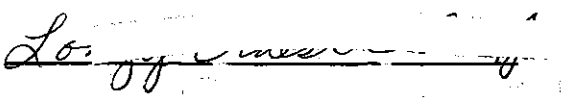
In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institute shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.



3/17/65
b

NONLINEAR OPTIMIZATION

A THESIS

Presented to

The Faculty of the Graduate Division

by

Lonzy Ernest Elder, Jr.

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Mathematics

Georgia Institute of Technology

April, 1968

NONLINEAR OPTIMIZATION

Approved:

Chairman

Date approved by Chairman: April 6, 1968

ACKNOWLEDGMENTS

I wish to express my appreciation to Dr. George L. Cain, Jr., my thesis advisor, for his guidance and help in the preparation of this thesis. I also wish to thank Dr. George C. Caldwell and Dr. D. T. Paris for their reading of the thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I

INTRODUCTION

The purpose of this paper is to study iterative methods for optimizing (that is, maximizing or minimizing) a real valued function of n variables, $n \geq 1$ (which may be subject to constraints). To be consistent in the discussion, function minimization will be discussed, since finding the maximum value of a function is equivalent to finding the minimum value of the negative of the function. In mathematical terms, for a given continuous function f which is defined by the equation $f(\bar{x}) = f(x_1,x_2,\ldots,x_n)$ on E_n, the problem is to find a point $\bar{y}$ in E_n (if one exists) such that $f(\bar{y})$ has a minimum value, that is, $f(\bar{y}) \leq f(\bar{x})$ for all $\bar{x}$ in E_n in a neighborhood of $\bar{y}$. The restrictions on the independent variables that will be considered can be written as

$$g_i(x_1,x_2,\ldots,x_n) \geq 0, \qquad i = 1,2,\ldots,m, \tag{1}$$

where the functions g_i, $i = 1,2,\ldots,m$, are continuous. A function subject to restrictions like those in (1) is said to be constrained. For an unconstrained function with continuous second-order partial derivatives, a minimum value of the function occurs at a point where the first-order partial derivatives of the function vanish and the matrix of second-order partial derivatives of the function is positive definite (see [1,p.152]). Determining such a point in this manner

requires the solution of at least n nonlinear equations simultaneously.

This paper will primarily examine sequential (or iterative) minimization procedures suitable for digital computers in which the points tested for a minimum value of the function are completely determined by a previous set of operations. Nonsequential types of function minimization (where previous test points do not determine where the next test point will be located) generally involve solving a system of nonlinear equations or else use random methods to pick test points in a region where the minimum is thought to be located. In the latter case, after a number of points has been chosen, the smallest value the function assumes in this set of points is taken to be the minimum, and the number of necessary test points for a certain accuracy can be determined from statistical theory. Nonsequential types of methods will not be considered in this paper.

In Chapter II, sequential methods for successively restricting the interval in which a point at which the minimum value of a function of one variable is attained is located are considered, along with methods of quadratic and cubic interpolation to approximate the minimum value of the function with only a few function evaluations. In Chapter III, methods for minimizing functions of n variables, $n > 1$, subject to no constraints are described and convergence features proved. Chapter IV is concerned with a few of the methods which have been developed to handle constrained functions; specifically those methods which convert the constrained function to an unconstrained form are studied. Final conclusions are made in Chapter V.

Notation

The components of n dimensional column vectors will be distinguished by subscripts on the letters while bars over letters will identify vectors; for example,

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} .$$

The word "direction" will be used synonymously with "vector." Iteration numbers will usually be denoted by subscripts on vectors and matrices will be indicated by capital letters (with or without subscripts). The vector $\bar{g}$ will denote the gradient vector of the function f. The letter n will indicate the number of function variables.

Preliminary Definitions and Theorems

The following definitions and theorems will be used throughout this paper.

Definition 1. For a function f defined on a set S in E_n, the statement, "$\bar{x}_o$ is a minimum," means that $\bar{x}_o$ is a point where a relative minimum value of the function occurs, that is, $f(\bar{x}_o) \leq f(\bar{x})$, for all $\bar{x}$ in a neighborhood of $\bar{x}_o$ and in S; and $f(\bar{x}_o)$ is said to be a minimum value of the function.

Definition 2. A vector $\bar{p}$ is called a *direction of search* if a number λ can be calculated so that $f(\bar{x} + \lambda\bar{p}) < f(\bar{x})$ for a fixed $\bar{x}$. The process of calculating this λ is called *searching* in the direction $\bar{p}$.

Definition 3. When a direction $\bar{p}$ is being searched, the vector $\lambda\bar{p}$, where λ is a scalar, is called a *step* in the direction $\bar{p}$, and the magnitude of this vector $|\lambda\bar{p}|$ is called the *step size* or *length of step*.

Definition 4. If c is a number in the range of a function f, then the set of all points $\bar{x}$ in E_n which satisfy the equation $f(\bar{x}) = c$ is called a *contour* of the function.

Definition 5. The matrix of second-order partial derivatives of a function f, that is, the matrix $A = [a_{ij}]$ with elements $a_{ij} = \dfrac{\partial^2 f}{\partial x_i \partial x_j}$ is called the *Hessian matrix* of the function f.

Definition 6. A *quadratic function* of n variables is a function f defined by the equation

$$f(\bar{x}) = \frac{1}{2}\bar{x}^t A\bar{x} + \bar{b}^t\bar{x} + c \ , \qquad (2)$$

where c is a scalar, $\bar{b}$ is a constant vector, and A is a constant matrix. A *positive definite quadratic function*, the only kind of quadratic function considered in this paper, is defined by Equation (2) and A is required to be a symmetric and positive definite matrix. Note that A is the Hessian matrix of the quadratic function.

Theorem 1. For a positive definite quadratic function f, there exists a unique point $\bar{x}$ in E_n such that $f(\bar{x}) < f(\bar{y})$ for all $\bar{x} \neq \bar{y}$ in E_n.

Proof. The existence and uniqueness of the point follow from Theorem 7-9 in Apostol [1,p.152]. ∎

Definition 7. An iterative method for minimizing a function of n variables is said to have *quadratic convergence* if for any positive definite quadratic function, it is guaranteed that the minimum value of the function will be located exactly, apart from rounding errors, within n iterations.

Definition 8. The directions $\bar{p}$ and $\bar{q}$ are defined to be *A-conjugate* (or simply *conjugate*) if $\bar{p}^t A\bar{q} = 0$ for $\bar{p} \neq \bar{q}$, $\bar{p}, \bar{q} \neq \bar{0}$.

Theorem 2. If A is a positive definite matrix and if $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_n$ are vectors such that

$$\bar{p}_i^t A \bar{p}_j = 0, \qquad i \neq j,$$

$$\bar{p}_i^t A \bar{p}_j > 0, \qquad i = j,$$

that is, if they are mutually conjugate, then $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_n$ form a linearly independent set of vectors.

Proof. Suppose that $\bar{p}_1,\bar{p}_2,\ldots,\bar{p}_n$ are linearly dependent; that is, there are scalars $c_1,c_2,\ldots,c_n$ not all zero such that

$$\sum_{j=1}^{n} c_j\bar{p}_j = \bar{0} .$$

Now

$$A \sum_{j=1}^{n} c_j\bar{p}_j = \sum_{j=1}^{n} c_jA\bar{p}_j = \bar{0} .$$

By hypothesis, at least one of the $c_j \neq 0$, say $c_i \neq 0$. Then

$$\bar{p}_i^t \sum_{j=1}^{n} c_jA\bar{p}_j = \sum_{j=1}^{n} c_j\bar{p}_i^tA\bar{p}_j = c_i\bar{p}_i^tA\bar{p}_i = 0.$$

Hence by contradiction, $\bar{p}_1,\bar{p}_2,\ldots,\bar{p}_n$ are linearly independent. □

Definition 9. A function is said to be *convex* if

$$f((1-\theta)\bar{x} + \theta\,\bar{y}) \leq (1-\theta)\,f(\bar{x}) + \theta\,f(\bar{y})$$

for $0 \leq \theta \leq 1$ and for all $\bar{x}$ and $\bar{y}$ in the domain of the function. It is *strictly convex* if

$$f((1-\theta)\,\bar{x} + \theta\,\bar{y}) < (1-\theta)\,f(\bar{x}) + \theta\,f(\bar{y})$$

holds for distinct $\bar{x}$ and $\bar{y}$ in the domain of the function and for

$0 < \theta < 1$. A function is *concave (strictly concave)* if $-f(\bar{x})$ is convex (strictly convex).

CHAPTER II

ONE VARIABLE SEARCH

Methods for finding the minimum value of a function of one variable, in addition to being important methods in themselves, are required in many methods for minimizing a function f of n variables, $n > 1$.

Definition 10. A function f of one variable is *unimodal* on an interval (a,b) if there is an x^* in (a,b) such that $f(x^*) < f(x)$ for all $x \neq x^*$ in (a,b), and $f(x) > f(y)$ if $a < x < y < x^*$ and $f(x) < f(y)$ if $x^* < x < y < b$. Note that the function f must only be defined on (a,b) and that there are no restrictions on the function such as continuity and differentiability.

Dichotomous Search

Let f be a unimodal function of one variable on an interval (a_o,b_o). Choose a number $\varepsilon > 0$ such that ε is less than the accuracy desired in the variable x. The kth iteration, $k = 0,1,\ldots$, of the dichotomous search that is to be repeated until the desired accuracy is obtained is given as follows.

(i) Let

$$x_k = \frac{a_k + b_k}{2} - \frac{\varepsilon}{2} \quad \text{and} \quad y_k = \frac{a_k + b_k}{2} + \frac{\varepsilon}{2}$$

and evaluate $f(x_k)$ and $f(y_k)$.

(ii) If $f(x_k) = f(y_k)$, let $a_{k+1} = x_k$, $b_{k+1} = y_k$, and stop.

If $f(x_k) < f(y_k)$, let $a_{k+1} = a_k$, $b_{k+1} = y_k$.

If $f(x_k) > f(y_k)$, let $a_{k+1} = x_k$, $b_{k+1} = b_k$.

After k iterations, the minimum is located within an interval of length $\frac{1}{2^k}(b_o - a_o) + \varepsilon\left(1 - \frac{1}{2^k}\right)$ since

$$b_k - a_k = \frac{1}{2}(b_{k-1} - a_{k-1}) + \frac{\varepsilon}{2}$$

$$= \frac{1}{2}(b_{k-2} - a_{k-2}) + \frac{\varepsilon}{2}) + \frac{\varepsilon}{2} = \ldots$$

$$= \frac{1}{2^k}(b_o - a_o) + \frac{\varepsilon}{2}\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{k-1}}\right)$$

$$= \frac{1}{2^k}(b_o - a_o) + \frac{\varepsilon}{2}\left[\frac{1 - \frac{1}{2^k}}{1 - \frac{1}{2}}\right]$$

$$= \frac{1}{2^k}(b_o - a_o) + \varepsilon\left(1 - \frac{1}{2^k}\right);$$

but if the first alternative is chosen in step (ii), the interval in which the minimum is located has length ε. Note that the function must always be evaluated two times per interation and that this method reduces the interval the maximum amount per iteration for any method using two points in the subinterval since if two arbitrary points are chosen, the amount of reduction is greater than one-half the length of the interval.

Fibonacci Search

Definition 11. The Fibonacci sequence is the sequence of integers $\{F_k\}$ defined as:

$$F_o = 1, \quad F_1 = 1, \quad F_k = F_{k-1} + F_{k-2} \text{ for } k \geq 2.$$

Let f be a unimodal function of one variable on (a_o, b_o) with a minimum value at x^* to be determined and let the number of function evaluations to be made be N + 1. (This will determine the accuracy that can be obtained as will be shown later.) The kth iteration, k = 0,1,..., N-1, can be stated as follows.

(i) Choose the points x_k and y_k in the interval (a_k,b_k) as

$$x_k = \frac{F_{N-k-1}}{F_{N-k+1}} (b_k - a_k) + a_k, \tag{3a}$$

$$y_k = \frac{F_{N-k}}{F_{N-k+1}} (b_k - a_k) + a_k \tag{3b}$$

$$= \frac{F_{N-k+1}}{F_{N-k+1}} (b_k - a_k) - \frac{F_{N-k-1}}{F_{N-k+1}} (b_k - a_k) + a_k$$

$$= b_k - \frac{F_{N-k-1}}{F_{N-k+1}} (b_k - a_k),$$

since $F_{N-k} = F_{N-k+1} - F_{N-k-1}$.

(Note that $x_k < y_k$ since $\frac{F_j}{F_{j+2}} < \frac{1}{2}$ for any integer $j \geq 1$.)

(ii) Calculate $f(x_k)$ if $(a_k,b_k) = (a_{k-1},y_{k-1})$. Note that $f(y_k) = f(x_{k-1})$ (see Theorem 3).
Otherwise, calculate $f(y_k)$ if $(a_k,b_k) = (x_{k-1}, b_{k-1})$.
Note that $f(x_k) = f(y_{k-1})$.

(iii) If $f(x_k) < f(y_k)$, then $a_k \leq x^* < y_k$, and let $(a_{k+1}, b_{k+1}) = (a_k,y_k)$.
If $f(x_k) > f(y_k)$, then $x_k < x^* \leq b_k$, and let $(a_{k+1}, b_{k+1}) = (x_k,b_k)$.
If $f(x_k) = f(y_k)$, then $x_k < x^* < y_k$, and let either $(a_{k+1}, b_{k+1}) = (a_k,y_k)$ or $(a_{k+1},b_{k+1}) = (x_k,b_k)$.

The length of the last interval is

$$b_N - a_N = \begin{Bmatrix} y_{N-1} - a_{N-1} \\ \text{or} \\ b_{N-1} - x_{N-1} \end{Bmatrix} = \frac{F_1}{F_2}(b_{N-1} - a_{N-1}) = \frac{F_2F_1}{F_3F_2}(b_{N-2} - a_{N-2})$$

$$= \cdots = \frac{F_1}{F_N}(b_o - a_o) = \frac{1}{F_N}(b_o - a_o) ,$$

where $(b_o - a_o)$ is the length of the original interval. Thus the number of iterations for a desired accuracy can be determined from the sequence of Fibonacci numbers.

Theorem 3. The function is evaluated once per iteration for k = 1,2,..., N-1.

Proof. Assume that at the kth iteration, the minimum is located between a_k and y_k. Then at the next iteration

$$y_{k+1} = \frac{F_{N-(k+1)}}{F_{N+1-(k+1)}} (y_k - a_k) + a_k,$$

since $y_k = b_{k+1}$ and $a_k = a_{k+1}$. If the values of y_k in (3b) is substituted in this equation, then

$$y_{k+1} = \frac{F_{N-(k+1)}}{F_{N-k}} \left[\frac{F_{N-k}}{F_{N-k+1}} (b_k - a_k) + a_k - a_k \right] + a_k$$

$$= \frac{F_{N-k-1}}{F_{N-k+1}} (b_k - a_k) + a_k = x_k;$$

and the function needs to be evaluated at x_{k+1} during the (k+1)th iteration. A similar proof can be developed for the case when x^* is between x_k and b_k. ∎

The kth term of the Fibonacci sequence is given by

$$F_k = \frac{1}{\sqrt{5}} \left[\left[\frac{1 + \sqrt{5}}{2} \right]^{k+1} - \left[\frac{1 - \sqrt{5}}{2} \right]^{k+1} \right];$$

from this it follows that the limiting values of ratios of Fibonacci numbers are

$$\lim_{k\to\infty} \frac{F_k}{F_{k+1}} = \frac{\sqrt{5} - 1}{2} \doteq 0.618,$$

$$\lim_{k\to\infty} \frac{F_{k-1}}{F_{k+1}} = \frac{3 - \sqrt{5}}{2} \doteq 0.382.$$

On this basis, the following approximate formulas can be used to obtain x_k and y_k ($k = 1,2,\ldots,N$) in step (i) (Equation (3)),

$$\begin{aligned} x_k &= 0.382(b_k - a_k) + a_k \\ y_k &= 0.618(b_k - a_k) + a_k. \end{aligned} \tag{4}$$

After the first two points, the function is evaluated one time per iteration and the test points at any iteration are independent of the total number of points contrary to the Fibonacci search where x_k, y_k depend on N by Equation (3). Each iteration is simplified, but the length of the final interval after k iterations will be

$$b_k - a_k = (0.618)^k (b_o - a_o)$$

which is slightly larger (by 13 per cent, as shown by Kiefer [19]) than the corresponding interval in the Fibonacci search. When the formulas in (4) are used instead of (3), then the method is called "golden section."

