A New Gradient Method with an Optimal Stepsize Property*

Y. H. Dai† and X. Q. Yang‡

Abstract

The gradient method for the symmetric positive definite linear system $Ax = b$ is as follows

$$x_{k+1} = x_k - \alpha_k g_k$$

where $g_k = Ax_k - b$ is the residual of the system at x_k and α_k is the stepsize. The stepsize $\alpha_k = \frac{2}{\lambda_1 + \lambda_n}$ is optimal in the sense that it minimizes the modulus $\|I - \alpha A\|_2$, where λ_1 and λ_n are the minimal and maximal eigenvalues of A respectively. Since λ_1 and λ_n are unknown to users, it is usual that the gradient method with the optimal stepsize is only mentioned in theory. In this paper, we will propose a new stepsize formula which tends to the optimal stepsize as $k \to \infty$. At the same time, the minimal and maximal eigenvalues, λ_1 and λ_n, of A can be obtained.

Key words: linear system, gradient method, steepest descent method, (shifted) power method.

Mathematics Subject classification: 65F10, 65F15.

1. Introduction

Consider the following symmetric positive definite (SPD) linear system

$$Ax = b, \tag{1.1}$$

*This research was initiated while the first author was visiting The Hong Kong Polytechnic University.

†State Key Laboratory of Scientific and Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, P. O. Box 2719, Beijing 100080, P. R. China. Email: *dyh@lsec.cc.ac.cn*. This author was supported by the Chinese NSF grants (No. 19801033 and 10171104) and an innovation fund of Chinese Academy of Sciences.

‡Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong. This author was supported by a grant from the Research Committee of the Hong Kong Polytechnic University (A-PC36).

where $A \in R^{n\times n}$ is an SPD matrix and $b \in R^n$. Denote by $g_k = Ax_k - b$ the residual of the system at x_k. The gradient method for solving (1.1) is of the form

$$x_{k+1} = x_k - \alpha_k g_k, \tag{1.2}$$

where x_1 is a starting point given by users and α_k is a stepsize determined by some rule. The classical steepest descent (SD) method, which can be dated back to [3], decides its stepsize as follows:

$$\alpha_k^{SD} = \frac{g_k^t g_k}{g_k^t A g_k}. \tag{1.3}$$

It is well known that the stepsize in (1.3) has the optimal property

$$\alpha_k^{SD} = \arg\min_{\alpha\in R^1} \|(I - \alpha A) g_k\|_{A^{-1}}, \tag{1.4}$$

where $\|x\|_{A^{-1}} = \sqrt{x^t A^{-1} x}$.

Here we note by (1.2) and the definition of g_k that

$$g_{k+1} = (I - \alpha_k A) g_k. \tag{1.5}$$

The SD method is a fundamental and probally the earliest method for nonlinear optimization so that it attracts many researchers' attention since its proposition. A recent and interesting report on the SD method can be found in [9].

Assume that λ_1 and λ_n are the minimal and maximal eigenvalues of A, respectively. Another choice for the stepsize is

$$\alpha_k^{OPT} \equiv \frac{2}{\lambda_1 + \lambda_n} \tag{1.6}$$

(for example, see [5]). This stepsize is such that the modulus $\|I - \alpha A\|_2$ is minimized, that is

$$\alpha_k^{OPT} = \arg\min_{\alpha\in R^1} \|I - \alpha A\|_2, \tag{1.7}$$

and hence gives the best convergence result while the analysis is made with the help of the relation

$$\|g_{k+1}\|_2 \le \|I - \alpha_k A\|_2 \|g_k\|_2. \tag{1.8}$$

We call this method as optimal stepsize gradient method in this paper. Since λ_1 and λ_n is normally unknown to users, it is usual that the stepsize (1.6) is only mentioned in theory. The performances of the optimal stepsize gradient method are then unknown to the community.

In this paper, we will consider a new choice for the stepsize α_k:

$$\alpha_k^{NEW} = \frac{\|g_k\|_2}{\|A g_k\|_2}. \tag{1.9}$$

It is known that the linear system (1.1) is equivalent to the minimization of the quadratic function $f(x) = \frac{1}{2}x^TAx - b^Tx$. Note that $g_k = \nabla f(x_k)$. The proposition of (1.9) is in connection with the estimate of the Lipschitz constant of the gradient $\nabla f(x)$:

$$\frac{1}{\alpha_k^{NEW}} = \frac{\|\nabla f(x_k - \alpha g_k) - \nabla f(x_k)\|_2}{\|(x_k - \alpha g_k) - x_k\|_2}, \quad \text{where } \alpha > 0. \tag{1.10}$$

Here we should note that, to minimize a general function $f(x)$, [10] tried to estimate the Lipschitz constant of $\nabla f(x)$ based on the two points x_{k-1} and x_k and considered the gradient method (1.2) with

$$\alpha_k = \frac{\|x_k - x_{k-1}\|_2}{2\|\nabla f(x_k) - \nabla f(x_{k-1})\|_2}. \tag{1.11}$$

However, in the quadratic case, the formula α_k will reduce to $\alpha_k = \frac{\|g_{k-1}\|_2}{2\|Ag_{k-1}\|_2}$, that is quite different from (1.9). As has been observed, the properties of (1.11) are similar to those of the Barzilai and Borwein gradient method [2, 6].

