On the Relative Efficiencies of Gradient Methods

By John Greenstadt

Summary. A comparison is made among various gradient methods for maximizing a function, based on a characterization by Crockett and Chernoff of the class of these methods. By defining the "efficiency" of a gradient step in a certain way, it becomes easy to compare the efficiencies of different schemes with that of Newton's method, which can be regarded as a particular gradient scheme. For quadratic functions, it is shown that Newton's method is the most efficient (a conclusion which may be approximately true for nonquadratic functions). For functions which are not concave (downward), it is shown that the Newton direction may be just the opposite of the most desirable one. A simple way of correcting this is explained.

In trying to maximize a function f of the N variables $\{x_1, x_2, \cdots, x_N\}$, iterative techniques of the class known as *gradient methods* have proved of great utility [1], [2].

For general functions, the analysis usually begins with an expansion of f in a Taylor series around some point x ($\equiv \{x_1, x_2, \cdots, x_N\}$):

$$(1) \qquad f(x+\delta) = f(x) + \sum_{i=1}^{N} \delta_i \frac{\partial f}{\partial x_i} + \frac{1}{2} \sum_{i,k=1}^{N} \delta_i \delta_k \frac{\partial^2 f}{\partial x_i \partial x_k} + O(\delta^3) .$$

It is assumed that the matrix:

$$\{-L_{ik}\} \equiv \{\partial^2 f / \partial x_i \partial x_k\}$$

is negative definite, which means that $\{L_{ik}\}$ itself is positive definite. The direction of $\{\delta_i\}$ is denoted now by $\{s_i\}$, and its length by h. The vector of first derivatives $\{\partial f/\partial x_i\}$ is denoted by $\{g_i\}$. In matrix notation, we then have:

$$(2) \qquad f(x+hs) \cong f(x) + hg^T s - \tfrac{1}{2} h^2 s^T L s$$

(where the superscript T indicates transposition).

The arguments which follow are exact for quadratic functions of the form in equation (2); they are hence approximately true for nonquadratic functions. For the time being, we shall regard f as quadratic.

It was shown by Crockett and Chernoff* [2] that the direction of maximum rate of increase of f, *assuming the normalization* $s^T G s = 1$ (where G is some positive definite matrix) is given by:

$$(3) \qquad s = G^{-1} g / (g^T G^{-1} g)^{1/2} .$$

If this direction is then followed until f begins to decrease, the step covered will then have, for a quadratic f, the length:

$$(4) \qquad h = g^T s / s^T L s .$$

Received May 18, 1966.

* See the Appendix for an outline of this analysis.

The total change in f due to taking this optimal step in the direction s is then:

$$\Delta f_s = f(x + hs) - f(x) = \tfrac{1}{2}\,(g^T s)^2/s^T Ls \tag{5}$$

or, substituting from (3):

$$\Delta f_G = \frac{1}{2}\frac{(g^T G^{-1} g)^2}{g^T G^{-1} L G^{-1} g}\,. \tag{6}$$

We wish now to compare this net increase with that due to taking an optimum step in the direction given by Newton's method. In this case, it is merely necessary to take $G = L$. We then have:

$$\Delta f_N = \tfrac{1}{2}\,(g^T L^{-1} g)\,. \tag{7}$$

We shall now examine the ratio of Δf_G to Δf_N, which we shall take to be a measure of the efficiency of the procedure:

$$\rho \equiv \frac{\Delta f_G}{\Delta f_N} = \frac{(g^T G^{-1} g)^2}{(g^T G^{-1} L G^{-1} g)(g^T L^{-1} g)}\,. \tag{8}$$

Since, in a quadratic form such as (1), there is no *a priori* connection between the first and second derivatives, we may assume that g is an arbitrary vector. Also, since G is assumed positive definite, we can find its symmetric positive-definite square root, and we may set:

$$p = G^{-1/2} g \tag{9}$$

where p is also an arbitrary vector.

