Global convergence of nonmonotone descent methods for unconstrained optimization problems

Wenyu Sun\textsuperscript{a,*}, Jiye Han\textsuperscript{b}, Jie Sun\textsuperscript{c}

\textsuperscript{a}School of Mathematics and Computer Science, Nanjing Normal University, Nanjing 210097, People’s Republic of China
\textsuperscript{b}Institute of Applied Mathematics, Chinese Academy of Science, Beijing 100080, People’s Republic of China
\textsuperscript{c}Faculty of Business Administration and Singapore-MIT Alliance, National University of Singapore, Singapore

Received 6 February 2001; received in revised form 26 June 2001

Abstract

Global convergence results are established for unconstrained optimization algorithms that utilize a nonmonotone line search procedure. This procedure allows the user to specify a flexible forcing function and includes the nonmonotone Armijo rule, the nonmonotone Goldstein rule, and the nonmonotone Wolfe rule as special cases. © 2002 Elsevier Science B.V. All rights reserved.

MSC: 65k05; 90c30

Keywords: Unconstrained optimization; Nonmonotone line search; Global convergence

1. Introduction

A large portion of optimization methods require monotonicity of the objective values to guarantee their global convergence. This target is usually achieved by a suitable line search technique even when the initial point is far away from the optimum. Among the most popular line search techniques are the Armijo rule, the Goldstein rule and the Wolfe rule (see [3,5,16]).

Recent research [6,7,10,17,19] indicates that the monotone line search technique may have some drawbacks. In particular, enforcing monotonicity may considerably reduce the rate of convergence when the iteration is trapped near a narrow curved valley, which can result in very short steps.
or zigzagging. Therefore, it might be advantageous to allow the iterative sequence to occasionally generate points with nonmonotone objective values. Grippo et al. [6] generalized the Armijo rule and proposed a nonmonotone line search technique for Newton’s method which permits increase in function value, while retaining global convergence of the minimization algorithm. Several numerical tests show that the nonmonotone line search technique for unconstrained optimization and constrained optimization is efficient and competitive [6,7,10,12,17,19]. Note that the famous watchdog technique for constrained optimization proposed in [1] can also be viewed as strategy of the nonmonotone type.

The forcing function introduced in [11] is an important class of functions which can be used to measure sufficiency of descent and prove convergence. In [11] a detailed steplength analysis with forcing function is given. Han and Liu [8] used the idea of forcing function and proposed a general line search rule. Liu et al. [10] applied a nonmonotone technique to BFGS method. In this paper, we combine forcing functions with the nonmonotone line search technique and give a general line search rule, called the nonmonotone F-rule, for unconstrained minimization problems. We show that some common nonmonotone line search rules such as the nonmonotone Armijo line search rule, the nonmonotone Goldstein line search rule, and the nonmonotone Wolfe line search rule are special cases of the nonmonotone F-rule. Finally, we prove the global convergence of the resulted nonmonotone descent methods under mild conditions.

The remainder of this paper is organized as follows. In Section 2 we describe our nonmonotone F-rule and show that the aforementioned common nonmonotone line search rules are particular cases of the nonmonotone F-rule. In Section 3, we establish the global convergence of nonmonotone descent methods for unconstrained optimization. Some conclusions are given in Section 4.

2. The nonmonotone line search technique

Consider the unconstrained optimization problem

\[ \min_{x \in \mathbb{R}^n} f(x). \]  

(2.1)

We use notation \( g(x) = \nabla f(x) \) and \( g_k = g(x_k) = \nabla f(x_k) \).

The following assumption is imposed throughout the paper.

**Assumption 2.1.** The function \( f: \mathbb{R}^n \to \mathbb{R} \) is a continuously differentiable function and the level set

\[ \Omega = \{x \mid f(x) \leq f(x_0)\} \]

is compact.

Under this assumption, \( f \) is bounded below and the gradient function \( g(x) \) is uniformly continuous in \( \Omega \).