For the method (1.2) and (1.9), we will prove that the stepsize (1.9) tends to the optimal stepsize (1.6) as $k \to \infty$. At the same time, the minimal and maximal eigenvalues, λ_1 and λ_n, of A can be obtained (see the next section). Numerical results will be reported on the linear systems arising from a two-point boundary value problem (see Section 3). They not only demonstrate our convergence theorem, but suggest that the method (1.2) and (1.9) slightly performs better than the steepest descent method and the optimal stepsize gradient method. Some discussions on how to improve the method (1.2) and (1.9) will be made in Section 4.

2. Theoretical Analysis

For the SPD matrix A, we denote by $\{\lambda_i : i = 1, 2, \ldots, n\}$ and $\{u_i : i = 1, 2, \ldots, n\}$ its eigenvalues and their associated orthonormal eigenvalues respectively. Suppose that λ_1 and λ_n are the minimal and maximal eigenvalues of A. There exist sequences of vector $\{d_k\}$ such that

$$g_k = \sum_i d_k^{(i)} u_i, \tag{2.1}$$

where and below $\sum_i$ means $\sum_{i=1}^n$. The main purpose of this section to establish for the gradient method (1.2) and (1.9) the following theorem:

Theorem 2.1 *Consider the SPD linear system (1.1). For any starting point x_1 satisfying*

$$d_1^{(1)} \neq 0 \quad \textit{and} \quad d_1^{(n)} \neq 0, \tag{2.2}$$

let $\{x_k\}$ be the iterations generated by the gradient method (1.2) and (1.9). Then we have that

$$\lim_{k\to\infty} \alpha_k^{NEW} = \frac{2}{\lambda_1 + \lambda_n}, \tag{2.3}$$

which means that the stepsize α_k^{NEW} tends to the optimal stepsize (1.6). Further, the vectors

$$\frac{g_k}{\|g_k\|_2} + \frac{g_{k+1}}{\|g_{k+1}\|_2} \quad and \quad \frac{g_k}{\|g_k\|_2} - \frac{g_{k+1}}{\|g_{k+1}\|_2} \tag{2.4}$$

tend to be eigenvectors corresponding to λ_1 and λ_n, respectively.

To prove the above theorem, we require several lemmas. A result similar to the following lemma was once mentioned in Akaike (1959) without proof. Here we provide a simple proof, that is due to Ya-xiang Yuan (private communications).

Lemma 2.2 *Assume that $\beta_1, \beta_2, \ldots, \beta_n$ are all different and $t_1, t_2, \ldots, t_n$ are non-negative. Denote*

$$\Gamma_j(\beta) = \sum_i \beta_i^j t_i, \quad j = 0, 1, 2, \cdots \tag{2.5}$$

Then for any positive integer m, the following matrix

$$B = (b_{ij})_{m\times m} = (\Gamma_{i+j-2}(\beta))_{m\times m} \tag{2.6}$$

is semi-positive definite, which implies that det$(B) \ge 0$. And det$(B) = 0$ iff there are at most $(m-1)$ indices i's such that $t_i > 0$.

Proof Let $T = \mathrm{diag}(t_1, t_2, \ldots, t_n)$ and

$$S = \begin{pmatrix} 1 & \beta_1 & \cdots & \beta_1^{m-1} \\ 1 & \beta_2 & \cdots & \beta_2^{m-2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \beta_n & \cdots & \beta_n^{m-1} \end{pmatrix}. \tag{2.7}$$

Then the matrix B can be expressed as

$$B = S^t T S, \tag{2.8}$$

which with $t_i \ge 0$ $(i = 1, \ldots, n)$ implies that B is semi-positive definite and hence $\det(B) \ge 0$. Since $\{\beta_i\}$ are all different, the matrix S has the rank of m. Thus $\det(B) = 0$ iff there exists at least some index i with $t_i = 0$. This completes our proof. □

Now we consider the sequence $\{d_k\}$ given in (2.1). We assume without loss of generality that

