Journal of Computational Mathematics, Vol.17, No.6, 1999, 645–652.

CONVERGENCE PROPERTIES OF A MODIFIED BFGS ALGORITHM FOR MINIMIZATION WITH ARMIJO-GOLDSTEIN STEPLENGTHS*[1)]

Nai-yang Deng Zheng-feng Li
(*Department of Mathematics, Beijing Agricultural Engineering University, Beijing 100083, China*)

Abstract

The line search strategy is crucial for an efficient unconstrained optimization algorithm. One of the reason why the Wolfe line searches is recommended lies in that it ensures positive definiteness of BFGS updates. When gradient information has to be obtained costly, the Armijo-Goldstein line searches may be preferred. To maintain positive difiniteness of BFGS updates based on the Armijo-Goldstein line searches, a slightly modified form of BFGS update is proposed by I.D. Coope and C.J. Price (Journal of Computational Mathematics, 13 (1995), 156–160), while its convergence properties is open up to now. This paper shows that the modified BFGS algorithm is globally and superlinearly convergent based on the Armijo-Goldstein line searches.

Key words: BFGS methods, Convergence, Superlinear convergence.

1. Introduction

Assume that we are finding the minimizer of the following unconstrained optimization problem

$$\min_{x\in R^n} f(x), \tag{1.1}$$

and assume the current point is x_k. To calculate x_{k+1} from x_k by a line search method, the following iteration

$$x_{k+1} = x_k + \lambda_k p_k, \quad k = 1, 2, \cdots \tag{1.2}$$

is applied. In the BFGS algorithm the search direction p_k is chosen so that $B_k p_k = g_k$, where $g_k = \nabla f(x_k)$, the matrices B_k are defined by the update formula recurrently

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k} \tag{1.3}$$

* Received December 5, 1995.
[1)] Work supported by the National Natural Science Foundation of China and the Natural Science Foundation of Beijing.

$$s_k = x_{k+1} - x_k \tag{1.4}$$

$$y_k = g_{k+1} - g_k \tag{1.5}$$

It is well known that if B_1 is positive definite and

$$s_k^T y_k > 0 \tag{1.6}$$

then all matrices B_k, $k = 1, 2, \cdots$, generated by (1.3) are positive definite. One of the line search strategies is the Wolfe line searches which require the steplength $\lambda_k > 0$ to satisfy the inequalities

$$f(x_k + \lambda_k p_k) \leq f(x_k) + \alpha \lambda_k g_k^T p_k \tag{1.7}$$

$$g(x_k + \lambda_k p_k)^T p_k \geq \beta g_k^T p_k \tag{1.8}$$

where α and β are constants that satisfy $0 < \alpha < \beta < 1$ and $\alpha < 1/2$. It is easy to show that condition (1.8) implies that

$$s_k^T y_k \geq (\beta - 1) s_k^T g_k > 0 \tag{1.9}$$

so that the BFGS updating formula can be applied with positive definiteness being maintained automatically. A disadvantage is that to test condition (1.8) requires an extragradient evaluation at each trial value for λ_k. When gradient information has to be obtained costly, the Armijo-Goldstein line searches

$$\alpha_2 \lambda_k p_k^T g_k \leq f(x_{k+1}) - f(x_k) \leq \alpha_1 \lambda_k p_k^T g_k \tag{1.10}$$

may be preferred, where $0 < \alpha_1 < 1/2 < \alpha_2 < 1$. However, condition (1.10) does not ensure that $s_k^T y_k > 0$. To maintain positive definiteness of BFGS updates based on the Armijo-Goldstein line searches, a slightly modified form of BFGS update is proposed by I.D. Coope and C.J. Price in [2]. They require the quadratic model, $q_k(x)$, interpolating the data $q_k(x_k) = f(x_k)$, $q_k(x_{k+1}) = f(x_{k+1})$, and $\nabla q_k(x_k) = g_k$. Let

$$z_k = y_k + \frac{2(f(x_{k+1}) - f(x_k) - s_k^T g_k) - s_k^T y_k}{s_k^T s_k} s_k \tag{1.11}$$

Applying the standard BFGS update (1.3), they derive their modified BFGS update with z_k replacing y_k

