Numerische Mathematik (2019) 142:941–971
https://doi.org/10.1007/s00211-019-01037-7

Numerische Mathematik

A Newton's method characterization for real eigenvalue problems

Yunho Kim[1]

Received: 24 March 2018 / Revised: 20 December 2018 / Published online: 5 April 2019

Abstract
The current work is a continuation of Kim (An unconstrained global optimization framework for real symmetric eigenvalue problems, submitted), where an unconstrained optimization problem was proposed and a first order method was shown to converge to a global minimizer that is an eigenvector corresponding to the smallest eigenvalue with no eigenvalue estimation given. In this second part, we provide local and global convergence analyses of the Newton's method for real symmetric matrices. Our proposed framework discovers a new eigenvalue update rule and shows that the errors in eigenvalue and eigenvector estimations are comparable, which extends to nonsymmetric diagonalizable matrices as well. At the end, we provide numerical experiments for generalized eigenvalue problems and for the trust region subproblem discussed in Adachi et al. (SIAM J Optim 27(1):269–291, 2017) to confirm efficiency and accuracy of our proposed method.

Mathematics Subject Classification 49M15 · 65F15

1 Introduction

Importance of eigenvalues and eigenvectors is manifest in various fields of research. Among the eigenvalues, we are particularly interested in the smallest eigenvalues from a finite dimensional and an infinite dimensional viewpoints. In a finite dimensional viewpoint, we know that finding the smallest eigenvalue generally requires matrix inversion, which is computationally challenging that various numerical techniques avoid matrix inversion and, instead, solve systems of linear equations. In an infinite dimensional viewpoint, as a simple example, the Laplacian operator $-\Delta$ on a compact manifold $\mathcal{M}$ not only encodes structural information of $\mathcal{M}$ in the eigenvalues $0 \leq \lambda_1 < \lambda_2 \leq \cdots$, but also allows for a function space decomposition based on the

✉ Yunho Kim
yunhokim@unist.ac.kr

[1] Ulsan National Institute of Science and Technology, UNIST-gil 50, Ulsan, South Korea

eigenvalues and eigenfunctions. There are interesting related works (e.g., [1–4]) in image processing to compute eigenvalues of $-\Delta$ on surfaces represented by point cloud data. We note that the first few smallest eigenvalues are particularly important because they are robust to noise in the data. With the two viewpoints in mind, an unconstrained minimization problem for real symmetric eigenvalue problems was proposed in [5], whose global minimizer is an eigenvector that can be computed by a first order method, "Gradient Descent", with its convergence guaranteed. In fact, it was shown in [5] that the gradient descent method applied to the proposed functional guarantees convergence to a global minimizer of the functional, which is an eigenvector corresponding to the smallest eigenvalue, only by matrix addition and multiplication without any eigenvalue estimation and that the norm of a global minimizer explicitly decides the smallest eigenvalue. Various applications such as (generalized) symmetric eigenvalue problems and comparisons with well-known methods such as BFGS can be found in [5]. More history including related works of eigenvalue problems in various settings can be found in the introduction of [5] as well as in the references therein.

The main interest of the present work lies in a situation when solving a system of linear equations of the form $A\mathbf{x} = b$ is, indeed, easy and efficient due to a particular structure of the matrix A, where faster algorithms of the Newton type can be expected for the same task as the authors of [6] pointed out. Since the analyzed functional in [5] has the form of difference of convex functionals, one of which is quadratically convex, and the other is almost linearly convex, one may naturally expect the Newton's method to work better than the gradient descent method at least locally near a global minimizer.

Therefore, we present local and global convergence results for a sequence $\{\mathbf{x}_k\}$ generated by the Newton's method applied to the functional in [5]. It is interesting to see that a sequence $\{\mathbf{x}_k\}$ converges if and only if the sequence $\{\|\mathbf{x}_k\|\}$ in $\mathbb{R}$ converges. Moreover, when the sequence converges, the limit of $\{\|\mathbf{x}_k\|\}$ determines the eigenvalue corresponding to the limit of $\{\mathbf{x}_k\}$. This suggests a new eigenvalue update rule using the norm $\|\mathbf{x}_k\|$. We also provide quantitative analyses on the error estimation of an eigenvector in terms of the error in the eigenvalue estimation via a particular system of linear equations arising from the Newton's method.

The rest of this manuscript is organized as follows. In Sect. 2, we analyze convergence of the Newton's method in the case of a symmetric matrix A under certain conditions. Inspired by the convergence analysis, we analyze quantitatively the errors in eigenvalue and eigenvector estimations and show that the errors are comparable with each other. The same analysis also applies to nonsymmetric diagonalizable matrices. In Sect. 3, we provide a few numerical experiments. Firstly, using two eigenvalue update rules, one from our approach and the other of the Rayleigh quotient type, the nature of our update rule is numerically confirmed, which is to find the smallest eigenvalue via an unconstrained minimization from a random initial guess. This is not observed with the other update rule of the Rayleigh quotient type as expected. Secondly, we discuss the trust region subproblem in [7] in the hard case where the authors of [7] considered an additional procedure of using a basis for the null space $\mathcal{N}(A+\lambda_* B)$. Our approach via the Newton's method turns out to implicitly incorporate a necessary component from the null space $\mathcal{N}(A+\lambda_* B)$ into the solution without relying on a basis, which

shows numerical efficiency as well as simplicity. In Sect. 4, we summarize our present work and provide future plans.

2 Analysis of the Newton's method

To begin with, the fundamental model proposed in [5] needs to be reminded. Denoting by $\mathbf{Sym}_N(\mathbb{R})$ the set of $N \times N$ real symmetric matrices, for a given matrix $A \in \mathbf{Sym}_N(\mathbb{R})$ with $\gamma > \max(0, -\lambda_1)$, where $\lambda_1 \leq \cdots \leq \lambda_N$ are the N eigenvalues of A, we define $F_A : \mathbb{R}^N \to \mathbb{R}$ by

$$F_A(\mathbf{x}) = \frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle + \frac{\gamma}{2}\|\mathbf{x}\|^2 - \gamma\|\mathbf{x}\|, \tag{1}$$

and solve

$$\min_{\mathbf{x}\in\mathbb{R}^N} F_A(\mathbf{x}),$$

where $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{y}^T\mathbf{x}$ and $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle}$. An element $\mathbf{x} \in \mathbb{R}^N$ will be considered as an $N \times 1$ vector. For a positive definite $A \in \mathbf{Sym}_N(\mathbb{R})$, any $\gamma > 0$ works. We can see that the form (1) applies not only to $A \in \mathbf{Sym}_N(\mathbb{R})$, but also to a complex Hermitian matrix A with a minor assumption on γ. We can further consider a general $M \times N$ matrix and compute singular values and singular vectors in a similar form to (1). All these cases can be understood once we understand the case of $A \in \mathbf{Sym}_N(\mathbb{R})$ in (1). As was seen in [5], the proposed minimization problem is equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^N} \left[\frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle + \frac{\gamma}{2}(\|\mathbf{x}\| - 1)^2\right].$$

2.1 Convergence of the Newton's method

Since the functional (1) is continuously twice differentiable at $\mathbf{x} \neq 0$ and its nonzero critical points are eigenvectors of A as was discussed in [5], we consider the Newton's method to find the critical points of F_A, which reads as follows.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (\nabla^2 F_A(\mathbf{x}_k))^{-1}\nabla F_A(\mathbf{x}_k), \quad k = 0, 1, 2, \ldots,$$

with an initial $\mathbf{x}_0 \neq 0$ unless $\nabla^2 F_A(\mathbf{x}_k)$ is singular. Indeed, the sequence $\{\mathbf{x}_k\}$ generated by the Newton's method satisfies

$$\left[\frac{1}{\gamma}A + \left(1 - \frac{1}{\|\mathbf{x}_k\|}\right)I + \frac{1}{\|\mathbf{x}_k\|}\mathbf{y}_k\mathbf{y}_k^T\right]\mathbf{x}_{k+1} = \mathbf{y}_k, \quad k = 0, 1, \ldots, \tag{2}$$

with $\mathbf{y}_k = \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|}$. Note that if γ is chosen to satisfy $\gamma + \lambda_j \neq 0$ for all j's, $\left(\frac{1}{\gamma}A + I\right)$ is nonsingular and (2) can be rewritten as

$$B_k\left(\frac{1}{\gamma}A+I\right)\mathbf{x}_{k+1}=\mathbf{y}_k,$$

where $B_k = I - \frac{1}{\|\mathbf{x}_k\|}(I-\mathbf{y}_k\mathbf{y}_k^T)\left(\frac{1}{\gamma}A+I\right)^{-1}$. Furthermore, B_k can be rewritten as

$$B_k = (I-\mathbf{y}_k\mathbf{y}_k^T)\left(I-\frac{1}{\|\mathbf{x}_k\|}\left(\frac{1}{\gamma}A+I\right)^{-1}\right)+\mathbf{y}_k\mathbf{y}_k^T. \tag{3}$$

The decomposition of B_k in (3) says that the orthogonal complement of $\mathbf{y}_k$, which we denote by V_k, is invariant under B_k.

When B_k is also nonsingular, by setting $\mathbf{z}_k = \left(\frac{1}{\gamma}A+I\right)\mathbf{x}_{k+1}$, we can see from $B_k\mathbf{z}_k = \mathbf{y}_k$ and (2) that $\mathbf{y}_k^T\mathbf{z}_k = 1$ and for some $\alpha_k \in \mathbb{R}$,

$$\left(I-\frac{1}{\|\mathbf{x}_k\|}\left(\frac{1}{\gamma}A+I\right)^{-1}\right)\mathbf{z}_k = \alpha_k\mathbf{y}_k \tag{4}$$

in which case it is possible to write $\alpha_k \in \mathbb{R}$ explicitly by $\frac{1}{\alpha_k} = \sum_{j=1}^N \frac{y_{k,j}^2}{1-\frac{1}{\|\mathbf{x}_k\|}(\frac{\gamma}{\gamma+\lambda_j})}$. Moreover, $\mathbf{y}_k^T\mathbf{z}_k = 1$ implies that

$$\|\mathbf{x}_{k+1}\| \geq \frac{1}{\left\|\left(\frac{1}{\gamma}A+I\right)\mathbf{y}_k\right\|} \geq \min_{j=1,\ldots,N}\left|\frac{\gamma}{\gamma+\lambda_j}\right| > 0. \tag{5}$$

Hence, we can see that as long as the system (2) is nonsingular, each iterate $\mathbf{x}_k$ stays away from the origin, the unique nondifferentiable point of F_A.

On the other hand, by recasting (2) using (3), (4), we can see that with $\mu_k = \frac{1}{\|\mathbf{x}_k\|}-1$,

$$\begin{cases}\left(\frac{1}{\gamma}A-\mu_k I\right)\mathbf{w} = \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|},\\ \mathbf{x}_{k+1} = \left(\frac{1}{(1+\mu_k)\left\langle\left(\frac{1}{\gamma}A+I\right)\mathbf{x}_k,\mathbf{w}\right\rangle}\right)\mathbf{w},\end{cases} \tag{6}$$

which suggests a completely different eigenvalue update μ_k from the famous Rayleigh quotient. However, this update is naturally derived from the Newton's method applied to the quadratic functional F_A.

It can also be expected from the recast (6) that the convergence of $\{\|\mathbf{x}_k\|\}$ to $\frac{\gamma}{\gamma+\lambda_{i_0}}$ for some eigenvalue λ_{i_0} of A determines the convergence of $\{\mathbf{x}_k\}$ to a corresponding eigenvector. More interesting to see in Theorem 1 is that when $\{\|\mathbf{x}_k\|\}$ converges, it must converge to $\frac{\gamma}{\gamma+\lambda_{i_0}}$ for some eigenvalue λ_{i_0} of A.

Theorem 1 *Let $A \in \mathbf{Sym}_N(\mathbb{R})$ have eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$. Let $\lambda_1 < \lambda_N$ and $\gamma > \max(0,-\lambda_1)$, and $\mathbf{x}_0 \neq 0$ with $\|\mathbf{x}_0\| \neq \frac{\gamma}{\gamma+\lambda_j}$ for all j's. Suppose that a sequence*

$\{\mathbf{x}_k\}$ can be generated by (2), i.e., $\mathbf{x}_k$ is uniquely computable for all $k \in \mathbb{N}$, and that $\|\mathbf{x}_k\|$ converges to $\eta > 0$.

If $\|\mathbf{x}_k\| \neq \frac{\gamma}{\gamma+\lambda_j}$ for any $1 \leq j \leq N$ and for all $k \in \mathbb{N}$, then the limit $\eta > 0$ must be $\frac{\gamma}{\gamma+\lambda_{i_0}}$ for some $1 \leq i_0 \leq N$ and $\{\mathbf{x}_k\}$ converges to an eigenvector $\mathbf{x}_*$ of A corresponding to the eigenvalue λ_{i_0} with $\|\mathbf{x}_*\| = \eta = \frac{\gamma}{\gamma+\lambda_{i_0}}$.

Proof It suffices to consider the case that A is a diagonal matrix, i.e.,

$$A = \operatorname{diag}(\lambda_1, \ldots, \lambda_N).$$

With a given sequence $\{\mathbf{x}_k\}$ by (2), we can see that for $j = 1, \ldots, N$,

$$\left(1 + \frac{\lambda_j}{\gamma} - \frac{1}{\|\mathbf{x}_k\|}\right) x_{k+1,j} = y_{k,j} \left(1 - \frac{\|\mathbf{x}_{k+1}\|}{\|\mathbf{x}_k\|}(\mathbf{y}_k^T \mathbf{y}_{k+1})\right), \tag{7}$$

where $\mathbf{x}_{k+1} = \begin{bmatrix} x_{k+1,1} \ \ldots \ x_{k+1,N} \end{bmatrix}^T$ and $\mathbf{y}_k = \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|} = \begin{bmatrix} y_{k,1} \ \ldots \ y_{k,N} \end{bmatrix}^T$. Note also that we have

$$\mathbf{y}_k^T \mathbf{y}_{k+1} = \left(\frac{\|\mathbf{x}_k\|}{\|\mathbf{x}_{k+1}\|} - (\mathbf{y}_k^T \mathbf{y}_{k+1})\right) \left(\sum_{j=1}^{N} \frac{y_{k,j}^2}{\|\mathbf{x}_k\| \left(1 + \frac{\lambda_j}{\gamma}\right) - 1}\right). \tag{8}$$

The assumption $\gamma > \max(0, -\lambda_1)$ guarantees that $\left(\frac{1}{\gamma} A + I\right)$ is nonsingular.

