A CLASS OF NONMONOTONE SPECTRAL MEMORY
GRADIENT METHOD

ZHENSHENG YU, JINSONG ZANG, AND JINGZHAO LIU

Abstract. In this paper, we develop a nonmonotone spectral memory
gradient method for unconstrained optimization, where the spectral step-
size and a class of memory gradient direction are combined efficiently.
The global convergence is obtained by using a nonmonotone line search
strategy and the numerical tests are also given to show the efficiency of
the proposed algorithm.

1. Introduction

In this paper, we consider the unconstrained optimization problem
\[ \min f(x), \quad x \in \mathbb{R}^n, \]
where \( f(x) : \mathbb{R}^n \to \mathbb{R} \) is continuously differentiable with gradient function
\[ \nabla f(x) = g(x). \]

The spectral gradient method was originally proposed by Barzilai and Bor-
wein [1] for quadratic function. The main feature of this method is that only
gradient directions are used at each line search whereas a non-monotone strat-
egy [6] guarantees global convergence. Compared with the classical steepest
descent method, the spectral gradient method requires less computational work
and speeds up the convergence greatly, hence it is suitable for large scale prob-
lems. Due to its simplicity and numerical efficiency, the spectral gradient
method has now received a good deal of attention in optimization commu-
nity. Recently, the method has been extended successfully to unconstrained
and constrained optimization, see for examples [2, 3, 4, 10].

In [2], Birgin and Martínez combined the spectral gradient and conjugate
gradient(CG) idea and proposed a spectral conjugate gradient method for un-
constrained optimization. The numerical performance showed that the method
is more efficient than the CG method. Memory gradient methods have the same

Received February 19, 2008; Revised November 24, 2008.
2000 Mathematics Subject Classification. 90C30, 65K05.
Key words and phrases. unconstrained optimization, spectral memory gradient method,
nonmonotone technique, global convergence.

This work is supported by National Natural Science Foundation of China (No.10671126)
and Shanghai Leading Academic Discipline Project (No. S30501).

©2010 The Korean Mathematical Society
property as that of CG methods, but compared with CG methods, the main
difference is that they can use the information of the previous iterations more
sufficiently and hence they are helpful to design algorithms with quick conver-
gence rate. The research for memory gradient method can be found in [5, 8,
9].

Motivated by the idea of [2], in this paper, we consider combining the spectra
gradient method with a class of memory gradient method [9] and propose a
spectral memory gradient method for problem (1). A nonmonotone Armijio-
type line search (which is motivated by the nonmonotone trust region technique
[12]) is employed to obtain the iterate sequence. Under certain conditions, the
strong global convergence of the proposed method is obtained, the numerical
tests are also given to show the efficiency of the proposed algorithm.

The paper is organized as follows: In Section 2, we propose our algorithm
and show its global convergence; In Section 3, we report the numerical tests.

2. Algorithm model and global convergence

In this section, we describe the nonmonotone spectral memory gradient al-
gorithm, we first give our memory gradient direction, where the parameter
β
k
comes from the definition in Y. Narushima and H. Yabe [9]:

Let \( d^1_k \) be defined by

\[
\begin{align*}
d^1_k &= -g_k + \beta_N Y_k d_{k-1},
\end{align*}
\]

where

\[
\beta^N Y_k = \begin{cases} 
0, & \text{if } g_k^T d_{k-1} \leq 0, \\
\frac{\|g_k\|^2}{g_k^T d_{k-1} + \|g_k\| \|d_{k-1}\|}, & \text{otherwise}
\end{cases}
\]

Our spectral memory gradient direction is defined as

(2)

\[
d_k = \theta_k d^1_k.
\]

The definition of \( \theta_k \) is defined in the following algorithm.

Algorithm 2.1.

Step 0. Given positive integer \( M \) and constant \( \varepsilon > 0, 1 > \sigma_2 \geq \sigma_1 > 0, \gamma \in (0, 1), \delta > 0, \theta_{\max} > \theta_{\min} > 0 \), choose \( \theta_0 \in [\theta_{\min}, \theta_{\max}], k = 0 \).

Step 1: Compute \( g_k \), if \( \|g_k\| \leq \varepsilon \), stop.

Step 2: Compute \( d_k \) by (2).