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{z_k z_k^T}{s_k^T z_k} \tag{1.12}$$

Notice the condition (1.10), we have

$$s_k^T z_k = s_k^T y_k + 2(f(x_{k+1}) - f(x_k)) - s_k^T g_k) - s_k^T y_k = 2(f(x_{k+1}) - f(x_k)) - s_k^T g_k)$$

$$\geq 2(\alpha_2 - 1)s_k^T g_k > 0 \tag{1.13}$$

So, positive definiteness of the update (1.12) is now maintained. Moreover, the updating formula (1.12) is equivalent to (1.3) when the objective function is a strictly convex quadratic function. Now we are in a position to state formally Coope and Price's Algorithm (Algorithm (CP))

Algorithm (CP)

1° Select x_1 and a symmetric and positive definite matrix B_1. Set $k = 1$;

2° Compuye $g_k = g(x_k) = \nabla f(x_k)$. If $\|g_k\| = 0$, stop; otherwise, go to step 3°;

3° Set $p_k = -B_k^{-1} g_k$;

4° Compute λ_k such that (1.10) are satisfied, beginning by the trial steplength $\lambda_k = 1$;

5° Set $x_{k+1} = x_k + \lambda_k p_k$;

6° Compute $g_{k+1} = g(x_{k+1})$. If $\|g_{k+1}\| = 0$, stop; otherwise, go to step 7°;

7° Set $s_k = x_{k+1} - x_k$, $z_k = y_k + \{[2(f(x_{k+1} - f(x_k) - s_k^T g_k) - s_k^T y_k]/s_k^T s_k\} s_k$ and

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{z_k z_k^T}{z_k^T s_k}$$

8° Increase k by one, and go to step 3°.

2. Main Results

Hereafter we will suppose that the following assumption on the objective function $f(x)$ holds.

Assumption A

(i) $f(x)$ is twice continuously differentiable and x_* is a minimizer of $f(x)$;

(ii) There exist positive constants m and M such that

$$m\|z\|^2 \leq z^T G(x) z \leq M\|z\|^2 \tag{2.1}$$

for all $z \in R^n$ and $x \in D$, where the level set $D = \{x \in R^n | f(x) \leq f(x_1)\}$ is convex, $G(x)$ is the Hessian of $f(x)$ at x;

(iii) The Hessian $G(x)$ is Lipschitz continuous on D, i.e. there exists a constant L such that

$$\|G(x) - G(x')\| \leq L\|x - x'\|, \quad \text{for all } x, x' \in D \tag{2.2}$$

An immediate consequence of Assumption A is that, if we define

$$s_k = x_{k+1} - x_k, \quad y_k = g_{k+1} - g_k,$$

then we have

$$\|y_k\| \leq M\|s_k\|, \tag{2.3}$$

$$m\|s_k\|^2 \leq s_k^T y_k \leq M\|s_k\|^2, \tag{2.4}$$

and

$$ms_k^T y_k \leq \|y_k\|^2 \leq Ms_k^T y_k, \tag{2.5}$$

To prove the global and superlinear convergence of Algorithm (CP), with the techniques due to Byrd and Nocedal[1], we only need to show that the analogues of the theorem 2.1 ([1], pp.729–730), the theorem 3.2 ([1], 734–735) and their respective propositions are true, the only differences lie in the fact we are replacing the difference y_k of gradients with z_k defined by (1.13). We shall first give some lemmas.

Lemma 2.1. Let Assumption A hold. *Consider the sequence* $\{x_k\}$ *beginning at* x_1 *and having the following property: for each* $k = 1, 2, \cdots$, *there exist* $p_k \in R^n \backslash \{0\}$ *such that*

(1) $p_k^T g_k < 0$;

(2) *The steplength* $\xi_k > 0$ *satisfies*

$$f(x_k + \xi_k p_k) \leq f(x_k) + \alpha \xi_k g_k^T p_k$$

where $0 < \alpha < 1$;

(3) $x_{k+1} = x_k + \xi_k p_k = x_k + s_k$

Then $\sum_{k=1}^{\infty} \|s_k\|^2 < +\infty$.

