THE DAVIDON-FLETCHER-POWELL METHOD AND FAMILIES OF VARIABLE METRIC METHODS FOR UNCONSTRAINED MINIMIZATION

By

ROSALEE JOY TAYLOR

Bachelor of Science in Education
Southwestern Oklahoma State University
Weatherford, Oklahoma
1966

Master of Arts
University of Missouri-Columbia
Columbia, Missouri
1968

Submitted to the Faculty of the Graduate College
of the Oklahoma State University
in partial fulfillment of the requirements
for the Degree of
DOCTOR OF EDUCATION
May, 1976

THE DAVIDON-FLETCHER-POWELL METHOD AND FAMILIES OF VARIABLE METRIC METHODS FOR UNCONSTRAINED MINIMIZATION

Thesis Approved:

Donald W. Grace
Thesis Adviser

J. P. Chandler

D B Aichele

H G. Burchard

E. K. McLachlan

N N Durham
Dean of the Graduate College

PREFACE

This paper is an expository study of Fletcher and Powell's version of Davidon's original variable metric method and generalizations of this method, that is, parametric families of variable metric methods which contain the Davidon-Fletcher-Powell method and have basic properties in common with this method. The main emphasis is on the motivation and basic ideas leading to the development of these methods and on the theoretical properties which form their foundation.

I wish to express my appreciation to my thesis adviser, Dr. Donald W. Grace, for his support and assistance in the preparation of this thesis. I would also like to thank the other members of my committee, Dr. J. P. Chandler, Dr. E. K. McLachlan, Dr. H. G. Burchard, and Dr. D. B. Aichele for their assistance and cooperation.

I am especially grateful to my husband, Dave, for his encouragement and understanding.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$f(x),\ x = (\xi_1, \ldots, \xi_n)^T$	- Function to be minimized
$g(x) = \left(\frac{\partial f(x)}{\partial \xi_1}, \ldots, \frac{\partial f(x)}{\partial \xi_n} \right)^T$	- Gradient vector of f(x)
$G(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial \xi_1 \partial \xi_1} & \cdots & \frac{\partial^2 f(x)}{\partial \xi_1 \partial \xi_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f(x)}{\partial \xi_n \partial \xi_1} & \cdots & \frac{\partial^2 f(x)}{\partial \xi_n \partial \xi_n} \end{bmatrix}$	- Hessian matrix of f(x)
H_k	- K-th approximation to $G^{-1}(x_k)$
x'	- A local minimum of f
$f_k = f(x_k)$	- Function value at x_k
$g_k = g(x_k)$	- Gradient vector at x_k
$d_k = - H_k^T g_k$	- Search direction
α_k	- Step size from linear search
$s_k = \alpha_k d_k = x_{k+1} - x_k$	- Step
$y_k = g_{k+1} - g_k$	- Gradient difference

Scalars are denoted by lower case Greek letters, vectors are denoted by lower case Latin letters, and matrices are denoted by upper case Latin letters.

CHAPTER I

INTRODUCTION

In 1959, W. C. Davidon [14] developed a numerical method for determining an unconstrained local minimum of a differentiable function f of n real variables, $\xi_1, \ldots, \xi_n$. This method generates a sequence of points $x = (\xi_1, \ldots, \xi_n)^T$ in an effort to locate a point at which the gradient vector g, given by

$$g(x) = \left(\frac{\partial f(x)}{\partial \xi_1}, \ \ldots, \ \frac{\partial f(x)}{\partial \xi_n} \right)^T,$$

is zero and at which the Hessian matrix G, whose ij-th element is given by

$$\frac{\partial^2 f(x)}{\partial \xi_i \partial \xi_j}, \ i, \ j = 1, \ \ldots, \ n,$$

is positive definite. If f has continuous second partial derivatives, then such an x is a strong local minimum of f.

The ideas which form the basis for Davidon's minimization procedure can be described by using geometrical concepts. The variables $\xi_1, \ldots, \xi_n$ are the coordinates of the point x in the n-dimensional space R^n. Consider the set $S = \{x \mid f(x) = \mu\}$ for some constant μ. If the point $w = (\omega_1, \ldots, \omega_{n-1}, \omega_n)$ belongs to the set S and $g(w) \neq 0$, without loss of generality suppose $(\partial f/\partial \xi_n)(w) \neq 0$, then by the Implicit

Function Theorem there exists a neighborhood of $(\omega_1, \ldots, \omega_{n-1})$ and a unique continuously differentiable function h from R^{n-1} to R defined on this neighborhood such that $\omega_n = h(\omega_1, \ldots, \omega_{n-1})$ and $f(\xi_1, \ldots, \xi_{n-1}, h(\xi_1, \ldots, \xi_{n-1})) = \mu$ for each point $(\xi_1, \ldots, \xi_{n-1})$ in this neighborhood. The graph of the function h forms an (n - 1)-dimensional surface in R^n. For n = 2, the one-dimensional surfaces are called contour lines of the function f. An illustration is given in Figure 1. The point x' is the minimum point of f.

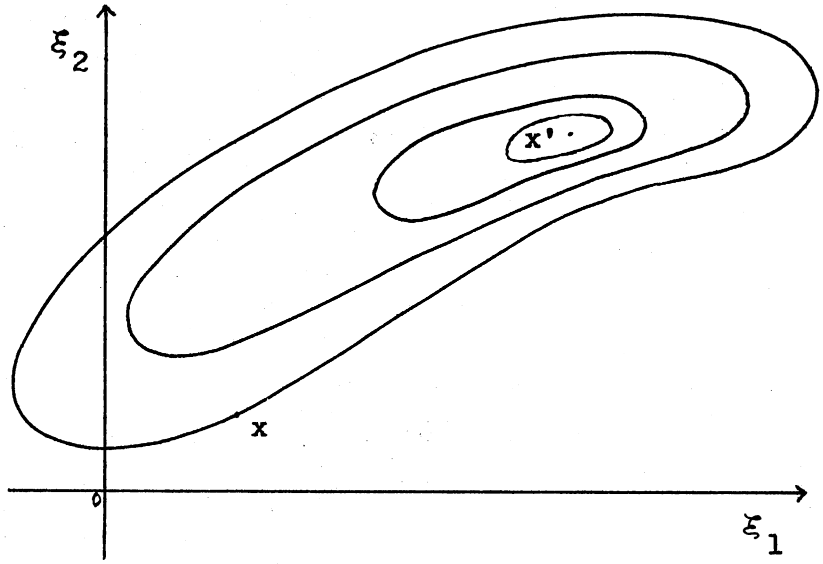

Figure 1. Geometrical Representation of x and the Contour Lines of f

If f has continuous second partial derivatives, then by the Taylor expansion of f about x, for a sufficiently small change Δx,

$$f(x + \Delta x) \doteq f(x) + g^T(x) \Delta x.$$

Differentiating with respect to x gives

$$g(x + \Delta x) \doteq g(x) + G(x)\,\Delta x.$$

Therefore, in a neighborhood of x, the change in gradient, $\Delta g = g(x + \Delta x) - g(x)$, caused by the change in x is approximated by

$$\Delta g \doteq G(x)\,\Delta x.$$

If f is a quadratic function, then the Hessian matrix G is constant and for any Δx,

$$f(x + \Delta x) = f(x) + g^T(x)\,\Delta x + \tfrac{1}{2}\,\Delta x^T\,G\,\Delta x,$$

which implies $\Delta g = G\,\Delta x$. If G is positive definite, then the value of the gradient at the one point x would suffice to determine the minimum. Since the desired change in g(x) is $-\,g(x)$, the equation $-\,g(x) = G\,\Delta x$ may be solved for Δx, which now represents the change in x needed to reach the minimum. However, in general, G is not constant and the minimum may not be obtained by the given single step from the point x. Instead, a sequence of points is generated, starting from the point x. Since explicit evaluation and inversion of G at points that could be far from the minimum might not be worth the amount of computation required, an initial positive definite trial matrix H is assumed for the matrix $[G(x)]^{-1}$. The change in x is then determined by minimizing f in the direction $-\,Hg(x)$. That is, the next point in the sequence, x*, is given by the expression

$$x^* = x - \alpha H g(x),$$

where the scalar $\alpha > 0$ is chosen to minimize $f(x - \alpha' Hg(x))$ with respect to α'. This one-dimensional minimization, called a linear search, is illustrated in Figure 2. After making this change in x, the trial matrix H is improved on the basis of the actual relations between changes in x and changes in the gradient. By iterating these steps, the sequence of points is generated.

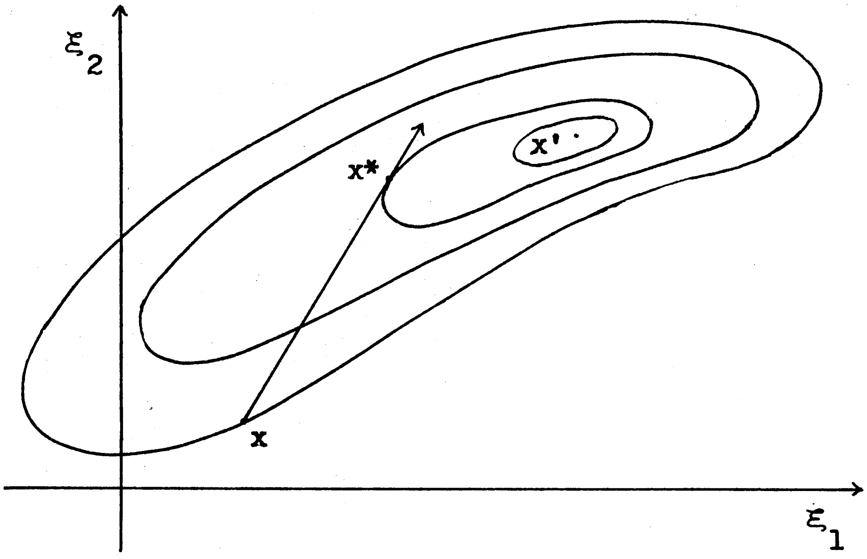

Figure 2. Minimization of f in the Direction - Hg(x)

Associated with the positive definite matrix H is the norm defined by $||\,x\,||_H = \sqrt{x^T H^{-1} x}$, for points x in the n-dimensional space R^n. Thus, H induces the metric $d(x,\ z) = ||\,x - z\,||_H$. Davidon called his method a variable metric method to reflect the fact that H is changed after each iteration.

The change in H at each iteration affects the direction of steepest descent from a given point x, because this direction depends upon how the distance between two points x and z in R^n is measured. In general, there is no reason to assume that a unit of distance along the ξ_i axis is equal to a unit of distance along the ξ_j axis, for $i \neq j$. The definition of distance, that is, the metric, implies a particular system of weighting these units.

If the distance between x and z is defined by $\| x - z \|_H$, then the set of all points z at a distance μ from x is given by the ellipsoid $\| z - x \|_H = \mu$, that is, $\left\{z \mid (z - x)^T H^{-1}(z - x) = \mu^2\right\}$. The direction of steepest descent in the neighborhood bounded by this ellipsoid may be defined as the direction from x to that point on the ellipsoid for which the value of the function f is smallest. It is shown in Appendix 1 of [12] that, as μ tends to zero, this direction approaches a limit which is the direction of the vector

$$d = - Hg(x).$$

Therefore, this direction is called the direction of steepest descent from x relative to H.

If H = I, then $\| x \|_H = \sqrt{x^T x}$ is the Euclidean norm and $d = - g(x)$ is called simply the direction of steepest descent from x. This is the most common usage of the term "steepest descent." In particular, it is the direction used in the classical method of steepest descent described by A. Cauchy [11] in 1847. This method often converges slowly because the direction of steepest descent and the direction to the minimum may be nearly perpendicular. An example is shown in Figure 3. This is to be expected since the direction of steepest descent depends not only upon the function being minimized, but also on the metric. The distinguishing feature of Davidon's method is that the metric is iteratively ?

adjusted in an effort to make the direction of steepest descent relative to the metric point toward a minimum.

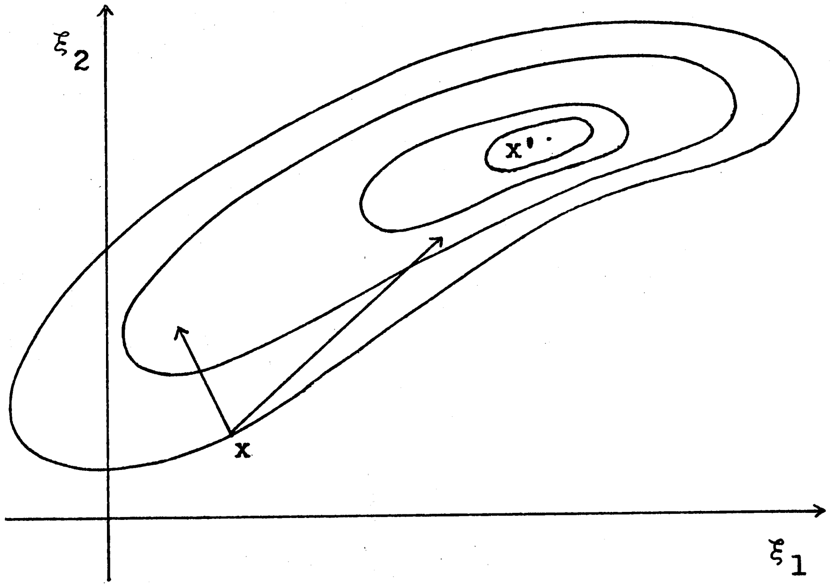

Figure 3. Direction of Steepest Descent at x Versus Direction to the Minimum

The effect of a variable metric and its advantage over a constant metric can be illustrated by a simple example in which the Hessian matrix G is constant. Let f be the function of two variables defined by

$$f(x) = 16\xi_1^2 + \xi_2^2.$$

The Hessian matrix G = diag (32, 2) is a constant positive definite matrix. The contour lines of f are elongated ellipses whose axes are the coordinate axes and whose centers are at the origin. Clearly, the minimum point is (0, 0).

Figure 4 shows the sequence of points generated by minimizing in the direction of steepest descent at each iteration. The metric is constant and given by the Euclidean norm. Note the inefficient zigzag behavior in the vicinity of the minimum.

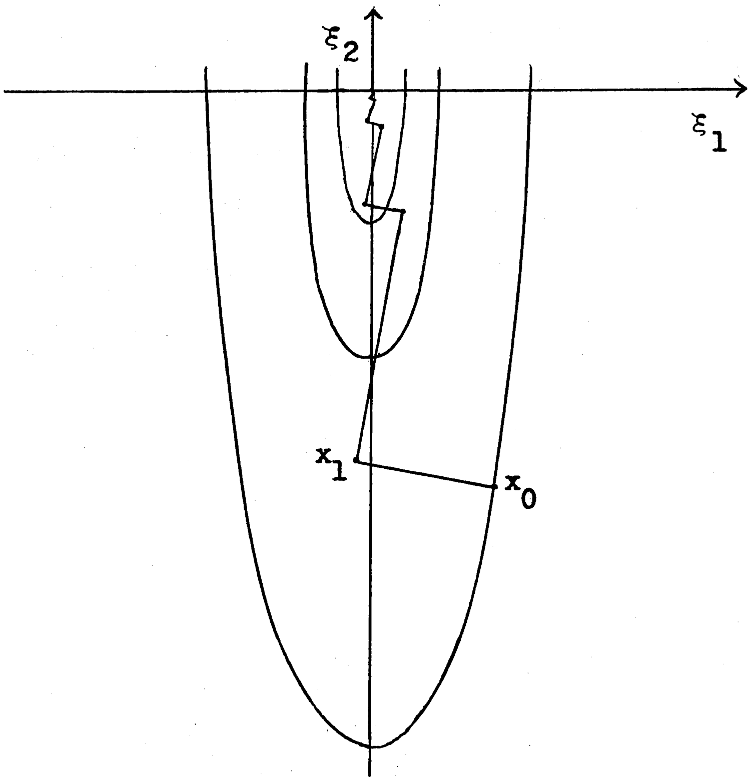

Figure 4. Minimization of $f(x) = 16\xi_1^2 + \xi_2^2$ in Which a Constant Metric Is Used

In Figure 5, the direction - Hg, the direction of steepest descent relative to the variable matrix H, is used at each iteration. In this case, the advantage of a variable metric can be seen, particularly as the minimum is approached.

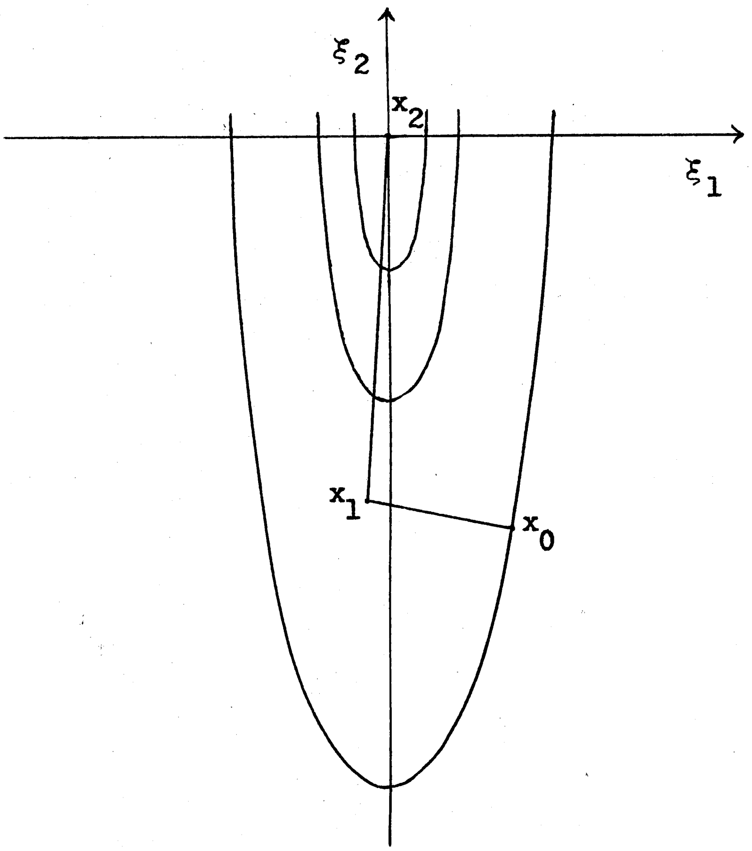

Figure 5. Minimization of $f(x) = 16\xi_1^2 + \xi_2^2$ in Which a Variable Metric Is Used

Since G^{-1} is constant and known, the effect of minimizing in the direction $- G^{-1}g$, the direction of steepest descent relative to G^{-1}, can be shown in Figure 6. Recall that in this case one step in sufficient to reach the minimum. In fact, in the metric space with metric given by $d(x, z) = || x - z ||_{G^{-1}}$, the equation of a circle with center at the origin and radius μ is $x^TGx = \mu^2$. Hence, in this metric space, the contour lines of $f(x) = 16\xi_1^2 + \xi_2^2 = \frac{1}{2}x^TGx$ are circular and the direction of steepest descent, $- g(x)$, points to the minimum, as shown in Figure 7.

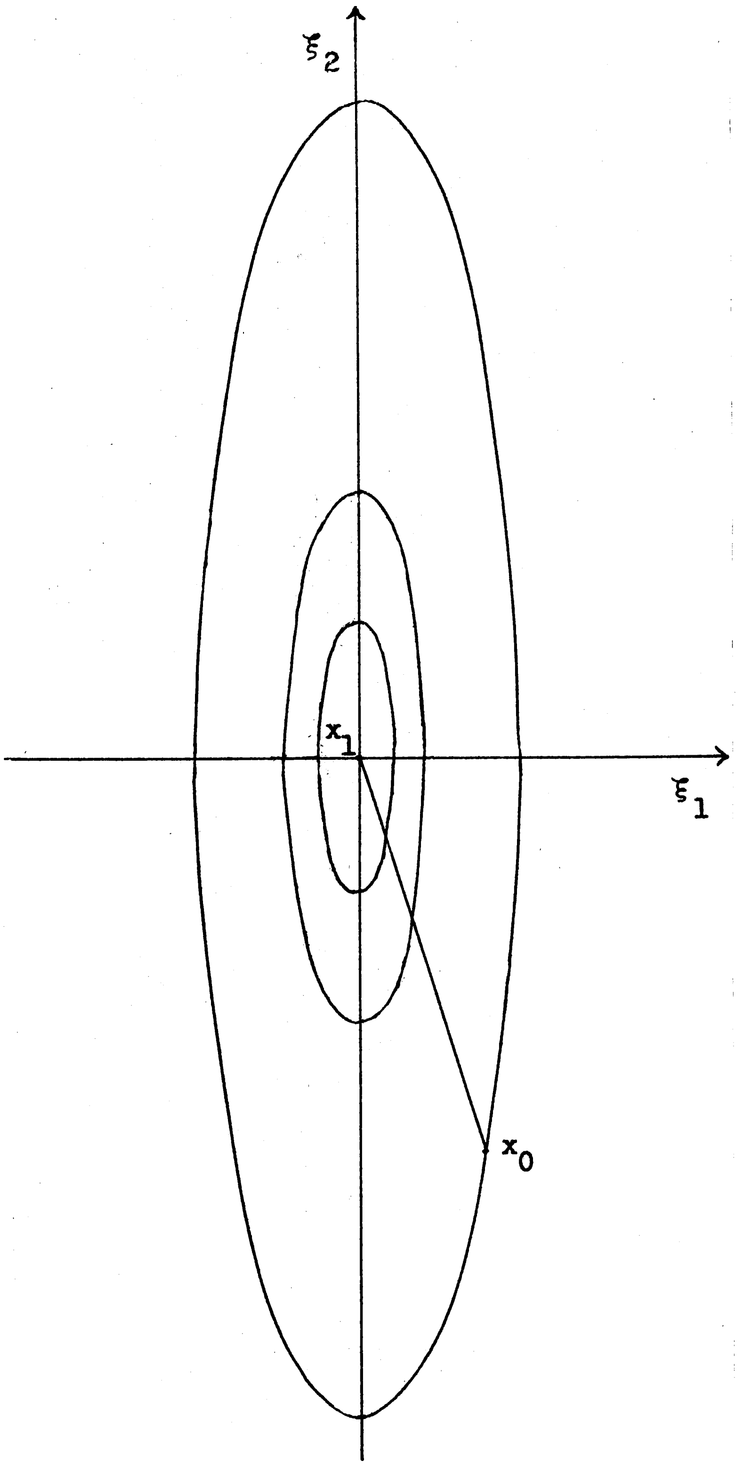

Figure 6. Minimization of $f(x) = 16\xi_1^2 + \xi_2^2$ in Which the Metric Induced by G^{-1} Is Used

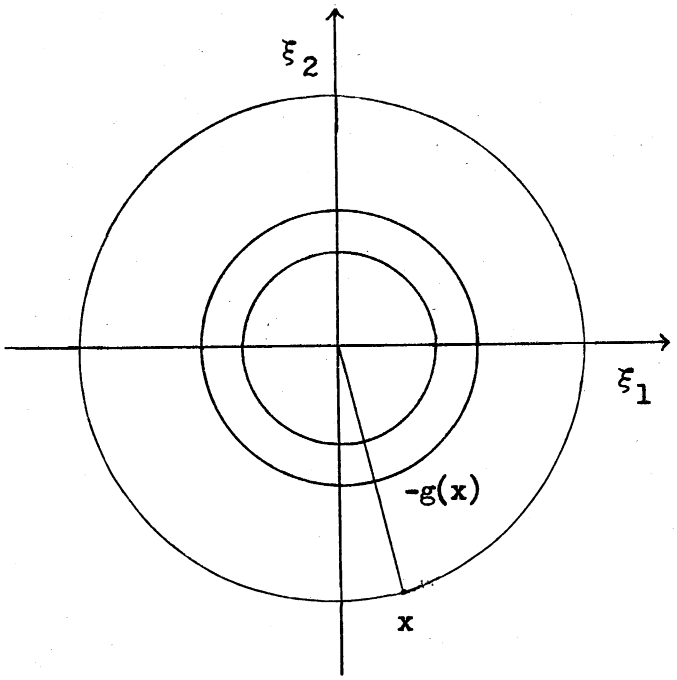

Figure 7. Contour Lines of $f(x) = \frac{1}{2}x^TGx$ in the Metric Space With Metric Induced by G^{-1}

Geometrically, the change in the contour lines from Figure 6 to Figure 7 is the result of a change in scale on the ξ_1 and ξ_2 axes. In Figure 6, a unit of distance along the ξ_1 axis is equal to a unit of distance on the ξ_2 axis; while in Figure 7, the metric has changed the weighting of these units so that the axes of the ellipse are of equal length.

The function f used in the above example is a strictly convex quadratic function, and some of the results illustrated are dependent upon that fact. However, the behavior of a method on such a function is important. Suppose the function f has continuous second partial derivatives and satisfies sufficient conditions for a strong local minimum at x'. Since the gradient of f vanishes at the minimum, the Taylor series expansion about x' gives

$$f(x) \doteq f(x') + \tfrac{1}{2}(x - x')^T G(x')(x - x'),$$

where G(x') is positive definite. Thus, the function f behaves like a strictly convex quadratic function in a neighborhood of x'. Therefore, the behavior of a minimization algorithm on a strictly convex quadratic function is indicative of its behavior in the neighborhood of the minimum of a more general function.

While Davidon's method was not widely publicized, it constituted a considerable advance over then current alternatives. In 1963, R. Fletcher and M. J. D. Powell [26] published a simplified version of Davidon's method, known as the Davidon-Fletcher-Powell, or DFP, method. As in Davidon's method, the next point in the iteration, x*, is found by minimizing f in the direction - Hg(x) from the current point x. However, while Davidon's method used some empirical devices when updating the variable matrix H, in the Fletcher and Powell version, H is updated by adding a symmetric matrix of rank two, defined in terms of H, the change in x, and the change in the gradient.

The DFP method may be applied to a general differentiable function, but proof that the sequence of points generated by this method will always converge to a local minimum of the function, if one exists, can be given only for a more restricted class of functions. In their original publication, Fletcher and Powell established convergence to the minimum of a strictly convex quadratic function. The convergence of the DFP method has since been extended by Powell [47, 49] to more general classes of functions.

Davidon's use of the term "variable metric" was based on the fact that the variable matrix H, as a positive definite matrix, could be used to define a metric. However, this term has been applied by some authors to methods in which the variable matrix is nondefinite. Therefore, the following general definition will be used.

Definition 1.1: A variable metric method is an iterative minimization method using the following iteration. Given the point x and the matrix H, let $d = - H^T g$, where g is the gradient of f at x. Compute the next point $x^* = x + \alpha d$, where α is chosen to minimize $f(x + \alpha' d)$ with respect to α', and update H to $H^* = H + C$, where C is a given correction matrix. Different variable metric methods are obtained from different correction matrices.

Parametric families of variable metric methods, containing the DFP method as a special case, have been developed from a number of different approaches. The first family was developed by C. G. Broyden [6] in 1967. His approach to the minimization of f by finding x such that $g(x) = 0$ is to use a quasi-Newton method for solving this equation. While Newton's method uses the inverse Hessian matrix at each point in the iteration, quasi-Newton methods use an approximation which is modified at each iteration. This modification is such that the new approximation to the inverse Hessian matrix satisfies an equation called the quasi-Newton equation. The purpose of this equation is to force the approximation to possess, to some extent, the properties of the inverse Hessian matrix. Since the modification is made by adding a correction matrix, quasi-Newton methods using linear searches are also variable metric methods. Families of methods can be obtained because these conditions do not uniquely determine the correction matrix. Broyden's family is based on a correction matrix satisfying the quasi-Newton equation and defined in terms of an arbitrary scalar parameter.

A similar approach was taken by D. F. Shanno [53] in 1970. However, his family of methods is based on a correction matrix which is a solution of a particular parametric separation of the quasi-Newton equation. This correction matrix depends upon the parameter introduced in the separation.

In 1970, D. Goldfarb [27] obtained a family of methods from a combination of two correction matrices belonging to a family derived by J. Greenstadt [28] using a variational approach. The variational problem formulated by Greenstadt was to find a symmetric correction matrix of minimum norm which also satisfies the quasi-Newton equation. The norm used was defined in terms of an arbitrary positive definite matrix. Thus, the solution yielded a family of correction matrices.

Although different approaches were used in the development of these one-parameter families, the families of Shanno and Goldfarb are equivalent to Broyden's 1967 family. In addition to containing the DFP method as a special case, this one-parameter family has important properties in common with the DFP method. Therefore, this family is a generalization of the DFP method.

Another family of correction matrices equivalent to Broyden's was published in 1970 by Fletcher [23]. It was developed as a combination of the DFP correction matrix and one derived by an inverse relationship to the DFP matrix. However, Fletcher is concerned with properties of the updating formula when used in an algorithm not requiring linear searches.

The DFP method is generally successful in practice, but numerical difficulties have been noted by Y. Bard [3] and Broyden [6], among others. In particular, the variable matrix H has exhibited a tendency toward singularity. Generalizations of the DFP method offer the possibility of choosing the parameter to eliminate this tendency, while still retaining the desirable characteristics of this method. This idea has been explored by Broyden [7, 8] and Shanno [53].

In 1972, L. C. W. Dixon [17] established a particularly useful result. He proved that, given the same initial conditions, the sequences of points generated by different members of Broyden's 1967 family are identical if the linear search is exact. Therefore, since the DFP method belongs to this family, Powell's general convergence theorem applies to the other members.

More general families of variable metric methods have also been developed. In 1969, J. D. Pearson [42] developed a class of variable metric methods based on the generalized solution of a set of underdetermined linear equations. This class was extended to a more general family of methods by N. Adachi [1] in 1971. Another general family was constructed by H. Huang [30] in 1970 using a unified approach based on the analysis of certain desired properties. Work in classifying these general families has been done by Huang [30], Dixon [18], and Adachi [2].

Thus, since Davidon's original algorithm in 1959, much research has been done on variable metric methods. The numerous papers published have simplified Davidon's method, developed general families, and established new theoretical results. The best known variable metric method is Fletcher and Powell's simplification of Davidon's method. General families offer a choice of parameters that may lead to improved algorithms. Also, the development of these families provides a general theoretical foundation that aids in the understanding of the members. Hence, the study of the DFP method and generalizations of this method is justified.

The primary purpose of this dissertation is to unify the various papers written in this area and to discuss and organize their results. The paper will be concerned mainly with the DFP method and the development of generalizations of this method. The major goals are explanation of these methods with an emphasis on the motivation and basic ideas leading to their development; discussion of their theoretical and numerical properties, concentrating on those principal results which form the foundation for these methods; and organization and classification of these methods based upon their relationships and common properties.

The following organization will be used. The DFP method will be presented first, in Chapter II. This method was the first widely used variable metric method, and as such, provided a basis and motivation for its generalizations. In addition, it will provide an introduction to the basic concepts and help the reader to develop a familiarity with the notation and terminology used. The one-parameter family of methods will be the topic of Chapter III, which will explain the different developments of this family and will examine the various relationships. The properties of this family and the search for an optimal parameter will also be investigated. Chapter IV will discuss the development, properties, and relationships of the more general families. The common properties and the interrelationships of the methods considered in this paper will be summarized in the last chapter, Chapter V.

Variable metric methods are a particular class of methods for finding an unconstrained local minimum of a differentiable function f of n real variables. Since a necessary condition for the point x' to be a local minimum of f is that g(x') = 0, the primary objective is to locate a point satisfying this condition. Thus, the problem of finding a local minimum of f leads to the general problem of solving a system of nonlinear equations

$$h_i(\xi_1, \ldots, \xi_n) = 0, \; i = 1, \ldots, m.$$

This system of m equations in n unknowns, $\xi_1, \ldots, \xi_n$, may be expressed as h(x) = 0, where $h(x) = (h_1(\xi_1, \ldots, \xi_n), \ldots, h_m(\xi_1, \ldots, \xi_n))^T$. For the minimization problem, m = n and $h_i = \partial f/\partial \xi_i$. Hence, any method used to solve a system of nonlinear equations may be applied to the minimization problem. In addition, it is possible to introduce refinements into the method to take account of the special nature of the system. For example, the method may be modified so that the value of f decreases at each iteration. Also, if f has continuous second partial derivatives, then the Jacobian matrix of g, being the Hessian matrix of f, must be symmetric.

Alternatively, the problem of solving the system h(x) = 0 can be converted into a minimization problem. Let p be a function defined on R^m with the property that the point x = 0 is the unique global minimum of p. For example, $p(x) = x^Tx$. Then define the function r by r(x) = p(h(x)). If the system h(x) = 0 has a solution, then x' is a global minimum of r if and only if h(x') = 0. Hence, in order to find x' it suffices to minimize r. In the case that h(x) = 0 has no solution and $p(x) = x^Tx$, a global minimum of r is called a least-squares solution of the system, since it minimizes

$$r(x) = \sum_{i=1}^{m} [h_i(x)]^2.$$

The minimization of a function which is a sum of squares of nonlinear functions is an important special case. General algorithms for unconstrained minimization can be applied to this function, but usually it is much more efficient to use an algorithm that takes account of the fact that the function is a sum of squares. The least-squares problem typically arises when attempting to estimate certain parameters in a functional relationship by means of experimental data. For example, suppose the quantity ψ is assumed to satisfy $\psi = u(\phi; x)$, where u is a known function of an independent variable ϕ and an unknown parameter vector $x = (\xi_1, \ldots, \xi_n)^T$. Then for various values ϕ_i, $i = 1, \ldots, m$, measurements ψ_i, $i = 1, \ldots m$, are made in order to determine x. If these measurements were exact, then the vector x would satisfy the system of m equations in the n unknowns, $\xi_1, \ldots, \xi_n$,

$$u(\phi_i; x) = \psi_i, \; i = 1, \ldots, m.$$

However, in general, the measurements are subject to error so that more measurements than the number of unknowns are taken, that is, $m > n$, and x is determined to minimize the sum of squares of the deviations $\psi_i - u(\phi; x)$. That is, the problem becomes that of minimizing the function

$$f(x) = \sum_{i=1}^{m} [\psi_i - u(\phi_i; x)]^2.$$

A comprehensive study of the iterative solution of systems of nonlinear equations may be found in the book by J. M. Ortega and

W. C. Rheinboldt [41]. Additional references include G. D. Byrne and C. A. Hall [10] and J. W. Daniel [13].

It is assumed that the reader of this paper has had an introduction to numerical optimization. A college level background in analysis and linear algebra will also be assumed. A good summary of the fundamentals of function minimization is given by W. Murray [38]. This book also contains an appendix reviewing some aspects of linear algebra relevant to optimization.

CHAPTER II

DAVIDON-FLETCHER-POWELL METHOD

Description

The DFP method for unconstrained function minimization, published by Fletcher and Powell [26] in 1963, is a simplification of the variable metric method developed by Davidon [14] in 1959. The basic concepts of this variable metric method, discussed in Chapter I, also apply to the DFP method.

The DFP method generates a sequence $\{x_k\}$, $k = 0, 1, 2, \ldots,$ of approximations to a local minimum of a differentiable function f according to the following algorithm.

