DEPARTMENT OF MATHEMATICS
UNIVERSITY OF HOUSTON HOUSTON, TEXAS

NASA CR 151637

QUASI-NEWTON METHODS

BY
HOMER F. WALKER

REPORT #67 FEBRUARY 1972

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-15[illegible]

HOUSTON, TEXAS 77004

QUASI-NEWTON METHODS

by

Homer F. Walker
Department of Mathematics
University of Houston
Houston, Texas

Report #67
NASA Contract NAS-9-15000

QUASI-NEWTON METHODS

by

Homer F. Walker
Department of Mathematics
University of Houston
Houston, Texas

1. Introduction

Systems of nonlinear equations can seldom be solved exactly. Usually, one must obtain approximations to the solutions of such systems by iteration. Quasi-Newton methods (also known as variable metric, variance, secant, update, or modification methods) constitute a class of iterative procedures which may be regarded as generalizations of the secant method for solving a single equation in one unknown. Indeed, not only is the quasi-Newton equation (the equation characteristically satisfied by the iterates produced by these methods) a direct extension of the equation which defines the iterates of the secant method, but also these procedures share many of the computational advantages of the secant method over Newton's method.

Quasi-Newton methods were first introduced in the papers of Davidon [2], Fletcher and Powell [4], and Broyden [1]. In spite of their recent origins, these methods have proved themselves in dealing with practical problems and have become the subject of a large amount of research. The paper of Dennis and Moré [3] provides both an excellent in-depth survey and an elegant unified development of quasi-Newton methods and their theory as understood in the mid-1970's. The main body of this note is a rearrangement and condensation of

material in [3].

In the following, we first formulate precisely the problem to be solved and motivate the introduction of quasi-Newton methods by considering the classical Newton and secant methods and their properties. We then survey three highly successful quasi-Newton methods: Broyden's method for the solution of general nonlinear equations, and the Davidon-Fletcher-Powell and Broyden-Fletcher-Goldfarb-Shanno procedures for unconstrained minimization. (The last two methods will henceforth be referred to as the DFP and BFGS methods, respectively.) Finally, we compare the properties of these methods to those of Newton's method and UHMLE in potential applications to maximum-likelihood estimation of parameters in mixture distributions.

2. The problem

We consider the problem of solving $F(x) = 0$ in an open convex subset D of R^n under the following assumptions on the mapping $F:D \to R^n$:

(a) F is continuously differentiable on D.

(b) There is an x^* in D such that $F(x^*) = 0$ and $F'(x^*)$ is nonsingular.

Newton's method for iteratively approximating the solution x^* begins with an initial approximation x_0 to x^* and attempts to obtain improved approximations by the iteration

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

The convergence properties of Newton's method which are important here are summarized in the following theorem.

Theorem: Whenever x_0 is sufficiently near x^*, there is a sequence $\{\alpha_k\}_{k=0,1,\ldots}$ of non-negative numbers which converges to zero and for which

$$|x_{k+1} - x^*| \leq \alpha_k |x_k - x^*| \qquad k = 0,1, \ldots \quad . \tag{1}$$

If, in addition to satisfying assumptions (a) and (b) above, F has a derivative which is Lipschitz continuous at x^*, i.e., there exists a κ for which $|F'(x) - F'(x^*)| \leq \kappa |x - x^*|$ for all x sufficiently near x^*, then there exists a constant β such that

$$|x_{k+1} - x^*| \leq \beta |x_k - x^*|^2 \qquad k = 0,1, \ldots \tag{2}$$

whenever x_0 is sufficiently near x^*.

A sequence which satisfies an inequality of the form (1) with a sequence $\{\alpha_k\}_{k=0,1,\ldots}$ which converges to zero is said to converge superlinearly. If a sequence satisfies an inequality of the form (2), then it is said to converge quadratically. Superlinear convergence is fast; quadratic convergence is very fast. Since Lipschitz continuity is a very weak assumption, one might say that the theorem asserts that the convergence exhibited by the Newton iterates is always fast and almost always very fast.

The rapid convergence of the Newton iterates is the major advantage of Newton's method. Another advantage is that Newton's method is "self-corrective" in the sense that x_{k+1} depends only on F and x_k so that bad effects of previous iterations are not carried along. (Quasi-Newton methods are not self-corrective in this sense.) Balanced against these advantages is the fact that Newton's method often requires a great deal of computation at each iteration. Indeed, the determination of each iterate requires $O(n^2)$ function evaluations

and $O(n^3)$ arithmetic operations. Thus one is led to ask whether there are methods which retain fast convergence while requiring fewer function evaluations and arithmetic operations at each iteration.

