ON THE ASYMPTOTIC CONVERGENCE AND ACCELERATION OF GRADIENT METHODS*

YAKUI HUANG†, YU-HONG DAI‡, XIN-WEI LIU§, AND HONGCHAO ZHANG¶

Abstract. We consider the asymptotic behavior of a family of gradient methods, which include the steepest descent and minimal gradient methods as special instances. It is proved that each method in the family will asymptotically zigzag between two directions. Asymptotic convergence results of the objective value, gradient norm, and stepsize are presented as well. To accelerate the family of gradient methods, we further exploit spectral properties of stepsizes to break the zigzagging pattern. In particular, a new stepsize is derived by imposing finite termination on minimizing two-dimensional strictly convex quadratic function. It is shown that, for the general quadratic function, the proposed stepsize asymptotically converges to the reciprocal of the largest eigenvalue of the Hessian. Furthermore, based on this spectral property, we propose a periodic gradient method by incorporating the Barzilai-Borwein method. Numerical comparisons with some recent successful gradient methods show that our new method is very promising.

Key words. gradient methods, asymptotic convergence, spectral property, acceleration of gradient methods, Barzilai-Borwein method, unconstrained optimization, quadratic optimization

AMS subject classifications. 90C20, 90C25, 90C30

1. Introduction.

The gradient method is well-known for solving the following unconstrained optimization

$$\min_{x\in\mathbb{R}^n} \ f(x), \tag{1.1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is continuously differentiable, especially when the dimension n is large. In particular, at k-th iteration gradient methods update the iterates by

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

where $g_k=\nabla f(x_k)$ and $\alpha_k>0$ is the stepsize determined by the method.

One simplest nontrivial nonlinear instance of (1.1) is the quadratic optimization

$$\min_{x\in\mathbb{R}^n} \ f(x)=\frac{1}{2}x^{\mathsf{T}}Ax-b^{\mathsf{T}}x, \tag{1.3}$$

where $b\in\mathbb{R}^n$ and $A\in\mathbb{R}^{n\times n}$ is symmetric and positive definite. Solving (1.3) efficiently is usually a pre-requisite for a method to be generalized to solve more general optimization. In addition, by Taylor's expansion, a general smooth function can be approximated by a quadratic function near the minimizer. So, the local convergence behaviors of gradient methods are often reflected by solving (1.3). Hence, in this

*August 19, 2019, This research was supported by the National Natural Science Foundation of China (11701137, 11631013, 11671116), by the National 973 Program of China (2015CB856002), by the China Scholarship Council (No. 201806705007), and by the USA National Science Foundation (1522654, 1819161).

†Institute of Mathematics, Hebei University of Technology, Tianjin 300401, China (huangyakui2006@gmail.com).

‡LSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China (dyh@lsec.cc.ac.cn, http://lsec.cc.ac.cn/~dyh/).

§Institute of Mathematics, Hebei University of Technology, Tianjin 300401, China (mathlxw@hebut.edu.cn).

¶Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803-4918, USA (hozhang@math.lsu.edu, https://www.math.lsu.edu/~hozhang/).

paper, we focus on studying the convergence behaviors and propose efficient gradient methods for solving (1.3) efficiently.

In [4], Cauchy proposed the steepest descent (SD) method that solves (1.3) by using the exact stepsize

$$\alpha_k^{SD} = \arg\min_{\alpha} \ f(x_k - \alpha g_k) = \frac{g_k^{\mathsf{T}} g_k}{g_k^{\mathsf{T}} A g_k}. \tag{1.4}$$

Although α_k^{SD} minimizes f along the steepest descent direction, the SD method often performs poorly in practice and has linear converge rate [1, 18] as

$$\frac{f(x_{k+1}) - f^*}{f(x_k) - f^*} \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^2, \tag{1.5}$$

where f^* is the optimal function value of (1.3) and $\kappa = \lambda_n/\lambda_1$ is the condition number of A with λ_1 and λ_n being the smallest and largest eigenvalues of A, respectively. Thus, if κ is large, the SD method may converge very slowly. In addition, Akaike [1] proved that the gradients will asymptotically alternate between two directions in the subspace spanned by the two eigenvectors corresponding to λ_1 and λ_n. So, the SD method often has zigzag phenomenon near the solution. In [18], Forsythe generalized Akaike's results to the so-called optimum s-gradient method and Pronzato et al. [27] further generalized the results to the so-called P-gradient methods in the Hilbert space. Recently, by employing Akaike's results, Nocedal et al. [26] presented some insights for asymptotic behaviors of the SD method on function values, stepsizes and gradient norms.

Contrary to the SD method, the minimal gradient (MG) method [10] computes its stepsize by minimizing the gradient norm,

$$\alpha_k^{MG} = \arg\min_{\alpha} \ \|g(x_k - \alpha g_k)\| = \frac{g_k^{\mathsf{T}} A g_k}{g_k^{\mathsf{T}} A^2 g_k}. \tag{1.6}$$

It is widely accepted that the MG method can also perform poorly and has similar asymptotic behavior as the SD method, i.e., it will asymptotically zigzag in a two-dimensional subspace. In [32], the authors provide some interesting analyses on α_k^{MG} for minimizing two-dimensional quadratics. However, rigorous asymptotic convergence results of the MG method for minimizing general quadratic function are very limit in literature.

In order to avoid the zigzagging pattern, it is useful to determine the stepsize without using the exact stepsize because it would yield a gradient perpendicular to the current one. Barzilai and Borwein [2] proposed the following two novel stepsizes:

$$\alpha_k^{BB1} = \frac{s_{k-1}^{\mathsf{T}} s_{k-1}}{s_{k-1}^{\mathsf{T}} y_{k-1}} \quad \text{and} \quad \alpha_k^{BB2} = \frac{s_{k-1}^{\mathsf{T}} y_{k-1}}{y_{k-1}^{\mathsf{T}} y_{k-1}}, \tag{1.7}$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$. The BB method (1.7) performs quite well in practice, though it generates a nonmonotone sequence of objective values. Due to its simplicity and efficiency, the BB method has been widely studied [6, 7, 8, 17, 28] and extended to general problems and various applications, see [3, 22, 23, 24, 25, 29]. Another line of research to break the zigzagging pattern and accelerate the convergence is occasionally applying short stepsizes that approximate $1/\lambda_n$ to

eliminate the corresponding component of the gradient. One seminal work is due to Yuan [30, 31], who derived the following stepsize:

$$\alpha_k^Y = \frac{2}{\frac{1}{\alpha_{k-1}^{SD}} + \frac{1}{\alpha_k^{SD}} + \sqrt{\left(\frac{1}{\alpha_{k-1}^{SD}} - \frac{1}{\alpha_k^{SD}}\right)^2 + \frac{4\|g_k\|^2}{(\alpha_{k-1}^{SD}\|g_{k-1}\|)^2}}}. \tag{1.8}$$

Dai and Yuan [11] further suggested a new gradient method with

$$\alpha_k^{DY} = \begin{cases} \alpha_k^{SD}, & \text{if } \mathrm{mod}(k,4) < 2; \\ \alpha_k^{Y}, & \text{otherwise.} \end{cases} \tag{1.9}$$

The DY method (1.9) is a monotone method and appears very competitive with the nonmonotone BB method. Recently, by employing the results in [1, 26], De Asmundis et al. [12] show that the stepsize α_k^Y converges to $1/\lambda_n$ if the SD method is applied to problem (1.3). This spectral property is the key to break the zigzagging pattern.

In [9], Dai and Yang developed the asymptotic optimal gradient (AOPT) method whose stepsize is given by

$$\alpha_k^{AOPT} = \frac{\|g_k\|}{\|Ag_k\|}. \tag{1.10}$$

Unlike the DY method, the AOPT method only has one stepsize. In addition, they show that α_k^{AOPT} asymptotically converges to $\frac{2}{\lambda_1+\lambda_n}$, which is in some sense an optimal stepsize since it minimizes $\|I - \alpha A\|$ over α [9, 16]. However, the AOPT method also asymptotically alternates between two directions. To accelerate the AOPT method, Huang et al. [21] derived a new stepsize that converges to $1/\lambda_n$ during the AOPT iterates and further suggested a gradient method to exploit spectral properties of the stepsizes. For the latest developments of exploiting spectral properties to accelerate gradient methods, see [12, 13, 14, 20, 21].

In this paper, we present the analysis on the asymptotic behaviors of gradient methods and the techniques for breaking the zigzagging pattern. For a uniform analysis, we consider the following stepsize

$$\alpha_k = \frac{g_k^{\mathsf{T}}\Psi(A)g_k}{g_k^{\mathsf{T}}\Psi(A)Ag_k}, \tag{1.11}$$

where Ψ is a real analytic function on $[\lambda_1, \lambda_n]$ and can be expressed by Laurent series

$$\Psi(z) = \sum_{k=-\infty}^{\infty} c_k z^k, \quad c_k \in \mathbb{R},$$

such that $0 < \sum_{k=-\infty}^{\infty} c_k z^k < +\infty$ for all $z \in [\lambda_1, \lambda_n]$. Apparently, α_k is a family of stepsizes that would give a family of gradient methods. When $\Psi(A) = A^u$ for some nonnegative integer u, we get the following stepsize

$$\alpha_k = \frac{g_k^{\mathsf{T}}A^u g_k}{g_k^{\mathsf{T}}A^{u+1}g_k}. \tag{1.12}$$

The α_k^{SD} and α_k^{MG} simply correspond to the cases $u = 0$ and $u = 1$, respectively.

We will present theoretical analysis on the asymptotic convergence on the family of gradient methods whose stepsize can be written in the form (1.11), which provides

justifications for the zigzag behaviors of all these gradient methods including the SD and MG methods. In particular, we show that each method in the family (1.11) will asymptotically alternate between two directions associated with the two eigenvectors corresponding to λ_1 and λ_n. Moreover, we analyze the asymptotic behaviors of the objective value, gradient norm, and stepsize. It is shown that, when $\Psi(A) \neq I$, the two sequences $\left\{\frac{\Delta_{2k+1}}{\Delta_{2k}}\right\}$ and $\left\{\frac{\Delta_{2k+2}}{\Delta_{2k+1}}\right\}$ may converge at different speeds, while the odd and even subsequences $\left\{\frac{\Delta_{2k+3}}{\Delta_{2k+1}}\right\}$ and $\left\{\frac{\Delta_{2k+2}}{\Delta_{2k}}\right\}$ converge at the same rate, where $\Delta_k = f(x_k) - f^*$. Similar property is also possessed by the gradient norm sequence. In addition, we show each method in (1.11) has the same worst asymptotic rate.

In order to accelerate the gradient methods (1.11), we investigate techniques for breaking the zigzagging pattern. We derive a new stepsize $\tilde{\alpha}_k$ based on finite termination for minimizing two-dimensional strictly convex quadratic function. For the n-dimensional case, we prove that $\tilde{\alpha}_k$ converges to $1/\lambda_n$ when gradient methods (1.11) are applied to problem (1.3). Furthermore, based on this spectral property, we propose a periodic gradient method, which, in a periodic mode, alternately uses the BB stepsize, stepsize (1.11) and our new stepsize $\tilde{\alpha}_k$. Numerical comparisons of the proposed method with the BB [2], DY [11], ABBmin2 [19], and SDC [12] methods show that the new gradient method is very efficient. Our theoretical results also significantly improve and generalize those in [1, 26], where only the SD method (i.e., $\Psi(A) = I$) is considered. We point out that [27] does not analyze the asymptotic behaviors of the objective value, gradient norm, and stepsize, though (1.11) is similar to the P-gradient methods in [27]. Moreover, we develop techniques for accelerating these zigzag methods with simpler analysis. Notice that α_k^{AOPT} can not be written in the form (1.11). Thus, our results are not applicable to the AOPT method. On the other hand, the analysis of the AOPT method presented in [9] can not be applied directly to the family of methods (1.11).

The paper is organized as follows. In Section 2, we analyze the asymptotic behaviors of the family of gradient methods (1.11). In Section 3, we accelerate the gradient methods (1.11) by developing techniques to break its zigzagging pattern and propose a new periodic gradient method. Numerical experiments are presented in Section 4. Finally, some conclusions and discussions are made in Section 5.

2. Asymptotic behavior of the family (1.11). In this section, we present a uniform analysis on the asymptotic behavior of the family of gradient methods (1.11) for general n-dimensional strictly convex quadratics.

Let $\{\lambda_1, \lambda_2, \cdots, \lambda_n\}$ be the eigenvalues of A, and $\{\xi_1, \xi_2, \ldots, \xi_n\}$ be the associated orthonormal eigenvectors. Noting that the gradient method is invariant under translations and rotations when applying to a quadratic function. For theoretical analysis, we can assume without loss of generality that

$$A = \mathrm{diag}\{\lambda_1, \lambda_2, \cdots, \lambda_n\}, \quad 0 < \lambda_1 < \lambda_2 < \cdots < \lambda_n. \tag{2.1}$$

Denoting the components of g_k along the eigenvectors ξ_i by $\mu_k^{(i)}$, $i = 1, \ldots, n$, i.e.,

$$g_k = \sum_{i=1}^{n} \mu_k^{(i)} \xi_i. \tag{2.2}$$

The above decomposition of gradient g_k together with the update rule (1.2) gives that

$$g_{k+1} = g_k - \alpha_k A g_k = \prod_{j=1}^{k} (I - \alpha_j A) g_0 = \sum_{i=1}^{n} \mu_{k+1}^{(i)} \xi_i, \tag{2.3}$$

where

$$\mu_{k+1}^{(i)} = (1-\alpha_k\lambda_i)\mu_k^{(i)} = \mu_0^{(i)}\prod_{j=1}^{k}(1-\alpha_j\lambda_i). \tag{2.4}$$

Defining the vector $q_k = \left(q_k^{(i)}\right)$ with

$$q_k^{(i)} = \frac{(\mu_k^{(i)})^2}{\|\mu_k\|^2} \tag{2.5}$$

and

$$\gamma_k = \frac{1}{\alpha_k} = \frac{g_k^{\mathsf{T}}\Psi(A)Ag_k}{g_k^{\mathsf{T}}\Psi(A)g_k} = \frac{\sum_{i=1}^n \Psi(\lambda_i)\lambda_i q_k^{(i)}}{\sum_{i=1}^n \Psi(\lambda_i)q_k^{(i)}}, \tag{2.6}$$

we can have from (2.4), (2.5) and (2.6) that

$$q_{k+1}^{(i)} = \frac{(\lambda_i-\gamma_k)^2 q_k^{(i)}}{\sum_{i=1}^n(\lambda_i-\gamma_k)^2 q_k^{(i)}}. \tag{2.7}$$

In addition, by the definition of q_k, we know that $q_k^{(i)} \geq 0$ for all i and

$$\sum_{i=1}^n q_k^{(i)} = 1, \quad \forall \quad k \geq 1.$$

Before establishing the asymptotic convergence of the family of gradient methods (1.11), we first give some lemmas on the properties of the sequence $\{q_k\}$.

