Alternate Minimization Gradient Method*

Yu-Hong Dai Ya-xiang Yuan

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, China. Email: {dyh,yyx}@lsec.cc.ac.cn

Abstract

It is well known that the minimization of a smooth function $f(x)$ is equivalent to minimizing its gradient norm $\|g(x)\|_2$ in some sense. In this paper, we propose a modified steepest descent method, whose stepsizes alternately minimize the function value and the gradient norm along the line of steepest descent. Hence the name "alternate minimization (AM) gradient method". For strictly convex quadratics, the AM method is proved to be Q-superlinearly convergent in two-dimensions, and Q-linearly convergent in any dimension. Numerical experiments are presented for symmetric and positive definite linear systems. They suggest that the AM method is much better than the classical steepest descent (SD) method and comparable with some existing gradient methods. They also show that the AM method is an efficient alternative if a solution with a low precision is required. Two variants of the AM method, named shortened SD step gradient methods, are also presented and analysed in this paper. By designing a new kind of line search, the two variants are extended to the field of unconstrained optimization.

Key words: linear system, unconstrained optimization, gradient method, monotone and nonmonotone, convergence rate, line search.

Mathematics subject classification: 49M37, 65F10, 65K05, 90C26.

1. Introduction

It is well known that the solution of the linear system

$$Ax = b, \quad \text{where } A \in R^{n\times n},\ b \in R^n \tag{1.1}$$

*This research was supported by the Chinese NSF grants 19731010, 19801033, 10171104, the CNPq in Brazil and an Innovation Fund of Chinese Academy of Sciences. This work was once reported on the International Workshop on "Numerical Linear Algebra, Numerical Method for PDE, and Mathematical Programming" (Curitiba, Brazil, August 20-23, 2001).

also solves the minimization problem

$$\min q(x) = \frac{1}{2}x^T Ax - b^T x, \tag{1.2}$$

if the matrix A is symmetric and positive definite (SPD). It was because of this fact that Fletcher and Reeves [7] extended the linear conjugate gradient method ([12]) to the field of unconstrained optimization

$$\min f(x), \quad x \in R^n. \tag{1.3}$$

In this paper, we are interested in the gradient method for problem (1.1) or (1.2), and we then try to extend the new gradient method to problem (1.3).

Letting $g_k = Ax_k - b$, the gradient method for problem (1.1) or (1.2) is of the form

$$x_{k+1} = x_k - \alpha_k g_k, \tag{1.4}$$

where x_1 is a starting point and α_k is a stepsize. The classical steepest descent (SD) method, that can be dated back to [3], determines α_k by

$$\alpha_k^{SD} = \frac{g_k^T g_k}{g_k^T A g_k}. \tag{1.5}$$

The stepsize (1.5) minimizes the function value along the line, namely,

$$\alpha_k = \text{argmin}\{f(x_k - \alpha g_k) : \ \alpha \in R^1\}. \tag{1.6}$$

However, the SD method is very slow in practical computations and produces zigzags. It is proved in [1, 8] that the directions in the SD method will tend to two different directions; namely,

$$\lim_{k\to\infty} \frac{g_{2k}}{\|g_{2k}\|} = \bar{d} \quad \text{and} \quad \lim_{k\to\infty} \frac{g_{2k+1}}{\|g_{2k+1}\|} = \hat{d} \quad (\text{with } \bar{d} \neq \hat{d}). \tag{1.7}$$

In 1988, Barzilai and Borwein [2] proposed another choice for the stepsize:

$$\alpha_k^{BB} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}} = \frac{g_{k-1}^T g_{k-1}}{g_{k-1}^T A g_{k-1}}, \tag{1.8}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$. The stepsize (1.8) is such that the matrix $D_k = \frac{1}{\alpha_k} I$ has a certain quasi-Newton property:

$$D_k = \arg \min_{D=\frac{1}{\alpha}I} \{\|Ds_{k-1} - y_{k-1}\|_2\}. \tag{1.9}$$

Although it cannot guarantee a descent in the function value at each iteration, the BB method (1.4)∼(1.8) converges to the solution of (1.2) (see [15]), and is much faster than the SD method (for example, see [6]). Regarding the convergence rate, the BB method is R-superlinearly convergent for two-dimensional strictly convex quadratics ([2]), and R-linearly convergent in any dimension ([5]). This method has now received many useful extensions and generalizations to unconstrained optimization, convex constrained optimization, nonlinear systems, optimal control, *etc.* (see [16, 10, 4] and the references therein).

Note that each zigzag generated by the SD method includes two steps. By replacing one of the two steps with the BB step, Dai [4] proposed the so-called alternate stepsize (AS) gradient method. More exactly, the stepsize in the AS method has the form

$$\alpha_k^{AS} = \begin{cases} \alpha_k^{SD}, & \text{for odd } k; \\ \alpha_k^{BB}, & \text{for even } k. \end{cases} \tag{1.10}$$

The AS method is proved to be two-step Q-superlinearly convergent for two-dimensional strictly convex quadratics and R-linearly convergent in any dimension. The numerical results for both linear systems and unconstrained optimization in [4] showed that the AS method is a promising alternative to the BB method. Noting that $\alpha_{2k}^{AS} = \alpha_{2k-1}^{AS} = \alpha_{2k-1}^{SD}$, this method can be regarded as a particular member of the cyclic SD stepsize method in [10].