Step 3.1: Set \( \lambda_k = \frac{-g_k^T d_k}{\|d_k\|^2} \).

Step 3.2: Set \( x_+ = x_k + \lambda_k d_k \).

Step 3.3: Compute the largest index \( l(k) \) such that

\[
f(x_{l(k)}) = \max_{\max\{k-M+1, 0\} \leq j \leq k} f(x_j).
\]
Compute

\[
\rho_{1,k} = \begin{cases} 
  \frac{f(x_l(k)) - f(x_+)}{k}, & \text{if } k > 0, \\
  \sum_{j=l(k)} -\lambda_j g_j^T d_j, & \text{otherwise}
\end{cases}
\]

(3)

and

\[
\rho_{2,k} = \frac{f(x_k) - f(x_+)}{-\lambda_k g_k^T d_k},
\]

(4)

set

\[
\rho_k = \max\{\rho_{1,k}, \rho_{2,k}\}.
\]

(5)

If \(\rho_k \geq \gamma\), then set \(x_{k+1} = x_+, \ s_k = x_{k+1} - x_k, \ y_k = g(x_{k+1}) - g(x_k)\), and go to Step 4.

If \(\rho_k \geq \gamma\) does not hold, define \(\lambda_{\text{new}} \in [\sigma_1 \lambda_k, \sigma_2 \lambda_k]\), set \(\lambda_k = \lambda_{\text{new}}\), and go to Step 3.2.

**Step 4.** Compute \(b_k = \langle s_k, y_k \rangle\), if \(b_k \leq 0\), set \(\theta_{k+1} = \theta_{\text{max}}\), else, compute

\[
a_k = \langle s_k, s_k \rangle \quad \text{and} \quad \theta_{k+1} = \min\{\theta_{\text{max}}, \max\{\theta_{\text{min}}, a_k/b_k\}\}.
\]

**Step 5:** Set \(k := k + 1\), go to Step 1.

To establish the global convergence, we make the following assumptions:

**Assumption 2.1.**

A1: \(f(x)\) is bounded below on the level set \(L = \{x \mid f(x) \leq f(x_0)\}\).

A2: The gradient function is Lipschitz continuous on an open set \(\Omega\) that contains \(L\), i.e., there exists a constant \(L\) such that for all \(x, y \in \Omega\)

\[
\|g(x) - g(y)\| \leq L\|x - y\|.
\]

The following result gives the descent property of the direction \(d_k\).

**Lemma 1.** Let \(d_k\) be generated by Algorithm 2.1. Then we have

\[
g_k^T d_k \leq -\frac{\theta_{\text{min}}}{2}\|g_k\|^2.
\]

(6)

**Proof.** If \(\beta_k = 0\), then \(d_k = -\theta_k g_k\), since \(\theta_k \geq \theta_{\text{min}}\), we have

\[
g_k^T d_k = -\theta_k\|g_k\|^2 \leq -\frac{\theta_{\text{min}}}{2}\|g_k\|^2.
\]
If $\beta_k > 0$, then it follows from the definition of $d_k$ that
\[
g_k^T d_k = \theta_k (-\|g_k\|^2 + \beta_k g_k^T d_{k-1})
\leq \theta_k \left(-\|g_k\|^2 + \frac{\|g_k\|^2}{2 g_k^T d_{k-1}} g_k^T d_{k-1}\right)
\leq -\frac{\theta_k}{2} \|g_k\|^2
\leq -\frac{\theta_{\min}}{2} \|g_k\|^2.
\]

**Lemma 2.** Let $\{x_k\}$ be generated by Algorithm 2.1. If Assumption 2.1 holds, then there exists a positive constant $\mu$ such that
\[
\lambda_k \geq \mu - g_k^T d_k \|d_k\|^2.
\]

**Proof.** If $\lambda_k = -\frac{\delta g_k^T d_k}{\|d_k\|^2}$, then conclusion is obvious.

If $\lambda_k \leq -\frac{\delta g_k^T d_k}{\|d_k\|^2}$, then there exists a $\pi \in \left[\frac{1}{\sigma_2}, \frac{1}{\sigma_1}\right]$ such that for $x_+ = x_k + \pi \lambda_k$,
\[
\rho_{1,k} \leq \rho_k \leq \gamma,
\]
which together with yields
\[
\gamma \pi \lambda_k g_k^T d_k \leq \pi^2 \lambda_k^2 L \|d_k\|^2 + \pi \lambda_k g_k^T d_k,
\]
i.e.,
\[
\pi^2 \lambda_k^2 L \|d_k\|^2 \geq (\gamma - 1) \pi \lambda_k g_k^T d_k.
\]
Hence
\[
\lambda_k \geq \frac{(1 - \gamma) - g_k^T d_k}{2 L \pi \|d_k\|^2}.
\]

