A Nonconvex Unconstrained Method for Eigenvalue Problems and A Nonsingular System for Eigenvector Estimation

Yunho Kim*

November 28, 2016

Abstract

We propose a nonconvex unconstrained minimization problem for eigenvalue problems. In this framework, given a symmetric matrix A, it turns out that any nonzero critical point is an eigenvector of A and any local minimizer is a global minimizer, an eigenvector of A corresponding to the smallest eigenvalue. Unlike the conventional way of estimating an eigenvector with a given eigenvalue, our proposed problem estimates the corresponding eigenvalue from the estimated eigenvector of A. We analyze two algorithms to solve the proposed problem, one of which guarantees convergence to a global minimizer. This makes our method applicable even to a singular matrix A, where conventional methods avoid finding an eigenvector corresponding to the eigenvalue 0. The other tries to find a critical point that may not be a global minimizer, however, is faster and gives rise to a nonsingular linear system, as a byproduct, whose unique solution is an eigenvector of A, which we believe is a new approach to eigenvector estimation. Our proposed method applies to nonsquare matrices, complex hermitian matrices and also to infinite dimensional cases such as self-adjoint elliptic operators, some of which will be presented as specific applications.

1 Introduction

Given an $N \times N$ matrix A, the eigenvalue problem of our interest is to find an eigenvalue and its corresponding eigenvector of A, that is, to solve $Ax = \lambda x$ for x and λ. This is one of the most fundamental problems in mathematics with applications to all other fields of science. Especially, one may be interested in estimating the largest and the smallest eigenvalues. If A is symmetric and positive definite, then finding the smallest eigenvalue of A is the same as finding the largest eigenvalue of A^{-1}. However, finding the smallest eigenvalue of A involves solving systems of linear equations of type $Ax = b$. For example, the inverse power method to estimate the smallest eigenvalue involves solving $Ax = b$ as well as matrix multiplication, whereas the power method to estimate the largest eigenvalue involves only matrix multiplication. From these observations, we asked ourselves a question, "Is it possible to compute the smallest eigenvalue of A through only basic matrix operations such as multiplication and addition, without solving $Ax = b$?"

Our interest of estimating the smallest eigenvalue and its corresponding eigenvector extends to the following infinite dimensional application, as well. On a compact manifold $\mathcal{M}$,

*Department of Mathematical Sciences, UNIST, South Korea. Email: yunhokim@unist.ac.kr

eigenvalues of the Laplacian $-\Delta$ reveals important structures of $\mathcal{M}$, which makes understanding the eigenvalues of $-\Delta$ on $\mathcal{M}$ very important. This has interesting applications. For example, in image processing there are a few interesting works (e.g. [7], [3], [9]) to distinguish reconstructed objects from point cloud data by evaluating $-\Delta$ on the surfaces of the objects. We can even consider general self-adjoint linear elliptic operators and find their eigenvalues and eigenfunctions. With these theoretical and numerical points of view in mind, our main discussion will be concentrated on finding the smallest eigenvalue and a corresponding eigenvector of a nonzero symmetric matrix, which leads us to begin with the following well-known constrained problem: given a symmetric and positive definite matrix A,

$$\min_{x\in\mathbb{R}^N} \langle x, Ax\rangle \quad \text{subject to} \quad \|x\|^2 = 1. \tag{1}$$

There have been a large number of works to solve (1) by the name of inverse iteration methods. In particular, we would like to mention the work [4] by J.E. Dennis and R.A. Tapia, which surveys historical developments of inverse and shifted inverse iterations and of Rayleigh quotient iteration, and which approaches the listed methods from the viewpoint of the Newton's method. The unconstrained version of (1) analyzed in [4] is

$$\min_{x\in\mathbb{R}^N} \langle x, Ax\rangle + \frac{\gamma}{2}(\|x\|^2 - 1)^2. \tag{2}$$

The authors of [4] explained why the inverse and the shifted inverse Rayleigh quotient iterations are fast and effective by showing the equivalence between the inverse Rayleigh quotient iteration and the Newton's method for (2), and also between the shifted inverse Rayleigh quotient iteration and the Newton's method for the shifted version of (2), when the given matrix A is symmetric and invertible. We refer to [4], and references therein, any interested reader in the developments of the inverse and shifted inverse Rayleigh quotient iteration methods.

There are, however, a few disadvantageous features of the functional in (2) that we paid attention to. First of all, [4] considered only nonsingular matrices for (2) just as all other conventional methods do. Second of all, in the simplest case when A is symmetric and positive definite, if $0 < \gamma < \lambda_1$, then the zero vector is the only critical point of the functional in (2) and, even if $\gamma \geq \lambda_1$, the critical points of the functional in (2) are only the eigenvalues of A less than γ. Hence, our goal is of two folds: 1) extending existing theories to singular matrices, and 2) removing the additional limitations imposed by the parameter γ in (2). Noting that the functional in (2) contains the term

$$(\|x\|^2 - 1)^2 = (\|x\| + 1)^2(\|x\| - 1)^2,$$

we believed that the factor $(\|x\| + 1)^2$, even though this is convex, is not desirable because the factor tries to reduce the norm $\|x\|$ to 0 during minimization. Therefore, our analysis begins without this factor.

The rest of this manuscript is organized as follows. In Section 2, we define and analyze our proposed minimization problem in the case of a nonzero real symmetric matrix A. In the analysis, we prove that our proposed method does not possess the disadvantageous features mentioned above. The same analysis applies to complex hermitian matrices, as well. Moreover, we analyze two minimization algorithms, the Gradient Descent method and the Newton's method, to find an eigenvalue and its corresponding eigenvector, where the former ensures convergence to a global minimizer and the latter ensures fast convergence to critical points. In fact, as for the Newton's method, we show that the a generated sequence $\{x_k\}$

converges if and only if the sequence of their norms $\{\|x_n\|\}$ converges and that the limit of $\{\|x_k\|\}$ determines the corresponding eigenvalue. What is more interesting is that we can provide a nonsingular system of linear equations whose unique solution is an eigenvector of A when an exact eigenvalue of A is known. The same is true for general diagonalizable matrices. This is unusual because an exact eigenvalue gives rise to a singular system of linear equations, which is why conventional methods assume nonsingularity of A and generate a sequence of approximate eigenvalues and justify that the sequence converges to an exact eigenvalue. In Section 3, we provide a few applications such as generalized eigenvalue problems and an infinite dimensional example for self-adjoint linear elliptic operators, which shows applicability of our method to various types of eigenvalue and eigenvector (or eigenfunction) estimation, not only in the finite dimensional setting, but also in the infinite dimensional setting.

2 Eigenvalue problem on a finite dimensional space

A general setting is the following. Let $\mathbb{F}$ be either $\mathbb{R}$ or $\mathbb{C}$. Given a matrix $A \in \mathbf{M}_{M\times N}(\mathbb{F})$, we define $F_A : \mathbb{F}^N \to \mathbb{F}$ by

$$F_A(x) = \frac{1}{2}\langle x, Bx\rangle + \frac{\gamma}{2}\|x\|^2 - \gamma\|x\|, \tag{3}$$

where B is

$$B = \begin{cases} A, & \text{if } A \text{ is real symmetric or complex hermitian,} \\ A^*A, & \text{if } A \text{ is neither real symmetric nor complex hermitian,} \end{cases}$$

and $\langle x, y\rangle = y^*x$ and $\|x\| = \sqrt{\langle x, x\rangle}$. As usual, $x \in \mathbb{F}^N$, $Ax \in \mathbb{F}^M$ are considered as $N \times 1$ and $M \times 1$ column vectors, respectively, and $|A| = \sqrt{A^*A}$. We use the convention that $x^* = x^T$ and $|A| = \sqrt{A^TA}$ if $\mathbb{F} = \mathbb{R}$.

Our discussion on the real symmetric case extends to the complex hermitian case without alteration. Moreover, the same discussion applies to the singular value decomposition of a nonsymmetric matrix A with $B = A^*A$ in the definition (3). Hence, and we will only present the real symmetric case, where the minimization problem that we will consider is

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

which is equivalent to

$$\min_{x\in\mathbb{R}^N} \frac{1}{2}\langle x, Ax\rangle + \frac{\gamma}{2}(\|x\| - 1)^2.$$

2.1 The Real Symmetric Case

To begin with, we will consider the case that $A \in \mathbf{M}_N(\mathbb{R})$ is nonzero and symmetric, where the functional that we minimize in (3) has the form

$$F_A(x) = \frac{1}{2}\langle x, Ax\rangle + \frac{\gamma}{2}\|x\|^2 - \gamma\|x\| \tag{4}$$

Lemma 1. *Let $\lambda_1 \in \mathbb{R}$ be the smallest eigenvalue of A. For $\gamma > \max(0, -\lambda_1)$, the set of nonzero critical points of F_A is*

$$\left\{x \in \mathbb{R}^n : x \text{ is an eigenvector of } A \text{ corresponding to an eigenvalue } \lambda \text{ with norm } \frac{\gamma}{\gamma+\lambda}\right\},$$

and

$$\min_{x\in\mathbb{R}^n} F_A(x) = -\frac{\gamma^2}{2(\gamma+\lambda_1)}.$$

Proof. Let $x_0 \neq 0$ be a critical point of F_A. For any unit vector $\theta \in \mathbb{R}^N$, we note that

$$\lim_{\epsilon\to 0} \frac{F_A(x_0+\epsilon\theta) - F_A(x_0)}{\epsilon} = \langle Ax_0, \theta\rangle + \gamma\langle x_0, \theta\rangle - \frac{\gamma}{\|x_0\|}\langle x_0, \theta\rangle = 0.$$

This implies

$$\Big[A + \gamma\Big(1 - \frac{1}{\|x_0\|}\Big)I\Big]x_0 = 0 \quad \Leftrightarrow \quad Ax_0 = \gamma\Big(\frac{1}{\|x_0\|} - 1\Big)x_0.$$

Hence, $x_0 \neq 0$ is a critical point of F_A if and only if x_0 is an eigenvector of A corresponding to the eigenvalue

$$\lambda_0 = \gamma\Big(\frac{1}{\|x_0\|} - 1\Big).$$

Moreover,

$$\|x_0\| = \frac{\gamma}{\gamma+\lambda_0} \quad \text{and} \quad F_A(x_0) = -\frac{\gamma}{2}\|x_0\| = -\frac{\gamma^2}{2(\gamma+\lambda_0)}$$

and the set of nonzero critical points of F_A is

$$\Big\{x \in \mathbb{R}^N : x \text{ is an eigenvector of } A \text{ corresponding to an eigenvalue } \lambda \text{ with norm } \frac{\gamma}{\gamma+\lambda}\Big\}.$$

It is easy to see that a global minimizer x_* of F_A exists and is an eigenvector of A corresponding to an eigenvalue λ_* with norm $\frac{\gamma}{\gamma+\lambda_*}$. Therefore, $F_A(x_*) = \min_{x\in\mathbb{R}^N} F_A(x)$ implies that $\lambda_* = \lambda_1$ and

$$\min_{x\in\mathbb{R}^N} F_A(x) = -\frac{\gamma^2}{2(\gamma+\lambda_1)} < 0.$$

□

Throughout the section, we will assume $\gamma > \max(0, -\lambda_1)$, where λ_1 is the smallest eigenvalue of A if no condition on γ is stated.

Theorem 1. *Any local minimizer x_* of F_A is a global minimizer.*

Proof. First of all, 0 is not a local minimizer of F_A because for any nonzero $x_0 \in \mathbb{R}$, we have

$$\lim_{t\to 0^+} \frac{F_A(tx_0) - F_A(0)}{t\|x_0\|} = -\gamma < 0.$$

Suppose that $x_* \neq 0$ is a local minimizer of F_A. Lemma 1 says that x_* is an eigenvector of A corresponding to an eigenvalue λ_* with norm $\frac{\gamma}{\gamma+\lambda_*}$ and that

$$F_A(x_*) = -\frac{\gamma^2}{2(\gamma+\lambda_*)} \geq -\frac{\gamma^2}{2(\gamma+\lambda_1)} = \min_{x\in\mathbb{R}} F_A(x),$$

where λ_1 is the smallest eigenvalue of A. We may diagonalize A such that $A = Q\Lambda Q^T$, where Λ is a diagonal matrix with nondecreasing diagonal entries $\lambda_1 \leq \cdots \leq \lambda_N$ and Q is an

orthogonal matrix having $\frac{x_*}{\|x_*\|}$ as the j^{th} column for some j implying $\lambda_* = \lambda_j$. Then, with $y = Q^T x$, (3) becomes

$$F_A(x) = \ F_A(Qy) = F_\Lambda(y). \tag{5}$$

For $k = 1, 2, \ldots, N$, we set $\mathbf{e}_k$ to be the k^{th} column of the identity matrix $I \in \mathbf{M}_N(\mathbb{R})$. Then, x_* being a local minimizer of F_A is equivalent to $\frac{\gamma}{\gamma+\lambda_j}\mathbf{e}_j$ being a local minimizer of F_Λ. Note that

$$F_\Lambda(y) = \frac{1}{2}\langle \Lambda y, y\rangle + \frac{\gamma}{2}\|y\|^2 - \gamma\|y\| = \Big(\frac{1}{2}\sum_{k=1}^{N}\Big(\lambda_k + \gamma\Big)y_k^2\Big) - \gamma\|y\| \tag{6}$$

and that $\frac{\gamma}{\gamma+\lambda_1}\mathbf{e}_1$ is a global minimizer of F_Λ. If we consider F_Λ on the subspace spanned by $\{\mathbf{e}_1, \mathbf{e}_j\}$ by defining $H : \mathbb{R}^2 \to \mathbb{R}$ by

$$H(y_1, y_2) = F_\Lambda(y_1\mathbf{e}_1 + y_2\mathbf{e}_j) = \frac{1}{2}\Big(\lambda_1 + \gamma\Big)y_1^2 + \frac{1}{2}\Big(\lambda_j + \gamma\Big)y_2^2 - \gamma\sqrt{y_1^2 + y_2^2},$$

then $\nabla^2 H$ at $\Big(0, \frac{\gamma}{\gamma+\lambda_j}\Big)$ must be positive semidefinite, that is,

$$\det\begin{bmatrix} (\gamma+\lambda_1) - \frac{\gamma y_2^2}{\sqrt{y_1^2+y_2^2}^3} & \frac{\gamma y_1 y_2}{\sqrt{y_1^2+y_2^2}^3} \\ \frac{\gamma y_1 y_2}{\sqrt{y_1^2+y_2^2}^3} & (\gamma+\lambda_j) - \frac{\gamma y_1^2}{\sqrt{y_1^2+y_2^2}^3} \end{bmatrix} = (\lambda_1 - \lambda_j)(\gamma + \lambda_j) \geq 0$$

at $\Big(0, \frac{\gamma}{\gamma+\lambda_j}\Big)$. This implies $\lambda_1 \geq \lambda_j = \lambda_*$, i.e., $\lambda_* = \lambda_1$ and

$$F_A(x_*) = -\frac{\gamma^2}{2(\gamma+\lambda_*)} = \min_{x\in\mathbb{R}^N} F_A(x).$$

Therefore, any local minimizer x_* of F_A is a global minimizer of F_A. □

In addition, we may be able to find all the eigenvalues and their corresponding eigenvectors of A.

Theorem 2. *Let A be symmetric with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_k$ be the first k eigenvectors of A corresponding to the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_k$. We consider the following problem*

$$\min_{x\in\mathbb{R}^N} F_A(x) \quad \textit{subject to} \quad \langle x, \mathbf{x}_i\rangle = 0, \ i = 1, 2, \ldots, k. \tag{7}$$

Then, any local minimizer of (7) is a global minimizer corresponding to the eigenvalue λ_{k+1} with norm $\frac{\gamma}{\gamma+\lambda_{k+1}}$.

Proof. With a diagonalization QDQ^T of A, where $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and the first k columns of Q are $\mathbf{x}_1, \ldots, \mathbf{x}_k$, and $y = Q^T x$, we have that $F_A(x) = F_D(y)$ and $\|x\| = \|y\|$, so (7) is equivalent to

$$\min\{F_D(y) : y = (0, \ldots, 0, y_{k+1}, \ldots, y_n) \in \mathbb{R}^N\}. \tag{8}$$

Let D_k be the last $(n-k)\times(n-k)$ block of D, i.e., $D_k = \mathrm{diag}(\lambda_{k+1}, \ldots, \lambda_N)$. Then, (8) is equivalent to

$$\min_{z\in\mathbb{R}^{n-k}} F_{D_k}(z).$$

Theorem 1 applies to F_{D_k} and we are done. □

2.1.1 Algorithm 1 : The Gradient Descent Method

We have seen in the previous section that our model does not possess a local minimizer that is not a global minimizer. To find a global minimizer of our model, we will consider and analyze the following gradient descent method with stepsize $\alpha_k > 0$:

$$x_{k+1} = x_k - \alpha_k \nabla F_A(x_k) = x_k - \alpha_k \Big(Ax_k + \gamma x_k - \frac{\gamma}{\|x_k\|} x_k \Big), \quad \text{with } x_0 \neq 0. \tag{9}$$

Let $\{q_1, \ldots, q_N\}$ be an orthonormal basis for $\mathbb{R}^N$ consisting of unit eigenvectors of A corresponding to the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$, respectively. If $x_k = \mu_{k,1} q_1 + \cdots + \mu_{k,N} q_N$, then

$$x_{k+1} = \mu_{k+1,1} q_1 + \cdots + \mu_{k+1,N} q_N = \sum_{i=1}^{N} \mu_{k,i} \Big[1 - \alpha_k \Big(\lambda_i + \gamma - \frac{\gamma}{\|x_k\|} \Big) \Big] q_i$$

implies

$$\mu_{k+1,i} = \mu_{k,i} \Big[1 - \alpha_k \Big(\lambda_i + \gamma - \frac{\gamma}{\|x_k\|} \Big) \Big], \quad i = 1, \ldots, N,$$

resulting in

$$x_{k+1} = \sum_{i=1}^{N} \mu_{0,i} \Pi_{j=0}^{k} \Big[1 - \alpha_j \Big(\lambda_i + \gamma - \frac{\gamma}{\|x_j\|} \Big) \Big] q_i. \tag{10}$$

For simplicity, we assume a fixed stepsize $0 < \alpha_k = \alpha < \frac{1}{\lambda_N + \gamma}$ for all $k \in \mathbb{N}$.

One very distinctive feature of our model is that (9) can find a global minimizer, i.e., an eigenvector of A corresponding to the smallest eigenvalue, even when A is singular. Theorem 10 below is given in a general form.