Proof. See [4] pp.105–122.

Lemma 2.2. *Let* $\{B_k\}_{k=1}^{\infty}$ *be generated by*

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{z_k z_k^T}{z_k^T s_k}$$

where B_1 *is symmetric and positive definite and where, for all* $k \geq 1$, z_k *and* s_k *satisfy*

$$\frac{z_k^T s_k}{s_k^T s_k} \geq c > 0, \quad \frac{\|z_k\|^2}{z_k^T s_k} \leq c'$$

where c *and* c' *are constants. Then for any* $\gamma \in (0, 1)$ *there exist constants* $c_1, c_2, c_3 > 0$ *such that, for any* $k > 1$, *the relations*

$$\cos\theta_j \geq c_1$$
$$c_2 \leq \frac{\|B_j s_j\|}{\|s_j\|} \leq \frac{c_3}{c_1}$$

holds for at least $[\gamma k]$ *values of* $j \in [1, k]$, *where* θ_j *is the angle between* s_j *and* $-g_j$.

Proof. See [1] pp.729–731.

Lemma 2.3. *Let* x_1 *be a starting point for which* f *satisfies Assumption A, and suppose that* $\{x_k\}$ *is generated by* $x_{k+1} = x_k - \lambda_k B_k^{-1} g_k$, *where* λ_k *is chosen so that*

(1.10) *is satisfied. Suppose in addition that the matrices* B_k *are positive definite and that there exist* $\gamma \in (0,1)$ *and* $\alpha_3, \alpha_4 > 0$, *such that for any* $k \geq 1$, *the inequalities*

$$\cos\theta_j \geq \alpha_3;$$
$$\frac{\|B_j s_j\|}{\|s_j\|} \leq \alpha_4;$$

hold for at least $[\gamma k]$ *values of* $j \in [1,k]$, *where* $s_j = x_{j+1} - x_j$ *and* θ_j *is defined as that in Lemma* 2.2. *Then* $\{x_k\} \to x_*$, *moreover*

$$\sum_{k=1}^{\infty} \|x_k - x_*\| < \infty,$$

and there is a constant $0 \leq \tau < 1$ *such that*

$$f(x_{k+1}) - f(x_*) \leq \tau^k [f(x_1) - f(x_*)]$$

holds for all k.

Proof. See [1] pp. 732–733.

By the above Lemmas, we can prove the following theorem.

Theorem 2.1. *Let* x_1 *be a starting point for which* f *satisfies Assumption A. Then for any positive definite matrix* B_1, *the sequence* $\{x_k\}$ *generated by Algorithm* (CP), *converges to* x_*. *Moreover*

$$\sum_{k=1}^{\infty} \|x_k - x_*\| < \infty,$$

and there is constant $\tau \in [0,1]$ *such that*

$$f(x_{k+1}) - f(x_*) \leq \tau^k [f(x_1) - f(x_*)]$$

holds for sufficiently large k.

Proof. We just need to show that the hypotheses of Lemma 2.2 are satisfied. lemma 2.1 ensures that $\lim_{k\to\infty} \|s_k\| = 0$, so there exists a constant k_1 such that for $k \geq k_1$

$$\|s_k\| \leq 1. \tag{2.6}$$

By the mean value theorem, we have

$$f(x_{k+1}) = f(x_k) + s_k^T g_k + \frac{1}{2} s_k^T \int_0^1 (1-t)\nabla^2 f(x_k + t(x_{k+1} - x_k))dt s_k$$
$$s_k^T y_k = s_k^T \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))dt s_k \tag{2.7}$$

which, together with (2.1) and (2.2) yields

$$\frac{z_k^T s_k}{s_k^T s_k} = \frac{2[f(x_{k+1}) - f(x_k) - s_k^T g_k]}{s_k^T s_k} = \frac{s_k^T \int_0^1 (1-t)\nabla^2 f(x_k + t(x_{k+1} - x_k))dt s_k}{s_k^T s_k} \geq m. \tag{2.8}$$