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n. \tag{2.9}$$

If there are duplicate eigenvalues, $\lambda_{i_1} = \lambda_{i_2}$ say, we can consider the sum $d_k^{(i_1)} + d_k^{(i_2)}$ instead of $d_k^{(i_1)}$ and $d_k^{(i_2)}$. By (1.5) and (2.1), we can get that

$$d_{k+1}^{(i)} = (1 - \alpha_k \lambda_i) d_k^{(i)}. \tag{2.10}$$

Further, defining

$$p_k^{(i)} = \frac{(d_k^{(i)})^2}{\|d_k\|^2} \tag{2.11}$$

and

$$\gamma_k = \alpha_k^{-1}, \tag{2.12}$$

it follows from (2.10), (2.1) and (1.9) that

$$p_{k+1}^{(i)} = \frac{(\lambda_i - \gamma_k)^2 p_k^{(i)}}{\sum_i (\lambda_i - \gamma_k)^2 p_k^{(i)}}, \quad \text{where } \gamma_k = \sqrt{\textstyle\sum_i \lambda_i^2 p_k^{(i)}}. \tag{2.13}$$

By the definition of p_k, it is obvious that $p_k^{(i)} \geq 0$ for all i and

$$\sum_i p_k^{(i)} = 1, \quad \text{for all } k. \tag{2.14}$$

Now we provide the following lemma.

Lemma 2.3 *Suppose that p is a vector in R^n such that (i) $p^{(i)} \geq 0$ for $i = 1, 2, \ldots, n$; (ii) there exist at least two i's with $p^{(i)} > 0$; and (iii) $\sum_i p^{(i)} = 1$. Suppose that $T : R^n \to R$ is the transformation as follows:*

$$(Tp)^{(i)} = \frac{(\lambda_i - \gamma(p))^2 p^{(i)}}{\sum_i (\lambda_i - \gamma(p))^2 p^{(i)}}, \quad \textit{where } \gamma(p) = \sqrt{\textstyle\sum_i \lambda_i^2 p^{(i)}}. \tag{2.15}$$

Define the function

$$\Delta(p) = \gamma(p) - \sum_i \lambda_i p^{(i)}. \tag{2.16}$$

Then we have that

$$\Delta(Tp) \geq \Delta(p). \tag{2.17}$$

Further, (2.17) holds with equality iff there are only two indices, i_1 and i_2 say, such that $p^{(i)} = 0$ if $i \notin \{i_1, i_2\}$ and

$$p^{(i_1)} = \frac{\lambda_{i_1} + 3\lambda_{i_2}}{4(\lambda_{i_1} + \lambda_{i_2})}, \quad p^{(i_2)} = \frac{3\lambda_{i_1} + \lambda_{i_2}}{4(\lambda_{i_1} + \lambda_{i_2})}. \tag{2.18}$$

Proof At first, we introduce the following notation:

$$M_j(p) = \sum_i (\lambda_i - \gamma(p))^j p^{(i)}, \quad \text{for } j = 0, 1, 2, \cdots \tag{2.19}$$

By the definition of $\gamma(p)$ and the property (iii) of p, we have that

$$M_2(p) = \sum_i \lambda_i^2 p^{(i)} - 2\gamma(p) \sum_i \lambda_i p^{(i)} + \gamma^2(p) = 2\gamma(p)[\gamma(p) - \sum_i \lambda_i p^{(i)}] = 2\gamma(p)\Delta(p). \tag{2.20}$$

In addition, by the properties (i) and (ii) of p, we can see that both $\gamma(p)$ and $\Delta(p)$ are positive. Thus by (2.20), $M_2(p) > 0$, which implies the well definition of the transformation T.

Now, noting that

$$\begin{aligned} \lambda_i &= (\lambda_i - \gamma(p)) + \gamma(p), && (2.21) \\ \lambda_i^2 &= (\lambda_i - \gamma(p))^2 + 2\gamma(p)(\lambda_i - \gamma(p)) + \gamma^2(p), && (2.22) \end{aligned}$$

we have by the definition of $M_j(p)$ that

$$\begin{aligned} \sum_i \lambda_i (\lambda_i - \gamma(p))^2 p^{(i)} &= M_3(p) + \gamma(p) M_2(p) && (2.23) \\ \sum_i \lambda_i^2 (\lambda_i - \gamma(p))^2 p^{(i)} &= M_4(p) + 2\gamma(p) M_3(p) + \gamma^2(p) M_2(p). && (2.24) \end{aligned}$$

Thus by (2.20), (2.23) and (2.24), the relation (2.17) is equivalent to

$$\sqrt{\frac{M_4(p) + 2\gamma(p)M_3(p) + \gamma^2(p)M_2(p)}{M_2(p)}} - \frac{M_3(p) + \gamma(p)M_2(p)}{M_2(p)} \geq \frac{M_2(p)}{2\gamma(p)}. \tag{2.25}$$

(2.25) can be rewritten as

$$\sqrt{\frac{M_4(p) + 2\gamma(p)M_3(p) + \gamma^2(p)M_2(p)}{M_2(p)}} \geq \frac{M_3(p) + \gamma(p)M_2(p)}{M_2(p)} + \frac{M_2(p)}{2\gamma(p)}. \tag{2.26}$$