The iterations of the general solution method for problem (2.1) are defined as

\[ x_{k+1} = x_k + \lambda_k d_k, \quad k = 0, 1, \ldots, \]  

(2.2)

where \( x_0 \in \mathbb{R}^n \) is a given starting point, \( \lambda_k \) is a stepsize with \( \lambda_k \geq 0 \), and \( d_k \) is a search direction which satisfies \( g_k^T d_k \leq 0 \) and is determined, in general, by some gradient-type methods or Newton-type methods.
Now we describe the nonmonotone Armijo rule. Let \( a > 0, \gamma \in (0, 1), \beta \in (0, 1) \) and let \( M \) be a nonnegative integer. For each \( k \), let \( m(k) \) satisfy

\[
m(0) = 0, \quad 0 \leq m(k) \leq \min \{ m(k - 1) + 1, M \} \quad \text{for} \ k \geq 1.
\] (2.3)

Let \( \lambda_k = \beta^p a \) and \( p_k \) be the smallest nonnegative integer \( p \) such that

\[
f(x_k + \beta^p a d_k) \leq \max_{0 \leq j \leq m(k)} [f(x_{k-j})] + \gamma \beta^p ag_k^T d_k.
\] (2.4)

Set

\[x_{k+1} = x_k + \lambda_k d_k.\]

Similarly, the nonmonotone Goldstein rule can be defined as follows:

\[
f(x_k + \lambda_k d_k) \leq \max_{0 \leq j \leq m(k)} [f(x_{k-j})] + \mu_1 \lambda_k g_k^T d_k,
\] (2.5)

\[
f(x_k + \lambda_k d_k) \geq \max_{0 \leq j \leq m(k)} [f(x_{k-j})] + \mu_2 \lambda_k g_k^T d_k,
\] (2.6)

where \( 0 < \mu_1 \leq \mu_2 < 1 \).

Finally, the nonmonotone Wolfe rule can be described as follows:

\[
f(x_k + \lambda_k d_k) \leq \max_{0 \leq j \leq m(k)} [f(x_{k-j})] + \gamma_1 \lambda_k g_k^T d_k,
\] (2.7)

\[
g(x_k + \lambda_k d_k)^T d_k \geq \gamma_2 g_k^T d_k,
\] (2.8)

where \( 0 < \gamma_1 \leq \gamma_2 < 1 \).

Next, we present the nonmonotone F-rule. We begin with two definitions describing the forcing function and the reverse modulus of continuity of gradient.

**Definition 2.2.** The function \( \sigma : [0, +\infty) \to [0, +\infty) \) is a forcing function (F-function), if for any sequence \( \{t_i\} \subset [0, +\infty) \),

\[
\lim_{i \to \infty} \sigma(t_i) = 0 \quad \text{implies} \quad \lim_{i \to \infty} t_i = 0.
\] (2.9)

**Definition 2.3.** Let

\[
\eta = \sup \{ ||g(x) - g(y)|| \mid x, y \in \Omega \} > 0.
\]

Then the mapping \( \delta : [0, \infty) \to [0, \infty) \) defined by

\[
\delta(t) = \begin{cases} 
\inf \{ ||x - y|| \mid ||g(x) - g(y)|| \geq t \}, \ t \in [0, \eta), \\
\lim_{s \to \eta^-} \delta(s), \quad \text{for} \ t \in [\eta, +\infty)
\end{cases}
\]

