SIAM J. OPTIM.
Vol. 14, No. 4, pp. 1043–1056

A NONMONOTONE LINE SEARCH TECHNIQUE AND ITS APPLICATION TO UNCONSTRAINED OPTIMIZATION*

HONGCHAO ZHANG† AND WILLIAM W. HAGER†

Abstract. A new nonmonotone line search algorithm is proposed and analyzed. In our scheme, we require that an average of the successive function values decreases, while the traditional nonmonotone approach of Grippo, Lampariello, and Lucidi [*SIAM J. Numer. Anal.*, 23 (1986), pp. 707–716] requires that a maximum of recent function values decreases. We prove global convergence for nonconvex, smooth functions, and R-linear convergence for strongly convex functions. For the L-BFGS method and the unconstrained optimization problems in the CUTE library, the new nonmonotone line search algorithm used fewer function and gradient evaluations, on average, than either the monotone or the traditional nonmonotone scheme.

Key words. nonmonotone line search, R-linear convergence, unconstrained optimization, L-BFGS method

AMS subject classifications. 90C06, 90C26, 65Y20

DOI. 10.1137/S1052623403428208

1. Introduction.

We consider the unconstrained optimization problem

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

where $f : \Re^n \mapsto \Re$ is continuously differentiable. Many iterative methods for (1.1) produce a sequence $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$, $\ldots$, where $\mathbf{x}_{k+1}$ is generated from $\mathbf{x}_k$, the current direction $\mathbf{d}_k$, and the stepsize $\alpha_k > 0$ by the rule

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k.$$

In monotone line search methods, α_k is chosen so that $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$. In nonmonotone line search methods, some growth in the function value is permitted. As pointed out by many researchers (for example, see [4, 16]), nonmonotone schemes can improve the likelihood of finding a global optimum; also, they can improve convergence speed in cases where a monotone scheme is forced to creep along the bottom of a narrow curved valley. Encouraging numerical results have been reported [6, 8, 11, 14, 15, 16] when nonmonotone schemes were applied to difficult nonlinear problems.

The earliest nonmonotone line search framework was developed by Grippo, Lampariello, and Lucidi in [7] for Newton's methods. Their approach was roughly the following: Parameters λ_1, λ_2, σ, and δ are introduced where $0 < \lambda_1 < \lambda_2$ and $\sigma, \delta \in (0, 1)$, and they set $\alpha_k = \bar{\alpha}_k \sigma^{h_k}$ where $\bar{\alpha}_k \in (\lambda_1, \lambda_2)$ is the "trial step" and h_k is the smallest nonnegative integer such that

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \le \max_{0\le j\le m_k} f(\mathbf{x}_{k-j}) + \delta\alpha_k \nabla f(\mathbf{x}_k)\mathbf{d}_k. \tag{1.2}$$

*Received by the editors May 20, 2003; accepted for publication (in revised form) October 2, 2003; published electronically May 25, 2004. This material is based upon work supported by National Science Foundation grant 0203270. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

http://www.siam.org/journals/siopt/14-4/42820.html

†PO Box 118105, Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (hzhang@math.ufl.edu, http://www.math.ufl.edu/∼hzhang; hager@math.ufl.edu, http://www.math.ufl.edu/∼hager).

Here the gradient of f at $\mathbf{x}_k$, $\nabla f(\mathbf{x}_k)$, is a row vector. The memory m_k at step k is a nondecreasing integer, bounded by some fixed integer M. More precisely,

$$m_0 = 0 \text{ and for } k > 0,\ 0 \le m_k \le \min\{m_{k-1} + 1, M\}.$$

Many subsequent papers, such as [2, 6, 8, 11, 15, 18], have exploited nonmonotone line search techniques of this nature.

Although these nonmonotone techniques based on (1.2) work well in many cases, there are some drawbacks. First, a good function value generated in any iteration is essentially discarded due to the max in (1.2). Second, in some cases, the numerical performance is very dependent on the choice of M (see [7, 15, 16]). Furthermore, it has been pointed out by Dai [4] that although an iterative method is generating R-linearly convergent iterations for a strongly convex function, the iterates may not satisfy the condition (1.2) for k sufficiently large, for any fixed bound M on the memory. Dai's example is

$$f(x) = \frac{1}{2}x^2, \quad x \in \Re, \quad x_0 \ne 0, \quad d_k = -x_k, \quad \text{and} \tag{1.3}$$

$$\alpha_k = \begin{cases} 1 - 2^{-k} & \text{if } k = i^2 \text{ for some integer } i, \\ 2 & \text{otherwise.} \end{cases}$$

The iterates converge R-superlinearly to the minimizer $x^* = 0$; however, condition (1.2) is not satisfied for k sufficiently large and any fixed M.

Our nonmonotone line search algorithm, which was partly studied in the first author's masters thesis [17], has the same general form as the scheme of Grippo, Lampariello, and Lucidi, except that their "max" is replaced by an average of function values. More precisely, our nonmonotone line search algorithm is the following:

Nonmonotone Line Search Algorithm (NLSA).

- **Initialization:** Choose starting guess $\mathbf{x}_0$, and parameters $0 \le \eta_{\min} \le \eta_{\max} \le 1$, $0 < \delta < \sigma < 1 < \rho$, and $\mu > 0$. Set $C_0 = f(\mathbf{x}_0)$, $Q_0 = 1$, and $k = 0$.
- **Convergence test:** If $\|\nabla f(\mathbf{x}_k)\|$ sufficiently small, then stop.
- **Line search update:** Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ where α_k satisfies either the (nonmonotone) Wolfe conditions:

$$f(\mathbf{x}_k + \alpha_k \mathbf{d}_k) \le C_k + \delta \alpha_k \nabla f(\mathbf{x}_k)\mathbf{d}_k, \tag{1.4}$$

$$\nabla f(\mathbf{x}_k + \alpha_k \mathbf{d}_k)\mathbf{d}_k \ge \sigma \nabla f(\mathbf{x}_k)\mathbf{d}_k, \tag{1.5}$$

  or the (nonmonotone) Armijo conditions: $\alpha_k = \bar{\alpha}_k \rho^{h_k}$, where $\bar{\alpha}_k > 0$ is the trial step, and h_k is the largest integer such that (1.4) holds and $\alpha_k \le \mu$.
- **Cost update:** Choose $\eta_k \in [\eta_{\min}, \eta_{\max}]$, and set

$$Q_{k+1} = \eta_k Q_k + 1, \qquad C_{k+1} = (\eta_k Q_k C_k + f(\mathbf{x}_{k+1}))/Q_{k+1}. \tag{1.6}$$

  Replace k by $k+1$ and return to the convergence test.

