Pacific Journal of Mathematics

GRADIENT METHODS OF MAXIMIZATION

JEAN BRONFENBRENNER CROCKETT AND HERMAN CHERNOFF

Vol. 5, No. 1 September 1955

GRADIENT METHODS OF MAXIMIZATION

JEAN BRONFENBRENNER CROCKETT AND HERMAN CHERNOFF

1. Introduction. We shall consider the computational problem of finding a point

$$c = (c_1, c_2, \cdots, c_n)$$

for which a given function of n variables

$$f(x) = f(x_1, x_2, \cdots, x_n)$$

attains its maximum. Frequently, the form of the function and the number of independent variables involved make it prohibitively difficult to determine the point c by direct methods, and methods of successive approximations are accordingly used.

Suppose that

$$x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)})$$

is an approximation to the point c. Assuming that $x^{(0)}$ is sufficiently close to c, we may obtain an improved approximation to c by considering the first few terms of the Taylor expansion of $f(x)$ about $x^{(0)}$. Presumably, the greater the number of terms of the expansion that we consider, the better will be our improved approximation and the more rapidly will the corresponding iterative procedure converge. On the other hand, increasing the number of terms of the expansion involves the calculation of higher order derivatives and increases considerably the computational cost of each iteration.

The methods of successive approximations to be discussed in this paper are (1) gradient methods using the first order derivatives only, and (2) the Newton method which uses first and second order derivatives. *In both cases, it is possible to obtain, from the successive approximations, certain relevant information about terms of order higher than those actually computed, and to conveniently use this information to improve the rate of convergence.*

Received July 17, 1953. This paper was prepared with the partial support of the Office of Naval Research. This paper is the completion of research carried out while the authors were with the Cowles Commission for Research in Economics.

Pacific J. Math., 5 (1955), 33-50

2. Assumptions and notation. We shall assume that $f(x)$ and its derivatives are sufficiently "well behaved" to permit the use of Taylor's expansions of sufficiently high order. (For most arguments the use of third or fourth order expansions will suffice.) We also assume that $f(x)$ has an isolated maximum at $x = c$, and that

$$L = -\left\| \frac{\partial^2 f(c)}{\partial x_i \, \partial x_j} \right\| \qquad (i, j = 1, 2, \cdots, n)$$

is positive definite. In the proof of convergence we shall assume that our initial approximation $x^{(0)}$ is in a sufficiently small neighborhood of c.

A vector x may be considered as a matrix consisting of one column. The transpose x' of x then consists of one row. Thus $x'y$ would represent the scalar product $\sum_{i=1}^{n} x_i y_i$ of x and y.

The large and small o notation will be occasionally used. The equation $y = o(x)$ will indicate that as $x \longrightarrow 0$, $|y|/|x| \longrightarrow 0$. The equation $y = O(x)$ will indicate that there is a constant k such that $|y| \leq k|x|$. This notation extends in an obvious fashion to vectors. For the benefit of those readers not accustomed to this notation the equations will be written so that the terms involving the small and large o are small compared to the remaining terms; that is, *the equations are approximately correct if the terms involving the o's are neglected.* The expression $x \approx y$ should be read "x and y are approximately equal."

3. Gradient methods. In this section the gradient method is introduced. In order to determine a convenient computational procedure we shall study, in particular, the rate of convergence and the behavior of the successive iterations when this method is used.

For a given initial approximation $x^{(0)}$ to c it is natural to select the next approximation $x^{(1)}$ in such a way that the step from $x^{(0)}$ to $x^{(1)}$ is in the direction of "steepest ascent" or gradient. The direction of steepest ascent depends, however, on the way in which one measures the distance between two points x and y in n-dimensional space. In general there is no reason to assume that a unit of distance along the x_1 axis is equivalent to a unit of distance along the x_2 axis. The definition of distance (that is, metric) to be used implies a particular system of weighting these units.

Let us suppose that the distance d from x to y is defined by

$$(1) \qquad d = \left[\sum_{i,j=1}^{n} b_{ij}(y_i - x_i)(y_j - x_j) \right]^{1/2},$$

where $B = ||b_{ij}||$ is a positive definite symmetric matrix. The locus of points at distance k from $x^{(0)}$ is given by the ellipsoid

$$\sum_{i,j=1}^{n} b_{ij}(x_i - x_i^{(0)})(x_j - x_j^{(0)}) = k^2$$

with center $x^{(0)}$. *The direction of steepest ascent in the k neighborhood of* $x^{(0)}$ may be defined as the direction from $x^{(0)}$ to that point of the above ellipsoid for which the value of the function f is greatest. In Appendix 1, it is shown that, as $k \longrightarrow 0$, this direction approaches a limit which is the direction of the vector

$$(2) \qquad \delta(x^{(0)}) = B^{-1} l(x^{(0)}),$$

where $l(x)$ represents the column vector whose ith component is $\partial f/\partial x_i$, and $l(x^{(0)})$ is assumed to be different from zero. Hereafter we shall call

$$\delta(x) = B^{-1} l(x)$$

the *gradient vector* (*at x relative to B*). One may ask what advantages one metric has over another. This question will be treated in § 5.

