MATEMATIKA, 2006, Volume 22, Number 1, pp. 17–24
©Department of Mathematics, UTM.

Convergence of A Modified BFGS Method

[1]Malik Abu Hassan, [2]Leong Wah June, [3]Mansor Monsi
Department of Mathematics, University Putra Malaysia
Serdang, Selangor 43400, Malaysia.
e-mail: [1]malik@fsas.upm.edu.my, [2]lwjune@fsas.upm.edu.my,[3]mansor@fsas.upm.edu.my

Abstract In this paper we discuss the convergence of a modified BFGS method. We prove that the modified BFGS method will terminate in n steps when minimizing $n-$dimensional quadratic functions with exact line searches.

Keywords Quadratic termination, modified BFGS.

1 Introduction

The quasi-Newton methods are very useful and efficient methods for solving the unconstrained minimization problem

$$\min f(x); x \in \Re^n. \tag{1}$$

Many of these methods share the properties of finite termination on strictly convex quadratic functions, a linear or superlinear rate of convergence on general convex functions, and no need to store or evaluate the second derivative matrix. In general, an approximation to the second derivative matrix is built by accumulating the results of earlier steps. Typically, given both an approximation H_k to $[\nabla^2 f(x_k)]^{-1}$ and g_k the gradient $\nabla f(x_k)$ at the current point x_k, a quasi-Newton algorithm starts each iteration by taking a step from the current

$$x_{k+1} = x_k - \lambda H_k g_k, \tag{2}$$

where the steplength $\lambda > 0$ is chosen so that

$$f(x_k) \geq f(x_k - \lambda H_k g_k) \tag{3}$$

are satisfied; and then to form H_{k+1} by using an updating formula satisfying the quasi-Newton condition

$$H_{k+1} y_k = s_k, \tag{4}$$

where $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$. Descriptions of many quasi-Newton algorithms can be found in books by Luenberger [4] and Dennis and Schnabel [3]. Although there are a large number of quasi-Newton methods, one method surpasses the others in popularity: the BFGS update of Broyden, Fletcher, Goldfarb, and Shanno; see, e.g., Dennis and Schnabel [3]:

$$H_{k+1} = H_k + \frac{1}{s_k^T y_k}\left(\left(1 + \frac{y_k^T H_k y_k}{s_k^T y_k}\right) s_k s_k^T - s_k y_k^T H_k - H_k y_k s_k^T\right). \tag{5}$$

This method exhibits more robust behavior than its relatives. Many attempts have been made to improve this robustness. Among them are the works by Yuan [7] and Biggs [1, 2], which give a modified BFGS update. In the following section, we will briefly describe this modified update. We also give some convergence properties for these methods in Section 3.

In this paper, the following notations are used: $span\{x_1, x_2, \ldots, x_k\}$ denotes the subspace spanned by $x_1, x_2, \ldots, x_k$. Whenever we refer to an $n-$dimensional strictly convex quadratic function, we assume it is of the form

$$f(x) = \frac{1}{2}x^T Ax - x^T b,$$

where A is a positive definite $n \times n$ matrix and b is an n vector.

2 A Modified BFGS Update

Assuming H_k non-singular, we define $B_k = H_k^{-1}$. It is easy to see that the *quasi-Newton* step

$$d_k = -H_k g_k \tag{6}$$

is a stationary point of the following problem:

$$min_{d \in \Re^n} \phi_k(d) = f(x_k) + d^T g_k + \frac{1}{2} d^T B_k d \tag{7}$$

which is an approximation to problem (1) near the current iterate x_k, since $\phi_k(d) \approx f(x_k+d)$ for small d. In fact, the definition of $\phi_k(\cdot)$ in (7) imples that

$$\phi_k(0) = f(x_k), \tag{8}$$

$$\nabla\phi_k(0) = g(x_k), \tag{9}$$

and the quasi-Newton condition (4) is equivalent to

$$\nabla\phi_k(x_{k-1} - x_k) = g(x_{k-1}). \tag{10}$$

Thus, $\phi_k(x - x_k)$ is a quadratic interpolation of $f(x)$ at x_k and x_{k-1}, satisfying conditions (8)$-$ (10). The matrix B_k (or H_k) can be updated so that the quasi-Newton equation is satisfied.