In this paper, we will propose a new gradient method, whose stepsizes alternately minimize the function value and the gradient norm along the line (see the next section). Hence the name "alternate minimization (AM) gradient method". For strictly convex quadratics, the AM method is proved to be Q-superlinearly convergent in two-dimensions, and Q-linearly convergent in any dimension (see §3). Numerical results for SPD linear systems are made in §4, which suggest that the AM method is much better than the SD method and comparable with the BB and AS methods. They also suggest that the AM method is an efficient alternative if a solution with a low precision is required. Two variants of the AM method, namely, SS1 and SS2, are also presented and analysed in this paper (see §§2-4 for some related statements). Both the SS1 and SS2 methods are extended to the field of unconstrained optimization by designing a new kind of line search (see §5). Conclusions and discussions are made in the last section.

2. Motivation and the method

In the previous section, we have described several gradient methods. Notice that the BB method is a nonmonotone method for it cannot guarantee a descent in the objective function at each iteration. The AS method is also nonmonotone but is not so nonmonotone because an SD step is used every other iteration. Specifically, when we tested the AS method for unconstrained optimization ([4]) with the strategy of requiring the function value at the BB step to be less than the maximal function value during the previous M iterations, we found that any integer $M \geq 2$ provided the same numerical results. In contrast, our numerical experience showed that the parameter M influenced the performance of the BB method and that the difference when using different values of M is sometimes significant when the problem is ill-conditioned (see also [16] for the latter point). The SD method is indeed a monotone method, but it is very slow in practice.

Suppose that the dimension of the problem is very large, so that each algorithm can compute only a few iterations. Then one should prefer a monotone algorithm to a nonmonotone one since the object is to minimize the function. Therefore the following question would be very interesting: does there also exist a monotone gradient algorithm, that is much faster than the SD method?

To give an answer to this question, our basic idea is alternately to minimize the gradient norm $\|g(x)\|_2$ and the function value $q(x)$ along the line. More exactly, for all $k \geq 1$, we choose the stepsizes such that

$$\alpha_{2k-1} = \operatorname{argmin}\{\|g(x_{2k-1} - \alpha g_{2k-1})\|_2 : \ \alpha \in R^1\} \tag{2.1}$$

and

$$\alpha_{2k} = \operatorname{argmin}\{q(x_{2k} - \alpha g_{2k}) : \ \alpha \in R^1\}. \tag{2.2}$$

Noting that the gradients of $\|g(x)\|_2$ and $q(x)$ are $Ag(x)/\|g(x)\|_2$ and $g(x)$, respectively, we have by (2.1) and (2.2) that

$$g_{2k}^T A g_{2k-1} = 0 \tag{2.3}$$

and

$$g_{2k+1}^T g_{2k} = 0. \tag{2.4}$$

The above relations imply that g_{2k} is conjugate to g_{2k-1}, whereas g_{2k+1} and g_{2k} are orthogonal. It follows by (1.4) and $g_k = Ax_k - b$ that

$$g_{k+1} = g_k - \alpha_k A g_k. \tag{2.5}$$

Then we can solve from (2.3) and (2.4) to obtain

$$\alpha_k^{AM} = \begin{cases} \frac{g_k^T A g_k}{g_k^T A^2 g_k}, & \text{for odd } k; \\ \frac{g_k^T g_k}{g_k^T A g_k}, & \text{for even } k. \end{cases} \tag{2.6}$$

We call the method (1.4)$\sim$(2.6) the alternate minimization (AM) gradient method. To be able to avoid zigzags and become a fast method like the BB and AS methods, the AM method is expected to have some kind of superlinear convergence for two-dimensional strictly convex quadratics. In the next section, we will show that, in this case, the convergence rate of the AM method is Q-superlinear.

By the Cauchy inequality and the assumption that A is a SPD matrix, we have

$$\frac{v^T A v}{v^T A^2 v} \leq \frac{v^T v}{v^T A v}, \quad \text{for any } v \in R^n \backslash \{0\}. \tag{2.7}$$

This shows that if the gradient does not vanish, the stepsize of the AM method at any odd iteration is less than or equal to the corresponding SD stepsize, namely

$$\alpha_{2k-1}^{AM} \leq \alpha_{2k-1}^{SD}. \tag{2.8}$$

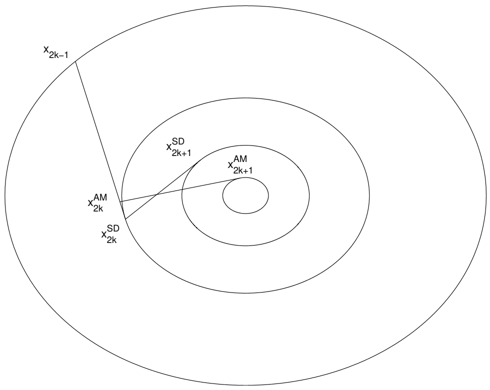
For the stepsizes in even iterations, we clearly have that

$$\alpha_{2k}^{AM} = \alpha_{2k}^{SD}. \tag{2.9}$$

The relations (2.8) and (2.9) indicate that the AM method carries out a full SD step after a shortened SD step. This property of the AM method can be illustrated by Figure 2.1, that is based on:

$$q(x) = \frac{1}{2} x^T \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix} x, \quad x_{2k-1} = \begin{pmatrix} -2 \\ 1 \end{pmatrix}. \tag{2.10}$$