LEMMA 2.1. *Suppose $p \in \mathbb{R}^n$ satisfies (i) $p^{(i)} \geq 0$ for all $i = 1, 2, \ldots, n$; (ii) there exist at least two i's with $p^{(i)} > 0$; and (iii) $\sum_{i=1}^n p^{(i)} = 1$. Define $T : \mathbb{R}^n \to \mathbb{R}$ be the following transformation:*

$$(Tp)^{(i)} = \frac{(\lambda_i-\gamma(p))^2 p^{(i)}}{\sum_{i=1}^n(\lambda_i-\gamma(p))^2 p^{(i)}}, \tag{2.8}$$

where

$$\gamma(p) = \frac{\sum_{i=1}^n \Psi(\lambda_i)\lambda_i p^{(i)}}{\sum_{i=1}^n \Psi(\lambda_i)p^{(i)}}. \tag{2.9}$$

Then we have

$$\Theta(Tp) \geq \Theta(p), \tag{2.10}$$

where

$$\Theta(p) = \frac{\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(p))^2 p^{(i)}}{\sum_{i=1}^n \Psi(\lambda_i)p^{(i)}}. \tag{2.11}$$

In addition, (2.10) holds with equality if and only if there are two indices, say i_1 and i_2, such that $p^{(i)} = 0$ for all $i \notin \{i_1, i_2\}$ and

$$\gamma(Tp) + \gamma(p) = \lambda_{i_1} + \lambda_{i_2}. \tag{2.12}$$

Proof. It follows from the definition of Tp that

$$\begin{aligned}\Theta(Tp) &= \frac{\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(Tp))^2(Tp)^{(i)}}{\sum_{i=1}^n \Psi(\lambda_i)(Tp)^{(i)}} \\ &= \frac{\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(Tp))^2(\lambda_i-\gamma(p))^2p^{(i)}}{\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(p))^2p^{(i)}}. \end{aligned} \tag{2.13}$$

Let us define two vectors $w=(w_i)\in\mathbb{R}^n$ and $z=(z_i)\in\mathbb{R}^n$ by

$$w_i = \sqrt{\Psi(\lambda_i)}(\lambda_i-\gamma(Tp))(\lambda_i-\gamma(p))\sqrt{p^{(i)}}$$

and

$$z_i = \sqrt{\Psi(\lambda_i)}\sqrt{p^{(i)}}.$$

Then, we have from the Cauchy-Schwarz inequality that

$$\begin{aligned}\|w\|^2\|z\|^2 &= \left(\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(Tp))^2(\lambda_i-\gamma(p))^2p^{(i)}\right)\left(\sum_{i=1}^n \Psi(\lambda_i)p^{(i)}\right) \\ &\geq (w^{\mathsf{T}}z)^2 = \left(\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(Tp))(\lambda_i-\gamma(p))p^{(i)}\right)^2. \end{aligned} \tag{2.14}$$

Using the definition of $\gamma(p)$, we can obtain that

$$\begin{aligned}&\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(Tp))(\lambda_i-\gamma(p))p^{(i)} - \sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(p))^2p^{(i)} \\ &= (\gamma(p)-\gamma(Tp))\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(p))p^{(i)} = 0, \end{aligned} \tag{2.15}$$

which together with (2.14) gives

$$\begin{aligned}&\left(\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(Tp))^2(\lambda_i-\gamma(p))^2p^{(i)}\right)\left(\sum_{i=1}^n \Psi(\lambda_i)p^{(i)}\right) \\ &\geq \left(\sum_{i=1}^n \Psi(\lambda_i)(\lambda_i-\gamma(p))^2p^{(i)}\right)^2. \end{aligned} \tag{2.16}$$

Then, the inequality (2.10) follows immediately.

The equality in (2.14) holds if and only if

$$\sqrt{\Psi(\lambda_i)}(\lambda_i-\gamma(Tp))(\lambda_i-\gamma(p))\sqrt{p^{(i)}} = C\sqrt{\Psi(\lambda_i)}\sqrt{p^{(i)}}, \quad i=1,\ldots,n \tag{2.17}$$

for some nonzero scalar C. Clearly, (2.17) holds when $p^{(i)}=0$. Suppose that there exist two indices i_1 and i_2 such that $p^{(i_1)}, p^{(i_2)}>0$. It follows from (2.17) that

$$(\lambda_{i_1}-\gamma(Tp))(\lambda_{i_1}-\gamma(p)) = (\lambda_{i_2}-\gamma(Tp))(\lambda_{i_2}-\gamma(p)).$$

So, by the assumption (2.1), we have

$$\lambda_{i_1}+\lambda_{i_2} = \gamma(Tp)+\gamma(p),$$

which again with assumption (2.1) imply that (2.17) holds if and only if p has only two nonzero components and (2.12) holds. ∎

LEMMA 2.2. *Let $p_* \in \mathbb{R}^n$ satisfy the conditions of Lemma 2.1 and T be the transformation (2.8). If p_* has only two nonzero components $p_*^{(i_1)}$ and $p_*^{(i_2)}$, we have*

$$(Tp_*)^{(i_1)} = \frac{\Psi^2(\lambda_{i_2})p_*^{(i_2)}}{\Psi^2(\lambda_{i_1})p_*^{(i_1)} + \Psi^2(\lambda_{i_2})p_*^{(i_2)}}, \tag{2.18}$$

$$(Tp_*)^{(i_2)} = \frac{\Psi^2(\lambda_{i_1})p_*^{(i_1)}}{\Psi^2(\lambda_{i_1})p_*^{(i_1)} + \Psi^2(\lambda_{i_2})p_*^{(i_2)}}, \tag{2.19}$$

$$(T^2p_*)^{(i_1)} = p_*^{(i_1)}, \quad (T^2p_*)^{(i_2)} = p_*^{(i_2)}, \tag{2.20}$$

and

$$\gamma(p_*) + \gamma(Tp_*) = \lambda_{i_1} + \lambda_{i_2}, \tag{2.21}$$

where the function γ is defined in (2.9). Moreover, $p_* = Tp_*$ if and only if

$$p_*^{(i_1)} = \frac{\Psi(\lambda_{i_2})}{\Psi(\lambda_{i_1}) + \Psi(\lambda_{i_2})} \quad \text{and} \quad p_*^{(i_2)} = \frac{\Psi(\lambda_{i_1})}{\Psi(\lambda_{i_1}) + \Psi(\lambda_{i_2})}. \tag{2.22}$$

Proof. By the definition of $\gamma(p)$, we have

$$\gamma(p_*) = \frac{\Psi(\lambda_{i_1})\lambda_{i_1}p_*^{(i_1)} + \Psi(\lambda_{i_2})\lambda_{i_2}p_*^{(i_2)}}{\Psi(\lambda_{i_1})p_*^{(i_1)} + \Psi(\lambda_{i_2})p_*^{(i_2)}}, \tag{2.23}$$

which indicates that

$$\lambda_{i_1} - \gamma(p_*) = \frac{\Psi(\lambda_{i_2})p_*^{(i_2)}(\lambda_{i_1} - \lambda_{i_2})}{\Psi(\lambda_{i_1})p_*^{(i_1)} + \Psi(\lambda_{i_2})p_*^{(i_2)}}, \quad \lambda_{i_2} - \gamma(p_*) = \frac{\Psi(\lambda_{i_1})p_*^{(i_1)}(\lambda_{i_2} - \lambda_{i_1})}{\Psi(\lambda_{i_1})p_*^{(i_1)} + \Psi(\lambda_{i_2})p_*^{(i_2)}}.$$

Then, it follows from the definition of transformation T that

$$\begin{aligned}(Tp_*)^{(i_1)} &= \frac{(\Psi(\lambda_{i_2})p_*^{(i_2)})^2 p_*^{(i_1)}}{(\Psi(\lambda_{i_2})p_*^{(i_2)})^2 p_*^{(i_1)} + (\Psi(\lambda_{i_1})p_*^{(i_1)})^2 p_*^{(i_2)}} \\ &= \frac{\Psi^2(\lambda_{i_2})p_*^{(i_2)}}{\Psi^2(\lambda_{i_1})p_*^{(i_1)} + \Psi^2(\lambda_{i_2})p_*^{(i_2)}}.\end{aligned}$$

This gives (2.18). (2.19) can be proved similarly. By (2.18) and (2.19), we have

$$\begin{aligned}(T^2p_*)^{(i_1)} &= \frac{\Psi^2(\lambda_{i_2})(Tp_*)^{(i_2)}}{\Psi^2(\lambda_{i_1})(Tp_*)^{(i_1)} + \Psi^2(\lambda_{i_2})(Tp_*)^{(i_2)}} \\ &= \frac{\Psi^2(\lambda_{i_1})\Psi^2(\lambda_{i_2})p_*^{(i_1)}}{\Psi^2(\lambda_{i_1})\Psi^2(\lambda_{i_2})p_*^{(i_2)} + \Psi^2(\lambda_{i_1})\Psi^2(\lambda_{i_2})p_*^{(i_1)}} \\ &= \frac{p_*^{(i_1)}}{p_*^{(i_1)} + p_*^{(i_2)}} = p_*^{(i_1)}.\end{aligned}$$

$(T^2p_*)^{(i_2)}$ follows similarly. This proves (2.20).

Again by (2.18), (2.19) and the definition of function γ in (2.9), we have

$$\gamma(Tp_*) = \frac{\lambda_{i_1}\Psi(\lambda_{i_2})p_*^{(i_2)} + \lambda_{i_2}\Psi(\lambda_{i_1})p_*^{(i_1)}}{\Psi(\lambda_{i_1})p_*^{(i_1)} + \Psi(\lambda_{i_2})p_*^{(i_2)}}. \tag{2.24}$$

Then, the equality (2.21) follows from (2.23) and (2.24). For (2.22), let

$$p_*^{(i_1)} = \frac{\Psi^2(\lambda_{i_2})p_*^{(i_2)}}{\Psi^2(\lambda_{i_1})p_*^{(i_1)} + \Psi^2(\lambda_{i_2})p_*^{(i_2)}}.$$

Rearranging terms and using $p_*^{(i_1)} + p_*^{(i_2)} = 1$, we have

$$\Psi^2(\lambda_{i_1})(p_*^{(i_1)})^2 = \Psi^2(\lambda_{i_2})(p_*^{(i_2)})^2,$$

which implies that

$$\Psi(\lambda_{i_1})p_*^{(i_1)} = \Psi(\lambda_{i_2})p_*^{(i_2)}.$$

This together with the fact $p_*^{(i_1)} + p_*^{(i_2)} = 1$ yields (2.22). ∎

LEMMA 2.3. *Let $p \in \mathbb{R}^n$ satisfy the conditions of Lemma 2.1 and T be the transformation (2.8). Then, there exists a p_* satisfying*

$$\lim_{k\to\infty} T^{2k}p = p_* \quad \textit{and} \quad \lim_{k\to\infty} T^{2k+1}p = Tp_*, \tag{2.25}$$

where p_* and Tp_* have only two nonzero components satisfying

$$p_*^{(i_1)} + p_*^{(i_2)} = 1, \quad p_*^{(i)} = 0, \quad i \neq i_1, i_2, \tag{2.26}$$

$$(Tp_*)^{(i_1)} + (Tp_*)^{(i_2)} = 1, \quad (Tp_*)^{(i)} = 0, \quad i \neq i_1, i_2, \tag{2.27}$$

for some $i_1, i_2 \in \{1, \ldots, n\}$. Hence, (2.18), (2.19), (2.20) and (2.21) hold.

Proof. Let $p_0 = T^0p = p$ and $p_k = Tp_{k-1} = T^kp_0$. Obviously, for all $k \geq 0$, p_k satisfies (i) and (iii) of Lemma 2.1. Let $i_{\min} = \min\{i \in \mathcal{N} : p_0^{(i)} > 0\}$ and $i_{\max} = \max\{i \in \mathcal{N} : p_0^{(i)} > 0\}$, where $\mathcal{N} = \{1, \ldots, n\}$. From the definition of γ, we know $\lambda_{i_{\min}} < \gamma(p) < \lambda_{i_{\max}}$. Thus, by the definition of T, we have $p_1^{(i_{\min})} > 0$ and $p_1^{(i_{\max})} > 0$. Then, by induction, for all $k \geq 0$, p_k satisfies (ii) of Lemma 2.1. So, by Lemma 2.1, $\{\Theta(p_k)\}$ is a monotonically increasing sequence. Since $\lambda_1 \leq \gamma(p) \leq \lambda_n$, we have $(\lambda_i - \gamma(p))^2 \leq (\lambda_n - \lambda_1)^2$. Hence, we have from the definition of Θ that $\Theta(p_k) \leq (\lambda_n - \lambda_1)^2$. Thus, $\{\Theta(p_k)\}$ is convergent. Let $\Theta_* = \lim_{k\to\infty}\Theta(p_k) > 0$.

Denote the set of all limit points of $\{p_k\}$ by P_* with cardinality $|P_*|$. Since $\{p_k\}$ is bounded, $|P_*| \geq 1$. For any subsequence $\{p_{k_j}\}$ converging to some $p_* \in P_*$, we have

$$\lim_{j\to\infty}\Theta(p_{k_j}) = \Theta(p_*) \quad \text{and} \quad \lim_{j\to\infty}\Theta(Tp_{k_j}) = \Theta(Tp_*),$$

by the continuity of Θ and T. Notice $p_{k_j+1} = Tp_{k_j}$, we have $\Theta_* = \Theta(p_*) = \Theta(Tp_*)$.

Since p_k satisfies (i)-(iii) of Lemma 2.1 for all $k \geq 0$, p_* must satisfy (i) and (iii). If p_* has only one positive component, we have $\Theta(p_*) = 0$ which contradicts $\Theta(p_*) = \Theta_* > 0$. Hence, by Lemma 2.1, Lemma 2.2 and $\Theta(p_*) = \Theta(Tp_*)$, p_* has only two nonzero components, say $p_*^{(i_1)}$ and $p_*^{(i_2)}$, and their values are uniquely determined by

the indices i_1, i_2 and the eigenvalues λ_{i_1} and λ_{i_2}. This implies $|P_*| < \infty$. Furthermore, by Lemma 2.2, for any $p_* \in P_*$, Tp_* is given by (2.18) and (2.19), and $Tp_* \in P_*$.

We now show that $|P_*| \leq 2$ by way of contradiction. Suppose $|P_*| \geq 3$. For any $p_* \in P_*$ and $Tp_* \in P_*$, denote δ_1 and δ_2 to be the distance from p_* to $P_* \setminus \{p_*\}$ and from Tp_* to $P_* \setminus \{Tp_*\}$, respectively. Since $3 \leq |P_*| < \infty$, we have $\delta_1 > 0$, $\delta_2 > 0$ and there exists an infinite subsequence $\{p_{k_j}\}$ such that

$$p_{k_j} \to p_*, \quad \text{and} \quad p_{k_j+1} = Tp_{k_j} \to Tp_*,$$

but $p_{k_j+2} \notin \mathcal{B}\left(p_*, \frac{1}{2}\delta\right) \cup \mathcal{B}\left(Tp_*, \frac{1}{2}\delta\right)$, where $\delta = \min\{\delta_1, \delta_2\}$ and $\mathcal{B}(p_*, r) = \{p : \|p - p_*\| \leq r\}$. However, by (2.20) we have $T^2 p_* = p_*$. Hence, by continuity of T,

$$\lim_{j\to\infty} p_{k_j+2} = \lim_{j\to\infty} Tp_{k_j+1} = \lim_{j\to\infty} T^2 p_{k_j} = p_*,$$

which contradicts the choice of $p_{k_j+2} \notin \mathcal{B}\left(p_*, \frac{1}{2}\delta\right)$. Thus, $\{p_k\}$ has at most two limit points p_* and Tp_*, and both have only two nonzero components.

Now, we assume that p_* is a limit point of $\{p_{2k}\}$. Since $T^2 p_* = p_*$, all subsequences of $\{p_{2k}\}$ have the same limit point, i.e., $p_{2k} = T^{2k} p \to p_*$. Similarly, we have $T^{2k+1} p \to Tp_*$. Then, (2.26) and (2.27) follow directly from the analysis. ∎

Based on the above analysis, we can show that each gradient method in (1.11) will asymptotically reduces its search in a two-dimensional subspace spanned by the two eigenvectors ξ_1 and ξ_n.

