COWLES COMMISSION DISCUSSION PAPER: STATISTICS: No. 332

Gradient Methods of Maximization

by Herman Chernoff
Jean Bronfenbrenner

1. Introduction

In estimating a set of parameters by maximum likelihood methods, it is ordinarily necessary to find the point in the parameter space which maximizes a certain function (the likelihood function). Frequently the form of the function and the number of independent variables involved make it prohibitively difficult to determine the maximizing point directly, and methods of successive approximations are accordingly used.

Suppose we have an approximation $x^{(0)}[=(x_1^{(0)}\ x_2^{(0)}\ldots x_n^{(0)})]$ to the point $c\ [=(c_1\ c_2\ \ldots c_n)]$ which maximizes a function $f(x)$. Then if we expand $f(x)$ in a Taylor series about $x^{(0)}$ we have a means for obtaining an improved approximation to c. Presumably the more terms of the expansion we consider, the greater the speed with which we may expect to approach c. However, calculation of terms involving higher order derivatives increases considerably the computational cost of each iteration.

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

2. Gradient Methods

Given an initial approximation $x^{(0)}[=(x_1^{(0)}\ x_2^{(0)}\ldots x_n^{(0)})]$ it is natural to choose 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 with respect to $f(x)$, i.e., in the direction of the gradient. The direction of steepest ascent depends, however, on the concept of distance used. In general there is no reason to assume that a unit of distance along the x_i-axis is in any sense equivalent to a unit of distance along the x_j-axis, $i \neq j$. The metric chosen implies a particular system for weighting these units.

Let $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$, and suppose that the distance from $x^{(0)}$ to the point $(x^{(0)} + \epsilon)$ be defined as

$$d = \left(\sum_{i,j} B_{ij} \epsilon_i \epsilon_j\right)^{\frac{1}{2}}$$

where $B[= \| B_{ij} \|]$ is a positive definite symmetric matrix which may or may not depend on the point $x^{(0)}$. Then the locus of points $(x^{(0)} + \epsilon)$ lying at a distance k from $x^{(0)}$ is given by the ellipsoid

$$\Sigma B_{ij} \epsilon_i \epsilon_j = k^2$$

with center at $x^{(0)}$.

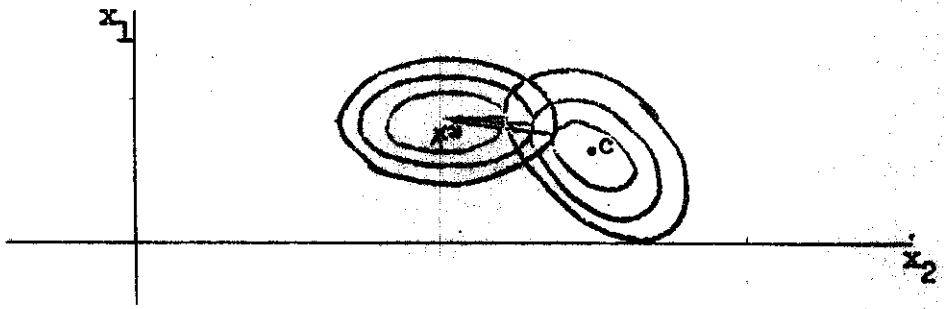
If we now let $\epsilon_i = k \delta_i$ this ellipsoid becomes

$$\Sigma B_{ij} \delta_i \delta_j = 1 \tag{1}$$

Then the direction of steepest ascent in the k-neighborhood of $x^{(0)}$ is the direction from $x^{(0)}$ to the point of the ellipsoid for which the function is a maximum. In the limiting case as k approaches zero it is easy to approximate this direction by writing the Taylor expansion of $f(x^{(0)} + k\delta)$, maximizing subject to the restraint (1) and disregarding higher order terms in k.

For a particular value of k, the maximizing values of the δ_i give the projection on the $x_1 x_2 \ldots x_n$ - space of the steepest secant that can be drawn through the $(n + 1)$ - dimensional point $[x^{(0)}, f(x^{(0)})]$ and any point $[x^{(0)} + k\delta, f(x^{(0)} + k\delta)]$ such that δ satisfies (1). In the limit as $k \to 0$, the maximizing values of the δ_i give the projection of the steepest

tangent line through $[x^{(0)}, f(x^{(0)})]$. The two dimensional case is illustrated in the graph below. The curves enclosing c are contour lines, each corresponding to f(x) = constant. The curves enclosing $x^{(0)}$ are ellipses corresponding to different values of k. The arrows indicate the directions of the steepest secants for the respective values of k.



