SIAM J. MATRIX ANAL. APPL.
Vol. 46, No. 1, pp. 45–73

EPIC: A PROVABLE ACCELERATED EIGENSOLVER BASED ON PRECONDITIONING AND IMPLICIT CONVEXITY*

NIAN SHAO†, WENBIN CHEN‡, AND ZHAOJUN BAI§

Abstract. This paper is concerned with the extraction of the smallest eigenvalue and its corresponding eigenvector of a symmetric positive definite matrix pencil. We reveal implicit convexity of the eigenvalue problem in Euclidean space. A provable accelerated eigensolver based on preconditioning and implicit convexity (EPIC) is proposed. Theoretical analysis shows the acceleration of EPIC with a rate of convergence resembling the conjectured rate of convergence of the well-known locally optimal preconditioned conjugate gradient. Numerical results confirm our theoretical findings of EPIC.

Key words. eigenvalue problem, convexity, preconditioning, acceleration

MSC codes. 15A08, 65F08, 65F15, 90C25

DOI. 10.1137/24M1641440

1. Introduction. Eigenvalue problems are cornerstones in scientific and engineering computations. In this paper, we consider the following generalized eigenvalue problem:

$$Au = Mu\lambda, \tag{1.1}$$

where A and M are given $n \times n$ symmetric positive definite matrices, and (λ, u) is a desired eigenpair. Numerous algorithms for computing eigenvalues and their associated eigenvectors have been developed [3, 10, 28, 30, 38]. Preconditioning techniques are often necessary for large-scale problems and have been well studied for solving linear systems of equations [4, 37]. For eigenvalue problems, preconditioning has also been investigated extensively. There are the preconditioned steepest descent (PSD) method [21, 31, 39] and preconditioned gradient-type methods [12, 16, 34, 40]. The convergence analysis of these gradient-type eigensolvers is studied in [2, 8, 15, 25] and the references therein. One of the most popular preconditioned iterative method for the eigenvalue problem (1.1) is locally optimal preconditioned conjugate gradient (LOPCG) method and its block variant LOBPCG [13]. Compared with the PSD method, which only uses a current approximation and a preconditioned residual, LOPCG involves a previous approximation, which is called the momentum term. Numerical results show that convergence of LOBPCG is satisfied under careful implementations [7, 14]. Despite its great success in practice, proving the conjectured rate of convergence and acceleration of LOPCG in [13, eq. (5.5)] is still elusive.

There are preconditioned eigensolvers with momentum from perspectives of differential equations; see [5] and the references therein. Numerical results show that

*Received by the editors February 23, 2024; accepted for publication (in revised form) by M. A. Freitag August 22, 2024; published electronically January 3, 2025.
https://doi.org/10.1137/24M1641440
Funding: The second author is supported by the National Natural Science Foundation of China (NSFC) 12471369 and 12241101.
†Institute of Mathematics, EPF Lausanne, 1015 Lausanne, Switzerland (nian.shao@epfl.ch).
‡School of Mathematical Sciences and Shanghai Key Laboratory for Contemporary Applied Mathematics, Fudan University, Shanghai, 200433, People's Republic of China (wbchen@fudan.edu.cn).
§Department of Computer Science and Department of Mathematics, University of California, Davis, CA 95616 USA (zbai@ucdavis.edu).

momentum terms can significantly improve convergence but, theoretically, an acceleration is hard to prove.

Momentum methods are widely used in convex optimization, with their origins dating back to the early 1960s [29]. One popular momentum method is Nesterov accelerated gradient (NAG) flow [19]. In theoretical analysis of NAG flows, two main approaches are prominent. The classical method, pioneered by Nesterov in [19], utilizes estimating sequences. In contrast, a more contemporary approach, as investigated in [36], involves the derivation of a second-order ordinary differential equation (ODE) to delve into the dynamics of NAG flows. The connection between NAG flows and ODEs has been studied extensively in recent years [17, 18, 26, 33]. For example, by combining NAG flows with preconditioning, a preconditioned accelerated gradient descent method for solving semilinear PDEs was proposed in [27].

The crux of the success of NAG flow approach is the convexity of the objective function. Unfortunately, for the eigenvalue problem (1.1), the associated Rayleigh quotient

$$\mathrm{Rq}(x) = \frac{x^{\mathsf{T}}Ax}{x^{\mathsf{T}}Mx}, \quad x \neq \mathbf{0},$$

is not (strongly) convex in Euclidean space, due to the homogeneity $\mathrm{Rq}(tx) = \mathrm{Rq}(x)$ for all nonzero scalar t. One approach to investigate convexity in eigenvalue computation is to consider the Rayleigh quotient on smooth manifolds [9]. Recently, a Riemannian acceleration with preconditioning (RAP) has been proposed in [32], providing a provable accelerated preconditioned eigensolver with an essentially similar convergence rate to the conjecture of LOPCG. From the Riemannian manifold viewpoint, the convexity structure, known as geodesic convexity, has been extensively studied in the absence of preconditioning [1]. However, when incorporating preconditioning, there is a need to transform the objective function from a quadratic to a rational form and introduce intricate complexities to the convexity. The introduction of new technical conditions for preconditioners, in addition to the traditional spectral condition number $\kappa(T^{-1}A)$ as in [15], where T is a symmetric positive preconditioner for A, becomes essential for the theoretical analysis of acceleration induced by operations on manifolds. While extra conditions can be verified for some popular preconditioners, such as domain decomposition, it would be preferable if acceleration could be achieved with only minimal requirements about the spectral condition number. A viable alternative strategy is to explore convexity structures in Euclidean space, a pursuit we will undertake in this work.

In this paper, we reveal a new structure, named as implicit convexity, of the eigenvalue problem (1.1) with respect to the smallest eigenvalue and its associated eigenvector. Compared with the treatment of geodesic convexity, the implicit convexity only involves analysis in Euclidean space as commonly encountered in matrix computations. A provable accelerated symmetric Eigensolver based on preconditioning and implicit convexity (EPIC) will be proposed. Theoretical analysis of EPIC is presented and shows that the rate of convergence of EPIC resembles the conjectured convergence of LOPCG in [13, eq. (5.5)]. Numerical results confirm our theoretical findings.

For ease of reference, the following proposition provides some characterizations of strongly convex functions. Taking into account the preconditioning to be discussed later, we consider a P inner-product as

$$\langle x, y\rangle_P = x^{\mathsf{T}}Py, \tag{1.2}$$

where P is a symmetric positive definite matrix. The preconditioners P defined here and T introduced previously are related but not the same, their relationships will be discussed in section 5. In the next proposition, we use $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ to denote a general inner-product and norm, respectively, such as the P inner-product and P norm.

PROPOSITION 1.1 (see [20, sect. 2.1]). *Suppose ϕ is a smooth function on a convex domain $\mathcal{Y}$, and $0<\mu\le L$ are positive scalars, the following three inequalities for characterizing the strong convexity and Lipschitz smoothness of ϕ are equivalent:*

$$\frac{\mu}{2}\|y_1-y_2\|^2 \le \phi(y_1)-\phi(y_2)-\langle\nabla\phi(y_2),y_1-y_2\rangle \le \frac{L}{2}\|y_1-y_2\|^2, \tag{1.3}$$

$$\mu\|y_1-y_2\| \le \|\nabla\phi(y_1)-\nabla\phi(y_2)\| \le L\|y_1-y_2\|, \tag{1.4}$$

$$\mu P \preceq \nabla^2\phi(y) \preceq LP, \tag{1.5}$$

where y, y_1, $y_2\in\mathcal{Y}$ and $M_1\preceq M_2$ means M_2-M_1 is a symmetric positive semidefinite matrix.

By convention in convex optimization [20, p. 77], the condition number of a strongly convex function ϕ is denoted by the ratio $\kappa = L/\mu$, where L and μ are the optimal bounds from Proposition 1.1. The condition number is closely tied to fundamental properties of algorithms. For example, the rate of convergence of gradient descent methods and accelerated gradient descent methods are bounded by $1-c\kappa$ and $1-c\kappa^{1/2}$, respectively, for unconstrained convex minimization, where c is a universal positive constant [20, Chap. 2.1].

The rest of this paper is organized as follows. In section 2, we introduce the implicit convexity of the smallest eigenvalue problem by constructing an auxiliary problem on the tangent plane of an approximation of an eigenvector on the M-sphere. A novel locally optimal scheme of NAG (LONAG) flow will be proposed and analyzed in section 3. In section 4, we will show that the auxiliary problem can be solved by LONAG implicitly on the M-sphere, which only involves some cheap operations. Such an implicit algorithm will be named as eigensolver based on implicit convexity (EIC). Compared with the steepest descent method, an acceleration of EIC will be proved. In section 5, a preconditioned version of EIC, which is called eigensolver based on preconditioning and implicit convexity (EPIC), will be given by involving a preconditioner P, which is associated with a copreconditioner T for A, for auxiliary problem. Theoretical analysis shows that EPIC can achieve acceleration, whose rate of convergence is faster than PSD and similar to the conjecture of LOPCG. Numerical results, including a test for theoretical results and comparison with LOPCG will be given in section 6.