THEOREM 2.4. *Assume that the starting point x_0 has the property that*

$$g_0^\mathsf{T} \xi_1 \neq 0 \quad \textit{and} \quad g_0^\mathsf{T} \xi_n \neq 0. \tag{2.28}$$

Let $\{x_k\}$ be the iterations generated by applying a method in (1.11) to solve problem (1.3). Then

$$\lim_{k\to\infty} \frac{(\mu_{2k}^{(i)})^2}{\sum_{j=1}^n (\mu_{2k}^{(j)})^2} = \begin{cases} \dfrac{1}{1+c^2}, & \textit{if } i = 1, \\ 0, & \textit{if } i = 2, \ldots, n-1, \\ \dfrac{c^2}{1+c^2}, & \textit{if } i = n, \end{cases} \tag{2.29}$$

and

$$\lim_{k\to\infty} \frac{(\mu_{2k+1}^{(i)})^2}{\sum_{j=1}^n (\mu_{2k+1}^{(j)})^2} = \begin{cases} \dfrac{c^2\Psi^2(\lambda_n)}{\Psi^2(\lambda_1) + c^2\Psi^2(\lambda_n)}, & \textit{if } i = 1, \\ 0, & \textit{if } i = 2, \ldots, n-1, \\ \dfrac{\Psi^2(\lambda_1)}{\Psi^2(\lambda_1) + c^2\Psi^2(\lambda_n)}, & \textit{if } i = n, \end{cases} \tag{2.30}$$

where c is a nonzero constant.

Proof. By the assumption (2.28), we know that q_0 satisfies (i)-(iii) of Lemma 2.1. Notice that $q_k = T^k q_0$. Then, by Lemma 2.3, there exists a p_* such that the sequences $\{q_{2k}\}$ and $\{q_{2k+1}\}$ converge to p_* and Tp_*, respectively, which have only two nonzero components satisfying (2.26), (2.27) for some $i_1, i_2 \in \{1, \ldots, n\}$, and (2.20) holds. Hence, if $1 \leq i_1 < i_2 < n$, we have

$$\lim_{k\to\infty} q_{2k}^{(n)} = 0, \qquad \lim_{k\to\infty} \frac{q_{2k}^{(i_2)}}{q_{2k+2}^{(i_2)}} = 1, \tag{2.31}$$

and

$$\lim_{k\to\infty}(\gamma(q_{2k})+\gamma(q_{2k+1}))=\gamma(p_*)+\gamma(Tp_*)=\lambda_{i_1}+\lambda_{i_2}.$$

In addition, since $q_0^{(1)}>0$ and $q_0^{(n)}>0$ by (2.28), we can see from the proof of Lemma 2.3 that $q_k^{(1)}>0$, $q_k^{(n)}>0$ for all $k\geq 0$. Thus, we have

$$\begin{aligned}\lim_{k\to\infty}\frac{q_{2k+2}^{(n)}}{q_{2k}^{(n)}}&=\lim_{k\to\infty}\frac{q_{2k+2}^{(n)}}{q_{2k}^{(n)}}\frac{q_{2k}^{(i_2)}}{q_{2k+2}^{(i_2)}}=\lim_{k\to\infty}\frac{(\lambda_n-\gamma(q_{2k+1}))^2(\lambda_n-\gamma(q_{2k}))^2}{(\lambda_{i_2}-\gamma(q_{2k+1}))^2(\lambda_{i_2}-\gamma(q_{2k}))^2}\\&=\lim_{k\to\infty}\left(\frac{\lambda_n^2-(\gamma(q_{2k})+\gamma(q_{2k+1}))\lambda_n+\gamma(q_{2k})\gamma(q_{2k+1})}{\lambda_{i_2}^2-(\gamma(q_{2k})+\gamma(q_{2k+1}))\lambda_{i_2}+\gamma(q_{2k})\gamma(q_{2k+1})}\right)^2\\&=\left(\frac{\lambda_n^2-(\lambda_{i_1}+\lambda_{i_2})\lambda_n+\tilde{\gamma}}{\lambda_{i_2}^2-(\lambda_{i_1}+\lambda_{i_2})\lambda_{i_2}+\gamma(p_*)\gamma(Tp_*)}\right)^2\\&=\left(1+\frac{(\lambda_n-\lambda_{i_1})(\lambda_n-\lambda_{i_2})}{\lambda_{i_2}^2-(\lambda_{i_1}+\lambda_{i_2})\lambda_{i_2}+\gamma(p_*)\gamma(Tp_*)}\right)^2=:\rho.\end{aligned}\tag{2.32}$$

Since $\lambda_{i_1}<\gamma(p_*)<\lambda_{i_2}$ and $\lambda_{i_1}<\gamma(Tp_*)<\lambda_{i_2}$, we have

$$\begin{aligned}\lambda_{i_2}^2-(\lambda_{i_1}+\lambda_{i_2})\lambda_{i_2}+\gamma(p_*)\gamma(Tp_*)&=\lambda_{i_2}^2-(\gamma(p_*)+\gamma(Tp_*))\lambda_{i_2}+\gamma(p_*)\gamma(Tp_*)\\&=(\lambda_{i_2}-\gamma(p_*))(\lambda_{i_2}-\gamma(Tp_*))>0.\end{aligned}$$

Hence, it follows from (2.32) that $\rho>1$. So, $q_{2k}^{(n)}\to+\infty$, which contradicts (2.31). Then, we must have $i_2=n$. In a similar way, we can show that $i_1=1$. Finally, the equalities in (2.29) and (2.30) follow directly from Lemma 2.2. ∎

In the following, we refer c as the same constant in Theorem 2.4. By Theorem 2.4 we can directly obtain the asymptotic behavior of the stepsize.

COROLLARY 2.5. *Under the conditions of Theorem 2.4, we have*

$$\lim_{k\to\infty}\alpha_{2k}=\frac{\Psi(\lambda_1)+c^2\Psi(\lambda_n)}{\lambda_1(\Psi(\lambda_1)+c^2\kappa\Psi(\lambda_n))}\tag{2.33}$$

and

$$\lim_{k\to\infty}\alpha_{2k+1}=\frac{\Psi(\lambda_1)+c^2\Psi(\lambda_n)}{\lambda_1(\kappa\Psi(\lambda_1)+c^2\Psi(\lambda_n))},\tag{2.34}$$

where α_k is defined in (1.11) and $\kappa=\lambda_n/\lambda_1$ is the condition number of A. Moreover,

$$\lim_{k\to\infty}\left(\frac{1}{\alpha_{2k}}+\frac{1}{\alpha_{2k+1}}\right)=\lambda_1+\lambda_n.\tag{2.35}$$

The next corollary interprets the constant c. A special result for the case $\Psi(A)=I$ (i.e., the SD method) can be found in Lemma 3.4 of [26].

COROLLARY 2.6. *Under the conditions of Theorem 2.4, we have*

$$c=\lim_{k\to\infty}\frac{\mu_{2k}^{(n)}}{\mu_{2k}^{(1)}}=-\frac{\Psi(\lambda_1)}{\Psi(\lambda_n)}\lim_{k\to\infty}\frac{\mu_{2k+1}^{(1)}}{\mu_{2k+1}^{(n)}}.\tag{2.36}$$

Proof. It follows from Theorem 2.4 that

$$\lim_{k\to\infty}\frac{(\mu_{2k}^{(n)})^2}{(\mu_{2k}^{(1)})^2}=\frac{\Psi^2(\lambda_1)}{\Psi^2(\lambda_n)}\lim_{k\to\infty}\frac{(\mu_{2k+1}^{(1)})^2}{(\mu_{2k+1}^{(n)})^2}=c^2. \tag{2.37}$$

Note that $1/\lambda_n<\alpha_k<1/\lambda_1$ by the assumption (2.28). And we have by (2.4) that

$$\mu_{2k+2}^{(1)}=\prod_{\ell=1}^{2}(1-\alpha_{2k+\ell}\lambda_1)\mu_{2k}^{(1)}\quad\text{and}\quad\mu_{2k+2}^{(n)}=\prod_{\ell=1}^{2}(1-\alpha_{2k+\ell}\lambda_n)\mu_{2k}^{(n)}.$$

Thus, the sequence $\left\{\frac{\mu_{2k}^{(n)}}{\mu_{2k}^{(1)}}\right\}$, and similarly for $\left\{\frac{\mu_{2k+1}^{(1)}}{\mu_{2k+1}^{(n)}}\right\}$, do not change its sign. Hence, without loss of generality, we can assume by (2.37) that

$$c=\lim_{k\to\infty}\mu_{2k}^{(n)}/\mu_{2k}^{(1)}. \tag{2.38}$$

Then, by (2.4), (2.33) and (2.38), we have

$$\lim_{k\to\infty}\frac{\mu_{2k+1}^{(1)}}{\mu_{2k+1}^{(n)}}=\lim_{k\to\infty}\frac{\mu_{2k}^{(1)}(1-\alpha_{2k}\lambda_1)}{\mu_{2k}^{(n)}(1-\alpha_{2k}\lambda_n)}=-c\frac{\Psi(\lambda_n)}{\Psi(\lambda_1)},$$

which gives (2.36). ∎

We have the following results on the asymptotic convergence of the function value.

THEOREM 2.7. *Under the conditions of Theorem 2.4, we have*

$$\lim_{k\to\infty}\frac{f(x_{2k+1})-f^*}{f(x_{2k})-f^*}=R_f^1\quad\text{and}\quad\lim_{k\to\infty}\frac{f(x_{2k+2})-f^*}{f(x_{2k+1})-f^*}=R_f^2, \tag{2.39}$$

where

$$R_f^1=\frac{c^2(\kappa-1)^2(\Psi^2(\lambda_1)+c^2\kappa\Psi^2(\lambda_n))}{(\Psi(\lambda_1)+c^2\kappa\Psi(\lambda_n))^2(c^2+\kappa)}, \tag{2.40}$$

$$R_f^2=\frac{c^2(\kappa-1)^2(c^2+\kappa)\Psi^2(\lambda_1)\Psi^2(\lambda_n)}{(c^2\Psi(\lambda_n)+\kappa\Psi(\lambda_1))^2(\Psi^2(\lambda_1)+c^2\kappa\Psi^2(\lambda_n))}. \tag{2.41}$$

In addition, if $\Psi(\lambda_n)=\Psi(\lambda_1)$ *or* $c^2=\Psi(\lambda_1)/\Psi(\lambda_n)$, *then* $R_f^1=R_f^2$.

Proof. Let $\epsilon_k=x_k-x^*$. Since $g_k=A\epsilon_k$, by (2.2), we have

$$\epsilon_k=\sum_{i=1}^{n}\lambda_i^{-1}\mu_k^{(i)}\xi_i.$$

By Theorem 2.4, we only need to consider the case $\mu_k^{(i)}=0$, $i=2,\dots,n-1$, that is,

$$\epsilon_k=\lambda_1^{-1}\mu_k^{(1)}\xi_1+\lambda_n^{-1}\mu_k^{(n)}\xi_n.$$

Thus,

$$f(x_k)-f^*=\frac{1}{2}\epsilon_k^{\mathsf{T}}A\epsilon_k=\frac{1}{2}\frac{\lambda_n(\mu_k^{(1)})^2+\lambda_1(\mu_k^{(n)})^2}{\lambda_1\lambda_n}. \tag{2.42}$$

Since

$$g_k = \mu_k^{(1)}\xi_1 + \mu_k^{(n)}\xi_n \quad \text{and} \quad \alpha_k = \frac{\Psi(\lambda_1)(\mu_k^{(1)})^2 + \Psi(\lambda_n)(\mu_k^{(n)})^2}{\lambda_1\Psi(\lambda_1)(\mu_k^{(1)})^2 + \lambda_n\Psi(\lambda_n)(\mu_k^{(n)})^2},$$

by the definition of ϵ_k and the update rule (1.2), we further have that

$$\begin{aligned}
\epsilon_{k+1} &= \epsilon_k - \alpha_k g_k = (\lambda_1^{-1} - \alpha_k)\mu_k^{(1)}\xi_1 + (\lambda_n^{-1} - \alpha_k)\mu_k^{(n)}\xi_n \\
&= \frac{\Psi(\lambda_n)(\lambda_n - \lambda_1)(\mu_k^{(n)})^2\mu_k^{(1)}}{\lambda_1\left(\lambda_1\Psi(\lambda_1)(\mu_k^{(1)})^2 + \lambda_n\Psi(\lambda_n)(\mu_k^{(n)})^2\right)}\xi_1 \\
&+ \frac{\Psi(\lambda_1)(\lambda_1 - \lambda_n)(\mu_k^{(1)})^2\mu_k^{(n)}}{\lambda_n\left(\lambda_1\Psi(\lambda_1)(\mu_k^{(1)})^2 + \lambda_n\Psi(\lambda_n)(\mu_k^{(n)})^2\right)}\xi_n \\
&= \frac{(\lambda_n - \lambda_1)\left(\lambda_n\Psi(\lambda_n)(\mu_k^{(n)})^2\mu_k^{(1)}\xi_1 - \lambda_1\Psi(\lambda_1)(\mu_k^{(1)})^2\mu_k^{(n)}\xi_n\right)}{\lambda_1\lambda_n\left(\lambda_1\Psi(\lambda_1)(\mu_k^{(1)})^2 + \lambda_n\Psi(\lambda_n)(\mu_k^{(n)})^2\right)}.
\end{aligned}$$

Hence, we obtain

$$\begin{aligned}
f(x_{k+1}) - f^* &= \frac{1}{2}\epsilon_{k+1}^{\mathsf{T}} A\epsilon_{k+1} \\
&= \frac{1}{2}\frac{(\lambda_n - \lambda_1)^2(\mu_k^{(1)})^2(\mu_k^{(n)})^2\left(\lambda_n\Psi^2(\lambda_n)(\mu_k^{(n)})^2 + \lambda_1\Psi^2(\lambda_1)(\mu_k^{(1)})^2\right)}{\lambda_1\lambda_n\left(\lambda_1\Psi(\lambda_1)(\mu_k^{(1)})^2 + \lambda_n\Psi(\lambda_n)(\mu_k^{(n)})^2\right)^2}.
\end{aligned} \tag{2.43}$$

Combining (2.42) with (2.43) yields that

$$\begin{aligned}
\frac{f(x_{k+1}) - f^*}{f(x_k) - f^*} &= \frac{\epsilon_{k+1}^{\mathsf{T}} A\epsilon_{k+1}}{\epsilon_k^{\mathsf{T}} A\epsilon_k} \\
&= \frac{(\mu_k^{(1)})^2(\mu_k^{(n)})^2(\kappa - 1)^2\left(\kappa\Psi^2(\lambda_n)(\mu_k^{(n)})^2 + \Psi^2(\lambda_1)(\mu_k^{(1)})^2\right)}{\left(\Psi(\lambda_1)(\mu_k^{(1)})^2 + \kappa\Psi(\lambda_n)(\mu_k^{(n)})^2\right)^2\left(\kappa(\mu_k^{(1)})^2 + (\mu_k^{(n)})^2\right)},
\end{aligned}$$

which gives (2.39) by substituting the limits of $(\mu_k^{(1)})^2$ and $(\mu_k^{(n)})^2$ in Theorem 2.4.