When the last three methods are compared, the lengths of the

last interval after k+1 function evaluations obey the following inequalities, corresponding respectively to dichotomous, golden section and Fibonacci searches:

$$\frac{1}{2^{[(k+1)/2]}}(b_o-a_o) > (0.618)^k(b_o-a_o) > \frac{1}{F_k}(b_o-a_o), \qquad (5)$$

where [ ] denotes the greatest integer function. The first inequality is true since

$$\frac{1}{2^{k+1}} > (0.618)^{2k}$$

or

$$\frac{1}{2}(.5)^k > (.381924)^k$$

or

$$(0.763858)^k < 0.5,\ k > 2.$$

A tabulation of the values used in the coefficients of (b_o-a_o) in (5) (as in Wilde [28], p. 29) bears out the relationship as does a graphical comparison in Boas [4]. The above statements establish the fact that the Fibonacci search gives greater accuracy per specified number of function evaluations, but the method of golden section is much easier to use and the accuracy is not much less.

For functions of n variables, n > 1, due to the methods themselves and the fact that the function is not necessarily a quadratic

function, it is only necessary to have an approximate value of t such that $h(t) = f(\bar{x} + t\bar{p})$ has a minimum value. The methods to be discussed use quadratic and cubic interpolation and are more efficient than the prior methods discussed in the sense that they find an approximate value for the minimum with fewer function evaluations. Here there are no restrictive intervals in which the search is conducted.

Quadratic Interpolation

For the function h of one variable, let (a,h(a)), (b,h(b)), (c,h(c)) be three distinct points. A quadratic function q through the points is defined by the equation

$$q(t) = k_1 t^2 + k_2 t + k_3,$$

which for the three points gives the set of linear equations

$$a^2 k_1 + a k_2 + k_3 = h(a)$$

$$b^2 k_1 + b k_2 + k_3 = h(b)$$

$$c^2 k_1 + c k_2 + k_3 = h(c),$$

from which k_1, k_2, and k_3 can be determined. The coefficient matrix is non-singular since the points are distinct. By elementary calculus, the minimum value of the function q (which is an approximation to the minimum value of the function h) is

$$k_3 - \frac{1}{4}\frac{k_2^2}{k_1}$$

at $t^* = -k_2/2k_1$ if $k_1 > 0$, or equivalently, the function q has a minimum value

$$\frac{h(a)bc(b-c) - h(b)ac(a-c) + h(c)ab(a-b)}{(a-b)(a-c)(b-c)}$$

$$- \frac{1}{4}\frac{[-h(a)(b^2-c^2) + h(b)(a^2-c^2) - h(c)(a^2-b^2)]^2}{h(a)(b-c) - h(b)(a-c) + h(c)(a-b)}$$

at

$$t^* = -\frac{1}{2}\frac{-h(a)(b^2-c^2) + h(b)(a^2-c^2) - h(c)(a^2-b^2)}{h(a)(b-c) - h(b)(a-c) + h(c)(a-b)} \qquad (6)$$

if $d > 0$, where

$$d = \frac{h(a)(b-c) - h(b)(a-c) + h(c)(a-b)}{(a-b)(a-c)(b-c)} .$$

If $k_1 < 0$, then $\lim_{t\to\infty} q(t) = -\infty$ and $\lim_{t\to-\infty} q(t) = -\infty$, while q is linear if $k_1 = 0$.

For the way in which Powell [22] uses quadratic interpolation to approximate the point on a line where the minimum value of a function of more than one variable occurs, let $h(t) = f(\bar{x} + t\bar{p})$ and let the magnitude q of the step length, an upper bound m on the step length

(q is assumed to be less than m), and the accuracy e be given. The algorithm is given as follows.

(i) Calculate h(a) and h(b) = h(a + q).

(ii) Calculate h(c) = h(a - q) if h(a) < h(b).
Otherwise, calculate h(c) = h(a + 2q).

(iii) Use (a,h(a)), (b,h(b)), (c,h(c)) to calculate t^* and d.

(iv) If $d \leq 0$ or if $t^* > a + m$ or $t^* < a - m$, then replace a by a + m if c > b or replace b by a - m if c < a and go to step (iii).
Otherwise, go to step (v).

(v) If any of the following hold:
$|t^* - a| < e$, $|t^* - b| < e$, $|t^* - c| < e$,
accept t^* as an approximation to the point where the minimum occurs. If not, calculate $h(t^*)$.

(vi) If h(a) > h(b) and h(a) > h(c), replace a by t^* and go to step (iii);
if h(b) > h(a) and h(b) > h(c), replace b by t^* and go to step (iii);
if h(c) > h(a) and h(c) > h(b), replace c by t^* and go to step (iii).

Example 1. To illustrate the algorithm given by Powell and stated above, consider the function

$$h(t) = -\frac{1}{1 + t^2},$$

which has a minimum value of -1 at $t = 0$. Choose $a = 1$, $q = 1$, $m = 3$, $e = 0.1$.

Now $h(1) = -0.5$, $h(2) = -0.2$; therefore $c = 0$, $h(0) = -1$.

$t^* = 3$ and $d = -0.1 < 0$.

Hence $b = 2$ is replaced by $a - m = 1 - 3 = -2$ and t^* is calculated using $h(1) = -0.5$, $h(-2) = -0.2$, $h(0) = -1$.

$t^* = -0.33$ and $d = 0.3 > 0$.

t^* is again calculated using $h(1) = -0.5$, $h(-0.33) = -0.9$, $h(0) = -1$.

$t^* = 0.083$ and $d = 0.2 > 0$.

This value is accepted for t^* since $|0 - 0.083| < 0.1 = e$.

Cubic Interpolation

If the values of a differentiable function h at two points and the derivatives of h at these two points are available, then cubic interpolation can be used to establish an approximation to the point at which the minimum value of the function occurs. If a, b, h(a), h(b), h'(a), h'(b), are assumed to be known, then the resulting cubic polynomial p is given by the equation

$$p(t) = h(b) + (t-b)h'(b) - \frac{(t-b)^2}{(a-b)}(h'(b) + z)$$

$$+ \frac{(t-b)^3}{3(a-b)^2} (h'(a) + h'(b) + 2z) ,$$

where

$$z = 3 \left[\frac{h(a) - h(b)}{b - a}\right] + h'(a) + h'(b) ,$$

and its slope for $a \leq t \leq b$ is given by

$$p'(t) = h'(b) - 2 \frac{(t-b)}{a-b} (h'(b) + z)$$

$$+ \frac{(t-b)^2}{(a-b)^2} (h'(a) + h'(b) + 2z) .$$

If $h'(a) + h'(b) + 2z = 0$, then a quadratic polynomial could have been assumed. If $h'(a) < 0$ and $h'(b) > 0$, then the root t^* of $p'(t)$ that corresponds to a minimum lies between a and b and $p''(t^*) > 0$, where

$$p''(t) = -2 \frac{(h(b) + z)}{a - b} + 2 \frac{(t-b)}{(a-b)^2} (h'(a) + h'(b) + 2z) .$$

The root t^* of $p'(t)$ is

$$t^* - b = (a-b) \left[\frac{h'(b) + z \pm w}{h'(a) + h'(b) + 2z}\right] ,$$

where $w = (z^2 - h'(a)\,h'(b))^{1/2}$, and since

$$p''(t^*) = -2\left[\frac{h(b) + z}{a - b}\right] + \frac{2}{a - b}\,(h'(a) + h'(b) + 2z)\left[\frac{h'(b) + z \pm w}{h'(a) + h'(b) + 2z}\right]$$

is greater than zero only when the negative sign is chosen, then

$$t^* - b = (a-b)\left[\frac{h'(b) + z - w}{h'(a) + h'(b) + 2z}\right]$$

$$= (a-b)\left[\frac{h'(b) + w - z}{h'(b) - h'(a) + 2w}\right] . \qquad (7)$$

Davidon [9] (see also [16]) gives an algorithm that uses this interpolation to approximate the value t_m where $h'(t) = 0$ if $h(t) = f(\bar{x} + t\bar{p})$. (Note that $h'(t) = \bar{p}^t\bar{g}(\bar{x}+t\bar{p})$ and that $h'(0) = \bar{p}^t\bar{g}(\bar{x})$.)

Let h_e be an estimate of the minimum value of the function h and s be a step size (s can be chosen equal to 1). Assume $h'(0) < 0$. (If $h'(0) > 0$ then a variation of the following algorithm can be made.) The algorithm can be stated as follows.

(i) Evaluate $h(0)$, $h'(0)$.

(ii) Let $k = 2\,\dfrac{(h_e - h(0))}{h'(0)}$.

(iii) Choose a step length $q = k$ if $0 < k < (s^2)^{-1/2}$.

Otherwise, let $q = (s^2)^{-1/2}$.

(Note: When this algorithm is applied to the function f then $(s^2)^{-1/2}$ is replaced by $(\bar{p}\bar{t}\bar{p})^{-1/2}$.)

(iv) For $i = 0,1,2,\ldots$, evaluate $h'(2^i q)$ until h' is non-negative. Let $a = 2^{i-1}q$ and $b = 2^i q$. Thus t_m is bounded in the interval $a < t_m \leq b$.

(v) Calculate an estimate t^* of t_m using Equation (7).

(vi) If $h(t^*) < h(a)$ and $h(t^*) < h(b)$, then accept t^* as an estimate of t_m.
Otherwise, if $h'(t^*) \geq 0$, replace b by t^* and go to step (v); if $h'(t^*) < 0$, replace a by t and go to step (v).

Example 2. For an example of this algorithm, let

$$h(t) = -\frac{1}{1 + (t-7)^2}$$

which has a minimum of -1 at $t = 7$ and for which

$$h'(t) = \frac{2(t-7)}{(1 + (t-7)^2)^2} .$$

Choose $h_e = -1$, $s = 1$. Then $k > 1$ and therefore $q = 1$.

$$h'(0) = -14/50^2,$$

$$h'(1) = -12/37^2,$$

$$h'(2) = -10/26^2,$$

$$h'(4) = -6/10^2,\ h(4) = -1/10,\ a = 4,$$

$$h'(8) = 1/2 > 0,\ h(8) = -1/2,\ b = 8.$$

Using Equation (7), $t^* - 8 = -1$, $t^* = 7$, $h(7) = -1$, and $h(7) < h(4)$, $h(7) < h(8)$.

$t^* = 7$ is accepted as an approximation to the minimum.

CHAPTER III

MULTI-VARIABLE SEARCH

Sequential examination of points for functions of n variables, $n > 1$, can be divided into two classes: gradient methods which make beneficial use of the gradient vector to determine directions of search and non-gradient methods which are only systemized methods to compare points. The latter methods will be considered first.

Non-gradient Methods

Direct Search

The phrase "direct search," as used by Hooke and Jeeves [18], describes "a sequential examination of trial solutions involving comparison of each trial solution with the 'best' obtained up to that time together with a strategy for determining (as a function of earlier results) what the next trial solution will be." The method of direct search employs no techniques of analysis.

The basic algorithm is as follows.

(i) Select an initial approximation $\bar{x}$ to the minimum as the first "base point."

(ii) Choose another point $\bar{z}$. If $f(\bar{z}) < f(\bar{x})$, then set $\bar{x} = \bar{z}$ to give a new base point.

(iii) Repeat step (ii) until the convergence criterion is satisfied. (See pages 72-73 for a discussion of the criterion that can be used.)

The strategy for selecting new trial points can be divided into two separate parts. The first part establishes a pattern of search by making exploratory moves and the second moves in the established pattern.

Let $\bar{\Delta}_j$ be the vector $(0,\ldots,0,\delta_j,0,\ldots,0)^t$, where $\delta_j \neq 0$ and let $\rho > 1$ be given. Then the kth iteration can be stated as follows.

(i) Let $\bar{x}_{k,o}$ be the current base point.

(ii) For $j = 1,2,\ldots,n$, in turn, use one of the following to find $\bar{x}_{k,j}$:

(a) if $f(\bar{x}_{k,j-1} + \bar{\Delta}_j) < f(\bar{x}_{k,j-1})$, set $\bar{x}_{k,j} = \bar{x}_{k,j-1} + \bar{\Delta}_j$ and replace $\bar{\Delta}_j$ by $\rho\bar{\Delta}_j$;

(b) if $f(\bar{x}_{k,j-1} - \bar{\Delta}_j) < f(\bar{x}_{k,j-1})$, set $\bar{x}_{k,j} = \bar{x}_{k,j-1} - \bar{\Delta}_j$ and replace $\bar{\Delta}_j$ by $\rho\bar{\Delta}_j$;

(c) if $f(\bar{x}_{k,j-1}) \leq \min\{f(\bar{x}_{k,j-1} + \bar{\Delta}_j), f(\bar{x}_{k,j-1} - \bar{\Delta}_j)\}$ set $\bar{x}_{k,j} = \bar{x}_{k,j-1}$ and replace $\bar{\Delta}_j$ by $1/\rho\ \bar{\Delta}_j$.

(iii) If $f(2\bar{x}_{k,n} - \bar{x}_{k,o}) < f(\bar{x}_{k,n})$, set $\bar{x}_{k+1,o} = \bar{x}_{k,n} + (\bar{x}_{k,n} - \bar{x}_{k,o})$. Otherwise, set $\bar{x}_{k+1,o} = \bar{x}_{k,n}$.

The step size of the pattern move made in step (iii), that is, the magnitude of the vector $(\bar{x}_{k,n} - \bar{x}_{k,o})$ may also be increased if the first alternative of (iii) is satisfied.

Example 3. As an illustration of the previous algorithm consider the function f defined by

$$f(x_1,x_2) = x_1^2 - 2x_1x_2 + 2\,x_2^2 ,$$

whose minimum value is zero at the origin. Initially choose $\bar{\Delta}_1^t$ = (1/2,0), $\bar{\Delta}_2^t$ = (0,1/2), and ρ = 2, and let the starting point be (1,2). Table 1 gives the steps and function values for four iterations.

Table 1. Direct Search

	i = 1	Function Value	i = 2	Function Value	i = 3	Function Value	i = 4	Function Value
$\bar{X}_{i,o}^t$	(1,2)	5	(2,1)	2	(1,1)	1	(1,0.5)	0.5
$\bar{X}_{i,1}^t$	(1.5,2)	4.25	(1,1)	1	(1,1)	1	(1,0.5)	0.5
$\bar{\Delta}_1^t$	(1,0)		(2,0)		(1,0)		(0.5,0)	
$\bar{X}_{i,2}^t$	(1.5,1.5)	2.25	(1,1)	1	(1,0.5)	0.5	(1,0.5)	0.5
$\bar{\Delta}_2^t$	(0,1)		(0,0.5)		(0,1)		(0,0.5)	

Rosenbrock's Method

A method developed by Rosenbrock [23] uses a stepping procedure in n orthogonal directions with cyclic changes of these directions. The algorithm may be stated as follows.

(i) Choose an initial approximation $\bar{x}_o$ and numbers α, β so that $\alpha > 1$, $0 < \beta < 1$.

(ii) Choose n orthonormal directions $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_n$ initially and step lengths $e_{1,j}$, $j = 1,2,\ldots,n$.

(iii) For $j = 1,2,\ldots,n$,

(a) if $f(\bar{x}_o + e_{i,j}\bar{p}_j) \leq f(\bar{x}_o)$, replace $\bar{x}_o$ by $\bar{x}_o + e_{i,j}\bar{p}_j$ and put $e_{i+1,j} = \alpha e_{i,j}$;

(b) if $f(\bar{x}_o + e_{i,j}\bar{p}_j) > f(\bar{x}_o)$, put $e_{i+1,j} = -\beta e_{i,j}$.

(iv) Repeat step (iii), increasing i by 1 each time, until both (a) and (b) have been used for each $j = 1,2,\ldots,n$. For each direction, if (a) is chosen first, then eventually (b) will be chosen since either the function will increase or the point where the minimum value occurs will be reached. If (b) is chosen first, then eventually (a) will be chosen since the step size $e_{i,j}$ will become so small that the numbers $f(\bar{x}_o)$ and $f(\bar{x}_o + e_{i,j}\bar{p}_j)$ are indistinguishable in the first λ significant digits.

(v) Set $d_j = \sum_i e_{i,j}$,

where the sum is over the step sizes used in alternative (a) in step (iii) in the direction $\bar{p}_j$, $j = 1,2,\ldots,n$. Note that $d_j \neq 0$ since eventually alternative (a) will be used.

(vi) Let

$$\bar{q}_1 = d_1\bar{p}_1 + d_2\bar{p}_2 + \cdots + d_n\bar{p}_n$$

$$\bar{q}_2 = \qquad d_2\bar{p}_2 + \cdots + d_n\bar{p}_n$$

$$\vdots$$

$$\bar{q}_n = \qquad\qquad\qquad d_n\bar{p}_n.$$

(vii) Orthonormalize the directions $\bar{q}_1, \bar{q}_2, \ldots, \bar{q}_n$ by the Gram-Schmidt process and call these new directions $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_n$.

(viii) Repeat steps (iii) through (vii) until the convergence criterion is satisfied.