Theorem 3. *A sequence $\{x_k\}$ generated by (9) with $x_0 = \mu_{0,1} q_1 + \cdots + \mu_{0,N} q_N \neq 0$ converges to a critical point x_* of F_A which is an eigenvector of A corresponding the eigenvalue λ_l with*

$$\|x_*\| = \frac{\gamma}{\gamma + \lambda_l},$$

where $l = \min\{j \in \{1, \ldots, N\} | \mu_{0,j} \neq 0\}$. More precisely, x_* is

$$\Big(\frac{\gamma}{\gamma + \lambda_l} \Big) \Big(\frac{1}{\sqrt{\sum_{m:\lambda_m = \lambda_l} \mu_{0,m}^2}} \Big) \sum_{m:\lambda_m = \lambda_l} \mu_{0,m} q_m.$$

Proof. For any $x \neq 0$, we get

$$\Big\langle \nabla F_A(x), \frac{x}{\|x\|} \Big\rangle = \Big(\frac{\langle Ax, x \rangle}{\|x\|^2} + \gamma \Big) \|x\| - \gamma = (\lambda + \gamma) \|x\| - \gamma, \tag{11}$$

where $\lambda_1 \leq \lambda \leq \lambda_N$. Then, since $\alpha < \frac{1}{\lambda_N + \gamma}$, we obtain

$$\|x_{k+1}\|^2 \geq \Big| \Big\langle x_k - \alpha \nabla F_A(x_k), \frac{x_k}{\|x_k\|} \Big\rangle \Big|^2 \geq \big(\|x_k\| - \alpha(\lambda + \gamma) \|x_k\| + \alpha\gamma \big)^2 > (\alpha\gamma)^2,$$

which implies $\|x_k\| > \alpha\gamma$ for all $k \in \mathbb{N}$. Note that for $x \neq 0$,

$$\nabla^2 F_A(x) = A + \gamma I - \frac{\gamma}{\|x\|} \Big[I - \Big(\frac{x}{\|x\|} \Big) \Big(\frac{x}{\|x\|} \Big)^T \Big].$$

Then, it is easy to see that $\|\nabla^2 F_A(x)\| \leq \max(\lambda_N + \gamma, \frac{1}{\alpha}) \leq \frac{1}{\alpha}$ for $\|x\| \geq \alpha\gamma$. Moreover, the line segment connecting x_k and x_{k+1} for $k \in \mathbb{N}$ lies entirely in $\{x \in \mathbb{R}^N : \|x\| \geq \alpha\gamma\}$. To see this, we take $0 < t < 1$ and observe that for $k \in \mathbb{N}$,

$$\|x_k - t\alpha\nabla F(x_k)\|^2 \geq \big(\|x_k\| - t\alpha(\lambda+\gamma)\|x_k\| + t\alpha\gamma\big)^2 > (\alpha\gamma)^2(1 - t\alpha(\lambda+\gamma) + t)^2 > (\alpha\gamma)^2.$$

Then, for each $k \geq 1$, we have that

$$\begin{aligned} F_A(x_{k+1}) &\leq F_A(x_k) + \langle \nabla F_A(x_k), x_{k+1} - x_k\rangle + \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 \\ &= F_A(x_k) - \frac{\alpha}{2}\|\nabla F_A(x_k)\|^2, \end{aligned}$$

which implies

$$\|\nabla F_A(x_k)\|^2 \leq \frac{2}{\alpha}(F_A(x_k) - F_A(x_{k+1})),$$

and that for any $K \geq 1$,

$$\frac{1}{\alpha^2}\sum_{k=1}^{K}\|x_{k+1} - x_k\|^2 = \sum_{k=1}^{K}\|\nabla F_A(x_k)\|^2 \leq \frac{2}{\alpha}(F_A(x_1) - (\min_{x\in\mathbb{R}^N} F_A(x))). \tag{12}$$

Since F_A is coercive and $F_A(x_k) \leq F_A(x_1) < \infty$ for all $k \geq 1$, $\{x_k\}$ must be a bounded sequence in

$$\{x \in \mathbb{R}^N : \|x\| \geq \alpha\gamma\}.$$

Choosing a subsequence, $\{x_{k_n}\}$, converging to x_*, (12) implies that $\nabla F_A(x_*) = 0$, i.e., according to Lemma 1, x_* is an eigenvector of A corresponding to an eigenvalue λ_i for some $1 \leq i \leq N$ with norm

$$\|x_*\| = \frac{\gamma}{\gamma + \lambda_i}.$$

Knowing that $\{F_A(x_k)\}$ is a decreasing and bounded sequence, we can easily derive that any subsequential limit $\tilde{x}$ of $\{x_k\}$ satisfies $F_A(\tilde{x}) = F_A(x_*)$ and $A\tilde{x} = \lambda_i\tilde{x}$ with norm

$$\|\tilde{x}\| = \frac{\gamma}{\gamma + \lambda_i}.$$

Hence, we can conclude that

$$\lim_{k\to\infty}\|x_k\| = \frac{\gamma}{\gamma + \lambda_i}.$$

Now, we set $l = \min\{j \in \{1, \ldots, N\} | \mu_{0,j} \neq 0\}$. Then, $\lambda_i \geq \lambda_l$. Suppose $\lambda_i > \lambda_l$. We note that for all $k \in \mathbb{N}$,

$$1 - \alpha\Big(\lambda_l + \gamma - \frac{\gamma}{\|x_k\|}\Big) > 1 - \frac{\lambda_l + \gamma}{\lambda_N + \gamma} + \frac{\alpha\gamma}{\|x_k\|} \geq \frac{\alpha\gamma}{\|x_k\|} > 0$$

and that as $k \to \infty$,

$$1 - \alpha\Big(\lambda_l + \gamma - \frac{\gamma}{\|x_k\|}\Big) \to 1 - \alpha(\lambda_l - \lambda_i) > 1.$$

This implies from (10) that

$$\mu_{0,l}\Pi_{j=0}^{k}\Big[1 - \alpha\Big(\lambda_l + \gamma - \frac{\gamma}{\|x_j\|}\Big)\Big] \to \infty \quad \text{as} \quad k \to \infty.$$

This is a contradiction because $\{x_k\}$ is a bounded sequence. Therefore, $\lambda_i = \lambda_l$.

Moreover, for $\lambda_p > \lambda_l$, we have

$$1 - \alpha\Big(\lambda_p + \gamma - \frac{\gamma}{\|x_k\|}\Big) \to 1 - \alpha(\lambda_p - \lambda_l) \in (0,1) \ \text{ as } \ k \to \infty,$$

implying

$$\mu_{0,p}\Pi_{j=0}^{k}\Big[1 - \alpha\Big(\lambda_p + \gamma - \frac{\gamma}{\|x_j\|}\Big)\Big] \to 0 \ \text{ as } \ k \to \infty.$$

Hence, referring to (10), we can see that

$$x_{k+1} - \Big(\sum_{m:\lambda_m=\lambda_l} \mu_{0,m} q_m\Big)\Pi_{j=0}^{k}\Big[1 - \alpha\Big(\lambda_l + \gamma - \frac{\gamma}{\|x_j\|}\Big)\Big] \to 0 \ \text{ as } \ k \to \infty.$$

Moreover, convergence of the norm $\|x_k\|$ implies that

$$\Pi_{j=0}^{\infty}\Big[1 - \alpha\Big(\lambda_l + \gamma - \frac{\gamma}{\|x_j\|}\Big)\Big] = \frac{\gamma}{\gamma + \lambda_l}\Big(\frac{1}{\sqrt{\sum_{m:\lambda_m=\lambda_l} \mu_{0,m}^2}}\Big).$$

and x_k converges to

$$\Big(\sum_{m:\lambda_m=\lambda_l} \mu_{0,m} q_m\Big)\Pi_{j=0}^{\infty}\Big[1 - \alpha\Big(\lambda_l + \gamma - \frac{\gamma}{\|x_j\|}\Big)\Big].$$

□

If λ_1 has multiplicity $p \geq 1$, i.e., $\lambda_1 = \cdots = \lambda_p < \lambda_{p+1}$, and if $|\mu_{0,1}| + \cdots + |\mu_{0,p}| > 0$, then the theorem says that the sequence generated by the gradient descent method converges to an eigenvector of A corresponding to the smallest eigenvalue λ_1 with norm $\frac{\gamma}{\gamma+\lambda_1}$. In fact, if we begin the algorithm with a randomly chosen initial guess x_0, then it is easy to see that it converges to a global minimizer with probability 1. Therefore, regardless of the definiteness of A and of the multiplicity of the smallest eigenvalue of A, the gradient descent method (9) finds a global minimizer.

When compared with the inverse power method, the two have the same assumption for convergence. However, our approach only requires matrix multiplications, whereas the inverse power method needs to solve a linear system at every iteration.

2.1.2 Strongly convex F_A around a global minimum

It is easy to see that F_A is strongly convex near a global minimum point x_* since $\nabla^2 F_A(x_*)$ is positive definite. In this section, we would like to see how much F_A could be strongly convex, and how large such a strongly convex region near a global minimum point could be, which will give us some idea about when the Newton's method, if ever applied, will converge.

Theorem 4. *Suppose that A is symmetric with the smallest eigenvalue λ_1 of multiplicity 1 and that $\gamma > \max(0, -\lambda_1)$. Let F_A attain the global minimum value at x_*. Let $\eta = \min(2(\gamma + \lambda_1), \lambda_2 - \lambda_1)$. Then, F_A is $\frac{\eta}{4}$-strongly convex in the region*

$$\Big\{x : \|x - x_*\| < \min\Big(\frac{\gamma(\lambda_2 - \lambda_1)}{4(\gamma + \lambda_2)(\gamma + \lambda_1)}, \frac{\gamma}{5\gamma + 4\lambda_2 + \lambda_1}\Big)\Big\},$$

where $\lambda_2 > \lambda_1$ is the second smallest eigenvalue of A.

Proof. For simplicity, we may assume $\lambda_1 \geq 0$ since the proof below will not change when $\lambda_1 < 0$ with $\gamma + \lambda_1 > 0$. Let $k \geq 2$. For some $0 < \epsilon < \frac{\|x_*\|}{k}$, we suppose that x satisfies $\|x - x_*\| < \epsilon$. Then, we have

$$\frac{\gamma(k-1)}{(\gamma+\lambda_1)k} = \|x_*\|\Big(1 - \frac{1}{k}\Big) \leq \|x\| \leq \|x_*\|\Big(1 + \frac{1}{k}\Big) = \frac{\gamma(k+1)}{(\gamma+\lambda_1)k}$$

and

$$\Big\langle \frac{x}{\|x\|}, \frac{x_*}{\|x_*\|} \Big\rangle > \frac{1}{2}\Big(\frac{\|x\|}{\|x_*\|} + \frac{\|x_*\|}{\|x\|}\Big) - \frac{\epsilon^2}{2\|x\|\|x_*\|} \geq 1 - \frac{1}{2k(k-1)} \geq 1 - \frac{1}{k^2}.$$

Noting that

$$\nabla^2 F(x) = A + \gamma\Big(1 - \frac{1}{\|x\|}\Big)I + \frac{\gamma}{\|x\|^3} x x^T,$$

we will estimate $\langle z, \nabla^2 F_A(x) z\rangle$ for $z \in \mathbb{R}^N$ with $\|z\| = 1$. Let $\{(x_i, \lambda_i) : i = 1, \ldots, N\}$ be a set of eigenpairs of A such that $\{x_1, \ldots, x_n\}$ with $x_1 = \frac{x_*}{\|x_*\|}$ is an orthonormal basis for $\mathbb{R}^n$ and $0 \leq \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N$. Let $V = \{cx_* : c \in \mathbb{R}\}$ and $W = \text{span}\{x_i : i = 2, \ldots, N\}$. Then, for any $z \in \mathbb{R}^N$, there exists a unique decomposition of z into $v + w$, where $v \in V$ and $w \in W$. Hence, with $\|z\| = 1$, we have

$$\begin{aligned}
\langle z, \nabla^2 F_A(x) z\rangle &= \langle v, Av\rangle + \langle w, Aw\rangle + \gamma\Big(1 - \frac{1}{\|x\|}\Big) + \frac{\gamma}{\|x\|}\Big(\Big\langle \frac{x}{\|x\|}, v\Big\rangle + \Big\langle \frac{x}{\|x\|}, w\Big\rangle\Big)^2 \\
&= a^2\lambda_1 + b^2\lambda + \gamma\Big(1 - \frac{1}{\|x\|}\Big) + \frac{\gamma}{\|x\|}\Big(a\Big\langle \frac{x}{\|x\|}, \frac{v}{\|v\|}\Big\rangle + b\Big\langle \frac{x}{\|x\|}, \frac{w}{\|w\|}\Big\rangle\Big)^2,
\end{aligned}$$

where $\|v\| = a$ and $\|w\| = b$ and $a^2 + b^2 = 1$ and $\lambda = \langle w, Aw\rangle \geq \lambda_2$.

If $a \geq b$, then $a \geq \frac{1}{\sqrt{2}} \geq b$ and we have that for $k \geq 2$,

$$\Big(a\Big\langle \frac{x}{\|x\|}, \frac{v}{\|v\|}\Big\rangle + b\Big\langle \frac{x}{\|x\|}, \frac{w}{\|w\|}\Big\rangle\Big)^2 \geq \Big(a\Big(1 - \frac{1}{k^2}\Big) - b\sqrt{\frac{2}{k^2} - \frac{1}{k^4}}\Big)^2,$$

which implies that for $k \in \mathbb{N}$ with $k \geq \frac{2(\gamma+\lambda_2)}{\lambda_2 - \lambda_1} > 1$,

$$\begin{aligned}
\langle z, \nabla^2 F_A(x) z\rangle &\geq a^2(\lambda_1 + \gamma) + b^2\Big(\lambda_2 + \gamma - \frac{(\gamma+\lambda_1)k}{k-1}\Big) + \frac{\gamma}{\|x\|}\Big[\frac{2(b^2 - a^2)}{k^2} - 2ab\sqrt{\frac{2}{k^2} - \frac{1}{k^4}}\Big] \\
&\geq a^2(\lambda_1 + \gamma) + b^2\frac{(\lambda_2 - \lambda_1)(\gamma + \lambda_2)}{2\gamma + \lambda_2 + \lambda_1} - \frac{(\gamma+\lambda_2)}{k-1}\Big[\frac{2(a^2 - b^2)}{k} + 2ab\sqrt{2 - \frac{1}{k^2}}\Big] \\
&> a^2(\lambda_1 + \gamma) + b^2\frac{(\lambda_2 - \lambda_1)}{2} - \frac{(\gamma+\lambda_2)}{k-1}\Big[\frac{2}{k} + \sqrt{2}\Big]a^2 \\
&= a^2\frac{(\lambda_1 + \gamma)}{2} + b^2\frac{(\lambda_2 - \lambda_1)}{2} + \Big(\frac{(\lambda_1 + \gamma)}{2} - \frac{(\gamma+\lambda_2)}{k-1}\Big[\frac{2}{k} + \sqrt{2}\Big]\Big)a^2.
\end{aligned}$$

In addition, if $k \geq \max(\frac{2(\gamma+\lambda_2)}{\lambda_2 - \lambda_1}, \frac{4(\gamma+\lambda_2)}{\gamma+\lambda_1} + 1)$, then

$$\langle z, \nabla^2 F_A(x) z\rangle > a^2\frac{(\lambda_1 + \gamma)}{2} + b^2\frac{(\lambda_2 - \lambda_1)}{2} \geq \min\Big(\frac{\gamma + \lambda_2}{4}, \frac{\gamma + \lambda_1}{2}\Big).$$

On the other hand, if $a < b$, then $a < \frac{1}{\sqrt{2}} < b$ and $a^2\lambda_1 + b^2\lambda_2 > \frac{\lambda_1 + \lambda_2}{2}$ and we have that, with $k > \frac{4\gamma + \lambda_2 + 3\lambda_1}{\lambda_2 - \lambda_1}$,

$$\langle z, \nabla^2 F_A(x) z\rangle \geq a^2\lambda_1 + b^2\lambda_2 + \gamma - \frac{(\gamma+\lambda_1)k}{k-1} > \frac{\lambda_1 + \lambda_2}{2} + \gamma - \frac{(\gamma+\lambda_1)k}{k-1} > \frac{\lambda_2 - \lambda_1}{4}.$$

Therefore, if $k \geq \max(\frac{4(\gamma+\lambda_2)}{\lambda_2-\lambda_1}, 5 + \frac{4(\lambda_2-\lambda_1)}{\gamma+\lambda_1})$, then $\nabla^2 F_A(x) - \frac{\eta}{4} I$ is positive definite for $\|x - x_*\| < \epsilon$, where $\eta = \min(2(\gamma + \lambda_1), \lambda_2 - \lambda_1)$, i.e., $\mathcal{F}_A$ is $\frac{\eta}{4}$-strongly convex in

$$\{x \in \mathbb{R} : \|x - x_*\| < \epsilon\}.$$

□

2.1.3 Algorithm 2 : The Newton's Method

We have analyzed the gradient descent method to minimize the model (3) and confirmed that the method finds a global minimizer of (3) if an initial guess x_0 is not orthogonal to the eigenspace corresponding to the smallest eigenvalue. However, it converges slowly. After realizing strongly convex regions near global minimizers in the previous sections, we now discuss the Newton's method applied to (3) to enhance the convergence rate. As a byproduct of the discussion, we will be able to provide a nonsingular linear system whose unique solution is an eigenvector of A when the corresponding eigenvalue of multiplicity 1 is known.

Since the functional (3) is continuously twice differentiable at $x \neq 0$, if we apply the Newton's method, we will get

$$x_{k+1} = x_k - (\nabla^2 F(x_k))^{-1} \nabla F(x_k), \ k = 0, 1, 2, \ldots, \tag{13}$$

with an initial $x_0 \neq 0$ unless $\nabla^2 F(x_k)$ is singular. Noting that

$$\begin{aligned}
(\nabla^2 F(x))^{-1} \nabla F(x) &= \Big[A + \gamma\Big(1 - \frac{1}{\|x\|}\Big) I + \frac{\gamma}{\|x\|^3} x x^T\Big]^{-1} \Big[A + \gamma\Big(1 - \frac{1}{\|x\|}\Big) I\Big] x \\
&= x - \Big[A + \gamma\Big(1 - \frac{1}{\|x\|}\Big) I + \frac{\gamma}{\|x\|^3} x x^T\Big]^{-1} \Big[\frac{\gamma}{\|x\|} x\Big] \\
&= x - \Big[\frac{1}{\gamma} A + \Big(1 - \frac{1}{\|x\|}\Big) I + \frac{1}{\|x\|} \Big(\frac{x}{\|x\|}\Big) \Big(\frac{x}{\|x\|}\Big)^T\Big]^{-1} \Big[\frac{x}{\|x\|}\Big],
\end{aligned}$$

we can observe that (13) becomes

$$x_{k+1} = \Big[\frac{1}{\gamma} A + \Big(1 - \frac{1}{\|x_k\|}\Big) I + \frac{1}{\|x_k\|} \Big(\frac{x_k}{\|x_k\|}\Big) \Big(\frac{x_k}{\|x_k\|}\Big)^T\Big]^{-1} \frac{x_k}{\|x_k\|}, \ k = 0, 1, 2, \ldots.$$

Hence, we propose the following scheme:

- Initialize $x_0 \neq 0$,
- For $k = 0, 1, 2, \ldots$, compute y_k and x_{k+1} by

$$y_k = \frac{x_k}{\|x_k\|},$$

and

$$\Big[\frac{1}{\gamma} A + \Big(1 - \frac{1}{\|x_k\|}\Big) I + \frac{1}{\|x_k\|} \Big(\frac{x_k}{\|x_k\|}\Big) \Big(\frac{x_k}{\|x_k\|}\Big)^T\Big] x_{k+1} = y_k. \tag{14}$$

We wrote (14) in the given form to enhance its similarity to the inverse power method. Under the assumptions of Theorem 4, choosing an initial x_0 in the strongly convex region near a global minimum point x_*, the above algorithm presents, at least, the quadratic convergence. Moreover, unlike the Inverse Rayleigh Quotient Iteration, the system of linear equations in

(14) that we need to solve near x_* is far from being singular because $\nabla^2 F(x)$ is $\frac{\eta}{4}$-strongly convex in the region where we apply the above algorithm.

We will now show that the Newton's method discussed above converges to a critical point under a certain condition. In fact, the following theorem says that convergence of the norm $\|x_k\|$ is equivalent to convergence of x_k.

