—NOTES—

THE METHOD OF STEEPEST DESCENT FOR NON-LINEAR MINIMIZATION PROBLEMS*

By HASKELL B. CURRY (*Frankford Arsenal*)

1. Introduction. The problem considered here is that of minimizing a function of n real variables, $G(x_1, \cdots, x_n)$. The object is to find a practical method for evaluating, approximately at least, a stationary point for G.

This problem includes as a special case that of solving a set of simultaneous equations

$$f_i(x_1, \cdots, x_n) = 0 \qquad (i = 1, 2, \cdots, m), \tag{1}$$

because the function

$$G(x_1, \cdots, x_n) = \sum_{k=1}^{m} f_k^2 \tag{2}$$

has a minimum at a solution of (1). It also includes that of determining the parameters $x_1, \cdots, x_n$ of a function $f(u; x_1, \cdots, x_n)$ so as to get the best approximation, in a least square sense, to a function $F(u)$ for certain values of u; the G in this case is of the form given by

$$G(x_1, \cdots, x_n) = \sum_{k=1}^{p} [F(u_k) - f(u_k; x_1, \cdots, x_n)]^2. \tag{3}$$

Certain engineering applications of the latter sort of problem arose in the work of the Engineering Research Section, Fire Control Design Division, at Frankford Arsenal. In these applications, the function $f(u; x_1, \cdots, x_n)$ was sufficiently complicated so that the standard method for dealing with non linear least square problems[1] failed to converge. Two techniques for dealing with this situation were developed by the section under the direction of J. G. Tappert. One of these was an original suggestion of my associate K. Levenberg.[2] The second method is the subject of this note.

This method is not new. Levenberg found it set forth in a paper by Cauchy dated 1847.[3] That it has become a standard procedure in analysis is clear from a recent paper by Courant.[4] Nevertheless it does not appear to be well known to authorities on nu-

---

[1] See, for example, W. E. Deming, *Some notes on least squares*, U. S. Dept. of Agriculture Graduate School, 1938, p. 31 ff., or E. T. Whittaker and G. Robinson, *The calculus of observations*, Blackie and Son, London, 1940, p. 214. Deming's treatment is also given in his book, *Statistical adjustment of data*, John Wiley & Sons, New York, 1943, p. 52 ff.

[2] K. Levenberg, *A method for the solution of certain non-linear problems in least squares*, Quarterly of Applied Mathematics, **2**, 164 (1944).

[3] A. L. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées*, Comptes rendus, Ac. Sci. Paris, **25**, 536–538 (1847).

[4] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc. **49**, 1–23 (1943). See especially pp. 17–20. Courant calls the method the "method of gradients" and ascribes its origin to a paper published by Hadamard in 1907.

merical computation. It was used for the case of linear equations by G. Temple;[5] but he gave no reference to Cauchy's work nor indeed to any previous use of the method. Accordingly there is room for an exposition of the method with emphasis on its practical aspects.

This note also contains an outline of a convergence proof. Cauchy stated that the process converged but gave no proof, at least in the paper cited. Temple's convergence proof applied only to the linear case. Courant (l.c.) gives references to papers dealing with the method; but some of these were not accessible to me under wartime conditions. The convergence proof, as outlined here, is an elementary one and gives a weak result.

The argument is, incidentally, capable of generalization to certain cases where there are infinitely many parameters, i.e., where G is a function of a vector $\mathbf{x}$ belonging to a suitable abstract space. The essential point is that there be a vector function $\mathbf{H}(\mathbf{x})$, the gradient, such that for vectors $\mathbf{x}$, $\mathbf{y}$ and scalar t

$$\frac{d}{dt}\left[G(\mathbf{x}+t\mathbf{y})\right] = \mathbf{H}(\mathbf{x}+t\mathbf{y})\cdot\mathbf{y},$$

where the dot indicates a scalar product.[6] Such generalizations will not be considered explicitly.

2. Explanation of the method. The letters $\mathbf{x}$, $\mathbf{z}$ will be used to stand for the n-tuples (vectors) $(x_1, \cdots, x_n)$ and $(z_1, \cdots, z_n)$ respectively. It will be convenient also to think of the vector $\mathbf{x}$ as a point and $\mathbf{z}$ as a set of direction numbers of a direction, viz. the direction $\mathbf{z}$, emanating from $\mathbf{x}$. Superscripts will be used systematically to distinguish different points and their corresponding directions.