It is assumed that $\mathbf{x}_k$ is uniquely computable for all $k \in \mathbb{N}$ and $\|\mathbf{x}_k\| \neq \frac{\gamma}{\gamma+\lambda_j}$ for all k's and j's and that $\|\mathbf{x}_k\| \to \eta$ for some $\eta > 0$. Then, noting from (5) and (7), we have

$$\inf_{k \in \mathbb{N}} \|\mathbf{x}_k\| \geq \frac{\gamma}{\gamma + \lambda_N}, \quad \text{and} \quad 1 - \frac{\|\mathbf{x}_{k+1}\|}{\|\mathbf{x}_k\|}(\mathbf{y}_k^T \mathbf{y}_{k+1}) \neq 0, \quad k \geq 0.$$

By setting $\mathcal{J}_0 := \{j \in \{1, 2, \ldots, N\} : x_{0,j} \neq 0\}$, we can easily see that for $k \geq 0$,

$$x_{k,j} \neq 0 \text{ if and only if } j \in \mathcal{J}_0.$$

We will now prove by contradiction that $\limsup_k \mathbf{y}_k^T \mathbf{y}_{k+1} = 1$. Suppose that

$$\limsup_k \mathbf{y}_k^T \mathbf{y}_{k+1} < 1.$$

Then, there exists $\epsilon < 1$ with $\limsup_k \mathbf{y}_k^T \mathbf{y}_{k+1} = \epsilon$. Given $0 < \delta < 1 - \epsilon$, we may choose $l_1 \in \mathbb{N}$ so that $k \geq l_1$ implies

$$|\|\mathbf{x}_k\| - \eta| < \frac{\delta}{2} \quad \text{and} \quad \left|\frac{\|\mathbf{x}_k\|}{\|\mathbf{x}_{k+1}\|} - 1\right| < \frac{\delta}{2} \quad \text{and} \quad \mathbf{y}_k^T \mathbf{y}_{k+1} < \epsilon + \frac{\delta}{2}. \tag{9}$$

We also choose $J \in \mathcal{J}_0$ satisfying

$$\left|\eta\left(1+\frac{\lambda_J}{\gamma}\right)-1\right| = \min_{j\in\mathcal{J}_0}\left|\eta\left(1+\frac{\lambda_j}{\gamma}\right)-1\right|.$$

From (7), we see that for $k \geq l_1$,

$$|y_{k+1,J}| := \frac{|x_{k+1,J}|}{\|\mathbf{x}_{k+1}\|} \geq \frac{1-\delta-\epsilon}{\left|\eta\left(1+\frac{\lambda_J}{\gamma}\right)-1\right| + \frac{\delta}{2}\left(1+\frac{\lambda_J}{\gamma}\right)}|y_{k,J}| \tag{10}$$

If $\left|\eta\left(1+\frac{\lambda_J}{\gamma}\right)-1\right| < 1-\epsilon$, since we can choose $\delta > 0$ to satisfy

$$0 < \left|\eta\left(1+\frac{\lambda_J}{\gamma}\right)-1\right| + \frac{\delta}{2}\left(1+\frac{\lambda_J}{\gamma}\right) < 1-\delta-\epsilon,$$

we can see from (10) that $1 \geq \lim_{k\to\infty}|y_{k+1,J}| = \infty$, which is impossible. Hence, we must have

$$\min_{j\in\mathcal{J}_0}\left|\eta\left(1+\frac{\lambda_j}{\gamma}\right)-1\right| \geq 1-\epsilon \quad \text{i.e.,} \quad 0 < \eta \leq \frac{\epsilon\gamma}{\gamma+\lambda^*} \text{ or } \eta \geq \frac{(2-\epsilon)\gamma}{\gamma+\lambda_*},$$

where $\lambda_* = \min_{j\in\mathcal{J}_0}\lambda_j$ and $\lambda^* = \max_{j\in\mathcal{J}_0}\lambda_j$.

Suppose that $0 < \eta \leq \frac{\epsilon\gamma}{\gamma+\lambda^*}$. For any $0 < \delta < \frac{(1-\epsilon)\gamma}{2\gamma+\lambda^*} < 1-\epsilon$, we can see from (9) that for each $j \in \mathcal{J}_0$, $k \geq l_1$ implies

$$\|\mathbf{x}_k\|\left(1+\frac{\lambda_j}{\gamma}\right)-1 < \left(\eta+\frac{\delta}{2}\right)\left(1+\frac{\lambda_j}{\gamma}\right)-1 \leq \frac{\epsilon(\gamma+\lambda_j)}{\gamma+\lambda^*}+\frac{\delta}{2}\left(1+\frac{\lambda_j}{\gamma}\right)-1 < 0,$$

and

$$\frac{\|\mathbf{x}_k\|}{\|\mathbf{x}_{k+1}\|} - (\mathbf{y}_k^T\mathbf{y}_{k+1}) > 1-\frac{\delta}{2}-(\mathbf{y}_k^T\mathbf{y}_{k+1}) > 0.$$

This results in, for $k \geq l_1$,

$$\begin{aligned}\mathbf{y}_k^T\mathbf{y}_{k+1} &= \left(\frac{\|\mathbf{x}_k\|}{\|\mathbf{x}_{k+1}\|} - (\mathbf{y}_k^T\mathbf{y}_{k+1})\right)\left(\sum_{j=1}^{N}\frac{y_{k,j}^2}{\|\mathbf{x}_k\|\left(1+\frac{\lambda_j}{\gamma}\right)-1}\right)\\ &\leq \left(1-\frac{\delta}{2}-(\mathbf{y}_k^T\mathbf{y}_{k+1})\right)\left(\sum_{j\in\mathcal{J}_0}\frac{y_{k,j}^2}{\|\mathbf{x}_k\|\left(1+\frac{\lambda_j}{\gamma}\right)-1}\right) < 0.\end{aligned}$$

This implies that $\epsilon \leq 0$, i.e., $\eta \leq 0$, which is a contradiction. Hence, we can see that only $\eta \geq \frac{(2-\epsilon)\gamma}{\gamma+\lambda_*}$ can be possible.

Then, choosing $\delta > 0$ to satisfy $0 < \delta < \frac{(1-\epsilon)\gamma}{2\gamma+\lambda_*} < 1-\epsilon$, we can also see that for $k \geq l_1$, and for each $j \in \mathcal{J}_0$,

$$\begin{aligned}\|\mathbf{x}_k\|\left(1+\frac{\lambda_j}{\gamma}\right)-1 &> \left(\eta-\frac{\delta}{2}\right)\left(1+\frac{\lambda_j}{\gamma}\right)-1\\ &\geq \frac{(2-\epsilon)(\gamma+\lambda_j)}{\gamma+\lambda_*}-\frac{\delta}{2}\left(1+\frac{\lambda_j}{\gamma}\right)-1\\ &> \frac{(2-\epsilon)(\gamma+\lambda_j)}{\gamma+\lambda_*}-\frac{(1-\epsilon)(\gamma+\lambda_j)}{2(2\gamma+\lambda_*)}-1>0,\end{aligned}$$

which implies that for $k \geq l_1$,

$$\begin{aligned}\mathbf{y}_k^T\mathbf{y}_{k+1} &= \left(\frac{\|\mathbf{x}_k\|}{\|\mathbf{x}_{k+1}\|}-(\mathbf{y}_k^T\mathbf{y}_{k+1})\right)\left(\sum_{j=1}^{N}\frac{y_{k,j}^2}{\|\mathbf{x}_k\|\left(1+\frac{\lambda_j}{\gamma}\right)-1}\right)\\ &\geq \left(1-\frac{\delta}{2}-(\mathbf{y}_k^T\mathbf{y}_{k+1})\right)\left(\sum_{j\in\mathcal{J}_0}\frac{y_{k,j}^2}{\|\mathbf{x}_k\|\left(1+\frac{\lambda_j}{\gamma}\right)-1}\right)>0.\end{aligned}$$

Hence, we have

$$0 \leq \epsilon < 1. \tag{11}$$

Now, extracting a subsequence $\{\mathbf{y}_{k_n}\}$ such that $\mathbf{y}_{k_n}^T\mathbf{y}_{k_n+1} \to \epsilon$ as $n \to \infty$, then using the form of (2) and knowing that

$$\liminf_{n\to\infty}\frac{\mathbf{y}_{k_n+1}^T}{\|\mathbf{x}_{k_n+1}\|}\frac{1}{\gamma}A\mathbf{x}_{k_n+1} = \liminf_{n\to\infty}\frac{1}{\gamma}\mathbf{y}_{k_n+1}^TA\mathbf{y}_{k_n+1} \geq \frac{\lambda_*}{\gamma}$$

we can see that

$$\begin{aligned}&\frac{\mathbf{y}_{k_n+1}^T}{\|\mathbf{x}_{k_n+1}\|}\left[\frac{1}{\gamma}A+\left(1-\frac{1}{\|\mathbf{x}_{k_n}\|}\right)I+\frac{1}{\|\mathbf{x}_{k_n}\|}\mathbf{y}_{k_n}\mathbf{y}_{k_n}^T\right]\mathbf{x}_{k_n+1} = \frac{\mathbf{y}_{k_n+1}^T\mathbf{y}_{k_n}}{\|\mathbf{x}_{k_n+1}\|}\\ \Leftrightarrow\ &\frac{1}{\gamma}\mathbf{y}_{k_n+1}^TA\mathbf{y}_{k_n+1}+\left(1-\frac{1}{\|\mathbf{x}_{k_n}\|}\right)+\frac{1}{\|\mathbf{x}_{k_n}\|}(\mathbf{y}_{k_n}^T\mathbf{y}_{k_n+1})^2 = \frac{\mathbf{y}_{k_n+1}^T\mathbf{y}_{k_n}}{\|\mathbf{x}_{k_n+1}\|}\end{aligned}$$

implies

$$\frac{\lambda_*}{\gamma}+\left(1-\frac{1}{\eta}\right)+\frac{\epsilon^2}{\eta} \leq \frac{\epsilon}{\eta} \quad\Leftrightarrow\quad \eta \leq \frac{(1+\epsilon-\epsilon^2)\gamma}{\gamma+\lambda_*}$$

from which we can see that $\eta > 0$ must satisfy

$$\frac{(2-\epsilon)\gamma}{\gamma+\lambda_*} \leq \eta \leq \frac{(1+\epsilon-\epsilon^2)\gamma}{\gamma+\lambda_*}. \tag{12}$$

That is, we have $2-\epsilon \le 1+\epsilon-\epsilon^2 \Leftrightarrow (1-\epsilon)^2 \le 0 \Leftrightarrow \epsilon = 1$. This is a contradiction due to the assumption $\epsilon < 1$.

Therefore, we conclude that $\epsilon < 1$ is impossible and we must have

$$\limsup_{k\to\infty} \mathbf{y}_k^T \mathbf{y}_{k+1} = 1.$$

We can now show that $\eta = \frac{\gamma}{\gamma+\lambda_{i_0}}$ for some $1 \le i_0 \le N$. Suppose that $\eta \ne \frac{\gamma}{\gamma+\lambda_j}$ for any $1 \le j \le N$. By considering a subsequence $\{\mathbf{y}_{k_n}\}$ with $\lim_{n\to\infty} \mathbf{y}_{k_n}^T \mathbf{y}_{k_n+1} = 1$, it is easy to see from (8) and from the Cauchy–Schwarz inequality that

$$\begin{aligned}
\infty &= \lim_{n\to\infty} \frac{(\mathbf{y}_{k_n}^T \mathbf{y}_{k_n+1})^2}{\left(\frac{\|\mathbf{x}_{k_n}\|}{\|\mathbf{x}_{k_n+1}\|} - (\mathbf{y}_{k_n}^T \mathbf{y}_{k_n+1})\right)^2} \\
&= \lim_{n\to\infty} \left[\sum_{j=1}^{N} y_{k_n,j} \left(\frac{y_{k_n,j}}{\|\mathbf{x}_{k_n}\|\left(1+\frac{\lambda_j}{\gamma}\right)-1}\right)\right]^2 \\
&\le \lim_{n\to\infty} \sum_{j=1}^{N} \frac{y_{k_n,j}^2}{\left(\|\mathbf{x}_{k_n}\|\left(1+\frac{\lambda_j}{\gamma}\right)-1\right)^2} \le \max_{1\le j\le N} \frac{1}{\left(\eta\left(1+\frac{\lambda_j}{\gamma}\right)-1\right)^2} < \infty,
\end{aligned}$$

which is a contradiction. Therefore, we can conclude that $\eta = \frac{\gamma}{\gamma+\lambda_{i_0}}$ for some $1 \le i_0 \le N$.

Next, we want to show that $\{\mathbf{x}_k\}$ converges to an eigenvector $\mathbf{x}_*$ of A corresponding to the eigenvalue λ_{i_0} with norm $\|\mathbf{x}_*\| = \eta = \frac{\gamma}{\gamma+\lambda_{i_0}}$. Indeed, there exists $j \in \mathcal{J}_0$ such that $\lambda_j = \lambda_{i_0}$ because the summation in (8) is over all $j \in \mathcal{J}_0$. That is, $\{j \in \mathcal{J}_0 : \lambda_j = \lambda_{i_0}\} \ne \emptyset$. Hence, without loss of generality we may assume that $i_0 \in \mathcal{J}_0$.

Let $k_0 \in \mathbb{N}$ be such that $k \ge k_0$ implies

$$\left|1+\frac{\lambda_{i_0}}{\gamma} - \frac{1}{\|\mathbf{x}_k\|}\right| < \min_{\lambda_j \ne \lambda_{i_0}} \frac{|\lambda_j - \lambda_{i_0}|}{3\gamma}.$$

Then, for $k \ge k_0$, and for $\lambda_j \ne \lambda_{i_0}$,

$$\left|\frac{1+\frac{\lambda_{i_0}}{\gamma} - \frac{1}{\|\mathbf{x}_k\|}}{1+\frac{\lambda_j}{\gamma} - \frac{1}{\|\mathbf{x}_k\|}}\right| < \frac{1}{2}.$$

Noting that for $1 \le j \le N$ and $K \ge 0$, we have

$$\frac{x_{K+1,j}}{\|\mathbf{x}_{K+1}\|} = \left(\Pi_{k=0}^{K} \left[\frac{\frac{1}{\|\mathbf{x}_{k+1}\|} - \frac{1}{\|x_k\|}(\mathbf{y}_k^T \mathbf{y}_{k+1})}{1+\frac{\lambda_j}{\gamma} - \frac{1}{\|\mathbf{x}_k\|}}\right]\right) \frac{x_{0,j}}{\|\mathbf{x}_0\|}, \tag{13}$$

we can see that for $\lambda_j \neq \lambda_{i_0}$,

$$\left|\frac{x_{K+1,j}}{x_{K+1,i_0}}\right| = \left(\Pi_{k=k_0}^{K}\left|\frac{1+\frac{\lambda_{i_0}}{\gamma}-\frac{1}{\|\mathbf{x}_k\|}}{1+\frac{\lambda_j}{\gamma}-\frac{1}{\|\mathbf{x}_k\|}}\right|\right)\left(\Pi_{k=0}^{k_0-1}\left|\frac{1+\frac{\lambda_{i_0}}{\gamma}-\frac{1}{\|\mathbf{x}_k\|}}{1+\frac{\lambda_j}{\gamma}-\frac{1}{\|\mathbf{x}_k\|}}\right|\right)\left|\frac{x_{0,j}}{x_{0,i_0}}\right|$$
$$\leq \frac{1}{2^{K-k_0+1}}\left(\Pi_{k=0}^{k_0-1}\left|\frac{1+\frac{\lambda_{i_0}}{\gamma}-\frac{1}{\|\mathbf{x}_k\|}}{1+\frac{\lambda_j}{\gamma}-\frac{1}{\|\mathbf{x}_k\|}}\right|\right)\left|\frac{x_{0,j}}{x_{0,i_0}}\right| \to 0 \text{ as } K \to \infty.$$

Since $\frac{x_{K+1,j}}{x_{K+1,i_0}} = \frac{x_{0,j}}{x_{0,i_0}}$ for $\lambda_j = \lambda_{i_0}$ and for all $K \geq 0$, we can see that

$$\frac{1}{|y_{k,i_0}|^2} = \frac{\|\mathbf{x}_k\|^2}{|x_{k,i_0}|^2} = \sum_{j=1}^{N}\left|\frac{x_{k,j}}{x_{k,i_0}}\right|^2 \to (m_{i_0})^2 \quad \text{as } k \to \infty,$$

where $m_{i_0} := \left(\sum_{\{j\in\mathcal{J}_0:\lambda_j=\lambda_{i_0}\}}\left|\frac{x_{0,j}}{x_{0,i_0}}\right|^2\right)^{\frac{1}{2}}$. That is,

$$\lim_{k\to\infty}|x_{k,j}| = \begin{cases}\frac{1}{m_{i_0}}\left(\frac{\gamma}{\gamma+\lambda_{i_0}}\right)\left|\frac{x_{0,j}}{x_{0,i_0}}\right|, & j \in \{j : \lambda_j = \lambda_{i_0}\},\\ 0, & j \in \{j : \lambda_j \neq \lambda_{i_0}\}.\end{cases}$$