Example 4. For the function in Example 3,

$$f(x_1,x_2) = x_1^2 - 2\,x_1x_2 + 2\,x_2^2,$$

let the starting point be (1,2) and choose $\alpha = 3$, $\beta = 0.5$, $\bar{p}_1^t = (1,0)$, $\bar{p}_2^t = (0,1)$, $e_{1,1} = 0.5$, $e_{1,2} = 0.5$. The following illustrates the method for one iteration.

$$\bar{p}_1^t = (1,0),\ \bar{p}_2^t = (0,1)$$

$$\bar{x}_o^t = (1,2),\ f(\bar{x}_o) = 5,$$

$$e_{1,1} = 0.5,$$

$$\bar{x}_o^t = (1.5,2),\ f(\bar{x}_o) = 4.25,$$

$$e_{2,1} = 1.5,$$

$$e_{3,1} = -0.75,$$

$$\bar{x}_o^t = (1.5,2.5),\ f(\bar{x}_o) = 1,$$

$$e_{2,2} = 1.5,$$

$$e_{3,2} = -0.75,$$

$$d_1 = 1.5,\ d_2 = 1.5,$$

$$\bar{q}_1^t = (1.5,\ 1.5),\ \bar{q}_2^t = (0,1.5),$$

$$\bar{p}_1^t = (0.7071,0.7071),\ \bar{p}_2^t = (0.7071,-0.7071).$$

The result of applying steps (vi) and (vii) several times ensures that $\bar{p}_1$ coincides with the directions of fastest advance, $\bar{p}_2$ along the best direction which can be found normal to $\bar{p}_1$, and so on. It is stated by Smith [25] that when using the method on a positive definite quadratic function, the directions $\bar{p}_i$, $i = 1,2,\ldots,n$, align themselves in the limit along the axes of the function (the eigenvectors of the matrix of second derivatives--a particular case of conjugate directions) and that although the method does not have quadratic convergence, it does have a similarity with other methods that have quadratic convergence in this limiting process.

A variation of the above procedure, due to Davies, Swann, and Campey [14], replaces steps (iii), (iv), and (v) by a minimization of the function in the direction $\bar{p}_j$. This can be stated as

(iiia) For $j = 1,2,\ldots,n$, find a number d_j such that the function h defined by the equation $h(d) = f(\bar{x}_o + d\bar{p}_j)$ has a minimum at $d = d_j$ and replace $\bar{x}_o$ by $\bar{x}_o + d_j\bar{p}_j$.

Note that if any $d_j = 0$, then the vectors $\bar{q}_1,\bar{q}_2,\ldots,\bar{q}_{j-1}$, $\bar{q}_{j+1},\ldots,\bar{q}_n$ are orthonormalized in step (vii) and $\bar{p}_j$ is added to the resulting vectors after they are renamed.

Example 5. This method is illustrated in Table 2 by the same function in Example 3 and the $\bar{p}$ directions in Example 4.

Table 2. Method of Davies, Swann, and Campey

	1	Function Value	2
$\bar{x}_o^t$	(1, 2)	5	(2, 2.25)
$\bar{p}_1^t$	(1, 0)		(0.9701, -0.2425)
$\bar{p}_2^t$	(0, 1)		(0.2425, 0.9701)
d_1	1		.0793
$\bar{x}_o^t$	(2, 2)	4	(2.077, 1.731)
d_2	-0.25		1.972
$\bar{x}_o^t$	(2, 2.25)	2.5	

Powell's Method

Powell [22] developed a nongradient method which is a variation of the previous methods. He "proves" the method has quadratic convergence, but Zangwill [29] gives an example of a positive definite quadratic function which will not converge in any number of iterations. The crucial hypothesis omitted was that the directions of search spanned the space. See the paragraph preceding Theorem 4 for further discussion. Nevertheless, the method is efficient and is now stated.

Initially choose $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_n$ to be the n coordinate directions and let $\bar{x}_o$ be the starting point. An iteration of the basic procedure

to be repeated until the convergence criterion is satisfied is as follows.

(i) For $k = 1,2,\ldots,n$, find λ_k such that the function h_k defined by the equation $h_k(\lambda) = f(\bar{x}_{k-1} + \lambda\bar{p}_k)$ has a minimum at $\lambda = \lambda_k$ and define $\bar{x}_k = \bar{x}_{k-1} + \lambda_k\bar{p}_k$.

(ii) For $k = 1,2,\ldots,n-1$, replace $\bar{p}_k$ by $\bar{p}_{k+1}$ and replace $\bar{p}_n$ by $(\bar{x}_n - \bar{x}_o)$.

(iii) Choose λ^* so that the function h defined by the equation $h(\lambda) = f(\bar{x}_n + \lambda(\bar{x}_n - \bar{x}_o))$ has a minimum at $\lambda = \lambda^*$ and replace $\bar{x}_o$ by $\bar{x}_n + \lambda^*(\bar{x}_n - \bar{x}_o)$.

The sequence of points $\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n$ is calculated by searching in the directions $\bar{p}_1,\bar{p}_2,\ldots,\bar{p}_n$ successively. New directions are established by deleting $\bar{p}_1$ and replacing $\bar{p}_k$ by $\bar{p}_{k+1}$, $k = 1,2,\ldots,n-1$ and $\bar{p}_n$ by $\bar{x}_n - \bar{x}_o$. This last direction is searched and the point at which the minimum occurs replaces the old $\bar{x}_o$ completing one cycle.

Example 6. Consider the function

$$f(x_1,x_2) = x_1^2 - 2\,x_1x_2 + 2\,x_2^2$$

from Example 3 with starting point $\bar{x}_o^t = (1,2)$. The algorithm is illustrated by Table 3.

Table 3. Powell's Method

	1	Function Value	2	Function Value	3	Function Value
$\bar{x}_o^t$	(1,2)	5	(1.8,1.2)	1.8	(0,0)	0
$\bar{p}_1^t$	(1,0)		(0,1)		(1,-1)	
$\bar{p}_2^t$	(0,1)		(1,-1)		(-0.18,-0.12)	
λ_1	1		-0.3			
$\bar{x}_1^t$	(2,2)	4	(1.8,0.9)	1.62		
λ_2	-1		-0.18			
$\bar{x}_2^t$	(2,1)	2	(1.62,1.08)	1.458		
λ^*	-0.2		9			

Powell's Second Method (Simplified)

Because the previous method sometimes gives dependent directions of search, Powell [22] gave an alternate method. Zangwill [29] gives a simplification of this procedure and proves that it converges. This modification will now be given.

Let $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_n$ be the normalized coordinate directions and $\bar{x}_o$ be an initial approximation to the minimum. Let a scalar ε, $0 < \varepsilon \leq 1$, be given and set $\delta = 1$. The kth iteration to be repeated until the convergence criterion is satisfied can be stated as follows.

(i) For $j = 1,2,\ldots,n$, find λ_j such that the function h_j defined by the equation $h_j(\lambda) = f(\bar{x}_{j-1} + \lambda\bar{p}_j)$ has a minimum at $\lambda = \lambda_j$ and define $\bar{x}_j = \bar{x}_{j-1} + \lambda_j\bar{p}_j$.

(ii) Define $\alpha = |\bar{x}_n - \bar{x}_o|$ and $\bar{p}_{n+1} = (\bar{x}_n - \bar{x}_o)/\alpha$.

(iii) Find λ_{n+1} such that the function h_{n+1} defined by the equation $h_{n+1}(\lambda) = f(\bar{x}_n + \lambda\bar{p}_{n+1})$ has a minimum at $\lambda = \lambda_{n+1}$ and set $\bar{x}_o = \bar{x}_n + \lambda_{n+1}\bar{p}_{n+1}$.

(iv) Let $\lambda_s = \max\{\lambda_j : j = 1,2,\ldots,n\}$.

(a) If $\lambda_s\delta/\alpha \geq \varepsilon$, replace $\bar{p}_s$ by $\bar{p}_{n+1}$ and δ by $\lambda_s\delta/\alpha$.

(b) Otherwise, retain the same $\bar{p}$ directions and δ.

Example 7. Consider the function

$$f(x_1,x_2) = x_1^2 - 2x_1x_2 + 2\,x_2^2$$

from Example 3. Table 4 illustrates the algorithm with starting point (1,2) and $\varepsilon = 0.1$.

Zangwill [29] establishes that this procedure must converge to a point at which the gradient of a function f is zero, ($\bar{g} = \bar{0}$) for f a strictly convex continuously differentiable function, but the necessary lemmas and theorems will not be proved here.

Zangwill's Method

Zangwill [29] introduces a procedure based on theorems proved by Powell [22] which converges in a finite number of iterations for a

Table 4. Powell's Second Method (Simplified)

	1	$f(\bar{x})$	2	$f(x)$	3	$f(\bar{x})$
$\bar{x}_o^t$	(1,2)	5	(1.8,1.2)	1.8	(1.8,0.9)	1.62
$\bar{p}_1^t$	(1,0)		(.7071,-.7071)		(.7071,-.7071)	
$\bar{p}_2^t$	(0,1)		(0,1)		(0,1)	
λ_1	1		0		2.2556	
$\bar{x}_1^t$	(2,2)	4	(1.8,1.2)	1.8	(1.62,1.08)	1.458
λ_2	-1		-0.3		-.23	
$\bar{x}_2^t$	(2,1)	2	(1.8,0.9)	1.62	(1.62,.85)	1.3154
$\bar{p}_3^t$	(.7071,-.7071)		(0,-1)		(-.9635,-.2676)	
λ_3	-.2828		0		1.3732	
$\bar{x}_o^t$	(1.8,1.2)	1.8	(1.8,0.9)	1.62	(.2969,.4825)	.2672

positive definite quadratic function. He also establishes theoretical convergence for strictly convex continuously differentiable functions. His method is stated as follows.

Let $\bar{e}_j$, $j = 1,2,\ldots,n$, be the normalized coordinate directions and let $\bar{p}_j$, $j = 1,2,\ldots,n$, be n normalized directions which are given. Let $\bar{x}_n$ be the starting point and choose the number $t = 1$.

(i) Find λ_n so that the function h defined by $h(\lambda) = f(\bar{x}_n + \lambda\bar{p}_n)$ has a minimum at $\lambda = \lambda_n$ and let $\bar{x}_{n+1} = \bar{x}_n + \lambda_n\bar{p}_n$.

For iterations $k = 1,2,\ldots$, repeat the following seven steps.

(ii) Find α so that the function h defined by $h(\lambda) = f(\bar{x}_{n+1} + \lambda\bar{e}_t)$ has a minimum at $\lambda = \alpha$.

(iii) Replace t by t+1 if $1 \leq t < n$ and by 1 if $t = n$.

(iv) If $\alpha = 0$, go to step (ii). If this alternative occurs n times in succession, the point $\bar{x}_{n+1}$ is a minimum. If $\alpha \neq 0$, let $\bar{x}_o = \bar{x}_{n+1} + \alpha\bar{e}_t$.

(v) For $j = 1,2,\ldots,n$, find λ_j so that the function h_j defined by $h_j(\lambda) = f(\bar{x}_{j-1} + \lambda\bar{p}_j)$ has a minimum at $\lambda = \lambda_j$ and define $\bar{x}_j = \bar{x}_{j-1} + \lambda_j\bar{p}_j$.

(vi) Let $\bar{p}_{n+1} = (\bar{x}_n - \bar{x}_{n+1})/|\bar{x}_n - \bar{x}_{n+1}|$.

(vii) Find λ_{n+1} so that the function h_{n+1} defined by $h_{n+1}(\lambda) = f(\bar{x}_n + \lambda\bar{p}_{n+1})$ has a minimum at $\lambda = \lambda_{n+1}$ and define $\bar{x}_{n+1} = \bar{x}_n + \lambda_{n+1}\bar{p}_{n+1}$.

(viii) Replace $\bar{p}_j$ by $\bar{p}_{j+1}$, $j = 1,2,\ldots,n$, and go to step (ii).

If steps (ii) - (iv) are repeated n times in succession, then all n coordinate directions have been searched and no change in the point has occurred. Such a situation can only occur if the gradient of the function at that point is zero. For a strictly convex continuously

differentiable function the point is the minimum. Steps (v) - (viii) are similar to the previous two methods. Observe that after at most n iterations, all coordinate directions have been searched.

Example 8. The positive definite quadratic function introduced in Example 3 will again be used to illustrate Zangwill's method in Table 5.

Table 5. Zangwill's Method

	k = 0	$f(\bar{x})$	k = 1	$f(\bar{x})$	k = 2
$\bar{p}_1^t$			(1,0)		(0,1)
$\bar{p}_2^t$			(0,1)		(-1.3416,-.4472)
$\bar{x}_o^t$			(.5,.5)	.25	
λ_1			0		
$\bar{x}_1^t$			(.5,.5)	.25	
λ_2	-1.5		-.25		
$\bar{x}_2^t$	(1,2)	5	(.5,.25)	.125	
$\bar{p}_3^t$			(-.8944,-.4472)		
λ			.5590		
$\bar{x}_3^t$	(1.5)	2.25	(0,0)		
α	-.5		0		

The same starting point $\bar{x}_2^t = (1,2)$ will be used. Note that the vectors $\bar{p}_1$ and $\bar{p}_2$ are mutually conjugate for k = 2.

The following theorem and lemma were proved by Powell [22] in his attempt to establish quadratic convergence for his method. Theorem 4 is revised to require that the vectors $\bar{q}_i$, i = 1,2,...,m, span the m-dimensional space.

Theorem 4. If $\bar{q}_1,\bar{q}_2,\ldots,\bar{q}_m$, $m \leq n$, are mutually conjugate directions, then the minimum of the quadratic function $f(\bar{x})$ in the m-dimensional space containing $\bar{x}_o$ and the directions $\bar{q}_1,\bar{q}_2,\ldots,\bar{q}_m$ may be found by searching along each of the directions once only.

Proof. The required minimum of the quadratic function is the point $\bar{x}_o + \sum_{i=1}^{m} a_i\bar{q}_i$, where the parameters a_i, i = 1,2,...,m, are chosen by minimizing the function in the direction $\bar{q}_i$. From the definition of a quadratic function in Equation (2),

$$f(\bar{x}_o + \sum_{i=1}^{m} a_i\bar{q}_i) =$$

$$\frac{1}{2}(\bar{x}_o + \sum_{i=1}^{m} a_i\bar{q}_i)t_A(\bar{x}_o + \sum_{i=1}^{m} a_i\bar{q}_i) + \bar{b}^t(\bar{x}_o + \sum_{i=1}^{m} a_i\bar{q}_i) + c$$

$$= \frac{1}{2}\bar{x}_o^tA\bar{x}_o + \bar{x}_o^tA(\sum_{i=1}^{m} a_i\bar{q}_i) + \frac{1}{2}(\sum_{i=1}^{m} a_i\bar{q}_i)t_A(\sum_{i=1}^{m} a_i\bar{q}_i)$$

$$+ \bar{b}^t\bar{x}_o + \bar{b}^t\Big(\sum_{i=1}^{m} a_i\bar{q}_i\Big) + c$$

$$= f(\bar{x}_o) + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j \bar{q}_i^t A\bar{q}_j + \bar{x}_o^t A\Big(\sum_{i=1}^{m} a_i\bar{q}_i\Big) + \bar{b}^t\Big(\sum_{i=1}^{m} a_i\bar{q}_i\Big)$$

$$= f(\bar{x}_o) + \sum_{i=1}^{m}\Big(\frac{1}{2} a_i^2\bar{q}_i^t A\bar{q}_i + a_i\bar{q}_i^t(A\bar{x}_o + \bar{b})\Big). \qquad (8)$$

There are no terms with $a_i a_j$, $i = j$, since the directions $\bar{q}_1,\bar{q}_2,\ldots,\bar{q}_m$ are mutually conjugate. Consequently, the effect of searching in the direction $\bar{q}_i$ is to find a_i to minimize

$$\frac{1}{2} a_i^2\bar{q}_i^t A\bar{q}_i + a_i\bar{q}_i^t(A\bar{x}_o + \bar{b})$$

in that direction. Since this value of a_i is independent of the other terms of the sum in (8), then searching in each of the directions $\bar{q}_1,\bar{q}_2,\ldots,\bar{q}_m$ once only will find the absolute minimum in the space determined by $\bar{x}_o,\bar{q}_1,\bar{q}_2,\ldots,\bar{q}_m$. ∎

An alternate proof for m = n is given on pages 44-45.

Lemma 5.1. Let $\bar{x}_o$ be a point so that the function h_o defined by $h_o(\lambda) = f(\bar{x}_o + \lambda\bar{q})$ has a minimum at $\lambda = 0$, and let $\bar{x}_1$ be a point so that the function h_1 defined by $h_1(\lambda) = f(\bar{x}_1 + \lambda\bar{q})$ has a minimum at $\lambda = 0$. Then the direction $\bar{x}_1 - \bar{x}_o$ is conjugate to $\bar{q}$.

