Quasi-Newton Methods, Motivation and Theory

John Dennis, Jorge Moré

QUASI-NEWTON METHODS, MOTIVATION AND THEORY

J. E. DENNIS, JR.† AND JORGE J. MORÉ‡

Abstract. This paper is an attempt to motivate and justify quasi-Newton methods as useful modifications of Newton's method for general and gradient nonlinear systems of equations. References are given to ample numerical justification; here we give an overview of many of the important theoretical results and each is accompanied by sufficient discussion to make the results and hence the methods plausible.

1. Introduction. Nonlinear problems in finite dimensions are generally solved by iteration. Davidon (1959), for the minimization problem, and Broyden (1965), for systems of equations, introduced new methods which although iterative in nature, were quite unlike any others in use at the time. These papers together with the very important modification and clarification of Davidon's work by Fletcher and Powell (1963) have sparked a large amount of research in the late sixties and early seventies. This work has led to a new class of algorithms which have been called by the names quasi-Newton, variable metric, variance, secant, update, or modification methods. Whatever one calls them (we will use quasi-Newton), they have proved themselves in dealing with practical problems of the two types mentioned; that is, systems of n equations in n unknowns, and the unconstrained minimization of functionals.

A predictable consequence of this research is that there has been a proliferation of quasi-Newton methods for unconstrained minimization. Moreover, the derivation and relationship between these methods has usually been obscured by appealing to certain idealized situations such as exact line searches and quadratic functionals. This has not happened in nonlinear equations since the only quasi-Newton method that has been seriously used is the one proposed by Broyden (1965).

In this paper we show that it is possible to derive all of the known practical quasi-Newton methods from very natural considerations and in such a way that the relationship between these methods is clear. In addition, this paper contains a survey of the theoretical results which yield insight into the behavior of quasi-Newton methods, and in order to motivate these methods, there is also some background material in §§ 2 and 6. In either case, we have only given those proofs which are either new, give insight, or are simpler than those previously published, but references are always given.

In §§ 4 and 7 we derive the various quasi-Newton updates. This is done by taking the point of view that these updates are methods for generating approximations to derivatives—Jacobians for nonlinear equations and Hessians in unconstrained minimization. This point of view suggests how to use quasi-Newton methods in other areas such as least squares and constrained optimization.

The theoretical results are contained in §§ 5 and 8. These results show, in particular, that there are four quasi-Newton updates which are globally and superlinearly convergent for linear problems (even in the absence of orthogonality assumptions or exact line searches), and locally and superlinearly convergent for nonlinear problems. These updates are Broyden's 1965 update for nonlinear equations, Powell's symmetric form of Broyden's update, the Davidon–Fletcher–Powell update, and the Broyden–Fletcher–Goldfarb–Shanno update. The theoretical results quoted tend to explain why these four updates are the ones most used in practical work.

In addition to the above material there are some rate of convergence results in § 3. In particular, we emphasize superlinear convergence and its geometric interpretation.

We use R^n to denote n-dimensional real Euclidean space with the usual inner product $\langle x, y\rangle = x^T y$ while $L(R^n)$ is the linear space of all real matrices of order n. Moreover, $\|\cdot\|$ stands for either the l_2 vector norm $\|x\| = \langle x, x\rangle^{1/2}$, or for any matrix norm which is consistent with (or subordinate to) the l_2 vector norm in the sense that $\|Ax\| \leqq \|A\|\,\|x\|$ for each x in R^n and A in $L(R^n)$. In particular, the l_2 operator norm and the Frobenius norm are consistent with the l_2 vector norm. For future reference we note that the Frobenius norm can be computed by

$$\|A\|_F^2 = \sum_{i=1}^{n} \|Av_i\|^2 = \operatorname{tr}(A^T A), \tag{1.1}$$

where $\{v_1, \cdots, v_n\}$ is any orthonormal set in R^n, and that for any pair A, B in $L(R^n)$,

$$\|AB\|_F \leqq \min\{\|A\|_2\|B\|_F, \|A\|_F\|B\|_2\}. \tag{1.2}$$

In addition to the above matrix norms, we also make use of the weighted norms

$$\|A\|_{M,2} = \|MAM\|_2, \qquad \|A\|_{M,F} = \|MAM\|_F, \tag{1.3}$$

where M is a nonsingular symmetric matrix in $L(R^n)$. These norms do not satisfy the submultiplicative property $\|AB\| \leqq \|A\|\,\|B\|$ which is usually satisfied by matrix norms, but are very useful because they can be used to measure the relative error of approximations to symmetric, positive definite matrices. To be specific, suppose that A is symmetric and positive definite, and let $A^{-1/2}$ be the symmetric positive definite square root of A^{-1}. Since

$$\frac{\|B-A\|}{\|A\|} \leqq \|A^{-1/2}(B-A)A^{-1/2}\|$$

for either the l_2 operator norm or the Frobenius norm, it is clear that if $M = A^{-1/2}$, then $\|B-A\|_{M,2}$ and $\|B-A\|_{M,F}$ measure the relative error of B as an approximation to A in the l_2 and Frobenius norms, respectively.

2. Variations on Newton's method for nonlinear equations. Let $F: R^n \to R^n$ be a mapping with domain and range in R^n and consider the problem of finding a solution to the system of n equations in n unknowns given by

$$f_i(x_1, \cdots, x_n) = 0, \qquad 1 \leqq i \leqq n,$$

where $f_1, \cdots, f_n$ are the component functions of F.

The best known method for attacking this problem is Newton's method, but sometimes it is modified so as to improve its computational efficiency. In this section we examine some of these variations and their corresponding advantages and disadvantages. This will help to motivate the introduction of quasi-Newton methods as variations of Newton's method.

For the purpose of analyzing the algorithms for solving $F(x) = 0$, the mapping F is assumed to have the following properties.

(2.1)
(a) The mapping F is continuously differentiable in an open convex set D.
(b) There is an x^* in D such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular.

The notation $F'(x)$ denotes the Jacobian matrix $(\partial_j f_i(x))$ evaluated at x so that (2.1) guarantees that x^* is a locally unique solution to the equations $F(x) = 0$.

In addition to (2.1) sometimes we will need the stronger requirement that F' satisfies a Lipschitz condition at x^*: There is a constant κ such that

$$\|F'(x) - F'(x^*)\| \leqq \kappa \|x - x^*\|, \qquad x \in D. \tag{2.2}$$

Note that if D is sufficiently small then (2.2) is satisfied if, for example, F is twice differentiable at x^*.

Newton's method for nonlinear equations can be derived by assuming that we have an approximation x_k to x^* and that in a neighborhood of x_k the linear mapping

$$L_k(x) = F(x_k) + F'(x_k)(x - x_k)$$

is a good approximation to F. If this is the case, then a presumably better approximation x_{k+1} to x^* can be obtained by solving the linear system $L_k(x) = 0$. Thus Newton's method takes an initial approximation x_0 to x^*, and attempts to improve x_0 by the iteration

$$x_{k+1} = x_k - F'(x_k)^{-1} F(x_k), \qquad k = 0, 1, \cdots.$$

Actually, this is the form of Newton's method which is convenient for analysis. The computational form consists of carrying out the following steps for $k = 0, 1, \cdots, m$ where m is the maximum number of iterations allowed.

(2.3)
(a) Compute $F(x_k)$ and if x_k is acceptable, stop. Otherwise, compute $F'(x_k)$.
(b) Solve the linear system $F'(x_k)s_k = -F(x_k)$ for s_k and set $x_{k+1} = x_k + s_k$.

The advantages of this algorithm are summarized in the following well-known result.

THEOREM 2.1. *Let $F: R^n \to R^n$ satisfy assumptions* (2.1). *Then there is an open set S which contains x^* such that for any $x_0 \in S$ the Newton iterates are well-defined, remain in S and converge to x^*. Moreover, there is a sequence $\{\alpha_k\}$ which converges to zero and with*

$$\|x_{k+1} - x^*\| \leqq \alpha_k \|x_k - x^*\|, \qquad k = 0, \cdots. \tag{2.4}$$

If, in addition, F satisfies (2.2) *then there is a constant β such that*

$$\|x_{k+1}-x^*\| \leqq \beta \|x_k - x^*\|^2, \qquad k = 0, \cdots. \tag{2.5}$$

For a proof of this result see, for example, Ortega and Rheinboldt (1970, p. 312). However, in § 5 we shall show that if F satisfies (2.2) then the convergence of Newton's method follows from a much more general result. Moreover (2.4) and (2.5) will follow from results in § 3.

Two advantages of Newton's method are expressed by Theorem 2.1. The first one is the existence of a domain of attraction S for Newton's method. The existence of this domain of attraction implies that if the Newton iterates ever land in S, then they will remain in S and eventually converge to x^*. This insures some measure of stability for the iteration.

The other advantage is expressed by (2.4) and is known as superlinear convergence. Moreover, if (2.2) holds then Theorem 2.1 shows that we obtain (at least) second order or quadratic convergence; that is, (2.5) holds. However, the example

$$f(x) = x + |x|^{1+\alpha}, \qquad \alpha \in (0, 1), \quad x^* = 0,$$

shows that in general (2.5) does not hold. If $\beta\|x^*\|$ is not too large, then an informal interpretation of (2.5) is that eventually each iteration doubles the number of significant digits in x_k as an approximation to x^*.

Also note that Newton's method is self-corrective; that is, x_{k+1} only depends upon F and x_k so that bad effects from previous iterations are not carried along. As we shall see, this is an advantage of Newton's method which is not shared by quasi-Newton methods.

The best known disadvantage of Newton's method is that a particular problem may require a very good initial approximation to x^* if the iteration is to converge. This is due to the fact that the set S in Theorem 2.1 can be very small. To overcome this disadvantage, special techniques (e.g. Powell's (1970a)) are needed.

On the other hand, for many problems the most important disadvantage of Newton's method is the requirement that $F'(x_k)$ be determined for each k. This involves the evaluation of n^2 scalar functions at each step and for most functions this is a very costly operation. It is usually taken to be equivalent to n evaluations of F, but the exact cost varies from problem to problem. If the Jacobian is relatively easy to obtain, then Newton's method is very attractive. If obtaining the Jacobian is relatively expensive, then this problem can be circumvented in some cases by using a finite difference approximation to the Jacobian matrix.

For example, $F'(x_k)$ could be replaced in (2.3) by the computation of $A(x_k, h_k) \in L(R^n)$ where

$$[A(x, h)]_{i,j} = [f_i(x + \eta_j e_j) - f_i(x)]/\eta_j, \tag{2.6}$$

and $h = (\eta_1, \cdots, \eta_n)$ is some suitably chosen vector. Of course, we now solve the system

$$A(x_h, h_k)s_k = -F(x_k) \tag{2.7}$$

for s_k.

There is a significant amount of theoretical and computational support for this approach. For example, if F satisfies assumptions (2.1) and (2.2), and at each iteration $\|h_k\| \leqq \gamma \|F(x_k)\|$ for some constant γ then all the conclusions of Theorem 2.1 also hold for the finite difference Newton's method. However the expense of computing n^2 scalar functions still remains. A popular technique for trying to reduce the overall computational effort of the Newton or the finite difference Newton's method is to hold the Jacobian fixed for a given number of iterations. This is particularly useful when the Jacobian is not changing very rapidly. However, it is always difficult to decide how long the Jacobian should be held fixed. Brent (1973) has shown that although this technique decreases the rate of convergence, it can increase a certain measure of efficiency.

Finally, note that all the modifications of Newton's method mentioned in this section require the solution of a system of linear equations and therefore $O(n^3)$ arithmetic operations per iteration. For some problems, the solution of these linear systems is the most expensive part of the iteration, and in these cases one should consider holding the Jacobian matrix fixed for a given number of iterations since in each such iteration this expense would be reduced to $O(n^2)$.

3. Rates of convergence. It is very important to understand something about the rate of convergence of different algorithms, since to a certain extent the rate of convergence of a method is as important as the fact that it converges; if it converges very slowly we may never be able to see it converge. Therefore, in this section we shall outline certain results which give insight into rates of convergence. In particular we emphasize the notion of superlinear convergence and mention its geometrical interpretation.

A reasonable algorithm should at least be linearly convergent in the sense that if $\{x_k\}$ is generated by the algorithm and $\{x_k\}$ converges to x^*, then for some norm $\|\cdot\|$ there is an $\alpha \in (0, 1)$ and $k_0 \geqq 0$ such that

$$\|x_{k+1} - x^*\| \leqq \alpha \|x_k - x^*\|, \qquad k \geqq k_0.$$

This guarantees that eventually the error will be decreased by the factor $\alpha < 1$. To be competitive an algorithm should be superlinearly convergent in the sense that (2.4) holds for some sequence $\{\alpha_k\}$ which converges to zero. As noted by Dennis and Moré (1974) one of the properties of superlinearly convergent methods is that

$$\lim_{k \to +\infty} \|x_{k+1} - x_k\| / \|x_k - x^*\| = 1 \tag{3.1}$$

provided, of course, that $x_k \neq x^*$ for $k \geqq 0$. That (3.1) holds is quite easy to prove and follows from (2.4) and the fact that

$$\big| \|x_{k+1} - x_k\| - \|x_k - x^*\| \big| \leqq \|x_{k+1} - x^*\|.$$

The importance of (3.1) is that it provides some justification for stopping the iteration when $\|x_{k+1} - x_k\| \leqq \varepsilon_1 \|x_k\|$ for some pre-specified ε_1. This termination criterion is often used together with one of the form $\|F(x_k)\| \leqq \varepsilon_2$. The reader can easily construct one dimensional examples to show the shortcomings of either criteria; a good routine should allow the user to select from several reasonable choices.

The following result of Dennis and Moré (1974) shows precisely when an iteration is superlinearly convergent.

THEOREM 3.1. *Let $F: R^n \to R^n$ satisfy assumptions* (2.1), *and let $\{B_k\}$ in $L(R^n)$ be a sequence of nonsingular matrices. Suppose that for some x_0 in D the sequence*

$$x_{k+1} = x_k - B_k^{-1}(x_k), \qquad k = 0, 1, \cdots, \tag{3.2}$$

remains in D, $x_k \neq x^*$ for $k \geqq 0$, and converges to x^*. Then $\{x_k\}$ converges superlinearly to x^* if and only if

$$\lim_{k \to +\infty} \frac{\|[B_k - F'(x^*)](x_{k+1} - x_k)\|}{\|x_{k+1} - x_k\|} = 0. \tag{3.3}$$

Clearly, if $\{B_k\}$ converges to $F'(x^*)$, then (3.3) holds and thus Theorem 3.1 explains why Newton's method and the finite-difference Newton's method with $\|h_k\| = O(\|F(x_k)\|)$ converges superlinearly. However, (3.3) only requires that $\{B_k\}$ converge to $F'(x^*)$ along the directions $s_k = x_{k+1} - x_k$ of the iterative method. As pointed out in §§ 5 and 8, this is the case for certain quasi-Newton methods, and yet for these methods $\{B_k\}$ does not, in general, converge to $F'(x^*)$.

An equivalent but more geometric formulation of (3.3) is that it requires $s_k = x_{k+1} - x_k$ in the iterative method to asymptotically approach the Newton correction $s_k^N = -F'(x_k)^{-1}F(x_k)$ in both length and direction. To see this note that since $F(x_k) = -B_k s_k$,

$$s_k - s_k^N = s_k + F'(x_k)^{-1}F(x_k) = F'(x_k)^{-1}[F'(x_k) - B_k]s_k,$$

and thus (3.3) is equivalent with

$$\lim_{k \to +\infty} \frac{\|s_k - s_k^N\|}{\|s_k\|} = 0. \tag{3.4}$$

Equation (3.4) shows that the relative error of s_k as an approximation to s_k^N approaches zero, and it is fairly easy to prove that this is equivalent to requiring that s_k approach s_k^N in both length and direction. For future reference, we state this formally.

LEMMA 3.2. *Let u, v belong to R^n with u, $v \neq 0$ and let $\alpha \in (0, 1)$. If $\|u - v\| \leqq \alpha \|u\|$, then $\langle u, v \rangle$ is positive and*

$$\left|1 - \frac{\|v\|}{\|u\|}\right| \leqq \alpha, \qquad 1 - \left(\frac{\langle u, v \rangle}{\|u\| \, \|v\|}\right)^2 \leqq \alpha^2. \tag{3.5}$$

Conversely, if $\langle u, v \rangle$ is positive and (3.5) *holds, then*

$$\|u - v\| \leqq 3\alpha \|u\|.$$

Proof. Assume first that $\|u - v\| \leqq \alpha \|u\|$. Then

$$\left|\frac{\|u\| - \|v\|}{\|u\|}\right| \leqq \frac{\|u - v\|}{\|u\|} \leqq \alpha,$$

and thus the first part of (3.5) holds. For the second part let $\omega = \langle u, v \rangle / (\|u\| \, \|v\|)$ and

note that

$$\|u-v\|^2=\|u\|^2-2\|u\|\,\|v\|\omega+\|v\|^2\geqq\|u\|^2(1-\omega^2).$$

This proves (3.5). Now note that if $\omega\leqq 0$ then the equality above shows that $\|u-v\|\geqq\|u\|$. Hence, $\alpha<1$ implies that $\langle u, v\rangle$ is positive. For the converse note that

$$\|u-v\|^2=(\|u\|-\|v\|)^2+2(1-\omega)\|u\|\,\|v\|\leqq\alpha^2\|u\|^2[1+2(1+\alpha)]$$

and since $\alpha<1$, it certainly follows that $\|u-v\|\leqq 3\alpha\|u\|$ as desired.