Knowing that $\{x_{k,i_0}\}_{k=1}^{\infty}$ has at most two subsequential limits $\pm\frac{1}{m_{i_0}}\left(\frac{\gamma}{\gamma+\lambda_{i_0}}\right)$, we can show that $\{x_{k,i_0}\}$ converges to either one of $\pm\frac{1}{m_{i_0}}\left(\frac{\gamma}{\gamma+\lambda_{i_0}}\right)$. Since

$$\lim_{k\to\infty}\left(\left|\sum_{\lambda_j=\lambda_{i_0}}\frac{y_{k,j}^2}{\|\mathbf{x}_k\|\left(1+\frac{\lambda_{i_0}}{\gamma}\right)-1}\right| - \left|\sum_{\lambda_j\neq\lambda_{i_0}}\frac{y_{k,j}^2}{\|\mathbf{x}_k\|\left(1+\frac{\lambda_j}{\gamma}\right)-1}\right|\right) = \infty$$

implies $\lim_{k\to\infty}\left|\sum_{j=1}^{N}\frac{y_{k,j}^2}{\|\mathbf{x}_k\|(1+\frac{\lambda_j}{\gamma})-1}\right| = \infty$, we can see from (8) that

$$1 \geq \left|\frac{\|\mathbf{x}_k\|}{\|\mathbf{x}_{k+1}\|} - (\mathbf{y}_k^T\mathbf{y}_{k+1})\right|\left|\sum_{j=1}^{N}\frac{y_{k,j}^2}{\|\mathbf{x}_k\|\left(1+\frac{\lambda_j}{\gamma}\right)-1}\right|$$

implies $\lim_{k\to\infty}\left|\frac{\|\mathbf{x}_k\|}{\|\mathbf{x}_{k+1}\|} - (\mathbf{y}_k^T\mathbf{y}_{k+1})\right| = 0$. Therefore, not only do we have $\limsup_{k\to\infty}\mathbf{y}_k^T\mathbf{y}_{k+1} = 1$, but also

$$\lim_{k\to\infty}\mathbf{y}_k^T\mathbf{y}_{k+1} = 1.$$

Since $\frac{x_{k,j}}{x_{k,i_0}} = \frac{x_{0,j}}{x_{0,i_0}}$ for $\lambda_j = \lambda_{i_0}$ and

$$\mathbf{y}_k^T \mathbf{y}_{k+1} = \frac{1}{\|\mathbf{x}_k\|\|\mathbf{x}_{k+1}\|} \left(\sum_{\lambda_j = \lambda_{i_0}} (x_{k,i_0} x_{k+1,i_0}) \left(\frac{x_{0,j}}{x_{0,i_0}} \right)^2 + \sum_{\lambda_j \neq \lambda_{i_0}} x_{k,j} x_{k+1,j} \right)$$
$$\to \left(\frac{\gamma + \lambda_{i_0}}{\gamma} \right)^2 (m_{i_0})^2 \lim_{k\to\infty} (x_{k,i_0} x_{k+1,i_0}),$$

we finally have $\lim_{k\to\infty}(x_{k,i_0} x_{k+1,i_0}) = \frac{1}{m_{i_0}^2} \left(\frac{\gamma}{\gamma+\lambda_{i_0}} \right)^2 > 0$, which guarantees that the sequence $\{x_{k,i_0}\}$ converges to either one of $\pm \frac{1}{m_{i_0}} \left(\frac{\gamma}{\gamma+\lambda_{i_0}} \right)$. This concludes that $\{\mathbf{x}_k\}$ converges to $\mathbf{x}_*$, where $\mathbf{x}_* = \begin{bmatrix} x_{*,1} & \dots & x_{*,N} \end{bmatrix}^T$ is given by

$$x_{*,j} = \begin{cases} \left(\frac{x_{0,j}}{x_{0,i_0}} \right) x_{*,i_0}, & \text{for } \lambda_j = \lambda_{i_0}, \\ 0, & \text{for } \lambda_j \neq \lambda_{i_0}, \end{cases}$$

and $x_{*,i_0} = \lim_{k\to\infty} x_{k,i_0} \in \left\{ \pm \frac{1}{m_{i_0}} \frac{\gamma}{\gamma+\lambda_{i_0}} \right\}$.

It is obvious to see that $\mathbf{x}_*$ is an eigenvector of A corresponding to the eigenvalue λ_{i_0} with norm $\|\mathbf{x}_*\| = \frac{\gamma}{\gamma+\lambda_{i_0}}$. □

We now see what happens if $\|\mathbf{x}_k\| = \frac{\gamma}{\gamma+\lambda_i}$ for some $1 \le i \le N$ in which case $\mathbf{x}_{k+1}$ may not be uniquely computable.

Theorem 2 *With the same assumptions on γ as in Theorem 1, we fix $1 \le i \le N$ and further assume that an eigenvalue λ_i has multiplicity 1 and $\mathbf{q}_i$ is a unit eigenvector of A corresponding to λ_i. We also assume that $\mathbf{x}_0$ is not a critical point of F_A, and yet, satisfies $\|\mathbf{x}_0\| = \frac{\gamma}{\gamma+\lambda_i}$.*

If $|\mathbf{q}_i^T \mathbf{x}_0| > 0$, then $\mathbf{x}_1$ obtained by (2) from $\mathbf{x}_0$ is an eigenvector of A corresponding to the eigenvalue λ_i. If we further assume $|\mathbf{q}_i^T \mathbf{x}_0| \neq \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ for all j's with $j < i$, then $\mathbf{x}_2$ obtained by (2) from $\mathbf{x}_1$ is a critical point of F_A, i.e., an eigenvector of A corresponding to the eigenvalue λ_i with norm $\|\mathbf{x}_2\| = \frac{\gamma}{\gamma+\lambda_i}$.

However, if $|\mathbf{q}_i^T \mathbf{x}_0| = \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ for some $j < i$, then the system becomes singular and $\mathbf{x}_2$ is not uniquely computable.

Proof It suffices to consider the case that A is a diagonal matrix, i.e.,

$$A = \mathrm{diag}(\lambda_1, \dots, \lambda_N)$$

and $\mathbf{q}_i = \mathbf{e}_i$, where $\{\mathbf{e}_1, \dots, \mathbf{e}_N\}$ is the standard basis for $\mathbb{R}^N$. Note that $\mathbf{x}_0$ is not a critical point of F_A and yet, satisfies $\|\mathbf{x}_0\| = \frac{\gamma}{\gamma+\lambda_i}$ for some i.

We first assume that $|\mathbf{e}_i^T \mathbf{x}_0| = |x_{0,i}| > 0$. This implies $|x_{0,i}| < \frac{\gamma}{\gamma+\lambda_i}$. Since the multiplicity of λ_i is 1, the system (2) is nonsingular, that is,

$$\frac{1}{\gamma} A + \left(1 - \frac{1}{\|\mathbf{x}_0\|}\right) I + \frac{1}{\|\mathbf{x}_0\|} \mathbf{y}_0 \mathbf{y}_0^T = \frac{1}{\gamma} \operatorname{diag}(\lambda_1 - \lambda_i, \ldots, \lambda_N - \lambda_i) + \frac{\gamma}{\gamma + \lambda_i} \mathbf{y}_0 \mathbf{y}_0^T$$

is nonsingular, and we can see from (7) that $\mathbf{x}_1$ satisfies $\frac{1}{\|\mathbf{x}_0\|}(\mathbf{y}_0^T \mathbf{x}_1) = 1$ and $\left(\frac{\lambda_j - \lambda_i}{\gamma}\right) x_{1,j} = 0$ for $j = 1, \ldots, N$. Indeed, we have that $\mathbf{x}_1 = \alpha \mathbf{e}_i$ with

$$\alpha = \frac{\gamma^2}{x_{0,i}(\gamma + \lambda_i)^2}, \tag{14}$$

which means that $\mathbf{x}_1$ is an eigenvector of A corresponding to the eigenvalue λ_i with $\|\mathbf{x}_1\| = |\alpha| > \frac{\gamma}{\gamma+\lambda_i}$. However, $\mathbf{x}_1$ is still not a critical point of F_A.

We can further see from (7) with $\mathbf{y}_1 = \pm \mathbf{e}_i$ that $\mathbf{x}_2$ satisfies for $1 \le j \le N$,

$$\begin{cases} \left(\frac{\lambda_j}{\gamma} + 1 - \frac{1}{|\alpha|}\right) x_{2,j} = \delta_{ij} \left(1 - \frac{1}{|\alpha|} x_{2,i}\right), & \text{if } \mathbf{y}_1 = \mathbf{e}_i, \\ \left(\frac{\lambda_j}{\gamma} + 1 - \frac{1}{|\alpha|}\right) x_{2,j} = -\delta_{ij} \left(1 + \frac{1}{|\alpha|} x_{2,i}\right), & \text{if } \mathbf{y}_1 = -\mathbf{e}_i. \end{cases} \tag{15}$$

Since $|\alpha| = \frac{\gamma}{\gamma+\lambda_j}$ is equivalent to $|x_{0,i}| = \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ from (14), we can see from $|x_{0,i}| < \frac{\gamma}{\gamma+\lambda_i}$ that $|\alpha| = \frac{\gamma}{\gamma+\lambda_j}$ is possible only when $\lambda_j < \lambda_i$, i.e., $j < i$.

Hence, if $|x_{0,i}| \neq \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ for $j < i$, then $|\alpha| \neq \frac{\gamma}{\gamma+\lambda_j}$ for $j < i$, and (15) is nonsingular and has a unique solution

$$\mathbf{x}_2 = \frac{\gamma}{\gamma + \lambda_i} \mathbf{y}_1$$

depending on $\mathbf{y}_1 = \pm \mathbf{e}_i$. That is, $\mathbf{x}_2$ is a critical point of F_A, an eigenvector of A corresponding to the eigenvalue λ_i with norm $\|\mathbf{x}_2\| = \frac{\gamma}{\gamma+\lambda_i}$.

On the other hand, if $|\mathbf{x}_{0,i}| = \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ for some $j < i$, then $|\alpha| = \frac{\gamma}{\gamma+\lambda_j}$ and the system (15) becomes singular and $\mathbf{x}_2$ cannot be uniquely computable. □

Theorem 2 says that the Newton's method (2) naturally suggests a nonsingular system for an eigenvector of A corresponding to a known eigenvalue λ of multiplicity 1, which will be discussed in a later section as an application.

What is more interesting is that when the Newton's method generates $\{\mathbf{x}_k\}_{k=1}^{\infty}$ uniquely and if $\{\|\mathbf{x}_k\|\}$ converges, then $\{\mathbf{x}_k\}$ converges as well regardless of the multiplicity of the corresponding eigenvalue λ of A just as [5] showed that the convergence of a generated sequence to an eigenvector by the gradient descent method is irrelevant to the multiplicity of the corresponding eigenvalue. Not to mention, an initial vector $\mathbf{x}_0$ determines the subspace of $\mathbb{R}^N$ where all $\mathbf{x}_k$'s lie.

2.2 The Newton's method (2) for $\gamma \in \mathbb{R}\backslash\{0\}$: global convergence

We want to emphasize that $\gamma > \max(0, -\lambda_1)$ was assumed for the functional F_A to be bounded below for the existence of a global minimizer. However, when considering the Newton's method (2), we realize that we are interested in not only a global minimizer, but also critical points of F_A.

In this section, we consider $\gamma \in \mathbb{R}\backslash\{0\}$ and $A \in \mathbf{Sym}_N(\mathbb{R})$ with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$ with $\lambda_1 < \lambda_N$. In addition, given $N \in \mathbb{N}$ and $\gamma \in \mathbb{R}\backslash\{0\}$, we set $J_N = \{1, 2, \ldots, N\}$ and

$$J^\gamma = \{j \in J_N : \gamma(\gamma + \lambda_j) > 0\}.$$

Due to our intention to analyze the Newton's method (2) for $\gamma \in \mathbb{R}\backslash\{0\}$, the functional in (1) is denoted by F_γ, not by F_A, to emphasize the role of γ.

Lemma 1 *Given $\gamma \in \mathbb{R}\backslash\{0\}$, the set of nonzero critical points of F_γ is the set of eigenvectors of A corresponding to the eigenvalues $\{\lambda_j\}_{j\in J^\gamma}$ with norm $\|\mathbf{x}\| = \frac{\gamma}{\gamma+\lambda_j}$, i.e.*

$$\bigcup_{j\in J^\gamma} \left\{ \mathbf{x} \in \mathbb{R}^N : A\mathbf{x} = \lambda_j \mathbf{x} \text{ with } \|\mathbf{x}\| = \frac{\gamma}{\gamma + \lambda_j} \right\}$$

Proof Since F_γ is differentiable at $\mathbf{x} \neq 0$, we can see that $\nabla F_\gamma(\mathbf{x}) = 0$ for $\mathbf{x} \neq 0$ is equivalent to $\mathbf{x}$ being an eigenvector of A corresponding to an eigenvalue $\lambda_j = \gamma(\frac{1}{\|\mathbf{x}\|} - 1)$ for some $j \in J$. Noting that $\|\mathbf{x}\| = \frac{\gamma}{\gamma+\lambda_j} > 0$, we can easily see that the set of nonzero critical points of F_γ is

$$\bigcup_{j\in J^\gamma} \left\{ \mathbf{x} \in \mathbb{R}^N : A\mathbf{x} = \lambda_j \mathbf{x} \text{ with } \|\mathbf{x}\| = \frac{\gamma}{\gamma + \lambda_j} \right\}.$$

□

In particular, if $\lambda_1 \leq 0 < \lambda_2$ and $-\lambda_2 < \gamma < 0$, since the multiplicity of λ_1 is 1, there are only two critical points of F_γ that are

$$\pm \frac{\gamma}{\gamma + \lambda_1} \mathbf{x}_* \in \{\mathbf{x} \in \mathbb{R}^N : \|\mathbf{x}\| = 1\}$$

with a unit eigenvector $\mathbf{x}_*$ of A corresponding to the eigenvalue λ_1 in which case we provide a global convergence result for the Newton's method. When we choose an initial $\mathbf{x}_0$ at random, $\mathbf{x}_0$ can be thought of as one realization of a Gaussian random vector.

Theorem 3 *We assume that $0 < -\lambda_1 < \lambda_2$ and $-\lambda_2 < \gamma < 0$ and $2\gamma + \lambda_N + \lambda_1 > 0$. Suppose that $\mathbf{x}_k$ is computable for all $k \in \mathbb{N}$ with $\|\mathbf{x}_k\|$ replaced by $\min(1, \|\mathbf{x}_k\|)$ in (2). We choose $\mathbf{x}_0$ at random. Then, the sequence $\{\mathbf{x}_k\}$ converges to one of the two critical points $\pm\frac{\gamma}{\gamma+\lambda_1}\mathbf{x}_*$ of F_γ, where $\mathbf{x}_*$ is a unit eigenvector of A corresponding to λ_1.*

Proof For simplicity, we consider a diagonal matrix $A = \text{diag}(\lambda_1, \ldots, \lambda_N)$ and show that $\{\mathbf{x}_k\}$ converges to either one of $\pm\frac{\gamma}{\gamma+\lambda_1}\mathbf{e}_1$. By assumption, $\frac{1}{\gamma}A+I$ has eigenvalues $\frac{\gamma+\lambda_1}{\gamma} > 1$ and $\frac{\gamma+\lambda_j}{\gamma} < 0$ for $j = 2, \ldots, N$. Since $\frac{1}{\gamma}A + I$ is nonsingular, we can see from (2), (3), (4), (5) that $\mathbf{x}_{k+1}$, $k \geq 0$, satisfies

$$\mathbf{x}_{k+1} = \left(\frac{1}{\gamma}A + I\right)^{-1}\mathbf{z}_k = \alpha_k \,\text{diag}(\omega_{k,1}, \ldots, \omega_{k,N})\mathbf{y}_k, \tag{16}$$

where $\omega_{k,j} = \frac{\frac{\gamma}{\gamma+\lambda_j}}{1-\frac{1}{\min(1,\|\mathbf{x}_k\|)}\frac{\gamma}{\gamma+\lambda_j}}$ for $j = 1, \ldots, N$, and $\inf_{k\geq 1}\|\mathbf{x}_k\| \geq \left|\frac{\gamma}{\gamma+\lambda_N}\right|$.