Proof. Now for a quadratic function

$$f(\bar{x}_o + \lambda\bar{q}) = \frac{1}{2}(\bar{x}_o + \lambda\bar{q})^t A(\bar{x}_o + \lambda\bar{q}) + \bar{b}^t(\bar{x}_o + \lambda\bar{q}) + c$$

$$= \frac{1}{2}\bar{x}_o^t A\bar{x}_o + \lambda\bar{q}^t A\bar{x}_o + \frac{1}{2}\lambda^2\bar{q}^t A\bar{q} + \bar{b}^t\bar{x}_o + \lambda\bar{b}^t\bar{q} + c.$$

Since $\bar{x}_o$ is a minimum in the direction $\bar{q}$, $\frac{\partial}{\partial\lambda}[f(\bar{x}_o + \lambda\bar{q})] = 0$ at $\lambda = 0$. Therefore,

$$\bar{q}^t(A\bar{x}_o + \bar{b}) = 0, \qquad \lambda = 0.$$

Also,

$$\bar{q}^t(A\bar{x}_1 + \bar{b}) = 0, \qquad \lambda = 0.$$

Hence

$$\bar{q}^t A(\bar{x}_1 - \bar{x}_o) = 0. \quad \blacksquare$$

The proof of quadratic convergence of Zangwill's method is given in the following theorem.

Theorem 5. For a positive definite quadratic function (Equation (2)) Zangwill's method stops at the minimum in step (iv) of iteration k where $k \leq n$.

Proof. The proof is by induction on k, the number of conjugate directions after k iterations.

Assume at the beginning of iteration k, $k \leq n-1$, that the directions $\bar{p}_{n-k+1}, \bar{p}_{n-k+2}, \ldots, \bar{p}_n$ are mutually conjugate and linearly independent. If the procedure does not stop in step (iv) of iteration k, $\alpha \neq 0$ and $\bar{x}_{n+1} \neq \bar{x}_o$ and $f(\bar{x}_o) \leq f(\bar{x}_{n+1})$. Since A is positive definite $f(\bar{x}_o) < f(\bar{x}_{n+1})$. From step (v), the point $\bar{x}_n$ is such that $f(\bar{x}_n) \leq f(\bar{x}_o)$ and thus $f(\bar{x}_n) \leq f(\bar{x}_o) < f(\bar{x}_{n+1})$ so that in step (vi), $\bar{x}_{n+1} \neq \bar{x}_n$ and $\bar{p}_{n+1} \neq \bar{0}$.

At iteration k-1, from step (viii), the last k directions to be employed were $\bar{p}_{n-k+1}, \ldots, \bar{p}_n$. Since these directions are assumed to be linearly independent, the point $\bar{x}_{n+1}$ in step (vii) of iteration k-1 is a minimum in the k-dimensional space containing $\bar{p}_{n-k+1}, \ldots, \bar{p}_n$ using Theorem 4. Similarly the point $\bar{x}_n$ in step (v) of iteration k is such a point. Thus from Lemma 5.1, $\bar{p}_{n+1}$ is mutually conjugate to $\bar{p}_{n-k+1}, \ldots, \bar{p}_n$. By the previous paragraph, $\bar{p}_{n+1}$ is non-zero; from induction, the directions $\bar{p}_{n-k+1}, \ldots, \bar{p}_n$ are all non-zero. By Theorem 2, they are linearly independent. Thus after iteration k the directions $\bar{p}_{n-(k+1)+1}, \ldots, \bar{p}_n$ are linearly independent and mutually conjugate.

The above argument holds for the first iteration which establishes the induction.

Thus if the procedure has not stopped by the beginning of iteration n, then n mutually conjugate and linearly independent directions have been generated. In step (v) - (vii) of iteration n - 1 the quadratic function has been minimized over these n directions, so that

the point $\bar{x}_{n+1}$ must be a minimum. The procedure will then stop in step (iv) of iteration n. ∎

Zangwill [29] establishes that this method converges to the minimum point of a strictly convex continuously differentiable function but the necessary lemmas will not be proved here.

Gradient Methods

When the gradient of a differentiable function f of n variables is known, it can be used to reduce the number of function evaluations in minimizing the function, but the computation of an n-component gradient is added. Gradient methods are now considered and the function will be assumed to be differentiable with continuous partial derivatives.

Steepest Descent

One of the oldest gradient methods was developed by Cauchy [7] in 1847 and is usually known as the method of steepest descent. After an initial point $\bar{x}_o$ is selected, the basic algorithm is as follows.

(i) Compute the gradient vector $\bar{g}_k$ at the point $\bar{x}_k$.

(ii) Let $\bar{p}_k = -\bar{g}_k$.

(iii) Find t_k such that the function h_k defined by $h_k(t) = f(\bar{x}_k + t\bar{p}_k)$ has a minimum at $t = t_k$.

(iv) Repeat steps (i) through (iii) for $k = 0,1,2,\ldots$, with $\bar{x}_{k+1} = \bar{x}_k + t_k\bar{p}_k$ as the next starting point until the convergence criterion is satisfied.

Note that the vector $\bar{p}_k$ in step (ii) can be normalized.

Example 9. The same function and starting point as in Examples 3-8 is used to illustrate this algorithm for four iterations in Table 6. Notice the orthogonality of consecutive $\bar{g}_i$'s and that $t_i > 0$.

Table 6. Steepest Descent

	i = 0	i = 1	i = 2	i = 3
$\bar{x}_i^t$	(1,2)	(1.4,0.8)	(0.2,0.4)	(0.28,0.16)
$f(\bar{x}_i)$	5	1	0.2	0.04
$\bar{g}_i^t$	(-2,6)	(1.2,0.4)	(-0.4,1.2)	(0.24,0.08)
$\bar{p}_i^t$	(2,-6)	(-1.2,-0.4)	(0.4,-1.2)	
t_i	0.2	1	0.2	

For the function f, the sequence $f(\bar{x}_o)$, $f(\bar{x}_1)$,... is a decreasing sequence since the function h_k defined by $h_k(t) = f(\bar{x}_k + t\bar{p}_k)$ has a negative derivative at $t = 0$, $h_k'(0) = \bar{g}_k^t\bar{p}_k = -|\bar{g}_k|$, and it is therefore possible to find a number $t > 0$ such that $h_k(t) < h_k(0)$, that is, such that $f(\bar{x}_k + t\bar{p}_k) < f(\bar{x}_k)$.

The method of steepest descent seems attractive, but the apparent advantage of searching in the direction of steepest descent is deceptive.

The vector $\bar{p}_1$ is the direction obtained after the function is minimized in the direction $\bar{p}_o$ and is perpendicular to p_o since $h_o'(t_o') = \bar{p}_o^t g_1 = 0$; similarly, $\bar{p}_2$ will be perpendicular to $\bar{p}_1$. With functions of two variables, $\bar{p}_2$ will be parallel to $\bar{p}_o$ and the method of steepest descent is basically the same as minimizing the function in the coordinate directions since in each case the directions of search are orthogonal at all stages, but the latter method is much simpler and easier. With more than two variables, the two methods are no longer necessarily equivalent; each $\bar{p}$ is perpendicular to the preceding direction. Since n successive directions of search do not necessarily form an orthogonal set of vectors (which they did for the direct search method), these directions do not necessarily span the domain space.

The method can be modified by selecting $\bar{x}_{k+1} = \bar{x}_k + at_k\bar{p}_k$ ($a \neq 1$, but a is near 1, for example, $a = 0.9$), meaning that successive $\bar{p}$'s are not perpendicular, but quadratic convergence is still not guaranteed.

Gradient Methods Using Conjugate Directions

Let $\bar{x}_o$ be an initial approximation to the point at which the minimum of a function occurs and let $\bar{p}_i$, $i = 1,2,\ldots,n$, be n linearly independent directions. For $i = 1,2,\ldots$, if $\bar{x}_{i+1}$ is the position of the minimum of the function $f(\bar{x})$ with respect to variations along the line through $\bar{x}_i$ in some specified direction $\bar{p}_{i+1}$, then

$$\bar{g}_{i+1}^t \bar{p}_{i+1} = 0, \tag{9}$$

where

$$\bar{x}_{i+1} = \bar{x}_i + a_{i+1}\bar{p}_{i+1} \tag{10}$$

for some scalar a_{i+1} since, for the function h defined by $h(t) = f(\bar{x}_i + t\bar{p}_{i+1})$, $h'(a_{i+1}) = \bar{g}_{i+1}^t\bar{p}_{i+1} = 0$. Consider the positive definite quadratic function defined by

$$f(\bar{x}) = \frac{1}{2}\bar{x}^t A\,\bar{x} + \bar{b}^t\bar{x} + c, \tag{2}$$

for which the gradient vector is

$$\bar{g}(\bar{x}) = A\,\bar{x} + \bar{b}. \tag{11}$$

An alternate proof of Theorem 4 will now be given for n = m.

Proof. Repeated use of Equation (10) gives

$$\bar{x}_n = \bar{x}_j + \sum_{i=j+1}^{n} a_i\bar{p}_i, \quad \text{for } 1 \le j \le n.$$

It then follows from Equation (11) that

$$\bar{g}_n = \bar{g}_j + \sum_{i=j+1}^{n} a_i A\bar{p}_i,$$

and therefore, from (9), that

$$\bar{g}_n^t\bar{p}_j = \sum_{i=j+1}^{n} a_i\bar{p}_i^t A\bar{p}_j.$$

Since the vectors $\bar{p}_1,\bar{p}_2,\ldots,\bar{p}_n$ are mutually conjugate by assumption,

$$\bar{g}_n^t\bar{p}_j = 0, \quad 1 \le j \le n,$$

and thus $\bar{g}_n = \bar{0}$ since $\bar{p}_1,\bar{p}_2,\ldots,\bar{p}_n$ form a basis by Theorem 2. The last equation implies that the minimum of the quadratic function has been found in E_n. ∎

General Partan

A method developed by Shah, et al. [24] is an alteration of the method of steepest descent and is called the method of parallel tangents or "partan."

Definition 12. In E_n, the set of points satisfying the linear equation

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = c, \tag{12}$$

where $a_1,a_2,\ldots,a_n,c$ are constants, is called a *hyperplane*. (It is an (n-1)-dimensional figure in n dimensions.) A hyperplane is tangent to a contour of the function f at $\bar{x}$ if $a_i = \partial f/\partial x_i$ and the point $\bar{x}$ satisfies Equation (12).

Definition 13. Three distinct points $\bar{x}$, $\bar{y}$, $\bar{z}$ in E_n are said to be *collinear* if $\bar{x} = a\bar{y} + b\bar{z}$ for constants a, b such that $a + b = 1$.

If Π_i denotes the hyperplane tangent to the contour of the function f at $\bar{x}_i$, then the algorithm for general partan can be stated as follows.

(i) Select an initial point $\bar{x}_o$ and any direction $\bar{q}_2$ such that $\bar{q}_2$ does not lie in the tangent hyperplane Π_o at $\bar{x}_o$, that is, so that $\bar{q}_2^t \bar{g}_o \neq 0$.

(ii) Calculate λ_1 so that the function h_1 defined by the equation $h_1(\lambda) = f(\bar{x}_o + \lambda \bar{q}_2)$ has a minimum at $\lambda = \lambda_1$ and let $\bar{p}_2 = \lambda_1 \bar{q}_2$ and $\bar{x}_2 = \bar{x}_o + \bar{p}_2$.

For $k = 1,2,3,\ldots$, repeat the next four steps until the convergence criterion is satisfied.

(iii) Find a vector $\bar{q}_{2k+1}$ such that

$$\bar{q}_{2k+1}^t \bar{g}_{2j} = 0, \quad j = 0,1,2,\ldots,k-1 \tag{13}$$

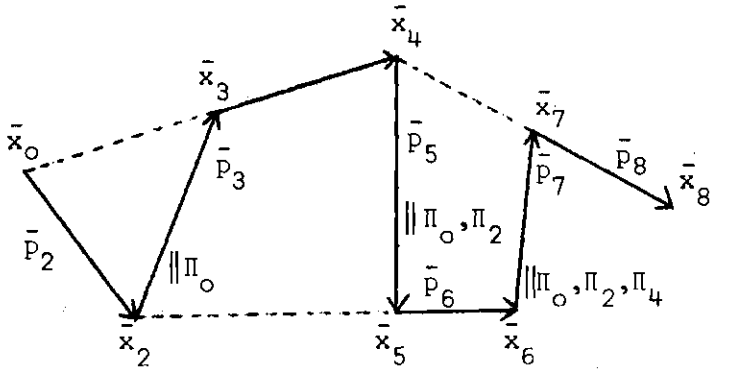
that is, $\bar{q}_{2k+1}$ is parallel to the planes $\Pi_o, \Pi_2, \ldots, \Pi_{2k-2}$. To find $\bar{q}_{2k+1}$ the dependent system of linear equations in (13) can be solved.

(iv) Calculate λ_{2k+1} so that the function h_{2k} defined by the equation $h_{2k}(\lambda) = f(\bar{x}_{2k} + \lambda \bar{q}_{2k+1})$ has a minimum at $\lambda = \lambda_{2k+1}$ and let $\bar{p}_{2k+1} = \lambda_{2k+1} \bar{q}_{2k+1}$ and $\bar{x}_{2k+1} = \bar{x}_{2k} + \bar{p}_{2k+1}$.

(v) Define $\bar{q}_{2k+2} = \bar{x}_{2k+1} - \bar{x}_{2k-2}$.

(vi) Calculate λ_{2k+2} so that the function h_{2k+1} defined by the equation $h_{2k+1}(\lambda) = f(\bar{x}_{2k+1} + \lambda \bar{q}_{2k+2})$ has a minimum at $\lambda = \lambda_{2k+2}$ and let $\bar{p}_{2k+2} = \lambda_{2k+2} \bar{q}_{2k+2}$ and $\bar{x}_{2k+2} = \bar{x}_{2k+1} + \bar{p}_{2k+2}$.

The algorithm of general partan (an alternative version called steepest descent partan will be stated later) is clarified by Figure 1 and the discussion that follows.



The solid line indicates the path of general partan. For steepest descent partan, the vectors $\bar{p}_2,\bar{p}_3,\bar{p}_5,\bar{p}_7$, ..., are steepest descent directions.

Figure 1. Schematic Diagram of Partan

From the initial point $\bar{x}_o$, proceed along a polygonal line determined by the points $\bar{x}_o,\bar{x}_2,\bar{x}_3,\bar{x}_4,\ldots$ such that each $\bar{x}_k$, $k = 2,3,\ldots$, is the minimum of the function f on the line through $\bar{x}_{k-1}$ in the direction $\bar{q}_k$. (At even numbered points directions of decreasing f must be chosen.) The direction $\bar{p}_2 = \bar{x}_2 - \bar{x}_o$ is otherwise arbitrary; $\bar{p}_3$ is any direction parallel to the plane Π_o; thereafter, for $k = 1,2,\ldots,\bar{x}_{2k+2}$ is collinear with $\bar{x}_{2k-2}$ and $\bar{x}_{2k+1}$ (by step (v) and (vi)), and for $k = 2,3,\ldots,\bar{p}_{2k+1}$ is parallel to $\Pi_o,\Pi_2,\Pi_4,\ldots,\Pi_{2k-2}$.

The fact that general partan reaches the minimum value of a positive definite quadratic function of n variables by the point $\bar{x}_{2n}$

will be proved after some initial observations of various properties of partan in relation to a quadratic function.