Observe that C_{k+1} is a convex combination of C_k and $f(\mathbf{x}_{k+1})$. Since $C_0 = f(\mathbf{x}_0)$, it follows that C_k is a convex combination of the function values $f(\mathbf{x}_0), f(\mathbf{x}_1), \ldots, f(\mathbf{x}_k)$. The choice of η_k controls the degree of nonmonotonicity. If $\eta_k = 0$ for each k, then the line search is the usual monotone Wolfe or Armijo line search. If $\eta_k = 1$ for each k, then $C_k = A_k$, where

$$A_k = \frac{1}{k+1}\sum_{i=0}^{k} f_i, \quad f_i = f(\mathbf{x}_i),$$

is the average function value. The scheme with $C_k = A_k$ was suggested to us by Yu-hong Dai. In [9], the possibility of comparing the current function value with an average of M previous function values was also analyzed; however, since M is fixed, not all previous function values are averaged together as in (1.6). As we show in Lemma 1.1, for any choice of $\eta_k \in [0,1]$, C_k lies between f_k and A_k, which implies that the line search update is well-defined. As η_k approaches 0, the line search closely approximates the usual monotone line search, and as η_k approaches 1, the scheme becomes more nonmonotone, treating all the previous function values with equal weight when we compute the average cost value C_k.

LEMMA 1.1. *If $\nabla f(\mathbf{x}_k)\mathbf{d}_k \le 0$ for each k, then for the iterates generated by the nonmonotone line search algorithm, we have $f_k \le C_k \le A_k$ for each k. Moreover, if $\nabla f(\mathbf{x}_k)\mathbf{d}_k < 0$ and $f(\mathbf{x})$ is bounded from below, then there exists α_k satisfying either the Wolfe or Armijo conditions of the line search update.*

Proof. Defining $D_k : \Re \to \Re$ by

$$D_k(t) = \frac{tC_{k-1} + f_k}{t+1},$$

we have

$$D_k'(t) = \frac{C_{k-1} - f_k}{(t+1)^2}.$$

Since $\nabla f(\mathbf{x}_k)\mathbf{d}_k \le 0$, it follows from (1.4) that $f_k \le C_{k-1}$, which implies that $D_k'(t) \ge 0$ for all $t \ge 0$. Hence, D_k is nondecreasing, and $f_k = D_k(0) \le D_k(t)$ for all $t \ge 0$. In particular, taking $t = \eta_{k-1}Q_{k-1}$ gives

$$f_k = D_k(0) \le D_k(\eta_{k-1}Q_{k-1}) = C_k. \tag{1.7}$$

This establishes the lower bound for C_k in Lemma 1.1.

The upper bound $C_k \le A_k$ is proved by induction. For $k = 0$, this holds by the initialization $C_0 = f(\mathbf{x}_0)$. Now assume that $C_j \le A_j$ for all $0 \le j < k$. By (1.6), the initialization $Q_0 = 1$, and the fact that $\eta_k \in [0,1]$, we have

$$Q_{j+1} = 1 + \sum_{i=0}^{j} \prod_{m=0}^{i} \eta_{j-m} \le j + 2. \tag{1.8}$$

Since D_k is monotone nondecreasing, (1.8) implies that

$$C_k = D_k(\eta_{k-1}Q_{k-1}) = D_k(Q_k - 1) \le D_k(k). \tag{1.9}$$

By the induction step,

$$D_k(k) = \frac{kC_{k-1} + f_k}{k+1} \le \frac{kA_{k-1} + f_k}{k+1} = A_k. \tag{1.10}$$

Relations (1.9) and (1.10) imply the upper bound of C_k in Lemma 1.1.

Since both the standard Wolfe and Armijo conditions can be satisfied when $\nabla f(\mathbf{x}_k)\mathbf{d}_k < 0$ and $f(\mathbf{x})$ is bounded from below, and since $f_k \le C_k$, it follows that for each k, α_k can be chosen to satisfy either the Wolfe or the Armijo line search conditions in the nonmonotone line search algorithm. $\square$

Our paper is organized as follows: In section 2 we prove global convergence under appropriate conditions on the search directions. In section 3 necessary and sufficient conditions for R-linear convergence are established. In section 4 we implement our scheme in the context of Nocedal's L-BFGS quasi-Newton method [10, 13], and we give numerical comparisons using the unconstrained problems in the CUTE test problem library [3].

2. Global convergence. To begin, we give a lower bound for the step generated by the nonmonotone line search algorithm. Here and elsewhere, $\|\cdot\|$ denotes the Euclidean norm, and $\mathbf{g}_k = \nabla f(\mathbf{x}_k)^\mathsf{T}$, a column vector.

LEMMA 2.1. *Suppose the nonmonotone line search algorithm is employed in a case where $\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \le 0$ and ∇f satisfies the following Lipschitz conditions with Lipschitz constant L:*

1. $\|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\| \le L\|\mathbf{x}_{k+1} - \mathbf{x}_k\|$ *if the Wolfe conditions are used, or*
2. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}_k)\| \le L\|\mathbf{x} - \mathbf{x}_k\|$ *for all $\mathbf{x}$ on the line segment connecting $\mathbf{x}_k$ and $\mathbf{x}_k + \alpha_k\rho\mathbf{d}_k$ if the Armijo condition is used and $\rho\alpha_k \le \mu$.*

If the Wolfe conditions are satisfied, then

$$\alpha_k \ge \left(\frac{1-\sigma}{L}\right)\frac{|\mathbf{g}_k^\mathsf{T}\mathbf{d}_k|}{\|\mathbf{d}_k\|^2}. \tag{2.1}$$

If the Armijo conditions are satisfied, then

$$\alpha_k \ge \min\left\{\frac{\mu}{\rho},\ \left(\frac{2(1-\delta)}{L\rho}\right)\frac{|\mathbf{g}_k^\mathsf{T}\mathbf{d}_k|}{\|\mathbf{d}_k\|^2}\right\}. \tag{2.2}$$

Proof. We consider the lower bounds (2.1) and (2.2) in the following two cases.

