IBM Research

EVALUATING A RESTART PROCEDURE FOR CONJUGATE GRADIENTS

M. F. McGuire/P. Wolfe

June 5, 1973

RC4382

Yorktown Heights, New York

San Jose, California

Zurich, Switzerland

EVALUATING A RESTART PROCEDURE FOR CONJUGATE GRADIENTS

by

Mary F. McGuire*
Philip Wolfe

Mathematical Sciences Department
IBM Watson Research Center
Yorktown Heights, New York

ABSTRACT: A recent promising variant on the use of the method of conjugate gradients for nonlinear optimization is studied both theoretically and, computationally, on a class of easily generated test problems. It turns out to give about the same results as standard methods do.

RC4382 (No. 19614)
June 5, 1973
Mathematics

* University of Alberta, Calgary, Alberta, Canada

1. INTRODUCTION

E. M. L. Beale recently made an interesting observation (communicated in his address "Conjugate Gradient Methods for Nonquadratic Problems" at the Conference on Nonlinear Optimization at Dundee, June 1971) on the method of conjugate gradients for optimization. His observation suggested a way in which the method as currently used for the minimization of nonquadratic functions might be improved; this note reports some observations on its use.

2. THE STANDARD CONJUGATE GRADIENT METHOD

The typical step of the conjugate gradient method may be described: given the point x_k in E^n, the gradient $g_k = \nabla f(x_k)$ of the differentiable function f at x_k, and the direction (a nonnull vector in E^n) d_k, define

$$x_{k+1} = x_k + t_k d_k , \tag{1}$$

where t_k is some value of t such that

$$\frac{d}{dt} f(x_k + t d_k) = \nabla f(x_k + td_k)^T d_k = 0 \tag{2}$$

and that $f(x_k + td_k)$ is decreasing from $t = 0$ to $t = t_k$. Define

$$d_{k+1} = -g_{k+1} + s_k d_k , \tag{3}$$

where s_k is chosen so that d_{k+1} is conjugate to d_k, which we here take to mean

$$d^T_{k+1} (g_{k+1} - g_k) = 0. \tag{4}$$

(This is one of the more useful relations reducing in the quadratic case $f(x) = p^T x + \frac{1}{2} x^T Qx$, whence $\nabla f(x) = p + Qx$ and hence, via

$$g_{k+1} - g_k = Q(x_{k+1} - x_k) = t_k Qd_k, \tag{5}$$

to the relation $d^T_{k+1} Q d_k$ as used by Hestenes and Stiefel [1] in their treatment of the quadratic case. The starting condition $d_0 > - g_0$ is standard and important.)

For nonquadratic problems it has been found valuable [2, 3, 4] to periodically restart the procedure: when $k+1 = K, 2K, \ldots$ (with $K \geq n$), to replace Formula 3 by just

$$d_{k+1} = -g_{k+1}. \tag{6}$$

The reason for this can be seen in the quadratic case: if the starting condition $d_0 = -g_0$ is not used, then the procedure does not terminate in the solution of the quadratic problem [3]: and it can be shown [4] that this behavior induces poor convergence in the use of the nonrestarted method on nonquadratic problems.

3. THE REVISED METHOD

One has, nevertheless, the feeling that there is some useful information in the direction d_{pK} $(p=1, 2, \ldots)$ which is being abandoned by the formula 6, and Beale's revision of the method is intended to preserve such information: formula 6 is replaced by formulas 3 and 4 for $k = K, 2K, \ldots,$ while

$$d_{k+1} = -g_{k+1} + s_k d_k + u_k d_{pK}$$

$$\text{for } k = pK+1, pK+2, \ldots, (p+1)K-1, \quad (7)$$

$$p = 1, 2, \ldots ,$$

where s_k and u_k are chosen so that d_{k+1} is conjugate to both d_k and d_{pK}, that is, so that both formula 4 and

$$d_{k+1}^T (g_{k+1} - g_K) = 0 \quad (8)$$

hold.

One can show (along the lines of the standard proof for the ordinary method [/] that this procedure has the desired termination properties: whatever be the vector d_{pK}, if f is quadratic then the iteration determined by formulas 2, 7, 4, 8 yields $g_k = 0$ for some $k \leq pK+n$. It should be firmly pointed out that the unmodified procedure does not have this property, as we have mentioned.