<u>Algorithm</u> 2.1 (Fletcher and Powell, 1963): Given an initial vector x_0 and an initial matrix $H_0 = I$ or any positive definite matrix.

```
For $k = 0, 1, 2, \ldots,$
    If $g_k = g(x_k) = 0$, then stop.
    Else, set $d_k = - H_k g_k$,
        find $\alpha_k > 0$ which minimizes $f(x_k + \alpha d_k)$ with respect to $\alpha$,
        set $s_k = \alpha_k d_k$,
            $x_{k+1} = x_k + s_k$,
            $y_k = g_{k+1} - g_k$,
            $H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{s_k^T y_k}$.
```

Since the algorithm is terminated when the gradient at the current point x_k becomes zero, this point x_k is a stationary point of f but not necessarily a local minimum. That is, if f has continuous first partial derivatives, then the point x_k satisfies necessary but not sufficient conditions for a local minimum. However, if the function f has continuous second partial derivatives, then the stationary point x_k is a local minimum if the Hessian matrix G at x_k is positive definite. In an implementation of Algorithm 2.1 the termination criterion would be $\| g_k \|_2^2 < \epsilon$ for some given tolerance $\epsilon > 0$ since, in general, g_k will not be exactly zero for any k.

Basic Properties

The step from x_k to x_{k+1} is in the direction $d_k = - H_k g_k$. The step size is chosen to minimize f in that direction, that is, to minimize $f(x_k + \alpha d_k)$ with respect to α. Hence,

$$\left.\frac{df(x_k + \alpha d_k)}{d\alpha}\right|_{\alpha=\alpha_k} = 0,$$

that is,

$$d_k^T g(x_k + \alpha_k d_k) = d_k^T g_{k+1} = 0. \tag{2.1}$$

It was established in Chapter I that, for H_k positive definite, this direction is the direction of steepest descent from x_k relative to H_k. Thus it is expected that f(x) decreases as x moves from x_k in the direction d_k. This is easily shown for a function f having continuous second partial derivatives. For a sufficiently small step $\alpha > 0$, the first order terms in the Taylor series for f give

$$f(x_k + \alpha d_k) \doteq f(x_k) + \alpha d_k^T g_k.$$

Since $\alpha > 0$, this implies that

$$f(x_k + \alpha d_k) < f(x_k) \text{ if and only if } d_k^T g_k < 0,$$

that is, the direction d_k is downhill if and only if $- g_k^T H_k g_k < 0$. Therefore, if H_k is positive definite and $g_k \neq 0$, there exists an $\alpha_k > 0$ such that

$$f(x_{k+1}) = f(x_k + \alpha_k d_k) < f(x_k).$$

This property, known as stability, is defined below.

Definition 2.1: An iterative minimization method is stable if the value of the function being minimized is decreased at each step. That is, if $\{x_k\}$, $k = 0, 1, 2, \ldots,$ is the sequence of points generated by the method and f is the function being minimized, then $f(x_{k+1}) < f(x_k)$ for each k.

Stability is a desirable property for variable metric methods since it guarantees that some progress in decreasing f is made at each step. However, it is not sufficient for convergence because the sequence of function values at the points generated by a stable method may be unbounded below.

The concept of stability may also be considered geometrically. The gradient g_k is normal to the surface $f(x) = f(x_k)$ at the point x_k. Hence the direction d_k will be downhill if and only if the angle ϕ between d_k and $- g_k$ is acute. This is illustrated in Figure 8 for $n = 2$. The angle ϕ between the vectors d_k and $- g_k$ is defined by

$$\cos \phi = \frac{-\, d_k^T g_k}{\| d_k \|_2 \| g_k \|_2}, \quad 0 \leq \phi \leq \pi.$$

Thus, the angle ϕ is acute if and only if $\cos \phi > 0$, that is, if and only if $d_k^T g_k < 0$.

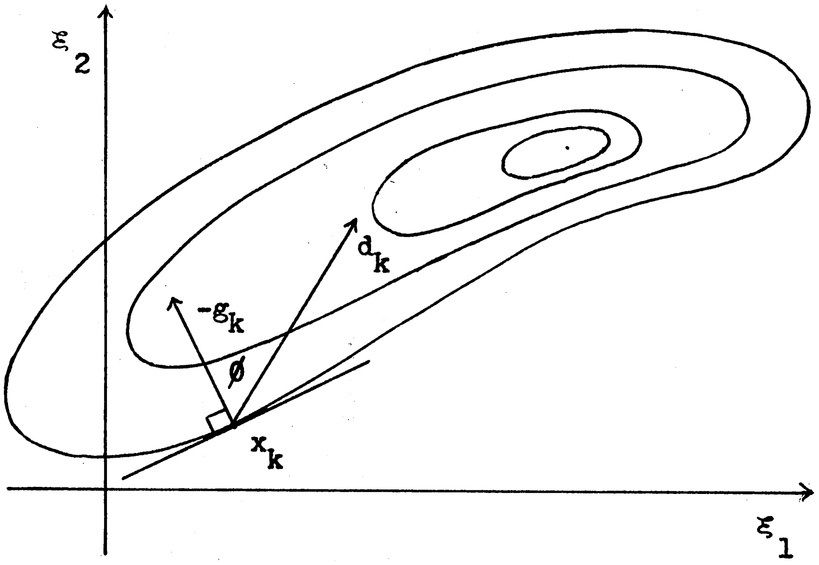

Figure 8. Downhill Direction d_k

Therefore, to establish that the DFP method is stable it must be shown that the variable matrix H_k is positive definite for each k. Since H_0 is positive definite, an inductive argument is used. The following theorem, first proved by Fletcher and Powell, will be proved as a special case of a general family in Chapter III.

Theorem 2.1: For each k, the variable matrix H_k in the DFP method defined in Algorithm 2.1 is positive definite.

Corollary 2.1: The DFP method is stable.

Fletcher and Powell also proved some properties of the DFP method when applied to quadratic functions. For the remainder of this section, let the function f be given by

$$f(x) = \tfrac{1}{2}x^TGx + a^Tx + \gamma, \tag{2.2}$$

where the Hessian matrix G is positive definite. Then, since f is a strictly convex quadratic function, f has a unique minimum. It was shown that the method, when applied to this function, finds the minimum in at most n iterations. Termination in less than n iterations would occur if $H_k = G^{-1}$ for some $k < n$, since, as shown in Chapter I, a search in the direction $d_k = - G^{-1}g_k$ would find the minimum. This property, called quadratic termination, is defined below. It is important because it assures rapid convergence in the final stages of minimization since, as shown in Chapter I, even a nonquadratic function behaves approximately quadratically in a neighborhood of a minimum.

Definition 2.2: An iterative minimization method is quadratically terminating if it finds the minimum of a strictly convex quadratic function of n variables in at most n iterations.

The term "quadratic convergence" is sometimes used for this property instead of "quadratic termination." Since the above definition does not mean that the sequence $\{x_k\}$, $k = 0, 1, \ldots$, converges quadratically, the term "quadratic convergence" will not be used to avoid confusion with the use of this term to mean rate of convergence.

Proof of the following theorem, establishing quadratic termination for the DFP method, will follow as a special case of a general family in Chapter IV.

<u>Theorem</u> 2.2: If f is a strictly convex quadratic function of n variables, then the DFP method finds the minimum of this function in at most n iterations.

Fletcher and Powell's proof of this theorem is an induction proof establishing

$$s_i^T G s_j = 0,\ 0 \le i < j \le k, \tag{2.3}$$

$$H_k G s_i = s_i,\ 0 \le i < k, \tag{2.4}$$

for $1 \le k \le n$. If the algorithm has not terminated due to $g_k = 0$ for some $0 \le k < n$, then $\alpha_k > 0$, $0 \le k < n$. It then follows from (2.3) and the definition of s_k that the search directions $d_0, d_1, \ldots, d_{n-1}$ are nonzero and

$$d_i^T G d_j = 0,\ 0 \le i < j \le n - 1.$$

This property, called conjugacy, is defined below.

<u>Definition</u> 2.3: The nonzero vectors $w_0, w_1, \ldots, w_k$ are conjugate with respect to the positive definite matrix A if

$$w_i^T A w_j = 0,\ 0 \le i < j \le k.$$

It is easily shown that this definition implies that the vectors $w_0, \ldots, w_k$ are linearly independent. The following theorem shows that termination can be obtained by performing linear searches in n conjugate directions.

<u>Theorem</u> 2.3: Let the iterative minimization method in which each iteration is a linear search in a given direction, that is,

$$x_{k+1} = x_k + \alpha_k d_k, \ k = 0, 1, \ldots,$$

where α_k is such that

$$d_k^T g_{k+1} = 0, \tag{2.5}$$

be applied to the function f defined by (2.2). If the search directions $d_0, d_1, \ldots, d_{n-1}$ are conjugate with respect to G, that is, are nonzero and satsify

$$d_i^T G d_j = 0, \ 0 \leq i < j \leq k, \ 1 \leq k \leq n - 1, \tag{2.6}$$

then the minimum will be found in at most n iterations.

<u>Proof</u>: For f defined by (2.2), the gradient g_{k+1} at x_{k+1} is given by

$$g_{k+1} = Gx_{k+1} + a.$$

Using the iteration formula repeatedly, it follows that

$$\begin{aligned} g_{k+1} &= G(x_k + \alpha_k d_k) + a \\ &\vdots \\ &= G(x_{i+1} + \alpha_{i+1} d_{i+1} + \ldots + \alpha_k d_k) + a \\ &= g_{i+1} + \textstyle\sum_{j=i+1}^{k} \alpha_j G d_j, \ 0 \leq i \leq k - 1. \end{aligned}$$

Thus, by (2.5) and (2.6),

$$d_i^T g_{k+1} = d_i^T g_{i+1} + \textstyle\sum_{j=i+1}^{k} \alpha_j d_i^T G d_j = 0, \ 0 \leq i \leq k - 1.$$

Combining this equation with (2.5) gives

$$d_i^T g_{k+1} = 0,\ 0 \leq i \leq k, \tag{2.7}$$

which, for k = n - 1, yields

$$d_i^T g_n = 0,\ 0 \leq i \leq n - 1.$$

The conjugacy of the vectors d_0, d_1, ..., d_{n-1} implies their linear independence. Hence, g_n is orthogonal to n linearly independent n-dimensional vectors which is possible only if $g_n = 0$. Since G is positive definite the stationary point x_n is the desired minimum.

Theorem 2.3 is the basis for a class of quadratically terminating methods, known as conjugate direction methods. It follows that the DFP method is also a conjugate direction method and obtains its quadratic termination on that basis.

Definition 2.4: A conjugate direction method is an iterative minimization method in which each iteration is a linear search in a given direction, with the property that the directions generated for a quadratic function with positive definite Hessian matrix are conjugate with respect to that matrix.

In the DFP method, as in Davidon's method, the variable matrix H is used to approximate G^{-1}, the inverse Hessian matrix. For the quadratic function f, an interesting result is that the modifications to this variable matrix, using only evaluations of the function and its gradient, are such that $H_n = G^{-1}$. That is, the n-th approximation is the exact inverse Hessian matrix of f. This result is obtained for the DFP method by modifying H_k so that for each k, s_0, s_1, ..., s_k are linearly independent eigenvectors of $H_{k+1}G$ with eigenvalue unity.

That is, from (2.4),

$$H_{k+1}Gs_i = s_i, \; 0 \le i < k + 1. \tag{2.8}$$

For $k = n - 1$, (2.8) gives

$$H_nGs_i = s_i, \; 0 \le i < n. \tag{2.9}$$

If $\alpha_i > 0$, $0 \le i < n$, then the vectors $s_0, s_1, \ldots, s_{n-1}$ are nonzero and hence by (2.3) are conjugate with respect to G. This implies that they are linearly independent, so that if E is the matrix

$$E = [s_0, s_1, \ldots, s_{n-1}]$$

then E^{-1} exists. Thus, from (2.9), $H_nGE = E$ which then implies $H_nG = I$.

At the k-th iteration, the matrix H_k is modified by adding to it the two matrices

$$A_k = \frac{-H_ky_ky_k^TH_k}{y_k^TH_ky_k}, \text{ and } B_k = \frac{s_ks_k^T}{s_k^Ty_k}.$$

The form of the matrix A_k can be deduced because equation (2.8) must be valid for $i = k$. That is, the equation

$$H_{k+1}Gs_k = s_k \tag{2.10}$$

must be satisfied. For f given by (2.2),

$$\begin{aligned} y_k &= g_{k+1} - g_k \\ &= (Gx_{k+1} + a) - (Gx_k + a) \\ &= Gs_k. \end{aligned} \tag{2.11}$$

Hence, equation (2.10) is equivalent to

$$H_{k+1}y_k = s_k, \tag{2.12}$$

which, using the definition of H_{k+1}, gives the equation

$$H_k y_k + A_k y_k + B_k y_k = s_k.$$

From the definition of B_k it is easily seen that $B_k y_k = s_k$, so that A_k must satsify the equation $A_k y_k = - H_k y_k$. This implies that the simplest form of A_k is given by

$$A_k = \frac{- H_k y_k z_k^T}{z_k^T y_k}$$

for some vector z_k. Since H_k, and thus A_k, is to be symmetric,

$$A_k = \frac{- H_k y_k y_k^T H_k}{y_k^T H_k y_k}.$$

B_k is the factor which makes H tend to G^{-1} in the sense that for the quadratic function f,

$$G^{-1} = \sum_{k=0}^{n-1} B_k. \tag{2.13}$$

This result can be proved from the conjugacy conditions (2.3) because these imply

$$\begin{aligned} E^T GE &= [s_0, s_1, \ldots, s_{n-1}]^T G[s_0, s_1, \ldots, s_{n-1}] \\ &= \text{diag}\,(s_0^T Gs_0, s_1^T Gs_1, \ldots, s_{n-1}^T Gs_{n-1}). \end{aligned}$$

Then, if D is this diagonal matrix, it follows that $G = (ED^{-1}E^T)^{-1}$.

Thus,

$$\begin{aligned} G^{-1} &= (ED^{-1})E^T \\ &= [(s_0^TGs_0)^{-1}s_0, \ldots, (s_{n-1}^TGs_{n-1})^{-1}s_{n-1}][s_0, \ldots, s_{n-1}]^T \\ &= \sum_{k=0}^{n-1} (s_k^TGs_k)^{-1}s_ks_k^T. \end{aligned} \tag{2.14}$$

Using (2.11) and the definition of B_k, equation (2.14) gives

$$\begin{aligned} G^{-1} &= \sum_{k=0}^{n-1} (s_k^Ty_k)^{-1}s_ks_k^T \\ &= \sum_{k=0}^{n-1} B_k \end{aligned}$$

and equation (2.13) is established.

Equation (2.12) is also true for nonquadratic functions. From the definition of H_{k+1},

$$H_{k+1}y_k = H_ky_k - \frac{H_ky_ky_k^TH_k}{y_k^TH_ky_k}\, y_k + \frac{s_ks_k^T}{s_k^Ty_k}\, y_k = s_k. \tag{2.15}$$

This result is significant because equation (2.12) is the quasi-Newton equation referred to in Chapter I. The derivation of this equation will be discussed in Chapter III where it will be used to define a quasi-Newton method. It will then follow from (2.15) that the DFP method is a quasi-Newton method

In 1968, G. E. Meyers [37] explored the eigenvalues and eigenvectors of the variable matrix H used in the DFP method with $H_0 = I$ for the quadratic function defined by (2.2) leading to a proof that the gradient vectors at each step are mutually orthogonal. From this, a geometric interpretation of the H matrix in terms of the projection of the negative of the gradient into a solution subspace was derived.

Since each matrix H_i is positive definite, its eigenvalues are positive real numbers. In particular, it will be shown that at least $n - i - 1$ of these are unity when the function is quadratic. It is assumed that $g_i \neq 0$, $0 \leq i < n$. The following lemma is needed.

Lemma 2.1: The scalar relation $g_j^T H_{i-1} g_i = 0$ holds for $0 < i < j \leq n$.

Proof: From (2.7)

$$g_j^T d_i = 0, \quad 0 < i < j \leq n. \tag{2.16}$$

Also, by the definitions of d_i and H_i,

$$d_i = -\left[H_{i-1} - \frac{H_{i-1} y_{i-1} y_{i-1}^T H_{i-1}}{y_{i-1}^T H_{i-1} y_{i-1}} + \frac{s_{i-1} s_{i-1}^T}{s_{i-1}^T y_{i-1}} \right] g_i .$$

Using the definition of y_{i-1} and (2.1) with the symmetry of H_{i-1},

$$y_{i-1}^T H_{i-1} y_{i-1} = g_i^T H_{i-1} g_i + g_{i-1}^T H_{i-1} g_{i-1},$$

$$y_{i-1}^T H_{i-1} g_i = g_i^T H_{i-1} g_i, \text{ and}$$

$$s_{i-1}^T g_i = 0.$$

Hence, the expression for d_i reduces to

$$d_i = -\left[H_{i-1} - \frac{H_{i-1} y_{i-1} y_{i-1}^T H_{i-1}}{g_i^T H_{i-1} g_i + g_{i-1}^T H_{i-1} g_{i-1}} + \frac{s_{i-1} s_{i-1}^T}{s_{i-1}^T y_{i-1}} \right] g_i$$

$$= - H_{i-1} g_i + \frac{H_{i-1} y_{i-1} (g_i^T H_{i-1} g_i)}{g_i^T H_{i-1} g_i + g_{i-1}^T H_{i-1} g_{i-1}}. \tag{2.17}$$

Substituting this expression for d_i into (2.16) gives

$$g_j^T H_{i-1} g_i - \frac{g_j^T H_{i-1} y_{i-1}(g_i^T H_{i-1} g_i)}{g_i^T H_{i-1} g_i + g_{i-1}^T H_{i-1} g_{i-1}} = 0.$$

But, by the definition of y_{i-1} and (2.16),

$$g_j^T H_{i-1} y_{i-1} = g_j^T H_{i-1} g_i,$$

implying that

$$g_j^T H_{i-1} g_i \left[1 - \frac{g_i^T H_{i-1} g_i}{g_i^T H_{i-1} g_i + g_{i-1}^T H_{i-1} g_{i-1}} \right] = 0.$$

For the factor in brackets on the right to be zero, $g_{i-1}^T H_{i-1} g_{i-1}$ must be zero. This is impossible since g_{i-1} is assumed to be nonzero and H_{i-1} is positive definite. Therefore, $g_j^T H_{i-1} g_i = 0$ and the lemma is proved.

Theorem 2.4: For $0 \le i < j < n$, the gradients g_j are eigenvectors of the matrix H_i with eigenvalue unity.

Proof: The definition of H_i, for $i > 0$, gives

$$H_i g_j = H_{i-1} g_j - \frac{H_{i-1} y_{i-1}(y_{i-1}^T H_{i-1} g_j)}{y_{i-1}^T H_{i-1} y_{i-1}} + \frac{s_{i-1}(s_{i-1}^T g_j)}{s_{i-1}^T y_{i-1}}.$$

But, by (2.7), $s_{i-1}^T g_j = 0$, and by Lemma 2.1 and (2.7),

$$y_{i-1}^T H_{i-1} g_j = g_i^T H_{i-1} g_j - g_{i-1}^T H_{i-1} g_j = 0,$$

implying that $H_i g_j = H_{i-1} g_j$. Repeated application of the above reasoning gives the result

$$H_i g_j = H_{i-1} g_j = \cdot \cdot \cdot = H_0 g_j = g_j$$

which establishes the theorem.

An immediate consequence of this theorem is the mutual orthogonality of the gradient vectors. The theorem shows that

$$H_i g_j = g_j, \; 0 \leq i < j < n,$$

so that $g_i^T g_j = g_i^T H_i g_j$. Then, by the symmetry of H_i and (2.7), it follows that $g_i^T g_j = - d_i^T g_j = 0$. Since mutual orthogonality of nonzero vectors implies their linear independence, it is confirmed that unity is an eigenvalue of H_i of multiplicity $n - i - 1$.

A further consequence of this theorem is that the expression for the search direction for a quadratic function can be reduced to a recursion formula. This formula is derived in the following corollary.

Corollary 2.2: For a quadratic function, the direction vectors of the DFP method can be given by the recursion formula

$$d_i = \frac{[g_i g_{i-1}^T + (g_i^T g_i) I] d_{i-1}}{g_i^T g_i + g_{i-1}^T d_{i-1}}, \; i > 0.$$

Proof: From (2.17),

$$d_i = - H_{i-1} g_i + \frac{(g_i^T H_{i-1} g_i) H_{i-1} (g_i - g_{i-1})}{g_i^T H_{i-1} g_i + g_{i-1}^T H_{i-1} g_{i-1}}.$$

By applying Theorem 2.4,

$$d_i = - g_i + \frac{(g_i^T g_i)(g_i + d_{i-1})}{g_i^T g_i - g_{i-1}^T d_{i-1}}.$$

Combining these two terms gives the equation

$$d_i = \frac{[g_i g_{i-1}^T + (g_i^T g_i)I]d_{i-1}}{g_i^T g_i - g_{i-1}^T d_{i-1}}$$

and the corollary is proved.

From these results, a geometric interpretation of the H matrix for a quadratic function can be given, namely that the matrix H_i projects the negative of the gradient g_i into the space spanned by d_i, ..., d_{n-1}. This projected gradient becomes the next direction of search for the minimization of the function in this space. Since $d_i = - H_i g_i$, the following theorem establishes this interpretation.

<u>Theorem</u> 2.5: The direction vector d_i, $0 \leq i < n$, in the DFP method, with $H_0 = I$, applied to the function f given by (2.2), is the projection of the negative of the gradient g_i in the space spanned by the vectors d_i, ..., d_{n-1}.

<u>Proof</u>: Let W be the space spanned by d_i, ..., d_{n-1}. Since the direction vectors are conjugate with respect to the Hessian matrix G, the vectors Gd_0, ..., Gd_{i-1} span V, the orthogonal complement of W. Hence it must be shown that

$$- g_i = d_i + q_i$$

where q_i is in V. Noting that, by (2.11),

$$\begin{aligned} Gd_j &= (1/\alpha_j)Gs_j \\ &= (1/\alpha_j)y_j, \end{aligned}$$

it is sufficient to show that

$$d_i = \begin{cases} - g_i, & \text{if } i = 0, \\ - g_i - \sum_{j=0}^{i-1} \gamma_j y_j, & \text{if } 0 < i < n, \end{cases}$$

for some scalars γ_j. Proof is by induction. Since $d_0 = - g_0$, the induction is valid for $i = 0$. Assume that

$$d_{i-1} = - g_{i-1} - \sum_{j=0}^{i-2} \delta_j y_j, \quad 0 < i < n$$

where the δ_j are scalars. Then

$$g_{i-1}^T d_{i-1} = - g_{i-1}^T g_{i-1} - \sum_{j=0}^{i-2} \delta_j g_{i-1}^T (g_{j+1} - g_j),$$

which, by the mutual orthogonality of the gradient vectors, reduces to

$$\begin{aligned} g_{i-1}^T d_{i-1} &= - g_{i-1}^T g_{i-1} - \delta_{i-2} g_{i-1}^T g_{i-1} \\ &= - (1 + \delta_{i-2}) g_{i-1}^T g_{i-1}. \end{aligned} \tag{2.18}$$

From Corollary 2.2,

$$d_i = \frac{g_i (g_{i-1}^T d_{i-1}) + (g_i^T g_i) d_{i-1}}{g_i^T g_i - g_{i-1}^T d_{i-1}}$$

and substituting (2.18) and the induction hypothesis gives

$$d_i = \frac{- g_i (1 + \delta_{i-2})(g_{i-1}^T g_{i-1}) - (g_i^T g_i)(g_{i-1} + \sum_{j=0}^{i-2} \delta_j y_j)}{g_i^T g_i + (1 + \delta_{i-2})(g_{i-1}^T g_{i-1})}$$

which can be rewritten as

$$d_i = \frac{- g_i(g_i^T g_i) \ - g_i(1 + \delta_{i-2})(g_{i-1}^T g_{i-1})}{g_i^T g_i \ + (1 + \delta_{i-2})(g_{i-1}^T g_{i-1})}$$

$$+ \frac{(g_i^T g_i)g_i - (g_i^T g_i)g_{i-1} - (g_i^T g_i) \sum_{j=0}^{i-2} \delta_j y_j}{g_i^T g_i \ + (1 + \delta_{i-2})(g_{i-1}^T g_{i-1})}.$$

Defining

$$\gamma_{i-1} = \frac{- g_i^T g_i}{g_i^T g_i \ + (1 + \delta_{i-2})(g_{i-1}^T g_{i-1})}, \text{ and}$$

$$\gamma_j = \gamma_{i-1}\delta_j, \ j = 0, \ \ldots, \ i - 2,$$

gives

$$\begin{aligned} d_i &= - g_i - \gamma_{i-1}y_{i-1} - \sum_{j=0}^{i-2} \gamma_j y_j \\ &= - g_i - \sum_{j=0}^{i-1} \gamma_j y_j \end{aligned}$$

and the theorem is proved.

Convergence

In the years following its publication in 1963, Fletcher and Powell's modification of Davidon's variable metric method became one of the most frequently used and most successful techniques for finding the minimum of a differentiable function of several real variables. However, until 1971, it had been proved only that the method is successful if the function is a strictly convex quadratic function, (Theorem 2.2); although in practice, it handled many types of functions successfully. It is difficult to prove convergence because the method is intended to be applied to general differentiable functions.

In 1971, Powell [47] extended convergence of the method to a class of functions more general than strictly convex quadratic functions. The conditions the function f must satisfy are:

1) f has continuous second partial derivatives, and

2) there exists a positive constant ϵ such that, for all x, the eigenvalues of G(x) are not less than ϵ, where G(x) is the Hessian matrix of f at x.

Condition 1) restricts the class of functions to which f belongs to one for which sufficient conditions on f at the minimum exist. Condition 2) is a very strict convexity condition called uniform convexity. Since it implies that G(x) is positive definite for all x, if f satisfies conditions 1) and 2) then x' is a strong local minimum if $g(x') = 0$. In other words, the sequence $\{x_k\}$, k = 0, 1, ..., converges to x' if the sequence $\{g_k\}$, k = 0, 1, ..., tends to zero. The convergence theorem established by Powell is stated below.

Theorem 2.6: If the function f satisfies conditions 1) and 2), then the sequence of points, $\{x_k\}$, k = 0, 1, ..., generated by the DFP method, converges to x', the point at which f is minimum.

Proof of this theorem is given as proof of Theorem 1 in [47]. The method of proof is to define T_k to be the matrix H_k^{-1}, to obtain an expression for the trace of T_k, and to show that this expression implies a contradiction unless the sequence of gradients $\{g_k\}$, k = 0, 1, ..., tends to zero.

By requiring one other condition on the function f, Powell also proves that the DFP method converges superlinearly. The condition required is the Lipschitz condition at the minimum x' given below.

3) There exists a constant δ such that, for all vectors x belonging to the set $S = \{x \mid f(x) \leq f(x_0)\}$, the inequality

$$\left| \frac{\partial^2 f(x)}{\partial\xi_i \partial\xi_j} - \frac{\partial^2 f(x')}{\partial\xi_i\, \partial\xi_j} \right| \leq \delta\ \| x - x' \|_2,\ i,\ j = 1,\ 2,\ \ldots,\ n, \quad (2.19)$$

is satisfied.

The Lipschitz condition (2.19) need only be satisfied on the set S since the stability of the DFP method implies that all points x_k generated by the method belong to this set. The following theorem then establishes the rate of convergence for the DFP method under these conditions. Proof of this theorem is found as proof of Theorem 4 in [47].

Theorem 2.7: If the function f satsifies conditions 1), 2), and 3), then

$$\frac{\| x_{k+1} - x' \|_2}{\| x_k - x' \|_2} \to 0 \text{ as } k \to \infty ,$$

where the vectors $\{x_k\}$, $k = 0, 1, 2, \ldots$, are the points generated by the DFP method and where $f(x')$ is the minimum value of f. That is, the DFP method converges superlinearly.

If f satsifies conditions 1) and 2), then for each vector x_0, the set S defined in 3) has additional properties established by the following lemma.

Lemma 2.2: $S = \{x \mid f(x) \leq f(x_0)\}$ is closed, convex, and bounded.

Proof: Since $S = f^{-1}(-\infty, f(x_0)]$, the closure of S follows from the continuity of the function f. The convexity of the set follows from the fact that f is a convex function. If x and z are in S and $0 \leq \delta \leq 1$ is a scalar, then, by the convexity of f and the definition of S,

$$f(\delta x + (1 - \delta)z) \leq \delta f(x) + (1 - \delta)f(z)$$

$$\leq \delta f(x_0) + (1 - \delta)f(x_0)$$

Thus, $f(\delta x + (1 - \delta)z) \leq f(x_0)$ which implies $\delta x + (1 - \delta)z$ is in S. Therefore, S is convex.

To show that S is bounded, let d be any direction through x_0 that is normalized, that is, $\| d \|_2 = 1$, and let h be the function of one variable defined by

$$h(\alpha) = f(x_0 + \alpha d).$$

Then

$$h'(\alpha) = d^T g(x_0 + \alpha d), \text{ and } h''(\alpha) = d^T G(x_0 + \alpha d)d.$$

If U is the orthonormal basis of eigenvectors corresponding to eigenvalues $\lambda_1, \ldots, \lambda_n$ of $G(x_0 + \alpha d)$, then for some vector c, d = Uc. Thus, by condition 2),

$$d^T G(x_0 + \alpha d)d = c^T U^T G(x_0 + \alpha d)Uc$$

$$= c^T \text{ diag } (\lambda_1, \ldots, \lambda_n) \ c$$

$$\geq \epsilon \ \| c \|_2^2$$

That is, $h''(\alpha) \geq \epsilon$, since the orthogonality of U implies that $\| c \|_2 = \| d \|_2 = 1$. Then the function r defined by

$$r(\alpha) = h(\alpha) - h(0) - \alpha d^T g(x_0) - \tfrac{1}{2}\alpha^2 \epsilon$$

is convex since

$$r''(\alpha) = h''(\alpha) - \epsilon \geq 0.$$

Also, $r(0) = r'(0) = 0$, so that, for each α, $r(\alpha) \geq 0$, and hence

$$h(\alpha) \geq h(0) + \alpha d^T g(x_0) + \tfrac{1}{2}\alpha^2\epsilon.$$

But, the right hand side of this inequality exceeds $h(0)$ if

$$|\alpha| > 2\| g(x_0) \|_2/\epsilon \geq 2| d^T g(x_0) |/\epsilon.$$

That is,

$$f(x_0 + \alpha d) > f(x_0) \text{ if } \| (x_0 + \alpha d) - x_0 \|_2 > 2\| g(x_0) \|_2/\epsilon.$$

Thus, since the direction of d is arbitrary,

$$f(x) > f(x_0) \text{ if } \| x - x_0 \|_2 > 2\| g(x_0) \|_2/\epsilon.$$

Therefore, the set of points x satisfying the condition $f(x) \leq f(x_0)$ is bounded and Lemma 2.2 is proved.

An important corollary of this lemma and the fact that f is continuous is that the minimum value of f is attained at some finite point x'. Moreover, the minimum value of f is attained at only one point. By the proof of the lemma,

$$f(x) > f(x') \text{ if } \| x - x' \|_2 > 2\| g(x') \|_2/\epsilon.$$

But, $g(x') = 0$, so if $x \neq x'$, $f(x) > f(x')$. In addition, this lemma and the definition of a derivative imply that if f is three times continuously differentiable at x', then f satisfies condition 3).

It should be noted that Theorems 2.6 and 2.7 are sometimes relevant to non-convex functions, because the conditions on f have to be obtained only for values of x that satsify the inequality $f(x) \leq f(x_0)$. Moreover, the structure of the algorithm is such that any calculated vector x_k can be regarded as a starting point for the later iterations. Therefore, if the algorithm is applied to a non-convex function, and if it happens that a point x_k is calculated, such that the derivative conditions are met for all x satisfying the condition $f(x) \leq f(x_k)$, then convergence to the minimum at a superlinear rate is implied. Moreover, if the sequence of points $\{x_k\}$, $k = 0, 1, 2, \ldots,$ converges to a local minimum of f that is not the global minimum, then it may also be possible to apply the theorems to infer superlinear convergence, by isolating the domain of x to a neighborhood of the local minimum. However, no conclusions about the behavior of the algorithm may be drawn when the estimates x_k are in a region where the second derivative matrices of f do not satisfy the required conditions.

In 1972, Powell [49] obtained some preliminary results that depend on much less restrictive conditions on f. The conditions imposed on f are:

1') $\{x \mid f(x) \leq f(x_0)\}$ is bounded, and

2') f has continuous second partial derivatives bounded by the inequality $\| G(x) \|_F \leq \nu$.

The following results can then be derived from these conditions and the conjecture stated below.

There exist functions f, satisfying conditions 1') and 2'), for which the sequence of numbers $\{\| g_k \|_2\}$, $k = 0, 1, \ldots,$ is bounded away from zero. That is, there exists a positive constant μ such that

$$\| g_k \|_2 \geq \mu, \; k = 0, 1, \ldots . \tag{2.20}$$

This conjecture has not been shown to be false for general functions f. Proofs of the lemmas may be found in Section 4 of [49].

<u>Lemma</u> 2.3: There exist positive constants μ_1 and μ_2 such that the trace of T_{k+1}, where $T_{k+1} = H_{k+1}^{-1}$, denoted by $\mathrm{Tr}(T_{k+1})$, is bounded by the inequality

$$\mu_1 \sum_{i=0}^{k} \frac{\| y_i \|_2^2}{s_i^T y_i} \leq \mathrm{Tr}(T_{k+1}) \leq \mu_2 \sum_{i=0}^{k} \frac{\| y_i \|_2^2}{s_i^T y_i}.$$

<u>Lemma</u> 2.4: There exists a constant μ_3 such that $\| H_{k+1} g_{k+1} \|_2$ is bounded by the inequality

$$\| H_{k+1} g_{k+1} \|_2 \leq \mu_3 + \sum_{i=0}^{k} \| s_i \|_2 .$$

<u>Lemma</u> 2.5: There exists a positive constant μ_4 such that the trace of T_{k+1} is bounded by the inequality

$$\frac{\| g_{k+1} \|_2^2}{g_{k+1}^T H_{k+1} g_{k+1}} < \mathrm{Tr}(T_{k+1}) \leq \frac{\mu_4 (k+1)^2}{g_k^T H_k g_k}.$$

<u>Lemma</u> 2.6: There exists a positive constant μ_5 such that

$$k^{3/2} g_k^T H_k g_k < \mu_5 .$$

<u>Lemma</u> 2.7: $\sum_{i=0}^{\infty} \| s_i \|_2$ diverges.

If the conjecture were false, it would follow that the limit points of the sequence $\{x_k\}$, k = 0, 1, ..., generated by the DFP method include at least one stationary point of f. The term "stationary point" must be used instead of "local minimum" because the conditions imposed on f are not sufficient for $g(x') = 0$ to imply that x' is a local minimum. Although some of the consequences of the conjecture given in the above lemmas are surprising, Powell was not able to show that they are contradictory. However, he does show that if the extra condition that f is convex is included, then inequality (2.20) leads to a contradiction. Thus, the DFP method converges for convex functions satisfying conditions 1') and 2'). This is an advance on Theorem 2.6 which requires f to be uniformly convex. The following lemma is also needed to prove the convergence theorem.

Lemma 2.8: If the function f, satisfying conditions 1') and 2') is convex, then the inequality

$$\frac{\| y_k \|_2^2}{s_k^T y_k} \leq \nu, \; k = 0, 1, \ldots,$$

holds, where ν is the bound of condition 2'). That is, for each x, $\| G(x) \|_F \leq \nu$, where the matrix norm is the Frobenius norm induced by the Euclidean vector norm.