In [7], approximate function $\phi_k(d)$ in (7) is required to satisfy the interpolation condition

$$\phi_k(x_{k-1} - x_k) = f(x_{k-1}) \tag{11}$$

instead of (10). This change was inspired from the fact that for one dimensional problem, using (11) give a slightly faster local convergence if we assume $\lambda_k = 1$ for all k. Equation (11) can be rewritten as

$$s_{k-1}^T B_k s_{k-1} = 2\left[f(x_{k-1}) - f(x_k) + s_{k-1}^T g_k\right], \tag{12}$$

In order to satisfy (12), the BFGS formula is modified as follows:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + t_k \frac{y_k y_k^T}{s_k^T y_k}, \tag{13}$$

where

$$t_k = \frac{2}{s_k^T y_k}\left[f(x_k) - f(x_{k+1}) + s_k^T g_{k+1}\right]. \tag{14}$$

The inverse update, H_{k+1} will be

$$H_{k+1} = H_k + \frac{1}{s_k^T y_k}\left(\left(\alpha_k + \frac{y_k^T H_k y_k}{s_k^T y_k}\right) s_k s_k^T - s_k y_k^T H_k - H_k y_k s_k^T\right), \tag{15}$$

with $\alpha_k = 1/t_k$.

Assume that B_k is positive definite and that $s_k^T y_k > 0$, B_{k+1} defined by (13) is positive definite if and only if $t_k > 0$. The inequality $t_k > 0$ is trivial if f is strictly convex, and it is also true if the steplength λ_k is chosen by an exact line search, which requires $s_k^T g_{k+1} = 0$. For a uniformly convex function, it can be easily shown that there exists a constant $\delta > 0$ such that $t_k \in [\delta, 2]$ for all k, and consequently global convergence proof of the BFGS method for convex functions with inexact line searches, which was given by Powell [5].

For a general nonlinear function, Yuan [7] truncated t_k to the interval $[0.01, 100]$, and showed that the global convergence of the modified BFGS algorithm is preserved for convex functions.

If the objective function f is cubic along the line segment between x_{k-1} and x_k then we have the following relation

$$s_{k-1}^T \nabla^2 f(x_k) s_{k-1} = 4 s_{k-1}^T g_k + 2 s_{k-1}^T g_{k-1} - 6\left[f(x_{k-1}) - f(x_k)\right], \tag{16}$$

by considering the Hermit interpolation on the line between x_{k-1} and x_k. Hence it is reasonable to require that the new approximate Hessian satisfy condition

$$s_{k-1}^T B_k s_{k-1} = 4 s_{k-1}^T g_k + 2 s_{k-1}^T g_{k-1} - 6\left[f(x_{k-1}) - f(x_k)\right]. \tag{17}$$

Biggs [1, 2] gives the inverse of update of (13) with the value t_k so chosen that (17) holds. The respected value of t_k is given by

$$t_k = \frac{6}{s_k^T y_k}\left[f(x_k) - f(x_{k+1}) + s_k^T g_{k+1}\right] - 2. \tag{18}$$

For one-dimensional problems, Wang and Yuan [6] showed that (13) with (18) and without line searches (that is $\lambda_k = 1$ for all k) implies $R-$quadratic convergence.

3 Convergence of the modified BFGS method

We will now describe new representations of the modified BFGS update and show that using this update, the quasi-Newton with exact line searches will terminte in n step when minimizing quadratic functions of n variables.

Let us consider quasi-Newton methods with an update of the form

$$H_{k+1} = P_k^T H_0 Q_k + \sum_{i=1}^{k} w_{ik} z_{ik}^T. \tag{19}$$

Here, we restrict ourselves to the following:

(i) H_0 is an $n \times n$ symmetric positive definite matrix denotes the initial approximation of the inverse Hessian. Mostly $H_0 = I$, the identity matrix is set;

(ii) P_k is an $n \times n$ matrix that is the product of projection matrices of the form

$$I - \frac{uv^T}{u^T v}, \tag{20}$$

where $u \in span\{y_0, \ldots, y_k\}$ and $v \in span\{s_0, \ldots, s_k\}$, and Q_k is an $n \times n$ matrix that is the product of projection matrices of the same form where u is any $n-$vector and $v \in span\{s_0, \ldots, s_k\}$;

(iii) w_{ik} $(i = 1, \ldots, k)$ is any $n-$vector, and z_{ik} $(i = 1, \ldots, k)$ is any vector in $span\{s_0, \ldots, s_k\}$.