Notice $\kappa > 1$ by our assumption. So, $R_f^1 = R_f^2$ is equivalent to

$$\frac{\Psi^2(\lambda_1) + c^2\kappa\Psi^2(\lambda_n)}{(\Psi(\lambda_1) + c^2\kappa\Psi(\lambda_n))^2(c^2 + \kappa)} = \frac{(c^2 + \kappa)\Psi^2(\lambda_1)\Psi^2(\lambda_n)}{(c^2\Psi(\lambda_n) + \kappa\Psi(\lambda_1))^2(\Psi^2(\lambda_1) + c^2\kappa\Psi^2(\lambda_n))},$$

which by rearranging terms gives

$$c^4\Psi^2(\lambda_n)(\Psi(\lambda_n) - \Psi(\lambda_1)) = \Psi^2(\lambda_1)(\Psi(\lambda_n) - \Psi(\lambda_1)).$$

Hence, $R_f^1 = R_f^2$ holds if $\Psi(\lambda_n) = \Psi(\lambda_1)$ or $c^2 = \Psi(\lambda_1)/\Psi(\lambda_n)$. ∎

Remark 2.8. Theorem 2.7 indicates that, when $\Psi(A) = I$ (i.e., the SD method), the two sequences $\left\{\frac{\Delta_{2k+1}}{\Delta_{2k}}\right\}$ and $\left\{\frac{\Delta_{2k+2}}{\Delta_{2k+1}}\right\}$ converge at the same speed, where $\Delta_k = f(x_k) - f^*$. Otherwise, the two sequences may converge at different rates.

To illustrate the results in Theorem 2.7, we apply gradient method (1.11) with $\Psi(A) = A$ (i.e., the MG method) to an instance of (1.3), where the vector of all ones was used as the initial point, the matrix A is diagonal with

$$A_{ii} = i\sqrt{i}, \quad i = 1, \ldots, n, \tag{2.44}$$

and $b = 0$. Figure 1 clearly shows the difference between R_f^1 and R_f^2.

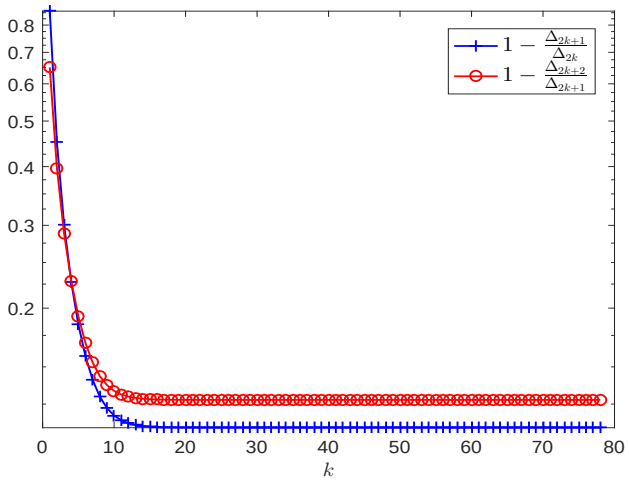

FIG. 1. *Problem (2.44) with $n = 10$: convergence history of the sequences $\{1 - \frac{\Delta_{2k+1}}{\Delta_{2k}}\}$ and $\{1 - \frac{\Delta_{2k+2}}{\Delta_{2k+1}}\}$ generated by gradient method (1.11) with $\Psi(A) = A$ (i.e., the MG method).*

The next theorem shows the asymptotic convergence of the gradient norm.

THEOREM 2.9. *Under the conditions of Theorem 2.4, the following limits hold,*

$$\lim_{k\to\infty} \frac{\|g_{2k+1}\|^2}{\|g_{2k}\|^2} = R_g^1 \quad \text{and} \quad \lim_{k\to\infty} \frac{\|g_{2k+2}\|^2}{\|g_{2k+1}\|^2} = R_g^2, \tag{2.45}$$

where

$$R_g^1 = \frac{c^2(\kappa-1)^2(\Psi^2(\lambda_1) + c^2\Psi^2(\lambda_n))}{(1+c^2)(\Psi(\lambda_1) + c^2\kappa\Psi(\lambda_n))^2}, \tag{2.46}$$

$$R_g^2 = \frac{c^2(1+c^2)(\kappa-1)^2\Psi^2(\lambda_1)\Psi^2(\lambda_n)}{(c^2\Psi(\lambda_n) + \kappa\Psi(\lambda_1))^2(\Psi^2(\lambda_1) + c^2\Psi^2(\lambda_n))}. \tag{2.47}$$

In addition, if $\Psi(\lambda_n) = \kappa\Psi(\lambda_1)$ or $c^2 = \Psi(\lambda_1)/\Psi(\lambda_n)$, then $R_g^1 = R_g^2$.

Proof. Using the same arguments as in Theorem 2.7, we have

$$\|g_k\|^2 = (\mu_k^{(1)})^2 + (\mu_k^{(n)})^2$$

and

$$\|g_{k+1}\|^2 = \epsilon_{k+1}^{\mathsf{T}} A^2 \epsilon_{k+1} = \frac{(\lambda_n - \lambda_1)^2(\mu_k^{(1)})^2(\mu_k^{(n)})^2\left(\Psi^2(\lambda_n)(\mu_k^{(n)})^2 + \Psi^2(\lambda_1)(\mu_k^{(1)})^2\right)}{\left(\lambda_1\Psi(\lambda_1)(\mu_k^{(1)})^2 + \lambda_n\Psi(\lambda_n)(\mu_k^{(n)})^2\right)^2},$$

which give that

$$\frac{\|g_{k+1}\|^2}{\|g_k\|^2} = \frac{(\kappa-1)^2(\mu_k^{(1)})^2(\mu_k^{(n)})^2\left(\Psi^2(\lambda_n)(\mu_k^{(n)})^2+\Psi^2(\lambda_1)(\mu_k^{(1)})^2\right)}{\left(\Psi(\lambda_1)(\mu_k^{(1)})^2+\kappa\Psi(\lambda_n)(\mu_k^{(n)})^2\right)^2\left((\mu_k^{(1)})^2+(\mu_k^{(n)})^2\right)}.$$

Thus, (2.45) follows by substituting the limits of $(\mu_k^{(1)})^2$ and $(\mu_k^{(n)})^2$ in Theorem 2.4.

Notice $\kappa > 1$ by our assumption. So, $R_g^1 = R_g^2$ is equivalent to

$$\frac{\Psi^2(\lambda_1)+c^2\Psi^2(\lambda_n)}{(1+c^2)(\Psi(\lambda_1)+c^2\kappa\Psi(\lambda_n))^2} = \frac{(1+c^2)\Psi^2(\lambda_1)\Psi^2(\lambda_n)}{(c^2\Psi(\lambda_n)+\kappa\Psi(\lambda_1))^2(\Psi^2(\lambda_1)+c^2\Psi^2(\lambda_n))},$$

which by rearranging terms gives

$$c^4\Psi^2(\lambda_n)(\kappa\Psi(\lambda_1)-\Psi(\lambda_n)) = \Psi^2(\lambda_1)(\kappa\Psi(\lambda_1)-\Psi(\lambda_n)).$$

Hence, $R_g^1 = R_g^2$ holds if $\Psi(\lambda_n) = \kappa\Psi(\lambda_1)$ or $c^2 = \Psi(\lambda_1)/\Psi(\lambda_n)$. ∎

Remark 2.10. Theorem 2.9 indicates that the two sequences $\left\{\frac{\|g_{2k+1}\|^2}{\|g_{2k}\|^2}\right\}$ and $\left\{\frac{\|g_{2k+2}\|^2}{\|g_{2k+1}\|^2}\right\}$ generated by the MG method (i.e., $\Psi(A) = A$) converge at the same rate. Otherwise, the two sequences may converge at different rates.

By Theorems 2.7 and 2.9, we can obtain the following corollary.

COROLLARY 2.11. *Under the conditions of Theorem 2.4, we have*

$$\lim_{k\to\infty}\frac{f(x_{2k+3})-f^*}{f(x_{2k+1})-f^*} = \lim_{k\to\infty}\frac{f(x_{2k+2})-f^*}{f(x_{2k})-f^*} = R_f^1R_f^2, \tag{2.48}$$

$$\lim_{k\to\infty}\frac{\|g_{2k+3}\|^2}{\|g_{2k+1}\|^2} = \lim_{k\to\infty}\frac{\|g_{2k+2}\|^2}{\|g_{2k}\|^2} = R_g^1R_g^2. \tag{2.49}$$

In addition,

$$R_f^1R_f^2 = R_g^1R_g^2 = \frac{c^4(\kappa-1)^4\Psi^2(\lambda_1)\Psi^2(\lambda_n)}{(\Psi(\lambda_1)+c^2\kappa\Psi(\lambda_n))^2(c^2\Psi(\lambda_n)+\kappa\Psi(\lambda_1))^2}. \tag{2.50}$$

Remark 2.12. Corollary 2.11 shows that the odd and even subsequences of objective values and gradient norms converge at the same rate. Moreover, we have

$$R_f^1R_f^2 = R_g^1R_g^2 = \frac{(\kappa-1)^4}{(1+\kappa/t+t\kappa+\kappa^2)^2} \le \left(\frac{\kappa-1}{\kappa+1}\right)^4, \tag{2.51}$$

where $t = c^2\Psi(\lambda_n)/\Psi(\lambda_1)$. Notice that the right side of (2.51) only depends on κ, which implies these odd and even subsequences generated by all the gradient methods (1.11) will have the same worst asymptotic rate independent of Ψ.

Now, as in [26], we define the *minimum deviation*

$$\sigma = \min_{i\in\mathcal{I}}\left|\frac{2\lambda_i-(\lambda_1+\lambda_n)}{\lambda_n-\lambda_1}\right|, \tag{2.52}$$

where

$$\mathcal{I} = \{i : \lambda_1 < \lambda_i < \lambda_n,\ g_0^{\mathsf{T}}\xi_i \neq 0,\ \text{and } \lambda_i \neq \alpha_k \text{ for all } k\}.$$

Clearly, $\sigma \in (0,1)$. We now close this section by deriving a bound on the constant c defined in Theorem 2.4. The following theorem generalizes the results in [1, 26], where only the case $\Psi(A) = I$ (i.e., the SD method) is considered.

THEOREM 2.13. *Under the conditions of Theorem 2.4, and assuming that $\mathcal{I}$ is nonempty, we have*

$$\frac{\Psi(\lambda_1)}{\Psi(\lambda_n)}\frac{1}{\phi_\sigma} \le c^2 \le \frac{\Psi(\lambda_1)}{\Psi(\lambda_n)}\phi_\sigma, \tag{2.53}$$

where

$$\phi_\sigma = \frac{2+\eta_\sigma+\sqrt{\eta_\sigma^2+4\eta_\sigma}}{2} \quad \text{and} \quad \eta_\sigma = 4\left(\frac{1+\sigma^2}{1-\sigma^2}\right). \tag{2.54}$$

Proof. Let $p = q_0$. By the definition of T, we have that

$$\frac{(T^{k+2}p)^{(i)}}{(T^{k+2}p)^{(1)}} = \frac{(T^kp)^{(i)}}{(T^kp)^{(1)}}\frac{(\lambda_i-\gamma(T^kp))^2(\lambda_i-\gamma(T^{k+1}p))^2}{(\lambda_1-\gamma(T^kp))^2(\lambda_1-\gamma(T^{k+1}p))^2}. \tag{2.55}$$

It follows from Theorem 2.4 and Lemma 2.3 that

$$\frac{(T^kp)^{(i)}}{(T^kp)^{(1)}} \to 0,\ i = 2,\dots,n-1. \tag{2.56}$$

By the continuity of T and (2.25) in Lemma 2.3, we always have that

$$\frac{(\lambda_i-\gamma(T^kp))^2(\lambda_i-\gamma(T^{k+1}p))^2}{(\lambda_1-\gamma(T^kp))^2(\lambda_1-\gamma(T^{k+1}p))^2} \to \frac{(\lambda_i-\gamma(p_*))^2(\lambda_i-\gamma(Tp_*))^2}{(\lambda_1-\gamma(p_*))^2(\lambda_1-\gamma(Tp_*))^2},$$

which together with (2.55) and (2.56) implies that

$$\frac{(\lambda_i-\gamma(p_*))^2(\lambda_i-\gamma(Tp_*))^2}{(\lambda_1-\gamma(p_*))^2(\lambda_1-\gamma(Tp_*))^2} \le 1,\ i = 2,\dots,n-1, \tag{2.57}$$

where p_* is the same vector as in Lemma 2.3. Clearly, (2.57) also holds for $i = 1$. As for $i = n$, it follows from (2.21) in Lemma 2.2 and Theorem 2.4 that

$$\gamma(p_*)+\gamma(Tp_*) = \lambda_1+\lambda_n, \tag{2.58}$$

which yields that

$$\frac{(\lambda_n-\gamma(p_*))^2(\lambda_n-\gamma(Tp_*))^2}{(\lambda_1-\gamma(p_*))^2(\lambda_1-\gamma(Tp_*))^2} = 1.$$

Thus, (2.57) holds for $i = 1,\dots,n$. Hence, we have

$$\begin{aligned}&(\lambda_i-\delta-(\gamma(p_*)-\delta))^2(\lambda_i-\delta-(\gamma(Tp_*)-\delta))^2\\ &\le (\lambda_1-\delta-(\gamma(p_*)-\delta))^2(\lambda_1-\delta-(\gamma(Tp_*)-\delta))^2,\end{aligned} \tag{2.59}$$

where $\delta = \frac{\lambda_1+\lambda_n}{2}$. By (2.58) and (2.59), we obtain

$$\begin{aligned}&(\lambda_i-\delta-(\gamma(p_*)-\delta))^2(\lambda_i-\delta+(\gamma(p_*)-\delta))^2\\ &\le \left(\frac{\lambda_1-\lambda_n}{2}-(\gamma(p_*)-\delta)\right)^2\left(\frac{\lambda_1-\lambda_n}{2}+(\gamma(p_*)-\delta)\right)^2,\end{aligned}$$

which implies that

$$\left(\frac{\lambda_1-\lambda_n}{2}\right)^2+(\lambda_i-\delta)^2\geq 2\left(\gamma(p_*)-\delta\right)^2. \tag{2.60}$$

By Lemma 2.2 and Theorem 2.4, we have that

$$\gamma(p_*)=\frac{\lambda_1\Psi(\lambda_1)p_*^{(1)}+\lambda_n\Psi(\lambda_n)p_*^{(n)}}{\Psi(\lambda_1)p_*^{(1)}+\Psi(\lambda_n)p_*^{(n)}}.$$

Substituting $\gamma(p_*)$ into (2.60), we obtain

$$\left(\frac{\lambda_1-\lambda_n}{2}\right)^2+(\lambda_i-\delta)^2\geq\frac{(\lambda_n-\lambda_1)^2(\Psi(\lambda_n)c^2-\Psi(\lambda_1))^2}{2(\Psi(\lambda_n)c^2+\Psi(\lambda_1))^2},$$

which gives

$$4\left(\frac{1+\sigma_i^2}{1-\sigma_i^2}\right)\geq\frac{(c^2\Psi(\lambda_n)-\Psi(\lambda_1))^2}{c^2\Psi(\lambda_1)\Psi(\lambda_n)},\quad\text{where}\quad\sigma_i=\frac{2\lambda_i-(\lambda_1+\lambda_n)}{\lambda_n-\lambda_1}. \tag{2.61}$$

Noting that (2.61) holds for all $i\in\mathcal{I}$. Thus, we have

$$\frac{(c^2\Psi(\lambda_n)-\Psi(\lambda_1))^2}{c^2\Psi(\lambda_1)\Psi(\lambda_n)}\leq\eta_\sigma, \tag{2.62}$$

which implies (2.53). This completes the proof. ∎

3. Techniques for breaking the zigzagging pattern. As shown in the previous section, all the gradient methods (1.11) asymptotically conduct its searches in the two-dimensional subspace spanned by ξ_1 and ξ_n. By (2.4), if either $\mu_k^{(1)}$ or $\mu_k^{(n)}$ equals to zero, the corresponding component will vanish at all subsequent iterations. Hence, in order to break the undesired zigzagging pattern, a good strategy is to employ some stepsize approximating $1/\lambda_1$ or $1/\lambda_n$. In this section, we will derive a new stepsize converging to $1/\lambda_n$ and propose a periodic gradient method using this new stepsize.