Theorem 5. *Let A be an $N \times N$ nonzero real symmetric matrix with eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$. Let $\gamma > \max(0, -\lambda_1)$, and $x_0 \neq 0$ with $\|x_0\| \neq \frac{\gamma}{\gamma+\lambda_j}$ for any $1 \leq j \leq N$.*

1. *Suppose that a sequence $\{x_k\}$ can be generated by (14) and that $\|x_k\|$ converges to $\eta > 0$. If $\|x_k\| \neq \frac{\gamma}{\gamma+\lambda_j}$ for any $1 \leq j \leq N$, $k \in \mathbb{N} \cup \{0\}$, then there exists $1 \leq i_0 \leq N$ such that $\eta = \frac{\gamma}{\gamma+\lambda_{i_0}}$ and x_k converges to an eigenvector x_* of A corresponding to the eigenvalue λ_{i_0} with $\|x_*\| = \eta$.*
2. *On the other hand, we assume that λ_i for some $1 \leq i \leq N$ has multiplicity 1 and q_i is a unit eigenvector of A corresponding to λ_i and $\|x_{k_0}\| = \frac{\gamma}{\gamma+\lambda_i}$ for some $k_0 \in \mathbb{N}$. If x_{k_0} is not a critical point of F_A with $|q_i^T x_{k_0}| > 0$, then x_{k_0+1} is an eigenvectors of A corresponding to the eigenvalue λ_i. If $|q_i^T x_{k_0}| \neq \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ for $j < i$, then x_{k_0+2} is a critical point of F_A, i.e., an eigenvector of A corresponding to the eigenvalue λ_i with norm $\|x_{k_0+2}\| = \frac{\gamma}{\gamma+\lambda_i}$. However, if $|q_i^T x_{k_0}| = \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ for some $j < i$, then the system becomes singular and we may not compute x_{k_0+2} uniquely. In any case, the algorithm terminates in $k_0 + 2$ iterations.*

Proof. It suffices to consider the case that A is a diagonal matrix with diagonal entries $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. Then, (14) becomes

$$\frac{1}{\gamma}\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix}\begin{bmatrix} x_{k+1,1} \\ \vdots \\ x_{k+1,N} \end{bmatrix} + \Big(1 - \frac{1}{\|x_k\|}\Big)\begin{bmatrix} x_{k+1,1} \\ \vdots \\ x_{k+1,N} \end{bmatrix} + \frac{1}{\|x_k\|}(y_k^T x_{k+1})y_k = y_k, \quad (15)$$

where $x_{k+1} = \begin{bmatrix} x_{k+1,1} & \cdots & x_{k+1,N} \end{bmatrix}^T$ and $y_k = \frac{x_k}{\|x_k\|}$. Let $\{x_k\}$ be a sequence generated by (15) with $\|x_k\| \neq \frac{\gamma}{\gamma+\lambda_j}$ for any $1 \leq j \leq N$ and for all $k \in \mathbb{N} \cup \{0\}$.

Firstly, we consider that $\|x_k\|$ converges to $\eta > 0$. Suppose that $\eta \neq \frac{\gamma}{\gamma+\lambda_j}$ for any $1 \leq j \leq N$. Note that for $1 \leq j \leq N$, we have

$$\begin{aligned} & \Big(1 + \frac{\lambda_j}{\gamma} - \frac{1}{\|x_k\|}\Big)x_{k+1,j} = y_{k,j}\Big(1 - \frac{\|x_{k+1}\|}{\|x_k\|}(y_k^T y_{k+1})\Big) \\ \Leftrightarrow \quad & \frac{x_{k+1,j}}{\|x_{k+1}\|} = \Big[\frac{\frac{\|x_k\|}{\|x_{k+1}\|} - (y_k^T y_{k+1})}{\|x_k\|(1 + \frac{\lambda_j}{\gamma}) - 1}\Big]\frac{x_{k,j}}{\|x_k\|}. \end{aligned} \quad (16)$$

Since $\eta > 0$ and $\Big(1 + \frac{\lambda_j}{\gamma} - \frac{1}{\|x_k\|}\Big) \neq 0$ for any $1 \leq j \leq N$ and $k \in \mathbb{N} \cup \{0\}$, we know that $\|x_k\| \neq 0$ for all $k \geq 0$. This implies that

$$1 - \frac{\|x_{k+1}\|}{\|x_k\|}(y_k^T y_{k+1}) \neq 0, \quad k \geq 0.$$

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

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

We will now show that $\limsup_k y_k^T y_{k+1} = 1$. Suppose that $\limsup_k y_k^T y_{k+1} = \epsilon < 1$. Given $0 < \delta < 1 - \epsilon$, we may choose $l_1 \in \mathbb{N}$ so that $k \geq l_1$ implies

$$|\|x_k\| - \eta| < \frac{\delta}{2} \quad \text{and} \quad \Big|\frac{\|x_k\|}{\|x_{k+1}\|} - 1\Big| < \frac{\delta}{2} \quad \text{and} \quad y_k^T y_{k+1} < \epsilon + \frac{\delta}{2}. \tag{17}$$

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

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

From (16), we see that for $k \geq l_1$,

$$\frac{|x_{k+1,J}|}{\|x_{k+1}\|} \geq \frac{1 - \delta - \epsilon}{\Big|\eta(1 + \frac{\lambda_J}{\gamma}) - 1\Big| + \frac{\delta}{2}(1 + \frac{\lambda_J}{\gamma})} \frac{|x_{k,J}|}{\|x_k\|}. \tag{18}$$

If $\Big|\eta\Big(1 + \frac{\lambda_J}{\gamma}\Big) - 1\Big| < 1 - \epsilon$, then with $\delta > 0$ satisfying

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

we can see from (18) that $\lim_{k\to\infty} \frac{|x_{k+1,J}|}{\|x_{k+1}\|} = \infty$, which is impossible. Hence,

$$\min_{j \in \mathcal{J}_0} \Big|\eta\Big(1 + \frac{\lambda_j}{\gamma}\Big) - 1\Big| \geq 1 - \epsilon$$

i.e.,

$$\eta \leq \frac{\epsilon\gamma}{\gamma + \lambda^*} \quad \text{or} \quad \eta \geq \frac{(2 - \epsilon)\gamma}{\gamma + \lambda_*},$$

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

Suppose that $\eta \leq \frac{\epsilon\gamma}{\gamma + \lambda^*}$. For any $0 < \delta < \frac{(1-\epsilon)\gamma}{2\gamma + \lambda^*} < 1 - \epsilon$, we can see from (17) that $k \geq l_1$ implies that for each $j \in \mathcal{J}_0$, we have

$$\|x_k\|\Big(1 + \frac{\lambda_j}{\gamma}\Big) - 1 < \Big(\eta + \frac{\delta}{2}\Big)\Big(1 + \frac{\lambda_j}{\gamma}\Big) - 1 \leq \frac{\epsilon(\gamma + \lambda_j)}{\gamma + \lambda^*} + \frac{\delta}{2}\Big(1 + \frac{\lambda_j}{\gamma}\Big) - 1 < 0. \tag{19}$$

In addition, from (17) we know that for $k \geq l_1$,

$$\frac{\|x_k\|}{\|x_{k+1}\|} - (y_k^T y_{k+1}) > 1 - \frac{\delta}{2} - (y_k^T y_{k+1}) > 0,$$

since $\epsilon + \delta < 1$ implies

$$y_k^T y_{k+1} < \epsilon + \frac{\delta}{2} < 1 - \frac{\delta}{2}.$$

This results in, together with (19),

$$\begin{aligned} y_k^T y_{k+1} &= \Big(\frac{\|x_k\|}{\|x_{k+1}\|} - (y_k^T y_{k+1})\Big)\Big(\sum_{j=1}^{N} \frac{y_{k,j}^2}{\|x_k\|(1 + \frac{\lambda_j}{\gamma}) - 1}\Big) \\ &\leq \Big(1 - \frac{\delta}{2} - (y_k^T y_{k+1})\Big)\Big(\sum_{j \in \mathcal{J}_0} \frac{y_{k,j}^2}{\|x_k\|(1 + \frac{\lambda_j}{\gamma}) - 1}\Big) < 0. \end{aligned}$$

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

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

$$\begin{aligned}\|x_k\|\Big(1+\frac{\lambda_j}{\gamma}\Big)-1 &> \Big(\eta-\frac{\delta}{2}\Big)\Big(1+\frac{\lambda_j}{\gamma}\Big)-1\\ &\geq \frac{(2-\epsilon)(\gamma+\lambda_j)}{\gamma+\lambda_*}-\frac{\delta}{2}\Big(1+\frac{\lambda_j}{\gamma}\Big)-1\\ &> \frac{(2-\epsilon)(\gamma+\lambda_j)}{\gamma+\lambda_*}-\frac{(1-\epsilon)(\gamma+\lambda_j)}{2(2\gamma+\lambda_*)}-1>0.\end{aligned}$$

Together with (17), we can conclude that for $k \geq l_1$, since $\epsilon+\delta<1$,

$$\begin{aligned}y_k^T y_{k+1} &= \Big(\frac{\|x_k\|}{\|x_{k+1}\|}-(y_k^T y_{k+1})\Big)\Big(\sum_{j=1}^N \frac{y_{k,j}^2}{\|x_k\|(1+\frac{\lambda_j}{\gamma})-1}\Big)\\ &\geq \Big(1-\frac{\delta}{2}-(y_k^T y_{k+1})\Big)\Big(\sum_{j\in\mathcal{J}_0} \frac{y_{k,j}^2}{\|x_k\|(1+\frac{\lambda_j}{\gamma})-1}\Big)>0.\end{aligned}$$

Hence, we have

$$0 \leq \epsilon < 1. \tag{20}$$

If we extract a subsequence y_{k_n} such that $y_{k_n}^T y_{k_n+1} \to \epsilon$ as $n \to \infty$, then using the form of (14) and knowing that

$$\liminf_{n\to\infty} \frac{y_{k_n+1}^T}{\|x_{k_n+1}\|}\frac{1}{\gamma}Ax_{k_n+1} = \liminf_{n\to\infty}\frac{1}{\gamma}y_{k_n+1}^T Ay_{k_n+1} \geq \frac{\lambda_*}{\gamma}$$

we can see that

$$\begin{aligned}&\frac{y_{k_n+1}^T}{\|x_{k_n+1}\|}\Big[\frac{1}{\gamma}A+\Big(1-\frac{1}{\|x_{k_n}\|}\Big)I+\frac{1}{\|x_{k_n}\|}\Big(\frac{x_{k_n}}{\|x_{k_n}\|}\Big)\Big(\frac{x_{k_n}}{\|x_{k_n}\|}\Big)^T\Big]x_{k_n+1} = \frac{y_{k_n+1}^T y_{k_n}}{\|x_{k_n+1}\|}\\ \Leftrightarrow\ &\frac{1}{\gamma}y_{k_n+1}^T Ay_{k_n+1}+\Big(1-\frac{1}{\|x_{k_n}\|}\Big)+\frac{1}{\|x_{k_n}\|}(y_{k_n}^T y_{k_n+1})^2 = \frac{y_{k_n+1}^T y_{k_n}}{\|x_{k_n+1}\|}\\ \Rightarrow\ &\frac{\lambda_*}{\gamma}+\Big(1-\frac{1}{\eta}\Big)+\frac{\epsilon^2}{\eta} \leq \frac{\epsilon}{\eta}\\ \Leftrightarrow\ &\eta \leq \frac{(1+\epsilon-\epsilon^2)\gamma}{\gamma+\lambda_*}.\end{aligned}$$

Hence, we have

$$\frac{(2-\epsilon)\gamma}{\gamma+\lambda_*} \leq \eta \leq \frac{(1+\epsilon-\epsilon^2)\gamma}{\gamma+\lambda_*}. \tag{21}$$

However, (20) implies $1+\epsilon-\epsilon^2 < 2-\epsilon$, which contradicts (21). Therefore, we conclude that $\epsilon = 1$, i.e., $\limsup_k y_k^T y_{k+1} = 1$.

By considering a subsequence y_{k_n} such that $\lim_{n\to\infty} y_{k_n}^T y_{k_n+1} = 1$, we can also see from (16) that

$$\begin{aligned}\infty = \lim_{n\to\infty}\frac{1}{(\frac{\|x_{k_n}\|}{\|x_{k_n+1}\|}-(y_{k_n}^T y_{k_n+1}))^2} &= \lim_{n\to\infty}\sum_{j=1}^N \frac{y_{k_n,j}^2}{(\|x_{k_n}\|(1+\frac{\lambda_j}{\gamma})-1)^2}\\ \leq \max_{1\leq j\leq N}\frac{1}{(\eta(1+\frac{\lambda_j}{\gamma})-1)^2} &< \infty,\end{aligned}$$

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

We will now show that x_k converges to an eigenvector x_* of A corresponding to λ_{i_0} with norm $\|x_*\| = \eta$. Firstly, we show that there exists $j \in \mathcal{J}_0$ such that $\lambda_j = \lambda_{i_0}$. It is easy to see that $\limsup_{k\to\infty} y_k^T y_{k+1} = 1$, whose proof is almost exactly the same as above and is easier using $\eta = \frac{\gamma}{\gamma+\lambda_{i_0}}$. So we will omit it. By choosing a subsequence y_{k_n} such that $y_{k_n}^T y_{k_n+1} \to 1$ a $n \to \infty$, we can also see that

$$\infty = \lim_{n\to\infty} \frac{1}{(\frac{\|x_{k_n}\|}{\|x_{k_n+1}\|} - (y_{k_n}^T y_{k_n+1}))^2} = \lim_{n\to\infty} \sum_{j=1}^{N} \frac{y_{k_n,j}^2}{(\|x_{k_n}\|(1+\frac{\lambda_j}{\gamma}) - 1)^2}.$$

Since $\|x_{k_n}\| \to \eta$, there must be $j \in \mathcal{J}_0$ with $\lambda_j = \lambda_{i_0}$. That is, $\{j \in \mathcal{J}_0 : \lambda_j = \lambda_{i_0}\} \neq \emptyset$. Hence, without loss of generality we will say that $i_0 \in \mathcal{J}_0$.

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

$$\Big|1 + \frac{\lambda_{i_0}}{\gamma} - \frac{1}{\|x_k\|}\Big| < \min_{\lambda_j \neq \lambda_{i_0}} \frac{|\lambda_j - \lambda_{i_0}|}{3\gamma}.$$

Note that $k \ge k_0$, and for $\lambda_j \neq \lambda_{i_0}$,

$$\Big|\frac{1 + \frac{\lambda_{i_0}}{\gamma} - \frac{1}{\|x_k\|}}{1 + \frac{\lambda_j}{\gamma} - \frac{1}{\|x_k\|}}\Big| < \frac{1}{2}.$$

Moreover, for $1 \le j \le N$, we can represent

$$\frac{x_{K+1,j}}{\|x_{K+1}\|} = \Big(\Pi_{k=0}^{K}\Big[\frac{\frac{1}{\|x_{k+1}\|} - \frac{1}{\|x_k\|}(y_k^T y_{k+1})}{1 + \frac{\lambda_j}{\gamma} - \frac{1}{\|x_k\|}}\Big]\Big)\frac{x_{0,j}}{\|x_0\|}, \quad K \ge 0. \tag{22}$$

Since $i_0 \in \mathcal{J}_0$, we have that for $\lambda_j \neq \lambda_{i_0}$, as $K \to \infty$,

$$\begin{aligned}\Big|\frac{x_{K+1,j}}{x_{K+1,i_0}}\Big| &= \Big(\Pi_{k=k_0}^{K}\Big|\frac{1 + \frac{\lambda_{i_0}}{\gamma} - \frac{1}{\|x_k\|}}{1 + \frac{\lambda_j}{\gamma} - \frac{1}{\|x_k\|}}\Big|\Big)\Big(\Pi_{k=0}^{k_0-1}\Big|\frac{1 + \frac{\lambda_{i_0}}{\gamma} - \frac{1}{\|x_k\|}}{1 + \frac{\lambda_j}{\gamma} - \frac{1}{\|x_k\|}}\Big|\Big)\Big|\frac{x_{0,j}}{x_{0,i_0}}\Big| \\ &\le \frac{1}{2^{K-k_0+1}}\Big|\frac{x_{0,j}}{x_{0,i_0}}\Big| \to 0.\end{aligned}$$

Therefore, as $K \to \infty$,

$$(x_{K+1,i_0})^2\Big(\sum_{\substack{j\in\mathcal{J}_0,\\ \lambda_j=\lambda_{i_0}}}\Big(\frac{x_{K+1,j}}{x_{K+1,i_0}}\Big)^2 + \sum_{\lambda_j\neq\lambda_{i_0}}\Big(\frac{x_{K+1,j}}{x_{K+1,i_0}}\Big)^2\Big) = \|x_{K+1}\|^2 \to \eta^2 = \Big(\frac{\gamma}{\gamma+\lambda_{i_0}}\Big)^2,$$

which implies that as $K \to \infty$, we have

$$(x_{K+1,i_0})^2 \sum_{\substack{j\in\mathcal{J}_0,\\ \lambda_j=\lambda_{i_0}}}\Big(\frac{x_{k_0,j}}{x_{k_0,i_0}}\Big)^2 \to \Big(\frac{\gamma}{\gamma+\lambda_{i_0}}\Big)^2.$$

Let $m = \Big(\sum_{\substack{j\in\mathcal{J}_0,\\ \lambda_j=\lambda_{i_0}}}(\frac{x_{k_0,j}}{x_{k_0,i_0}})^2\Big)^{\frac{1}{2}} \ge 1$. Then, we can see that as $K \to \infty$,

$$(x_{K+1,i_0})^2 \to \frac{1}{m^2}\Big(\frac{\gamma}{\gamma+\lambda_{i_0}}\Big)^2. \tag{23}$$

This implies that $y_{K+1,i_0}^2 \to \frac{1}{m^2}$ as $K \to \infty$. Noting that for all k,

$$\begin{aligned} 1 \geq |y_k^T y_{k+1}| &= \Big| \frac{\|x_k\|}{\|x_{k+1}\|} - (y_k^T y_{k+1}) \Big| \Big| \sum_{j=1}^{N} \frac{y_{k,j}^2}{\|x_k\|(1+\frac{\lambda_j}{\gamma})-1} \Big| \\ &\geq \Big| \frac{\|x_k\|}{\|x_{k+1}\|} - (y_k^T y_{k+1}) \Big| \Big(\Big| \sum_{\lambda_j=\lambda_{i_0}} \frac{y_{k,j}^2}{\|x_k\|(1+\frac{\lambda_{i_0}}{\gamma})-1} \Big| - \Big| \sum_{\lambda_j \neq \lambda_{i_0}} \frac{y_{k,j}^2}{\|x_k\|(1+\frac{\lambda_j}{\gamma})-1} \Big| \Big), \end{aligned}$$

and

$$\lim_{k\to\infty} \Big(\Big| \sum_{\lambda_j=\lambda_{i_0}} \frac{y_{k,j}^2}{\|x_k\|(1+\frac{\lambda_{i_0}}{\gamma})-1} \Big| - \Big| \sum_{\lambda_j \neq \lambda_{i_0}} \frac{y_{k,j}^2}{\|x_k\|(1+\frac{\lambda_j}{\gamma})-1} \Big| \Big) = \infty,$$

we can see that

$$\lim_{k\to\infty} \Big| \frac{\|x_k\|}{\|x_{k+1}\|} - (y_k^T y_{k+1}) \Big| = 0,$$

i.e., $y_k^T y_{k+1}$ converges to 1 as $k \to \infty$. Since (22) implies that for $j \in \mathcal{J}_0$ with $\lambda_j = \lambda_{i_0}$,

$$x_{k,j} = x_{k,i_0} \Big(\frac{x_{0,j}}{x_{0,i_0}} \Big),$$

we have that as $k \to \infty$,

$$\begin{aligned} y_k^T y_{k+1} &= \frac{1}{\|x_k\|\|x_{k+1}\|} \sum_{j=1}^{N} x_{k,j} x_{k+1,j} \\ &= \frac{1}{\|x_k\|\|x_{k+1}\|} \Big(\sum_{\substack{j\in\mathcal{J}_0, \\ \lambda_j=\lambda_{i_0}}} x_{k,i_0} x_{k+1,i_0} \Big(\frac{x_{0,j}}{x_{0,i_0}} \Big)^2 + \sum_{\substack{j\in\mathcal{J}_0, \\ \lambda_j\neq\lambda_{i_0}}} x_{k,j} x_{k+1,j} \Big) \\ &\to \Big(\frac{\gamma+\lambda_{i_0}}{\gamma} \Big)^2 m^2 \lim_{k\to\infty} (x_{k,i_0} x_{k+1,i_0}) = 1. \end{aligned}$$

Together with (23), we can easily see that x_{k,i_0} converges either to $\frac{1}{m}\frac{\gamma}{\gamma+\lambda_{i_0}}$ or $-\frac{1}{m}\frac{\gamma}{\gamma+\lambda_{i_0}}$.