By squaring both sides of (2.26) and multiplying with $4\gamma^2(p)M_2(p)$, we know that (2.26) is equivalent to

$$4\gamma^2(p)[M_2(p)M_4(p) - M_3^2(p) - M_2^3(p)] - M_2^4(p) - 4\gamma(p)M_2^2(p)M_3(p) \geq 0. \tag{2.27}$$

By the definition of $\gamma(p)$, (2.22) and $\sum_i p^{(i)} = 1$, we can also get that

$$\gamma^2(p) = \sum_i \lambda_i^2 p^{(i)} = M_2(p) + 2\gamma(p)M_1(p) + \gamma^2(p), \tag{2.28}$$

which gives

$$\gamma(p) = -\frac{M_2(p)}{2M_1(p)}. \tag{2.29}$$

Substituting (2.29) into (2.27) and noting that $M_1(p) < 0$, we see that (2.27) is further equivalent to

$$M_2(p)M_4(p) - M_3^2(p) - M_2^3(p) + 2M_1(p)M_2(p)M_3(p) - M_1^2(p)M_2^2(p) \geq 0. \tag{2.30}$$

Now we consider the matrix $B = (M_{i+j-2}(p))_{3\times 3}$:

$$B = \begin{pmatrix} M_0(p) & M_1(p) & M_2(p) \\ M_1(p) & M_2(p) & M_3(p) \\ M_2(p) & M_3(p) & M_4(p) \end{pmatrix}. \tag{2.31}$$

We have by Lemma 2.2 and $M_0(p) = \sum_i p^{(i)} = 1$ that

$$\det(B) = M_2(p)M_4(p) - M_3^2(p) - M_2^3(p) + 2M_1(p)M_2(p)M_3(p) - M_1^2(p)M_4(p) \geq 0. \tag{2.32}$$

In addition, it is obvious that

$$M_4(p) \geq M_2^2(p). \tag{2.33}$$

Thus by (2.32) and (2.33), we have (2.30) and hence (2.17). Further, we see that the $=$ in (2.17) holds iff the $=$'s in (2.32) and (2.33) hold simultaneously. By Lemma 2.2, the $=$ in (2.32) holds iff there are only two i's with $p^{(i)} > 0$. Denote i_1 and i_2 to be the two indices. In this case, we can see that the $=$ in (2.33) holds iff

$$(\lambda_{i_1} - \gamma(p))^2 = (\lambda_{i_2} - \gamma(p))^2. \tag{2.34}$$

Since in this case $\gamma(p) = \sqrt{\lambda_{i_1}^2 p^{(i_1)} + \lambda_{i_2}^2 p^{(i_2)}}$, we can get from this and (2.34) that

$$\lambda_{i_1}^2 p^{(i_1)} + \lambda_{(i_2)}^2 p^{(i_2)} = \left(\frac{\lambda_{i_1} + \lambda_{i_2}}{2}\right)^2. \tag{2.35}$$

Therefore we can solve from (2.35) and $p^{(i_1)} + p^{(i_2)} = 1$ that (2.18) holds. This completes our proof. □

Lemma 2.4 *Let $p = p_1 \in R^n$ satisfying the conditions of Lemma 2.3 and T be the transformation in (2.15). Then the sequence $\{T^k p; k = 1, 2, \cdots\}$ is convergent to some p_*. This p_* has only two nonzero components*

$$p_*^{(i_1)} = \frac{\lambda_{i_1} + 3\lambda_{i_2}}{4(\lambda_{i_1} + \lambda_{i_2})} \quad \textit{and} \quad p_*^{(i_2)} = \frac{3\lambda_{i_1} + \lambda_{i_2}}{4(\lambda_{i_1} + \lambda_{i_2})}. \tag{2.36}$$

Proof Let $p_{k+1} = T^k p = T^k p_1$. Note by Lemma 2.3 that $\Delta(p_k)$ is monotonical increasing. We know by this, $\Delta(p_1) > 0$ and $\gamma(p_k) > 0$ for all k that $M_2(p_k) = 2\gamma(p_k)\Delta(p_k) > 0$ for all k. Hence the sequence $\{p_{k+1}\}$ is well defined. Let us denote

$$P_* = \{\text{all cluster points of } \{p_k;\ k = 0, 1, 2, \cdots\}\ \} \tag{2.37}$$

and $|P_*|$ to be the numbers of elements in P_*. It is obvious that $|P_*| \geq 1$ since p_k is bounded. For any $p_* \in P_*$, we know by Lemma 2.3 that $\Delta(p_*) = \Delta(Tp_*)$. Thus p_* has only two nonzero components, $p_*^{(\mu)}$ and $p_*^{(\nu)}$ say, and their values are uniquely determined by the indices μ, ν and the eigenvalues λ_μ and λ_ν. This means that p_* is characterized by the indices μ and ν, showing that $|P_*| \leq C_n^2 = \frac{1}{2}n(n-1)$.