Consider a positive definite quadratic function. There is a vector $\bar{h}$ such that $A\bar{h} = \bar{b}$; since A is nonsingular, $\bar{h} = A^{-1}\bar{b}$. Then

$$f(\bar{x}) = \frac{1}{2}\bar{x}^t A\bar{x} + \bar{x}^t\bar{b} + c$$

$$= \frac{1}{2}\bar{x}^t A\bar{x} + \bar{x}^t A\bar{h} + \frac{1}{2}\bar{h}^t A\bar{h} - \frac{1}{2}\bar{h}^t A\bar{h} + c$$

$$= \frac{1}{2}(\bar{x}+\bar{h})^t A(\bar{x}+\bar{h}) + c - \frac{1}{2}\bar{h}^t A\bar{h}$$

and hence, with a transformation that replaces $\bar{x}+\bar{h}$ with $\bar{x}$, the quadratic function can be written as

$$f(\bar{x}) = \frac{1}{2}\bar{x}^t A\bar{x} + (c - \frac{1}{2}\bar{h}^t A\bar{h}), \tag{14}$$

where A is symmetric and positive definite and this function has a minimum value at the origin. The gradient vector is $\bar{g}_k = A\bar{x}_k$. Define

$$c_{i,j} = c_{j,i} = \bar{x}_i^t A\,\bar{x}_j = \bar{x}_i^t\,\bar{g}_j = \bar{x}_j^t\,\bar{g}_i. \tag{15}$$

The requirement that the vector from $\bar{x}_2$ to $\bar{x}_3$ be parallel to Π_o (as in step (iii)) can be expressed as

$$(\bar{x}_3-\bar{x}_2)^t\bar{g}_o = \bar{p}_3^t\,\bar{g}_o = 0, \quad \text{or} \quad c_{o,2} = c_{o,3}.$$

In general, Equation (13) can be stated as

$$c_{2j,2k} = c_{2j,2k+1}, \quad j = 0,1,2,\ldots,k-1, \quad k = 1,2,\ldots,n-1 \tag{16}$$

The collinearity of $\bar{x}_{2j}$, $\bar{x}_{2j-1}$, and $\bar{x}_{2j-4}$ in step (vi) gives

$$\bar{x}_{2j} = (1+\lambda_{2j})\,\bar{x}_{2j-1} - \lambda_{2j}\bar{x}_{2j-4} = \alpha_j\bar{x}_{2j-4} + \beta_j\bar{x}_{2j-1} \tag{17}$$

where $\alpha_j + \beta_j = 1$, $j = 2,3,\ldots,n$. Thus for any k,

$$c_{2j,k} = \alpha_j c_{2j-4,k} + \beta_j c_{2j-1,k}, \quad j = 2,3,\ldots,n. \tag{18}$$

Since $\bar{x}_2$ is the point where the minimum value of the function occurs on the line through $\bar{x}_o$ in the direction $\bar{q}_2$, $\bar{x}_o - \bar{x}_2$ is parallel to the hyperplane Π_2 and thus normal to $\bar{g}_2$, whence $(\bar{x}_o - \bar{x}_2)^t\bar{g}_2 = c_{o,2} - c_{2,2} = 0$, and in general,

$$c_{o,2} = c_{2,2}, \tag{19}$$

$$c_{j-1,j} = c_{j,j}, \quad j = 3,4,5,\ldots,2n, \tag{20}$$

since f is minimized in each of the directions $\bar{q}_j$, $j = 2,3,4,\ldots,2n$. Furthermore, since $\bar{x}_{2j}$, $\bar{x}_{2j-1}$, $\bar{x}_{2j-4}$ are collinear, then $\bar{x}_{2j} - \bar{x}_{2j-1} = \lambda_{2j}\bar{q}_{2j}$ is parallel to $\bar{x}_{2j-1} - \bar{x}_{2j-4} = \bar{q}_{2j}$, both of which are orthogonal to $\bar{g}_{2j}$ since h_{2j-1} has a minimum at $\lambda = \lambda_{2j}$, and the equations

$$c_{2j-4,2j} = c_{2j-1,2j} = c_{2j,2j}, \quad j = 2,3,\ldots,n, \tag{21}$$

holds.

Lemma 6.1. For $k = 1,2,\ldots,n$,

$$c_{2k,o} = c_{2k,2} = c_{2k,4} = \cdots = c_{2k,2k}. \tag{22}$$

Proof. By (19) the result is true for $k = 1$. Now assume (22) holds and use mathematical induction to prove

$$c_{2k+2,o} = c_{2k+2,2} = c_{2k+2,4} = \cdots = c_{2k+2,2k+2}. \tag{23}$$

Now

$$c_{2k+2,2k+2} - c_{2k+2,2k} = c_{2k+2,2k+1} - c_{2k+2,2k}, \qquad \text{by (20)},$$

$$= \alpha_{k+1}c_{2k-2,2k+1} + \beta_{k+1}c_{2k+1,2k+1} - \alpha_{k+1}c_{2k-2,2k} - \beta_{k+1}c_{2k+1,2k}, \qquad \text{by (18)},$$

$$= \alpha_{k+1}(c_{2k-2,2k+1} - c_{2k-2,2k}) + \beta_{k+1}(c_{2k+1,2k+1} - c_{2k+1,2k})$$

$$= 0, \quad \text{by (16) and by (20)}. \tag{24}$$

From (21),

$$c_{2k-2,2k+2} = c_{2k+2,2k+2}. \tag{25}$$

Now (24) and (25) establish the equality of the last three c's in (23). Equality of earlier terms can be established by taking $j \leq k-1$ and writing, by (18),

$$c_{2k+2,2k-2} - c_{2k+2,2j} = \alpha_{k+1}c_{2k-2,2k-2} + \beta_{k+1}c_{2k+1,2k-2}$$

$$- \alpha_{k+1}c_{2k-2,2j} - \beta_{k+1}c_{2k+1,2j}$$

$$= \alpha_{k+1}(c_{2k-2,2k-2} - c_{2k-2,2j}) + \beta_{k+1}(c_{2k+1,2k-2} - c_{2k+1,2j}). \tag{26}$$

The coefficient of α_{k+1} is zero by the induction hypothesis (22). If Equation (16) is applied to both terms of the coefficient of β_{k+1} in (26) above ($j \leq k-1$ is used in the second member) and Equation (22) is used again, then

$$c_{2k+1,2k-2} - c_{2k+1,2j} = c_{2k,2k-2} - c_{2k,2j} = 0$$

so that (26) is zero and (23) is established. ∎

An immediate consequence of the above lemma is

Lemma 6.2. The vectors $\bar{w}_1 = \bar{x}_2 - \bar{x}_o$, $\bar{w}_2 = \bar{x}_4 - \bar{x}_2, \ldots, \bar{w}_n = \bar{x}_{2n} - \bar{x}_{2n-2}$ are mutually conjugate.

Theorem 6. When general partan is used, the minimum value of a positive definite quadratic function is reached prior to the calculation of $\bar{x}_{2n+1}$.

Proof. The proof follows immediately upon application of Lemma 6.2 and Theorem 4. Alternatively, note that if the origin is not reached prior to the calculation of $\bar{x}_{2n+1}$, then the construction leads to non-null vectors $\bar{w}_1,\bar{w}_2,\ldots,\bar{w}_n$ for which by Lemma 6.1,

$$\bar{g}_{2n}^t\bar{w}_j = c_{2n,2j} - c_{2n,2j-2} = 0, \quad j = 1,2,\ldots,n.$$

The vectors $\bar{w}_1,\bar{w}_2,\ldots,\bar{w}_n$ are nonzero since A is positive definite and $\bar{w}_j^tA\bar{w}_j = \bar{x}_{2j}^tA\bar{x}_{2j} + \bar{x}_{2j-2}^tA\bar{x}_{2j} > 0$, $j = 1,2,\ldots,n$, unless $\bar{x}_{2j} = \bar{0}$ contradicting the fact that the minimum was not found prior to the calculation of $\bar{x}_{2n+1}$. But in n dimensions, only the null vector can be orthogonal to n mutually conjugate directions. Thus $\bar{g}_{2n} = \bar{0}$ and $\bar{x}_{2n}$ is the minimum. ∎

All points on the extended line through the points $\bar{x}_o$ and $\bar{x}_2$ correspond to vectors $a_o\bar{x}_o + a_2\bar{x}_2$ where $a_o + a_2 = 1$. Similarly, a k-dimensional set E_k of vectors is the collection of vectors of the form $\sum a_i\bar{x}_i$ where $\sum a_i = 1$, and the summations may be over $i = 0,2,4,\ldots,2k$, or equivalently, in view of the collinarity relation (17), over $i = 0,2,3,4,5,\ldots,2k$.

Corollary. $\bar{x}_{2k}$ occurs at the minimum of f in the set of vectors E_k; and if for $i = 0,2,3,4,\ldots,2k$, $\sum b_i = 0$, then $\bar{g}_{2k}^t(\sum b_i\bar{x}_i) = 0$.

Proof. These results follow from Theorem 6 by restricting attention to the set of vectors E_k. Alternately, note that a necessary and sufficient condition for a quadratic function to have a minimum value is that the gradient $\bar{g}_{2k} = \text{grad } f(\bar{x}_{2k})$ be normal to every vector parallel to the space E_k. But such vectors will always be the difference of two vectors of the form $\sum a_i \bar{x}_i$ where $\sum a_i = 1$, and hence have the form $\sum b_i \bar{x}_i$, where $\sum b_i = 0$. Any such vector may be represented using even-numbered indices only, and by Lemma 6.1,

$$\bar{g}_{2k}^t(\sum b_i \bar{x}_i) = (\sum b_i)c_{2k,o} = 0. \quad \blacksquare$$

Steepest Descent Partan

General partan allows arbitrary choices, within restrictions, of the directions $\bar{q}_o, \bar{q}_2, \bar{q}_4, \ldots, \bar{q}_{2n-4}$. An alternative method, called steepest descent partan, uses the method of steepest descent to determine even-numbered directions of search. That it has finite convergence for quadratic functions is proved by establishing a relationship between the two methods.

The previous algorithm for general partan is changed by substituting the following alternatives for steps (i) and (iii):

(ia) Select an initial point $\bar{x}_o$ and let $\bar{q}_2 = -\bar{g}_o$.

(iiia) Let $\bar{q}_{2k+1} = -\bar{g}_{2k}$, $k = 1,2,\ldots$.

Example 10. The steps in this alternate algorithm are given in Table 7, using the same positive definite function

$$f(x_1,x_2) = x_1^2 - 2x_1x_2 + 2x_2^2$$

considered in Examples 3-9.

Table 7. Steepest Descent Partan

		i = 2	i = 3	i=4
$\bar{x}_i^t$	(1,2)	(1.4,0.8)	(0.2,0.4)	(0,0)
$f(\bar{x}_i)$	5	1	0.2	0
$\bar{g}_i^t$	(-2,6)	(1.2,0.4)		
$\bar{q}_{i+1}^t$	(2,-6)	(-1.2,-0.4)	(-0.8,-1.6)	
λ_{i+1}	0.2	1	0.25	

Theorem 7. If f is a positive definite quadratic function of n variables (Equation (14)), then the minimum value of the function is reached prior to the calculation of $\bar{x}_{2n+1}$ when steepest descent partan is used.

Proof. By construction,

$$\bar{x}_o - \bar{x}_2 = \lambda_o \bar{g}_o, \tag{27}$$

$$\bar{x}_{2j} - \bar{x}_{2j+1} = \lambda_j \bar{g}_{2j}, \qquad j = 1,2,\ldots,n-1, \tag{28}$$

where $\lambda_j \neq 0$ when $\bar{x}_{2j}$ is not the minimum. With j = 1, Equations (27)

and (28) give

$$c_{o,2} - c_{o,3} = \bar{g}_o^t(\bar{x}_2 - \bar{x}_3) = \frac{\lambda_1}{\lambda_o}(\bar{x}_o - \bar{x}_2)^t\bar{g}_2,$$

which is zero by step (ii) and thus the two methods are consistent up to $\bar{x}_4$, that is, $\bar{w}_1^t A\bar{w}_2 = 0$ for both methods.

Now assume they are consistent up to $\bar{x}_{2k}$, that is, that $\bar{w}_1,\bar{w}_2,\ldots,\bar{w}_k$ are mutually conjugate. Then for $j = 0,1,\ldots,k-1$, by (28),

$$c_{2j,2k} - c_{2j,2k+1} = \bar{g}_{2j}^t(\bar{x}_{2k} - \bar{x}_{2k+1}) = \frac{a_k}{a_j}(\bar{x}_{2j} - \bar{x}_{2j+1})^t\bar{g}_{2k},$$

which is zero by the Corollary (to Theorem 6) which can be applied because of the induction hypothesis. Hence the equations required for reaching $\bar{x}_{2k+2}$ are satisfied and the induction is established. ∎

For a geometric proof of Theorems 6 and 7, refer to [24] and [28] in the Bibliography. Reference [24] also lists several alternatives.

If the procedure in either general partan or steepest descent partan has not been terminated before the calculation of $\bar{x}_{2n+1}$, then $\bar{x}_{2n}$ may be taken as a new initial $\bar{x}_o$, and the procedure started again (called *iterated partan*). Alternatively with steepest descent partan, the established pattern of alternating steepest descent and acceleration steps is continued (called *continued partan*). This is not practical for general partan due to storage requirements.

Conjugate Gradient Method

The method of conjugate gradients developed by Hestenes and Stiefel [17] is an n-step procedure for solving a set of simultaneous linear equations having a symmetric positive definite matrix of coefficients. The equivalence of that problem and the minimization of a quadratic function is clear from Equations (2) and (11) since the gradient vanishes if, and only if, $A\bar{x}_n = \bar{b}$ (that is, $\bar{x}_n$ is the solution to the system of equations).

This equivalence suggests the following minimization algorithm as stated by Fletcher and Reeves [16].

(i) Choose an initial point $\bar{x}_o$ and let $\bar{g}_o = \bar{g}(\bar{x}_o)$, $\bar{p}_o = -\bar{g}_o$.

(ii) Find a_i so that the function h_i defined by the equation $h_i(a) = f(\bar{x}_i + a\bar{p}_i)$ has a minimum at $a = a_i$, and set

$$\bar{x}_{i+1} = \bar{x}_i + a_i\bar{p}_i, \tag{29}$$

and $\bar{g}_{i+1} = \bar{g}(\bar{x}_{i+1})$.

(iii) Set $\beta_i = \bar{g}_{i+1}^2/\bar{g}_i^2$.

(iv) Set

$$\bar{p}_{i+1} = -\bar{g}_{i+1} + \beta_i\bar{p}_i. \tag{30}$$

(v) Repeat steps (ii) through (iv) until the convergence criterion is satisfied.

Example 11. The quadratic convergence of this method for the same positive definite function considered in Examples 3-10 is illustrated in Table 8.

Table 8. Conjugate Gradients

	i=o	i=1	i=2
$\bar{x}_i^t$	(1,2)	(1.4,0.8)	(0,0)
$f(\bar{x}_i)$	5	1	0
$\bar{g}_i^t$	(-2,6)	(1.2,0.4)	
$\bar{p}_i^t$	(2,-6)	(-1.12,-0.64)	
a_i	0.2	1.25	

Theorem 8. The method of conjugate gradients has quadratic convergence.

Proof. By Theorem 4, this theorem will be proved if the directions of search $\bar{p}_i$, $i = 0,1,\ldots,n-1$, are shown to be mutually conjugate for a positive definite quadratic function as defined in Equation (2). Since by (29), $\bar{x}_{i+1} = \bar{x}_i + a_i\bar{p}_i$,

$$\bar{g}_{i+1} = \bar{g}_i + a_iA\bar{p}_i, \quad i = 0,1,2,\ldots,n-1. \tag{31}$$

Now

$$f(\bar{x} + a\bar{p}) = \frac{1}{2}(\bar{x} + a\bar{p})^t A(\bar{x} + a\bar{p}) + \bar{b}^t(\bar{x} + a\bar{p}) + c$$

$$= \frac{1}{2}\bar{x}^t A\bar{x} + a\bar{p}^t A\bar{x} + \frac{1}{2}a^2\bar{p}^t A\bar{p} + \bar{b}^t\bar{x} + a\bar{p}^t\bar{b} + c$$

and

$$\frac{d}{da}h(a) = \frac{d}{da}f(\bar{x} + a\bar{p}) = a\bar{p}^t A\bar{p} + \bar{p}^t A\bar{x} + \bar{p}^t\bar{b}$$

$$= a\bar{p}^t A\bar{p} + \bar{p}^t(A\bar{x} + \bar{b})$$

$$= a\bar{p}^t A\bar{p} + \bar{p}^t\bar{g}.$$

Since $h'(a_i) = 0$,

$$a_i = -\frac{\bar{p}_i^t\bar{g}_i}{\bar{p}_i^t A\bar{p}}. \tag{32}$$

The vector $\bar{p}_i \neq \bar{0}$ unless either the minimum has been reached or the function was not minimized each time in step (ii). If Equation (30) is used repeatedly, then

$$\bar{p}_k = -\bar{g}_k - \bar{g}_k^2\left[\frac{\bar{g}_{k-1}}{\bar{g}_{k-1}^2} + \frac{\bar{g}_{k-2}}{\bar{g}_{k-1}^2} + \cdots + \frac{\bar{g}_i}{\bar{g}_i^2} + \cdots + \frac{\bar{g}_o}{\bar{g}_o^2}\right] = -\bar{g}_k^2\sum_{j=0}^{k}\frac{\bar{g}_j}{\bar{g}_j^2}. \tag{33}$$

That the vectors $\bar{g}_o, \bar{g}_1, \ldots,$ and $\bar{p}_o, \bar{p}_1, \ldots,$ satisfy

$$\bar{g}_i^t \bar{g}_j = 0, \quad i \neq j, \tag{34}$$

$$\bar{p}_i^t A \bar{p}_j = 0, \quad i \neq j, \tag{35}$$

$$\bar{p}_i^t \bar{p}_j = 0, \quad i < j, \quad \bar{p}_i^t \bar{g}_j = -\bar{g}_i^t, \quad i \geq j, \tag{36}$$

$$\bar{g}_i^t A \bar{p}_i = \bar{p}_i^t A \bar{p}_i, \quad \bar{g}_i^t A \bar{p}_j = 0, \quad i \neq j, \quad i \neq j+1, \tag{37}$$

will be proved by mathematical induction. Now

$$\bar{g}_o^t \bar{g}_1 = \bar{g}_o^t \bar{g}_o + a_o \bar{g}_o^t A \bar{p}_o = \bar{g}_o^t \bar{g}_o - \frac{\bar{p}_o^t \bar{g}_o}{\bar{p}_o^t A \bar{p}} \bar{g}_o^t A \bar{p}_o, \qquad \text{by (32)},$$

$$= \bar{g}_o^t \bar{g}_o - \bar{g}_o^t \bar{g}_o, \quad \text{since } \bar{p}_o = -\bar{g}_o \text{ in step (i)},$$

$$= 0.$$

Assume Equations (34), (35), (36), (37) hold for the vectors $\bar{g}_o, \bar{g}_1, \ldots, \bar{g}_k$ and $\bar{p}_o, \bar{p}_1, \ldots, \bar{p}_{k-1}$. To show that $\bar{p}_k$ can be adjoined to this set it is sufficient to show that

$$\bar{g}_i^t \bar{p}_k = -\bar{g}_k^2, \quad i \leq k, \tag{38}$$

$$\bar{p}_i^t A \bar{p}_k = 0, \quad i < k, \tag{39}$$

$$\bar{g}_k^t A \bar{p}_i = -\bar{p}_k^t A \bar{p}_i, \quad i \leq k, \quad i \neq k-1. \tag{40}$$

