ANL-AMD-TM-282 (Rev.)

A Relationship between the BFGS and Conjugate Gradient Algorithms

Larry Nazareth

MASTER

U of C-AUA USERDA

ARGONNE NATIONAL LABORATORY, ARGONNE, ILLINOIS

operated under contract W-31-109-Eng-38 for the

U. S. ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATIO

DISCLAIMER

The facilities of Argonne National Laboratory are owned by the United States Government. Under the terms of a contract (W-31-109-Eng-38) between the U. S. Energy Research and Development Administration, Argonne Universities Association and The University of Chicago, the University employs the staff and operates the Laboratory in accordance with policies and programs formulated, approved and reviewed by the Association.

ANL-AMD-TM-282 (Rev.)

Applied Mathematics Division
Argonne National Laboratory
Argonne, Illinois 60439

A Relationship between the BFGS and Conjugate Gradient Algorithms*

by

Larry Nazareth

This report was prepared primarily for internal distribution.

*Work performed under the auspices of the U.S. Energy Research and Development Administration. This paper constitutes a revised version of Tech. Memo 282. It may be referenced as ANL-AMD Tech. Memo 282 (Rev.), Applied Mathematics Division, Argonne National Laboratory (1977).

A Relationship between the BFGS and Conjugate Gradient Algorithms

by

Larry Nazareth

Abstract

Based upon analysis and numerical experience, the BFGS algorithm is currently considered to be one of the most effective algorithms for finding a minimum of an unconstrained function, $f(x)$, $x \in \mathbb{R}^n$. However, when computer storage is at a premium, the usual alternative is to use a Conjugate Gradient (CG) method. In this paper we show that the two algorithms are related to one another in a particularly close way. Based upon these observations a new family of algorithms is proposed.

1. Introduction

We first give a concise statement of the algorithms under consideration and summarize briefly some of their well known properties. We then show, in Section 2, an exact correspondence between the search vectors developed by the BFGS and CG algorithms, when applied to quadratic functions. For arbitrary differentiable functions we give an interpretation of the BFGS algorithm as a CG algorithm with variable metric, chosen at each step from the Broyden β-class. These observations then lead us to a family of algorithms termed Variable Storage Generalized Conjugate Gradient Methods, introduced in Section 4.

The Conjugate Gradient Method [1] in a fixed metric defined by the positive definite symmetric matrix H and started from a given point x_1, develops successive search directions d_j^{CG}, iterates x_j and gradients $g_j = \nabla f(x_j)$ as follows:

$$d_1^{CG} = -Hg_1$$

$$d_j^{CG} = -Hg_j + \left[\frac{y_{j-1}^T Hg_j}{y_{j-1}^T d_{j-1}^{CG}}\right] d_{j-1}^{CG} \qquad j > 1 \tag{1}$$

$$x_{j+1} = x_j + \lambda_j d_j^{CG}, \quad \text{where} \quad \lambda_j = \min_{\lambda} f(x_j + \lambda d_j^{CG})$$

and $y_{j-1} \triangleq (g_j - g_{j-1})$.

The above follows the Hestenes-Stiefel formulation originally proposed for solving linear systems and extended to non-linear optimization by Fletcher & Reeves [2]. Various formulations of the algorithm (see [2],[3]) are equivalent when applied to quadratic functions, but differ for arbitrary functions. In the basic form of the CG algorithm, H is set to the identity, and it is well known that using an arbitrary positive definite symmetric matrix H corresponds to applying the change of variables $y = H^{-\frac{1}{2}}x$ to the basic algorithm.

<u>Variable Metric Methods</u> [4],[5] in Broyden's β-class, started with a positive definite symmetric matrix H and initial point x_1, develop successive positive definite and symmetric approximations H_j^β to the inverse Hessian, successive search direction d_j^β, and iterates x_j as follows:

$$H_1^\beta = H$$

$$H_j^\beta = H_{j-1}^\beta - \frac{H_{j-1}^\beta y_{j-1} y_{j-1}^T H_{j-1}}{y_{j-1}^T H_{j-1}^\beta y_{j-1}} + \frac{s_{j-1} s_{j-1}^T}{s_{j-1}^T y_{j-1}} \tag{2}$$

$$+ \beta_{j-1}\left(H_{j-1}^\beta y_{j-1} - \theta_{j-1}^\beta s_{j-1}\right)\left(H_{j-1}^\beta y_{j-1} - \theta_{j-1}^\beta s_{j-1}\right)^T \qquad j > 1$$

where $\beta_{j-1} \geq 0$

$$\theta^{\beta}_{j-1} \triangleq y^T_{j-1} H^{\beta}_{j-1} y_{j-1} / s^T_{j-1} y_{j-1}$$

$$s_{j-1} \triangleq (x_j - x_{j-1})$$

$$d^{\beta}_j = -H^{\beta}_j g_j$$

and $x_{j+1} = x_j + \lambda_j d^{\beta}_j$ with $\lambda_j = \min_{\lambda} f(x_j + \lambda d^{\beta}_j)$.