Proof: Differentiation gives the equation

$$\frac{d[g(x_k + \phi s_k)]}{d\phi} = G(x_k + \phi s_k)s_k$$

which implies, from the definition of y_k, the identity

$$y_k = \int_0^1 G(x_k + \phi s_k)s_k \, d\phi = [\int_0^1 G(x_k + \phi s_k) \, d\phi]s_k.$$

That is,

$$y_k = \bar{G}_k s_k \tag{2.21}$$

where the ij-th element of the matrix $\bar{G}_k$ is

$$\int_0^1 \frac{\partial^2 f(x_k + \phi s_k)}{\partial \xi_i \partial \xi_j} d\phi .$$

For any vector $w \neq 0$,

$$w^T \bar{G}_k w = \int_0^1 w^T G(x_k + \phi s_k) w \, d\phi \geq 0$$

since f convex implies that

$$w^T G(x_k + \phi s_k) w \geq 0, \; 0 \leq \phi \leq 1.$$

Thus, $\bar{G}_k$ is positive definite or positive semi-definite and therefore it has a square root. Let z_k be the vector

$$z_k = \bar{G}_k^{1/2} s_k .$$

Condition 2') and the definition of $\bar{G}_k$ give the bound $\| \bar{G}_k \|_F \leq \nu$ which implies the inequality

$$\begin{aligned} z_k^T \bar{G}_k z_k &\leq \| z_k \|_2 \| \bar{G}_k z_k \|_2 \\ &\leq \| z_k \|_2 \| \bar{G}_k \|_F \| z_k \|_2 \\ &\leq \nu \| z_k \|_2^2 . \end{aligned}$$

Substituting the definition of z_k in this expression gives

$$s_k^T \bar{G}_k \bar{G}_k s_k \leq \nu s_k^T \bar{G}_k s_k .$$

Then, by using equation (2.21), the inequality

$$y_k^T y_k \leq \nu s_k^T y_k$$

is obtained and, since $s_k^T y_k > 0$, the lemma is proved.

Theorem 2.8: If f is a convex function, having continuous second partial derivatives bounded by the inequality $\| G(x) \|_F \leq \nu$, and if the set $\{x \mid f(x) \leq f(x_0)\}$ is bounded, then if the DFP algorithm is applied to f, the sequence of function values $\{f(x_k)\}$, k = 0, 1, ..., terminates at, or converges to, the least value of f.

Proof: It will be shown first that the conjectured inequality (2.20) gives a contradiction. If this inequality were true, then Lemmas 2.3 and 2.8 would imply

$$\mathrm{Tr}(T_{k+1}) \leq \mu_2 \sum_{i=0}^{k} \frac{\| y_i \|_2^2}{s_i^T y_i} \leq \mu_2 (k+1)\nu$$

and therefore, from inequality (2.20) and Lemma 2.5 the inequality

$$\begin{aligned} \frac{\mu^2}{g_{k+1}^T H_{k+1} g_{k+1}} &\leq \frac{\| g_{k+1} \|_2^2}{g_{k+1}^T H_{k+1} g_{k+1}} \\ &< \mathrm{Tr}(T_{k+1}) \\ &\leq \mu_2 (k+1)\nu \end{aligned}$$

is obtained. This gives the bound

$$g_{k+1}^T H_{k+1} g_{k+1} > \frac{\mu^2}{\mu_2 (k+1)\nu}. \qquad (2.22)$$

However, Lemma 2.6 implies the bound

$$g_{k+1}^T H_{k+1} g_{k+1} < \frac{\mu_5}{(k+1)^{3/2}}$$

which contradicts expression (2.22) when k becomes large. Therefore, the sequence $\{ \| g_k \|_2 \}$, k = 0, 1, ..., is not bounded away from zero, so that the algorithm terminates because some g_k is zero or

$$\liminf_{k \to \infty} \| g_k \|_2 = 0.$$

In the latter case, there exists a subsequence $\{ g_{k_j} \}$, j = 1, 2, ..., such that

$$\operatorname*{limit}_{j \to \infty} g_{k_j} = 0. \qquad (2.23)$$

Because the sequence $\{x_k\}$, k = 0, 1, ..., and hence $\{x_{k_j}\}$, j = 1, 2, .., is in a compact set, namely $\{x \mid f(x) \leq f(x_0)\}$, the subsequence has a limit point, x' say. Without loss of generality, it may be assumed the subsequence $\{x_{k_j}\}$, j = 1, 2, ..., converges to x'. That is,

$$\operatorname*{limit}_{j \to \infty} x_{k_j} = x'. \qquad (2.24)$$

Then, since g is continuous,

$$\operatorname*{limit}_{j \to \infty} g_{k_j} = g(x')$$

and by (2.23), g(x') = 0.

In the other case, if the iterations of the algorithm terminate, it is convenient to also denote by x' the point x_k at which $g_k = 0$. Moreover, f is continuous and the algorithm ensures that the sequence $\{f(x_k)\}$, k = 0, 1, ..., decreases monotonically, so that (2.24) implies

$$\lim_{k \to \infty} f(x_k) = f(x').$$

Since f is convex and $g(x') = 0$, $f(x')$ is the least value of f and the theorem has been proved.

If f is least at only one point, x' say, which is the case if f is strictly convex, then the above theorem implies that the sequence $\{x_k\}$, $k = 0, 1, \ldots,$ converges to x'. However, if it happens that f is least for a set of two or more points, X say, then the theorem implies that every limit point of the sequence is in X.

Numerical Difficulties

The previous theorem guarantees, in theory, the convergence of the DFP method for a restricted class of functions. Knowledge of its behavior on more general functions must be based on numerical experience. Also, the theoretical results assume exact arithmetic which is not possible when implementing the method on a computer. For example, if t significant digits are carried, the product of two numbers will generally require 2t digits for its representation and hence will be represented inexactly. The error introduced by the inexactness of the computer arithmetic operations is called rounding error. In an extensive calculation, rounding errors will accumulate and contaminate the results, possibly to an intolerable degree.

In practice, the DFP algorithm has been generally successful, however numerical difficulties have been reported. Broyden [6] notes that negative steps have to be taken occasionally, implying that some calculated matrices H_k are not positive definite. McCormick and Pearson [36] state that for some problems, the algorithm can get "stuck", that is, changes in the current approximation to the minimum can become negligibly small, and that resetting the matrix H_k to a constant positive definite matrix after every n iterations improves the method's performance. Powell [45] notes that occasionally the slow progress happens when a steepest descent step would cause a substantial decrease in the value of the function. Bard [3] reports encountering similar behavior in his work.

The loss of positive definiteness, contrary to Theorem 2.1, is serious because it suggests that a calculated matrix H_k may happen to be singular, or nearly singular. That is, H_k remains positive definite but one or more of its eigenvalues becomes arbitrarily small and in practical computation, it is then effectively singular. In fact, Bard states that he found his difficulties invariably the result of the matrix turning singular. Broyden [9] shows that the behavior observed by McCormick and Pearson could also be caused by a singular H_k. In the DFP algorithm, $s_k = - \alpha_k H_k g_k$, so that H_{k+1} may be written as

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{\alpha_k^2 H_k g_k g_k^T H_k}{s_k^T y_k} = H_k M_k,$$

where

$$M_k = I - \frac{y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{\alpha_k^2 g_k g_k^T H_k}{s_k^T y_k}.$$

Thus, by induction,

$$H_{k+r} = H_k M_k M_{k+1} \cdots M_{k+r-1}, \; r \geq 1,$$

so that

$$s_{k+r} = - \alpha_{k+r}H_{k+r}g_{k+r} = H_k v, \qquad (2.25)$$

where $v = - \alpha_{k+r}M_k M_{k+1} \cdots M_{k+r-1}g_{k+r}$. Suppose now that H_k is singular, so that $H_k w = 0$ for some nonzero vector w. It follows from (2.25) that, for $r \geq 1$, $w^T s_{k+r} = w^T H_k v = 0$. Hence, once a particular H_k becomes singular, all subsequent steps are orthogonal to some fixed vector and are thus restricted to lie in a subspace of R^n. Unless the minimum also lies in this subspace, and in general it will not, the algorithm is "stuck" in this subspace. This would explain the improvement obtained by periodically resetting H_k to some positive definite matrix, commonly the identity matrix. A nearly singular H_k could also result in the search direction, $d_k = - H_k g_k$, and the negative gradient, $- g_k$, being nearly orthogonal. As illustrated in Figure 9, a minimization in this direction would allow only a small step while a steepest descent step would give a larger decrease in the value of the function.

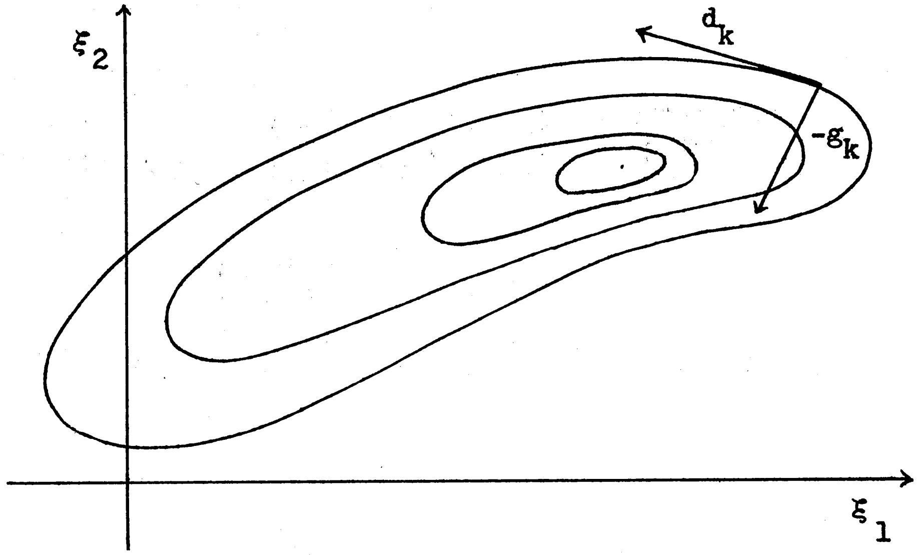

Figure 9. Search Direction d_k Nearly Orthogonal to $- g_k$

Various explanations have been offered for this departure from the theoretical positive definiteness and nonsingularity of H_k. Broyden [6] attributes this reported loss of positive definiteness, and hence stability, to computer rounding error. From his experiments, he concludes that stability depends critically upon the accuracy to which each successive value of α_k is obtained.

Bard [3] shows how poor scaling can cause H_k to become singular. For k = 0, 1, ..., let

$$A_k = \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}, \text{ and } B_k = \frac{s_k s_k^T}{s_k^T y_k}$$

so that $H_{k+1} = H_k - A_k + B_k$. If $H_0 = I$, then the elements of H_0 are of the order of magnitude of unity. If $y_0 = (\eta_1, \ldots, \eta_n)$, then the elements of A_0 are given by

$$\frac{\eta_i \eta_j}{\eta_1^2 + \cdots + \eta_n^2}, \quad i, j = 1, 2, \ldots, n.$$

Using the inequalities, $(\eta_i - \eta_j)^2 \geq 0$ and $(\eta_i + \eta_j)^2 \geq 0$, it can easily be shown that

$$\left| \frac{\eta_i \eta_j}{\eta_1^2 + \cdots + \eta_n^2} \right| < 1. \tag{2.26}$$

Hence, the elements of A_0 are also of the order of magnitude of unity. The matrix B_0 may be expressed

$$B_0 = \left(\frac{s_0^T s_0}{s_0^T y_0}\right)\left(\frac{s_0 s_0^T}{s_0^T s_0}\right).$$

Then, by (2.26) with y_0 replaced by s_0, the magnitude of the elements of B_0 is bounded by $| s_0^Ts_0/s_0^Ty_0 |$. Assuming that the algorithm has not terminated due to $\| g_0 \|_2 \leq \epsilon$ for some given small positive number ϵ,

$$\begin{aligned} | s_0^Ty_0 | &= | s_0^T(g_1 - g_0) | \\ &= | - s_0^Tg_0 | \\ &= \alpha_0 \| g_0 \|_2^2 \\ &> \alpha_0 \epsilon^2, \end{aligned}$$

that is, $| s_0^Ty_0 |$ is bounded away from zero. Since

$$| s_0^Ty_0 | = \| s_0 \|_2 \| y_0 \|_2 \cos \phi,$$

where ϕ is the angle between s_0 and y_0, this implies that $\cos \phi > \sigma$ for some positive constant σ. Thus,

$$\begin{aligned} \left| \frac{s_0^Ts_0}{s_0^Ty_0} \right| &= \frac{\| s_0 \|_2^2}{\| s_0 \|_2 \| y_0 \|_2 \cos \phi} \\ &< \frac{1}{\sigma} \frac{\| s_0 \|_2}{\| y_0 \|_2}. \end{aligned}$$

Hence, the elements of B_0 are of the order of magnitude of $\| s_0 \|_2/\| y_0 \|_2$. Suppose that f is scaled by a factor of δ, a positive constant. This leaves x and s unchanged, but g and y will be scaled by a factor of δ. Thus, all elements of B_0 will be scaled by $1/\delta$. Or, suppose x is scaled by a factor of γ, a positive constant, while the value of f is unchanged, so that the function under consideration is $f(x/\gamma)$. Then s will also be scaled by γ, but g and y will be scaled by

$1/\gamma$. In this case, the elements of B_0 will be scaled by γ^2. Therefore, the magnitude of the elements of B_0 depends on the scales chosen for f and x. In particular, if the scaling is such that $\| y_0 \|_2 >> \| s_0 \|_2$, the elements of B_0 will be very small compared to those of $H_0 - A_0$, so that

$$H_1 \doteq H_0 - A_0 = H_0 - \frac{H_0 y_0 y_0^T H_0}{y_0^T H_0 y_0}.$$

Since $H_1 y_0 = 0$, the matrix H_1 is singular. Conversely, if $\| y_0 \|_2 << \| s_0 \|_2$, the matrix B_0 will dominate $H_0 - A_0$, and so

$$H_1 \doteq B_0 = \frac{s_0 s_0^T}{s_0^T y_0}.$$

Again, H_1 is singular, being of rank one.

Once an H_k has turned singular, there is virtually no hope of recovery. If H_k is singular, it has a null vector z. That is, $H_k z = 0$. Then, as is easily seen from the definition of A_k, both z and y_k will be null vectors of $H_k - A_k$, so that except in the improbable case of z and y_k being linearly dependent, the rank of $H_k - A_k$ will be at most $n - 2$. Since B_k has rank one,

$$\begin{aligned} \text{rank } H_{k+1} &= \text{rank } (H_k - A_k + B_k) \\ &\leq \text{rank } (H_k - A_k) + \text{rank } B_k \\ &\leq n - 1. \end{aligned}$$

Thus, if H_1 is singular, all subsequent H_k are also likely to be singular.

It must be observed that the singularity is only approximate. However, if t significant digits are carried, and if $\| s_0 \|_2/\| y_0 \|_2 = 10^{-t}$ or 10^t, the matrices will be singular to the precision of the calculations. To overcome this problem, Bard recommended using double precision or scaling the variables so that the diagonal elements of B_0 are approximately unity. However, if the character of the function changes drastically from one region to another, then a rescaling of x and reinitialization of H_k whenever the process seems to get stuck at a nonstationary point is suggested.

A nearly singular or poorly scaled H_k can increase the influence of computer rounding errors made when multiplying a vector by this matrix. Let z' be the computed value of a vector z, that is, $z' = z + e$, where e is the error made in computing z. If $w = H_k z$, then w', the computed value of $H_k z$, is given by

$$w' = H_k z' = w + H_k e.$$

Hence, the relative error in this product is given by

$$\frac{\| w' - w \|_2}{\| w \|_2} = \frac{\| H_k e \|_2}{\| w \|_2}$$

To bound this error, note that

$$\| H_k e \|_2 \leq \| H_k \|_2 \| e \|_2$$

and

$$\| z \|_2 = \| H_k^{-1} w \|_2 \leq \| H_k^{-1} \|_2 \| w \|_2$$

which implies that

$$\frac{1}{\| w \|_2} \leq \frac{\| H_k^{-1} \|_2}{\| z \|_2}$$

Therefore,

$$\frac{\| w' - w \|_2}{\| w \|_2} \leq \| H_k \|_2 \| H_k^{-1} \|_2 \frac{\| z' - z \|_2}{\| z \|_2}. \qquad (2.27)$$

This inequality means that the relative error in z may be magnified by as much as

$$\chi(H_k) = \| H_k \|_2 \| H_k^{-1} \|_2$$

when computing $H_k z$. For this reason, $\chi(H_k)$ is called the condition number of H_k, with respect to this operation. If this number is large, then $H_k z$ and $H_k z'$ may differ greatly and the matrix H_k is said to be ill-conditioned. The condition number of a matrix bounds the degree of its ill-conditioning. If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the positive definite matrix H_k, by Corollary 5.2 of [56, p. 308],

$$\| H_k \|_2 = \max \left\{ \lambda_i, \; i = 1, \ldots, n \right\}.$$

Therefore,

$$\chi(H_k) = \frac{\lambda_{max}}{\lambda_{min}},$$

where λ_{max} is the largest eigenvalue of H_k and λ_{min} is the smallest eigenvalue of H_k. Thus, a nearly singular or poorly scaled H_k would be ill-conditioned.

By deriving a recursion formula for the determinant of H_k, Pearson [42] shows directly that H_k tends to become singular when the Hessian matrix G of f is ill-conditioned. The following lemma is needed to establish the effect of a rank two perturbation on the determinant of the identity matrix I. Proof of this lemma is given in Appendix B of [42].

Lemma 2.9: For any vectors u, w, and independent vectors v and z,

$$\det (I + uv^T + wz^T) = (1 + u^Tv)(1 + w^Tz) - (z^Tu)(v^Tw).$$

Since

$$H_{k+1} = H_k\left[I + (-\,y_k)\left(\frac{H_ky_k}{y_k^TH_ky_k}\right)^T + (H_k^{-1}s_k)\left(\frac{s_k}{s_k^Ty_k}\right)^T\right],$$

Lemma 2.9 yields the equation

$$\det H_{k+1} = (\det H_k)\left(\frac{s_k^Ty_k}{y_k^TH_ky_k}\right). \qquad (2.28)$$

Because $\Delta g \doteq G(x)\,\Delta x$ locally, inequality (2.27) implies that, in a region where G is ill-conditioned, a small change in x can cause a large change in g. Thus, it is possible for a small s_k to result in a large y_k, so that $s_k^Ty_k$ could be small and $y_k^TH_ky_k$ large. Then, by recursion formula (2.28), the matrix H_k would rapidly become singular. This type of problem occurs when minimizing a penalty function, that is, when f(x) includes a term to constrain the range of x, because the Hessian matrix at points where one or more constraints are binding is excessively ill-conditioned. Numerical examples given by Pearson indicate that resetting is not beneficial with simple functions but that it is especially valuable for penalty functions.

The best explanation of why the numerical difficulties described can occur with the DFP method is given by Powell [48]. It is based on the following result.

<u>Lemma</u> 2.10: The sequence of numbers $\left\{g_k^T H_k g_k\right\}$, k = 0, 1, ..., generated by the DFP algorithm, decreases strictly monotonically.

<u>Proof</u>: Using the definitions of H_{k+1} and y_k and equation (2.1),

$$g_{k+1}^T H_{k+1} g_{k+1} = g_{k+1}^T \left[H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} \right] g_{k+1}$$

$$= g_k^T \left[H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} \right] g_k . \qquad (2.29)$$

By the definition of y_k and (2.1),

$$g_k^T (H_k y_k y_k^T H_k) g_k = g_k^T (H_k g_k g_k^T H_k) g_k , \text{ and}$$

$$y_k^T H_k y_k = g_{k+1}^T H_k g_{k+1} + g_k^T H_k g_k .$$

Substituting into (2.29) yields

$$g_{k+1}^T H_{k+1} g_{k+1} = g_k^T \left[H_k - \frac{H_k g_k g_k^T H_k}{y_k^T H_k y_k} \right] g_k$$

$$= \frac{(g_k^T H_k g_k)(g_{k+1}^T H_k g_{k+1})}{g_k^T H_k g_k + g_{k+1}^T H_k g_{k+1}} .$$

By inverting both sides of this equation, the identity

$$\frac{1}{g_{k+1}^T H_{k+1} g_{k+1}} = \frac{1}{g_{k+1}^T H_k g_{k+1}} + \frac{1}{g_k^T H_k g_k}$$

is obtained. Then the positive definiteness of H_k implies

$$\frac{1}{g_{k+1}^T H_{k+1} g_{k+1}} > \frac{1}{g_k^T H_k g_k}$$

or equivalently, $g_{k+1}^T H_{k+1} g_{k+1} < g_k^T H_k g_k$, and the lemma is proved.

The decreasing monotonicity of this sequence can be detrimental to the progress of the algorithm, For instance, if an unfortunate choice of the initial matrix H_0 causes $g_0^T H_0 g_0$ to be small, then on every iteration, $g_k^T H_k g_k$ has to be small also. This result supports the importance of the scaling of H_0 expressed by Bard.

Another frequently occurring event can cause $g_k^T H_k g_k$ to be small prematurely. If the function f has a saddle point, that is, a non-optimal stationary point, it may appear to the algorithm to be like a local minimum. In that case, a point, x_j say, would be calculated that is close to the saddle point and therefore g_j is small, and presumably $g_j^T H_j g_j$ is small also. The latter iterations usually cause the sequence of points $\{x_k\}$, $k = j + 1, j + 2, \ldots,$ to leave the saddle point, but nevertheless, the values of $g_k^T H_k g_k$ are forced to be small, due to the smallness of $g_j^T H_j g_j$.

The number $g_k^T H_k g_k$ is important because it is the magnitude of the scalar product of the gradient at x_k and the search direction d_k from the point x_k. If it is small when H_k and g_k are moderate in size, then either the search direction is almost orthogonal to the gradient, or there is much cancellation in the evaluation of the vector $H_k g_k$. Each of these cases can cause difficulty, because the first is a bias away from the direction of steepest descent, and the second increases the influence of computer rounding errors. Moreover, in both cases, the matrix H_k is ill-conditioned.

In general, therefore, the numerical difficulties encountered with the DFP method are related to the condition of the variable matrix H. Development of generalizations of this method then naturally suggest the possibility of choosing the parameter(s) to improve the condition of the corresponding variable matrix H. Indeed, analysis with this goal was done by Broyden [7] and Shanno [53]. Their work will be discussed in Chapter III.

CHAPTER III

ONE-PARAMETER FAMILY

Parametric families of variable metric methods, containing the DFP method as a special case, have been developed from a number of different approaches. These families can be divided into a family containing one parameter and more general families having several parameters. The one-parameter family is the subject of this chapter. The more general families will be discussed in Chapter IV.

This one-parameter family was first developed by Broyden [6] in 1967. The family of correction matrices obtained by Broyden was also developed independently by Shanno [53], Goldfarb [27], and Fletcher [23]. It is particularly interesting that several quite different approaches used by these authors all lead to the development of the same family. In addition, the different developments identify various characteristics of this family of matrices. For these reasons, the development and analysis by each of these authors will be discussed.

Broyden

Broyden's approach to the minimization of f is to use a quasi-Newton method to solve the equation

$$g(x) = 0, \tag{3.1}$$

that is, to find a stationary point of f. Recall that a necessary condition for x' to be a local minimum of a function f having continuous first partial derivatives is that x' is a stationary point. Quasi-Newton methods are iterative methods based on Newton's method for solving a set of nonlinear equations. In this case, the set of equations, equivalent to equation (3.1), to be solved is

$$h_1(x) = 0,\ h_2(x) = 0,\ \ldots,\ h_n(x) = 0,$$

where

$$h_i = \frac{\partial f}{\partial \xi_i},\ i = 1, 2, \ldots, n.$$

If the k-th approximation to the solution is $x_k = (\xi_{k1}, \xi_{k2}, \ldots, \xi_{kn})$ and the (k + 1)-st approximation is $x_{k+1} = (\xi_{k+1,1}, \xi_{k+1,2}, \ldots, \xi_{k+1,n})$ then, for $i = 1, 2, \ldots, n$, the Taylor expansion of h_i about x_k gives

$$h_i(x_{k+1}) \doteq h_i(x_k) + \sum_{j=1}^{n} \frac{\partial h_i}{\partial \xi_j}(x_k)(\xi_{k+1,j} - \xi_{k,j}).$$

This set of equations is equivalent to the matrix equation

$$g(x_{k+1}) \doteq g(x_k) + G(x_k)(x_{k+1} - x_k), \tag{3.2}$$

where g is the gradient vector and G is the Hessian matrix. Since the objective is to find x such that $g(x) = 0$, $g(x_{k+1})$ is set to zero and (3.2) then gives the basic iteration in Newton's method,

$$\begin{aligned} x_{k+1} &= x_k - [G(x_k)]^{-1} g(x_k) \\ &= x_k - G_k^{-1} g_k. \end{aligned}$$

Because this form of the method often fails to converge to a solution from a poor initial estimate, a scalar parameter α_k is sometimes added to give the iteration

$$x_{k+1} = x_k - \alpha_k G_k^{-1} g_k,$$

where α_k is chosen so that $f(x_{k+1}) < f(x_k)$. The disadvantage of evaluating and inverting the second derivative matrix of f at each iteration in Newton's method provides the underlying motivation for the quasi-Newton methods.

In quasi-Newton methods, the inverse Hessian matrix G_k^{-1} is replaced by an approximation H_k, leading to the iteration

$$x_{k+1} = x_k + \alpha_k d_k, \tag{3.3}$$

where $d_k = - H_k g_k$ and α_k is a scalar parameter. This approximation is modified at each iteration so that it possesses, to some extent, the properties of the inverse Hessian matrix. The equation on which this modification is based is derived by considering the special case in which the function f is defined by

$$f(x) = \tfrac{1}{2}x^T Gx + a^T x + \gamma, \tag{3.4}$$

where the matrix G is symmetric and nonsingular. The Hessian matrix of f is G and the gradient of f is $g(x) = Gx + a$. Thus, if s_k and y_k are defined by the equations

$$s_k = x_{k+1} - x_k, \; y_k = g_{k+1} - g_k,$$

then the Hessian matrix G satisfies the equation

$$y_k = (Gx_{k+1} + a) - (Gx_k + a) = Gs_k. \tag{3.5}$$

Since H_k is to approximate G_k^{-1}, it would be desirable for H_k to satisfy the equation $H_k y_k = s_k$. But, y_k depends on g_{k+1} which depends on x_{k+1} which in turn depends on H_k, so this equation cannot be used to determine H_k. However, the next approximation H_{k+1} can be required to satisfy the equation

$$H_{k+1} y_k = s_k. \tag{3.6}$$

Equation (3.6) is called the quasi-Newton equation and is the equation underlying all quasi-Newton methods. However, the quasi-Newton equation is not sufficient to define H_{k+1} or to give any indication of how it may be derived. Since H_k is available and possesses, to some extent, the properties of G_k^{-1}, it seems reasonable to obtain H_{k+1} by adding some correction matrix C_k to H_k, that is,

$$H_{k+1} = H_k + C_k. \tag{3.7}$$

This development of quasi-Newton methods as applied to function minimization is summarized in the following definition.

Definition 3.1: A quasi-Newton method when applied to the minimization of a differentiable function f is an iterative method which generates a sequence $\{x_k\}$, $k = 0, 1, \ldots,$ of approximations to the minimum. At each iteration, given the vector x_k and the matrix H_k, the next approximation is given by (3.3) and the matrix H_k is then updated by (3.7) for some given correction matrix C_k chosen so that the quasi-Newton equation (3.6) is satisfied.

The conditions imposed on the correction matrix C_k in the above definition imply that C_k must satisfy the equation

$$C_k y_k = (H_{k+1} - H_k) y_k$$

$$= s_k - H_k y_k .$$

This equation does not uniquely determine the correction matrix C_k. One general solution of the equation is

$$C_k = s_k q_k^T - H_k y_k z_k^T, \qquad (3.8)$$

where q_k and z_k are arbitrary vectors except for the condition that

$$q_k^T y_k = z_k^T y_k = 1. \qquad (3.9)$$

Some additional criteria are needed to more precisely determine C_k. If a quasi-Newton method is to solve effectively a general set of nonlinear equations it is reasonable to require that it solve a general set of linear equations in a finite number of iterations. For a quasi-Newton method applied to function minimization, this means the method should minimize a quadratic function in a finite number of iterations. Examination of sufficient conditions on C_k to achieve this property leads to Broyden's one-parameter family of correction matrices.

Let r be a positive integer denoting the number of iterations, t a nonnegative integer, and define the matrices

$$Y_r = [y_0, y_1, \ldots, y_{r-1}],$$

$$Z_r = [z_0, z_1, \ldots, z_{r-1}],$$

$$B_{tr} = (I - y_t z_t^T)(I - y_{t+1} z_{t+1}^T) \cdots (I - y_{t+r-1} z_{t+r-1}^T),$$

$$S_r = [s_0, s_1, \ldots, s_{r-1}], \text{ and}$$

$$W_r = [B^T_{1,r-1}q_0, B^T_{2,r-2}q_1, \ldots, B^T_{r-1,1}q_{r-2}, q_{r-1}]. \qquad (3.10)$$

The following sequence of steps is obtained by repeated application of (3.7) with $k = r - 1, r - 2, \ldots, 1, 0$, and C_k given by (3.8).

$$\begin{aligned} H_r &= H_{r-1} - H_{r-1}y_{r-1}z^T_{r-1} + s_{r-1}q^T_{r-1} \\ &= H_{r-1}(I - y_{r-1}z^T_{r-1}) + s_{r-1}q^T_{r-1} \\ &= (H_{r-2} - H_{r-2}y_{r-2}z^T_{r-2} + s_{r-2}q^T_{r-2})(I - y_{r-1}z^T_{r-1}) + s_{r-1}q^T_{r-1} \\ &= H_{r-2}(I - y_{r-2}z^T_{r-2})(I - y_{r-1}z^T_{r-1}) \\ &\qquad + s_{r-2}q^T_{r-2}(I - y_{r-1}z^T_{r-1}) + s_{r-1}q^T_{r-1} \\ &= \cdot\ \cdot\ \cdot \\ &= H_0(I - y_0z^T_0)(I - y_1z^T_1)\cdots(I - y_{r-1}z^T_{r-1}) \\ &\qquad + s_0q^T_0(I - y_1z^T_1)\cdots(I - y_{r-1}z^T_{r-1}) \\ &\qquad + s_1q^T_1(I - y_2z^T_2)\cdots(I - y_{r-1}z^T_{r-1}) \\ &\qquad + \cdot\ \cdot\ \cdot \\ &\qquad + s_{r-2}q^T_{r-2}(I - y_{r-1}z^T_{r-1}) + s_{r-1}q^T_{r-1}. \end{aligned}$$

The definitions in (3.10) then imply

$$H_r = H_0B_{0r} + S_rW^T_r. \qquad (3.11)$$

The first term on the right hand side of this equation consists of H_0 modified by postmultiplication by B_{0r}, and the second term consists solely of information derived from the r iterations and the choice of q_k and z_k, $k = 0, 1, \ldots, r - 1$. Since it is reasonable to require that H_r consists of the latest information derived from the r-th iteration, the first term on the right hand side of (3.11), which represents essentially old information, should tend to the null matrix as r increases. If H_0 is nonsingular, this is achieved if and only if B_{0r} tends to the null matrix as r tends to infinity. A stronger requirement is that B_{0r} becomes the null matrix after a finite number of iterations. It will be shown that B_{0r} cannot be null for $r < n$, and necessary and sufficient conditions for its nullity will be established.

<u>Theorem</u> 3.1: If Y_n, Z_n, and B_{0n} are as defined in (3.10), then the necessary and sufficient condition for B_{0n} to be null is that the matrix $Z_n^T Y_n$ is unit upper triangular, that is,

$$z_k^T y_k = 1, \quad k = 0, 1, \ldots, n - 1,$$

$$z_k^T y_j = 0, \quad 0 \leq j < k \leq n - 1. \qquad (3.12)$$

<u>Proof</u>: If $Z_n^T Y_n$ is unit upper triangular, then, since both Y_n and Z_n are square, Y_n is nonsingular. From the definition of B_{0n} and (3.12), it follows that

$$B_{0n} y_k = (I - y_0 z_0^T)(I - y_1 z_1^T) \cdots (I - y_{n-1} z_{n-1}^T) y_k = 0,$$

for $k = 0, 1, \ldots, n - 1$, that is,

$$B_{0n} Y_n = 0.$$

Therefore, since Y_n is nonsingular, B_{0n} is null, and sufficiency has been proved. If B_{0n} is null, then expansion of the right hand side of the definition of B_{0n} gives

$$\begin{aligned} 0 &= I - y_{n-1}z_{n-1}^T - y_{n-2}z_{n-2}^T(I - y_{n-1}z_{n-1}^T) \\ &\quad - \cdot\cdot\cdot \\ &\quad - y_1z_1^T(I - y_2z_2^T)\cdots(I - y_{n-1}z_{n-1}^T) \\ &\quad - y_0z_0^T(I - y_1z_1^T)\cdots(I - y_{n-1}z_{n-1}^T) \\ &= I - Y_nV_nZ_n^T, \end{aligned} \tag{3.13}$$

where V_n is unit upper triangular. By (3.13), Y_n is nonsingular, so that premultiplication by Y_n^{-1} and postmultiplication by Y_n of (3.13) implies $V_n(Z_n^TY_n) = I$. Since the inverse of the unit upper triangular matrix V_n is itself unit upper triangular, it follows that $Z_n^TY_n$ is unit upper triangular, and necessity has been proved.

Corollary 3.1: B_{0r} cannot be null for $r < n$.

Proof: If $r < n$, then since rank $Y_r \leq r$, there exists a vector $w \neq 0$ such that $w^TY_r = 0$. Since $B_{0r} = I - Y_rV_rZ_r^T$, this implies $w^TB_{0r} = w^T$, and thus B_{0r} is not null, completing the proof of the corollary.

Equation (3.9) and Theorem 3.1 imply that the vector z_k should satisfy the conditions

$$\begin{aligned} &z_k^Ty_k = 1,\ k = 0, 1, \ldots, r - 1, \\ &z_k^Ty_j = 0,\ 0 \leq j < k \leq r - 1, \end{aligned} \tag{3.14}$$

for $1 \leq r \leq p$, $1 \leq p \leq n$. Then, by (3.11) and Theorem 3.1,

$$H_n = H_0B_{0n} + S_nW_n^T = S_nW_n^T.$$

If the function f is defined by (3.4), then the above equation and the desire that H_k approximate G_k^{-1} suggest the possibility of choosing C_k so that the n-th approximation H_n is exactly equal to G^{-1} and leads to the following definition.

<u>Definition</u> 3.2: The quasi-Newton method defined by Definition 3.1 is exact if $H_n = G^{-1}$ when the method is applied to the function f defined by (3.4).

If the quasi-Newton method used to minimize f is exact and the matrix G is positive definite so that the solution of

$$g(x) = Gx + a = 0$$

is the minimum of f, then this minimum will be found in a finite number of iterations since

$$\begin{aligned} x_{n+1} &= x_n - \alpha_n H_n g_n \\ &= x_n - \alpha_n G^{-1}(Gx_n + a) \\ &= - G^{-1}a \end{aligned} \tag{3.15}$$

for $\alpha_n = 1$.

