y}_j - A\hat{s}_j \| \, 3\gamma}{\| \hat{s}_j \|}$$

$$\leq 2\gamma \frac{\| (B_i-A)\hat{s}_i \|}{\| \hat{s}_i \|} + \sum_{j=\ell_i}^{i-1} 3\gamma \, \sigma_j \, m_j$$

$$\leq 2\gamma \frac{\| (B_i-A)\hat{s}_i \|}{\| \hat{s}_i \|} + \sigma_i \, m_{i-1} \tag{9.3.17}$$

Since $\lim_{i\to\infty} m_i = 0$, $\sigma_i < \sigma$ and σ, γ are constants, (9.3.16) and (9.3.17) imply (9.3.15). This completes the proof of Q-superlinear convergence of $\{x_i\}$ to x^*.

<u>Lemma 9.3.4</u> Let F satisfy the conditions of Lemma 9.3.1. Let $(x_0, x_1, \ldots)$ be a sequence ϵR^n, let τ, ℓ_i, s_i, y_i, $\hat{s}_i$, $\hat{y}_i$ be defined as in Algorithm 9.1.7, and let A, m_i, γ, σ_i, σ be defined as in Lemma 9.3.1. Let the sequence $(B_0, B_1, \ldots) \epsilon L(R^n)$ be generated from a symmetric B_0 using Update 9.1.14 and points $(x_0, x_1, \ldots)$. Define

$$\omega_k = (1+3\gamma\tau)^{k-1} - 1.$$

Then for any non-negative integer i, if $m_i \leq (2\sigma \| A^{-1} \|)^{-1}$ then

$$\| B_i\hat{s}_j - \hat{y}_j \| \leq \| \hat{s}_j \| (K+\sigma_j)\omega_{i+j-2\ell_i}\, m_{i-1}, \quad j = \ell_i, \ldots, i-1 \qquad (9.3.18)$$

<u>Proof</u>: The proof is by induction on i, and like the proof of Lemma 9.3.1 uses messy but straightforward algebra. For i = 0 it is vacuously true. Now assume (9.3.18) holds for i = 0,...,k. If $\ell_{k+1} = k+1$, it is again vacuously true. Now assume $\ell_{k+1} < k+1$. Define $\alpha = (1+3\gamma\tau)$. We first prove (9.3.18) for $j \epsilon [\ell_{k+1}, k-1]$ (which may be any empty set) and then for j = k.

For $\ell_{k+1} \leq j \leq k-1$,

$$B_{k+1}\hat{s}_j = B_k\hat{s}_j + \frac{\hat{y}_k \langle y_k - B_k s_k, \hat{s}_j \rangle}{\langle \hat{y}_k, \hat{s}_k \rangle} \qquad (9.3.19)$$

since $\langle \hat{y}_k, \hat{s}_j \rangle = 0$. Rewriting the upper inner product term as

$$\langle y_k - B_k s_k, \hat{s}_j \rangle = \langle y_k - A s_k, \hat{s}_j \rangle - \langle s_k, \hat{y}_j - A\hat{s}_j \rangle - \langle s_k, B_k\hat{s}_j - \hat{y}_j \rangle$$

and substituting into (9.3.19) shows

$$\| B_{k+1}\hat{s}_j-\hat{y}_j\| \leq \| B_k\hat{s}_j-\hat{y}_j\| + \frac{\| \hat{y}_k\|[\| y_k-As_k\| \|\hat{s}_j\|+\| s_k\| \|\hat{y}_j-A\hat{s}_j\|+\| s_k\| \|B_k\hat{s}_j-\hat{y}_j\|]}{<\hat{y}_k,\hat{s}_k>}$$

Now using equation (9.1.B), Lemmas 9.2.1, 9.3.1, the induction hypothesis and the fact that $m_j \leq m_k$ for $\ell_k \leq j \leq k$,

$$\| B_{k+1}\hat{s}_j-\hat{y}_j\| \leq \| B_k\hat{s}_j-\hat{y}_j\|(1+3\gamma\tau) + 3\gamma\tau[\| \hat{s}_j\|Km_k+\| \hat{s}_j\|\sigma_j m_j]$$

$$\leq \| \hat{s}_j\|(K+\sigma_j)m_k[\alpha^{k+j-2\ell_{k+1}} - \alpha + \alpha -1]$$

$$= \| \hat{s}_j\|(K+\sigma_j)m_k\ \omega_{k+1-j-2\ell_{k+1}}$$

which proves (9.3.1B) for $j < k$.

Finally for $j = k$,

$$B_{k+1}\hat{s}_k-\hat{y}_k = B_{k+1}s_k-y_k + \sum_{q=\ell_{k+1}}^{k-1} \frac{\hat{y}_q<\hat{s}_q,y_k>-B_{k+1}\hat{s}_q<\hat{y}_q,s_k>}{<\hat{s}_q,\hat{y}_q>} \quad (9.3.20)$$

Rewriting the numerator of the last term as

$$\hat{y}_q<\hat{s}_q,y_k> - B_{k+1}\hat{s}_q<\hat{y}_q,s_k> = \hat{y}_q[<\hat{s}_q,y_k-As_k>-<\hat{y}_q-A\hat{s}_q,s_k>]$$

$$+ (\hat{y}_q-B_{k+1}\hat{s}_q)<\hat{y}_q,s_k>$$

and substituting this into (9.3.20), recalling that $B_{k+1}s_k = y_k$, gives

$$\| B_{k+1}\hat{s}_k-\hat{y}_k\| \leq$$

$$\sum_{q=\ell_{k+1}}^{k-1} \frac{\| \hat{y}_q \| [\| \hat{s}_q\| \|y_k-As_k\| + \| \hat{y}_q-A\hat{s}_q\| \|s_k\|] + \| \hat{y}_q-B_{k+1}\hat{s}_q\| \|\hat{y}_q\| \|s_k\|}{\langle \hat{s}_q,\hat{y}_q\rangle}$$

Using (9.1.8), Lemmas 9.2.1 and 9.3.1 and the induction hypotheses just proven

$$\| B_{k+1}\hat{s}_k-\hat{y}_k\| \leq \| \hat{s}_k\| \, 3\gamma\tau \sum_{q=\ell_{k+1}}^{k-1} [Km_k+\sigma_q m_q+(K+\sigma_q)m_k\ell_{k+1+q-2\ell_{k+1}}]$$

$$\leq \| \hat{s}_k\| \, 3\gamma\tau \, (K+\sigma_{k-1})m_k \sum_{q=\ell_{k+1}}^{k-1} \alpha^{k+q-2\ell_{k+1}}$$

$$= \| \hat{s}_k\| \, (K+\sigma_{k-1})m_k \, \alpha^k(\alpha^k-1)$$

$$\leq \| \hat{s}_k\| \, (K+\sigma_k)m_k \, \omega_{2k+1-2\ell_{k+1}}$$

which proves (9.3.18) for $j = k$, and completes the proof.

It is interesting to note that by the techniques of Lemma 9.3.4, we can also show that

$$\| B_i s_j-y_j\| \leq \| s_j\| \, 2K \, m_{i-1} \, \omega_{i-j} \qquad j = \ell_i,\ldots,i-1 \tag{9.3.21}$$

where the various symbols are defined as in Lemma 9.3.4. This is a smaller bound than the one Lemma 9.3.4 gives for $\| B_i\hat{s}_j-\hat{y}_j\|$. There are two reasons. The fact that $B_i s_j = y_j$ while $B_i\hat{s}_j$ only approximately equals $\hat{y}_j$ accounts for the ω_{i-j} term in (9.3.21) as opposed to the larger $\omega_{i+j-2\ell_i}$ in (9.3.18); while the fact that the bound on $\|y_j-F'(x^*)s_j\|$ from Lemma 9.1.1 is less than or equal to the bound on $\|\hat{y}_j-F'(x^*)\hat{s}_j\|$ from Lemma 9.3.1 accounts for the 2K term in (9.3.21) replacing the larger $K+\sigma_k$ term in (9.3.18). Although equation (9.3.21) is not used in any of our proofs, it is of considerable interest, because our real aim in using projected updates is to have $B_i s_j = y_j$ for $j = \ell_i,\ldots,i-1$.