Figure 2.1 Illustrating the AM, and SD methods



In Figure 2.1, x_{2k}^{SD} and x_{2k+1}^{SD} denote the iterations generated by the SD method from x_{2k-1}, while x_{2k}^{AM} and x_{2k+1}^{AM} stand for the iterations by the AM method. From the figure, we see that $f(x_{2k+1}^{AM}) < f(x_{2k}^{AM}) < f(x_{2k-1})$, which implies that the AM is a monotone method (this will be strictly established by Theorem 3.2). We can also see from the figure that

$$f(x_{2k}^{AM}) > f(x_{2k}^{SD}) \quad \text{but} \quad f(x_{2k+1}^{AM}) < f(x_{2k+1}^{SD}). \tag{2.11}$$

Although the second inequality in (2.11) is not always true for all strictly convex quadratics, our random experiments show that it holds for most of the cases. In theory, we can show that for any strictly convex quadratic, there must exist a stepsize $\tilde{\alpha}_{2k-1} < \alpha_{2k-1}^{SD}$ such that a full SD step at the point $x_{2k-1} - \tilde{\alpha}_{2k-1} g_{2k-1}$ will provide the minimizer of the function in the two-dimensional subspace $x_{2k-1} + \text{Span}\{g_{2k-1}, Ag_{2k-1}\}$. In other words, a suitable reduction in the stepsize α_{2k-1}^{SD} will lead to a better function value $f(x_{2k+1})$. Since a simple way to reduce the stepsize is to multiply it by some positive constant less than 1, we also propose the following two ways to shorten the SD step:

$$\alpha_k^{SS1} = \gamma_1 \alpha_k^{SD} \tag{2.12}$$

and

$$\alpha_k^{SS2} = \begin{cases} \gamma_2 \alpha_k^{SD}, & \text{for odd } k; \\ \alpha_k^{SD}, & \text{for even } k, \end{cases} \tag{2.13}$$

where γ_1 and γ_2 are positive constants less than 1. For convenience, we call the methods (1.4)$\sim$(2.12) and (1.4)$\sim$(2.13) shortened SD step gradient methods (I) and (II), and abbreviate them as SS1 and SS2 methods.

As will be shown in Table 4.3, the performances of the SS1 and SS2 methods are similar to those of the AM method. However, one advantage of the SS1 and SS2 methods over the AM method is that they can easily be extended to the field of unconstrained optimization (see §5). In addition, the SS1 method with $\gamma_1 = 0.8$ has successfully been applied by Professor Zhiming Chen (private communications) to a real large-scale saddle-point problem. For his problem, the SS1 method only requires 220 iterations whereas his previous method based on [13] needs 290 iterations.

3. Convergence rate analysis

In this section, we analyse the convergence rate of the AM method for strictly convex quadratics. These analyses can also be extended to the (not necessarily strictly) convex case by reducing the case to the strictly convex one. More details can be found in [9], where a proof of the convergence of the BB method in the convex case is presented.

At first, we have the following theorem for the AM method in the two-dimensional case.

Theorem 3.1 *Consider the minimization problem (1.2), where $n = 2$. Let $\{x_k\}$ be the iterations generated by the AM method (1.4)~(2.6). If $g_2 \neq 0$, the sequence of gradient norms $\{\|g_k\|_2\}$ is Q-superlinearly convergent; namely, we have that*

$$\lim_{k\to\infty} \frac{\|g_{k+1}\|_2}{\|g_k\|_2} = 0. \tag{3.1}$$

Further, the Q-superlinear convergence order is only 1, namely, for any constant $\varepsilon > 0$ we have that

$$\lim_{k\to\infty} \frac{\|g_{k+1}\|_2}{\|g_k\|_2^{1+\varepsilon}} = +\infty. \tag{3.2}$$

Proof Noting that the method (1.4)~(2.6) is invariable under orthogonal transformations, we assume without loss of generality that

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}, \tag{3.3}$$

where $\lambda \geq 1$ is constant. Let $g_k^{(1)}$ and $g_k^{(2)}$ be the two components of g_k; namely,

$$g_k = (g_k^{(1)}, g_k^{(2)})^T. \tag{3.4}$$

Then we have by direct calculations that

$$\left\{ \begin{array}{rcl} g_2^{(1)} & = & \frac{\lambda(\lambda-1)(g_1^{(2)})^2}{(g_1^{(1)})^2+\lambda(g_1^{(2)})^2} g_1^{(1)}, \\ g_2^{(2)} & = & \frac{(1-\lambda)(g_1^{(1)})^2}{(g_1^{(1)})^2+\lambda(g_1^{(2)})^2} g_1^{(2)}. \end{array} \right. \tag{3.5}$$

Thus $g_2^{(1)} = 0$ ($g_2^{(2)} = 0$) if and only if at least one of the relations $\lambda = 1$, $g_1^{(1)} = 0$ and $g_1^{(2)} = 0$ holds. As a result, if $g_2 \neq 0$, we must have that $\lambda > 1$, $g_2^{(1)} \neq 0$ and $g_2^{(2)} \neq 0$. By induction, we can then prove that

$$g_k^{(1)} \neq 0, \quad g_k^{(2)} \neq 0, \quad \text{for all } k. \tag{3.6}$$

Now, let us define t_k and r_k by

$$t_k = \frac{g_{2k-1}^{(1)}}{g_{2k-1}^{(2)}} \quad \text{and} \quad r_k = \frac{g_{2k}^{(1)}}{g_{2k}^{(2)}}. \tag{3.7}$$

By the relations (3.7), (2.3) and (2.4), we can see that

$$r_k = -\lambda t_k^{-1} \tag{3.8}$$

and

$$t_{k+1} = -r_k^{-1}. \tag{3.9}$$

Then it follows from (3.8) and (3.9) that

$$t_{k+1} = \lambda^{-1} t_k = \cdots = \lambda^{-k} t_1, \tag{3.10}$$

which, with $t_1 \neq 0$, $\lambda > 1$ and (3.8), yields

$$\lim_{k\to\infty} t_k = 0 \quad \text{and} \quad \lim_{k\to\infty} |r_k| = +\infty. \tag{3.11}$$