Lemma 3.2 shows that (3.4) is equivalent to

$$\lim_{k\to\infty}\frac{\|s_k^N\|}{\|s_k\|}=\lim_{k\to\infty}\left\langle\frac{s_k}{\|s_k\|},\frac{s_k^N}{\|s_k^N\|}\right\rangle=1,$$

and thus an iterative method is superlinearly convergent if and only if its directions asymptotically approach the Newton direction in both length and direction.

We would also like to explore second order convergence and for this we need the following estimate.

LEMMA 3.3. *Let $F:R^n\to R^n$ satisfy assumptions* (2.1) (*a*) *and* (2.2). *Then for any x^* in D,*

$$\|F(v)-F(u)-F'(x^*)(v-u)\|\leqq\kappa\max\{\|v-x^*\|,\|u-x^*\|\}\|v-u\| \tag{3.6}$$

for all v and u in D.

The proof of this result follows immediately from Theorem 3.2.5 of Ortega and Rheinboldt (1970); note that the assumption $F(x^*)=0$ is not necessary for Lemma 3.3 nor is the invertibility of $F'(x^*)$.

Using Lemma 3.3 it is not difficult to modify the proof of Theorem 3.1 as given by Dennis and Moré (1974) and show that if the assumptions of Theorem 3.1 are satisfied and (2.2) holds then there is a constant μ_1 such that

$$\|x_{k+1}-x^*\|\leqq\mu_1\|x_k-x^*\|^p,\qquad k=0,1,\cdots,$$

for some $p\in(1,2)$ if and only if there is a constant μ_2 such that

$$\|[B_k-F'(x^*)](x_{k+1}-x_k)\|\leqq\mu_2\|x_{k+1}-x_k\|^p,\qquad k=0,1,\cdots$$

However, we have not found any use for this result. The following well-known result is much easier to prove and is apparently just as useful.

THEOREM 3.4. *Let $F:R^n\to R^n$ satisfy assumptions* (2.1) *and* (2.2), *and let $\{B_k\}$ be a sequence of nonsingular matrices. Assume that for some x_0 in D the sequence* (3.2) *remains in D and converges to x^*. If*

$$\|B_k-F'(x^*)\|\leqq\eta\|x_k-x^*\|,\qquad k=0,1,\cdots, \tag{3.7}$$

then $\{x_k\}$ converges quadratically to x^*.

Proof. Since $\{x_k\}$ converges to x^*, inequality (3.7) and the Banach lemma (e.g. Ortega and Rheinboldt (1970, p. 45)) imply that there is a constant γ such that $\|B_k^{-1}\|\leqq\gamma$ for k sufficiently large. Since

$$x_{k+1}-x^*=-B_k^{-1}\{[F(x_k)-F(x^*)-F'(x^*)(x_k-x^*)]+(F'(x^*)-B_k)(x_k-x^*)\},$$

Lemma 3.3 together with (3.7) show that

$$\|x_{k+1}-x^*\| \leqq \gamma\{\kappa\|x_k-x^*\|^2+\eta\|x_k-x^*\|^2\},$$

and it follows that $\{x_k\}$ converges quadratically to x^*.

The most natural way to guarantee that (3.7) holds is to require that

$$\|B_k-F'(x_k)\| \leqq \eta_1\|F(x_k)\|, \qquad k \geqq 0. \tag{3.8}$$

If this is the case then

$$\|B_k-F'(x^*)\| \leqq \eta_1\|F(x_k)\|+\kappa\|x_k-x^*\|,$$

and Lemma 3.3 implies that (3.7) holds. Note that Newton's method and the finite difference Newton's method with $\|h_k\| = O(\|F(x_k)\|)$ satisfy (3.8).

4. Broyden's method. In § 2 we saw that two disadvantages of Newton's method were its need for n^2+n scalar function evaluations and its use of $O(n^3)$ arithmetic operations at each iteration. We shall now derive Broyden's method (1965) and show how it affects an order of magnitude reduction in each of these expenses. The price paid is a reduction from second order to superlinear convergence.

From the point of view taken here Broyden's 1965 proposal is a method for approximating Jacobian matrices. As pointed out in § 2, one of the major expenses of Newton's method is the calculation of $F'(x_k)$; let us now show how Broyden derived an approximation B_k to $F'(x_k)$ such that B_{k+1} can be obtained from B_k in $O(n^2)$ arithmetic operations per iteration and evaluating F at only x_k and x_{k+1}.

To derive this method, assume that $F: R^n \to R^n$ is continuously differentiable in an open convex set D and that for given x in D and $s \neq 0$, the vector $\bar{x} = x+s$ belongs to D. You should associate x with x_k and $\bar{x}$ with x_{k+1}, so that what we want is a good approximation to $F'(\bar{x})$.

Since F' is continuous at $\bar{x}$, given $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\|F(x)-F(\bar{x})-F'(\bar{x})(x-\bar{x})\| \leqq \varepsilon\|x-\bar{x}\|$$

provided $\|\bar{x}-x\| < \delta$. It follows that

$$F(x) \cong F(\bar{x})+F'(\bar{x})(x-\bar{x}),$$

the degree of approximation increasing as $\|x-\bar{x}\|$ decreases. Hence, if $\bar{B}$ is to denote our approximation to $F'(\bar{x})$, it seems reasonable to require that $\bar{B}$ satisfy the equation

$$F(x) = F(\bar{x})+\bar{B}(x-\bar{x}).$$

This is generally written

$$\bar{B}s = y \equiv F(\bar{x})-F(x), \tag{4.1}$$

where $s = \bar{x}-x$.

In the case of $n = 1$, equation (4.1) completely determines $\bar{B}$ and the secant method would result from using this approximate derivative in a Newton-like iteration. For $n > 1$, we can still argue that the only new information about F has

been gained in the direction determined by s. Now suppose we had an approximation B to $F'(x)$. Broyden reasoned that there really is no justification for having $\bar{B}$ differ from B on the orthogonal complement of s. This can be expressed as the requirement

$$\bar{B}z = Bz \quad \text{if } \langle z, s\rangle = 0. \tag{4.2}$$

Clearly (4.1) and (4.2) uniquely determine $\bar{B}$ from B and in fact

$$\bar{B} = B + \frac{(y - Bs)s^T}{\langle s, s\rangle}. \tag{4.3}$$

Equation (4.1) is central to the development of quasi-Newton methods, and therefore it has often been called the *quasi-Newton equation.* In fact, it also plays a role in a second derivation of Broyden's update.

The second derivation again starts from the assumption that any matrix that satisfies the quasi-Newton equation (4.1) is a good candidate for $\bar{B}$. However, now it is argued that out of all the matrices that satisfy the quasi-Newton equation, $\bar{B}$ should be the closest to B. The next result establishes that this matrix is again given by (4.3) if "closest" is measured by the Frobenius norm.

THEOREM 4.1. *Given $B \in L(R^n)$, $y \in R^n$ and some nonzero $s \in R^n$, define $\bar{B}$ by* (4.3). *Then $\bar{B}$ is the unique solution to the problem*

$$\min\{\|\hat{B} - B\|_F : \hat{B}s = y\}.$$

Proof. To show that $\bar{B}$ is a solution note that if $y = \hat{B}s$ then

$$\|\bar{B} - B\|_F = \left\|(\hat{B} - B)\frac{ss^T}{\langle s, s\rangle}\right\|_F \leqq \|\hat{B} - B\|_F$$

That $\bar{B}$ is the unique solution follows from the fact that the mapping $f: L(R^n) \to R$ defined by $f(A) = \|B - A\|_F$ is strictly convex in $L(R^n)$ and that the set of $\hat{B} \in L(R^n)$ such that $\hat{B}s = y$ is convex.

By now it should be clear how (4.3) can be used in an iterative method. For example, in its most basic form Broyden's method is defined by

$$x_{k+1} - x_k - B_k^{-1}F(x_k), \qquad k = 0, 1, \cdots, \tag{4.4}$$

where the matrices $B_k \in L(R^n)$ are generated by

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T}{\langle s_k, s_k\rangle}, \qquad k = 0, 1, \cdots, \tag{4.5}$$

with

$$y_k = F(x_{k+1}) - F(x_k), \quad \text{and} \quad s_k = x_{k+1} - x_k. \tag{4.6}$$

As it stands, it is clear that given x_0 and B_0, Broyden's method can be carried out with n scalar function evaluations per iteration. However, (4.4) and (4.5) seem to indicate that the solution of the linear system $B_k s_k = -F(x_k)$ is required. One way to overcome this difficulty requires the following result which is due to Sherman and Morrison (1949).

LEMMA 4.2. *Let $u, v \in R^n$ and assume that $A \in L(R^n)$ is nonsingular. Then $A + uv^T$ is nonsingular if and only if $\sigma = 1 + \langle v, A^{-1}u \rangle \neq 0$. If $\sigma \neq 0$, then*

$$(A + uv^T)^{-1} = A^{-1} - (1/\sigma)A^{-1}uv^TA^{-1}. \tag{4.7}$$

Proof. That $A + uv^T$ is nonsingular if and only if $\sigma \neq 0$ follows from Lemma 4.4 which will be proved later. It is easy to verify (4.7) because if the matrix on the right-hand side is multiplied by $A + uv^T$ then the result is the identity matrix.

From Lemma 4.2 it follows that if $H_k = B_k^{-1}$, then $H_{k+1} = B_{k+1}^{-1}$ is defined by

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)s_k^T H_k}{\langle s_k, H_k y_k \rangle} \tag{4.8}$$

provided $\langle s_k, H_k y_k \rangle \neq 0$. Therefore, Broyden's method can also be implemented as

$$x_{k+1} = x_k - H_k F(x_k),$$

where $\{H_k\}$ is generated by (4.8), and in this form Broyden's method only requires n scalar function evaluations and $O(n^2)$ arithmetic operations per iteration.

It is also possible to implement (4.5) and use only $O(n^2)$ arithmetic operations per iteration. For example, Gill and Murray (1972) describe a method by which if $B_k = Q_k R_k$ where Q_k is orthogonal and R_k is upper triangular, then the corresponding factorization of B_{k+1} can be obtained in $O(n^2)$ operations. Of course, if $B_k = Q_k R_k$ is given, then the solution of the linear system $B_k s_k = -F(x_k)$ only involves $O(n^2)$ operations. One reason why this approach would be preferable over (4.8) is because in (4.5) there are no matrix-vector multiplications; the term $B_k s_k$ is just $-F(x_k)$. Another reason is that the analysis of § 5 shows that (4.5) is more stable.

Note that we don't need to choose $s_k = x_{k+1} - x_k$ (see, however, the remarks after Theorem 5.4) in either (4.5) or (4.8). It is entirely reasonable to choose s_k to be any vector such that F is defined at $x_k + s_k$ and then set $y_k = F(x_k + s_k) - F(x_k)$. For example, if we set $s_k = \eta e^j$ for some scalar η, then (4.5) shows that B_{k+1} only differs from B_k in the jth column, and that this column is now

$$[F(x + \eta e^j) - F(x)]/\eta.$$

Of course, if $s_k \neq x_{k+1} - x_k$, then each iteration requires two function evaluations instead of one.

As theoretical justification for his method, Broyden only offered the fact that for affine functions it is norm-reducing with respect to the l_2 operator norm. The following well-known result shows that a slightly stronger result holds in the Frobenius norm.

THEOREM 4.3. *Let $A \in L(R^n)$ satisfy $y = As$ for some nonzero $s \in R^n$ and $y \in R^n$. Moreover, given $B \in L(R^n)$ define $\bar{B}$ by* (4.3). *Then*

$$\|\bar{B} - A\|_F \leqq \|B - A\|_F$$

with equality if and only if $\bar{B} = B$.

Proof. Since A lies in the affine subspace $\{B : y = Bs\}$ and since by Theorem 4.1, the matrix $\bar{B}$ is the orthogonal projection of B onto this subspace,

$$\|B - A\|_F^2 = \|\bar{B} - B\|_F^2 + \|\bar{B} - A\|_F^2.$$

The result follows from this relationship.

If $\{x_k\}$ is any sequence, and s_k, y_k are defined by (4.6), then $y_k = As_k$ for

$$A = \int_0^1 F'(x_k + \theta s_k)\, d\theta.$$

Thus, Theorem 4.3 guarantees that in the Frobenius norm, B_{k+1} is a better approximation than B_k to the average of F' on the line segment from x_k to x_{k+1}. Of course, if $F: R^n \to R^n$ is affine, then A is the coefficient matrix, and therefore for affine functions Broyden's method is norm-reducing in the Frobenius norm.

To conclude this section we point out that Broyden's method is sometimes implemented in the form

$$\bar{B} = B + \theta \frac{(y - Bs)s^T}{\langle s, s \rangle} \tag{4.9}$$

where θ is chosen so as to avoid singularity in $\bar{B}$. The following result can be used to decide how to choose θ.

LEMMA 4.4. *Let v, w in R^n be given. Then*

$$\det (I + vw^T) = 1 + \langle v, w \rangle. \tag{4.10}$$

Proof. Let $P = I + vw^T$ and assume that $v \neq 0$ for otherwise the result is trivial. Then any eigenvector of P is either orthogonal to w or a multiple of v. If the eigenvector is orthogonal to w, then the eigenvalue is unity while if it is parallel to v then the eigenvalue is $1 + \langle v, w \rangle$. Equation (4.10) follows.

To avoid singularity in $\bar{B}$ note that if $\bar{B}$ is defined by (4.9), then Lemma 4.4 yields

$$\det \bar{B} = \det B \left[(1 - \theta) + \theta \frac{\langle y, B^{-1}s \rangle}{\langle s, s \rangle} \right].$$

We can now follow a suggestion of Powell (1970a) and choose θ as a number closest to unity such that $|\det \bar{B}| \geqq \sigma |\det B|$ for some σ in $(0, 1)$; Powell uses $\sigma = 0.1$.

5. Local convergence results. We now would like to present a local convergence result that is available for Broyden's method and some of its variations. The importance of this result lies in the fact that the techniques used in its proof are applicable to other methods and in particular, to the double-rank updates of § 7.

In this analysis it is assumed that x_0 and B_0 are sufficiently close to x^* and $F'(x^*)$, respectively, where F satisfies assumptions (2.1) and (2.2). The convergence follows from a very general theorem due to Broyden, Dennis and Moré (1973). This result was developed to extend, to other quasi-Newton methods, the analysis given by Dennis (1971) for Broyden's method.

To describe the algorithms that this result handles, we shall need the concept of an update function. Update functions are only a means to denote the various Jacobian approximations which might be used in iterative processes. For example, consider iteration (3.2) where the matrices $\{B_k\}$ lie in a set D_M in $L(R^n)$ and F is defined on a set D. The method for generating $\{B_k\}$ can then be described by

specifying for each (x_k, B_k) a nonempty set $U(x_k, B_k)$ of possible candidates for B_{k+1}. Iteration (3.2) then becomes

$$\begin{aligned} x_{k+1} &= x_k - B_k^{-1}F(x_k), \\ B_{k+1} &\in U(x_k, B_k), \qquad k = 0, 1, \cdots. \end{aligned} \tag{5.1}$$

Thus U is a set-valued mapping whose domain, dom U, is a subset of $D \times D_M$ and whose range is contained in D_M. Note that D_M denotes the matrix part of the domain of U.

To illustrate these concepts note that for Newton's method $U(x, B) = \{F'(\bar{x})\}$ where $\bar{x} = x - B^{-1}F(x)$, while for Broyden's method $U(x, B) = \bar{B}$ where $\bar{B}$ is defined by (4.3) with $y = F(\bar{x})$ and $s = \bar{x} - x$. Also note that the finite difference form of Newton's method defined by (2.4) and (2.5) can be described by $U(x, B) = \{A(x, h): \|h\| \leqq \gamma \|F(x)\|\}$ where γ is a fixed nonnegative constant. This description has the advantage of not requiring a precise specification of the choice of h. Another illustration of the ease of description furnished by update functions is the following. Let U be given and for $(x, B) \in \text{dom } U$, set $\hat{U}(x, B) = U(x, B) \cup \{B\}$. Then $\hat{U}$ defines the modification to (5.1) in which B_k is not necessarily changed at each iteration. Finally note that in the above examples we can take $D_M = L(R^n)$.

Update functions also apply to the minimization algorithms of §§ 6 and 7. These algorithms are of the form (5.1), at least in a neighborhood of a local minimizer, where U is an update function for a gradient mapping. In this case D_M is usually the set of all symmetric matrices in $L(R^n)$.

The above examples lead us to the following definition. Given a set D_M in $L(R^n)$ and a mapping $F: R^n \to R^n$ defined on a set D, an *update function U for F on* D is a set-valued mapping from $D \times D_M$ into D_M. Thus $U(x, B)$ is a nonempty subset of D_M for each (x, B) in dom U.

The domain of U depends on the particular algorithm. For Newton's method the domain of the update function consists of all (x, B) in $D \times L(R^n)$ such that B is nonsingular and $\bar{x} = x - B^{-1}F(x)$ belongs to D. The domain of U for Broyden's method has the traditional restriction that $\bar{x} \neq x$. Of course, if $\bar{x} = x$, then $F(x) = 0$ and the algorithm stops. For iteration (5.1) it is convenient to define dom U as the set of all (x, B) in $D \times D_M$ such that B is nonsingular and $\bar{x} = x - B^{-1}F(x)$ belongs to D and differs from x.