Particular cases are given by $\beta_{j-1} = 0$ (DFP), and $\beta_{j-1} = 1/y^T_{j-1} H^{\beta}_{j-1} y_{j-1}$ (BFGS). In the latter case (2) simplifies to

$$H^{BFGS}_j = H^{BFGS}_{j-1} + \frac{1}{s^T_{j-1} y_{j-1}} \left[1 + \frac{y^T_{j-1} H^{BFGS}_{j-1} y_{j-1}}{s^T_{j-1} y_{j-1}} \right] s_{j-1} s^T_{j-1}$$

$$- \frac{1}{s^T_{j-1} y_{j-1}} \left(s_{j-1} y^T_{j-1} H^{BFGS}_{j-1} + H^{BFGS}_{j-1} y_{j-1} s^T_{j-1} \right) \qquad (3)$$

$$d^{BFGS}_j = -H^{BFGS}_j g_j \ .$$

The following properties of the conjugate gradient and variable metric algorithms are well known.

When applied to quadratic functions $\psi(x) = a + b^T x + \frac{1}{2} x^T A x$, with A positive definite and symmetric ($A > 0$), we have (i) termination in at most n steps, (ii) search vectors are conjugate, (iii) $g^T_i H g_j = 0$, $i \neq j$, (iv) the j'th direction lies in the subspace spanned by $Hg_1, \ldots, Hg_j$, (v) since there is no flexibility in choice of directions given the above conditions, d_j^{cg} and d_j^{β} must be linearly dependent, (vi) $H_j^{\beta}A(d_1^{\beta},\ldots,d_{j-1}^{\beta}) = (d_1^{\beta},\ldots,d_{j-1}^{\beta})$, (vii) provided premature termination does not occur $H_{n+1}^{\beta} = A^{-1}$.

Furthermore, Dixon's Theorem demonstrates that for quite general continuously differentiable objective functions, the corresponding search directions $d_j^{\beta'}$ and $d_j^{\beta''}$ developed by any two members of Broyden's β-class (using the same starting point x_1 and initial approximation H) are linearly dependent and successive iterates are identical, when line searches are exact and unambiguously defined.

2. A Result for Quadratics

We now strengthen property (v) of Section 1 to show that for one member of the β-class (the BFGS update), the search vectors d_j^{CG} and d_j^{BFGS} are precisely the same. This correspondence is, we feel, indicative of underlying structure, and is developed further in Section 3, for arbitrary functions.

Lemma: When the CG and BFGS algorithms are applied to a quadratic function $\psi(x) = a + b^Tx + \frac{1}{2}x^TAx$, $A > 0$, using the same starting point x_1 and positive definite symmetric H, then

$$d_j^{CG} = d_j^{BFGS} \qquad j=1,2,\ldots,n \,.$$

Proof

Fact 1: $g_{j+1}^T s_j = 0$. Further a well known result is that

$$g_k^T s_j = 0, \quad k > j \,.$$

Fact 2: $H_j^{BFGS} g_k = Hg_k \qquad j < k \le n+1, \quad 1 \le j \le n \,.$

□Proof of Fact 2: This may be shown by induction on j.

Assume true for H_{j-1}^{BFGS}, i.e.,

$$H_{j-1}^{BFGS} g_k = Hg_k, \quad j-1 < k \leq n \quad .$$

Now combining (3), Fact 1 above, property (iii) of Section 1, and the induction hypothesis we have

$$H_j^{BFGS} g_k = H_{j-1}^{BFGS} g_k = Hg_k \qquad j < k \leq n \ .$$

Since $H_1 g_k = Hg_k$ for $1 \leq k \leq n$, the result follows by induction.□

Returning to the proof of Lemma 1, we have

$$d_j^{BFGS} = -H_j^{BFGS} g_j \ .$$

Using (3) and the fact that line searches are exact, we have

$$d_j^{BFGS} = -H_{j-1}^{BFGS} g_j + \left[\frac{y_{j-1}^T H_{j-1}^{BFGS} g_j}{y_{j-1}^T d_{j-1}^{BFGS}} \right] d_{j-1}^{BFGS}$$