Now we proceed by contradiction and assume that $|P_*| \geq 2$. Pick any $p_* \in P_*$ and denote δ to be the distance between p_* and $P_* \backslash \{p_*\}$. Since P_* has only finite elements, we know that $\delta > 0$. Defining $\mathcal{B}(p_*, r) = \{p : \|p - p_*\|_2 \leq r\}$, we have by the definition of P_* and $|P_*| \geq 2$ that there exists an infinite subsequence $\{k_i\}$ such that

$$p_{k_i} \in \mathcal{B}(p_*, \frac{1}{3}\delta), \quad \text{but} \quad p_{k_i+1} \in R^n \backslash \mathcal{B}(p_*, \frac{2}{3}\delta). \tag{2.38}$$

However, since the transformation is continuous, we have by this and $p_{k_i} \to p_*$ that $p_{k_i+1} = Tp_{k_i} \to Tp_* = p_*$, which implies that

$$\lim_{i\to\infty} \|p_{k_i+1} - p_{k_i}\| = 0. \tag{2.39}$$

(2.38) and (2.39) give a contradiction. Thus we have that $|P_*| = 1$, which means that $\{p_k\}$ is convergent.

Assume that $p_k \to p_*$. Then we have that $Tp_* = p_*$. Thus by Lemma 2.3, p_* has only two nonzero components satisfying (2.36). $\square$

Lemma 2.5 *Under the conditions of Lemma 2.4, if $p_1^{(1)} > 0$ and $p_1^{(n)} > 0$, we have that*

$$\{i_1, i_2\} = \{1, n\}. \tag{2.40}$$

Proof It is obvious that

$$\lambda_1 < \gamma(p) < \lambda_n, \quad \text{if } p^{(1)} > 0 \text{ and } p^{(n)} > 0. \tag{2.41}$$

It follows from this and the definition of T that

$$(Tp)^{(1)} > 0 \text{ and } \ (Tp)^{(n)} > 0, \quad \text{if } p^{(1)} > 0 \text{ and } \ p^{(n)} > 0. \tag{2.42}$$

Then by (2.42), $p_1^{(1)} > 0$, $p_1^{(n)} > 0$ and the induction principle, we see that

$$p_k^{(1)} > 0,\ \ p_k^{(n)} > 0 \quad \text{for } k = 1, 2, \cdots \tag{2.43}$$

Now we assume without loss of generality that $i_1 < i_2$. If $i_2 < n$, we have by Lemma 2.4 and the definition of $\gamma(p)$ that

$$\lim_{k\to\infty} p_k^{(i_2)}/p_{k+1}^{(i_2)} = 1, \tag{2.44}$$

$$\lim_{k\to\infty} p_k^{(n)} = 0, \tag{2.45}$$

$$\lim_{k\to\infty} \gamma(p_k) = \frac{\lambda_{i_1}+\lambda_{i_2}}{2}. \tag{2.46}$$

By (2.13), (2.44) and (2.46), we have that

$$\begin{aligned}
\lim_{k\to\infty} \frac{p_{k+1}^{(n)}}{p_k^{(n)}} &= \lim_{k\to\infty} \frac{p_{k+1}^{(n)}}{p_k^{(n)}}\frac{p_k^{(i_2)}}{p_{k+1}^{(i_2)}} \\
&= \lim_{k\to\infty} \left(\frac{\lambda_n-\gamma(p_k)}{\lambda_{i_2}-\gamma(p_k)}\right)^2 \\
&= \lim_{k\to\infty} \left(\frac{2\lambda_n-\lambda_{i_1}-\lambda_{i_2}}{\lambda_{i_2}-\lambda_{i_1}}\right)^2 \\
&= \lim_{k\to\infty} \left(1+2\,\frac{\lambda_n-\lambda_{i_2}}{\lambda_{i_2}-\lambda_{i_1}}\right)^2 > 1.
\end{aligned} \tag{2.47}$$

The above relation implies that $\lim_{k\to\infty} p_k^{(n)} = +\infty$, contracting (2.45). Thus we must have that $i_2 = n$. In a similar way, we can show that $i_1 = 1$. $\square$

Now we are able to give a proof to Theorem 2.1.

Proof to Theorem 2.1. Assume without loss of generality that (2.9) holds. By (2.2) and the definition (2.13) of p_k, we see that $p = p_1$ satisfies the conditions of Lemmas 2.4 and 2.5. Thus by the two lemmas, we know that

$$\lim_{k\to\infty} p_k = (a, 0, \cdots, 0, b)^t, \tag{2.48}$$

where

$$a = \frac{\lambda_1+3\lambda_n}{4(\lambda_1+\lambda_n)} \quad \text{and} \quad b = \frac{3\lambda_1+\lambda_n}{4(\lambda_1+\lambda_n)}. \tag{2.49}$$

By (2.12), (2.13), (2.48) and (2.49), we can get that

$$\lim_{k\to\infty} \alpha_k^{NEW} = \frac{1}{\sqrt{\lambda_1^2 a + \lambda_n^2 b}} = \frac{2}{\lambda_1+\lambda_n}. \tag{2.50}$$

