AN ALTERNATIVE VARIATIONAL PRINCIPLE FOR VARIABLE METRIC UPDATING

Larry Nazareth

January 1983
WP-83-12

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

ABSTRACT

We describe a variational principle based upon minimizing the extent to which the inverse hessian approximation, say H, violates the quasi-Newton relation, on the step immediately prior to the step used to construct H. It suggests use of the BFGS update.

AN ALTERNATIVE VARIATIONAL PRINCIPLE FOR VARIABLE METRIC UPDATING

Larry Nazareth

1. Introduction

The problem under consideration here is that of minimizing an unconstrained function $f(x)$, $x \in R^n$, by means of a variable metric method. The original method of this type is due to Davidon, 1959, whose work was subsequently clarified and extended by Fletcher & Powell, 1963. The method is thus popularly known as the DFP method.

Given an approximation B to the hessian of $f(x)$, a step δx and the gradient δg corresponding to this step, with $\delta x^T \delta g \neq 0$, a new approximation B_+, which satisfies the quasi-Newton relation $B_+ \delta x = \delta g$, is defined as follows:

$$B_+ = (I - \rho \delta g \, \delta x^T) B (I - \rho \delta g \, \delta x^T)^T + \rho \delta g \, \delta g^T \qquad (1.1a)$$

where $\rho = 1 / \delta g^T \delta x$. This is the DFP update.

If $H = B^{-1}$, the new approximation $H_+ = B_+^{-1}$ to the inverse hessian obtained using the DFP update, satisfies $H_+\delta g = \delta x$ and is given by:

$$H_+ = H - \frac{H\delta g\,\delta g^T H}{\delta g^T H \delta g} + \frac{\delta x\,\delta x^T}{\delta g^T \delta x} \tag{1.1b}$$

By interchanging δx and δg and interchanging H and B in (1.1), we obtain the complementary DFP update, known popularly as the BFGS update. This is widely believed to be the most effective variable metric update, and is defined by:

$$H_+ = (I - \rho\delta x\,\delta g^T)H(I - \rho\delta x\,\delta g^T)^T + \rho\delta x\,\delta x^T \tag{1.2a}$$

$$B_+ = B - \frac{B\delta x\,\delta x^T B}{\delta x^T B \delta x} + \frac{\delta g\,\delta g^T}{\delta g^T \delta x} \tag{1.2b}$$

The DFP and BFGS updates are both members of a single parameter family known as the Broyden class, Broyden, 1970. There are a number of equivalent expressions for it. A convenient one is:

$$H_+^\beta = (I - \rho\delta x\,\delta g^T)H(I - \rho\delta x\,\delta g^T)^T + \rho\delta x\,\delta x^T + \beta w w^T \tag{1.3}$$

where

$$w = H\delta g - \frac{\delta g^T H \delta g}{\delta g^T \delta x}\delta x \tag{1.4}$$

β is a real number, and ρ is defined as in (1.1a). There is a corresponding expression for B_+^β.

In (1.3) we can think of H_+^β as being obtained by adding to H suitable rank-1 matrices composed from the vectors $H\delta g$ and δx, or equivalently from the vectors $(\delta x - H\delta g)$ and δx. The significance of this remark is that when variable metric methods that use *exact* line searches are

applied to a quadratic function, the vectors $(\delta x - H\delta g)$ and δx at the current iterate can be shown to be conjugate to all previous steps. Thus H_+^β has what is known as the hereditary property i.e. it will satisfy the quasi-Newton relation on previous steps. (We assume that the reader is reasonably familiar with the terminology and literature on variable metric methods, see Murray, 1972.) Similar statements can be made about B_+^β.

We shall use the notation of update functions, see Dennis & More, 1977, to write (1.2a) as $H_+ = U_{BFGS}(\delta x, \delta g, H)$ and (1.3) as $H_+^\beta = U_\beta(\delta x, \delta g, H)$. Similar expressions are used for the DFP update. Also $H > 0$ means H is positive definite, and $u // v$ means the vector u is parallel to the vector v.