THEOREM 5.1. *Let $F: R^n \to R^n$ satisfy assumptions* (2.1) *and* (2.2), *and let U be an update function for F such that for all* $(x, B) \in \text{dom } U$ *and* $\bar{B} \in U(x, B)$,

$$\|\bar{B} - F'(x^*)\| \leqq [1 + \alpha_1 \sigma(x, \bar{x})]\|B - F'(x^*)\| + \alpha_2 \sigma(x, \bar{x}) \tag{5.2}$$

for some constants α_1 and α_2 where $\bar{x} = x - B^{-1}F(x)$ and

$$\sigma(x, \bar{x}) = \max\{\|\bar{x} - x^*\|, \|x - x^*\|\}. \tag{5.3}$$

Then there are positive constants ε and δ such that if $x_0 \in D$ and $B_0 \in D_M$ satisfy $\|x_0 - x^*\| < \varepsilon$ and $\|B_0 - F'(x^*)\| < \delta$, then iteration (5.1) *is well-defined and converges linearly to x^*.*

By definition, iteration (5.1) is *locally convergent* at x^* if there is an $\varepsilon > 0$ and a $\delta > 0$ such that whenever $x_0 \in D$ and $B_0 \in D_M$ satisfy $\|x_0 - x^*\| < \varepsilon$ and

$\|B_0 - F'(x^*)\| < \delta$, then $\{x_k\}$ is well-defined and converges to x^*. Thus Theorem 5.1 guarantees the local and linear convergence of (5.1). Note that local convergence depends on D_M but since D_M is usually $L(R^n)$ or the set of symmetric matrices, D_M is large enough to make Theorem 5.1 meaningful. Also note that there is no restriction on the matrix norm in (5.2) because given any matrix norm $\|\cdot\|$ there is a constant $\gamma > 0$ such that the matrix norm $\gamma\|\cdot\|$ is consistent with the l_2 vector norm. Hence, if (5.2) holds for some matrix norm, it also holds for a matrix norm consistent with the l_2 vector norm.

Now obviously Theorem 5.1 cannot guarantee better than linear convergence since the stationary iteration $U(x, B) = \{B\}$ satisfies (5.2) with $\alpha_1 = \alpha_2 = 0$. The usual procedure is to use this theorem to prove the existence and convergence of $\{x_k\}$ and then apply Theorem 3.1 or Theorem 3.4 to make a more precise statement about the rate of convergence. We illustrate this below.

If F satisfies (2.2), then for Newton's method, $U(x, B) = \{F'(\bar{x})\}$ satisfies (5.2) with $\alpha_1 = 0$, $\alpha_2 = \kappa$ and $D_M = L(R^n)$. This proves the local convergence of Newton's method. The quadratic convergence follows from Theorem 3.4.

The proof of Theorem 2.1 that we have just given generalizes quite readily to the finite difference Newton's method defined by $x_{k+1} = x_k + s_k$ where s_k satisfies (2.6) and (2.7) with $\|h\| \leqq \gamma\|F(x)\|$ for some constant γ. We now turn to the application of Theorem 5.1 to Broyden's method.

THEOREM 5.2. *Let $F: R^n \to R^n$ satisfy assumptions* (2.1) *and* (2.2), *and consider Broyden's method as defined by equations* (4.4), (4.5) *and* (4.6). *Then Broyden's method is locally and superlinearly convergent at x^*.*

Proof. We shall prove that Broyden's method is locally convergent at x^* by showing that (5.2) is satisfied with $D_M = L(R^n)$. For this note that (4.5) implies that

$$\bar{B} - F'(x^*) = [B - F'(x^*)]\left[I - \frac{ss^T}{\langle s, s\rangle}\right] + \frac{(y - F'(x^*)s)s^T}{\langle s, s\rangle}.$$

In particular,

$$\|\bar{B} - F'(x^*)\| \leqq \|B - F'(x^*)\| + \frac{\|y - F'(x^*)s\|}{\|s\|}, \tag{5.4}$$

where the matrix norm is either the l_2 operator norm or the Frobenius norm. Therefore Lemma 3.3 implies that (5.2) is satisfied with $\alpha_1 = 0$ and $\alpha_2 = \kappa$. This proves the linear convergence of Broyden's method.

Like Newton's method, the more precise rate of convergence requires further work. In fact, we shall show that (3.3) holds. For this, note that direct computation using $\|A\|_F^2 = \operatorname{tr}(A^T A)$ shows that

$$\left\|E\left[I - \frac{ss^T}{\langle s, s\rangle}\right]\right\|_F^2 = \|E\|_F^2 - \left(\frac{\|Es\|}{\|s\|}\right)^2 \tag{5.5}$$

for any $E \in L(R^n)$, and since $(\alpha^2 - \beta^2)^{1/2} \leqq \alpha - (2\alpha)^{-1}\beta^2$,

$$\left\|E\left[I - \frac{ss^T}{\langle s, s\rangle}\right]\right\|_F \leqq \|E\|_F - (2\|E\|_F)^{-1}\left(\frac{\|Es\|}{\|s\|}\right)^2.$$

Now define $\eta_k = \|B_k - F'(x^*)\|_F$ and use the above inequality and Lemma 3.3 in (5.4) to obtain that

$$\eta_{k+1} \leqq [1 - (2\eta_k^2)^{-1}\psi_k^2]\eta_k + \kappa\sigma_k,$$

where $\sigma_k = \max\{\|x_{k+1} - x^*\|, \|x_k - x^*\|\}$ and

$$\psi_k = \frac{\|[B_k - F'(x^*)]s_k\|}{\|s_k\|}.$$

Since $\eta_{k+1} \leqq \eta_k + \kappa\sigma_k$ and $\{x_k\}$ is linearly convergent, it follows that $\{\eta_k\}$ is bounded, and if η is an upper bound, then

$$(2\eta)^{-1}\psi_k^2 \leqq \eta_k - \eta_{k+1} + \kappa\sigma_k.$$

Thus

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

forcing $\{\psi_k\}$ to converge to zero. Hence, (3.3) holds and this concludes the proof.

There are several interesting points about the proof of Theorem 5.2. The first is that although (3.3) holds, it does not necessarily follow that $\{B_k\}$ converges to $F'(x^*)$.

Example 5.3. Let $F: R^2 \to R^2$ be defined by $x = (\xi_1, \xi_2)^T$ and $F(x) = (\xi_1, \xi_2 + \xi_2^3)^T$, and consider Broyden's method with $x_0 = (0, \varepsilon)^T$ and

$$B_0 = \begin{pmatrix} 1+\delta & 0 \\ 0 & 1 \end{pmatrix}.$$

It is easy to verify that the (1, 1) element of B_k is always $1+\delta$ and thus $\{B_k\}$ does not converge to $F'(x^*)$.

The above example points out that one of the disadvantages of Broyden's method is that it is not self-correcting. In particular, B_k depends upon each B_j with $j < k$ and thus it may retain information which is irrelevant or even harmful.

Another point of interest about this proof is that it generalizes to the modification of Broyden's method given by (4.9). Thus Moré and Trangenstein (1976) prove that a parameter θ_k can be chosen so that if (4.5) is replaced by

$$B_{k+1} = B_k + \theta_k \frac{(y_k - B_k s_k)s_k^T}{\langle s_k, s_k \rangle}, \tag{5.6a}$$

$$B_{k+1} \text{ nonsingular} \quad |\theta_k - 1| \leqq \hat{\theta} \quad \text{and} \quad \hat{\theta} \in (0, 1), \tag{5.6b}$$

then Theorem 5.2 holds. They also noted that if F is affine, then for this modification the ε and δ in Theorem 5.1 are infinite.

THEOREM 5.4. *Let $F: R^n \to R^n$ be defined by $F(x) = Ax - b$, where $A \in L(R^n)$ is nonsingular and $b \in R^n$, and consider Broyden's method as defined by* (4.4), (4.6) *and* (5.6). *Then Broyden's method is globally and superlinearly convergent to $A^{-1}b$.*

The proof of Theorem 5.2 shows that Broyden's method is *linearly* convergent even if $s_k \neq x_{k+1} - x_k$. Thus, if we decide that $s_k = x_{k+1} - x_k$ is not a suitable direction, we can use (4.4), (4.5) but replace (4.6) by $y_k = F(x_k + s_k) - F(x_k)$,

where s_k is any nonzero vector such that

$$\|s_k\| \leqq \eta \max \{\|x_{k+1} - x^*\|, \|x_k - x^*\|\}$$

for some constant η. For example, the choice $s_k = \|F(x_{k+1})\| e_j$ is suitable for each j. Of course, if $s_k \neq x_{k+1} - x_k$, then the computation of y_k involves two evaluations of F and moreover, superlinear convergence will be lost unless, for example, the $\{s_k\}$ are uniformly linearly independent. For a discussion of this point, see Moré and Trangenstein (1976).

There is a variation of Broyden's method which is of interest in the case that $F'(x)$ is sparse. In this variation equations (4.4), (4.5) and (4.6) are used to define B_{k+1} from B_k but before it is used, B_{k+1} is forced to have the same sparsity pattern as $F'(x)$. That Theorem 5.2 holds follows from the observation that forcing B_k to have the same sparsity pattern as $F'(x)$ decreases $\|B_k - F'(x^*)\|$. Schubert (1970) has proposed an algorithm along these lines and Broyden (1971a) has shown that it is locally convergent. A graduate student at Cornell University, E. Marwil, has recently shown that Schubert's algorithm is superlinearly convergent.

We conclude this section by discussing two important variations of Theorem 5.2. The following variation arises because for some algorithms it is more natural to think of them as generating approximations to the inverse of the Jacobian. In this case dom U will be the set of all (x, H) in $D \times D_M$ such that $\bar{x} = x - HF(x)$ belongs to D and differs from x.

THEOREM 5.5. *Let $F: R^n \to R^n$ satisfy assumptions* (2.1) *and* (2.2), *and let U be an update function for F such that for all $(x, H) \in$ dom U and $\bar{H} \in U(x, H)$,*

$$\|\bar{H} - F'(x^*)^{-1}\| \leqq [1 + \alpha_1 \sigma(x, \bar{x})] \|H - F'(x^*)^{-1}\| + \alpha_2 \sigma(x, \bar{x}) \tag{5.7}$$

for some constants α_1 and α_2 where $\bar{x} = x - HF(x)$ and $\sigma(x, \bar{x})$ is defined by (5.3). *Then there are $\varepsilon > 0$ and $\delta > 0$ such that if $x_0 \in D$ and $H_0 \in D_M$ satisfy $\|x_0 - x^*\| < \varepsilon$ and $\|H_0 - F'(x^*)\| < \delta$, then the iteration*

$$\begin{aligned} x_{k+1} &= x_k - H_k F(x_k), \\ H_{k+1} &\in U(x_k, H_k), \qquad k = 0, 1, \cdots \end{aligned} \tag{5.8}$$

is well-defined and converges linearly to x^*.

The same remarks that we made after Theorem 5.2 for iteration (5.1) also apply, with suitable modifications, to (5.8). In particular, if (5.8) satisfies the conclusions of Theorem 5.5, then by definition (5.8) is locally and, of course, linearly convergent at x^*.

We also note that although Theorems 5.1 and 5.5 as well as their proofs are very similar, the two results are independent of each other. In fact, in § 8 we will discuss two important algorithms and show that the local convergence of one of these algorithms follows from Theorem 5.1 while the other needs Theorem 5.5.

Finally we note that it is possible to generalize both these theorems by showing that the conclusions still hold if instead of (5.1) and (5.8) we consider the sequence

$$x_{k+1} = x_k - \lambda_k B_k^{-1} F(x_k) \equiv x_k - \lambda_k H_k F(x_k)$$

provided the sequence $\{\lambda_k\}$ satisfies $|\lambda_k - 1| < \hat{\lambda}$ for some $\hat{\lambda} \in (0, 1)$.

6. Variations of Newton's method for unconstrained minimization. Let $f: R^n \to R$ be a functional defined on an open set D and consider the problem of finding a z in D such that $f(z) \leqq f(x)$ for each x in D. In this case z is a *global minimizer* of f and even if it is known to exist, finding it is usually an intractable task. Generally, one seeks z among the *local minimizers* of f; that is, find x^* in D such that for some $\delta > 0$,

$$f(x^*) \leqq f(x), \quad \|x - x^*\| \leqq \delta, \quad x \in D. \tag{6.1}$$

In this section we provide some background material and outline some of the methods that are used to solve (6.1). In particular, we stress the differences and analogies between the methods considered here and those in previous sections. This will help to motivate the introduction of quasi-Newton methods for unconstrained minimization.

We only consider the solution of (6.1) if f is differentiable. In this case (6.1) is usually attacked by trying to find a zero of ∇f—the gradient of f. This approach is based on the fact that if x^* is a local minimizer of f in the open set D and f is differentiable at x^*, then $\nabla f(x^*) = 0$. Moreover, in this section only descent methods are considered.

A *descent method* for solving (6.1) generates for each iterate x_k a direction p_k of local descent in the sense that there is a λ_k^* such that $f(x_k + \lambda p_k) < f(x_k)$ for $\lambda \in (0, \lambda_k^*]$. The next iterate is of the form $x_{k+1} = x_k + \lambda_k p_k$, where the parameter λ_k is chosen so that $f(x_{k+1}) < f(x_k)$. The directions p_k and the parameters should be chosen in such a way that $\{\nabla f(x_k)\}$ converges to zero. If $\|\nabla f(x_k)\|$ is small then usually x_k is near a zero of ∇f while the fact that $\{f(x_k)\}$ is decreasing indicates that this zero of ∇f is probably a local minimizer of f.

The simplest example of a descent method is the method of steepest descent. In this method we ask for the vector $\hat{p}$ of unit length (in the l_2 norm) such that for some $\hat{\lambda} > 0$,

$$f(x + \lambda \hat{p}) < f(x + \lambda p), \qquad \lambda \in (0, \hat{\lambda}), \quad p \neq \hat{p},$$

for all $\|p\| = 1$. It is not difficult to show that if $\nabla f(x) \neq 0$ then $\hat{p} = -\nabla f(x)/\|\nabla f(x)\|$. Therefore, the method of steepest descent is given by

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k), \qquad k = 0, 1, \cdots, \tag{6.2}$$

where the parameter λ_k is needed to guarantee that $f(x_{k+1}) < f(x_k)$; that such a parameter exists is a consequence of the following simple result.

LEMMA 6.1. *Let $f: R^n \to R$ be defined in an open set D and differentiable at x in D. If $\langle \nabla f(x), p \rangle < 0$ for some p in R^n then there is a $\lambda^* = \lambda^*(x, p)$ such that $\lambda^* > 0$ and*

$$f(x + \lambda p) < f(x), \qquad \lambda \in (0, \lambda^*).$$

The proof of this well-known result is quite easy and follows from the fact that

$$\lim_{\lambda \to 0^+} [f(x + \lambda p) - f(x)]/\lambda = \langle \nabla f(x), p \rangle.$$

Lemma 6.1 guarantees, in particular, that the parameter λ_k in the steepest descent method can be chosen so that $f(x_{k+1}) < f(x_k)$. This is not sufficient to show

that $\{x_k\}$ gets close to a zero of ∇f since λ_k may be arbitrarily small. In fact, $\lambda_k > 0$ can be chosen so that $\|x_{k+1} - x_k\| \leqq \varepsilon/2^k$ and therefore $\{x_k\}$ converges to a point x^* with $\|x_0 - x^*\| \leqq 2\varepsilon$. If $\nabla f(x_0) \neq 0$ and ∇f is continuous at x_0, then ε can be chosen so that $\nabla f(x^*) \neq 0$. At the end of this section we discuss a specific method for choosing λ_k which avoids this problem, and note that if λ_k is chosen appropriately, then the following result holds.

THEOREM 6.2. *Let $f: R^n \to R$ be continuously differentiable and bounded below on R^n, and assume that x_0 is such that ∇f is uniformly continuous on the level set*

$$L(x_0) = \{x \in R^n : f(x) \leqq f(x_0)\}.$$

Then there is a sequence $\{\lambda_k\}$ such that the steepest descent sequence (6.2) *is well defined, $\{f(x_k)\}$ is decreasing, and $\{\nabla f(x_k)\}$ converges to zero.*

One of the first proofs of Theorem 6.2 is that of Goldstein (1965). Since then this result has been generalized and refined; most of these extensions are discussed by Ortega and Rheinboldt (1970, Chap. 14) and Daniel (1971, Chap. 4 and 6).

If f is continuously differentiable on R^n and $L(x_0)$ is compact, then the rest of the assumptions of Theorem 6.2 are automatically satisfied and in addition, f has a global minimizer and $\{\nabla f(x_k)\}$ converges to zero. However, not even in this case is the steepest descent sequence guaranteed to converge to a local minimizer of f. An example reported by Wolfe (1971) shows that the steepest descent sequence may converge to a saddle point of f. Nevertheless, Theorem 6.2 is quite a strong convergence result. The fact that $\{\nabla f(x_k)\}$ converges to zero implies that any limit point of $\{x_k\}$ is a zero of ∇f and that for any $\varepsilon > 0$ the stopping criterion $\|\nabla f(x_k)\| < \varepsilon$ will be satisfied in a finite number of steps. Unfortunately, steepest descent usually converges linearly.

The slow rate of convergence of steepest descent can be improved by switching to a faster method in a neighborhood of a zero of ∇f. Since $F = \nabla f$ is a mapping from R^n to R^n, any of the methods discussed in §§ 2 and 4 could be used. For example, if f is twice differentiable, then Newton's method is given by

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k), \qquad k = 0, 1, \cdots, \tag{6.3}$$

where $\nabla^2 f(x)$ is the Hessian matrix of f at x; that is, $\nabla^2 f(x)$ is just the Jacobian matrix of ∇f. It should be clear that Theorem 2.1 applies to (6.3) with $F = \nabla f$, and that under the appropriate conditions we obtain local and quadratic convergence of (6.3) to a zero of ∇f.

In view of the global convergence of steepest descent and the fast local convergence of Newton's method, it would be desirable to have a method that behaves like Newton's method near a local minimizer but like steepest descent far from a local minimizer. Most descent methods of this type are of the form

$$x_{k+1} = x_k - \lambda_k B_k^{-1} \nabla f(x_k), \qquad k = 0, 1, \cdots, \tag{6.4}$$

where B_k is a symmetric, positive definite matrix which resembles $\nabla^2 f(x_k)$, at least in a neighborhood of a local minimizer.