Equation (38) follows from (33) and (34). To prove (39), Equation (33) and

$$\bar{g}_{i+1}^t \bar{p}_k = \bar{g}_i^t \bar{p}_k + a_i \bar{p}_i^t A \bar{p}_k$$

are used. By (38) this becomes $-\bar{g}_k^2 = -\bar{g}_k^2 + a_i \bar{p}_i^t A \bar{p}_k$, $i < k$. Since $a_i > 0$ by (32), Equation (39) holds. In order to establish (40), (30) and (39) are used to obtain

$$\bar{p}_k^t A \bar{p}_i = -\bar{g}_k^t A \bar{p}_i + \beta_{k-1} \bar{p}_{k-1}^t A \bar{p}_i = -\bar{g}_k^t A \bar{p}_i, \qquad i \neq k-1.$$

It follows that (40) holds and hence that (34), (35), (36), (37) hold for the vectors $\bar{g}_o, \bar{g}_1, \ldots, \bar{g}_k$ and $\bar{p}_o, \bar{p}_1, \ldots, \bar{p}_k$. It remains to show that $\bar{g}_{k+1}$ can be adjoined to this set. This will be done by showing that

$$\bar{g}_i^t \bar{g}_{k+1} = 0, \qquad i \leq k, \tag{41}$$

$$\bar{p}_i^t A \bar{p}_{k+1} = 0, \qquad i < k, \tag{42}$$

$$\bar{p}_i^t \bar{g}_{k+1} = 0, \qquad i \leq k. \tag{43}$$

By (31)

$$\bar{g}_i^t \bar{g}_{k+1} = \bar{g}_i^t \bar{g}_k + a_k \bar{g}_i^t A \bar{p}_k.$$

If $i < k$, the terms on the right are zero by (34) and (37) and thus (41) holds. If $i = k$, the right member is zero by (37), (32), and (38). If (31) is used again, then for $i < k$,

$$0 < \bar{g}_{k+1}^t \bar{g}_{i+1} = \bar{g}_{k+1}^t \bar{g}_i + a_i \bar{g}_{k+1}^t A\bar{p}_i = a_i \bar{g}_{k+1}^t A\bar{p}_i.$$

Hence (42) holds. The Equation (43) follows from (41) and (33).

In the process of establishing (34) - (37), conjugate directions have been obtained from Equation (35). Hence by Theorem 4, the method of conjugate gradients has quadratic convergence. ∎

Davidon's Method

An efficient gradient method, originally developed by Davidon [9] and clarified by Fletcher and Powell [15] in 1963, uses a sequence of positive definite symmetric matrices $\{H_i\}$ which converges to the inverse of the Hessian matrix A of the function evaluated at the minimum. If the vector $\bar{p}_i$ is defined by $\bar{p}_i = -H_i\bar{g}_i$, $i = 1,2,\ldots$, then $\bar{s}_i$, the step taken to the minimum function value of a positive definite quadratic function in the direction $\bar{p}_i$ is an eigenvector of the matrix $H_{i+1}A$ which insures that H_i tends to A^{-1} evaluated at the minimum as the procedure converges. These ideas will be expanded on and proved later.

Since it is convenient to start with the unit matrix for H_o, meaning that the first iteration coincides with the method of steepest descent, the algorithm can be stated in the following form, where the current point is $\bar{x}_i$ with gradient $\bar{g}_i$ and the matrix is H_i.

(i) Set

$$\bar{p}_i = -H_i \bar{g}_i. \tag{44}$$

(ii) Obtain a_i such that the function h defined by the equation $h(a) = f(\bar{x}_i + a\bar{p}_i)$ has a minimum at $a = a_i$.

(iii) Set $\bar{x}_{i+1} = \bar{x}_i + \bar{s}_i$, where

$$\bar{s}_i = a_i \bar{p}_i. \tag{45}$$

(iv) Evaluate $f(\bar{x}_{i+1})$ and $\bar{g}_{i+1}$ and note that

$$\bar{g}_{i+1}^t \bar{s}_i = 0, \tag{46}$$

since $h'(a_i) = 0$.

(v) Set

$$\bar{y}_i = \bar{g}_{i+1} - \bar{g}_i. \tag{47}$$

(vi) Set

$$H_{i+1} = H_i + B_i + C_i, \tag{48}$$

where

$$B_i = \frac{\bar{s}_i \bar{s}_i^t}{\bar{s}_i^t \bar{y}_i} \quad \text{and} \quad C_i = -\frac{H_i \bar{y}_i \bar{y}_i^t H_i}{\bar{y}_i^t H_i \bar{y}_i}.$$

(Neither denominator is zero since H_i is a positive definite matrix for all i as will be proved later.)

The theoretical justification for the manner in which the matrices are modified and the proof of quadratic convergence will be given later.

Example 12. The algorithm for the method is illustrated in Table 9 for the same positive definite quadratic function,

$$f(x_1,x_2) > x_1^2 - 2x_1x_2 + 2x_2^2,$$

considered in Examples 3-11.

Theorem 9. In step (ii) of the algorithm there is an $a_i > 0$.

Proof. For the function $h(c) = f(\bar{x}_i + c\bar{p}_i)$, $h'(0) = \bar{g}_i^t \bar{p}_i$. If $h'(0) < 0$, then there is a member $c > 0$ such that $h(c) < h(0)$ and $h'(c) = 0$. Let this number be a_i. Thus $f(\bar{x}_i + a_i\bar{p}_i) < f(\bar{x}_i)$, for $a_i > 0$, will hold for $\bar{p}_i$ if $-\bar{p}_i^t\bar{g}_i = \bar{g}_i^t H_i \bar{g}_i$ is positive which is true for all possible $\bar{g}_i \neq \bar{0}$ if H_i is a positive definite matrix. In view of the fact that the initial H_o has been chosen to be positive definite (and symmetric), the proof will use an inductive argument.

Table 9. Davidon's Method

	i = o	i=1	i=2
$\bar{x}_i^t$	(1,2)	(1.4,0.8)	(0,0)
$f(\bar{x}_i)$	5	1	0
$\bar{g}_i^t$	(-2,6)	(1.2,0.4)	(0,0)
H_i	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} .77384 & .37077 \\ .37077 & .42615 \end{bmatrix}$	$\begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$
$\bar{p}_i^t$	(2,-6)	(1.07672,.61538)	
a_i	0.2	-1.3	
$\bar{s}_i^t$	(0.4,-1.2)	(-1.4,0.8)	

Assume that H_i is positive definite and consequently that a_i is positive to show that $\bar{x}^t H_{i+1}\bar{x} > 0$ for any $\bar{x} \neq \bar{0}$. Since H_i is assumed positive definite and symmetric (see step (vi)), then there is a positive definite matrix U_i such that $(U_i)^2 = H_i$ or $U_i = (H_i)^{1/2}$ (see [10], p. 91). Define $\bar{u} = (H_i)^{1/2}\bar{x}$ and $\bar{v} = (H_i)^{1/2}\bar{y}_i$. By the definition of H_{i+1} in step (vi)

$$\bar{x}^t H_{i+1}\bar{x} = \bar{x}^t H_i \bar{x} + \frac{\bar{x}^t \bar{s}_i \bar{s}_i^t \bar{x}}{\bar{s}_i^t \bar{y}_i} - \frac{\bar{x}^t H_i \bar{y}\ \bar{y}_i^t H_i \bar{x}}{\bar{y}_i^t\ H_i \bar{y}_i}$$

$$= \frac{\bar{u}^t\bar{u}\ \bar{v}^t\bar{v} - (\bar{u}^t\bar{v})^2}{\bar{v}^t\bar{v}} + \frac{(\bar{x}^t\bar{s}_i)^2}{\bar{s}_i^t\bar{y}_i} \geq \frac{(\bar{x}^t\bar{s}_i)^2}{\bar{s}_i^t\bar{y}_i}$$

by the Schwarz inequality. But

$$\bar{s}_i^t\bar{y}_i = \bar{s}_i^t\bar{g}_{i+1} - \bar{s}_i^t\bar{g}_i, \qquad \text{by step (v)},$$

$$= -\bar{s}_i^t\bar{g}_i, \qquad \text{by (46)},$$

$$= a_i\ \bar{g}_i^t H_i \bar{g}_i, \qquad \text{by definition of } \bar{s}_i,$$

$$> 0, \qquad \text{by assumption.}$$

Hence $\bar{x}^t H_{i+1}\ \bar{x} > 0$ for all $\bar{x} \neq \bar{0}$. Therefore, H_{i+1} is positive definite and the function is decreased at each iteration. ∎

For the proof of quadratic convergence, let f be a positive definite quadratic function as given in Equation (2) with gradient given by Equation (11). At the minimum $\bar{x}_o$, $\bar{0} = A\bar{x}_o + \bar{b}$. Subtracting Equation (11) from this gives the equation $A\bar{x}_o - A\bar{x} = -\bar{g}$, which gives the difference between a point $\bar{x}$ and the minimum $\bar{x}_o$ as

$$\bar{x}_o - \bar{x} = -A^{-1}\bar{g}. \tag{49}$$

The following lemma will be used later.

Lemma 10.1. The directions $\bar{s}_0, \bar{s}_1, \ldots, \bar{s}_k, k < n$ (defined in step (iii)) are linearly independent eigenvectors of $H_{k+1}A$ with eigenvalues unity.

Proof. By definition, for $i = 0,1,\ldots,n-1$,

$$\bar{y}_i = \bar{g}_{i+1} - \bar{g}_i = A\bar{x}_{i+1} - A\bar{x}_i, \qquad \text{by (11)},$$

$$= A\bar{s}_i. \qquad (50)$$

If (50) is premultiplied by H_{i+1}, then

$$H_{i+1}A\bar{s}_i = H_{i+1}\bar{y}_i$$

$$= H_i\bar{y}_i + \bar{s}_i - H_i\bar{y}_i, \qquad \text{by (18)},$$

$$= \bar{s}_i. \qquad (51)$$

Consider the equations

$$\bar{s}_i^t A\bar{s}_j = 0, \quad 0 \le i < j < k, \qquad (52)$$

$$H_k A\bar{s}_i = \bar{s}_i, \quad 0 \le i < k, \qquad (53)$$

and use mathematical induction on k. For k = 1,

$$H_1 A\bar{s}_o = \bar{s}_o, \qquad \text{by (51)},$$

and for k = 2,

$$\bar{s}_o^t A \bar{s}_1 = a_1 \bar{s}_o^t A \bar{p}_1, \qquad \text{by (45)},$$

$$= -a_1 \bar{s}_o^t A H_1 \bar{g}_1, \qquad \text{by (44)},$$

$$= -a_1 \bar{s}_o^t \bar{g}_1, \qquad \text{by (51)},$$

$$= 0, \qquad \text{by (46)}.$$

Now assume that (52) and (53) hold for k. From (11),

$$\bar{g}_k = \bar{b} + A\bar{x}_k$$

$$= \bar{b} + A(\bar{x}_{i+1} + \bar{s}_{i+1} + \bar{s}_{i+2} + \cdots + \bar{x}_{k-1}),$$

for $0 \le i \le k-1$,

$$= \bar{g}_{i+1} + A(\bar{s}_{i+1} + \bar{s}_{i+2} + \cdots + \bar{x}_{k-1}), \qquad \text{by (11)}.$$

Multiply by $\bar{s}_i^t$ and use assumption (52) to get

$$\bar{s}_i^t \bar{g}_k = \bar{s}_i^t \bar{g}_{i+1}$$

$$= 0 \qquad (54)$$

for $0 \le i < k$, by (46).

From assumption (53),

$$\bar{g}_k^t H_k A \bar{s}_i = \bar{s}_i^t A H_k g_k = \bar{g}_k^t \bar{s}_i = 0, \qquad \text{by (54)},$$

and by (44), (45) and the fact that $a_i > 0$ by the previous theorem,

$$\bar{s}_i^t A \bar{s}_k = 0, \qquad 0 \le i < k. \tag{55}$$

concluding the induction for Equation (52). (Notice that Equation (55) states that the directions $\bar{s}_0, \bar{s}_1, \ldots, \bar{s}_k, k<n$, are mutually conjugate.)

Now

$$\bar{y}_k^t H_k A \bar{s}_i = \bar{y}_k^t \bar{s}_i, \qquad \text{by (53)},$$

$$= \bar{s}_k^t A \bar{s}_i, \qquad \text{by (50)},$$

$$= 0 \tag{56}$$

$$0 \le i < k, \qquad \text{by (55)}.$$

From (48),

$$H_{k+1} A \bar{s}_i = (H_k + B_k + C_k) A \bar{s}_i$$

$$= H_k A \bar{s}_i + \frac{\bar{s}_k (\bar{s}_k^t A \bar{s}_i)}{\bar{s}_k^t \bar{y}_k} - \frac{H_k \bar{y}_k (\bar{y}_k^t H_k A \bar{s}_i)}{\bar{y}_k^t H_k \bar{y}_k}.$$

But since

$$\bar{s}_k^t A \bar{s}_i = 0, \quad 0 \le i < k, \qquad \text{by (55)},$$

and

$$\bar{y}_k^t H_k A \bar{s}_i = 0, \quad 0 \le i < k, \qquad \text{by (56)},$$

then

$$H_{k+1} A \bar{s}_i = H_k A \bar{s}_i$$

$$= \bar{s}_i, \qquad (57)$$

$$0 \le i < k, \qquad \text{By assumption (53)},$$

concluding the induction for Equation (53).

The vectors $\bar{s}_o, \bar{s}_1, \ldots, \bar{s}_k, k < n$, are mutually conjugate by Equation (55), nonzero unless $\bar{g}_i = \bar{0}$ since H_i is positive definite and $a_i > 0$, and linearly independent by Theorem 2, while Equation (53) implies they are eigenvectors of the matrix $H_{k+1}A$ with eigenvalue unity. ∎

Theorem 10. Davidon's method has quadratic convergence.

Proof. The proof is obvious from Lemma 10.1 and Theorem 4. ∎

By Lemma 10.1, $H_n A$ has eigenvalues unity with linearly independent eigenvectors $\bar{s}_o, \bar{s}_1, \ldots, \bar{s}_{n-1}$. This implies that $H_n A$ is the identity matrix, that is, $H_n A = I$, and hence $H_n = A^{-1}$. When A^{-1} has been found,

then the minimum of a quadratic can be found from any starting point $\bar{x}$ by the simple use of Equation (49).