Case 1. Suppose that α_k satisfies the Wolfe conditions. By (1.5), we have

$$(\nabla f(\mathbf{x}_k + \alpha_k\mathbf{d}_k) - \nabla f(\mathbf{x}_k))\mathbf{d}_k \ge (\sigma - 1)\nabla f(\mathbf{x}_k)\mathbf{d}_k.$$

Since $\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \le 0$ and $\sigma < 1$, $(\sigma-1)\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \ge 0$, and by the Lipschitz continuity of f,

$$\alpha_k L\|\mathbf{d}_k\|^2 \ge (\sigma-1)\mathbf{g}_k^\mathsf{T}\mathbf{d}_k,$$

which implies (2.1).

Case 2. Suppose that α_k satisfies the Armijo conditions. If $\rho\alpha_k \ge \mu$, then $\alpha_k \ge \mu/\rho$, which gives (2.2). Conversely, if $\rho\alpha_k < \mu$, then since h_k is the largest integer such that $\alpha_k = \bar{\alpha}_k\rho^{h_k}$ satisfies (1.4) and since $f_k \le C_k$, we have

$$f(\mathbf{x}_k + \rho\alpha_k\mathbf{d}_k) > C_k + \delta\rho\alpha_k\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \ge f(\mathbf{x}_k) + \delta\rho\alpha_k\mathbf{g}_k^\mathsf{T}\mathbf{d}_k. \tag{2.3}$$

When ∇f is Lipschitz continuous,

$$\begin{aligned} f(\mathbf{x}_k + \alpha\mathbf{d}_k) - f(\mathbf{x}_k) &= \alpha\mathbf{g}_k^\mathsf{T}\mathbf{d}_k + \int_0^\alpha [\nabla f(\mathbf{x}_k + t\mathbf{d}_k) - \nabla f(\mathbf{x}_k)]\mathbf{d}_k\, dt \\ &\le \alpha\mathbf{g}_k^\mathsf{T}\mathbf{d}_k + \int_0^\alpha tL\|\mathbf{d}_k\|^2\, dt \\ &= \alpha\mathbf{g}_k^\mathsf{T}\mathbf{d}_k + \frac{1}{2}L\alpha^2\|\mathbf{d}_k\|^2. \end{aligned}$$

Combining this with (2.3) gives (2.2). ∎

Our global convergence result utilizes the following assumption (see, for example, [4, 7]) concerning the search directions.

Direction Assumption. There exist positive constants c_1 and c_2 such that

$$\mathbf{g}_k^\mathsf{T}\mathbf{d}_k \le -c_1\|\mathbf{g}_k\|^2, \tag{2.4}$$

and

$$\|\mathbf{d}_k\| \le c_2 \|\mathbf{g}_k\| \tag{2.5}$$

for all sufficiently large k.

THEOREM 2.2. *Suppose $f(\mathbf{x})$ is bounded from below and the direction assumption holds. Moreover, if the Wolfe conditions are used, we assume that ∇f is Lipschitz continuous, with Lipschitz constant L, on the level set*

$$\mathcal{L} = \{\mathbf{x} \in \Re^n : f(\mathbf{x}) \le f(\mathbf{x}_0)\}.$$

Let $\bar{\mathcal{L}}$ denote the collection of $\mathbf{x} \in \Re^n$ whose distance to $\mathcal{L}$ is at most $\mu d_{\max}$, where $d_{\max} = \sup_k \|\mathbf{d}_k\|$. If the Armijo conditions are used, we assume that ∇f is Lipschitz continuous, with Lipschitz constant L, on $\bar{\mathcal{L}}$. Then the iterates $\mathbf{x}_k$ generated by the nonmonotone line search algorithm have the property that

$$\liminf_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| = 0. \tag{2.6}$$

Moreover, if $\eta_{\max} < 1$, then

$$\lim_{k\to\infty} \nabla f(\mathbf{x}_k) = \mathbf{0}. \tag{2.7}$$

Hence, every convergent subsequence of the iterates approaches a point $\mathbf{x}^*$, where $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Proof. We first show that

$$f_{k+1} \le C_k - \beta \|\mathbf{g}_k\|^2, \tag{2.8}$$

where

$$\beta = \min\left\{ \frac{\delta\mu c_1}{\rho},\ \frac{2\delta(1-\delta)c_1^2}{L\rho c_2^2},\ \frac{\delta(1-\sigma)c_1^2}{L c_2^2} \right\}. \tag{2.9}$$

Case 1. If the Armijo conditions are used and $\rho\alpha_k \ge \mu$, then $\alpha_k \ge \mu/\rho$. By (1.4) and (2.4), it follows that

$$f_{k+1} \le C_k + \delta\alpha_k \mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k \le C_k - \delta\alpha_k c_1 \|\mathbf{g}_k^{\mathsf{T}}\|^2 \le C_k - \frac{\delta\mu c_1}{\rho}\|\mathbf{g}_k\|^2,$$

which implies (2.8).

Case 2. If the Armijo conditions are used and $\rho\alpha_k \le \mu$, then by (2.2),

$$\alpha_k \ge \left(\frac{2(1-\delta)}{L\rho}\right) \frac{|\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k|}{\|\mathbf{d}_k\|^2}, \tag{2.10}$$

and by (1.4), we have

$$f_{k+1} \le C_k - \left(\frac{2\delta(1-\delta)}{L\rho}\right)\left(\frac{\mathbf{g}_k^{\mathsf{T}}\mathbf{d}_k}{\|\mathbf{d}_k\|}\right)^2. \tag{2.11}$$

Finally, by (2.4) and (2.5),

$$f_{k+1} \le C_k - \left(\frac{2\delta(1-\delta)c_1^2}{L\rho c_2^2}\right)\|\mathbf{g}_k\|^2, \tag{2.12}$$

which implies (2.8).