Let us suppose, then, that we start at a point $\mathbf{x}^0$ and determine the direction in which G decreases most rapidly. This direction is given by $z_i = -\lambda\, \partial G/\partial x_i$ or, in vector form, $\mathbf{z}^0 = -\lambda\ \mathrm{grad}\ G$, where λ is an arbitrary positive factor of proportionality. (In practice we should either take $\lambda = 1$ or choose λ so that the vector $\mathbf{z}$ is of unit length.) Then the function $g(t) = G(\mathbf{x}^0 + t\mathbf{z}^0)$ has a negative derivative at $t = 0$. It will therefore be possible to find a $t > 0$ such that

$$g(t) < g(0). \tag{4}$$

With such a t we can take $\mathbf{x}^1 = \mathbf{x}^0 + t\mathbf{z}^0$ as a new starting point and continue. We should then have a sequence of points $\mathbf{x}^0$, $\mathbf{x}^1$, $\mathbf{x}^2$, $\cdots$ such that $G(\mathbf{x}^{k+1}) < G(\mathbf{x}^k)$. Under suitable restrictions (to be considered later) the sequence will attain or converge to a stationary point of G.

The determination of t can be accomplished by trial. If no other indication is availble we can take as first trial value the intercept of the tangent to the curve $y = g(t)$ on the t-axis.[7] If this fails to satisfy (4), it is too large; we can then take half of it, and so on. In this process we can draw a rough graph of $g(t)$, and after a few trials it is

[5] G. Temple, *The general theory of relaxation methods applied to linear systems*, Proc. Roy. Soc. London (A) **169**, 476–500 (1939). For this reference I am indebted to H. Hotelling.

[6] Even the case where such a gradient does not exist, there being only a total differential, can probably be handled by a method which bears the same relation to the present method that Temple's method for gyrostatic systems does to his method of steepest descent.

[7] This was done by Cauchy (l.c.). It represents the approximation by Newton's method to the smallest positive zero of $g(t)$. This is a reasonable guess for a G given by (2), where the numerical value is zero. In the least square cases (where $G \geqq 0$) the guess is often many times too large.

usually possible to locate a t which is at or near a minimum of $g(t)$. Experience will presumably disclose many ways to shorten the process in individual cases.

If we take for t precisely the smallest positive root of

$$g'(t) = 0, \tag{5}$$

the process has the following geometrical interpretation. Starting at $\mathbf{x}^0$, we determine the direction in which the surface

$$y = G(\mathbf{x}) = G(x_1, \cdots, x_n) \tag{6}$$

is descending most rapidly. We continue in that direction until we find ourselves going along a contour (i.e. a horizontal section of the surface). Then we stop, take a new direction of steepest descent, and so continue. Since the direction of steepest descent is always normal to the contour it follows that the directions $\mathbf{z}^k$ and $\mathbf{z}^{k+1}$ are at right angles.[8] This is important in the convergence proof.

3. Proof of Convergence. Let us suppose now that $G(x_1, \cdots, x_n)$ is defined and has continuous first partial derivatives at all points within or on the boundary of a region S. Let $\mathbf{x}^0$ be a point within S. Let C be the broken line path starting at $\mathbf{x}^0$ and going in the direction of steepest descent at $\mathbf{x}^0$ until it reaches either the boundary of S or the next approximation $\mathbf{x}^1$ determined as in §3 with t the least positive root of (5); in the latter case the broken line goes in the direction of steepest descent at $\mathbf{x}^1$ until it reaches the boundary of S or the next approximation $\mathbf{x}^2$ determined in the same way; and so on. Then G is monotone decreasing along C. There are three possibilities: (1) The path C may run into the boundary of S. (2) The path C may terminate at a point where the direction of steepest descent does not exist, i.e. at a stationary point of G. (3) The process may continue indefinitely. The first possibility will certainly be excluded if the value of G at $\mathbf{x}^0$ is less than at any point on the boundary of S. If the second possibility occurs the case is trivial. I shall make the limitation just stated in regard to $G(\mathbf{x}^0)$ and shall suppose that the process continues indefinitely.

Under these presuppositions let $\mathbf{x}^\infty$ be a limit point of $\mathbf{x}^0, \mathbf{x}^1, \cdots$. Then it is clear that

$$G(\mathbf{x}^\infty) < G(\mathbf{x}^k) \qquad (k = 0, 1, 2, \cdots). \tag{7}$$

It will now be shown that $\mathbf{x}^\infty$ is a stationary point of G.