A possible troublesome point should be noted: in the ordinary algorithm (3), the direction d_{k+1} is assuredly a descent direction, since $d_{k+1}^T g_{k+1} = - g_k^2 < 0$, while it is not clear that this property should hold in the revised procedure for nonquadratics. Accordingly, we will replace d_{k+1} by $-d_{k+1}$ when necessary to have a descent direction. It is not clear whether this somewhat agricultural provision is sounder than just restarting at such a place; in any case, it seems to be a rare event.

4. COMMENTARY

In considering use of the revised conjugate gradient procedure we were encouraged by the observation that, to a certain extent, conjugation can't hurt. Specifically, it is not difficult to compare the decreases in function values obtained in taking steps of these two different kinds when the function is convex and quadratic:

(a) from x_k to the minimum in the direction $-g_k$ -- the ordinary steepest descent step;

(b) from x_k to the minimum in the direction d, where $d = -g_k + s\,c$, d is conjugate to c, and c is arbitrary excepting that $g_k^T c = 0$ (so that x_k is the minimum of f along the line through it in the direction c).

It can be shown [3] that the decrease obtained in step (b) is never less than that in (a), so that any step conjugate to a given direction is at least as good as a steepest descent step. Thus the step pK of the revised procedure, using Formula 3, is always at least as good as that of the ordinary restarted procedure, using Formula 6, in the quadratic case.

The revised procedure has, however, this drawback in the non-quadratic case: the fact that each d_{k+1} for $k = pK+1, \ldots, pK+K$ is conjugate to d_{pK} means that all points x for those values of k lie on the linear manifold

$$M = \{ x : d_{pK}^T (x - x_{pK}) = 0 \}. \tag{9}$$

While in the quadratic case the solution of the problem does lie on M, in the general case that cannot be expected; and to the extent to which it does not, the procedure is prevented from getting close to the solution.

We can make a crude estimate of the trouble this should cause. Assuming the problem to have the solution x^* and the Hessian of the function to be smooth--i.e., $|| \nabla^2 f(y) - \nabla^2 f(x)|| < L\,|y-x|$ for an appropriate matrix norm $||\cdot||$ --we can expect the direction d_k to differ from the "correct" direction (one for which M contains the solution) by a term of the order of $|x_k - x^*|$; and thus expect that for $x_{(p+1)K}$, which is crudely approximated by the closest point on the manfold to x^*, we have

$$|x_{(p+1)K} - x^*| \cong |x_{kP} - x^*|^2. \qquad (10)$$

Now the ordinary restarted conjugate gradient method gives quadratic convergence for suitable convex functions [5, 6], so the error introduced here, which is of the same kind, does not alter the type of convergence; but it does prevent us from doing any better. Anyway, even in the quadratic case the greatest benefit one could reasonably expect from a good choice of d_{pK} would be the reduction by one dimension of the manifold that one must search, which is not a great contribution.

We conclude from this discussion that the revised procedure should have about the same convergence properties as the ordinary restarted procedure. The next section presents the results of some numerical tests of this contention. In all fairness it should be admitted that the experiments came first; the drawbacks mentioned did not occur to us until our experimental results required explanation.

5. EXPERIMENTAL RESULTS

We have designed a particular kind of function for computational experiments in unconstrained optimization methods. We write

$$f(x) = \frac{1}{2} \Sigma_i Q_i x_i^2 + \frac{1}{3} \Sigma_{i,j,k} R_{ijk} x_i x_j x_k , \qquad (11)$$

where all indices run $1, \ldots, n$. The numbers Q_i are personally chosen, while for $i \leq j \leq k$ each R_{ijk} is chosen at random with uniform distribution from the interval $[-\epsilon, \epsilon]$, the remaining R_{ijk} being determined by complete symmetry--that is, that the value is independent of the ordering of the indices i, j, k.