Let $\underline{\imath}$ be a row vector whose i^{th} element is $\frac{\partial f}{\partial x_i}$ $[= f_i]$ and let L be a matrix whose ij^{th} element is $-\frac{\partial^2 f}{\partial x_i \partial x_j}$ $[= -f_{ij}]$. Then

$$f(x^{(0)} + k\delta) = f(x^{(0)}) + k\,\delta\,\underline{\imath}'(x^{(0)}) - \frac{k^2}{2}\,\delta\,L(x^{(0)})\,\delta' + \ldots \quad 1/$$

Maximizing subject to the restraint (1) we have

$$k\,\underline{\imath}'(x^{(0)}) - k^2\,L(x^{(0)})\,\delta' + \ldots = 2\lambda B\,\delta',$$

where λ is a Lagrange multiplier, and for very small k

$$\delta' \approx B^{-1}\,\underline{\imath}'(x^{(0)})$$

except for a scale factor. This indicates in effect the proper direction for a step, in the limit as the size of the step approaches zero. The problem then arises as to how large a step may profitably be taken in the limiting direction indicated. This clearly depends on how fast the slope of f(x) changes as we depart from $x^{(0)}$ in various directions and so cannot be determined without some consideration of higher order terms.

We observe that if

$$x^{(1)\prime} = x^{(0)\prime} + h\delta' = x^{(0)\prime} + h\,B^{-1}\,\underline{\imath}'(x^{(0)}),$$

1/ A vector x is considered to be a matrix with only one row and x', the transpose of x, would therefore consist of one column. Thus $\delta\,\underline{\imath}'$ is $\Sigma\,\delta_i\,\underline{\imath}_i$ or the scalar product of δ and $\underline{\imath}$.

then for small enough $h > 0$ we necessarily have

$$f(x^{(1)}) > f(x^{(0)}).$$

For

$$f(x^{(1)}) = f(x^{(0)} + h\delta) = f(x^{(0)}) + h\,\delta\,\lambda'(x^{(0)}) + \ldots$$
$$= f(x^{(0)}) + h\,\lambda(x^{(0)})\,B^{-1}\,\lambda'(x^{(0)}) + \ldots$$

But

$$h\,\lambda(x^{(0)})\,B^{-1}\,\lambda'(x^{(0)}) > 0$$

since $h > 0$ and B is positive definite.

It may further be shown that if a sequence of successive approximations $x^{(0)}, x^{(1)}, \ldots, x^{(m)}, \ldots$ is formed, where

$$x^{(m+1)'} = x^{(m)'} + h\,B^{-1}\,\lambda'(x^{(m)}) \tag{2}$$

then if the initial approximation is in a sufficiently small neighborhood of c this sequence converges to c for sufficiently small values of h. To establish this result we expand $\lambda(x)$ about c. Letting $x - c = e$ we have

$$\lambda(x)' = \lambda(c + e)' = \lambda(c)' - L(c)\,e' + \ldots = -L(c)\,e' + \ldots \tag{3}$$

since $\lambda(c) = 0$. If $x^{(m)} - c = e^{(m)}$ we have from (2) and (3)

$$x^{(m+1)'} - c' = x^{(m)'} - c' - h\,B^{-1}L(c)\,e^{(m)'} + \ldots$$

$$e^{(m+1)'} \approx [I - h\,B^{-1}\,L(c)]\,e^{(m)'} \approx [I - h\,B^{-1}L(c)]^{m+1}\,e^{(0)'} \tag{4}$$

Now if c is an isolated maximum, $L(c)$ is a positive definite matrix and $B^{-1}L(c)$ is therefore the product of two positive definite matrices and itself positive definite. Let λ_i and μ_i, $i = 1, 2, \ldots n$ be respectively the characteristic values and vectors of $B^{-1}\,L(c)$. Then $\lambda_i > 0$ for all i. Indeed we assume $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0$. We observe that the μ_i are also the characteristic vectors of $I - h\,B^{-1}\,L(c)$ while $1 - h\lambda_i$, $i = 1, 2, \ldots, n$, are the characteristic values. For by definition

$$[B^{-1} L(c) - \lambda_i I]\mu_i' = 0$$

$$B^{-1} L(c)\mu_i' = \lambda_i \mu_i'$$

$$\mu_i' - h B^{-1} L(c)\mu_i' = \mu_i' - h\lambda_i \mu_i'$$

$$[I - h B^{-1} L(c) - (1 - h\lambda_i)I]\mu_i' = 0$$

For sufficiently small values of h, we have $0 < 1 - h\lambda_i < 1$[1/] since $\lambda_i > 0$.