One would naturally expect that $f(x)$ *increases as x moves from* $x^{(0)}$ *in the direction of the gradient.* Indeed, the proof of the following theorem is left to the reader.

THEOREM 1. *For positive h small enough,*

$$f(x^{(0)} + h\delta(x^{(0)})) > f(x^{(0)})$$

if $l(x^{(0)}) \neq 0$.

The problem now arises as to how large a step may profitably be taken in the direction of the gradient. If h is taken too small, $x^{(0)} + h\delta(x^{(0)})$ will not be much closer to c than $x^{(0)}$. If h is taken too large, $x^{(0)} + h\delta(x^{(0)})$ may overshoot c and even lead to diminishing the value of f. Clearly an optimal procedure should depend at least in part on how fast the slope of $f(x)$ changes as x moves from $x^{(0)}$ in the direction of the gradient, and thus cannot be determined

without some consideration of second order derivatives. To help solve the problems of selecting B and h we shall now study the rate of convergence and the behavior of the successive approximations when the gradient method is used.

Let $x^{(0)}, x^{(1)}, \cdots, x^{(m)}, \cdots$ be a sequence of approximations formed so that

$$x^{(m+1)} = x^{(m)} + h_m \delta(x^{(m)}) = x^{(m)} + h_m B^{-1} l(x^{(m)}), \tag{3}$$

and let

$$e^{(m)} = x^{(m)} - c \tag{4}$$

represent the error of the mth approximation. We abbreviate $\delta(x^{(m)})$ to $\delta^{(m)}$ and set

$$L(x) = -\left\| \frac{\partial^2 f(x)}{\partial x_i \, \partial x_j} \right\| \qquad (L = L(c)). \tag{5}$$

The following theorem is established in Appendix 2.

THEOREM 2.

$$e^{(m+1)} = (I - h_m B^{-1} L) e^{(m)} + o(e^{(m)}), \tag{6}$$

$$\delta^{(m)} = -B^{-1} L e^{(m)} + o(e^{(m)}), \tag{7}$$

$$\delta^{(m+1)} = (I - h_m B^{-1} L) \delta^{(m)} + o(e^{(m)}). \tag{8}$$

The results in Theorem 2 may be written in another form. A classical result of matrix theory [1] tells us that since B^{-1} and L are positive definite symmetric matrices, $B^{-1}L$ has n positive characteristic values which we may label λ_i in order of magnitude; that is,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0.$$

There is also a linearly independent set $\mu_1, \mu_2, \cdots, \mu_n$ of corresponding characteristic vectors satisfying

$$B^{-1} L \mu_i = \lambda_i \mu_i. \tag{9}$$

(It is important to realize that the λ_i depend on L which is unknown.) Then any

n-dimensional vector may be uniquely represented as a linear combination of the μ_i, and we may write

$$(10)\qquad \begin{cases} e^{(m)} = k_1^{(m)}\mu_1 + k_2^{(m)}\mu_2 + \cdots + k_n^{(m)}\mu_n, \\ \delta^{(m)} = \bar{k}_1^{(m)}\mu_1 + \bar{k}_2^{(m)}\mu_2 + \cdots + \bar{k}_n^{(m)}\mu_n. \end{cases}$$

Thus Theorem 2 gives us

$$(11)\qquad k_i^{(m+1)} = (1 - h_m\lambda_i)k_i^{(m)} + o(e^{(m)}),$$

$$(12)\qquad \bar{k}_i^{(m)} = -\lambda_i k_i^{(m)} + o(e^{(m)}),$$

$$(13)\qquad \bar{k}_i^{(m+1)} = (1 - h_m\lambda_i)\bar{k}_i^{(m)} + o(e^{(m)}) \qquad (i = 1, 2, \cdots, n).$$

Theorem 2 suggests conjectures of the following sort. If the h_m are selected so that

$$-1 < 1 - h_m\lambda_1 \leq 1 - h_m\lambda_2 \leq \cdots \leq 1 - h_m\lambda_n < 1,$$

then $k_i^{(m)} \longrightarrow 0$ and $x^{(m)} \longrightarrow c$. If in addition the h_m are selected so that

$$0 < h_m < \frac{2}{\lambda_1 + \lambda_n},$$

then

$$1 > |1 - h_m\lambda_n| \geq |1 - h_m\lambda_i| \text{ for all } i,$$

and of all the coefficients $k_i^{(m)}$ $(i = 1, 2, \cdots, n)$ the coefficient $k_n^{(m)}$ of μ_n is diminished least rapidly. After a large number of iterations the coefficient of μ_n becomes dominant and we have

$$\delta^{(m)} \approx \bar{k}_n^{(m)}\mu_n$$

and

$$\delta^{(m+1)} \approx (1 - h_m\lambda_n)\delta^{(m)}.$$

The precise results are proved in Appendix 3 and stated in Theorems 3 and 4.