By (3.5) and the definitions of Y_r and S_r given in (3.10), $Y_r = GS_r$ so that

$$S_r = G^{-1}Y_r. \tag{3.16}$$

Hence, if

$$H_rY_r = S_r,\ 1 \le r \le p,\ 1 \le p \le n, \tag{3.17}$$

then $H_nY_n = G^{-1}Y_n$ which would imply, by the nonsingularity of Y_n, that

$$H_n = G^{-1}.$$

By the definition of B_{0r} given in (3.10) and conditions (3.14), for $j = 0, 1, \ldots, r - 1$,

$$B_{0r}y_j = (I - y_0z_0^T)(I - y_1z_1^T)\cdots(I - y_{r-1}z_{r-1}^T)y_j = 0,$$

that is, $B_{0r}Y_r = 0$. Thus, (3.11) implies

$$H_rY_r = (H_0B_{0r} + S_rW_r^T)Y_r$$

$$= S_rW_r^TY_r,\ 1 \le r \le p,\ 1 \le p \le n, \tag{3.18}$$

so that equation (3.17) would be satisfied if

$$W_r^TY_r = I,\ 1 \le r \le p,\ 1 \le p \le n.$$

By the definitions of W_r and Y_r given in (3.10), a simple multiplication shows that $W_r^TY_r$ is the r x r matrix with the ij-th element given by $q_{i-1}^TB_{i,r-i}y_{j-1}$ if $i < r$ and the j-th element in the r-th row given by $q_{r-1}^Ty_{j-1}$. If the vector z_k satisfies conditions (3.14) then, by the definition of $B_{k,r-k}$, for $k = 1, 2, \ldots, r - 1$ and $j = k, k + 1, \ldots, r - 1$,

$$B_{k,r-k}y_j = (I - y_kz_k^T)(I - y_{k+1}z_{k+1}^T)\cdots(I - y_{r-1}z_{r-1}^T)y_j = 0;$$

and for $k = 2, 3, \ldots, r - 1$ and $j = 0, 1, \ldots, k - 1$,

$$B_{k,r-k}y_j = (I - y_kz_k^T)(I - y_{k+1}z_{k+1}^T)\cdots(I - y_{r-1}z_{r-1}^T)y_j = y_j.$$

Hence, the matrix $W_r^T Y_r$ may be expressed as

$$W_r^T Y_r = \begin{bmatrix} q_0^T y_0 & 0 & \cdots & 0 & 0 \\ q_1^T y_0 & q_1^T y_1 & 0 \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ q_{r-2}^T y_0 & q_{r-2}^T y_1 & \cdots & q_{r-2}^T y_{r-2} & 0 \\ q_{r-1}^T y_0 & q_{r-1}^T y_1 & \cdots & q_{r-1}^T y_{r-2} & q_{r-1}^T y_{r-1} \end{bmatrix}$$

so that for $W_r^T Y_r = I$, the vector q_k must satisfy the conditions

$$q_k^T y_k = 1, \; k = 0, 1, \ldots, r - 1,$$

$$q_k y_j = 0, \; 0 \leq j < k \leq r - 1,$$

for $1 \leq r \leq p$, $1 \leq p \leq n$. Then, by (3.18) and (3.16),

$$H_r Y_r = S_r = G^{-1} Y_r, \; 1 \leq r \leq p, \; 1 \leq p \leq n.$$

Combining the conditions which have been placed on the correction matrix C_k leads to the quasi-Newton method:

```
    Given a vector $x_0$ and a nonsingular matrix $H_0$.
    For $k = 0, 1, \ldots, p - 1$,
        set $d_k = - H_k g_k$,
            $s_k = \alpha_k d_k$,
            $x_{k+1} = x_k + s_k$,
            $y_k = g_{k+1} - g_k$,
            $H_{k+1} = H_k + s_k q_k^T - H_k y_k z_k^T$,                    (3.19)
```

where α_k is an arbitrary nonzero scalar, and q_k and z_k are arbitrary vectors except for the conditions

$$q_k^T y_k = z_k^T y_k = 1, \; k = 0, 1, \ldots, p - 1, \tag{3.20}$$

$$q_k^T y_j = z_k^T y_j = 0, \; 0 \leq j < k \leq p - 1, \tag{3.21}$$

where $1 \leq p \leq n$.

The analysis which led to these conditions establishes the following theorem and corollary.

<u>Theorem</u> 3.2: If the quasi-Newton method given by equations (3.19)-(3.21) is applied to the function f defined by (3.4), then

$$H_r Y_r = G^{-1} Y_r, \; 1 \leq r \leq p, \; 1 \leq p \leq n.$$

<u>Corollary</u> 3.2: The quasi-Newton method given by equations (3.19)-(3.21) is exact.

<u>Corollary</u> 3.3: Under the hypotheses of Theorem 3.2,

$$(y_j^T H_r - \alpha_j d_j^T) Y_r = 0, \; 1 \leq r \leq p, \; j = 0, 1, \ldots .$$

<u>Proof</u>: From the theorem and the symmetry of G

$$Y_r^T H_r^T = Y_r^T G^{-1}$$

which implies that

$$Y_r^T (H_r^T G - I) s_j = 0.$$

Since $Gs_j = y_j$, it follows that

$$Y_r^T (H_r^T y_j - s_j) = 0.$$

Hence, by transposing and substituting the definition of s_j, the conclusion of the corollary follows.

By Corollary 3.2, the quasi-Newton method given by equations (3.19)-(3.21) is exact. Hence q_k and z_k must be chosen so that equations (3.20) and (3.21) are satisfied. Because equation (3.20) implies that the quasi-Newton equation is satisfied, (3.20) must be satisfied when the method is applied to any differentiable function f. Assuming that q_k and z_k have been so chosen, then (3.21) must be satisfied only when the method is applied to the function f defined by (3.4) since exactness depends only on properties of the method for this special case. If the vectors q_k and z_k are not chosen specifically to satisfy (3.21) but are chosen in such a way that these equations are satisfied automatically when the method is applied to the function f defined by (3.4), then the method is thus exact. The following theorems establish some further properties of the quasi-Newton method given by equations (3.19)-(3.21) when applied to this function which suggest the vectors q_k and z_k chosen for Broyden's one-parameter family of methods.

<u>Theorem</u> 3.3: If the quasi-Newton method given by equations (3.19)-(3.21) is applied to the function f defined by (3.4) and H_r is symmetric for $1 \leq r \leq p$, then

$$d_{r+1}^T y_j = (\delta_r + \gamma_r \alpha_r) d_r^T y_j, \quad 0 \leq j \leq r - 1,$$

where

$$\delta_r = 1 - \alpha_r(1 + q_r^T g_r), \text{ and } \gamma_r = z_r^T g_r.$$

<u>Proof</u>: From (3.19),

$$\begin{aligned} d_{r+1} &= - H_{r+1}g_{r+1} \\ &= - (H_r + \alpha_r d_r q_r^T - H_r y_r z_r^T)(y_r + g_r) \\ &= d_r\delta_r + H_r y_r \gamma_r . \end{aligned}$$

Thus, from the symmetry of H_r, for $0 \leq j \leq r - 1$,

$$d_{r+1}^T y_j = \delta_r d_r^T y_j + \gamma_r y_r^T H_r y_j,$$

that is,

$$d_{r+1}^T Y_r = (\delta_r d_r^T + \gamma_r y_r^T H_r) Y_r .$$

From Corollary 3.3 with $j = r$, it follows that

$$d_{r+1}^T Y_r = (\delta_r + \gamma_r \alpha_r) d_r^T Y_r,$$

and the theorem is proved.

<u>Corollary</u> 3.4: Under the hypotheses of Theorem 3.3, if $d_{m+1}^T y_m = 0$ for $0 \leq m \leq p - 1$, and H_{j+1} is symmetric for $j = m, m + 1, \ldots, p - 1$, then $d_{j+2}^T y_m = 0$ for $j = m, m + 1, \ldots, p - 1$.

<u>Proof</u>: Repeated application of Theorem 3.3 with $r = m + 1, \ldots, p$ and $j = m$ gives the result.

<u>Theorem</u> 3.4: If the quasi-Newton method given by equations (3.19)-(3.21) is applied to the function f defined by (3.4) and if H_{j+1} is symmetric and $d_{j+1}^T y_j = 0$ for $j = 0, 1, \ldots, p - 1$, then

$$d_k^T y_j = 0, \text{ and}$$

$$y_k^T H_k y_j = 0, \ 0 \leq j < k \leq p. \qquad (3.22)$$

Proof: By Corollary 3.4, $d_{i+2}y_j = 0$ for $i = j, j + 1, \ldots, p - 1$ and $0 \leq j \leq p - 1$. Combining this equation with the hypothesis $d^T_{j+1}y_j = 0$ for $0 \leq j \leq p - 1$ gives the relation

$$d^T_k y_j = 0, \; 0 \leq j < k \leq p,$$

that is, $d^T_k Y_k = 0$. Thus, by Corollary 3.3 with $r = k$ and $j = k$,

$$y^T_k H_k Y_k = \alpha_k d^T_k Y_k = 0,$$

that is,

$$y^T_k H_k y_j = 0, \; 0 \leq j < k \leq p,$$

and the proof is complete.

Corollary 3.5: Under the hypotheses of Theorem 3.4

$$d^T_k G d_j = 0, \; 0 \leq j < k \leq p.$$

Proof: By Theorem 3.2 with $r = k$ and Theorem 3.4,

$$y^T_k G^{-1} Y_k = y^T_k H_k Y_k = 0.$$

Substituting (3.5) and (3.16) with $r = k$ into this equation yields $s^T_k G S_k = 0$, that is,

$$s^T_k G s_j = 0, \; 0 \leq j < k \leq p,$$

which implies, since no α_k is zero,

$$d^T_k G d_j = 0, \; 0 \leq j < k \leq p,$$

completing the proof of the corollary.

If the quasi-Newton method given by equations (3.19)-(3.21) is applied to the function f defined by (3.4) where G is positive definite, then Corollary 3.5 implies that the search directions d_0, d_1, ..., d_{n-1} are conjugate with respect to G provided they are nonzero. Then, as in the DFP method, $g_n = 0$, so that the exact step given by (3.15) will not be taken. Therefore, if H_k, $k = 0, 1, \ldots, n - 1$, is nonsingular, then $d_k = - H_k g_k$ is not zero for $g_k \neq 0$ and the method is thus quadratically terminating.

If the additional hypotheses of Theorem 3.4 are satisfied, then Theorem 3.4 implies that equation (3.21), with p - 1 replaced by p, will be satisfied when the method given by equations (3.19)-(3.21) is applied to the function f defined by (3.4) if the vectors q_k^T and z_k^T are taken to be linear combinations of s_k^T and $y_k^T H_k$. Therefore, the vectors q_k and z_k and the scalar α_k will be chosen so that the additional hypotheses of Theorem 3.4 and equation (3.20) are satisfied. In addition, the vectors q_k and z_k will be defined in terms of the quantities in (3.19) and an arbitrary scalar parameter β_k. The requirements that q_k^T and z_k^T be linear combinations of s_k^T and $y_k^T H_k$ and (3.20) be satisfied lead to the simple choices

$$q_k^T = \frac{s_k^T}{s_k^T y_k}, \text{ and } z_k^T = \frac{y_k^T H_k}{y_k^T H_k y_k}. \qquad (3.23)$$

With this choice, by (3.19),

$$H_{k+1} = H_k + \frac{s_k s_k^T}{s_k^T y_k} - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k}$$

which is the DFP formula. To obtain a family of methods, for which the

DFP method is a special case, a scalar parameter β_k is introduced into equations (3.23), in such a way that q_k^T and z_k^T are more general linear combinations of s_k^T and $y_k^TH_k$ and equation (3.20) remains satisfied. This is accomplished by taking

$$q_k^T = \frac{(1 - \beta_k y_k^T H_k y_k) s_k^T}{s_k^T y_k} + \beta_k y_k^T H_k, \text{ and}$$

$$z_k^T = \frac{(1 - \beta_k s_k^T y_k) y_k^T H_k}{y_k^T H_k y_k} + \beta_k s_k^T.$$

The additional hypotheses of Theorem 3.4 must also be satisfied. If H_k is symmetric, then H_{k+1} is symmetric if C_k is symmetric. By (3.19) and the above choices for q_k^T and z_k^T,

$$\begin{aligned} C_k &= s_k q_k^T - H_k y_k z_k^T \\ &= \frac{(1 - \beta_k y_k^T H_k y_k) s_k s_k^T}{s_k^T y_k} + \beta_k s_k y_k^T H_k - \frac{(1 - \beta_k s_k^T y_k) H_k y_k y_k^T H_k}{y_k^T H_k y_k} - \beta_k H_k y_k s_k^T. \end{aligned}$$

Hence, changing the sign on β_k in q_k^T would make C_k symmetric and q_k would still satisfy (3.20). By the definition of d_{k+1}, the symmetry of H_{k+1}, and the quasi-Newton equation (3.6),

$$d_{k+1}^T y_k = - g_{k+1}^T H_{k+1} y_k = - g_{k+1}^T s_k.$$

Hence, if α_k is chosen to minimize $f(x_k + \alpha d_k)$ with respect to α, then $g_{k+1}^T d_k = 0$ which implies

$$g_{k+1}^T s_k = 0. \tag{3.24}$$

and the remaining hypothesis of Theorem 3.4 is satisfied.

The preceding analysis leads to Broyden's one-parameter family of algorithms given below.

Algorithm 3.1 (Broyden, 1967): Given an initial vector x_0 and an initial symmetric nonsingular matrix H_0.

```
For k = 0, 1, 2, ...,
    If $g_k = g(x_k) = 0$, then stop.
    Else, set $d_k = - H_k g_k$,
        find $\alpha_k$ which minimizes $f(x_k + \alpha d_k)$ with respect to $\alpha$,
        set $s_k = \alpha_k d_k$,
            $x_{k+1} = x_k + s_k$,
            $y_k = g_{k+1} - g_k$,
            $q_k^T = \frac{(1 + \beta_k y_k^T H_k y_k) s_k^T}{s_k^T y_k} - \beta_k y_k^T H_k$,
            $z_k^T = \frac{(1 - \beta_k s_k^T y_k) y_k^T H_k}{y_k^T H_k y_k} + \beta_k s_k^T$,
            $H_{k+1} = H_k + s_k q_k^T - H_k y_k z_k^T$,
        where $\beta_k$ is an arbitrary scalar parameter.
```

Since equations (3.22) were obtained under the assumption that q_k and z_k satisfied (3.20) and (3.21), it can be shown by induction and the discussion leading to Algorithm 3.1 that this algorithm is exact.

Theorem 3.5: Algorithm 3.1 is exact.

Proof: Let the algorithm be applied to the function f defined by (3.4). The theorem will follow from Corollary 3.2 if it is shown that (3.20) and (3.21) are satisfied for $1 \leq p \leq n$. The proof is by induction.

Assume that

$$q_k^T y_k = z_k^T y_k = 1, \quad 0 \leq k \leq p - 1, \tag{3.25}$$

$$q_k^T y_j = z_k^T y_j = 0, \quad 0 \leq j < k \leq p - 1, \tag{3.26}$$

$$H_{k+1}^T = H_{k+1}, \quad 0 \leq k \leq p - 1. \tag{3.27}$$

Since α_j, $0 \leq j \leq p - 1$, is chosen to minimize $f(x_j + \alpha d_j)$ with respect to α, it follows from the definition of d_{j+1}, the symmetry of H_{j+1}, and (3.6) that

$$\begin{aligned} d_{j+1}^T y_j &= - g_{j+1}^T H_{j+1} y_j \\ &= - g_{j+1}^T s_j \\ &= 0, \quad 0 \leq j \leq p - 1. \end{aligned} \tag{3.28}$$

Then, by Theorem 3.4 with $k = p$,

$$y_p^T H_p y_j = d_p^T y_j = 0, \quad 0 \leq j \leq p - 1. \tag{3.29}$$

Hence, for q_p and z_p defined by Algorithm 3.1, (3.29) implies that

$$q_p^T y_j = \frac{(1 + \beta_p y_p^T H_p y_p) s_p^T y_j}{s_p^T y_p} - \beta_p y_p^T H_p y_j = 0, \text{ and}$$

$$z_p^T y_j = \frac{(1 - \beta_p s_p^T y_p) y_p^T H_p y_j}{y_p^T H_p y_p} + \beta_p s_p^T y_j = 0,$$

for $0 \leq j \leq p - 1$. Thus (3.26) is valid with $p - 1$ replaced by p.

By the discussion which led to the definitions of q_k and z_k in Algorithm 3.1, equations (3.25) and (3.27) are valid for all p,

$1 \leq p \leq n$. Thus the induction is complete if (3.26) is valid for $p = 2$, that is, for $j = 0$ and $k = 1$. By (3.28) with $j = 0$,

$$s_1^T y_0 = \alpha_1 d_1^T y_0 = 0$$

so that by Corollary 3.3 with $r = 1$ and $j = 1$,

$$y_1^T H_1 y_0 = \alpha_1 d_1^T y_0 = 0.$$

The definitions of q_1 and z_1 then imply $q_1^T y_0 = z_1^T y_0 = 0$, completing the induction.

The proof of Theorem 3.5 shows that Algorithm 3.1 is a quasi-Newton method of the form given by equations (3.19)-(3.21) when applied to the function f defined by (3.4). Hence the discussion following Corollary 3.5 establishes the following corollary.

<u>Corollary</u> 3.6: Algorithm 3.1 is quadratically terminating provided no H_k, $k = 0, 1, \ldots, n - 1$, is singular or undefined due to a denominator being zero.

To ensure that the algorithm can be applied to an arbitrary differentiable function without breaking down, nonsingularity and nonzero denominators must be guaranteed for all H_k. The denominators in the iteration formula for H_k are $y_k^T H_k y_k$ and $s_k^T y_k$. By the definitions of y_k and s_k and (3.24),

$$s_k^T y_k = s_k^T g_{k+1} - s_k^T g_k = \alpha_k g_k^T H_k g_k. \tag{3.30}$$

If H_k is positive definite then, as in the DFP method, $d_k = - H_k g_k$ is downhill and it is thus always possible to choose α_k positive. Hence it is sufficient to show that H_k is positive definite for all k. The following sufficient condition on β_k for H_k positive definite to imply H_{k+1} positive definite was established by Broyden.

Theorem 3.6: If H_k is positive definite and β_k is nonnegative, then H_{k+1} as given by Algorithm 3.1 is positive definite.

Proof: Since H_k is positive definite, there exists a real nonsingular matrix L such that $H_k = LL^T$. Let c be an arbitrary nonzero vector and define u, v, and w by

$$u = L^T g_k, \quad v = L^T c, \quad w = L^T y_k.$$

Note that u, v, and w are not null if $g_k \neq 0$. Then, using (3.30) and the definition of s_k, the iteration formula for H_k gives

$$c^T H_{k+1} c = v^T v - \frac{(v^T w)^2}{w^T w} + \alpha_k \frac{(u^T v)^2}{u^T u}$$

$$+ \frac{\beta_k \alpha_k}{u^T u w^T w}(u^T u v^T w + w^T w v^T u)^2. \qquad (3.31)$$

Since α_k is positive, if β_k is positive, then the last term on the right hand side of (3.31) is nonnegative, so that if $c^T H_{k+1} c$ is positive for $\beta_k = 0$ then certainly $c^T H_{k+1} c$ is positive for β_k positive. Hence it is sufficient to prove that H_{k+1} is positive definite for $\beta_k = 0$, that is, for the DFP formula. For $\beta_k = 0$, (3.31) becomes

$$c^T H_{k+1} c = v^T v - \frac{(v^T w)^2}{w^T w} + \frac{\alpha_k (u^T v)^2}{u^T u}.$$

Now, by the Schwarz inequality,

$$v^T v - \frac{(v^T w)^2}{w^T w} \geq 0$$

with equality only if v and w are linearly dependent. Furthermore, since α_k is positive,

$$\frac{\alpha_k (u^T v)^2}{u^T u} \geq 0$$

with equality only if u and v are orthogonal. Thus, $c^T H_{k+1} c > 0$ unless $u^T w = 0$. But, by the definitions of y_k and s_k and (3.24),

$$u^T w = g_k^T H_k y_k = - g_k^T H_k g_k$$

which, by the positive definiteness of H_k, is nonzero if $g_k \neq 0$. Since the algorithm is terminated if $g_k = 0$, the proof is complete.

<u>Corollary</u> 3.7: If H_0 is positive definite and β_k is nonnegative for $k = 0, 1, \ldots$, then H_k, $k = 0, 1, \ldots$, is positive definite.

Since $\beta_k = 0$ for each k yields the DFP method, Corollary 3.7 implies Theorem 2.1 and, as in the DFP method, Corollary 3.7 also implies the following corollary.

<u>Corollary</u> 3.8: Algorithm 3.1 is stable if the parameter β_k is chosen to be nonnegative at each iteration.

To simplify notation in the next three sections, the subscript k on the quantities C, H, s, y, g, α, and the parameters will be omitted and the subscript k + 1 will be denoted by the superscript *.

Shanno

Shanno's [53] method of obtaining a family of matrices is similar to Broyden's since it is also based on finding a correction matrix C defined in terms of a scalar parameter which satisfies the equation $Cy = s - Hy$. However, Shanno introduces the parameter τ initially into this equation by the parametric separation

$$Cy = \tau s + [(1 - \tau)s - Hy].$$

By grouping as indicated, this equation yields the solution

$$C = \frac{\tau ss^T}{s^Ty} + \frac{[(1 - \tau)s - Hy][(1 - \tau)s - Hy]^T}{[(1 - \tau)s - Hy]^Ty}.$$

After expanding and regrouping, C may be expressed as

$$C = \left[1 + \frac{(1 - \tau)y^THy}{(1 - \tau)s^Ty - y^THy}\right]\frac{ss^T}{s^Ty} - \left[\frac{1 - \tau}{(1 - \tau)s^Ty - y^THy}\right]sy^TH$$
$$- \left[1 - \frac{(1 - \tau)s^Ty}{(1 - \tau)s^Ty - y^THy}\right]\frac{Hyy^TH}{y^THy} - \left[\frac{1 - \tau}{(1 - \tau)s^Ty - y^THy}\right]Hys^T.$$

This form of C shows that Shanno's one-parameter family of correction matrices is equivalent to Broyden's family. For if

$$\frac{1 - \tau}{(1 - \tau)s^Ty - y^THy} = \beta,$$

that is,

$$\tau = 1 + \frac{\beta y^THy}{1 - \beta s^Ty}, \tag{3.32}$$

then Broyden's correction matrix is obtained.

To provide insight into the significance of the parameter τ, Shanno shows that a particular choice of τ leads to a zero search vector when the gradient is nonzero, that is, a singular H. Consider the case when $\tau = 0$, that is, when C degenerates to the rank one matrix.

$$C = \frac{(s - Hy)(s - Hy)^T}{(s - Hy)^T y}.$$

If $\alpha = 1$, then

$$s - Hy = - Hg - H(g^* - g) = - Hg^*,$$

and, by (3.24),

$$- g^{*T} Hg = g^{*T} s = 0.$$

Hence,

$$C = - \frac{Hg^* g^{*T} H}{g^{*T} H(g^* - g)} = - \frac{Hg^* g^{*T} H}{g^{*T} Hg^*}$$

which implies that

$$d^* = - (H + C)g^* = 0,$$

independent of the magnitude of g^*.

Computation shows that the composition of

$$\hat{H} = H + \frac{\tau ss^T}{s^T y} \qquad (3.33)$$

and

$$H^* = \hat{H} + \frac{(s - \hat{H}y)(s - \hat{H}y)^T}{(s - \hat{H}y)^T y}$$

is identical to the equation

$$H^* = H + \frac{\tau ss^T}{s^T y} + \frac{[(1 - \tau)s - Hy][(1 - \tau)s - Hy]^T}{[(1 - \tau)s - Hy]^T y}. \qquad (3.34)$$

Thus, the value of τ for which the search direction $d^* = 0$ is that value for which (3.33) gives $\hat{\alpha} = 1$. Using the definitions of y and s and (3.24), equation (3.33) gives

$$\hat{H}g = Hg - \frac{\tau \alpha Hg(s^T g)}{s^T(g^* - g)} = (1 + \tau\alpha)Hg$$

which implies that $\hat{\alpha} = 1$ if $1 + \tau\alpha = \alpha$. Thus, if $\tau = (\alpha - 1)/\alpha$, then $d^* = - H^*g^* = 0$ so that H^* is singular if $g^* \neq 0$.

Shanno further restricts the choice of τ by the following theorem which shows how positive definiteness of the variable matrix depends on the choice of the parameter τ. Proof of this theorem is given as proof of Theorem 2 in [53]. It will also follow from a more general theorem which will be established in this section.

Theorem 3.7: If H is positive definite and $\tau > (\alpha - 1)/\alpha$, then H^* given by (3.34) is positive definite.

Theorem 3.7 establishes that the condition on the parameter τ, $\tau > (\alpha - 1)/\alpha$, is sufficient for H positive definite to imply H^* positive definite. Thus, if H_0 is positive definite, then the method is stable. By a further analysis of this family of methods as developed by Shanno, Shanno and Kettler [54] derive necessary and sufficient conditions on the range of the parameter τ to guarantee stability of the method. The following theorem used in establishing these conditions is significant in itself because it shows that the parameter affects only the length, not the direction, of the search vector at each iteration.

Theorem 3.8: The search direction $d^* = - H^*g^*$ can be represented as $d^* = h(\tau)$, where $h(\tau)$ is a scalar function of τ and r is a vector independent of τ. In particular,

$$d^* = - h(\tau)(\delta Hg^* + \gamma Hg),$$

where

$$\delta = g^T Hg, \ \gamma = g^{*T}Hg^*, \text{ and}$$

$$h(\tau) = \frac{(\alpha\tau - \alpha + 1)}{(\alpha\tau - \alpha + 1)\delta + \gamma}.$$

Proof: Using the definitions of y and d and (3.24), equation (3.34) yields

$$H^*g^* = Hg^* - \frac{(g^{*T}Hg^*)[(1 - \tau)s - Hy]}{[(1 - \tau)s - Hy]^T y}. \qquad (3.35)$$

By the definitions of s and y, the denominator can be expressed as

$$[(1 - \tau)s - Hy]^T y = (- 1 + \alpha - \alpha\tau)g^T Hg - g^{*T}Hg^*.$$

Substituting this expression into (3.35) and combining the two terms on the right hand side gives, using the definitions of s and y,

$$H^*g^* = \frac{-(\alpha\tau - \alpha + 1)\delta Hg - (\alpha\tau - \alpha + 1)\gamma Hg}{-(\alpha\tau - \alpha + 1)\delta - \gamma}$$

$$= \left[\frac{\alpha\tau - \alpha + 1}{(\alpha\tau - \alpha + 1)\delta + \gamma}\right](\delta Hg^* + \gamma Hg). \tag{3.36}$$

Since d* = - H*g*, this establishes the theorem.

To prove that h(τ) > 0 is a necessary and sufficient condition for H positive definite to imply H* positive definite, the following lemma is required.

Lemma 3.1: H positive definite implies $g^{*T}H^*g^* > 0$ if and only if h(τ) > 0.

Proof: Premultiplying (3.36) by g^{*T} and applying (3.24) gives

$$g^{*T}H^*g^* = h(\tau)\delta\gamma. \tag{3.37}$$

Since H is positive definite, δ and γ are positive unless either g or g* is zero, at which point the algortihm is terminated. Thus, by (3.37), the lemma is proved.

Theorem 3.9: If H is positive definite, H* is positive definite if and only if h(τ) is positive.

Proof: Any set of n vectors which are conjugate with respect to the positive definite matrix H are linearly independent and hence form a basis for R^n. Since $g \neq 0$, $g^* \neq 0$, and $g^THg^* = 0$, let g, g*, and any n - 2 vectors $z_1, \ldots, z_{n-2}$ which are conjugate with respect to H and which satisfy $z_i^THg = 0$ and $z_i^THg^* = 0$, $i = 1, \ldots, n - 2$, be a basis for

R^n. From (3.34), the conjugacy of these vectors implies

$$z_i^T H^* z_i = z_i^T H z_i, \quad z_i^T H^* z_j = 0,$$

$$z_i^T H^* g = 0, \text{ and } z_i^T H^* g^* = 0. \tag{3.38}$$

for $i \neq j$, i, $j = 1, \ldots, n - 2$. Let w be an arbitrary nonzero vector. Since $z_1, \ldots, z_{n-2}, g, g^*$ form a basis, w can be expressed as

$$w = \sum_{i=1}^{n-2} \mu_i z_i + \mu_{n-1} g + \mu_n g^*$$

for some scalars μ_i, $i = 1, \ldots, n$. Then, by (3.38),

$$w^T H^* w = \left[\sum_{i=1}^{n-2} \mu_i z_i^T + \mu_{n-1} g^T + \mu_n g^{*T} \right] H^* \left[\sum_{i=1}^{n-2} \mu_i z_i + \mu_{n-1} g + \mu_n g^* \right]$$

$$= \sum_{i=1}^{n-2} \mu_i^2 z_i^T H z_i + \mu_{n-1}^2 g^T H^* g + 2\mu_{n-1}\mu_n g^T H^* g^* + \mu_n^2 g^{*T} H^* g^*. \tag{3.39}$$

Using the definition of y and the quasi-Newton equation (3.6), (3.24) implies

$$g^{*T} H^* g^* = g^{*T} H^*(y + g) = g^{*T} H^* g$$

and the definition of s gives

$$g^T H^* g = g^T H^*(g^* - y) = g^T H^* g^* + \alpha g^T H g.$$

Hence, (3.39) becomes

$$w^T H^* w = \sum_{i=1}^{n-2} \mu_i^2 z_i^T H z_i + \mu_{n-1}^2 \alpha g^T H g + (\mu_{n-1}^2 + 2\mu_{n-1}\mu_n + \mu_n^2) g^{*T} H^* g^*.$$

Since H is positive definite, $w^T H^* w$ is positive if and only if $g^{*T} H^* g^*$ is positive. Therefore, the theorem follows from Lemma 3.1.

From the definition of $h(\tau)$ given in Theorem 3.8, it follows that if H is positive definite, $h(\tau)$ is positive if and only if $\tau > (\alpha - 1)/\alpha$ or $\tau < (\alpha - 1)/\alpha - \gamma/\alpha\delta$. Thus, Theorem 3.9 establishes Theorem 3.7. Theorem 3.6 proves that the positive definiteness of H is retained if β is nonnegative. Theorem 3.9 extends this range. By (3.32), $\tau > (\alpha - 1)/\alpha$ if and only if

$$\frac{\beta y^T Hy}{1 - \beta s^T y} > - \frac{1}{\alpha}. \tag{3.40}$$

Since

$$s^T y + \alpha g^T Hg \text{ and } y^T Hy = g*^T Hg* + g^T Hg, \tag{3.41}$$

if H is positive definite and $1 - \beta s^T y$ is positive, then (3.40) implies $\beta > - 1/(\alpha g*^T Hg*)$. Similarly, if H is positive definite and $1 - \beta s^T y$ is negative, then (3.40) implies $\beta < - 1/(\alpha g*^T Hg*)$. Therefore, if H is positive definite then

$$\tau > \frac{\alpha - 1}{\alpha} \text{ if and only if } - \frac{1}{\alpha g*^T Hg*} < \beta < \frac{1}{s^T y}. \tag{3.42}$$

By (3.32), $\tau < (\alpha - 1)/\alpha - \gamma/\alpha\delta$ if and only if

$$\frac{\beta y^T Hy}{1 - \beta s^T y} < - \frac{1}{\alpha} - \frac{\gamma}{\alpha\delta}$$

which, by (3.41) and the definitions of δ and γ, is equivalent to

$$\beta > \frac{1}{s^T y}. \tag{3.43}$$

Thus, by combining the results of Theorems 3.6 and 3.9, the following corollary extending the range of β for which retention of positive definiteness is guaranteed, is established.

Corollary 3.9: If H is positive definite and $\beta > -\mu$, where μ is the positive number given by $\mu = 1/(\alpha g^{*T}Hg^{*})$, then H* is positive definite.

Theorem 3.9 also shows that this range of β for which H is positive definite implies H* is positive definite cannot be extended. If H is positive definite then H* positive definite implies, by Theorem 3.9, (3.42) and (3.43), that

$$-\frac{1}{\alpha g^{*T}Hg^{*}} < \beta < \frac{1}{s^Ty} \text{ or } \beta > \frac{1}{s^Ty}$$

which implies $\beta > -\mu$.

Goldfarb

Goldfarb [27] develops a one-parameter family of variable metric methods from a combination of two correction matrices belonging to a family derived by Greenstadt [28] using a variational approach. As did Broyden and Shanno, Greenstadt wishes to find a correction C to the estimate H of the inverse Hessian matrix so that the quasi-Newton equation is satisfied. Since C is not uniquely determined by this condition, Greenstadt chooses to look for the "best" correction C. In particular, he wishes to find the smallest correction C in the sense of some norm, because this would tend to keep the elements of H from growing too large, which might cause difficulty.

The norm chosen should be simple and lead to simple solutions for C. These criteria suggest a simple quadratic form in the elements of

C, that is,

$$\| C \|_F^2 = \sum_{i,j=1}^{n} \gamma_{ij}^2,$$

where γ_{ij} represents the ij-th element of C. Because minimizing $\| C \|_F$ is equivalent to minimizing $\| C \|_F^2$ and $\sum_{i,j} \gamma_{ij}^2 = \mathrm{Tr}(CC^T)$, the problem is to minimize N(C), where $N(C) = \mathrm{Tr}(CC^T)$. However, this is too specialized, so C is transformed to

$$C' = ACA^T,$$

where A is a nonsingular matrix. Then, by the properties of the trace,

$$\begin{aligned} N(C') &= \mathrm{Tr}[(ACA^TAC^T)A^T] \\ &= \mathrm{Tr}[A^T(ACA^TAC^T)] \\ &= \mathrm{Tr}(WCWC^T), \end{aligned}$$

where W is the positive definite matrix A^TA. Thus, the problem is to find the symmetric correction matrix C which minimizes $\mathrm{Tr}(WCWC^T)$ subject to the quasi-Newton equation. The symmetry condition, which will preserve the symmetry of H if the initial matrix H_0 is symmetric, is required because the Hessian matrix is symmetric if the function f has continuous second partial derivatives. The variational formulation of this problem is

$$\begin{aligned} &\underset{C}{\text{minimize}}\ \mathrm{Tr}(WCWC^T), \\ &\text{subject to } Cy - r = 0, \text{ and} \\ &\qquad\qquad C^T - C = 0, \end{aligned}$$

where r = s - Hy.