On the other hand, we have by (2.5), (2.6), (3.7) and direct calculations that

$$\frac{\|g_{2k}\|_2^2}{\|g_{2k-1}\|_2^2} = 1 - \frac{(g_{2k-1}^T A g_{2k-1})^2}{(g_{2k-1}^T g_{2k-1})(g_{2k-1}^T A^2 g_{2k-1})} = \frac{(\lambda-1)^2 t_k^2}{(1+t_k^2)(\lambda^2+t_k^2)} \tag{3.12}$$

and

$$\frac{\|g_{2k+1}\|_2^2}{\|g_{2k}\|_2^2} = -1 + \frac{(g_{2k}^T g_{2k})(g_{2k}^T A^2 g_{2k})}{(g_{2k}^T A g_{2k})^2} = \frac{(\lambda-1)^2 r_k^2}{(\lambda+r_k^2)^2}. \tag{3.13}$$

Thus by (3.11), (3.12) and (3.13), we know that the relation (3.1) holds.

Further, substituting (3.8) into (3.13), we find that

$$\frac{\|g_{2k+1}\|_2^2}{\|g_{2k}\|_2^2} = \frac{(\lambda-1)^2 t_k^2}{(\lambda+t_k^2)^2}. \tag{3.14}$$

By (3.12), (3.14), (3.10) and $\lambda > 1$, it is easy to show that

$$\frac{\lambda-1}{\lambda^2+t_1^2}|t_k| \leq \frac{\|g_{k+1}\|_2}{\|g_k\|_2} \leq (\lambda-1)|t_k|, \quad \text{for all } k. \tag{3.15}$$

By the second inequality in (3.15) and (3.10), we can show that

$$\|g_k\|_2 \leq (\lambda-1)^{k-1} \prod_{i=1}^{k-1} |t_i| = (\lambda-1)^{k-1} |t_1|^{k-1} \lambda^{-\frac{1}{2}k(k-1)}. \tag{3.16}$$

Then we know from the first inequality in (3.15), (3.10) and (3.16) that the relation (3.2) holds for any constant $\varepsilon > 0$. Consequently, the sequence $\{\|g_k\|_2\}$ is Q-superlinearly convergent to zero with order of only 1. $\square$

Theorem 3.1 indicates that for two-dimensional strictly convex quadratics, the convergence rate of the AM method is Q-superlinear. Consequently, the AM method can avoid zigzags and improve the numerical behaviour of the SD method. It is known that the SD method can produce zigzags and it is shown in [1, 8] that the search directions generated by the SD method will tend to two different directions. The BB and AS methods can avoid zigzags, as mentioned in §1. However, only R-superlinear convergence and two-step Q-superlinear convergence are established for the BB and AS methods, respectively.

Now we consider the case of any dimension. Let $x^* = -A^{-1}b$ be the unique minimizer of the quadratic in (1.2) and define $\|x\|_A = \sqrt{x^T A x}$. Then we have the following Q-linear result for the AM method.

Theorem 3.2 *Consider the minimization problem (1.2). Let $\{x_k\}$ be the iterations generated by the AM method (1.4)∼(2.6). Denote*

$$c := \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}, \tag{3.17}$$

where λ_1 and λ_n are the maximal and minimal eigenvalues of A, respectively. Then we have that

$$\|x_{k+1} - x^*\|_A \le c\sqrt{1 + c^2 - c^4}\, \|x_k - x^*\|_A, \quad \textit{for all } k. \tag{3.18}$$

Thus the convergence rate of the AM method is Q-linear.

Proof If

$$\bar{\alpha}_k = \frac{g_k^T g_k}{g_k^T A g_k}, \tag{3.19}$$

the stepsize α_k in (2.6) can be written as

$$\alpha_k = \tau_k \bar{\alpha}_k, \tag{3.20}$$

where

$$\tau_k = \begin{cases} \frac{(g_k^T A g_k)^2}{(g_k^T g_k)(g_k^T A^2 g_k)}, & \text{for odd } k; \\ 1, & \text{for even } k. \end{cases} \tag{3.21}$$

Define $e_k = x_k - x^*$ and $E(x) = \|x - x^*\|_A^2$. It is easy to see that

$$g_k = A e_k \tag{3.22}$$

and

$$e_{k+1} = e_k - \tau_k \bar{\alpha}_k g_k. \tag{3.23}$$

Then we have by the definition of $E(x)$ and the relations (3.22) and (3.23) that

$$\begin{aligned} \frac{E(x_k) - E(x_{k+1})}{E(x_k)} &= \frac{e_k^T A e_k - e_{k+1}^T A e_{k+1}}{e_k^T A e_k} \\ &= \frac{2\tau_k \bar{\alpha}_k g_k^T A e_k - \tau_k^2 \bar{\alpha}_k^2 g_k^T A g_k}{e_k^T A e_k} \\ &= \frac{(2\tau_k - \tau_k^2)(g_k^T g_k)^2}{(g_k^T A g_k)(g_k^T A^{-1} g_k)}. \end{aligned} \tag{3.24}$$

By the Kantorovich inequality and the definition of c in (3.17), we have that

$$\frac{(v^T A v)^2}{(v^T v)(v^T A^2 v)} \ge \frac{4\lambda_1 \lambda_n}{(\lambda_1 + \lambda_n)^2} = 1 - c^2. \tag{3.25}$$