If $\mu = \min\{\delta, \frac{(1-\gamma)}{2L\pi}\}$, we get (7). \qed

Define
\[
p(k) = \begin{cases} 
l(k), & \text{if } \rho_k = \rho_{1,k} \\
k, & \text{if } \rho_k = \rho_{2,k}.
\end{cases}
\]
We will call iteration $p(k)$ the reference iteration associated with iteration $k$, then it is easy to obtain the following result.
Lemma 3. For each \( k \), we have
\[
    f(x_{p(k)}) - f(x_{k+1}) \geq -\gamma \sum_{j=p(k)}^{k} \lambda_j g_j^T d_j.
\]

Theorem 1. Let \( \{x_k\} \) be an infinite sequence generated by Algorithm 2.1, and Assumption 2.1 holds, then we have
\[
    \lim_{k \to \infty} \|g_k\| = 0.
\]

Proof. Considering the \( k \)th iteration, we see that this iteration has a reference iteration \( p(k) \), in turn, the \( p(k) \)th iteration has a reference iteration \( p(p(k)) \), \ldots, up to the point where \( x_0 \) is reached by this backward referencing process. Hence, we may construct, for each \( k \), a sequence of iteration indexed by \( p_0, p_1, \ldots, p_q \), such that
\[
    x_0 = x_{p_0}, \ x_{p_j+1} = x_{p(p_j)}, \ j = 1, 2, \ldots, q, \ x_{p_q+1} = x_{p(k)}.
\]

Note that
\[
    f(x_0) - f(x_{k+1}) = f(x_0) - f(x_{p_0+1}) + \sum_{j=1}^{q} (f(x_{p_j-1+1}) - f(x_{p_j+1})) + f(x_{p(k)}) - f(x_{k+1}).
\]

Applying Lemma 3 to each item in the right side of the above equation together with (11), we have
\[
    f(x_0) - f(x_{k+1}) \geq -\mu\gamma \sum_{j=0}^{k} \lambda_j g_j^T d_j \geq \mu\gamma \sum_{j=0}^{k} (g_j^T d_j)^2.
\]

Since \( f(x_k) \) is bounded below on \( \mathcal{L} \), we have
\[
    \sum_{j=0}^{k} (g_j^T d_j)^2 < \infty.
\]

On the other hand, from the definition of \( d_k \), we have
\[
    d_k + \theta_k g_k = \theta_k \beta_k d_{k-1},
\]

and squaring both sides of the above equation, we have
\[
    \|d_k\|^2 = \|\theta_k \beta_k d_{k-1}\|^2 - 2\theta_k g_k^T d_k - \theta_k^2 \|g_k\|^2.
\]

Therefore, we obtain that
\[
    \frac{\|d_k\|^2}{(g_k^T d_k)^2} = \frac{\|\theta_k \beta_k d_{k-1}\|^2}{(g_k^T d_k)^2} - \frac{2\theta_k g_k^T d_k}{(g_k^T d_k)^2} - \frac{\theta_k^2 \|g_k\|^2}{(g_k^T d_k)^2}
\]
\[
    = \frac{\|\theta_k \beta_k d_{k-1}\|^2}{(g_k^T d_k)^2} - \frac{2\theta_k}{g_k^T d_k} - \frac{\theta_k^2 \|g_k\|^2}{g_k^T d_k}
\]
\[
    = \frac{\|\theta_k \beta_k d_{k-1}\|^2}{(g_k^T d_k)^2} - \theta_k \left( \frac{1}{\|g_k\|} + \frac{\|g_k\|}{g_k^T d_k} \right)^2 + \theta_k \frac{1}{\|g_k\|}.
\]
$$\leq \frac{\|\theta_k \beta_k d_k - 1\|^2}{(-g_k^T d_k)^2} + \frac{\theta_k}{\|g_k\|^2}. $$