Since the proofs are involved and not of especial interest, the authors suggest that readers with an elementary background in mathematics leave this appendix for last.

THEOREM 3. *For any* $\epsilon > 0$, *there is a neighborhood of* c *such that if* $x^{(0)}$ *lies in this neighborhood, and*

$$|1 - h_m \lambda_i| \leq 1 - \epsilon$$

for all m *and* i, *then* $\lim_{m \to \infty} x^{(m)} = c$.

Let

$$(14) \quad \begin{cases} R(m) = (k_1^{(m)2} + k_2^{(m)2} + \cdots + k_r^{(m)2})/(k_1^{(m)2} + k_2^{(m)2} + \cdots + k_n^{(m)2}), \\ \bar{R}(m) = (\bar{k}_1^{(m)2} + \bar{k}_2^{(m)2} + \cdots + \bar{k}_r^{(m)2})/(\bar{k}_1^{(m)2} + \bar{k}_2^{(m)2} + \cdots + \bar{k}_n^{(m)2}), \end{cases}$$

where r is defined by

$$(15) \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_n.$$

$R(m)$ and $\bar{R}(m)$ measure those parts of $e^{(m)}$ and $\delta^{(m)}$ which correspond to characteristic values which are greater than λ_n.

THEOREM 4. *For any* $\epsilon > 0$, *and* $\eta > 0$, *there is a neighborhood of* c *such that if* $x^{(0)}$ *lies in this neighborhood,* $R(0) < 1 - \eta$, *and*

$$\frac{2 - \epsilon}{\lambda_1 + \lambda_n} \geq h_m \geq \frac{\epsilon}{\lambda_n}$$

for all m, *then*

$$(16) \quad \begin{cases} \lim_{n \to \infty} R(m) = 0, \\ \lim_{m \to \infty} \bar{R}(m) = 0. \end{cases}$$

4. Selection of step size, h_m. In this section we shall apply the results in § 3 to investigate the problem of finding a good choice of h_m. We mentioned previously that an intelligent choice of h_m should depend on some information regarding higher order terms in the Taylor expansion of $f(x)$. From (11) it seems apparent that the relevant information would be the λ_i. For a good choice

of h's would be one where the h_m take various values close to the various $1/\lambda_i$, thus diminishing

$$k_i^{(m+1)} = (1 - h_m \lambda_i) k_i^{(m)}$$

very rapidly. It should be noted that if $1/\lambda_i$ is large compared to $1/\lambda_i$ the use of $h_m = 1/\lambda_i$ will tend to magnify the coefficient of μ_1. Knowledge of λ_1 and λ_n would be especially valuable since good choices of h should be between $1/\lambda_1$ and $1/\lambda_n$.

Theorem 4 furnishes us with a procedure for approximating λ_n. For if a *small* value of h, that is,

$$h < \frac{2}{\lambda_1 + \lambda_n},$$

is used for many approximations, $\delta^{(m+1)}$ *will be approximately proportional to* $\delta^{(m)}$ *and the ratio will be* $1 - h_m \lambda_n$. Since h_m, $\delta^{(m)}$, and $\delta^{(m+1)}$ are known or computed, λ_n may be approximated. This method would seem to be inapplicable without some knowledge of what constitutes a small value of h. Fortunately, Theorem 2 may be invoked to tell us that if h_m is very small, then $\delta^{(m+1)} \approx \delta^{(m)}$. If h_m is very large, $\delta^{(m+1)}$ will tend to have the direction opposite to that of $\delta^{(m)}$, and $\delta^{(m+1)}$ may possibly be of larger magnitude than $\delta^{(m)}$. Hence a very large or small value of h_m will be revealed by the results obtained from using this value.

The value of λ_1 may be estimated by a similar approach. However, the estimate of λ_1 so obtained is very sensitive to higher order terms, and it is very difficult to obtain a good estimate with a reasonable number of iterations.

In view of these remarks, the following system of choosing the h's seems to have merit. Begin with a small value of h and gradually increase h. Note that h should increase slowly if the $\delta^{(m+1)}$ are in very different directions from the $\delta^{(m)}$. Note also that it is not efficient to repeat values of h. Continue until $\delta^{(m+1)} \approx \rho \delta^{(m)}$, where ρ is not close to one. Then use

$$h_{m+1} = \frac{h_m}{1 - \rho},$$

the reciprocal of the estimated λ_n. Since this h_{m+1} may be so large as to revitalize components corresponding to large λ_i, start a new round by applying a small value of h and repeating the above procedure.