This constrained minimization problem will be solved by the use of Lagrange multipliers. Denote the matrix C and the vectors y and r by

$$C = (\gamma_{ij}),\ i,\ j = 1,\ \ldots,\ n$$

$$y = (\eta_1,\ \ldots,\ \eta_n)^T,\ \text{and}$$

$$r = (\rho_1,\ \ldots,\ \rho_n)^T.$$

Then the constraints are equivalent to

$$\sum_{j=1}^{n} \gamma_{ij}\eta_j - \rho_i = 0,\ i = 1,\ \ldots,\ n,\ \text{and}$$

$$\gamma_{ji} - \gamma_{ij} = 0,\ i,\ j = 1,\ \ldots,\ n.$$

Thus, the composite function Φ is formed as

$$\Phi = \tfrac{1}{2}\mathrm{Tr}(WCWC^T) + \sum_{i=1}^{n} \pi_i\Big(\sum_{j=1}^{n} \gamma_{ij}\eta_j - \rho_i\Big) + \sum_{i,j=1}^{n} \delta_{ij}(\gamma_{ji} - \gamma_{ij})$$

which may be expressed

$$\begin{aligned}\Phi &= \tfrac{1}{2}\mathrm{Tr}(WCWC^T) + \mathrm{Tr}[(Cy - r)p^T] + \mathrm{Tr}[D(C - C^T)] \\ &= \tfrac{1}{2}\mathrm{Tr}(C^TWCW) + \mathrm{Tr}(Cyp^T) - \mathrm{Tr}(rp^T) + \mathrm{Tr}(CD) - \mathrm{Tr}(C^TD),\end{aligned}$$

where the multipliers are $p = (\pi_1,\ \ldots,\ \pi_n)$ and $D = (\delta_{ij})$, $i,\ j = 1,\ \ldots,\ n$. The use of $\tfrac{1}{2}\mathrm{Tr}(WCWC^T)$ instead of $\mathrm{Tr}(WCWC^T)$ is for computational ease and does not affect the result. In order to differentiate Φ with respect to C, note that for any matrix $A = (\alpha_{ij})$, $i,\ j = 1,\ \ldots,\ n$, the partial with respect to C of Tr(CA) is the matrix whose km-th element is given by

$$\frac{\partial}{\partial\gamma_{km}}\sum_{i,j=1}^{n} \gamma_{ij}\alpha_{ji} = \alpha_{mk}.$$

That is,

$$\frac{\partial}{\partial C}\mathrm{Tr}(CA) = A^T.$$

Similarly,

$$\frac{\partial}{\partial C}\mathrm{Tr}(C^TA) = A.$$

Also, if $WCW = (\beta_{ij})$ and $W = (\omega_{ij})$, $i, j = 1, \ldots, n$, the km-th element of the partial with respect to C of $\frac{1}{2}\mathrm{Tr}(C^T(WCW))$ is given by

$$\frac{1}{2}\frac{\partial}{\partial\gamma_{km}}\sum_{i,j=1}^{n}\gamma_{ji}\beta_{ji}.$$

Application of the rule for differentiating products and substitution for the element β_{ji} then yields the expression

$$\frac{1}{2}\left\{\sum_{i,j=1}^{n}\gamma_{ji}\frac{\partial}{\partial\gamma_{km}}\sum_{q,p=1}^{n}\omega_{jq}\gamma_{qp}\omega_{pi} + \sum_{i,j=1}^{n}\beta_{ji}\frac{\partial}{\partial\gamma_{km}}\gamma_{ji}\right\}.$$

By taking the indicated partials, this expression reduces to

$$\tfrac{1}{2}\left(\sum_{i,j=1}^{n}\gamma_{ji}\omega_{jk}\omega_{mi} + \beta_{km}\right).$$

Then, noting that W is symmetric and that $\beta_{km} = \sum_{i,j}\omega_{kj}\gamma_{ji}\omega_{im}$, the above expression reduces to β_{km}. Therefore,

$$\frac{\partial}{\partial C}\tfrac{1}{2}\mathrm{Tr}(C^TWCW) = WCW.$$

Thus, using the above relations,

$$\frac{\partial\Phi}{\partial C} = WCW + py^T + D^T - D,$$

and setting this partial derivative equal to zero implies that

$$C = - M(py^T + D^T - D)M, \qquad (3.44)$$

where $M = W^{-1}$. The multipliers p and D must now be eliminated from this expression for C. The constraint $C^T - C = 0$ gives

$$- M(yp^T - py^T + 2D - 2D^T)M = 0$$

which implies

$$D^T - D = \tfrac{1}{2}(yp^T - py^T).$$

Substituting this expression into (3.44) yields

$$C = - M[py^T + \tfrac{1}{2}(yp^T - py^T)]M$$

$$= - \tfrac{1}{2}M(yp^T + py^T)M. \qquad (3.45)$$

Then the constraint $Cy - r = 0$ implies

$$- \tfrac{1}{2}M(yp^T + py^T)My - r = 0.$$

Premultiplying by $- 2W$ gives

$$(yp^T + py^T)My + 2Wr = 0$$

and then solving for the p which is free from the inner product yields

$$p = - [2Wr + y(p^TMy)]/(y^TMy). \qquad (3.46)$$

Premultiplying this expression by y^TM gives

$$y^TMp = - [2y^Tr + (y^TMy)(p^TMy)]/(y^TMy).$$

Since $y^TMp = p^TMy$, this equation may be solved for p^TMy, getting

$$p^TMy = - (y^Tr)/(y^TMy).$$

Substituting this expression back into (3.46) yields

$$p = - (y^TMy)^{-1}[2Wr - (y^TMy)^{-1}(y^Tr)y].$$

Thus, substituting this expression for p into (3.45) completes the elimination of the multipliers in C, giving

$$C = \frac{1}{y^TMy}\left[ry^TM + Myr^T - \left(\frac{y^Tr}{y^TMy}\right)Myy^TM\right].$$

Finally, replacing r by s - Hy gives the solution

$$C = \frac{1}{y^TMy}\left[sy^TM + Mys^T - Hyy^TM - Myy^TH - \frac{1}{y^TMy}(s^Ty - y^THy)Myy^TM\right]. \tag{3.47}$$

One obvious choice for the weighting matrix W which will lead to a relatively simple formula for C is $W^{-1} = M = H$. The result, denoted C_H, is

$$C_H = \frac{1}{y^THy}\left[sy^TH + Hys^T - \left(1 + \frac{s^Ty}{y^THy}\right)Hyy^TH\right]$$

which resembles, to some extent, the DFP correction matrix. In fact, the resemblance between these two correction matrices goes deeper than mere appearance. It is shown by Bard in the appendix of [28] that the variable metric method using C_H is also quadratically terminating and exact. The proof follows exactly the argument presented by Fletcher and

Powell for the DFP method. However, the correction matrix C_H does not preserve the positive definiteness of H as does the DFP correction matrix, since numerical experiments by Greenstadt show that is was frequently necessary to take a negative step in order to make f decrease.

Goldfarb obtains a variable metric method which, in addition to being quadratically terminating and exact, preserves the positive definiteness of the variable matrix by using the correction matrix obtained by substituting H* for M in (3.47). Using the quasi-Newton equation, this correction matrix, denoted C_{H*}, can be expressed as

$$C_{H*} = \frac{1}{s^T y}\left[- sy^T H - Hys^T + \left(1 + \frac{y^T Hy}{s^T y} \right) ss^T \right].$$

To show that the variable metric method with correction matrix C_{H*} is quadratically terminating and exact, Bard's proof may be followed almost entirely, except for some obvious and trivial changes. Proof that $H^* = H + C_{H*}$ is positive definite if H is positive definite follows from observing that H* may be expressed as

$$H^* = (H + C_{DFP}) + (C_{H*} - C_{DFP}),$$

where C_{DFP} is the DFP correction matrix,

$$C_{DFP} = \frac{ss^T}{s^T y} - \frac{Hyy^T H}{y^T Hy}.$$

And, for an arbitrary nonzero vector w, the definitions of C_{H*} and C_{DFP} give

$$w^T(C_{H*} - C_{DFP})w = \frac{[(y^T Hy)(w^T s) - (s^T y)(w^T Hy)]^2}{(s^T y)^2 (y^T Hy)} \geq 0 \qquad (3.48)$$

Thus, by Theorem 2.1 and (3.48), H* is positive definite since it is the sum of a positive definite matrix and a positive semi-definite matrix.

The two variationally derived correction matrices C_H and C_{H*} are combined by Goldfarb to obtain the one-parameter family of correction matrices

$$C = \gamma C_H + (1 - \gamma)C_{H*} \tag{3.49}$$

By substituting the given expressions for C_H and C_{H*}, this family may be expressed as

$$C = \left[1 - \gamma + \frac{(1 - \gamma)y^THy}{s^Ty}\right]\frac{ss^T}{s^Ty} - \left[\frac{-\gamma}{y^THy} + \frac{1 - \gamma}{s^Ty}\right]sy^TH$$

$$- \left[1 + \frac{\gamma(s^Ty)}{y^THy} - 1 + \gamma\right]\frac{Hyy^TH}{y^THy} - \left[\frac{-\gamma}{y^THy} + \frac{1 - \gamma}{s^Ty}\right]Hys^T.$$

By setting

$$\frac{-\gamma}{y^THy} + \frac{1 - \gamma}{s^Ty} = \beta,$$

that is,

$$\gamma = \frac{(1 - \beta s^Ty)y^THy}{y^THy + s^Ty}$$

it is clear that this family is also equivalent to Broyden's family.

As noted in the first section of this chapter, if $\beta = 0$, then the DFP correction matrix is obtained. Thus, if

$$\gamma = \frac{y^THy}{y^THy + s^Ty},$$

by (3.49), the DFP correction matrix can be expressed directly as a weighted sum of C_H and C_{H*}, namely as

$$C_{DFP} = \frac{(y^THy)C_H + (s^Ty)C_{H*}}{y^THy + s^Ty}.$$

It is also possible to obtain C_{DFP} directly from (3.47) by choice of a suitable M. Goldfarb gives several forms of M, all of which may be shown by substitution to give the DFP correction matrix. One example is

$$M = (y^THy)^{\frac{1}{2}}H* - (s^Ty)^{\frac{1}{2}}H.$$

Although the given matrices M are, in general, nonsingular, they and hence, the corresponding $W = M^{-1}$ are not necessarily positive definite. Thus, their substitution in (4.33) is somewhat contrived so that their role in the variational derivation of the DFP method is not clear.

Fletcher

Fletcher [23] generates a class of updating formulae for the variable matrix H by taking any linear combination of the DFP updating formula and a new formula, such that the coefficients sum to unity. This new formula is based upon a very simple idea. The DFP formula forces the relationship $H*y = s$ to hold. If $T = H^{-1}$ and $T*^{-1} = H*^{-1}$, then by applying Householder's modification rule twice sequentially, as in the proof of Theorem 3.13, to the DFP formula,

$$H^* = H + \frac{ss^T}{s^Ty} - \frac{Hyy^TH}{y^THy}, \qquad (3.50)$$

T and T* corresponding to H and H* of the DFP formula are related by

$$T^* = T - \frac{ys^TT}{s^Ty} - \frac{Tsy^T}{s^Ty} + \left(1 + \frac{s^TTs}{s^Ty}\right)\frac{yy^T}{s^Ty} \qquad (3.51)$$

Since T*s = y, (3.51) gives a mapping of s into y. By the simple interchange of s and y in this equation, a formula is obtained which maps y into s. Thus, the equation

$$H^* = H - \frac{sy^TH}{s^Ty} - \frac{Hys^T}{s^Ty} + \left(1 + \frac{y^THy}{s^Ty}\right)\frac{ss^T}{s^Ty} \qquad (3.52)$$

could be used as a formula to update H. If H is updated by (3.52), then the corresponding updating formula for T is obtained by performing the interchange of s and y in the DFP formula, that is,

$$T^* = T + \frac{yy^T}{s^Ty} - \frac{Tss^TT}{s^TTs}. \qquad (3.53)$$

Thus, the formulae (3.50), (3.51) and (3.52), (3.53) may be considered as dual in this sense. Equation (3.52) is also called the complementary DFP formula. In addition, the correction matrix in this formula is identical to Goldfarb's correction matrix C_{H*}.

Denoting the H* in (3.50) and (3.52) by H^*_{DFP} and $H^*_{DFP'}$, respectively, Fletcher's class of formulae is given by

$$H^* = (1 - \phi)H^*_{DFP} + \phi H^*_{DFP'}. \qquad (3.54)$$

Substituting for H^*_{DFP} and H^*_{DFP}, gives

$$H^* = H + (1 - \phi)\left[\frac{ss^T}{s^Ty} - \frac{Hyy^TH}{y^THy}\right] + \phi\left[-\frac{sy^TH}{s^Ty} - \frac{Hys^T}{s^Ty} + \left(1 + \frac{y^THy}{s^Ty}\right)\frac{ss^T}{s^Ty}\right]$$

$$= H + \left(1 + \frac{\phi y^THy}{s^Ty}\right)\frac{ss^T}{s^Ty} - \frac{\phi}{s^Ty}(sy^TH + Hys^T) - (1 - \phi)\frac{Hyy^TH}{y^THy}$$

which shows that this class is also equivalent to Broyden's family through the relationship

$$\phi = \beta s^Ty.$$

An important new result given by Fletcher is that (3.54) can be rearranged as

$$H^* = H^*_{DFP} + \phi(y^THy)\left[\frac{ss^T}{(s^Ty)^2} - \frac{sy^TH}{(s^Ty)(y^THy)} - \frac{Hys^T}{(s^Ty)(y^THy)} + \frac{Hyy^TH}{(y^THy)^2}\right]$$

$$= H^*_{DFP} + \phi vv^T, \qquad (3.55)$$

where

$$v = (y^THy)^{\frac{1}{2}}\left[\frac{s}{s^Ty} - \frac{Hy}{y^THy}\right].$$

Thus, the difference between any two formulae $H^*_{\phi_1}$ and $H^*_{\phi_2}$ in the class is given by

$$(\phi_1 - \phi_2)vv^T$$

which is a matrix of rank one. Equivalently, any formula H^*_{ϕ} in the class differs from the DFP formula by the rank one matrix ϕvv^T. In particular, the rank one property enables the following lemma to be applied. This lemma is established in [58, pp. 94-98].

Lemma 3.2: If $A' = A + \sigma ww^T$, $\sigma = \pm 1$, and if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the eigenvalues of A and $\lambda_1' \geq \lambda_2' \geq \cdots \geq \lambda_n'$ are the eigenvalues of A', then

i) if $\sigma = +1$, $\lambda_1' \geq \lambda_1 \geq \lambda_2' \geq \lambda_2 \geq \cdots \geq \lambda_n' \geq \lambda_n$, and

ii) if $\sigma = -1$, $\lambda_1 \geq \lambda_1' \geq \lambda_2 \geq \lambda_2' \geq \cdots \geq \lambda_n \geq \lambda_n'$.

By (3.54) and (3.55),

$$H^*_{DFP'} = H^*_{DFP} + vv^T$$

so that Lemma 3.2 implies

$$\det H^*_{DFP'} = \prod_{i=1}^{n} \lambda_i' \geq \prod_{i=1}^{n} \lambda_i = \det H^*_{DFP}.$$

Thus, $H^*_{DFP'}$ is "less singular" than H^*_{DFP}, indicating that the use of $H^*_{DFP'}$ in a variable metric algorithm might counteract the tendency toward singularity of H^*_{DFP} discussed in Chapter II.

Choice of Parameter

Broyden's one-parameter family of variable metric methods contains the DFP method as a special case. In addition, this family possesses the important properties, quadratic termination, exactness, and stability for $\beta \geq 0$, of the DFP method. It was shown in the last section of Chapter II that the numerical difficulties encountered with the DFP method are related to the condition of the variable matrix H. Thus, this family of methods offers the possibility of choosing the parameter β to improve the condition of the corresponding variable matrix H, while still retaining the desirable characteristics of the DFP method.

In Algorithm 3.1, the matrix H_i, $i = 0, 1, \ldots,$ is updated at each iteration by the equation

$$H_{i+1} = H_i + \mu_i s_i s_i^T - \beta_i s_i y_i^T H_i - \nu_i H_i y_i y_i^T H_i - \beta_i H_i y_i s_i^T, \qquad (3.56)$$

where

$$\mu_i = \frac{1 + \beta_i y_i H_i y_i}{s_i^T y_i}, \text{ and } \nu_i = \frac{1 - \beta_i s_i^T y_i}{y_i^T H_i y_i}. \qquad (3.57)$$

Since this updating equation is somewhat complicated, Broyden [7] analyzes its properties when the function to be minimized is a strictly convex quadratic function, that is, the function f is given by (3.4), where G is positive definite. He also transforms the problem so that the inverse Hessian matrix G^{-1} is the identity matrix and the approximation to it is

$$K_i = G^{\frac{1}{2}} H_i G^{\frac{1}{2}}.$$

Broyden's suggestion of a value for β_i to obtain an algorithm having better numerical properties is the result of examining the updating procedure and the dependence of the matrix sequence $\{K_i\}$, $i = 0, 1, \ldots,$ upon the parameters β_i.

By (3.56) and (3.5), the updating procedure for K_i is given by

$$\begin{aligned} K_{i+1} = K_i &+ \mu_i G^{\frac{1}{2}} s_i s_i^T G^{\frac{1}{2}} - \beta_i G^{\frac{1}{2}} s_i s_i^T G H_i G^{\frac{1}{2}} \\ &- \nu_i G^{\frac{1}{2}} H_i G s_i s_i^T G H_i G^{\frac{1}{2}} - \beta_i G^{\frac{1}{2}} H_i G s_i s_i^T G^{\frac{1}{2}}. \end{aligned} \qquad (3.58)$$

Since $x' = - G^{-1} a$ is the minimum of f, the error at x_i is given by

$e_i = x_i - x'$, so that the gradient at x_i is given by

$$g_i = Gx_i + a = Ge_i. \tag{3.59}$$

Then, using the definition of s_i,

$$G^{\frac{1}{2}}s_i = -\alpha_i(G^{\frac{1}{2}}H_iG^{\frac{1}{2}})(G^{\frac{1}{2}}e_i) = -\alpha_iK_iz_i \tag{3.60}$$

and

$$G^{\frac{1}{2}}H_iGs_i = (G^{\frac{1}{2}}H_iG^{\frac{1}{2}})(G^{\frac{1}{2}}s_i) = -\alpha_iK_i^2z_i,$$

where $z_i = G^{\frac{1}{2}}e_i$, so that (3.58) becomes

$$\begin{aligned} K_{i+1} = K_i &+ K_iz_i\alpha_i^2(\mu_iz_i^TK_i - \beta_iz_i^TK_i^2) \\ &- K_i^2z_i\alpha_i^2(\nu_iz_i^TK_i^2 + \beta_iz_i^TK_i). \end{aligned} \tag{3.61}$$

Analysis of the sequence $\{K_i\}$, $i = 0, 1, \ldots$, also requires that the properties given by Theorem 3.2 and Corollary 3.5 be expressed in terms of K_i and z_i. Corollary 3.5 with $k = i$ and $p = n - 1$ gives

$$d_i^TGd_j = 0, \; 0 \leq j < i < n.$$

Applying the definition of d_i and (3.59) to this equation yields

$$z_i^TK_iK_jz_j = 0, \; 0 \leq j < i < n. \tag{3.62}$$

By Theorem 3.2 with $r = i$ and $p = n - 1$,

$$GH_iy_j = y_j, \; 0 \leq j < i < n,$$

which implies, by (3.5) and (3.60), if $\alpha_i \neq 0$,

$$K_iK_jz_j = K_jz_j, \; 0 \leq j < i < n. \tag{3.63}$$

Equation (3.61) shows that K_{i+1} depends on both K_i and β_i, and since β_i is arbitrary, K_{i+1} is also, to a certain extent, arbitrary. But K_i is itself arbitrary, depending upon the choice of β_{i-1}, and it might therefore be thought that K_{i+1} would also depend, through K_i, upon β_{i-1}. Broyden proves that this is not the case. The following theorem and corollary are needed.

Theorem 3.10: If K_i and z_i are as previously defined, K_i is positive definite, and $\beta_i \geq 0$, then

$$K_{i+1}^m z_{i+1} = \sum_{k=1}^{m+1} \delta_k K_i^k z_i, \tag{3.64}$$

where the δ_k are scalars, $\delta_{m+1} \neq 0$ if $z_{i+1} \neq 0$, and m is any positive integer.

Proof: Proof is by induction. Using the appropriate definitions and (3.60),

$$z_{i+1} = z_i - \alpha_i K_i z_i. \tag{3.65}$$

Then, by (3.61),

$$K_{i+1} z_{i+1} = K_i z_i \psi_i + K_i^2 z_i \phi_i \tag{3.66}$$

for appropriate scalars ψ_i and ϕ_i. If $\phi_i \neq 0$, then (3.66) establishes (3.64) for m = 1. By (3.62) and (3.63),

$$z_{i+1}^T K_i z_i = z_{i+1}^T K_{i+1} K_i z_i = 0 \tag{3.67}$$

so that (3.66) implies

$$z_{i+1}^T K_i^2 z_i \phi_i = z_{i+1}^T K_{i+1} z_{i+1}. \tag{3.68}$$

From the definition of K_i, it follows that K_i is positive definite if and only if H_i is positive definite. By hypothesis, K_i is positive definite and $\beta_i \geq 0$ so that by Theorem 3.6, H_{i+1} is positive definite. Hence, K_{i+1} is positive definite and (3.68) implies that $\emptyset_i \neq 0$ if $z_{i+1} \neq 0$.

Assume the validity of (3.64). Then, from (3.61),

$$\begin{aligned} K_{i+1}^{m+1} z_{i+1} &= K_{i+1} \sum_{k=1}^{m+1} \delta_k K_i^k z_i \\ &= K_i \sum_{k=1}^{m+1} \delta_k K_i^k z_i + \delta_i' K_i z_i + \delta_i'' K_i^2 z_i \\ &= \sum_{k=1}^{m+2} \delta_k' K_i^k z_i \end{aligned}$$

for appropriate scalars δ_i', δ_i'', and δ_k'. Since $\delta_{m+1} \neq 0$ implies $\delta_{m+2}' \neq 0$, the proof is complete.

Corollary 3.10: If $z_j \neq 0$, $j = 0, 1, \ldots, i$, where $0 \leq i \leq m \leq n - 1$, $\beta_j \geq 0$, $j = 0, 1, \ldots, i - 1$, and K_0 is positive definite, then

$$K_i z_i = \sum_{k=1}^{i+1} \delta_k K_0^k z_0 \tag{3.69}$$

where the δ_k are scalars, $\delta_{i+1} \neq 0$.

Proof: Since (3.69) is obviously true for $i = 0$, assume $i \geq 1$. Replacing i by i - 1 in (3.66) gives

$$K_i z_i = \sum_{k=1}^{2} \delta_k K_{i-1}^k z_{i-1}.$$

Applying Theorem 3.10 to each term of this sum yields

$$\begin{aligned} K_i z_i &= \sum_{k=1}^{2} \delta_k \left(\sum_{p=1}^{k+1} \delta_p' K_{i-2}^p z_{i-2} \right) \\ &= \sum_{k=1}^{3} \delta_k'' K_{i-2}^k z_{i-2}. \end{aligned}$$

By successively applying Theorem 3.10 in this way, the corollary is established.

For the remainder of this section, assume that the hypotheses of Corollary 3.10 are satisfied, that is, the algorithm has not terminated with the i-th iteration and K_i is positive definite. Then $K_i z_i \neq 0$, $0 \leq i \leq m$. Corollary 3.10 shows that $K_i z_i$ may be expressed as a linear combination of the vectors $K_0 z_0$, $K_0^2 z_0$, ..., $K_0^{i+1} z_0$. Equation (3.69) may be written as

$$K_i z_i = [K_0 z_0, K_0^2 z_0, \ldots, K_0^{i+1} z_0] w_i, \tag{3.70}$$

where the elements of w_i are the scalars δ_k, $k = 1, 2, \ldots, i + 1$. Define the matrix M by

$$M = [K_0 z_0, K_0^2 z_0, \ldots, K_0^{m+1} z_0].$$

Equation (3.70) may now be written as

$$K_i z_i = M v_i, \quad 0 \leq i \leq m, \tag{3.71}$$

where v_i is a vector whose first $i + 1$ elements are the same as those of w_i and whose remaining elements are zero. If V is the $(m + 1)x(m + 1)$ upper triangular matrix whose $(i + 1)$-st column is v_i, then (3.71) is equivalent to

$$[K_0 z_0, K_1 z_1, \ldots, K_m z_m] = MV. \tag{3.72}$$

By (3.62), the vectors $K_i z_i$, $i = 0, 1, \ldots, m$, are mutually orthogonal, and hence linearly independent since they are not null. If q_i denotes the normalized form of $K_i z_i$, then

$$[K_0z_0, K_1z_1, \ldots, K_mz_m]R = Q, \qquad (3.73)$$

where

$$Q = [q_0, q_1, \ldots, q_m]$$

and R is diagonal and chosen so that

$$Q^TQ = I. \qquad (3.74)$$

It follows from (3.72) and (3.73) that

$$Q = MU, \qquad (3.75)$$

where $U = VR$. The choice of R is unique apart from the signs of its nonzero diagonal elements and since V is upper triangular with nonzero diagonal elements, these may be chosen to make the diagonal elements of U positive. Then U is an upper triangular matrix with positive diagonal elements and hence is nonsingular. Since Q has rank $m + 1$, (3.75) implies that M also has rank $m + 1$. Equations (3.74) and (3.75) imply $U^TM^TMU = I$, so that

$$UU^T = (M^TM)^{-1}. \qquad (3.76)$$

Since M has rank $m + 1$, M^TM is positive definite and hence, by Theorem 3.8 of [56, p. 140], has a unique Cholesky decomposition, that is, there is a unique lower triangular matrix L with positive diagonal elements such that $M^TM = LL^T$. Therefore, (3.75) and (3.76) and the definition of M imply that, subject to the sign convention adopted, Q is uniquely determined by K_0 and z_0.

Using the above results, the following theorem shows that despite the fact that the matrix K_i depends upon i arbitrary parameters β_j, $j = 0, 1, \ldots, i - 1$, there is only one arbitrary term in its composition.

Theorem 3.11: The matrix K_i depends only upon the initial matrix K_0 and vector z_0 apart from a single arbitrary additive term of rank one.

Proof: By Theorem 3.10 with $m = 1$, and appropriate scalars ϕ and ψ,

$$K_j^2 z_j = \phi K_j z_j + \psi K_{j+1} z_{j+1}.$$

By substituting this expression into (3.61), it follows that

$$K_{j+1} = K_j + [q_j, q_{j+1}] \begin{bmatrix} \sigma_j & \delta_j \\ \delta_j & \gamma_j \end{bmatrix} \begin{bmatrix} q_j^T \\ q_{j+1}^T \end{bmatrix}, \tag{3.77}$$

where σ_j, δ_j, and γ_j are scalars which will be determined subsequently and the vectors q_j are as previously defined. Applying (3.77) consecutively with $j = i - 1, i, \ldots, 0$, gives

$$K_i = K_i' + q_i \gamma_{i-1} q_i^T, \tag{3.78}$$

where

$$K_i' = K_0 + q_0 \sigma_0 q_0^T + \sum_{j=1}^{i-1} q_j(\sigma_j + \gamma_{j-1}) q_j^T$$
$$+ \sum_{j=1}^{i-1} \delta_j (q_j q_{j+1}^T + q_{j+1} q_j^T). \tag{3.79}$$

Then by the orthonormality of the vectors q_j,

$$K_i q_0 = K_0 q_0 + \sigma_0 q_0 + \delta_0 q_1, \text{ and} \tag{3.80}$$

$$K_i q_j = K_0 q_j + (\sigma_j + \gamma_{j-1}) q_j + \delta_{j-1} q_{j-1} + \delta_j q_{j+1},\ 1 \le j \le i - 1. \tag{3.81}$$

Equation (3.63) and the definition of q_j imply that

$$K_i q_j = q_j, \; 0 \leq j < i.$$

Hence, premultiplying (3.80) by q_0^T and q_1^T and (3.81) by q_j^T and q_{j+1}^T, respectively, and using the orthonormality of the q_j gives

$$1 = q_0^T K_0 q_0 + \sigma_0,$$

$$0 = q_1^T K_0 q_0 + \delta_0,$$

$$1 = q_j^T K_0 q_j + \sigma_j + \gamma_{j-1}, \; 1 \leq j \leq i - 1, \text{ and}$$

$$0 = q_{j+1}^T K_0 q_j + \delta_j, \; 1 \leq j \leq i - 1.$$

Since the vectors q_j depend only upon K_0 and z_0, these equations together with (3.79) imply that K_i' is determined solely by K_0 and z_0, completing the proof of the theorem.

It follows from the preceding theorem and the fact that K_{i+1} depends upon β_i that γ_i must also depend upon β_i. To derive the precise form of this dependence note that, by (3.66) and the definition of q_{i+1}, the term $\gamma_i q_{i+1}^T q_{i+1}^T$ in (3.78), with i replaced by i + 1, gives rise to a term of the form $K_i^2 z_i z_i^T K_i^2$ which must correspond to the term, $- \alpha_i^2 \nu_i K_i^2 z_i z_i^T K_i^2$, in (3.61). Denote $z_i^T K_i z_i$ by θ_j, j = 1, 2, 3, and 4, to simplify notation. Then by (3.57), (3.5), and (3.60),

$$\nu_i \alpha_i^2 = \frac{1 - \alpha_i^2 \beta_i \theta_2}{\theta_3} \qquad (3.82)$$

By (3.61) and (3.65),

$$K_{i+1}z_{i+1} = K_iz_i[1 + \alpha_i^2(\mu_i\theta_1 - \beta_i\theta_2) - \alpha_i^3(\mu_i\theta_2 - \beta_i\theta_3)]$$

$$- K_i^2z_i[\alpha_i^2(\nu_i\theta_2 + \beta_i\theta_1) + \alpha_i - \alpha_i^3(\nu_i\theta_3 + \beta_i\theta_2)]. \qquad (3.83)$$

Premultiplying (3.65) by $z_i^TK_i$ and using (3.67) yields

$$\alpha_i\theta_2 = \theta_1, \qquad (3.84)$$

so that (3.83) reduces to

$$K_{i+1}z_{i+1} = K_iz_i[1 - \beta_i\alpha_i^2(\theta_2 - \theta_3\alpha_i)] - K_i^2z_i[\alpha_i + \nu_i\alpha_i^2(\theta_2 - \theta_3\alpha_i)]$$

which, after substituting for $\nu_i\alpha_i^2$ given by (3.82), becomes

$$K_{i+1}z_{i+1} = (K_iz_i - K_i^2z_i\theta_2/\theta_3)\varrho_i,$$

where

$$\varrho_i = 1 - \beta_i\alpha_i^2(\theta_2 - \theta_3\alpha_i).$$

It follows immediately that

$$z_{i+1}^TK_{i+1}^2z_{i+1} = (\theta_4\theta_2^2/\theta_3^2 - \theta_2)\varrho_i^2.$$

Therefore, the coefficient of the term $K_i^2z_iz_i^TK_i^2$ obtained from

$$\gamma_iq_{i+1}q_{i+1}^T = \frac{\gamma_iK_{i+1}z_{i+1}z_{i+1}^TK_{i+1}}{z_{i+1}^TK_{i+1}^2z_{i+1}}$$

is given by

$$\frac{\gamma_i\theta_2^2/\theta_3^2}{\theta_4\theta_2^2/\theta_3^2 - \theta_2} \qquad (3.85)$$

Since the coefficients of $K_i^2z_iz_i^TK_i^2$ from (3.61) and (3.78) must be equal, (3.82) and (3.85) imply

$$\gamma_i = \theta(\beta_i\alpha_i^2\theta_2 - 1), \tag{3.86}$$

where $\theta = (\theta_4\theta_2 - \theta_3^2)/\theta_3\theta_2$. Hence γ_i varies linearly with β_i. Since $\theta_2 > 0$ and $\theta_3 > 0$ by the positive definiteness of K_i, it follows from the Schwarz inequality that $\theta \geq 0$. Thus, in general, γ_i increases with β_i and $\beta_i \geq 0$ if and only if $\gamma_i \geq -\theta$.

The parameter γ_i is essentially arbitrary in that it depends upon β_i, and if $\gamma_i \geq -\theta$ then the resulting variable metric method will be stable. By Theorem 3.11, K_{i+1} depends only upon the initial matrix K_0 and vector z_0 apart from the additive term $\gamma_i q_{i+1}q_{i+1}^T$ where the vector q_{i+1} is uniquely determined by K_0 and z_0. Thus, to obtain a stable method which might avoid the numerical difficulties encountered with the DFP method, it is logical to choose $\gamma_i > -\theta$ having regard for the condition number of K_{i+1}. Although choosing γ_i to minimize this condition number would require excessive computation, elementary considerations of this nature lead Broyden to suggest that a reasonable value for γ_i is zero. If γ_i were negative, λ the smallest eigenvalue of K_{i+1} with eigenvector x, and λ' the smallest eigenvalue of K'_{i+1} with eigenvector x', then by (3.78) and Theorem 5.7 of [56, p. 312],

$$\lambda \leq \frac{x'^T K_{i+1} x'}{x'^T x'} = \frac{x'^T K'_{i+1} x'}{x'^T x'} + \frac{\gamma_i x'^T q_{i+1} q_{i+1}^T x'}{x'^T x'}$$

$$< \frac{x'^T K'_{i+1} x'}{x'^T x'} = \lambda'.$$

That is, the smallest eigenvalue of K_{i+1} would be less than that of

K'_{i+1}. Similarly, if γ_i were positive, the largest eigenvalue of K_{i+1} would be greater than that of K'_{i+1}. In either case, the matrix K_{i+1} could be more ill-conditioned than the matrix K'_{i+1} which is determined solely by K_0 and z_0. From the definition of K_i and the consistency of the matrix 2-norm,

$$\| H_i \|_2 \leq \| G^{-\frac{1}{2}} \|_2^2 \| K_i \|_2, \text{ and}$$

$$\| H_i^{-1} \|_2 \leq \| G^{\frac{1}{2}} \|_2^2 \| K_i^{-1} \|_2,$$

so that

$$\chi(H_i) \leq [\chi(G^{\frac{1}{2}})]^2 \chi(K_i).$$

Hence if K_i is ill-conditioned, H_i might also be ill-conditioned. Note that for the DFP method, $\beta_i = 0$ for all i so that by (3.86), γ_i is in general negative for all i. Thus, the reported behavior of the DFP method supports the above reasoning.

If γ_i is set equal to zero, it follows from (3.86) and (3.84) that β_i must be chosen to satisfy

$$\beta_i \alpha_i z_i^T K_i z_i = 1.$$

By (3.60), the definition of z_i, and (3.59), this equation implies

$$\beta_i = \frac{-1}{e_i^T G^{\frac{1}{2}} G^{\frac{1}{2}} s_i} = \frac{-1}{g_i^T s_i}$$

which, by (3.24), is equivalent to

$$\beta_i = \frac{1}{s_i^T y_i}.$$

If this value of β_i is substituted into the general matrix updating formula of Algorithm 3.1, the updating formula for the variable matrix H_i is

$$H_{i+1} = H_i + \frac{1}{s_i^T y_i}\left[- s_i y_i^T H_i - H_i y_i s_i^T + \left(1 + \frac{y_i^T H_i y_i}{s_i^T y_i} \right) s_i s_i^T \right] . (3.87)$$

This formula is identical to (3.52), the complementary DFP formula obtained by Fletcher, who showed that the matrix H_{i+1} obtained by this formula is "less singular" than that obtained by the DFP formula. In addition, the correction matrix in the formula given by (3.87) is identical to Goldfarb's correction matrix C_{H*} which minimizes the norm defined by $Tr(WCWC^T)$ where $W^{-1} = H_{i+1}$.

This special case of Algorithm 3.1 possesses in theory all the properties that made the DFP algorithm so successful. To determine if this choice of the parameter β_i has improved the numerical properties, Broyden [8] compares the performance of the new algorithm with that of the DFP algorithm on a variety of standard test functions. Results of the computation for a representative sample are summarized in Table I. Computation for each function was terminated when $\| g_k \|_2 < \epsilon$, where ϵ is the specified tolerance. These test functions are documented in the Appendix.

Table I reveals little significant difference between the methods except for the third and fifth functions. On the third function, the new method is markedly inferior. This behavior is explained by Theorem 2 of [8] which proves that if the new algorithm is applied to f given by (3.4), where G is positive definite, and if $e_{i+1} \neq 0$, then

$$\| K_{i+1} - I \|_F < \| K_i - I \|_F .$$

Since reducing the matrix error norm makes the iteration matrix look more like the inverse Hessian matrix, the new algorithm, in this sense, resembles Newton's method more closely than does the DFP algorithm. Thus its performance is expected to reflect that of Newton's method, so that it might perform comparatively badly if the Hessian matrix is singular at the minimum point, as in Powell's function.