Note on the Construction of B_i in Equation (48). B_i is the factor which makes H tend to A^{-1} in the sense that for a quadratic function $A^{-1} = \sum_{i=0}^{n-1} B_i$. To establish this fact, define the matrix S to be $S = [\bar{s}\ \bar{s}_1 \ \ldots\ \bar{s}_{n-1}]$, where $\bar{s}_i$ are column vectors. Because of Equation (52), $S^tAS = D$ where D is a diagonal matrix with diagonal elements $d_{ii} = \bar{s}_{i-1}^t A \bar{s}_{i-1}$, $i = 1,2,\ldots,n$. D^{-1} exists since A was assumed positive definite and S^{-1} exists since the vectors $\bar{s}_o, \bar{s}_1, \ldots, \bar{s}_{n-1}$ are linearly independent. Hence

$$A = (S^t)^{-1} D S^{-1} = (S D^{-1} S^t)^{-1}$$

and therefore

$$A^{-1} = S D^{-1} S^t.$$

Since D is a diagonal matrix,

$$A^{-1} = \sum_{i=1}^{n} \frac{1}{d_{ii}} (\bar{s}_{i-1} \bar{s}_{i-1}^t) = \sum_{i=0}^{n-1} \frac{\bar{s}_i \bar{s}_i^t}{\bar{s}_i^t A \bar{s}_i}.$$

Therefore from Equation (50) and the definition of B_i,

$$A^{-1} = \sum_{i=0}^{n-1} B_i.$$

Note on the Construction of C_i in Equation (48). The reason for the way C_i is chosen can be deduced from the fact that Equation (51) must be valid. From (51) and (48),

$$H_{i+1}A\bar{s}_i = \bar{s}_i = H_iA\bar{s}_i + B_iA\bar{s}_i + C_iA\bar{s}_i.$$

Since

$$B_iA\bar{s}_i = B_i\bar{g}_i = \bar{s}_i,$$

the equation

$$C_iA\bar{s}_i = C_i\bar{y}_i = -H_iA\bar{s}_i = -H_i\bar{y}_i, \qquad \text{by (58) and (50),}$$

must be satisfied. From this equation, the simplest form for C_i is given by the equation

$$C_i = -\frac{H_i\bar{y}_i\bar{z}^t}{\bar{z}^t\bar{y}_i}$$

for some vector $\bar{z}$, but since C_i is to be a symmetric matrix, this gives

$$C_i = -\frac{(H_i\bar{y}_i)(\bar{y}_i^tH_i)}{\bar{y}_i^tH_i\bar{y}_i}.$$

Criterion for Convergence

Quadratic convergence is assured when some methods are used on positive definite quadratic functions, but if the function is not quadratic or if the method does not have quadratic convergence, then the number of iterations necessary to attain a minimum value of the function is not necessarily finite and therefore some convergence criterion must be established to determine when the iterating should be stopped. Ideally this criterion would be to stop the iterating when the absolute value of the differences between the predicted component value of the variables x_i and the actual component values at a true minimum were less than given small amounts ϵ_i, $i = 1,2,\ldots,n$. In any event, when a criterion is chosen, a compromise has to be made between stopping the iterative procedure too soon and calculating $f(\bar{x})$ an unnecessarily large number of times.

If the gradient $\bar{g}$ of the function is available, then having the value of $\bar{g}^t\bar{g}$ at the current point be less than a specified number is one suggestion for a criterion; this uses the fact that the partial derivatives of a function are zero at a stationary point. Alternatively, if the change of the function value per iteration or the change of each of the variables per iteration is less than some predetermined number, then the iterating could be stopped. The latter is easy to use and usually has the desired result. After a criterion is chosen, it is usually possible to find a function for which the iterating does not stop at the minimum when this criterion is used. Powell [22] gives a procedure for his method which he claims has never failed to yield the required accuracy. Essentially the procedure uses his method to

minimize the function; another point is chosen near this resulting approximation to the minimum and then Powell's method is used again on the function with this starting value. From the two points resulting from the two uses of Powell's method, a choice is made for the point where the minimum occurs.

CHAPTER IV

CONSTRAINTS

Previously, the problem of minimizing a function without constraints on the independent variables was discussed. Now methods for minimizing $f(\bar{x})$ subject to the constraints

$$g_i(\bar{x}) \geq 0 \qquad i = 1,\ldots,m \tag{1}$$

will be considered. Since the constraint $r(\bar{x}) \leq 0$ can be written as $g(\bar{x}) = -r(\bar{x}) \geq 0$ and $r(\bar{x}) = 0$ as $g_1(\bar{x}) = r(\bar{x}) \geq 0$ and $g_2(x) = r(x) \geq 0$, all possibilities are covered.

If the function f and its constraints are linear, this problem can be solved by linear programming techniques or similar methods in a finite number of steps. If f is nonlinear, then methods of minimization which use tangent plane approximations to the constraints (if they are also non-linear) have been developed. The approach of converting the constrained problem into one which is not constrained will be considered here.

Transformations

By transformations of independent variables from an x-coordinate system to a y-coordinate system, it is possible to incorporate some types of constraints into the function, possibly giving an unconstrained problem. Thus for the function f subject to constraints (1) (these

define a "region" R in the x-space, that is, the subsets of E_n with coordinate variables indicated by x_i), it may be possible to find a transformation T (not necessarily linear) which maps the y-space, that is, E_n, onto R, T: $E_n \rightarrow R$; and then the y-space is searched for a minimum of $f(T(\bar{y}))$ which accomplishes the same result as searching the region R in the x-space. Some if not all the constraints may be dealt with in this way. The decision as to which transformation T, if any, to use depends on various factors, such as the kind of constraints in Equation (1), the minimization method used, whether the method requires a gradient, the ease of finding the inverse transformation, and the behavior of the function on the boundary of the region, that is, where equality holds in (1).

As an example of the type of constraints which can be eliminated, suppose that three independent variables x_1, x_2, x_3 are to be constrained by $0 \leq x_1 \leq x_2 \leq x_3$. Then by the transformation

$$x_1 = y_1^2$$

$$x_2 = y_1^2 + y_2^2$$

$$x_3 = y_1^2 + y_2^2 + y_3^2,$$

it is seen that a minimization procedure with no provision for incorporating constraints can now be used to minimize the function in the y-space. It is obvious that this transformation is nonlinear.

Other suggested transformations are

$$\text{(i)} \quad x_i = y_i^2,$$

$$\text{(ii)} \quad x_i = e^{y_i},$$

$$\text{(iii)} \quad x_i = |y_i|,$$

$$\text{(iv)} \quad x_i = \sin^2 y_i.$$

The transformations in (i) and (iii) constrain the variable x_i to non-negative values while transformation (ii) restricts it to strictly positive values. The use of (iii) destroys the differentiability for some values of y_i of the function on which the transformation is used. Transformation (iv) restricts x_i to the range $0 \leq x_i \leq 1$. If each independent variable is subject to constant lower and upper constraints, for example, $a_i \leq x_i \leq b_i$, $i = 1,2,\ldots,n$, then the permissible region consists of a rectilinear "box" in n dimensions. If the transformation $x_i = a_i + (b_i - a_i) \sin^2 y_i$ is used, then an unconstrained optimum in y-space can be sought. The periodicity of the solutions in y-space should not cause any difficulty provided the method of minimization in use does not take steps so large that it jumps from peak to peak.

Created Response Surface Technique

The created response surface technique developed by Carroll [6] will now be described. If the constraints are never allowed to be

violated during the minimization process, then the resulting minimum value of the function is sure to be a feasible one, that is, one where no constraints are violated, assuming that an initial feasible starting point is known. This requirement can be satisfied by devising a penalty which is added to the value of the function to give a new function h called a created response function and which becomes increasingly severe as constraint boundaries are approached.

If f is the function to be minimized and if the m constraints are expressed in the form

$$g_i(\bar{x}) \geq 0, \qquad i = 1,2,\ldots,m, \tag{1}$$

the created response function h can be written as

$$h(\bar{x},r_k) = f(\bar{x}) + r_k \sum_{i=1}^{m} \frac{w_i}{g_i(\bar{x})},$$

where $w_i > 0$, $i = 1,2,\ldots,m$, and $r_k \geq 0$, $k = 1,2,\ldots$. The summation represents the "penalty" in the sense that as any constraint $g_i(\bar{x})$ approaches its limiting value (zero), h approaches positive infinity (since (1) holds). In this way a progressively severe penalty is imposed as the limiting value of a constraint is approached. The w_i's (a footnote in [12] indicates these might just as well be chosen equal to one as will be done here) weigh the individual constraints among themselves while r_k weighs the sum of these penalties in relation to the function f.

The iterative procedure on r_k as given in [13] is described by the following algorithm. A detailed analysis of each step is given in [13], but the main ideas will be expanded in this paper after the algorithm is stated. Let a number $c > 1$ be given.

(i) Select a point $\bar{x}_o$ such that $g_i(\bar{x}_o) > 0$, $i = 1,2,\ldots,m$.

(ii) Select an initial value of r_1, $r_1 > 0$.

For $k = 1,2,\ldots$, repeat the following four steps.

(iii) Determine a minimum $\bar{x}(r_k)$ of $h(\bar{x},r_k)$ for the current value of r_k using $\bar{x}(r_{k-1})$ as a starting point and check the convergence criterion (to be discussed on p. 82).

(iv) If $k > 1$, estimate the minimum $\bar{x}^*$ of f subject to the constraints by the approximation

$$\bar{x}^* \doteq \frac{c^{1/2}\bar{x}(r_k) - \bar{x}(r_{k-1})}{c^{1/2} - 1} \tag{58}$$

(v) If the convergence criterion is satisfied, then terminate computations. If not, select $r_{k+1} = r_k/c$.

(vi) If $k > 1$, estimate the minimum $\bar{x}(r_{k+1})$ of $h(\bar{x},r_{k+1})$ using the approximation

$$\bar{x}(r_{k+1}) \doteq \bar{x}(r_k) + \frac{1}{c^2}[\bar{x}(r_k) - \bar{x}(r_{k-1})] \tag{59}$$

and check the convergence criterion.

Example 13. To illustrate the created response surface technique, let the function f be defined by

$$f(x_1,x_2) = x_1^2 + 2\,x_2^2$$

and subject to $x_1 \geq 0$, $x_2 \geq 0$; it has a minimum value of 0 at the point (0,0). Since the constraints can be written as

$$g_1(\bar{x}) = x_1 \geq 0$$

$$g_2(\bar{x}) = x_2 \geq 0$$

then let

$$h(\bar{x},r) = x_1^2 + 2\,x_2^2 + r\left[\frac{1}{x_1} + \frac{1}{x_2}\right].$$

Let $\bar{x}_o^t = (2,3)$ and let c = 10; using Equation (60), r_1 is calculated as 26.7216. Define the error function

$$E(\bar{x},r) = r\left[\frac{1}{x_1} + \frac{1}{x_2}\right]$$

and stop when $E(\bar{x},r) < 1$. The calculations are given in Table 10.

Table 10. Created Response Surface Technique

i		r_i	$x_{1,i}$	$x_{2,i}$	$f(\bar{x}_i)$	$E(\bar{x}_i, r_i)$
0			2	3	22	
1	$\bar{x}(r_1)$	26.7216	2.3729	1.8834	12.7255	25.4487
2	$\bar{x}(r_2)$	2.67216	1.1014	.8741	2.7412	5.4831
	$\bar{x}^*$		.5133	.4073	.5924	11.7652
	$\bar{x}_{est}$		1.0887	.8640	2.6796	5.5471
3	$\bar{x}(r_3)$	.267216	.5112	.4057	.5906	1.1813
	$\bar{x}^*$		.2383	.1891	.1283	2.5342
	$\bar{x}_{est}$		.5053	.4011	.5771	1.1951
4	$x(r_4)$	.0267216	.2373	.1883	.1272	.2545
	x^*		.1106	.0878	.0277	.5459

For the method to compute the initial interior point that satisfies the inequalities $g_i(\bar{x}) > 0$, $i = 1,2,\ldots,m$, let $\bar{x}_k$ be a given point and define the sets $S_k = \{s: g_s(\bar{x}_k) \leq 0\}$ and $T_k = \{t: g_t(\bar{x}_k) > 0\}$. A sequence of points is generated that increases the value of $g_i(\bar{x})$ for $i \in S_k$ until $i \in T_k$ without violating any of the inequalities already satisfied. For the computational procedure the following algorithm is given.

(i) Define

$$h(\bar{x},r) = -g_{s_1}(\bar{x}) + r \sum_{t\epsilon T_k} \frac{1}{g_t(\bar{x})},$$

where $s_1 \epsilon S_k$.

(ii) Find a point $\bar{x}_{k+1}$ such that $h(\bar{x}_{k+1},r) < h(\bar{x}_k,r)$. (Any of the methods in the previous chapter can be initiated.)

(iii) Evaluate $g_s(\bar{x}_{k+1})$ for all $s \epsilon S_k$ and define the sets S_{k+1} and T_{k+1}.

(iv) If S_{k+1} is nonempty, go to step (i).
Otherwise, an initial interior point has been found.

In reference [13] Fiacco and McCormick give a rationale for computing r_1, the initial value of r_k, consistent with attempting to reduce the effort of minimizing $h(\bar{x},r_k)$. They also state that choosing r_1 extremely large or extremely small (these extremes depend on x_o) results in an increase in the number of iterations. If $p(\bar{x}) = \sum_{i=1}^{m} \frac{1}{g_i(\bar{x})}$ and $\bar{g}(\bar{x}_o)$ and $\bar{q}(\bar{x}_o)$ denote the gradient of $f(\bar{x})$ and $p(\bar{x})$, respectively, evaluated at the initial point $\bar{x}_o$, then r_1 is given by

$$r_1 = - \frac{\bar{g}(\bar{x}_o)^t \bar{q}(\bar{x}_o)}{|\bar{q}(\bar{x}_o)|^2} . \qquad (60)$$

If $\bar{g}(\bar{x}_o)^t\bar{q}(\bar{x}_o) \geq 0$, giving $r_1 \leq 0$, then proceed by taking a sequence of steps in the direction $-\bar{g}(\bar{x}_o)$ and recomputing r_1 using Equation (60) at

each new point until a positive r_1 is obtained or until an unconstrained minimum of f is achieved (see pp. 72-73). That one of these alternatives will prevail and other ways of choosing r_1 are given in [13] as well as methods for reducing r_k after each h minimization.

The method of minimizing $h(\bar{x},r_k)$ in step (iii) can be any of those discussed in the previous chapter or the first and second-order gradient methods which are summarized in [13].

Fiacco and McCormick [13] use extrapolation formulas based on the fact that the decrease in $\bar{x}(r_k)$ each time is approximately linear in $r^{1/2}$ (derived from experience) to get a "first-order" estimate of the point where the minima $\bar{x}^*$ (for (58)) and $\bar{x}(r_{k+1})$ (for (59)) occur.

If any point whose functional value $f(\bar{x})$ is within $\varepsilon > 0$ of a true minimum value is acceptable as the solution to the constrained minimization problem, then a useful convergence criterion is to terminate the algorithm above when

$$0 < r \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r))} < \varepsilon$$

for any $r = r_k$. This criterion is suggested by the Corollary 12.2 to be stated later.

To prove the convergence of the sequence of values of $h(\bar{x},r_k)$ to a minimum value of the function (if r_k is a strictly monotonic decreasing sequence and $r_k \to 0$ as $k \to \infty$), that is, that $f(\bar{x}(r_k)) \to f(\bar{x}^*)$ (where x^* is the point where a minimum value of the function f occurs subject to $g_i(\bar{x}) \geq 0$, $i = 1,2,\ldots,m$) as $r_k \to 0$, the following conditions are imposed:

(a) $R^{\circ} = \{\bar{x}: g_i(\bar{x}) > 0,\ i = 1,2,\ldots,m\}$ is nonempty.

(b) The functions $f, g_1, \ldots, g_m$ are twice continuously differentiable.

(c) For every finite k, the set $D = \{\bar{x} : f(\bar{x}) \leq k,\ x \in R\}$ is a bounded set (hence is compact) where $R = \{\bar{x} : g_i(\bar{x}) \geq 0,\ i = 1,2,\ldots,m\}$.

Lemma 11.1. There is a point $\bar{x}^* \in R$ such that $f(\bar{x}^*) \leq f(\bar{x})$ for any $\bar{x} \in R$.

Proof. Let k_1 be a number such that D is nonempty. For the points $\bar{x}$ in R but not in D, $f(\bar{x}) > k_1$ and $D \subset R$. Since f is a continuous function on a compact set D then (by Theorem 4-20, p. 73, Apostol [1]) there is a point $\bar{x}^* \in D$ $(\bar{x}^* \in R)$ such that $f(x^*) \leq f(\bar{x})$ for $\bar{x} \in D$. For any $\bar{x} \in R - D$, $f(\bar{x}^*) \leq k_1 < f(\bar{x})$. ∎

Let $$f_o = f(\bar{x}^*) = \inf_{\bar{x} \in R} f(\bar{x}).$$

Lemma 11.2. If R° is not empty, there is a point $\bar{x}(r_k) \in R^{\circ}$ such that $h(\bar{x}(r_k), r_k) \leq h(\bar{x}, r_k)$ for any $\bar{x} \in R^{\circ}$.

Proof. Let $\bar{x}_o \in R^{\circ}$ be the point where the minimization procedure begins and let $M_o = h(\bar{x}_o, r_k)$. Define the sets $S_o = \{\bar{x} : f(\bar{x}) \leq M_o,\ \bar{x} \in R\}$, $S_i = \{\bar{x} : r_k / g_i(\bar{x}) \leq M_o - f_o,\ \bar{x} \in R\}$ for $i = 1,2,\ldots,m$. Finally, let $S = \bigcap_{i=1}^{n} S_i$. S is nonempty because $\bar{x}_o \in S_o$ since $f(\bar{x}_o) \leq h(\bar{x}_o, r_k)$ and

$\bar{x}_o \in S_i$, $i = 1,\ldots,m$, since $M_o = h(\bar{x}_o,r_k) \geq f_o + \frac{r_k}{g_i(\bar{x}_o)}$. S_o is closed and bounded by choosing $k = M_o$ in assumption (c) above. For the sets S_i, $i = 1,2,\ldots,m$, $g_i(\bar{x}) \geq \frac{r_k}{M_o - f_o} > 0$ and the sets S_i are closed since the functions g_i, $i = 1,2,\ldots,m$, are continuous. Therefore S is closed and S is bounded and hence S is compact.