2. Implicit convexity of symmetric eigenvalue problem.

Suppose A and M are $n\times n$ symmetric positive definite matrices, $0<\lambda_1<\lambda_2\le\cdots\le\lambda_n$ are eigenvalues of (A,M), and $u_1,\ldots,u_n$ are an M-orthonormal set of eigenvectors. We consider the computation of the smallest eigenvalue and its associated eigenvector (λ_1,u_1) of (A,M):

$$Au_1 = Mu_1\lambda_1. \tag{2.1}$$

It is well known [10, p. 441] that u_1 is a global minimizer of Rayleigh quotient:

$$\lambda_1 = \mathrm{Rq}(u_1) = \min_{x\neq\mathbf{0}}\mathrm{Rq}(x).$$

2.1. Auxiliary problem.

In this section, we will construct an auxiliary problem of the eigenvalue problem (2.1) and then convert the eigenvalue problem (2.1) into an optimization problem of a convex function over a convex domain. Let q be an approximation of the eigenvector u_1 satisfying $q^{\mathsf{T}} M u_1 > 0$, such that $\|q\|_M = 1$ and

$$\lambda_1 \leq \mathrm{Rq}(q) < \frac{\lambda_1 + \lambda_2}{2}. \tag{2.2}$$

Let $\mathcal{S}_q^{n-1}$ be a hemisphere in $\mathbb{R}^n$ determined by q:

$$\mathcal{S}_q^{n-1} := \left\{ x \in \mathbb{R}^n \;\middle|\; \|x\|_M = 1,\, q^{\mathsf{T}} M x > 0 \right\}.$$

Define an A-spherical cap $\mathcal{X}$ of $\mathcal{S}_q^{n-1}$ as

$$\mathcal{X} = \left\{ x \in \mathcal{S}_q^{n-1} \;\middle|\; \mathrm{Rq}(x) \leq \mathrm{Rq}(q) \right\} \subset \mathcal{S}_q^{n-1}. \tag{2.3}$$

It is obvious that $\mathcal{X}$ is nonempty since $u_1 \in \mathcal{X}$. Define operators $\psi \colon \mathcal{S}_q^{n-1} \mapsto \mathbb{R}^{n-1}$ and $\psi^{\dagger} \colon \mathbb{R}^{n-1} \mapsto \mathcal{S}_q^{n-1}$ as

$$\psi(x) := \frac{Q_{\perp}^{\mathsf{T}} M x}{q^{\mathsf{T}} M x} \quad \text{and} \quad \psi^{\dagger}(y) := \frac{Q_{\perp} y + q}{\|Q_{\perp} y + q\|_M}, \tag{2.4}$$

where $Q_{\perp}$ is an orthonormal basis of the M-orthogonal complement of q, i.e., the matrix $[q, Q_{\perp}]$ is M-orthonormal. Then denominators of ψ and $\psi^{\dagger}$ are both nonzero. The following lemma shows that $\psi^{\dagger}$ is the inverse of ψ.

LEMMA 2.1. *For operators ψ and $\psi^{\dagger}$ defined in* (2.4),
1. *ψ and $\psi^{\dagger}$ are injections;*
2. *$\psi^{\dagger}\big(\psi(x)\big) = x$ holds for all $x \in \mathcal{S}_q^{n-1}$;*
3. *$\psi\big(\psi^{\dagger}(y)\big) = y$ holds for all $y \in \mathbb{R}^{n-1}$.*

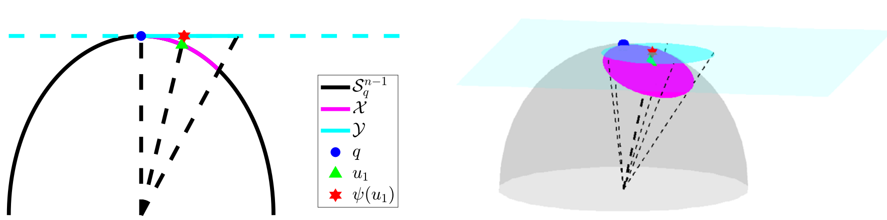
Proof. See Appendix A. ∎

Define a projected A-spherical cap $\mathcal{Y}$ of $\mathcal{X}$ as

$$\mathcal{Y} := \left\{ y \in \mathbb{R}^{n-1} \;\middle|\; y = \psi(x),\, x \in \mathcal{X} \right\}. \tag{2.5}$$

Relationships of $\mathcal{S}_q^{n-1}$, $\mathcal{X}$, $\mathcal{Y}$, q, u_1, and $\psi(u_1)$ are illustrated in Figure 1. The tangent space of $\mathcal{S}_q^{n-1}$ at q in M inner-product is $\left\{ Q_{\perp} y + q \;\middle|\; y \in \mathbb{R}^{n-1} \right\} \subset \mathbb{R}^n$. For any $x \in \mathcal{S}_q^{n-1}$, since $[q, Q_{\perp}]$ is M-orthonormal,

$$Q_{\perp} \psi(x) + q = \frac{Q_{\perp} Q_{\perp}^{\mathsf{T}} M x + q q^{\mathsf{T}} M x}{q^{\mathsf{T}} M x} = \frac{x}{q^{\mathsf{T}} M x}. \tag{2.6}$$



(a) 2D case (b) 3D case

FIG. 1. *Relationships of $\mathcal{S}_q^{n-1}$, $\mathcal{X}$, $\mathcal{Y}$, q, u_1, and $\psi(u_1)$. Note: color appears only in the online article.*

Therefore, $Q_\perp \psi(x) + q$ is a projection of $x \in \mathcal{X}$ onto the tangent space at q. The operator ψ maps a point $x \in \mathcal{X}$ to the coordinates of its projection in the tangent space with the basis $Q_\perp$. The operator $Q_\perp \mathcal{Y} + q$ is a projection of $\mathcal{X}$ from the origin.

Let $\phi\colon \mathbb{R}^{n-1} \mapsto \mathbb{R}$ be defined by

$$\phi(y) := \mathrm{Rq}(Q_\perp y + q) = \frac{y^\mathsf{T} B y + 2y^\mathsf{T} b + \mathrm{Rq}(q)}{\|y\|^2 + 1}, \tag{2.7}$$

where $B = Q_\perp^\mathsf{T} A Q_\perp$ and $b = Q_\perp^\mathsf{T} A q$. It is obvious that ϕ is a smooth function. A connection between Rayleigh quotient $\mathrm{Rq}(\cdot)$ and auxiliary function $\phi(\cdot)$ can be established as follows.

Proposition 2.1. *Let $x \in \mathcal{S}_q^{n-1}$ and $y = \psi(x)$. Then*

$$\mathrm{Rq}(x) = \phi(y). \tag{2.8}$$

Proof. The result is a direct consequence of the identities

$$\mathrm{Rq}(x) = \mathrm{Rq}\big(\psi^\dagger(y)\big) = \mathrm{Rq}(Q_\perp y + q) = \phi(y),$$

where we use Lemma 2.1, the homogeneity of the Rayleigh quotient, and (2.7), respectively. ∎

An auxiliary problem of the eigenvalue problem (2.1) is defined by

$$\min_{y \in \mathcal{Y}} \phi(y). \tag{2.9}$$

In the rest of this section, we will show that if $\mathrm{Rq}(q)$ is chosen sufficiently close to λ_1, the region $\mathcal{Y}$ is convex and the auxiliary function ϕ is strongly convex on $\mathcal{Y}$. Consequently, by the theory of convex optimization concerning existence and uniqueness of the solution (for example see [22, Thm. 2.4]), and properties of $\psi^\dagger$ in Lemma 2.1, we can conclude that the auxiliary problem (2.9) has a unique minimizer y_*, and the eigenvector u_1 of the eigenvalue problem (2.1) is given by $u_1 = \psi^\dagger(y_*)$.

2.2. Convexity of $\mathcal{Y}$ and ϕ. First, we recall the following lemma from Notay [23, Lem. 3.1].

Lemma 2.2. *Let A be a real symmetric $n \times n$ matrix with eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$. For any vector u with norm unity,*

$$\min_{z \perp u, \|z\| = 1} z^\mathsf{T} A z \ge \lambda_1 + \lambda_2 - u^\mathsf{T} A u.$$

By Lemma 2.2, it is easy to see that for a symmetric matrix pair (A, M) with eigenvalues $\lambda_1 < \lambda_2 \le \cdots \le \lambda_n$, and an M-orthonormal matrix $[q, Q_\perp]$, if $\mathrm{Rq}(q) < (\lambda_1 + \lambda_2)/2$, then

$$\lambda_{\min}(B) \ge \lambda_1 + \lambda_2 - \mathrm{Rq}(q) > \mathrm{Rq}(q), \tag{2.10}$$

where $B = Q_\perp^\mathsf{T} A Q_\perp$.