This form of update fits many known quasi-Newton methods, including the Broyden family and BFGS method. The modified BFGS update (15) is also equivalent to the (19) with

$$P_k = Q_k = \prod_{j=0}^{k}\left(I - \frac{y_j s_j^T}{s_j^T y_j}\right), w_{ik} = z_{ik} = \frac{\prod_{j=i}^{k}\left(I - (y_j s_j^T)/(s_j^T y_j)\right)^T s_i}{\sqrt{t_i s_i^T y_i}}. \tag{21}$$

It is trivial that P_k, Q_k and z_{ik} all obey the constraints imposed on them.

We now show that the modified BFGS method of the form (19) with (21) produce conjugate search directions and terminate in n iterations.

Theorem 1 *Suppose that we apply a quasi-Newton method with an update of the form (19) with (21) to minimize an n-dimensional strictly convex quadratic function. Then for each k before termination (i.e., $g_{k+1} \neq 0$),*

$$g_{k+1}^T s_j = 0, \text{ for all } j = 0, 1, \ldots, k, \tag{22}$$

$$s_{k+1}^T A s_j = 0, \text{ for all } j = 0, 1, \ldots, k, \text{ and} \tag{23}$$

$$span\{s_0, \ldots, s_{k+1}\} = span\{H_0 g_0, \ldots, H_0 g_{k+1}\}, \tag{24}$$

Proof Since

$$P_k y_i = \begin{cases} 0, & \text{if } i = 1, \ldots, k \\ y_i, & \text{if } i = 0. \end{cases}$$

we will first show that

$$P_j y_i \in span\{y_0, \ldots, y_{j-1}\} \text{ for all } i = 0, 1, \ldots, k, j = 1, \ldots, k. \tag{25}$$

Note that

$$P_j y_i = \prod_{i=1}^{j}\left(I - \frac{y_i s_i^T}{s_i^T y_i}\right) y_i. \tag{26}$$

We will prove (22)-(24) by induction. Since the line searches are exact, g_1 is orthogonal to s_0. Using the fact that $P_0 y_0 = 0$ from (25) and the fact that $z_{i0} \in span\{s_0\}$ implies $g_1^T z_{i0} = 0$, $i = 1, \ldots, k$, we see that s_1 is conjugate to s_0 since

$$\begin{aligned} s_1^T A s_0 &= \lambda_1 d_1^T y_0 \\ &= -\lambda_1 g_1^T H_1^T y_0 \\ &= -\lambda_1 g_1^T \left(Q_0^T H_0 P_0 + z_{1,0} w_{1,0}^T\right) y_0 \\ &= 0. \end{aligned}$$

Finally, $span\{s_0\} = span\{H_0 g_0\}$, and so the base case is established.

We will now assume that claims (22)-(24) hold for $k = 0, 1, \ldots, \hat{k} - 1$ and prove that they also hold for $k = \hat{k}$.

The vector $g_{\hat{k}+1}$ is orthogonal to $s_{\hat{k}}$ since the line search is exact. Using the induction hypothesis that $g_{\hat{k}}$ is orthogonal to $\{s_0, \ldots, s_{\hat{k}-1}\}$ and $s_{\hat{k}}$ is conjugate to $\{s_0, \ldots, s_{\hat{k}-1}\}$, we see that, for $j < \hat{k}$,

$$g_{\hat{k}+1}^T s_j = (g_{\hat{k}} + y_{\hat{k}})^T s_j = (g_{\hat{k}} + A s_{\hat{k}})^T s_j = 0.$$

Hence, (22) holds for $k = \hat{k}$.

To prove (23), we note that

$$s_{\hat{k}+1}^T A s_j = -\lambda_{\hat{k}+1} g_{\hat{k}+1}^T H_{\hat{k}+1}^T y_j,$$

so it is sufficient to prove that $g_{\hat{k}+1}^T H_{\hat{k}+1}^T y_j = 0$ for $j = 0, 1, \ldots, \hat{k}$. We will use the following facts:

(i) $g_{\hat{k}+1}^T Q_{\hat{k}}^T = g_{\hat{k}+1}^T$ since each s used to form $Q_{\hat{k}}$ is in $span\{s_0, \ldots, s_{\hat{k}}\}$, and $g_{\hat{k}+1}^T$ is orthogonal to that span.

(ii) $g_{\hat{k}+1}^T z_{i\hat{k}} = 0$ for $i = 1, \ldots, \hat{k}$ since each $z_{i\hat{k}}$ is in $span\{s_0, \ldots, s_{\hat{k}}\}$, and again $g_{\hat{k}+1}^T$ is orthogonal to that span.