Since h is a continuous function on a compact set S, there is a point $\bar{x}(r_k) \in S$ such that $h(\bar{x}(r_k),r_k) \leq h(\bar{x},r_k)$ for any $\bar{x} \in S$.

Let $\bar{x} \in R^o$. If $\bar{x} \notin S$, then either $f(\bar{x}) \geq M_o = h(\bar{x}_o,r_k)$ or $f_o + \frac{r_k}{g_i(\bar{x})} \geq M_o = h(\bar{x}_o,r_k)$. For the former, $h(\bar{x}_o,r_k) \geq h(\bar{x}(r_k),r_k)$ which gives $h(\bar{x},r_k) \geq h(\bar{x}(r_k),r_k)$ and the latter likewise gives $h(\bar{x},r_k) \geq h(\bar{x}(r_k),r_k)$ for any $\bar{x} \in R^o$. ∎

Theorem 11. If the closure of R^o is R, then $\lim_{k\to\infty} h(\bar{x}(r_k),r_k) = f_o$.

Proof. Let $\varepsilon > 0$ be any positive number. There is a $\bar{y} \in R$ such that $f(\bar{y}) < f_o + \frac{\varepsilon}{2}$; otherwise, the inequality $f(\bar{x}) \geq f_o + \frac{\varepsilon}{2}$ would hold for all $\bar{x} \in R$ implying that $f_o + \frac{\varepsilon}{2}$ is the minimum value of f for $\bar{x} \in R$ instead of f_o. For such a $\bar{y} \in R$ by hypothesis and the fact that f is continuous in R, there is a neighborhood of $\bar{y}$ containing a point $\bar{x}^+ \in R^o$ such that $f(\bar{x}^+) < f_o + \frac{\varepsilon}{2}$. Select k^+ such that $r_k + \frac{\varepsilon}{2m} \min_i\{g_i(\bar{x}^+)\}$. Then for $k > k^+$,

$$f_o \leq \inf_{\bar{x}\in R^o} h(\bar{x},r_k) = h(\bar{x}(r_k),r_k) \leq h(x(\bar{r}_k+),r_k) < h(x(\bar{r}_k+),r_k+)$$

$$\leq h(\bar{x}^+,r_k+) < f_o + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} < f_o + \varepsilon$$

which proves the theorem. ∎

Theorem 12. Every subsequence of $\{\bar{x}(r_k)\}$ has a subsequence that converges to some point $\bar{x}^*$ which is such that $f(\bar{x}^*) = f_o$.

Proof. Let

$$K = h(\bar{x}(r_1),r_1) = f(\bar{x}(r_1)) + r_1 \sum_{j=1}^{m} \frac{1}{g_j(\bar{x}(r_1))} .$$

Let $A = \{\bar{x} : f(\bar{x}) \leq K, \bar{x} \in R\}$. Now $h(\bar{x}(r_k),r_k) \leq h(\bar{x}(r_{k-1}),r_k) \leq h(\bar{x}(r_{k-1}),r_{k-1})$ for $k > 2$. Therefore $h(\bar{x}(r_k),r_k) \leq h(\bar{x}(r_1),r_1)$ and

$$f(\bar{x}(r_k)) \leq f(\bar{x}(r_k)) + r_k \sum_{j=1}^{m} \frac{1}{g_j(\bar{x}(r_k))} \leq f(\bar{x}(r_1)) + r_1 \sum_{j=1}^{m} \frac{1}{g_j(\bar{x}(r_1))}$$

Thus $\bar{x}(r_k) \in A$.

Let $\{\bar{x}_k\}$ be a convergent subsequence of $\{\bar{x}(r_k)\}$ with limit $\bar{x}^*$. Suppose $f(\bar{x}^*) \neq f_o$, and let $\delta = \dfrac{f(\bar{x}^*) - f_o}{2}$. Since f is continuous at $\bar{x}^*$, $\lim_{k\to\infty} f(\bar{x}_k) = f(\bar{x}^*)$. Let k_1 be such that $|f(\bar{x}_k) - f(\bar{x}^*)| < \delta$ for all $k \geq k_1$. From Theorem 11 there is k_2 such that $|h(\bar{x}_k,r_k) - f_o| < \delta$ for all $k \geq k_2$. Choose $k \geq \max(k_1,k_2)$. Then

$$f(\bar{x}_k) + r_k \sum_{j=1}^{m} \frac{1}{g_j(\bar{x}_k)} - f_o < \delta$$

and

$$|f_o - f(\bar{x}^*)| \leq |f_o - f(\bar{x}_k)| + |f(\bar{x}_k) - f(\bar{x}^*)|,$$

$$2\,\delta < |f_o - f(x_k)| + \delta,$$

$$\delta < f(\bar{x}_k) - f_o.$$

Then

$$\delta < f(\bar{x}_k) - f_o < f(\bar{x}_k) - f_o + r_k \sum_{j=1}^{m} \frac{1}{g_j(\bar{x}_k)} < \delta,$$

a contradiction. Thus $f(\bar{x}^*) = f_o$. ∎

Corollary 12.1. If there is only one point $\bar{x}^* \in R$ such that $f(\bar{x}^*) = f_o$, then $\lim_{k\to\infty} \bar{x}(r_k) = \bar{x}^*$.

Proof. Suppose $\lim_{k\to\infty} \bar{x}(r_k) \neq \bar{x}^*$. By Theorem 12 $\{\bar{x}(r_k)\}$ has a subsequence $\{\bar{x}_k\}$ that converges to a point $\bar{x}^*$ such that $f(\bar{x}^*) = f_o$. Consider an open neighborhood about $\bar{x}^*$ such that there is an infinite number of points of the sequence $\{\bar{x}(r_k)\}$ outside the neighborhood. This is possible if $\lim_{k\to\infty} \bar{x}(r_k) \neq \bar{x}^*$. Let $\{\bar{y}_k\}$ be a convergent subsequence of these points converging to a point $\bar{y}$. Now $\bar{y} \neq \bar{x}^*$ from the construction of $\{\bar{y}_k\}$, but by Theorem 12 the sequence $\{\bar{y}_k\}$ converges to a point such that

$f(\bar{y}) = f_o$. Thus there are two distinct points, $\bar{y}$ and $\bar{x}^*$ such that $f(\bar{y}) = f(\bar{x}^*) = f_o$, but by hypothesis there is only one--a contradiction. Thus $\lim_{k\to\infty} \bar{x}(r_k) = \bar{x}^*$. ∎

Corollary 12.2. If $\{\bar{x}(r_k)\}$ is the sequence obtained from the algorithm on page 78, then

$$\lim_{k\to\infty} f(\bar{x}(r_k)) = f_o,$$

$$\lim_{k\to\infty} r_k \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_k))} = 0.$$

Proof. Since f is a continuous function on R, then by Corollary 12.1,

$$\lim_{k\to\infty} f(\bar{x}(r_k)) = f(\lim_{k\to\infty} \bar{x}(r_k)) = f(\bar{x}^*) = f_o.$$

By Theorem 11 and the previous sentence,

$$\lim_{k\to\infty} r_k \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_k))} = \lim_{k\to\infty} [h(\bar{x}(r_k),r_k) - f(\bar{x}(r_k))]$$

$$= \lim_{k\to\infty} h(\bar{x}(r_k),r_k) - \lim_{k\to\infty} (\bar{x}(r_k)) = f_o - f_o = 0.$$

Therefore $\lim_{k\to\infty} r_k \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_k))}$ exists and equals zero. ∎

Theorem 13.

$$\sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_k))} \leq \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_{k+1}))} \tag{63}$$

and

$$f(\bar{x}(r_k)) \geq f(\bar{x}(r_{k+1})). \tag{64}$$

Proof. The following inequalities are true.

$$f(\bar{x}(r_{k+1})) + r_{k+1} \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_{k+1}))} \leq f(\bar{x}(r_k)) + r_{k+1} \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_k))}$$

and

$$f(\bar{x}(r_k)) + r_k \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_k))} \leq f(\bar{x}(r_{k+1})) + r_k \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_{k+1}))}.$$

When the above inequalities are combined, then

$$(r_{k+1} - r_k) \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_{k+1}))} \leq (r_{k+1} - r_k) \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_k))}$$

or

$$\sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_{k+1}))} \geq \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_k))}$$

which is (63). By the first inequality in the proof and (63)

$$f(\bar{x}(r_{k+1})) + r_{k+1} \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_{k+1}))} \leq f(\bar{x}(r_k)) + r_{k+1} \sum_{i=1}^{m} \frac{1}{g_i(\bar{x}(r_{k+1}))},$$

from which (64) is obvious. ∎

CHAPTER V

CONCLUSION

In the preceding parts of this paper, some of the more significant iterative methods of minimizing a function of n variables and ways of dealing with constraints have been discussed. A comparison of the various methods to determine which one is "best" under any criterion is difficult since there will always be some particular function for which a given method is best suited. However, there are certain fundamental characteristics of the methods which affect their performance.

Many of the methods considered have quadratic convergence, meaning that the minimum value of a positive definite quadratic function is found, apart from rounding errors, in n iterations. A logical extension of this result is that any method with quadratic convergence would find the minimum value of any function in fewer iterations than one without it because, near the minimum, the second order terms of a Taylor's series expansion of the function dominate and the only methods which will converge quickly for a general function are those which will guarantee to find the minimum of a general quadratic speedily. This is due to the fact that the "curvature" of the function (as measured by the Hessian matrix of second order partial derivatives) is relatively stable near the minimum. (Note that Davidon's method gives a good estimate of this matrix that can be used for this purpose.) The superiority of methods with quadratic convergence is upheld in the studies of Box [5],

Fletcher and Powell [15], and Fletcher and Reeves [16]. The type of function and the behavior of this function near the minimum still has an effect on the number of necessary iterations; the "nearer" the function is to a quadratic in a neighborhood of the minimum, the more effective a method with quadratic convergence will be and the fewer the number of iterations.

However, the above statements should be qualified to some extent due to the influences of various other factors. The starting point, its nearness to the minimum and the criterion for convergence will have an effect on the number of iterations. In choosing a method to minimize a function, the number and ease of computations, the number of function evaluations per iteration and whether the gradient of the function is available should be considered. If the gradient is not available, then Powell's method or Zangwill's method might be preferred; the latter has quadratic convergence. In a comparison of several nongradient methods, Fletcher [14] states that Powell's method is certainly the most efficient on the basis of the number of function evaluations and that it has rapid convergence near the minimum. (Zangwill's method was not included in this comparison.) When the gradient is available, Davidon's method usually gives faster convergence even though other gradient methods have quadratic convergence. Fletcher and Powell [15] state that Davidon's method is probably the most powerful general procedure available for finding a local minimum and Box [5] states that it was the most consistently successful in his comparison of several procedures. If the gradient cannot be determined analytically, then a finite difference approximation can be made for the partial derivative such as

$$\frac{\partial f}{\partial x_i} = \frac{f(x_1,\ldots,x_i + h,x_{i+1},\ldots,x_n) - f(\bar{x})}{h}, \quad i = 1,2,\ldots,n,$$

where h is a small number, or

$$\frac{\partial f}{\partial x_i} = \frac{f(x_1,\ldots,x_i + h,\ldots,x_n) - f(x_1,\ldots,x_i - h,\ldots,x_n)}{2h}.$$

Methods for dealing with constraints have been discussed. Other methods are known, but most of them are restricted to using a specific minimization method or work only for constraints of a certain form. Box [5] states that for a general constrained minimization problem, Davidon's method combined with Carroll's created response surface technique has been successfully used. Fiacco and McCormick [13] state that Carroll's method has worked orderly most of the time they have used it.

There are other types of methods of minimizing a function and other ways of dealing with constraints as well as methods which combine the two problems which have not been discussed. The reader is referred to an extensive bibliography in reference [20] and to the Journal of Industrial Engineering [21] which gives a flow chart indicating ways to decide the approach to use in optimizing a function subject to constraints.

BIBLIOGRAPHY

1. T. M. Apostol, *Mathematical Analysis, A Modern Approach to Advanced Calculus,* Addison-Wesley, Inc., Reading, Mass. (1957).

2. K. J. Arrow and L. Hurwicz, *Reduction of Constrained Maximum to Saddle Point Problems,* Proc. Berkeley Symposium on Mathematical Statistics and Probability, Vol. 5, Univ. of California Press, Berkeley (1956).

3. F. S. Beckman, *The Solution of Linear Equations by the Conjugate Gradient Method,* Mathematical Methods for Digital Computers, Ralston, A., and Wilf, H. S. (Eds.) Wiley (1960).

4. H. A. Boas, *Modern Mathematical Tools for Optimization,* Chemical Engineering, Dec., 1962, Jan., Feb., Mar., Apr., 1963.

5. M. J. Box, *A Comparison of Several Current Optimization Methods and the Use of Transformations in Constrained Problems.* The Computer Journal, Vol. 9, pp. 67-77 (1966).

6. C. W. Carroll, *The Created Response Surface Technique for Optimizing Nonlinear, Restrained Systems,* Operations Research Society of America Journal, Vol. 9, pp. 169-184 (1961).

7. A. L. Cauchy, *Méthode Générale Pour la Résolution des Systémes d'Équations Simultanées,* C. R. Read Sci., Paris, Vol. 25, pp. 536-538 (1847).

8. H. B. Curry, *The Method of Steepest Descent for Nonlinear Minimization Problems,* Quarterly of Applied Mathematics, Vol. 2, pp. 250-261 (1944).

9. W. C. Davidon, *Variable Metric Method for Minimization,* Argonne Nat'l Lab. (ANL-5990 Rev.), Univ. of Chicago, 21 pp. (1959).

10. D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra,* W. H. Freeman and Co., San Francisco, Calif. (1963).

11. A. V. Fiacco, *Comments on the Paper by C. W. Carroll,* J. Oper. Res. Soc. Am., Vol. 9, pp. 184-185 (1962).

12. A. V. Fiacco and G. P. McCormick, *The Sequential Unconstrained Minimization Technique for Nonlinear Programming a Primal-Dual Method,* Management Science, Vol. 10, pp. 360-366 (1964).

13. A. V. Fiacco and G. P. McCormick, *Computational Algorithm for the Sequential Unconstrained Minimization Technique for Nonlinear Programming,* Manage. Sci., Vol. 10, pp. 601-617 (1964).

14. R. Fletcher, *Function Minimization Without Evaluating Derivatives--A Review,* The Computer Journal, Vol. 8, pp. 33-41 (1965).

15. R. Fletcher and M. J. D. Powell, *A Rapidly Convergent Descent Method for Minimization,* The Computer Journal, Vol. 6, pp. 163-168 (1963).

16. R. Fletcher and C. M. Reeves, *Function Minimization by Conjugate Gradients,* The Computer Journal, Vol. 7, pp. 149-154 (1964).

17. M. R. Hestenes and E. Stiefel, *Methods of Conjugate Gradients for Solving Linear Systems,* Journal of Research, National Bureau of Standards, Vol. 49, pp. 409-436 (1952).

18. R. Hooke and T. A. Jeeves, *'Direct Search' Solution of Numerical and Statistical Problems,* Association for Computing Machinery Journal, pp. 212-229 (1961).

19. J. Kiefer, *Optimum Sequential Search and Approximation Methods under Minimum Regularity Assumptions,* S.I.A.M. Journal, Vol. 5, pp. 105-136 (1957).

20. A. Lavi and T. Vogl, (Eds.), Proc. Symp. on *Recent Advances in Optimization Techniques,* John Wiley (1965).

21. J. E. Mulligan, *Basic Optimization Techniques--A Brief Survey,* The Journal of Industrial Engineering, Vol. 16, No. 3, pp. 192-197 (1965).

22. M. J. D. Powell, *An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives,* The Computer Journal, Vol. 7, pp. 155-162 (1964).

23. H. H. Rosenbrock, *An Automatic Method for Finding the Greatest or Least Value of a Function,* The Computer Journal, Vol. 3, pp. 175-184 (1960).

24. B. V. Shah, R. J. Buehler, and O. Kempthorne, *Some Algorithms for Minimizing a Function of Several Variables,* S.I.A.M. Journal, Vol. 12, pp. 74-92 (1964).

25. C. S. Smith, *The Automatic Computation of Maximum Likelihood Estimates,* N. C. B. Scientific Dept. Report S. C. 846/MR/40 (1960).

26. H. A. Spang, *A Review of Minimization Techniques for Nonlinear Functions*, S.I.A.M. Review, Vol. 4, pp. 343-365 (1962).

27. A. W. Tucker, *Linear and Nonlinear Programming*, J. Oper. Res. Soc. Am., Vol. 5, pp. 245-257 (1957).

28. D. J. Wilde, *Optimum Seeking Methods*, Prentice-Hall, Inc., Englewood Cliffs, N. J. (1964).

29. W. I. Zangwill, *Minimizing a Function Without Calculating Derivatives*, The Computer Journal, Vol. 10, pp. 293-296 (1967).