As an example of such a method, Goldfeldt, Quandt and Trotter (1966)

suggested the iteration

$$x_{k+1} = x_k - \lambda_k(\nabla^2 f(x_k) + \mu_k I)^{-1}\nabla f(x_k), \qquad k = 0, 1, \cdots, \tag{6.5}$$

where the scalar $\mu_k \geqq 0$ is chosen so that $\nabla^2 f(x_k) + \mu_k I$ is positive definite. To justify the claim that (6.5) behaves like Newton's method in a neighborhood of a local minimizer, recall that if f is differentiable in an open set D and twice differentiable at a local minimizer x^* of f in D then $\nabla^2 f(x^*)$ is positive semidefinite. Therefore if x_k is in a neighborhood of a local minimizer, then very small values of μ_k will suffice to make $\nabla^2 f(x_k) + \mu_k I$ positive definite. Also note that if

$$s(\mu) = -(\nabla^2 f(x) + \mu I)^{-1}\nabla f(x),$$

then $s(0)$ is the Newton direction while as $\mu \to +\infty$ the angle between $s(\mu)$ and $-\nabla f(x)$ decreases monotonically to zero. Thus for large μ iteration (6.5) behaves like steepest descent.

In order to preserve, in (6.5), the good local properties of Newton's method, one has to choose μ_k and λ_k with some care. It is easy to see from Theorem 3.1 that as long as $\{\mu_k\}$ and $\{\lambda_k\}$ converge to zero and unity, respectively, iteration (6.5) is superlinearly convergent. Moreover, Theorem 3.4 shows that if $\mu_k \leqq \eta\|\nabla f(x_k)\|$ for some constant η and $\lambda_k = 1$ for all sufficiently large k, then (6.5) converges quadratically. Unfortunately, these results do not indicate how to choose $\{\mu_k\}$ globally, and in fact, this has turned out to be a hard problem.

There is a method of the form (6.4) which avoids the problem of choosing μ_k in (6.5) and yet resembles (6.5). In this method we try to obtain a Cholesky decomposition of $\nabla^2 f(x_k)$; that is, we try to find a nonsingular, lower triangular matrix L_k such that $\nabla^2 f(x_k) = L_k L_k^T$. Of course, if $\nabla^2 f(x_k)$ is not positive definite then this decomposition does not even exist, but the idea is that as the decomposition proceeds it is possible to add to the diagonal of $\nabla^2 f(x_k)$ and ensure that we obtain the Cholesky decomposition of a well-conditioned, positive definite matrix which differs from $\nabla^2 f(x_k)$ in some minimal way. In particular, if $\nabla^2 f(x_k)$ is a well-conditioned positive definite matrix then $\nabla^2 f(x_k) = L_k L_k^T$. The details are given by Murray (1972, p. 64). A more sophisticated version of the algorithm is given by Gill and Murray (1974), but for a factorization of the form $\tilde{L}_k D_k \tilde{L}_k^T$ where $\tilde{L}_k$ is a unit lower triangular matrix and D_k is a diagonal matrix with positive diagonal elements. Of course, the Cholesky decomposition can be obtained by realizing that $L_k = \tilde{L}_k D_k^{1/2}$.

In the remainder of this section we describe some of the selection rules for λ_k which are used in methods of the form (6.4) and more generally, in any descent method of the form

$$x_{k+1} = x_k + \lambda_k p_k, \qquad k = 0, 1, \cdots, \tag{6.6}$$

where $\langle \nabla f(x_k), p_k \rangle < 0$. The development of these particular rules are due to the initial work of Goldstein (1965) and Armijo (1966). Other selection rules for λ_k are discussed by Ortega and Rheinboldt (1970, pp. 249–258) and Jacoby, Kowalik and Pizzo (1972, Chap. 3).

In a descent method λ_k should satisfy $f(x_{k+1}) < f(x_k)$ but we have already noted that this requirement can be satisfied by arbitrarily small λ_k and then $\{x_k\}$ may converge to a point at which ∇f is not zero. A more reasonable requirement

(see Fig. 1) is that

$$f(x_k+\lambda_k p_k)\leqq f(x_k)+\alpha\lambda_k\langle\nabla f(x_k), p_k\rangle, \qquad \alpha\in(0, 1/2). \tag{6.7}$$

The reason for choosing $\alpha<1/2$ is that with this choice, Theorem 6.4 shows that if $\{x_k\}$ converges to a local minimizer of f at which $\nabla^2 f(x^*)$ is positive definite, and $\{p_k\}$ converges to the Newton step $-\nabla^2 f(x_k)^{-1}\nabla f(x_k)$ in both length and direction, then $\lambda_k=1$ will satisfy (6.7) for all sufficiently large k.

If α is close to zero then (6.7) is not a very stringent requirement, and α is generally chosen in this way with $[10^{-4}, 10^{-1}]$ being the usual range. However, it is not a good idea to fix λ_k by just requiring that it satisfy (6.7) since, for instance, $\lambda_k=0$ is then admissible. In general, unreasonably small λ_k are ruled out by the numerical search procedure but theoretically we need to impose another requirement. One such requirement is that

$$\langle\nabla f(x_k+\lambda_k p_k), p_k\rangle\geqq\beta\langle\nabla f(x_k), p_k\rangle, \qquad \beta\in(\alpha, 1). \tag{6.8}$$

The λ_k which satisfy (6.7) and (6.8) in Fig. 1 lie in the intervals J_1 and J_2. At the left endpoint of each of these intervals equality holds in (6.8) while at the right endpoint equality holds in (6.7). To show that there are λ_k which satisfy (6.7) and (6.8) assume that f is defined on R^n and $f(x_k+\lambda p_k)$ is bounded below for $\lambda\geqq 0$. It is then geometrically obvious that there are $\lambda_k>0$ for which equality holds in (6.7). If $\hat{\lambda}_k$ is the first such λ_k then the mean value theorem implies that

$$\hat{\lambda}_k\langle\nabla f(x_k+\theta_k\hat{\lambda}_k p_k), p_k\rangle=f(x_k+\hat{\lambda}_k p_k)-f(x_k)=\alpha\hat{\lambda}_k\langle\nabla f(x_k), p_k\rangle$$

for some $\theta_k\in(0, 1)$, and since $\alpha<\beta$,

$$\langle\nabla f(x_k+\theta_k\hat{\lambda}_k p_k), p_k\rangle\geqq\beta\langle\nabla f(x_k), p_k\rangle.$$

Thus $\lambda_k=\theta_k\hat{\lambda}_k$ satisfies (6.7) and (6.8). However, we emphasize that a search routine for λ should not necessarily try to satisfy (6.7) and (6.8). In fact, the intervals which satisfy these two conditions can be quite small (as for example, interval J_2 in Fig. 1) and therefore difficult to find. Moreover, to test whether or not (6.8) is satisfied requires the evaluation of ∇f. Instead, the search routine

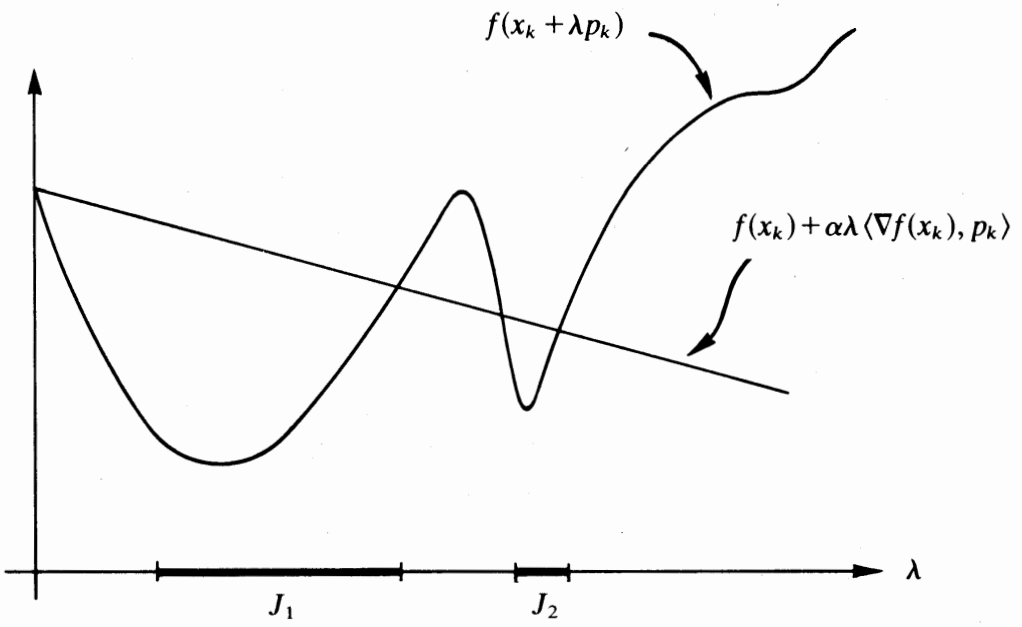

FIG. 1

should produce a λ_k which satisfies (6.7) and is not too small; (6.8) just guarantees that λ_k is not too small.

THEOREM 6.3. *Let $f: R^n \to R$ satisfy the assumptions of Theorem* 6.2, *and consider an iteration of the form* (6.6) *where the search directions p_k satisfy $\langle \nabla f(x_k), p_k \rangle < 0$. Then there is a sequence $\{\lambda_k\}$ which satisfies* (6.7) *and* (6.8) *and*

$$\lim_{k \to +\infty} \left\langle \nabla f(x_k), \frac{p_k}{\|p_k\|} \right\rangle = 0. \tag{6.9}$$

Theorem 6.3 is due to Wolfe (1969) who also pointed out that for many iterations (6.9) implies that $\{\|\nabla f(x_k)\|\}$ converges to zero; it is only necessary to verify that the angle between p_k and $\nabla f(x_k)$ stays bounded away from ninety degrees. For example, if $p_k = -\nabla f(x_k)$, or more generally, if $p_k = -B_k^{-1} \nabla f(x_k)$ where $\{B_k\}$ is a sequence of symmetric, positive definite matrices with uniformly bounded condition numbers, then

$$-\left\langle \nabla f(x_k), \frac{p_k}{\|p_k\|} \right\rangle \geqq \mu \|\nabla f(x_k)\|,$$

where μ^{-1} is an upper bound on the condition number of B_k. Hence, (6.9) ensures that $\{\|\nabla f(x_k)\|\}$ converges to zero.

To conclude this section we assume that the vectors p_k converge in direction and length to the Newton step and show that $\lambda_k = 1$ will eventually satisfy (6.7) and (6.8).

THEOREM 6.4. *Let $f: R^n \to R$ be twice continuously differentiable in an open set D and consider iteration* (6.6), *where $\langle \nabla f(x_k), p_k \rangle < 0$ and λ_k is chosen to satisfy* (6.7) *and* (6.8). *If $\{x_k\}$ converges to a point x^* in D at which $\nabla^2 f(x^*)$ is positive definite and*

$$\lim_{k \to \infty} \frac{\|\nabla f(x_k) + \nabla^2 f(x_k) p_k\|}{\|p_k\|} = 0, \tag{6.10}$$

then there is an index $k_0 \geqq 0$ such that $\lambda_k = 1$ is admissible for $k \geqq k_0$. Moreover, $\nabla f(x^*) = 0$ and $\{x_k\}$ converges superlinearly to x^*.

Proof. As a first step note that a consequence of (6.10) is that there is an $\eta > 0$ such that

$$-\langle \nabla f(x_k), p_k \rangle \geqq \eta \|p_k\|^2 \tag{6.11}$$

for all k large enough. This follows since

$$-\langle \nabla f(x_k), p_k \rangle = \langle \nabla^2 f(x_k) p_k, p_k \rangle - \langle \nabla^2 f(x_k) p_k + \nabla f(x_k), p_k \rangle,$$

so that (6.11) follows from (6.10) and the fact that $\nabla^2 f(x)$ is positive definite for all x close enough to x^*.

To show that (6.7) is eventually satisfied by $\lambda_k = 1$ use the mean value theorem to obtain u_k in the line segment from x_k to $x_k + p_k$ such that

$$f(x_k + p_k) - f(x_k) - \tfrac{1}{2} \langle \nabla f(x_k), p_k \rangle = \tfrac{1}{2} \langle \nabla^2 f(u_k) p_k + \nabla f(x_k), p_k \rangle.$$

Now (6.9) and (6.11) show that $\{p_k\}$ converges to zero; therefore (6.10) implies

that for all k sufficiently large

$$f(x_k+p_k)-f(x_k)-\tfrac{1}{2}\langle\nabla f(x_k), p_k\rangle \leqq (\tfrac{1}{2}-\alpha)\eta\|p_k\|^2, \tag{6.12}$$

and thus (6.11) and (6.12) show that (6.7) is satisfied by $\lambda_k = 1$. To prove that (6.8) is also eventually satisfied by $\lambda_k = 1$ we again use the mean value theorem to show that there is a v_k such that

$$\langle\nabla f(x_k+p_k), p_k\rangle = \langle\nabla f(x_k)+\nabla^2 f(v_k)p_k, p_k\rangle.$$

Thus (6.10) and (6.11) imply that for all k large enough,

$$\langle\nabla f(x_k+p_k), p_k\rangle \leqq \eta\beta\|p_k\|^2 \leqq -\beta\langle\nabla f(x_k), p_k\rangle.$$

Hence $\lambda_k = 1$ satisfies (6.8) and this concludes the first part of the proof. For the remainder, note that since $\{p_k\}$ converges to zero, (6.10) shows that $\nabla f(x^*) = 0$. The superlinear convergence of $\{x_k\}$ follows from Theorem 3.1.

7. Quasi-Newton methods for unconstrained minimization. The derivation of updates suitable for unconstrained optimization proceeds along lines similar to the development in § 4. For nonlinear equations only Broyden's method appears to be satisfactory, but here some notable differences, motivated by the discussion in § 6, will lead us to single out four reasonable update formulas.

One important consideration is the desire to have the quasi-Newton step $-B_k^{-1}\nabla f(x_k)$ define a descent direction. In fact, the most widespread use of these methods is in conjunction with iterations of the form (6.4). In this context the update formula should generate a sequence of symmetric positive definite matrices $\{B_k\}$ such that B_k resembles $\nabla^2 f(x_k)$, at least when x_k is near a local minimizer of f. We will examine these updates in § 7.2.

In § 7.1 we examine quasi-Newton methods which can be used to approximate the Hessian in such a way that the direction $p_k = -B_k^{-1}\nabla f(x_k)$ resembles the true Newton direction. In this case p_k may not be a descent direction, so that the direction is usually further modified. For example, it may be modified by adding to B_k a suitable multiple of the identity matrix as in iteration (6.5).

It is also possible to look at the updates of §§ 7.1 and 7.2 from an "inverse" point of view in which we try to generate approximations to the inverse of the Hessian. It turns out that this gives rise to at least one other important update. These inverse updates and their relationship to the updates of §§ 7.1 and 7.2 are examined in § 7.3.

Throughout this section we assume $f\colon R^n \to R$ to be twice differentiable in the open convex set D, and that we have an approximation B to $\nabla^2 f(x)$ for some x in D, and a direction s such that $x+s$ belongs to D. We now want to obtain a good approximation $\bar{B}$ to $\nabla^2 f(\bar{x})$ where $\bar{x} = x+s$.

7.1 Symmetry and the quasi-Newton equation. In view of the above discussion, and since the Hessian is symmetric, we want the update formula to have the property of hereditary symmetry; that is,

$$B \text{ symmetric} \quad \text{implies} \quad \bar{B} \text{ symmetric.} \tag{7.1}$$

Moreover, because of our desire to approximate the Hessian, arguments similar to those in § 4 lead us to require that $\bar{B}$ satisfy

$$\bar{B}s = y \equiv \nabla f(\bar{x}) - \nabla f(x). \tag{7.2}$$

Note that (7.2) is just the quasi-Newton equation (4.1) for $F = \nabla f$.

It is natural to ask whether it is possible to satisfy (7.1) and (7.2) with a rank one update formula. To see whether this can be done, first note that the general single-rank update that satisfies the quasi-Newton equation (7.2) is given by

$$\bar{B} = B + \frac{(y - Bs)c^T}{\langle c, s\rangle} \tag{7.3}$$

for $c \in R^n$ with $\langle c, s\rangle \neq 0$. If $\bar{B}$ is to satisfy (7.1), then it is easy to show that

$$\bar{B} = B + \frac{(y - Bs)(y - Bs)^T}{\langle y - Bs, s\rangle} \tag{7.4}$$

is the only solution provided $\langle y - Bs, s\rangle \neq 0$. If $y = Bs$, then $\bar{B} = B$ is the solution while if $y \neq Bs$ but $\langle y - Bs, s\rangle = 0$, then there is no solution.

This update is known as the *symmetric single-rank* formula. It seems to have been first published by Davidon (1959, Appendix), although Broyden (1967) and others discovered it independently later on. If $H = B^{-1}$ and $\bar{H} = \bar{B}^{-1}$ both exist and B is symmetric then the inverse relation

$$\bar{H} = H + \frac{(s - Hy)(s - Hy)^T}{\langle s - Hy, y\rangle} \tag{7.5}$$

holds. The following theorem, essentially due to Fiacco and McCormick (1968), shows that this method has very interesting behavior when it is applied to a quadratic functional.