We note that if $\|\mathbf{x}_k\| > \frac{\gamma}{\gamma+\lambda_1}$,

$$|\omega_{k,1}| = \left|\frac{\frac{\gamma}{\gamma+\lambda_1}}{1-\frac{1}{\min(1,\|\mathbf{x}_k\|)}\frac{\gamma}{\gamma+\lambda_1}}\right| \geq \frac{\gamma}{\lambda_1} > \left|\frac{\gamma}{\lambda_2}\right| \geq \max_{j=2,\ldots,N}(|\omega_{k,j}|),$$

and if $\left|\frac{\gamma}{\gamma+\lambda_N}\right| \leq \|\mathbf{x}_k\| < \frac{\gamma}{\gamma+\lambda_1}$,

$$\left|\frac{\omega_{k,1}}{\|\mathbf{x}_k\|}\right| = 1 + \frac{1}{\frac{1}{\|\mathbf{x}_k\|}\frac{\gamma}{\gamma+\lambda_1} - 1} \geq 1 + \frac{1}{\left|\frac{\gamma+\lambda_N}{\gamma+\lambda_1}\right| - 1} > 1 \geq \max_{j=2,\ldots,N}\left|\frac{\omega_{k,j}}{\|\mathbf{x}_k\|}\right|.$$

This implies that if $\left|\frac{\gamma}{\gamma+\lambda_N}\right| \leq \|\mathbf{x}_k\|$ and $\|\mathbf{x}_k\| \neq \frac{\gamma}{\gamma+\lambda_1}$, for $j = 2, \ldots, N$,

$$\left|\frac{\omega_{k,j}}{\omega_{k,1}}\right| \leq c := \max\left(\left|\frac{\lambda_1}{\lambda_2}\right|, \frac{\left|\frac{\gamma+\lambda_N}{\gamma+\lambda_1}\right| - 1}{\left|\frac{\gamma+\lambda_N}{\gamma+\lambda_1}\right|}\right) < 1.$$

With $\mathbf{y}_{k+1} = \frac{\mathbf{x}_{k+1}}{\|\mathbf{x}_{k+1}\|} = [y_{k+1,1} \;\cdots\; y_{k+1,N}]^T$, we have

$$|y_{k+1,1}| = \frac{|y_{k,1}|}{\sqrt{y_{k,1}^2 + \sum_{j=2}^N \left(\frac{\omega_{k,j}}{\omega_{k,1}}\right)^2 y_{k,j}^2}} \geq \frac{|y_{k,1}|}{\sqrt{y_{k,1}^2 + c^2\sum_{j=2}^N y_{k,j}^2}} \geq |y_{k,1}|.$$

That is, $\{|y_{k,1}|\}$ is an increasing sequence bounded above by 1. Since $\lim_{k\to\infty}|y_{k,1}|$ exists, we can see that together with $c < 1$,

$$\sqrt{y_{k,1}^2 + c^2\sum_{j=2}^N y_{k,j}^2} \to 1 = \sqrt{\sum_{j=1}^N y_{k,j}^2}$$

implies that $y_{k,j} \to 0$ for $j = 2, \ldots, N$, and $|y_{k,1}| \to 1$. Therefore, we have

$$|\mathbf{y}_k| := [|y_{k,1}| \;\cdots\; |y_{k,N}|]^T \to \mathbf{e}_1 \tag{17}$$

In other words, $\{\mathbf{y}_k\}$ has at most two subsequential limits $\pm\mathbf{e}_1$.

We now show that $\{\mathbf{y}_k\}$ converges to either one of $\pm\mathbf{e}_1$ and $\|\mathbf{x}_k\| \to \frac{\gamma}{\gamma+\lambda_1}$. Note from (4) and (17) that

$$\liminf_{k\to\infty} \|\mathbf{x}_k\| \geq \liminf_{k\to\infty} \frac{1}{\left\|\left(\frac{1}{\gamma}A + I\right)\mathbf{y}_k\right\|} = \frac{\gamma}{\gamma+\lambda_1}. \tag{18}$$

On the other hand, we have that

$$0 \leq \lim_{k\to\infty} \sum_{j=2}^{N} \frac{y_{k,j}^2}{1 - \frac{1}{\min(1,\|\mathbf{x}_k\|)}\left(\frac{\gamma}{\gamma+\lambda_j}\right)} \leq \lim_{k\to\infty} \sum_{j=2}^{N} y_{k,j}^2 = 0,$$

and that $\|\mathbf{x}_k\| \geq \left|\frac{\gamma}{\gamma+\lambda_N}\right|$ implies

$$\left|\frac{y_{k,1}^2}{1 - \frac{1}{\min(1,\|\mathbf{x}_k\|)}\left(\frac{\gamma}{\gamma+\lambda_1}\right)}\right| \geq \frac{y_{k,1}^2}{1 + \left|\frac{\gamma+\lambda_N}{\gamma+\lambda_1}\right|}.$$

Therefore, we have

$$\liminf_{k\to\infty} \left|\frac{1}{\alpha_k}\right| \geq \liminf_{k\to\infty} \left|\frac{y_{k,1}^2}{1 - \frac{1}{\min(1,\|\mathbf{x}_k\|)}\left(\frac{\gamma}{\gamma+\lambda_1}\right)}\right| \geq \frac{1}{1 + \left|\frac{\gamma+\lambda_N}{\gamma+\lambda_1}\right|},$$

i.e., $\limsup_{k\to\infty} |\alpha_k| \leq 1 + \left|\frac{\gamma+\lambda_N}{\gamma+\lambda_1}\right|$, which shows

$$\begin{aligned}
&\limsup_{k\to\infty} \left|1 - \frac{\alpha_k y_{k,1}^2}{1 - \frac{1}{\min(1,\|\mathbf{x}_k\|)}\left(\frac{\gamma}{\gamma+\lambda_1}\right)}\right| \\
&= \limsup_{k\to\infty} \left|\alpha_k \sum_{j=2}^{N} \frac{y_{k,j}^2}{1 - \frac{1}{\min(1,\|\mathbf{x}_k\|)}\left(\frac{\gamma}{\gamma+\lambda_j}\right)}\right| = 0.
\end{aligned}$$

Knowing that $|y_{k,1}| \to 1$, we have $\frac{\alpha_k}{1 - \frac{1}{\min(1,\|\mathbf{x}_k\|)}\left(\frac{\gamma}{\gamma+\lambda_1}\right)} \to 1$ and conclude that

$$\begin{aligned}
\limsup_{k\to\infty} \|\mathbf{x}_{k+1}\|^2 &= \limsup_{k\to\infty} \sum_{j=1}^{N} \frac{\alpha_k^2 \left(\frac{\gamma}{\gamma+\lambda_j}\right)^2 y_{k,j}^2}{\left(1 - \frac{1}{\min(1,\|\mathbf{x}_k\|)}\frac{\gamma}{\gamma+\lambda_j}\right)^2} \\
&= \limsup_{k\to\infty} \frac{\alpha_k^2 \left(\frac{\gamma}{\gamma+\lambda_1}\right)^2 y_{k,1}^2}{\left(1 - \frac{1}{\min(1,\|\mathbf{x}_k\|)}\frac{\gamma}{\gamma+\lambda_1}\right)^2} = \left(\frac{\gamma}{\gamma+\lambda_1}\right)^2.
\end{aligned}$$

Together with (18), we obtain that $\lim_{k\to\infty}\|\mathbf{x}_k\| = \frac{\gamma}{\gamma+\lambda_1}$. We can also see from $\alpha_k\omega_{k,1} \to \frac{\gamma}{\gamma+\lambda_1}$ that

$$\lim_{k\to\infty}\mathbf{y}_k^T\mathbf{y}_{k+1} = \lim_{k\to\infty}\frac{1}{\|\mathbf{x}_{k+1}\|}\sum_{j=1}^{N}\alpha_k w_{k,j}y_{k,j}^2 = \lim_{k\to\infty}\frac{\alpha_k w_{k,1}y_{k,1}^2}{\|\mathbf{x}_{k+1}\|} = 1.$$

This implies that $\{\mathbf{y}_k\}$ converges to either one of $\pm\mathbf{e}_1$ just as we saw in Theorem 1. Therefore, we conclude that $\{\mathbf{x}_k\}$ converges to either one of $\pm\frac{\gamma}{\gamma+\lambda_1}\mathbf{e}_1$. □

As was seen in Theorem 1, the multiplicity of λ_1 is irrelevant in Theorem 3. That is, if λ_1 has multiplicity $1 < p < N$ so that we have

$$\lambda_1 = \cdots = \lambda_p < 0 < \lambda_{p+1} \le \cdots \le \lambda_N$$

and if γ satisfies $-\lambda_{p+1} < \gamma < 0$ and $2\gamma + \lambda_N + \lambda_1 > 0$, then whenever $\mathbf{x}_k$ is computable for all $k \in \mathbb{N}$ as in Theorem 3, we can prove in the same way that $\{\mathbf{x}_k\}$ satisfies that $\{|y_{k,j}|\}_{k\in\mathbb{N}}$ for each $j = 1, \ldots, p$ is an increasing sequence bounded above by 1 and

$$\lim_{k\to\infty}\|\mathbf{x}_k\| = \frac{\gamma}{\gamma+\lambda_1} \quad\text{and}\quad \lim_{k\to\infty}\mathbf{y}_k^T\mathbf{y}_{k+1} = 1 \quad\text{and}\quad \lim_{k\to\infty}(A-\lambda_1 I)\mathbf{x}_k = 0,$$

which eventually guarantees that $\{\mathbf{x}_k\}$ converges to an eigenvector of A corresponding to λ_1 with norm $\frac{\gamma}{\gamma+\lambda_1}$.

It is interesting to see that even when $0 < -\lambda_1 = \lambda_2$, by choosing $\epsilon \ll 1$ and replacing $\|\mathbf{x}_k\|$ by $\min(1-\epsilon, \|\mathbf{x}_k\|)$ in (2), we can modify the proof of Theorem 3 slightly to show that $\{\mathbf{x}_k\}$ converges to either one of $\pm\frac{\gamma}{\gamma+\lambda_1}\mathbf{x}_*$, where $\mathbf{x}_*$ is a unit eigenvector of A corresponding to λ_1.

In the case that A is positive semidefinite, we can consider $\gamma < 0$ and obtain a global convergence result as a corollary of Theorem 3.

Corollary 1 *Suppose that* $\lambda_1 = 0 < \lambda_2 \le \cdots \le \lambda_N$. *Let* $\gamma < 0$ *be such that* $2\gamma + \lambda_2 > 0$. *Let* $\{\mathbf{x}_k\}$ *be generated with* $\|\mathbf{x}_k\|$ *replaced by* $\min(1+\epsilon, \|\mathbf{x}_k\|)$ *in* (2), *with* $\mathbf{x}_0$ *chosen at random and* $0 < \epsilon \ll 1$. *Then, the sequence* $\{\mathbf{x}_k\}$ *converges to one of the two critical points* $\pm\mathbf{x}_*$ *of* F_γ, *where* $\mathbf{x}_*$ *is a unit eigenvector of* A *corresponding to* $\lambda_1 = 0$.

Proof We believe that it suffices to indicate only the part of the proof of Theorem 3 that needs a slight modification. For any $0 < \epsilon < 2$, we can see that

$$\left|\frac{(1+\epsilon)\gamma}{\epsilon\gamma}\right| > \left|\frac{(1+\epsilon)\gamma}{\lambda_2+\epsilon(\gamma+\lambda_2)}\right|.$$

Furthermore, it is easy to see that for $0 < \epsilon \ll 1$, if $\|\mathbf{x}_k\| > 1$, we have

$$|\omega_{k,1}| = \left|\frac{1}{1-\frac{1}{\min(1+\epsilon,\|\mathbf{x}_k\|)}}\right| \ge \left|\frac{(1+\epsilon)\gamma}{\epsilon\gamma}\right| > \left|\frac{(1+\epsilon)\gamma}{\lambda_2+\epsilon(\gamma+\lambda_2)}\right| > \max_{j=2,\ldots,N}(|\omega_{k,j}|),$$

and if $\|\mathbf{x}_k\| < 1$, we have

$$\left|\frac{\omega_{k,1}}{\|\mathbf{x}_k\|}\right| = \frac{1}{\frac{1}{\|\mathbf{x}_k\|} - 1} > \|\mathbf{x}_k\| > \frac{1}{\frac{1}{\|\mathbf{x}_k\|} - \frac{\gamma+\lambda_j}{\gamma}} = \left|\frac{\omega_{k,j}}{\|\mathbf{x}_k\|}\right|.$$

As long as $\|\mathbf{x}_k\| \neq 1$ for all k's, since $\inf_{k\in\mathbb{N}} \|\mathbf{x}_k\| > 0$, we can easily see that

$$\sup_{k\in\mathbb{N}, 1\leq j\leq N} \left|\frac{\omega_{k,j}}{\omega_{k,1}}\right| < 1.$$

The rest of the proof is the same as that of Theorem 3. □

2.3 Error estimation

2.3.1 The symmetric case

It is well known that an eigenvector of A is a solution to a singular linear system, which makes it difficult to find a corresponding eigenvector, and iterative methods are designed to approximate eigenvectors. However, the nonsingular system that we are going to see guarantees to find an eigenvector by solving the system once if the corresponding eigenvalue has multiplicity 1. Moreover, even if we have an estimated eigenvalue, the error in the eigenvector estimation turns out to be comparable with the error from the estimated eigenvalue. For eigenvalues with multiplicity greater than 1, the same form of a nonsingular system provides an estimated eigenvector within an arbitrarily small error.

One Step Eigenvector Estimation *Given an $N \times N$ symmetric matrix A, and an eigenvalue $\tilde{\lambda}$ of A, and $\gamma > 0$ with $\gamma \neq -\tilde{\lambda}$, we choose $\mathbf{x}_0$ uniformly at random from S^{N-1} and solve for $\mathbf{x}$,*

$$(A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})\mathbf{x}_0\mathbf{x}_0^T)\mathbf{x} = \gamma\mathbf{x}_0. \tag{19}$$

In fact, the case $\tilde{\lambda} = 0$ turns (19) into $(A + \gamma\mathbf{x}_0\mathbf{x}_0^T)\mathbf{x} = \gamma\mathbf{x}_0$, i.e., if we set $B = A - \tilde{\lambda} I$, then (19) is equivalent to $(B + \gamma\mathbf{x}_0\mathbf{x}_0^T)\mathbf{x} = \gamma\mathbf{x}_0$. By choosing $\gamma = 1$, we have $(B + \mathbf{x}_0\mathbf{x}_0^T)\mathbf{x} = \mathbf{x}_0$, i.e., $(A - \tilde{\lambda} I + \mathbf{x}_0\mathbf{x}_0^T)\mathbf{x} = \mathbf{x}_0$. Hence, instead of (19), we may use $(A - \tilde{\lambda} I + \mathbf{x}_0\mathbf{x}_0^T)\mathbf{x} = \mathbf{x}_0$. However, we decide to keep the form (19) because it is not only an equation that the Newton's method (2) converges to in the limit, but also will be used to approximate an eigenvector in the case of a non-exact eigenvalue.

Before further proceeding, we want to mention a work [8] of Peters and Wilkinson, which was further explained in [6]. In [8], the authors discussed an idea of computing an approximate eigenvector $\mathbf{x}_\lambda$ when an approximate eigenvalue λ is given, i.e., when $A - \lambda I$ is very ill-conditioned, or near singular, by considering

$$(A - \lambda I + \mathbf{x}_i p^T)\mathbf{x} = \mathbf{x}_i, \tag{20}$$

with a random vector p, inspired by the inverse iteration, i.e., by $(A - \lambda I)\mathbf{x}_{i+1} = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$. The authors noticed that $A - \lambda I + \mathbf{x}_i p^T$ can be well-conditioned and the solution to (20)

is nothing but a constant multiple of the solution to $(A - \lambda I)\mathbf{x} = \mathbf{x}_i$, but provided reasons why (20) is not in their favor. In short, the main reason is because $A-\lambda I+\mathbf{x}_i\, p^T$ changes its form at every iteration making computations inefficient. However, with λ fixed, even though a limit exists for the inverse iteration, the convergence is still slow. Further discussions can be found in [9], and the references therein, in relation to the (shifted) inverse iteration and the Rayleigh quotient iteration.