is the reverse modulus of continuity of gradient \( g(x) \).
Now we give the nonmonotone F-rule for line searches as follows. Let $M$ be a nonnegative integer. For each $k$, let $m(k)$ satisfy
\begin{equation}
    m(0) = 0, \quad 0 \leq m(k) \leq \min[m(k-1) + 1, M] \quad \text{for } k \geq 1.
\end{equation}
Let $\lambda_k \geq 0$ be bounded above and satisfy
\begin{equation}
    f(x_k + \lambda_k d_k) \leq \max_{0 \leq j \leq m(k)} [f(x_{k-j})] - \sigma(t_k),
\end{equation}
where $\sigma$ is a forcing function and $t_k = -g_k^T d_k / \|d_k\|$. Set
\[ x_{k+1} = x_k + \lambda_k d_k. \]
Obviously, if $M = 0$, the above nonmonotone F-rule is just the rule of sufficient decrease in [11]. Note also that any nondecreasing function $\sigma : [0, \infty) \to [0, \infty)$ such that $\sigma(0) = 0$ and $\sigma(t) > 0$ for $t > 0$ is necessarily an F-function. Hence, the presented rule is quite general.

For convenience, in the following, let:
\begin{equation}
    f(x_l(k)) = \max_{0 \leq j \leq m(k)} [f(x_{k-j})],
\end{equation}
where
\[ k - m(k) \leq l(k) \leq k. \]

We now show that the nonmonotone line search rules (2.4), (2.5)–(2.6) and (2.7)–(2.8) satisfy the nonmonotone F-rule (2.11), respectively. To be concise, we only give a proof for nonmonotone Goldstein rule (2.5)–(2.6). The proofs for other schemes are similar.

**Proposition 2.4.** (1) The nonmonotone Armijo rule (2.4) satisfies the nonmonotone F-rule (2.11), where $\sigma(t) = \gamma \beta t \delta((1 - \gamma)t)$.

(2) The nonmonotone Goldstein rule (2.5)–(2.6) satisfies the nonmonotone F-rule (2.11), where $\sigma(t) = \mu_1 t \delta((1 - \mu_2)t)$.

(3) The nonmonotone Wolfe rule (2.7)–(2.8) satisfies the nonmonotone F-rule (2.11), where $\sigma(t) = \gamma_1 t \delta((1 - \gamma_2)t)$.

**Proof.** We only prove (2). The proofs of (1) and (3) are similar. From (2.6),
\begin{align*}
    f(x_k + \lambda_k d_k) &\geq f(x_l(k)) + \mu_2 \lambda_k g_k^T d_k \\
    &\geq f(x_k) + \mu_2 \lambda_k g_k^T d_k,
\end{align*}
which implies that
\begin{equation}
    g(x_k + \theta_k \lambda_k d_k)^T d_k \geq \mu_2 g_k^T d_k, \quad \theta_k \in (0, 1).
\end{equation}
Then
\begin{equation}
    (\mu_2 - 1) \frac{g_k^T d_k}{\|d_k\|} \leq \|g(x_k + \theta_k \lambda_k d_k) - g(x_k)\|.
\end{equation}
Using Definition 2.3, we have

\[ \theta_k \lambda_k \|d_k\| \geq \delta \left( (\mu_2 - 1) \left( \frac{g_k^T d_k}{\|d_k\|} \right) \right), \]

which means

\[ \lambda_k \|d_k\| \geq \delta \left( (1 - \mu_2) \left( \frac{-g_k^T d_k}{\|d_k\|} \right) \right), \quad (2.13) \]

where \( \delta(\cdot) \) is the reverse modulus. So, it follows from (2.5) and (2.13) that

\[ f(x_k + \lambda_k d_k) \leq f(x_{l(k)}) + \mu_1 \lambda_k g_k^T d_k \]

\[ = f(x_{l(k)}) - \mu_1 \lambda_k \|d_k\| \left( \frac{-g_k^T d_k}{\|d_k\|} \right) \]

\[ \leq f(x_{l(k)}) - \mu_1 \left( \frac{-g_k^T d_k}{\|d_k\|} \right) \delta \left( (1 - \mu_2) \left( \frac{-g_k^T d_k}{\|d_k\|} \right) \right) \]

\[ \leq f(x_{l(k)}) - \sigma \left( \frac{g_k^T d_k}{\|d_k\|} \right), \quad (2.14) \]

where \( \sigma(t) = \mu_1 t \delta((1 - \mu_2)t), \ t \geq 0 \). Clearly, \( \sigma(t) \) is a forcing function. This indicates that the rule (2.5)–(2.6) satisfies the nonmonotone F-rule (2.11).