We may decompose the vector $(x^{(0)} - c) = e^{(0)}$ into components along the n orthogonal characteristic vectors μ_i. Then (4) gives for m = 0

$$e^{(1)'} = [I - h B^{-1} L(C)]\,[k_1 \mu_1' + k_2 \mu_2' + \ldots + k_n \mu_n'] + O(e^{(0)2})$$[2/]

$$= (1 - h\lambda_1)\, k_1 \mu_1' + \ldots + (1 - h\lambda_n)\, k_n \mu_n' + O(e^{(0)2})$$

From this the proof of convergence is evident for an initial approximation close enough to c. We have furthermore

$$e^{(m+1)'} = (1 - h\,\lambda_1)^{m+1} k_1 \mu_1' + \ldots + (1 - h\,\lambda_n)^{m+1} k_n \mu_n' \tag{5}$$

$$\delta^{(m)'} = \frac{x^{(m+1)'} - x^{(m)'}}{h} \approx -B^{-1} L(c)\, e^{(m)'} \approx -B^{-1} L(c)[I - h B^{-1} L(c)]^m e^{(0)'} \tag{6}$$

$$= -\sum_i \lambda_i (1 - h\lambda_i)^m k_i \mu_i'$$

3. Choice of h

We mentioned above that an intelligent choice of h depends on some information regarding higher order terms in the Taylor expansion of f(x). It appears from (5) that the relevant information concerns the characteristic values of $B^{-1} L(c)$, for by letting $h = \frac{1}{\lambda_i}$ we may eliminate the i^{th} terms on the right in (5). In particular it will be shown below how we may, without explicitly evaluating L, use successive iterations with any sufficiently small h to estimate λ_1 and λ_n, the largest and the smallest of the characteristic values.

1/ If B is not strictly constant, we may assume that it approaches a certain value as x approaches c_1. Then λ_i behave in a similar fashion and the inequality above is valid for x sufficiently close to c.

2/ $y = O(x)$ indicates that y is at most of the order of magnitude of x. That is there exists a K so that $|y| \leq K|x|$ as $x \to 0$. The extension of this notation to vectors is trivial.

Now if we let $h = \frac{1}{\lambda_n}$ the last term in (5) vanishes but the coefficient of the first becomes $(1 - \frac{\lambda_1}{\lambda_n})^{m+1}$ and this may be very large if λ_1 is much greater than λ_n. On the other hand if $h = \frac{1}{\lambda_1}$ the first term vanishes but the coefficient of the last becomes $(1 - \frac{\lambda_n}{\lambda_1})^{m+1}$, and this, while less than unity, may decrease very slowly as m increases. Therefore, if a constant h is to be used the value

$$h = \frac{2}{\lambda_1 + \lambda_n}$$

is recommended. This minimizes the maximum valve of $|1 - h\lambda_i|$ by making the largest positive and the largest negative valve numerically equal. I.e., it makes

$$-(1 - h\lambda_1) = 1 - h\lambda_n = 1 - \frac{2\lambda_n}{\lambda_1 + \lambda_n}$$

$$= \frac{\lambda_1 - \lambda_n}{1 + n}$$

$$= \frac{1 - \frac{n}{1}}{1 + \frac{n}{1}} \quad 1/$$

However there is no reason why a constant valve of h should be used in successive iterations. On the contrary it would seem desirable to vary h over the range between $1/\lambda_1$ and $1/\lambda_n$ in order to provide opportunities for coming close to the receprocals of as many of the λ_i as possible. In this case we have instead of (5) and (6)

$$(7) \qquad e^{(m+1)'} = \sum_i k_i \mu_i' \left[\prod_{k=0}^{m} (1 - h_k \lambda_i) \right]$$

$$(8) \qquad \delta^{(m)'} = \frac{e^{(m+1)'} - e^{(m)'}}{h_m} = -\sum_i \lambda_i \prod_{k=0}^{m-1} (1 - h_k \lambda_i) k_i \mu_i'$$

1/ Since this is less than unity, it follows from (5) that such an h is small enough to assure convergence if the original approximation is close enough to the maximum.

The procedure for estimating the smallest root λ_n from the successive approximations $x^{(m)}$ is as follows, assuming $\lambda_i > \lambda_n$ for all $i \neq n$. Let $x_j^{(m)}$ be the j^{th} element of $x^{(m)}$, $\delta_j^{(m)}$ the j^{th} element of $\delta^{(m)}$ and $(\mu_i)_j$ the j^{th} element of μ_i. Then from (8)

$$\frac{\delta_j^{(m)}}{\delta_j^{(m-1)}} = \frac{-\sum_i k_i (\mu_i)_j \left[\lambda_i \prod_{k=0}^{m-1} (1 - h_k \lambda_i) \right]}{-\sum_i k_i (\mu_i)_j \left[\lambda_i \prod_{k=0}^{m-2} (1 - h_k \lambda_i) \right]}$$