THEOREM 7.1. *Let $A \in L(R^n)$ be a nonsingular symmetric matrix, and set $y_k = As_k$ for $0 \leqq k \leqq m$ where $\{s_0, \cdots, s_m\}$ spans R^n. Let H_0 be symmetric and for $k = 0, \cdots, m$ generate the matrices*

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{\langle s_k - H_k y_k, y_k\rangle}, \tag{7.6}$$

where it is assumed that

$$\langle s_k - H_k y_k, y_k\rangle \neq 0. \tag{7.7}$$

Then $H_{m+1} = A^{-1}$.

The proof of this result consists of verifying, by induction, that

$$H_k y_j = s_j, \quad 0 \leqq j < k, \quad \text{for } k = 1, \cdots, m+1.$$

Once this is done,

$$H_{m+1} y_j = H_{m+1} A s_j = s_j, \qquad 0 \leqq j \leqq m,$$

and the result follows from the assumption that $\{s_0, \cdots, s_m\}$ spans R^n.

The gist of Theorem 7.1 lies in the fact that if we have an iteration of the form $x_{k+1} = x_k + s_k$ and (7.7) holds, then the use of (7.6) allows one to minimize a

quadratic functional in a finite number of steps. Unfortunately, there is no guarantee that (7.7) will hold although it is not difficult to show (Goldfarb (1969)) that if $A^{-1}-H_0$ is semidefinite (positive or negative) and if $\{H_k\}$ is generated by (7.6) when (7.7) holds, and $H_{k+1}=H_k$ otherwise, then $H_{m+1}=A^{-1}$.

The fact that the vectors $s-Hy$ and y can be orthogonal forces a certain amount of numerical instability on the symmetric single-rank method. In particular, update (7.4) does not satisfy (5.2) or (5.7). These difficulties have led to several improvements in the basic algorithm, and in its modified form the method has been quite successful. See, for example, the numerical results of Dixon (1972b).

The numerical difficulties with the symmetric single-rank method have led to a whole class of updates which satisfy (7.1) and (7.2). The technique used to derive this class is due to Powell (1970d) who used it to obtain a double-rank version of Broyden's method. Dennis (1972) then showed that Powell's technique could be used to derive most of the well-known quasi-Newton updates.

In this derivation we begin with a symmetric $B\in L(R^n)$ and consider

$$C_1=B+\frac{(y-Bs)c^T}{\langle c,s\rangle}$$

as a possible candidate for $\bar{B}$. In general C_1 is not symmetric, so consider

$$C_2=(C_1+C_1^T)/2.$$

However, since C_2 does not satisfy the quasi-Newton equation, we repeat the process. In this way a sequence $\{C_k\}$ is generated by

$$\begin{aligned} C_{2k+1}&=C_{2k}+\frac{(y-C_{2k}s)c^T}{\langle c,s\rangle},\\ C_{2k+2}&=(C_{2k+1}+C_{2k+1}^T)/2, \end{aligned} \qquad k=0,1,\cdots, \tag{7.8}$$

where $C_0=B$. It turns out that $\{C_k\}$ has a limit $\bar{B}$ given by

$$\bar{B}=B+\frac{(y-Bs)c^T+c(y-Bs)^T}{\langle c,s\rangle}-\frac{\langle y-Bs,s\rangle}{\langle c,s\rangle^2}cc^T \tag{7.9}$$

and it is clear that this update satisfies (7.1) and (7.2).

LEMMA 7.2. *Let $B\in L(R^n)$ be symmetric and let c, s, and y be in R^n with $\langle c,s\rangle\neq 0$. If the sequence $\{C_k\}$ is defined by* (7.8) *with $C_0=B$, then $\{C_k\}$ converges to $\bar{B}$ as defined by* (7.9).

Proof. We only need to prove that the sequence $\{C_{2k}\}$ converges. If $G_k=C_{2k}$, then (7.8) shows that

$$G_{k+1}=G_k+\frac{1}{2}\frac{w_kc^T+cw_k^T}{\langle c,s\rangle} \tag{7.10}$$

where $w_k=y-G_ks$. In particular,

$$w_{k+1}=Pw_k, \qquad P=\frac{1}{2}\left[I-\frac{cs^T}{\langle c,s\rangle}\right].$$

It is clear that P has one zero eigenvalue and all other eigenvalues equal to $\frac{1}{2}$, so

that the Neumann lemma (e.g. Ortega and Rheinboldt (1970, p. 45) implies that

$$\sum_{k=0}^{\infty} w_k = \sum_{k=0}^{\infty} P^k(y - Bs) = (I-P)^{-1}(y - Bs). \tag{7.11}$$

Since

$$\lim_{k\to\infty} G_k = B + \sum_{k=0}^{\infty} (G_{k+1} - G_k),$$

it follows from (7.10) and (7.11) that $\{G_k\}$ converges. Thus since Lemma 4.2 shows that

$$(I-P)^{-1} = 2\left[I - (1/2)\frac{cs^T}{\langle c, s\rangle}\right],$$

equations (7.10) and (7.11) also imply that the limit of $\{G_k\}$ is $\bar{B}$ as defined by (7.9).

Once c is chosen, (7.9) is a rank two update which satisfies (7.1) and (7.2). Before looking at special cases of (7.9), we show that this update solves a problem similar to the one specified in Theorem 4.1.

THEOREM 7.3. *Let $B \in L(R^n)$ be symmetric, and let c, s, and y be in R^n with $\langle c, s\rangle > 0$. Assume that $M \in L(R^n)$ is any nonsingular, symmetric matrix such that $Mc = M^{-1}s$. Then $\bar{B}$ as defined by* (7.9) *is the unique solution to the problem*

$$\min\{\|\hat{B} - B\|_{M,F} : \hat{B} \text{ symmetric}, \hat{B}s = y\} \tag{7.12}$$

where $\|\cdot\|_{M,F}$ is defined by (1.3).

Proof. Let $\hat{B}$ be any symmetric matrix such that $y = \hat{B}s$, and pre- and post-multiply (7.9) by M. If $Mc = M^{-1}s = z$ it follows that

$$\bar{E} = \frac{Ezz^T + zz^TE}{\langle z, z\rangle} - \frac{\langle Ez, z\rangle}{\langle z, z\rangle^2} zz^T,$$

where $E = M(\hat{B} - B)M$ and $\bar{E} = M(\bar{B} - B)M$. Now it is clear that $\|\bar{E}z\| = \|\bar{E}z\|$, and that if v is orthogonal to z then $\|\bar{E}v\| \leqq \|Ev\|$. Thus $\|\bar{E}\|_F \leqq \|E\|_F$ as desired. To show uniqueness just note that the mapping $f : L(R^n) \to R$ defined by $f(A) = \|B - A\|_{M,F}$ is strictly convex on the convex set of symmetric $\hat{B}$ such that $\hat{B}s = y$.

Theorem 7.3 was inspired and is closely related to the results of Greenstadt (1970) and Goldfarb (1970) and it shows that the updates obtained by Greenstadt (1970) could also have been obtained by the symmetrization argument of Lemma 7.2. Also note that a minor modification of the proof of Theorem 7.3 shows that the solution to the problem

$$\min\{\|\hat{B} - B\|_{M,F} : \hat{B}s = y\}$$

is given by (7.3).

Powell (1970d) used the argument of Lemma 7.2 to obtain formula (7.9) in the case $c = s$. Since in this case the underlying single-rank method is Broyden's, the double-rank formula is often called the *Powell symmetric Broyden update*, or

the PSB update:

$$\bar{B}_{\mathrm{PSB}} = B + \frac{(y - Bs)s^T + s(y - Bs)^T}{\langle s, s\rangle} - \frac{\langle y - Bs, s\rangle ss^T}{\langle s, s\rangle^2}. \tag{7.13}$$

Theorem 7.3 implies that $\bar{B}_{\mathrm{PSB}}$ is the unique solution to the problem

$$\min\{\|\hat{B} - B\|_F : \hat{B} \text{ symmetric}, \hat{B}s = y\}$$

and this property is reminiscent of Theorem 4.1. Because of this property it follows that if A is any symmetric matrix with $y = As$, then

$$\|B - A\|_F^2 = \|\bar{B}_{\mathrm{PSB}} - B\|_F^2 + \|\bar{B}_{\mathrm{PSB}} - A\|_F^2.$$

These considerations lead us to believe that $\bar{B}_{\mathrm{PSB}}$ is a good approximation to the Hessian. To further justify this claim note that (7.13) implies that for any symmetric A and B in $L(R^n)$,

$$\bar{B}_{\mathrm{PSB}} - A = P^T(B - A)P + [(y - As)s^T + s(y - As)^T P]/\langle s, s\rangle$$

with $P = I - ss^T/\langle s, s\rangle$. Therefore (1.2) shows that

$$\|\bar{B}_{\mathrm{PSB}} - A\|_F \leqq \|B - A\|_F + 2\frac{\|y - As\|}{\|s\|}.$$

If $A = \nabla^2 f(x)$ and $\nabla^2 f$ is Lipschitz continuous (with constant κ) in the open convex set D, then Lemma 3.3 implies that

$$\|\bar{B}_{\mathrm{PSB}} - \nabla^2 f(\bar{x})\|_F \leqq \|B - \nabla^2 f(x)\|_F + 3\kappa\|s\|$$

whenever x and $\bar{x}$ lie in D. This relationship shows that the absolute error of B_k as an approximation to $\nabla^2 f(x_k)$ grows linearly with $\|s_k\|$, and that this holds independent of the position of x in D.

7.2. Positive definiteness. We now turn to updates which in addition to satisfying (7.1) and (7.2) generate positive definite matrices. For this, let us investigate the property of hereditary positive definiteness; that is,

$$B \text{ positive definite} \quad \text{implies} \quad \bar{B} \text{ positive definite.} \tag{7.14}$$

Note that if an update satisfies (7.2) and (7.14), then $y = \bar{B}s$ and therefore $\langle y, s\rangle > 0$ whenever B is positive definite. This imposes a restriction on the angle between y and s, which although not severe, must be kept in mind. In fact, if $\langle \nabla f(x), s\rangle < 0$ then $\langle y, s\rangle > 0$ is equivalent to the existence of a $\beta \in (0, 1)$ such that $\langle \nabla f(\bar{x}), s\rangle \geqq \beta\langle \nabla f(x), s\rangle$. For this reason the requirement (6.8) is very natural for quasi-Newton methods.

To investigate the property of hereditary positive definiteness, we need a result from the perturbation theory of symmetric matrices, e.g. Wilkinson (1965, pp. 95–98):

LEMMA 7.4. *Let $A \in L(R^n)$ be symmetric with eigenvalues*

$$\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_n,$$

and let $A^* = A + \sigma uu^T$ *for some* $u \in R^n$. *If* $\sigma \geqq 0$ *then* A^* *has eigenvalues* λ_i^* *such that*

$$\lambda_1 \leqq \lambda_1^* \leqq \lambda_2 \leqq \cdots \leqq \lambda_n \leqq \lambda_n^*,$$

while if $\sigma \leqq 0$ *then the eigenvalues of* A^* *can be arranged so that*

$$\lambda_1^* \leqq \lambda_1 \leqq \lambda_2^* \leqq \cdots \leqq \lambda_n^* \leqq \lambda_n.$$

Lemma 7.4 and the next two results will lead us to a choice of c in (7.9) which naturally satisfies (7.14). This development is a bit long, but it gives a lot of insight.

THEOREM 7.5. *Let* $B \in L(R^n)$ *be symmetric and positive definite, and let* c, s, *and* y *be in* R^n *with* $\langle c, s\rangle \neq 0$. *Then* $\bar{B}$ *as defined by* (7.9) *is positive definite if and only if* $\det \bar{B} > 0$.

Proof. If $\bar{B}$ is positive definite, then clearly $\det \bar{B} > 0$. For the converse first note that we can write

$$\bar{B} = B + vw^T + wv^T,$$

where $w = c$ and

$$v = \frac{y - Bs}{\langle c, s\rangle} - \frac{1}{2}\frac{\langle y - Bs, s\rangle}{\langle c, s\rangle^2} c.$$

Therefore,

$$\bar{B} = B + \tfrac{1}{2}[(v+w)(v+w)^T - (v-w)(v-w)^T],$$

and thus we have written $\bar{B}$ as B plus the sum of two symmetric rank one matrices. If B is positive definite then Lemma 7.4 implies that $\bar{B}$ can have at most one nonpositive eigenvalue. Therefore, if $\det \bar{B} > 0$, then all the eigenvalues must be positive and thus $\bar{B}$ is positive definite.

In view of Theorem 7.5, conditions (7.1) and (7.14) for the updates defined by (7.9) require that if B is symmetric and positive definite then $\det \bar{B} > 0$. To find out what choices of c satisfy this requirement we need an expression for $\det \bar{B}$.

LEMMA 7.6. *Let* $u_i \in R^n$ *for* $i = 1, 2, 3, 4$. *Then*

$$\det (I + u_1 u_2^T + u_3 u_4^T) = (1 + \langle u_1, u_2\rangle)(1 + \langle u_3, u_4\rangle) - \langle u_1, u_4\rangle \langle u_2, u_3\rangle.$$

Proof. A proof of this result can be found in Pearson (1969); the following is an alternative argument.

Assume for the moment that $\langle u_1, u_2\rangle \neq -1$. Then $I + u_1 u_2^T$ is nonsingular and

$$I + u_1 u_2^T + u_3 u_4^T = (I + u_1 u_2^T)(I + (I + u_1 u_2^T)^{-1} u_3 u_4^T).$$

The result now follows by using Lemmas 4.2 and 4.4. Since it holds for $\langle u_1, u_2\rangle \neq -1$, a continuity argument shows that it holds in general.

Now apply Lemma 7.6 to (7.9). After some algebra it follows that

$$\det \bar{B} = \det B[(\langle c, Hy\rangle^2 - \langle c, Hc\rangle \langle y, Hy\rangle + \langle c, Hc\rangle \langle y, s\rangle)/\langle c, s\rangle^2], \tag{7.15}$$

where $H = B^{-1}$. If we assume that B is positive definite and let $v = H^{1/2} y$ and $w = H^{1/2} c$, then

$$\det \bar{B} = \det B[(\langle v, w\rangle^2 - \|v\|^2 \|w\|^2 + \|w\|^2 \langle y, s\rangle)/\langle c, s\rangle^2], \tag{7.16}$$

and Theorem 7.5 implies that $\bar{B}$ is positive definite if and only if

$$\|w\|^2\langle y, s\rangle > \|v\|^2\|w\|^2 - \langle v, w\rangle^2. \tag{7.17}$$

It is now apparent that the most natural way to satisfy (7.17) is to choose w to be a multiple of v so that (7.17) only requires that $\langle y, s\rangle$ be positive. In this case c is a multiple of y and then (7.9) reduces to an update introduced by Davidon (1959), and later clarified and improved by Fletcher and Powell (1963). The DFP *update* is then given by

$$\begin{aligned}\bar{B}_{\mathrm{DFP}} &= B + \frac{(y - Bs)y^T + y(y - Bs)^T}{\langle y, s\rangle} - \frac{\langle y - Bs, s\rangle yy^T}{\langle y, s\rangle^2} \\ &= \left(I - \frac{ys^T}{\langle y, s\rangle}\right)B\left(I - \frac{sy^T}{\langle y, s\rangle}\right) + \frac{yy^T}{\langle y, s\rangle}.\end{aligned} \tag{7.18}$$

Some of its properties are given in the following result, but first we note that the underlying single-rank formula (7.3) where c is a multiple of y is an update due to Pearson (1969).

THEOREM 7.7. *Let $B \in L(R^n)$ be a nonsingular, symmetric matrix and define $\bar{B}_{\mathrm{DFP}} \in L(R^n)$ by (7.18) for any vectors y and s in R^n with $\langle y, s\rangle \neq 0$. Then $\bar{B}_{\mathrm{DFP}}$ is nonsingular if and only if $\langle y, Hy\rangle \neq 0$, where $H = B^{-1}$. If $\bar{B}_{\mathrm{DFP}}$ is nonsingular, then $\bar{H}_{\mathrm{DFP}} = \bar{B}_{\mathrm{DFP}}^{-1}$ can be expressed as*

$$\bar{H}_{\mathrm{DFP}} = H + \frac{ss^T}{\langle s, y\rangle} - \frac{Hyy^TH}{\langle y, Hy\rangle}. \tag{7.19}$$

Furthermore, if B is positive definite, then $\bar{B}_{\mathrm{DFP}}$ is positive definite if and only if $\langle y, s\rangle > 0$.

Proof. Recall that for the DFP update c is a multiple of y so that (7.16) reduces to

$$\det \bar{B}_{\mathrm{DFP}} = \det B\left[\frac{\langle y, Hy\rangle}{\langle y, s\rangle}\right]. \tag{7.20}$$

Thus $\bar{B}_{\mathrm{DFP}}$ is nonsingular if and only if $\langle y, Hy\rangle \neq 0$. To verify that $\bar{H}_{\mathrm{DFP}}$ is given by (7.19) one can either show that $\bar{H}_{\mathrm{DFP}}\bar{B}_{\mathrm{DFP}} = I$ or one can use Lemma 4.2 twice on (7.18). In either case the proof is straightforward but tedious and is therefore omitted. Finally, assume that B is positive definite. If $\langle y,s\rangle$ is positive, then (7.20) shows that $\det \bar{B}_{\mathrm{DFP}}$ is also positive and thus Theorem 7.5 implies that $\bar{B}_{\mathrm{DFP}}$ is positive definite. Conversely, if $\bar{B}_{\mathrm{DFP}}$ is positive definite, then

$$\langle y, s\rangle = \langle \bar{B}_{\mathrm{DFP}}s, s\rangle > 0$$

which is the desired result.