Therefore, x_k converges to x_* with

$$x_{*,j} = \begin{cases} (\frac{x_{0,j}}{x_{0,i_0}}) x_{*,i_0}, & \text{for } j \in \mathcal{J}_0 \text{ and } \lambda_j = \lambda_{i_0}, \\ 0, & \text{otherwise,} \end{cases}$$

where $x_{*,i_0} = \pm\frac{1}{m}\frac{\gamma}{\gamma+\lambda_{i_0}}$. Note that x_* is an eigenvector of A corresponding to λ_{i_0} with norm $\|x_*\| = \eta = \frac{\gamma}{\gamma+\lambda_{i_0}}$.

Secondly, we consider that λ_i for some $1 \leq i \leq N$ has multiplicity 1 and $\|x_{k_0}\| = \frac{\gamma}{\gamma+\lambda_i}$ for some $k_0 \in \mathbb{N}$. If x_{k_0} is not a critical point of F_A and $x_{k_0,i} \neq 0$, then we have that $|x_{k_0,i}| < \frac{\gamma}{\gamma+\lambda_i}$ and $\frac{1}{\|x_{k_0}\|}(y_k^T x_{k_0+1}) = 1$ in (15), which leads to

$$\frac{1}{\gamma} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_N \end{bmatrix} \begin{bmatrix} x_{k_0+1,1} \\ \vdots \\ x_{k_0+1,N} \end{bmatrix} = \frac{\lambda_i}{\gamma} \begin{bmatrix} x_{k_0+1,1} \\ \vdots \\ x_{k_0+1,N} \end{bmatrix}. \tag{24}$$

Hence, there exists a unique solution x_{k_0+1}, which is $x_{k_0+1} = \alpha \mathbf{e}_i$, where $\mathbf{e}_i$ is the standard basis element in $\mathbb{R}^N$ with $\mathbf{e}_{i,j} = \delta_{ij}$ and $\alpha = \frac{\gamma^2}{x_{k_0,i}(\gamma+\lambda_i)^2}$. Since $|x_{k_0,i}| < \frac{\gamma}{\gamma+\lambda_i}$, we have that

$|\alpha| > \frac{\gamma}{\gamma+\lambda_i}$ and x_{k_0+1} is an eigenvector of A corresponding to the eigenvalue λ_i, which is not a critical point of F_A.

Since x_{k_0+2} satisfies (15) and y_{k_0+1} is either $\mathbf{e}_i$ or $-\mathbf{e}_i$, if we multiply (15) by $\mathbf{e}_j^T$ from the left, then we have

$$\begin{cases} (\frac{\lambda_j}{\gamma} + 1 - \frac{1}{|\alpha|})x_{k_0+2,j} + \frac{\delta_{ij}}{|\alpha|}x_{k_0+2,i} = \delta_{ij}, & \text{if } y_{k_0+1} = \mathbf{e}_i, \\ (\frac{\lambda_j}{\gamma} + 1 - \frac{1}{|\alpha|})x_{k_0+2,j} + \frac{\delta_{ij}}{|\alpha|}x_{k_0+2,i} = -\delta_{ij}, & \text{if } y_{k_0+1} = -\mathbf{e}_i. \end{cases}$$

Note that $|\alpha| = \frac{\gamma}{\gamma+\lambda_j}$ is equivalent to $|x_{k_0,i}| = \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$. Since $|x_{k_0,i}| < \frac{\gamma}{\gamma+\lambda_i}$, we know that it is possible to have $|\alpha| = \frac{\gamma}{\gamma+\lambda_j}$ only if $j < i$.

Hence, if $|x_{k_0,i}| \neq \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ for $j < i$, then $|\alpha| \neq \frac{\gamma}{\gamma+\lambda_j}$ for $j \neq i$, and we have

$$x_{k_0+2,j} = \begin{cases} \pm\frac{\gamma}{\gamma+\lambda_i}, & \text{if } j = i, \\ 0, & \text{if } j \neq i. \end{cases}$$

depending on $y_{k_0+1} = \pm\mathbf{e}_i$. That is, x_{k_0+2} is a critical point of A, that is, an eigenvector of A corresponding to the eigenvalue λ_i with norm $\|x_{k_0+2}\| = \frac{\gamma}{\gamma+\lambda_i}$ and the algorithm terminates.

On the other hand, if $|x_{k_0,i}| = \frac{\gamma(\gamma+\lambda_j)}{(\gamma+\lambda_i)^2}$ for some $j < i$, then $|\alpha| = \frac{\gamma}{\gamma+\lambda_j}$ and the system becomes singular and the algorithm terminates. □

2.1.4 A Nonsingular Linear System for Eigenvector Estimation

Inspired by the Newton's method applied to our proposed model (3) with the convergence results, we realize that we can propose a nonsingular linear system for eigenvector estimation.

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, our nonsingular system guarantees to find a corresponding eigenvector by solving the system once when knowing that the eigenvalue has multiplicity 1. Moreover, even if we have an estimated eigenvalue, we can guaranteed that the error in the eigenvector estimation is 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 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 x_0 uniformly at random from S^{N-1} and solve

$$(A - \tilde{\lambda}I + (\gamma + \tilde{\lambda})x_0x_0^T)x = \gamma x_0. \tag{25}$$

The equation (25), which is what we propose to solve for an eigenvector of A corresponding to $\tilde{\lambda}$, is inspired by (14) and the proof of the second part of Theorem 5. By dividing (25) by γ, we have the same form as in (14). The difference is the condition on γ. In solving (25), any $\gamma > 0$ with $\gamma \neq -\tilde{\lambda}$ works. The following proposition says that (25) is, indeed, a nonsingular linear system with probability 1, whose unique solution is a corresponding eigenvector.

Proposition 1. *Suppose that $\tilde{\lambda}$ has multiplicity 1. With probability 1, the equation* (25) *has a unique solution $\tilde{x}$ that is an eigenvector of A corresponding to the eigenvalue $\tilde{\lambda}$.*

Proof. Let q be a unit eigenvector of A corresponding to $\tilde{\lambda}$. Note that if we choose $x_0 \in S^{N-1}$ uniformly at random, then we have $q^Tx_0 \neq 0$ with probability 1. Moreover, if $q^Tx_0 \neq 0$, then $(A - \tilde{\lambda}I + (\gamma + \tilde{\lambda})(x_0x_0^T))z = 0$ implies $x_0^Tz = 0$. Hence, we have $Az = \tilde{\lambda}z$. Since $\tilde{\lambda}$ has

multiplicity 1, $z = aq$ for some $a \in \mathbb{R}$. In addtion, since $0 = z^T x_0 = aq^T x_0$ and $q^T x_0 \neq 0$, we must have $a = 0$. That is, $z = 0$. Hence, $A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})(x_0 x_0^T)$ is nonsingular and there exists a unique solution $\tilde{x}$ to (25). By multiplying (25) by q^T, we have $(\gamma + \tilde{\lambda}) x_0^T \tilde{x} = \gamma$, which implies that $\tilde{x}$ also satisfies

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

□

For practical reasons, when only approximate eigenvalues are available, a good estimate of an eigenvalue guarantees a good estimate of a corresponding eigenvalue using a nonsingular linear system in the same form as (25).

Proposition 2. *Let A be an $N \times N$ symmetric matrix. Let $\tilde{\lambda}$ be an eigenvalue of A with multiplicity 1. Let $\gamma > 0$ be such that $\gamma \neq -\tilde{\lambda}$. Let $x_0 \in S^{N-1}$ be such that $0 < |q^T x_0| < 1$, where q is a unit eigenvector of A corresponding to $\tilde{\lambda}$. Then, for λ close enough to $\tilde{\lambda}$, but not equal, $A - \lambda I + (\gamma + \lambda)(x_0 x_0^T)$ is nonsingular and there exist $0 < c \leq C$ such that*

$$c|\lambda - \tilde{\lambda}| < \|x - \tilde{x}\| < C|\lambda - \tilde{\lambda}|,$$

where x and $\tilde{x}$ are the solutions to (25), respectively, i.e.,

$$(A - \lambda I + (\gamma + \lambda)(x_0 x_0^T))x = \gamma x_0 \quad \text{and} \quad (A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})(x_0 x_0^T))\tilde{x} = \gamma x_0.$$

Proof. Since $q^T x_0 \neq 0$ implies, by Proposition 1, that $A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})(x_0 x_0^T)$ is nonsingular, there exists $\epsilon > 0$ such that if $|\lambda - \tilde{\lambda}| < \epsilon$,

$$A - \lambda I + (\gamma + \lambda)(x_0 x_0^T)$$

is nonsingular.

Let $\alpha, \beta > 0$ be the smallest and the largest singular values of $A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})(x_0 x_0^T)$. Note that

$$\alpha \|\tilde{x}\| \leq \|(A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})(x_0 x_0^T))\tilde{x}\| = \gamma \leq \beta \|\tilde{x}\| \;\Rightarrow\; \alpha \leq \frac{\gamma}{\|\tilde{x}\|} \leq \beta,$$

and

$$\alpha \|x\| \leq \|(A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})(x_0 x_0^T))x\| = \|\gamma x_0 + (\lambda - \tilde{\lambda})(I - x_0 x_0^T)x\| \leq \gamma + |\lambda - \tilde{\lambda}| \|x\|.$$

So, if $|\lambda - \tilde{\lambda}| < \min(\frac{\alpha}{2}, \epsilon)$, then $\|x\| < \frac{2\gamma}{\alpha}$. In addition, noting that

$$\Big(A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})(x_0 x_0^T)\Big)(x - \tilde{x}) = (\lambda - \tilde{\lambda})(I - x_0 x_0^T)x, \tag{26}$$

we obtain that for $\|x\| < \frac{2\gamma}{\alpha}$,

$$\alpha \|x - \tilde{x}\| \leq \|(A - \tilde{\lambda} I + (\gamma + \tilde{\lambda})(x_0 x_0^T))(x - \tilde{x})\| \leq |\lambda - \tilde{\lambda}| \|x\| < \frac{2\gamma}{\alpha} |\lambda - \tilde{\lambda}|.$$

Hence, if we set

$$C = \frac{2\gamma}{\alpha^2} > 0,$$

and if $|\lambda - \tilde{\lambda}| < \min(\epsilon, \frac{\alpha}{2})$, then we have

$$\|x - \tilde{x}\| < C|\lambda - \tilde{\lambda}|$$

Secondly, since (26) is the same as

$$(\lambda-\tilde{\lambda})(I-x_0x_0^T)(\tilde{x}-x)+(A-\tilde{\lambda}I+(\gamma+\tilde{\lambda}(x_0x_0^T)))(x-\tilde{x})=(\lambda-\tilde{\lambda})(I-x_0x_0^T)\tilde{x},$$

we have

$$(|\lambda-\tilde{\lambda}|+\beta)\|x-\tilde{x}\|\geq|\lambda-\tilde{\lambda}|\|\tilde{x}-(x_0^T\tilde{x})x_0\|.$$

Therefore, if $|\lambda-\tilde{\lambda}|<\min(\epsilon,\frac{\alpha}{2})$, then we have

$$\Big(\frac{\alpha}{2}+\beta\Big)\|x-\tilde{x}\|>|\lambda-\tilde{\lambda}|\|\tilde{x}-(x_0^T\tilde{x})x_0\|$$

implying

$$c|\lambda-\tilde{\lambda}|<\|x-\tilde{x}\|$$

with

$$c=\frac{2\|\tilde{x}-(x_0^T\tilde{x})x_0\|}{(\alpha+2\beta)}>0$$

since $|q^Tx_0|<1$ implies $\tilde{x}\neq(x_0^T\tilde{x})x_0$. □

When the multiplicity of an eigenvalue $\tilde{\lambda}$ whose eigenvector is being sought is greater than 1, we can see that $A-\tilde{\lambda}I+(\gamma+\tilde{\lambda})(x_0x_0^T)$ is singular and may not apply the previous propositions. However, we can still estimate a corresponding eigenvector with an estimated eigenvalue.

Proposition 3. *Let A be an $N\times N$ symmetric matrix with eigenvalues $\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_N$ and linearly independent corresponding unit eigenvectors $q_1,\ldots,q_N$, respectively. Let $\tilde{\lambda}$ be an eigenvalue λ_{k_0}, for some $1\leq k_0\leq N$, of A with multiplicity $m>1$ so that $\tilde{\lambda}=\lambda_{k_0}=\cdots=\lambda_{k_0+m-1}$. Let $\gamma>0$ be such that $\gamma\neq-\tilde{\lambda}$. Let $x_0\in S^{N-1}$ be such that*

$$\zeta_0:=\sqrt{\sum_{j=0}^{m-1}|q_{k_0+j}^Tx_0|^2}>0.$$

For λ close enough to $\tilde{\lambda}$, but not equal, there exists a unique solution x_λ to

$$(A-\lambda I+(\gamma+\lambda)(x_0x_0^T))x_\lambda=\gamma x_0.$$

Moreover, there exists $\eta>0$ such that

$$\|(A-\tilde{\lambda}I)x_\lambda\|\leq\eta|\tilde{\lambda}-\lambda|$$

and $\{\frac{x_\lambda}{\|x_\lambda\|}\}$ has exactly two limit points $\tilde{x}$ and $-\tilde{x}$, where $\tilde{x}$ satisfies

$$q_i^T\tilde{x}=\begin{cases}\frac{q_i^Tx_0}{\zeta_0}, & \text{for } k_0\leq i\leq k_0+m-1,\\ 0, & \text{otherwise.}\end{cases}$$

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 $m>1$ and $\tilde{\lambda}=\lambda_{k_0}=\cdots=\lambda_{k_0+m-1}$ for some $1\leq k_0\leq N-m+1$. We assume that

$$\zeta_0:=\sqrt{\sum_{j=0}^{m-1}|x_{0,k_0+j}|^2}>0.$$

Note that if $\lambda \neq \lambda_j$ for all $1 \leq j \leq N$, $A - \lambda I$ is nonsingular and x_λ must satisfy

$$(A-\lambda I)x_\lambda = \begin{bmatrix} (\lambda_1-\lambda)x_{\lambda,1} \\ \vdots \\ (\lambda_N-\lambda)x_{\lambda,N} \end{bmatrix} = (\gamma - (\gamma+\lambda)(x_0^T x_\lambda)) \begin{bmatrix} x_{0,1} \\ \vdots \\ v_{0,N} \end{bmatrix},$$

i.e., $(A-\lambda I)x_\lambda = \alpha_\lambda x_0$ for some $\alpha_\lambda \in \mathbb{R}$. Since $x_\lambda = 0$ cannot be a solution, we know that α_λ must be nonzero and

$$x_\lambda = \alpha_\lambda \begin{bmatrix} \frac{1}{\lambda_1-\lambda}x_{0,1} \\ \vdots \\ \frac{1}{\lambda_N-\lambda}x_{0,N} \end{bmatrix}.$$

Since $\alpha_\lambda = \gamma - (\gamma+\lambda)(x_0^T x_\lambda)$, we have

$$\alpha_\lambda = \gamma - \alpha_\lambda\Big(\frac{\gamma+\lambda}{\lambda_1-\lambda}x_{0,1}^2 + \cdots + \frac{\gamma+\lambda}{\lambda_N-\lambda}x_{0,N}^2\Big)$$

implying

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

For λ close enough to $\tilde{\lambda}$, not equal to, it is easy to see that $\alpha_\lambda \neq 0$ exists and is unique, i.e. x_λ exists and is unique. In fact, since

$$\begin{aligned} \Big|\frac{\gamma+\tilde{\lambda}}{\tilde{\lambda}-\lambda}\Big|\Big(\sum_{j=0}^{m-1} x_{0,k_0+j}^2\Big) - \Big|\sum_{j\notin\{k_0,\ldots,k_0+m-1\}} \Big(\frac{\gamma+\lambda_j}{\lambda_j-\lambda}\Big)x_{0,j}^2\Big| \\ \leq \Big|\sum_{j=1}^N \Big(\frac{\gamma+\lambda_j}{\lambda_j-\lambda}\Big)x_{0,j}^2\Big| \\ \leq \Big|\frac{\gamma+\tilde{\lambda}}{\tilde{\lambda}-\lambda}\Big|\Big(\sum_{j=0}^{m-1} x_{0,k_0+j}^2\Big) + \Big|\sum_{j\notin\{k_0,\ldots,k_0+m-1\}} \Big(\frac{\gamma+\lambda_j}{\lambda_j-\lambda}\Big)x_{0,j}^2\Big|, \end{aligned}$$

we can see that there exists $0 < \epsilon$ such that $0 < |\lambda - \tilde{\lambda}| < \epsilon$ implies

$$\frac{1}{2}\Big|\frac{\gamma+\tilde{\lambda}}{\tilde{\lambda}-\lambda}\Big|\Big(\sum_{j=0}^{m-1} x_{0,k_0+j}^2\Big) \leq \Big|\sum_{j=1}^N \Big(\frac{\gamma+\lambda_j}{\lambda_j-\lambda}\Big)x_{0,j}^2\Big| \leq \frac{3}{2}\Big|\frac{\gamma+\tilde{\lambda}}{\tilde{\lambda}-\lambda}\Big|\Big(\sum_{j=0}^{m-1} x_{0,k_0+j}^2\Big),$$

that is,

$$\frac{2\gamma}{3(\gamma+\tilde{\lambda})\zeta_0^2}|\lambda-\tilde{\lambda}| \leq |\alpha_\lambda| \leq \frac{2\gamma}{(\gamma+\tilde{\lambda})\zeta_0^2}|\lambda-\tilde{\lambda}|.$$

Let $c = \min(|\lambda_{k_0-1} - \tilde{\lambda}|, |\lambda_{k_0+m} - \tilde{\lambda}|)$. We may assume that $\epsilon < \min(c, 1)$. Therefore, for $0 < |\lambda - \tilde{\lambda}| < \epsilon$, x_λ exists and is unique and

$$A - \lambda I + (\gamma+\lambda)(x_0 x_0^T)$$

is nonsingular, as well.