3. The global convergence

In this section we establish the global convergence properties of optimization methods with non-
monotone F-rule. Note that to establish our result, we need some additional mild conditions.

**Theorem 3.1.** Let function \( f : \mathbb{R}^n \to \mathbb{R} \) satisfy Assumption 2.1. Let the sequence \( \{x_k\} \) be defined by

\[ x_{k+1} = x_k + \lambda_k d_k, \quad k = 0, 1, \ldots , \]

where the steplength \( \lambda_k \) is defined by the nonmonotone F-rule (2.11). If the direction \( d_k \) satisfies

\[ \left| \frac{-g_k^T d_k}{\|d_k\|} \right| \geq \sigma(\|g_k\|), \quad k = 0, 1, \ldots , \quad (3.1) \]

(gradient relatedness) and

\[ \|d_k\| \leq c_2 \|g_k\|, \quad (3.2) \]

where \( \sigma(\cdot) \) is a forcing function and \( c_2 > 0 \). Then the sequence \( \{x_k\} \subset \Omega \) and every accumulation point of \( \{x_k\} \) is a stationary point.
Proof. Since \( m(k + 1) \leq m(k) + 1 \), we have
\[
\begin{align*}
\max f(x_{l(k)+1}) &= \max_{0 \leq j \leq m(k+1)} \left[ f(x_{k+1-j}) \right] \\
&\leq \max_{0 \leq j \leq m(k)+1} \left[ f(x_{k+1-j}) \right] \\
&= \max \left\{ f(x_{l(k)}), f(x_{k+1}) \right\} \\
&= f(x_{l(k)}). 
\end{align*}
\tag{3.3}
\]