Case 3. If the Wolfe conditions are used, then the analysis is the same as in Case 2, except that the lower bound (2.10) is replaced by the corresponding lower bound (2.1).

Combining the cost update relation (1.6) and the upper bound (2.8),

$$\begin{aligned} C_{k+1} &= \frac{\eta_k Q_k C_k + f_{k+1}}{Q_{k+1}} \\ &\le \frac{\eta_k Q_k C_k + C_k - \beta\|\mathbf{g}_k\|^2}{Q_{k+1}} = C_k - \frac{\beta\|\mathbf{g}_k\|^2}{Q_{k+1}}. \end{aligned} \tag{2.13}$$

Since f is bounded from below and $f_k \le C_k$ for all k, we conclude that C_k is bounded from below. It follows from (2.13) that

$$\sum_{k=0}^{\infty} \frac{\|\mathbf{g}_k\|^2}{Q_{k+1}} < \infty. \tag{2.14}$$

If $\|\mathbf{g}_k\|$ were bounded away from 0, (2.14) would be violated since $Q_{k+1} \le k+2$ by (1.8). Hence, (2.6) holds. If $\eta_{\max} < 1$, then by (1.8),

$$Q_{k+1} = 1 + \sum_{j=0}^{k} \prod_{i=0}^{j} \eta_{k-i} \le 1 + \sum_{j=0}^{k} \eta_{\max}^{j+1} \le \sum_{j=0}^{\infty} \eta_{\max}^{j} = \frac{1}{1-\eta_{\max}}. \tag{2.15}$$

Consequently, (2.14) implies (2.7). □

REMARK. *The bound condition $\alpha_k \le \mu$ in the Armijo conditions of the line search update can be removed if ∇f satisfies the Lipschitz condition slightly outside of $\mathcal{L}$. In the proof of Theorem* 2.2, *this bound ensures that when $\rho\alpha_k < \mu$, the point $\mathbf{x}_k + \rho\alpha_k\mathbf{d}_k$ lies in the region $\bar{\mathcal{L}}$, where ∇f is Lipschitz continuous, which is required for establishing Lemma* 2.1.

Similar to [4], a slightly different global convergence result is obtained when (2.5) is replaced by the following growth condition on $\mathbf{d}_k$: There exist positive constants τ_1 and τ_2 such that

$$\|\mathbf{d}_k\|^2 \le \tau_1 + \tau_2 k \tag{2.16}$$

for each k.

COROLLARY 2.3. *Suppose $\eta_{\max} < 1$ and all the assumptions of Theorem* 2.2 *are in effect except the direction assumption which is replaced by* (2.4) *and* (2.16). *If $\tau_2 \ne 0$, then*

$$\liminf_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| = 0. \tag{2.17}$$

If $\tau_2 = 0$, then

$$\lim_{k\to\infty} \|\nabla f(\mathbf{x}_k)\| = 0. \tag{2.18}$$

Proof. We assume, without loss of generality, that $\tau_1 \ge 1$. The analysis is identical to that given in the proof of Theorem 2.2 except that the bound $\|\mathbf{d}_k\| \le c_2\|\mathbf{g}_k\|$ used in the transition from (2.11) to (2.12) is replaced by the bound (2.16). As a result, the inequality (2.8) is replaced by

$$f_{k+1} \le C_k - \left(\frac{\beta_1}{\tau_1 + \tau_2 k}\right)\|\mathbf{g}_k\|^{l_k}, \tag{2.19}$$

where $l_k = 2$ in Case 1, $l_k = 4$ in Cases 2 and 3, and

$$\beta_1 = \min\left\{\frac{\delta\mu c_1}{\rho},\ \frac{2\delta(1-\delta)c_1^2}{L\rho},\ \frac{\delta(1-\sigma)c_1^2}{L}\right\}.$$

Using the upper bound (2.19) for $f(\mathbf{x}_{k+1})$ in the series of inequalities (2.13) gives

$$C_{k+1} \le C_k - \left(\frac{\beta_1}{Q_k(\tau_1+\tau_2 k)}\right)\|\mathbf{g}_k\|^{l_k}.$$

By (2.15),

$$C_{k+1} \le C_k - \left(\frac{\beta_1(1-\eta_{\max})}{\tau_1+\tau_2 k}\right)\|\mathbf{g}_k\|^{l_k}. \tag{2.20}$$

Since f is bounded from below and $C_k \ge f_k$, we obtain (2.17) when $\tau_2 \ne 0$ and (2.18) when $\tau_2 = 0$. This completes the proof. $\square$

3. Linear convergence. In [4] Dai proves R-linear convergence for the nonmonotone max-based line search scheme (1.2), when the cost function is strongly convex. Similar to [4], we now establish R-linear convergence for our nonmonotone line search algorithm when f is strongly convex. Recall that f is strongly convex if there exists a scalar $\gamma > 0$ such that

$$f(\mathbf{x}) \ge f(\mathbf{y}) + \nabla f(\mathbf{y})(\mathbf{x}-\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{y}\|^2 \tag{3.1}$$

for all $\mathbf{x}$ and $\mathbf{y} \in \Re^n$. After interchanging $\mathbf{x}$ and $\mathbf{y}$ and adding,

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))(\mathbf{x}-\mathbf{y}) \ge \frac{1}{\gamma}\|\mathbf{x}-\mathbf{y}\|^2. \tag{3.2}$$

If $\mathbf{x}^*$ denotes the unique minimizer of f, it follows from (3.2), with $\mathbf{y} = \mathbf{x}^*$, that

$$\|\mathbf{x}-\mathbf{x}^*\| \le \gamma\|\nabla f(\mathbf{x})\|. \tag{3.3}$$

For $t \in [0,1]$, define $\mathbf{x}(t) = \mathbf{x}^* + t(\mathbf{x}-\mathbf{x}^*)$. Since f is convex, $f(\mathbf{x}(t))$ is a convex function of t, and the derivative $f'(\mathbf{x}(t))$ is an increasing function of $t \in [0,1]$ with $f'(\mathbf{x}(0)) = 0$. Hence, for $t \in [0,1]$, $f'(\mathbf{x}(t))$ attains its maximum value at $t = 1$. This observation combined with (3.3) gives