Now from Facts 1 and 2 above, and property (iii) of Section 1, this gives

$$d_j^{BFGS} = -Hg_j + \left[\frac{y_{j-1}^T Hg_j}{y_{j-1}^T d_{j-1}^{BFGS}} \right] d_{j-1}^{BFGS}$$

$$= -Hg_j + \frac{y_{j-1}^T Hg_j}{y_{j-1}^T d_{j-1}^{CG}} d_{j-1}^{CG} \quad \text{using property (v) of Section 1.}$$

$$= d_j^{CG} \quad \text{using (1), and this is the desired result.}$$

3. Interpretation of the BFGS Algorithm for Arbitrary Differentiable Functions

We employ the following theorem due to Powell. This is pharaphrased below, and for the proof we refer the reader to [6].

Theorem: Let the variable metric method of Section 1 be applied to a differentiable function $f(x)$, and assume that all line searches are exact and that the λ_j are chosen unambiguously. Let $x_1,\ldots,x_j$ be the sequence of iterates and $H_1^\beta,\ldots,H_{j-1}^\beta$ the sequence of matrices developed prior to the j'th iteration, and assume that no search vector d_j^β vanishes. Then, if the choice of β corresponding to the BFGS update is used at iteration j, the matrix H_j^{BFGS} obtained is independent of the parameter values β used during previous iterations.

Invoking this theorem, setting $\beta_{j-1} = 1/y_{j-1}^T H_{j-1}^\beta y_{j-1}$ in (2), and using the fact that line searches are exact, we can state the BFGS algorithm as follows:

$$d_1^{BFGS} = -H_1 g_1$$

$$d_j^{BFGS} = -H_{j-1}^\beta g_j + \left[\frac{y_{j-1}^T H_{j-1}^\beta g_j}{y_{j-1}^T d_{j-1}^{BFGS}}\right] d_{j-1}^{BFGS} \tag{4}$$

and

$x_{j+1} = x_j + \lambda_j d_j^{BFGS}$ where $\lambda_j = \min_\lambda f(x_j+\lambda d_j^{BFGS})$. H_j^β is developed from H_{j-1}^β using (2) and x_1 and H_1 are specified.

By comparing (4) and (1) we see that the BFGS algorithm may be interpreted as a CG algorithm for which the metric, instead of being fixed as in (1), is updated at each step to be any member of the Broyden β-class.

This interpretation is of value because it motivates techniques for using limited storage to improve the Conjugate Gradient Method, discussed in the next section.

4. Variable Storage Generalized Conjugate Gradient Algorithms

Conjugate Gradient algorithms require the storage of only a few vectors, typically four. Variable Metric Methods on the other hand require $O(n^2)$ storage. As Fletcher states ([7], p. 82) "practical experience with the Fletcher-Reeves Conjugate Gradient method is that more iterations have usually been required for convergence as against variable metric algorithms -- a factor of two is typical. This has been ascribed to the fact that less information is stored in the Fletcher-Reeves method about the behaviour of the function." Therefore, by using more information about the function one might hope to accelerate the convergence of the conjugate gradient method. For example, in a problem with 10^3 variables, a user may not be able to provide $10^6/2$ words of working storage, thus ruling out variable metric codes implemented in the standard way. However, it may be quite feasible for him to provide $2*10^5$ words well above the $4*10^3$ words required by conjugate gradient methods.

The observations made in earlier sections lead us to suggest the following family of algorithms which can exploit additional storage and form a continuum between the BFGS and CG methods.

The following algorithm describes the family in general terms. We also explain the possible options and discuss them. For a particular implementation see [8].

On Input

n dimension of problem.

$\underline{x}_1$ starting point.

$\underline{\delta}$ vector giving diagonal elements of initial diagonal approximation to inverse Hessian H_0. Note in particular that the symbol H_j represents the n×n Hessian inverse approximation at step j. This is not stored. Instead it is defined implicitly by storing vectors and scalers defining the rank-1 or rank-2 updates at Step 5B below.