TABLE I

COMPARISON OF THE DFP METHOD AND THE COMPLEMENTARY DFP METHOD, BROYDEN

Function	n	$\| g_0 \|_2$	ϵ	DFP'		DFP	
				Iter.	Eval.	Iter.	Eval.
Rosenbrock	2	$2.30*10^2$	10^{-6}	19	188	23	239
Helical Valley	3	$1.90*10^3$	10^{-6}	21	167	21	167
Powell	4	$3.60*10^3$	10^{-6}	26	231	18	182
Trigonometric	45	$1.63*10^{11}$	$6.71*10^{-5}$	63	480	63	499
Sum of Exponentials	6	$5.70*10$	10^{-6}	33	404	65	886

For the fifth function in Table I, the new method is much better. The behavior of the DFP algorithm on this and similar functions is particularly interesting. Broyden reports that for the fifth function, after 33 iterations $\| g_k \|_2$ had been reduced to approximately 10^{-3}, and it then hovered around this value until the 60-th iteration when it was reduced to about 10^{-4}. Subsequent iterations then reduced $\| g_k \|_2$ steadily until at the 65-th iteration, it fell below 10^{-6} and the program was terminated. Thus, the DFP algorithm appeared to get reasonably close to the solution in only a few more iterations than required by the new algorithm and it then proceeded to "mark time" for perhaps 20 iterations or so.

These numerical results suggest that the performance of the new algorithm is substantially the same as that of the DFP algorithm in the initial stages of the minimization, but that the characteristics of the algorithms during the final stages are markedly different. A consideration of the values of β_i for the two algorithms shows this to be reasonable. By Theorem 5.7 of [56, p. 312], if $g_i \neq 0$, then

$$\lambda_{\min} \| g_i \|_2^2 \leq g_i^T H_i g_i \leq \lambda_{\max} \| g_i \|_2^2,$$

where $\lambda_{\min}$ is the smallest eigenvalue of H_i and $\lambda_{\max}$ is the largest eigenvalue of H_i. Since the gradients at the beginning of the minimization are usually large, the value for the new algorithm, $\beta_i = 1/\alpha_i g_i^T H_i g_i$, may well approach zero, the value for the DFP algorithm, provided α_i is not too small. Thus, the two algorithms become effectively identical. On the other hand, as the minimum is approached, β_i for the new method becomes extremely large and the maximum discrepancy between the two methods occurs. Broyden reports a range of values of β_i for the new algorithm from 10^{-4} initially to 10^4. Another situation that could give rise to a large value of β_i for the new algorithm is the occurrence of a nearly singular H_i. In this case, it would be possible for both $g_i^T H_i g_i$ and α_i to be small despite a large value of $\| g_i \|_2$. Thus, the DFP algorithm would differ markedly from the new algorithm and, by the discussion following Theorem 3.11, the DFP algorithm could yield a new value of H_i that would be much more badly conditioned than that given by the new algorithm. It was shown in the last section of Chapter II that the observed poor performance of the DFP algorithm could be explained by the occurrence of a nearly singular H_i. Therefore, the difference between the two algorithms in this case is highly encouraging. On the basis of his numerical experiments, Broyden concludes that the observed behavior of the DFP algorithm is probably due to a tendency toward singularity for the matrices H_i as hypothesized from the negative values of γ_i, and that the strategy of choosing β_i to eliminate this tendency appears to have been largely successful.

With the exception of the third function in Table I in which the DFP algorithm was significantly superior, for all of the functions tested by Broyden the number of iterations required by the new algorithm was comparable to, or substantially less than, that required by the DFP algorithm. In addition, if the number of function evaluations per iteration is taken as the measure, then the new algorithm is slightly better in terms of work done during each iteration. Broyden reports an average ratio of function evaluations to iterations of 8.55 for the new method and 9.88 for the DFP method. Although these results represent only a limited amount of numerical experience on a restricted set of functions and to this extent will not necessarily reflect the overall merit of the two algorithms, they do indicate that the new algorithm is worth further consideration, especially for difficult problems or those for which existing methods are either unsuccessful or slow in converging.

Shanno [53] investigates the conditioning of the family of matrices (3.34) as a function of the scalar parameter τ. As noted in the last section of Chapter II, computational difficulties can arise when the smallest eigenvalue of the variable matrix H goes to zero, since this causes the condition number of H to become large. By Theorem 3.7, if H_k is positive definite and $\tau > (\alpha_k - 1)/\alpha_k$, then H_{k+1} is positive definite. Hence, at no finite step k does the smallest eigenvalue of H_{k+1} ever become zero. However, it is possible for λ to approach zero as k approaches infinity, where λ is the smallest eigenvalue of H_{k+1}. Thus, Shanno elects to condition the matrix H_{k+1} by choosing τ at each step in such a way as to maximize the smallest eigenvalue of H_{k+1}. If λ is the smallest eigenvalue of H_{k+1} with eigenvector x such that $\|x\|_2 = 1$, then $\lambda = x^TH_{k+1}x$. Hence λ is maximized by choosing τ to maximize $w^TH_{k+1}w$ for an arbitrary nonzero vector w. To determine the value of τ which maximizes $w^TH_{k+1}w$, the following lemma is needed.

<u>Lemma</u> 3.3: For $\tau > (\alpha_k - 1)/\alpha_k$ and H_k positive definite, $g_{k+1}^TH_{k+1}g_{k+1}$ is a monotonically increasing function of τ.

<u>Proof</u>: By (3.37),

$$g_{k+1}^TH_{k+1}g_{k+1} = \frac{(\alpha_k\tau - \alpha_k + 1)g_k^TH_kg_kg_{k+1}^TH_kg_{k+1}}{(\alpha_k\tau - \alpha_k + 1)g_k^TH_kg_k + g_{k+1}^TH_kg_{k+1}}.$$

Differentiating with respect to τ yields

$$\frac{d(g_{k+1}^TH_{k+1}g_{k+1})}{d\tau} = \frac{\alpha_kg_k^TH_kg_k(g_{k+1}^TH_kg_{k+1})^2}{[(\alpha_k\tau - \alpha_k + 1)g_k^TH_kg_k + g_{k+1}^TH_kg_{k+1}]^2}.$$

Since H_k is positive definite, this derivative is positive and the lemma is proved.

Theorem 3.12: Let w be an arbitrary nonzero vector. For $\tau > (\alpha_k - 1)/\alpha_k$, $w^TH_{k+1}w$ is a monotonically increasing function of τ.

Proof: As in the proof of Theorem 3.9, there exists a basis for R^n composed of the vectors g_k, g_{k+1}, z_1, ..., z_{n-2} which are conjugate with respect to H_k. From (3.38), the vectors z_1, ..., z_{n-2} are also conjugate with respect to H_{k+1} and satisfy the conditions $z_i^TH_{k+1}g_k = 0$ and $z_i^TH_{k+1}g_{k+1} = 0$. Also, $z_i^TH_{k+1}z_i = z_i^TH_kz_i$ and hence is independent of τ. Then $z_i^TH_{k+1}y_k = 0$ and from the quasi-Newton equation and (3.24), $g_{k+1}^TH_{k+1}y_k = g_{k+1}^Ts_k = 0$. Therefore, since for $\tau > (\alpha_k - 1)/\alpha_k$, H_{k+1} is positive definite, the vectors g_{k+1}, y_k, z_1, ..., z_{n-2} form a basis for R^n. Thus, w can be written as a linear combination of these vectors,

$$w = \sum_{i=1}^{n-2} \mu_i z_i + \mu_{n-1}g_{k+1} + \mu_n y_k,$$

and from the conjugacy of these vectors,

$$w^TH_{k+1}w = \sum_{i=1}^{n-2} \mu_i^2 z_i^TH_{k+1}z_i + \mu_{n-1}^2 g_{k+1}^TH_{k+1}g_{k+1} + \mu_n^2 y_k^TH_{k+1}y_k.$$

The first term on the right hand side is independent of τ and

$$y_k^TH_{k+1}y_k = s_k^T(g_{k+1} - g_k) = \alpha_k g_k^TH_kg_k$$

so the third term is also independent of τ. Hence

$$\frac{d(w^TH_{k+1}w)}{d\tau} = \mu_{n-1}^2 \frac{d(g_{k+1}^TH_{k+1}g_{k+1})}{d\tau},$$

and, by Lemma 3.3, this is positive if $\mu_{n-1} \neq 0$. This completes the proof of the theorem.

Theorem 3.12 shows that the condition of H_{k+1} as represented by $w^T H_{k+1} w$ improves monotonically with $\mathcal{T}$. Thus, it is necessary to find a closed form representation of H_{k+1} for $\mathcal{T} = \infty$. This is done in the following theorem.

<u>Theorem</u> 3.13: Let H_{k+1} be defined by (3.34). Then

$$\lim_{\mathcal{T} \to \infty} H_{k+1} = H_k + \frac{(s_k - \mu H_k y_k)(s_k - \mu H_k y_k)^T}{(s_k - \mu H_k y_k)^T y_k} + (\mu - 1)\frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k},$$

where

$$\mu = \frac{s_k^T y_k}{s_k^T y_k + y_k^T H_k y_k}.$$

<u>Proof</u>: By Householder's modification rule [29, pp. 123-124], if A is a nonsingular matrix, σ a scalar and w an arbitrary vector such that $A + \sigma w w^T$ is nonsingular, then

$$(A + \sigma w w^T)^{-1} = A^{-1} - \frac{\sigma}{1 + \sigma w^T A^{-1} w} A^{-1} w w^T A^{-1}. \qquad (3.88)$$

Applying (3.88) to

$$H_{k+1} = \left[H_k + \left(\frac{\mathcal{T}}{s_k^T y_k} \right) s_k s_k^T \right] + \delta b b^T,$$

where

$$\delta = \frac{1}{(1 - \mathcal{T}) s_k^T y_k - y_k^T H_k y_k}, \text{ and } b = (1 - \mathcal{T}) s_k - H_k y_k,$$

yields

$$H_{k+1}^{-1} = B^{-1} - \gamma B^{-1} b b^T B^{-1},$$

where

$$B = H_k + \left(\frac{\tau}{s_k^T y_k} \right) s_k s_k^T, \text{ and } \gamma = \frac{\delta}{1 + \delta b^T B^{-1} b},$$

which after applying (3.88) to B and simplifying becomes

$$H_{k+1}^{-1} = H_k^{-1} + \frac{\alpha_k \tau}{1 + \alpha_k \tau} \frac{\alpha_k g_k g_k^T}{g_k^T s_k} + \frac{\left(\frac{\alpha_k}{1 + \alpha_k \tau} g_k + y_k \right)\left(\frac{\alpha_k}{1 + \alpha_k \tau} g_k + y_k \right)^T}{\left(\frac{\alpha_k}{1 + \alpha_k \tau} g_k + y_k \right)^T s_k}$$

Then,

$$\lim_{\tau \to \infty} H_{k+1}^{-1} = H_k^{-1} + \frac{\alpha_k g_k g_k^T}{g_k^T s_k} + \frac{y_k y_k^T}{y_k^T s_k}. \tag{3.89}$$

Since

$$\left(\lim_{\tau \to \infty} H_{k+1} \right)\left(\lim_{\tau \to \infty} H_{k+1} \right) = I,$$

$$\lim_{\tau \to \infty} H_{k+1} = \left(\lim_{\tau \to \infty} H_{k+1}^{-1} \right)^{-1}$$

which can be obtained by applying (3.88) to (3.89) twice sequentially, as indicated below. After some tedious manipulations,

$$\lim_{\tau \to \infty} H_{k+1} = \left[\left(H_k^{-1} + \frac{y_k y_k^T}{y_k^T s_k} \right) + \frac{\alpha_k g_k g_k^T}{g_k^T s_k} \right]^{-1}$$

yields the desired result. This completes the proof of the theorem.

Computation shows that the updating formula obtained in Theorem 3.13 is identical to Broyden's special case given by (3.87). This result also follows from (3.32) which relates the parameters β_k and and shows that $\beta_k \to 1/s_k^T y_k$ as $\tau \to \infty$.

Shanno tests the methods corresponding to $\tau = \infty$ and $\tau = 1$ for various initial estimates on four standard test functions which are documented in the Appendix. A representative sample of his results is given in Table II. Computation is terminated, that is, convergence is assumed, when $|\text{i-th component of } s_k| \leq 10^{-5}|\text{i-th component of } x_k|$, and $|\text{i-th component of } g_k| \leq 10^{-5}|\text{i-th component of } x_k|$.

TABLE II

COMPARISON OF THE DFP METHOD AND THE COMPLEMENTARY DFP METHOD, SHANNO

Function	x_0	DFP'		DFP	
		Iter.	Eval.	Iter.	Eval.
Sum of Two Exponentials	(5, 20)	8	33	9	49
Rosenbrock	(-1.2, 1)	14	56	13	65
Rosenbrock	(1.489, -2.547)	18	77	20	134
Wood	(-3, -1, -3, -1)	21	97	22	114
Weibull	(100, 3, 12.5)	28	149	(*)	(*)

*Convergence had not been attained in 50 iterations and 499 evaluations.

Shanno reports that, in virtually all cases, the DFP' method corresponding to $\tau = \infty$ outperformed the DFP method corresponding to $\tau = 1$, and the difference became more notable as the complexity of the function increased. His conclusion is that the new method is preferable to the DFP method.

Convergence

Dixon [17] establishes a result which allows Powell's general convergence theorem, Theorem 2.7, for the DFP method to be applied to other members of Broyden's one-parameter family of methods. Essentially, this result shows that under the same initial conditions, the sequence of points generated by Algorithm 3.1 is independent of the choice of parameter at each iteration, provided the linear search is exact. However, two other conditions must be met.

The value of α_k at each iteration in Algorithm 3.1 is determined by a linear search from the point x_k in the direction $\pm d_k$. If the search is exact, then the gradient at $x_{k+1} = x_k + \alpha_k d_k$ is orthogonal to the step $s_k = \alpha_k d_k$ taken, that is, $g_{k+1}^T s_k = 0$. To ensure that, given x_k and d_k, the value of α_k is uniquely defined, it will also be assumed that the search locates the nearest local minimum in the downhill direction, $\pm d_k$, from x_k. Such a search will be called a perfect linear search.

It was shown in the second section of this chapter that the value of $\tau = (\alpha_k - 1)/\alpha_k$, or equivalently, $\beta_k = - 1/(\alpha_k y_k^T H_k y_k - s_k^T y_k)$ makes d_{k+1} identically zero. In addition, d_{k+1} will be undefined if $d_k^T g_k = 0$ since this causes H_{k+1} to be undefined due to zero denominators. Hence, it must be assumed that d_k is defined and nonzero, that is, degeneracy has not occurred.

Dixon's main result is derived from the following theorem. The proof of this theorem will follow from a general result in Chapter IV.

Theorem 3.14: If a sequence of points $\{x_k\}$, $k = 0, 1, \ldots,$ is generated by Algorithm 3.1, starting at a given point x_0 with a given symmetric nonsingular matrix H_0 and using a perfect linear search at each iteration, then provided degeneracy does not occur, the sequence of search directions generated can be represented by $d_k = \mu_k p_k$ for some scalar μ_k, where

$$p_0 = H_0 g_0,$$

$$p_1 = \left(I - \frac{s_0 y_0^T}{s_0^T y_0} \right) H_0 g_1,$$

and for $k > 1$,

$$p_k = \left[\prod_{j=0}^{k-1} \left(I - \frac{s_j y_j^T}{s_j^T y_j} \right) \right] H_0 g_k$$

$$+ \sum_{j=0}^{k-2} \left[\prod_{m=j+1}^{k-1} \left(I - \frac{s_m y_m^T}{s_m^T y_m} \right) \right] \frac{s_j^T g_k}{s_j^T y_j} s_j.$$

For a given point x_0 and matrix H_0, a perfect linear search in the direction $d_0 = - H_0 g_0$ yields the same point x_1, and hence the same values of s_0 and y_0 for all members of Broyden's family. Then, by Theorem 3.14, $d_1 = \mu_1 p_1$ where p_1 is the same for all members so that a search in the direction d_1 yields the same point x_2, and hence the same values of s_1 and y_1, independent of the choice of β_0. Assuming that the same points x_0, x_1, $\ldots$, x_k, and hence the same values of

$s_0, s_1, \ldots, s_{k-1}$, and $y_0, y_1, \ldots, y_{k-1}$, have been generated, the expression for p_k given by Theorem 3.14 is the same for all members. Hence, a search in the direction $d_k = \mu_k p_k$ yields the same point x_{k+1}, independent of the choice of β_{k-1}. Thus, Theorem 3.14 implies the desired result.

Theorem 3.15: Under the conditions of Theorem 3.14, the sequence $\{x_k\}$, $k = 0, 1, \ldots$, is independent of the choice of the parameter at each iteration.

Since the DFP method is a member of Broyden's family, Theorem 3.15 extends Powell's convergence theorem to the other members under the stated conditions. In particular, since degeneracy does not occur if $\beta_k = 1/s_k^T y_k$, the variable metric method using the complementary DFP formula and perfect linear searches converges to the minimum of a convex function satisfying the conditions of Theorem 2.7. By Theorem 3.15, the DFP' algorithm and DFP algorithm will generate the same sequence of points if perfect linear searches are used. Since most implementations do not undertake accurate linear searches, this implies that the improved numerical performance of the DFP' algorithm over the DFP algorithm is crucially dependent on the nonaccurate linear search strategy used in the implementations of each of these algorithms. This conclusion is supported by a careful numerical study by Dixon [19], which compares the performance of these formulas when used in conjunction with different strategies for determining the step length.

CHAPTER IV

GENERAL FAMILIES

Huang

A general family of variable metric methods is obtained by Huang [30] using a unified approach to construct a minimization algorithm having the following properties:

i) the algorithm uses linear searches only;

ii) the algorithm is quadratically terminating;

iii) the algorithm requires function and gradient values only; and

iv) the algorithm employs only information from the present and immediately preceding iterations.

In constructing this algorithm, by property ii), it is assumed that the function f to be minimized is defined by

$$f(x) = \tfrac{1}{2}x^T Gx + a^T x + \gamma, \tag{4.1}$$

where G is an n x n positive definite matrix, a an arbitrary n-vector, and γ a scalar.

The algorithm will generate a sequence of points $\{x_k\}$, $k = 0, 1, \ldots,$ by the iteration formula

$$x_{k+1} = x_k + s_k \tag{4.2}$$

with

$$s_k = \alpha_k d_k, \tag{4.3}$$

where the vector d_k denotes the search direction and the scalar α_k is the step size. Then $f(x_{k+1}) = f(x_k + \alpha_k d_k)$ depends on α_k and d_k. Hence, by property i), d_k must be defined so that $f(x_{k+1})$ becomes a function of α_k only. In that case, α_k is determined by a linear search along the direction $\pm d_k$ from x_k and

$$g_{k+1}^T d_k = 0, \tag{4.4}$$

where g_{k+1} is the gradient $g(x_{k+1})$. For f defined by (4.1),

$$g_{k+1} = g_k + Gs_k. \tag{4.5}$$

Then, by (4.4) and (4.3),

$$(g_k + \alpha_k G d_k)^T d_k = 0$$

which implies

$$\alpha_k = \frac{-\, g_k^T d_k}{d_k^T G d_k}. \tag{4.6}$$

Thus, from the definition of f and equations (4.2) and (4.3),

$$f(x_{k+1}) - f(x_k) = \tfrac{1}{2} s_k^T G s_k + g_k^T s_k = \frac{-\,(g_k^T d_k)^2}{2 d_k^T G d_k}. \tag{4.7}$$

Since the matrix G is positive definite, $f(x_{k+1}) < f(x_k)$ if

$$g_k^T d_k \neq 0. \tag{4.8}$$

Equation (4.8) states that d_k should not be orthogonal to g_k and thus will be called the nonorthogonality condition.

From (4.4) with k replaced by k - 1,

$$g_k^T d_{k-1} = 0$$

which with (4.8) implies that d_k should not be parallel to the previous search direction d_{k-1}. In fact, by Theorem 2.3, property ii) will be obtained if the search direction d_k is conjugate to all previous search directions d_j with respect to the matrix G, that is, if

$$d_k^T G d_j = 0, \ 0 \leq j < k \leq p, \ 1 \leq p \leq n - 1. \tag{4.9}$$

Therefore, if the search direction is defined by

$$d_k = - H_k^T g_k,$$

where H_k is a matrix to be determined, then the conjugacy condition (4.9) is equivalent to

$$g_k^T H_k G d_j = 0, \ 0 \leq j \leq k - 1. \tag{4.10}$$

Also, by the proof of Theorem 2.3, if all previous search directions d_j are chosen so that the conjugacy condition

$$d_i^T G d_j = 0, \ 0 \leq i < j \leq k - 1,$$

is satisfied, then the gradient g_k has the property

$$g_k^T d_j = 0, \ 0 \leq j \leq k - 1. \tag{4.11}$$

Comparison of equations (4.10) and (4.11) shows that (4.10), and hence (4.9), can be satisfied if the matrix H_k is chosen such that

$$H_kGd_j = \sigma d_j,\ 0 \leq j \leq k - 1, \qquad (4.12)$$

where σ is an arbitrary scalar. If y_k is defined by

$$y_k = g_{k+1} - g_k,$$

then by (4.5),

$$y_k = Gs_k. \qquad (4.13)$$

Hence, the matrix G may be eliminated from (4.12) by multiplying by α_j and using (4.3) and (4.13). The resulting equation,

$$H_ky_j = \sigma s_j,\ 0 \leq j \leq k - 1, \qquad (4.14)$$

may be separated into

$$H_ky_j = \sigma s_j,\ 0 \leq j \leq k - 2, \qquad (4.15)$$

and

$$H_ky_{k-1} = \sigma s_{k-1}. \qquad (4.16)$$

Subtracting (4.14) with k replaced by k - 1, that is,

$$H_{k-1}y_j = \sigma s_j,\ 0 \leq j \leq k - 2, \qquad (4.17)$$

from (4.15) yields

$$(H_k - H_{k-1})y_j = 0,\ 0 \leq j \leq k - 2.$$

Therefore, if the matrix H_k is obtained from H_{k-1} by

$$H_k = H_{k-1} + C_{k-1} \qquad (4.18)$$

for some matrix C_{k-1}, then (4.15) can be satisfied if C_{k-1} has the property

$$C_{k-1}y_j = 0, \ 0 \leq j \leq k - 2. \tag{4.19}$$

Also, (4.16) is satisfied if C_{k-1} has the additional property

$$C_{k-1}y_{k-1} = \sigma s_{k-1} - H_{k-1}y_{k-1}. \tag{4.20}$$

Equations (4.19) and (4.20) are satisfied if C_{k-1} is given by

$$C_{k-1} = \sigma \frac{s_{k-1}q_{k-1}^T}{q_{k-1}^T y_{k-1}} - \frac{H_{k-1}y_{k-1}z_{k-1}^T}{z_{k-1}^T y_{k-1}}, \tag{4.21}$$

where q_{k-1} and z_{k-1} are n-vectors having the properties

$$q_{k-1}^T y_j = 0, \text{ and}$$

$$z_{k-1}^T y_j = 0, \ 0 \leq j \leq k - 2. \tag{4.22}$$

Property iv) implies that the vectors q_{k-1} and z_{k-1} must be defined using only information from the present and immediately preceding iterations. The conjugacy condition (4.9) for the previous iteration gives

$$d_{k-1}^T G d_j = 0, \ 0 \leq j \leq k - 2,$$

which, using (4.3) and (4.13), yields

$$s_{k-1}^T y_j = 0, \ 0 \leq j \leq k - 2. \tag{4.23}$$

Equation (4.11) and the same relation for the previous iteration implies

$$y_{k-1}^T d_j = 0, \ 0 \leq j \leq k - 2,$$

or, by (4.3),

$$y_{k-1}^T s_j = 0, \; 0 \leq j \leq k - 2.$$

This equation then implies, using (4.17), that

$$y_{k-1}^T H_{k-1} y_j = 0, \; 0 \leq j \leq k - 2. \tag{4.24}$$

Hence, by (4.23) and (4.24), the properties given by (4.22) will be satisfied if q_{k-1} and z_{k-1} are chosen as

$$q_{k-1} = \gamma_1 s_{k-1} + \gamma_2 H_{k-1}^T y_{k-1}, \text{ and}$$

$$z_{k-1} = \delta_1 s_{k-1} + \delta_2 H_{k-1}^T y_{k-1}, \tag{4.25}$$

where γ_1, γ_2, δ_1, and δ_2 are arbitrary scalars. Thus, for $k \geq 1$, H_k is given by (4.18), (4.21), and (4.25). It remains to choose the initial matrix H_0. The following lemmas are used.

<u>Lemma</u> 4.1: If H_k, for $k \geq 1$, is given by (4.18), (4.21), and (4.25), then the search direction $d_k = - H_k^T g_k$ can be expressed as

$$d_k = \mu_k p_k, \tag{4.26}$$

where μ_k is a scalar and

$$p_k = \left[I - \frac{s_{k-1} y_{k-1}^T}{s_{k-1}^T y_{k-1}} \right] H_{k-1}^T g_k. \tag{4.27}$$

<u>Proof</u>: By (4.18), (4.21), and (4.25),

$$d_k = - H_{k-1} g_k - \sigma \frac{q_{k-1} s_{k-1}^T g_k}{q_{k-1}^T y_{k-1}} + \frac{(\delta_1 s_{k-1} + \delta_2 H_{k-1}^T y_{k-1}) y_{k-1}^T H_{k-1}^T g_k}{z_{k-1}^T y_{k-1}}. \tag{4.28}$$

Since (4.3) and (4.4) imply $s_{k-1}^T g_k = 0$ and the definitions of y_{k-1} and s_{k-1} give

$$H_{k-1}^T y_{k-1} = H_{k-1}^T g_k + \frac{s_{k-1}}{\alpha_{k-1}}, \tag{4.29}$$

equation (4.28) becomes

$$\begin{aligned} d_k = {} & \left[\frac{\delta_2 y_{k-1}^T H_{k-1}^T g_k}{z_{k-1}^T y_{k-1}} - 1 \right] H_{k-1}^T g_k \\ & - \left[\left(- \delta_1 - \frac{\delta_2}{\alpha_{k-1}} \right) \frac{s_{k-1}^T y_{k-1}}{z_{k-1}^T y_{k-1}} \right] \frac{s_{k-1} y_{k-1}^T}{s_{k-1}^T y_{k-1}} H_{k-1}^T g_k. \end{aligned} \tag{4.30}$$

Using the definitions of s_{k-1}, y_{k-1}, and z_{k-1},

$$\begin{aligned} \left(- \delta_1 - \frac{\delta_2}{\alpha_{k-1}} \right) \frac{s_{k-1}^T y_{k-1}}{z_{k-1}^T y_{k-1}} &= (- \delta_1 s_{k-1}^T y_{k-1} + \delta_2 g_{k-1}^T H_{k-1} y_{k-1} \\ &\quad - \delta_2 g_k^T H_{k-1} y_{k-1} + \delta_2 y_{k-1}^T H_{k-1}^T g_k) / z_{k-1}^T y_{k-1} \\ &= \frac{\delta_2 y_{k-1}^T H_{k-1}^T g_k}{z_{k-1}^T y_{k-1}} - 1. \end{aligned}$$

Hence, the result follows from (4.30).

<u>Lemma</u> 4.2: Under the hypothesis of Lemma 4.1, if the vectors $s_0, s_1, \ldots, s_{k-1}$ are defined and nonzero, the vector p_k given by this lemma may be expressed as

$$p_k = \left[I - \sum_{j=0}^{k-1} \frac{s_j y_j^T}{s_j^T y_j} \right] H_0^T g_k, \; k \geq 1. \tag{4.31}$$

The proof of Lemma 4.2 will follow as a special case of a more general result to be established later in this section.

Since the vector p_k is independent of the parameters σ, γ_1, γ_2, δ_1, and δ_2, Lemma 4.1 implies that the search directions d_k generated by different choices of these parameters are parallel to each other if the matrix H_{k-1} used at the point x_{k-1} is the same. Hence, the vector p_k can be regarded as the search direction for all the algorithms. Then, since $s_k = \alpha_k \mu_k p_k$, the optimum stepsize along the direction p_k is given by $\alpha_k \mu_k$. By (4.6) and (4.26),

$$\alpha_k \mu_k = \frac{- g_k^T p_k}{p_k^T G p_k}$$

which is clearly independent of the parameters. Thus, by (4.26), for all the algorithms, equation (4.7) becomes

$$f(x_{k+1}) - f(x_k) = \frac{- (g_k^T p_k)^2}{2 p_k^T G p_k}.$$

Hence, the nonorthogonality condition (4.8) is replaced by

$$g_k^T p_k \neq 0, \quad 0 \leq k \leq n - 1. \qquad (4.32)$$

Premultiplying the expression given by (4.31) for p_k by g_k^T and applying (4.11) yields

$$g_k^T p_k = g_k^T H_0^T g_k$$

which implies that the nonorthogonality condition (4.32) can be satisfied if $g_k^T H_0 g_k$ is nonzero. Thus, if H_0 is chosen such that $\frac{1}{2}(H_0 + H_0^T)$ is positive definite or negative definite, then for $g_k \neq 0$,

$$\tfrac{1}{2}g_k(H_0 + H_0^T)g_k = g_k^T H_0^T g_k$$

is nonzero. In particular, if H_0 is symmetric, this implies that H_0 must be positive definite or negative definite.

The preceding analysis has constructed the following general algorithm having the desired properties.

<u>Algorithm</u> 4.1 (Huang, 1970): Given an initial vector x_0 and an initial matrix H_0 such that $\tfrac{1}{2}(H_0 + H_0^T)$ is positive definite or negative definite.

```
For k = 0, 1, 2, ...,
    If $g_k = g(x_k) = 0$, then stop.
    Else, set $d_k = - H_k^T g_k$,
          find $\alpha_k$ which minimizes $f(x_k + \alpha d_k)$ with respect to $\alpha$,
          set $s_k = \alpha_k d_k$,
              $x_{k+1} = x_k + s_k$,
              $y_k = g_{k+1} - g_k$,
              $q_k = \gamma_1 s_k + \gamma_2 H_k^T y_k$,
              $z_k = \delta_1 s_k + \delta_2 H_k^T y_k$,
              $H_{k+1} = H_k + \sigma \frac{s_k q_k^T}{q_k^T y_k} - \frac{H_k y_k z_k^T}{z_k^T y_k}$,
          where $\sigma$, $\gamma_1$, $\gamma_2$, $\delta_1$, and $\delta_2$ are arbitrary scalars except
          for the conditions that $\gamma_1$ and $\gamma_2$, and $\delta_1$ and $\delta_2$, must
          not vanish simultaneously.
```

Basic properties established by the development of this algorithm are summarized in the following theorems.

Theorem 4.1: Let Algorithm 4.1 be applied to the function f defined by (4.1). If the search directions $d_0, d_1, \ldots, d_{n-1}$ are all nonzero, then

i) $d_0, d_1, \ldots, d_{n-1}$ are conjugate with respect to G, and

ii) $g_n = 0$, that is, the algorithm is quadratically terminating.

Theorem 4.2: Under the hypotheses of Theorem 4.1, $H_n = \sigma G^{-1}$.

Proof: By (4.12) with k = n,

$$H_n G d_j = \sigma d_j, \quad 0 \leq j \leq n - 1,$$

and by i) of Theorem 4.1, the vectors d_j, $0 \leq j \leq n - 1$, are linearly independent. Therefore, $H_n G = \sigma I$, and the theorem is proved.

The general family of variable metric methods given by Algorithm 4.1 contains the DFP method as a special case. It is easily seen that the DFP iteration formula will be obtained if $\sigma = 1$, $\gamma_1 = 1$, $\gamma_2 = 0$, $\delta_1 = 0$, and $\delta_2 = 1$. Therefore, Theorem 4.1 establishes Theorem 2.2.

The similiarities in the developments of Huang's family and Broyden's family suggest a direct relationship. This relationship can be determined by considering the differences between these two families. Since the iteration formula of Algorithm 4.1 can be expressed formally as

$$H_{k+1} = H_k + \sigma \frac{s_k(\gamma s_k + H_k^T y_k)^T}{(\gamma s_k + H_k^T y_k)^T y_k} - \frac{H_k y_k (s_k + \delta H_k^T y_k)^T}{(s_k + \delta H_k^T y_k)^T y_k}, \qquad (4.33)$$

where $\gamma = \gamma_1/\gamma_2$ and $\delta = \delta_2/\delta_1$, Huang's family contains three arbitrary scalar parameters, σ, γ, and δ.

Broyden's family was developed as a quasi-Newton method so that the iteration matrix satisfies the equation

$$H_{k+1}y_k = s_k.$$

By (4.16), Huang's iteration matrix is chosen to satisfy

$$H_{k+1}y_k = \sigma s_k.$$

Therefore, set $\sigma = 1$. Also, Broyden's iteration matrix is symmetric while Huang's matrix is not necessarily symmetric. If H_k is symmetric and $\sigma = 1$, then (4.33) with subscripts omitted, becomes

$$H^* = H + \frac{\gamma ss^T}{\gamma s^T y + y^T Hy} + \frac{sy^T H}{\gamma s^T y + y^T Hy}$$

$$- \frac{Hys^T}{s^T y + \delta y^T Hy} - \frac{\delta Hyy^T H}{s^T y + \delta y^T Hy}. \tag{4.34}$$

This equation implies that, if H is symmetric, then H* will be symmetric provided

$$\frac{1}{\gamma s^T y + y^T Hy} = \frac{-1}{s^T y + \delta y^T Hy},$$

that is, provided

$$\delta = \frac{(\gamma + 1)s^T y}{- y^T Hy} - 1. \tag{4.35}$$

Hence, the conditions that the iteration matrix be symmetric and satisfy the quasi-Newton equation result in an iteration formula based on the one parameter γ.

A comparison of (4.34) and Broyden's iteration formula given by Algorithm 3.1, in particular, the first term of the correction matrix, suggests the relation

$$\beta = \frac{-1}{\gamma s^T y + y^T Hy}.$$

Then,

$$\gamma = \frac{1 + \beta y^T Hy}{-\beta s^T y}$$

and, from (4.35),

$$\delta = \frac{1 - \beta s^T y}{\beta y^T Hy}.$$

Since $\gamma = \gamma_1/\gamma_2$ and $\delta = \delta_2/\delta_1$, let

$$\gamma_1 = 1 + \beta y^T Hy,\ \gamma_2 = -\beta s^T y,$$

$$\delta_1 = \beta y^T Hy, \text{ and } \delta_2 = 1 - \beta s^T y. \tag{4.36}$$

Substitution shows that if $\sigma = 1$ and γ_1, γ_2, δ_1, and δ_2 are given by (4.36), then the iteration formula of Algorithm 4.1 is equivalent to Broyden's formula. Therefore, Broyden's one-parameter family may be characterized as the subset of Huang's family for which the iteration matrix is symmetric and satisfies the quasi-Newton equation.

By Theorem 3.15, all members of Broyden's family generate the same sequence of points, under the conditions of Theorem 3.14. Since Broyden's family is a subset of Huang's family, it is natural to ask whether this result can be extended to Huang's family when applied to a certain class of functions or whether this family can be divided into subsets that generate identical sequences of points when applied to a general differentiable function. Huang shows, as a result of Lemmas 4.1 and 4.2, that for a strictly convex quadratic function, all members of Algorithm 4.1 generate the same sequence of points. These lemmas establish that the search direction d_k, $k \geq 1$, can be expressed as $d_k = \mu_k p_k$, where μ_k is a scalar and p_k is the vector given by (4.31). If p_0 is defined by

$$p_0 = H_0^T g_0, \tag{4.37}$$

then equations (4.31) and (4.37) determine the sequence of search directions p_k, $k \geq 0$. By the same reasoning used for Theorem 3.15, it can be concluded that, for a given initial point x_0 and initial matrix H_0, the sequence of points x_0, x_1, ..., x_n is the same for all the algorithms, that is, it is independent of the parameters σ, γ_1, γ_2, δ_1, and δ_2.