With this question in mind, consider the secant method in the case $n = 1$. This method begins with an initial approximation x_0 to x^* and defines successive approximations by the iteration

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

One may regard the secant method as being obtained from Newton's method by replacing the derivative $F'(x_k)$ by a finite-difference approximation. A particular consequence is that the number of function evaluations per iteration is reduced from two for Newton's method to one for the secant method while the number of arithmetic operations per iteration is not significantly increased. It can be proved that, for x_0 sufficiently near x^*, the iterates produced by the secant method exhibit superlinear convergence rather than quadratic convergence as in the case of the Newton iterates. Nevertheless, superlinear convergence is still fast, and experience has shown that, as a general-purpose algorithm, the secant method is more efficient in total computation time than Newton's method. This suggests that generalizations of the secant method to higher dimensions might be similarly successful.

3. Quasi-Newton methods

Quasi-Newton methods are generalizations of the secant method which are applicable to problems of the type at hand involving an arbitrary number of independent variables. The key properties of these methods are that the

iterates exhibit superlinear local convergence and that each iteration requires n function evaluations and $O(n^2)$ arithmetic operations. In spite of the fact that quasi-Newton methods do not have the quadratic convergence property of Newton's method, the comparatively small number of function evaluations and arithmetic operations make them preferable to Newton's method in many applications.

Quasi-Newton methods have the general form

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

where B_k satisfies the quasi-Newton equation

$$B_k(x_k - x_{k-1}) = F(x_k) - F(x_{k-1}) . \tag{3}$$

Note that B_k has the action of a finite-difference approximation to $F'(x_{k-1})$ in the direction $(x_k - x_{k-1})$. Thus quasi-Newton methods in general bear the same relation to Newton's method as the secant method in the case $n = 1$.

It is clear that the secant method is a quasi-Newton method. In fact, if $n = 1$, then the quasi-Newton equation determines the scalar B_k exactly, and so the secant method is the only quasi-Newton method in this case. If $n > 1$, then the quasi-Newton equation alone does not determine B_k uniquely; hence, there is no unique natural extension of the secant method to the case of an arbitrary number of independent variables. This lack of uniqueness in the general case may be regarded as an advantage, for it allows a variety of quasi-Newton algorithms which may be drawn upon to take advantage of any special structure which may be present in specific problems of interest.

When $n > 1$, one must impose relations between successive matrices B_k and their predecessors which, together with the quasi-Newton equation, uniquely determine these matrices inductively. In general, those relations are chosen with an eye toward minimizing the computational complexity of the resulting update formula for determining B_{k+1} from B_k, x_k, and F while taking maximal advantage of whatever special structure may be shared by the particular problems under consideration. Of the three quasi-Newton methods presented below, the first (Broyden's method) is intended to be a general purpose algorithm which can be applied to all problems without regard to special structure. Consequently, in Broyden's method, B_{k+1} is obtained by adding a rank-one "correction term" to B_k in such a way that the quasi-Newton equation is satisfied and B_{k+1} agrees with B_k on the orthogonal complement of $(x_{k+1} - x_k)$. In a sense, this may be regarded as the "simplest" way to obtain B_{k+1} from B_k in such a way that the quasi-Newton equation is satisfied. On the other hand, the second two methods (the DFP and BFGS methods) are designed for unconstrained minimization problems, in which the Jacobian $F'(x)$ can be expected to be symmetric and positive-definite. Thus the update formulas for these methods are such that the successive B_k's "inherit" symmetry and positive-definiteness from the preceding ones. Not surprisingly, these formulas are more complex than the update formula of Broyden's method. In fact, in order to guarantee hereditary symmetry and positive-definiteness, it is necessary in these formulas to determine B_{k+1} from B_k with a correction term of rank two.

4. Broyden's method for general nonlinear equations

Broyden's method is, in a sense, the "simplest" of the most popular quasi-Newton methods and is intended to be a general-purpose algorithm for solving arbitrary nonlinear equations. To derive the formula used in Broyden's method to update the matrices B_k, suppose that, for some $k \geq 0$, one has arrived at x_k and B_k. Then x_{k+1} can be generated by the formula

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

Our objective is to use x_k, x_{k+1}, B_k and F to update B_k in the "simplest" way to obtain a matrix B_{k+1} which satisfies the quasi-Newton equation.