Given $\delta x, \delta g$ with $\delta x^T \delta g > 0$ and $H = H^T$, Dennis & More, 1977, show that the update $H_+ = U_{BFGS}(\delta x, \delta g, H)$ is the minimum norm update in the following sense

$$H_+ = \arg\min_{\bar{H}} \{ ||\bar{H} - H||_{W,F} : \bar{H} \text{ symmetric and } \bar{H}\delta g = \delta x \} \quad (1.5)$$

where $||\cdot||_{W,F}$ is a weighted Frobenius norm defined for any square symmetric matrix as

$$||M||_{W,F}^2 = \sum_{i,j} (\tilde{H}^{-1/2} M \tilde{H}^{-1/2})_{ij}^2 \quad (1.6)$$

and $\tilde{H}$ satisfies $\tilde{H}\delta g = \delta x, \tilde{H} > 0$.

Using the above weighted norm represents a very natural rescaling of the problem using a positive definite matrix which satisfies the quasi-Newton relation on the current step. The originators of this approach include Greenstalt, 1970 and Goldfarb, 1970.

Here we study an alternative variational principle. Suppose δx_- represents the step immediately prior to δx and δg_- the corresponding gradient change, with $H\delta g_- = \delta x_-$. In general $H_+\delta g_- \neq \delta x_-$. Since the purpose behind the formation of H on this prior iteration was to satisfy the quasi-Newton relation on δx_-, it seems reasonable to ask which update H_+^β from among those of the form (1.3) minimizes $||H_+^\beta \delta g_- - \delta x_-||_W$, where W is a suitable vector norm. We show that for different choices of W, solutions correspond to the BFGS and DFP updates. In particular, the BFGS is, in a sense, the "best" solution, because the associated choice of W is the most natural one. In the discussion we compare the new variational principle with (1.5).

2. Alternative Variational Principle

Let us first study the preliminary question of when an inverse hessian approximation H can satisfy a quasi-Newton relation simultaneously over several steps.

Theorem 2.1: Given linearly independent δx_i and linearly independent δg_i, $i = 1,2,\ldots,k$ which satisfy $\delta x_i^T \delta g_i \neq 0$, then there is a symmetric matrix H such that $H\delta g_i = \delta x_i$, $i = 1,2,\ldots,k$ if and only if $\delta x_i^T \delta g_j = \delta g_i^T \delta x_j$, $i \neq j$, $1 \leq i, j \leq k$.

Proof: (i) Suppose there exists a symmetric matrix H such that $H\delta g_i = \delta x_i$, and $H\delta g_j = \delta x_j$, $i \neq j$, $1 \leq i, j \leq k$.

Then

$$\delta g_j^T H \delta g_i = \delta g_j^T \delta x_i \tag{2.1}$$

and

$$\delta g_i^T H \delta g_j = \delta g_i^T \delta x_j \quad .$$

Since H is symmetric, $\delta x_i^T \delta g_j = \delta g_i^T \delta x_j$, $i \neq j$, $1 \leq i, j \leq k$.

(ii) Suppose now

$$\delta x_i^T \delta g_j = \delta g_i^T \delta x_j, \; i \neq j, \; 1 \leq i, j \leq k \tag{2.2}$$

Assume that there exists an index $m < k$ and a symmetric H_m such that

$$H_m \delta g_j = \delta x_j \; , \; j = 1,2,\ldots,m-1 \tag{2.3}$$

Let

$$H_{m+1} = H_m + \frac{(\delta x_m - H_m \delta g_m) u_m^T + u_m (\delta x_m - H_m \delta g_m)^T}{u_m^T \delta g_m}$$

$$- \frac{(\delta x_m - H_m \delta g_m)^T \delta g_m}{(u_m^T \delta g_m)^2} u_m u_m^T$$

where $u_m \neq 0$ is chosen orthogonal to $\delta g_1, \ldots, \delta g_{m-1}$ and $u_m^T \delta g_m \neq 0$. The latter can always be satisfied since $\delta g_1, \ldots, \delta g_m$ are linearly independent. Then

$$H_m \delta g_m = \delta x_m \quad .$$

Also for $j < m$ we have

$$(\delta x_m - H_m \delta g_m)^T \delta g_j = \delta x_m^T \delta g_j - \delta g_m^T (H_m \delta g_j)$$

$$= \delta x_m^T \delta g_j - \delta g_m^T \delta x_j$$

by the induction hypothesis (2.3).

$$= 0 \quad \text{by (2.2).}$$

Because of this, and the way u_m is defined

$$H_{m+1}\delta g_j = \delta x_j \text{ for } j = 1,2,\ldots,m$$

The induction hypothesis therefore holds for $j = 1,2,\ldots,m$.