One will frequently find that the last h of the round tends to increase from round to round. For an example see [6, p. 284 or p. 294]. This "effect" may be explained by the following type of argument. Suppose that initially $k_n^{(0)}$ is very small compared to $k_{n-1}^{(0)}$. After several iterations, $k_{n-1}^{(m)}$ may become large compared to $k_{n-2}^{(m)}$, $k_{n-3}^{(m)}, \cdots$, and so on, and $k_n^{(m)}$ may still be small compared to $k_{n-1}^{(m)}$. In that event we will have $\delta^{(m+1)} \approx \rho\, \delta^{(m)}$, where $\rho \approx 1 - h_m \lambda_{n-1}$. The use of

$$h_{m+1} = \frac{h_m}{1-\rho} \approx \frac{1}{\lambda_{n-1}}$$

will then tend to make $k_{n-1}^{(m+1)}$ small. In the next round the k_n may come to dominate all other k's, and the last h of this round will be close to $1/\lambda_n > 1/\lambda_{n-1}$.

5. Selection of B. Newton method and modification. In this section we shall consider a measure of the efficiency of a metric B. It will be evident that the Newton method may be considered as a most efficient gradient method. The high computational cost of computing second order derivatives leads one occasionally to make use of a simple modification of the Newton method. These concepts are of special interest in a large class of statistical problems.

To construct a measure of the efficiency of a metric B, let us suppose that for some reason or other it is desirable to use a constant value h of h_m. In this case, the rate of convergence obviously depends on

$$M = \max_{1 \le i \le n} |1 - h\lambda_i|. \tag{17}$$

Thus the best value of h would be that for which M is the smallest possible. Then

$$h = \frac{2}{\lambda_1 + \lambda_n}, \tag{18}$$

$$M = \frac{1 - (\lambda_n/\lambda_1)}{1 + (\lambda_n/\lambda_1)}. \tag{19}$$

From this point of view, λ_n/λ_1 can be considered as an indication of the convergence rate per iteration when the metric B is used. The closer this ratio is to one the more rapid the rate of convergence. (Note that n, the number of

components of x, is relatively unimportant in measuring the convergence rate per iteration although it influences greatly the amount of work per iteration.)

If $B^{-1}L$ is close to the identity matrix then the characteristic values λ_i are all close to unity (and λ_n/λ_1 is close to one). Theorem 2 tells us that in this case a value of h_m close to one will make $e^{(m+1)}$ very small.

With this in mind, we see that it would appear to be desirable to set $B = L(x^{(m)})$, since in the neighborhood of c we may expect $L^{-1}(x^{(m)})L$ to be close to the identity matrix. Then we have

$$x^{(m+1)} = x^{(m)} + h_m L^{-1}(x^{(m)})\, l(x^{(m)}) \tag{20}$$

which is Newton's method if $h_m = 1$ (see Appendix 4).

It should be noted that the speed of convergence per iteration of Newton's method increases as $x^{(m)}$ gets closer to c. This is obvious since the characteristic values λ_i are pushed close to one as $L^{-1}(x^{(m)})L$ approaches the identity matrix. In fact it can easily be shown that

$$e^{(m+1)} = O(|e^{(m)}|^2). \tag{21}$$

This property makes the Newton method especially valuable when $x^{(m)}$ is very close to c. The Newton method seems to differ from the gradient method not only in regard to the rate of convergence but also in that B does not remain fixed. However, the results obtained for the gradient methods can easily be extended to those cases where B varies from iteration to iteration, the variations being subject to certain mild restrictions.

The great speed of convergence of the Newton method is offset by the cost of computing second order derivatives, which is often extremely high. To lessen this cost while still retaining some of the advantages of the Newton method, the following modification has often been used. Instead of using $B = L(x^{(m)})$ for the mth iteration, we may use $B = L(x^{(r)})$ for the $r, r+1, \cdots, r+k$ iterations, thus avoiding the calculation of $L(x^{(r+1)}), \cdots, L(x^{(r+k)})$. However, the tendency of the λ_i to get closer to one, thereby accelerating the convergence, will not be present during these interludes when the metric is not changed.

The notions used in this section are of special relevance to statisticians. To employ their language, in problems of maximum-likelihood estimation where $f(x)$ is the logarithm of the likelihood function, $-L^{-1}(c)$ represents an estimate of Σ, the covariance matrix of the maximum-likelihood estimate c. Frequently, the statistician can use this fact to find an easily computed matrix

$R(x)$ which is close to $L(x)$ for reasonably large samples. For example, the information matrix may be applied for this purpose. This idea has been used in probit analysis [2].