3.1. A new stepsize. Our new stepsize will be derived by imposing finite termination on minimizing two-dimensional strictly convex quadratic function, see [30] for the case of $\Psi(A)=I$ (i.e., the SD method). We mention that the key property used by Yuan [30] is that two consecutive gradients generated by the SD method are perpendicular to each other, which may not be true for all the gradient methods (1.11). However, we have by the stepsize definition (1.11) that

$$g_k^{\mathsf{T}}\Psi(A)g_{k+1}=g_k^{\mathsf{T}}\Psi(A)g_k-\alpha_k g_k^{\mathsf{T}}\Psi(A)Ag_k=0. \tag{3.1}$$

Consider the two-dimensional case. Suppose we want to find the minimizer of (1.3) with $n=2$ after the following 3 iterations:

$$x_1=x_0-\alpha_0g_0,\quad x_2=x_1-\alpha_1g_1,\quad x_3=x_2-\alpha_2g_2,$$

where $g_i\neq 0$, $i=0,1,2$, α_0 and α_2 are stepsizes given by (1.11), and α_1 is to be derived by ensuring x_3 is the solution.

By (3.1), we have $g_0^{\mathsf{T}}\Psi(A)g_1=0$. Hence, all vectors x_k can be expressed by the linear combination of $\frac{\Psi^r(A)g_0}{\|\Psi^r(A)g_0\|}$ and $\frac{\Psi^{1-r}(A)g_1}{\|\Psi^{1-r}(A)g_1\|}$ for any given $r\in\mathbb{R}$. Now, consider

$$\varphi(t,l):=f\left(x_1+t\frac{\Psi^r(A)g_0}{\|\Psi^r(A)g_0\|}+l\frac{\Psi^{1-r}(A)g_1}{\|\Psi^{1-r}(A)g_1\|}\right) \tag{3.2}$$

$$= f(x_1) + G^{\mathsf{T}} \begin{pmatrix} t \\ l \end{pmatrix} + \frac{1}{2} \begin{pmatrix} t \\ l \end{pmatrix}^{\mathsf{T}} H \begin{pmatrix} t \\ l \end{pmatrix},$$

where

$$G = Bg_1 = \begin{pmatrix} \frac{g_1^{\mathsf{T}} \Psi^r(A) g_0}{\|\Psi^r(A) g_0\|} \\ \frac{g_1^{\mathsf{T}} \Psi^{1-r}(A) g_1}{\|\Psi^{1-r}(A) g_1\|} \end{pmatrix} \text{ with } B = \left(\frac{\Psi^r(A) g_0}{\|\Psi^r(A) g_0\|}, \frac{\Psi^{1-r}(A) g_1}{\|\Psi^{1-r}(A) g_1\|} \right)^{\mathsf{T}} \tag{3.3}$$

and

$$H = BAB^{\mathsf{T}} = \begin{pmatrix} \frac{g_0^{\mathsf{T}} \Psi^{2r}(A) A g_0}{\|\Psi^r(A) g_0\|^2} & \frac{g_0^{\mathsf{T}} \Psi(A) A g_1}{\|\Psi^r(A) g_0\| \|\Psi^{1-r}(A) g_1\|} \\ \frac{g_0^{\mathsf{T}} \Psi(A) A g_1}{\|\Psi^r(A) g_0\| \|\Psi^{1-r}(A) g_1\|} & \frac{g_1^{\mathsf{T}} \Psi^{2(1-r)}(A) A g_1}{\|\Psi^{1-r}(A) g_1\|^2} \end{pmatrix}. \tag{3.4}$$

Note that $B^{\mathsf{T}} B = BB^{\mathsf{T}} = I$ since $n = 2$. The minimizer (t^*, l^*) of φ satisfy

$$G + H \begin{pmatrix} t^* \\ l^* \end{pmatrix} = 0, \quad \Longrightarrow \quad \begin{pmatrix} t^* \\ l^* \end{pmatrix} = -H^{-1} G.$$

Suppose x_3 is the solution, that is

$$x_3 = x_1 + t^* \frac{\Psi^r(A) g_0}{\|\Psi^r(A) g_0\|} + l^* \frac{\Psi^{1-r}(A) g_1}{\|\Psi^{1-r}(A) g_1\|}.$$

Then, since $x_3 = x_2 - \alpha_2 g_2$, we have $x_3 - x_2$ is parallel to g_2, i.e.,

$$B^{\mathsf{T}} \begin{pmatrix} t^* \\ l^* \end{pmatrix} + \alpha_1 g_1 \quad \text{ is parallel to } \quad g_2, \tag{3.5}$$

which is equivalent to

$$\begin{pmatrix} t^* \\ l^* \end{pmatrix} - (-\alpha_1 G) = -(H^{-1} G - \alpha_1 G) \quad \text{and} \quad G + H(-\alpha_1 G) \tag{3.6}$$

are parallel. Denote the components of G by G_i, and the components of H by H_{ij}, $i, j = 1, 2$. By (3.6), we would have

$$\begin{pmatrix} H_{22} G_1 - H_{12} G_2 - \alpha_1 \Delta G_1 \\ H_{11} G_2 - H_{12} G_1 - \alpha_1 \Delta G_2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} G_1 - \alpha_1 (H_{11} G_1 + H_{12} G_2) \\ G_2 - \alpha_1 (H_{12} G_1 + H_{22} G_2) \end{pmatrix}.$$

are parallel, where $\Delta = \det(H) = \det(A) > 0$. It follows that

$$\begin{aligned} &(H_{22} G_1 - H_{12} G_2 - \alpha_1 \Delta G_1)[G_2 - \alpha_1 (H_{12} G_1 + H_{22} G_2)] \\ &= (H_{11} G_2 - H_{12} G_1 - \alpha_1 \Delta G_2)[G_1 - \alpha_1 (H_{11} G_1 + H_{12} G_2)], \end{aligned}$$

which gives

$$\alpha_1^2 \Delta \Gamma - \alpha_1 (H_{11} + H_{22}) \Gamma + \Gamma = 0, \tag{3.7}$$

where

$$\Gamma = (H_{12} G_1 + H_{22} G_2) G_1 - (H_{11} G_1 + H_{12} G_2) G_2.$$

On the other hand, if (3.7) holds, we have (3.5) holds, which by (3.3), $H^{-1} = BA^{-1}B^{\mathsf{T}}$ and $B^{\mathsf{T}} B = I$ implies that

$$-B^{\mathsf{T}} H^{-1} G + \alpha_1 g_1 = -A^{-1} g_1 + \alpha_1 g_1 = -A^{-1}(g_1 - \alpha_1 A g_1) = -A^{-1} g_2$$

is parallel to g_2. Hence, g_2 is an eigenvector of A, i.e. $Ag_2 = \lambda g_2$ for some $\lambda > 0$, since $g_2 \neq 0$. So, by (1.11), $\alpha_2 = \Psi(\lambda)g_2^{\mathsf{T}}g_2/(\lambda\Psi(\lambda)g_2^{\mathsf{T}}g_2) = 1/\lambda$. Therefore, $g_3 = g_2 - \alpha_2 Ag_2 = g_2 - \alpha_2\lambda g_2 = 0$, which implies x_3 is the solution. So, (3.7) guarantees x_3 is the minimizer.

Hence, to ensure x_3 is the minimizer, by (3.7), we only need to choose α_1 satisfying

$$\alpha_1^2\Delta - \alpha_1(H_{11} + H_{22}) + 1 = 0, \tag{3.8}$$

whose two positive roots are

$$\frac{(H_{11} + H_{22}) \pm \sqrt{(H_{11} + H_{22})^2 - 4\Delta}}{2\Delta}.$$

These two roots are $1/\lambda_1$ and $1/\lambda_2$, where $0 < \lambda_1 < \lambda_2$ are two eigenvalues of A (Note that A and H have same eigenvalues). For numerical reasons (see next subsection), we would like to choose α_1 to be the smaller one $1/\lambda_2$, which can be calculated as

$$\begin{aligned} \alpha_1 &= \frac{2}{(H_{11} + H_{22}) + \sqrt{(H_{11} + H_{22})^2 - 4\Delta}} \\ &= \frac{2}{(H_{11} + H_{22}) + \sqrt{(H_{11} - H_{22})^2 + 4H_{12}^2}}. \end{aligned} \tag{3.9}$$

To check this finite termination property, we applied the above described method with α_1 given by (3.9), and $\Psi(A) = A$ in (1.11), (i.e., α_0 and α_2 use the MG stepsize) to minimize two-dimensional quadratic function (1.3) with

$$A = \mathrm{diag}\{1, \lambda\} \quad \text{and} \quad b = 0. \tag{3.10}$$

We run the algorithm for 3 iterations using ten random starting points and the averaged values of $\|g_3\|$ and $f(x_3)$ are presented in Table 1. We can observe that for different values of λ, the $\|g_3\|$ and $f(x_3)$ obtained by the method in three iterations are numerically very close to zero. This coincides with our analysis.

TABLE 1
Averaged results for problem (3.10) with different condition numbers.

λ	$\|g_3\|$	$f(x_3)$
10	4.8789e-18	8.0933e-36
100	4.1994e-18	2.2854e-37
1000	1.2001e-18	2.8083e-39
10000	1.0621e-18	5.3460e-40

3.2. Spectral property of the new stepsize. Notice that $g_1 = g_0 - \alpha_0 Ag_0$ and $g_0^{\mathsf{T}}\Psi(A)g_1 = 0$. So, we have

$$g_0^{\mathsf{T}}\Psi(A)Ag_1 = -(g_1^{\mathsf{T}}\Psi(A)g_1)/\alpha_0.$$

Hence, the matrix H given in (3.4) can be also written as

$$H = \begin{pmatrix} \frac{g_0^{\mathsf{T}}\Psi^{2r}(A)Ag_0}{\|\Psi^r(A)g_0\|^2} & -\frac{g_1^{\mathsf{T}}\Psi(A)g_1}{\alpha_0\|\Psi^r(A)g_0\|\|\Psi^{1-r}(A)g_1\|} \\ -\frac{g_1^{\mathsf{T}}\Psi(A)g_1}{\alpha_0\|\Psi^r(A)g_0\|\|\Psi^{1-r}(A)g_1\|} & \frac{g_1^{\mathsf{T}}\Psi^{2(1-r)}(A)Ag_1}{\|\Psi^{1-r}(A)g_1\|^2} \end{pmatrix}. \tag{3.11}$$

So, for general case, we could propose our new stepsize at the k-th iteration as

$$\tilde{\alpha}_k = \frac{2}{(H_{11}^k + H_{22}^k) + \sqrt{(H_{11}^k - H_{22}^k)^2 + 4(H_{12}^k)^2}}, \tag{3.12}$$

where H_{ij}^k is the component of H^k:

$$H^k = \begin{pmatrix} \frac{g_{k-1}^{\mathsf{T}}\Psi^{2r}(A)Ag_{k-1}}{\|\Psi^r(A)g_{k-1}\|^2} & -\frac{g_k^{\mathsf{T}}\Psi(A)g_k}{\alpha_{k-1}\|\Psi^r(A)g_{k-1}\|\|\Psi^{1-r}(A)g_k\|} \\ -\frac{g_k^{\mathsf{T}}\Psi(A)g_k}{\alpha_{k-1}\|\Psi^r(A)g_{k-1}\|\|\Psi^{1-r}(A)g_k\|} & \frac{g_k^{\mathsf{T}}\Psi^{2(1-r)}(A)Ag_k}{\|\Psi^{1-r}(A)g_k\|^2} \end{pmatrix} \tag{3.13}$$

and α_{k-1} is given by (1.11). Clearly, α_k^Y in (1.8) can be obtained by by setting $\Psi(A) = I$ in (3.13). In addition, by (3.12) we have that

$$\frac{1}{H_{11}^k + H_{22}^k} \le \tilde{\alpha}_k \le \frac{1}{\max\{H_{11}^k, H_{22}^k\}}. \tag{3.14}$$

The next theorem shows that the stepsize $\tilde{\alpha}_k$ enjoys desirable spectral property.

THEOREM 3.1. *Suppose that the conditions of Theorem 2.4 hold. Let $\{x_k\}$ be the iterations generated by any gradient method in (1.11) to solve problem (1.3). Then*

$$\lim_{k\to\infty} \tilde{\alpha}_k = \frac{1}{\lambda_n}. \tag{3.15}$$

Proof. It follows from (2.29) and (2.30) of Theorem 2.4 that

$$\begin{aligned} \lim_{k\to\infty} H_{11}^k &= \lim_{k\to\infty} \frac{g_{k-1}^{\mathsf{T}}\Psi^{2r}(A)Ag_{k-1}}{\|g_{k-1}\|^2} \frac{\|g_{k-1}\|^2}{\|\Psi^r(A)g_{k-1}\|^2} \\ &= \frac{\lambda_1(c^2\Psi^{2r}(\lambda_1)\Psi^2(\lambda_n) + \kappa\Psi^{2r}(\lambda_n)\Psi^2(\lambda_1))}{c^2\Psi^{2r}(\lambda_1)\Psi^2(\lambda_n) + \Psi^{2r}(\lambda_n)\Psi^2(\lambda_1)} \end{aligned}$$

and

$$\begin{aligned} \lim_{k\to\infty} H_{22}^k &= \frac{g_k^{\mathsf{T}}\Psi^{2(1-r)}(A)Ag_k}{\|g_k\|^2} \frac{\|g_k\|^2}{\|\Psi^{1-r}(A)g_k\|^2} \\ &= \frac{\lambda_1(\Psi^{2(1-r)}(\lambda_1) + \kappa c^2\Psi^{2(1-r)}(\lambda_n))}{\Psi^{2(1-r)}(\lambda_1) + c^2\Psi^{2(1-r)}(\lambda_n)} \\ &= \frac{\lambda_1(\Psi^2(\lambda_1)\Psi^{2r}(\lambda_n) + \kappa c^2\Psi^2(\lambda_n)\Psi^{2r}(\lambda_1))}{\Psi^2(\lambda_1)\Psi^{2r}(\lambda_n) + c^2\Psi^2(\lambda_n)\Psi^{2r}(\lambda_1)}, \end{aligned}$$

which give

$$\lim_{k\to\infty} (H_{11}^k + H_{22}^k) = \lambda_1(\kappa + 1) \tag{3.16}$$

and

$$\lim_{k\to\infty} (H_{11}^k - H_{22}^k) = \frac{\lambda_1(\kappa - 1)(\Psi^2(\lambda_1)\Psi^{2r}(\lambda_n) - c^2\Psi^2(\lambda_n)\Psi^{2r}(\lambda_1))}{\Psi^2(\lambda_1)\Psi^{2r}(\lambda_n) + c^2\Psi^2(\lambda_n)\Psi^{2r}(\lambda_1)}. \tag{3.17}$$