Since λ_n is the smallest of the λ_i it follows that in both numerator and denominator the last term in the summation becomes dominant after many iterations with small enough positive h. For if we divide both numerator and denominator by $\prod_{k=0}^{m-2} (1 - h_k \lambda_n)$ then the i^{th} term of each will contain as a factor

$$\frac{\prod_{k=0}^{m-2} (1 - h_k \lambda_i)}{\prod_{k=0}^{m-2} (1 - h_k \lambda_n)} = \prod_{k=0}^{m-2} \frac{1 - h_k \lambda_i}{1 - h_k \lambda_n} \leq \left(\frac{1 - \bar{h} \lambda_i}{1 - \bar{h} \lambda_n} \right)^{m-1}$$

where h is the smallest of the h_k. For m sufficiently large this becomes arbitrarily small, $i \neq n$. Since the remaining factors are bounded, it follows that for large m we may neglect all terms except the last in both numerator and denominator. Thus we have

$$(9) \qquad \frac{\delta_j^{(m)}}{\delta_j^{(m-1)}} \approx \frac{(k_n \mu_n)_j \prod_{k=0}^{m-1} (1 - h_k \lambda_n)}{(k_n \mu_n)_j \prod_{k=0}^{m-2} (1 - h_k \lambda_n)} = (1 - h_{m-1} \lambda_n)$$

Since only λ_n is unknown we may solve this equation for the desired estimate of λ_n.

The estimation of λ_1 is more difficult and much less precise. Given an approximation to $[B^{-1} L(c)]^m e^{(0)'}$, we may make use of the fact that for large m

$$\begin{aligned}[B^{-1} L(c)]^m e^{(0)'} &= [B^{-1} L(c)]^m (k_1 \mu_1' + k_2 \mu_2' + \cdots + k_n \mu_n') \\ &= \lambda_1^m k_1 \mu_1' + \lambda_2^m k_2 \mu_2' + \cdots + \lambda_n^m k_n \mu_n' \\ &\approx \lambda_1^m k_1 \mu_1'\end{aligned}$$

since λ_1 is the largest characteristic value. Then

$$(10) \qquad \frac{j^{th} \text{ element of } [B^{-1} L(c)]^m e^{(0)'}}{j^{th} \text{ element of } [B^{-1} L(c)]^{m-1} e^{(0)'}} \approx \frac{\lambda_1^m (\mu_1)_j}{\lambda_1^{m-1} (\mu_1)_j} = \lambda_1$$

The approximation to $[B^{-1} L(c)]^m e^{(0)'}$ may be obtained by a recursive process. We have

$$(11a) \qquad (x^{(1)} - x^{(0)})' = e^{(1)'} - e^{(0)'} \approx [I - h_0 B^{-1} L(c)] e^{(0)'} - e^{(0)'} = -h_0 B^{-1} L(c) e^{(0)'}$$

$$(x^{(2)} - x^{(1)})' \approx -h_1 B^{-1} L(c) e^{(1)'} = -h_1 B^{-1} L(c) [I - h_0 B^{-1} L(c)] e^{(0)'}$$

$$(11b) \qquad = -h_1 B^{-1} (L(c) e^{(0)'} + h_0 h_1 [B^{-1} L(c)]^2 e^{(0)'}$$

$$(x^{(3)} - x^{(2)})' \approx -h_2 B^{-1} L(c) [I - h_1 B^{-1} L(c)][I - h_0 B^{-1} L(c)] e^{(0)'}$$

$$(11c) \qquad = -h_2 B^{-1} L(c) e^{(0)'} + h_2 (h_0 + h_1) [B^{-1} L9c)]^2 e^{(0)'}$$

$$- h_0 h_1 h_2 [B^{-1} L(c)]^3 e^{(0)'}$$

Etc.

From these we obtain the following relations by repeated substitution.

$$(12a) \qquad B^{-1} L(c) e^{(0)'} \approx -\frac{1}{h_0} (x^{(1)} - x^{(0)})'$$

$$(12b) \qquad [B^{-1}L(c)]^2\, e^{(0)'} \approx -\frac{1}{h_0^2 h_1}\left[h_1(x^{(1)} - x^{(0)}) - h_0(x^{(2)} - x^{(1)})\right]'$$

$$(12c) \qquad [B^{-1}L(c)]^3\, e^{(0)'} \approx \frac{1}{h_0^3 h_1^2 h_2}\left[-h_0^2 h_1(x^{(3)}-x^{(2)})' + h_0 h_2(h_0+h_1)(x^{(2)}-x^{(1)})' - h_1^2 h_2(x^{(1)}-x^{(0)})'\right]$$

Etc.