The following result shows that the region $\mathcal{Y}$ defined in (2.5) is convex.

Theorem 2.1. *Under the condition* (2.2), *i.e.*, $\lambda_1 \le \mathrm{Rq}(q) < (\lambda_1 + \lambda_2)/2$,

1. $\phi(y) \le \mathrm{Rq}(q)$ *if and only if* $y \in \mathcal{Y}$;
2. *the set* $\mathcal{Y}$ *is convex.*

Proof. For item 1, let $x = \psi^\dagger(y)$, by the definitions of $\mathcal{X}$ and $\mathcal{Y}$ in (2.3) and (2.5) and Proposition 2.1, we have

$$\phi(y) \le \mathrm{Rq}(q) \iff \mathrm{Rq}(x) \le \mathrm{Rq}(q) \iff x \in \mathcal{X} \iff y \in \mathcal{Y}.$$

For item 2, we consider an equivalent definition of $\mathcal{Y}$:

$$\mathcal{Y} = \{y \in \mathbb{R}^{n-1} \mid \phi(y) \le \mathrm{Rq}(q)\}.$$

Combining the estimation of eigenvalues of B in (2.10) and the assumption (2.2), we have

$$\lambda_{\min}(B) - \mathrm{Rq}(q) \ge \lambda_1 + \lambda_2 - 2\,\mathrm{Rq}(q) > 0,$$

which means that $B - \mathrm{Rq}(q)I$ is a symmetric positive definite matrix. Since

$$\begin{aligned} \phi(y) \le \mathrm{Rq}(q) &\iff \frac{y^\top By + 2y^\top b + \mathrm{Rq}(q)}{y^\top y + 1} \le \mathrm{Rq}(q) \\ &\iff y^\top(B - \mathrm{Rq}(q)I)y + 2y^\top b \le 0 \\ &\iff (y+z)^\top(B - \mathrm{Rq}(q)I)(y+z) \le z^\top(B - \mathrm{Rq}(q)I)z, \end{aligned}$$

where $z = (B - \mathrm{Rq}(q)I)^{-1}b$, we know that $\mathcal{Y}$ is a closed ball in $(B - \mathrm{Rq}(q)I)$ inner-product with center $(-z)$ and radius $\left(z^\top(B - \mathrm{Rq}(q)I)z\right)^{1/2}$. Therefore, $\mathcal{Y}$ is a convex set. ∎

Next we show that the auxiliary function ϕ is convex on $\mathcal{Y}$ by proving that ϕ is a strongly convex function satisfying the second-order characterization (1.5).

THEOREM 2.2. *Given a vector q for the auxiliary problem* (2.9)*, let $Q_\perp$ be an M-orthonormal basis of q's M-orthogonal complement, and $B = Q_\perp^\top A Q_\perp$. For any $(n-1)\times(n-1)$ symmetric positive definite matrix P, let*

$$\chi_P = \frac{8\lambda_2}{\lambda_1}\frac{\xi_{\max}}{\xi_{\min}}\left(\frac{\lambda_2 + \lambda_1}{2(\lambda_2 - \lambda_1)}\right)^{1/2} > 0, \tag{2.11}$$

where $\xi_{\min}$ and $\xi_{\max}$ are the smallest and largest eigenvalues of $P^{-1}B$, respectively. Suppose that the Rayleigh quotient of q satisfies

$$\lambda_1 \le \mathrm{Rq}(q) < \lambda_1 + \frac{\lambda_2 - \lambda_1}{2 + \chi_P}. \tag{2.12}$$

Then the second-order characterization of the convexity of ϕ in the auxiliary problem

$$\mu_P P \preceq \nabla^2\phi(y) \preceq L_P P \tag{2.13}$$

holds for all $y \in \mathcal{Y}$, where

$$\mu_P = 2\xi_{\min}\left(1 - \frac{2\big(\mathrm{Rq}(q) - \lambda_1\big)}{\lambda_2 - \lambda_1}\right)^2\left(1 - \frac{\lambda_1}{\lambda_2} - \frac{(2+\chi_P)}{\lambda_2}(\mathrm{Rq}(q) - \lambda_1)\right) > 0, \tag{2.14a}$$

$$L_P = 2\xi_{\max}\left(1 - \frac{\lambda_1}{\lambda_n} + \frac{\chi_P}{\lambda_2}\frac{\xi_{\min}}{\xi_{\max}}(\mathrm{Rq}(q) - \lambda_1)\right). \tag{2.14b}$$

For the proof of Theorem 2.2, we note that for any $y \in \mathbb{R}^{n-1}$, the Hessian of ϕ is given by

$$\nabla^2\phi(y) = \frac{2}{\|y\|^2+1}\Big(B - \phi(y)I - y\big(\nabla\phi(y)\big)^\mathsf{T} - \nabla\phi(y)y^\mathsf{T}\Big). \tag{2.15}$$

By the lower bound of $\lambda_{\min}(B)$ established in (2.10) and the upper bound of $\phi(y)$ given in Theorem 2.1, we know that the matrix $B - \phi(y)I$ is positive definite. The following lemma provides bounds of the term $y\big(\nabla\phi(y)\big)^\mathsf{T} + \nabla\phi(y)y^\mathsf{T}$.

LEMMA 2.3. *Under the condition* (2.2), *for any* $y \in \mathcal{Y}$, *relationships*

$$\frac{1}{1+\|y\|^2} \geq \left(1 - \frac{2\big(\mathrm{Rq}(q)-\lambda_1\big)}{\lambda_2-\lambda_1}\right)^2 \tag{2.16}$$

and

$$-\chi_g B \preceq \nabla\phi(y)y^\mathsf{T} + y\big(\nabla\phi(y)\big)^\mathsf{T} \preceq \chi_g B \tag{2.17}$$

hold, where

$$\chi_g = \frac{8(\mathrm{Rq}(q)-\lambda_1)}{\lambda_1}\Big(\frac{\lambda_1+\lambda_2}{2(\lambda_2-\lambda_1)}\Big)^{1/2}. \tag{2.18}$$

Proof. See Appendix B. ◻

Proof of Theorem 2.2. First, we prove the positive definiteness of the matrix $B-\phi(y)I$ in P inner-product. In fact, by $\phi(y) \leq \mathrm{Rq}(q)$, we have

$$\max_{s\in\mathbb{R}^{n-1}} \frac{s^\mathsf{T}(B-\phi(y)I)s}{\|s\|_P^2} \leq \big(1-\phi(y)\lambda_{\max}^{-1}(B)\big)\xi_{\max} \leq \Big(1-\frac{\lambda_1}{\lambda_n}\Big)\xi_{\max},$$
$$\min_{s\in\mathbb{R}^{n-1}} \frac{s^\mathsf{T}(B-\phi(y)I)s}{\|s\|_P^2} \geq \big(1-\phi(y)\lambda_{\min}^{-1}(B)\big)\xi_{\min} \geq \Big(1-\frac{\mathrm{Rq}(q)}{\lambda_1+\lambda_2-\mathrm{Rq}(q)}\Big)\xi_{\min},$$

where we use $\lambda_{\max}(B) \leq \lambda_n$ and $\lambda_{\min}(B) \geq \lambda_1+\lambda_2-\mathrm{Rq}(q)$ in (2.10). Since $\mathrm{Rq}(q) < (\lambda_1+\lambda_2)/2$, we have

$$\frac{\mathrm{Rq}(q)}{\lambda_1+\lambda_2-\mathrm{Rq}(q)} = \frac{\lambda_1}{\lambda_2} + \frac{(\lambda_2+\lambda_1)(\mathrm{Rq}(q)-\lambda_1)}{\lambda_2(\lambda_1+\lambda_2-\mathrm{Rq}(q))} \leq \frac{\lambda_1}{\lambda_2} + \frac{2(\mathrm{Rq}(q)-\lambda_1)}{\lambda_2} < 1.$$

Consequently, the positive definiteness of $B-\phi(y)I$ is verified by the following bounds:

$$\left(1-\frac{\lambda_1}{\lambda_2}-\frac{2(\mathrm{Rq}(q)-\lambda_1)}{\lambda_2}\right)\xi_{\min}P \preceq B-\phi(y)I \preceq \Big(1-\frac{\lambda_1}{\lambda_n}\Big)\xi_{\max}P. \tag{2.19}$$

Second, by Lemma 2.3 and the definition of $\xi_{\max}$, we immediately have

$$-\chi_g\xi_{\max}P \preceq -\chi_g B \preceq \nabla\phi(y)y^\mathsf{T} + y\big(\nabla\phi(y)\big)^\mathsf{T} \preceq \chi_g B \preceq \chi_g\xi_{\max}P, \tag{2.20}$$

where χ_g is defined in (2.18). Combining (2.15), (2.19), and (2.20), we know the following bounds for the Hessian $\nabla^2\phi(y)$:

$$\frac{2\xi_{\min}}{\|y\|^2+1}\Big(1-\frac{\lambda_1}{\lambda_2}-\frac{2(\mathrm{Rq}(q)-\lambda_1)}{\lambda_2}-\frac{\chi_g\xi_{\max}}{\xi_{\min}}\Big)P \preceq \nabla^2\phi(y) \preceq \frac{2\xi_{\max}}{\|y\|^2+1}\Big(1-\frac{\lambda_1}{\lambda_n}+\chi_g\Big)P.$$