In addition, it follows from (2.10) and $\lambda_n^{-1} < \alpha_k < \lambda_1^{-1}$ that

$$\mathrm{sign}(d_{k+1}^{(1)}) = \mathrm{sign}(d_k^{(1)}) \quad \text{and} \quad \mathrm{sign}(d_{k+1}^{(n)}) = -\mathrm{sign}(d_k^{(n)}). \tag{2.51}$$

Then we know by (2.48), (2.11) and (2.51) that

$$\lim_{k\to\infty}\frac{d_{2k-1}}{\|d_{2k-1}\|_2}=(\mathrm{sign}(d_1^{(1)})\sqrt{a},0,\cdots,0,\mathrm{sign}(d_1^{(n)})\sqrt{b})^t, \tag{2.52}$$

$$\lim_{k\to\infty}\frac{d_{2k}}{\|d_{2k}\|_2}=(\mathrm{sign}(d_1^{(1)})\sqrt{a},0,\cdots,0,-\mathrm{sign}(d_1^{(n)})\sqrt{b})^t. \tag{2.53}$$

By (2.1) and $\|g_k\|_2=\|d_k\|_2$, we then know that

$$\lim_{k\to\infty}\frac{g_k}{\|g_k\|_2}+\frac{g_{k+1}}{\|g_{k+1}\|_2}=2\mathrm{sign}(d_1^{(1)})\sqrt{a}\,u_1, \tag{2.54}$$

$$\lim_{k\to\infty}\frac{g_k}{\|g_k\|_2}-\frac{g_{k+1}}{\|g_{k+1}\|_2}=\pm2\sqrt{b}\,u_n. \tag{2.55}$$

The relations (2.50), (2.54) and (2.55) show the truth of Theorem 2.1. □

By Theorem 2.1, it is easy to conclude that the sequence $\{x_k\}$ generated by the method (1.2) and (1.9) converges to the unique solution x_* of system (1.1) and the asymptotic order of Q-linear convergence is $\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}$. In theory, Theorem 2.1 requires the assumption (2.2), namely, the initial residual g_1 has two nonzero components at the eigenvectors u_1 and u_n. If this is not the case, we see that the stepsize α_k^{NEW} will tend to $\frac{2}{\lambda_{i_1}+\lambda_{i_2}}$, where i_1 and i_2 are the smallest and largest indicies such that $d_1^{(i)}\neq 0$, respectively. In practical computations, however, the assumption (2.2) is of less importance. This is because, the components $d_k^{(1)}$ and $d_k^{(n)}$ will become nonzero due to the disturbance in computations, and then eventually dominant the components in the vector d_k.

Remark Assumption (2.2) is not restrictive. Indeed, let B be the inverse of the matrix $[u_1\cdots u_n]$. Choose a vector d_1 satisfying (2.2) and let

$$x_1=A^{-1}\left(B^{-1}[d_1^1\cdots d_1^{(n)}]^\top+b\right).$$

Then x_1 is a starting point that satisfies (2.2).

3. Numerical experiments

We tested the new gradient method (1.2) and (1.9) on one kind of linear system (1.1) arising from two-point boundary value problems, where the coefficient matrix $A=(a_{ij})_{n\times n}$ is as follows:

$$a_{ij}=\begin{cases}2, & \text{if } i=j;\\ -1, & \text{if } |i-j|=1;\\ 0, & \text{otherwise.}\end{cases} \tag{3.1}$$

It is known that the above matrix A has the eigenvalues

$$\lambda_i = 4\sin^2 \frac{i\pi}{2(n+1)}, \quad i = 1, 2, \dots, n, \tag{3.2}$$

and the vectors $\{u_i : i = 1, 2, \dots, n\}$ with

$$u_i^{(j)} = \sin \frac{ij\pi}{n+1}, \quad j = 1, 2, \dots, n \tag{3.3}$$

are their corresponding eigenvectors. It is easy to see that

$$\alpha_k^{OPT} \equiv \frac{2}{\lambda_1 + \lambda_n} = \frac{1}{2} \tag{3.4}$$

and

$$\|u_i\|_2 \equiv \sqrt{\frac{n+1}{2}}, \quad \text{for all } i. \tag{3.5}$$

The condition number of A is $\cot^2 \frac{\pi}{2(n+1)} \approx \frac{4(n+1)^2}{\pi^2}$. Based on the property (3.5), we choose the right hand term b of the system (1.1) as

$$b = \sum_i u_i \tag{3.6}$$

that treats each eigenvector u_i all the same. For different values of n, we always choose the zero vector as the starting point and terminate the solution procedure while

$$\|g_k\|_2 \le 10^{-6}\|g_1\|_2 \tag{3.7}$$

or the iteration number exceeds the limit 9999. Our numerical experiments were made on an SGI Indigo workstation with MATLAB 6.0.