For convenience, we adopt the following notation:

$$x_k = x, \quad B_k = B, \quad B_{k+1} = \bar{B}, \quad x_{k+1} - x_k = s, \quad F(x_{k+1}) - F(x_k) = y.$$

In this notation, the quasi-Newton equation which we wish B_{k+1} to satisfy is $\bar{B}s = y$. This equation uniquely specifies the action of $\bar{B}$ in the direction of s. Since there is no apparent reason for $\bar{B}$ to differ from B on the orthogonal complement of s, it seems reasonable to impose on $\bar{B}$ the condition that $Bz = \bar{B}z$ for all z such that $z^Ts = 0$. It is easily verified that there is a unique $\bar{B}$ which satisfies both this condition and the quasi-Newton equation. This $\bar{B}$ is given by the formula

$$\bar{B} = B + \frac{(y - Bs)s^T}{|s|^2} .$$

Note that $\bar{B}$ and B differ by a rank-one operator. Restoring subscripts, we obtain the iteration formulas for Broyden's method:

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

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^T}{|s_k|^2} \quad ,$$

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

Does Broyden's method exhibit the key properties attributed to quasi-Newton methods in the preceding section? It can be shown that if x_0 and B_0 are sufficiently near $x^\#$ and $F'(x^\#)$, respectively, then the Broyden iterates are well-defined and converge superlinearly to $x^\#$. (The proof is very involved, and we omit it.) Also, it is clear that, for a given value of k, the determination of x_{k+1} and B_{k+1} requires only the n function evaluations necessary to specify $F(x_{k+1})$, assuming that $F(x_k)$ can be provided from storage. Finally, it is evident that, for a given k, x_{k+1} and B_{k+1} can be determined with $O(n^2)$ arithmetic operations if $B_k^{-1}F(x_k)$ can be evaluated with $O(n^2)$ arithmetic operations.

There are two ways of evaluating $B_k^{-1}F(x_k)$ with $O(n^2)$ arithmetic operations, both of which require information about B_{k-1}. The first way is based on the Sherman-Morrison formula [8] and produces $\bar{B}^{-1}$ from B^{-1} with $O(n^2)$ arithmetic operations in the following way: write

$$\bar{B} = B + \frac{(y - Bs)s^T}{|s|^2} = B + uv^T \quad ,$$

where $u = (y - Bs)$, $v = \dfrac{s^T}{|s|^2}$; then

$$\bar{B}^{-1} = B^{-1} - \frac{1}{1 + \langle v, B^{-1}u \rangle} B^{-1}uv^TB^{-1} \quad .$$

The second way is based on a special factorization procedure due to Gill and Murray [5] which begins with a factorization $B = QR$ and yields a factorization $\bar{B} = \bar{Q}\bar{R}$ with $O(n^2)$ arithmetic operations. (Here, Q and $\bar{Q}$ are orthogonal and R and $\bar{R}$ are upper-triangular.) Since an n-dimensional linear system whose coefficient matrix is factored in this way can be solved with $O(n^2)$ arithmetic operations, this allows the evaluation of the terms $B_k^{-1}F(x_k)$ with $O(n^2)$ arithmetic operations as desired. For reasons of numerical stability, the Gill-Murray factorization procedure is generally preferable to the method using the Sherman-Morrison formula.

5. The DFP and BFGS methods for unconstrained minimization

For the purposes of this note, the basic problem of unconstrained minimization may be regarded as the problem of solving $\nabla f(x) = 0$ in an open convex subset D of R^n, where f is a nonlinear functional from D to R^1. Clearly, this problem is of the type introduced in Section 2, with ∇f playing the role of F. The special feature of this problem is that the Jacobian of the function whose zero is being sought is actually the Hessian $\nabla^2 f$, a matrix which is certainly symmetric. In fact, in most problems of practical interest, $\nabla^2 f$ is positive-definite near the minimum of f.