This reduces the formula for ρ to:

$$\rho = \frac{(p^T p)^2}{(p^T G^{-1/2} L G^{-1/2} p)(p^T G^{1/2} L^{-1} G^{1/2} p)} \tag{10}$$

and, defining A by:

$$A \equiv G^{-1/2} L G^{-1/2} \tag{11}$$

Eq. (10) becomes:

$$\rho = \frac{(p^T p)^2}{(p^T A p)(p^T A^{-1} p)}\,, \tag{12}$$

We are now in a position to apply the Kantorovich inequality [3] to ρ^{-1}:

$$1 \le \rho^{-1} \le \tfrac{1}{4}\{(\alpha_1/\alpha_N)^{1/2} + (\alpha_N/\alpha_1)^{1/2}\}^2\,, \tag{13}$$

where α_1 and α_N are the smallest and largest eigenvalues of A, respectively. Defining $\theta \equiv \alpha_N/\alpha_1$, and rearranging (13), we obtain:

$$4\theta/(1 + \theta)^2 \le \rho \le 1\,, \tag{14}$$

Since ρ is no greater than unity, we see that the Newton step is most efficient (if g is arbitrary). For other choices of G than L, it is clear that the condition of A (expressed as the ratio θ) [9] is critical. The better conditioned A can be made, the more efficient will be the procedure. This result corresponds to that of Crockett and Chernoff [2] (expressed in a somewhat different form).

One of the major drawbacks to the use of Newton's method, is the fact that, unless L is positive definite, there is no assurance of a positive Δf_N, i.e., the quadratic form in equation (7) might have a negative value. Clearly, it is necessary to replace L in equation (7) in some fashion so as to guarantee an increase in f. The alternative is to abandon the Newton method in this case.

In what follows, we shall present a heuristic argument which indicates that, to some extent, the advantages of the Newton step direction may be retained, at the same time guaranteeing a rise in f (at least initially).

We must now assume that $f(x + hs)$ is no longer quadratic in h, but contains higher derivatives in its Taylor expansion. This implies a change in the spectrum of L as x moves from point to point. However, the directional derivative of f at x is still given by (cf. equation (2)):

$$\dot{f} \equiv (df/dh)_{h=0} = g^T s \tag{15}$$

and the Newton step is given by:

$$\delta\ (\equiv hs) = L^{-1} g \tag{16}$$

even if L is not positive definite. Hence, we have, combining equations (15) and (16):

$$\dot{f} = (g^T L^{-1} g)/h \tag{17}$$

We now analyze this quadratic form by referring all quantities to the normalized eigenvectors of L, defined by:

$$L\xi_k = \lambda_k \xi_k\,,\ \xi_k^T \xi_k = 1\ ; \qquad k = 1, \cdots, N\,, \tag{18}$$

where $\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$ are the eigenvalues of L in ascending order, and ξ_k are the corresponding eigenvectors. Since L is symmetric, the latter form a complete system. Hence, we have [4]:

$$L = \sum_{k=1}^{N} \lambda_k \xi_k \xi_k^T\,, \tag{19}$$

$$g = \sum_{k=1}^{N} \gamma_k \xi_k\,, \tag{20}$$

and $\dot{f}$ becomes:

$$\dot{f} = \frac{1}{h} \sum_k \lambda_k^{-1} \gamma_k^2\,. \tag{21}$$

In the very difficult problems, in which most gradient sequences involve rapid oscillations of direction across a "ridge" with very little overall progress ("hemstitching"), one may attribute this to the "ill-condition" of L in the absence of the appropriate choice of G. Eq. (14) indicates the very slow convergence one may expect in this case if the minimum ρ should occur. For simplicity, we shall assume initially that λ_1 is positive but very much smaller than $\lambda_2, \lambda_3, \cdots, \lambda_N$. Thus:

$$0 < \lambda_1 \ll \lambda_2 < \lambda_3 < \cdots < \lambda_N\,. \tag{22}$$

(Under these circumstances, the ridge would be in the direction of ξ_1.)