<u>Theorem 9.3.5</u> Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^*\varepsilon D$ and $p > 0$, $K \geq 0$ that (9.2.1) holds where $F(x^*) = 0$ and $F'(x^*)$ is symmetric and positive definite. Define $A = F'(x^*)$. Let the sequence $(x_0,x_1,\ldots)\varepsilon R^n$ be generated from $x_0\varepsilon R^n$ and a symmetric $B_0\varepsilon L(R^n)$ by the quasi-Newton algorithm

$$x_{i+1} = x_i-B_i^{-1}\, F(x_i)$$

$$B_{i+1} = U(x_i,\ldots,x_{i-\ell_i},B_i) \text{ given by Update 9.1.14.}$$

Then there exist constants ϵ, $\delta > 0$ such that if $\| x_0-x^*\| \leq \epsilon$ and $\| B_0-A\|_{A^{-1/2}} \leq \delta$, the sequence $\{x_i\}$ is well-defined and converges Q-superlinearly to x^*.

<u>Proof</u>: Let B_{i+1} be the update given by Update 9.1.14, and $\overline{B}_{i+1}$ the update given by Update 9.1.13 for the same $x_i,\ldots,x_{i-q},B_i$. We use Lemmas 9.3.1 [illegible] show that if $m_i \leq (3\sigma\| A^{-1}\|)^{-1}$ (m_i, σ as defined in Lemma

9.3.1) then $\| B_{i+1}-\bar{B}_{i+1}\| \le \mu m_i$ for some constant μ. Combined with the bound on $\| \bar{B}_{i+1}-A\|_{A^{-1/2}}$ from Theorem 9.3.3, this shows that Update 9.1.14 obeys the conditions of Theorem 9.2.2, so that the algorithm is Q-linearly convergent. The proof of Q-superlinear convergence is almost the same as that in Theorem 9.3.3, but uses also Lemma 9.3.4.

We let τ, ℓ_i, s_i, y_i, $\hat{s}_i$, $\hat{y}_i$, P_i be defined by Algorithm 9.1.7, and γ, σ_i as in Lemma 9.3.1, ω_k as in Lemma 9.3.4. Define $u_i = y_i-B_is_i$ and $\hat{u}_i = \hat{y}_i-B_i\hat{s}_i$. Then, using $<\hat{y}_i,s_i> = <\hat{y}_i,\hat{s}_i>$ from Lemma 9.1.1

$$B_{i+1}-\bar{B}_{i+1} = \frac{(u_i-\hat{u}_i)\hat{y}_i^T+\hat{y}_i(u_i-\hat{u}_i)^T}{<\hat{y}_i,\hat{s}_i>} - \frac{(<u_i-\hat{u}_i,s_i>+<\hat{u}_i,s_i-\hat{s}_i>)\hat{y}_i\hat{y}_i^T}{<\hat{y}_i,\hat{s}_i>^2}$$

$$\| B_{i+1}-\bar{B}_{i+1}\| \le \frac{2\| u_i-\hat{u}_i\| \, \|\hat{y}_i\|}{<\hat{y}_i,\hat{s}_i>} + \frac{\| u_i-\hat{u}_i\| \, \|s_i\| \, \|\hat{y}_i\|^2}{<\hat{y}_i,\hat{s}_i>^2}$$

$$+ \frac{|<\hat{u}_i,s_i-\hat{s}_i>| \, \|\hat{y}_i\|^2}{<\hat{y}_i,\hat{s}_i>^2} \qquad (9.3.22)$$

We now bound $\| u_i-\hat{u}_i\|$ and $|<\hat{u}_i,s_i-\hat{s}_i>|$. Using the definitions of $\hat{s}_i$ and $\hat{y}_i$

$$u_i-\hat{u}_i = \sum_{j=\ell_i}^{i-1} \frac{\hat{y}_j<\hat{s}_j,y_i> - B_i\hat{s}_j<\hat{y}_j,s_i>}{<\hat{s}_j,\hat{y}_j>}$$

$$= \sum_{j=\ell_i}^{i-1} \frac{\hat{y}_j[<\hat{s}_j,y_i-As_i>-<\hat{y}_j-A\hat{s}_j,s_i>] + (\hat{y}_j-B_i\hat{s}_j)<\hat{y}_j,s_i>)}{<\hat{s}_j,\hat{y}_j>} .$$

Now using Lemmas 9.3.1, 9.3.4 and equation (9.1.8),

$$\| u_i-\hat{u}_i \| \le \| \hat{s}_i \| \, 3\gamma\tau \sum_{j=\ell_i}^{i-1} [Km_i+\sigma_j m_j+(K+\sigma_j)\omega_{i+j-2\ell_i} \, m_{i-1}]$$

$$\le \| \hat{s}_i \| \, 3\gamma\tau \sum_{j=\ell_i}^{i-1} (K+\sigma_j)(1+3\gamma\tau)^{i+j-2\ell_i-1} \, m_i$$

$$\le \| \hat{s}_i \| \, (K+\sigma_{i-1})(1+3\gamma\tau)^{2i-2\ell_i-1} \, m_i$$

$$\le \| \hat{s}_i \| \, (K+\sigma)(1+3\gamma\tau)^{2n-3} \, m_i \qquad (9.3.23)$$

Define $\beta = (K+\sigma)(1+3\gamma\tau)^{2n-3}$. We now bound $|<\hat{u}_i,s_i-\hat{s}_i>|$. Using Lemmas. 9.3.1, 9.3.4, equation (9.1.8), and the fact from Lemma 9.1.1 that $<\hat{y}_i,\hat{s}_j> = <\hat{s}_i,\hat{y}_j> = 0$ for $\ell_i \le j < i$,

$$<\hat{u}_i,s_i-\hat{s}_i> = <\hat{y}_i-B_i\hat{s}_i,s_i-\hat{s}_i> = -<B_i\hat{s}_i,s_i-\hat{s}_i>$$

$$= \sum_{j=\ell_i}^{i-1} \frac{-<\hat{s}_i,B_i\hat{s}_j><\hat{y}_j,s_i>}{<\hat{s}_j,\hat{y}_j>}$$

$$= \sum_{j=\ell_i}^{i-1} \frac{<\hat{s}_i,\hat{y}_j-B_i\hat{s}_j><\hat{y}_j,s_i>}{<\hat{s}_j,\hat{y}_j>}$$

$$|<\hat{u}_i,s_i-\hat{s}_i>| \le \| \hat{s}_i \|^2 \, 3\gamma\tau \sum_{j=\ell_i}^{i-1} (K+\sigma_j)(1+3\gamma\tau)^{i+j-2\ell_i-1} \, m_{i-1}$$

$$\leq \| \hat{s}_i \|^2 \, (K+\sigma_{i-1})(1+3\gamma\tau)^{2i-2\ell_i-1} \, m_{i-1}$$

$$\leq \| \hat{s}_i \|^2 \, \beta \, m_{i-1} \tag{9.3.24}$$

Now substituting (9.3.23) and (9.3.24) into (9.3.22) and using Lemma 9.3.1,

$$\| B_{i+1}-\overline{B}_{i+1} \| \leq 6\gamma \, \beta \, m_i + 9\gamma^2 \, \tau\beta \, m_i + 9\gamma^2 \, \beta \, m_i \triangleq \mu m_i$$