It seems reasonable to require that the matrices B_k appearing in a quasi-Newton method applied to an unconstrained minimization problem be symmetric and positive-definite. Since each B_k is to be determined from its predecessor by an update formula, it is reasonable to impose conditions on the update formula which guarantee that symmetry and positive-definiteness are inherited by the successive matrices B_k. Unfortunately, imposing hereditary symmetry as well as the quasi-Newton equation completely determines a rank-one update formula, and

this formula does not guarantee hereditary positive-definiteness. Consequently, one is led to look for rank-two update formulas which insure that the successive matrices B_k inherit symmetry and positive-definiteness.

A general rank-two update formula which guarantees hereditary symmetry is the following:

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

where c is any vector in R^n such that $\langle c,s\rangle \neq 0$. A "natural" choice of c which insures hereditary positive-definiteness whenever $\langle y,s\rangle > 0$ is $c = y$. (Since $\langle y,s\rangle \approx \langle \nabla^2 f(x^*)s,s\rangle$ near x^*, one expects $\langle y,s\rangle$ to be positive near x^*.) The resulting update formula is that used in the Davidon-Fletcher-Powell (DFP) method. Denoting by $\bar{B}_{DFP}$ the updated matrix obtained from B by applying this formula, one has

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

$$= (I - \frac{ys^T}{\langle y,s\rangle})B(I - \frac{sy^T}{\langle y,s\rangle}) + \frac{yy^T}{\langle y,s\rangle} .$$

As with Broyden's method, one can show that the DFP iterates converge superlinearly to x^* whenever x_0 and B_0 are sufficiently near x^* and $\nabla^2 f(x^*)$, respectively, and that each iteration requires n function evaluations and $O(n^2)$ arithmetic operations. Although the DFP update formula is a bit more complicated than the Broyden update formula, experience has shown that the DFP method is generally superior to Broyden's method for problems in unconstrained minimization.

At the k^{th} iteration, both Broyden's method and the DFP method require first the determination of $B_k^{-1}F(x_k)$ and then the updating of B_k. It is natural to ask whether a more efficient method might be obtained by applying an update formula directly to B_k^{-1}. If we denote B^{-1} by H and $\bar{B}^{-1}$ by $\bar{H}$, the quasi-Newton equation $\bar{B}s = y$ becomes $s = \bar{H}y$. Carrying out a development completely analogous to that leading to the DFP update formula yields the update formula of the Broyden-Fletcher-Shanno-Goldfarb (BFGS) method. Denoting by $\bar{H}_{BFGS}$ the updated matrix obtained from H by applying this formula, one has

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

It is not difficult to see that, as in the case of the DFP update, this update adds a rank-two correction term to H and guarantees hereditary symmetry and, if $\langle y,s\rangle > 0$, positive-definiteness. Again, it can be shown that the BFGS iterates converge superlinearly to x^* wherever x_0 and H_0 are sufficiently near x^* and $\nabla^2 f(x^*)^{-1}$, respectively. It is clear that each iteration requires n function evaluations and $O(n^2)$ arithmetic operations.

The BFGS method is not the same as the DFP method. In fact,

$$\bar{H}_{BFGS} = (\bar{B}_{DFP})^{-1} + vv^T$$

where $v = \langle y,Hy\rangle^{1/2}[\frac{s}{\langle s,y\rangle} - \frac{Hy}{\langle y,Hy\rangle}]$. According to [3], there is "growing evidence that BFGS is the best current update formula for use in unconstrained minimization".

6. A potential application

We conclude this note by comparing the properties of quasi-Newton methods to those of Newton's method and UHMLE in a potential application to the problem of obtaining maximum-likelihood estimates of the parameters in mixture distributions. Such estimates, of course, play a fundamental role in certain approaches to signature extension, estimation of proportions, and clustering. For a description of the UHMLE algorithm, see [6] and [7].

Let X be an n-dimensional random variable with probability density function

$$p(x) = \sum_{i=1}^{m} \alpha_i^0 \, p_i(x) \ ,$$

where

$$p_i(x) = \frac{1}{(2\pi)^{n/2} |\Sigma_i^0|^{1/2}} \ e^{-1/2 (x-\mu_i^0)^T \Sigma_i^{0-1} (x-\mu_i^0)}$$

and the proportions α_i^0 are positive and sum to 1. Suppose that $\{x_k\}_{k=1,\ldots,N}$ is a sample of independent observations on X. By a maximum-likelihood estimate of the parameters $\{\alpha_i^0, \mu_i^0, \Sigma_i^0\}_{i=1,\ldots,m}$, we mean a choice of parameters $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1,\ldots,m}$ which locally maximizes the log-likelihood function

$$L = \sum_{k=1}^{N} \log p(x_k) \ ,$$

regarded as a function of the parameters $\{\alpha_i, \mu_i, \Sigma_i\}_{i=1,\ldots,m}$. It is known that, loosely speaking, there is a unique strongly-consistent maximum-likelihood estimate. (See [7] for a clarification and proof of this statement.)