Note that

$$\chi_P = \frac{\chi_g \lambda_2 \xi_{\max}}{(\mathrm{Rq}(q) - \lambda_1)\xi_{\min}}.$$

Therefore, we have the coefficient L_P of an upper bound of $\nabla^2 \phi(y)$:

$$\frac{2\xi_{\max}}{\|y\|^2+1}\left(1 - \frac{\lambda_1}{\lambda_n} + \chi_g\right) \leq 2\xi_{\max}\left(1 - \frac{\lambda_1}{\lambda_n} + \frac{\chi_P \xi_{\min}}{\lambda_2 \xi_{\max}}(\mathrm{Rq}(q) - \lambda_1)\right) := L_P.$$

On the other hand, by Lemma 2.3 again, we have the coefficient μ_P of a lower bound of $\nabla^2 \phi(y)$:

$$\frac{2\xi_{\min}}{\|y\|^2+1}\left(1 - \frac{\lambda_1}{\lambda_2} - \frac{2(\mathrm{Rq}(q) - \lambda_1)}{\lambda_2} - \frac{\chi_g \xi_{\max}}{\xi_{\min}}\right)$$
$$\geq 2\xi_{\min}\left(1 - \frac{2\big(\mathrm{Rq}(q) - \lambda_1\big)}{\lambda_2 - \lambda_1}\right)^2 \left(1 - \frac{\lambda_1}{\lambda_2} - \frac{(2+\chi_P)(\mathrm{Rq}(q) - \lambda_1)}{\lambda_2}\right) := \mu_P. \qquad \square$$

We have the following two corollaries of Theorem 2.2.

COROLLARY 2.1. *Up to the first-order of* $\mathrm{Rq}(q) - \lambda_1$, μ_P *and* L_P *are*

$$\mu_P = 2\xi_{\min}\Big(1 - \frac{\lambda_1}{\lambda_2}\Big) + \mathcal{O}(\mathrm{Rq}(q) - \lambda_1) \quad \text{and} \quad L_P = 2\xi_{\max}\Big(1 - \frac{\lambda_1}{\lambda_n}\Big) + \mathcal{O}(\mathrm{Rq}(q) - \lambda_1).$$

The condition number $\kappa_P = L_P/\mu_P$ *of the auxiliary function is given by*

$$\kappa_P = \frac{L_P}{\mu_P} = \iota_\xi \frac{1 - \lambda_1/\lambda_n}{1 - \lambda_1/\lambda_2} + \mathcal{O}(\mathrm{Rq}(q) - \lambda_1),$$

where $\iota_\xi = \xi_{\max}/\xi_{\min}$. *When the standard inner-product is applied, i.e.,* $P = I$, *by the estimation of the smallest eigenvalue of* B *in* (2.10), *we have*

$$\kappa_I = \frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1} + \mathcal{O}(\mathrm{Rq}(q) - \lambda_1).$$

Proof. The proof is shown by a direct expansion of μ_P and L_P from (2.14) in terms of $\mathrm{Rq}(q) - \lambda_1$. $\square$

COROLLARY 2.2. *Under condition* (2.12), *for any* $y \in \mathcal{Y}$,

$$\phi(y) - \phi(y_*) \leq \frac{L_P}{2}\|y - y_*\|_P^2. \tag{2.21}$$

Reversely, for any $y \in \mathbb{R}^{n-1}$, *the relationship* $y \in \mathcal{Y}$ *holds if*

$$\|y - y_*\|_P^2 \leq \frac{2(\mathrm{Rq}(q) - \lambda_1)}{L_P}. \tag{2.22}$$

Proof. See Appendix C. $\square$

2.3. Implicit convexity of eigenvalue problem. The main result on implicit convexity of the eigenvalue problem (2.1) is stated in the following theorem.

THEOREM 2.3. *Given a positive definite matrix pair (A, M) with a simple smallest eigenvalue, i.e., $0 < \lambda_1 < \lambda_2$, suppose the Rayleigh quotient of q satisfies*

$$\lambda_1 \le \mathrm{Rq}(q) < \lambda_1 + \frac{\lambda_2 - \lambda_1}{2 + \chi_P}, \tag{2.23}$$

where χ_P is defined in (2.11)*; then*

1. *the region $\mathcal{Y}$ defined in* (2.5) *is a convex set;*
2. *the auxiliary function ϕ defined in* (2.7) *is convex in $\mathcal{Y}$ with P inner-product;*
3. *the auxiliary problem* (2.9) *has a unique minimizer y_*;*
4. *the eigenvector u_1 of the eigenvalue problem* (2.1) *is given by $u_1 = \psi^\dagger(y_*)$.*

Proof. The first two items have been proved in Theorems 2.1 and 2.2. For item 3, according to the theory of existence and uniqueness of an optimizer for convex optimization [22, Thm. 2.4], the auxiliary problem (2.9) has a unique solution $y_* = \psi(x_*)$. For item 4, let $y_{**} = \psi(u_1)$, by $u_1 \in \mathcal{X}$ and the connection of the auxiliary function and Rayleigh quotient in Proposition 2.1,

$$\phi(y_*) \le \phi(y_{**}) = \mathrm{Rq}(u_1) \le \mathrm{Rq}(x_*) = \phi(y_*),$$

which implies $y_* = y_{**} = \psi(u_1)$. The theorem is proved by $u_1 = \psi^\dagger(\psi(u_1)) = \psi^\dagger(y_*)$. ∎

3. LONAG descent methods for convex optimization. In this section, we discuss the following general convex optimization problem:

$$\min_{y \in \mathcal{Y}} \phi(y), \tag{3.1}$$

where $\phi(y)$ is a smooth strongly convex function defined on a convex set $\mathcal{Y}$. Similarly to Proposition 1.1, we use $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ to denote a general inner-product and norm, such as P inner-product and P norm. Following the presentation in [17], we will review some results about NAG methods with a dynamical system analogy first proposed in [36]. Then, we propose a new discretization scheme and analyze its convergence.

3.1. NAG methods with a dynamical system analogy. Consider the following first-order dynamical system of $\big(y(t), s(t)\big)$:

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = s(t) - y(t), \tag{3.2a}$$

$$\frac{\mathrm{d}s(t)}{\mathrm{d}t} = y(t) - s(t) - \frac{1}{\mu}\nabla\phi(y(t)) \tag{3.2b}$$

with initial conditions $y(0) = y_0$ and $s(0) = s_0$, where $t > 0$, ϕ and μ satisfy (1.3). To establish a connection between solution of the optimization problem (3.1) and the dynamical system (3.2), let us consider the following so-called Lyapunov function:

$$\mathcal{L}(t) = \phi\big(y(t)\big) - \phi(y_*) + \frac{\mu}{2}\|s(t) - y_*\|^2 \ge 0, \tag{3.3}$$

where y_* is the unique minimizer of (3.1). Note that $\mathcal{L}(t) \ge 0$. It is shown in [17, Lem. 2] that the Lyapunov function exponentially decays:

$$\mathcal{L}(t) \le e^{-t}\mathcal{L}(0). \tag{3.4}$$

Combining (3.3) and (3.4), we know

$$\lim_{t\to\infty} \phi\big(y(t)\big) - \phi(y_*) \le \lim_{t\to\infty} \mathcal{L}(t) = 0.$$

Consequently, $y(t) \to y_*$ as $t \to \infty$ since y_* is the unique minimizer.

There are a number of discretization schemes for the dynamical system (3.2) [17, 18, 26, 33, 36]. To balance efficiency and stability, we have chosen the following corrected semi-implicit scheme from [17, eqs. (96–97)]. Given initial $(s_0, y_0) \in (\mathcal{Y}, \mathcal{Y})$, μ and L as defined in (1.3), step size $\tau > 0$, the corrected semi-implicit scheme generates the iterates (s_k, y_k) for $k = 0, 1, 2, \ldots$, by the recursions

$$\frac{\overline{y}_k - y_k}{\tau} = s_k - \overline{y}_k, \tag{3.5a}$$

$$\frac{s_{k+1} - s_k}{\tau} = (\overline{y}_k - s_k) - \frac{1}{\mu}\nabla\phi(\overline{y}_k), \tag{3.5b}$$

$$\text{update } y_{k+1} \text{ satisfying } \phi(y_{k+1}) \le \phi(\overline{y}_k) - \frac{1}{2L}\|\nabla\phi(\overline{y}_k)\|^2. \tag{3.5c}$$

When the first-order characterization (1.3) of the convexity of ϕ holds globally, a popular choice for y_{k+1} in step (3.5c) is the following gradient step [20, eq. (2.2.19)]:

$$y_{k+1} = \overline{y}_k - \frac{1}{L}\nabla\phi(\overline{y}_k).$$

The following theorem from [17, Thm. 7] proves a convergence rate of the scheme (3.5).