Then, by the definition of α_k, we have

$$g_k^{\mathsf{T}}\Psi(A)g_k = -\alpha_{k-1}g_{k-1}^{\mathsf{T}}\Psi(A)Ag_{k-1} + \alpha_{k-1}^2 g_{k-1}^{\mathsf{T}}\Psi(A)A^2g_{k-1},$$

which together with (2.30) in Theorem 2.4 and (2.34) in Corollary 2.5 yields that

$$\begin{aligned}
&\lim_{k\to\infty}(H_{12}^k)^2\\
=&\lim_{k\to\infty}\frac{g_k^{\mathsf{T}}\Psi(A)g_k}{\alpha_{k-1}^2\|\Psi^r(A)g_{k-1}\|^2}\frac{g_k^{\mathsf{T}}\Psi(A)g_k}{\|\Psi^{1-r}(A)g_k\|^2}\\
=&\lim_{k\to\infty}\left(-\frac{1}{\alpha_{k-1}}\frac{g_{k-1}^{\mathsf{T}}\Psi(A)Ag_{k-1}}{\|\Psi^r(A)g_{k-1}\|^2}+\frac{g_{k-1}^{\mathsf{T}}\Psi(A)A^2g_{k-1}}{\|\Psi^r(A)g_{k-1}\|^2}\right)\frac{g_k^{\mathsf{T}}\Psi(A)g_k}{\|\Psi^{1-r}(A)g_k\|^2}\\
=&\left[-\frac{\lambda_1(\kappa\Psi(\lambda_1)+c^2\Psi(\lambda_n))}{\Psi(\lambda_1)+c^2\Psi(\lambda_n)}\frac{\lambda_1(c^2\Psi(\lambda_1)\Psi^2(\lambda_n)+\kappa\Psi(\lambda_n)\Psi^2(\lambda_1))}{c^2\Psi^{2r}(\lambda_1)\Psi^2(\lambda_n)+\Psi^{2r}(\lambda_n)\Psi^2(\lambda_1)}+\right.\\
&\left.\frac{\lambda_1^2(c^2\Psi(\lambda_1)\Psi^2(\lambda_n)+\kappa^2\Psi(\lambda_n)\Psi^2(\lambda_1))}{c^2\Psi^{2r}(\lambda_1)\Psi^2(\lambda_n)+\Psi^{2r}(\lambda_n)\Psi^2(\lambda_1)}\right]\frac{(\Psi(\lambda_1)+c^2\Psi(\lambda_n))\Psi^{2r}(\lambda_1)\Psi^{2r}(\lambda_n)}{\Psi^2(\lambda_1)\Psi^{2r}(\lambda_n)+c^2\Psi^2(\lambda_n)\Psi^{2r}(\lambda_1)}\\
=&\frac{\lambda_1^2c^2(\kappa-1)^2\Psi^{2+2v}(\lambda_1)\Psi^{2+2v}(\lambda_n)}{(\Psi^2(\lambda_1)\Psi^{2r}(\lambda_n)+c^2\Psi^2(\lambda_n)\Psi^{2r}(\lambda_1))^2}.
\end{aligned}$$

Then, from the above equality and (3.17), we obtain that

$$\lim_{k\to\infty}\sqrt{(H_{11}^k-H_{22}^k)^2+4(H_{12}^k)^2}=\lambda_1(\kappa-1). \tag{3.18}$$

Combining (3.16) and (3.18), we have that

$$\lim_{k\to\infty}\tilde{\alpha}_k=\frac{2}{\lambda_1(\kappa+1)+\lambda_1(\kappa-1)}=\frac{1}{\lambda_n}.$$

This completes the proof. ∎

Remark 3.2. When $r=1$, we have from (3.14) that $\tilde{\alpha}_k\le 1/H_{22}^k=\alpha_k^{SD}$. Hence, using this stepsize $\tilde{\alpha}_k$ will give a monotone gradient method. Theorem 3.1 indicates that the general $\tilde{\alpha}_k$ will have the asymptotic spectral property (3.15), and hence will be asymptotically be smaller than α_k^{SD} independent of r. But a proper choice r will facilitate the calculation of $\tilde{\alpha}_k$. This will be more clear in the next section.

Using the similar arguments, we can also show the larger stepsize derived in subsection 3.1 converges to $1/\lambda_1$.

THEOREM 3.3. *Let*

$$\bar{\alpha}_k=\frac{2}{(H_{11}^k+H_{22}^k)-\sqrt{(H_{11}^k-H_{22}^k)^2+4(H_{12}^k)^2}}.$$

Under the conditions of Theorem 3.1, we have

$$\lim_{k\to\infty}\bar{\alpha}_k=\frac{1}{\lambda_1}.$$

To present an intuitive illustration of the asymptotic behaviors of $\tilde{\alpha}_k$ and $\bar{\alpha}_k$, we applied the gradient method (1.11) with $\Psi(A)=A$ (i.e., the MG method) to minimize the quadratic function (1.3) with

$$A=\mathrm{diag}\{a_1,a_2,\ldots,a_n\}\quad\text{and}\quad b=0, \tag{3.19}$$

where $a_1=1$, $a_n=n$ and a_i is randomly generated between 1 and n for $i=2,\ldots,n-1$. From Figure 2, we can see that $\tilde{\alpha}_k$ approximates $1/\lambda_n$ with satisfactory accuracy in a few iterations. However, $\bar{\alpha}_k$ converges to $1/\lambda_1$ even slower than the decreasing of gradient norm. This, to some extent, explains the reason why we prefer $\tilde{\alpha}_k$ to the short stepsize.

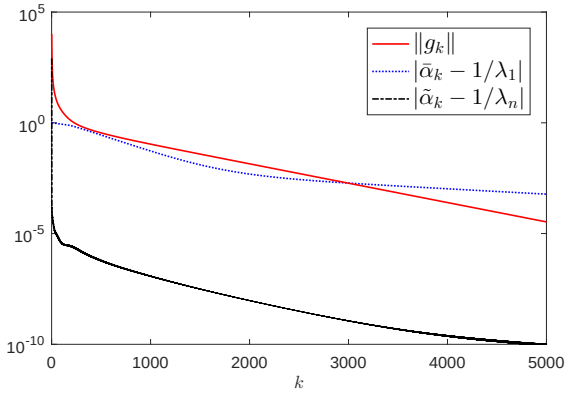

FIG. 2. *Problem* (3.19) *with* $n = 1,000$*: convergence history of the sequences* $\{\tilde{\alpha}_k\}$ *and* $\{\bar{\alpha}_k\}$ *for the first 5,000 iterations of the gradient method* (1.11) *with* $\Psi(A) = A$ *(i.e., the MG method).*

3.3. A periodic gradient method. A method alternately using α_k in (1.11) and $\tilde{\alpha}_k$ to minimize a 2-dimensional quadratic function will monotonically decrease the objective value, and terminates in 3 iterations. However, for minimizing a general n-dimensional quadratic function, this alternating scheme may not be efficient for the purpose of vanishing the component $\mu_k^{(n)}$. One possible reason is that, as shown in Figure 2, it needs tens of iterations before $\tilde{\alpha}_k$ being a good approximation of $1/\lambda_n$ with satisfactory accuracy. In what follows, by incorporating the BB method, we develop an efficient periodic gradient method using $\tilde{\alpha}_k$.

Figure 3 illustrates a comparison of the gradient method (1.11) using $\Psi(A) = A$ (i.e., the MG method) with a method using 20 BB2 steps first and then MG steps on solving problem (2.44). We can see that by using some BB2 steps, the modified MG method is accelerated and the stepsize $\tilde{\alpha}_k$ will approximate $1/\lambda_n$ with a better accuracy. Thus, our method will run some BB steps first. Now, we investigate the affect of reusing a short stepsize on the performance of the gradient method (1.11). Suppose that we have a good approximation of $1/\lambda_n$, say $\alpha = \frac{1}{\lambda_n + 10^{-6}}$. We compare MG method with its two variants by applying (i) $\alpha_0 = \alpha$ or (ii) $\alpha_0 = \ldots = \alpha_9 = \alpha$ before using the MG stepsize. Figure 4 shows that reusing α will accelerate the MG method. Hence, we prefer to reuse $\tilde{\alpha}_k$ for some consecutive steps when $\tilde{\alpha}_k$ is a good approximation of $1/\lambda_n$. Finally, our new method is summarized in Algorithm 3.1, which periodically applies the BB stepsize, α_k in (1.11) and $\tilde{\alpha}_k$. The R-linear global convergence of Algorithm 3.1 for solving (1.3) can be established by showing that it satisfies the property in [5], see Theorem 3 of [7] for example.

Algorithm 3.1 Periodic gradient method

Choose an initial point $x_0 \in \mathbb{R}^n$, initial stepsize α_0, positive integers K_b, K_m, K_s, and termination tolerance $\epsilon > 0$.
Take one gradient step with α_0
while $\|g_k\| > \epsilon$ **do**
  Take K_b BB steps
  Take K_m gradient steps with α_k in (1.11)
  Take K_s short steps with $\tilde{\alpha}_t$, where $\tilde{\alpha}_t$ is the first stepsize after α_k-steps
end while

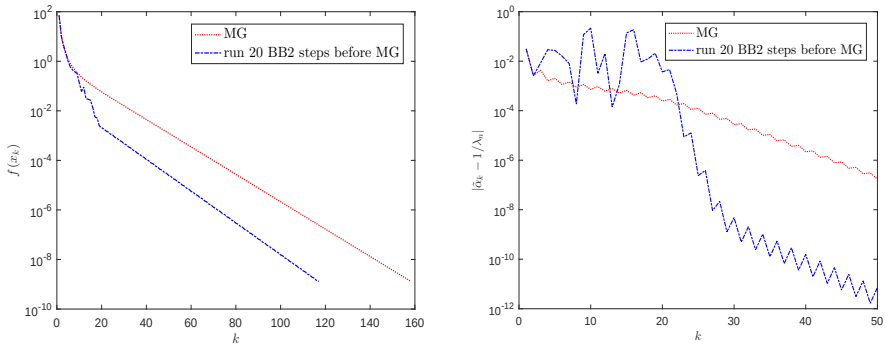

FIG. 3. *Problem* (2.44) *with* $n = 10$*: convergence history of objective values and stepsizes.*

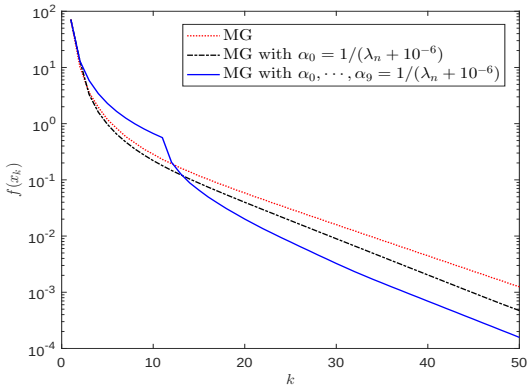

FIG. 4. *Problem* (2.44) *with* $n = 10$*: the MG method (i.e.,* $\Psi(A) = A$*) with different stepsizes.*

Remark 3.4. The BB steps in Algorithm 3.1 can either employ the BB1 or BB2 stepsize in (1.7). The idea of using short stepsizes to eliminate the component $\mu_k^{(n)}$ has been investigated in [12, 13, 20]. However, these methods are based on the SD method, that is, occasionally applying short steps during the iterates of the SD method. One exception is given by [21], where a method is developed by employing new stepsizes during the iterates of the AOPT method. But our method periodically uses three different stepsizes: the nonmonotone BB method, the gradient method (1.11) and the new stepsize $\tilde{\alpha}_k$.

4. Numerical experiments. In this section, we present numerical comparisons of Algorithm 3.1 and the following methods: BB with α_k^{BB1} [2], Dai-Yuan (DY) [11], ABBmin2 [19], and SDC [12].

Notice that the stepsize rule for Algorithm 3.1 can be written as

$$\alpha_k = \begin{cases} \alpha_k^{BB}, & \text{if } \mathrm{mod}(k, K_b + K_m + K_s) < K_b; \\ \alpha_k(\Psi(A)), & \text{if } K_b \le \mathrm{mod}(k, K_b + K_m + K_s) < K_b + K_m; \\ \tilde{\alpha}_k(\Psi(A)), & \text{if } \mathrm{mod}(k, K_b + K_m + K_s) = K_b + K_m; \\ \alpha_{k-1}, & \text{otherwise}, \end{cases} \tag{4.1}$$

where α_k^{BB} can either be α_k^{BB1} or α_k^{BB2}, $\alpha_k(\Psi(A))$ and $\tilde{\alpha}_k(\Psi(A))$ are the stepsizes given by (1.11) and (3.12), respectively. We tested the following four variants of Algorithm 3.1 using combinations of the two BB stepsizes and $\Psi(A) = I$ or A:

- BB1SD: α_k^{BB1} and $\Psi(A) = I$ in (4.1)
- BB1MG: α_k^{BB1} and $\Psi(A) = A$ in (4.1)
- BB2SD: α_k^{BB2} and $\Psi(A) = I$ in (4.1)
- BB2MG: α_k^{BB2} and $\Psi(A) = A$ in (4.1)

Now we derive a formula for the case $\Psi(A) = A$, i.e., $\alpha_k(\Psi(A)) = \alpha_k^{MG}$. If we set $r = 0$, by (3.12), we have

$$\tilde{\alpha}_k = \frac{2}{\left(\frac{1}{\alpha_{k-1}^{SD}} + \frac{g_k^{\mathsf{T}} A^3 g_k}{g_k^{\mathsf{T}} A^2 g_k}\right) + \sqrt{\left(\frac{1}{\alpha_{k-1}^{SD}} - \frac{g_k^{\mathsf{T}} A^3 g_k}{g_k^{\mathsf{T}} A^2 g_k}\right)^2 + \frac{4(g_k^{\mathsf{T}} A g_k)^2}{(\alpha_{k-1}^{MG})^2 \|g_{k-1}\|^2 g_k^{\mathsf{T}} A^2 g_k}}}, \tag{4.2}$$

which is expensive to compute directly. However, if we set $r = 1/2$, we get

$$\tilde{\alpha}_k = \frac{2}{\frac{1}{\alpha_{k-1}^{MG}} + \frac{1}{\alpha_k^{MG}} + \sqrt{\left(\frac{1}{\alpha_{k-1}^{MG}} - \frac{1}{\alpha_k^{MG}}\right)^2 + \frac{4 g_k^{\mathsf{T}} A g_k}{(\alpha_{k-1}^{MG})^2 g_{k-1}^{\mathsf{T}} A g_{k-1}}}}. \tag{4.3}$$

This formula can be computed without additional cost because $g_{k-1}^{\mathsf{T}} A g_{k-1}$ and $g_k^{\mathsf{T}} A g_k$ have been obtained when computing the stepsizes α_{k-1}^{MG} and α_k^{MG}.

All the methods in consideration were implemented in Matlab (v.9.0-R2016a) and carried out on a PC with an Intel Core i7, 2.9 GHz processor and 8 GB of RAM running Windows 10 system. We stopped the algorithm if the number of iteration exceeds 20,000 or the gradient norm reduces by a factor of ϵ.

We randomly generated quadratic problems (1.1) proposed in [7], where $A = QVQ^{\mathsf{T}}$ with

$$Q = (I - 2w_3 w_3^{\mathsf{T}})(I - 2w_2 w_2^{\mathsf{T}})(I - 2w_1 w_1^{\mathsf{T}}),$$

w_1, w_2, and w_3 are unitary random vectors, and $V = diag(v_1, \ldots, v_n)$ is a diagonal matrix where $v_1 = 1$, $v_n = \kappa$, and v_j, $j = 2, \ldots, n-1$, are randomly generated between 1 and κ by the *rand* function in Matlab. We tested seven sets of different distributions of v_j as shown in Table 2 with different values of the condition number κ and tolerance ϵ. In particular, κ were set to $10^4, 10^5, 10^6$ and ϵ were set to $10^{-6}, 10^{-9}, 10^{-12}$. For each value of κ or ϵ, 10 instances were generated and there are totally 630 instances. For each instance, the entries of b were randomly generated in $[-10, 10]$ and $e = (1, \ldots, 1)^{\mathsf{T}}$ was used as the starting point.