Let

$$\phi_k = \frac{2[(f(x_{k+1} - f(x_k) - s_k^T g_k] - s_k^T y_k}{s_k^T s_k} \tag{2.9}$$

then we have by (2.7)

$$\begin{aligned}
|\phi_k| =& \Big| \frac{s_k^T \int_0^1 (1-t)\nabla^2 f(x_k + t(x_{k+1} - x_k))dt s_k - s_k^T \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))dt s_k}{s_k^T s_k} \Big| \\
=& \frac{s_k^T \int_0^1 (1-t)\nabla^2 f + t(x_{k+1} - x_k))dt - \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))dt s_k}{s_k^T s_k} \\
\leq& \frac{L\|s_k\|^3}{s_k^T s_k} = L\|s_k\|.
\end{aligned} \tag{2.10}$$

Therefore, noticing (2.3), (2.4) and (2.6), we get

$$\begin{aligned}
\frac{\|z_k\|^2}{z_k^\tau s_k} =& \frac{\|y_k\|^2 + 2\phi_k y_k^T s_k + \phi_k^2 \|s_k\|^2}{z_k^T s_k} \leq \frac{\|y_k\|^2}{m s_k^T s_k} + 2\phi_k \frac{s_k^T y_k}{m s_k^T s_k} + \frac{1}{m}\phi_k^2 \\
\leq& \frac{M^2}{m} + 2\frac{ML}{m}\|s_k\| + \frac{L^2}{m}\|s_k\|^2 \leq \frac{M^2}{m} + 2\frac{ML}{m} + \frac{L^2}{m}
\end{aligned} \tag{2.11}$$

for all $k \geq k_1$.

Then by lemma 2.2 we know that the matrices B_k satisfy the hypotheses of Lemma 2.2 and the result follows.

We now discuss the superlinear convergence of Algorithm (CP). First, we give a lemma.

Lemma 2.4. *Let $\{B_k\}_{k=1}^{\infty}$ be generated by*

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{z_k z_k^T}{z_k^T s_k}$$

where $z_k^T s_k > 0$ for all k. Furthermore assume that $\{s_k\}$ and $\{z_k\}$ are such that

$$\frac{\|z_k - G_* s_k\|}{\|s_k\|} \leq \varepsilon_k \tag{2.12}$$

for some symmetric and positive definite matrix G_*, and for some sequence $\{\varepsilon_k\}$ with the property $\sum_{k=1}^{\infty} \varepsilon_k < \infty$. Then

$$\lim_{k\to\infty} \frac{\|(B_k - G_*)s_k\|}{\|s_k\|} = 0 \tag{2.13}$$

and the sequences $\{\|B_k\|\}$, $\{\|B_k^{-1}\|\}$ are bounded.

Proof. See [1] pp.734–735.

The following theorem is crucial to prove the superlinear convergence of Algorithm (CP).

Theorem 2.2. *Let x_1 be a starting point for which f satisfies Assumption A. Then for any positive definite matrix B_1, the sequence $\{B_k\}$ generated by Algorithm (CP), satisfies the condition*

$$\lim_{k\to\infty} \frac{\|(B_k - G(x_*))s_k\|}{\|s_k\|} = 0 \tag{2.14}$$

and the sequences $\{\|B_k\|\}$, $\{\|B_k^{-1}\|\}$ are bounded.

Proof. To prove that (2.14) is true, according to Lemma 2.4, we only need to show that there exists a sequence $\{\varepsilon_k\}$ such that

$$\frac{\|z_k - G_* s_k\|}{\|s_k\|} \le \varepsilon_k \tag{2.15}$$

and

$$\sum_{k=1}^{\infty} \varepsilon_k < \infty \tag{2.16}$$

In fact

$$\frac{\|z_k - G(x_*)s_k\|}{\|s_k\|} = \frac{\|y_k - G(x_*)s_k + \phi_k s_k\|}{\|s_k\|} \le \frac{\|y_k - G(x_*)s_k\|}{\|s_k\|} + \frac{\|\phi_k s_k\|}{\|s_k\|}$$

where ϕ_k is defined by (2.9). By Assumption A, we have

$$\frac{\|y_k - G(x_*)s_k\|}{\|s_k\|} \le L \max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\}$$

which, together with (1.11), yields

$$\frac{\|z_k - G(x_*)s_k\|}{\|s_k\|} \le L' \max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\} \tag{2.17}$$

where $L' = 2L$.