THEOREM 3.1 (see [17, Thm. 7]). *Let $y_0 \in \mathcal{Y}$ and $s_0 \in \mathcal{Y}$ be initials and $\tau > 0$ be a step size of the corrected semi-implicit scheme* (3.5). *Assume that*

- *the step size τ satisfies $0 < \tau \le \kappa^{-1/2}$, where $\kappa = L/\mu$, and μ and L are as defined in* (1.3)*;*
- *all iterates (s_k, y_k) lie in $\mathcal{Y}$.*

Then

$$\mathcal{L}_{k+1} \le (1-\tau)\mathcal{L}_k, \quad \textit{where} \quad \mathcal{L}_k = \phi(y_k) - \phi(y_*) + \frac{\mu}{2}\|s_k - y_*\|^2. \tag{3.6}$$

By the inequality (3.6), the convergence of discrete Lyapounov function $\mathcal{L}_k$ implies the convergence of y_k to y_*. Taking an optimal step size as $\tau = \kappa^{-1/2}$, an acceleration is achieved by improving the rate of convergence from $1 - 2(\kappa + 1)^{-1}$ of gradient methods [20, Thm. 2.1.15] to $1 - \kappa^{-1/2}$.

3.2. LONAG scheme and convergence analysis. There are two issues with the corrected semi-implicit scheme (3.5): there is no guarantee for a monotonically decreasing property of $\phi(y_k)$, which actually may fluctuate [27], and assumptions about all iterates (s_k, y_k) in $\mathcal{Y}$ are necessary for proving convergence. In this part, we propose a new scheme to guarantee $\phi(y_k)$ decreasing monotonically and analyze its convergence with conditions only about initial values (s_0, y_0).

Given initials $(s_0, y_0) \in (\mathcal{Y}, \mathcal{Y})$, μ and L as defined in (1.3), step size $\tau > 0$, we propose to replace the update (3.5c) with a locally optimal correction, and generate (s_k, y_k) for $k = 0, 1, 2, \ldots$, by the recursions

$$\frac{\overline{y}_k - y_k}{\tau} = s_k - \overline{y}_k, \tag{3.7a}$$

$$\frac{s_{k+1} - s_k}{\tau} = (\overline{y}_k - s_k) - \frac{1}{\mu}\nabla\phi(\overline{y}_k), \tag{3.7b}$$

$$y_{k+1} = \underset{y \in \operatorname{span}\{y_k, \overline{y}_k, \nabla\phi(\overline{y}_k)\}}{\arg\min} \phi(y). \tag{3.7c}$$

The update (3.7c) for y_{k+1} is inspired by LOPCG [13]. Since the scheme (3.7) is a combination of a locally optimal step and an NAG flow, we name it LONAG. The following result states the monotonicity of the function values $\phi(y_k)$.

PROPOSITION 3.1. *The iterates* $\{y_{k+1}\}$ *from the LONAG* (3.7) *satisfy*

$$\phi(y_{k+1}) \le \phi(y_k) \le \cdots \le \phi(y_0).$$

Proof. The result is a direct consequence of the locally optimal step (3.7c). ◻

Remark 3.1. As a consequence of Proposition 3.1, when the level set property $\{y \mid \phi(y) \le \phi(y_0)\} \subset \mathcal{Y}$ holds with a proper choice of the initial y_0, the locally optimal step (3.7c) is equivalent to

$$y_{k+1} = \underset{y \in \mathcal{Y} \cap \operatorname{span}\{y_k, \overline{y}_k, \nabla\phi(\overline{y}_k)\}}{\arg\min} \phi(y).$$

For the convergence of the LONAG, we would like to use the convergence of the corrected semi-implicit scheme (3.5) in Theorem 3.1. The challenge arises from the absence of prior assumptions on containment of iterates (s_k, y_k). However, we know that once the convergence of $\mathcal{L}_k$ is proved as (3.6), iterates y_k and s_k cannot be too far from the minimizer y_*. Fortunately, the following theorem shows that we can prove these two properties, i.e., containment and convergence, recursively when initials (y_0, s_0) and step size τ are properly selected.

THEOREM 3.2. *Given* L *and* μ *defined in* (1.3). *Let* $\mathcal{L}_0 = \phi(y_0) - \phi(y_*) + \frac{\mu}{2}\|s_0 - y_*\|^2$ *and*

$$R_1 = (2\mathcal{L}_0/\mu)^{1/2} \quad \textit{and} \quad R_2 = \max\{2R_1, (1+\tau\kappa)R_1\}. \tag{3.8}$$

Assume that initials (s_0, y_0) *satisfy*

$$s_0 \in \mathcal{B}_{R_1} \quad \textit{and} \quad \{y \mid \phi(y) \le \phi(y_0)\} \subset \mathcal{B}_{R_1} \subset \mathcal{B}_{R_2} \subset \mathcal{Y}, \tag{3.9}$$

where $\mathcal{B}_R$ *is a closed ball with center* y_**, the unique optimizer of* (3.1)*, and radius* R*:*

$$\mathcal{B}_R := \{y \mid \|y - y_*\| \le R\}.$$

If the step size τ *satisfies* $0 < \tau \le \kappa^{-1/2}$*, where* $\kappa = L/\mu$*, then iterates* (s_k, y_k) *with* $k > 0$ *generated by LONAG* (3.7) *satisfy*

(a) $\overline{y}_k \in \mathcal{B}_{R_1}$ $y_{k+1} \in \mathcal{B}_{R_1}$*;*
(b) y_{k+1} *satisfies the sufficient decrease property* (3.5c)*;*
(c) $\mathcal{L}_{k+1} \le (1-\tau)\mathcal{L}_k$*, where* $\mathcal{L}_k$ *is defined in* (3.6)*;*
(d) $s_{k+1} \in \mathcal{B}_{R_1}$.

Proof. The conclusions will be proved recursively. Let us assume that both y_k and s_k are in $\mathcal{B}_{R_1}$, which are satisfied if $k=0$. Then by (3.7a) and $\|\nabla\phi(y_*)\|=0$, we have

$$\begin{aligned} &\overline{y}_k = \frac{y_k}{1+\tau} + \frac{\tau s_k}{1+\tau} \in \mathcal{B}_{R_1} \quad \text{and} \\ &\|\nabla\phi(\overline{y}_k)\| = \|\nabla\phi(\overline{y}_k) - \nabla\phi(y_*)\| \le L\|\overline{y}_k - y_*\| \le LR_1. \end{aligned} \tag{3.10}$$

According to the recurrence of s_{k+1} in (3.7b), we have

$$\begin{aligned} \|s_{k+1} - y_*\| &= \|(1-\tau)s_k + \tau\overline{y}_k - \frac{\tau}{\mu}\nabla\phi(\overline{y}_k) - y_*\| \\ &= \|(1-\tau)(s_k - y_*) + \tau(\overline{y}_k - y_*) - \frac{\tau}{\mu}\nabla\phi(\overline{y}_k)\| \\ &\le (1-\tau)\|s_k - y_*\| + \tau\|\overline{y}_k - y_*\| + \frac{\tau}{\mu}\|\nabla\phi(\overline{y}_k)\| \\ &\le (1-\tau)R_1 + \tau R_1 + \tau\kappa R_1 \le R_2, \end{aligned}$$

where for the second inequality, we use the inequality (3.10) and $\kappa = L/\mu$, and for the last inequality we use (3.8). Therefore, $s_{k+1} \in \mathcal{B}_{R_2} \subset \mathcal{Y}$.

Now according to the monotonically decreasing property in Proposition 3.1, we know that

$$\phi(y_{k+1}) \le \phi(y_k) \le \phi(y_0). \tag{3.11}$$

Therefore, by containment property (3.9), $y_{k+1} \in \mathcal{B}_{R_1}$ is true.

We can also show that y_{k+1} of LONAG satisfies the sufficient decrease of $\phi(y_k)$ in the corrected semi-implicit scheme (3.5c). Let

$$\widetilde{y}_k = \overline{y}_k - \frac{1}{L}\nabla\phi(\overline{y}_k) \in \operatorname{span}\{y_k, \overline{y}_k, \nabla\phi(\overline{y}_k)\}.$$

First, since $\overline{y}_k \in \mathcal{B}_{R_1}$, by (3.10), we have

$$\|\widetilde{y}_k - y_*\| \le \|\overline{y}_k - y_*\| + \frac{1}{L}\|\nabla\phi(\overline{y}_k)\| \le 2R_1 \le R_2,$$

which means $\widetilde{y}_k \in \mathcal{B}_{2R_1} \subset \mathcal{B}_{R_2} \subset \mathcal{Y}$. Now using (1.3), we have

$$\begin{aligned} \phi(y_{k+1}) &\le \phi(\widetilde{y}_k) \le \phi(\overline{y}_k) + \langle\nabla\phi(\overline{y}_k), \widetilde{y}_k - \overline{y}_k\rangle + \frac{L}{2}\|\widetilde{y}_k - \overline{y}_k\|^2 \\ &= \phi(\overline{y}_k) - \frac{1}{L}\|\nabla\phi(\overline{y}_k)\|^2 + \frac{1}{2L}\|\nabla\phi(\overline{y}_k)\|^2 = \phi(\overline{y}_k) - \frac{1}{2L}\|\nabla\phi(\overline{y}_k)\|^2, \end{aligned} \tag{3.12}$$

where the first inequality comes from the locally optimal step (3.7c) and the second inequality comes from the first-order characterization (1.3).

