

February 24, 1948

Gradient Methods of Maximization

by Herman Chernoff

1. The P_n methods used are essentially gradient methods in that steps are taken in the direction of the gradient (uphill) with respect to a certain metric. While a gradient method will converge under comparatively simple assumptions for small enough h , it is necessary for speed to use h large enough but not too large. It was shown that for the non diagonal and diagonal cases $h \leq 1$, $h \leq 2$ would give convergence (with some question as to the speed if the original approximation is in a sufficiently small neighborhood of the absolute maximum. However, it may well be that in a particular example $h > 2$ may have better convergence properties.

We develop a local theory which will indicate what values of h to be used in actual iterations.

2. Suppose

$$f(x) = f(0) - \frac{1}{2} \sum a_{ij} x_i x_j \quad \text{neglecting higher order terms}$$

where $A = \| a_{ij} \|$ is a positive definite matrix. Suppose x^0 is an initial vector. Then

$$\begin{aligned} (1) \quad f(x^0 + \delta) &= f(x^0) + \sum \delta_i \frac{\partial f}{\partial x_i} + \frac{1}{2} \sum \delta_i \delta_j \frac{\partial^2 f}{\partial x_i \partial x_j} \\ &= f(x^0) - \sum \delta_i x_j^0 a_{ij} - \frac{1}{2} \sum \delta_i \delta_j a_{ij} \end{aligned}$$

$$(2) \quad f(x^0 + \delta) - f(x^0) = - \sum \delta_i x_j^0 a_{ij} + \text{higher order terms.}$$

If a matrix $B > 0$ determines the metric of importance of errors in estimation; i.e. we wish to minimize $x^0 B x^1$; then of the components of x , it is reasonable to attempt to consider the gradient to be determined by the direction δ_1^0 in which $f(x^0 + \delta) - f(x^0)$ increases most for $\frac{1}{2} \delta B \delta' = \text{const.}$

Using Lagrange multiplier

$$(3) \quad - \sum_{j=1}^k a_{ij} x_j^0 + \lambda_0 \sum_{j=1}^k b_{ij} \delta_j^0 = 0$$

Except for a scale factor $B \delta^0 = A x^0$

$$(4) \quad \delta^0 = -B^{-1} A x^0 \quad ; \quad x^1 = x^0 + h_0 \delta^0 = (I - h_0 B^{-1} A) x^0$$

and

$$f(x^{(1)}) - f(x^0) = f(x^0 + h \delta^0) - f(x^0) = h \sum_i \delta_i^0 \sum_j B_{ij} + \text{higher order terms}$$

$$f(x^{(1)}) - f(x^0) = f(x^0 + h \delta^0) - f(x^0) > 0 \quad \text{for } h \text{ sufficiently small.}$$

3. By this method of iteration we obtain

$$x^{(1)} = (I - h_0 B^{-1} A) x^0$$

$$x^{(2)} = (I - h B^{-1} A) (I - h_0 B^{-1} A) x^0$$

$$(5) \quad x^{(n)} = \prod_{i=1}^{n-1} (I - h_i B^{-1} A) x^0$$

where in the general case B, A may vary from pt. to pt., i.e. we could put in B_i, A_i to indicate dependence on higher order terms. It is desired that $x^{nB} x^{n'} \rightarrow 0$ as rapidly as possible. This seems to offer complications. Consider on the other hand just $x^{n'} \rightarrow 0$ rapidly.

Let us assume that $C = B^{-1} A$ is constant and h is left fixed; then we have merely to investigate the characteristic values of $C > 0$; suppose they range from $0 < \lambda_1 < \lambda_2 < \dots < \lambda_k$. Then the char. values of $(I - h C)$ are $1 - h \lambda_i$.