Nevertheless, we would like to make full use of the nonsingular system (19) to further analyze quantitatively the error in eigenvector estimation regardless of the multiplicities of the corresponding eigenvalues helping understand the Newton's method (2). Indeed, Proposition 2 shows that the solution of (19) is a well-approximated eigenvector through a quantitative analysis for the limit, including the form of the limit, as an inexact λ approaches an exact one $\tilde{\lambda}$.

Firstly, Proposition 1 confirms that (19) is a nonsingular linear system whose unique solution is a corresponding eigenvector.

Proposition 1 *Suppose that $\tilde{\lambda}$ has multiplicity 1. With $\mathbf{x}_0 \in S^{N-1}$ chosen uniformly at random, Eq. (19) has a unique nonzero solution $\tilde{\mathbf{x}}$ that is an eigenvector of A corresponding to the eigenvalue $\tilde{\lambda}$ with probability 1.*

Proof Let $\mathbf{q}$ be a unit eigenvector of A corresponding to $\tilde{\lambda}$. Note that if we choose $\mathbf{x}_0 \in S^{N-1}$ uniformly at random, then we have $\mathbf{q}^T\mathbf{x}_0 \neq 0$ with probability 1. Moreover, if $\mathbf{q}^T\mathbf{x}_0 \neq 0$, then $(A - \tilde{\lambda}I + (\gamma + \tilde{\lambda})(\mathbf{x}_0\mathbf{x}_0^T))\mathbf{z} = 0$ implies $\mathbf{x}_0^T\mathbf{z} = 0$. Hence, we have $A\mathbf{z} = \tilde{\lambda}\mathbf{z}$. Since $\tilde{\lambda}$ has multiplicity 1, $\mathbf{z} = a\mathbf{q}$ for some $a \in \mathbb{R}$. In addition, since $0 = \mathbf{z}^T\mathbf{x}_0 = a\mathbf{q}^T\mathbf{x}_0$ and $\mathbf{q}^T\mathbf{x}_0 \neq 0$, we must have $a = 0$. That is, $\mathbf{z} = 0$. Hence, $A - \tilde{\lambda}I + (\gamma + \tilde{\lambda})(\mathbf{x}_0\mathbf{x}_0^T)$ is nonsingular and there exists a unique nonzero solution $\tilde{\mathbf{x}}$ to (19). By multiplying (19) by $\mathbf{q}^T$, we have $(\gamma + \tilde{\lambda})\mathbf{x}_0^T\tilde{\mathbf{x}} = \gamma$, which implies that $\tilde{\mathbf{x}}$ also satisfies

$$(A - \tilde{\lambda}I)\tilde{\mathbf{x}} = 0.$$

□

If the multiplicity of an eigenvalue $\tilde{\lambda}$ is greater than 1, then (19) becomes singular and Proposition 1 does not apply. However, when the multiplicity $m > 1$ is known, we can construct another nonsingular system as given in Corollary 2. In addition, Proposition 2 says that, regardless of the multiplicity, a good estimate of the eigenvalue guarantees a good estimate of a corresponding eigenvector through the nonsingular linear system (19).

Corollary 2 *Suppose that an eigenvalue $\tilde{\lambda}$ of A has multiplicity $m > 1$. We choose $\mathbf{x}_0, \ldots, \mathbf{x}_{m-1}$ uniformly at random from S^{N-1} and set an $N \times m$ matrix $X_0 = [\mathbf{x}_0 \ \ldots \mathbf{x}_{m-1}]$. With probability 1, the equation*

$$(A - \tilde{\lambda}I + (\gamma + \tilde{\lambda})X_0X_0^T)X = \gamma X_0$$

has a unique nonzero solution $\tilde{X} = [\tilde{\mathbf{x}}_0 \ \ldots \tilde{\mathbf{x}}_{m-1}]$, an $N \times m$ matrix, whose columns $\tilde{\mathbf{x}}_0, \ldots, \tilde{\mathbf{x}}_{m-1}$ constitute a basis for the eigenspace corresponding to the eigenvalue $\tilde{\lambda}$.

Proposition 2 *Let A be an $N \times N$ symmetric matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ and $\lambda_1 < \lambda_N$. Let $\{\mathbf{q}_1, \ldots, \mathbf{q}_N\}$ be an orthonormal basis for $\mathbb{R}^N$ consisting of eigenvectors of A such that each $(\lambda_i, \mathbf{q}_i)$, $i = 1, \ldots, N$, is an eigenpair of A. Let $\tilde{\lambda}$ be an eigenvalue λ_{k_0}, for some $1 \leq k_0 \leq N$, of A with multiplicity $m \geq 1$ so that $\tilde{\lambda} = \lambda_{k_0} = \cdots = \lambda_{k_0+m-1}$. Let $\gamma > 0$ be such that $\gamma \neq -\tilde{\lambda}$. Let $\mathbf{x}_0 \in S^{N-1}$ be such that*

$$\zeta_0 := \sqrt{\sum\nolimits_{j=0}^{m-1} |\mathbf{q}_{k_0+j}^T \mathbf{x}_0|^2} \in (0, 1).$$

Let $c_1 = \min(|\lambda_{k_0-1} - \tilde{\lambda}|, |\lambda_{k_0+m} - \tilde{\lambda}|) > 0$ and $c_2 = \max_{1 \leq j \leq N}(|\lambda_j - \tilde{\lambda}|) > 0$, and $d = \max_{1 \leq j \leq N}(|\gamma + \lambda_j|) > 0$ and $\omega = \min(\frac{|\gamma+\tilde{\lambda}|\zeta_0^2}{2d(1-\zeta_0^2)}, 1) > 0$. Let $\tilde{\mathbf{x}}$ be an eigenvector of A corresponding to $\tilde{\lambda}$ given by

$$\tilde{\mathbf{x}} = \frac{\gamma}{(\gamma + \tilde{\lambda})\zeta_0^2} \sum_{j=0}^{m-1} \left(\mathbf{q}_{k_0+j}^T \mathbf{x}_0\right) \mathbf{q}_{k_0+j}.$$

Then, there exists $0 < \epsilon < \min(\frac{c_1\omega}{2}, 1)$ such that $0 < |\lambda - \tilde{\lambda}| < \epsilon$ gives rise to a unique nonzero $\mathbf{x}_\lambda$ which satisfies

$$\left(A - \lambda I + (\gamma + \lambda)(\mathbf{x}_0 \mathbf{x}_0^T)\right) \mathbf{x}_\lambda = \gamma \mathbf{x}_0, \tag{21}$$

and

$$\left(\eta_1 \sqrt{1 - \zeta_0^2}\right) |\lambda - \tilde{\lambda}| < \|\mathbf{x}_\lambda - \tilde{\mathbf{x}}\| < \left(\eta_2 \sqrt{1 - \zeta_0^2}\right) |\lambda - \tilde{\lambda}| \tag{22}$$

for some $0 < \eta_1 < \eta_2 < \infty$ depending only on $\gamma, \mathbf{x}_0, \tilde{\lambda}$.

Proof It suffices to consider a diagonal matrix A with diagonal entries $\lambda_1 \leq \cdots \leq \lambda_N$. Let $\tilde{\lambda}$ be of multiplicity $m \geq 1$, i.e., $\tilde{\lambda} = \lambda_{k_0} = \cdots = \lambda_{k_0+m-1}$ for some $1 \leq k_0 \leq N - m + 1$. We choose $\mathbf{x}_0 = [x_{0,1}\ x_{0,2}\ \ldots\ x_{0,N}]^T \in S^{N-1}$ and set

$$\zeta_0 := \sqrt{\sum\nolimits_{j=0}^{m-1} |x_{0,k_0+j}|^2}. \tag{23}$$

Note that the condition $\zeta_0 \in (0, 1)$ holds with probability 1 when choosing $\mathbf{x}_0 \in S^{N-1}$ uniformly at random.

We also set $\tilde{\mathbf{x}} = [\tilde{x}_1\ \ldots\ \tilde{x}_N]^T \in \mathbb{R}^N$ by

$$\tilde{\mathbf{x}} = \frac{\gamma}{(\gamma + \tilde{\lambda})\zeta_0^2} \sum_{j=k_0}^{k_0+m-1} x_{0,j} \mathbf{e}_j, \quad \text{i.e. } \tilde{x}_i = \begin{cases} \left(\frac{\gamma}{\gamma+\tilde{\lambda}}\right) \frac{x_{0,i}}{\zeta_0^2}, & k_0 \leq i < k_0 + m, \\ 0, & \text{otherwise,} \end{cases}$$

and note that $\tilde{\mathbf{x}}$ is an eigenvector of A corresponding to $\tilde{\lambda}$.

Defining a polynomial $p(\lambda) = \det(A-\lambda I+(\gamma+\lambda)(\mathbf{x}_0\mathbf{x}_0^T))$ of degree N, we can see from Proposition 1 that $p(\tilde{\lambda}) \neq 0$ when $m = 1$. So, there exists $0 < \epsilon < \min(\frac{c_1\omega}{2}, 1)$ such that $p(\lambda) \neq 0$ for $|\lambda - \tilde{\lambda}| < \epsilon$. When $m > 1$, even though we have $p(\tilde{\lambda}) = 0$, we can still find $0 < \epsilon < \min(\frac{c_1\omega}{2}, 1)$ such that $p(\lambda) \neq 0$ for $0 < |\lambda - \tilde{\lambda}| < \epsilon$ due to the discreteness of the distinct zeros of p. Therefore, in both cases, there exists $0 < \epsilon < \min(\frac{c_1\omega}{2}, 1)$ such that $A - \lambda I + (\gamma + \lambda)(\mathbf{x}_0\mathbf{x}_0^T)$ is nonsingular, i.e., there exists a unique nonzero $\mathbf{x}_\lambda$ satisfying (21) for $0 < |\lambda - \tilde{\lambda}| < \epsilon$.

We now proceed to find an explicit representation of $\mathbf{x}_\lambda$ for $0 < |\lambda - \tilde{\lambda}| < \epsilon$. First of all, $\mathbf{x}_\lambda$ must satisfy

$$(A - \lambda I)\mathbf{x}_\lambda = \begin{bmatrix} (\lambda_1 - \lambda)x_{\lambda,1} \\ \vdots \\ (\lambda_N - \lambda)x_{\lambda,N} \end{bmatrix} = (\gamma - (\gamma + \lambda)(\mathbf{x}_0^T\mathbf{x}_\lambda)) \begin{bmatrix} x_{0,1} \\ \vdots \\ x_{0,N} \end{bmatrix},$$

i.e., $(A - \lambda I)\mathbf{x}_\lambda = \alpha_\lambda \mathbf{x}_0$ with $\alpha_\lambda = \gamma - (\gamma + \lambda)(\mathbf{x}_0^T\mathbf{x}_\lambda)$ and

$$\mathbf{x}_\lambda = (A - \lambda I)^{-1}(\alpha_\lambda \mathbf{x}_0) = \alpha_\lambda \begin{bmatrix} \frac{1}{\lambda_1-\lambda}x_{0,1} \\ \vdots \\ \frac{1}{\lambda_N-\lambda}x_{0,N} \end{bmatrix}.$$

Moreover, we can see that α_λ satisfies

$$\alpha_\lambda = \gamma - (\gamma + \lambda)(\mathbf{x}_0^T\mathbf{x}_\lambda) = \gamma - (\gamma + \lambda)(\mathbf{x}_0^T[(A - \lambda I)^{-1}(\alpha_\lambda \mathbf{x}_0)]),$$

that is,

$$\alpha_\lambda = \frac{\gamma}{1 + \frac{\gamma+\lambda}{\lambda_1-\lambda}x_{0,1}^2 + \cdots + \frac{\gamma+\lambda}{\lambda_N-\lambda}x_{0,N}^2} = \frac{\gamma}{\sum_{j=1}^N \left(\frac{\gamma+\lambda_j}{\lambda_j-\lambda}\right) x_{0,j}^2}. \tag{24}$$

Therefore, for $0 < |\lambda - \tilde{\lambda}| < \epsilon$, we have

$$\mathbf{x}_\lambda = \left(\frac{\gamma}{\sum_{j=1}^N \left(\frac{\gamma+\lambda_j}{\lambda_j-\lambda}\right) x_{0,j}^2}\right) \begin{bmatrix} \frac{1}{\lambda_1-\lambda}x_{0,1} \\ \vdots \\ \frac{1}{\lambda_N-\lambda}x_{0,N} \end{bmatrix}. \tag{25}$$

Knowing that

$$\left| \sum_{j \notin \{k_0,\ldots,k_0+m-1\}} \left(\frac{\gamma + \lambda_j}{\lambda_j - \lambda}\right) x_{0,j}^2 \right| < \frac{2d}{c_1}(1 - \zeta_0^2)$$

and that $0 < |\lambda - \tilde{\lambda}| < \epsilon$ implies $0 < |\lambda - \tilde{\lambda}| < \frac{c_1\omega}{2}$, i.e., $\frac{2d(1-\zeta_0^2)}{c_1} < \frac{1}{2}\left|\frac{\gamma+\tilde{\lambda}}{\tilde{\lambda}-\lambda}\right|\zeta_0^2$, we can easily see that $0 < |\lambda - \tilde{\lambda}| < \epsilon$ implies

$$\frac{1}{2}\left|\frac{\gamma+\tilde{\lambda}}{\tilde{\lambda}-\lambda}\right|\zeta_0^2 < \left|\sum_{j=1}^{N}\left(\frac{\gamma+\lambda_j}{\lambda_j-\lambda}\right)x_{0,j}^2\right| < \frac{3}{2}\left|\frac{\gamma+\tilde{\lambda}}{\tilde{\lambda}-\lambda}\right|\zeta_0^2. \tag{26}$$

Then, we can see from (25), (26) and $|\lambda_j - \lambda| > \frac{c_1}{2}$ for $j \notin \{k_0, \ldots, k_0+m-1\}$ that for $k_0 \le i < k_0+m$, i.e., $\lambda_i = \tilde{\lambda}$, and $0 < |\lambda - \tilde{\lambda}| < \epsilon$,

$$\begin{aligned}
|x_{\lambda,i} - \tilde{x}_i| &= \left|\frac{\frac{\gamma}{\lambda_i-\lambda}x_{0,i}}{\sum_{j=1}^{N}\left(\frac{\gamma+\lambda_j}{\lambda_j-\lambda}\right)x_{0,j}^2} - \frac{\gamma x_{0,i}}{(\gamma+\tilde{\lambda})\zeta_0^2}\right| \\
&= \left|\frac{\gamma x_{0,i}}{(\gamma+\tilde{\lambda})\zeta_0^2}\left(\frac{\sum_{j\notin\{k_0,\ldots,k_0+m-1\}}(\frac{\gamma+\lambda_j}{\lambda_j-\lambda})x_{0,j}^2}{\sum_{j=1}^{N}(\frac{\gamma+\lambda_j}{\lambda_j-\lambda})x_{0,j}^2}\right)\right| \\
&\le \frac{4\gamma|x_{0,i}|}{c_1(\gamma+\tilde{\lambda})^2\zeta_0^4}|\lambda-\tilde{\lambda}| \sum_{j\notin\{k_0,\ldots,k_0+m-1\}}|\gamma+\lambda_j|x_{0,j}^2 \\
&\le \frac{4d\gamma|x_{0,i}|(1-\zeta_0^2)}{c_1(\gamma+\tilde{\lambda})^2\zeta_0^4}|\lambda-\tilde{\lambda}|.
\end{aligned} \tag{27}$$