We have the convenient formulas

$$[\nabla f(x)]_j = [Qx]_j + x^T R_j x ,$$
$$\nabla^2 f(x) = Q + 2 \Sigma_i R_i x_i , \qquad (12)$$

where Q is the n-by-n diagonal matrix for which $Q_{jj} = Q_j$ and R_i is the n-by-n matrix whose j, k element is R_{ijk}. The form (12) for f is reasonably general, since almost any cubic function of n variables may be set in that form by an orthogonal transformation (which diagonalizes the quadratic terms) followed by a translation (removing the linear terms--provided no quadratic term vanishes, which we suppose) and altered by a constant. Our function then has a stationary point at $x = 0$. It will be convex in a neighborhood of the origin if all $Q_i > 0$; indeed, the Gerschgorin circle theorem assure that $\nabla^2 f(x)$ will be positive definite if $Q_j > 2 \Sigma_k | [\Sigma_i R_i x_i]_{jk} |$, which will be true for all x satisfying $|x_i| < \Delta$, all i, if

$$\Delta = \frac{1}{2} \operatorname{Min}_j Q_j / \Sigma_{i,k} |R_{ijk}| . \tag{13}$$

In our experiments the starting point, chosen more or less at random, is required to satisfy the above condition quite strictly. The procedure, minimizing f, generates a sequence of points which generally also satisfy the condition, so the function remains convex in the region of interest. This form of problem has the advantage that the linear optimization problem (2) is easy to carry out, requiring usually the solution of a quadratic equation and a little testing for the precise determination of t_k. We find that the optimality requirement $d^T_{k+1} g_k = 0$ is satisfied to about ten decimal places, rather than to considerably less than one decimal place, as has occurred with some routines we have investigated. While of course such accuracy is hardly a feasible or even desirable goal when trying to solve serious problems efficiently, we think that it helps our work in eliminating one possibly troublesome experimental variable.

We performed our experiments using $Q_i = i$, $i = 1, 2, 3$, with three different choices of R and five choices of the starting x_0, running each problem with both the revised and the original procedure. The results were collectively indistinguishable.

Here is a typical problem: the components of R (lexicographically ordered for $i \leq j \leq k$) were -0.048, -0.100, -0.082, -0.170, -0.051, -0.193, 0.119, 0.098, 0.026, -0.040, so that Δ of formula 13 was 0.570. The starting point was (0.0069, 0.84, 0.0083). Here $k = 3$, so that restarting occured for $k = 3, 6, 9$; the procedure was terminated when

$f(x_k) < 10^{-50}$. (A couple of runs were also tried using $n = K = 6$, with generally the same results.)

k	$f(x_k)$ (standard procedure)	$f(x_k)$ (revised procedure)
0	7.290E¯1	7.290E¯1
1	3.854E¯3	3.854E¯3
2	7.618E¯4	7.618E¯4
3	1.348E¯10	2.753E¯5
4	5.655E¯12	2.753E¯5
5	1.039E¯13	1.820E¯7
6	5.036E¯25	3.039E¯12
7	1.090E¯26	3.006E¯12
8	4.003E¯28	1.266E¯13
9	2.795E¯30	2.129E¯27
10	2.189E¯32	1.027E¯27
11	1.356E¯36	1.336E¯30
12	6.756E¯60	1.587E¯56

REFERENCES

1. Hestenes, M. R. and E. Stiefel, "Methods of Conjugate Gradients for Solving Linear Systems", J. Research National Bureau of Standards 49, 409-436 (1952).

2. Fletcher, R. and C. M. Reeves, "Function Minimization by Conjugate Gradients". The Computer Journal 7, 149-156 (1964).

3. Crowder, Harlan and Philip Wolfe, "Linear Convergence of the Conjugate Gradient Method". IBM J. Research and Development 16, 431-433 (1972).

4. Philip Wolfe, "Convergence of Continued Conjugate Gradient Methods". IBM Research Center Report, in preparation.

5. Cohen, Arthur I., "Rate of Convergence of Several Conjugate Gradient Algorithms". Undated manuscript, Dept. of Electrical Engineering, University of California, Berkeley.

6. McCormick, Garth P. and Klaus Ritter, "On the Convergence and Rate of Convergence of the Conjugate Gradient Method". MRC Technical Summary Report No. 1118, June, 1971, Mathematics Research Center, The University of Wisconsin.