We minimize $\text{Max } |1 - h \lambda_i|$ by setting $1 - h \lambda_1 = -(1 - h \lambda_k)$

$$\text{i.e. } h = \frac{2}{\lambda_1 + \lambda_k}$$

$$\text{Max } |1 - h \lambda_i| = 1 - \frac{2 \lambda_1}{\lambda_1 + \lambda_k} = \frac{\lambda_k - \lambda_1}{\lambda_k + \lambda_1} = \frac{1 - \lambda_1/\lambda_k}{1 + \lambda_1/\lambda_k}$$

At this point we see B from another point of view. If the char values of C are such that λ/λ_k is close to one there is an h which will make $x^{(n)} \rightarrow 0$ very rapidly. If λ/λ_k is close to zero $x^{(n)} \rightarrow 0$ slowly by the P_n method. λ/λ_k close to one means that $C \approx CI$. λ/λ_k close to zero means that C is almost singular when normalized. Thus a good B to use would be one such that

B is easy to invert and

B is very close to CA, i.e. $C = B^{-1} A \approx CI$

4. Determination of h in Practice.

For the nondiagonal case $h = 1$ is smaller than optimum in the neighborhood of the max.

For the diagonal case $h = 1/2$

However there has been no indication of what h is an optimum. The following development will indicate methods of improving h.

Our data never gives $x^{(n)}$ but only gives $d^{(n)} = x^{(n)} - x^{(n-1)}$.

a) λ_1 : From this data we can determine λ_1 .

for $d^{(n)} = x^{(n)} - x^{(n-1)} = [(I - hC)^n - (I - hC)^{n-1}] x^{(0)} = (I - hC)^{n-1} (-hCx^{(0)})$

Suppose h is smaller than it should be. Then the largest charact. value of $I - hC = 1 - h\lambda_1$ will show up by comparing the ratios $d_i^{(n+1)}/d_i^{(n)} \approx 1 - h\lambda_1$ for sufficiently large n.

It should be kept in mind that this method is not too exact if h is very small, for then $1 - h\lambda_k$ are all close to one. However in this case $1 - h\lambda_1$ will not be overestimated and λ_1 not underestimated.

b) λ_k

The evaluation of λ_k is rather delicate.

$$d^{(n)} = (I - hC)^{n-1} (-hCx^0)$$

$$(hC)^n = \left[I - (I-hC) \right]^n = I - n(I-hC) + \binom{n}{2} (I-hC)^2 \dots$$

$$-(hC)^{n+1} x^0 = d^{(1)} - nd^{(2)} + \binom{n}{2} d^{(3)} + \dots$$

Because evaluation of $(hC)^{n+1}$ involves $d^1, d^2, \dots, d^{(n)}$, this is quite susceptible to accumulation of errors and also the effects of higher order terms. Furthermore we can not let the result of the 10th iteration be x^0 and work from there (as may be done in the λ evaluation). This would give misleading results because x^0 would be the charact. vector of $(I - h\lambda_k)$ of C . Instead we must be content with just a comparison of hC, h^2C^2, h^3C^3 to get a rough idea of λ_k . One must keep in mind the danger of underestimating λ_k .

c) Reevaluation.

It seems as though a wise procedure would be to

- 1) take about 3 iterations with a small enough h ;
- 2) Observe the differences $d^{(1)}, d^{(2)}, d^{(3)}$. If they are very nearly equal, h is too small.

Estimate (λ_1, λ_k)

[λ_1 is close to one and reasonably easy to calculate]

3) Re-evaluate h ;

Here it may be permissible to select h_4 so that $1 - h_4\lambda_1 \approx 0$ and then to let $h_5 = h_6 = \dots = h$ a value safer in that $1 - h\lambda_k$ would not be too negative. (h_4) would serve to eliminate the effects of the char vector of C corresponding to (λ_1) .

4) With our new safe value perform about 3 or 4 more iterations keeping on alert for too small an h or too large an h . If h is too small a positive char value of $(I-hC)$ will assert itself in that we shall have almost

$$d^{(i+1)} = \lambda d^{(i)} \qquad \lambda \approx 1 - h\lambda_1$$

If h is too large, a negative charact value of $(I - hC)$ will assert itself.

$$\lambda = 1 - h\lambda_k$$

If h is approximately right we should find that

$$d^{(i+2)}/d^{(i)} \approx \lambda^2$$

but not $d^{(i+1)}/d^{(i)} \approx \lambda^2$

$$\lambda^2 = (1 - h \lambda_i)^2 = (1 - h \lambda_k)^2$$

In all three of these cases we can eliminate the major factors by taking for one step $h = 1/\lambda_i$, $h = 1/\lambda_k$ or in the just right case first $h = 1/\lambda_i$ and then $h = 1/\lambda_k$

5) Finally we go back to the good value of h with a little improvement.