Huang and A. V. Levy [31] test the behavior of some particular algorithms belonging to Huang's family on a quadratic function and several nonquadratic functions. On the quadratic function, the results show that, if high-precision arithmetic and high accuracy in the linear search are used, all the algorithms behave identically for a given initial point and initial matrix. That is, they all produce the same sequence of points and lead to the minimum in no more than n iterations, where n is the number of variables. For the nonquadratic functions, the results show that some of the algorithms tested behave identically. It is concluded that this family could be divided into subsets that also generate identical sequences of points on more general functions.

Dixon [18] establishes a necessary and sufficient condition for members of Huang's family to generate identical sequences of points when applied to the same general nonquadratic function. The same conditions as in Theorem 3.14 are needed to ensure that given x_k and d_k, the value of α_k is uniquely defined, and that d_k is defined and nonzero.

Theorem 4.3: If sequences of points $\{x_k\}$, $k = 0, 1, \ldots$, are generated by Algorithm 4.1 applied to the same differentiable function, starting at a given initial point x_0 with a given initial matrix H_0 and using a perfect linear search at each iteration, then provided degeneracy does not occur, the necessary and sufficient condition for all the sequences to be identical is that the iteration formulas used possess the same value of σ at each iteration.

Proof: If an initial point x_0 and initial matrix H_0 are given, since $d_0 = - H_0^T g_0$, the point x_1 and hence the values of s_0 and y_0 are the same for all members of Huang's family. Since no quadratic properties were used in proving Lemma 4.1, it is also valid for nonquadratic functions. It then follows from (4.27) that p_1 and hence x_2, s_1, and y_1 are the same for all members. Assume that the same points $x_0, x_1, \ldots, x_k$, and hence the same values of $s_0, s_1, \ldots, s_{k-1}$, and $y_0, y_1, \ldots, y_{k-1}$, have been generated. It remains to show that all members of Huang's family generate the same direction p_k, $k > 1$, given by (4.27) if and only if they all possess the same value of σ. In the expression for p_k, the quantity dependent upon the parameters is the vector $H_{k-1}^T g_k$. Hence, the method of proof is to derive a substitution for $H_{k-1}^T g_k$, then $H_{k-2}^T g_k$, and so on, back to $H_0^T g_k$.

If w is an arbitrary vector, then from Algorithm 4.1,

$$H_k^T w = H_{k-1}^T w + \sigma \frac{s_{k-1}^T w}{q_{k-1}^T y_{k-1}} (\gamma_1 s_{k-1} + \gamma_2 H_{k-1}^T y_{k-1})$$

$$- \frac{y_{k-1}^T H_{k-1}^T w}{z_{k-1}^T y_{k-1}} (\delta_1 s_{k-1} + \delta_2 H_{k-1}^T y_{k-1}).$$

Substituting for $H_{k-1}^T y_{k-1}$ from (4.29) gives

$$H_k^T w = H_{k-1}^T w + \sigma \frac{s_{k-1}^T w}{q_{k-1}^T y_{k-1}} \left[\left(\gamma_1 + \frac{\gamma_2}{\alpha_{k-1}} \right) s_{k-1} + \gamma_2 H_{k-1}^T g_k \right]$$

$$- \frac{y_{k-1}^T H_{k-1}^T w}{z_{k-1}^T y_{k-1}} \left[\left(\delta_1 + \frac{\delta_2}{\alpha_{k-1}} \right) s_{k-1} + \delta_2 H_{k-1}^T g_k \right]. \quad (4.38)$$

From Lemma 4.1,

$$H_k^T g_k = - \mu_k \left[H_{k-1}^T g_k - \frac{s_{k-1} y_{k-1}^T H_{k-1}^T g_k}{s_{k-1}^T y_{k-1}} \right]$$

so that

$$H_{k-1}^T g_k = - \frac{1}{\mu_k} H_k^T g_k + \frac{y_{k-1}^T H_{k-1}^T g_k}{s_{k-1}^T y_{k-1}} s_{k-1}. \quad (4.39)$$

Substituting (4.39) into (4.38) and then simplifying by the use of the definitions of q_{k-1}, s_{k-1}, and y_{k-1} gives the expression

$$H_k^T w = H_{k-1}^T w + \sigma s_{k-1} w \left[\frac{s_{k-1}}{s_{k-1}^T y_{k-1}} - \frac{\gamma_2}{\mu_k q_{k-1}^T y_{k-1}} H_k^T g_k \right]$$

$$- y_{k-1}^T H_{k-1}^T w \left[\frac{s_{k-1}}{s_{k-1}^T y_{k-1}} - \frac{\delta_2}{\mu_k z_{k-1}^T y_{k-1}} H_k^T g_k \right].$$

This equation can be written as

$$H_k^T w = \left[I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}} + \psi_k s_k y_{k-1}^T \right] H_{k-1}^T w$$

$$+ \sigma \left[\frac{s_{k-1}}{s_{k-1}^T y_{k-1}} - \phi_k s_k \right] s_{k-1}^T w, \qquad (4.40)$$

where

$$\phi_k = \frac{-\gamma_2}{\mu_k \alpha_k q_{k-1}^T y_{k-1}}, \text{ and } \psi_k = \frac{-\delta_2}{\mu_k \alpha_k z_{k-1}^T y_{k-1}}.$$

Thus, substituting from (4.40) for $H_{k-1}^T g_k$ in the expression for p_k given by (4.27) and then simplifying yields

$$p_k = \left[I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}} \right] \left[I - \frac{s_{k-2}y_{k-2}^T}{s_{k-2}y_{k-2}} \right] H_{k-2}^T g_k$$

$$+ \left[I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}} \right] \left(\sigma \frac{s_{k-2}^T g_k}{s_{k-2}^T y_{k-2}} \right) s_{k-2},$$

since

$$\left[I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}} \right] \psi_{k-1} s_{k-1} y_{k-2}^T H_{k-2}^T g_k = 0, \text{ and}$$

$$\left[I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}} \right] \phi_{k-1} \sigma s_{k-1} s_{k-2}^T g_k = 0.$$

If the substitution from (4.40) for $H_{k-2}^T g_k$ were now made, a similar simplification would occur. This process can be continued back to

$H_0^T g_k$, giving

$$p_k = \left[\prod_{j=0}^{k-1}\left(I - \frac{s_j y_j^T}{s_j^T y_j}\right)\right] H_0^T g_k$$

$$+ \sum_{j=0}^{k-2}\left[\prod_{m=j+1}^{k-1}\left(I - \frac{s_m y_m^T}{s_m^T y_m}\right)\right] \sigma \frac{s_j^T g_k}{s_j^T y_j} s_j. \tag{4.41}$$

Thus, it follows from the induction hypothesis that all members of Huang's family generate the same direction p_k if and only if they all possess the same value of σ, and the proof is complete.

Since Broyden's family is the symmetric subset of Huang's family with $\sigma = 1$, Theorem 3.14 is established by (4.27) of Lemma 4.1 for $k = 1$ and (4.41) of Theorem 4.3 for $k > 1$ with $H_0^T = H_0$. Theorem 3.15, which follows from Theorem 3.14, can also be obtained directly from Theorem 4.3.

In the special case of a positive definite quadratic function, the conjugacy of the search directions implies, by (4.11), that

$$g_k^T s_j = 0, \; 0 \leq j \leq k - 1, \tag{4.42}$$

so that (4.41) reduces to

$$p_k = \left[\prod_{j=0}^{k-1}\left(I - \frac{s_j y_j^T}{s_j^T y_j}\right)\right] H_0^T g_k. \tag{4.43}$$

Expanding the first two factors of the product in brackets gives

$$\left[I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}} \right] \left[I - \frac{s_{k-2}y_{k-2}^T}{s_{k-2}^T y_{k-2}} \right] = I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}} - \frac{s_{k-2}y_{k-2}^T}{s_{k-2}^T y_{k-2}}$$

$$+ \frac{s_{k-1}(y_{k-1}^T s_{k-2})y_{k-2}^T}{(s_{k-1}^T y_{k-1})(s_{k-2}^T y_{k-2})}.$$

The definition of y_{k-1} and (4.42) implies

$$y_{k-1}^T s_j = g_k^T s_j - g_{k-1}^T s_j = 0, \; 0 \le j \le k - 2.$$

Hence, the product of the first two factors of (4.43) reduces to

$$I - \frac{s_{k-1}y_{k-1}^T}{s_{k-1}^T y_{k-1}} - \frac{s_{k-2}y_{k-2}^T}{s_{k-2}^T y_{k-2}}.$$

If the product of this factor and the third factor of (4.43) were now expanded, a similar reduction would occur. This process can be continued until (4.31) is obtained. Thus Lemma 4.2 is established.

Pearson and Adachi

Pearson [42] develops a class of variable metric algorithms which includes the DFP algorithm. The problem considered is to find the minimum of the function f defined by (4.1). His approach is to obtain a class of matrices H_k such that for $d_k = - H_k^T g_k$, the search directions $d_0, d_1, \ldots, d_{n-1}$ are conjugate with respect to G, since this will give quadratic termination. In addition, if n iterations are needed, it is required that $H_n = G^{-1}$.

Suppose the conjugate directions $d_0, d_1, \ldots, d_{k-1}$ have been generated, that is,

$$d_i^T G d_j = 0, \ 0 \leq i < j \leq k - 1. \qquad (4.44)$$

Since $s_i = \alpha_i d_i$, if $\alpha_i \neq 0$, these conjugate directions result in conjugate steps, $s_0, s_1, \ldots, s_{k-1}$, that is,

$$s_i^T G s_j = 0, \ 0 \leq i < j \leq k - 1.$$

Define the n x k matrices

$$Y_k = [y_0, y_1, \ldots, y_{k-1}], \text{ and } S_k = [s_0, s_1, \ldots, s_{k-1}]. \qquad (4.45)$$

For f given by (4.1),

$$y_i = Gs_i \qquad (4.46)$$

so that

$$G^{-1} Y_k = S_k. \qquad (4.47)$$

So, suppose H_k is a matrix satisfying

$$H_k Y_k = S_k. \qquad (4.48)$$

By the proof of Theorem 2.3, (4.44) implies $d_i^T g_k = 0$, $0 \leq i < k$, which by the definition of s_i, implies $s_i^T g_k = 0$, $0 \leq i < k$, that is,

$$S_k^T g_k = 0. \qquad (4.49)$$

Then by the definitions of s_k and d_k, and by (4.48),

$$\begin{aligned} Y_k^T s_k &= - \alpha_k Y_k^T H_k^T g_k \\ &= - \alpha_k S_k^T g_k \\ &= 0, \end{aligned} \qquad (4.50)$$

that is, $y_j^T s_k = 0$, $0 \leq j \leq k - 1$, so that by (4.46),

$$s_j^T G s_k = 0, \quad 0 \leq j \leq k - 1. \tag{4.51}$$

Therefore, if $s_k \neq 0$, the new step s_k is conjugate to the previous ones. Equation (4.51) also implies, by (4.46), that $s_j^T y_k = 0$, $0 \leq j \leq k - 1$, that is,

$$S_k^T y_k = 0. \tag{4.52}$$

In addition, for $k = n$, (4.49) gives $S_n^T g_n = 0$ and (4.48) and (4.47) give

$$S_n = H_n Y_n = H_n G S_n$$

which implies $g_n = 0$ and $H_n = G^{-1}$ if S_n is nonsingular, that is, if s_0, s_1, ..., s_{n-1} are all nonzero. Note that if H_k satisfies (4.48) for every k, then $H_{k+1} Y_{k+1} = S_{k+1}$ implies, by the definitions of Y_{k+1} and S_{k+1}, that $H_{k+1} y_k = s_k$, that is, the iteration matrix satisfies the quasi-Newton equation.

To obtain a general solution of (4.48), the following lemma is needed. This lemma is established by Theorem 2 of [43].

Lemma 4.3: A necessary and sufficient condition for the matrix equation

$$CXD = E$$

to have a solution is that

$$CC^+ED^+D = E,$$

where C^+ and D^+ are matrices which satisfy the relations

$$CC^+C = C, \text{ and } DD^+D = D.$$

In this case, the general solution of the equation is

$$X = C^{+}ED^{+} + Y - C^{+}CYDD^{+},$$

where Y is an arbitrary matrix of the same size as X.

Applying this lemma, the general solution of (4.48) is

$$H_k = S_k Y_k' + R(I - Y_k Y_k''), \tag{4.53}$$

where R is an arbitrary n x n matrix and Y_k' and Y_k'' are k x n matrices which satisfy the relations

$$Y_k Y_k' Y_k = Y_k, \text{ and } Y_k Y_k'' Y_k = Y_k. \tag{4.54}$$

The condition that the equation (4.48) be solvable is that $S_k Y_k' Y_k = S_k$. By (4.47) and (4.54),

$$\begin{aligned} S_k Y_k' Y_k &= G^{-1} Y_k Y_k' Y_k \\ &= G^{-1} Y_k \\ &= S_k, \end{aligned}$$

so this condition is always satisfied.

Pearson restricts R to be a positive definite matrix and Y_k' and Y_k'' to have the form $(Y_k^T M Y_k)^{-1} Y_k^T M$ for positive definite matrices $M = G^{-1}$ or $M = R$, independently for each term. This leads to a class of four algorithms. Given the vector x_k and gradient $g_k \neq 0$, the k-th iteration of the general algorithm sets $d_k = - H_k^T g_k$, where H_k, $k \geq 1$, is defined by (4.53) with Y_k' and Y_k'' as specified above, and $H_0 = R$. Then the vector $x_{k+1} = x_k + s_k$, where $s_k = \alpha_k d_k$ with α_k chosen to minimize $f(x_k + \alpha d_k)$ with respect to α. Each of the four algorithm, if applied to the function f defined by (4.1), will find the minimum in at most n iterations. Also, if n iterations are required, then $H_n = G^{-1}$. These properties are established by Theorem 2 of [42]. Particular algorithms are obtained by alternate choices of M in Y_k' and Y_k''. Recursion formulas for H_{k+1} in terms of H_k, y_k, and s_k can then be found for three of these algorithms by applying a recursion formula established in Appendix A of [42] and using (4.50) and (4.52). The DFP formula is obtained by substituting $M = G^{-1}$ in Y_k' and $M = R$ in Y_k''. Based on the same idea, Adachi [1] develops a general variable metric algorithm. However, he obtains a more general recursion formula for H_k given by (4.53) and Y_k' and Y_k'' satisfying (4.54). This recursion formula includes those derived by Pearson.

Define the k x n matrix E_{1k} and n-vector e_{1k} by

$$E_{1k} = \begin{cases} \dfrac{- Y_k' y_k b_{1k}^T (I - Y_k Y_k')}{b_{1k}^T (I - Y_k Y_k') y_k}, & \text{if } I - Y_k Y_k' \neq 0, \\ 0, & \text{if } I - Y_k Y_k' = 0, \end{cases}$$

$$e_{1k}^T = \begin{cases} \dfrac{c_{1k}^T (I - Y_k Y_k')}{c_{1k} (I - Y_k Y_k') y_k}, & \text{if } I - Y_k Y_k' \neq 0, \\ 0, & \text{if } I - Y_k Y_k' = 0, \end{cases} \qquad (4.55)$$

where the vectors b_{1k} and c_{1k} are such that

$$b_{1k}^T (I - Y_k Y_k') y_k \neq 0, \text{ and } c_{1k}^T (I - Y_k Y_k') y_k \neq 0.$$

Then the following lemma defines recursively a matrix Y_k' which satisfies equation (4.54).

Lemma 4.4: The k x n matrices Y'_k, $k \geq 1$, defined recursively by

$$Y'_1 = \frac{c_{10}^T}{c_{10}^T y_0}, \text{ and } Y'_{k+1} = \begin{bmatrix} Y'_k \\ 0 \end{bmatrix} + \begin{bmatrix} E_{1k} \\ e_{1k}^T \end{bmatrix}, \ k \geq 1,$$

satisfy the relation

$$Y_k Y'_k Y_k = Y_k, \ k = 1, 2, \ldots. \tag{4.56}$$

Proof: The proof is by induction on k. Clearly, by the definitions of Y_1 and Y'_1, (4.56) is true for k = 1. Assume (4.56) is true. Then

$$(I - Y_k Y'_k) Y_k = 0. \tag{4.57}$$

Using the definitions of Y_{k+1} and Y'_{k+1},

$$Y_{k+1} Y'_{k+1} Y_{k+1} = [Y_k \quad y_k] \begin{bmatrix} Y'_k Y_k + E_{1k} y_k & Y'_k y_k + E_{1k} y_k \\ e_{1k}^T Y_k & e_{1k}^T y_k \end{bmatrix}$$

which, by (4.57) and the definitions of E_{1k} and e_{1k}^T, reduces to

$$Y_{k+1} Y'_{k+1} Y_{k+1} = [Y_k \quad y_k] \begin{bmatrix} Y'_k Y_k & 0 \\ 0 & 1 \end{bmatrix} = [Y_k Y'_k Y_k \quad y_k].$$

Then, by the induction hypothesis and the definition of Y_{k+1}, it follows that (4.56) is true for k replaced by k + 1.

Let Y'_k and Y''_k in (4.53) be defined by the recursive formulas

$$Y'_1 = \frac{c_{10}^T}{c_{10}^T y_0}, \ Y'_{k+1} = \begin{bmatrix} Y'_k + E_{1k} \\ e_{1k}^T \end{bmatrix}, \ k \geq 1,$$

$$Y''_1 = \frac{c_{20}^T}{c_{20}^T y_0}, \quad Y''_{k+1} = \begin{bmatrix} Y''_k + E_{2k} \\ \\ e_{2k}^T \end{bmatrix}, \quad k \geq 1, \tag{4.58}$$

with E_{1k} and e_{1k}^T given by (4.55) and E_{2k} and e_{2k}^T given by

$$E_{2k} = \begin{cases} \dfrac{- Y''_k y_k b_{2k}^T (I - Y_k Y''_k)}{b_{2k}^T (I - Y_k Y''_k) y_k}, & \text{if } I - Y_k Y''_k \neq 0, \\ \\ 0, & \text{if } I - Y_k Y''_k = 0, \end{cases}$$

$$e_{2k}^T = \begin{cases} \dfrac{c_{2k}^T (I - Y_k Y''_k)}{c_{2k}^T (I - Y_k Y''_k) y_k}, & \text{if } I - Y_k Y''_k \neq 0, \\ \\ 0, & \text{if } I - Y_k Y''_k = 0, \end{cases}$$

where the vectors b_{2k} and c_{2k} are such that

$$b_{2k}^T (I - Y_k Y''_k) y_k \neq 0, \text{ and } c_{2k}^T (I - Y_k Y''_k) y_k \neq 0.$$

Then, by (4.53),

$$\begin{aligned} H_{k+1} &= S_{k+1} Y'_{k+1} + R(I - Y_{k+1} Y''_{k+1}) \\ &= [S_k Y'_k + S_k E_{1k} + s_k e_{1k}^T] + R(I - [Y_k Y''_k + Y_k E_{2k} + y_k e_{2k}^T]) \\ &= H_k + \frac{s_k c_{1k}^T (I - Y_k Y'_k)}{c_{1k}^T (I - Y_k Y'_k) y_k} - \frac{S_k Y'_k y_k b_{1k}^T (I - Y_k Y'_k)}{b_{1k}^T (I - Y_k Y'_k) y_k} \\ &\quad - R \left[\frac{y_k c_{2k}^T (I - Y_k Y''_k)}{c_{2k}^T (I - Y_k Y''_k) y_k} - \frac{Y_k Y''_k y_k b_{2k}^T (I - Y_k Y''_k)}{b_{2k}^T (I - Y_k Y''_k) y_k} \right]. \end{aligned}$$

Denote the matrices $I - Y_k Y'_k$ and $I - Y_k Y''_k$ by A_k and B_k, respectively. Then,

$$A_{k+1} = I - [Y_k Y_k' + Y_k E_{1k} + y_k e_{1k}^T]$$

$$= A_k - Y_k E_{1k} - y_k e_{1k}^T.$$

Similarly,

$$B_{k+1} = B_k - Y_k E_{2k} - y_k e_{2k}^T.$$

Also, by (4.53),

$$S_k Y_k' = H_k - RB_k. \qquad (4.59)$$

Therefore, the recursion formula for H_k can be expressed as

$$H_{k+1} = H_k + \frac{s_k c_{1k}^T A_k}{c_{1k}^T A_k y_k} - \frac{(H_k - RB_k) y_k b_{1k}^T A_k}{b_{1k}^T A_k y_k} - R\left[\frac{y_k c_{2k}^T B_k}{c_{2k}^T B_k y_k} - \frac{(I - B_k) y_k b_{2k}^T B_k}{b_{2k}^T B_k y_k}\right], \qquad (4.60)$$

where

$$A_{k+1} = A_k + \frac{(I - A_k) y_k b_{1k}^T A_k}{b_{1k}^T A_k y_k} - \frac{y_k c_{1k}^T A_k}{c_{1k}^T A_k y_k}, \text{ and} \qquad (4.61)$$

$$B_{k+1} = B_k + \frac{(I - B_k) y_k b_{2k}^T B_k}{b_{2k}^T B_k y_k} - \frac{y_k c_{2k}^T B_k}{c_{2k}^T B_k y_k}, \qquad (4.62)$$

where the vectors b_{1k}, c_{1k}, b_{2k}, and c_{2k} are chosen so that

$$b_{1k}^T A_k y_k \neq 0, \; c_{1k}^T A_k y_k \neq 0,$$

$$b_{2k}^T B_k y_k \neq 0, \text{ and } c_{2k}^T B_k y_k \neq 0.$$

This recursion formula is used to define the following general variable metric algorithm.

Algorithm 4.2 (Adachi, 1971): Given an initial vector x_0 and initial symmetric matrices $H_0 = R$, $A_0 = I$, and $B_0 = I$.

```
For $k = 0, 1, 2, \ldots,$
    If $g_k = g(x_k) = 0$, then stop.
    Else, set $d_k = - H_k^T g_k$,
        find $\alpha_k$ which minimizes $f(x_k + \alpha d_k)$ with respect to $\alpha$,
        set $s_k = \alpha_k d_k$,
            $x_{k+1} = x_k + s_k$,
            $y_k = g_{k+1} - g_k$,
        update $H_k$ by equations (4.60)-(4.62).
```

The properties of this algorithm established by the discussion preceding Lemma 4.3 are summarized in the following theorem.

Theorem 4.4: Let Algorithm 4.2 be applied to the function f defined by (4.1). If the vectors $s_0, s_1, \ldots, s_{n-1}$ are all nonzero, then

i) $s_0, s_1, \ldots, s_{n-1}$ are conjugate with respect to G,

ii) $g_n = 0$, that is, the algorithm is quadratically terminating, and

iii) $H_n = G^{-1}$, that is, the algorithm is exact.

Particular algorithms are obtained from the general variable metric algorithm given by Algorithm 4.2 by appropriate choices for the vectors b_{1k}, c_{1k}, b_{2k}, and c_{2k}. If equations (4.50) and (4.52) are then applied to the resulting iteration formula, various known formulas, including the DFP formula, can be derived. Since these equations depend on

Algorithm 4.2 being applied to f defined by (4.1) and are not true in general, this relationship holds only in this case. However, another general algorithm can be obtained from Algorithm 4.2 by choosing b_{1k}, c_{1k}, b_{2k}, and c_{2k} in (4.60) as linear combinations of s_k and $H_k^T y_k$ and then applying equations (4.50) and (4.52). Let

$$\begin{aligned} b_{jk} &= \gamma_{jk} s_k + \delta_{jk} H_k^T y_k, \text{ and} \\ c_{jk} &= \phi_{jk} s_k + \psi_{jk} H_k^T y_k, \ j = 1, 2, \end{aligned} \tag{4.63}$$

where γ_{jk}, δ_{jk}, ϕ_{jk}, and ψ_{jk}, $j = 1, 2$, are scalars. Since $H_k Y_k = S_k$, by (4.50) and (4.52),

$$c_{1k}^T A_k = (\phi_{1k} s_k + \psi_{1k} H_k^T y_k)^T - \phi_{1k} s_k^T Y_k Y_k' - \psi_{1k} y_k^T H_k Y_k Y_k' = c_{1k}^T. \tag{4.64}$$

Similarly,

$$\begin{aligned} &b_{1k}^T A_k = b_{1k}^T, \ c_{2k}^T B_k = c_{2k}^T, \text{ and} \\ &b_{2k}^T B_k = b_{2k}^T. \end{aligned} \tag{4.65}$$

Hence the recursion formula for H_k given by (4.60)-(4.62) reduces to

$$\begin{aligned} H_{k+1} = H_k &+ \frac{s_k c_{1k}^T}{c_{1k}^T y_k} - \frac{(H_k - RB_k) y_k b_{1k}^T}{b_{1k}^T y_k} \\ &- R \left[\frac{y_k c_{2k}^T}{c_{2k}^T y_k} - \frac{(I - B_k) y_k b_{2k}^T}{b_{2k}^T y_k} \right], \end{aligned} \tag{4.66}$$

where

$$B_{k+1} = B_k + \frac{(I - B_k) y_k b_{2k}^T}{b_{2k}^T y_k} - \frac{y_k c_{2k}^T}{c_{2k}^T y_k} \tag{4.67}$$

This leads to Algorithm 4.2' in which equations (4.63), (4.66), and (4.67) are used in Algorithm 4.2 instead of equations (4.60)-(4.62).

The DFP algorithm is derived from Algorithm 4.2' by letting $b_{1k} = c_{1k} = s_k$ and $b_{2k} = c_{2k} = H_k^T y_k$ in (4.66) and choosing R to be positive definite. Equation (4.66) then reduces to

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k + RB_k y_k) s_k^T}{s_k^T y_k} - \frac{RB_k y_k y_k^T H_k}{y_k^T H_k y_k}.$$

Substituting for RB_k given by (4.59) yields

$$H_{k+1} = H_k + \frac{(s_k - S_k Y_k' y_k) s_k^T}{s_k^T y_k} - \frac{(H_k y_k - S_k Y_k' y_k) y_k^T H_k}{y_k^T H_k y_k}. \qquad (4.68)$$

In this case, it can be shown by induction that the matrices $S_k Y_k'$, $k \geq 1$, are symmetric. Clearly, by the definitions of S_1 and Y_1',

$$S_1 Y_1' = \frac{s_0 s_0^T}{s_0^T y_0}$$

is symmetric. Assume $S_k Y_k'$ is symmetric. By the appropriate definitions and equations (4.64) and (4.65),

$$S_{k+1} Y_{k+1}' = \begin{bmatrix} S_k & s_k \end{bmatrix} \begin{bmatrix} Y_k' - \dfrac{Y_k y_k s_k^T}{s_k^T y_k} \\ \\ \dfrac{s_k^T}{s_k^T y_k} \end{bmatrix} = S_k Y_k' - \frac{S_k Y_k' y_k s_k^T + s_k s_k^T}{s_k^T y_k}.$$

The symmetry of $S_k Y_k'$ and (4.52) imply

$$(S_k Y_k' y_k)^T = y_k^T S_k Y_k' = 0. \qquad (4.69)$$

Thus, $S_{k+1}Y'_{k+1}$ is symmetric and the induction is complete. Equation (4.69) then reduces (4.68) to the DFP formula.

Adachi [2] proves that, given the same initial point x_0 and initial matrix $H_0 = R$, the sequence of points $x_0, x_1, \ldots,$ generated by Algorithm 4.2 with b_{jk} and c_{jk}, $j = 1, 2$, defined by (4.63) is uniquely determined, if it is defined in fact, independently of the parameters γ_{jk}, δ_{jk}, ϕ_{jk}, and ψ_{jk}, $j = 1, 2$, when the algorithm is applied to the quadratic function defined by (4.1). It is first shown, under these conditions, that given a matrix H_k, the $(k + 1)$-st search direction is uniquely determined independently of the parameters. This result is established by the following theorem which shows that the parameters affect only the magnitude, not the direction, of d_{k+1}.

Theorem 4.5: If Algorithm 4.2 with b_{jk} and c_{jk}, $j = 1, 2$, given by (4.63) is applied to the function f defined by (4.1), then the search direction

$$d_{k+1} = - H^T_{k+1}g_{k+1}$$

can be expressed as

$$d_{k+1} = \theta_{k+1}p_{k+1},$$

where θ_{k+1} is a scalar which depends on the parameters γ_{jk}, δ_{jk}, ϕ_{jk}, and ψ_{jk}, $j = 1, 2$, and p_{k+1} is the vector defined by

$$p_{k+1} = \left[I - \frac{s_k y_k^T}{s_k^T y_k} \right] H_k^T g_{k+1}.$$

Proof: To simplify notation, the subscript k will be omitted and the subscript k + 1 will be denoted by the superscript *. Under the stated hypotheses, the iteration formula for H_k used in Algorithm 4.2 given by (4.60) reduces to the formula given by (4.66). Using this formula,

$$H^{*T}g^* = H^Tg^* + \frac{c_1s^Tg^*}{c_1^Ty} - \frac{b_1y^T(H - RB)^Tg^*}{b_1^Ty} - \left[\frac{c_2y^T}{c_2^Ty} - \frac{b_2y^T(I - B)^T}{b_2^Ty}\right]R^Tg^*$$

$$= \left[H^Tg^* - \frac{b_2y^TH^Tg^*}{b_2^Ty}\right] + y^TR^Tg^*\left[\frac{b_2}{b_2^Ty} - \frac{c_2}{c_2^Ty}\right]$$

$$+ y^T(H - RB)^Tg^*\left[\frac{b_2}{b_2^Ty} - \frac{b_1}{b_1^Ty}\right]. \qquad (4.70)$$

Substituting for b_2 given by (4.63) and using the definitions of y and s, the first term in the right hand side of (4.70) may be expressed as

$$H^Tg^* - \frac{b_2y^TH^Tg^*}{b_2^Ty} = H^Tg^* - \frac{\gamma_2sy^TH^Tg^*}{b_2^Ty} - \frac{\delta_2H^Tg^*y^TH^Tg^*}{b_2^Ty} - \frac{\delta_2sy^TH^Tg^*}{\alpha b_2^Ty}$$

$$= \left[1 - \frac{\delta_2y^TH^Tg^*}{b_2^Ty}\right]H^Tg^* - \left[\frac{\gamma_2s^Ty}{b_2^Ty} + \frac{\delta_2s^Ty}{\alpha b_2^Ty}\right]\frac{sy^TH^Tg^*}{s^Ty}.$$

Since

$$\frac{\gamma_2s^Ty}{b_2^Ty} + \frac{\delta_2s^Ty}{\alpha b_2^Ty} = \frac{\gamma_2s^Ty}{b_2^Ty} - \frac{\delta_2y^T[\alpha H^T(g^* - y)]}{\alpha b_2^Ty} = 1 - \frac{\delta_2y^TH^Tg^*}{b_2^Ty},$$

the first term of (4.70) is a scalar multiple of the vector p*. Similarly, by substituting for b_2 and c_2, the second term may be expressed as

$$y^TR^Tg^*\left[\frac{b_2}{b_2^Ty} - \frac{c_2}{c_2^Ty}\right] = \frac{(y^TR^Tg^*)(s^Ty)(\delta_2\phi_2 - \psi_2\gamma_2)}{(b_2^Ty)(c_2^Ty)}\left[H^Ty - \frac{sy^TH^Ty}{s^Ty}\right].$$

Using the definitions of y and s, the factor in square brackets in the right hand side reduces to

$$H^Tg^* - H^Tg - \frac{sy^TH^Tg^*}{s^Ty} - \frac{sy^TH^Tg}{\alpha y^TH^Tg} = H^Tg^* - \frac{sy^TH^Tg^*}{s^Ty},$$

so that the second term in (4.70) is also a multiple of p*. Since equations (4.49) and (4.52) are valid for the given function f, and H - RB = SY' by (4.59),

$$g^{*T}(H - RB) = (g^T + y^T)SY' = 0$$

and the third term of (4.70) is zero. Hence, by (4.70),

$$d^* = - H^{*T}g^* = \theta^*p^*$$

for an appropriate scalar θ*.

<u>Theorem</u> 4.6: Under the hypothesis of Theorem 4.5, if the vectors $s_0, s_1, \ldots, s_k$ are defined and nonzero, then the (k + 1)-st search direction p_{k+1} defined by Theorem 4.5 may be expressed as

$$p_{k+1} = \left[I - \sum_{r=0}^{k} \frac{s_ry_r^T}{s_r^Ty_r} \right] Rg_{k+1}. \qquad (4.71)$$

<u>Proof</u>: From Theorem 4.5,

$$p_{k+1} = H_k^Tg_{k+1} - s_k\left(\frac{y_k^TH_k^Tg_{k+1}}{s_k^Ty_k} \right). \qquad (4.72)$$

Equation (4.70) with k replaced by k - 1 and then g_k replaced by g_{k+1} implies that $H_k^Tg_{k+1}$ is equal to $H_{k-1}^Tg_{k+1}$ plus a linear combination of the vectors $b_{1,k-1}$, $b_{2,k-1}$, and $c_{2,k-1}$. Each of these vectors as defined by (4.63) is a linear combination of s_{k-1} and $H_{k-1}^T y_{k-1}$. Using the definitions of y_k and s_k, for an appropriate scalar σ,

$$H_k^T y_k = \left[I - \frac{s_k y_k^T}{s_k^T y_k} \right] H_k^T g_{k+1} + s_k \left(\frac{y_k^T H_k^T g_{k+1}}{s_k^T y_k} \right) - H_k^T g_k = p_{k+1} + \sigma s_k .$$

By Theorem 4.5, p_{k+1} is a scalar multiple of d_{k+1} which is a multiple of s_{k+1}. Thus, $H_k^T y_k$ is a linear combination of s_k and s_{k+1}. It then follows from (4.72) that

$$p_{k+1} = H_{k-1}^T g_{k+1} + \sigma_k s_k + \sigma_{k-1} s_{k-1} \tag{4.73}$$

for appropriate scalars σ_k and σ_{k-1}. Since p_{k+1} is a multiple of s_{k+1} and, for the given function f, $y_k = G s_k$ and $y_{k-1} = G s_{k-1}$, if $s_{k+1} \neq 0$ then the conjugacy of the vectors $s_0, s_1, \ldots, s_{k+1}$ implies

$$y_k^T p_{k+1} = 0, \text{ and } y_{k-1}^T p_{k+1} = 0.$$

These equations are also true if $s_{k+1} = 0$. Therefore, by (4.73),

$$y_k^T H_{k-1}^T g_{k+1} + \sigma_k s_k^T y_k = 0, \text{ and}$$

$$y_{k-1}^T H_{k-1}^T g_{k+1} + \sigma_{k-1} s_{k-1}^T y_{k-1} = 0.$$

Solving the above equations for σ_k and σ_{k-1}, respectively, and substituting into (4.73) gives

$$p_{k+1} = \left[I - \frac{s_{k-1} y_{k-1}^T}{s_{k-1}^T y_{k-1}} - \frac{s_k y_k^T}{s_k^T y_k} \right] H_{k-1}^T g_{k+1} .$$

This same procedure may be repeated until (4.71) is obtained.

The following corollary is the result of (4.49) and Theorem 4.6. If the initial matrix R is positive definite or negative definite, it implies that the algorithms are stable for positive definite quadratic functions.