From (2.11), we have for \( k > M \)
\[
\begin{align*}
f(x_{l(k)}) &\leq f(x_{l(k)-1}) - \sigma(t_{l(k)-1}).
\end{align*}
\tag{3.4}
\]
The Assumption 2.1 implies that \( f \) is bounded below. Since \( f(x_k) \leq f(x_0), \forall k, \{x_k\} \subset \Omega \), so that \( \{f(x_{l(k)})\} \) converges. Therefore,
\[
\lim_{k \to \infty} \sigma(t_{l(k)-1}) = 0,
\tag{3.5}
\]
which means from Definition 2.2 that
\[
\lim_{k \to \infty} t_{l(k)-1} = \lim_{k \to \infty} \frac{-g_{l(k)-1}^T d_{l(k)-1}^T}{\|d_{l(k)-1}\|} = 0.
\tag{3.6}
\]
Using condition (3.1), we deduce
\[
\lim_{k \to \infty} \sigma(\|g_{l(k)-1}\|) = 0,
\]
which implies
\[
\lim_{k \to \infty} \|g_{l(k)-1}\| = 0
\]
from Definition 2.2. Then it follows from (3.2) that
\[
\lim_{k \to \infty} \|d_{l(k)-1}\| = 0.
\tag{3.7}
\]
Let
\[
\hat{l}(k) = l(k + M + 2).
\tag{3.8}
\]
We prove by induction that for any given \( j \geq 1 \),
\[
\lim_{k \to \infty} \|d_{l(k)-j}\| = 0
\tag{3.9}
\]
and
\[
\lim_{k \to \infty} f(x_{l(k)-j}) = \lim_{k \to \infty} f(x_{l(k)}).
\tag{3.10}
\]
If \( j = 1 \), since \( \{\hat{l}(k)\} \subset \{l(k)\} \), (3.9) and (3.10) follow from (3.7). Assume that (3.9) and (3.10) hold for a given \( j \). We consider the case of \( j + 1 \). Since
\[
f(x_{l(k)-j}) \leq f(x_{l(\hat{l}(k)-(j+1))}) - \sigma(y_{l(k)-(j+1)}),
\]
we have
\[
f(x_{l(k)-j}) \leq f(x_{l(\hat{l}(k)-(j+1))}) - \sigma(y_{l(k)-(j+1)}).
\]
using the same argument for deriving (3.7), we deduce
\[ \lim_{k \to \infty} \|d_{\hat{l}(k)-(j+1)}\| = 0. \] (3.11)
Noting that \( \Omega = \{x \mid f(x) \leq f(x_0)\} \) is compact, \( x_k + \lambda_k d_k \in \Omega, \forall k, \) and that \( \lambda_k \) stay bounded, we have \( \|x_{\hat{l}(k)-(j)} - x_{\hat{l}(k)-(j+1)}\| \to 0, \) which implies
\[ \lim_{k \to \infty} f(x_{\hat{l}(k)-(j+1)}) = \lim_{k \to \infty} f(x_{\hat{l}(k)}) \] (3.12)
from uniform continuity of \( f \) and (3.10). Therefore, (3.9) and (3.10) hold for any given \( j \geq 1. \) Now for any \( k, \)
\[ x_{k+1} = x_{\hat{l}(k)} - \sum_{j=1}^{\hat{l}(k)-k-1} \lambda_{\hat{l}(k)-j} d_{\hat{l}(k)-j}. \] (3.13)
Note that \( \hat{l}(k) - k - 1 \leq M + 1 \) and by (3.9), we obtain
\[ \lim_{k \to \infty} \|x_{k+1} - x_{\hat{l}(k)}\| = 0. \] (3.14)
Since \( \{f(x_{\hat{l}(k)})\} \) admits a limit, it follows from the uniform continuity of \( f \) on \( \Omega \) that
\[ \lim_{k \to \infty} f(x_k) = \lim_{k \to \infty} f(x_{\hat{l}(k)}). \] (3.15)
So, for
\[ f(x_{k+1}) \leq f(x_{\hat{l}(k)}) - \sigma \left( -\frac{g_k^T d_k}{\|d_k\|} \right) \]
taking limits for \( k \to \infty, \) we get
\[ \lim_{k \to \infty} \sigma \left( -\frac{g_k^T d_k}{\|d_k\|} \right) = 0, \] (3.16)
which means
\[ \lim_{k \to \infty} \|g_k\| = 0. \]

From this theorem, we obtain immediately the following corollaries.

**Corollary 3.2.** Under the assumptions of Theorem 3.1, the sequence \( \{x_k\} \) generated from optimization methods employing either the nonmonotone Armijo rule, the nonmonotone Goldstein rule, or the nonmonotone Wolfe rule remains in \( \Omega \) and every accumulation point of \( \{x_k\} \) is a stationary point.

**Proof.** This follows directly from Proposition 2.4 and Theorem 3.1.

**Corollary 3.3.** Let the function \( f : \mathbb{R}^n \to \mathbb{R} \) satisfy Assumption 2.1. Let the sequence \( \{x_k\} \) be defined by
\[ x_{k+1} = x_k + \lambda_k d_k, \quad k = 0, 1, \ldots, \]
where the steplength $\lambda_k$ is defined by nonmonotone $F$-rule (2.11) with $F$-function $\sigma(t) = (c_1/c_2)t$ and the direction $d_k$ satisfies

$$g_k^Td_k \leq -c_1\|g_k\|^2$$

(3.17)

and

$$\|d_k\| \leq c_2\|g_k\|,$$

(3.18)

where $c_1, c_2 > 0$. Then the sequence $\{x_k\} \subset \Omega$ and every accumulation point of $\{x_k\}$ is a stationary point.