It should be noted that (12b) depends on both 11b and 12a; (12c) depends on (11c) (12a) and (12b); and so on. Since all these expressions neglect terms in higher powers of the elements of $e^{(0)}$, this provides opportunity for the cumulation of errors as m increases and limits the accuracy with which λ_1 may be estimated by this process.

4. Newton Method

We have not yet discussed the choice of the matrix B. There may be some a priori reason for weighting errors in one parameter more strongly than errors in another. Barring this situation it is desirable to choose B in such a way as to speed convergence as much as possible. This may involve varying B between iterations. If B^{-1} L(c) is close to the identity matrix then its characteristic values, the λ_i, will all be close to unity. Since from (5)

$$e^{(m+1)'} = \sum_{i=1}^{n} (1-h\,\lambda_i)^{m+1} k_i \mu_i' + O(e^{(m)2})$$

it is clear that a choice of h close to unity will then quickly reduce all the terms in the summation on the right, permitting $x^{(m)}$ to converge rapidly to c.

With this in mind it would seem desirable to set $B = L(x^{(m)})$ since in the neighborhood of c we may expect $L^{-1}(x^{(m)})$ L(c) to be close to the

identity matrix. Then we have

$$x^{(m+1)'} = x^{(m)'} + L^{-1}(x^{(m)})\, \nabla'(x^{(m)})$$

or

$$\delta^{(m)'} = x^{(m+1)'} - x^{(m)'} = L^{-1}(x^{(m)})\, \nabla'(x^{(m)}) \tag{13}$$

But this is equivalent to the Newton method of successive approximations, in which we approximate $f(x)$ to second order terms of the Taylor expansion (about $x^{(m)}$) and then chose $x^{(m+1)}$ so as to maximize the quadratic thus obtained. For if we write

$$f(x) = f(x^{(m)} + \delta^{(m)}) \approx f(x^{(m)}) + \delta^{(m)} \nabla'(x^{(m)}) - \frac{1}{2}\, \delta^{(m)} L(x^{(m)})\, \delta^{(m)'}$$

and maximize with respect to $\delta^{(m)}$ we obtain

$$\nabla'(x^{(m)}) - L(x^{(m)})\, \delta^{(m)'} = 0$$

and (13) follows immediately.

It should be noted that speed of convergence increases as we move closer to the maximum. This is obvious since the characteristic values, λ_i, are pushed together as $L^{-1}(x^{(m)})\, L(c)$ approaches the identity matrix. In fact it may be shown that in the neighborhood of (c), the elements of $e^{(m+1)}$ are of the order of magnitude of the squares of the elements of $e^{(m)}$. This property makes the Newton method valuable when we are very close to the maximum.

However, the computational cost of finding the second order terms for each iteration may be considerable. To lessen this cost while still retaining some of the advantages of the Newton method, the following modification may be introduced. The metric B (i.e., the second order matrix L) may be held constant for several iterations. That is, we may use $L(x^{(r)})$ to compute $\delta^{(r+1)}, \delta^{(r+2)}, \ldots, \delta^{(r+k)}$, thus avoiding calculation of $L(x^{(r+1)}), L(x^{(r+2)}), \ldots, L(x^{(r+k)})$. However there will be no change in the λ_i and no increase in the speed of convergence so long as the metric is held constant.

5. The "Finagle Factor"

Much of our discussion of simple gradient methods centered around the size of the step to be taken in the chosen direction, i.e., the choice of h. The Newton method in its ordinary form implies a value of h equal to 1. So long as this is true it will be found that the iterations tend to undershoot or overshoot the mark in a systematic fashion depending on third order terms. By observing successive iterations we may correct for this systematic tendency without actually computing the third order terms. The following procedure is applicable to the modified Newton method. We deal first with the one-dimensional case.

Let $x^{(r)}$ be the last point for which the second order term was evaluated. Then using the expansions of f and its derivatives about c, we have for $\delta^{(m)}$

$$\delta^{(m)} = -\frac{f'(x^{(m)})}{f''(x^{(r)})} = -\frac{e^{(m)}f''(c) + \frac{1}{2}e^{(m)2}f'''(c) + O(e^{(m)3})}{f''(c) + e^{(r)}f'''(c) + O(e^{(r)2})} \tag{14}$$

If $x^{(m+1)} = x^{(m)} + \delta^{(m)}$, we have

$$e^{(m+1)} = e^{(m)} + \delta^{(m)} = \frac{\frac{1}{2}e^{(m)}(2e^{(r)} - e^{(m)})f'''(c) + O(e^{(r)2}e^{(m)})}{f''(c) + e^{(r)}f'''(c) + O(e^{(r)2})}$$

$$= e^{(m)}(2e^{(r)} - e^{(m)})\frac{f'''(c)}{2f''(c)}[1 + O(e^{(r)})] \tag{15}$$