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{x}^*) &= \int_0^1 f'(\mathbf{x}(t))dt \le f'(\mathbf{x}(1)) = \nabla f(\mathbf{x})(\mathbf{x}-\mathbf{x}^*) \\ &\le \|\nabla f(\mathbf{x})\|\|\mathbf{x}-\mathbf{x}^*\| \le \gamma\|\nabla f(\mathbf{x})\|^2. \end{aligned} \tag{3.4}$$

THEOREM 3.1. *Suppose that f is strongly convex with unique minimizer $\mathbf{x}^*$, the search directions $\mathbf{d}_k$ in the nonmonotone line search algorithm satisfy the direction assumption, there exist $\mu > 0$ such that $\alpha_k \le \mu$ for all k, $\eta_{\max} < 1$, and ∇f is Lipschitz continuous on bounded sets. Then there exists $\theta \in (0,1)$ such that*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \theta^k(f(\mathbf{x}_0) - f(\mathbf{x}^*)) \tag{3.5}$$

for each k.

Proof. Since $f(\mathbf{x}_{k+1}) \le C_k$ and C_{k+1} is a convex combination of C_k and $f(\mathbf{x}_{k+1})$, we have $C_{k+1} \le C_k$ for each k. Hence,

$$f(\mathbf{x}_{k+1}) \le C_k \le C_{k-1} \le \cdots \le C_0 = f(\mathbf{x}_0),$$

which implies that all the iterates $\mathbf{x}_k$ are contained in the level set

$$\mathcal{L} = \{\mathbf{x} \in \Re^n : f(\mathbf{x}) \le f(\mathbf{x}_0)\}.$$

Since f is strongly convex, it follows that $\mathcal{L}$ is bounded and ∇f is Lipschitz continuous on $\mathcal{L}$. By the direction assumption and the fact that $\|\nabla f(\mathbf{x})\|$ is bounded on $\mathcal{L}$, $d_{\max} = \sup_k \|\mathbf{d}_k\| < \infty$. Let $\bar{\mathcal{L}}$ denote the collection of $\mathbf{x} \in \Re^n$ whose distance to $\mathcal{L}$ is at most $\mu d_{\max}$ and let L be a Lipschitz constant for ∇f on the $\bar{\mathcal{L}}$.

As shown in the proof of Theorem 2.2,

$$f(\mathbf{x}_{k+1}) \le C_k - \beta\|\mathbf{g}_k\|^2, \tag{3.6}$$

where β is given in (2.9). Also, by the direction assumption and the upper bound μ on α_k, $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$ satisfies

$$\|\mathbf{x}_{k+1} - \mathbf{x}_k\| = \alpha_k \|\mathbf{d}_k\| \le \mu c_2 \|\mathbf{g}_k\|.$$

Combining this with the Lipschitz continuity of ∇f gives

$$\|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)\| = \|\mathbf{g}_{k+1} - \mathbf{g}_k\| \le L\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \le \mu c_2 L \|\mathbf{g}_k\|,$$

from which it follows that

$$\|\mathbf{g}_{k+1}\| \le \|\mathbf{g}_{k+1} - \mathbf{g}_k\| + \|\mathbf{g}_k\| \le b\|\mathbf{g}_k\|, \quad b = 1 + \mu c_2 L. \tag{3.7}$$

We now show that for each k,

$$C_{k+1} - f(\mathbf{x}^*) \le \theta(C_k - f(\mathbf{x}^*)), \tag{3.8}$$

where

$$\theta = 1 - \beta b_2 (1 - \eta_{\max}) \quad \text{and} \quad b_2 = \frac{1}{\beta + \gamma b^2}.$$

This immediately yields (3.5) since $f(\mathbf{x}_k) \le C_k$ and $C_0 = f(\mathbf{x}_0)$.

Case 1. $\|\mathbf{g}_k\|^2 \ge b_2(C_k - f(\mathbf{x}^*)$. By the cost update formula (1.6), we have

$$C_{k+1} - f(\mathbf{x}^*) = \frac{\eta_k Q_k (C_k - f(\mathbf{x}^*)) + (f_{k+1} - f(\mathbf{x}^*))}{1 + \eta_k Q_k}. \tag{3.9}$$

Utilizing (3.6) gives

$$\begin{aligned} C_{k+1} - f(\mathbf{x}^*) &\le \frac{\eta_k Q_k (C_k - f(\mathbf{x}^*)) + (C_k - f(\mathbf{x}^*)) - \beta\|\mathbf{g}_k\|^2}{1 + \eta_k Q_k} \\ &= C_k - f(\mathbf{x}^*) - \frac{\beta\|\mathbf{g}_k\|^2}{Q_{k+1}}. \end{aligned}$$

Since $Q_{k+1} \le 1/(1 - \eta_{\max})$ by (2.15), it follows that

$$C_{k+1} - f(\mathbf{x}^*) \le C_k - f(\mathbf{x}^*) - \beta(1 - \eta_{\max})\|\mathbf{g}_k\|^2.$$

Since $\|\mathbf{g}_k\|^2 \ge b_2(C_k - f(\mathbf{x}^*))$, (3.8) has been established in Case 1.

Case 2. $\|\mathbf{g}_k\|^2 < b_2(C_k - f(\mathbf{x}^*))$. By (3.4) and (3.7), we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \le \gamma\|\mathbf{g}_{k+1}\|^2 \le \gamma b^2\|\mathbf{g}_k\|^2.$$

And by the Case 2 bound for $\|\mathbf{g}_k\|$, this gives

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \le \gamma b^2 b_2(C_k - f(\mathbf{x}^*)).$$

Inserting this bound for $f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)$ in (3.9) yields

$$\begin{aligned} C_{k+1} - f(\mathbf{x}^*) &\le \frac{(\eta_k Q_k + \gamma b^2 b_2)(C_k - f(\mathbf{x}^*))}{1+\eta_k Q_k} \\ &= \left(1 - \frac{1-\gamma b^2 b_2}{Q_{k+1}}\right)(C_k - f(\mathbf{x}^*)). \end{aligned} \tag{3.10}$$

Rearranging the expression for b_2, we have $\gamma b^2 b_2 = 1 - \beta b_2$. Inserting this relation in (3.10) and again utilizing the bound (2.15), we obtain (3.8).