The parameter K_b for Algorithm 3.1 was set to 100 for the first and fifth sets and 30 for other sets. Other two parameters K_m and K_s were selected from $\{9, 13, 15\}$. As in [19], the parameter τ of the ABBmin2 method was set to 0.9 for all instances. The parameter pair (h, s) used for the SDC method was set to $(8, 6)$, which is more efficient than other choices for this test.

Table 3 shows the averaged number of iterations of BB1SD and other four compared methods for the seven sets of problems listed in Table 2. We can see that, for the first problem set, our BB1SD method performs much better than the BB, DY and SDC methods, although the ABBmin2 method seems surprisingly efficient among the compared methods. For the second to the last problem sets, our method with different settings performs better than the BB, DY, ABBmin2 and SDC methods. Moreover, for all the settings and different tolerance levels, our method outperforms all the compared four methods in terms of total number of iterations.

TABLE 2
Distributions of v_j.

Set	Spectrum
1	$\{v_2, \ldots, v_{n-1}\} \subset (1, \kappa)$
2	$\{v_2, \ldots, v_{n/5}\} \subset (1, 100)$ $\{v_{n/5+1}, \ldots, v_{n-1}\} \subset (\frac{\kappa}{2}, \kappa)$
3	$\{v_2, \ldots, v_{n/2}\} \subset (1, 100)$ $\{v_{n/2+1}, \ldots, v_{n-1}\} \subset (\frac{\kappa}{2}, \kappa)$
4	$\{v_2, \ldots, v_{4n/5}\} \subset (1, 100)$ $\{v_{4n/5+1}, \ldots, v_{n-1}\} \subset (\frac{\kappa}{2}, \kappa)$
5	$\{v_2, \ldots, v_{n/5}\} \subset (1, 100)$ $\{v_{n/5+1}, \ldots, v_{4n/5}\} \subset (100, \frac{\kappa}{2})$ $\{v_{4n/5+1}, \ldots, v_{n-1}\} \subset (\frac{\kappa}{2}, \kappa)$
6	$\{v_2, \ldots, v_{10}\} \subset (1, 100)$ $\{v_{11}, \ldots, v_{n-1}\} \subset (\frac{\kappa}{2}, \kappa)$
7	$\{v_2, \ldots, v_{n-10}\} \subset (1, 100)$ $\{v_{n-9}, \ldots, v_{n-1}\} \subset (\frac{\kappa}{2}, \kappa)$

Tables 4, 5 and 6 present the averaged number of iterations of BB1MG, BB2SD and BB2MG, respectively. For comparison purposes, the results of the BB, DY, ABBmin 2 and SDC methods are also listed in those tables. As compared with the BB, DY, ABBmin 2 and SDC methods, similar results to those in Table 3 can be seen from these three tables. For the comparison of BB1SD and BB1MG, we can see from Tables 3 and 4 that BB1MG is slightly better than BB1SD for the second to fourth, sixth, and the last problem sets. In addition, BB1MG is comparable to BB1SD for the first and the fifth problem sets. The results in Tables 5 and 6 do not show much difference between BB2SD and BB2MG. In general, BB1MG performs slightly better than BB1SD, BB2SD and BB2MG for most of the problem sets.

We further compared these methods in Figures 5 and 6 by using the performance profiles of Dolan and Moré [15] on the iteration metric. In these figures, the vertical axis shows the percentage of the problems the method solves within the factor ρ of the metric used by the most effective method in this comparison. We select the results of our four methods corresponding to the column $(15, 15)$ in the above tables. It can be seen that all our methods BB1SD, BB1MG, BB2SD and BB2MG clearly outperform the other compared methods. For comparison of BB1SD, BB1MG, BB2SD and BB2MG, Figure 7 shows that BB1MG is slightly better than the other three methods, while BB1SD, BB2SD and BB2MG do not show much difference in this test.

5. Conclusions and discussions. We present theoretical analyses on the asymptotic behaviors of a family of gradient methods whose stepsize is given by (1.11), which includes the steepest descent and minimal gradient methods as special cases. It is shown that each method in this family will asymptotically zigzag in a two-dimensional subspace spanned by the two eigenvectors corresponding to the largest and smallest eigenvalues of the Hessian. In order to accelerate the gradient methods, we exploit the spectral property of a new stepsize to break the zigzagging pattern. This new stepsize is derived by imposing finite termination on minimizing two-dimensional strongly convex quadratics and is proved to converge to the reciprocal of the largest eigenvalue of the Hessian for general n-dimensional case. Finally, we propose a very efficient periodic gradient method that alternately uses the BB stepsize, α_k in (1.11) and our new stepsize. Our numerical results indicate that, by exploiting the asymptotic behavior and spectral properties of stepsizes, gradient methods can be greatly accelerated to outperform the BB method and other recently developed state-of-the-

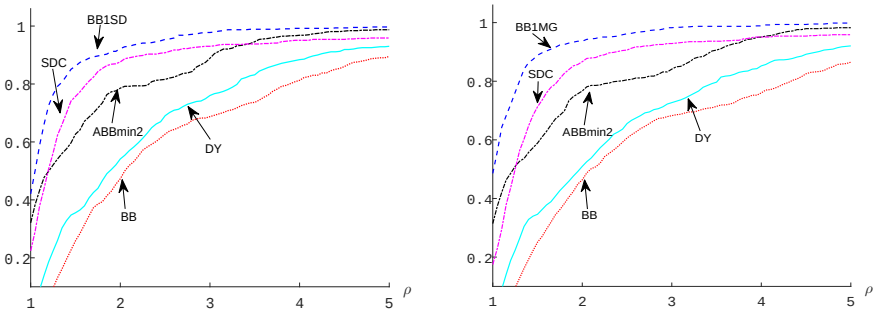

FIG. 5. *Performance profiles for BB1SD (left)/BB1MG (right), and BB, DY, ABBmin2 and SDC, iteration metric, 630 instances of the problems in Table 2.*

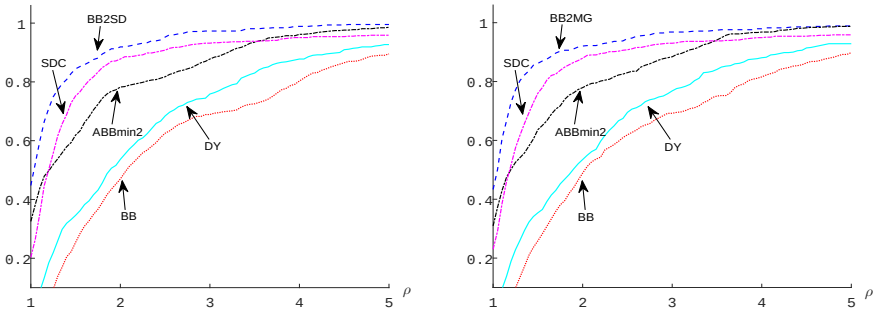

FIG. 6. *Performance profiles for BB2SD (left)/BB2MG (right), and BB, DY, ABBmin2 and SDC, iteration metric, 630 instances of the problems in Table 2.*

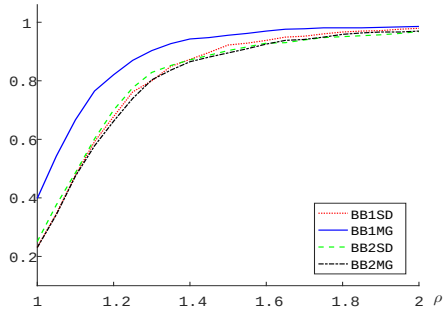

FIG. 7. *Performance profiles for BB1SD, BB1MG, BB2SD and BB2MG, iteration metric, 630 instances of the problems in Table 2.*

art gradient methods.

As a final remark, one may also break the zigzagging pattern by employing the

spectral property in (2.35). In particular, we could use the following stepsize

$$\hat{\alpha}_k = \left(\frac{1}{\alpha_{2k}} + \frac{1}{\alpha_{2k+1}} \right)^{-1}, \tag{5.1}$$

to break the zigzagging pattern. By (2.35), $\hat{\alpha}_k$ satisfies

$$\lim_{k\to\infty} \hat{\alpha}_k = \frac{1}{\lambda_1 + \lambda_n}.$$

Hence, $\hat{\alpha}_k$ is also a good approximation of $1/\lambda_n$ when the condition number $\kappa = \lambda_n/\lambda_1$ is large. One may see the strategy used in [13] for the case of the SD method.

Appendix A. Tables.

TABLE 3
Number of averaged iterations of BB1SD, BB, DY, ABBmin2 and SDC on the problems in Table 2.

Set	ϵ	(K_m, K_s)									BB	DY	ABBmin2	SDC
		(9, 9)	(9, 13)	(9, 15)	(13, 9)	(13, 13)	(13, 15)	(15, 9)	(15, 13)	(15, 15)				
1	10^{-6}	367.6	339.7	372.3	346.1	352.0	344.9	336.8	317.6	368.0	458.7	350.0	258.5	394.1
	10^{-9}	1232.3	983.7	1312.7	1149.2	1149.5	1281.4	1011.4	1086.7	1150.6	3694.4	3520.9	511.2	2410.6
	10^{-12}	1849.9	1514.1	1812.8	1760.6	1780.3	1792.6	1518.1	1605.0	1465.4	6825.4	6561.6	678.2	4917.2
2	10^{-6}	242.1	244.4	235.8	238.2	229.3	236.7	249.6	233.9	240.9	455.7	406.7	380.0	234.1
	10^{-9}	816.3	790.9	765.5	840.0	729.8	737.0	758.9	750.3	746.8	1882.0	1682.6	1425.7	879.8
	10^{-12}	1255.9	1222.6	1207.4	1305.8	1179.1	1154.8	1211.5	1187.4	1178.1	3149.5	2761.6	2255.9	1436.3
3	10^{-6}	297.1	288.5	275.9	284.6	283.1	273.3	283.6	279.7	270.0	495.6	435.8	487.4	298.4
	10^{-9}	829.0	816.2	796.5	848.0	758.5	763.2	796.9	791.7	743.7	1859.9	1678.7	1509.2	926.1
	10^{-12}	1330.5	1241.0	1252.8	1345.9	1224.3	1176.1	1275.2	1189.9	1178.9	3230.2	2747.0	2492.1	1402.4
4	10^{-6}	358.0	331.8	343.5	331.6	331.6	318.4	331.5	326.6	342.6	715.0	585.0	679.9	345.7
	10^{-9}	882.6	823.2	825.3	917.8	814.2	817.4	860.3	808.6	832.4	2097.1	1927.2	1749.7	969.2
	10^{-12}	1422.0	1327.9	1347.9	1324.3	1232.8	1271.9	1318.4	1288.9	1258.0	3355.7	3140.5	2673.9	1451.2
5	10^{-6}	838.4	829.4	850.8	851.9	836.4	855.5	874.1	856.5	844.5	1091.5	849.1	1043.1	861.7
	10^{-9}	3147.9	3086.3	2985.3	2932.6	3004.0	3062.6	3093.1	3086.7	3094.2	5262.6	4606.2	3542.8	4075.9
	10^{-12}	4942.5	4996.4	4688.7	4542.9	5020.5	4921.7	4900.0	4845.1	4868.1	7803.1	8048.4	5518.2	6279.4
6	10^{-6}	155.1	140.8	140.3	138.9	139.4	137.8	132.8	137.9	137.3	257.0	186.1	151.8	143.8
	10^{-9}	554.3	557.4	541.4	590.8	513.8	500.1	559.9	539.4	512.9	1574.2	1265.4	617.8	639.2
	10^{-12}	905.9	883.1	897.7	939.9	801.1	824.3	925.9	895.9	814.9	2603.9	2419.3	894.6	1129.3
7	10^{-6}	455.6	437.0	430.8	457.9	432.9	424.0	445.8	411.3	424.8	893.7	800.3	772.7	470.5
	10^{-9}	905.8	876.0	828.2	922.4	870.5	869.8	925.6	851.0	859.6	2110.7	1868.1	1613.9	936.6
	10^{-12}	1349.8	1323.1	1265.4	1374.2	1278.5	1267.2	1319.2	1252.1	1240.3	3252.1	2748.7	2372.9	1331.5
total	10^{-6}	2713.9	2611.6	2649.4	2649.2	2604.7	2590.6	2654.2	2563.5	2628.1	4367.2	3613.0	3773.4	2748.3
	10^{-9}	8368.2	7933.7	8054.9	8200.8	7840.3	8031.5	8006.1	7914.4	7940.2	18480.9	16549.1	10970.3	10837.4
	10^{-12}	13056.5	12508.2	12472.7	12593.6	12516.6	12408.6	12468.3	12264.3	12003.7	30219.9	28427.1	16885.8	17947.3

TABLE 4
Number of averaged iterations of BB1MG, BB, DY, ABBmin2 and SDC on the problems in Table 2.

Set	ϵ	(K_m, K_s)									BB	DY	ABBmin2	SDC
		(9, 9)	(9, 13)	(9, 15)	(13, 9)	(13, 13)	(13, 15)	(15, 9)	(15, 13)	(15, 15)				
1	10^{-6}	378.0	366.2	344.9	354.3	364.5	341.7	338.1	374.1	362.1	458.7	350.0	258.5	394.1
	10^{-9}	1187.6	1369.2	1192.8	1029.0	1297.6	1040.6	1124.6	1201.2	1095.8	3694.4	3520.9	511.2	2410.6
	10^{-12}	1909.2	1809.4	1666.2	1558.3	1784.7	1577.6	1578.7	1862.7	1485.3	6825.4	6561.6	678.2	4917.2
2	10^{-6}	216.5	211.0	227.0	218.2	211.2	228.5	223.3	225.5	230.2	455.7	406.7	380.0	234.1
	10^{-9}	729.7	679.9	703.0	665.9	674.4	686.0	675.6	665.8	680.9	1882.0	1682.6	1425.7	879.8
	10^{-12}	1199.7	1079.8	1130.7	1076.6	1076.3	1067.1	1096.8	1081.7	1059.3	3149.5	2761.6	2255.9	1436.3
3	10^{-6}	258.3	265.4	273.7	273.3	249.2	254.1	253.1	246.2	252.3	495.6	435.8	487.4	298.4
	10^{-9}	810.6	743.8	756.7	707.6	720.5	694.0	731.2	723.2	701.6	1859.9	1678.7	1509.2	926.1
	10^{-12}	1208.6	1137.4	1182.6	1112.4	1128.5	1102.7	1153.6	1108.9	1099.7	3230.2	2747.0	2492.1	1402.4
4	10^{-6}	309.8	325.1	305.4	315.2	309.3	312.6	315.1	304.9	315.8	715.0	585.0	679.9	345.7
	10^{-9}	871.1	753.9	764.6	771.7	766.2	748.4	766.8	749.2	768.3	2097.1	1927.2	1749.7	969.2
	10^{-12}	1268.6	1186.6	1203.9	1164.3	1162.0	1140.8	1200.9	1159.1	1181.2	3355.7	3140.5	2673.9	1451.2
5	10^{-6}	856.8	833.5	847.7	862.7	847.2	848.3	843.7	906.7	865.1	1091.5	849.1	1043.1	861.7
	10^{-9}	3197.5	3014.6	3216.2	2988.8	3015.1	3088.4	3137.5	3155.4	3042.1	5262.6	4606.2	3542.8	4075.9
	10^{-12}	4937.7	4769.0	4986.6	4933.8	4709.7	4861.1	4944.6	5167.5	4869.2	7803.1	8048.4	5518.2	6279.4
6	10^{-6}	129.1	125.6	126.0	132.5	126.1	135.4	128.6	127.0	137.3	257.0	186.1	151.8	143.8
	10^{-9}	510.8	498.9	510.1	496.3	452.1	471.3	461.6	487.2	447.6	1574.2	1265.4	617.8	639.2
	10^{-12}	841.4	799.5	789.0	808.8	712.1	780.5	754.2	748.2	699.8	2603.9	2419.3	894.6	1129.3
7	10^{-6}	400.6	417.1	382.8	423.1	407.0	405.6	402.0	415.8	402.7	893.7	800.3	772.7	470.5
	10^{-9}	841.3	815.6	788.3	832.9	820.8	794.4	825.4	844.7	814.5	2110.7	1868.1	1613.9	936.6
	10^{-12}	1245.0	1193.1	1161.9	1218.1	1202.7	1190.3	1210.3	1238.0	1167.7	3252.1	2748.7	2372.9	1331.5
total	10^{-6}	2549.1	2543.9	2507.5	2579.3	2514.5	2526.2	2503.9	2600.2	2565.5	4367.2	3613.0	3773.4	2748.3
	10^{-9}	8148.6	7875.9	7931.7	7492.2	7746.7	7523.1	7722.7	7826.7	7550.8	18480.9	16549.1	10970.3	10837.4
	10^{-12}	12610.2	11974.8	12120.9	11872.3	11776.0	11720.1	11939.1	12366.1	11562.2	30219.9	28427.1	16885.8	17947.3

TABLE 5

Number of averaged iterations of BB2SD, BB, DY, ABBmin2 and SDC on the problems in Table 2.