Then, making the reasonable assumption that $|\gamma_1|$ is not minute compared to

$|\gamma_2|$, $|\gamma_3|$, etc., we have, for f:

$$f \cong (1/h)\gamma_1^2/\lambda_1 \tag{23}$$

and also,

$$s \cong (1/h)\,(1/\lambda_1)\,\xi_1 \tag{24}$$

which shows that the Newton step direction is *along* the ridge, even if x is not situated exactly *on* the ridge.

Now we shall move the initial point x roughly parallel to the ridge, away from the maximum. As we do this, the values of g and L will change. We may regard the (small) change in L as a perturbation on L, and examine the effect on its eigenvalues and eigenvectors. Let the change be expressed by:

$$L = L_0 + \epsilon M \tag{25}$$

where M is of the same order of magnitude as L_0, and ϵ is very small. The new eigenvalues and eigenvectors of L are denoted again by λ and ξ. From first-order perturbation theory [5], we then have:

$$\lambda_k = \lambda_{0k} + \lambda_{1k}\,, \tag{26}$$

$$\xi_k = \xi_{0k} + \xi_{1k}\,, \tag{27}$$

with

$$\lambda_{1k} = \xi_{0k}^T\, M\xi_{0k}\ , \tag{28}$$

$$\xi_{1k} = -\sum_{m\neq k} \xi_{0m} \frac{(\xi_{0m}^T\, M\xi_{1k})}{\lambda_{0m} - \lambda_{0k}}\,, \tag{29}$$

These formulas show that the first-order changes in λ_k and ξ_k are of order ϵ, provided that the eigenvalues are well separated. In particular:

$$\lambda_{11} = \xi_{01}^T\, M\xi_{01} \tag{30}$$

$$\xi_{11} = -\sum_{m\neq 1} \xi_{0m} \frac{(\xi_{0m} M^T \xi_{01})}{\lambda_{0m} - \lambda_{01}}\ ; \qquad (\xi_1 = \xi_{01} + \epsilon\xi_{11})\ , \tag{31}$$

where we note that λ_{01} is, by hypothesis, well separated from the other eigenvalues; hence, no denominator is small.

We shall assume that the change in g is not such as accidentally to render γ_1 very much smaller than the other γ's.

Let us now consider that x has been moved from its initial position (where $\gamma_1 > 0$) sufficiently far from the domain of positive-definiteness of L, that the change in L (namely ϵM) is sufficient to have shifted λ_1 to being negative. However, since we assume a magnitude for the perturbation ϵM only sufficient to do this, and since λ_1 is very much smaller than all other λ's we may reasonably suppose that for "ordinary" surfaces, all the other λ's are still large and positive, as suggested by equation (26).* Hence, we may again approximate f as before:

* It would be possible to prove these suppositions, if one placed suitable restrictions on M, etc., but these would be equivalent to corresponding restrictive assumptions about the class of surfaces being discussed.

$$\dot{f} \cong (1/h)\gamma_1^2/\lambda_1 \tag{32}$$

compared with

$$\dot{f}_0 \cong (1/h_0)(\gamma_{01}^2/\lambda_{01}) \tag{23'}$$

the difference being that $\dot{f}_0 > 0$, while $\dot{f} < 0$.

Further, the estimated step directions are:

$$s = (1/h)\cdot(1/\lambda_1)\cdot\xi_1 \tag{33}$$

as compared with

$$s_0 = (1/h_0)\,(1/\lambda_{01})\,\xi_{01}\,. \tag{34}$$

Now, since we have shown (Eq. (31)) that ξ_1 differs from ξ_{01} only by a small (vectorial) amount, (33) and (34) show that the direction of the step has been substantially *reversed,* even though *it should not have been.*

This typical situation shows us, in effect, what may be done to repair the defect in Newton's method. We simply *force* all eigenvalues of L to be *positive.* The exact directional derivative (as given in Eq. (21)) then becomes:

$$\dot{f} = \frac{1}{h}\sum_k |\lambda_k|^{-1}\gamma_k^2 \tag{35}$$

which is clearly positive.