Since $\|x_\lambda\| = |\alpha_\lambda|\sqrt{\sum_{j=1}^N (\frac{1}{\lambda_j-\lambda})^2 x_{0,j}^2} \geq \frac{|\alpha_\lambda|}{|\lambda-\tilde{\lambda}|}\zeta_0$ and $0 < \epsilon < c$, we have that

$$\frac{2\gamma}{3(\gamma+\tilde{\lambda})\zeta_0} \leq \|x_\lambda\| \leq \frac{2\gamma}{(\gamma+\tilde{\lambda})\zeta_0^2}\sqrt{\frac{\epsilon^2}{c^2}+\zeta_0^2} \leq \frac{2\gamma\sqrt{1+\zeta_0^2}}{(\gamma+\tilde{\lambda})\zeta_0^2} \leq \frac{2\sqrt{2}\gamma}{(\gamma+\tilde{\lambda})\zeta_0^2}$$

for $0 < |\lambda - \tilde{\lambda}| < \epsilon$. Together with

$$\|(A - \lambda I)x_\lambda\| = \|\alpha_\lambda x_0\| = |\alpha_\lambda| \leq \frac{2\gamma}{(\gamma + \tilde{\lambda})\zeta_0^2}|\lambda - \tilde{\lambda}|,$$

we have that $0 < |\tilde{\lambda} - \lambda| < \epsilon$ implies

$$\|(A - \tilde{\lambda} I)x_\lambda\| \leq \|(A - \lambda I)x_\lambda\| + \|x_\lambda\||\tilde{\lambda} - \lambda| \leq \eta|\tilde{\lambda} - \lambda|,$$

where $\eta = \frac{2(1+\sqrt{2})\gamma}{(\gamma+\tilde{\lambda})\zeta_0^2}$.

Moreover,

$$y_\lambda := \frac{x_\lambda}{\|x_\lambda\|} = \begin{bmatrix} \frac{\frac{1}{\lambda_1 - \lambda}x_{0,1}}{\sqrt{\sum_{j=1}^N (\frac{1}{\lambda_j - \lambda})^2 x_{0,j}^2}} \\ \vdots \\ \frac{\frac{1}{\lambda_N - \lambda}x_{0,N}}{\sqrt{\sum_{j=1}^N (\frac{1}{\lambda_j - \lambda})^2 x_{0,j}^2}} \end{bmatrix}$$

and if $j \notin \{k_0, k_0 + 1, \ldots, k_0 + m - 1\}$, then as $\lambda \to \tilde{\lambda} = \lambda_{k_0}$, we have

$$\begin{aligned} y_{\lambda,j} &= \frac{\frac{1}{\lambda_j - \lambda}x_{0,j}}{\sqrt{\sum_{j=1}^N (\frac{1}{\lambda_j - \lambda})^2 x_{0,j}^2}} \\ &= \frac{\frac{|\lambda_{k_0} - \lambda|}{\lambda_j - \lambda}x_{0,j}}{\sqrt{\sum_{j \notin \{k_0, \cdots k_0 + m - 1\}} (\frac{\lambda_{k_0} - \lambda}{\lambda_j - \lambda})^2 x_{0,j}^2 + x_{0,k_0}^2 + \cdots + x_{0,k_0+m-1}^2}} \\ &\to 0. \end{aligned}$$

If $j \in \{k_0, k_0 + 1, \ldots, k_0 + m - 1\}$, then it is easy to see that

$$y_{\lambda,j} \to \begin{cases} \frac{x_{0,j}}{\zeta_0}, & \text{as } \lambda \uparrow \lambda_{k_0}, \\ -\frac{x_{0,j}}{\zeta_0}, & \text{as } \lambda \downarrow \lambda_{k_0}. \end{cases}$$

Hence,

$$y_\lambda \to \begin{cases} \tilde{x}, & \text{as } \lambda \uparrow \lambda_{k_0}, \\ -\tilde{x}, & \text{as } \lambda \downarrow \lambda_{k_0}, \end{cases}$$

where $\tilde{x} = \begin{cases} \frac{x_{0,i}}{\zeta_0}, & \text{for } k_0 \leq i < k_0 + m, \\ 0, & \text{otherwise.} \end{cases}$. □

Surprisingly, the propositions above say that the error in estimating an eigenvector $\tilde{x}$ is no worse than the error in estimating an eigenvalue $\tilde{\lambda}$. That is, if we have a very accurate estimation of an eigenvalue, then we can estimate a corresponding eigenvector within the same accuracy using (25), which is a new observation. For example, there is a recently proposed method to estimate eigenvalues using contour integrals in [1] and similar previous works, therein, where the authors mentioned that if more accurate estimation is needed, then part of their algorithm could be run again with another randomly set of vectors. However, simply solving our one step estimation method (25) with the estimated eigenvalues provides better eigenvalue estimation since an $O(\epsilon)$ error in eigenvector estimation gives rise to an $O(\epsilon^2)$ error in eigenvalue estimation (Rayleigh quotient effect), and this error in eigenvalue guarantees the same amount of error $O(\epsilon^2)$ in eigenvector using the simple one step (25).

2.1.5 One Step Estimation For Nonsymmetric Diagonalizable Matrices

We can also observe that the above propositions provide us with a method either to find or to approximate eigenvectors of a nonsymmetric diagonalizable matrix A 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 will provide the following corollary.

Corollary 1. *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 a unit vector x_0 uniformly at random, we have that with probability 1,*
$$(A + \gamma x_0 x_0^T)x = \gamma x_0$$
has a unique solution $\tilde{x}$ that spans $\mathcal{N}(A)$.
2. *If $\mathcal{N}(A)$ has dimension greater that 1, then for small enough $\lambda > 0$, choosing a unit vector x_0 uniformly at random, we have that with probability 1,*
$$(A - \lambda I + (\gamma + \lambda)x_0 x_0^T)x = \gamma x_0 \tag{27}$$
has a unique solution x_λ satisfying $\|Ax_\lambda\| \leq \eta\lambda$ with η depending on x_0.

Proof. The first part can be proven in the same way as we proved Proposition 1 with a choice of x_0 satisfying $q^T x_0 \neq 0$ where q is a unit vector spanning $\mathcal{N}(A)$. Since the second part can be proven with a minor modification in the proof of Proposition 3, we will only provide a sketch of this proof.

Note that $A - \lambda I$ for small enough $\lambda > 0$ is invertible and that if we choose a unit vector x_0 uniformly at random, then x_0 is not orthogonal to $\mathcal{N}(A)$ with probability 1. In fact, since $\mathbb{R}^N = \text{eigenspace}(0) \oplus \bigoplus_{\lambda_j \neq 0} \text{eigenspace}(\lambda_j)$, we can represent x_0 as $q_0 + r_0$, where $0 \neq q_0 \in \text{eigenspace}(0) = \mathcal{N}(A)$ and $r_0 \in \bigoplus_{\lambda_j \neq 0} \text{eigenspace}(\lambda_j)$.

With such an x_0, we need to see if we can solve $(A-\lambda I)x = (\gamma x_0 - (\gamma+\lambda)x_0 x_0^T x) = \alpha_\lambda x_0$, with
$$\alpha_\lambda = \gamma - (\gamma + \lambda)x_0^T x.$$
Noting that if there exists such a solution x_λ, we have
$$x_\lambda = \alpha_\lambda (A - \lambda I)^{-1} x_0$$
and α_λ satisfies
$$\alpha_\lambda = \gamma - \alpha_\lambda(\gamma + \lambda)x_0^T (A - \lambda I)^{-1} x_0,$$
i.e.,
$$\alpha_\lambda(1 + (\gamma + \lambda)x_0^T (A - \lambda I)^{-1} x_0) = \gamma \;\Leftrightarrow\; \alpha_\lambda\Big(1 + \Big(\frac{\gamma}{\lambda} + 1\Big)x_0^T \Big(\frac{1}{\lambda}A - I\Big)^{-1} x_0\Big) = \gamma.$$
It is not difficult to see that
$$\Big\|\Big(\frac{1}{\lambda}A - I\Big)^{-1} r_0\Big\| \to 0 \quad \text{as} \quad \lambda \downarrow 0.$$
Hence, there exists $\lambda_0 < \min(\|A\|, \frac{\gamma|x_0^T q_0|}{2(2-|x_0^T q_0|)})$ such that $0 < \lambda < \lambda_0$ implies

- $\|(\frac{1}{\lambda}A - I)^{-1} r_0\| < \frac{|x_0^T q_0|}{2}$,
- $\frac{\gamma}{\lambda} + 1 > \frac{2}{|x_0^T q_0|}$.

This implies that for $0 < \lambda < \lambda_0$, we have

$$\frac{1}{2}|x_0^T q_0| < \Big|x_0^T \Big(\frac{1}{\lambda}A - I\Big)^{-1} x_0\Big| = \Big| - x_0^T q_0 + x_0^T \Big(\frac{1}{\lambda}A - I\Big)^{-1} r_0\Big| < \frac{3}{2}|x_0^T q_0|.$$

It is now easy to see that for $0 < \lambda < \lambda_0$, $\alpha_\lambda = \frac{\gamma}{(1+(\gamma+\lambda)x_0^T(A-\lambda I)^{-1}x_0)}$ exists uniquely and satisfies

$$|\alpha_\lambda| < \frac{\gamma}{\frac{1}{2}\Big(\frac{\gamma}{\lambda}+1\Big)|x_0^T q_0| - 1} = \Big(\frac{2\gamma}{\gamma|x_0^T q_0| - \lambda(2 - |x_0^T q_0|)}\Big)\lambda < \frac{4\lambda}{|x_0^T q_0|}$$

and

$$|\alpha_\lambda| > \frac{\gamma}{\frac{3}{2}\Big(\frac{\gamma}{\lambda}+1\Big)|x_0^T q_0| + 1} = \Big(\frac{2\gamma}{3\gamma|x_0^T q_0| + \lambda(2 + 3|x_0^T q_0|)}\Big)\lambda > \frac{4\lambda}{11|x_0^T q_0|}$$

That is, for $0 < \lambda < \lambda_0$, there exists a unique solution

$$x_\lambda = \alpha_\lambda (A - \lambda I)^{-1} x_0 = \frac{\gamma (A - \lambda I)^{-1} x_0}{(1 + (\gamma + \lambda) x_0^T (A - \lambda I)^{-1} x_0)}$$

with

$$\frac{4}{11|x_0^T q_0|}\|\lambda(A - \lambda I)^{-1} x_0\| < \|x_\lambda\| < \frac{4}{|x_0^T q_0|}\|\lambda(A - \lambda I)^{-1} x_0\|$$

Note that for $0 < \lambda < \lambda_0$, we have

$$\frac{\|q_0\|}{2} \le \|\lambda(A - \lambda I)^{-1} x_0\| = \Big\|\Big(\frac{1}{\lambda}A - I\Big)^{-1} x_0\Big\| = \Big\| - q_0 + \Big(\frac{1}{\lambda}A - I\Big)^{-1} r_0\Big\| \le \frac{3\|q_0\|}{2},$$

implying that for $0 < \lambda < \lambda_0$,

$$\frac{2\|q_0\|}{11|x_0^T q_0|} < \|x_\lambda\| < \frac{6\|q_0\|}{|x_0^T q_0|}.$$

Finally, we note that $0 < \lambda < \lambda_0$ implies

$$\begin{aligned}
\|Ax_\lambda\| &= \Big\|\frac{\gamma A(A - \lambda I)^{-1} x_0}{(1 + (\gamma + \lambda) x_0^T (A - \lambda I)^{-1} x_0)}\Big\| \\
&= \Big\|\alpha_\lambda x_0 + \lambda \alpha_\lambda (A - \lambda I)^{-1} x_0\Big\| \\
&\le |\alpha_\lambda|(1 + \|\lambda(A - \lambda I)^{-1} x_0\|) \\
&= \eta\lambda,
\end{aligned}$$

where $\eta = \frac{4+6\|q_0\|}{|x_0^T q_0|}$. □

3 Applications

So far, we have analyzed our proposed model (3) when the given matrix A is real symmetric. As was mentioned earlier, the same analysis can be carried over to complex hermitian matrices. When A is real and nonsymmetric, our proposed model (3) with $B = A^T A$ will present a singular value decomposition of A. Since B is symmetric, the previous discussion applies without alteration. For this reason, we will not give a detailed presentation for singular value estimation, but rather provide other applications such as generalized eigenvalue problems and eigenfunctions of self-adjoint elliptic operators.

3.1 Generalized Eigenvalue Problem : $Ax = \lambda Bx$

Given that $A, B \in \mathbf{M}_N(\mathbb{R})$ are symmetric matrices, and B is positive definite with the smallest eigenvalue $\mu_{(B,1)} > 0$, and A has the smallest eigenvalue $\mu_{(A,1)}$, we would like to consider the generalized eigenvalue problem:

$$Ax = \lambda Bx. \tag{28}$$

We will denote by $\mu_{(A,j)}, \mu_{(B,j)}$ the j^{th} smallest eigenvalues of A and B, respectively. In addition, it is enough to consider that $\mu_{(B,1)} = 1$ and that either $-\mu_{(A,N)} \leq \mu_{(A,1)} < 0$ or $\mu_{(A,1)} \geq 0$ holds because (28) is equivalent to

$$\pm Ax = (\pm\lambda \cdot \mu_{(B,1)})\Big(\frac{1}{\mu_{(B,1)}}B\Big)x.$$

With $\gamma > \max(0, -\mu_{(A,1)})$, we know that $A + \gamma B$ is positive definite and we can solve (28) by minimizing

$$F_{A,B}(x) = \frac{1}{2}\langle Ax, x\rangle + \frac{\gamma}{2}\langle Bx, x\rangle - \gamma\sqrt{\langle Bx, x\rangle}, \tag{29}$$

since we have

$$\nabla F_{A,B}(x) = Ax + \gamma\Big(1 - \frac{1}{\sqrt{\langle Bx, x\rangle}}\Big)Bx = 0$$

if and only if

$$Ax = \gamma\Big(\frac{1}{\sqrt{\langle Bx, x\rangle}} - 1\Big)Bx.$$

Therefore, if $x_* \neq 0$ is a critical point of $F_{A,B}$, then (x_*, λ_*) is a solution pair of (28), i.e., $Ax_* = \lambda_* Bx_*$ with

$$\lambda_* = \gamma\Big(\frac{1}{\sqrt{\langle Bx_*, x_*\rangle}} - 1\Big).$$

Moreover, (y_*, μ_*) with $y_* = \sqrt{B}x_*$ and $\mu_* = \gamma(\frac{1}{\sqrt{\langle Bx_*, x_*\rangle}} - 1)$ is an eigenpair of the matrix $C = \sqrt{B}^{-1}A\sqrt{B}^{-1}$.

In this section, we set $r_1, \ldots, r_N$ to be linearly independent unit eigenvectors of C corresponding to the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$. We also let $q_j = \sqrt{B}^{-1}r_j$ for $1 \leq j \leq N$. Then, it is easy to see that $\{q_1, \ldots, q_N\}$ is an orthonormal basis for $\mathbb{R}^N$ with respect to the inner product $\langle x, y\rangle_B := x^T By$.

We can also apply the Gradient Descent method to (29) to solve (28) just as we did in Theorem 3.

Proposition 4. *Let A, B be as mentioned above. If we choose*

$$0 < \alpha < \frac{1}{(\mu_{(A,N)} + \gamma)(\mu_{(B,N)})^3},$$

then with a randomly chosen $x_0 \neq 0$,

$$x_{k+1} = x_k - \alpha\nabla F_{A,B}(x_k),\ k = 0, 1, 2, \ldots, \tag{30}$$

produce a sequence $\{x_k\}$ converging to x_* with $\langle Bx_*, x_*\rangle = \frac{\gamma}{\gamma + \lambda_*}$, where (x_*, λ_*) is a solution pair of (28) with $\lambda_* = \lambda_1 = \min\{\lambda \in \mathbb{R} : A - \lambda B$ is singular$\}$.

Proof. This proof mimics that of Theorem 3 with minor differences in detail. Let $\{x_k\}$ be the sequence generated by (30). First of all, we note that for any $x_k \in \mathbb{R}^N$,

$$\Big|\frac{\langle Ax_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle}\Big| \le \mu_{(A,N)}\mu_{(B,N)} \quad \text{and} \quad 1 = \mu_{(B,1)} \le \frac{\langle Bx_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle} \le \mu_{(B,N)},$$

which implies that if $0 < \alpha < \frac{1}{(\mu_{(A,N)}+\gamma)(\mu_{(B,N)})^3}$, then with $0 < \epsilon := 1 - \frac{1}{(\mu_{(B,N)})^2}$, we have

$$1 - \alpha\Big[\frac{\langle Ax_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle} + \gamma\frac{\langle Bx_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle}\Big] \ge 1 - \alpha(\mu_{(A,N)} + \gamma)\mu_{(B,N)} > \epsilon.$$

We can see that for $k = 1, 2, \ldots,$

$$\begin{aligned}
\langle Bx_{k+1}, x_{k+1}\rangle &\ge \frac{\langle Bx_{k+1}, x_k\rangle^2}{\langle Bx_k, x_k\rangle} = \Big(\frac{\langle x_k - \alpha\nabla F_A(x_k), Bx_k\rangle}{\sqrt{\langle Bx_k, x_k\rangle}}\Big)^2 \\
&= \Big(\sqrt{\langle Bx_k, x_k\rangle} - \alpha\Big[\frac{\langle Ax_k, Bx_k\rangle}{\sqrt{\langle Bx_k, x_k\rangle}} + \gamma\frac{\langle Bx_k, Bx_k\rangle}{\sqrt{\langle Bx_k, x_k\rangle}}\Big] + \alpha\gamma\frac{\langle Bx_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle}\Big)^2 \\
&= \Big[\sqrt{\langle Bx_k, x_k\rangle}\Big(1 - \alpha\Big[\frac{\langle Ax_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle} + \gamma\frac{\langle Bx_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle}\Big]\Big) + \alpha\gamma\frac{\langle Bx_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle}\Big]^2 \\
&\ge (\sqrt{\langle Bx_k, x_k\rangle}\epsilon + \alpha\gamma)^2 = \langle Bx_k, x_k\rangle\epsilon^2 + 2\sqrt{\langle Bx_k, x_k\rangle}\alpha\gamma\epsilon + (\alpha\gamma)^2. \qquad (31)
\end{aligned}$$

This proves that $\langle Bx_k, x_k\rangle \ge (\alpha\gamma)^2$ for $k \ge 1$, which enables us to improve the inequality in (31) as follows: for $k \ge 1$,

$$\begin{aligned}
\langle Bx_{k+1}, x_{k+1}\rangle &\ge (\sqrt{\langle Bx_k, x_k\rangle}\epsilon + \alpha\gamma)^2 = \langle Bx_k, x_k\rangle\epsilon^2 + 2\sqrt{\langle Bx_k, x_k\rangle}\alpha\gamma\epsilon + (\alpha\gamma)^2 \\
&\ge (\alpha\gamma)^2(1 + 2\epsilon) + \langle Bx_k, x_k\rangle\epsilon^2,
\end{aligned}$$

resulting in

$$\langle Bx_{k+1}, x_{k+1}\rangle \ge (\alpha\gamma)^2(1 + 2\epsilon)\sum_{l=0}^{k-1}\epsilon^{2l} + (\alpha\gamma)^2\epsilon^{2k}$$