6. The adjustment of h_m in the Newton method. Much of our discussion of gradient methods centered about the size of the step to be taken in the direction of the gradient. The Newton method in its ordinary form implies the use of $h_m = 1$. So long as $h_m = 1$ is used, it will be found that the iterations tend to undershoot the mark in a systematic fashion depending on third order derivatives. *By observing successive iterations we may correct for this systematic tendency without actually computing the third order derivatives.* The following procedure is applicable to the modified Newton method.

Using $B = L(x^{(r)})$ for the $r, r+1, \cdots, m, \cdots$ iterations and assuming that $e^{(m)}$ is small compared to $e^{(r)}$ for $m > r$, we show in Appendix 5 that

$$(22)\qquad \begin{cases} e^{(r+1)} \approx \left[I - h_r\left(I - \dfrac{U}{2}\right)\right] e^{(r)}, \\ e^{(m+1)} \approx [I - h_m(I - U)]\, e^{(m)} \qquad (m = r+1, r+2, \cdots), \end{cases}$$

where U is a matrix whose elements are homogeneous linear functions of the elements of $e^{(r)}$, indeed $U = O(e^{(r)})$. The one-dimensional case is fundamentally simpler than the n-dimensional case, and additional results have been established for this case in Appendix 6.

In equation 22, the terms multiplied by h replace the $B^{-1}L$ of the standard gradient method (see (6)). *Hence the role of the characteristic values of $B^{-1}L$ is here replaced by those of $I - U/2$ and $I - U$ for $m = r$ and $m > r$, respectively.* Since $U = O(e^{(r)})$, the λ_i will be close to 1. The treatment of the modified Newton method differs from that of the standard gradient method in that immediately after the rth iteration, the λ_i tend to double their distance from one. This tendency may be taken into account as follows. If $h_r = 1$ is used, and

$$\delta^{(r+1)} \approx 0.1\,\delta^{(r)},$$

the standard gradient method would suggest

$$h_{r+1} = \frac{1}{1 - .1} = \frac{10}{9}.$$

But the spread of the λ_i imply that the h should correspondingly spread, and $h_{r+1} = 11/9$ would be preferable.

7. Bibliographical remarks. The gradient method with Euclidean metric was suggested by Cauchy [3]. This method and extensions to functions on more general spaces was treated by Curry [4], who mentioned that the method was not invariant when changes in scale are made on the variables. (Such changes in scale correspond to a change in the metric.)

In 1946, gradient methods with non-Euclidean metrics were found appropriate and applied by Koopmans, Rubin, and Leipnik [5, pp. 153-238] to several problems at the Cowles Commission for Research in Economics. The authors found bounds on the λ_i and used the fact that a fixed value of $h < 2/\lambda_1$ leads to convergence. Apparently they were not aware of the fact that these methods were gradient methods, nor of the possibility of using the successive iterations to accelerate convergence. They experimented with a variation where, in each iteration, h was selected so as to maximize the quadratic approximation to the function along the gradient vector. This variation, lately called the optimum gradient method, did not work particularly well. (This method lacks optimality because it ignores the relevance of the λ_i.) In 1948 and 1949, Bronfenbrenner and Chernoff developed and applied the results in the present paper to the problems of the Cowles Commission. Some typical computations were presented by Chernoff and Divinsky [6, pp. 236-302].

In 1939, Temple [7] applied the optimum gradient method with Euclidean metric to maximizing quadratic functions (or equivalently to solving linear equations). He also extended this approach from n-dimensional space to Hilbert space.

An extensive bibliography on the extensions and developments of this approach for solving linear problems is contained in a paper by Forsythe [8]. In particular, there is a method due to Forsythe and Motzkin of accelerating the optimum gradient method by using the results of previous iterations. Also discussed is the conjugate-gradient method, an important variation in the case of linear problems. This method is due to Hestenes, Lanczos, and Stiefel.

Appendix 1. *We apply the method of Lagrange multipliers* to maximize $f(x^{(0)} + \delta)$, subject to the restriction

$$\sum_{i,j} b_{ij}\delta_i\delta_j = \delta' B\delta = k^2.$$

We obtain

$$\frac{\partial f(x^{(0)} + \delta)}{\partial x_i} = 2\lambda \sum_j b_{ij}\delta_j, \quad l(x^{(0)} + \delta) = 2\lambda B\delta,$$

$$\delta = \frac{1}{2\lambda} B^{-1} l(x^{(0)} + \delta), \quad \delta' B \delta = \frac{1}{4\lambda^2} l(x^{(0)} + \delta)' B^{-1} l(x^{(0)} + \delta) = k^2,$$

$$\delta = \frac{kB^{-1} l(x^{(0)} + \delta)}{\sqrt{l(x^{(0)} + \delta)' B^{-1} l(x^{(0)} + \delta)}}.$$

Letting $k \longrightarrow 0$, we have

$$\delta = O(k), \quad l(x^{(0)} + \delta) = l(x^{(0)}) + O(k).$$

Since

$$l(x^{(0)}) \neq 0, \quad B^{-1} l(x^{(0)}) \neq 0,$$

we get

$$\delta = \frac{k}{\sqrt{l(x^{(0)})' B^{-1} l(x^{(0)}) + O(k)}} [B^{-1} l(x^{(0)}) + O(k)],$$

and the direction of $B^{-1} l(x^{(0)}) + O(k)$ approaches that of $B^{-1} l(x^{(0)})$.