This simple procedure for determining s, which has worked very well in practice [6] for extremely ill-conditioned problems, may be summarized as follows:

(a) Analyze L into eigenvalues and eigenvectors (Jacobi's method was used in the practical computations). The result is Eq. (19).

(b) Set all eigenvalues positive, and take for G the matrix L^* defined by:

$$L^* \equiv \sum_k |\lambda_k|\xi_k\xi_k^T\,. \tag{36}$$

(c) Calculate an unnormalized direction from:

$$s^* = (L^*)^{-1}g = \sum_k |\lambda_k|^{-1}\gamma_k\xi_k\,. \tag{37}$$

(d) Use an extrapolation-interpolation search procedure to find the point of maximum f along s^*, since formula (16) no longer applies, even approximately.

It can be worked out, for cubic $f(x)$, that a small cubic term, acting as a perturbation on a quadratic, may easily disturb the maximum point only slightly, while inducing the change of step direction indicated above. This suggests that if one followed a direction s^*, which was substantially the same as the undisturbed direction s_0, the search procedure referred to would lead to substantially the same gain in f. Since the gain in f by the steepest ascent method ($G = I$) would be, from equation (6) (in terms of ξ_k-coordinates):

$$\Delta f_{SA} = \frac{\frac{1}{2}\sum_k \gamma_k^2}{\sum_k \lambda_k\gamma_k^2} \tag{38}$$

and since the λ_k are perturbed very slightly, as shown before, the proportional change in $(\Delta f)_{SA}$ would be negligible, regardless of the sign of λ_1. These arguments indicate that the lower bound of relative efficiency given in Eq. (14) is still more or less in force, suggesting in turn, that the modified Newton method here suggested is also (very nearly) optimally efficient.

The two major objections to this method are:

(a) It is necessary to compute second derivatives.

(b) It is necessary to do a complete eigenvalue-eigenvector analysis of L.

These objections must be taken seriously, to the extent that the modified Newton method might be rather inefficient (in the sense of computer time) in mildly ill-conditioned problems, as compared with a more straightforward gradient method. However, on the basis of the efficiency estimate in Eq. (14), if the value of θ is, e.g., 10^4, it would have to take roughly 10^4 as much computation per step for the modified Newton method as for the simpler method to make the latter faster.

This certainly suggests that Newton's method is worth considering in ill-conditioned cases, where a gradient method hemstitches badly.

The method of Davidon [7], [8], appears to be a successful attempt to combine the best of both worlds, by constructing L from values of g at different points along the path to the maximum (or minimum). It should be noted, however, that in very ill-conditioned problems, an accurate direction s is quite critical, so that an approximation to L might not suffice for high efficiency.

Appendix. We wish to maximize the directional derivative of f, viz.,

(A-1) $$\dot{f}_0 \equiv (df/dh)_{h=0} = g^T s$$

subject to:

(A-2) $$s^T G s = 1 .$$

We use Lagrange's method of multipliers to form:

(A-3) $$\Phi = g^T s - \lambda(s^T G s - 1)$$

and seek a stationary value of Φ:

(A-4a) $$\partial\Phi/\partial s = g - 2\lambda G s = 0 ,$$

(A-4b) $$\partial\Phi/\partial\lambda = s^T G s - 1 = 0 .$$

From (A-4a), we have:

(A-5) $$s = G^{-1} g/2\lambda .$$

Substituting this into (A-4b), we obtain:

(A-6) $$(g^T G^{-1}/2\lambda) G (G^{-1} g/2\lambda) = g^T G^{-1} g/4\lambda^2 = 1$$

so that

(A-7) $$2\lambda = (g^T G^{-1} g)^{1/2} .$$

Hence,

(A-8) $$s = G^{-1} g/(g^T G^{-1} g)^{1/2}$$

which coincides with Eq. (3). The foregoing analysis differs from that of Crockett and Chernoff somewhat, but the result is the same. The same applies to the remarks to follow.