Let $\varepsilon_k = L' \max\{\|x_{k+1} - x_*\|,\ \|x_k - x_*\|\}$. The inequality (2.15) is obviously true by (2.17). On the other hand, the validity of (2.16) is a conclusion of Theorem 2.1. Thus the theorem is completed.

Notice that the first trial steplength is $\lambda = 1$ in Algorithm (CP). We will show that the steplength $\lambda_k = 1$ always satisfies the condition (1.10) when k is sufficiently large. For simplicity of notation, we use the Landau symbol $a = o(\omega)$, which means that there exists a positive sequence $\{e_k\}$ with $\lim\limits_{k-\infty} e_k = 0$ such that $|a| \le e_k\omega$ for any small $\omega > 0$.

Noticing that $B_k s_k = -\lambda_k g_k$, we have from (2.14)

$$\lim_{k\to\infty} \frac{\|g_k - G(x_*)B_k^{-1}g_k\|}{\|B_k^{-1}g_k\|} = \lim_{k\to\infty} \frac{\|(B_k - G(x_*))s_k\|}{\|s_k\|} = 0 \tag{2.18}$$

Thus

$$g_k^T B_k^{-1} g_k - (B_k^{-1} g_k)^T G(x_*)(B_k^{-1} g_k) = (g_k - G(x_*) B_k^{-1} g_k)^T (B_k^{-1} g_k) = o(\|B_k^{-1} g_k\|^2)$$

and

$$g_k^T B_k^{-1} g_k = (B_k^{-1} g_k)^T G(x_*)((B_k^{-1} g_k)) + o(\|B_k^{-1} g_k\|^2) \tag{2.19}$$

Since $\|B_k^{-1}\|$ is bounded from above and $g_k \to 0$, so the value $o(\|B_k^{-1} g_k\|)$ is valid.

Therefore, by Assumption A, there exists a constant $\eta > 0$ such that for sufficiently large k

$$g_k^T B_k^{-1} g_k \geq \eta \|B_k^{-1} g_k\|^2 \tag{2.20}$$

By Taylor' formula and Assumption A,

$$\begin{aligned} f(x_k - B_k^{-1} g_k) - f(x_k) &= -g_k^T B_k^{-1} g_k \\ &\quad + \frac{1}{2}(B_k^{-1} g_k)^T \int_0^1 (1-t) \nabla^2 f(x_k + t B^{-1} g_k) dt ((B_k^{-1} g_k)) \\ &= -\frac{1}{2} g_k^T B_k^{-1} g_k + o(\|B_k^{-1} g_k\|^2) \end{aligned} \tag{2.21}$$

for some u_k between x_{k+1} and x_k. So for sufficiently large k, $\lambda_k = 1$ satisfies the condition (1.10).

Since we have shown that $\lim_{k \to \infty} \lambda_k = 1$, then applying Theorem 2.1, together with the results of Dennis and More [3, Corollary 2.3], we get the superlinear convergence of Algorithm (CP), stated as the following.

Theorem 2.3. *Let x_1 be a starting point for which f satisfies Assumption A. Then for any positive definite matrix B_1, the sequence sequence $\{x_k\}$ generated by Algorithm (CP) converges to the minimizer of $f(x)$ superlinearly.*

References

[1] R.H. Byrd, J. Nocedal, A tool for the analysis of quasi-Newton methods with application to unconstrained minimization, *SIAM J. Numer. Anal.,* **26** (1989), 727–749.

[2] I.D. Coope, C.J. Price, A modified BFGS formula maintaining positive definiteness with Armijo-Goldstein steplengths, *Journal of Computational Mathematics,* **13** (1995), 156–160.

[3] J.E. Dennis, Jr, J.J. Moré, A characterization of superlinear convergence and its application to quasi-Newton methods, *Mathematics of Computation,* **28**:9 (1974), 549–560.

[4] N.Y. Deng, Z.F. Li, Some global convergence properties of a conic-variable metric algorithm for minimization with inexact line searches, *Optimization methods and Softwares,* **5** (1995), 105–122.