The problem which we consider here is to approximate numerically the strongly-consistent maximum-likelihood estimate. This is potentially a very

difficult problem. Indeed, the number of independent variables is $(m - 1) + mn + m\frac{n(n+1)}{2}$, a number which may be very large. Furthermore, the evaluation of functions derived from the log-likelihood function usually involves summation over the entire sample of N observations and, hence, is a source of computational difficulty when the sample is large. In the table below, we list the key properties of UHMLE, Newton's method, and quasi-Newton methods when applied to solving likelihood equations obtained by differentiating the log-likelihood function. It should be noted that, in addition to the arithmetic operations listed in the table, each method requires at each iteration the evaluation of the functions $p_i(x_k)$, $i = 1,\ldots,m$, $k = 1,\ldots,N$.

METHOD	CONVERGENCE	ARITHMETIC OPERATIONS PER ITERATION
UHMLE	Linear	$O(mn^2N)$
Newton's Method	Quadratic	$O_1(m^2n^4N) + O_2(m^3n^6)$
Quasi-Newton Methods	Superlinear	$O_1(mn^2N) + O_2(m^2n^4)$

Of course, many factors must be considered in addition to convergence rates and the amount of arithmetic per iteration when deciding what sort of algorithm is best suited in a particular instance for application to the problem under consideration. For example, UHMLE is a type of gradient method; hence, one might expect UHMLE to enjoy the relatively good global convergence behavior usually associated with gradient methods. Furthermore, gradient methods are often competitive in speed of convergence to Newton's method and quasi-Newton methods when only "ball-park" approximations to the

solution are desired. Since the nearness of the maximum-likelihood estimate to the true parameters will be limited by the variance of the sample observations, "ball-park" approximations will certainly suffice except, perhaps, in the case of a very large sample.

It is difficult to predict circumstances in which the advantage of fast convergence for Newton's method and quasi-Newton methods will outweigh the disadvantage of having to perform a great many arithmetic operations at each iteration with these methods. However, it should be noted that if N is very large relative to m and n, then the number of arithmetic operations per iteration required by quasi-Newton methods is comparable to the number required by UHMLE. Also, if N is very large, one might reasonably want to obtain very accurate approximations of the maximum-likelihood estimate, in which case the superlinear convergence of quasi-Newton methods is clearly preferable to the linear convergence of UHMLE. Consequently, if N is very large relative to m and n and if particularly accurate approximations of the maximum-likelihood estimate are desired, then quasi-Newton methods appear to have a clear-cut advantage over UHMLE. In such circumstances, one might retain the good global properties of UHMLE by employing a hybrid method which initially behaves like UHMLE and then behaves increasingly like a quasi-Newton method as the iteration proceeds.

BIBLIOGRAPHY

1. C. G. Broyden, "A class of methods for solving nonlinear simultaneous equations," Math. Comp. 19 (1965), pp. 577-593.

2. W. C. Davidon, "Variable metric method for minimization," Rep. ANL-5990 Rev. (1959), Argonne National Laboratories, Argonne, Illinois.

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

4. R. Fletcher and M. J. D. Powell, "A rapidly convergent descent method for minimization," Comput. J. 6 (1963), pp. 163-168.

5. P. E. Gill and W. Murray, "Quasi-Newton methods for unconstrained minimization," J. Inst. Math. Appl. 9 (1972), pp. 91-108.

6. B. C. Peters, Jr. and H. F. Walker, "The numerical evaluation of the maximum-likelihood estimate of a subset of mixture proportions," University of Houston Math. Dept. Tech. Report No. 50, Contract NAS-9-12777 (1976).

7. B. C. Peters, Jr., and H. F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions, II," University of Houston Math. Dept. Tech. Report No. 51, Contract NAS-9-12777 (1976).

8. J. Sherman and W. J. Morrison, "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 (1949), p. 621.