On the other hand, for $i < k_0$ or $i \ge k_0+m$, we have

$$\frac{4\gamma|x_{0,i}|}{3(c_1+c_2)|\gamma+\tilde{\lambda}|\zeta_0^2}|\lambda-\tilde{\lambda}| < |x_{\lambda,i}| < \frac{4\gamma|x_{0,i}|}{c_1|\gamma+\tilde{\lambda}|\zeta_0^2}|\lambda-\tilde{\lambda}|. \tag{28}$$

Combining (27) and (28), we obtain that

$$\begin{aligned}
&\left(\frac{4\gamma}{3(c_1+c_2)(\gamma+\tilde{\lambda})\zeta_0^2}\right)^2(1-\zeta_0^2)|\lambda-\tilde{\lambda}|^2 < \|\mathbf{x}_\lambda - \tilde{\mathbf{x}}\|^2 \\
&= \left(\frac{4d\gamma(1-\zeta_0^2)}{c_1(\gamma+\tilde{\lambda})^2\zeta_0^4}\right)^2\zeta_0^2|\lambda-\tilde{\lambda}|^2 + \left(\frac{4\gamma}{c_1(\gamma+\tilde{\lambda})\zeta_0^2}\right)^2(1-\zeta_0^2)|\lambda-\tilde{\lambda}|^2.
\end{aligned}$$

Therefore, we have

$$\left(\eta_1\sqrt{1-\zeta_0^2}\right)|\lambda-\tilde{\lambda}| < \|\mathbf{x}_\lambda - \tilde{\mathbf{x}}\| \le \left(\eta_2\sqrt{1-\zeta_0^2}\right)|\lambda-\tilde{\lambda}|$$

with

$$\eta_1 = \frac{4\gamma}{3(c_1+c_2)|\gamma+\tilde{\lambda}|\zeta_0^2} \text{ and } \eta_2 = \frac{4\gamma}{c_1|\gamma+\tilde{\lambda}|\zeta_0^2}\sqrt{\frac{(d\zeta_0)^2(1-\zeta_0^2)}{(\gamma+\tilde{\lambda})^2\zeta_0^4}+1}.$$

□

Proposition 2 makes it possible to explicitly estimate how far the $k+1^{st}$ iterate $\mathbf{x}_{k+1}$ by the Newton's method (2) is from the eigenspace $\text{eig}(\tilde{\lambda})$ in terms of the error in the eigenvalue estimate at the kth iterate, i.e., $|\lambda_k - \tilde{\lambda}|$. As a generalization of (2), if we consider two sequences $\{\lambda_k\}$ and $\{\mathbf{x}_k\}$ satisfying for each $k = 0, 1, 2, \ldots,$

$$\left[A - \lambda_k I + (\gamma + \lambda_k)\mathbf{y}_k \mathbf{y}_k^T\right]\mathbf{x}_{k+1} = \gamma \mathbf{y}_k, \tag{29}$$

with $\mathbf{y}_k = \frac{\mathbf{x}_k}{\|\mathbf{x}_k\|}$, and assume that $\{\lambda_k\}_{k=0}^{\infty}$ converges to an eigenvalue $\tilde{\lambda}$ of A and $\{\mathbf{x}_k\}_{k=0}^{\infty}$ converges to a corresponding eigenvector of A, then Proposition 2 says that there exist $\epsilon > 0$ depending only on the matrix A, and $K \in \mathbb{N}$ such that $k > K$ implies $0 < |\lambda_k - \tilde{\lambda}| < \epsilon$ and

$$\left(\eta_{1,k}\sqrt{1-\zeta_{0,k}^2}\right)|\lambda_k - \tilde{\lambda}| < \|\mathbf{x}_{k+1} - \tilde{\mathbf{x}}_k\| < \left(\eta_{2,k}\sqrt{1-\zeta_{0,k}^2}\right)|\lambda_k - \tilde{\lambda}|, \tag{30}$$

where $\tilde{\mathbf{x}}_k$ is an eigenvector of A corresponding to $\tilde{\lambda}$, which is the vector $\tilde{\mathbf{x}}$ in Proposition 2, and $\eta_{1,k}$, $\eta_{2,k}$, $\zeta_{0,k}$ are determined by $\tilde{\lambda}$ and $\mathbf{x}_k$ as in Proposition 2 with $\mathbf{x}_0$ replaced by $\mathbf{y}_k$. Therefore, due to $\zeta_{0,k} \to 1$ as $k \to \infty$ and $\sup_{k\in\mathbb{N}} \eta_{2,k} < \infty$, we can see that

$$\text{dist}(\mathbf{x}_{k+1}, \text{eig}(\tilde{\lambda})) = O\left(\left(\sqrt{1-\zeta_{0,k}^2}\right)|\lambda_k - \tilde{\lambda}|\right) = o(|\lambda_k - \tilde{\lambda}|), \tag{31}$$

where $\text{dist}(\mathbf{x}_{k+1}, \text{eig}(\tilde{\lambda}))$ is the distance from $\mathbf{x}_{k+1}$ to the eigenspace $\text{eig}(\tilde{\lambda})$. In this case, (31) says that $\mathbf{x}_{k+1}$ is much more accurate to an eigenvector than λ_k is to the corresponding eigenvalue $\tilde{\lambda}$ as $k \to \infty$.

2.3.2 Nonsymmetric diagonalizable matrices

Interestingly, we can also discuss the propositions in the previous section for nonsymmetric diagonalizable matrices within a prescribed error. Since finding an eigenvector of A corresponding to an eigenvalue λ is equivalent to finding a nonzero vector in the null space of $A - \lambda I$, we provide the following corollary. We can see that the loss of symmetry results in loose error estimation. More precisely, the extra factor $\sqrt{1-\zeta_0^2}$ in (22) is no longer present.

Corollary 3 *Let A be an $N \times N$ nonzero real diagonalizable matrix. Suppose that A has a nontrivial null space $\mathcal{N}(A)$. Let $\gamma > 0$.*

1. *If $\mathcal{N}(A)$ has dimension 1, then choosing $\mathbf{x}_0 \in S^{N-1}$ uniformly at random, we have that with probability 1,*

$$(A + \gamma \mathbf{x}_0 \mathbf{x}_0^T)\mathbf{x} = \gamma \mathbf{x}_0$$

has a unique nonzero solution $\tilde{\mathbf{x}}$ that spans $\mathcal{N}(A)$.

2. *Regardless of the dimension of* $\mathcal{N}(A)$*, choosing* $\mathbf{x}_0 \in S^{N-1}$ *uniformly at random, we can see, with probability 1, that for* $0 < |\lambda| \ll 1$*,*

$$(A - \lambda I + (\gamma + \lambda)\mathbf{x}_0\mathbf{x}_0^T)\mathbf{x} = \gamma\mathbf{x}_0 \tag{32}$$

has a unique nonzero solution $\mathbf{x}_\lambda$ *satisfying*

$$\eta_1|\lambda| < \|\mathbf{x}_\lambda - \tilde{\mathbf{x}}\| < \eta_2|\lambda|$$

for some $0 < \eta_1 < \eta_2 < \infty$ *and* $\tilde{\mathbf{x}} \in \mathcal{N}(A)$*, where* η_1, η_2 *and* $\tilde{\mathbf{x}}$ *are determined by* $\mathbf{x}_0$*.*

Proof The first part can be proven in the same way as we proved Proposition 1 with a choice of $\mathbf{x}_0 \in S^{N-1}$ satisfying $q^T\mathbf{x}_0 \neq 0$ and $\tilde{q}^T\mathbf{x}_0 \neq 0$, where q and $\tilde{q}$ are unit vectors spanning $\mathcal{N}(A)$ and $\mathcal{N}(A^T)$, respectively.

For the second part of the theorem, we will only provide a sketch of this proof due to its similarity to the symmetric case. We will also consider $\lambda > 0$ for simplicity.

Note that $A - \lambda I$ for $0 < \lambda < \min_{\lambda_j \neq 0}(|\lambda_j|)$, is invertible and that if we choose $\mathbf{x}_0 \in S^{N-1}$ uniformly at random, then $\mathbf{x}_0$ is not orthogonal to $\mathcal{N}(A)$ with probability 1. In fact, since $\mathbb{R}^N = \text{eigenspace}(0) \oplus V$ with $V = \bigoplus_{\lambda_j \neq 0} \text{eigenspace}(\lambda_j)$, we can represent $\mathbf{x}_0$ uniquely as $q_0 + r_0$, where $0 \neq q_0 \in \text{eigenspace}(0) = \mathcal{N}(A)$ and $0 \neq r_0 \in V$. We also note that for any $0 < \delta < \min_{\lambda_j \neq 0}(|\lambda_j|)$, there exists $K > 0$ such that

$$\sup_{\lambda \in [-\delta, \delta] \setminus \{0\}} \|(A - \lambda I)^{-1}\|_V < K,$$

where $\|(A - \lambda I)^{-1}\|_V$ is the norm of the restriction $(A - \lambda I)^{-1} : V \to V$. If we also consider A^{-1} as being restricted to V, then we have

$$\sup_{\lambda \in [-\delta, \delta]} \|(A - \lambda I)^{-1}\|_V < K. \tag{33}$$

So, we will fix $0 < \delta < \min_{\lambda_j \neq 0}(|\lambda_j|)$ and consider $(A - \lambda I)^{-1}$ as being restricted to V for $\lambda \in [-\delta, \delta]$.

With such an $\mathbf{x}_0$, we will see if we can solve for $\mathbf{x}$

$$(A - \lambda I)\mathbf{x} = (\gamma\mathbf{x}_0 - (\gamma + \lambda)\mathbf{x}_0\mathbf{x}_0^T\mathbf{x}) = \alpha_\lambda\mathbf{x}_0,$$

with $\alpha_\lambda = \gamma - (\gamma + \lambda)\mathbf{x}_0^T\mathbf{x}$. As in Proposition 2, we can see that α_λ must satisfy

$$\alpha_\lambda = \gamma - \alpha_\lambda(\gamma + \lambda)\mathbf{x}_0^T(A - \lambda I)^{-1}\mathbf{x}_0.$$

Together with (33), it is not difficult to see that for $0 < \lambda < \delta$,

$$\frac{\|r_0\|\lambda}{\|A\| + \delta} < \left\| \left(\frac{1}{\lambda}A - I\right)^{-1} r_0 \right\| < K\|r_0\|\lambda.$$

Furthermore, we can see that there exists $0 < \lambda_0 < \delta$ such that for all $0 < \lambda < \lambda_0$, $\alpha_\lambda := \frac{\gamma}{(1+(\gamma+\lambda)\mathbf{x}_0^T(A-\lambda I)^{-1}\mathbf{x}_0)}$ exists and the unique nonzero solution $\mathbf{x}_\lambda$ is represented as

$$\mathbf{x}_\lambda = \alpha_\lambda (A-\lambda I)^{-1}\mathbf{x}_0 = \frac{\gamma(A-\lambda I)^{-1}\mathbf{x}_0}{(1+(\gamma+\lambda)\mathbf{x}_0^T(A-\lambda I)^{-1}\mathbf{x}_0)}$$

and

$$\left|\frac{\alpha_\lambda}{\lambda} + \frac{1}{(\mathbf{x}_0^T q_0)}\right| < \omega\lambda,$$

where $\omega > 0$ depends only on $\mathbf{x}_0$ and δ.

Finally, we set $\tilde{\mathbf{x}} = \frac{1}{(\mathbf{x}_0^T q_0)} q_0$. For $0 < \lambda < \lambda_0$, since $(A-\lambda I)^{-1} q_0 = -\lambda q_0$, we can see that

$$\begin{aligned}\|\mathbf{x}_\lambda - \tilde{\mathbf{x}}\| &= \left\| \alpha_\lambda (A-\lambda I)^{-1}(q_0 + r_0) - \frac{1}{(\mathbf{x}_0^T q_0)} q_0 \right\| \\ &\le \left\| \left(\frac{\alpha_\lambda}{\lambda} + \frac{1}{(\mathbf{x}_0^T q_0)} \right) q_0 \right\| + \|\alpha_\lambda (A-\lambda I)^{-1} r_0\| < \eta_2 \lambda,\end{aligned}$$

where $\eta_2 = \omega\|q_0\| + K\|r_0\|(\frac{1}{|\mathbf{x}_0^T q_0|} + \omega\delta)$. On the hand, we let $\mathbf{z}_0 \in V$ be such that $A\mathbf{z}_0 = r_0$ and set $\mathbf{y}_0 := \frac{\mathbf{z}_0}{\|\mathbf{z}_0\|}$. Due to the continuity of $(A-\lambda I)^{-1} r_0 : [-\delta, \delta] \to V$ in λ, it is not difficult to see that there exists $0 < \lambda_1 \le \lambda_0$ such that $0 < \lambda < \lambda_1$ implies

$$|\mathbf{y}_0^T (A-\lambda I)^{-1} r_0| > \frac{\|\mathbf{z}_0\|}{2} \quad \text{and} \quad |\alpha_\lambda| > \frac{\lambda}{2|\mathbf{x}_0^T q_0|}.$$

i.e.,

$$\|\mathbf{x}_\lambda - \tilde{\mathbf{x}}\| \ge |\mathbf{y}_0^T(\mathbf{x}_\lambda - \tilde{\mathbf{x}})| = |\alpha_\lambda \mathbf{y}_0^T (A-\lambda I)^{-1} r_0| > \eta_1 \lambda,$$

with $\eta_1 = \frac{\|\mathbf{z}_0\|}{4|\mathbf{x}_0^T q_0|} > 0$. Therefore, $0 < \lambda < \lambda_1$ implies

$$\eta_1 \lambda < \|\mathbf{x}_\lambda - \tilde{\mathbf{x}}\| < \eta_2 \lambda.$$

□

3 Numerical experiments

3.1 Generalized eigenvalue problem $A\mathbf{x} = \lambda B\mathbf{x}$

Since a generalized eigenvalue problem was solved in [5] as an application, we also provide how to apply the Newton's method (2) to solve $A\mathbf{x} = \lambda B\mathbf{x}$ via

$$\min_{\mathbf{x}\in\mathbb{R}^N} F_{A,B}(\mathbf{x}), \tag{34}$$

where A, B are symmetric and B is positive definite and $F_{A,B}$ is given by

$$F_{A,B}(\mathbf{x}) = \frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle + \frac{\gamma}{2}\langle \mathbf{x}, B\mathbf{x}\rangle - \gamma\sqrt{\langle \mathbf{x}, B\mathbf{x}\rangle},$$

with some $\gamma > 0$ making $A+\gamma B$ positive definite. In fact, as was seen in Sect. 2.2, we can see that the constant $\gamma \in \mathbb{R}\backslash\{0\}$ determines the set of critical points that are computable using $F_{A,B}$. That is, given $\gamma \in \mathbb{R}\backslash\{0\}$, only those eigenvectors corresponding to the eigenvalues λ of the pencil $A - \lambda B$ satisfying $\gamma(\gamma + \lambda) > 0$ can be critical points of $F_{A,B}$ since a critical point $\mathbf{x}_*$ satisfies

$$A\mathbf{x}_* + \gamma\left(1 - \frac{1}{\|\mathbf{x}_*\|_B}\right)B\mathbf{x}_* = 0 \quad \Rightarrow \quad \|\mathbf{x}_*\|_B = \frac{\gamma}{\gamma+\lambda} > 0.$$

The Newton's method applied to $F_{A,B}$ with $\|\mathbf{x}\|_B := \sqrt{\langle \mathbf{x}, B\mathbf{x}\rangle}$ generates a sequence $\{\mathbf{x}_k\}$ satisfying

$$\left[\frac{1}{\gamma}A + \left(1 - \frac{1}{\|\mathbf{x}_k\|_B}\right)B + \frac{1}{\|\mathbf{x}_k\|_B}\left(\frac{B\mathbf{x}_k}{\|\mathbf{x}_k\|_B}\right)\left(\frac{B\mathbf{x}_k}{\|\mathbf{x}_k\|_B}\right)^T\right]\mathbf{x}_{k+1} = \frac{B\mathbf{x}_k}{\|\mathbf{x}_k\|_B}, \tag{35}$$

which can be rewritten as

$$\left[(A - \lambda_k B) + (\gamma + \lambda_k)\left(\frac{B\mathbf{x}_k}{\|\mathbf{x}_k\|_B}\right)\left(\frac{B\mathbf{x}_k}{\|\mathbf{x}_k\|_B}\right)^T\right]\mathbf{x}_{k+1} = \gamma\frac{B\mathbf{x}_k}{\|\mathbf{x}_k\|_B} \tag{36}$$

with $\lambda_k = \gamma(\frac{1}{\|\mathbf{x}_k\|_B} - 1)$. Then, depending on how to update λ_k in (36), either by $\lambda_k = \gamma(\frac{1}{\|\mathbf{x}_k\|_B} - 1)$ or by $\lambda_k = \frac{\mathbf{x}_k^T A\mathbf{x}_k}{\mathbf{x}_k^T B\mathbf{x}_k}$ or maybe by some other clever ways, we end up with either (35) derived by the Newton's method (2) or a type of the Rayleigh quotient iteration or maybe another clever scheme.