From the definition of $d_k$ we have
$$-g_k^T d_k = \theta(k\|g_k\|^2 - \beta_k d_k - 1) \geq \theta_k \beta_k \|g_k\| \|d_k - 1\|,$$
and therefore we have
$$\|d_k\|^2 \leq \frac{1 + \theta_k}{\|g_k\|^2} \leq \frac{1 + \theta_{\text{max}}}{\|g_k\|^2},$$
which means
$$\frac{\|g_k\|^2}{\|d_k\|^2} \geq \frac{1 + \theta_{\text{max}}}{\|g_k\|^2}.$$
Hence by (12) we have
$$\sum_{k=0}^{\infty} \frac{\|g_k\|^2}{1 + \theta_{\text{max}}} \leq \sum_{k=0}^{\infty} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty,$$
which implies (10). \qed

### Table 1

<table>
<thead>
<tr>
<th>Problem</th>
<th>$\theta_k = \theta_{\text{max}}$</th>
<th>$\theta_k = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_g/I_f$</td>
<td>$L_g/I_f$</td>
<td>Time</td>
</tr>
<tr>
<td>HS2</td>
<td>8/34</td>
<td>0.0448</td>
</tr>
<tr>
<td>HS26</td>
<td>9/13</td>
<td>0.1060</td>
</tr>
<tr>
<td>HS32</td>
<td>11/12</td>
<td>0.0397</td>
</tr>
<tr>
<td>HS42</td>
<td>5/9</td>
<td>0.4472</td>
</tr>
<tr>
<td>HS46</td>
<td>15/63</td>
<td>2.2526</td>
</tr>
<tr>
<td>HS47</td>
<td>31/56</td>
<td>0.0794</td>
</tr>
<tr>
<td>HS52</td>
<td>26/27</td>
<td>1.0566</td>
</tr>
<tr>
<td>HS57</td>
<td>6/13</td>
<td>0.1136</td>
</tr>
<tr>
<td>HS63</td>
<td>59/110</td>
<td>0.1079</td>
</tr>
<tr>
<td>HS76</td>
<td>3/14</td>
<td>0.7355</td>
</tr>
<tr>
<td>HS100</td>
<td>14/37</td>
<td>1.9874</td>
</tr>
<tr>
<td>HS113</td>
<td>17/30</td>
<td>1.1462</td>
</tr>
<tr>
<td>HS213</td>
<td>24/55</td>
<td>1.6886</td>
</tr>
<tr>
<td>HS240</td>
<td>12/15</td>
<td>0.3252</td>
</tr>
<tr>
<td>HS246</td>
<td>195/2108</td>
<td>0.0378</td>
</tr>
<tr>
<td>HS256</td>
<td>959/1080</td>
<td>0.4326</td>
</tr>
<tr>
<td>HS257</td>
<td>448/479</td>
<td>0.6675</td>
</tr>
<tr>
<td>HS272</td>
<td>19/27</td>
<td>0.0439</td>
</tr>
<tr>
<td>HS273</td>
<td>19/50</td>
<td>0.0434</td>
</tr>
<tr>
<td>HS280</td>
<td>387/432</td>
<td>0.2666</td>
</tr>
</tbody>
</table>
3. Numerical test

In this section, we test our nonmonotone spectral memory gradient method with MATLAB7.0, the parameters are set as follows: $M = 5$, $\sigma_1 = 0.1$, $\sigma_2 = 1.2$, $\gamma = 10^{-3}$, $\delta = 1$, $\theta_{\text{min}} = 10^{-30}$, $\theta_{\text{max}} = 10^{30}$.

We test 20 examples taken from [7, 11], the numerical results are shown in Table 1. Here HS$i$ denotes the $i$th example in [7, 11], we report the number of function evaluations ($I_f$), the number of gradient evaluations ($I_g$), CUP time (Time), and denote * if the iteration large than 5000.

We compare the method with spectral stepsize ($\theta_k = \theta_k^{sg}$) with the non-spectral stepsize ($\theta_k = 1$), the numerical results show that the performance of spectral method is more efficiency than that of non-spectral method.

Acknowledgments. The authors would like to express their thanks to the anonymous referees for his helpful comments and suggestions.

References