**Proof.** It is enough to note that (3.17)–(3.18) satisfies (3.1)–(3.2). □

As a conclusion of this section, we consider nonmonotone Newton-type method

$$x_{k+1} = x_k + \lambda_k d_k,$$

(3.19)

where $\lambda_k$ is generated by nonmonotone $F$-rule (2.11), and

$$d_k = -H_k g_k,$$

(3.20)

where $H_k$ is an $n \times n$ symmetric positive definite matrix. Assume that there exist constants $\underline{\tau}_k > 0$ and $\bar{\tau}_k > 0$ such that

$$\underline{\tau}_k\|d\|^2 \leq d^TH_kd \leq \bar{\tau}_k\|d\|^2, \quad \forall k \quad \text{and} \quad d \in \mathbb{R}^n.$$

(3.21)

This class of methods includes Newton’s method and the quasi-Newton methods [14–16,18]. As to the convergence of this class of method (3.19)–(3.21), we have the following theorem.

**Theorem 3.4.** Let $f : \mathbb{R}^n \to \mathbb{R}$ satisfy Assumption 2.1. Consider nonmonotone Newton-type method (3.19)–(3.21). Then the sequence $\{x_k\} \subset \Omega$ and every accumulation point of $\{x_k\}$ is a stationary point.

**Proof.** Analogous to the proof of Theorem 3.1, we know that $\{f(x_l(k))\}$ admits a limit and

$$\lim_{k \to \infty} \frac{-g_{l(k)-1}^Td_{l(k)-1}}{\|d_{l(k)-1}\|} = 0,$$

(3.22)

which implies that

$$\lim_{k \to \infty} g_{l(k)-1} = 0$$

(3.23)

from (3.20)–(3.21).

Let $\hat{l}(k) = l(k + M + 2)$. Similarly, we also have

$$\lim_{k \to \infty} g_{\hat{l}(k)-j} = 0.$$
We note that
\[ \|x_{k+1} - \hat{x}^{(k)}\| \leq \sum_{j=1}^{l(k)-1} \lambda^{(k)-j} \|d^{(k)-j}\| \]
\[ \leq \sum_{j=1}^{l(k)-1} \lambda^{(k)-j} \|H^{(k)-j}\| \|g^{(k)-j}\| \]
\[ \leq \sum_{j=1}^{l(k)-1} \lambda^{(k)-j} \tilde{\tau}^{(k)-j} \|g^{(k)-j}\|. \]  
(3.25)

It follows from (3.24) that
\[ \lim_{k \to \infty} \|x_{k+1} - \hat{x}^{(k)}\| = 0. \]  
(3.26)

The following argument is the same as one in Theorem 3.1. □

**Remark 3.5.** It is easy to see that, for nonmonotone Newton-type method (3.19)–(3.21), the conditions of Theorem 3.1 hold. In fact, from (3.20)–(3.21), there exists \( \omega > 0 \) such that
\[ d_k^T g_k \leq -\omega \|g_k\| \|d_k\|, \]
which satisfies just the gradient relatedness. In addition, by (3.20)–(3.21), the vector \( d_k \) satisfies (3.2) with \( c_2 = \tilde{\tau}_k \). Therefore, the conditions of Theorem 3.1 hold.

4. Conclusion

We provide a general framework for nonmonotone descent methods to globally converge to a stationary point. This is done by simply following the nonmonotone F-rule in line searches. Convergence results are established under mild assumptions such as the boundedness of a level set and Newton-type search directions. Note that nonmonotone techniques can be used in trust region methods as well (see [2]). It would be interesting to develop some general rules for the trust region cases, which might be a topic of further study. Overall, we feel that nonmonotone method is a useful technique for optimization, and it can also be generalized to the nonquadratic model case and other optimization cases like the minimax problem, and variational inequality problems [4,9,13].

**Acknowledgements**

We are thankful to Professor Liqun Qi and two referees for their helpful comments. Their comments greatly improved our paper.
References