We also compared the performances of our new method with the steepest descent method and the optimal stepsize gradient method. See Table 3.1, where SD, OPT and NEW stand for the steepest descent method, the optimal stepsize gradient method, and the method (1.2)$\sim$(1.9), respectively. The columns 'iter' give the iteration numbers required by the methods, and the columns 'error' provide the final values of $\|g_k\|/\|g_1\|$. From Table 3.1, we see that the optimal stepsize gradient method is the worst, although its stepsize has the optimal property (1.7). We can also see that the new method performs slightly better than the steepest descent method. One possible explanation to this is that, the stepsize in (1.9) is always shorter than or equal to the steepest descent one:

$$\alpha_k^{NEW} \le \alpha_k^{SD}, \tag{3.8}$$

that follows directly from the Cauchy-Schwarz inequality. Suppose that the current point is x_k. An advantage to shorten the steepest descent stepsize α_k^{SD} is that, the residual norm $\|g_{k+2}\|_{A^{-1}}$ will become better much probably (see [4]).

Table 3.1. Comparing different gradient methods

n	SD		OPT		NEW	
	iter	error	iter	error	iter	error
20	702	9.8440e-07	1142	9.9516e-07	696	9.9311e-07
30	1338	9.9695e-07	2453	9.9555e-07	1324	9.9798e-07
50	2966	9.9921e-07	6508	9.9895e-07	2921	9.9895e-07
100	8122	9.9984e-07	9999	1.2965e-03	7904	9.9985e-07

Table 3.2 provides the final values of α_k in the new method and the approximate minimal and maximal eigenvalues $\tilde{\lambda}_1$ and $\tilde{\lambda}_n$ estimated by the final residuals g_{k-1} and g_k. Relative errors of these values with the true values are also given. This table clearly demonstrated the convergence result stated in Theorem 2.1.

Table 3.2. Drawing more information for the new method

n	α_k^{NEW}	$\lvert 1-\frac{\alpha_k^{NEW}}{\alpha_k^{OPT}}\rvert$	$\tilde{\lambda}_1$	$\lvert 1-\frac{\tilde{\lambda}_1}{\lambda_1}\rvert$	$\tilde{\lambda}_n$	$\lvert 1-\frac{\tilde{\lambda}_n}{\lambda_n}\rvert$
20	5.000e-1	2.2204e-16	2.2338e-02	3.1063e-16	3.9777e+00	2.2329e-16
30	5.000e-1	0	1.0261e-02	1.6905e-16	3.9897e+00	4.4523e-16
50	5.000e-1	2.2204e-16	3.7933e-03	1.6440e-13	3.9962e+00	5.5564e-16
100	5.000e-1	1.1843e-12	9.6744e-04	5.2008e-09	3.9990e+00	7.8734e-14

4. Some discussions

In this paper, we have proposed a new gradient method, namely, the method (1.2) and (1.9), for SPD linear system. The proposition of this method makes it possible to approximate the optimal stepsize gradient method (1.2) and (1.6). By this method, we can also solve the linear system and calculate the minimal and maximal eigenvalues (and the corresponding eigenvectors) of its coefficient matrix simultaneously. As pointed out by Jinyun Yuan (private communications), since the Chebyshev iterative method, which is 2-order convergence, needs to estimate $\frac{2}{\lambda_1+\lambda_n}$, namely the value in (1.6), it may be worthwhile to consider a combination of the Chebyshev iteration method and the method (1.2) and (1.9). The method (1.2) and (1.9) can be also used to improve other methods for TLS problems.

The numerical experiments in Section 3 showed that, the new method performs better than the steepest descent method and the optimal stepsize gradient method. For any of the three methods, however, we know that the solution procedure tends to proceed in some two-dimensional subspaces and hence zigzags occur, and the asymptotic order of Q-linear convergence is $\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}$. There have been various ways to accelerate the steepest descent method (for example, see

[2, 4, 7, 8, 10]). Here we point out that, one possible approach is to combine both the steepest descent method and the new method, choosing

$$\alpha_k = \begin{cases} \alpha_k^{SD}, & \text{for } k \text{ odd;} \\ \alpha_k^{NEW}, & \text{for } k \text{ even.} \end{cases} \tag{4.1}$$

It can be shown that, the resulting method of (4.1) is two-step Q-superlinearly convergent for two-dimensional SPD linear system, and hence can avoid zigzags. Our numerical experiments showed that, this method requires about one half of the iteration numbers required by the new method. For example, for the problem with $n = 100$ in Section 3, to reach (3.7), the method (4.1) only needs 3921 iterations.