Set	ϵ	(K_m, K_s)									BB	DY	ABBmin2	SDC
		(9, 9)	(9, 13)	(9, 15)	(13, 9)	(13, 13)	(13, 15)	(15, 9)	(15, 13)	(15, 15)				
1	10^{-6}	347.9	357.2	365.1	349.4	344.3	325.0	338.1	349.4	369.2	458.7	350.0	258.5	394.1
	10^{-9}	1132.2	1454.1	1247.4	1192.7	1224.4	1274.7	1237.7	1291.9	1209.6	3694.4	3520.9	511.2	2410.6
	10^{-12}	1985.3	2429.8	1986.8	1838.2	2062.1	2181.2	1958.2	1961.0	1927.2	6825.4	6561.6	678.2	4917.2
2	10^{-6}	219.4	223.9	220.5	226.0	229.3	224.4	217.8	220.4	226.4	455.7	406.7	380.0	234.1
	10^{-9}	749.4	723.3	713.9	746.6	720.1	711.2	728.1	729.4	713.3	1882.0	1682.6	1425.7	879.8
	10^{-12}	1235.9	1188.4	1168.4	1167.9	1158.1	1158.3	1165.2	1186.0	1130.9	3149.5	2761.6	2255.9	1436.3
3	10^{-6}	248.5	259.0	253.8	254.0	246.3	261.6	252.6	262.8	267.4	495.6	435.8	487.4	298.4
	10^{-9}	780.5	757.1	754.2	759.3	738.4	767.2	793.6	774.4	759.3	1859.9	1678.7	1509.2	926.1
	10^{-12}	1229.4	1230.7	1227.8	1216.0	1214.8	1182.3	1215.2	1227.7	1210.6	3230.2	2747.0	2492.1	1402.4
4	10^{-6}	320.8	315.1	305.5	313.6	315.9	310.9	318.4	307.5	317.1	715.0	585.0	679.9	345.7
	10^{-9}	805.0	823.3	813.4	819.5	813.5	789.0	779.5	836.1	802.5	2097.1	1927.2	1749.7	969.2
	10^{-12}	1348.7	1298.3	1244.4	1242.8	1276.1	1238.6	1250.0	1269.9	1246.3	3355.7	3140.5	2673.9	1451.2
5	10^{-6}	860.0	847.3	848.7	831.2	799.3	825.5	804.4	809.5	862.0	1091.5	849.1	1043.1	861.7
	10^{-9}	3066.6	3191.0	2998.8	2918.1	3049.0	3038.7	2995.5	2995.7	3095.7	5262.6	4606.2	3542.8	4075.9
	10^{-12}	5272.4	5133.8	5106.8	4962.9	4867.3	4894.1	5083.6	4775.5	5100.4	7803.1	8048.4	5518.2	6279.4
6	10^{-6}	129.1	138.8	124.8	128.4	135.3	133.7	122.2	130.8	133.4	257.0	186.1	151.8	143.8
	10^{-9}	560.3	549.5	531.5	514.6	520.9	538.9	516.5	530.4	525.1	1574.2	1265.4	617.8	639.2
	10^{-12}	912.8	892.1	940.0	913.5	928.1	873.5	892.5	873.3	845.2	2603.9	2419.3	894.6	1129.3
7	10^{-6}	418.4	393.6	406.6	410.6	409.7	418.4	394.8	429.7	405.9	893.7	800.3	772.7	470.5
	10^{-9}	898.0	835.8	849.0	852.9	847.6	847.8	868.4	873.3	848.4	2110.7	1868.1	1613.9	936.6
	10^{-12}	1324.7	1238.8	1221.1	1290.1	1263.3	1265.1	1302.7	1279.2	1267.4	3252.1	2748.7	2372.9	1331.5
total	10^{-6}	2544.1	2534.9	2525.0	2513.2	2480.1	2499.5	2448.3	2510.1	2581.4	4367.2	3613.0	3773.4	2748.3
	10^{-9}	7992.0	8334.1	7908.2	7803.7	7913.9	7967.5	7919.3	8031.2	7953.9	18480.9	16549.1	10970.3	10837.4
	10^{-12}	13309.2	13411.9	12895.3	12631.4	12769.8	12793.1	12867.4	12572.6	12728.0	30219.9	28427.1	16885.8	17947.3

TABLE 6

Number of averaged iterations of BB2MG, BB, DY, ABBmin2 and SDC on the problems in Table 2.

Set	ϵ	(K_m, K_s)									BB	DY	ABBmin2	SDC
		(9, 9)	(9, 13)	(9, 15)	(13, 9)	(13, 13)	(13, 15)	(15, 9)	(15, 13)	(15, 15)				
1	10^{-6}	355.7	365.1	341.9	322.6	350.7	327.5	337.9	313.6	321.1	458.7	350.0	258.5	394.1
	10^{-9}	1209.5	1327.4	908.0	1064.7	1206.9	1209.7	965.6	1255.1	1351.1	3694.4	3520.9	511.2	2410.6
	10^{-12}	1858.7	1772.7	1477.3	1640.8	1701.6	1877.9	1651.6	1889.2	1751.7	6825.4	6561.6	678.2	4917.2
2	10^{-6}	235.1	237.9	238.2	233.0	229.2	239.2	236.4	235.2	238.0	455.7	406.7	380.0	234.1
	10^{-9}	822.7	778.9	752.8	805.0	747.0	762.7	785.7	748.0	737.0	1882.0	1682.6	1425.7	879.8
	10^{-12}	1273.8	1233.0	1212.6	1294.3	1144.2	1193.2	1248.0	1178.3	1167.0	3149.5	2761.6	2255.9	1436.3
3	10^{-6}	273.8	265.6	287.9	264.6	271.2	274.4	275.2	263.1	281.9	495.6	435.8	487.4	298.4
	10^{-9}	866.7	831.4	793.5	862.2	777.6	789.1	804.3	786.0	786.6	1859.9	1678.7	1509.2	926.1
	10^{-12}	1313.6	1318.9	1244.3	1361.4	1219.6	1234.7	1313.4	1271.2	1251.8	3230.2	2747.0	2492.1	1402.4
4	10^{-6}	333.7	335.8	341.9	353.0	319.9	317.4	331.7	333.0	329.1	715.0	585.0	679.9	345.7
	10^{-9}	876.9	877.7	853.3	863.8	844.5	836.6	881.4	804.5	800.1	2097.1	1927.2	1749.7	969.2
	10^{-12}	1364.3	1329.9	1307.0	1351.1	1296.9	1259.4	1337.0	1275.7	1286.7	3355.7	3140.5	2673.9	1451.2
5	10^{-6}	806.4	836.7	837.7	807.1	842.2	862.9	817.8	814.9	819.9	1091.5	849.1	1043.1	861.7
	10^{-9}	3106.8	3101.1	3008.3	3102.0	3169.6	3058.9	3073.8	2997.6	3097.9	5262.6	4606.2	3542.8	4075.9
	10^{-12}	4996.6	5100.9	4749.5	5079.1	5012.9	5004.8	5090.7	5094.0	4708.6	7803.1	8048.4	5518.2	6279.4
6	10^{-6}	137.1	138.9	135.9	143.4	135.1	139.0	135.1	136.9	138.9	257.0	186.1	151.8	143.8
	10^{-9}	612.6	571.2	535.3	588.6	543.6	523.8	504.2	569.0	523.0	1574.2	1265.4	617.8	639.2
	10^{-12}	933.9	874.6	870.0	1026.1	864.7	830.9	862.3	910.9	861.2	2603.9	2419.3	894.6	1129.3
7	10^{-6}	462.7	430.8	434.4	454.2	428.2	438.8	440.8	437.9	435.1	893.7	800.3	772.7	470.5
	10^{-9}	957.1	932.7	904.4	935.3	868.1	889.4	933.5	917.1	869.6	2110.7	1868.1	1613.9	936.6
	10^{-12}	1383.7	1337.3	1281.5	1344.8	1288.7	1323.3	1373.0	1310.1	1277.8	3252.1	2748.7	2372.9	1331.5
total	10^{-6}	2604.5	2610.8	2617.9	2577.9	2576.5	2599.2	2574.9	2534.6	2564.0	4367.2	3613.0	3773.4	2748.3
	10^{-9}	8452.3	8420.4	7755.6	8221.6	8157.3	8070.2	7948.5	8077.3	8165.3	18480.9	16549.1	10970.3	10837.4
	10^{-12}	13124.6	12967.3	12142.2	13097.6	12528.6	12724.2	12876.0	12929.4	12304.8	30219.9	28427.1	16885.8	17947.3

REFERENCES

[1] H. AKAIKE, *On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method*, Ann. Inst. Stat. Math., 11 (1959), pp. 1–16.

[2] J. BARZILAI AND J. M. BORWEIN, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

[3] E. G. BIRGIN, J. M. MARTÍNEZ, AND M. RAYDAN, *Nonmonotone spectral projected gradient methods on convex sets*, SIAM J. Optim., 10 (2000), pp. 1196–1211.

[4] A. CAUCHY, *Méthode générale pour la résolution des systemes d'équations simultanées*, Comp. Rend. Sci. Paris, 25 (1847), pp. 536–538.

[5] Y.-H. DAI, *Alternate step gradient method*, Optimization, 52 (2003), pp. 395–415.

[6] Y.-H. DAI AND R. FLETCHER, *On the asymptotic behaviour of some new gradient methods*, Math. Program., 103 (2005), pp. 541–559.

[7] Y.-H. DAI, Y. HUANG, AND X.-W. LIU, *A family of spectral gradient methods for optimization*, Comp. Optim. Appl., 74 (2019), pp. 43–65.

[8] 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.

[9] Y.-H. DAI AND X. YANG, *A new gradient method with an optimal stepsize property*, Comp.

Optim. Appl., 33 (2006), pp. 73–88.
[10] Y.-H. Dai and Y.-X. Yuan, *Alternate minimization gradient method*, IMA J. Numer. Anal., 23 (2003), pp. 377–393.
[11] Y.-H. Dai and Y.-X. Yuan, *Analysis of monotone gradient methods*, J. Ind. Mang. Optim., 1 (2005), p. 181.
[12] R. De Asmundis, D. Di Serafino, W. W. Hager, G. Toraldo, and H. Zhang, *An efficient gradient method using the Yuan steplength*, Comp. Optim. Appl., 59 (2014), pp. 541–563.
[13] R. De Asmundis, D. di Serafino, F. Riccio, and G. Toraldo, *On spectral properties of steepest descent methods*, IMA J. Numer. Anal., 33 (2013), pp. 1416–1435.
[14] D. Di Serafino, V. Ruggiero, G. Toraldo, and L. Zanni, *On the steplength selection in gradient methods for unconstrained optimization*, Appl. Math. Comput., 318 (2018), pp. 176–195.
[15] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Program., 91 (2002), pp. 201–213.
[16] H. C. Elman and G. H. Golub, *Inexact and preconditioned Uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.
[17] R. Fletcher, *On the Barzilai–Borwein method*, Optimization and control with applications, (2005), pp. 235–256.
[18] G. E. Forsythe, *On the asymptotic directions of the s-dimensional optimum gradient method*, Numer. Math., 11 (1968), pp. 57–76.
[19] G. Frassoldati, L. Zanni, and G. Zanghirati, *New adaptive stepsize selections in gradient methods*, J. Ind. Mang. Optim., 4 (2008), p. 299.
[20] C. C. Gonzaga and R. M. Schneider, *On the steepest descent algorithm for quadratic functions*, Comp. Optim. Appl., 63 (2016), pp. 523–542.
[21] Y. Huang, Y.-H. Dai, X.-W. Liu, and H. Zhang, *Gradient methods exploiting spectral properties*, arXiv preprint arXiv:1905.03870, (2019).
[22] Y. Huang and H. Liu, *Smoothing projected Barzilai–Borwein method for constrained non-Lipschitz optimization*, Comp. Optim. Appl., 65 (2016), pp. 671–698.
[23] Y. Huang, H. Liu, and S. Zhou, *Quadratic regularization projected Barzilai–Borwein method for nonnegative matrix factorization*, Data Min. Knowl. Disc., 29 (2015), pp. 1665–1684.
[24] B. Jiang and Y.-H. Dai, *Feasible Barzilai–Borwein-like methods for extreme symmetric eigenvalue problems*, Optim. Method Softw., 28 (2013), pp. 756–784.
[25] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, *Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms*, IEEE Trans. Signal Process., 59 (2011), pp. 1142–1157.
[26] J. Nocedal, A. Sartenaer, and C. Zhu, *On the behavior of the gradient norm in the steepest descent method*, Comp. Optim. Appl., 22 (2002), pp. 5–35.
[27] L. Pronzato, H. P. Wynn, and A. A. Zhigljavsky, *Asymptotic behaviour of a family of gradient algorithms in R^d and Hilbert spaces*, Math. Program., 107 (2006), pp. 409–438.
[28] M. Raydan, *On the Barzilai and Borwein choice of steplength for the gradient method*, IMA J. Numer. Anal., 13 (1993), pp. 321–326.
[29] M. Raydan, *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*, SIAM J. Optim., 7 (1997), pp. 26–33.
[30] Y.-X. Yuan, *A new stepsize for the steepest descent method*, J. Comput. Math., (2006), pp. 149–156.
[31] Y.-X. Yuan, *Step-sizes for the gradient method*, AMS IP Studies in Advanced Mathematics, 42 (2008), pp. 785–796.
[32] B. Zhou, L. Gao, and Y.-H. Dai, *Gradient methods with adaptive step-sizes*, Comp. Optim. Appl., 35 (2006), pp. 69–86.