Hence,

$$\liminf_{k\to\infty}\langle Bx_k, x_k\rangle \ge (\alpha\gamma)^2\frac{1+2\epsilon}{1-\epsilon^2} > \frac{(\alpha\gamma)^2}{1-\epsilon} = (\alpha\gamma\mu_{(B,N)})^2$$

and there exists $K \in \mathbb{N}$ such that $k \ge K$ implies

$$\langle Bx_k, x_k\rangle > (\alpha\gamma\mu_{(B,N)})^2.$$

Moreover, we can estimate $\langle B(x_k + t(x_{k+1} - x_k)), (x_k + t(x_{k+1} - x_k))\rangle$ for $t \in (0, 1)$ as follows: for $k > 1, 2, \ldots,$ and $0 < t < 1$,

$$\begin{aligned}
&\langle B(x_k + t(x_{k+1} - x_k)), (x_k + t(x_{k+1} - x_k))\rangle = \|\sqrt{B}(x_k - t\alpha\nabla F_A(x_k))\|^2 \\
&\ge \Big\langle\sqrt{B}(x_k - t\alpha\nabla F_{A,B}(x_k)), \frac{\sqrt{B}x_k}{\|\sqrt{B}x_k\|}\Big\rangle^2 = \Big(\frac{\langle x_k - t\alpha\nabla F_A(x_k), Bx_k\rangle}{\sqrt{\langle Bx_k, x_k\rangle}}\Big)^2 \\
&= \Big[\sqrt{\langle Bx_k, x_k\rangle}\Big(1 - t\alpha\Big[\frac{\langle Ax_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle} + \gamma\frac{\langle Bx_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle}\Big]\Big) + t\alpha\gamma\frac{\langle Bx_k, Bx_k\rangle}{\langle Bx_k, x_k\rangle}\Big]^2 \\
&\ge (\sqrt{\langle Bx_k, x_k\rangle}(1 - t\epsilon_1) + t\alpha\gamma)^2 = (\sqrt{\langle Bx_k, x_k\rangle} + t(\alpha\gamma - \sqrt{\langle Bx_k, x_k\rangle}\epsilon_1)^2 \\
&\ge \begin{cases} \langle Bx_k, x_k\rangle, & \text{if } \alpha\gamma \ge \sqrt{\langle Bx_k, x_k\rangle}\epsilon_1, \\ (\sqrt{\langle Bx_k, x_k\rangle}(1 - \epsilon_1) + \alpha\gamma)^2, & \text{if } \alpha\gamma < \sqrt{\langle Bx_k, x_k\rangle}\epsilon_1. \end{cases}
\end{aligned}$$

where $\epsilon_1 = 1 - \epsilon$. Noting that $\alpha\gamma < \sqrt{\langle Bx_k, x_k\rangle}\epsilon_1$ implies

$$\sqrt{\langle Bx_k, x_k\rangle}(1-\epsilon_1) + \alpha\gamma > \frac{\alpha\gamma(1-\epsilon_1)}{\epsilon_1} + \alpha\gamma = \frac{\alpha\gamma}{1-\epsilon},$$

we have that for $k \geq K$,

$$\min_{t\in[0,1]} \langle B(x_k + t(x_{k+1} - x_k)), (x_k + t(x_{k+1} - x_k))\rangle > (\alpha\gamma\mu_{(B,N)})^2.$$

Therefore, we have

$$\{x_k + t(x_{k+1} - x_k) : t \in [0,1], k \geq K\} \subset \{x \in \mathbb{R}^N : \langle Bx, x\rangle \geq (\alpha\gamma\mu_{(B,N)})^2\}.$$

Since $\sqrt{B}\Big(\frac{\gamma}{\sqrt{\langle Bx,x\rangle}}\Big[I - \Big(\frac{\sqrt{B}x}{\|\sqrt{B}x\|}\Big)\Big(\frac{\sqrt{B}x}{\|\sqrt{B}x\|}\Big)^T\Big]\Big)\sqrt{B}$ is positive semidefinite and

$$\nabla^2 F_{A,B}(x) = A + \gamma B - \sqrt{B}\Big(\frac{\gamma}{\sqrt{\langle Bx, x\rangle}}\Big[I - \Big(\frac{\sqrt{B}x}{\|\sqrt{B}x\|}\Big)\Big(\frac{\sqrt{B}x}{\|\sqrt{B}x\|}\Big)^T\Big]\Big)\sqrt{B},$$

we can see that

$$\|\nabla^2 F_{A,B}(x)\| \leq \max\Big\{\|A + \gamma B\|, \Big\|\sqrt{B}\Big(\frac{\gamma}{\sqrt{\langle Bx, x\rangle}}\Big[I - \Big(\frac{\sqrt{B}x}{\|\sqrt{B}x\|}\Big)\Big(\frac{\sqrt{B}x}{\|\sqrt{B}x\|}\Big)^T\Big]\Big)\sqrt{B}, \Big\|\Big\}.$$

We also note that

$$\|A + \gamma B\| \leq \mu_{(A,N)} + \gamma\mu_{(B,N)} < \frac{1}{\alpha},$$

and that for $x \in \{z \in \mathbb{R}^N : \langle Bz, z\rangle \geq (\alpha\gamma\mu_{(B,N)})^2\}$

$$\Big\|\sqrt{B}\Big(\frac{\gamma}{\sqrt{\langle Bx, x\rangle}}\Big[I - \Big(\frac{\sqrt{B}x}{\|\sqrt{B}x\|}\Big)\Big(\frac{\sqrt{B}x}{\|\sqrt{B}x\|}\Big)^T\Big]\Big)\sqrt{B}\Big\| \leq \frac{\mu_{(B,N)}}{\alpha\mu_{(B,N)}} = \frac{1}{\alpha}.$$

Therefore, for each $k \geq K$,

$$\sup_{t\in[0,1]} \|\nabla^2 F_{A,B}(x_k + t(x_{k+1} - x_k))\| \leq \frac{1}{\alpha}$$

and we finally have that for $k \geq K$,

$$\begin{aligned} F_{A,B}(x_{k+1}) &\leq F_{A,B}(x_k) + \langle \nabla F_{A,B}(x_k), x_{k+1} - x_k\rangle + \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 \\ &= F_{A,B}(x_k) - \frac{\alpha}{2}\|\nabla F_{A,B}(x_k)\|^2, \end{aligned}$$

which implies

$$\sum_{k=K}^{\infty} \|\nabla F_{A,B}(x_k)\|^2 \leq \frac{2}{\alpha}(F_{A,B}(x_K) - \min_{x\in\mathbb{R}^N}(F_{A,B}(x))).$$

We omit the rest of the proof because we can now follow the proof of Theorem 3 for the convergence of the sequence $\{x_k\}_{k>K}$ to x_* satisfying

$$\sqrt{\langle Bx_*, x_*\rangle} = \|\sqrt{B}x_*\| = \frac{\gamma}{\gamma + \lambda_*},$$

where (x_*, λ_*) is a solution pair of (28).

Let $r_1, \ldots, r_N$ be linearly independent unit eigenvectors of $C = \sqrt{B}^{-1} A \sqrt{B}^{-1}$ corresponding to the eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$. If we define q_j, $1 \leq j \leq N$ by $q_j = \sqrt{B}^{-1} r_j$ and represent x_k, $k \geq 1$, as

$$x_k = \mu_{k,1} q_1 + \cdots + \mu_{k,N} q_N,$$

then (30) becomes

$$\sum_{j=1}^{N} \mu_{k+1,j} q_j = \sum_{j=1}^{N} \mu_{k,j} q_j - \alpha \Big[\sum_{j=1}^{N} \mu_{k,j} A q_j + \gamma \Big(1 - \frac{1}{\|\sqrt{B} x_k\|} \Big) \sum_{j=1}^{N} \mu_{k,j} B q_j \Big].$$

Noting that $\{q_1, \ldots, q_N\}$ is an orthonormal basis for $\mathbb{R}^N$ with respect to the inner product $\langle x, x \rangle_B := x^T B x$, by taking the inner product $\Big\langle q_j, \sum_{j=1}^{N} \mu_{k+1,j} q_j \Big\rangle$ for each $j = 1, 2, \ldots, N$, we have

$$\mu_{k+1,j} = \mu_{k,j} \Big(1 - \alpha \Big(\lambda_j + \gamma - \frac{\gamma}{\|\sqrt{B} x_k\|} \Big) \Big) = \mu_{0,j} \Pi_{m=0}^{k} \Big(1 - \alpha \Big(\lambda_j + \gamma - \frac{\gamma}{\|\sqrt{B} x_m\|} \Big) \Big)$$

since $\langle q_j, A q_l \rangle = \langle r_j, C r_l \rangle = \lambda_j \delta_{jl}$. The rest of the proof is the same as that of Theorem 3. Hence, we have that

$$\lambda_* = \lambda_1 = \min\{\lambda \in \mathbb{R} : A - \lambda B \text{ is singular}\}.$$

□

If it is easy to compute B^{-1}, then we can do a little better than Proposition 4. Due to the equivalence between $Ax = \lambda B x$ and $B^{-1} A x = \lambda x$, all we need to do is to find eigenvectors and their corresponding eigenvalues of $E = B^{-1} A$. Since E may not be symmetric, we cannot use Theorem 3. However, realizing that $F_{A,B}$ can be rewritten using a different inner product and proposing a gradient descent method with respect to the new inner product, we can find an eigenvector of E corresponding to the smallest eigenvalue. We will also see that this method allows for a bigger stepsize.

Proposition 5. *With $F_{A,B}$ defined as in (29) and $\gamma > \max(0, -\mu_{(A,1)})$, a sequence $\{x_k\}$ generated by*

$$x_{k+1} = x_k - \alpha B^{-1} \nabla F_{A,B}(x_k), \ k = 0, 1, 2 \ldots, \tag{32}$$

with a randomly chosen $x_0 \neq 0$ and $0 < \alpha < \frac{1}{\mu_{(A,N)} + \gamma}$, converges to x_*, where (x_*, λ_*) is a solution pair of (28) satisfying

$$\sqrt{\langle B x_*, x_* \rangle} = \|\sqrt{B} x_*\| = \frac{\gamma}{\gamma + \lambda_*}$$

and

$$\lambda_* = \min\{\lambda : A - \lambda B \text{ is singular}\}.$$

In fact, (32) is the gradient descent of $F_{A,B}$ with respect to the inner product $\langle x, y \rangle_B := x^T B y$.

Proof. First of all, as we mentioned, (28) is equivalent to

$$B^{-1} A x = \lambda x,$$

and to

$$C y = \lambda y, \quad \text{with } C = \sqrt{B}^{-1} A \sqrt{B}^{-1}, \ y = \sqrt{B} x.$$

Therefore, (x, λ) is a solution pair of (28) if and only if x is an eigenvector of $B^{-1}A$ corresponding to the eigenvalue λ if and only if $\sqrt{B}x$ is an eigenvector of C corresponding to the eigenvalue λ. Note that

$$\lambda_1 = \min_{\|y\|=1} \langle y, Cy\rangle = \frac{\langle x, Ax\rangle}{\|x\|^2}\|x\|^2 \geq \mu_{(A,1)}\|x\|^2 \geq \begin{cases} \mu_{(A,1)}, & \text{if } \mu_{(A,1)} < 0, \\ 0, & \text{if } \mu_{(A,1)} \geq 0, \end{cases}$$

where $x = \sqrt{B}^{-1}y$ and $\|x\| = \|\sqrt{B}^{-1}y\| \leq \frac{1}{\sqrt{\mu_{(B,1)}}} = 1$. Since $\gamma > \max(0, -\mu_{(A,1)})$, if we set $G_C(y)$ to be

$$G(y) = \frac{1}{2}\langle Cy, y\rangle + \frac{\gamma}{2}\|y\|^2 - \gamma\|y\|,$$

and apply Theorem 3 to G, then since λ_N, the largest eigenvalue of C, is at most $\mu_{(A,N)} > 0$, a sequence $\{y_k\}$ generated by

$$y_{k+1} = y_k - \alpha\nabla G(y_k),\ k = 0, 1, 2, \ldots,$$

with a randomly chosen $y_0 \neq 0$, and any $0 < \alpha < \frac{1}{\lambda_N+\gamma} \leq \frac{1}{\mu_{(A,N)}+\gamma}$, converges to y_*, an eigenvector of C corresponding to the smallest eigenvalue $\lambda_* = \lambda_1$ with norm $\|y_*\| = \frac{\gamma}{\gamma+\lambda_1}$.

On the other hand, with $y_k = \sqrt{B}x_k$, $k = 0, 1, \ldots$, we have

$$\begin{aligned}\nabla G_C(y_k) &= Cy_k + \gamma\Big(1 - \frac{1}{\|y_k\|}\Big)y_k = (\sqrt{B})^{-1}Ax_k + +\gamma\Big(1 - \frac{1}{\sqrt{\langle Bx_k, x_k\rangle}}\Big)\sqrt{B}x_k \\ &= (\sqrt{B})^{-1}\nabla F_{A,B}(x_k),\end{aligned}$$

which implies that

$$y_{k+1} = y_k - \alpha\nabla G_C(y_k)$$

is equivalent to (32), i.e., $x_{k+1} = x_k - \alpha B^{-1}\nabla F_{A,B}(x_k)$.

Hence, the sequence generated by (32), with a randomly chosen $x_0 \neq 0$, converges to x_*, where (x_*, λ_*) is a solution pair of (28) satisfying $\sqrt{\langle Bx_*, x_*\rangle} = \|y_*\| = \frac{\gamma}{\gamma+\lambda_*}$ and

$$\lambda_* = \lambda_1 = \min\{\lambda : A - \lambda B \text{ is singular}\}.$$

In addition, with respect to the inner product $\langle x, y\rangle_B := x^T By$, we note that

$$B^{-1}\nabla F_{A,B}(x) = B^{-1}Ax + \gamma\Big(1 - \frac{1}{\sqrt{\langle Bx, x\rangle}}\Big)x = B^{-1}Ax + \gamma\Big(1 - \frac{1}{\sqrt{\langle x, x\rangle_B}}\Big)x,$$

which is the gradient of $F_{A,B}$ with respect to the inner product $\langle\cdot,\cdot\rangle_B$ because

$$F_{A,B}(x) = \frac{1}{2}\langle x, B^{-1}Ax\rangle_B + \frac{\gamma}{2}\|x\|_B^2 - \gamma\|x\|_B,$$

where $\|x\|_B^2 = \langle x, x\rangle_B$. Note that $B^{-1}A$ is self-adjoint with respect to $\langle\cdot,\cdot\rangle_B$. Therefore, (32) is the gradient descent of $F_{A,B}$ with respect to $\langle\cdot,\cdot\rangle_B$. □

In general, with a nonsymmetrix square matrix E, minimizing the functional F_E in any of the forms in (3) does not guarantee to find an eigenvector of E. However, Proposition 4 and Proposition 5 make it possible in certain cases when E decomposes into $E = B^{-1}A$ with a symmetric matrix A and a symmetric positive definite matrix B. If we know such matrices A, B, then Proposition 4 applies to find a global minimizer of F_E. Moreover, if it is easy to compute B^{-1}, then Proposition 5 applies to find a global minimizer of F_E with a better stepsize. We can also find subsequent eigenvectors in the same way as presented in Theorem 2.

Proposition 6. *Let* $(\mathbf{x}_1, \lambda_1), \ldots, (\mathbf{x}_m, \lambda_m)$ *be solution pairs of* (28) *where* $\lambda_1 \leq \cdots \leq \lambda_m$ *are the first* $m \geq 1$ *smallest ones in* $\{\lambda \in \mathbb{R} : A - \lambda B \ \text{ is singular}\}$.

We consider the following problem

$$\min_{x \in \mathbb{R}^N} F_{A,B}(x) \quad \text{subject to} \quad \langle x, \mathbf{x}_i \rangle_B = 0, \ i = 1, 2, \ldots, m. \tag{33}$$

Then, any local minimizer x_* *of* (7) *is a global minimizer with* $\sqrt{\langle Bx_*, x_* \rangle} = \frac{\gamma}{\gamma + \lambda_*}$ *and*

$$\lambda_* = \min(\{\lambda \in \mathbb{R} : A - \lambda B \ \text{ is singular}\} \setminus \{\lambda_1, \ldots, \lambda_m\}).$$

Proof. We omit the proof because the proof is not so much different from that of Theorem 2. □

As was done with the real symmetric case, we can also consider a similar nonsingular system whose unique solution solves (28). Please note that the following proposition is, in fact, the same as Corollary 1.

Proposition 7. *Let* $\tilde{\lambda}$ *be a generalized eigenvalue of the pencil* $A - \lambda B$ *of multiplicity 1. Let* $\gamma > 0$. *Let* x_0 *be chosen uniformly at random from* S^{N-1}. *Then, with probability 1, there is a unique solution* $\tilde{x}$ *to*

$$(A - \tilde{\lambda} B + \gamma x_0 x_0^T) x = \gamma x_0$$

so that $(\tilde{x}, \tilde{\lambda})$ *is a solution to* (28).

When a generalized eigenvalue has multiplicity more than 1, we can also apply Corollary 1 to find approximate eigenvalues.

3.2 Eigenvalue problems on infinite dimensional spaces

It is interesting to see that the same framework as (3) applies to eigenvalue problems on infinite dimensional spaces such as the Sturm-Liouville eigenvalue problem, the eigenvalue problem of self-adjoint elliptic operators, etc. We will present one such application, that is, finding an eigenfunction of a self-adjoint uniformly elliptic operator corresponding to the smallest eigenvalue.

3.2.1 Symmetric Uniformly Elliptic Operators

Let Ω be a bounded open subset of $\mathbb{R}^d$ with Lipschitz boundary $\partial\Omega$. We will denote by L a symmetric uniformly elliptic operator defined by

$$Lu = -\sum_{i,j=1}^{d} \partial_i (a_{i,j} \partial_j u) + cu, \tag{34}$$

where $a_{i,j}, c \in L^\infty(\Omega)$, $i, j = 1, \ldots, d$, are such that $a_{i,j}(x) = a_{j,i}(x)$ a.e., and $c(x) \geq 0$ a.e., and there exists $0 < \alpha \leq \beta < \infty$ such that for a.e. $x \in \Omega$ and for $\xi \in \mathbb{R}^d$,

$$\alpha |\xi|^2 \leq \sum_{i,j=1}^{d} a_{i,j}(x) \xi_i \xi_j \leq \beta |\xi|^2.$$

The problem that we are interested in is to find $\varphi \in H_0^1(\Omega)$ that solves

$$\begin{cases} L\varphi = \lambda \varphi, & \text{in } \ \Omega, \\ \varphi = 0, & \text{on } \partial\Omega. \end{cases} \tag{35}$$

It is known that the eigenvalues $\lambda_1, \lambda_2, \dots$ of L are nonnegative and we may have them in an increasing order, that is,

$$0 < \lambda_1 < \lambda_2 \le \cdots \le \lambda_3 \le \cdots .$$

Therefore, a natural question to ask is to find the smallest eigenvalue λ_1 of L and its corresponding eigenfuction. It will be apparent that the same analysis applies with the Neumann boundary condition. We would like to stress that a fundamental problem that our method can solve is to find eigenvalues and eigenfunctions of the Laplace operator $-\Delta$ on any domain Ω with Lipschitz boundary.