The gradient $g(x + hs)$ is given by:

$$g(x + hs) = g(x) - hLs\,. \tag{A-9}$$

If we denote the new value of g by g^*, we may write:

$$g^* = g - hLs\,. \tag{A-10}$$

Replacing s according to (A-8), we have:

$$g^* = g - (h/(g^TG^{-1}g)^{1/2})LG^{-1}g = (I - \rho LG^{-1})g\,, \tag{A-11}$$

where

$$\rho \equiv h/(g^TG^{-1}g)^{1/2}\,. \tag{A-12}$$

While we are finding the maximum point iteratively, we are trying to make g vanish. We may thus think of (A-11) as an iterative formula:

$$g_{k+1} = (I - \rho_k LG^{-1})g_k \equiv Ag_k\,. \tag{A-13}$$

From Schwarz' inequality, we have:

$$||g_{k+1}|| \leq ||I - \rho_k LG^{-1}||\ ||g_k|| = ||A||\ ||g_k|| \tag{A-14}$$

and we hope that the norm of A is less than unity.

As Crockett and Chernoff point out, if L and G are both symmetric positive definite, then LG^{-1} has positive eigenvalues $\{\beta_i\}$, such that:

$$0 < \beta_1 < \beta_2 < \cdots < \beta_N\,. \tag{A-15}$$

Thus, if we take for the norm of A the following:

$$\nu_A = \max_{1\leq i\leq N} |1 - \rho_k\beta_i, \tag{A-16}$$

we wish to choose ρ_k so as to minimize ν_A.

It can be shown that the correct value of ρ_k is:

$$\rho_k = 2/(\beta_1 + \beta_N) \tag{A-17}$$

from which we obtain the Crockett and Chernoff result:

$$\nu_A = \left|\frac{1 - \beta_1/\beta_N}{1 + \beta_1/\beta_N}\right|. \tag{A-18}$$

This shows that the rate of convergence, which depends directly on ν_A, has a functional dependence on the condition number of LG^{-1}. To maximize the rate of convergence, it is clear that $G = L$ is the best choice.

International Business Machines Corporation
New York Scientific Center
New York, New York 10022

1. H. A. Spang, III, "A review of minimization techniques for nonlinear functions," *SIAM Rev.*, v. 4, 1962, pp. 343–365. MR **26** #3171.

2. J. B. Crockett & H. Chernoff, "Gradient methods of maximization," *Pacific J. Math.*, v. 5, 1955, pp. 33–50. MR **17,** 790.

3. M. Marcus & H. Minc, *A Survey of Matrix Theory and Matrix Inequalities,* Allyn and Bacon, Boston, Mass., 1964, p. 117. MR **29** #112.

4. E. Bodewig, *Matrix Calculus,* North-Holland, Amsterdam, 1956, p. 45 et seq. MR **18,** 235.

5. R. Courant & D. Hilbert, *Methods of Mathematical Physics,* Vol. I, Interscience, New York, 1953, p. 343 et seq. MR **16,** 426.

6. H. Eisenpress & J. Greenstadt, "The estimation of non-linear econometric systems," *Econometrica* (To appear.)

7. W. C. Davidon, *Variable Metric Method for Minimization,* AEC Research and Development Report ANL-5990, 1959.

8. R. Fletcher & M. J. D. Powell, "A rapidly convergent descent method for minimization," *Comput. J.*, v. 6, 1963/1964, pp. 163–168. MR **27** #2096.

9. D. K. Faddeev & V. N. Faddeeva, *Computational Methods of Linear Algebra,* Freeman, San Francisco, Calif., 1963, p. 126. MR **28** #1742.