(iii) Since we have already showed that (25) holds true, for each $j = 0, 1, \ldots, \hat{k}$ there exist $\nu_0, \ldots, \nu_{\hat{k}-1}$ such that $P_{\hat{k}} y_j$ can be express as $\sum_{i=0}^{\hat{k}-1} \nu_i y_i$.

(iv) For $i = 0, 1, \ldots, \hat{k} - 1$, $g_{\hat{k}+1}$ is orthogonal to $H_0 y_i$ because $g_{\hat{k}+1}$ is orthogonal to $span\{s_0, \ldots, s_{\hat{k}}\}$ and $H_0 y_i \in span\{s_0, \ldots, s_{\hat{k}}\}$ from (24).

Thus,

$$\begin{aligned}
g_{\hat{k}+1}^T H_{\hat{k}+1}^T y_j &= g_{\hat{k}+1}^T \left(Q_{\hat{k}}^T H_0 P_{\hat{k}} + \sum_{i=1}^{\hat{k}} z_{i\hat{k}} w_{i\hat{k}}^T \right) y_j \\
&= g_{\hat{k}+1}^T Q_{\hat{k}}^T H_0 P_{\hat{k}} y_j + \sum_{i=1}^{\hat{k}} g_{\hat{k}+1}^T z_{i\hat{k}} w_{i\hat{k}}^T y_j \\
&= g_{\hat{k}+1}^T H_0 P_{\hat{k}} y_j \\
&= g_{\hat{k}+1}^T H_0 \left(\sum_{i=1}^{\hat{k}-1} \nu_i y_i \right) \\
&= \left(\sum_{i=1}^{\hat{k}-1} \nu_i g_{\hat{k}+1}^T H_0 y_i \right) \\
&= 0.
\end{aligned}$$

Therefore, (23) holds for $k = \hat{k}$.

Finally, using (i) and (ii) from above,

$$\begin{aligned}
s_{\hat{k}+1} &= -\lambda_{\hat{k}+1} H_{\hat{k}+1} g_{\hat{k}+1} \\
&= -\lambda_{\hat{k}+1} \left(P_{\hat{k}} H_0 Q_{\hat{k}} g_{\hat{k}+1} + \sum_{i=1}^{\hat{k}} w_{i\hat{k}} z_{i\hat{k}}^T g_{\hat{k}+1} \right) \\
&= -\lambda_{\hat{k}+1} P_{\hat{k}}^T H_0 g_{\hat{k}+1}.
\end{aligned}$$

Since $P_{\hat{k}}^T$ maps any $n-$vector v into $span\{v, s_0, \ldots, s_{\hat{k}+1}\}$ by its construction, there exist $\mu_0, \ldots, \mu_{\hat{k}+1}$ such that

$$s_{\hat{k}+1} = -\lambda_{\hat{k}+1} \left(H_0 g_{\hat{k}+1} + \sum_{i=0}^{\hat{k}+1} mu_i s_i \right).$$

Hence,

$$H_0 g_{\hat{k}+1} \in span\{s_0, \ldots, s_{\hat{k}+1}\},$$

so

$$span\{H_0 g_0, \ldots, H_0 g_{\hat{k}+1}\} \subseteq span\{s_0, \ldots, s_{\hat{k}+1}\}.$$

To show equality of the above sets, we will show that $H_0 g_{\hat{k}+1}$ is linearly independent of $\{H_0 g_0, \ldots, H_0 g_{\hat{k}}\}$. (We already have that the vector $H_0 g_0, \ldots, H_0 g_{\hat{k}}$ are linearly independent since they span the same space as the linear independent set $s_0, \ldots, s_{\hat{k}}$.) Suppose that $H_0 g_{\hat{k}+1}$ is not linearly independent. Then there exist $\beta_0, \ldots, \beta_{\hat{k}}$, not all zero, such that

$$H_0 g_{\hat{k}+1} = \sum_{i=0}^{\hat{k}} \beta_i H_0 g_i.$$

Since $g_{\hat{k}+1}$ is orthogonal to $\{s_0, \ldots, s_{\hat{k}}\}$ and by our induction assumption, this implies that $g_{\hat{k}+1}$ is also orthogonal to $\{H_0 g_0, \ldots, H_0 g_{\hat{k}}\}$. Thus, for any j between 0 and $\hat{k}$,

$$0 = g_{\hat{k}+1}^T H_0 g_j = \left(\sum_{i=0}^{\hat{k}} \beta_i H_0 g_i \right)^T g_j = \sum_{i=0}^{\hat{k}} \beta_i g_i^T H_0 g_j = \beta_j g_j^T H_0 g_j.$$

Since H_0 is positive definite and g_j is nonzero, we conclude that β_j must be zero. Since this is true for every j between 0 and $\hat{k}$, we have a contradiction. Thus, the set $\{H_0 g_0, \ldots, H_0 g_{\hat{k}+1}\}$ is linearly independent. Hence, (24) holds for $k = \hat{k}$.