Since the induction hypothesis is obviously true for $m = 2$, the result follows with H given by $H = H_{k+1}$. □

In general it is clear that the conditions of the above theorem will not hold. Reverting to our simpler notation of Section 1, $\delta x_-^T \delta g = \delta g^T \delta x_-$ will usually not hold. It is then natural to ask which update solves the problem:

$$\min_{\bar{H}} \left\{ ||\bar{H}\delta g_- - \delta x_-||_W : \bar{H} \in U_\beta(\delta x, \delta g, H) \right\} \tag{2.4}$$

for some suitable choice of vector norm $||\cdot||_W$.

Theorem 2.2: Given $H > 0$, $\delta x_- = (x - x_-)$ and corresponding $\delta g_- = (g - g_-)$, let $H\delta g_- = \delta x_-$. Let δx be a non-zero step satisfying $\delta x^T \delta g_- = 0$, and δg be the corresponding gradient change with $\delta x^T \delta g > 0$. Then:

(i) The BFGS update $H_+^B = U_{BFGS}(\delta x, \delta g, H)$ solves the problem (2.4) where $||\cdot||_W$ is the vector norm defined by $W = \tilde{H}^{-1}$ such that $\tilde{H}\delta g = \delta x$, $\tilde{H} > 0$.

(ii) The DFP update $H_+^D = U_{DFP}(\delta x, \delta g, H)$ solves the problem (2.4), where $||\cdot||_W$ is taken to be the vector norm defined by $W = H^{-1}$.

Proof: We now use the definition of the BFGS and DFP updates and the Broyden family given by (1.2a), (1.1b) and (1.3) and henceforth we affix

the symbols B and D to H_+ to distinguish the BFGS and DFP updates. It follows from (1.2a) and $\delta x^T \delta g_- = 0$ that

$$H_+^B \delta g_- = -\rho(\delta g^T \delta x_-)\delta x + \delta x_-$$

We can assume that $H_+^B \delta g_- \neq \delta x_-$ or the result would follow immediately. Hence

$$(H_+^+ \delta g_- - \delta x_-) // \delta x$$

Now from (1.3)

$$(H_+^\beta \delta g_- - \delta x_-) = (H_+^B \delta g_- - \delta x_-) + \beta w(w^T \delta g_-)$$

Also

$$(H_+^B \delta g_- - \delta x_-)^T(\tilde{H}^{-1})w = -\rho(\delta g^T \delta x_-)\delta x^T(\tilde{H}^{-1})w$$

$$= -\rho(\delta g^T \delta x_-)\delta g^T w$$

$$= 0 \text{ using (1.4), since } w^T \delta g = 0 \quad .$$

Thus $||H_+^\beta \delta g_- - \delta x_-||_W \geq ||H_+^B \delta g_- - \delta x_-||_W$ for all β where $||\cdot||_W$ is defined as in (i) in the statement of the theorem.

(ii) The Broyden family can also be written

$$H_+^\varphi = H_+^D + \varphi w w^T \quad , \ \varphi \text{ a scalar.}$$

Now from (1.1b)

$$(H_+^D \delta g_- - \delta x_-) = -(\delta g^T H \delta g_- / \delta g^T H \delta g) H \delta g \qquad (2.5)$$

and

$$(H_+^\varphi \delta g_- - \delta x_-) = (H_+^D \delta g_- - \delta x_-) + \varphi(w^T \delta g_-)w$$

Also

$$(H_+^D \delta g_- - \delta x_-)^T H^{-1} w = -(\delta g^T H \delta g_- / \delta g^T H \delta g) \delta g^T w = 0$$

again from (1.4).

It follows that $||H_+^\varphi \delta g_- - \delta x_-||_{H^{-1}} \geq ||H_+^D \delta g_- - \delta x_-||_{H^{-1}}$ for all φ.