Define $M = A^{-1/2}$. Therefore we have

$$\| B_{i+1}-A \|_M \leq \| \overline{B}_{i+1}-A \|_M + \| B_{i+1}-\overline{B}_{i+1} \|_M$$

$$\leq \| \overline{B}_{i+1}-A \|_M + \| A^{-1} \| \sqrt{n} \, \mu \, m_i \tag{9.3.25}$$

Combining formula (9.3.25) with the bound on $\| B_{i+1}-A \|_M$ in (9.3.14)(which did not depend in any way on previous updates) shows that if $m_i \leq (3\sigma \| A^{-1} \|)^{-1}$, then Update 9.1.14 satisfies (9.2.4). Thus there exist positive constants ϵ, δ such that for $\| x_0-x^* \| \leq \epsilon$ and $\| B_0-A \|_M \leq \delta$, $\{x_i\}$ is well-defined and converges Q-linearly to x^*. The only change to the proof of Q-superlinear convergence in Theorem 9.3.3 is that since $B_i \hat{s}_j$ no longer $= \hat{y}_j$ for $j = \ell_i,\ldots,i-1$, equation (9.3.17) is replaced by

$$\frac{\| (B_i-A)s_i \|}{\| s_i \|} \leq \frac{\| (B_i-A)\hat{s}_i \|}{\| s_i \|} + \sum_{j=\ell_i}^{i-1} \frac{\| (B_i-A)\hat{s}_j \| \, |\langle \hat{y}_j, s_i \rangle|}{\| s_i \| \langle \hat{s}_j, \hat{y}_j \rangle}$$

$$\leq \frac{\| (B_i - A)\hat{s}_i \|}{\| \hat{s}_i \| / \| P_i \|} + \sum_{j=\ell_i}^{i-1} \frac{(\| \hat{y}_j - A\hat{s}_j \| + \| B_i \hat{s}_j - \hat{y}_j \|) 3\gamma}{\| \hat{s}_j \|}$$

$$\leq 2\gamma \frac{\| (B_i - A)\hat{s}_i \|}{\| \hat{s}_i \|} + 3\gamma \sum_{j=\ell_i}^{i-1} \sigma_j m_j$$

$$+ 3\gamma \sum_{j=\ell_i}^{i-1} (K + \sigma_j) \omega_{i+j-2\ell_i} m_{i-1}$$

$$\leq 2\gamma \frac{\| (B_i - A)\hat{s}_i \|}{\| \hat{s}_i \|} + \sigma_i m_{i-1} + (K + \sigma_{i-1}) \omega_{2i-2\ell_i} m_{i-1}$$

$$\leq 2\gamma \frac{\| (B_i - A)\hat{s}_i \|}{\| \hat{s}_i \|} + (\sigma + \beta) m_{i-1}$$

The remainder of the proof proceeds as before.

Finally we state the analogs of Theorems 9.3.3 and 9.3.5 for methods using the dual versions of Updates 9.1.13 and 9.1.14. The second of these is likely to be the most promising update analyzed in this section.

<u>Theorem 9.3.6</u> Let $F: R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^* \varepsilon D$ and $p > 0$, $K \geq 0$ that (9.2.1) holds where $F(x^*) = 0$ and $F'(x^*)$ is symmetric and positive definite. Define $A = F'(x^*)$. Let the sequence $(x_0, x_1, \ldots) \varepsilon R^n$ be generated from $x_0 \varepsilon R^n$ and a symmetric [illegible] R^n) by the quasi-Newton algorithm

$$x_{i+1} = x_i - H_i F(x_i)$$

$$H_{i+1} = H_i + \frac{(\hat{s}_i - H_i\hat{y}_i)\hat{s}_i^T + \hat{s}_i(\hat{s}_i - H_i\hat{y}_i)^T}{\langle \hat{s}_i, \hat{y}_i \rangle} - \frac{\langle \hat{s}_i - H_i\hat{y}_i, \hat{y}_i \rangle \hat{s}_i\hat{s}_i^T}{\langle \hat{s}_i, \hat{y}_i \rangle^2}$$

or

$$H_{i+1} = H_i + \frac{(s_i - H_i y_i)\hat{s}_i^T + \hat{s}_i(s_i - H_i y_i)^T}{\langle \hat{s}_i, y_i \rangle} - \frac{\langle s_i - H_i y_i, y_i \rangle \hat{s}_i\hat{s}_i^T}{\langle \hat{s}_i, y_i \rangle^2}$$

where $\hat{s}_i$, $\hat{y}_i$ are defined by Algorithm 9.1.7, except that the phrase "if $\| s_i \| \geq \tau \| s_i - Q_i^T s_i \|$" is replaced by "if $\| y_i \| \geq \tau \| y_i - Q_i y_i \|$." Then there exist constants ε, $\delta > 0$ such that if $\| x_0 - x^* \| \leq \varepsilon$ and $\| H_0 - A^{-1} \|_{A^{1/2}} \leq \delta$, the sequence $\{x_i\}$ is well-defined and converges Q-superlinearly to x^*.

CHAPTER X

THE EFFECTS OF HUANG'S AND OREN'S PARAMETERS ON Q-SUPERLINEAR CONVERGENCE

In this chapter we characterize the effect of two additional parameters on the convergence of some quasi-Newton methods. The first parameter, ρ_i, was introduced by Huang (1970) to change the quasi-Newton equation from $B_{i+1}s_i = y_i$ to $B_{i+1}s_i = \rho_i y_i$. This may be done to avoid singularity or help introduce higher derivative information. The second parameter, γ_i, was introduced by Oren (1972) to make the intermediate change $B_i' = \gamma_i B_i$ before updating B_i' by some standard quasi-Newton formula. It can be useful in getting B_i to have the right size early in an algorithm, and also has interesting application in recent quasi-Newton algorithms for the nonlinear least squares problem (see Dennis, Gay and Welsch (1977)).

Our main point in this chapter is that ultimately, either of these parameters must converge to the standard value one in order to preserve Q-superlinear convergence. In Section 10.1 we show that if $\sum_{i=0}^{j} |\rho_i - 1|$ or $\sum_{i=0}^{j} |\gamma_i - 1|$ are uniformly bounded for all j, then the DFP or BFGS algorithms modified to use parameters ρ_i or γ_i retain their eventual Q-superlinear convergence. It is clear from our proofs that if we make a Huang or Oren modification to any other algorithm which has been proven Q-superlinearly convergent by the techniques of Broyden, Dennis and Moré (1973) or their generalizations, that these same conditions will be sufficient to retain Q-superlinear convergence.

In Section 10.2 we discuss weaker necessary conditions on the parameters ρ_i and γ_i for Q-superlinear convergence. For the Huang parameter

ρ_i, we show that any direct prediction quasi-Newton algorithm having $B_{i+1}s_i = \rho_i y_i$ must have $\lim_{i\to\infty} \rho_i = 1$ in order to be Q-superlinearly convergent for all problems satisfying our normal assumptions. For the Oren parameter γ_i, it is clear that we must specify the subsequent quasi-Newton update before we can say what effect the change $B_i' = \gamma_i B_i$ will have on the algorithm's convergence properties. We select a modification of Broyden's (1965) method (which is Q-superlinearly convergent under our normal assumptions), and show that in order for it to have Q-superlinearly convergence with Oren parameter γ_i for all problems satisfying our assumptions we must also have $\lim_{i\to\infty} \gamma_i = 1$. We suspect the same result holds for other methods, but the analysis is quite messy.

As in Chapter 9, we write our theorems to apply to the general problem of finding a zero of $F(x)$, $F:R^n \to R^n$, and denote the Jacobian of F at x by $F'(x)$. Also we again use $\|\cdot\|$ to stand for $\|\cdot\|_2$, and $\|\cdot\|_M$ to stand for $\|\cdot\|_{Fr,M}$.