Appendix 2. *Proof of Theorem* 2. Expanding the first derivative of f in a Taylor expansion about c, we obtain

$$l(x^{(m)}) = l(c + e^{(m)}) = l(c) - Le^{(m)} + o(e^{(m)}).$$

Since c is the point at which f attains its maximum, we have $l(c) = 0$ and

$$\delta^{(m)} = B^{-1} l(x^{(m)}) = - B^{-1} L e^{(m)} + o(e^{(m)}),$$

$$e^{(m+1)} = x^{(m+1)} - c = x^{(m)} - c + h_m \delta^{(m)}, = e^{(m)} - h_m B^{-1} L e^{(m)} + o(e^{(m)}),$$

$$= (I - h_m B^{-1} L) e^{(m)} + o(e^{(m)}).$$

Finally,

$$\delta^{(m+1)} = - B^{-1} L e^{(m+1)} + o(e^{(m+1)}) = (I - h_m B^{-1} L)(-B^{-1} L e^{(m)}) + o(e^{(m)})$$

$$= (I - h_m B^{-1} L) \delta^{(m)} + o(e^{(m)}).$$

This completes the proof of Theorem 2.

Appendix 3. *Proof of Theorem* 3. Let

$$k(m) = k_1^{(m)^2} + k_2^{(m)^2} + \cdots + k_n^{(m)^2}.$$

From (11), we have

$$k_i^{(m+1)} = (1 - h_m \lambda_i) k_i^{(m)} + o(\sqrt{k(m)}),$$

$$k(m+1) = \sum (1 - h_m \lambda_i)^2 k_i^{(m)^2} + o(k(m)).$$

Under the conditions of Theorem 3, we obtain

$$k(m+1) \leq (1 - \epsilon)^2 k(m) + o(k(m)).$$

It follows that there is a number a such that, for $k(m) \leq a$, we have

$$k(m+1) \leq \left(1 - \frac{\epsilon}{2}\right)^2 k(m) \leq a.$$

Hence if $k(0) \leq a$, then

$$k(m) \leq \left(1 - \frac{\epsilon}{2}\right)^{2m} k(0),$$

$e^{(m)} \longrightarrow 0$ and $x^{(m)} \longrightarrow c$.

Proof of Theorem 4. Let

$$\sum_{i=1}^{r} (1 - h_m \lambda_i)^2 k_i^{(m)^2} = \beta_m^2 \sum_{i=1}^{r} k_i^{(m)^2}$$

and

$$\sum_{i=r+1}^{n} (1 - h_m \lambda_i)^2 k_i^{(m)^2} = \gamma_m^2 \sum_{i=r+1}^{n} k_i^{(m)^2}.$$

Then

$$\gamma_m = 1 - h_m \lambda_n,\ 1 - \epsilon \geq \gamma_m \geq \frac{1 - \lambda_n/\lambda_1}{1 + \lambda_n/\lambda_1} = M > 0,$$

$$0 \leq \beta_m \leq \max\,(1 - h_m \lambda_r, |1 - h_m \lambda_1|) \leq \max\,(1 - h_m \lambda_r, \gamma_m - \epsilon).$$

It follows that β_m/γ_m is bounded away from 1; that is, there is an $\epsilon^* > 0$ such that

$$\frac{\beta_m}{\gamma_m} \leq 1 - \epsilon^*.$$

Now

$$\sum_{i=1}^{r} k_i^{(m+1)^2} = \beta_m^2 \sum_{i=1}^{r} k_i^{(m)^2} + o(k(m)),$$

$$\sum_{i=r+1}^{n} k_i^{(m+1)^2} = \gamma_m^2 \sum_{i=r+1}^{n} k_i^{(m)^2} + o(k(m)),$$

$$R(m+1) = \frac{\beta_m^2 \sum_{i=1}^{r} k_i^{(m)^2} + o(k(m))}{\beta_m^2 \sum_{i=1}^{r} k_i^{(m)^2} + \gamma_m^2 \sum_{i=r+1}^{n} k_i^{(m)^2} + o(k(m))}.$$