Applying this and (2.8) in (3.21), we get that

$$1 - c^2 \le \tau_k \le 1. \tag{3.26}$$

It follows from (3.26) and $c \in (0, 1)$ that

$$2\tau_k - \tau_k^2 = 1 - (1 - \tau_k)^2 \ge 1 - c^4. \tag{3.27}$$

Letting $v = A^{-\frac{1}{2}} g_k$ in (3.25), we also have that

$$\frac{(g_k^T g_k)^2}{(g_k^T A g_k)(g_k^T A^{-1} g_k)} \geq 1 - c^2. \tag{3.28}$$

Hence, by (3.24), (3.27) and (3.28), we obtain

$$\frac{E(x_k) - E(x_{k+1})}{E(x_k)} \geq (1 - c^2)(1 - c^4), \tag{3.29}$$

or equivalently,

$$E(x_{k+1}) \leq [1 - (1 - c^2)(1 - c^4)] E(x_k) = c^2(1 + c^2 - c^4) E(x_k). \tag{3.30}$$

The relation (3.30) and the definition of $E(x_k)$ implies the truth of (3.18). Noting that

$$c\sqrt{1 + c^2 - c^4} < c\sqrt{1 + c^2 - c^4 + (1 - c^2)^2} = c\sqrt{2 - c^2} < 1, \tag{3.31}$$

we know from (3.18) that the AM method is Q-linearly convergent. $\square$

Thus we have proved that the AM method is globally convergent for any-dimensional strictly convex quadratic and its convergence rate is Q-linear. Noting that

$$f(x) - f(x^*) = \frac{1}{2}\|x - x^*\|_A^2, \tag{3.32}$$

the relation (3.18) indicates that $\{f(x_k);\ k = 1, 2, \ldots\}$ is monotonically decreasing.

If we consider the iterations with odd and even indices, respectively, we can improve (3.18) to

$$\|x_{k+1} - x^*\|_A \leq \begin{cases} c\sqrt{1 + c^2 - c^4}\,\|x_k - x^*\|_A, & \text{for odd } k; \\ c\,\|x_k - x^*\|_A, & \text{for even } k. \end{cases} \tag{3.33}$$

For the SD method, however, we can establish the relation

$$\|x_{k+1} - x^*\|_A \leq c\,\|x_k - x^*\|_A. \tag{3.34}$$

The convergence relation (3.34) is better than (3.18) and (3.33) for $1 + c^2 - c^4 > 1$. Nevertheless, our numerical experiments in §4 show that the AM method performs much better than the SD method. Thus it still remains to study how to provide theoretical evidences showing that the AM method is better than the SD method in any dimension.

Now let us discuss the convergence properties of the SS1 and SS2 methods (see (2.12) and (2.13) for their formulae of stepsize), as suggested by one of the referees. Firstly, if we define

$$\tau_k = \begin{cases} \gamma_2, & \text{for odd } k; \\ 1, & \text{for even } k, \end{cases} \tag{3.35}$$

the relation (3.20) still holds for the SS2 method. Further, denoting c be the constant $\sqrt{1 - \gamma_2}$, which is less than 1, we can see that the proof of Theorem 3.2 still applies to the SS2 method. Hence the SS2 method is also Q-linearly convergent for any dimensional strictly convex quadratic. This analysis can be extended to the SS1 method by letting $\tau_k \equiv \gamma_1$ and $c = \sqrt{1-\gamma_1}$.

Secondly, we present here an analysis of the SS2 method for two-dimensional quadratics. In this case, we do not expect some kind of superlinear convergence result for the SS2 method. However, in our preliminary numerical experiments, we did not find that the SS2 method produces any zigzag, unlike the SD method. Noting that the SS2 method (1.4)∼(2.13) is also invariable under orthogonal transformations, we assume that the Hessian of the quadratic has the form (3.3). Assume also that the gradient $g_k = (g_k^{(1)}, g_k^{(2)})^T$ has nonzero components for all k, and t_k and r_k are given by (3.7). Noting that (2.5) is still available and the stepsize of the SS2 method at any even iteration is identical to the SD stepsize, we can obtain, by calculating the components of g_{2k+1} and the definitions of t_k and r_k,

$$t_{k+1} = -r_k^{-1}. \tag{3.36}$$

If another SD step is carried out at the next iteration, then we will get

$$r_{k+1} = -t_{k+1}^{-1} = r_k, \tag{3.37}$$

which shows that the SD method can produce zigzags in the two-dimensional case, one major disadvantage of the SD method. However, for the SS2 method, we have by (2.13) and direct calculations, that

$$r_{k+1} = \frac{(\lambda-\gamma_2) + (1-\gamma_2)t_{k+1}^2}{\lambda(1-\gamma_2) + (1-\lambda\gamma_2)t_{k+1}^2}\, t_{k+1} \tag{3.38}$$

(if the denominator in (3.38) is zero, we will have $g_{2k+2}^{(2)} = 0$). Substituting (3.36) into the above relation, we get

$$r_{k+1} = r_k - \frac{(1-\gamma_2)[1 + (\lambda+1)r_k^2 + \lambda r_k^4]}{r_k[(1-\lambda\gamma_2) + \lambda(1-\gamma_2)r_k^2]}, \tag{3.39}$$

which together with $\gamma_2 < 1$ indicates that $r_{k+1} \neq r_k$ for the SS2 method. Thus we can see that one role of the shortened SD step is that, it prevents the occurrence of zigzags. Further, in our preliminary numerical experiments, we did not find any cycle in the sequences of $\{r_k\}$ and $\{t_k\}$. This property of avoiding zigzags might explain from a different angle why the SS2 method generally performs better than the SD method, as will be shown in the next section. A similar property has been found in our implementations of the SS1 method.