One way to use the DFP update to generate a quasi-Newton direction and only use $O(n^2)$ arithmetic operations per iteration would be to generate $B_k^{-1} = H_k$ via equation (7.19). Another approach is based on the fact that if A is positive definite and $A = LL^T$ for some lower triangular matrix, then the corresponding decomposition of

$$\bar{A} = A + \alpha zz^T$$

can be obtained in $O(n^2)$ operations provided $\bar{A}$ is positive definite. Methods for doing this are surveyed by Gill, Golub, Murray and Saunders (1974). That these techniques apply to (7.18) follows from the proof of Theorem 7.5 which shows that (7.18) can be written as

$$\bar{B}_{\mathrm{DFP}} = B + \tfrac{1}{2} z_1 z_1^T - \tfrac{1}{2} z_2 z_2^T,$$

where z_1 and z_2 are linear combinations of Bs and y. If the DFP update is used in a method of the form (6.4) then an advantage of the latter approach is that (7.18) requires no matrix-vector products.

Finally we remark that the matrices generated by the DFP formula are good approximations to the Hessian. In fact in § 8 (see (8.16)) we shall derive a general result which can be interpreted as follows: If $\|s\|$ is small then the relative error (as measured in § 1) of $\bar{B}_{\mathrm{DFP}}$ as an approximation to a positive definite $\nabla^2 f(x)$ cannot be much larger than the corresponding error in B. Moreover the possible increase in this error is governed by a relative measure of how much f differs from a quadratic on D.

7.3. Inverse updates. So far we have been thinking in terms of obtaining an approximation to the Hessian, but it is perhaps equally reasonable to try to obtain an approximation to the inverse of the Hessian. In particular, it should be clear that it is possible to use the techniques that we have been discussing to develop updating formulas for the inverse. These updates are sometimes called inverse updates while the updates developed in §§ 7.1 and 7.2 could be called direct updates.

To develop inverse updates, assume that we have an approximation H to $\nabla^2 f(x)^{-1}$ and try to obtain a good approximation $\bar{H}$ to $\nabla^2 f(\bar{x})^{-1}$ where $\bar{x} = x + s$. For inverse updates the analogue of the quasi-Newton equation is

$$\bar{H} y = s, \tag{7.21}$$

and therefore, the general single rank formula which satisfies (7.21) is

$$\bar{H} = H + \frac{(s - Hy)d^T}{\langle d, y \rangle} \tag{7.22}$$

for any $d \in R^n$ with $\langle d, y \rangle \neq 0$.

It is important to realize the relationship between (7.3) and (7.22). If Lemma 4.2 is applied to (7.3) we obtain

$$\bar{B}^{-1} = B^{-1} + \frac{(s - B^{-1}y)c^T B^{-1}}{\langle c, B^{-1}y \rangle}.$$

Therefore, (7.3) and (7.22) represent the same update if $c = B^T d$.

Just as in § 7.1, it is possible to study the property of hereditary symmetry, which for inverse updates is

$$H \text{ symmetric} \quad \text{implies} \quad \bar{H} \text{ symmetric}. \tag{7.23}$$

It is easy to verify that the only single rank formula which satisfies the quasi-Newton equation (7.21) and the hereditary symmetric property (7.23) is again given by the symmetric single rank formula (7.5).

To obtain other inverse updates which satisfy (7.21) and (7.23) we carry out the symmetrization argument of Lemma 7.2 on (7.22) to obtain

$$\bar{H} = H + \frac{(s - Hy)d^T + d(s - Hy)^T}{\langle d, y\rangle} - \frac{\langle s - Hy, y\rangle}{\langle d, y\rangle^2} dd^T. \tag{7.24}$$

This result is due to Dennis (1972) who also noted that if $\bar{B}$ and $\bar{H}$ are defined by (7.9) and (7.24), respectively, then in general $\bar{B}\bar{H} \neq I$ even if B is symmetric, $BH = I$ and $c = Bd$. At first this is surprising because under these assumptions (7.3) and (7.22) represent the same update; however, in the argument of Lemma 7.3 we used the symmetrization operation $(A + A^T)/2$, and in general, the symmetrization and inversion operations do not commute.

It is also possible to prove an analogue of Theorem 7.3 for updates (7.24). In particular, if H is symmetric, then the unique solution to the problem

$$\min \{\|\hat{H} - H\|_F : \hat{H} \text{ symmetric}, \hat{H}y = s\}$$

is given by (7.24) with $d = y$. This update was proposed by Greenstadt (1970), but it has not received any more attention in the literature since it does not perform as well as the PSB update (7.13). It is interesting that the underlying single rank method was obtained by Broyden (1965), but that this update has also been neglected because of its poor numerical performance.

The most important instance of (7.24) was given by Broyden (1969), (1970), and independently by Fletcher (1970), Goldfarb (1970) and Shanno (1970). This update can be obtained by asking for the update of the general form (7.24) which "naturally" has the property of hereditary positive definiteness for inverse updates; that is, H positive definite implies $\bar{H}$ positive definite. It should be clear from the development in Section 7.2 that this update corresponds to choosing $d = s$ in (7.24) and therefore the *Broyden–Fletcher–Goldfarb–Shanno* or BFGS update can be written in the form

$$\bar{H}_{\text{BFGS}} = \left(I - \frac{sy^T}{\langle y, s\rangle}\right) H \left(I - \frac{ys^T}{\langle y, s\rangle}\right) + \frac{ss^T}{\langle y, s\rangle} \tag{7.25}$$

At this point we note that the BFGS update is sometimes called the *complementary* DFP *update* and that the underlying single rank method (7.22) in which $d = s$ was proposed by G. McCormick (see Pearson (1969)).

There is growing evidence that the BFGS is the best current update formula for use in unconstrained minimization. For example, see the results of Dixon (1972b). For this reason, and for future reference we state the following analogue of Theorem 7.7.

THEOREM 7.8. *Let $H \in L(R^n)$ be a nonsingular symmetric matrix, and define $\bar{H}_{\text{BFGS}} \in L(R^n)$ by (7.25) for any vectors y and s in R^n with $\langle y, s\rangle \neq 0$. Then $\bar{H}_{\text{BFGS}}$ is nonsingular if and only if $\langle s, Bs\rangle \neq 0$, where $B = H^{-1}$. If $\bar{H}_{\text{BFGS}}$ is nonsingular, then $\bar{B}_{\text{BFGS}} = \bar{H}_{\text{BFGS}}^{-1}$ can be expressed as*

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

Furthermore, if H is positive definite, then $\bar{H}_{\text{BFGS}}$ is positive definite if and only if $\langle y, s\rangle > 0$.

The remark at the end of § 7.2 about the behavior of $\bar{B}_{\text{DFP}}$ as a relative approximation to the Hessian holds for $\bar{H}_{\text{BFGS}}$ as a relative approximation to the inverse Hessian. (See (8.18)). Also note that there is a close relationship between the matrices generated by the DFP and BFGS updates for it is easy to verify that if H is positive definite, then

$$\bar{H}_{\text{BFGS}} = \bar{H}_{\text{DFP}} + vv^T, \tag{7.26}$$

where v is the vector

$$v = \langle y, Hy\rangle^{1/2}\left[\frac{s}{\langle s, y\rangle} - \frac{Hy}{\langle y, Hy\rangle}\right], \tag{7.27}$$

while if B is positive definite, then

$$\bar{B}_{\text{DFP}} = \bar{B}_{\text{BFGS}} + ww^T, \tag{7.28}$$

where w is the vector

$$w = \langle s, Bs\rangle^{1/2}\left[\frac{y}{\langle s, y\rangle} - \frac{Bs}{\langle s, Bs\rangle}\right]. \tag{7.29}$$

By virtue of Lemma 7.4, relations (7.26) and (7.28) imply that the eigenvalues of $\bar{H}_{\text{BFGS}}(\bar{B}_{\text{BFGS}})$ are larger (smaller) than the eigenvalues of $\bar{H}_{\text{DFP}}(\bar{B}_{\text{DFP}})$. However, there does not seem to be any relationship between the condition number of $\bar{H}_{\text{BFGS}}$ and the condition number of $\bar{H}_{\text{DFP}}$.

From a purely algebraic point of view, the developments of §§ 7.1 and 7.2 are identical to those in § 7.3. This follows from the fact that (7.22) and (7.24) can be obtained from (7.3) and (7.9), respectively, by interchanging y and s, replacing B's by H's and c by d. In particular Theorems 7.7 and 7.8 are identical since both of them follow from a more general result which relates $\bar{A}$ and A, where

$$\bar{A} = \left(I - \frac{uv^T}{\langle u, v\rangle}\right)A\left(I - \frac{vu^T}{\langle u, v\rangle}\right) + \frac{uu^T}{\langle u, v\rangle}$$

and $\langle u, v\rangle \neq 0$. In spite of these remarks we have opted for a separate development for expository purposes. Nevertheless, it is useful to note that the DFP and BFGS updates are related by the transformation

$$s \leftrightarrow y, \quad B \leftrightarrow H, \quad \bar{B} \leftrightarrow \bar{H}. \tag{7.30}$$

In fact, Fletcher's (1970) derivation of the BFGS update was through this transformation.

Finally, we note that if a direct and inverse update are related by the transformation (7.30) then these updates are sometimes called "dual" or "complementary" updates, and this is the reason why the BFGS is also called the complementary DFP formula.

8. Convergence results for rank-two quasi-Newton methods. Let $f: R^n \to R$ be continuously differentiable in an open set D and consider a method of the form

$$x_{k+1} = x_k - \lambda_k H_k \nabla f(x_k), \qquad k = 0, 1, \cdots, \tag{8.1}$$

where the matrices H_k are generated by one of the methods of § 7 and λ_k is suitably chosen. In this section we examine some of the convergence and rate of convergence results that are available for (8.1).

In a lot of theoretical work sufficient conditions are assumed so that λ_k can be chosen by an *exact line search*. This usually means that either

$$\lambda_k = \min\{\lambda > 0 : \langle \nabla f(x_k + \lambda p_k), p_k \rangle = 0\}, \tag{8.2}$$

where $p_k = -H_k \nabla f(x_k)$, or that λ_k is the first local minimizer of $f(x_k + \lambda p_k)$ for $\lambda \geqq 0$. Either choice is unrealistic as usually it is not possible to find λ_k to much accuracy in a reasonable amount of time unless, for example, f is a quadratic, positive definite functional. In this case

$$f(x) = \tfrac{1}{2}\langle x, Ax \rangle - \langle x, b \rangle + c \tag{8.3}$$

for some symmetric positive definite $A \in L(R^n)$, and the λ_k which satisfies (8.2) is given by

$$\lambda_k = -\langle Ax_k - b, p_k \rangle / \langle Ap_k, p_k \rangle. \tag{8.4}$$

The earlier convergence results for quasi-Newton methods were given for f defined by (8.3) and λ_k chosen by (8.4). It was shown that if $\{x_k\}$ is generated by (8.1) and H_k correspond to, say the DFP or BFGS updates, then $x_l = A^{-1}b$ for some $0 \leqq l \leqq n$, and if $l = n$ then $H_n = A^{-1}$. This type of finite termination property has sometimes been called *quadratic termination*. The relevance of the quadratic termination property to the general nonlinear problem was originally based on the assumption that if a method terminates in a finite number of steps for a quadratic then this implies superlinear convergence for nonlinear functionals. There has never been any theoretical or numerical support for this belief. (See, however, the discussion following Theorem 8.10). Nevertheless, quadratic termination seems to be a desirable property although as Broyden's method shows, it is not indispensable for superlinear convergence.

In order to describe the quadratic termination properties for symmetric rank two quasi-Newton methods, consider the following class of updates:

$$\bar{H}_\phi = (1 - \phi)\bar{H}_{\mathrm{DFP}} + \phi \bar{H}_{\mathrm{BFGS}}, \tag{8.5}$$

where ϕ is a parameter which may depend on s, y, H and the iteration counter. This class of updates was introduced by Broyden (1967) although not in the form (8.5). It was Fletcher (1970) who showed that Broyden's class, which had been given in terms of a parameter β, could be written in the form (8.5) and that the relationship between ϕ and β is that $\phi = \beta\langle y, s \rangle$. Fletcher also noted that if H is positive definite, then equation (7.26) implies that update (8.5) can be written as

$$\bar{H}_\phi = \bar{H}_{\mathrm{DFP}} + \phi v v^T,$$

where the vector v is defined by (7.27). It is immediate from this expression that if $\phi \geqq 0$, then $\bar{H}_\phi$ shares the property of hereditary positive definiteness with $\bar{H}_{\mathrm{DFP}}$.

Another interesting consequence of this expression is that

$$\bar{B}_\phi = \bar{B}_{\mathrm{BFGS}} + \Phi w w^T = (1-\Phi)\bar{B}_{\mathrm{BFGS}} + \Phi \bar{B}_{\mathrm{DFP}}$$

where w is defined by (7.29) and

$$\Phi = \Phi(\phi) = \frac{(1-\phi)\langle s, y\rangle^2}{\langle s, y\rangle^2 + \phi[\langle y, Hy\rangle\langle s, Bs\rangle - \langle y, s\rangle^2]}.$$

This result can be obtained by using Lemma 4.2 to express $\bar{B}_\phi$ in terms of $\bar{B}_{\mathrm{DFP}}$, $\bar{B}_{\mathrm{DFP}} v$ and ϕ, and then noting that (7.18) and (7.29) relate these quantities to $\bar{B}_{\mathrm{BFGS}}$, w and ϕ.

For future reference and to state the quadratic termination properties of (8.5), note that Broyden's class is generated by

(8.6a)
$$H_{k+1} = H_k + \frac{s_k s_k^T}{\langle s_k, y_k\rangle} - \frac{H_y y_k y_k^T H_k}{\langle y_k, H_k y_k\rangle} + \phi_k v_k v_k^T,$$
$$v_k = \langle y_k, H_k y_k\rangle^{1/2}\left[\frac{s_k}{\langle s_k, y_k\rangle} - \frac{H_k y_k}{\langle y_k, H_k y_k\rangle}\right],$$

and where the vectors s_k and y_k are usually defined by

(8.6b)
$$s_k = x_{k+1} - x_k, \qquad y_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

THEOREM 8.1. *Assume that $f: R^n \to R$ is the positive definite quadratic functional* (8.3) *and that $H_0 \in L(R^n)$ is symmetric and positive definite. For any given $x_0 \in R^n$, let $\{x_k\}$ be generated by* (8.1) *where λ_k, H_k satisfy* (8.4) *and* (8.6), *respectively, and ϕ_k may depend on s_k, y_k and H_k. If $\phi_k \geqq 0$, then there is an integer $0 \leqq l \leqq n$ such that $x_l = A^{-1}b$ and if $l = n$, then $H_n = A^{-1}$.*

A typical proof of Theorem 8.1 proceeds by induction to show that the directions s_k are A-conjugate in the sense that

$$\langle s_i, A s_j\rangle = 0, \qquad i > j,$$

and that also

$$H_i y_j = s_j, \qquad i > j.$$

This was the argument used by Broyden (1967); it shows that x_{k+1} minimizes f in the hyperplane $x_0 + L$ where L is the linear span of $s_0, \cdots, s_k$. Broyden (1971b) and Powell (1973), (1974) have extended and refined Theorem 8.1; in particular, Powell (1974) shows that $A^{1/2} H_k A^{1/2}$ has at least k unit eigenvalues. However, in all these results finite termination depends on choosing λ_k by (8.4). Also note that if

$$\phi_k = \frac{\langle y_k, s_k\rangle}{\langle s_k - H_k y_k, y_k\rangle},$$

then (8.6a) reduces to the symmetric rank one formula (7.6) but Theorems 7.1 and 8.1 are not comparable.

Powell (1971), (1972), (1975) has proved two analogues of Theorem 8.1 for nonlinear functionals. In the following result recall that $L(x_0)$ was defined in Theorem 6.2.

THEOREM 8.2. *Let $f: R^n \to R$ be twice differentiable and convex on R^n and assume that for a given $x_0 \in R^n$ the level set $L(x_0)$ is bounded. Suppose that $\{x_k\}$ is generated by* (8.1) *and that λ_k, H_k are chosen in one of the following two ways*:

(a) *An exact line search and the DFP update.*

(b) *The line search of Theorem* 6.3 *and the BFGS update.*

Then for any symmetric, positive definite $H_0 \in L(R^n)$ and $\varepsilon > 0$ there is an index k such that $\|\nabla f(x_k)\| < \varepsilon$.

Part (a) of Theorem 8.2 was proved by Powell (1971), (1972). Part (b) is of course, much more interesting, and was only proved recently by Powell (1976). Moreover, the proof of this second part is radically different from any that we have seen before and interestingly enough, does not seem to generalize to the DFP method. That part (a) of Theorem 8.2 extends to other methods in the Broyden class (8.5) follows from the following remarkable result of Dixon (1972a).

THEOREM 8.3. *Let $f: R^n \to R$ be differentiable on R^n, and assume that for a given $x_0 \in R^n$ the level set $L(x_0)$ is bounded. Given a symmetric positive definite $H_0 \in L(R^n)$ suppose that $\{x_k\}$ is generated by* (8.1) *where λ_k, H_k are chosen according to* (8.2) *and* (8.6), *respectively. If $\phi_k \geqq 0$, then the sequence $\{x_k\}$ is independent of $\{\phi_k\}$.*

Dixon's result is actually more general than Theorem 8.3 since it allows negative values of ϕ_k. However, the above formulation suffices for our purposes, and moreover, the more general formulation requires additional assumptions on $\{H_k\}$.

All the results presented so far on the convergence of rank-two quasi-Newton methods choose $\phi_k \in [0, \infty)$ since, if the line search guarantees that $\langle s_k, y_k \rangle$ is positive, then each H_k is positive definite. Of course, if f is strictly convex then $\langle s_k, y_k \rangle$ is always positive regardless of the line search. The following result of Fletcher (1970) shows that for quadratic functionals the updates with $\phi_k \in [0, 1]$ have a very desirable property which does not depend on any line searches.