This completes the proof of (3.8), and as indicated above, the linear convergence estimate (3.5) follows directly. ∎

In the introduction, example (1.3) revealed that linearly convergent iterates may not satisfy (1.2) for any fixed choice of the memory M. We now show that with our choice for C_k, we can always satisfy (1.4), when k is sufficiently large, provided η_k is close enough to 1. We begin with a lower bound for $f(\mathbf{x}) - f(\mathbf{x}^*)$, analogous to the upper bound (3.4). By (3.1) with $\mathbf{y} = \mathbf{x}^*$, we have

$$f(\mathbf{x}) - f(\mathbf{x}^*) \ge \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}^*\|^2. \tag{3.11}$$

If ∇f satisfies the Lipschitz condition

$$\|\nabla f(\mathbf{x})\| = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\| \le L\|\mathbf{x} - \mathbf{x}^*\|,$$

then (3.11) gives

$$f(\mathbf{x}) - f(\mathbf{x}^*) \ge \frac{1}{2\gamma L^2}\|\nabla f(\mathbf{x})\|^2. \tag{3.12}$$

THEOREM 3.2. *Let $\mathbf{x}^*$ denote a minimizer of f and suppose that the sequence $f(\mathbf{x}_k)$, $k = 0, 1, \ldots,$ converges R-linearly to $f(\mathbf{x}^*)$; that is, there exist constants $\theta \in (0,1)$ and c such that $f(\mathbf{x}_k) - f(\mathbf{x}^*) \le c\theta^k$. Assume that the $\mathbf{x}_k$ are contained in a closed, bounded convex set K, f is strongly convex on K, satisfying* (3.1), *∇f is Lipschitz continuous on K, with Lipschitz constant L, the direction assumption holds, and the stepsize α_k is bounded by a constant μ. If $\eta_{\min} > \theta$, then* (1.4) *is satisfied for k sufficiently large, where C_k is given by the recursion* (1.6).

Proof. By (3.9) and the bound $Q_k \le k+1$ (see (1.8)), we have

$$\begin{aligned} C_k - f(\mathbf{x}^*) &= \frac{\sum_{i=0}^{k}\left[(\prod_{j=i}^{k-1}\eta_j)(f(\mathbf{x}_i) - f(\mathbf{x}^*))\right]}{Q_k} \\ &\ge \frac{\prod_{j=0}^{k-1}\eta_j}{k+1}\sum_{i=0}^{k}\left[\frac{f(\mathbf{x}_i) - f(\mathbf{x}^*)}{\prod_{j=0}^{i-1}\eta_j}\right] \\ &\ge \frac{(\eta_{\min})^k}{k+1}\phi_k, \text{ where } \phi_k = \sum_{i=0}^{k}\frac{f(\mathbf{x}_i) - f(\mathbf{x}^*)}{\prod_{j=0}^{i-1}\eta_j}. \end{aligned} \tag{3.13}$$

Here we define a product $\prod_{j=i}^{k-1} \eta_j$ to be 1 whenever the range of indices is vacuous; in particular, $\prod_{j=k}^{k-1} \eta_j = 1$. Let Φ denote the limit (possibly $+\infty$) of the positive, monotone increasing sequence $\phi_0, \phi_1, \ldots$.

By the direction assumption and (3.12), we have

$$\alpha_k \mathbf{g}_k^\mathsf{T} \mathbf{d}_k \geq -\mu c_2 \|\mathbf{g}_k\|^2 \geq -2\gamma\mu c_2 L^2 (f(\mathbf{x}_k) - f(\mathbf{x}^*)). \tag{3.14}$$

Combining the R-linear convergence of $f(\mathbf{x}_k)$ to $f(\mathbf{x}^*)$ with (3.14) gives

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) - \delta\alpha_k \mathbf{g}_k^\mathsf{T} \mathbf{d}_k &\leq c\theta^{k+1} - \delta\alpha_k \mathbf{g}_k^\mathsf{T} \mathbf{d}_k \\ &\leq c\theta^k (\theta + 2\gamma\mu c_2 L^2). \end{aligned} \tag{3.15}$$

Comparing (3.13) with (3.15), it follows that when

$$\frac{\Phi}{k+1} \geq c \left(\frac{\theta}{\eta_{\min}} \right)^k (\theta + 2\gamma\mu c_2 L^2), \tag{3.16}$$

(1.4) is satisfied. Since $\eta_{\min} > \theta$, the inequality (3.16) holds for k sufficiently large, and the proof is complete. ▯

As a consequence of Theorem 3.2, the iterates of example (1.3) satisfy the Wolfe condition (1.4) for k sufficiently large, when $\eta_k = 1$ for all k.

4. Numerical comparisons.

In this section we compare three methods:

(i) the monotone line search, corresponding to $\eta_k = 0$ in the nonmonotone line search algorithm;

(ii) the nonmonotone scheme [7] based on a maximum of recent function values;

(iii) the new nonmonotone line search algorithm based on an average function value.

In our implementation, we chose the stepsize α_k to satisfy the Wolfe conditions with $\delta = 10^{-4}$ and $\sigma = .9$. For the monotone line search scheme (i), C_k in (1.4) is replaced by $f(\mathbf{x}_k)$; in the nonmonotone scheme (ii) based on the maximum of recent function values, C_k in (1.4) is replaced by

$$\max_{0 \leq j \leq m_k} f(\mathbf{x}_{k-j}).$$

As recommended in [7], we set $m_0 = 0$ and $m_k = \min\{m_{k-1} + 1, 10\}$ for $k > 0$. Although our best convergence results were obtained by dynamically varying η_k, using values closer to 1 when the iterates were far from the optimum, and using values closer to 0 when the iterates were near an optimum, the numerical experiments reported here employ a fixed value $\eta_k = .85$, which seemed to work reasonably well for a broad class of problems.

The search directions were generated by the L-BFGS method developed by Nocedal in [13] and Liu and Nocedal in [10]; their software is available from the web page http://www.ece.northwestern.edu/∼nocedal/software.html.