Section 10.1 - Sufficient Conditions for Q-Superlinear Convergence

In this section we show that there exists a constant $\sigma > 0$ such that if the direct prediction DFP or BFGS algorithms are modified to include Huang and Oren parameters ρ_i or γ_i, and

$$\sum_{i=0}^{j} |\rho_i - 1|, \quad \sum_{i=0}^{j} |\gamma_i - 1| \leq \sigma \text{ for all } j \geq 0 \tag{10.1.1}$$

then the modified algorithm still possesses local Q-superlinear convergence. We interpret this to mean that if $\sum_{i=0}^{j} |\rho_i - 1|$ or $\sum_{i=0}^{j} |\gamma_i - 1|$ are uni-[illegible] bounded by [illegible] constant, then some tail of the sequence will [illegible]

[illegible]

We prove the Q-superlinear convergence of the DFP method modified by Huang parameter ρ_i in Theorem 10.1.1, and modified by Oren parameter γ_i in Theorem 10.1.2. The proofs use the term ω_i introduced in our general convergence theorem in Section 9.2. However it may be that any application of these theorems will actually satisfy (10.1.1) because

$$\rho_i,\ \gamma_i = h(\|x_{i+1}-x^*\|,\ \|x_i-x^*\|,\ldots,\|x_{i-q}-x^*\|$$

for some convex function h and some constant q. In this case the effect of the parameters ρ_i or γ_i could be absorbed by the $\alpha_2 m_i$ term in (9.2.4).

Theorem 10.1.3 states the analog of Theorems 10.1.1 and 10.1.2 for the BFGS method. It is clear from the simple proofs of Theorems 10.1.1 and 10.1.2 that almost any method proven convergent by the techniques of Broyden, Dennis and Moré (1973) and their generalizations can be modified by parameters ρ_i and γ_i without losing its Q-superlinear convergence if (10.1.1) is satisfied. Therefore the theorems of this section could also be stated for the projected rank two methods analyzed in Section 9.3.

<u>Theorem 10.1.1</u> Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^* \epsilon D$ and $p > 0$, $K \geq 0$ that (9.2.1) holds, where $F(x^*) = 0$ and $F'(x^*)$ is non-singular. Define $A = F'(x^*)$. Consider the sequences $(x_0,x_1,\ldots)\epsilon R^n$ and $(B_0,B_1,\ldots)\epsilon L(R^n)$ generated from $x_0 \epsilon R^n$ and a symmetric $B_0 \epsilon L(R^n)$ by the DFP algorithm with Huang parameters ρ_i,

$$x_{i+1} = x_i - B_i^{-1} F(x_i)$$

$$s_i = x_{i+1} - x_i$$

$$y_i = F(x_{i+1}) - F(x_i)$$

$$B_{i+1} = B_{i+1}(\rho_i) = B_i + \frac{(\rho_i y_i - B_i s_i)y_i^T + y_i(\rho_i y_i - B_i s_i)^T}{\langle y_i, s_i\rangle}$$

$$- \frac{\langle \rho_i y_i - B_i s_i, s_i\rangle y_i y_i^T}{\langle y_i, s_i\rangle^2} \qquad (10.1.2)$$

where each ρ_i is a nonzero real number. Then there exist positive constants ε, δ, σ such that if $\| x_0 - x^*\| \leq \varepsilon$, $\| B_0 - A\|_{A^{-1/2}} \leq \delta$ and $\sum_{i=0}^{j} |\rho_i - 1| \leq \sigma$ for all $j \geq 0$, then the sequence $\{x_i\}$ is well-defined and converges Q-superlinearly to x^*.

Proof: The proof consists of using the work of Broyden, Dennis and Moré (1973) for the case $\rho_i = 0$ to show that update (10.1.2) satisfies (9.2.4) with $q = 0$, $\xi_i = 0$, $\omega_i = (B/3)|\rho_i - 1|$. This proves the linear convergence of the algorithm, and the work of Broyden, Dennis and Moré then proves the Q-superlinear convergence.

Broyden, Dennis and Moré have shown that under the assumptions of this theorem, there exist open neighborhoods N_1 around x^* and N_2 around A, N_2 containing only non-singular matrices, such that if x_i, $x_{i+1} \in N_1$ and $B_i \in N_2$ is symmetric, then there exist non-negative constants α_1, α_2 such that $B_{i+1}(0)$ satisfies

$$\| B_{i+1}(0) - A\|_M \leq (1 + \alpha_1 m_i)\| B_i - A\|_M + \alpha_2 m_i \qquad (10.1.3)$$

where $M = A^{-1/2}$ and $m_i = \max\{\| x_{i+1} - x^*\|^p, \| x_i - x^*\|^p\}$. They show in addition that

$$\langle y_i, s_i\rangle \geq (2/3)\| A^{1/2} s_i\|^2 \qquad (10.1.4)$$

$$\| A^{-1/2}y_i - A^{1/2}s_i \| \leq (1/3) \| A^{1/2}s_i \| \tag{10.1.5}$$

Thus from (10.1.5)

$$\| A^{-1/2}y_i \| \leq \| A^{-1/2}y_i - A^{1/2}s_i \| + \| A^{1/2}s_i \| \leq (4/3) \| A^{1/2}s_i \| \tag{10.1.6}$$

so that combining (10.1.4) and (10.1.6)

$$\langle y_i, s_i \rangle \geq (3/B) \| A^{-1/2}y_i \|^2$$

Now update (10.1.2) can be rewritten

$$B_{i+1}(\rho_i) = B_{i+1}(0) + (\rho_i - 1) \frac{y_i y_i^T}{\langle y_i, s_i \rangle}$$

Thus since B_i is symmetric, there exist open neighborhoods N_1 around x^* and N_2 around A such that if $x_1, x_{i+1} \varepsilon N_1$ and $B_i \varepsilon N_2$,

$$\| B_{i+1}(\rho_i) - A \|_M \leq \| B_{i+1}(0) - A \|_M + |\rho_i - 1| \frac{\| y_i y_i^T \|_M}{\langle y_i, s_i \rangle}$$

$$= \| B_{i+1}(0) - A \|_M + |\rho_i - 1| \frac{\| A^{-1/2}y_i \|^2}{\langle y_i, s_i \rangle}$$

$$\leq \| B_{i+1}(0) - A \|_M + (B/3)|\rho_i - 1|$$

Combining this with (10.1.3),

$$\| B_{i+1} - A \|_M \leq (1 + \alpha_1 m_i) \| B_i - A \|_M + \alpha_2 m_i + (B/3)|\rho_i - 1|.$$

Therefore by Theorem 9.2.2, there exist constants ε, δ, $\sigma > 0$ such that if $\| x_0 - x^* \| \leq \varepsilon$, $\| B_0 - A \|_{A^{-1/2}} \leq \delta$ and $\sum_{i=0}^{j} |\rho_i - 1| \leq \sigma$ for all $j \geq 0$, the sequence $\{x_i\}$ is well-defined and converges Q-linearly to x^*. The proof of Q-superlinear convergence is then identical to that of Broyden, Dennis and Moré (1973) for the case $\rho_i = 0$.