Thus we have proved that $\overline{y}_k$, $y_{k+1} \in \mathcal{B}_{R_1} \subset \mathcal{Y}$, $s_{k+1} \in \mathcal{B}_{R_2} \subset \mathcal{Y}$, and y_{k+1} of LONAG satisfies the sufficient decrease of $\phi(y_1)$ in (3.5c). Now we are ready to apply the convergence of the corrected semi-implicit scheme in Theorem 3.1 to obtain

$$\mathcal{L}_{k+1} = \phi(y_{k+1}) - \phi(y_*) + \frac{\mu}{2}\|s_{k+1} - y_*\|^2 \le (1-\tau)\mathcal{L}_k \le \cdots \le (1-\tau)^{k+1}\mathcal{L}_0. \tag{3.13}$$

Note that $\phi(y_{k+1}) - \phi(y_*) \ge 0$ always holds since y_* is the minimizer of ϕ. Therefore, we have

$$\frac{\mu}{2}\|s_{k+1}-y_*\|^2 \le \phi(y_{k+1})-\phi(y_*)+\frac{\mu}{2}\|s_{k+1}-y_*\|^2 = \mathcal{L}_{k+1} \le \mathcal{L}_0,$$

which implies that $s_{k+1} \in \mathcal{B}_{R_1}$. This completes the proof. ∎

Remark 3.2. Let us highlight a major difference between Theorems 3.1 and 3.2. In Theorem 3.1, iterates (s_k, y_k) are assumed to lie within $\mathcal{Y}$, which is a domain where ϕ satisfies the first-order characterization (1.3). However, for the auxiliary problem (2.9), the objective function is locally convex with respect to the choice of q, an approximate eigenvector of u_1. There is no prior assumption about locations of iterates (s_k, y_k). In this scenario, we need to prove a containment similar to (3.9), which is inspired by the work of Park, Salgado, and Wise [27] on a preconditioned NAG method for solving semilinear partial differential equations.

By the monotonical decrease property in Proposition 3.1 and convergence in Theorem 3.2, we have the following results on the convergence of the LONAG scheme.

THEOREM 3.3. *With the assumptions of Theorem* 3.2*, the sequence* $\{y_k\}$ *generated by LONAG* (3.7) *satisfies that*

$$\phi(y_k) \le \phi(y_{k-1}) \le \cdots \le \phi(y_0) \tag{3.14}$$

and

$$\phi(y_k)-\phi(y_*) \le (1-\tau)^k \mathcal{L}_0, \tag{3.15}$$

where $0<\tau\le\kappa^{-1/2}$*,* y_* *is the minimizer of* (3.1)*, and* $\mathcal{L}_0=\phi(y_0)-\phi(y_*)+\frac{\mu}{2}\|s_0-y_*\|^2$.

It is clear that LONAG, similarly to the corrected semi-implicit scheme in Theorem 3.1, achieves an acceleration by improving the rate of convergence to $1-\kappa^{-1/2}$.

4. EIC: A symmetric eigensolver based on implicit convexity. In this section, we propose an algorithm for solving the original eigenvalue problem (2.1) by transforming LONAG (3.7) for the auxiliary function (2.9) on $\mathcal{Y}$ onto $\mathcal{X}$. The new algorithm is called EIC. In addition, we will discuss convergence of EIC and needs of preconditioning.

4.1. EIC. Let us return to the auxiliary problem (2.9). As shown in Theorem 2.3, the auxiliary problem (2.9) is a locally convex optimization problem, and we can apply LONAG (3.7) for solving (2.9). With initials $(s_0, y_0) \in (\mathcal{Y}, \mathcal{Y})$, LONAG (3.7) generates iterates (s_k, y_k) by the following recursions:

$$\overline{y}_k = \frac{y_k+\tau s_k}{1+\tau}, \tag{4.1a}$$

$$s_{k+1} = (1-\tau)s_k + \tau\overline{y}_k - \frac{\tau}{\mu}\nabla\phi(\overline{y}_k), \tag{4.1b}$$

$$y_{k+1} = \underset{y\in\mathcal{Y}\cap\operatorname{span}\{y_k,\overline{y}_k,\nabla\phi(\overline{y}_k)\}}{\arg\min}\ \phi(y), \tag{4.1c}$$

where the step size τ satisfies $0<\tau\le\kappa^{-1/2}$, $\kappa=L/\mu$, μ and L are convexity parameters of $\phi(y)$ defined in Theorem 2.2 with $P=I$.

In (4.1), we solve the auxiliary problem (2.9) on $\mathcal{Y}$, and assume that $Q_\perp$ is explicitly available. This is impractical since using $Q_\perp$ is too expensive. To circumvent $Q_\perp$, we propose a scheme by transforming computation on $\mathcal{Y}$ into $\mathcal{X}$ without using $Q_\perp$. To do so, for $k\ge 0$, denote

$$z_k = \psi^\dagger(s_k),\ x_k=\psi^\dagger(y_k),\ \overline{x}_k=\psi^\dagger(\overline{y}_k), \tag{4.2}$$

where the operator $\psi^\dagger$ is defined as in (2.4). By Lemma 2.1, it is clear that

$$z_k, x_k, \overline{x}_k \in \mathcal{X}.$$

The following proposition, which can be easily verified, shows that the explicit reference to $Q_\perp$ can be avoided after applying $\psi^\dagger$ due to the relationship $\mathcal{Y} = \psi(\mathcal{X})$ established in Lemma 2.1.

PROPOSITION 4.1. *Let $[q, Q_\perp]$ be M-orthonormal. For any $x \in \mathbb{R}^n$,*

$$\psi^\dagger(Q_\perp^\mathsf{T} M x) = \frac{q + (I - qq^\mathsf{T} M)x}{\left\| q + (I - qq^\mathsf{T} M)x \right\|_M}.$$

The gradients of $\phi(\cdot)$ and $\mathrm{Rq}(\cdot)$ are connected by

$$\nabla \phi\big(\psi(x)\big) = \frac{Q_\perp^\mathsf{T} \nabla \mathrm{Rq}(x)}{q^\mathsf{T} M x}.$$

Now let us reveal expressions of $\overline{x}_k, z_{k+1}, x_{k+1}$ without explicit reference to $Q_\perp$. First, for $\overline{x}_k$, by Proposition 4.1 and the definition of $\psi^\dagger$ in (2.4),

$$\begin{aligned} \overline{x}_k &= \psi^\dagger(\overline{y}_k) = \psi^\dagger\Big(\frac{y_k + \tau s_k}{1+\tau}\Big) = \psi^\dagger\Big(\frac{\psi(x_k) + \tau\psi(z_k)}{1+\tau}\Big) \\ &= \psi^\dagger\bigg(Q_\perp^\mathsf{T} M\Big(\frac{x_k}{(1+\tau)(q^\mathsf{T} M x_k)} + \frac{\tau z_k}{(1+\tau)(q^\mathsf{T} M z_k)}\Big)\bigg) \\ &= \frac{1}{\eta_1}\bigg(\frac{x_k}{q^\mathsf{T} M x_k} + \frac{\tau z_k}{q^\mathsf{T} M z_k}\bigg), \end{aligned} \tag{4.3}$$

where η_1 ensures $\|\overline{x}_k\|_M = 1$.

Next, consider z_{k+1}. According to the definition of ψ in (2.4), and using Proposition 4.1 and $z_{k+1} = \psi^\dagger(s_{k+1})$, we have

$$\begin{aligned} z_{k+1} &= \psi^\dagger(s_{k+1}) = \psi^\dagger\Big((1-\tau)s_k + \tau\overline{y}_k - (\tau/\mu)\nabla\phi(\overline{y}_k)\Big) \\ &= \psi^\dagger\Big((1-\tau)\psi(z_k) + \tau\psi(\overline{x}_k) - \frac{\tau(q^\mathsf{T} M \overline{x}_k)}{\mu} Q_\perp^\mathsf{T} r_k\Big) \\ &= \psi^\dagger\bigg(Q_\perp^\mathsf{T} M\Big(\frac{(1-\tau)z_k}{q^\mathsf{T} M z_k} + \frac{\tau\overline{x}_k}{q^\mathsf{T} M \overline{x}_k} - \frac{\tau(q^\mathsf{T} M \overline{x}_k) M^{-1} r_k}{\mu}\Big)\bigg) \\ &= \frac{1}{\eta_2}\bigg(\frac{(1-\tau)z_k}{q^\mathsf{T} M z_k} + \frac{\tau\overline{x}_k}{q^\mathsf{T} M \overline{x}_k} - \frac{\tau(q^\mathsf{T} M \overline{x}_k)(I - qq^\mathsf{T} M) M^{-1} r_k}{\mu}\bigg), \end{aligned} \tag{4.4}$$

where $r_k = \nabla \mathrm{Rq}(\overline{x}_k) = 2\big(A\overline{x}_k - \mathrm{Rq}(\overline{x}_k) M \overline{x}_k\big)$ and η_2 ensures $\|z_{k+1}\|_M = 1$.