Since $e^{(m)}$ is very close to $-\delta^{(m)}$ it is clear that we could substantially reduce the right hand side of (15) by adding to it

$$\delta^{(m)}(2e^{(r)} - e^{(m)})\frac{f'''(c)}{2f''(c)}$$

Thus convergence would be speeded if we formed $x^{(m+1)}$ by adding to $x^{(m)}$, not $\delta^{(m)}$, but $\delta^{(m)}$ multiplied by an appropriate "finagle factor":

$$\tilde{x}^{(m+1)} = x^{(m)} + (1 + \varphi^{(m)})\delta^{(m)}$$

where

$$\varphi^{(m)} = (2e^{(r)} - e^{(m)}) \frac{f'''(c)}{2f''(c)} \tag{16}$$

Then

$$\tilde{e}^{(m+1)} = e^{(m)} + \delta^{(m)}(1+\varphi^{(m)}) = O(e^{(m)}\, e^{(r)2})$$

Thus far $\varphi^{(m)}$ is unknown. A relatively unsophisticated method of approximating this factor is the following. Compute $\delta^{(r)}$, $x^{(r+1)} = x^{(r)} + \delta^{(r)}$ and $\delta^{(r+1)}$.

$$-\delta^{(r+1)} \approx e^{(r+1)} \approx e^{(r)} \cdot e^{(r)} \frac{f'''(c)}{2f''(c)} \approx \delta^{(r)2} \frac{f'''(c)}{2f''(c)} \tag{17}$$

This can be used to estimate $f'''(c)/2f''(c)$.

We note that if $x^{(r)}$ is close to c, $e^{(r+1)}$, $e^{(r+2)}$... will be small compared to $e^{(r)}$ and while $\varphi^{(r)} = e^{(r)} \frac{f'''(c)}{2f''(c)}$, $\varphi^{(r+1)} \approx \varphi^{(r+2)} \approx \ldots \approx 2e^{(r)} \frac{f'''(c)}{2f''(c)}$.

Approximating $e^{(r)}$ by $-\delta^{(r)}$ we have from 17

$$\varphi^{(r+1)} \approx (-2\delta^{(r)} - \frac{\delta^{(r+1)}}{\delta^{(r)2}} = \frac{2\delta^{(r+1)}}{\delta^{(r)}} \tag{18}$$

If $x^{(r)}$ is not very close to c, $\varphi^{(m)}$ will tend to change a little less rapidly from $e^{(r)} \frac{f'''(c)}{2f''(c)}$ to $2e^{(r)} \frac{f'''(c)}{f''(c)}$

For this reason one may find the following slightly more sophisticated approach useful especially in those cases where the work per iteration is considerable. This approach also makes it possible to estimate $\frac{f'''(c)}{2f''(c)}$ without performing——as required by (17)——an iteration in which no "finagle factor" is used.

Suppose that $\hat{\varphi}^{(m-1)}$ was used as an approximation to $\varphi^{(m-1)}$ in the m^{th} iteration. ($\varphi^{(m-1)}$ is never known exactly and may be approximated with varying degrees of accuracy. $\hat{\varphi}^{(m-1)}$ may even be zero if no previous knowledge concerning third order terms has been obtained). Thus $x^{(m)} = x^{(m-1)} + \delta^{(m-1)}(1+\hat{\varphi}^{(m-1)})$

Then for $m \geq r+1$, we use (15) to obtain

$$(19) \qquad e^{(m)} = e^{(m-1)} + \delta^{(m-1)}(1+\hat{\varphi}^{(m-1)}) = e^{(m-1)}\left\{[2e^{(r)} - e^{(m-1)}]\frac{f'''(c)}{2f''(c)}\right\} + \hat{\varphi}^{(m-1)}\,\delta^{(m-1)} + O(e^{(r)2}\, e^{(m)})$$

Solving this to find an approximate value for $\frac{f'''(c)}{2f''(c)}$ and substituting in (16) we have

$$\varphi^{(m)} \approx \frac{2e^{(r)} - e^{(m)}}{[2e^{(r)} - e^{(m-1)}]}\left[\frac{e^{(m)} - \hat{\varphi}^{(m-1)}\,\delta^{(m-1)}}{e^{(m-1)}}\right]$$

The error terms can be approximated by various methods of somewhat different accuracy. For example we may use (15) to note that

$$e^{(m)} = -\delta^{(m)} + O(e^{(m)}\, e^{(r)})$$

$$e^{(m-1)} = e^{(m-1)} - e^{(m)} + e^{(m)} = x^{(m-1)} - x^{(m)} - \delta^{(m)} + O(e^{(m)}\, e^{(r)})$$