4. Numerical experiments

We have implemented the AM method (1.4)∼(2.6) with MATLAB 6.0 on a notebook ASUS L8400-B. Our numerical experiments are divided into three parts. Firstly, to check the Q-superlinear result stated in Theorem 3.1, we tested the AM method for the following two-dimensional convex quadratic:

$$f(x) = \frac{1}{2}x^T \begin{pmatrix} 0.2 & 0 \\ 0 & 2 \end{pmatrix} x, \quad x \in R^2. \tag{4.1}$$

Starting with the point $x_1 = (1000, 1000)^T$, the function values and the stepsizes produced by the AM method are listed in Table 4.1.

Table 4.1 Numerical results of AM for (4.1)

k	$f(x_k)$	${\alpha_k}^{-1}$
1	1.10000000e+06	1.99820180e+00
2	8.09846123e+04	2.00179982e–01
3	6.55313486e+01	1.99998200e+00
4	5.30272643e–02	2.00001800e–01
5	4.29516502e–07	1.99999982e–01
6	3.47904890e–12	2.00000018e–01
7	2.81802933e–19	

From Table 4.1, we can see that the AM method is indeed Q-superlinearly convergent for the two-dimensional convex quadratic, which demonstrates the result in Theorem 3.1. In addition, we also tested the SD, BB, and AS methods for this problem. To get $f(x_k) \leq 10^{-16}$, the numbers of iterations required by the three methods are 20, 9, and 8, respectively. Thus this example shows that the SD method is clearly the slowest, whereas the BB, AS, and AM methods are comparable.

Secondly, we tested a real problem ([10]). Consider the linear system $Ax = b$ with the coefficient matrix $A \equiv (a_{ij})$ given by

$$a_{ij} = \begin{cases} \frac{2}{h^2}, & \text{if } i = j; \\ -\frac{1}{h^2}, & \text{if } i = j \pm 1; \\ 0, & \text{otherwise,} \end{cases} \tag{4.2}$$

where $h = 11/n$ and $n = 1000$. Linear systems of this kind appear frequently in the numerical solution of two-point boundary-value problems. For the choice of b, we adopted the same as in [10]. That is to say, we generated a random solution x^* with components between -10 and 10 and computed $b = Ax^*$. We used thirty initial points x_1. The stopping condition in this example is

$$\|Ax_k - b\|_2 \leq \theta \, \|Ax_1 - b\|_2 \tag{4.3}$$

with different values of θ. Listed in Table 4.2 are the numbers of average iterations required by the conjugate gradient (CG), SD, BB, AS, and AM methods. For the tested θ, we see from the table that the BB, AS, and AM methods are much better than the SD method and competitive with the CG method. For this real problem, the AM method gave the least average number of iterations for all tested values of θ.

Table 4.2. Numerical comparisons for a real problem

θ	10^{-1}	10^{-2}	10^{-3}	10^{-4}
CG	4	10	27	95
SD	4	16	71	528
BB	4	10	29	73
AS	4	11	26	78
AM	4	10	22	68

Thirdly, we compared the CG, BB, AS, and AM methods for random problems ([10]). Consider the linear system $Ax = b$, where $A = QDQ^T$,

$$Q = (I - 2w_3w_3^T)(1 - 2w_2w_2^T)(I - 2w_1w_1^T), \tag{4.4}$$

and w_1, w_2, and w_3 are unitary random vectors. $D = diag(\sigma_1, \ldots, \sigma_n)$ is a diagonal matrix, in which $\sigma_1 = 1$, $\sigma_n = cond$, and σ_j is randomly generated between 1 and *cond* for $j = 2, \ldots, n-1$. The entries of the right-hand-side b are randomly generated between -10 and 10. The initial point is the null vector of R^n and, in these tests, we used $n = 5000$. Again, the stopping condition (4.3) is used here. For the tested θ, the numbers of iterations required by each method are listed in Table 4.3.

From Table 4.3, we see that if $\theta \leq 10^{-2}$ (namely, $\theta = 10^{-2}$, 10^{-3} and 10^{-4}), then CG is clearly the winner. In this case, if the condition number is small ($cond = 10^1, 10^2$), the AM method is better than the BB and AS methods; if the condition number becomes large ($cond = 10^3, 10^4$), then the BB and AS methods require less numbers of iterations than the AM method. Thus, this example shows that the AM method is more influenced by the problem condition than the BB and AS methods.

If $\theta = 10^{-1}$, namely if only a solution with a low precision is required, we see that the BB, AS, and AM methods are competitive with the CG method even when the problem condition is large. In this case, we can see that the AM method always requires fewer number of iterations than the BB and AS methods. This behaviour favours the application of gradient methods, even the AM method, in the context where a high precision is neither necessary nor recommended, for example in an inexact-Newton context (see [10]).