Finally for x_{k+1}, consider the local optimization problem (4.1c),

$$y_{k+1} = \operatorname*{arg\,min}_{y \in \mathcal{V}_\mathcal{Y}} \phi(y),$$

where $\mathcal{V}_\mathcal{Y} = \mathcal{Y} \cap \mathrm{span}\{y_k, \overline{y}_k, \nabla\phi(\overline{y}_k)\}$. Let

$$\mathcal{V}_\mathcal{X} = \mathcal{X} \cap \mathrm{span}\{q, x_k, \overline{x}_k, M^{-1}\nabla \mathrm{Rq}(\overline{x}_k)\}.$$

Then from the connection between gradients in Proposition 4.1, we know

$$\psi(\mathcal{V}_\mathcal{X}) \subset \mathcal{V}_\mathcal{Y} \quad \text{and} \quad \psi^\dagger(\mathcal{V}_\mathcal{Y}) \subset \mathcal{V}_\mathcal{X},$$

which means $\psi(\mathcal{V}_{\mathcal{X}}) = \mathcal{V}_{\mathcal{Y}}$. Now consider the following local optimization problem on $\mathcal{X}$:

$$x_* = \arg\min_{x\in\mathcal{V}_{\mathcal{X}}} \mathrm{Rq}(x).$$

By the minimization property of x_* and y_{k+1}, and Proposition 2.1, we have

$$\mathrm{Rq}(x_*) \le \mathrm{Rq}\big(\psi^\dagger(y_{k+1})\big) = \phi(y_{k+1}) \le \phi\big(\psi(x_*)\big) = \mathrm{Rq}(x_*).$$

Due to the uniqueness of x_* and y_{k+1}, we obtain

$$x_{k+1} = \psi^\dagger(y_{k+1}) = x_* = \arg\min_{x\in\mathcal{V}_{\mathcal{X}}} \mathrm{Rq}(x). \tag{4.5}$$

Combining (4.3)–(4.5), we derive an equivalent iteration of (4.1) with all computations on $\mathcal{X}$. The recursions (4.3)–(4.5) with initials $z_0, x_0 \in \mathcal{X}$ are called EIC.

4.2. Convergence analysis of EIC.

Theorem 4.1. *Given a positive definite matrix pair (A, M) with a simple smallest eigenvalue λ_1, i.e., $0 < \lambda_1 < \lambda_2$, assume that the initial vector $z_0 = x_0 \in \mathcal{X}$ of EIC is chosen such that*

$$0 \le \mathrm{Rq}(x_0) - \lambda_1 \le \frac{1}{\max\{8\kappa, 2\kappa(1+\tau\kappa)^2\}}(\mathrm{Rq}(q) - \lambda_1), \tag{4.6}$$

where $\mathrm{Rq}(q)$ satisfies the condition (2.23), and q is the vector in (2.2) to define the auxiliary problem (2.9), $\kappa = L/\mu$, μ and L are convexity parameters of $\phi(y)$ defined in Theorem 2.2 with $P = I$. If the step size satisfies $0 < \tau \le \kappa^{-1/2}$, then the Rayleigh quotient sequence of x_k generated by EIC satisfies

$$\mathrm{Rq}(x_k) \le \mathrm{Rq}(x_{k-1}) \le \cdots \le \mathrm{Rq}(x_0) \tag{4.7}$$

and

$$\mathrm{Rq}(x_k) - \lambda_1 \le 2(1-\tau)^k \big(\mathrm{Rq}(x_0) - \lambda_1\big). \tag{4.8}$$

Proof. The proof is based on the verification of all conditions of Theorem 3.2. Details are presented in Appendix D. ∎

Combining the convergence analysis of EIC in Theorem 4.1 with the estimation for the condition number of the auxiliary function in Corollary 2.1, neglecting the term $\mathcal{O}(\mathrm{Rq}(q) - \lambda_1)$, the rate of convergence of EIC is

$$\mathrm{Rq}(x_k) - \lambda_1 \le 2\left(1 - \left(\frac{\lambda_2 - \lambda_1}{\lambda_n - \lambda_1}\right)^{1/2}\right)^k \big(\mathrm{Rq}(x_0) - \lambda_1\big). \tag{4.9}$$

Compared with the convergence rate of the steepest descent method [11, Thm. 2.1],

$$\tan\Theta(x_k, u_1) \le \left(1 - \frac{\lambda_2 - \lambda_1}{\lambda_n - \lambda_1}\right)^k \tan\Theta(x_0, u_1),$$

EIC achieves an acceleration by improving the exponent of $\frac{\lambda_2-\lambda_1}{\lambda_n-\lambda_1}$ from 1 to 1/2. However, the bound (4.9) is not satisfactory in practice. When the relative spectral gap $\frac{\lambda_2-\lambda_1}{\lambda_n-\lambda_1}$ is small, such as the relative spectral gap $\mathcal{O}(h^2)$ of the discrete Laplacian operator Δ^h, where h is mesh size, the convergence rate of EIC is close to 1, which leads to slow convergence of EIC. Meanwhile, we observe that in Corollary 2.1, the condition number κ_P will be improved to $\lambda_2/(\lambda_2 - \lambda_1)$ when P is a good spectral approximation of B such that the ratio ι_ξ defined in (2.2) is close to 1, which leads to fast convergence of EIC. In the next section, we will derive a preconditioning technique in EIC to improve the condition number κ_P by using a properly chosen preconditioner P. The resulting algorithm is called EPIC.

5. EPIC. Let us again start with the auxiliary problem (2.9) on $\mathcal{Y}$. Given a symmetric positive definite matrix P and initials s_0 and y_0 in $\mathcal{Y}$, LONAG (4.1) in P inner-product is as follows:

$$\overline{y}_k = \frac{y_k + \tau_P s_k}{1+\tau_P}, \tag{5.1a}$$

$$s_{k+1} = (1-\tau_P)s_k + \tau_P \overline{y}_k - \frac{\tau_P}{\mu_P} P^{-1}\nabla\phi(\overline{y}_k), \tag{5.1b}$$

$$y_{k+1} = \mathop{\arg\min}_{y\in\mathcal{Y}\cap\operatorname{span}\{y_k,\overline{y}_k,P^{-1}\nabla\phi(\overline{y}_k)\}} \phi(y), \tag{5.1c}$$

where the step size τ_P satisfies $0<\tau_P\le\kappa_P^{-1/2}$, $\kappa_P = L_P/\mu_P$, μ_P and L_P are convexity parameters of $\phi(y)$ defined in Theorem 2.2. In [27], such a strategy is called preconditioning since the level sets of objective ϕ look more circular when a good P is applied. Throughout this section, we will also call P as a preconditioner and the scheme (5.1) as a preconditioned LONAG.

Similarly to subsection 4.1, we would like to compute the preconditioned LONAG flow (5.1) on $\mathcal{X}$. Recall variables defined in (4.2) as

$$z_k = \psi^\dagger(s_k),\ x_k = \psi^\dagger(y_k),\ \overline{x}_k = \psi^\dagger(\overline{y}_k).$$

By the identity (4.3), $\overline{x}_k$ can be updated as

$$\overline{x}_k = \frac{1}{\eta_1}\Big(\frac{x_k}{q^\mathsf{T} M x_k} + \frac{\tau_P z_k}{q^\mathsf{T} M z_k}\Big), \tag{5.2}$$

where η_1 ensures $\|\overline{x}_k\|_M = 1$.