When a method produces conjugate search directions, we can say something about termination.

Corollary *Suppose we have a method satisfying all conditions in Theorem 1, then this method will terminates in no more than n iterations.*

Proof Let k be such that $g_0, \ldots, g_k$ are all nonzero and such that $H_i g_i \neq 0$ for $i = 0, \ldots, k$. Since we have a method satisfying all conditions in Theorem 1, we claim that the $(k+1)$-subspace of search directions, $span\{s_0, \ldots, s_k\}$ is equal to the $(k+1)$-Krylov subspace, $span\{H_0 g_0, \ldots, (H_0 A)^k H_0 g_0\}$.

From (24), we know that $span\{s_0, \ldots, s_k\} = span\{H_0 g_0, \ldots, H_0 g_k\}$. We will show via induction that $span\{H_0 g_0, \ldots, H_0 g_k\} = span\{H_0 g_0, \ldots, (H_0 A)^k H_0 g_0\}$. This base case is trivial since $(H_0 A)^0 = I$. So assume that

$$span\{H_0 g_0, \ldots, H_0 g_i\} = span\{H_0 g_0, \ldots, (H_0 A)^i H_0 g_0\}$$

for some $i < k$. Now,

$$g_{i+1} = A x_{i+1} - b = A(x_i + s_i) - b = A s_i + g_i,$$

and from (24) and the induction hypothesis,

$$s_i \in span\{H_0 g_0, \ldots, H_0 g_i\} = span\{H_0 g_0, \ldots, (H_0 A)^i H_0 g_0\},$$

which implies that $H_0 A s_i \in span\{(H_0 A) H_0 g_0, \ldots, (H_0 A)^{i+1} H_0 g_0\}$. So,

$$H_0 g_{i+1} \in span\{H_0 g_0, \ldots, (H_0 A)^{i+1} H_0 g_0\}.$$

Hence, the search directions span the Krylov subspace and are conjugate. Then the iterates are the same as those produces by conjugate gradient methods with preconditioner H_0 (or classical conjugate gradients with $H_0 = I$).

The conjugate gradient method is well known to terminate within n iterations, we can conclude that the given modified BFGS scheme terminates in at most n iterations. □

Note that we require that $H_k g_k$ be nonzero whenever g_k is nonzero; this requirement is equivalent to positive definite updates and is trivial if $t_k > 0$.

4 Conclusions

We have shown that the modified BFGS method fitting a form (19) with (21) have the property of producing conjugate search directions on convex quadratics. This method will terminate in at most n iterations. This type of finite termination property has sometimes been called *quadratic termination.* The relevance of the quadratic termination property to the general nonlinear functions was originally based on the assumption that if a method terminates in a finite number of steps for a quadratic then this implies superlinear convergence for nonlinear functions.

References

[1] M. C. Biggs, *Minimization algorithms making use of non-quadratic properties of the objective function*, J. Institute of Mathematics and Its Appl., 8 (1971), 315–327.

[2] M. C. Biggs, *A note on minimization algorithms making use of non-quadratic properties of the objective function*, J. Institute of Mathematics and Its Appl., 12 (1973), 337–338.

[3] J. Dennis and R. B. Schnabel *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Series in Computational Mathematics, Prentice-Hall, Englewood Cliffs, NJ, 1983. Reprinted by Society for Industrial and Applied Mathematics, Philadelphia, PA, 1996.

[4] D. C. Luenberger, *Linear and Nonlinear Programming*, Second ed., Addison-Wesley, Reading, MA, 1984.

[5] M. J.D. Powell, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, in Nonlinear Programming SIAM-AMS Proceeding, Vol. 9, R.W. Cottle and C.E. Lemke, eds., SIAM Publications, 1976, 53–72.

[6] H. J. Wang and Y. Yuan, *A quadratic convergence method for one-dimensional optimization*, Chinese J. Operations Research, 11 (1992), 1–10.

[7] Y. Yuan, *A modified BFGS algorithm for unconstrained optimization*, IMA J. Numerical Analysis, 11 (1991), 325–332.