THEOREM 8.4. *Assume that $f: R^n \to R$ is the positive definite quadratic functional* (8.3) *and that $H_0 \in L(R^n)$ is symmetric and positive definite. Let $\{s_k\}$ be any sequence of nonzero vectors, let $\{H_k\}$ be generated by* (8.6a) *with $y_k = As_k$, and let $\lambda_i^{(k)}$, $i = 1, \cdots, n$, be the eigenvalues of $A^{1/2} H_k A^{1/2}$ in increasing order of magnitude. If $\phi_k \in [0, 1]$ then*

$$\min \{\lambda_i^{(k)}, 1\} \leqq \lambda_i^{(k+1)} \leqq \max \{\lambda_i^{(k)}, 1\}.$$

Since Fletcher (1970) showed that Theorem 8.4 fails if $\phi_k \notin [0, 1]$, this result indicates that the most reasonable updates in Broyden's class (8.5) correspond to $\phi \in [0, 1]$. In fact, numerical results of Dixon (1972b), and Gill, Murray and Pitfield (1972) suggest that $\phi = 1$ is to be preferred over $\phi = 0$. Of course, Theorem 8.4 does not say anything about the PSB update since it is not of the form (8.5).

We have surveyed the global convergence results for rank-two quasi-Newton methods; since the analysis of the asymptotic rate of convergence is of major importance, we now investigate this topic as well as the local convergence properties of the BFGS and DFP methods.

For the local convergence of these updates we show how to choose a norm so that (5.2) and (5.7) hold, respectively. First consider the DFP method and recall

that $\bar{B}_{\mathrm{DFP}}$ and B are related by (7.18). It follows that for any symmetric A and B in $L(R^n)$,

$$\bar{B}_{\mathrm{DFP}} - A = P^T(B-A)P + [(y-As)y^T + y(y-As)^T P]/\langle y, s\rangle, \tag{8.7}$$

where

$$P = I - \frac{sy^T}{\langle y, s\rangle}. \tag{8.8}$$

A similar relationship holds between H and $\bar{H}_{\mathrm{BFGS}}$ for the BFGS update. In this case $\bar{H}_{\mathrm{BFGS}}$ and H are related by (7.25) so that if A and H are symmetric, and A is nonsingular, then

$$\bar{H}_{\mathrm{BFGS}} - A^{-1} = P(H - A^{-1})P^T + [(s - A^{-1}y)s^T + s(s-A^{-1}y)^T P^T]/\langle y, s\rangle \tag{8.9}$$

where P is again defined by (8.8). In order to show that (8.7) satisfies (5.2), and (8.9) satisfies (5.7), we need the following result of Broyden (1970).

LEMMA 8.5. *If $Q \in L(R^n)$ is defined by*

$$Q = I - \frac{uv^T}{\langle u, v\rangle} \tag{8.10}$$

with u, v in R^n and $\langle u, v\rangle \neq 0$, then

$$\|Q\|_2 = \frac{\|u\|\,\|v\|}{|\langle u, v\rangle|}.$$

Proof. The most straightforward way to verify this result is to recall that $\|Q\|_2^2$ is the largest eigenvalue of Q^TQ and to calculate the eigenvalues of Q^TQ with Lemma 7.6.

Lemma 8.5 shows that $\|P\|_2$ is the secant of the angle between y and s, and since y and s are not in general parallel, $\|P\|_2$ may be arbitrarily large. Therefore the l_2 norm does not seem to be suitable for estimating (8.7) or (8.9). However, near x^* we do have that $A^{-1/2}y$ and $A^{1/2}s$ are nearly parallel if $A = \nabla^2 f(x^*)$ and this suggests the use of a weighted norm. For the DFP method an appropriate norm is defined by

$$\|E\|_{\mathrm{DFP}} = \|A^{-1/2}EA^{-1/2}\|_F. \tag{8.11}$$

Then Lemma 8.5 and (1.2) imply that

$$\|P^T(B-A)P\|_{\mathrm{DFP}} \leqq \|A^{1/2}PA^{-1/2}\|_2^2 \|B-A\|_{\mathrm{DFP}} \equiv \frac{1}{\omega^2}\|B-A\|_{\mathrm{DFP}}, \tag{8.12}$$

where

$$\omega = \frac{\langle y, s\rangle}{\|A^{-1/2}y\|\,\|A^{1/2}s\|} = \frac{\langle A^{-1/2}y, A^{1/2}s\rangle}{\|A^{-1/2}y\|\,\|A^{1/2}s\|}. \tag{8.13}$$

Similar estimates of the other terms in (8.7) yield

$$\left\| \frac{y(y-As)^T P}{\langle y, s\rangle} \right\|_{\mathrm{DFP}} \leqq \frac{1}{\omega^2} \frac{\|A^{-1/2}y - A^{1/2}s\|}{\|A^{1/2}s\|}, \tag{8.14}$$

$$\left\| \frac{(y-As)y^T}{\langle y, s\rangle} \right\|_{\mathrm{DFP}} \leqq \frac{1}{\omega} \frac{\|A^{-1/2}y - A^{1/2}s\|}{\|A^{1/2}s\|}. \tag{8.15}$$

Now place (8.12), (8.14) and (8.15) together to obtain

$$\|\bar{B}_{\mathrm{DFP}} - A\|_{\mathrm{DFP}} \leqq \frac{1}{\omega^2}\|B - A\|_{\mathrm{DFP}} + \frac{2}{\omega^2} \frac{\|A^{-1/2}y - A^{1/2}s\|}{\|A^{1/2}s\|}. \tag{8.16}$$

An analogous relationship holds for the BFGS update. In this case the appropriate norm is defined by

$$\|E\|_{\mathrm{BFGS}} = \|A^{1/2}EA^{1/2}\|_F, \tag{8.17}$$

and it is not difficult to verify that the analogue of (8.16) is

$$\|\bar{H}_{\mathrm{BFGS}} - A^{-1}\|_{\mathrm{BFGS}} \leqq \frac{1}{\omega^2}\|H - A^{-1}\|_{\mathrm{BFGS}} + \frac{2}{\omega^2} \frac{\|A^{1/2}s - A^{-1/2}y\|}{\|A^{-1/2}y\|}. \tag{8.18}$$

As noted in § 7, an interpretation of (8.18) is that if $A = \nabla^2 f(x)$ is positive definite and $\|s\|$ is small, then the relative error of $\bar{H}_{\mathrm{BFGS}}$ as an approximation to $\nabla^2 f(x)^{-1}$ is not too much larger than the relative error of H as an approximation to $\nabla^2 f(x)^{-1}$. Furthermore, the possible growth in this relative error is determined by how much f differs on the points x and $\bar{x}$ from the quadratic whose Hessian is A. This difference is measured in two ways but both have to do with how well $A^{-1/2}y$ is approximated by $A^{1/2}s$; there is an additive term which is the relative error in this approximation and a multiplicative term which is the square of the secant of the angle between these two vectors. Of course, we easily see that the additive term does not exceed the product of the square root of the condition number of A and the relative error in the approximation of y by As. An analogous discussion holds for (8.16).

Another consequence of (8.16) and (8.18) is the local convergence of the DFP and BFGS methods as given by Broyden, Dennis and Moré (1973).

THEOREM 8.6. *Let $f: R^n \to R$ be twice continuously differentiable in an open convex set D, and assume that $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite for some x^* in D. Suppose, in addition, that*

$$\|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leqq \kappa\|x - x^*\|, \qquad x \in D, \tag{8.19}$$

and consider the DFP and BFGS methods as defined by (8.1) *with $\lambda_k \equiv 1$. Then the DFP and BFGS methods are locally and linearly convergent at x^*.*

Proof. To prove that these methods are locally and linearly convergent at x^*, we only need to show that (5.2) and (5.7) are satisfied when D_M is the set of all symmetric matrices and $F = \nabla f$. For the DFP method, first note that (8.13) with $A = \nabla^2 f(x^*)$ and Lemmas 3.2 and 3.3 imply that

$$1 - \omega^2 \leqq \left[\mu \frac{\|y - \nabla^2 f(x^*)s\|}{\|s\|}\right]^2 \leqq [\mu\kappa\sigma(x, \bar{x})]^2,$$

where $\sigma(x, \bar{x})$ is defined by (4.3) and $\mu = \|\nabla^2 f(x^*)^{-1}\|$. Thus if x and $\bar{x}$ lie in a neighborhood N_1 of x^* such that $\sigma(x, \bar{x}) \leqq (2\mu\kappa)^{-1}$, then $\omega^2 \geqq \frac{1}{2}$. In particular,

$$\frac{1}{\omega^2} = 1 + \frac{1-\omega^2}{\omega^2} \leqq 1 + \kappa\mu\sigma(x, \bar{x}). \tag{8.20}$$

Therefore (8.16) with $A = \nabla^2 f(x^*)$ and Lemma 3.2 imply that (5.2) holds with $\alpha_1 = \kappa\mu$ and $\alpha_2 = 4\kappa\mu$ and where $\|\cdot\|$ is the matrix norm defined by (8.11). This proves the local convergence of the DFP method. For the BFGS first let ε be a positive lower bound for the eigenvalues of $\nabla^2 f(x)$ in a neighborhood N_2 of x^* so that $\|y\| \geqq \varepsilon\|s\|$ provided x and $\bar{x}$ lie in N_2. If $N_2 \subset N_1$, then (8.18) with $A = \nabla^2 f(x^*)$, (8.20) and Lemma 3.3 imply that (5.7) holds with $\alpha_1 = \kappa\mu$ and $\alpha_2 = 4(\mu\rho)^{1/2}\kappa/\varepsilon$. Here $\rho = \|\nabla^2 f(x^*)\|$ and $\|\cdot\|$ is the norm defined by (8.17).

As noted at the end of § 5, Theorem 8.6 still holds if instead of choosing $\lambda_k = 1$, it is chosen by a line search which guarantees that $|\lambda_k - 1| < \hat{\lambda}$ for some $\hat{\lambda} \in (0, 1)$. However, Stoer (1975) has recently been able to prove a much more powerful result.

THEOREM 8.7. *Let $f: R^n \to R$ satisfy the assumptions of Theorem* 8.6 *and consider iteration* (8.1) *with H_k generated by* (8.6) *with $\phi_k \in [0, 1]$ and*

$$\lambda_k = \min\{\lambda > 0 : \langle \nabla f(x_k + \lambda p_k), p_k\rangle = \mu_k \langle \nabla f(x_k), p_k\rangle\},$$

where $0 \leqq \mu_k \leqq \bar{\mu} < 1$ and $p_k = -H_k \nabla f(x_k)$. Then for any $\tau > 0$ there is an $\varepsilon > 0$ such that if the condition number of H_0 does not exceed τ and $\|x_0 - x^*\| < \varepsilon$, then $\{x_k\}$ is well-defined and converges linearly to x^*.

The line search of Theorem 8.7 is not very practical, but the proof of Theorem 8.7 clearly shows that the purpose of this particular search is to enforce a sufficient decrease on f and to guarantee that the iterates remain in a compact convex set $D_0 \subset D$ with x^* in its interior and on which f is uniformly convex. Thus, if f is uniformly convex on R^n then any reasonable line search could be used in Theorem 8.7.

Theorem 8.7 represents a type of result that is somewhat different from a local result like Theorem 8.6 and a global result like Theorem 8.2. Even if we restrict ourselves to the DFP and BFGS updates, Theorems 8.6 and 8.7 are not comparable because $\lambda_k = 1$ will not be an admissible value for the line search of Theorem 8.7 unless it can be established, a priori, that p_k is sufficiently close to the Newton step in both length and direction. On the other hand, Theorems 8.2 and 8.7 are comparable for the DFP and BFGS updates if we assume that f satisfies the hypothesis of Theorem 8.6 with $D = R^n$ and in addition, assume that f is convex on R^n. In this case, given any symmetric positive definite H_0, Theorem 8.2 implies that the DFP and BFGS methods converge to x^* for any x_0 while Theorem 8.7 imposes some restrictions on x_0.

To prove that the DFP and BFGS methods are superlinearly convergent requires careful estimation of the terms $\|P^T(B-A)P\|_{\text{DFP}}$ and $\|P(H-A^{-1})P^T\|_{\text{BFGS}}$ respectively. These estimates were obtained by Broyden, Dennis and Moré (1973) and then used by Dennis and Moré (1974) to prove superlinear convergence for various choices of λ_k in (8.1).

THEOREM 8.8. *Let $f: R^n \to R$ satisfy the assumptions of Theorem* 8.6, *and*

suppose that $\{x_k\}$ *is a sequence in* D *such that*

$$\sum_{k=0}^{\infty} \|x_k - x^*\| < +\infty. \tag{8.21}$$

If the sequence $\{H_k\}$ *is defined by* (8.6) *with either* $\phi_k \equiv 0$ *or* $\phi_k \equiv 1$ *and* $\langle y_k, s_k \rangle$ *is positive for* $k \geqq 0$, *then for any symmetric positive definite* H_0 *in* $L(R^n)$ *the matrices* H_k *are well-defined and positive definite with uniformly bounded condition numbers. Moreover, if* $B_k = H_k^{-1}$, *then*

$$\lim_{k\to\infty} \frac{\|[B_k - \nabla^2 f(x^*)]s_k\|}{\|s_k\|} = 0. \tag{8.22}$$

Since $\langle y_k, s_k \rangle$ is positive for $k \geqq 0$, Theorems 7.7 and 7.8 imply that in either case H_k is well-defined and positive definite. The remainder of the proof is somewhat long, so we only outline it. First it is shown that $\{\|B_k\|\}$ is bounded and (8.22) holds for the DFP method. A similar argument for BFGS shows that $\{\|H_k\|\}$ is bounded and instead of (8.22),

$$\lim_{k\to\infty} \frac{\|[H_k - \nabla^2 f(x^*)^{-1}]y_k\|}{\|y_k\|} = 0. \tag{8.23}$$

However, *if* $\{\|B_k\|\}$ is bounded, then

$$[B_k - \nabla^2 f(x^*)]s_k = [I - B_k \nabla^2 f(x^*)^{-1}](y_k - \nabla^2 f(x^*)s_k) - B_k[H_k - \nabla^2 f(x^*)^{-1}]y_k$$

shows that (8.23) implies (8.22). Hence, the final step in the proof consists of using the techniques of Powell (1971, pp. 31–32) to prove that $\{\|B_k\|\}$ and $\{\|H_k\|\}$ are bounded for the BFGS and DFP methods respectively.

Dennis and Moré (1974) elaborate on Theorem 8.8 and give examples due to Powell which show that (8.22) does not necessarily imply that $\{B_k\}$ converges to $\nabla^2 f(x^*)$. Also note that Theorem 8.8 implies that the DFP and BFGS methods of Theorem 8.6 are superlinearly convergent. This is also a consequence of the following more general result.

THEOREM 8.9. *Let* $f: R^n \to R$ *satisfy the assumptions of Theorem* 8.6 *and consider the DFP and BFGS methods as defined by* (8.1) *with* λ_k *determined by any strategy such that* (8.22) *implies that* $\{\lambda_k\}$ *converges to unity. If the sequence* $\{x_k\}$ *generated by the DFP and BFGS method satisfies* (8.21), *then* $\{x_k\}$ *converges superlinearly to* x^*.

Proof. Theorem 3.1 implies that $\{x_k\}$ converges superlinearly to x^* if

$$\lim_{k\to\infty} \frac{\|[\lambda_k^{-1} B_k - \nabla^2 f(x^*)]s_k\|}{\|s_k\|} = 0. \tag{8.24}$$

On the other hand, Theorem 8.8 and our assumptions show that (8.22) holds and hence $\{\lambda_k\}$ converges to unity. Thus (8.24) also holds and hence $\{x_k\}$ converges superlinearly to x^*.

There are a number of line searches which satisfy the requirements of

Theorem 8.9:

(a) Exact line searches. This result is due to Dennis and Moré (1974).
(b) The line search of Theorem 6.3. This is the content of Theorem 6.4.
(c) The line search of Theorem 8.7 provided $\{\mu_k\}$ converges to zero. This is not difficult to prove and in fact, the proof is very similar to that of Theorem 6.4.

The above remarks and Theorems 8.8 and 8.9 show that the DFP and BFGS methods of Theorems 8.2, 8.7 and 8.8 are superlinearly convergent. It is an open problem to prove, under the conditions of Theorem 8.6, that iteration (8.1) is superlinearly convergent when $\{H_k\}$ is defined by (8.6) for $\{\phi_k\} \subset [0, 1]$ and λ_k chosen in some reasonable (certainly not by exact line searches) manner.

It would be interesting to prove Theorem 8.8 assuming only that $\{x_k\}$ converges to x^* instead of (8.21). However, as it stands Theorem 8.9 shows that either the DFP and BFGS methods converge superlinearly or they converge sublinearly in the sense that

$$\limsup_{k \to +\infty} \|x_k - x^*\|^{1/k} = 1.$$

In practice, sublinear convergence is essentially equivalent to nonconvergence, so Theorem 8.9 covers all the computationally interesting cases.

If the line search is appropriately restricted then more is known about the rate of convergence.