Let us suppose the contrary. We write $\mathbf{H}(\mathbf{x}) = \operatorname{grad} G$, $h(\mathbf{x}) = |\mathbf{H}(\mathbf{x})|$, and let $\mathbf{z}(\mathbf{x})$ be a unit vector; thus

$$\mathbf{H}(\mathbf{x}) = - h(\mathbf{x})\mathbf{z}(\mathbf{x}). \tag{8}$$

According to the supposition, $h(\mathbf{x}^\infty) \neq 0$. Hence it will be possible to find a spherical neighborhood U of $\mathbf{x}^\infty$ such that for $\mathbf{x}$ in U,

$$|\mathbf{H}(\mathbf{x}) - \mathbf{H}(\mathbf{x}^\infty)| < \epsilon h(\mathbf{x}^\infty).$$

Then it will follow that $|h(\mathbf{x}) - h(\mathbf{x}^\infty)| < \epsilon h(\mathbf{x}^\infty)$, $|\mathbf{z} - \mathbf{z}^\infty| < 2\epsilon$. Hence, from the fact that $\mathbf{z}^k$ and $\mathbf{z}^{k+1}$ are at right angles, one can conclude that if $\mathbf{x}^k$ is in U, $\mathbf{x}^{k+1}$ is certainly not in U (provided ϵ is not too large).

Next, let K be the conical sector of U for which

$$\cos \theta = \frac{\mathbf{x} - \mathbf{x}^\infty}{|\mathbf{x} - \mathbf{x}^\infty|} \cdot \mathbf{z}^\infty > \epsilon.$$

_____
[8] This is also easily proved analytically.

Then it may be shown, by reasonably straight-forward methods, that for $\mathbf{x}$ in K,

$$G(\mathbf{x}) < G(\mathbf{x}^\infty). \tag{9}$$

Let V be a subneighborhood of U such that for $\mathbf{x}$ in V the ray in the direction $\mathbf{z}$ (as given by (8)) from $\mathbf{x}$ intersects K. Such a V exists if $\epsilon < \frac{1}{3}$. Since $\mathbf{x}^\infty$ is a limit point, there exists an $\mathbf{x}^n$ in V. Then the ray in the direction $\mathbf{z}^n$ from $\mathbf{x}^n$ will have a point $\mathbf{y}$ in K. This $\mathbf{y}$ cannot be beyond $\mathbf{x}^{n+1}$, since $\mathbf{x}^{n+1}$ is not in U and U is convex. Hence, by the monotonic character of G on C and by (9), $G(\mathbf{x}^{k+1}) < G(\mathbf{y}) < G(\mathbf{x}^\infty)$, which contradicts (7). This contradiction came from the assumption that $h(\mathbf{x}^\infty) \neq 0$.

The following example shows that we cannot expect a better result without further restrictions on G. Let $G(x, y) = 0$ on the unit circle and $G(x, y) > 0$ elsewhere. Outside the unit circle let the surface have a spiral gully making infinitely many turns about the circle. Then the path C will evidently follow the gully and have all points of the unit circle as limit points.

In a practical problem however we often know in advance that there is a unique minimum of G within S; in these cases convergence is assured. If G is given by (2) and the Jacobian of the f's does not vanish in S, then every stationary point of G is a solution of (1); if there is only one such solution the process converges to it.

4. Concluding Remarks. In regard to the practical aspects of the method the following points are to be noted: (1) It does not require any calculation of second derivatives. This is important for the application mentioned, where these second derivatives are numerous and complicated. (2) It involves only direct calculations of G and its first derivatives. (3) The approach to the limit, if any, is along a path C consisting of straight line segments, adjacent segments being approximately at right angles.

A comparison with Levenberg's method in regard to these three respects is now in order. In the first respect the two methods are alike. In the second respect Levenberg's method is more complicated because each stage requires the solution of a set of "normal" equations, as in the traditional method of least squares. In the third respect Levenberg's method is like the present one in that it involves approach along a broken path C; but the individual pieces of C are curved. There is evidence that in practical problems these curves follow the natural valleys of the surface, so that each step brings us further toward the goal. As to which of these opposing characteristics is the more important is not yet settled.[9]

Another point is that the process is not invariant under certain elementary transformations, e.g., a change of scale of one or more of the x_i. Thus, if the surface (6) (for $n = 2$) is a hemispherical bowl, the direction of steepest descent is along the meridian, and the minimum is reached in one step. If, however, the scale on the x_i-axis is changed so that the surface becomes ellipsoidal, the direction of steepest descent is no longer directed toward the minimum. The suggestions inherent in this lack of invariance have not yet been fully worked out.

---

[9] The engineers at Frankford Arsenal prefer the Levenberg method for the problems which have confronted them; but I do not know to what extent they have exploited the present method. Since the Levenberg method is confined to a G of the form (3) and makes use of that representation, it would not be surprising if it should prove superior for that case. On the other hand it is not difficult to concoct artificial examples for which the method of steepest descent is superior in the third respect as well as the second, at least for certain determinations of the weighting factors.