While the problem is only to calculate the minimal and maximal eigenvalues of an SPD matrix A, instead of the method (1.2) and (1.9), we may develop its following variation:

$$\left.\begin{array}{l} x_{k+1} = x_k - \frac{Ax_k}{\|Ax_k\|_2}, \\ x_{k+1} = \frac{x_{k+1}}{\|x_{k+1}\|_2}, \end{array}\right\} \quad k = 1, 2, \cdots \tag{4.2}$$

where $x_1 \in R^n$ is a starting point with $\|x_1\|_2 = 1$. In this case, the quantity $\|Ax_k\|_2$ will tend to $\frac{\lambda_1+\lambda_n}{2}$, and hence we can use the difference $|1 - \frac{\|Ax_k\|_2}{\|Ax_{k-1}\|_2}| \le \epsilon$ as the stopping rule. By choosing $n = 50$, $\epsilon = 10^{-10}$ and $x_1 = \frac{b}{\|b\|_2}$ with b given in (3.6), we found that an approximate minimal eigenvalue $\tilde{\lambda}_1 = 3.7958 \times 10^{-3}$ with relative error 6.3894×10^{-4} can be obtained in 742 iterations. At the same time, an approximate maximal eigenvalue $\tilde{\lambda}_n = 3.9962$ was obtained with relative error 6.0721×10^{-7}. By post error analysis, the relative errors of $\tilde{\lambda}_1$ and $\tilde{\lambda}_n$ can be improved to 1.3597×10^{-7} and 1.2369×10^{-10}, respectively. It is easy to see that, this method to compute minimal and maximal eigenvalues is similar to the shifted power method with an asymptotic shift being $\frac{\lambda_1+\lambda_n}{2}$.

Finally, we would like to mention that the analysis of Section 2 proceeds in the same line as in Akaike [1], while the major difficulty is to find the monotonical increasing quantity $\Delta(p)$ defined in (2.16). For the steepest descent method, Akaike [1] found that the quantity $M_2(p)$ is monotonically increasing. In this case, it follows from $\gamma(p) = \sum_i \lambda_i p^{(i)}$ and $\sum_i p^{(i)} = 1$ that

$$M_2(p) = \sum_i \lambda_i^2 p^{(i)} - (\sum_i \lambda_i p^{(i)})^2. \tag{4.3}$$

For the new method, the quantity $M_2(p)$ is not a monotonically increasing quantity any more. However, we have by $\gamma(p) = \sqrt{\sum_i \lambda_i^2 p^{(i)}}$ that

$$M_2(p) = 2\gamma(p) \left[\sqrt{\sum_i \lambda_i^2 p^{(i)}} - \sum_i \lambda_i p^{(i)} \right]. \tag{4.4}$$

A comparison between (4.3) and (4.4) hints us to use the quantity $\Delta(p) = \frac{M_2(p)}{2\gamma(p)}$ as the monotonical increasing quantity. The authors guess that there is another way to establish Lemma 2.3, for our numerical tests shows that this lemma holds for any transformation T in (2.15) with $\gamma(p)$ replaced by

$$\gamma(p) = (\sum_i \lambda_i^r p^{(i)})^{\frac{1}{r}}, \quad \text{where } r \text{ is any number greater than 1.} \tag{4.5}$$

Acknowledgements. The authors are indebted to Prof. Da-yong Cai in Tsinghua University, who suggested us the use of the linear systems with the coefficient matrix given in (3.1). They also thank Prof. Ya-xiang Yuan in Chinese Academy of Sciences for providing a simple and nice proof to Lemma 2.2, and Profs. Roger Fletcher in University of Dundee and Jinyun Yuan in Federal University of Parana for their many useful discussions and comments on an early version of this paper.

References

[1] H. Akaike, *On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method*, Ann. Inst. Statist. Math. Tokyo 11 (1959) 1-17.

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

[3] A. Cauchy, *Méthode générale pour la résolution des systèmes d'équations simultanées,* Comp. Rend. Acad. Sci. Paris 25 (1847) 536-538.

[4] Y. H. Dai and Y. Yuan, *Alternate minimization gradient method*, Research report-2001-043, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, 2001.

[5] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal. 31 (1994) 1645-1661.

[6] R. Fletcher, *On the Barzilai-Borwein method*, Research report, Unversity of Dundee, 2001.

[7] A. Friedlander, J. M. Martínez, B. Molina, and M. Raydan, *Gradient method with retards and generalizations*, SIAM J. Numer. Anal. 36 (1999) 275-289.

[8] Q. Hu and J. Zou, *An iterative method with variable relaxation parameters for saddle-point problems*, SIAM J. Matrix Anal. Appl. 23 (2001) 317-338.

[9] J. Nocedal, A. Sartenaer, and C. Zhu, *On the behavior of the gradient norm in the steepest descent method*, Research report, Northwestern University, 2000.

[10] M. N. Vrahatis, G. S. Androulakis, J. N. Lambrinos, G. D. Magoulas, *A class of gradient unconstrained minimization algorithms with adaptive stepsize*, Journal of Computational and Applied Mathematics 114 (2000) 367-386.