Dividing numerator and denominator by $k(m)$, we get

$$R(m+1) = \frac{\beta_m^2 R(m) + \tau_m}{\beta_m^2 R(m) + \gamma_m^2 (1 - R(m)) + \nu_m},$$

where τ_m and ν_m approach zero as $k(m)$ approaches zero. Thus

$$R(m+1) = \phi_m R(m) + \psi_m,$$

where

$$\phi_m = \frac{\beta_m^2}{\beta_m^2 R(m) + \gamma_m^2 (1 - R(m)) + \nu(m)},$$

$$\psi_m = \frac{\tau_m}{\beta_m^2 R(m) + \gamma_m^2 (1 - R(m)) + \nu(m)}.$$

If

$$R(m) < 1-\eta,\ M^2\eta \le \beta_m^2 R(m) + \gamma_m^2 (1-R(m)) \le (1-\epsilon)^2,$$

then

$$\phi_m = \frac{\beta_m^2}{\beta_m^2 R(m) + \gamma_m^2 (1-R(m))} + O(\nu_m)$$

and

$$\psi_m = O(\tau_m).$$

The expression

$$\frac{\beta_m^2}{\beta_m^2 R(m) + \gamma_m^2 (1-R(m))}$$

is monotonic increasing in $R(m)$ and in β_m/γ_m. Therefore this expression is bounded by

$$\frac{(1-\epsilon^*)^2}{(1-\epsilon^*)^2(1-\eta)+\eta} < 1.$$

It follows that there is a number $b \le a$ (a occurs in the proof of Theorem 3) such that for $k(m) \le b$, there is an $\eta^* > 0$ such that

$$\phi_m \le 1-\eta^*,\ \psi_m \le (1-\eta)\eta^*,$$

and thus

$$R(m+1) \le 1-\eta.$$

Hence, if $R(0) \le 1-\eta$ and $k(0) \le b$, it follows that for all m, we have

$$R(m) \le 1-\eta,\ k(m) \le b,\ \phi_m \le 1-\eta^*,\ \psi_m \le (1-\eta)\eta^*,$$

and thus $k(m) \longrightarrow 0$ and $\psi_m \longrightarrow 0$. Now

$$R(m) = \psi_{m-1} + \psi_{m-2}\phi_{m-1} + \psi_{m-3}\phi_{m-1}\phi_{m-2} + \cdots + (\psi_0 \phi_{m-1} \cdots \phi_1)$$

$$+(\phi_{m-1}\cdots\phi_1)R(0).$$

The right side may be decomposed into the sum of the first s terms and the remaining $m-s$ terms. Each of these sums approaches zero as $s\longrightarrow\infty$ and $m-s\longrightarrow\infty$. Applying equation 12 we readily show that $\overline{R}(m)\longrightarrow 0$.

Appendix 4. *The Newton method* is ordinarily introduced as the method where $x^{(m+1)}$ is obtained by finding that value of x which maximizes

$$f(x^{(m)})+(x-x^{(m)})'l(x^{(m)})-\frac{1}{2}(x-x^{(m)})'L(x^{(m)})(x-x^{(m)}),$$

the second-order Taylor expansion of f about $x^{(m)}$.

Appendix 5. *Error term in the modified Newton method.* We use subscripts with f to denote partial derivatives. Let $e=x-c$. Then

$$f(x)=f(c)+\frac{1}{2}\sum_{i,j}f_{ij}(c)e_ie_j+\frac{1}{6}\sum_{i,j,k}f_{ijk}e_ie_je_k+o(|e|^3),$$

$$f_i(x)=\sum_j f_{ij}(c)e_j+\frac{1}{2}\sum_{j,k}f_{ijk}e_je_k+o(|e|^2),$$

$$f_{ij}(x)=f_{ij}(c)+\sum_k f_{ijk}e_k+o(|e|).$$

Let $f^{ij}(x)$ represent the (i,j) element of the inverse of the matrix $||f_{ij}(x)||$. Then

$$f^{ij}(x)=f^{ij}(c)-\sum_{s,k,t}f^{is}(c)f_{skt}(c)f^{kj}(c)e_t+o(|e|),$$

$$\delta^{(m)}=L^{-1}(x^{(r)})l(x^{(m)}),$$

$$\delta_i^{(m)}=-e_i^{(m)}+\frac{1}{2}\sum_{j,k,t}f^{ij}(c)f_{jkt}(c)e_k^{(m)}(2e_t^{(r)}-e_t^{(m)})+o(|e^{(r)}|\cdot|e^{(m)}|).$$

Let

$$U_{ik} = \sum_{j,t} f^{ij}(c) f_{jkt}(c) e_t^{(r)}.$$

Then

$$\delta^{(r)} \approx -e^{(r)} + \frac{1}{2} U e^{(r)},$$

and assuming that $e^{(m)}$ is small compared to $e^{(r)}$ for $m > r$, we get

$$\delta^{(m)} \approx -e^{(m)} + U e^{(m)} \qquad \text{for } m > r.$$

Equations 22 are thus established.