<u>Theorem 10.1.2</u> Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^* \varepsilon D$ and $p > 0$, $K \geq 0$ that (9.2.1) holds, where $F(x^*) = 0$ and $F'(x^*)$ is symmetric and positive definite. Define $A = F'(x^*)$. Consider the sequences $(x_0, x_1, \ldots) \varepsilon R^n$ and $(B_0, B_1, \ldots) \varepsilon L(R^n)$ generated from $x_0 \varepsilon R^n$ and a symmetric $B_0 \varepsilon L(R^n)$ by the DFP algorithm with Oren parameters γ_i,

$$x_{i+1} = x_i - B_i^{-1} F(x_i)$$

$$s_i = x_{i+1} - x_i$$

$$y_i = F(x_{i+1}) - F(x_i)$$

$$B_{i+1} = B_{i+1}(\gamma_i) = \gamma_i B_i + \frac{(y_i - \gamma_i B_i s_i) y_i^T + y_i (y_i - \gamma_i B_i s_i)^T}{\langle y_i, s_i \rangle}$$

$$- \frac{\langle y_i - \gamma_i B_i s_i, s_i \rangle y_i y_i^T}{\langle y_i, s_i \rangle^2} \tag{10.1.7}$$

where each γ_i is a nonzero real number. Then there exist positive constants ε, δ, σ such that if $\| x_0 - x^* \| \leq \varepsilon$, $\| B_0 - A \|_{A^{-1/2}} \leq \delta$ and $\sum^{j} |\gamma_i - 1| \leq \sigma$ for all $j \geq 0$, then the sequence $\{x_i\}$ is well-defined and converges Q-superlinearly to x^*.

Proof: The proof is very similar to that of Theorem 10.1.1, with the only difference that this time we apply Theorem 9.2.2 with $\omega_i = 9\| B_i\|_{A^{-1/2}} |\gamma_i - 1|$.

As was shown in the proof of Theorem 10.1.1, we know from Broyden, Dennis and Moré (1973) that under the assumptions of Theorem 10.1.2, there exist open neighborhoods N_1 around x^* and N_2 around A, N_2 containing only non-singular matrices, such that if x_i, $x_{i+1} \epsilon N_1$ and $B_i \epsilon N_2$ is symmetric, then there exist non-negative constants α_1, α_2 such that $B_{i+1}(0)$ satisfies

$$\| B_{i+1}(0) - A\|_M \leq (1+\alpha_1 m_i)\| B_i - A\|_M + \alpha_2 m_i \qquad (10.1.8)$$

where $M = A^{-1/2}$ and $m_i = \max\{\| x_{i+1} - x^*\|^p, \| x_i - x^*\|^p\}$; and in addition

$$\langle y_i, s_i\rangle \geq (1/2)\| A^{1/2} s_i\| \, \| A^{-1/2} y_i\| .$$

Now update (10.1.7) can be rewritten

$$B_{i+1}(\gamma_i) = B_{i+1}(0) + (\gamma_i - 1)\left[B_i - \frac{B_i s_i y_i^T + y_i s_i^T B_i}{\langle y_i, s_i\rangle} + \frac{\langle s_i, B_i s_i\rangle y_i y_i^T}{\langle y_i, s_i\rangle^2}\right]$$

Thus since B_i is symmetric, there exist open neighborhoods N_1 around x^* and N_2 around A such that if $x_i, x_{i+1} \epsilon N_1$ and $B_i \epsilon N_2$,

$$\| B_{i+1}(\gamma_i) - A\|_M \leq \| B_{i+1}(0) - A\|_M +$$

$$|\gamma_i - 1|\left[\| B_i\| + \frac{2\| B_i\|_M \, \| A^{1/2} s_i\| \, \| A^{-1/2} y_i\|}{\langle y_i, s_i\rangle} + \frac{\| B_i\|_M \, \| A^{1/2} s_i\|^2 \| A^{-1/2} y_i\|^2}{\langle y_i, s_i\rangle^2}\right]$$

$$\leq \| B_{i+1}(0)-A\|_M + 9\| B_i\|_M \, |\gamma_i-1|$$

Combining this with (10.1.B),

$$\| B_{i+1}-A\|_M \leq (1+\alpha_1 m_i)\| B_i-A\|_M + \alpha_2 m_i + 9\| B_i\|_M \, |\gamma_i-1|$$

For B_i in neighborhood N_2, $\| B_i\|_M$ is bounded above. Therefore by Theorem 9.2.2, there exist constants ε, δ, $\sigma > 0$ such that if $\| x_0-x^*\| \leq \varepsilon$, $\| B_0-A\| \leq \delta$ and $\sum_{i=0}^{j} |\gamma_i-1| \leq \sigma$ for all $j \geq 0$, the sequence $\{x_i\}$ is well-defined and converges Q-linearly to x^*. The proof of Q-superlinear convergence is again identical to that of Broyden, Dennis and Moré for the case $\gamma_i = 0$.

Theorem 10.1.3 Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^* \varepsilon D$ and $p > 0$, $K \geq 0$ that (9.2.1) holds, where $F(x^*) = 0$ and $F'(x^*)$ is symmetric and positive definite. Consider the sequences $(x_0,x_1,\ldots)\varepsilon R^n$ and $(H_0,H_1,\ldots)\varepsilon L(R^n)$ generated from $x_0 \varepsilon R^n$ and a symmetric $H_0 \varepsilon L(R^n)$ by the BFGS algorithm with Huang parameters ρ_i and Oren parameters γ_i,

$$x_{i+1} = x_i - H_i F(x_i)$$

$$s_i = x_{i+1} - x_i$$

$$y_i = F(x_{i+1}) - F(x_i)$$

$$H_{i+1} = \gamma_i H_i + \frac{(\rho_i s_i - \gamma_i H_i y_i)s_i^T + s_i(\rho_i s_i - \gamma_i H_i y_i)^T}{\langle y_i, s_i\rangle}$$

$$- \frac{\langle \rho_i s_i - \gamma_i H_i y_i, y_i \rangle s_i s_i^T}{\langle y_i, s_i \rangle^2}$$

where ρ_i, γ_i are nonzero real numbers. Then there exist positive constants ε, δ, σ_1, σ_2 such that if $\| x_0 - x^* \| \le \varepsilon$, $\| H_0 - F'(x^*) \|_{F'(x^*)^{1/2}} \le \delta$, $\sum_{i=0}^{j} |\rho_i - 1| \le \sigma_1$ and $\sum_{i=0}^{j} |\gamma_i - 1| \le \sigma_2$ for all $j \ge 0$, then the sequence $\{x_i\}$ is well-defined and converges Q-superlinearly to x^*.

Section 10.2 - Necessary Conditions for Q-Superlinear Convergence

In this section we derive necessary conditions for quasi-Newton methods using Huang or Oren parameters ρ_i or γ_i to be Q-superlinearly convergent. The conditions are rather weak, but indicate the basic premise that these parameters should tend to 1 in practice. First we show, by a trivial example, that for any direct prediction quasi-Newton method which has $B_{i+1} s_i = \rho_i y_i$ to converge locally and Q-superlinearly for all problems satisfying our normal assumptions, $\lim_{i\to\infty} \rho_i$ must be 1. We do this in Theorem 10.2.1 by considering a one-dimensional problem, where the quasi-Newton equation completely characterizes the update.

For the parameter γ_i, such simple examples are not helpful, and it is also clear that we can only analyze the effect of this parameter if we know the underlying update it is used with. Hence we attempt to show by example that the local Q-superlinear convergence of some quasi-Newton method on some problem will be destroyed if the method is amended to include the intermediate step $B_i' = \gamma_i B_i$, and $\lim_{i\to\infty} \gamma_i \ne 1$. This seems to be a complicated task, and we select a modification of Broyden's (1965) method proposed by Gay and Schnabel (1977), and shown to be locally Q-superlinearly convergent under our normal assumptions, as it seems the easiest to analyze for its behavior with Oren parameter γ_i. We show by example in Theorem 10.2.2 that this method loses its Q-superlinear convergence if $\lim_{i\to\infty} \gamma_i \neq 1$. Most likely the same results can be obtained for other leading methods.