We now briefly summarize how the search directions are generated: $\mathbf{d}_k = -\mathbf{B}_k^{-1} \mathbf{g}_k$, where the matrices $\mathbf{B}_k$ are given by the update

$$\begin{aligned} \mathbf{B}_{k-1}^{(0)} &= \gamma_k \mathbf{I}, \\ \mathbf{B}_{k-1}^{(l+1)} &= \mathbf{B}_{k-1}^{(l)} - \frac{\mathbf{B}_{k-1}^{(l)} \mathbf{s}_l \mathbf{s}_l^\mathsf{T} \mathbf{B}_{k-1}^{(l)}}{\mathbf{s}_l^\mathsf{T} \mathbf{B}_{k-1}^{(l)} \mathbf{s}_l} + \frac{\mathbf{y}_l^\mathsf{T} \mathbf{y}_l}{\mathbf{y}_l \mathbf{s}_l}, \quad l = 0, 1, \ldots, M_k - 1, \\ \mathbf{B}_k &= \mathbf{B}_{k-1}^{M_k}. \end{aligned}$$

We took $M_k = \min\{k, 5\}$,

$$\mathbf{y}_l = \mathbf{g}_{j_l+1} - \mathbf{g}_{j_l}, \quad \mathbf{s}_l = \mathbf{x}_{j_l+1} - \mathbf{x}_{j_l}, \quad j_l = k - M_k + l,$$

and

$$\gamma_k = \begin{cases} \frac{\|\mathbf{y}_{k-1}\|^2}{\mathbf{y}_{k-1}\mathbf{s}_{k-1}} & \text{if} \quad k > 0, \\ 1 & \text{if} \quad k = 0. \end{cases}$$

The analysis in [10] reveals that when f is twice continuously differentiable and strongly convex, with the norm of the Hessian uniformly bounded, $\mathbf{B}_k^{-1}$ is uniformly bounded, which implies that the direction assumption is satisfied.

Our numerical experiments use double precision versions of the unconstrained optimization problems in the CUTE library [3]. Altogether, there were 80 problems. Our stopping criterion was

$$\|\nabla f(\mathbf{x}_k)\|_\infty \le 10^{-6}(1 + |f(\mathbf{x}_k)|), \quad \|\mathbf{y}\|_\infty = \max_{1 \le i \le n} |y_i|,$$

except for problems PENALTY1, PENALTY2, and QUARTC, which would stop at $k = 0$ with this criterion. For these three problems, the stopping criterion was

$$\|\nabla f(\mathbf{x}_k)\|_\infty \le 10^{-8}\|\nabla f(\mathbf{x}_0)\|_\infty.$$

In Tables 4.1 and 4.2, we give the dimension (Dim) of each test problem, the number n_i of iterations, and the number n_f of function or gradient evaluations. An "F" in the table means that the line search could not be satisfied. The line search routine in the L-BFGS code, according to the documentation, is a slight modification of the code CSRCH of Moré and Thuente. In the cases where the line search failed, it reported that "Rounding errors prevent further progress. There may not be a step which satisfies the sufficient decrease and curvature conditions. Tolerances may be too small." Basically, it was not possible to satisfy the first Wolfe condition (1.4) due to rounding errors. With our nonmonotone line search algorithm, on the other hand, the value of C_k was a bit larger than either the function value $f(\mathbf{x}_k)$ used in the monotone scheme (i) or the local maximum used in (ii). As a result, we were able to satisfy (1.4) using the Moré and Thuente code, despite rounding errors, in cases where the other schemes were not successful.

We now give an overview of the numerical results reported in Tables 4.1 and 4.2. First, in many cases, the numbers of function and gradient evaluations of the three line search algorithms are identical. When comparing the monotone scheme (i) to the nonmonotone schemes (ii) and (iii), we see that either of the nonmonotone schemes was superior to the monotone scheme. In particular, there were

- 20 problems where monotone (i) was superior to nonmonotone (ii),
- 35 problems where nonmonotone (ii) was superior to monotone (i),
- 15 problems where monotone (i) was superior to nonmonotone (iii),
- 43 problems where nonmonotone (iii) was superior to monotone (i).

When comparing the nonmonotone schemes, we see that the new nonmonotone line search algorithm (iii) was superior to the previous, max-based scheme (ii). In particular, there were

- 10 problems where (ii) was superior to (iii),
- 20 problems where (iii) was superior to (ii).

As the test problems were solved, we tabulated the number of iterations where the function increased in value. We found that for either of the nonmonotone schemes (ii) or (iii), in roughly 7% of the iterations, the function value increased.

TABLE 4.1
Numerical comparisons.

Problem name	Dim	Monotone (i) n_i	Monotone (i) n_f	Maximum (ii) n_i	Maximum (ii) n_f	Average (iii) n_i	Average (iii) n_f
ARGLINA	500	2	4	2	4	2	4
ARGLINB	500	F	F	F	F	35	44
ARGLINC	500	F	F	F	F	74	111
ARWHEAD	10000	12	15	12	14	12	14
BDQRTIC	5000	129	156	180	200	162	175
BROWNAL	400	6	14	6	14	6	14
BROYDN7D	2000	662	668	660	662	660	662
BRYBND	5000	29	32	38	41	38	41
CHAINWOO	800	3578	3811	3503	3530	3223	3258
CHNROSNB	50	295	308	313	315	298	300
COSINE	1000	11	16	12	16	12	16
CRAGGLVY	5000	61	68	59	63	59	63
CURLY10	1000	990	1024	1302	1310	1482	1488
CURLY20	1000	2392	2462	2019	2025	2322	2325
CURLY30	1000	3034	3123	3052	3060	2677	2683
DECONVU	61	605	634	324	326	324	326
DIXMAANA	3000	11	13	11	13	11	13
DIXMAANB	3000	11	13	11	13	11	13
DIXMAANC	6000	12	14	12	14	12	14
DIXMAAND	6000	14	16	14	16	14	16
DIXMAANE	6000	355	368	341	343	341	343
DIXMAANF	6000	284	295	258	260	258	260
DIXMAANG	6000	300	307	297	299	297	299
DIXMAANH	6000	294	305	303	305	303	305
DIXMAANI	6000	2355	2426	2616	2618	2576	2579
DIXMAANJ	6000	251	259	272	274	272	274
DIXMAANK	6000	258	266	220	222	220	222
DIXMAANL	6000	215	220	190	192	190	192
DIXON3DQ	800	4733	4874	4515	4516	4353	4356
DQDRTIC	10000	14	23	11	17	11	17
EDENSCH	5000	22	27	28	31	28	31
EG2	1000	4	5	4	5	4	5
EIGENALS	420	4377	4549	4016	4031	4381	4396
EIGENBLS	420	4572	4698	4214	4226	4288	4301
EIGENCLS	462	3327	3416	3615	3623	3615	3623
ENGVAL1	10000	14	17	14	17	14	17
ERRINROS	50	160	176	184	191	154	162
EXTROSNB	50	13789	17217	10128	10658	10606	11427
FLETCBV2	1000	1223	1265	1419	1420	1284	1286
FLETCBV3	1000	3	11	3	11	3	11