Step 1. Initialize

$f_1 \leftarrow f(\underline{x}_1)$, $\underline{g}_1 \leftarrow g(x_1)$, $\underline{y}_0 \leftarrow 0$, $\underline{d}_0 \leftarrow \underline{0}$, H_0 and H_1 are diagonal matrices defined by $\underline{\delta}$, $j \leftarrow 0$.

Step 2. Develop search direction

$$\underline{d}_{j+1} \leftarrow -H_j\underline{g}_{j+1} + \left[\frac{\underline{y}_j^T H_j \underline{g}_{j+1}}{\underline{y}_j^T \underline{d}_j}\right] \underline{d}_j$$

Comment. Relation (4) of Section 3 is used to define search derivations. When j = 0 the multiplier for the second term above is indefinite and is taken to be zero.

Step 3. Search

$j \leftarrow j+1$ if $g_j^T d_j > 0$ then restart;

$\underline{x}_{j+1} \leftarrow \underline{x}_j + \lambda_j \underline{d}_j$

where $\lambda_j = \min\limits_{\mu} f(\underline{x}_j + \mu \underline{d}_j)$.

Comment. For purposes of analysis line searches are taken to be exact. In practice they will not be. In the usual CG method, a fairly accurate line search is required. With VSGCG algorithms we can expect this requirement to be somewhat relaxed.

Step 4. Test for convergence

Stop if convergence criterion is met.

Step 5. If available storage is exceeded.

Step 5A. then employ a Reset Option

Comment. Possible reset options are H_j reset to the diagonal matrix defined by δ and goto Step 5B or H_{j+1} fixed at value of approximation when storage ran out and goto Step 6.

Step 5B. else update H_j to H_{j+1} using a member of the β-class.

Comment. There are a number of options here -- what member of the β-class to use, whether to employ projected vectors (see [9]), and how frequently to perform the update, i.e. whether to udpate whenever possible or every k iterations where k is some fraction of n determined by the amount of storage available. Note also that as discussed above, only the vectors and scalars defining the update are stored.

Step 6. If restart criterion not satisfied then goto Step 2

else employ suitable restart option

Comment. Possible restart criteria are to restart as suggested by Fletcher and Reeves [2] every n or n+1 iterations or to use techniques suggested by Powell [10]. The restart option is also linked to the choice for the reset option.

Remarks

The amount of storage provided is optional. When minimal storage is provided, so Step 5B is never executed, then the method is the standard Conjugate Gradient Method. If n^2 storage is provided and updating performed at every iteration then it is the BFGS algorithm with resetting.

Also, one can easily show that, provided the algorithm does not break down due to instabilities associated with the update, that it has quadratic termination. This is in contrast to the corresponding variable metric algorithm, which holds the approximation H_k fixed at some stage k, when $k < n$.

5. Concluding Remarks

The correspondence between methods proven in this paper requires that line searches be exact. Whether or not line searches are exact, expression (4) determines a family of generalized conjugate gradient algorithms, elaborated upon in Section 4. The properties of such algorithms are currently being investigated.

References

[1] Hestenes, M. R. and Stiefel, E. (1952). *Methods of conjugate gradients for solving linear systems*, J. Res. Natl. Bur. Stand., 49, 409-536.

[2] Fletcher, R. and Reeves, C. M. (1964). *Function minimization by conjugate gradients*, Comput. J., 7, 149-154.

[3] Polak, E. (1971). *Computational Methods in Optimization: a unified approach*, Academic Press.

[4] Davidon, W. C. (1959). *Variable metric method for minimization*, AEC Research and Development Report, ANL-5990 (Rev.).

[5] Broyden, C. G. (1970). *The convergence of a class of double-rank minimization algorithms*, J. Inst. of Math. and Applics., 6, 76-90.

[6] Powell, M.J.D. (1972). *Unconstrained minimization and extension for constraints*, T.P. 495, U.K.A.E.A. Research Group, Atomic Energy Research Establishment, Harwell, England.

[7] Murray, W., Ed. (1972). *Numerical methods for unconstrained optimization*, Academic Press, New York and London.

[8] Nazareth, L. (1977). *Minkit - An Optimization System*, presented at ORSA/TIMS Meeting, San Francisco, May, 1977.

[9] Davidon, W. C. (1975). *Optimally Conditioned Optimization algorithms without line searches*, Math. Prog. 9, 1-30.

[10] Powell, M.J.D. (1975). *Restart Procedures for the Conjugate Gradient Method*, C.S.S. 24, A.E.R.E., Harwell, England.