<u>Theorem 10.2.1</u> Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^* \epsilon D$ and $p > 0$, $K \geq 0$ that (9.2.1) holds, where $F(x^*) = 0$. Consider the sequences $(x_0, x_1, \ldots) \epsilon R^n$ and $(B_0, B_1, \ldots) \epsilon L(R^n)$ generated by a quasi-Newton algorithm of the form

$$x_{i+1} = x_i - B_i^{-1} F(x_i)$$

$$B_{i+1} = \text{a non-singular matrix such that } B_{i+1}(x_{i+1} - x_i)$$

$$= \rho_i (F(x_{i+1}) - F(x_i)) \qquad (10.2.1)$$

where each ρ_i is a nonzero real number. Then for some such F, if $B_0 \neq F'(x_0)$ and $x_0 \neq x^*$, $\{x_i\}$ converges Q-superlinearly to x^* only if $\lim_{i\to\infty} \rho_i = 1$.

<u>Proof</u>: The proof is by example. Let $n = 1$ and $F(x) = x$. Then $x^* = 0$, $F'(x) = 1$ for all x, F clearly obeys (9.2.1) (with $K = 0$), and scalar B_{i+1} is completely defined by quasi-Newton equation (10.2.1): $B_{i+1} = \rho_i$ as long as $x_{i+1} \neq x_i$. Thus $B_0 \neq 1$ and $x_0 \neq 0$ imply $x_1 \neq 0$, and for $i = 1, 2, \ldots$

$$x_i - B_i^{-1} F(x_i)$$

and

$$\| x_{i+1} - x^* \| = |1 - \frac{1}{\rho_{i-1}}| \ \| x_i - x^* \|$$

Therefore x_i converges Q-superlinearly to x^* if and only if $\lim\limits_{i \to \infty} (1 - \frac{1}{\rho_{i-1}})$ $= 0$, which occurs if and only if $\lim\limits_{i \to \infty} \rho_i = 1$.

To analyze the effect of Oren parameter γ_i on Q-superlinear convergence, we consider a simple projected update: one that always keeps

$$B_{i+1} s_i = y_i \quad \text{and} \quad B_{i+1} s_{i-1} = B_i s_{i-1} \qquad (10.2.2)$$

($s_i = x_{i+1} - x_i$, $y_i = F(x_{i+1}) - F(x_i)$). Such an update is possible if s_{i-1} and s_i are linearly independent. Gay and Schnabel (1977) propose a projected Broyden's update which obeys (10.2.2),

$$B_{i+1} = B_i + \frac{(y_i - B_i s_i) v_i^T}{\langle v_i, s_i \rangle}$$

$$(10.2.3)$$

$$v_i = s_i - \frac{s_{i-1} \langle s_{i-1}, s_i \rangle}{\langle s_{i-1}, s_{i-1} \rangle}$$

and their techniques show that it yields a locally Q-superlinearly convergent algorithm when used with direct quasi-Newton steps (as does Broyden's update, which has $v_i = s_i$ in (10.2.3)). In a two-dimensional example, equation (10.2.2) iteratively characterizes B_i (given B_0), and we show in Theorem 10.2.2 that if any such update is used with intermediate step $= \gamma_i B_i$ (so that now $B_{i+1} s_i = y_i$ and $B_{i+1} s_{i-1} = \gamma_i B_i s_{i-1}$) then the corresponding direct prediction quasi-Newton method does not retain its

Q-superlinear convergence if $\lim_{i\to\infty} \gamma_i \neq 1$. The proof is rather complicated, although its techniques may have further application in analyzing quasi-Newton examples. We work with the inverse form of the update (with $H_i' = \gamma_i H_i$) as it simplifies the algebra slightly.

<u>Theorem 10.2.2</u> Let $F:R^n \to R^n$ be differentiable in the open convex set D, and assume for some $x^*\epsilon D$ and $p > 0$, $K \geq 0$ that (9.2.1) holds, where $F(x^*) = 0$. Consider the sequences $(x_0, x_1, \ldots)\epsilon R^n$ and $(H_0, H_1, \ldots)\epsilon L(R^n)$ generated by a quasi-Newton algorithm of the form

$$x_{i+1} = x_i - H_i F(x_i)$$

$$s_i = x_{i+1} - x_i$$

$$y_i = F(x_{i+1}) - F(x_i)$$

H_{i+1} a matrix $\epsilon L(R^n)$ such that $H_{i+1}y_i = s_i$, and for $i > 0$,

$$H_{i+1}y_{i-1} = \gamma_i H_i y_{i-1}$$

where each γ_i is a nonzero real number. Then for some such F, if $H_0 \neq F'(x_0)^{-1}$ and $x_0 \neq x^*$, then $\{x_i\}$ converges Q-superlinearly to x^* only if $\lim_{i\to\infty} \gamma_i = 1$.

<u>Proof</u>: The proof is by example. Let $n = 2$ and $F(x) = x$. Then $x^* = 0$, $F'(x) = I$, and F clearly obeys (9.2.1) with $K = 0$. Also $y_i = s_i$ at each iteration. Define $\delta_i = \gamma_i - 1$. We will assume $|\delta_i|$ is bounded below by a positive constant, and show that no matter how small $\| x_0 - x^* \|$ and

$\| H_0 - I \|$ are (as long as they are both nonzero), x_i does not converge Q-superlinearly to x^*.

At any iteration i, $i > 0$, let the angle from s_i to $s_{i-1} = \phi_i$, and the angle from x_{i+1} to $s_i = \omega_i$ (see example, Figure 10.2.1a). If $\phi_i \neq 0°$ or $180°$, there exist unique α_i, β_i such that $x_{i+1} = \alpha_i s_i + \beta_i s_{i-1}$, and straightforward algebra shows that

$$\beta_i = \frac{\| x_{i+1} \| \sin \omega_i}{\| s_{i-1} \| \sin \phi_i} . \tag{10.2.4}$$

Hence, since $H_{i+1} y_i = H_{i+1} s_i = s_i$ and $H_{i+1} y_{i-1} = H_{i+1} s_{i-1} = (1+\delta_i) s_{i-1}$,

$$x_{i+2} = x_{i+1} - H_{i+1} x_{i+1} = -\delta_i \beta_i s_{i-1}, \tag{10.2.5}$$

which is nonzero unless $\delta_i = 0$ or x_{i+1}, s_i are linearly dependent (see example, Figure 10.2.1b). Thus it is clear that we can choose x_0, H_0 such that s_0, s_1 and x_2 are pairwise linearly independent no matter how small the positive quantities $\| x_0 - x^* \|$ and $\| H_0 - I \|$ are. (We also need to completely specify H_1 - any standard update will do.) Then we can choose $(\gamma_0, \gamma_1, \ldots)$ such that for all $i \geq 0$, $|\delta_i|$ is greater than or equal to some positive constant and s_i, s_{i+1}, x_{i+2} are pairwise linearly independent. Thus neither ϕ_i nor ω_i is $0°$ or $180°$.

Now assume that for some positive constant μ, $|\delta_i| \geq \mu$ for all $i \geq 0$, and $\{x_i\}$ is Q-superlinearly convergent to x^*. Thus from (10.2.4) and (10.2.5)

$$\limsup_{i \to \infty} \frac{\| x_{i+2} \|}{\| x_{i+1} \|} = \limsup_{i \to \infty} \left| \delta_i \frac{\sin \omega_i}{\sin \phi_i} \right| = 0$$

Therefore for any $\varepsilon > 0$, there exists $i(\varepsilon)$ such that for all $i > i(\varepsilon)$,

$\| x_{i+2} \| / \| x_{i+1} \| < \varepsilon\mu$, and so for all $i > i(\varepsilon)$

$$\left| \frac{\sin \omega_i}{\sin \phi_i} \right| < \varepsilon. \tag{10.2.6}$$

We show that (10.2.6) leads to a contradiction.