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

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

Then

$$(20) \qquad \varphi^{(m)} \approx \hat{\varphi}^{(m)} = \frac{2x^{(r)} - 2x^{(m)} - \delta^{(m)}}{2x^{(r)} - x^{(m)} - x^{(m-1)} - \delta^{(m)}} \cdot \frac{\delta^{(m)} + \hat{\varphi}^{(m-1)}\,\delta^{(m-1)}}{x^{(m)} - x^{(m-1)} + \delta^{(m)}}$$

If after obtaining $x^{(r_1)}$ $(r_1 > r)$, it is decided to recalculate second order terms we may proceed in the same manner to obtain

$$(20a) \qquad \hat{\varphi}^{(r_1)} = \frac{-\delta^{(r_1)}}{2x^{(r)} - x^{(r_1)} - x^{(r_1-1)} - \delta^{(r_1)}} \cdot \frac{\delta^{(r_1)} + \hat{\varphi}^{(r_1-1)}\,\delta^{(r_1-1)}}{x^{(r_1)} - x^{(r_1-1)} + \delta^{(r_1)}}$$

In the n-dimensional case we are not in a position to determine the "finagle factor" so definitely or uniquely, since we have a vector rather than a scalar to eliminate. The best we can do is make use of an appropriate succession of h's, which, as we might expect, are related to the characteristic roots of a certain matrix. Again let $x^{(r)}$ be the last point for which the second order terms were evaluated.

Expanding $f_i(x)$ and $f_{ij}(x)$ in Taylor series about c we have

$$f_i(x) = \sum_k e_k f_{ik}(c) + \frac{1}{2} \sum_{n,p} e_n e_p f_{inp}(c) + \cdots$$

$$f_{ij}(x) = f_{ij}(c) + \sum_n e_n f_{ijn}(c) + \cdots$$

We have furthermore[1]

$$f^{ij}(x) = f^{ij}(c) - \sum_k f^{ik}(c) \sum_q e_q \sum_s f_{kqs}(c) f^{sj}(c) + \cdots$$

$$\delta_j^{(m)} = - \sum_i f_i(x^{(m)}) f^{ij}(x^{(r)})$$

$$= - \sum_{i,k} e_k^{(m)} f_{ik}(c) f^{ij}(c) - \frac{1}{2} \sum_{i,n,p} e_n^{(m)} e_p^{(m)} f_{inp}(c) f^{ij}(c)$$

$$+ \sum_{i,k} e^{(m)} f_{ik}(c) \sum_k f^{ik}(c) \sum_q e_q^{(r)} \sum_s f_{kqs}(c) f^{sj}(c) + \cdots$$

$$= - e_j^{(m)} - \frac{1}{2} \sum_{i,n,p} e_n^{(m)} e_p^{(m)} f_{inp}(c) f^{ij}(c) + \sum_{k,q,s} e_k^{(m)} e_q^{(r)} f_{kqs}(c) f^{sj}(c) + \cdots$$

$$= - e_j^{(m)} + \frac{1}{2} \sum_{i,n,p} e_n^{(m)} (2e_p^{(r)} - e_p^{(m)}) f_{inp}(c) f^{ij}(c) + \cdots$$

$$e_j^{(m+1)} = e_j^{(m)} + h_m \delta_j^{(m)}$$

$$= (1-h_m) e_j^{(m)} + \frac{h_m}{2} \sum_{i,n,p} e_n^{(m)} (2e_p^{(r)} - e_p^{(m)}) f_{inp}(c) f^{ij}(c) + \cdots$$

Suppose m = r. Then

$$\delta_j^{(r)} = -e_j^{(r)} + \frac{1}{2} \sum_{i,n,p} e_n^{(r)} e_p^{(r)} f_{inp}(c) f^{ij}(c) + \cdots$$

$$e_j^{(r+1)} = (1-h_r) e_j^{(r)} + \frac{h_r}{2} \sum_{i,n,p} e_n^{(r)} e_p^{(r)} f_{inp}(c) f^{ij}(c) + \cdots$$

If we let $U = \| U_{jn} \|$ where

$$U_{jp} = \sum_{in} e_n^{(r)} f_{inp}(c) f^{ij}(c)$$

then we may write in vector notation

---

[1] Here we are making use of the following facts. If $L(x) = \| f_{ij}(x) \|$ then $L(c+e) = L(c) + D(e)$ where $D(e) = \| \sum_n e_n f_{ijn}(c) + O(e_p e_q) + \cdots \|$
For (e) sufficiently small
$$L^{-1}(x) = L^{-1}(c) - L^{-1}(c) D(e) L^{-1}(c) + O(D^2)$$