Table 4.3 Numerical comparisons for random problems

$cond$	θ	CG	BB	AS	AM	$SS1$	$SS2$
10^1	10^{-1}	5	7	7	5	7	5
	10^{-2}	9	16	14	10	10	11
	10^{-3}	12	22	17	15	13	15
	10^{-4}	16	23	20	21	19	20
10^2	10^{-1}	11	18	15	11	10	10
	10^{-2}	23	30	30	25	33	34
	10^{-3}	34	47	47	44	60	56
	10^{-4}	46	61	63	63	85	78
10^3	10^{-1}	24	23	18	14	13	14
	10^{-2}	63	68	82	80	90	80
	10^{-3}	98	150	143	173	175	173
	10^{-4}	132	205	197	272	261	272
10^4	10^{-1}	51	21	19	16	14	13
	10^{-2}	154	171	182	273	217	210
	10^{-3}	213	434	447	791	577	522
	10^{-4}	267	628	667	1281	883	811

Listed in Table 4.3 are also the numerical results of the SS1 method (2.12) and the SS2 method (2.13). We found that both methods can give good numerical results while $\gamma_1, \gamma_2 \in [0.7, 0.9]$, whereas the best choices for γ_1 and γ_2 seem to be

$$\gamma_1^* = 0.8, \qquad \gamma_2^* = 0.75. \tag{4.5}$$

The fact that γ_1^* is slightly greater than γ_2^* is largely in agreement with our intuition since the SS1 method never takes a full SD step but SS2 does every other iteration. These values of γ_1^* and γ_2^* are used in generating the results of the SS1 and SS2 methods in Table 4.3. From the table, we can see that both the SS1 and SS2 methods with the above choices perform as well as the AM method. If the condition number of the matrix is 10^4, the SS1 and SS2 methods are much better than the AM method. This might hint that the SS1 and SS2 methods are less affected by the problem condition than the AM method.

5. Generalizations to unconstrained optimization

In this section, we consider how to extend the AM method or its variants to the field of unconstrained optimization. In this part, we will only compare these monotone gradient methods for we believe where there are indeed some cases where only a very limited iterations are required, and the user is only interested in some monotone method. For example, in the context of global optimization, where a few inexpensive line searches are preferred around most of the sample points (for instance, see [11] and some references therein). Generally, if we are allowed to compute many numbers of iterations, then we still prefer the nonmonotone methods. It has been demonstrated by a large number of numerical experiments that the BB method is very efficient with some kind of nonmonotone line search (see [16]).

Following its basic idea for linear system, a natural extension of the AM method to unconstrained optimization is alternately to minimize the function value and the gradient norm along the line. However, the step of minimizing the gradient norm is more expensive than the one of minimizing the function value. This is because, for a general function, the gradient of $\|g(x)\|_2$ at x_k is unknown, and hence one quadratic interpolation of the gradient norm requires at least three points along the line. Since the SS1 and SS2 methods perform as well as the AM method, as shown in Table 4.3, we consider in the following their extensions to unconstrained optimization.

To do so, we first see that in case of convex quadratics, it follows from (1.4) and (2.12) that

$$g_{k+1}^T d_k = (1 - \gamma_1) g_k^T d_k < 0, \tag{5.1}$$

where $d_k = -g_k$. For a general function, assume that d_k is a descent search direction at the current point x_k and $\phi_k(\alpha)$ is some quadratic or cubic interpolation function. In this case, for the exact line search or the strong Wolfe line search *etc.*, a new trial stepsize α^* is calculated such that the derivative $\phi_k'(\alpha^*) = 0$. Based on its property (5.1) in the case of convex quadratics, for the SS1 method we require the new trial stepsize α^* to satisfy $\phi_k'(\alpha^*) = (1 - \gamma_1) g_k^T d_k$. Further, letting $\sigma = 1 - \gamma_1$, we require the final value of the stepsize α_k to satisfy the line search conditions:

$$f(x_k + \alpha d_k) \le f(x_k) + \delta \alpha g_k^T d_k \tag{5.2}$$

and

$$(\sigma+\tau)g_k^T d_k \le g(x_k+\alpha d_k)^T d_k \le (\sigma-\tau)g_k^T d_k, \tag{5.3}$$

where $\delta > 0$ and $\tau \ge 0$ are parameters. If $\sigma = 0$, (5.2)$\sim$(5.3) reduce to the strong Wolfe conditions exactly. This kind of line search can also be used in the shortened SD step of the SS2 method.

Using the above idea, we tested the SD method, the SS1 method, and the SS2 method with the line search (5.2)$\sim$(5.3). For the SD method, the parameters are $\sigma = 0$ and $\tau = 0.1$. In other words, the strong line search is used for the SD method. For the SS1 method, based on (4.5) and $\sigma = 1 - \gamma_1$, we choose $\sigma = 0.2$. The value of τ is set to 0.15 such that the corresponding line search strictly excludes the one-dimensional exact minimizer. In a similar way, for the SS2 method, we used $\sigma = 0$, $\tau = 0.1$ at odd iterations and $\sigma = 0.25$, $\tau = 0.2$ at even iterations. In addition, the value of δ was always set to 10^{-4}.

We tested the above three methods on an SGI indigo workstation in double precision. Since the BB and AS methods are nonmonotone, they are not tested in this kind of numerical experiment. All algorithms were written in FORTRAN 77. The test problems were taken from Moré *et al.* [14], except "Strictly Convex 1" and "Strictly Convex 2" that are provided in [16]. The test problems from [14] are numbered in the following way: "MGHi" means the i-th problem in [14]. The total number of test problems is 26. For all the algorithms, the stopping condition is

$$\|g_k\|_2/(1+|f_k|) \le \varepsilon \|g_1\|_2/(1+|f_1|). \tag{5.4}$$

The value of ε is set to 10^{-3}, except for 10^{-2} for problem 22 and 10^{-1} for problems 7 and 8. The maximum number of function evaluations is set to 9999. During each line search, we compute at most 20 trial points. We mark it with "$Failed$" if either of these limits is exceeded. The numerical results are listed in the form I-F-G in Table 5.1, where I, F, G denote the number of iterations, the number of function evaluations, and the number of gradient evaluations, respectively.