Let $\varepsilon < \min\{1/3, 1/2\mu^{1/2}\}$ be a positive number, and let $i > i(\varepsilon)$ be a positive integer. Let $(180°+\sigma_{i+1})$ be the angle from s_{i+1} to x_{i+1} (see example, Figure 10.2.1c). Then ϕ_{i+1}, the angle from s_{i+1} to s_i, is $180°+\sigma_{i+1}+\omega_i$. Therefore

$$\left| \frac{\sin \phi_{i+1}}{\sin \omega_{i+1}} \right| \leq \left| \frac{\sin \sigma_{i+1}}{\sin \omega_{i+1}} \right| + \left| \frac{\sin \omega_i}{\sin \omega_{i+1}} \right| \tag{10.2.7}$$

where ω_{i+1}, the angle from x_{i+2} to s_{i+1}, is $180° - (\sigma_{i+1}+\omega_i+\phi_i)$ if $\delta_i\beta_i < 0$ and $-(\sigma_{i+1}+\omega_i+\phi_i)$ if $\delta_i\beta_i > 0$. By the law of sines

$$\left| \frac{\sin \sigma_{i+1}}{\sin \omega_{i+1}} \right| = \frac{\| x_{i+2} \|}{\| x_{i+1} \|} < \varepsilon\mu \tag{10.2.8}$$

Using $\sin(\omega_{i+1}+\sigma_{i+1}) = \sin \omega_{i+1}\cos\sigma_{i+1}+\cos\omega_{i+1}\sin\sigma_{i+1}$ and (10.2.8),

$$|\sin(\omega_{i+1}+\sigma_{i+1})| < |\sin\omega_{i+1}|(1+\varepsilon\mu). \tag{10.2.9}$$

Also $|\sin(\omega_{i+1}+\sigma_{i+1})| = |\sin(\omega_i+\phi_i)|$, and since (10.2.6) implies $|\sin\omega_i| < \varepsilon$,

$$|\sin(\omega_i+\phi_i)| = \left| \sin\phi_i(\cos\omega_i + \frac{\sin\omega_i}{\sin\phi_i}\cos\phi_i) \right| > |\sin\phi_i|(\sqrt{1-\varepsilon^2}-\varepsilon) \tag{10.2.10}$$

Thus by (10.2.9) and (10.2.10),

$$|\sin \omega_{i+1}| > |\sin \phi_i| \frac{(\sqrt{1-\varepsilon^2}-\varepsilon)}{1+\varepsilon\mu} \qquad (10.2.11)$$

and combining (10.2.11) and (10.2.6)

$$\left|\frac{\sin \omega_i}{\sin \omega_{i+1}}\right| = \left|\frac{\sin \omega_i}{\sin \phi_i} \cdot \frac{\sin \phi_i}{\sin \omega_{i+1}}\right| < \frac{\varepsilon(1+\varepsilon\mu)}{\sqrt{1-\varepsilon^2}-\varepsilon} \qquad (10.2.12)$$

Now substituting (10.2.8) and (10.2.12) into (10.2.7),

$$\left|\frac{\sin \phi_{i+1}}{\sin \omega_{i+1}}\right| < \varepsilon(\mu + \frac{1+\varepsilon\mu}{\sqrt{1-\varepsilon^2}-\varepsilon})$$

which is $< 1/\varepsilon$ by $\varepsilon < \min\{1/3, 1/2\mu^{1/2}\}$. This contradicts the assumption that $|\sin \omega_j/\sin \phi_j| < \varepsilon$ for all $j > i(\varepsilon)$, and shows that $\limsup_{j\to\infty} (\|x_{j+1}\|/\|x_j\|) \neq 0$ if $|\delta_j| \geq \mu > 0$ for all j.

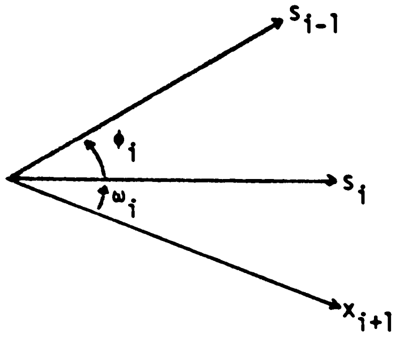

Figure 10.2.1a

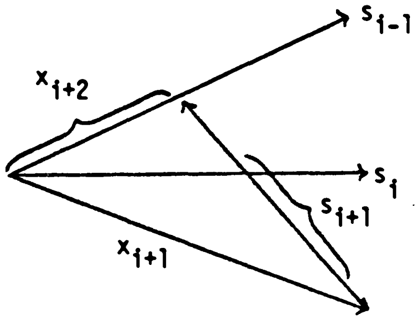

Figure 10.2.1b

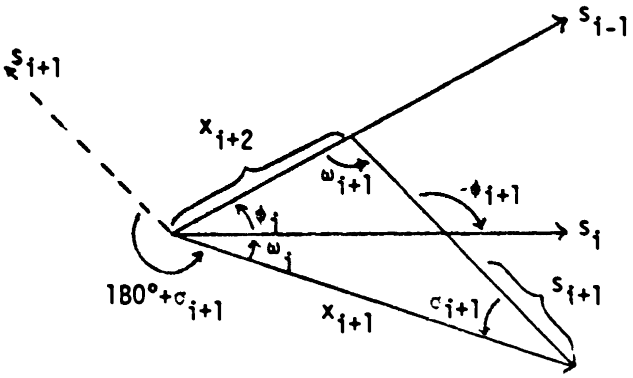

Figure 10.2.1c

CHAPTER XI

SUMMARY AND CONCLUSIONS

In this thesis we have analyzed, and to some extent tested, some of the leading new developments in quasi-Newton methods for unconstrained optimization. Foremost among these are the ideas of preserving past update information and of optimal conditioning which were introduced by Davidon (1975). We try to explain Davidon's work in Chapter 2. In conjunction we introduce the concepts of restricted updates and the restricted DFP-BFGS update class, which are an aid throughout the thesis. We analyze optimal conditioning extensively in both theory and practice in Chapters 7 and 8, and in Chapter 9 discuss different ways to implement Davidon's basic algorithm, some of which we show to be locally Q-superlinearly convergent.

We are helped by Chapters 3 - 6, which extend much of the existing analysis of symmetric rank-two quasi-Newton updates to include the more general updates we consider. Chapters 3 and 4 show how to analyze the important properties of general rank-two quasi-Newton updates, including hereditary positive definiteness, and Chapter 5 shows how to derive the minimum norm change update subject to a more general set of conditions. This work is directly applicable to Davidon's new update, helps us understand it and suggests methods for possibly improving upon it. Chapter 6 extends the analysis of quasi-Newton algorithms using perfect line search to a general class which includes Davidon's update class.

In Chapter 9 we also present a general convergence theorem for quasi-Newton methods which is somewhat broader than those previously in existance. It allows us to analyze updates which preserve past derivative information, and also aids in Chapter 10, where we discuss the effect on the convergence of quasi-Newton algorithms of two parameters, those of Huang (1970) and Oren (1972), whose value is only starting to be appreciated.

Thus we hope the effects of this thesis will fall into two main categories. By our work in Chapters 3 - 6 and Chapters 9 and 10, we hope we might facilitate the development of future quasi-Newton algorithms for various applications, by offering techniques for deriving updates with desired properties, and analyzing these updates. We think we are now at a point where quasi-Newton methods are just beginning to see their full spectrum of application, and so such work should be helpful. Secondly, by our work analyzing Davidon's algorithm and ideas, we hope to have improved the performance of quasi-Newton algorithms for unconstrained optimization, and to have made Davidon's ideas more accessible to other applications. Specifically for unconstrained optimization algorithms, we think we have demonstrated that the procedure called optimal conditioning can in theory be restricted to updates in the convex class, and in practice may not be helpful for most problems, except perhaps when used only at a few selected iterations. Our limited tests support these claims and demonstrate a need for more extensive testing. We have also raised interesting alternatives in the implementation of Davidon's basic algorithm for preserving past derivative information. Also we are beginning to witness the application of Davidon's idea of preserving past derivative information to other problems (such as systems of nonlinear equations) and hope to be of some assistance in this process.