then we may write in vector notation

$$e^{(r+1)'} = (I - h_r I)e^{(r)'} + \frac{h_r}{2} U\, e^{(r)'} + \cdots$$

$$= \left[I - h_r(I - \frac{U}{2}) \right] e^{(r)'} + \cdots$$

Since the elements of $e^{(r+1)}$ are small compared with those of $e^{(r)}$ when we are close to the maximum we have for the next iteration

$$e_j^{(r+2)'} \approx (1 - h_{r+1})e_j^{(r+1)} + \frac{h_{r+1}}{2} \sum_{i,n,p} e_n^{(r+1)}\, 2e_{ip}^{(r)}\, f_{inp}(c)\, f^{ij}(c) + \cdots$$

$$e^{(r+2)'} \approx \left[I - h_{r+1}(I - U) \right] e^{(r+1)'} + \cdots$$

Similarly, for $m > r+1$, so long as $L(x^{(r)})$ is used

$$e^{(m+1)'} \approx \left[I - h_m(I - U) \right] e^{(m)'} \approx \prod_{p=r+1}^{m} \left[I - h_p(I - U) \right] e^{(r+1)'}$$

In this case we may knock out the major part of $e^{(m+1)'}$, thus speeding convergence, if we choose h's close to the reciprocals of the characteristic values of I-U, where $U = O(e^{(r)})$ is a matrix involving third order terms of the expansion of f. The procedure for determining appropriate values of h from successive iterations is very similar to that discussed earlier for other gradient methods. Whenever second order terms are recomputed U will, of course, change, and the relevant characteristic values will change accordingly. This was to be expected since, in the forms used in our discussion of gradient methods, a recomputation of second order terms means a change in the metric B. We have the additional information that the subsequent iterations without recalculation of second order terms have the effect of spreading the relevant characteristic values in a systematic fashion. For if ν_i, $i = 1, 2, \ldots, n$ are the characteristic values of U, then $1 - \frac{\nu_i}{2}$ are the characteristic values of $I - \frac{U}{2}$ as compared with $1 - \nu_i$ for I-U.

6. Concluding Remarks

The method proposed to find the smallest λ can be modified to eliminate dominating characteristic components with small λ. If several comparatively small h's are used the relative importance of characteristic vectors with large λ in the error terms is diminished. If initially the component with the smallest λ is very small it may be that after a few such iterations the main components will be those with λ close to a certain value, say λ_0. This will be apparent by noting that $\delta^{(m)}$ is very close to a multiple of $\delta^{(m-1)}$.

$$\delta^{(m)} \approx (1-h_{m-1}\lambda_0)\,\delta^{(m-1)}$$

Then it would be wise to use $h_m = \frac{1}{\lambda_0}$. This would tend to eliminate the largest components of the error. If this h is quite large, it may tend to revive components with large λ's for $1-h_m\lambda$ would be highly negative. It is therefore wise to follow a large h with several small ones. A heuristic method of estimating λ_0 is illustrated by the following. If $\delta^{(m)} = .6\ \delta^{(m-1)}$, h_{m-1} was only .4 as large as it should have been, therefore, $h_m = \frac{10}{4} h_{m-1}$. This method is equivalent to the above for if $1-h_{m-1}\lambda_0 = .6$, $\lambda_0 = .4/h_{m-1}$, $h_m = 1/\lambda_0 = \frac{10}{4} h_{m-1}$. If this method is applied to the Modified Newton Method, one should allow for the spread of the λ about 1 immediately after the second order terms are calculated. For example, if $h = 1$ is used when the second order terms are calculated and one obtains $\delta^{(r+1)} = .1\ \delta^{(r)}$; $h_r = 1$ is only .9 as large as it should be; thus, we may prefer to use $h_{r+1} = \frac{10}{9}$. However, to allow for the spread h_{r+1} close to $1+\frac{2}{9}$ would be preferable.

The method of finding the largest characteristic root of a matrix on which is based the method of finding the smallest λ is a well known method. The Newton method and the gradient method with Euclidean Metric

$B = I$ have been known and used frequently in the past.[1] Indeed Curry has noted that the iterations with different scales along the axes would give different results. Generalizations of these methods have been applied to functions defined on spaces which are not Euclidean.[2]

The Cowles Commission has been using these methods for several years in Statistical problems of Econometrics.[3] Indeed in these problems one can make use of the fact that for reasonably large samples, certain matrices which are relatively easy to compute are practically equivalent to the matrix of second order terms.

1/ H. B. Curry, "The Method of Steepest Descent for Non-Linear Maximization Problems," Quarterly of Applied Mathematics, Oct. 1944. In this paper Curry also refers to papers of Cauchy, Courant, and Hadamard in which these methods have been used.

2/ Wolfowitz has applied the generalization of the Newton method to a problem in Calculus of Variations in an unpublished paper.

3/ Cowles Commission Monograph 10, Section 4 (to be published).