Appendix 6. *The modified Newton method in the one-dimensional case.* There is an advantage in the one-dimensional case that derives basically from the fact that for $n > 1$ the directions of the characteristic vectors may vary, while for $n = 1$ there is only one possible direction. We indicate for $n = 1$ a relatively sensitive application of the notion that the results of previous iterations may be used to obtain relevant information concerning higher order derivatives and to obtain good values of h_m. *It should be noted that the less sensitive method suggested in* § 6 *is also applicable and easier to apply.* We have

$$\delta^{(m)} = -e^{(m)}[1 - (2e^{(r)} - e^{(m)} f'''(c)/2f''(c)] + o(|e^{(r)}| \cdot |e^{(m)}|).$$

The desirable value to use for h_m would be approximately

$$\hat{h}_m = 1 + (2e^{(r)} - e^{(m)}) \frac{f'''(c)}{2f''(c)}.$$

Suppose that h_{m-1} were used in place of $\hat{h}_{m-1}$, and $\delta^{(m)}$ is computed using $L(x^{(r)})$. We now make use of the basic approximations

$$\delta^{(m)} \approx -e^{(m)}, \quad x^{(m-1)} + \hat{h}_{m-1} \delta^{(m-1)} \approx c.$$

Then

$$(h_{m-1} - \hat{h}_{m-1}) \delta^{(m-1)} \approx e^{(m)} \approx -\delta^{(m)}.$$

$$\hat{h}_{m-1} - 1 = \frac{[2e^{(r)} - e^{(m-1)}] f'''(c)}{2f''(c)} \approx \frac{\delta^{(m)} + (h_{m-1} - 1)\delta^{(m-1)}}{\delta^{(m-1)}},$$

$$\hat{h}_m - 1 = \frac{2e^{(r)} - e^{(m)}}{2e^{(r)} - e^{(m-1)}} (\hat{h}_{m-1} - 1).$$

Now

$$2e^{(r)} - e^{(m)} = 2e^{(r)} - 2e^{(m)} + e^{(m)} \approx 2x^{(r)} - 2x^{(m)} - \delta^{(m)},$$

$$2e^{(r)} - e^{(m-1)} = 2e^{(r)} - e^{(m-1)} - e^{(m)} + e^{(m)} \approx 2x^{(r)} - x^{(m-1)} - x^{(m)} - \delta^{(m)}.$$

Hence

$$\hat{h}_m = 1 + \frac{2x^{(r)} - 2x^{(m)} - \delta^{(m)}}{2x^{(r)} - x^{(m-1)} - x^{(m)} - \delta^{(m)}} \quad \frac{\delta^{(m)} + (h_{m-1} - 1)\delta^{(m-1)}}{\delta^{(m-1)}}.$$

We may treat the case where $\delta^{(m)}$ is computed using $L(x^{(m)})$ in place of $L(x^{(r)})$ in a similar manner, to obtain

$$\hat{h}_m = 1 - \frac{\delta^{(m)}}{2x^{(r)} - x^{(m)} - x^{(m-1)} - \delta^{(m)}} \quad \frac{\delta^{(m)} + (h_{m-1} - 1)\delta^{(m-1)}}{\delta^{(m-1)}}.$$

References

1. M. Bôcher, *Introduction to higher algebra*, The Macmillan Co., New York, 1907.

2. D. J. Finney, *Probit analysis*, Cambridge University Press, Cambridge, England, 1952.

3. A. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, C. R. Acad. Sci. Paris **25** (1847), 536-538.

4. H. B. Curry, *The method of steepest descent for non-linear minimization problems*, Quart. Appl. Math. **2** (1944), 258-261.

5. T. C. Koopmans, H. Rubin, and R. B. Leipnik, *Measuring the equations systems of dynamic economics*, Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph 10, edited by T. C. Koopmans, John Wiley and Sons, New York, 1950.

6. H. Chernoff and N. Divinsky, *The computation of maximum-likelihood estimates of linear structural equations*, Chapter 10 in Studies in Econometric Method, Cowles Commission Monograph 14, Wm. C. Hood and T. C. Koopmans editors, John Wiley and Sons, Inc. New York, 1953.

7. G. Temple, *The general theory of relaxation methods applied to linear systems*, Proc. Roy. Soc. London, Ser. A, **169** (1938-1939), 476-500.

8. G. E. Forsythe, *Solving linear algebraic equations can be interesting*, Bull. Amer. Math. Soc. **59** (1953), 299-329.

DEPARTMENT OF COMMERCE AND
STANFORD UNIVERSITY

Pacific Journal of Mathematics
Vol. 5, No. 1 September, 1955