In this experiment, we would like to compare the two eigenvalue update rules, one by $\lambda_k = \gamma(\frac{1}{\|\mathbf{x}_k\|_B} - 1)$ and another by $\lambda_k = \frac{\mathbf{x}_k^T A\mathbf{x}_k}{\mathbf{x}_k^T B\mathbf{x}_k}$ to confirm that the update rule $\lambda_k = \gamma(\frac{1}{\|\mathbf{x}_k\|_B} - 1)$ obtained from the explicit minimization framework (34) behaves differently from that by $\lambda_k = \frac{\mathbf{x}_k^T A\mathbf{x}_k}{\mathbf{x}_k^T B\mathbf{x}_k}$.

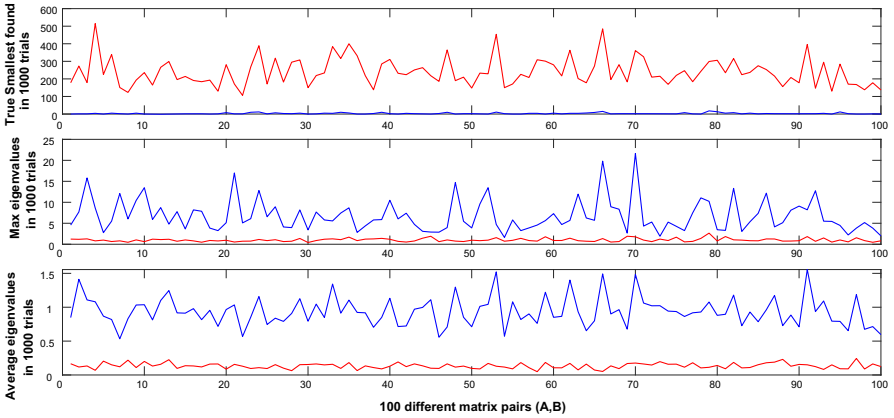

Fig. 1 100 trials with random positive definite symmetric pairs (A_i, B_i), $i = 1, 2, \ldots, 100$. A_i's and B_i's are of size 10×10. For each pair (A_i, B_i), we test the two eigenvalue update rules 1000 times starting from random vectors. The red and blue colors indicate the two update rules: $\lambda = \gamma(\frac{1}{\|\mathbf{x}\|_B} - 1)$ in red and $\lambda = \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}}$ in blue. Top row: For each $i = 1, \ldots, 100$, we plot the number of times each update rule finds the true smallest eigenvalue λ_{min}. We counted the number of eigenvalues computed whose difference from λ_{min} is less than 10^{-13}. Middle row: For each $i = 1, \ldots, 100$, we plot the maximum eigenvalue among the 1000 times trials. Bottom row: For each $i = 1, \ldots, 100$, we plot the mean eigenvalue among the 1000 times trials (color figure online)

When testing the two different eigenvalue update rules above with randomly selected positive definite matrices A, B, we observe numerically that the update rule $\lambda = \gamma(\frac{1}{\|\mathbf{x}\|_B} - 1)$ tends to find small eigenvalues much more often than the update rule $\lambda = \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}}$. In fact, $\lambda = \gamma(\frac{1}{\|\mathbf{x}\|_B} - 1)$ finds the true smallest eigenvalues quite well starting from a random initial vector as shown in Fig. 1.

In Fig. 2, we performed the same experiment as in Fig. 1 with different sizes, i.e., A_i's and B_i's are of size 50×50 shown in blue and of size 100×100 shown in red. Solid lines indicate results using the update rule $\lambda = \gamma(\frac{1}{\|\mathbf{x}\|_B} - 1)$ and lines with circular dots and crosses indicate results using the other update rule $\lambda = \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}}$.

From the discussion about (29) as a generalization of the Newton's method (2), the convergence of $\{\lambda_k\}$ to an eigenvalue $\tilde{\lambda}$ of A as well as that of $\{\mathbf{x}_k\}$ imply not only $\mathbf{y}_k^T \mathbf{y}_{k+1} \to 1$ and $\|A\mathbf{x}_{k+1} - \lambda_k \mathbf{x}_{k+1}\| \to 0$, but also another explicit connection between $\mathbf{y}_k^T \mathbf{y}_{k+1}$ and $\|A\mathbf{x}_{k+1} - \lambda_k \mathbf{x}_{k+1}\|$ at the $k + 1^{st}$ iterate $\mathbf{x}_{k+1}$, that is,

$$\|A\mathbf{x}_{k+1} - \lambda_k \mathbf{x}_{k+1}\| = \left| \gamma \left(1 - \left(1 + \frac{\lambda_k}{\gamma} \right) (\mathbf{y}_k^T \mathbf{x}_{k+1}) \right) \right|.$$

That is, the error $\|A\mathbf{x}_{k+1} - \lambda_k \mathbf{x}_{k+1}\|$ at the $k+1^{st}$ iterate $\mathbf{x}_{k+1}$ can be measured exactly using the angle $\mathbf{y}_k^T \mathbf{y}_{k+1}$, which allows for consideration of a new stopping criterion

$$\left| 1 - \left(1 + \frac{\lambda_k}{\gamma} \right) (\mathbf{y}_k^T \mathbf{x}_{k+1}) \right| \sim 0.$$

We used this stopping criterion for the experiments above.

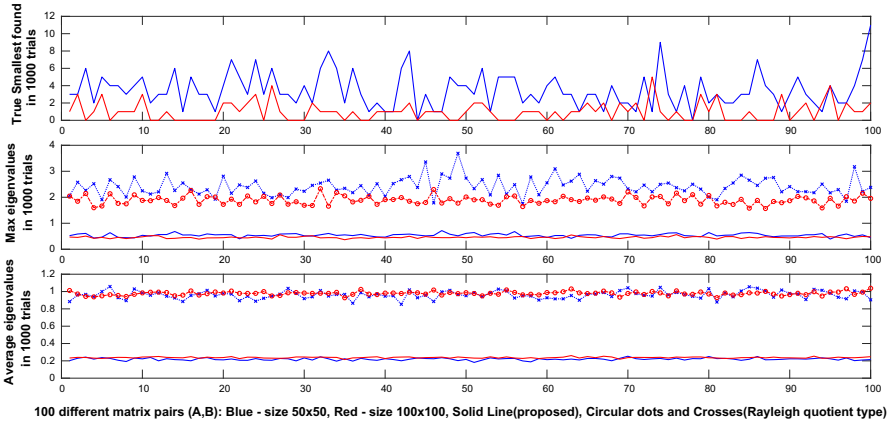

Fig. 2 The same experiments as in Fig. 1 with different matrix sizes. The blue and red colors indicate the two different sizes: 50×50 in blue and 100×100 in red. Solid lines are from the update rule $\lambda = \gamma(\frac{1}{\|\mathbf{x}\|_B} - 1)$ and the lines with circular dots or crosses are from the update rule $\lambda = \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}}$. Top row: We counted the number of times that the true smallest eigenvalues were computed among the 1000 times trials for each random positive definite pair $(A_i, B_i), i = 1, \ldots, 100$. Only the counts using the update rule $\lambda = \gamma(\frac{1}{\|\mathbf{x}\|_B} - 1)$ are shown because the update rule $\lambda = \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T B\mathbf{x}}$ never found the smallest eigenvalues. Middle row: The maximum eigenvalue among the 1000 times trials for each (A_i, B_i). Bottom row: The mean eigenvalue among the 1000 times trials for each (A_i, B_i) (color figure online)

3.2 Trust-region subproblem via a generalized eigenvalue problem

The same idea discussed in this work applies to solve

$$\min_{\|\mathbf{p}\|_B \le \Delta} \frac{1}{2}\langle \mathbf{p}, A\mathbf{p}\rangle + \langle g, \mathbf{p}\rangle, \tag{37}$$

which arises in trust region problems. For example, the authors of [7] proposed a generalized eigenvalue problem to solve the trust-region subproblem (37) where $A \in \mathbf{Sym}_N(\mathbb{R})$, $B \in \mathbf{Sym}_{N,p}(\mathbb{R})$ and g is an $N \times 1$ vector. When the minimum is attained at $\mathbf{p}_*$ with $\|\mathbf{p}_*\|_B < \Delta$, we can see that $\mathbf{p}_*$ solves $A\mathbf{p}_* + g = 0$.

In this section, we pay attention to the case where the minimum is attained on the boundary, i.e., at some $\mathbf{p}_*$ with $\|\mathbf{p}_*\|_B = \Delta$. By the optimality condition, the authors of [7] realized that there must be $\lambda_* \ge 0$ such that $A + \lambda_* B$ is positive semidefinite and

$$(A + \lambda_* B)\mathbf{p}_* + g = 0, \ \|\mathbf{p}_*\|_B = \Delta. \tag{38}$$

and solved

$$\begin{bmatrix} A & -\frac{gg^T}{\Delta^2} \\ -B & A \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = -\lambda \begin{bmatrix} B & O \\ O & B \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \Leftrightarrow \begin{bmatrix} A + \lambda B & -\frac{gg^T}{\Delta^2} \\ -B & A + \lambda B \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = O \tag{39}$$

for the largest real eigenvalue λ and its corresponding eigenvector $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2)$ and proved that λ_* for (38) is $\lambda_* = \text{argmax}\{\Re(\lambda) : \lambda \in \mathbb{C} \text{ solves } (39)\} \in \mathbb{R}$, and $\lambda_* \geq \mu_N$, where μ_N is the largest eigenvalue of the pencil $A+\mu B$. After solving (39), the authors check whether or not the hard case arises, which happens if $\langle g, \mathbf{y}_2\rangle = 0$, or equivalently, if $\lambda_* = \mu_N$. Note that if $\langle g, \mathbf{y}_2\rangle \neq 0$, then $\mathbf{p}_* = -\frac{\Delta^2}{\langle g, \mathbf{y}_2\rangle}\mathbf{y}_1$ satisfies (38). Hence, when $\langle g, \mathbf{y}_2\rangle = 0$, Sect. 4 in [7] provides an additional procedure to find $\mathbf{p}_*$ for (38) and its analysis to resolve the hard case estimating the accuracy of the solution of (39) using the known accuracy given in [10]. This additional procedure is unavoidable due to the round-off errors and [7] proposes further to estimate a basis for $\mathcal{N}(A + \lambda_* B)$, the null space of $A + \lambda_* B$, to find a solution in the hard case.

It turns out that our proposed minimization framework via the Newton's method can resolve the hard case in a very simple way without computing more eigenvalues and gaps between them as described in [7]. This is simply because the norm $\|\mathbf{x}_k\|$ carries the information about an eigenvalue when minimizing the functional (1). Indeed, after estimating λ_* in (39), we propose to minimize the functional

$$F(\mathbf{x}) = \left[\frac{1}{2}\langle \mathbf{x}, (A + \lambda_* B)\mathbf{x}\rangle + \langle g, \mathbf{x}\rangle\right] + \frac{\gamma}{2}\langle \mathbf{x}, B\mathbf{x}\rangle - \gamma\Delta\sqrt{\langle \mathbf{x}, B\mathbf{x}\rangle} \tag{40}$$

with $\gamma > 0$. Since $\lambda_* \geq \mu_N$ implies that $\frac{1}{2}\langle \mathbf{x}, (A + \lambda_* B)\mathbf{x}\rangle + \langle g, \mathbf{x}\rangle$ is convex in $\mathbf{x}$ and due to $\{\mathbf{x} : (A + \lambda_* B)\mathbf{x} + g = 0,\ \|\mathbf{x}\|_B = \Delta\} \neq \emptyset$ as given in (38), we can see that the set of all global minimizers of F is

$$\{\mathbf{x} : (A + \lambda_* B)\mathbf{x} + g = 0,\ \|\mathbf{x}\|_B = \Delta\}$$

by observing that $\mathbf{x}_* \in \{\mathbf{x} : (A + \lambda_* B)\mathbf{x} + g = 0,\ \|\mathbf{x}\|_B = \Delta\}$ if and only if

$$\begin{aligned} F(\mathbf{x}_*) &\geq \min_{\mathbf{x}} F(\mathbf{x}) \\ &= \min_{\mathbf{x}} \left[\left(\frac{1}{2}\langle \mathbf{x}, (A + \lambda_* B)\mathbf{x}\rangle + \langle g, \mathbf{x}\rangle\right) + \frac{\gamma}{2}\langle \mathbf{x}, B\mathbf{x}\rangle - \gamma\Delta\sqrt{\langle \mathbf{x}, B\mathbf{x}\rangle}\right] \\ &\geq \min_{\mathbf{x}} \left[\frac{1}{2}\langle \mathbf{x}, (A + \lambda_* B)\mathbf{x}\rangle + \langle g, \mathbf{x}\rangle\right] + \min_{\mathbf{x}} \left[\frac{\gamma}{2}\langle \mathbf{x}, B\mathbf{x}\rangle - \gamma\Delta\sqrt{\langle \mathbf{x}, B\mathbf{x}\rangle}\right] \\ &= \left[\frac{1}{2}\langle \mathbf{x}_*, (A + \lambda_* B)\mathbf{x}_*\rangle + \langle g, \mathbf{x}_*\rangle\right] - \frac{\gamma\Delta^2}{2} = F(\mathbf{x}_*). \end{aligned}$$

Especially, with $\gamma := \lambda_*$, a global minimizer $\mathbf{x}_* \in \{\mathbf{x} : (A + \lambda_* B)\mathbf{x} + g = 0,\ \|\mathbf{x}\|_B = \Delta\}$ of F satisfies

$$\begin{aligned} F(\mathbf{x}_*) &= \frac{1}{2}\langle \mathbf{x}_*, A\mathbf{x}_*\rangle + \langle g, \mathbf{x}_*\rangle = \min_{\mathbf{x}} F(\mathbf{x}) \\ &= \min_{\mathbf{x}} \left[\frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle + \langle g, \mathbf{x}\rangle + \lambda_*\sqrt{\langle \mathbf{x}, B\mathbf{x}\rangle}\left(\sqrt{\langle \mathbf{x}, B\mathbf{x}\rangle} - \Delta\right)\right] \\ &\leq \min_{\|\mathbf{x}\|_B \leq \Delta} \left[\frac{1}{2}\langle \mathbf{x}, A\mathbf{x}\rangle + \langle g, \mathbf{x}\rangle\right] \leq \frac{1}{2}\langle \mathbf{x}_*, A\mathbf{x}_*\rangle + \langle g, \mathbf{x}_*\rangle. \end{aligned}$$

That is, the global minimizers of F in (40) with $\gamma := \lambda_*$ solve (37). The Newton's method applied to (40) reads

$$\left[(A+\lambda_* B)+\gamma\left(1-\frac{\Delta}{\|\mathbf{x}_k\|_B}\right)B+\frac{\gamma\Delta}{\|\mathbf{x}_k\|_B}\mathbf{y}_k\mathbf{y}_k^T\right]\mathbf{x}_{k+1}=(\gamma\Delta)\mathbf{y}_k-g, \tag{41}$$

where $\mathbf{y}_k=\frac{B\mathbf{x}_k}{\|\mathbf{x}_k\|_B}$. We can further replace $\|\mathbf{x}_k\|_B$ in (40) by $\min(\|\mathbf{x}_k\|_B, \Delta+\epsilon)$ for $\epsilon \ll 1$. Therefore, whether or not the hard case in solving (39) arises, with $\gamma=\lambda_*$ estimated from (39), the global minimizers of (40) always provide solutions to (37).