From Table 5.1, we can see that both the SS1 and SS2 methods clearly perform much better than the SD method.

6. Conclusions and discussions

In this paper, we have proposed an alternate minimization gradient (AM) method, namely, (1.4)$\sim$(2.6) . The stepsize in the method alternately minimizes the function value and the gradient norm along the line. The AM method is a monotone method like the SD method, but is faster than the latter. Specifically, we have proved that the AM method is Q-superlinearly convergent for two-dimensional strictly convex quadratics. Hence the AM method can avoid zigzags, that are very likely to occur in the SD method.

We have also exposed the fact that the AM method carries out a full SD step after a shortened SD step. Following this, we proposed two shortened SD step gradient methods, namely, the SS1 method (1.4)$\sim$(2.12) and the SS2 method (1.4)$\sim$(2.13). Our numerical experiments for linear systems show that the two methods perform as well as the AM method. Further, by designing a new kind of inexact line search, we extended the SS1 and SS2 methods to unconstrained optimization. It was found that their

Table 5.1. Numerical Comparisons of gradient algorithms

Problem	n	SD	SS1	SS2
MGH11	3	337-1011-393	*Failed*	436-1268-529
MGH14	4	2227-6723-2265	41-147-60	1827-6007-2376
MGH18	6	3200-6491-4546	2876-6074-3892	2272-4879-3047
MGH22	16	2460-7388-2468	264-760-300	226-668-260
MGH24	20	322-975-362	123-355-153	14-59-29
	40	2254-6760-2321	439-1290-504	503-1507-575
MGH28	20	10-125-118	10-129-123	6-49-43
	50	9-99-91	6-42-36	6-50-44
MGH30	50	45-140-50	24-76-30	23-72-72
	500	53-163-57	28-88-34	27-86-33
MGH31	50	30-114-53	15-52-21	27-115-62
	500	23-79-32	20-64-23	16-54-26
MGH22	100	*Failed*	427-1228-482	682-1983-746
	500	*Failed*	1285-3685-1458	1130-3269-1248
MGH25	100	6-48-18	23-147-52	7-51-18
	1000	8-77-26	31-257-81	15-126-37
MGH21	1000	161-243-243	71-140-127	77-155-142
	10000	105-211-165	81-162-150	86-179-169
MGH23	1000	*Failed*	350-1098-423	587-1795-651
	10000	*Failed*	331-1018-416	559-1720-635
MGH26	1000	2522-5084-5069	510-878-821	455-935-906
	10000	1573-3128-3076	257-589-514	432-918-861
Strictly Convex 1	1000	3-5-5	5-11-11	3-6-6
	10000	3-5-5	5-11-11	3-6-6
Strictly Convex 2	1000	20-69-30	21-66-27	19-61-24
	10000	20-93-39	22-83-32	18-69-26

numerical performance is much better than that of the SD method. The new kind of line search includes the strong Wolfe line search as a special case. However, if $\sigma \geq \tau > 0$, the line search tries to find a stepsize that is shorter than the exact one-dimensional minimizer.

We have also proved that, for any-dimensional strictly convex quadratic, the convergence rate of the AM method is Q-linear. However, either the convergence relation (3.18) or (3.33) for the AM method is not so good as the relation (3.34) for the SD method. Although it is easy to show that a suitable reduction in the k-th SD stepsize will yield a lower function value at the $(k+2)$-th iteration, it still remains to study how to provide theoretical evidence showing that the AM method, or the SS1 and SS2 methods, are better than the SD method. We believe that an answer to this question will also be very helpful in understanding why the BB and AS methods are faster than the SD method in practical computations.

Acknowledgements. Some numerical experiments in this paper were done while the first author was visiting the Department of Mathematics at the Federal University of Parana. He would like to thank the CNPq in Brazil for supporting his visit. The authors also thank Hongchao Zhang for his careful reading of an early version of this paper, and Professor Iain Duff and the two anonymous referees for their many valuable suggestions and comments.

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), pp. 1-17.

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

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

[4] Y. H. Dai, *Alternate stepsize gradient method*, Research report AMSS-2001-041, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, 2001.

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

[6] R. Fletcher, *Low storage methods for unconstrained optimization,* Lectures in Applied Mathematics (AMS), 26 (1990), pp. 165-179.

[7] R. Fletcher and C. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149-154.

[8] G. E. Forsythe, *On the asymptotic directions of the s-dimensional optimum gradient method*, Numerische Mathematik, 11 (1968), pp. 57-76.

[9] A. Friedlander, J. M. Martínez, and M. Raydan, *A new method for large-scale box constrained convex quadratic minimization problems*, Optimization Methods and Software, 5 (1995), pp. 57-74.

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

[11] F. J. Hickernell and Y. Yuan, *A simple multistart algorithm for global optimization*, OR Transactions, 1:2 (1997), pp. 1-11.

[12] M. R. Hestenes and E. L. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards Sect., 5:49 (1952), pp. 409-436.

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

[14] J. J. Moré, B. S. Garbow and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Transactions on Mathematical Software, 7 (1981), pp. 17-41.

[15] M. Raydan, *On the Barzilai and Borwein choice of steplength for the gradient method*, IMA J. Numer. Anal., 13 (1993), pp. 321-326.

[16] M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7:1 (1997), pp. 26-33.