3. Discussion

The condition $\delta x^T \delta g_- = 0$ in Theorem 2.2 holds when x is a minimizng point of $f(x)$ along the direction δx_- (i.e. the line search is exact) and $\delta x \,//\, Hg$. This follows because

$$g^T H \delta g_- = g^T \delta x_- = 0 \quad , \text{ hence } \delta x^T \delta g_- = 0 \tag{3.1}$$

Now when line searches are exact (and conflicts are unambiguously resolved) we know from Dixon's, 1972, theorem that variable metric methods based upon (1.3) give iterates that are independent of the values of the parameter β. Furthermore, under these conditions, Powell has shown that the single parameter family (1.3) is the most general family of updates that leaves iterates unaltered, see Dixon, 1972. We therefore have a very natural context within which to ask the question: Which updates H_+^β solves (2.4)? From Theorem 2.2 we see that for what seems to be the most natural choice of W, the solution is the BFGS update.

The requirement that line searches be exact can be dropped, by modifying the way in which search directions are defined. If for given $H_- > 0$, $\delta x_- = x - x_-$ and $\delta g_- = g - g_-$ with $\delta g_-^T \delta x_- > 0$ we replace $\delta x \,//\, U_\beta(\delta x_-, \delta g_-, H_-)g$ by

$$\delta x \,//\, \left(H_- \delta g_- - \frac{\delta g_-^T H_- \delta g_-}{\delta g_-^T \delta x_-} \delta x_-\right) \tag{3.2}$$

Then $\delta x^T \delta g_- = 0$. Also Dixon's theorem extends to variable metric

methods based on (1.3) and (3.2), see Nazareth, 1982. This too is a natural setting for Theorem 2.2, which again suggest the BFGS update is the appropriate choice.

We have noted that the problem (1.5) has been extensively studied. A problem similar to the one quoted can be formulated for minimizing $||B_+ - B||_{W,F}$ in a suitable weighted Frobenius norm, and in this case the DFP update is the solution. Again, see Dennis & More, 1977. It is therefore open to question whether the BFGS update is singled out. It has been argued that since variable metric methods use the inverse hessian, the reasonable thing to do is to minimize $||H_+ - H||_{W,F}$. But search directions can equally well be defined in terms of the hessian B, and one could argue with equal conviction that one should minimize $||B_+ - B||_{W,F}$.

In contrast, the result that would be complementary to Theorem 2.2 does not go through in an analogous manner. The reason for this is that the proof of Theorem 2.2 uses the fact that $\delta x^T \delta g_- = 0$. When working with the complementary form, this relation would transform to $\delta g^T \delta x_- = 0$. This is not necessarily true for the first case one would consider, namely, the case when line searches are exact. If line searches are exact, and $\delta g^T \delta x_- = 0$, then the conditions of Theorem 2.1 hold, and the quasi-Newton relation would be simultaneously satisfied on δx and δx_-. Theorem 2.2 singles out the BFGS update since the scaling by $\tilde{H}^{-1}$ is to be preferred to the scaling by H^{-1}.

Finally we point out that Theorem 2.2 can quite easily be generalized to the extended family of updates of Davidon, 1975, and that alternative

variational principles to (1.5), e.g. (2.4), may provide useful guides for choosing suitable updates within other contexts, for example, quasi-Newton methods for solving systems of nonlinear equations.

REFERENCES

Broyden, C.G. (1970), "The convergence of a class of double-rank minimization algorithms", *Journal of the Institute of Mathematics and its Applications*, 6, 76-90.

Davidon, W.C. (1959), "Variable Metric Method for Minimization", AEC Research and Development Report, ANL-5990 (REV.), Argonne National Laboratory, Argonne, Illinois.

Davidon, W.C. (1975), "Optimally conditioned optimization algorithms without line searches", *Mathematical Programming*, 9, 1-30.

Dennis, J.E. and J.J. More (1977), "Quasi-Newton methods, motivation and theory", *SIAM Review*, 19, 46-89.

Dixon, L.C.W. (1972), "Quasi-Newton algorithms generate identical points", *Mathematical Programming*, 2, 383-387.

Fletcher, R. and M.J.D. Powell (1963), "A rapidly convergent descent

method for minimization", *Computer Journal*, 6, 163-168.

Goldfarb, D. (1970), "A family of variable metric methods derived by variational means", *Mathematics of Computation*, 24, 23-26.

Greenstadt, J. (1970), "Variations on variable metric methods", *Mathematics of Computation*, 24, 1-18.

Murray, W. (Ed.) (1972), *Numerical Methods for Unconstrained Optimization*, Academic Press, London and New York.

Nazareth, L. (1982), "Analogues of Dixon's and Powell's theorems for unconstrained minimization with inexact line searches", IIASA Working Paper, WP-82-100.