THEOREM 8.10. *Let $f: R^n \to R$ and x_0 satisfy the assumptions of Theorem* 8.2 *and let $\{x_k\}$ be generated by the DFP method with perfect line searches. If $\{x_k\}$ converges to a point x^* at which $\nabla^2 f(x^*)$ is positive definite and* (8.19) *holds, then $\nabla f(x^*) = 0$ and $\{x_k\}$ converges superlinearly to x^*. In addition, there is an $\eta > 0$ such that*

$$\|x_{k+n} - x^*\| \leqq \eta \|x_k - x^*\|^2, \qquad k \geqq 0. \tag{8.25}$$

That $\{x_k\}$ converges superlinearly to x^* is due to Powell (1971), but (8.25)—which is known as n-step quadratic convergence— is due to Burmeister (1973) and later but independently, to Schuller and Stoer (1974). The proofs of these two results are completely different; the proof of (8.25) depends on the finite termination property of DFP while, as pointed out by Dennis and Moré (1974), Powell's result can be proved by showing that (8.21) holds and then applying Theorem 8.9. Note that Theorem 8.10 extends to other methods in Broyden's class with $\phi_k \geqq 0$ via Dixon's result (Theorem 8.3). If $\phi_k \in [0, 1]$ it is possible to extend Theorem 8.10 for nonexact line searches: Stoer (1975) has shown that if in Theorem 8.10 we have

$$0 \leqq \mu_k \leqq \min \{\bar{\mu}, \gamma \|\nabla f(x_k)\|\}$$

for some constant γ, then (8.25) holds.

It should be appreciated that a sequence may converge n-step quadratically but not be superlinearly convergent and conversely. However, n-step quadratic

convergence does imply that

$$\lim_{k\to\infty} \|x_k - x^*\|^{1/k} = 0,$$

and thus $\{x_k\}$ is R-superlinearly convergent in the terminology of Ortega and Rheinboldt (1970); Q-superlinear convergence corresponds to the notion used in this paper.

THEOREM 8.11. *Let $f: R^n \to R$ be the strictly convex quadratic functional (8.3). Then for $\lambda_k \equiv 1$ the DFP and BFGS methods converge globally and superlinearly to $A^{-1}b$.*

Proof. It is clear that the iterations are well defined. To prove the result for the DFP method note that since $y = As$, equation (8.7) implies that

$$\|\bar{B}_{\mathrm{DFP}} - A\|_{\mathrm{DFP}} = \|Q^T[A^{-1/2}(B-A)A^{-1/2}]Q\|_F, \tag{8.26}$$

where $z = A^{1/2}s$ and

$$Q = I - \frac{zz^T}{\langle z, z\rangle}.$$

However, in the proof of Theorem 5.2 we showed that for any $E \in L(R^n)$ and $z \in R^n$,

$$\|EQ\|_F \leqq \|E\|_F - (2\|E\|_F)^{-1}\left(\frac{\|Ez\|}{\|z\|}\right)^2.$$

Thus, if we let $\eta_k = \|B_k - A\|_{\mathrm{DFP}}$, and use (1.2) and the above estimate in (8.26), then

$$\eta_{k+1} \leqq [1 - (2\eta_k^2)^{-1}\psi_k^2]\eta_k,$$

where

$$\psi_k = \frac{\|A^{-1/2}(B_k - A)s_k\|}{\|A^{1/2}s_k\|}.$$

It is now clear that $\{\eta_k\}$ is monotone decreasing and hence convergent. If η is an upper bound for $\{\eta_k\}$, then

$$(2\eta)^{-1}\psi_k^2 \leqq \eta_k - \eta_{k+1},$$

and since $\{\eta_k\}$ is converging it follows that $\{\psi_k\}$ tends to zero. Consequently,

$$\lim_{k\to\infty} \frac{\|(B_k - A)s_k\|}{\|s_k\|} = 0. \tag{8.27}$$

Moreover,

$$A(x_{k+1} - x^*) = \nabla f(x_{k+1}) = \nabla f(x_k) + As_k = (A - B_k)s_k$$

implies that

$$\frac{\|x_{k+1} - x^*\|}{\|s_k\|} \leqq \|A^{-1}\| \frac{\|(B_k - A)s_k\|}{\|s_k\|},$$

and thus (8.27) shows that $\{x_k\}$ converges superlinearly to x^*.

For the BFGS method similar calculations with (8.9) yield

$$\lim_{k\to\infty} \frac{\|[H_k - A^{-1}]y_k\|}{\|y_k\|} = 0. \tag{8.28}$$

Moreover, Theorem 8.4 shows that $\{\|B_k\|\}$ is bounded. Thus, as noted after Theorem 8.8, (8.28) implies that (8.27) holds and now the proof is completed as before.

Theorem 8.11 seems to be just a curiosity since if λ_k is chosen by an exact line search, then convergence will take place in at most n steps. However, it does give an indication of the stability of the DFP and BFGS updates without exact line searches.

We have now finished our study of the asymptotic behavior of the DFP and BFGS methods. It is also possible to study the PSB update, but since it does not generate positive definite matrices, results like Theorem 8.9 have to be modified for the PSB update.

The PSB update is not generally used in a descent implementation, but Powell (1970c), 1970d) has described and analyzed a quite competitive algorithm which uses the PSB algorithm in a hybrid implementation, and has shown that if certain "special iterations" are taken then the algorithm converges globally and superlinearly. These special iterations guarantee that the directions used by the PSB update are uniformly linearly independent and therefore, that the sequence $\{B_k\}$ generated by the PSB update converges to the Jacobian evaluated at the solution—for a discussion of the concept of uniform linear independence and its relationship to the Broyden and PSB update see Moré and Trangenstein (1976). Powell (1975) later proved that in theory the algorithm converged globally and superlinearly even if these special iterations are not used. In practice however they cannot be taken away from the algorithm, without a significant loss in efficiency.

The above results of Powell deserve further investigation. In fact, the whole question of how to globalize an algorithm is very important and represents an open field of research.

In its simplest form the PSB method is given by

$$x_{k+1} = x_k - B_k^{-1}\nabla f(x_k), \tag{8.29}$$

where $\{B_k\}$ is generated by

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T + s_k(y_k - B_k s_k)^T}{\langle s_k, s_k\rangle} - \frac{\langle y_k - B_k s_k, s_k\rangle}{\langle s_k, s_k\rangle^2} s_k s_k^T, \tag{8.30}$$

and as usual

$$y_k = \nabla f(x_{k+1}) - \nabla f(x_k), \qquad s_k = x_{k+1} - x_k. \tag{8.31}$$

Note that since the matrices $\{B_k\}$ are not necessarily positive definite it is not possible to carry out the above iteration by updating an LDL^T decomposition of B_k. To avoid $O(n^3)$ operations per step it is usual to generate $H_k = B_k^{-1}$. An alternative would be to update a factorization of the form QTQ^T, where Q is orthogonal and T is tridiagonal, but this approach has not been investigated.

The following result of Broyden, Dennis and Moré (1973) covers the above iteration.

THEOREM 8.12. *Let $f: R^n \to R$ satisfy the assumptions of Theorem* 8.6 *except that now $\nabla^2 f(x^*)$ is not required to be positive definite, and consider the PSB method as defined by* (8.29), (8.30) *and* (8.31). *Then the PSB method is locally and superlinearly convergent.*

Since the proof of this result is so similar to that of Theorem 5.2 we omit it. It is also worthwhile noting that the remarks made about Broyden's method after Theorem 4.2 apply, with obvious modifications, to the PSB method, and that Dennis (1971), (1972) has given Kantorovich theorems for the Broyden and PSB methods.

To conclude this section we point out that the PSB method, if properly modified, is globally and superlinearly convergent for the quadratic functional (8.3) if A is any nonsingular, symmetric matrix. It is only necessary to modify B_{k+1} so that it is nonsingular. For example, if instead of (8.10) we define

$$B_{k+1} = B_k + \theta_k \frac{(y_k - B_k s_k)s_k^T + s_k(y_k - B_k s_k)^T}{\langle s_k, s_k\rangle} - \theta_k^2 \frac{\langle y_k - B_k s_k, s_k\rangle}{\langle s_k, s_k\rangle^2} s_k s_k^T \tag{8.32}$$

and B_k is nonsingular, then it is possible to choose θ_k so that

$$B_{k+1} \text{ is nonsingular,} \quad |\theta_k - 1| \leqq \hat{\theta} \quad \text{for some } \hat{\theta} \in (0, 1). \tag{8.33}$$

Moré and Trangenstein (1976) elaborate on how this can be done, and also prove the following result.

THEOREM 8.13. *Let $f: R^n \to R$ be given by* (8.3) *where $A \in L(R^n)$ is any nonsingular, symmetric matrix and consider the PSB method* (8.29) *where $\{B_k\}$ is generated by* (8.32), (8.31) *and θ_k satisfies* (8.33). *Then the PSB method is globally and superlinearly convergent to $A^{-1}b$.*

Moré and Trangenstein (1976) also point out that Theorem 8.12 holds if the PSB method (8.29) is defined by (8.32), (8.31) and θ_k satisfies (8.33).

9. Concluding remarks. We have tried to write this survey in such a way that the important problems suggest themselves, so instead of ending with remarks about directions for future research, we end with an admission of certain omissions.

Although we have indicated several approaches to the computation of the updates, all these approaches are based on an additive correction of rank at most two. Other approaches are possible; Brodlie, Gourlay and Greenstadt (1973) discuss multiplicative corrections so that their direct updates are of the form

$$\bar{B} = (I + uv^T)B(I + vu^T),$$

and show that the DFP and BFGS can be written in this factored form.

We have not mentioned any particular implementation because there are a number of very promising algorithms (e.g. Davidon (1975)) now being tested and such remarks would likely be out of date before their publication. See, however, the paper of Fletcher (1972), which discusses several of the currently available algorithms.

We have also not mentioned several excellent papers which only deal with quasi-Newton methods as applied to strictly convex quadratic functionals. In particular, the paper of Huang (1970) introduces a class of updates which has many of the properties of the Broyden class. We have restricted our attention to the Broyden class since it is that subclass of the Huang class which satisfies the quasi-Newton equation and has the hereditary symmetry property.

Finally, we point out that this survey was not meant to be exhaustive even in matters related to quasi-Newton methods; in particular, no mention has been made of the use of quasi-Newton methods in connection with constrained optimization. Readers interested in this connection should consult the book edited by Gill and Murray (1974).

REFERENCES

L. ARMIJO (1966) *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16, pp. 1–3.

R. P. BRENT (1973), *Some efficient algorithms for solving systems of nonlinear equations*, SIAM J. Numer. Anal., 10, pp. 327–344.

K. W. BRODLIE, A. R. GOURLAY AND J. GREENSTADT (1973), *Rank-one and rank-two corrections to positive definite matrices expressed in product form*, J. Inst. Math. Appl., 11, pp. 73–82.

C. G. BROYDEN (1965), *A class of methods for solving nonlinear simultaneous equations*, Math. Comp., 19, pp. 577–593.

——— (1967), *Quasi-Newton methods and their application to function minimization*, Ibid., 21, 368–381.

——— (1969), *A new double-rank minimization algorithm*, Notices Amer. Math. Soc., 16, p. 670.

——— (1970), *The convergence of single-rank quasi-Newton methods*, Math. Comp., 24, pp. 365–382.

——— (1971a), *The convergence of an algorithm for solving sparse nonlinear systems*, Ibid., 25, pp. 285–294.

——— (1971b), *The convergence of a class of double-rank minimization algorithms, Parts I and II*, J. Inst. Math. Appl., 7, pp. 76–90, 222–236.

G. G. BROYDEN, J. E. DENNIS AND J. J. MORÉ (1973), *On the local and superlinear convergence of quasi-Newton methods*, Ibid., 12, pp. 223–246.

W. BURMEISTER (1973), *Die konvergenzordnung des Fletcher–Powell Algorithmus*, Z. Angew. Math. Mech., 53, pp. 693–699.

J. W. DANIEL (1971), *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, N.J.

W. C. DAVIDON (1959), *Variable metric method for minimization*, Rep. ANL-5990 Rev, Argonne National Laboratories, Argonne, Ill.

——— (1975), *Optimally conditioned optimization algorithms without line searches*, Math. Programming, 9, pp. 1–30.

J. E. DENNIS (1971), *On the convergence of Broyden's method for nonlinear systems of equations*, Math. Comp., 25, pp. 559–567.

——— (1972), *On some methods based on Broyden's secant approximation to the Hessian*, Numerical Methods for Nonlinear Optimization, F. A. Lootsma, ed., Academic Press, London.

J. E. DENNIS AND J. J. MORÉ (1974), *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28, pp. 549–560.

L. C. W. DIXON (1972a), *Variable metric algorithms: Necessary and sufficient conditions for identical behavior on nonquadratic functions*, J. Optimization Theory Appl., 10, pp. 34–40.

——— (1972b), *The choice of step length, a crucial factor in the performance of variable metric algorithms*, Numerical Methods for Nonlinear Optimization, F. A. Lootsma, ed., Academic Press, London.

A. V. FIACCO AND G. P. MCCORMICK (1968), *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York.

R. FLETCHER AND M. J. D. POWELL (1963), *A rapidly convergent descent method for minimization*, Comput. J., 6, pp. 163–168.

R. FLETCHER (1970), *A new approach to variable metric algorithms*, Ibid., 13, pp. 317–322.

——— (1972), *A survey of algorithms for unconstrained optimization*, Numerical Methods for Unconstrained Optimization, W. Murray, ed. Academic Press, London.

P. E. GILL, G. GOLUB, W. MURRAY AND M. A. SAUNDERS (1974), *Methods for modifying matrix factorizations*, Math. Comp., 28, pp. 505–536.

P. E. GILL AND W. MURRAY (1972), *Quasi-Newton methods for unconstrained minimization*, J. Inst. Math. Appl., 9, pp. 91–108.

——— (1974), *Newton-type methods for unconstrained and linearly constrained optimization*, Math. Programming, 7, pp. 311–350.

———, eds., (1974), *Numerical Methods for Constrained Optimization*, Academic Press, London.

P. E. GILL, W. MURRAY AND R. A. PITFIELD (1972), *The implementation of two revised quasi-Newton algorithms for unconstrained optimization*, Rep. NAC 11, National Physical Laboratory, Teddington, England.

D. GOLDFARB (1969), *Sufficient conditions for the convergence of a variable metric method*, Optimization, R. Fletcher, ed., Academic Press, London.

——— (1970), *A family of variable-metric methods derived by variational means*, Math. Comp., 24, pp. 23–26.

S. M. GOLDFELDT, R. E. QUANDT AND H. F. TROTTER (1966), *Maximization by quadratic hill-climbing*, Econometrics, 34, pp. 541–551.

A. A. GOLDSTEIN (1965), *On steepest descent*, SIAM J. Control, 3, pp. 147–151.

J. GREENSTADT (1970), *Variations on variable-metric methods*, Math. Comp., 24, pp. 1–18.

H. Y. HUANG (1970), *Unified approach to quadratically convergent algorithms for function minimization*, J. Optimization Theory Appl., 5, pp. 405–423.

S. L. S. JACOBY, J. S. KOWALIK AND J. T. PIZZO (1972), *Iterative Methods for Nonlinear Optimization Problems*, Prentice-Hall, Englewood Cliffs, N.J.

J. J. MORÉ AND J. A. TRANGENSTEIN (1976), *On the global convergence of Broyden's method*, Math. Comp., 30, pp. 523–540.

W. MURRAY (1972), *Second derivative methods*, Numerical Methods for Unconstrained Optimization, W. Murray, ed., Academic Press, London.

J. M. ORTEGA AND W. C. RHEINBOLDT (1970), *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York.

J. D. PEARSON (1969), *Variable metric methods of minimization*, Comput. J., 12, pp. 171–178.

M. J. D. POWELL (1970a), *A hybrid method for nonlinear equations*, Numerical Methods for Non-linear Algebraic Equations, P. Rabinowitz, ed., Gordon and Breach, London.

——— (1970b), *A FORTRAN subroutine for solving systems of nonlinear algebraic equations*, Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz, ed., Gordon and Breach, London.

——— (1970c), *A FORTRAN subroutine for unconstrained minimization, requiring first derivatives of the objective function*, Rep. 6469, Atomic Energy Research Establishment, Harwell, England.

——— (1970d), *A new algorithm for unconstrained optimization*, Nonlinear Programming, J. B. Rosen, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York.

——— (1971), *On the convergence of the variable metric algorithm*, J. Inst. Math. Appl., 7, pp. 21–36.

——— (1972), *Some properties of the variable metric method*, Numerical Methods for Non-linear Optimization, F. A. Lootsma, ed., Academic Press, London.

——— (1973), *Quadratic termination properties of minimization algorithms, Parts I and II*, J. Inst. Math. Appl., 10, pp. 333–342, 343–357.

——— (1974), *Unconstrained minimization and extensions for constraints*, Mathematical Programming in Theory and Practice, P. L. Hammer and G. Zoutendijk, eds., North-Holland, Amsterdam.

——— (1975), *Convergence properties of a class of minimization algorithms*, Nonlinear Programming 2, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York.

——— (1976), *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, Nonlinear Programming, SIAM-AMS Proceedings, Vol. 9, American Mathematical Society, Providence, R.I.

L. K. SCHUBERT (1970), *Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian*, Math. Comp., 24, pp. 27–30.

G. SCHULLER AND J. STOER (1974), *Uber die konvergenzordnung gewisser rang-2 verfahren zur minimierung von funktionen*, International Series of Numerical Mathematics, 23, pp. 125–147.

D. F. SHANNO (1970), *Conditioning of quasi-Newton methods for function minimization*, Math. Comp., 24, pp. 647–656.

J. SHERMAN AND W. J. MORRISON (1949), *Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix*, Ann. Math. Statist., 20, p. 621.

J. STOER (1975), *On the convergence rate of imperfect minimization algorithms in Broyden's β-class*, Math. Programming, 9, pp. 313–335.

J. H. WILKINSON (1965), *The Algebraic Eigenvalue Problem*, Oxford University Press, London.

P. WOLFE (1969), *Convergence conditions for ascent methods*, this Journal, 11, pp. 226–235.

——— (1971), *Convergence conditions for ascent methods II: Some corrections*, this Journal, 13, pp. 185–188.