Definition 1. $\varphi \in H_0^1(\Omega)$ is a weak solution of (35) if for any $\psi \in C_0^\infty(\Omega)$,

$$\int_\Omega \sum_{i,j=1}^d a_{i,j}(x)\partial_j\varphi(x)\partial_i\psi(x)dx + \int_\Omega (c(x)-\lambda)\varphi(x)\psi(x)dx = 0.$$

Then, it is natural from the discussion in the previous sections that we want to define a functional F_L with $\gamma > 0$ by

$$\begin{aligned} F_L(u) = \frac{1}{2}\int_\Omega \sum_{i,j=1}^d a_{i,j}(x)\partial_j u(x)\partial_i u(x)dx + \frac{1}{2}\int_\Omega c(x)|u(x)|^2dx \\ + \frac{\gamma}{2}\int_\Omega |u(x)|^2dx - \gamma\Big(\int_\Omega |u(x)|^2dx\Big)^{\frac{1}{2}} \end{aligned} \tag{36}$$

and solve the following minimization problem

$$\min_{u\in H_0^1(\Omega)} F_L(u) \tag{37}$$

and investigation the relation between (35) and (37). The existence of a minimizer of the problem (37) is obvious by the standard method using the compact embedding theorem by Rellich-Kondrachov. Moreover, we can observe the same characteristics of (36) as those of (3). For theorems and lemmas that follow, we will omit their proofs because they are essentially the same as what we presented in the finite dimensional case.

Lemma 2. *The set of nonzero critical points of F_L is*

$$\Big\{\varphi \in H_0^1(\Omega) : L\varphi = \lambda\varphi \text{ in } \Omega \quad \text{and} \quad \|\varphi\|_{L^2(\Omega)} = \frac{\gamma}{\gamma+\lambda}\Big\}.$$

and

$$\min_{u\in H_0^1(\Omega)} F_L(u) = -\frac{\gamma^2}{2(\gamma+\lambda_*)},$$

where $\lambda_* \ge 0$ is the smallest eigenvalue of L.

Theorem 6. *Any local minimizer of* (36) *is a global minimizer.*

3.2.2 Corresponding parabolic PDEs

Corresponding to the uniformly elliptic PDEs of the form (34), we will consider the following parabolic PDE: for $T \in (0,\infty)$, we solve

$$\begin{cases} \frac{\partial u}{\partial t} &= -Lu - \gamma\Big(1 - \frac{1}{\|u(t)\|_2}\Big)u \ \text{ in } \ \Omega_T := \Omega\times(0,T], \\ u &= 0 \ \text{ on } \ \partial\Omega_T := \partial\Omega\times[0,T], \\ u(0) &= u_0 \neq 0 \text{ in } H_0^1(\Omega), \end{cases} \tag{38}$$

where $\|u(t)\|_2 = (\int_\Omega |u(x,t)|^2 dx)^{\frac{1}{2}}$. Due to the condition $c \geq 0$ in Ω, L is positive definite. In the context, we will use $\langle \cdot, \cdot \rangle$ for both inner products in $L^2(\Omega)$ and in $\mathbb{R}^N$ for $N \in \mathbb{N}$. Note that the partial differential equation in (38) is the formal gradient flow of a functional F_L in (36), which is a difference of convex functionals, i.e.,

$$\frac{\partial u}{\partial t} = -\nabla F_L(u).$$

Theorem 7. *The problem* (38) *has a unique weak solution u for any $T \in (0, \infty)$. Moreover, if ψ_1 is an eigenfunction of L corresponding to the smallest eigenvalue $\lambda_1 > 0$ with $\|\psi_1\|_2 = 1$, and if $\langle u_0, \psi_1 \rangle \neq 0$, then the solution u satisfies*

$$u(t) \to \frac{\gamma}{\gamma + \lambda_1} \mathbf{v} \text{ in } L^2(\Omega) \text{ as } t \to \infty \quad \text{for some } \mathbf{v} \in \{\pm \psi_1\}.$$

Proof. The proof is inspired by the Galerkin's method. Note that we can find an orthonormal basis $\{\psi_k\}_{k \in \mathbb{N}}$ for $L^2(\Omega)$ such that ψ_k is an eigenfunction of L corresponding to the k^{th} smallest eigenvalue λ_k with $\psi_k \in H_0^1(\Omega)$. Then, we let V_N, $N \in \mathbb{N}$, be the subspace of $L^2(\Omega)$ spanned by $\{\psi_1, \ldots, \psi_N\}$ and let P_N be the projection of $L^2(\Omega)$ onto V_N.

Suppose that $u_0 \in H_0^1(\Omega)$ is given and satisfies $\langle u_0, \psi_1 \rangle \neq 0$. We then consider

$$\begin{cases} \frac{\partial u}{\partial t} &= -Lu - \gamma\Big(1 - \frac{1}{\|u(t)\|_2}\Big)u \text{ in } \Omega_T, \\ u &= 0 \text{ on } \partial\Omega_T, \\ u(0) &= P_N(u_0). \end{cases} \tag{39}$$

If $u_N \in L^2([0,T]; H_0^1(\Omega))$ with $\frac{d}{dt} u_N \in L^2([0,T]; H^{-1}(\Omega))$ is a solution of (39), then for $k = 1, \ldots, N$, we have

$$\frac{d}{dt}\langle u_N, \psi_k \rangle = -(\lambda_k + \gamma)\langle u_N, \psi_k \rangle + \frac{\gamma}{\|u_N(t)\|_2}\langle u_N, \psi_k \rangle$$

and

$$\frac{1}{2}\frac{d}{dt}\langle u_N, \psi_k \rangle^2 = -(\lambda_k + \gamma)\langle u_N, \psi_k \rangle^2 + \frac{\gamma}{\|u_N(t)\|_2}\langle u_N, \psi_k \rangle^2$$

with $\langle u_N(0), \psi_k \rangle = \langle P_N(u_0), \psi_k \rangle$. Considering $\phi_k(t) = \langle u_N(t), \psi_k \rangle$, $k = 1, \ldots, N$, we can see that $\phi_1, \ldots, \phi_N$ solve (54). Since $\Phi_N = [\phi_1 \ \cdots \ \phi_N]^T$ satisfies that for $t \in [0, \infty)$,

$$\|u_N(t)\|_2 \geq \|\Phi_N(t)\| > \omega = \frac{1}{2}\min\Big(\frac{\gamma}{\gamma + \lambda_N}, \|\Phi_N(0)\|\Big). \tag{40}$$

Therefore, the existence of a solution to (39) can be obtained by a linear system of ODEs (54) given in the Appendix below, with the initial condition

$$\Phi_N(0) = [\langle P_N(u_0), \psi_1 \rangle \ \cdots \ \langle P_N(u_0), \psi_N \rangle]^T,$$

and setting

$$u_N(x,t) = \phi_1(t)\psi_1(x) + \cdots + \phi_N(t)\psi_N(x).$$

Then, u_N is a weak solution of (39) such that

$$u_N(t) \to \frac{\gamma}{\gamma + \lambda_1}\mathbf{v} \text{ in } L^2(\Omega) \text{ as } t \to \infty,$$

where $\phi_1, \ldots, \phi_N$ satisfy (54) and $\mathbf{v} \in L^2(\Omega)$ is either ψ_1 or $-\psi_1$.

Suppose now that $u_{N,1}, u_{N,2}$ are two such solutions of (39). Let $v_N = u_{N,1} - u_{N,2}$. Then, we have

$$\frac{\partial v_N}{\partial t} = -Lv_N - \gamma(f(u_{N,1}) - f(u_{N,2})), \tag{41}$$

where $f(u(x,t)) = u(x,t) - \frac{u(x,t)}{\|u(t)\|_2}$. Since f on $L^2(\Omega)$ is Lipschitz with a Lipschitz constant $\mu = (\frac{2}{\omega} - 1)$ on $\{u : \|u\|_2 \geq \omega\}$ and the two solutions $u_{N,1}, u_{N,2}$ satisfy (40) for all $t \geq 0$, by taking the inner product on $L^2(\Omega)$ with v_N on both sides of (41), we have

$$\frac{1}{2}\frac{d}{dt}\langle v_N(t), v_N(t)\rangle \leq \langle -Lv_N(t), v_N(t)\rangle + \gamma\mu\langle v_N(t), v_N(t)\rangle,$$

which implies

$$\frac{1}{2}\frac{d}{dt}\langle v_N, v_N\rangle - \gamma\mu\langle v_N, v_N\rangle \leq -\langle Lv_N, v_N\rangle \leq 0. \tag{42}$$

Therefore, $\|v_N(0)\|_2 = 0$ implies $u_{N,1} = u_{N,2}$ for a.e. $(x,t) \in \Omega \times [0,T]$ for any $T \in (0,\infty)$. In fact, we know that

$$u_N \in C^1([0,\infty); H_0^1(\Omega))$$

as well as

$$u_N \in L^2([0,\infty); H_0^1(\Omega)), \quad \frac{d}{dt}u_N \in L^2([0,\infty); H^{-1}(\Omega)).$$

We now solve (38). Firstly, we fix $k_0 \in \mathbb{N}$ so that $\frac{\|u_0\|_2}{\sqrt{2}} < \|P_{k_0}(u_0)\|_2$ with $\lambda_{k_0} < \lambda_{k_0+1}$ and obtain the solution u_N of (39) with $N > k_0$. We may write u_N as

$$u_N(t) = \phi_{1,N}(t)\psi_1 + \cdots + \phi_{N,N}(t)\psi_N.$$

Let $\varphi_{1,N}(t) = \phi_{1,N}^2(t) + \cdots + \phi_{k_0,N}^2(t)$ and $\varphi_{2,N}(t) = \phi_{k_0+1,N}^2(t) + \cdots + \phi_{N,N}^2(t)$. Then, we have

$$\frac{1}{2}\frac{d\varphi_{1,N}}{dt} \geq -(\lambda_{k_0} + \gamma)\varphi_{1,N} + \frac{\gamma}{\|\Phi_N(t)\|}\varphi_{1,N}, \tag{43}$$

$$\frac{1}{2}\frac{d\varphi_{2,N}}{dt} \leq -(\lambda_{k_0+1} + \gamma)\varphi_{2,N} + \frac{\gamma}{\|\Phi_N(t)\|}\varphi_{2,N}. \tag{44}$$

This implies that

$$\varphi_{1,N}(t) \geq \varphi_{1,N}(0)e^{-2(\lambda_{k_0}+\gamma)t + \int_0^t \frac{2\gamma}{\|\Phi_N(s)\|}ds} \geq \varphi_{2,N}(0)e^{-2(\lambda_{k_0+1}+\gamma)t + \int_0^t \frac{2\gamma}{\|\Phi_N(s)\|}ds} \geq \varphi_{2,N}(t)$$

i.e.,

$$\frac{\varphi_{1,N}(t)}{\varphi_{2,N}(t)} \geq \frac{\varphi_{1,N}(0)}{\varphi_{2,N}(0)}e^{2(\lambda_{k_0+1}-\lambda_{k_0})t} > e^{2(\lambda_{k_0+1}-\lambda_{k_0})t} \tag{45}$$

due to $\varphi_{1,N}(0) > \frac{\|u_0\|_2^2}{2} > \varphi_{2,N}(0)$. Let

$$\begin{cases} M_1 &= 2\max\left(\frac{\gamma}{\gamma+\lambda_1}, \|u_0\|_2\right), \\ M_2 &= \frac{1}{\sqrt{2}}\min\left(\frac{\gamma}{\lambda_{k_0}+\gamma}, \|u_0\|_2\right). \end{cases}$$

Note that

$$\frac{1}{2}\frac{d\varphi_{1,N}}{dt} \leq -(\lambda_1 + \gamma)\varphi_{1,N} + \frac{\gamma}{\|\Phi_N(t)\|}\varphi_{1,N}$$

implies $\varphi_{1,N}(t) \le \|\Phi_N(t)\|^2 < M_1^2$. And (45) implies

$$\varphi_{2,N}(t) < M_1^2 e^{-2(\lambda_{k_0+1}-\lambda_{k_0})t}. \tag{46}$$

Since we already saw that $\phi_{j,N}(t)$, $j = 1, 2, \dots, N$, exist for all $t \in (0, \infty)$, if we suppose

$$t_0 = \inf\{t \in (0,\infty) | \varphi_{1,N}(t) = M_2^2\} < \infty,$$

then from (43) and (45), we have

$$\|\Phi_N(t_0)\|^2 = \varphi_{1,N}(t_0) + \varphi_{2,N}(t_0) < M_2^2(1 + e^{-2(\lambda_{k_0+1}-\lambda_{k_0})t_0}) < 2M_2^2 \le \Big(\frac{\gamma}{\lambda_{k_0}+\gamma}\Big)^2.$$

Therefore, there exists $\delta > 0$ such that for $t \in (t_0 - \delta, t_0)$,

$$\|\Phi_N(t)\| < \frac{\gamma}{\lambda_{k_0}+\gamma},$$

which implies that

$$\frac{1}{2}\frac{d\varphi_{1,N}}{dt}(t) \ge -(\lambda_{k_0}+\gamma)\varphi_{1,N}(t) + \frac{\gamma}{\|\Phi_N\|}\varphi_{1,N}(t) > 0 \ \text{ on } (t_0-\delta, t_0).$$

This is a contradiction since $\varphi_{1,N}(t) > M_2^2$ for $t \in [0, t_0)$. Therefore, we end up with

$$M_2^2 < \varphi_{1,N}(t) < M_1^2 \ \text{ for } t \ge 0. \tag{47}$$

Note that (47) implies

$$\inf_{t\ge 0} \|\Phi_N(t)\| \ge M_2$$

That is, the solution u_N for $N > k_0$ satisfies

$$\inf_{t\ge 0} \|u_N(t)\|_2 \ge M_2. \tag{48}$$

We fix $T \in (0, \infty)$ and consider a sequence of solutions $\{u_N(t)\}_{N>k_0}$, where u_N is the solution of (39) with $u_N(0) = P_N(u_0)$. Using $v = u_N - u_M$ with $N, M > k_0$ in place of v_N in (42), and noting that we may choose $\omega = M_2$ for $\mu = (\frac{2}{\omega} - 1)$, we obtain

$$\|u_N(t) - u_M(t)\|_2 \le e^{\gamma\mu t}\|u_N(0) - u_M(0)\|_2 \le e^{\gamma\mu T}\|u_N(0) - u_M(0)\|_2 \ \text{ for } t \in [0,T], \tag{49}$$

which implies that for any $m \in \mathbb{N}$, $\{\phi_{m,N}(t)\}_{N>k_0}$ converges uniformly to $\phi_{m,*}(t)$ in $[0,T]$ and

$$\|\Phi_N(t)\| \to \|\Phi_*(t)\| \text{ uniformly on } [0,T] \text{ as } \ N \to \infty,$$

where $\phi_{m,N}(t) = \langle u_N(t), \psi_m\rangle$ and $\|\Phi_*(t)\| = \sqrt{\sum_{m=1}^{\infty}\phi_{m,*}^2(t)}$. Hence, for all m, we have

$$\frac{d}{dt}\phi_{m,*} = -(\lambda_m+\gamma)\phi_{m,*} + \frac{\gamma}{\|\Phi_*\|}\phi_{m,*}.$$

In addition, for $k_0 < m \le N$, if we consider

$$\omega_{m,N}(t) = \phi_{m,N}^2(t) + \cdots + \phi_{N,N}^2(t),$$

then we can obtain, by the same argument for (46),

$$\omega_{m,N}(t) < M_1^2 e^{-2(\lambda_m-\lambda_{k_0})t}, \ t \ge 0 \tag{50}$$

implying $|\phi_{m,N}(t)| < M_1 e^{-(\lambda_m - \lambda_{k_0})t}$ on $[0,T]$ and, eventually, on $(0,\infty)$. By a slight modification on (46), we can see that even for each $2 \le m \le k_0$, there exists $\zeta_m > 0$ such that $|\phi_{m,N}(t)| < \zeta_m M_1 e^{-(\lambda_m - \lambda_1)t}$ on $[0,\infty)$, i.e., with $\eta_{k_0} = \max_{k=2,\dots,k_0}(\zeta_k)$,

$$|\phi_{m,N}(t)| < \begin{cases} \eta_{k_0} M_1 e^{-(\lambda_m - \lambda_1)t}, & 2 \le m \le k_0, \\ M_1 e^{-(\lambda_m - \lambda_{k_0})t}, & m > k_0. \end{cases} \tag{51}$$

This implies that

$$\Big\{ \sum_{m=1}^{N} \lambda_m \phi_{m,N}^2 \Big\}_{N\in\mathbb{N}} \text{ converges uniformly to } \sum_{m=1}^{\infty} \lambda_m \phi_{m,*}^2 \text{ in } [0,T] \text{ as } N \to \infty.$$

Hence, by defining

$$u_*(x,t) = \sum_{m=1}^{\infty} \phi_{m,*}(t)\psi_m(x),$$

we can easily see that $u_* \in L^2([0,T]; H_0^1(\Omega))$ and

$$\frac{d}{dt}u_* = \sum_{m=1}^{\infty} \frac{d\phi_{m,*}}{dt}(t)\psi_m(x) \in L^2([0,T]; H^{-1}(\Omega))$$

and that for a.e. $t \in [0,T]$, and for each $v \in H_0^1(\Omega)$,

$$\int_\Omega \frac{\partial}{\partial t} u_*(x,t)v(x)dx + \int_\Omega \sum_{i,j=1}^{d} a_{i,j}(x)\partial_j u_*(x)\partial_i v(x)dx + \int_\Omega c(x)u_*(x)v(x)dx$$
$$\gamma \int_\Omega u_*(x)v(x)dx - \frac{\gamma}{\|u_*(t)\|_2}\int_\Omega u_*(x)v(x)dx = 0$$

with $u_*(0) = u_0$ in $L^2(\Omega)$. Therefore, u_* is a weak solution of (38) for any $T \in (0,\infty)$. Next, using the functional in (36), we define $f(t)$ for $t \in (0,\infty)$ by

$$f(t) = F_L(u_*) = \frac{1}{2}\int_\Omega \sum_{i,j=1}^{d} a_{ij}(x)\partial_j u_*(x,t)\partial_i u_*(x,t)dx + \frac{1}{2}\int_\Omega c(x)|u_*(x,t)|^2 dx$$
$$+ \frac{\gamma}{2}\int_\Omega |u_*(x,t)|^2 dx - \gamma\Big(\int_\Omega |u_*(x,t)|^2 dx\Big)^{\frac{1}{2}},$$

Note that

$$f(t) = \frac{1}{2}\sum_{m=1}^{\infty} \lambda_m \phi_{m,*}^2(t) + \frac{\gamma}{2}\sum_{m=1}^{\infty} \phi_{m,*}^2(t) - \gamma\|\Phi_*(t)\|.$$

Since $\phi_{m,*}^2(t)$ decays exponentially to 0 in $(0,\infty)$ as $m \to \infty$, we can see that

$$\begin{aligned} \frac{df}{dt}(t) &= \sum_{m=1}^{\infty} \Big(\lambda_m \phi_{m,*}(t) + \gamma\phi_{m,*}(t) - \gamma\frac{\phi_{m,*}(t)}{\|\Phi_*(t)\|}\Big)\frac{d\phi_{m,*}}{dt}(t) \\ &= -\sum_{m=1}^{\infty} \Big|\frac{d\phi_{m,*}}{dt}(t)\Big|^2. \end{aligned}$$

Therefore, f is non-increasing and bounded below, i.e., $a = \lim_{t\to\infty} f(t)$ exists. From the fact that for all $N \in \mathbb{N}$,

$$\|\Phi_N(t)\| \to \frac{\gamma}{\gamma+\lambda_1} \quad \text{as } t \to \infty,$$

together with (51), we can see that

$$\|\Phi_*(t)\| \to \frac{\gamma}{\gamma+\lambda_1} \quad \text{as} \quad t \to \infty.$$

which eventually implies that for each $m \in \mathbb{N}$, $|\frac{d\phi_{m,*}}{dt}(t)| \to 0$ as $t \to \infty$, i.e.,

$$\lim_{t\to\infty} \Big((\lambda_m + \gamma) - \frac{\gamma}{\|\Phi_*(t)\|}\Big)\phi_{m,*}(t) = 0. \tag{52}$$

Moreover, since we have $|\phi_{m,*}(t)| \le \max_{k=1,\dots,k_0}\{M_1\zeta_k e^{-(\lambda_2-\lambda_1)t}\}$ for all $m \ge 2$ with $\zeta_1 = 1$ and

$$\inf_{t\ge 0} \|\Phi_*(t)\| \ge M_2,$$

we finally obtain, together with (52),

$$\lim_{t\to\infty} \|\Phi_*(t)\| = \frac{\gamma}{\lambda_1+\gamma},$$

which implies that $\lim_{t\to\infty} |\phi_{1,*}(t)| = \frac{\gamma}{\lambda_1+\gamma}$, and

$$u_*(t) \to \frac{\gamma}{\lambda_1+\gamma}\mathbf{v} \text{ in } L^2(\Omega) \quad \text{as} \quad t \to \infty$$

for some $\mathbf{v} \in \{\pm\psi_1\}$.