Corollary 4.1: Under the hypothesis of Theorem 4.6,

$$g_k^T p_k = g_k^T H_{k-1}^T g_k = \cdot \cdot \cdot \cdot = g_k^T R g_k .$$

Since Algorithm 4.2 with the parameters given by (4.63) reduces to Algorithm 4.2' when applied to the quadratic function f defined by (4.1), Theorems 4.5 and 4.6 and Corollary 4.1 are also valid for Algorithm 4.2'. It follows from Theorem 4.6 that, for a given initial matrix R, all the particular algorithms derived by Algorithm 4.2 with the parameters given by (4.63) or by Algorithm 4.2' generate a unique sequence of search directions p_0, p_1, ..., and a corresponding unique sequence of points x_0, x_1, ... However, Theorem 4.6 does not imply that the minimum x' of the function f defined by (4.1) is reached by all of these algorithms after at most n iterations, only that if the point x' is obtained by these algorithms after n iterations for a given initial point x_0 and an initial matrix R, then the sequence x_0, x_1, ..., x_{n-1}, x' is the same for all the algorithms. Some algorithms may stop at a nonstationary point or may not be defined at a certain step of the iterations.

Algorithms 4.2 and 4.2' may be applied to a nonquadratic differentiable function. However, the proofs of Theorems 4.5 and 4.6 do not hold in general since the quadratic properties of the function were used. In the proof of Theorem 4.5, quadratic properties are used only to reduce the recursion formula for H_k given by (4.60) used in Algorithm 4.2 to that given by (4.66) used in Algorithm 4.2' and to show that $(H_k - RB_k)^T g_{k+1} = 0$, so that the third term in the right hand side of (4.70) is zero. But, by the same method as used in the second term, this term may be expressed as

$$y^T(H - RB)^T g* \left[\frac{b_2}{b_2^T y} - \frac{b_1}{b_1^T y} \right] = \frac{y^T(H - RB)^T g*(\gamma_1\delta_2 - \gamma_2\delta_1)}{(b_2^T y)(b_1^T y)} \left[I - \frac{sy^T}{s^T y} \right] H^T g*.$$

Therefore, Theorem 4.5 is valid for Algorithm 4.2' applied to non-quadratic functions.

Theorem 4.7: If Algorithm 4.2' is applied to the differentiable function f, then the search direction d_{k+1} can be expressed as

$$d_{k+1} = \theta'_{k+1} p_{k+1},$$

where θ'_{k+1} is a scalar which depends on the parameters γ_{jk}, δ_{jk}, ϕ_{jk}, and ψ_{jk}, $j = 1, 2$, and p_{k+1} is the vector defined by Theorem 4.5.

The relationship between Adachi's general family of variable metric algorithms given by Algorithm 4.2 or Algorithm 4.2' and Huang's family given by Algorithm 4.1 can be determined by comparing the criterion used to derive the iteration matrix H_k. For Huang's family, the matrix H_k is chosen to satisfy (4.41), that is,

$$H_k y_j = \sigma s_j, \; 0 \leq j \leq k - 1.$$

Using the definitions of the matrices Y_k and S_k given by (4.45), this equation is equivalent to the equation $H_k Y_k = \sigma S_k$. Hence, for $\sigma = 1$, Huang's iteration matrix satisfies the equation

$$H_kY_k = S_k. \tag{4.74}$$

Adachi's iteration matrix is the general solution of this equation, given by (4.53), that is, $H_k = S_kY_k' + R(I - Y_kY_k'')$, where Y_k' and Y_k'' are defined by (4.58) and satisfy $Y_kY_k'Y_k = Y_k$ and $Y_kY_k''Y_k = Y_k$. Since $R(I - Y_kY_k'')Y_k = 0$,

$$H_k = S_kY_k' \tag{4.75}$$

is a particular solution of (4.74). Applying the method used by Adachi to obtain recursion formula (4.60) for the general solution (4.53), to the particular solution (4.75), yields

$$H_{k+1} = H_k + \frac{s_kc_{1k}^TA_k}{c_{1k}^TA_ky_k} - \frac{H_ky_kb_{1k}^TA_k}{b_{1k}^TA_ky_k}, \tag{4.76}$$

where A_k is given by (4.62). If

$$c_{1k} = \gamma_1s_k + \gamma_2H_k^Ty_k, \text{ and}$$

$$b_{1k} = \delta_1s_k + \delta_2H_k^Ty_k,$$

then, by (4.64) and (4.65), equation (4.76) reduces to the general iteration formula used by Huang's family in Algorithm 4.1. Therefore, in the case of $\sigma = 1$, Huang's general family can be obtained from a particular solution of $H_kY_k = S_k$, while Adachi's general family is derived from the general solution of this equation. In this sense, Theorem 4.5, Theorem 4.6, and Theorem 4.7 are generalizations of Lemma 4.1 and Lemma 4.2. However, a result corresponding to Theorem 4.3 has not been proved.

CHAPTER V

SUMMARY

This paper is an expository study of Fletcher and Powell's version of Davidon's original variable metric method and generalizations of this method, that is, parametric families of variable metric methods which contain the DFP method and have basic properties in common with this method. The main emphasis has been on the motivation and basic ideas leading to their development and on the theoretical properties which form the foundation of these methods.

Davidon's variable metric method introduced a variable metric into the direction of steepest descent, leading to the search direction $d_k = - H_k g_k$, where the variable matrix H_k approximates the inverse Hessian matrix at the point x_k. The basic concepts of this method were discussed in Chapter I. Fletcher and Powell simplified this method and established the properties of quadratic termination and exactness. That is, for a quadratic function f of n variables with positive definite Hessian matrix G, $g_n = 0$ and $H_n = G^{-1}$, if n iterations are required. In addition, they proved that the method was stable by showing that H_k was positive definite for each k. Powell's general convergence theorem extended convergence to convex functions. Chapter II, which covered the DFP method, concluded with a discussion and possible explanation of the numerical difficulties encountered with this method.

The first parametric family, the topic of Chapter III, was developed by Broyden as a quasi-Newton method. This one-parameter family was derived by modifying H_k so that the quasi-Newton equation, $H_{k+1}y_k = s_k$, is satisfied and $H_n = G^{-1}$ for a quadratic function f with positive definite Hessian matrix G in order to obtain finite termination. Symmetry of H_k was also required. A range on the parameter β_k which guarantees stability was established. Shanno's development of the same iteration formula was also based on the quasi-Newton equation. However, his formulation extended the range of β_k which ensures stability. The development by Goldfarb showed that the correction matrix could be expressed as a combination of two correction matrices of minimum norm obtained from a formula derived by Greenstadt. Fletcher's derivation of this same family showed that any member differed from the DFP matrix by a matrix of rank one. The analysis of Broyden and Shanno in their search for an optimal parameter led to the complementary DFP formula. Dixon's theorem extended Powell's convergence theorem to other members of this family. Table III gives the different formulations of the iteration formula for the variable matrix H. Table IV summarizes the relationships among the different formulations and gives the values of the parameters leading to the DFP formula and the complementary DFP formula.

TABLE III

FORMULATIONS OF THE ONE-PARAMETER FAMILY

Author	Iteration Formula
Broyden	$H^* = H + sq^T - Hyz^T$ $q^T = \frac{(1 + \beta y^T Hy)s^T}{s^T y} - \beta y^T H$ $z^T = \frac{(1 - \beta s^T y)y^T H}{y^T Hy} + \beta s^T$
Shanno	$H^* = H + \frac{\tau ss^T}{s^T y} + \frac{[(1 - \tau)s - Hy][(1 - \tau)s - Hy]^T}{[(1 - \tau)s - Hy]^T y}$
Goldfarb	$H^* = H + \gamma C_H + (1 - \gamma)C_{H*}$ $C_H = \frac{1}{y^T Hy}\left[sy^T H + Hys^T - \left(1 + \frac{s^T y}{y^T Hy} \right) Hyy^T H \right]$ $C_{H*} = \frac{1}{s^T y}\left[- sy^T H - Hys^T + \left(1 + \frac{y^T Hy}{s^T y} \right) ss^T \right]$
Fletcher	$H^* = (1 - \phi)H^*_{DFP} + \phi H^*_{DFP'}$ $H^*_{DFP} = H + \frac{ss^T}{s^T y} - \frac{Hyy^T H}{y^T Hy}$ $H^*_{DFP'} = H + C_{H*}$

TABLE IV

VALUES OF PARAMETERS LEADING TO ONE-PARAMETER FAMILY AND PARTICULAR ALGORITHMS

Author	Broyden	DFP	DFP'
Broyden	——	$\beta = 0$	$\beta = \frac{1}{s^T y}$
Shanno	$\tau = 1 + \frac{\beta y^T H y}{1 - \beta s^T y}$	$\tau = 1$	$\tau = \infty$
Goldfarb	$\gamma = \frac{(1 - \beta s^T y) y^T H y}{y^T H y + s^T y}$	$\gamma = \frac{y^T H y}{y^T H y + s^T y}$	$\gamma = 0$
Fletcher	$\phi = \beta s^T y$	$\phi = 0$	$\phi = 1$

The general families of Chapter IV were obtained by Huang, Pearson, and Adachi by not restricting H_k to be symmetric. In this case, the search direction was given by $d_k = - H_k^T g_k$. Since Huang's objective was to develop quadratically terminating algorithms, the variable matrix H_k was chosen so that, for a quadratic function f with positive definite Hessian matrix G, directions conjugate with respect to G would be generated. Adachi's family was based on the fact that, for a quadratic function f with positive definite Hessian matrix G, the directions would conjugate with respect to G and H_n would be equal to G^{-1} if the variable matrix H_k was a general solution of $H_k Y_k = S_k$. Huang, Dixon, and Adachi showed that these general families could be classified on the basis of identical behavior on certain classes of functions. The relationships among the parametric families of Broyden, Huang, and Adachi are summarized in Table V.

TABLE V

RELATIONSHIPS AMONG PARAMETRIC FAMILIES

	Huang ($\sigma = 1$)	Broyden
Huang		$\sigma = 1$
$H^* = H + \sigma \dfrac{s(\gamma_1 s + \gamma_2 H^T y)^T}{(\gamma_1 s + \gamma_2 H^T y)^T y}$ $- \dfrac{Hy(\delta_1 s + \delta_2 H^T y)^T}{(\delta_1 s + \delta_2 H^T y)^T y}$	——	$\gamma_1 = 1 + \beta y^T Hy$ $\gamma_2 = - \beta s^T y$ $\delta_1 = \beta y^T Hy$ $\delta_2 = 1 - \beta s^T y$
Adachi (using $H = SY'$)	$c_1 = \gamma_1 s + \gamma_2 H^T y$	$c_1 = (1 + \beta y^T Hy)s$ $+ (- \beta s^T y)H^T y$
$H^* = H + \dfrac{sc_1^T}{c_1^T y} - \dfrac{Hyb_1^T}{b_1^T y}$	$b_1 = \delta_1 s + \delta_2 H^T y$	$b_1 = (\beta y^T Hy)s$ $+ (1 - \beta s^T y)H^T y$

Table VI summarizes the basic properties of the DFP method and the conditions under which the parametric families studied also possess these properties.

TABLE VI

BASIC PROPERTIES OF THE DFP METHOD AND PARAMETRIC FAMILIES WHICH CONTAIN THIS METHOD

Property	DFP	Broyden	Huang	Adachi
Conjugate direction method	X	X*	X*	X*
Quadratically terminating	X	X*	X*	X*
Quasi-Newton method	X	X	$\sigma = 1$	X
Exact	X	X*	$\sigma = 1$*	X*
Stable[1]	X	X***	X**	X**

[1] provided H_0 is positive definite

*provided degeneracy does not occur

**for positive definite quadratic functions, provided degeneracy does not occur

***provided $\beta_k > - 1/(\alpha_k g_{k+1}^T H_k g_{k+1})$

This paper supplies the necessary background and suggests some related topics for other expository papers or further research. Since exact linear searches are basic to the development of the methods studied, the theoretical convergence properties presented are dependent upon this condition. However, some analysis on the convergence of certain algorithms using less than exact linear searches has been done. The convergence properties of the DFP method applied to a convex function are examined by M. L. Lenard [34]. Powell [50, 51] studies finite termination properties for Broyden's one-parameter family applied to a positive definite quadratic function.

The complementary DFP formula derived as an optimally conditioned member of Broyden's one-parameter family is also used by Fletcher [26] in a different algorithm. Since det $H_{DFP'} \geq$ det H_{DFP}, the use of $H_{DFP'}$ in a variable metric algorithm might counteract the tendency toward singularity of H_{DFP}. However, since Lemma 3.2 also implies that $\| H_{DFP'} \|_2 \geq \| H_{DFP} \|_2$, the use of $H_{DFP'}$ alone might cause H to tend to become unbounded. Fletcher's algorithm suggests a way to counter both singularity and unboundednes. If $s_k^T y_k \geq y_k^T H_k y_k$, then H_k is updated by the DFP' formula; otherwise, the DFP formula is used. The interpretation of this test is based on the fact that for a quadratic function with Hessian matrix G, $s_k = G^{-1} y_k$. Hence the "larger" DFP' formula is used whenever H_k is "smaller" than G^{-1} in the sense $y_k^T G^{-1} y_k \geq y_k^T H_k y_k$. In addition, Fletcher chooses not to carry out a full linear search on each iteration. Instead he uses a strategy that usually requires only one function and gradient evaluation on each iteration.

Linear searches are usually done by evaluating the function and gradient for a number of different step sizes and interpolating according to some strategy, until a sufficiently accurate minimum is obtained. Thus, considerable computing effort, as measured by the number of function and gradient evaluations, is required. Another disadvantage is the possibility that a minimum along the search direction may not exist at all. Fletcher's algorithm is based on the theory that it may not be worthwhile to calculate the optimal step size very accurately. Another approach is to consider whether the linear search can be avoided completely. The importance of the linear search is that the minimum of a quadratic function f with positive definite Hessian matrix G may be found in a finite number of iterations if the search directions are conjugate with respect to G. However, for one member of Broyden's one-parameter family of correction matrices, finite termination can be proved by showing $H_n = G^{-1}$ for a variable metric algorithm without linear searches. This member is the symmetric rank one matrix

$$C_k = \frac{(s_k - H_k y_k)(s_k - H_k y_k)^T}{(s_k - H_k y_k)^T y_k} \qquad (5.1)$$

obtained when

$$\beta_k = \frac{1}{s_k^T y_k - y_k^T H_k y_k}.$$

The use of this rank one correction matrix in a variable metric method was first suggested by Davidon [14]. It has also been suggested independently by Broyden [6], A. V. Fiacco and G. P. McCormick [21], B. A. Murtagh and R. W. H. Sargent [39], and P. Wolfe [59]. The property $H_n = G^{-1}$ is established by Broyden in Theorem 6 of [6] for an algorithm using the following iteration. Given the vector x_k, the gradient g_k, and the matrix H_k,

$$x_{k+1} = x_k - \alpha_k H_k g_k,$$

$$H_{k+1} = H_k + C_k, \qquad (5.2)$$

where α_k is an arbitrary nonzero scalar except that it must not cause H_{k+1} to be singular or undefined and where C_k is the matrix given by (5.1) with $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$. Although the use of (5.1) eliminates the need for a linear search, it presents some other problems. One is that H_k positive definite need not imply H_{k+1} positive definite. Hence stability cannot be guaranteed in a basic algorithm and H_{k+1} may be singular or undefined due to a zero denominator. For example, if $\alpha_k = 1$ happens to minimize $f(x_k - \alpha H_k g_k)$ with respect to α, then as shown in Chapter III, H_{k+1} is singular. Thus, many additions to the basic algorithm are required if this rank one updating formula is used. Davidon's [15, 16] rank one algorithm always uses $\alpha_k = 1$, that is, $x_{k+1} = x_k - H_k g_k$. If the resulting vectors s_k and y_k are such that $H_{k+1} = H_k + C_k$, with C_k given by (5.1), is not positive definite, then H_{k+1} is defined by adding a different multiple of $(s_k - H_k y_k)(s_k - H_k y_k)^T$ to H_k so that positive definiteness is obtained. After H_{k+1} has been calculated, if $f(x_{k+1}) > f(x_k)$, then the next iteration begins at x_k instead of x_{k+1}. Murtagh and Sargent [40] also propose algorithms in which $x_{k+1} = x_k - \alpha_k H_k g_k$ for some α_k and the positive definiteness of H_k is maintained.

An important property first noted by Wolfe [59] is that the rank one correction given by (5.1) can yield $H_n = G^{-1}$, for a quadratic function with positive definite Hessian matrix G, without the restriction that x_{k+1} be calculated by (5.2). Recall that this property follows from

$$H_k y_j = s_j, \; 0 \le j \le k - 1, \tag{5.3}$$

for k = n, if the vectors $s_0, s_1, \ldots, s_{n-1}$ are linearly independent since, in this case, $y_j = Gs_j$. Equation (5.3) is true for k = 1 because the rank one updating formula satisfies the quasi-Newton equation. Assuming (5.3) to be true and using the relation $y_j = Gs_j$, which is true for any j, gives

$$(s_k - H_k y_k)^T y_j = s_k^T G s_j - s_k^T G s_j = 0, \quad 0 \leq j \leq k - 1.$$

Thus, it follows from the iteration formula $H_{k+1} = H_k + C_k$ with C_k given by (5.1) and the induction hypothesis that

$$H_{k+1} y_j = s_j, \quad 0 \leq j \leq k - 1,$$

if the vectors s_j are such that H_{k+1} is defined. Since the quasi-Newton equation implies that the above equation is true for j = k, the induction is complete. Algorithms which attempt to take advantage of this flexibility in the choice of s_k have been proposed by Powell [46] and Bard [4]. In these algorithms, the matrix H_k need not be positive definite and so is always updated by (5.1) and s_k is not always a multiple of $- H_k g_k$.

Routines, in particular FORTRAN subroutines and ALGOL procedures, implementing the variable metric methods discussed in this paper are available. Implementations of the DFP method include FLEPOMIN by M. Wells [57] and FMFP from International Business Machines Corporation [32]. The complementary DFP formula is used in BROMIN by K. Fielding [22]. DAPODMIN by S. A. Lill [35] is an implementation of a modification of the DFP method suggested by G. W. Stewart [55] which uses difference approximations for the first partial derivatives. The derivatives are also estimated by differences in ZXMIN from International Mathematical and Statistical Libraries, Incorporated [33] which is based on VA10A by Fletcher [25]. Surveys of additonal existing implementations are given by Dixon [20] and Fletcher [24].

BIBLIOGRAPHY

(1) Adachi, N. "On Variable-Metric Algorithms." *J. Optim. Theory Applns.*, Vol. 7 (1971), 391-410.

(2) _________. "On the Uniqueness of Search Directions in Variable-Metric Algorithms." *J. Optim. Theory Applns.*, Vol. 11 (1973), 590-604.

(3) Bard, Y. "On a Numerical Instability of Davidon-Like Methods." *Math. Prog.*, Vol. 22 (1968), 665-666.

(4) _________. "Comparison of Gradient Methods for the Solution of Non-linear Parameter Estimation Problems." *SIAM J. Numer. Anal.*, Vol. 7 (1970), 157-186.

(5) Box, M. J. "A Comparison of Several Current Optimization Methods, and the Use of Transformations in Constrained Problems." *Comput. J.*, Vol. 9 (1966), 67-77.

(6) Broyden, C. G. "Quasi-Newton Methods and Their Application to Function Minimisation." *Maths. Comput.*, Vol 21 (1967), 368-381.

(7) _________. "The Convergence of a Class of Double-rank Minimization Algorithms, 1. General Considerations." *J. Inst. Maths. Applics.*, Vol. 6 (1970), 76-90.

(8) _________. "The Convergence of a Class of Double-rank Minimization Algorithms, 2. The New Algorithm." *J. Inst. Maths. Applics.*, Vol. 6 (1970), 222-231.

(9) _________. "Quasi-Newton Methods." *Numerical Methods for Unconstrained Optimization*. Ed. W. Murray. London and New York: Academic Press, 1972, 87-106.

(10) Byrne, G. D. and C. A. Hall, eds. *Numerical Solution of Systems of Nonlinear Algebraic Equations*. London and New York: Academic Press, 1973.

(11) Cauchy, A. "Methods Generale pour la Resolution des Systemes d'Equations Simultanees." *C. R. Acad. Sci. Paris*, Vol. 25 (1847), 536.

(12) Crockett, J. B. and H. Chernoff. "Gradient Methods of Maximization." *Pacific J. Math.*, Vol. 5 (1955), 33-50.

(13) Daniel, J. W. The Approximate Minimization of Functionals. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1971.

(14) Davidon, W. C. "Variable Metric Method for Minimization." AEC Research and Development Report. ANL-5990 (Rev.). Lemont, Illinois: Argonne National Laboratory, 1959.

(15) __________. "Variance Algorithm for Minimization." Comput. J., Vol. 10 (1968), 406-410.

(16) __________. "Variance Algorithms for Minimization." Optimization. Ed. R. Fletcher. London and New York: Academic Press, 1969, 13-20.

(17) Dixon, L. C. W. "Quasi-Newton Techniques Generate Identical Points II: The Proofs of Four New Theorems." Math. Prog., Vol. 3 (1972), 345-358.

(18) __________. "Variable Metric Algorithms: Necessary and Sufficient Conditions for Identical Behavior of Nonquadratic Functions." J. Optim. Theory Applns., Vol. 10 (1972), 34-40.

(19) __________. "Choice of Step Length, a Crucial Factor in the Performance of Variable Metric Algorithms." Numerical Methods for Non-linear Optimization. Ed. F. A. Lootsma. London and New York: Academic Press, 1972, 149-170.

(20) __________. "Nonlinear Optimisation: A Survey of the State of the Art." Software for Numerical Mathematics. Ed. D. J. Evans. London and New York: Academic Press, 1974, 193-218.

(21) Fiacco, A. V. and G. P. McCormick. Nonlinear Programming: Sequential Unconstrained Minimization Techniques. New York: John Wiley and Sons, Inc., 1968.

(22) Fielding, K. "Algorithm 387, Function Minimization and Linear Search." Comm. ACM, Vol. 13 (1970), 509-510.

(23) Fletcher, R. "A New Approach to Variable Metric Algorithms." Comput. J., Vol. 13 (1970), 317-322.

(24) __________. "A Survey of Algorithms for Unconstrained Optimization." Numerical Methods for Unconstrained Optimization. Ed. W. Murray. London and New York: Academic Press, 1972, 123-129.

(25) __________. "FORTRAN Subroutines for Minimisation by Quasi-Newton Methods." A. E. R. E. Report R 7125. Harwell, England: Atomic Energy Research Establishment, 1972.

(26) Fletcher, R. and M. J. D. Powell. "A Rapidly Convergent Descent Method for Minimization." Comput. J., Vol. 6 (1963), 163-168.

(27) Goldfarb, D. "A Family of Variable-Metric Methods Derived by Variational Means." *Maths. Comput.*, Vol. 24 (1970), 23-26.

(28) Greenstadt, J. "Variations on Variable-Metric Methods." *Maths. Comput.*, Vol. 24 (1970), 1-22.

(29) Householder, A. S. *The Theory of Matrices in Numerical Analysis.* New York: Blaisdell Publishing Co., 1964.

(30) Huang, H. Y. "Unified Approach to Quadratically Convergent Algorithms for Function Minimization." *J. Optim. Theory Applns.*, Vol. 5 (1970), 405-423.

(31) Huang, H. Y. and A. V. Levy. "Numerical Experiments on Quadratically Convergent Algorithms for Function Minimization." *J. Optim. Theory Applns.*, Vol. 6 (1970), 269-282.

(32) International Business Machines Corp. *IBM System/360 Scientific Subroutine Package: Programmer's Manual.* H20-0205-3. White Plains, New York: International Business Machines Corp., 1968, 221-224.

(33) International Mathematical and Statistical Libraries, Inc. *IMSL Library 1, Reference Manual.* 5th ed. Document IMSL LIB1-0005 (Rev.). Houston: International Mathematical and Statistical Libraries, Inc., 1975.

(34) Lenard, M. L. "Practical Convergence Conditions for the Davidon-Fletcher-Powell Method." *Math. Prog.*, Vol. 9 (1975), 69-86.

(35) Lill, S. A. "Algorithm 46, A Modified Davidon Method for Finding the Minimum of a Function Using Difference Approximations for Derivatives." *Comput. J.*, Vol. 13 (1970), 111-113.

(36) McCormick, G. P. and J. D. Pearson. "Variable Metric Methods and Unconstrained Optimization." *Optimization.* Ed. R. Fletcher. London and New York: Academic Press, 1969, 307-325.

(37) Meyers, G. E. "Properties of the Conjugate-Gradient and Davidon Methods." *J. Optim. Theory Applns.*, Vol. 2 (1968), 209-219.

(38) Murray, W. "Fundamentals." *Numerical Methods for Unconstrained Optimization.* Ed. W. Murray. London and New York: Academic Press, 1972, 1-12.

(39) Murtagh, B. A. and R. W. H. Sargent. "A Constrained Minimization Method With Quadratic Convergence." *Optimization.* Ed. R. Fletcher. London and New York: Academic Press, 1969, 215-246.

(40) __________. "Computational Experience with Quadratically Convergent Minimisation Methods." *Comput. J.*, Vol. 13 (1970), 185-194.

(41) Ortega, J. M. and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. New York and London: Academic Press, 1970.

(42) Pearson, J. D. "Variable Metric Methods of Minimisation." *Comput. J.*, Vol. 12 (1969), 171-178.

(43) Penrose, R. "A Generalized Inverse for Matrices." *Proc. Cambridge Philos. Soc.*, Vol. 51 (1954), 406-413.

(44) Powell, M. J. D. "An Iterative Method for Finding Stationary Values of a Function of Several Variables." *Comput. J.*, Vol. 5 (1962), 147-151.

(45) __________. "A Survey of Numerical Methods for Unconstrained Optimization." *SIAM Review*, Vol. 12 (1970), 79-97.

(46) __________. "Rank One Methods for Unconstrained Optimization." *Integer and Nonlinear Programming*. Ed. J. Abadie. Amsterdam: North-Holland Publishing Co., 1970, 139-156.

(47) __________. "On the Convergence of the Variable Metric Algorithm." *J. Inst. Maths. Applics.*, Vol. 7 (1971), 21-36.

(48) __________. "Recent Advances in Unconstrained Optimization." *Math. Prog.*, Vol. 1 (1971), 26-57.

(49) __________. "Some Properties of the Variable Metric Algorithm." *Numerical Methods for Non-Linear Optimization*. Ed. F. A. Lootsma. London and New York: Academic Press, 1972, 1-17.

(50) __________. "Quadratic Termination Properties of a Class of Double-Rank Minimization Algorithms." *A. E. R. E. Report TP 471*. Harwell, England: Atomic Energy Research Establishment, 1972.

(51) __________. "Some Theorems on Quadratic Termination Properties of Minimization Algorithms." *A. E. R. E. Report TP 472*. Harwell, England: Atomic Energy Research Establishment, 1972.

(52) Rosenbrock, H. H. "An Automatic Method for Finding the Greatest or Least Value of a Function." *Comput. J.*, Vol. 3 (1960), 175-184.

(53) Shanno, D. F. "Conditioning of Quasi-Newton Methods for Function Minimization." *Maths. Comput.*, Vol. 24 (1970), 647-656.

(54) Shanno, D. F. and P. C. Kettler. "Optimal Conditioning of Quasi-Newton Methods." *Maths. Comput.*, Vol. 24 (1970), 657-664.

(55) Stewart, G. W. "A Modification of Davidon's Minimization Method to Accept Difference Approximation of Derivatives." *J. ACM*, Vol. 14 (1967), 72-83.

(56) Stewart, G. W. Introduction to Matrix Computations. London and New York: Academic Press, 1973.

(57) Wells, M. "Algorithm 251, Function Minimization." Comm. ACM, Vol. 8 (1965), 169-170.

(58) Wilkinson, J. H. The Algebraic Eigenvalue Problem. Oxford: Clarendon Press, 1965.

(59) Wolfe, P. "Another Variable Metric Method." IBM Working Paper. International Business Machines Corp., Yorktown Heights, New York, 1967. (unpublished).

APPENDIX

Rosenbrock

This function, introduced by Rosenbrock [52], is defined by

$$f(\xi_1, \xi_2) = 100(\xi_2 - \xi_1^2)^2 + (1 - \xi_1)^2$$

with the suggested initial point (-1.2, 1). The minimum value of zero occurs at the point (1, 1). It is difficult to minimize because it has a steep-sided valley following the curve $\xi_1^2 = \xi_2$.

Helical Valley

This function, given by Fletcher and Powell [26], is defined by

$$f(\xi_1, \xi_2, \xi_3) = 100\{[\xi_3 - 10\theta(\xi_1, \xi_2)]^2 + [r(\xi_1, \xi_2) - 1]^2\} + \xi_3^2,$$

where

$$2\pi\theta(\xi_1, \xi_2) = \begin{cases} \arctan \xi_2/\xi_1, & \xi_1 > 0, \\ \pi + \arctan \xi_2/\xi_1, & \xi_1 < 0, \ -\pi/2 < 2\pi\theta < 3\pi/2, \end{cases}$$

and

$$r(\xi_1, \xi_2) = (\xi_1^2 + \xi_2^2)^{\frac{1}{2}}.$$

It has a steep-sided helical valley in the ξ_3 direction with pitch 10 and radius one. The initial point is (-1, 0, 0) and the point (1, 0, 0) gives the minimum value of zero.

Powell

This function, introduced by Powell [44], is given by

$$f(\xi_1, \xi_2, \xi_3, \xi_4) = (\xi_1 + 10\xi_2)^2 + 5(\xi_3 - \xi_4)^2 + (\xi_2 - 2\xi_3)^4 + 10(\xi_1 - \xi_4)^4.$$

The initial point (3, 1, 0, -1) is used and the minimum of zero occurs at the point (0, 0, 0, 0). This function is a severe test since the Hessian matrix is singular at the minimum point.

Trigonometric

Fletcher and Powell [26] defined these functions to test whether a method is suitable for finding the minimum of a function of a large number of variables. The problem is to solve the set of simultaneous nonlinear equations

$$\sum_{j=1}^{n} (\gamma_{ij}\sin \xi_j + \delta_{ij}\cos \xi_j) = \rho_i, \ i = 1, \ldots, n,$$

where the coefficients γ_{ij} and δ_{ij}, i, j = 1, ..., n, are generated as random integers between - 100 and + 100 and the right hand sides ρ_i, i = 1, ..., n, are calculated for values of the variables ξ_j, j = 1, ..., n, generated randomly between $-\pi$ and π. Hence, the function of n variables to be minimized is

$$f(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{n} [\rho_i - \sum_{j=1}^{n} (\gamma_{ij}\sin \xi_j + \delta_{ij}\cos \xi_j)]^2$$

with the minimum value of zero at the point $(\xi_1, \ldots, \xi_n)$ generated. The initial point is $(\xi_1 + 0.1\sigma_1, \ldots, \xi_n + 0.1\sigma_n)$, where σ_j, j = 1, ..., n, are also generated as random numbers between $-\pi$ and π.

Sum of Exponentials

Broyden [8] designed these functions to fit m data points (ϕ_i, ψ_i), $i = 1, \ldots, m$, by a sum of q exponentials in order to combine maximum scope for testing with minimum extra programming. The function to be minimized is defined as

$$f(\xi_1, \ldots, \xi_n) = \sum_{i=1}^{m} [\psi_i - \sum_{j=1}^{q} \xi_j \exp(- \xi_{j+q}\phi_i)]^2$$

where $n = 2q$. The minimum is dependent upon the way in which the data are obtained. For the function reported, $q = 3$ and the values of ψ_i were the sum of three exponentials evaluated at 13 values of ϕ_i.

Sum of Two Exponentials

Box [5] introduced this function which is defined by

$$f(\xi_1, \xi_2) = \sum_{i=1}^{10} [(\exp(- \xi_1 i/10) - \exp(- \xi_2 i/10) - (\exp(- i/10) - \exp(- i))]^2.$$

This function fits 10 data points (ϕ_i, ψ_i), $i = 1, \ldots, 10$, where ϕ_i ranges from 0.1 to 1 in steps of 0.1, and $\psi_i = \exp(- \phi_i) - \exp(- 10\phi_i)$, by a sum of two exponentials. The point (1, 10) gives the minimum value of zero. The suggested initial points are (0, 0), (0, 20), (5, 0), (5, 20), and (2.5, 10).

Wood

This function, credited to C. F. Wood and documented by Pearson [42], is given by

$$f(\xi_1, \xi_2, \xi_3, \xi_4) = 100(\xi_2 - \xi_1^2)^2 + (1 - \xi_1)^2 + 90(\xi_4 - \xi_3^2)^2 + (1 - \xi_3)^2$$
$$+ 10.1[(\xi_2 - 1)^2 + (\xi_4 - 1)^2] + 19.8(\xi_2 - 1)(\xi_4 - 1).$$

The initial point is (-3, -1, -3, -1) and the minimum value is zero at (1, 1, 1, 1). The function has a nonoptimal stationary point at (-0.9679, 0.9471, -0.9695, 0.9512) which can cause an algorithm to converge to this nonminimal point.

Weibull

The Weibull function, introduced by Shanno [53], is defined by

$$f(\xi_1, \xi_2, \xi_3) = \sum_{i=1}^{99} \left\{ \exp\left[- \frac{(\phi_i - \xi_3)^{\xi_2}}{\xi_1} \right] - \psi_i \right\}^2,$$

where $\phi_i = 25 + [50 \log_e(1/\psi_i)]^{2/3}$ and $\psi_i = i/100$. That is, the ϕ_i and ψ_i, i = 1, ..., 99, are perfect data generated for ψ_i ranging from 0.01 to 0.99 in steps of 0.01 for the values $\xi_1 = 50$, $\xi_2 = 1.5$, and $\xi_3 = 25$. The minimum value of zero occurs at the point (50, 1.5, 25). Different initial points may be used.

VITA

Rosalee Joy Taylor

Candidate for the Degree of

Doctor of Education

Thesis: THE DAVIDON-FLETCHER-POWELL METHOD AND FAMILIES OF VARIABLE METRIC METHODS FOR UNCONSTRAINED MINIMIZATION

Major Field: Higher Education

Biographical:

Personal Data: Born in Clinton, Oklahoma, on February 19, 1945, the daughter of Mrs. Hulda R. Tugwell; married to David Hartman Taylor on December 26, 1966.

Education: Graduated from Cordell High School, Cordell, Oklahoma, in May, 1963; received the Bachelor of Science in Education degree with a major in mathematics from Southwestern Oklahoma State University, Weatherford, Oklahoma, in July, 1966; received the Master of Arts degree in mathematics from the University of Missouri-Columbia, Columbia, Missouri, in June, 1968; completed requirements for the Doctor of Education degree at Oklahoma State University, Stillwater, Oklahoma, in May, 1976.

Professional Experience: Graduate assistant in Mathematics Department at the University of Missouri-Columbia, Columbia, Missouri, September, 1966-January, 1967, September, 1967-May, 1968; instructor in Business Department at Southwestern Oklahoma State University, Weatherford, Oklahoma, September, 1968-May, 1969; instructor in Mathematics Department at Southwestern Oklahoma State University, Weatherford, Oklahoma, June, 1969-May, 1971; graduate assistant in Mathematics Department at Oklahoma State University, Stillwater, Oklahoma, August, 1971-December, 1971; graduate assistant in Computing and Information Sciences Department at Oklahoma State University, Stillwater, Oklahoma, August, 1972-May, 1974.