Finally we comment upon the major unfinished tasks which we think are evident from this thesis. The main ones are computer testing of proposed modifications to Davidon's algorithm which we have either only suggested, or tested only incompletely. We list some of our suggestions for further work:

1) Testing of the alternative updates for preserving past derivative information which are proposed in Chapter 9.

2) Further testing, especially on singular problems, of the hybrid BFGS- optimal conditioning algorithms introduced in Chapter 8.

3) Convergence analysis of Davidon's (1975) algorithm or any competitive algorithms.

4) Application of Davidon's idea of preserving past derivative information to quasi-Newton methods for other problems, such as nonlinear least squares and constrained optimization.

5) Utilization of the parameters of Huang (1970) and Oren (1972) to improve the performance of quasi-Newton algorithms by incorporating higher derivative information or information about the general size of the matrix approximant.

6) More precise characterization of the effects of Huang's and Oren's parameters upon the convergence of existing methods.

7) Application of many of the techniques of this thesis to large problems where the derivative matrix has a significant spacity structure.

BIBLIOGRAPHY

Brodlie, K.W. (1976), "Unconstrained minimization," presented at the York State-of-the-Art in Numerical Analysis Conference, University of York, to appear in Modern Numerical Analysis, D. Jacobs, ed., Academic Press, London.

Brodlie, K.W., A.R. Gourlay and J. Greenstadt (1973), "Rank-one and rank-two corrections to positive definite matrices expressed in product form," Journal of the Institute of Mathematics and its Applications 11, 73-82.

Broyden, C.G. (1965), "A class of methods for solving nonlinear simultaneous equations," Mathematics of Computation 19, 577-593.

Broyden, C.G. (1967), "Quasi-Newton methods and their application to function minimization," Mathematics of Computation 21, 368-381.

Broyden, C.G. (1970), "The convergence of a class of double-rank minimization algorithms," Journal of the Institute of Mathematics and its Applications 6, 76-90.

Broyden, C.G., J.E. Dennis and J.J. Moré (1973), "On the local and superlinear convergence of quasi-Newton methods," Journal of the Institute of Mathematics and its Applications 12, 223-246.

Chien, A.S.A. (1974), "A class of derivative-free algorithms for unconstrained minimization," Ph.D. thesis, Department of Computer Science, Cornell University.

Davidon, W.C. (1959), "Variable metric method for minimization," Argonne National Laboratory Report ANL-5990 Rev.

Davidon, W.C. (1975), "Optimally conditioned optimization algorithms without line searches," Mathematical Programming 9, 1-30.

Dennis, J.E. (1972), "On some methods based on Broyden's secant approximation to the Hessian," in Numerical Methods for Nonlinear Optimization, F.A. Lootsma, ed., Academic Press, London.

Dennis, J.E. (1976), "Nonlinear least squares and equations," AERE Harwell Report C.S.S. 32, presented at the York State-of-the-Art in Numerical Analysis Conference, University of York, to appear in Modern Numerical Analysis, D. Jacobs, ed., Academic Press, London.

Dennis, J.E., D.M. Gay and R.E. Welsch (1977), "An adaptive nonlinear least squares algorithm," to appear.

Dennis, J.E. and H.H.W. Mei (1975), "An unconstrained optimization algorithm which uses function and gradient values," Cornell University Computer Science Technical Report 75-246, to appear in Journal of Optimization Theory and its Applications.

Dennis, J.E. and J.J. Moré (1974), "A characterization of superlinear convergence and its application to quasi-Newton methods," Mathematics of Computation 28, 549-560.

Dennis, J.E. and J.J. Moré (1977), "Quasi-Newton methods, motivation and theory," SIAM Review 19, 46-89.

Dixon, L.C.W. (1972a), "Quasi-Newton techniques generate identical points II: the proofs of four new theorems," Mathematical Programming 3, 345-358.

Dixon, L.C.W. (1972b), "Variable metric algorithms: necessary and sufficient conditions for identical behavior on non-quadratic functions," Journal of Optimization Theory and its Applications 10, 34-40.

Fletcher, R. (1970), "A new approach to variable metric algorithms," Computer Journal 13, 317-322.

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

Garcia-Palomares, U.M. (1973), "Superlinearly convergent quasi-Newton methods for nonlinear programming," Ph.D. thesis, Computer Sciences Department, University of Wisconsin, Madison.

Gay, D.M. (1976), "Representing symmetric rank 2 updates," National Bureau of Economic Research Working Paper No. 124.

Gay, D.M. and R.B. Schnabel (1977), "Solving systems of nonlinear equations by Broyden's method with projected updates," National Bureau of Economic Research Working Paper No. 169.

Goldfarb, D. (1970), "A family of variable-metric methods derived by variational means," Mathematics of Computation 24, 23-26.

Han, S.P. (1974), "Superlinearly convergent variable metric methods for general nonlinear programming problems," Ph.D. thesis, Computer Sciences Department, University of Wisconsin, Madison.

Huang, H.Y. (1970), "Unified approach to quadratically convergent algorithms for function minimization," Journal of Optimization Theory and its Applications 5, 405-423.

Mei, H.H.W. (1977), "An analysis and implementation of Davidon's techniques for unconstrained optimization," Ph.D. thesis, Department of Computer Science, Cornell University.

Oren, S.S. (1972), "Self-scaling variable metric methods for unconstrained minimization," Ph.D. thesis, Department of Engineering-Economic Systems, Stanford University.

Oren, S.S. and E. Spedicato (1976), "Optimal conditioning of self-scaling variable metric algorithms," Mathematical Programming 10, 70-90.

Ortega, J.M. and W.C. Rheinboldt (1970), Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York.

Powell, M.J.D. (1970a), "A hybrid method for nonlinear equations," in Numerical Methods for Non-Linear Algebraic Equations, P. Rabinowitz, ed., Gordon and Breach, London.

Powell, M.J.D. (1970b), "A new algorithm for unconstrained optimization," in Nonlinear Programming, J.B. Rosen, O.L. Mangasarian, and K. Ritter, eds., Academic Press, New York.

Powell, M.J.D. (1972), "Unconstrained minimization and extensions for constraints," AERE Harwell Report T.P. 495.

Powell, M.J.D. (1976a), "Some global convergence properties of a variable metric algorithm for minimization without exact line searches," in Nonlinear Programming, SIAM-AMS proceedings, Vol. IX, R.W. Cottle and C.E. Lemke, eds.

Powell, M.J.D. (1976b), "Quadratic termination properties of Davidon's new variable metric algorithm," AERE Harwell Report C.S.S. 15, to appear in Mathematical Programming.

Schnabel, R.B. (1976), "Optimal conditioning in the convex class of rank-two updates," Cornell University Computer Science Technical Report 76-288.

Schnabel, R.B. (1977), "Minimum norm symmetric quasi-Newton updates restricted to subspaces," Cornell University Computer Science Technical Report 77-300.

Shanno, D.F. (1970), "Conditioning of quasi-Newton methods for function minimization," Mathematics of Computation 24, 647-656.

Spedicato, E. (1975), "Computational experience with quasi-Newton algorithms for minimization problems of moderately large size," Computer Center, CISE, Segrate, Milano, Italy, Report No. CISE-N-175.

Stewart, G.W. (1973), Introduction to Matrix Computations, Academic Press, New York.

Stoer, J. (1975), "On the convergence rate of imperfect minimization algorithms in Broyden's β-class," Mathematical Programming 9, 313-335.

Wilkinson, J.H. and C. Reinsch (1971), Handbook of Automatic Computation, Volume II, Linear Algebra, Springer-Verlag, New York.