□

4 Appendix: Corresponding ODEs

For a given symmetric positive semidefinite matrix A in $\mathbf{M}_N(\mathbb{R})$, as we saw in (53) and in (6), with an orthogonal matrix Q and a diagonal matrix $\Lambda_N = diag(\lambda_1, \dots, \lambda_N)$, $0 \le \lambda_1 = \cdots = \lambda_p < \lambda_{p+1} \le \cdots \le \cdots \le \lambda_N$ for some $1 \le p < N$, we have $A = Q\Lambda_N Q^T$ and

$$F_A(\mathbf{x}) = F_{\Lambda_N}(\mathbf{y}), \quad \text{with } \mathbf{y} = Q^T\mathbf{x}. \tag{53}$$

The gradient descent flow associated with the functional F_{Λ_N} is

$$\frac{d}{dt}\Phi_N(t) = -\nabla F_{\Lambda_N}(\Phi_N(t)),$$

and we are interested in the existence of a solution of (54). So we solve on $(0,\infty)$

$$\begin{cases} \frac{d}{dt}\Phi_N(t) &= -\nabla F_{\Lambda_N}(\Phi_N(t)) = -(\Lambda_N + \gamma I)\Phi_N(t) + \frac{\gamma}{\|\Phi_N(t)\|}\Phi_N(t), \\ \Phi_N(0) &\in \mathbb{R}^N,\ \phi_j(0) \ne 0 \ \text{ for some } 1 \le j \le p, \end{cases} \tag{54}$$

where $\Phi_N = \Phi_N(t) = [\phi_1(t) \ \cdots \ \phi_N(t)]^T$ and $\|\Phi_N(t)\| = \sqrt{\sum_{k=1}^N \phi_k^2(t)}$.

Theorem 8. *There exist a unique solution* $\Phi_N \in C^1([0,\infty))$ *of* (54) *and* $\mathbf{v} \in \mathbb{R}^N$ *such that*

$$\Phi_N(t) \to \mathbf{v} \quad as \quad t \to \infty,$$

where $\mathbf{v}$ *is an eigenvector of* Λ_N *corresponding to the eigenvalue* λ_1 *with* $\|\mathbf{v}\| = \frac{\gamma}{\gamma+\lambda_1}$ *and* $\langle \mathbf{v}, \mathbf{e}_j \rangle \neq 0$. *Note that* $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ *is the standard basis for* $\mathbb{R}^N$.

Proof. The existence and uniqueness of a solution Φ_N on $[0,\infty)$ is easily guaranteed by the theory of ODEs once we establish a lower bound $\omega > 0$ for $\|\Phi_N(t)\|$, $t \in [0,\infty)$. Due to $\phi_j(0) \neq 0$, i.e., $\|\Phi_N(0)\| \neq 0$, a solution exists and is unique on $[0,\epsilon]$ with $\epsilon << 1$. Then, by setting

$$T_{max} = \sup\{T \in (0,\infty) : \text{a solution } \Phi_N \text{ exists and is unique on } [0,T]\},$$

we know that $T_{max} \geq \epsilon$. Note that for $t \in (0, T_{max})$ and for $1 \leq k \leq N$, we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\phi_k^2(t) = & -(\lambda_k + \gamma)\phi_k^2(t) + \frac{\gamma}{\|\Phi_N(t)\|}\phi_k^2(t) \\ > & -(\lambda_N + \gamma)\phi_k^2(t), \end{aligned}$$

resulting in $\phi_k(t) \neq 0$ for $t \geq 0$ if and only if $\phi_k(0) \neq 0$. With $\omega = \frac{1}{2}\min(\frac{\gamma}{\gamma+\lambda_N}, \|\Phi_N(0)\|)$, we suppose $\{t \in (0, T_{max}) : \|\Phi_N(t)\| \leq \omega\} \neq \emptyset$. Then, $T_\omega = \inf\{t \in (0,\epsilon) : \|\Phi_N(t)\| \leq \omega\}$ exists in $(0, T_{max})$. Since there exists $\delta > 0$ such that $\omega < \|\Phi_N(t)\| < \frac{3}{2}\omega$ on $(T_\omega - \delta, T_\omega)$, we have

$$\frac{1}{2}\frac{d}{dt}\|\Phi_N(t)\|^2 \geq \frac{1}{3}(\lambda_N + \gamma)\|\Phi_N(t)\|^2 > 0 \quad \text{on } (T_\omega - \delta, T_\omega),$$

that is, $\|\Phi_N(t)\|$ is increasing in $(T_\eta - \delta, T_\eta)$ and $\lim_{t \to T_\omega} \|\Phi_N(t)\| > \omega = \|\Phi_N(T_\omega)\|$. This is a contradiction. Therefore, we have

$$\inf_{t \in (0, T_{max})} \|\Phi_N(t)\| > \omega,$$

which eventually implies that $T_{max} = \infty$.

In addition, we note that

$$\frac{d}{dt}F_{\Lambda_N}(\Phi_N) = \Big\langle \frac{d}{dt}\Phi_N, \Lambda_N\Phi_N + \gamma\Phi_N - \frac{\gamma}{\|\Phi_N\|}\Phi_N \Big\rangle = -\Big\| \frac{d\Phi_N}{dt} \Big\|^2$$

and that since $F_{\Lambda_N}(\Phi_N)$ is bounded below, there exists $y \in \mathbb{R}$ such that

$$\lim_{t \to \infty} F_{\Lambda_N}(\Phi_N(t)) = y \tag{55}$$

implying that

$$\lim_{t \to \infty} \Big\| \frac{d\Phi_N}{dt} \Big\|^2 = 0 \;\Leftrightarrow\; \lim_{t \to \infty} \Big\| \Lambda_N\Phi_N + \gamma\Big(1 - \frac{1}{\|\Phi_N\|}\Big)\Phi_N \Big\|^2 = 0.$$

This implies that for each $k = 1, 2, \ldots, N$,

$$\lim_{t \to \infty} \Big(\lambda_k + \gamma - \frac{\gamma}{\|\Phi_N(t)\|}\Big)^2 \phi_k^2(t) = 0. \tag{56}$$

Since

$$0 = \lim_{t \to \infty} \Big\| \Lambda_N\Phi_N + \gamma\Big(1 - \frac{1}{\|\Phi_N\|}\Big)\Phi_N \Big\| \geq \lim_{t \to \infty} \Big| \|(\Lambda_N + \gamma I)\Phi_N\| - \gamma \Big|,$$

we have

$$\lim_{t\to\infty}\sum_{k=1}^{N}(\lambda_k+\gamma)^2\phi_k^2(t)=\gamma^2, \tag{57}$$

which implies that there exist k_1 and $\{t_n\}$ such that $t_n\to\infty$ as $n\to\infty$ and

$$\lim_{n\to\infty}\phi_{k_1}^2(t_n)>0.$$

Unless $\lim_{n\to\infty}\phi_{k_1}^2(t_n)=\frac{\gamma^2}{(\gamma+\lambda_{k_1})^2}$, there exists $k_2\neq k_1$ such that $\lim_{n\to\infty}\phi_{k_2}^2(t_n)>0$ by taking a subsequence of $\{t_n\}$ if necessary. Then, (56) implies

$$\lambda_{k_1}=\lambda_{k_2}\quad\text{and}\quad\lim_{n\to\infty}\|\Phi_N(t_n)\|=\frac{\gamma}{\gamma+\lambda_{k_1}}.$$

We may repeat this process finitely many times to have

$$\lim_{n\to\infty}\sum_{\{k:\lambda_k=\lambda_{k_1}\}}\phi_k^2(t_n)=\frac{\gamma^2}{(\gamma+\lambda_{k_1})^2},\quad\lim_{n\to\infty}\sum_{\{k:\lambda_k\neq\lambda_{k_1}\}}\phi_k^2(t_n)=0.$$

Suppose that there exist l and $\{s_n\}$ such that $\lambda_l\neq\lambda_{k_1}$ and $s_n\to\infty$ as $n\to\infty$ and

$$\lim_{n\to\infty}\phi_l^2(s_n)>0.$$

Then, as was done above, with a subsequence of $\{s_n\}$ if necessary, we have

$$\lim_{n\to\infty}\|\Phi_N(s_n)\|=\frac{\gamma}{\gamma+\lambda_l},\quad\lim_{n\to\infty}\sum_{\{k:\lambda_k=\lambda_l\}}\phi_k^2(s_n)=\frac{\gamma^2}{(\gamma+\lambda_l)^2},\quad\lim_{n\to\infty}\sum_{\{k:\lambda_k\neq\lambda_l\}}\phi_k^2(s_n)=0$$

and (55) implies that

$$\lim_{n\to\infty}F_{\Lambda_N}(\Phi_N(t_n))=-\frac{\gamma^2}{2(\gamma+\lambda_{k_1})}=y=\lim_{n\to\infty}F_{\Lambda_N}(\Phi_N(s_n))=-\frac{\gamma^2}{2(\gamma+\lambda_l)},$$

which is a contradiction. Therefore, we conclude that

$$\lim_{t\to\infty}\|\Phi_N(t)\|=\frac{\gamma}{\gamma+\lambda_{k_1}},\quad\lim_{t\to\infty}\sum_{\{k:\lambda_k=\lambda_{k_1}\}}\phi_k^2(t)=\frac{\gamma^2}{(\gamma+\lambda_{k_1})^2},\quad\lim_{t\to\infty}\sum_{\{k:\lambda_k\neq\lambda_{k_1}\}}\phi_k^2(t)=0.$$

We will now claim that $\lambda_{k_1}=\lambda_1$. Suppose that $\lambda_{k_1}>\lambda_1$, i.e., $k_1>p$. Then, since there exists $T>0$ such that $\|\Phi_N(t)\|<\frac{\gamma(2\gamma+\lambda_1+\lambda_{k_1})}{2(\gamma+\lambda_1)(\gamma+\lambda_{k_1})}$ for $t>T$, we have

$$\frac{1}{2}\frac{d}{dt}\sum_{k=1}^{p}\phi_k^2(t)=\Big(-(\lambda_1+\gamma)+\frac{\gamma}{\|\Phi_N(t)\|}\Big)\sum_{k=1}^{p}\phi_k^2(t)\geq\Big(\frac{\lambda_{k_1}-\lambda_1}{2\gamma+\lambda_1+\lambda_{k_1}}\Big)\sum_{k=1}^{p}\phi_k^2(t)$$

for $t>T$, which results in $\|\Phi_N(t)\|\to\infty$ as $t\to\infty$. This is a contradiction. Therefore, we have $\lambda_{k_1}=\lambda_1$ and

$$\lim_{t\to\infty}\|\Phi_N(t)\|^2=\lim_{t\to\infty}\sum_{k=1}^{p}\phi_k^2(t)=\frac{\gamma^2}{(\gamma+\lambda_1)^2},\quad\text{and}\quad\lim_{t\to\infty}\sum_{k=p+1}^{N}\phi_k^2(t)=0. \tag{58}$$

Lastly, we will claim that there exists $\mathbf{v} \in \mathbb{R}^N$ such that $\|\mathbf{v}\| = \frac{\gamma}{\gamma+\lambda_1}$, and $\langle \mathbf{v}, \mathbf{e}_j \rangle \neq 0$, and

$$\Phi_N(t) \to \mathbf{v} \quad \text{as} \quad t \to \infty.$$

Note that for $k = 1, \ldots, p$, if $\phi_k(0) \neq 0$, then

$$\frac{1}{2\phi_k^2(t)} \frac{d}{dt} \phi_k^2(t) = -(\lambda_1 + \gamma) + \frac{\gamma}{\|\Phi_N(t)\|} \quad \text{for} \quad t \geq 0,$$

implying that for $T \in (0, \infty)$,

$$\ln(\phi_k^2(T)) - \ln(\phi_k^2(0)) = 2 \int_0^T \Big(-(\lambda_1 + \gamma) + \frac{\gamma}{\|\Phi_N(t)\|} \Big) dt := 2\Psi(T).$$

Together with (58), we have

$$\sum_{k=1}^p \phi_k^2(T) = e^{2\Psi(T)} \sum_{k=1}^p \phi_k^2(0) \to \Big(\frac{\gamma}{\gamma + \lambda_1} \Big)^2 \quad \text{as} \quad T \to \infty.$$

Note that

$$\Psi_\infty = \lim_{t \to \infty} \Psi(t) = \int_0^\infty \Big(-(\lambda_1 + \gamma) + \frac{\gamma}{\|\Phi_N(t)\|} \Big) dt = \ln \Big(\frac{\gamma}{\gamma + \lambda_1} \Big) - \frac{1}{2} \ln \Big(\sum_{k=1}^p \phi_k^2(0) \Big).$$

Hence, for $1 \leq k \leq p$,

$$\phi_k^2(t) \to v_k^2 \quad \text{as} \quad t \to \infty,$$

where $v_k = \phi_k(0) e^{\Psi_\infty}$. Noting that $\phi_k(t)\phi_k(0) > 0$ for all $t \geq 0$ unless $\phi_k(0) = 0$, we can easily see that

$$\phi_k(t) \to v_k \quad \text{as} \quad t \to \infty.$$

If we set

$$\mathbf{v} = [v_1 \ \cdots \ v_p \ 0 \ \cdots \ 0]^T,$$

then $\mathbf{v}$ is an eigenvector of Λ_N corresponding to λ_1 and $\langle \mathbf{v}, \mathbf{e}_j \rangle = v_j \neq 0$ since $\phi_j(0) \neq 0$ for some $1 \leq j \leq p$, and

$$\Phi_N(t) \to \mathbf{v} \quad \text{as} \quad t \to \infty$$

and $\|\mathbf{v}\| = \frac{\gamma}{\gamma+\lambda_1}$. □

Remark 1. As was proved in Theorem 8, even though a symmetric positive semidefinite matrix $A \in M_N(\mathbb{R})$ has the smallest eigenvalue of multiplicity greater than 1, the solution of (54) converges to a corresponding eigenvector just as the solution of (9) does. So the discrete and continuous settings present the same behavior. Moreover, due to the equivalence (53), the solution of the gradient descent flow associated to F_A converges to $Q\mathbf{v}$, which is an eigenvector of A corresponding to the smallest eigenvalue λ_1.

5 Conclusion

In this paper, we have proposed and analyzed to minimize the functional (3) using two well-known methods: The Gradient Descent and The Newton's methods. The two algorithms have shown their advantages in various situations. More interestingly, these methods revealed what conventional methods haven't observed.

More precisely, applying the Gradient Descent method to minimize (3) guarantees that we can find a global minimizer of our proposed functional and allows us to find eigenvectors in the increasing order of their corresponding eigenvalues without matrix inversion and produces many applications such as generalized eigenvalue problems and finding eigenfunctions of self-adjoint operators.

On the other hand, besides its faster convergence, the Newton's method proposes a new nonsingular linear system that we can solve once to find an exact eigenvector when an exact eigenvalue is given. Even with an approximate eigenvalue, the nonsingular linear system 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 reveals that the error in eigenvalue estimation and the error in eigenvector estimation are comparable.

Moreover, it turns out that the same framework can extend to nonlinear operators as well. One such example is the p-Laplacian operator $-\Delta_p$, $p > 2$. As before, we want to find φ such that

$$\begin{cases} -\Delta_p \varphi = \lambda \varphi, & \text{in } \Omega, \\ \varphi = 0, & \text{on } \partial\Omega, \end{cases} \tag{59}$$

where $-\Delta_p \varphi = \operatorname{div}(|\nabla \varphi|^{p-2} \nabla \varphi)$ by minimizing the functional F_p defined by

$$F_p(u) = \frac{1}{2} \int_\Omega |\nabla u(x)|^p dx + \frac{\gamma}{2} \int_\Omega |u(x)|^2 dx - \gamma \Big(\int_\Omega |u(x)|^2 dx \Big)^{\frac{1}{2}} \tag{60}$$

Note that

$$\min_{u \in W_0^{1,p}(\Omega)} F_p(u)$$

exists and a minimizer satisfies (59).

We will provide detailed analyses and numerical computations in a subsequent paper on applications of our proposed framework, including those mentioned in this paper, to prove efficiency and usefulness of our proposed method by revealing what have not been known through conventional methods in finding eigenvalues and eigenvectors.

References

[1] A.P. Austin and L.N. Trefethen, *Computing eigenvalues of real symmetric matrices with rational filters in real arithmetic*, SIAM J. Sci. Comput., 37(3), pp. A1365-A1387, 2015

[2] C.A. Beattie, M. Embree and D.C. Sorensen, *Convergence of polynomial restart Krylov methods for eigenvalue computations*, SIAM Review, 47(3), pp. 492-515, 2005

[3] M. Belkin, J. Sun and Y. Wang, *Constructing Laplace operator from point clouds in* $\mathbb{R}^d$, In Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1031-1040, Philadelphia, PA, USA, 2009

[4] J.E. Dennis and R.A. Tapia, *Inverse, shifted inverse, and Rayleigh quotient iteration as Newton's method*, http://www.caam.rice.edu/~rat/cv/RQI.pdf

[5] X.B. Gao, G.H. Golub and L.Z. Liao, *Continuous methods for sysmmetric generalized eigenvalue problems*, Linear Alg. Appl., 428, pp. 676-696, 2008

[6] J.M. Gedicke, *On the numerical analysis of eigenvalue problems*, http://edoc.hu-berlin.de/dissertationen/gedicke-joscha-micha-2013-06-10/PDF/gedicke.pdf, Dissertation, 2013

[7] R. Kolluri, J.R. Shewchuk and J.F. O'Brien, *Spectral surface reconstruction from noisy point clouds*, Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing, SGP '04, pp. 11-21, 2004

[8] V. Kuleshov, *Fast algorithms for sparse principal component analysis based on Rayleigh quotient iteration*, JMLR W& CP 28(3), pp. 1418-1425, 2013

[9] R. Lai, J. Liang and H. Zhao, *A local mesh method for solving PDEs on point clouds*, Inverse Probl. Imaging, 7(3), pp. 737-755

[10] E. Mengi, E.A. Yildirim and M. Kilic, *Numerical optimization of eigenvalues of hermitian matrix functions*, SIAM J. Matrix Anal. Appl., 35(2), pp. 699-724, 2014

[11] Y. Notay, *Convergence analysis of inexact Rayleigh quotient iteration*, SIAM J. Matrix Anal. Appl., 24(3), pp. 627-644, 2003

[12] G.L.G. Sleijpen and H.A.Van der Vorst, *A Jacobi-Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17(2), pp. 401-425, 1996

[13] G. Still, *Computable Bounds for Eigenvalues and Eigenfunctions of Elliptic Differential Operators*, Numer. Math., 54, pp. 201-223, 1988