TABLE 4.2
Numerical comparisons (continued).

Problem name	Dim	Monotone (i)		Maximum (ii)		Average (iii)	
		n_i	n_f	n_i	n_f	n_i	n_f
FLETCHBV	500	2	10	2	10	2	10
FLETCHCR	5000	25245	27605	26449	26553	26257	26515
FMINSRF2	10000	385	395	387	389	387	389
FMINSURF	10000	601	611	686	688	686	688
FREUROTH	5000	16	23	16	22	16	22
GENHUMPS	1000	1892	2418	1978	2168	1944	2187
GENROSE	2000	4169	4510	4387	4444	4309	4380
HILBERTA	200	356	388	237	243	365	371
HILBERTB	200	7	9	7	9	7	9
INDEF	500	2	10	2	10	2	10
JIMACK	82	4423	4644	5531	5552	3892	3912
LIARWHD	10000	26	30	28	32	31	34
MANCINO	100	11	15	11	15	11	15
MOREBV	10000	74	77	77	79	77	79
NCB20	3010	429	474	337	347	316	323
NONCVXU2	1000	1227	1262	1583	1591	1583	1591
NONCVXUN	1000	1936	1987	1657	1664	1657	1664
NONDIA	10000	21	27	21	26	21	26
NONDQUAR	10000	3331	3685	3625	3751	3315	3444
PENALTY1	10000	23	31	23	31	23	31
PENALTY2	200	F	F	131	136	130	133
PENALTY3	200	F	F	F	F	73	107
POWELLSG	10000	55	63	59	62	68	71
POWER	5000	297	305	302	304	302	304
QUARTC	10000	23	31	23	31	23	31
SCHMVETT	10000	20	25	21	23	21	23
SENSORS	200	25	29	26	29	26	29
SINQUAD	5000	267	329	319	371	366	431
SPARSINE	1000	6692	6989	7173	7176	6220	6227
SPARSQUR	10000	34	39	35	37	35	37
SPMSRTLS	10000	245	260	243	250	243	250
SROSENBR	10000	17	20	17	20	18	20
TESTQUAD	2000	6431	6628	4549	4551	4456	4462
TOINTGOR	50	88	94	92	93	92	93
TOINTGSS	10000	17	22	17	22	17	22
TQUARTIC	10000	24	29	25	29	25	29
TRIDIA	10000	2781	2860	2977	2980	2637	2641
VARDIM	10000	1	2	1	2	1	2
VAREIGVL	5000	18	21	18	20	18	20
WOODS	10000	15	20	21	24	21	24

Acknowledgments. The first author thanks Yu-hong Dai for interesting discussions concerning the relations between monotone and nonmonotone optimization methods. The authors gratefully acknowledge the comments and suggestions of the referees.

REFERENCES

[1] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

[2] E. G. Birgin, J. M. Martinez, and M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.

[3] I. Bongartz, A. R. Conn, N. I. M. Gould, and P. L. Toint, *CUTE: Constrained and unconstrained testing environments*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.

[4] Y. H. Dai, *On the nonmonotone line search*, J. Optim. Theory Appl., 112 (2002), pp. 315–330.

[5] Y. H. Dai, *R-linear convergence of the Barzilai and Borwein gradient method*, IMA J. Numer. Anal., 22 (2002), pp. 1–10.

[6] Y. H. Dai, *A nonmonotone conjugate gradient algorithm for unconstrained optimization*, J. Syst. Sci. Complex., 15 (2002), pp. 139–145.

[7] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM J. Numer. Anal., 23 (1986), pp. 707–716.

[8] L. Grippo, F. Lampariello, and S. Lucidi, *A truncated Newton method with nonmonotone line search for unconstrained optimization*, J. Optim. Theory Appl., 60 (1989), pp. 401–419.

[9] L. Grippo, F. Lampariello, and S. Lucidi, *A class of nonmonotone stabilization methods in unconstrained optimization*, Numer. Math., 59 (1991), pp. 779–805.

[10] D. C. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Prog., 45 (1989), pp. 503–528.

[11] S. Lucidi, F. Rochetich, and M. Roma, *Curvilinear stabilization techniques for truncated Newton methods in large-scale unconstrained optimization*, SIAM J. Optim., 8 (1998), pp. 916–939.

[12] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.

[13] J. Nocedal, *Updating quasi-Newton matrices with limited storage*, Math. Comp., 35 (1980), pp. 773–782.

[14] E. R. Panier and A. L. Tits, *Avoiding the Maratos effect by means of a nonmonotone line-search*, SIAM J. Numer. Anal., 28 (1991), pp. 1183–1195.

[15] M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7 (1997), pp. 26–33.

[16] P. L. Toint, *An assessment of non-monotone line search techniques for unconstrained optimization*, SIAM J. Sci. Comput., 17 (1996), pp. 725–739.

[17] H. Zhang, *A Few Nonmonotone Methods in Nonlinear Optimization*, Master's Thesis, ICMSEC, Chinese Academy of Sciences, Beijing, China, 2001 (in Chinese).

[18] J. L. Zhou and A. L. Tits, *Nonmonotone line search for minimax problem*, J. Optim. Theory Appl., 76 (1993), pp. 455–476.