More precisely, when $A+\lambda_* B$ is positive definite, a global minimizer $\mathbf{x}_*$ of (40) is unique. In the hard case, if $\Delta=\min\{\|\mathbf{x}\|_B : (A+\lambda_* B)\mathbf{x}+g=0\}$, we also have a unique global minimizer of (40). By interpreting (41) as the system

$$\begin{cases}\left((A+\lambda_* B)+\gamma\left(1-\frac{\Delta}{\|\mathbf{x}_k\|_B}\right)B+\frac{\gamma\Delta}{\|\mathbf{x}_k\|_B}\mathbf{y}_k\mathbf{y}_k^T\right)\mathbf{z}_{k+1}=(\gamma\Delta)\mathbf{y}_k,\\ \left((A+\lambda_* B)+\gamma\left(1-\frac{\Delta}{\|\mathbf{x}_k\|_B}\right)B+\frac{\gamma\Delta}{\|\mathbf{x}_k\|_B}\mathbf{y}_k\mathbf{y}_k^T\right)\mathbf{w}_{k+1}=-g,\\ \mathbf{x}_{k+1}=\mathbf{z}_{k+1}+\mathbf{w}_{k+1},\end{cases} \tag{42}$$

if $\Delta > \min\{\|\mathbf{x}\|_B : (A+\lambda_* B)\mathbf{x}+g=0\}$ in the hard case, as $\{\|\mathbf{x}_k\|_B\}_{k\in\mathbb{N}}$ converges to Δ, it is unlikely that $\mathbf{x}_k$ is orthogonal to $\mathcal{N}(A+\lambda_* B)$ with respect to $\langle\cdot,\cdot\rangle_B$. We can see from the discussion in Sect. 2.3.1 that for any $\mathbf{x}_k$ not orthogonal to $\mathcal{N}(A+\lambda_* B)$ with respect to $\langle\cdot,\cdot\rangle_B$, $\mathbf{z}_{k+1}$ is close to $\mathcal{N}(A+\lambda_* B)$ as much as $\|\mathbf{x}_k\|_B$ is close to Δ. In addition, $\mathbf{w}_{k+1}$ becomes orthogonal to $\mathbf{x}_k$ with respect to $\langle\cdot,\cdot\rangle_B$ as $\|\mathbf{x}_k\|_B \to \Delta$. This allows us to understand (41) from the viewpoint of (42) in the sense that $\{\mathbf{w}_k\}$ converges to satisfy $(A+\lambda_* B)\mathbf{w}_k+g\simeq 0$ and $\mathbf{z}_k$ compensates for the deviation from the norm Δ by approaching the null space $\mathcal{N}(A+\lambda_* B)$ to minimize its influence on the equation $(A+\lambda_* B)\mathbf{w}_k+g\simeq 0$, so that we eventually have $\|\mathbf{x}_*\|_B=\|\mathbf{w}_*+\mathbf{z}_*\|_B=\Delta$ with $(A+\lambda_* B)\mathbf{w}_*+g=0$ and $\mathbf{z}_*\in\mathcal{N}(A+\lambda_* B)$ and $\langle\mathbf{x}_*,\mathbf{w}_*\rangle_B=0$. This is numerically confirmed in Fig. 3.

Even though, both our approach and the additional procedure given in [7] compute a null space compensation $\mathbf{z}_*$ in the hard case, our approach implicitly encodes the null space correction term $\mathbf{z}_*$ in $\mathbf{x}_*$, whereas the additional step in [7] selects the null space correction term $\mathbf{z}_*$ manually using a basis for $\mathcal{N}(A+\lambda_* B)$. In Fig. 4, we repeat the same experiment shown in Fig. 3 with different null space dimensions of $A+\lambda_* B$. That is, we consider

$$\Lambda_k=\mathrm{diag}(-1,-1,\ldots,-1,k+1,k+2,\ldots,n)$$

where Λ_k has the first k entry values equal to -1. For each $k=2,\ldots,10$, we choose an orthogonal matrix Q randomly and set $A_k=Q\Lambda_k Q^T$. We use $B=I$ and $g_k=Qg_{(0,k+1)}$, where $g_{(0,k+1)}=-3\alpha\Delta\mathbf{e}_{k+1}$, which is the same hard case as given in Fig. 3. Note that for each $k=2,\ldots,10$, we have $\lambda_*=1$ and $\dim(\mathcal{N}(A_k+\lambda_* B))=k$ and g_k is orthogonal to $\mathcal{N}(A_k+\lambda_* B)$.

Since our approach and [7] can directly deal with $B\neq I$ unlike other methods such as FRW, RSS compared in [7], we tried an experiment with relatively dense symmetric

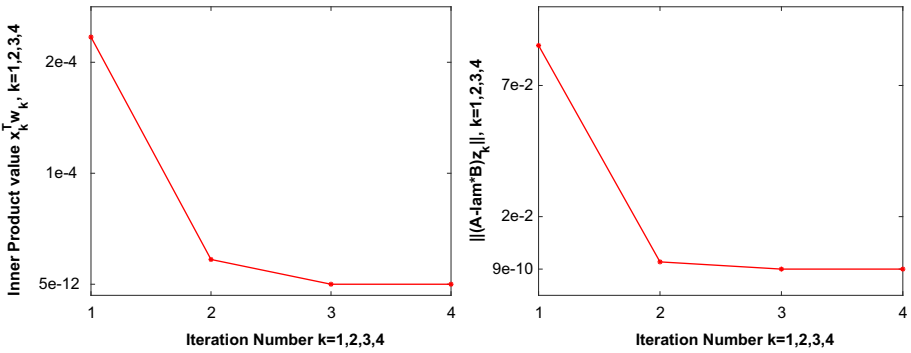

Fig. 3 This is the hard case in [7] with the known exact solution for $A = Q\Lambda Q^T$, $B = I$, $g = Qg_0$, where Q is a randomly chosen orthogonal matrix and $\Lambda = \text{diag}(-1, 2, \ldots, n)$ and $g_0 = -3\alpha\Delta\mathbf{e}_2$. Here, $n = 10^3$ and $\Delta = 1$ and $\alpha = 10^{-2}$ and $\mathbf{e}_2 = [0\ 1\ 0\ \ldots 0]^T \in \mathbb{R}^n$. The stopping criterion for (41) when $n \geq 10^3$ was set to be $\|\nabla F(\mathbf{x})\| < 10^{-9}$. In fact, it stopped at the 4th iterate $\mathbf{x}_4$ with $\|\nabla F(\mathbf{x}_4)\| \simeq 5 \times 10^{-13}$. Left: $\langle \mathbf{x}_k, \mathbf{w}_k \rangle = \mathbf{x}_k^T \mathbf{w}_k$ for $k = 1, 2, 3, 4$, Right: $\|(A + \lambda_* B)\mathbf{z}_k\|$, for $k = 1, 2, 3, 4$. This experiment confirms our understanding of (41) via the system (42)

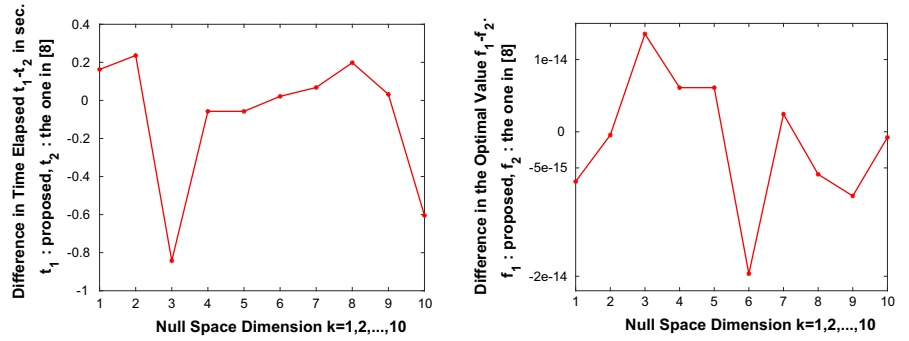

Fig. 4 An extension of the experiment shown in Fig. 3 when $\mathcal{N}(A+\lambda_* B)$ has dimension k for $k = 1, \ldots, 10$. The horizontal axis represents the dimension k of $\mathcal{N}(A + \lambda_* B)$. Left: difference in time elapsed, i.e., $t_1(k) - t_2(k)$, where t_1 is from our method and t_2 is from [7]. We can see that solving (41) is no slower in general than the additional step in [7]. For all $k = 1, \ldots, 10$, (41) was stopped at only the 4th iterate $\mathbf{x}_k$ satisfying the tolerance as shown in Fig. 3. Right: difference in the optimal value, i.e., $f_1(k) - f_2(k)$, where f_1 is from our method and f_2 is from [7]. Our method in the hard case even with $\dim(\mathcal{N}(A + \lambda_* B)) > 1$ is shown to be reliable

matrices A, B shown in Fig. 5. In MATLAB, using $n = 10^3$, we generate two $n \times n$ symmetric matrices A and B with approximately $\frac{n^2}{2}$ nonzero entries, which can be done by

$$A = \text{sprandsym}(n, 0.5), \quad B = \text{sprandsym}(n, 0.5).$$

Then, we replace B by $\frac{1}{\sqrt{n}}(I + B^T B)$ to impose positive definiteness on B. To consider the hard case, we estimate the largest real eigenvalue μ_N of the pencil $A + \lambda B$ and set $g = (A + \mu_N B)g_0$ with a random vector $g_0 \in \mathbb{R}^n$. Since this random vector g_0 does not belong to $\mathcal{N}(A + \mu_N B)$ with probability 1, g is orthogonal to $\mathcal{N}(A + \mu_N B)$, i.e.,

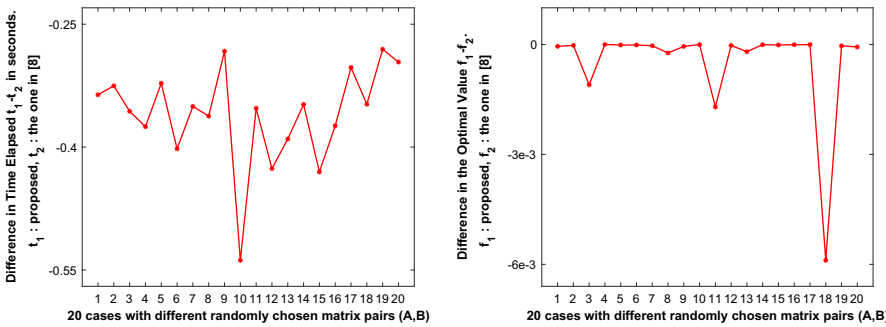

Fig. 5 20 hard case trials with relatively dense randomly generated $n \times n$ symmetric matrices A, B for $n = 10^3$. Note that B is replaced by $\frac{1}{\sqrt{n}}(I + B^T B)$ to satisfy positive definiteness and g is chosen to bring about the hard case. Left: difference in time elapsed, i.e., $t_1 - t_2$, where t_1 is from our approach and t_2 is from [7]. In this relatively dense scenario, during the 20 trials, we observed that the average time spent for our approach is 0.7 s, and the average time spent for the one in [7] is 1 s. Right: difference in the optimal values, i.e., $f_1 - f_2$, where f_1 is from our approach and f_2 is from [7]. It seems that our approach is better in accuracy in general, however, the difference in the optimal values could be negligible because the optimal values in all 20 trials were on the order of 10^8, which could mean the three noticeable trials at 3rd, 11th, 18th are due to system round-off errors

$\mathbf{z}^T g = 0$ for all $\mathbf{z} \in \mathcal{N}(A + \mu_N B)$. With A, B, g prepared as described, we solve (37) with $\Delta = 10^3$ to further enforce the hard case just as [7] described. We checked that A, B, g described as above give rise to the hard case, i.e., $\lambda_* = \mu_N$, and applied both our approach and the additional step in [7]. We generated 20 such sets of (A, B, g) at random as described and measured the time elapsed as well as the computed optimal value (37) for both our approach and the one in [7] in Fig. 5. This experiment in Fig. 5 shows that our approach using (41) performs slightly better and faster in general than the additional step in [7] with relatively dense A, B in the hard case.

4 Conclusion

In this paper, we analyzed the Newton's method for real eigenvalue problems. We provided local and global convergence results of a generated sequence $\{\mathbf{x}_k\}$ by the Newton's method and an interesting equivalence in convergence between $\{\mathbf{x}_k\}$ and $\{\|\mathbf{x}_k\|\}$.

Besides its faster convergence, our quantitative analysis shows that even with an approximate eigenvalue, the nonsingular linear system arising from the Newton's method guarantees that the unique solution is close to an exact eigenvector as much the estimated eigenvalue is close to an exact eigenvalue as possible, which allows for explicit error estimations revealing comparability in the amount of errors in eigenvalue and eigenvector estimations. Our analysis applies to both symmetric and nonsymmetric cases and various numerical experiments confirms the nature and efficiency of our proposed method.

As was pointed out in [5], our proposed Newton's method approach has a few advantages over conventional methods. For example, when computing eigenfuctions

of self-adjoint operators on a given domain, since we can formulate the problem as an unconstrained problem, we can compute them numerically easily on various shapes of the domain such as spheres, tori, etc., with efficiency. Moreover, noting from the example of the trust region subproblem in [7], where the main advantage of our approach is that our method provides a simpler analysis as well as a simpler algorithm and is easy to understand, we expect that our approach can find other interesting applications in various fields of research, where a minor modification of our framework computes desired solutions efficiently.

Acknowledgements This work was supported partially by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (NRF-2014R1A1A1002667) and supported partially by U-K Brand Future-core Research Fund (1.180016.01) of UNIST(Ulsan National Institute of Science & Technology). Moreover, I would like to thank my friend and colleague, Ernie Esser, Ph.D., with whom I had fruitful discussions on various topics in image processing. I could not have initiated this project without his inspiration.

References

1. Kolluri, R., Shewchuk, J., O'Brien, J.: Spectral surface reconstruction from noisy point clouds. In: Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, no. 11 in SGP '04, pp. 11–21. ACM, New York (2004)
2. Belkin, M., Sun, J., Wang, Y.: Constructing Laplace operator from point clouds in $\mathbb{R}^d$. In: Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '09, pp. 1031–1040. SIAM, Philadelphia, PA, USA (2009)
3. Lai, R., Liang, J., Zhao, H.K.: A local mesh method for solving PDEs on point clouds. Inverse Problems Imaging **7**(3), 737–755 (2013)
4. Lai, R., Liang, J., Zhao, H.: A local mesh method for solving PDEs on point clouds. Inverse Probl. Imaging **7**(3), 737–755 (2016)
5. Kim, Y.: An unconstrained global optimization framework for real symmetric eigenvalue problems, submitted
6. Tapia, R.A., Dennis, J.E., Schafermeyer, J.P.: Inverse, shifted inverse, and Rayleigh quotient iteration as Newton's method. SIAM Rev. **60**(1), 3–55 (2018)
7. Adachi, S., Iwata, S., Nakatsukasa, Y., Takeda, A.: Solving the trust-region subproblem by a generalized eigenvalue problem. SIAM J. Optim. **27**(1), 269–291 (2017)
8. Peters, G., Wilkinson, J.H.: Inverse iteration, ill-conditioned equations and Newton's method. SIAM Rev. **21**(3), 339–360 (1979)
9. Ipsen, I.C.F.: Computing an eigenvector with inverse iteration. SIAM Rev. **39**(2), 254–291 (1997)
10. Stewart, G.W., Sun, J.-G.: Matrix Perturbation Theory. Computer Science and Scientific Computing. Academic Press, Boston (1990)