For z_{k+1}, similar to the identity (4.4), we know

$$\begin{aligned} z_{k+1} &= \frac{1}{\eta_2}\Big(\frac{(1-\tau_P)z_k}{q^\mathsf{T} M z_k} + \frac{\tau_P\overline{x}_k}{q^\mathsf{T} M\overline{x}_k} - \frac{\tau_P(q^\mathsf{T} M\overline{x}_k)(I-qq^\mathsf{T} M)Q_\perp P^{-1}Q_\perp^\mathsf{T} r_k}{\mu_P}\Big) \\ &= \frac{1}{\eta_2}\Big(\frac{(1-\tau_P)z_k}{q^\mathsf{T} M z_k} + \frac{\tau_P\overline{x}_k}{q^\mathsf{T} M\overline{x}_k} - \frac{\tau_P(q^\mathsf{T} M\overline{x}_k)Q_\perp P^{-1}Q_\perp^\mathsf{T} r_k}{\mu_P}\Big), \end{aligned} \tag{5.3}$$

where η_2 ensures $\|z_{k+1}\|_M = 1$ and

$$r_k = \nabla\mathrm{Rq}(\overline{x}_k) = 2\big(A\overline{x}_k - \mathrm{Rq}(\overline{x}_k)M\overline{x}_k\big). \tag{5.4}$$

Note that computation for the vector z_{k+1} of (5.3) is unattainable due to the term $Q_\perp P^{-1}Q_\perp^\mathsf{T} r_k$ involving $Q_\perp$. To circumvent $Q_\perp$, we introduce a symmetric positive definite copreconditioner T of P, where $T\in\mathbb{R}^{n\times n}$, and enforce the form of P as

$$P = Q_\perp^\mathsf{T} T Q_\perp. \tag{5.5}$$

The following lemma shows that the term $Q_\perp P^{-1}Q_\perp^\mathsf{T} r_k$ can be computed without explicit reference to $Q_\perp$.

LEMMA 5.1. *Suppose T is symmetric positive definite and $P = Q_\perp^\mathsf{T} T Q_\perp$. Then for any $z\in\mathbb{R}^n$,*

$$Q_\perp P^{-1} Q_\perp^\mathsf{T} z = \Pi T^{-1} z, \tag{5.6}$$

where Π is a complementation of the oblique projector $\widetilde{q}q^\mathsf{T} M/(q^\mathsf{T} M\widetilde{q})$ defined as

$$\Pi = I - \frac{\widetilde{q}q^\mathsf{T} M}{q^\mathsf{T} M\widetilde{q}}, \tag{5.7}$$

and $\widetilde{q} = T^{-1}Mq$.

Proof. Since $[q, Q_\perp]$ is M-orthonormal, it is sufficient to prove

$$q^\mathsf{T} M\big(Q_\perp P^{-1} Q_\perp^\mathsf{T} z\big) = q^\mathsf{T} M\Pi T^{-1} z, \tag{5.8}$$

$$Q_\perp^\mathsf{T} M\big(Q_\perp P^{-1} Q_\perp^\mathsf{T} z\big) = Q_\perp^\mathsf{T} M\Pi T^{-1} z. \tag{5.9}$$

For (5.8), the left side is zero due to the M-orthogonality of $[q, Q_\perp]$, and the right side is also zero due to $\Pi^\mathsf{T} M q = \mathbf{0}$.

For (5.9), multiplying P on both sides, it is sufficient to prove

$$Q_\perp^\mathsf{T} z = P Q_\perp^\mathsf{T} M\Big(I - \frac{\widetilde{q} q^\mathsf{T} M}{q^\mathsf{T} M \widetilde{q}}\Big) T^{-1} z.$$

By $P = Q_\perp^\mathsf{T} T Q_\perp$ and $\Pi^\mathsf{T} M q = \mathbf{0}$, we have

$$\begin{aligned} P Q_\perp^\mathsf{T} M \Pi T^{-1} z &= Q_\perp^\mathsf{T} T Q_\perp Q_\perp^\mathsf{T} M \Pi T^{-1} z = Q_\perp^\mathsf{T} T (I - q q^\mathsf{T} M) \Pi T^{-1} z \\ &= Q_\perp^\mathsf{T} z - \frac{z^\mathsf{T} \widetilde{q}}{q^\mathsf{T} M \widetilde{q}} Q_\perp^\mathsf{T} T T^{-1} M q = Q_\perp^\mathsf{T} z, \end{aligned}$$

which means (5.9) holds. Then the lemma is proved by the identities (5.8) and (5.9). ∎

By Lemma 5.1, the updating formula (5.3) can be rewritten as

$$z_{k+1} = \frac{1}{\eta_2}\left(\frac{(1-\tau_P) z_k}{q^\mathsf{T} M z_k} + \frac{\tau_P \overline{x}_k}{q^\mathsf{T} M \overline{x}_k} - \frac{\tau_P (q^\mathsf{T} M \overline{x}_k) \widetilde{r}_k}{\mu_P}\right), \tag{5.10}$$

where $\widetilde{r}_k = Q_\perp P^{-1} Q_\perp^\mathsf{T} r_k = \Pi T^{-1} r_k$, Π is defined in (5.7), and η_2 ensures $\|z_{k+1}\|_M = 1$.

Finally, for the vector x_{k+1}, let us consider the local optimization problem

$$y_{k+1} = \arg\min_{y \in \mathcal{V}_\mathcal{Y}} \phi(y),$$

where

$$\mathcal{V}_\mathcal{Y} = \mathcal{Y} \cap \operatorname{span}\{y_k, \overline{y}_k, P^{-1} \nabla \phi(\overline{y}_k)\}.$$

First by Proposition 4.1 and (5.4), we have

$$P^{-1} \nabla \phi(\overline{y}_k) = \frac{P^{-1} Q_\perp^\mathsf{T} \nabla \operatorname{Rq}(\overline{x}_k)}{\sqrt{1 + \|\overline{y}_k\|^2}} = \frac{P^{-1} Q_\perp^\mathsf{T} r_k}{\sqrt{1 + \|\overline{y}_k\|^2}}.$$

Combining the above equation with Lemma 5.1 and (2.4), we have

$$\operatorname{span}\big\{q, \psi^\dagger\big(P^{-1} \nabla \phi(\overline{y}_k)\big)\big\} = \operatorname{span}\{q, Q_\perp P^{-1} \nabla \phi(\overline{y}_k)\} = \operatorname{span}\{q, \widetilde{r}_k\}.$$

Using the same arguments as (4.5), let

$$\mathcal{V}_\mathcal{X} = \mathcal{X} \cap \operatorname{span}\{q, x_k, \overline{x}_k, \widetilde{r}_k\}.$$

We know $\psi(\mathcal{V}_\mathcal{X}) = \mathcal{V}_\mathcal{Y}$ and the expression of x_{k+1} is

$$x_{k+1} = \psi^\dagger(y_{k+1}) = \arg\min_{x \in \mathcal{V}_\mathcal{X}} \operatorname{Rq}(x). \tag{5.11}$$

Combining (5.2), (5.3), and (5.11), we have a preconditioned LONAG on $\mathcal{X}$ outlined in Algorithm 1, which is called EPIC.

Algorithm 1: EPIC.

Input: Matrices A, M, a vector q, a preconditioner T, an initial vector x_0, and parameters $0 < \mu_P \leq L_P$.

Compute $T\widetilde{q} = Mq$ for $\widetilde{q}$ and $\tau_P = \sqrt{\mu_P/L_P}$;

Set $z_0 = x_0$ and $\alpha_0 = \gamma_0 = q^\mathsf{T} M x_0$;

for $k = 0, 1, 2, \ldots,$ **do**

  Compute $\overline{x}_k = \dfrac{x_k}{\alpha_k} + \dfrac{\tau_P z_k}{\gamma_k}$;

  Normalize $\overline{x}_k$ by $\overline{x}_k = \overline{x}_k / \|\overline{x}_k\|_M$;

  Compute $\beta_k = q^\mathsf{T} M \overline{x}_k$, $\rho_k = \mathrm{Rq}(\overline{x}_k)$ and $r_k = 2(A\overline{x}_k - \rho_k M \overline{x}_k)$;

  Compute $\widetilde{r}_k = \Pi T^{-1} r_k$, where $\Pi = I - \frac{\widetilde{q} q^\mathsf{T} M}{q^\mathsf{T} M \widetilde{q}}$;

  Compute $z_{k+1} = \dfrac{(1-\tau_P) z_k}{\gamma_k} + \dfrac{\tau_P \overline{x}_k}{\beta_k} - \dfrac{\tau_P \beta_k \widetilde{r}_k}{\mu_P}$;

  Normalize z_{k+1} by $z_{k+1} = z_{k+1} / \|z_{k+1}\|_M$;

  Compute $\gamma_{k+1} = q^\mathsf{T} M z_{k+1}$;

  Solve a local optimization problem $x_{k+1} = \displaystyle\operatorname*{arg\,min}_{x \in \mathcal{X} \cap \operatorname{span}\{q, x_k, \overline{x}_k, \widetilde{r}_k\}} \mathrm{Rq}(x)$;

  Compute $\alpha_{k+1} = q^\mathsf{T} M x_{k+1}$;

end

Remark 5.1. According to Stewart's analysis of oblique projectors in [35], the cancellation may happen during computing the complementation Π. A remedy is to repeat the process, which is called recomplementation.

Each iteration of EPIC involves one matrix-vector multiplication of A for computing the residual vector $r_k = 2(A\overline{x}_k - \rho_k M \overline{x}_k)$, one preconditioned linear system $T^{-1} r_k$, and one Rayleigh–Ritz procedure. The main difference with LOPCG is that the Rayleigh–Ritz procedure of LOPCG is carried out in a three-dimensional subspace while EPIC is in a four-dimensional subspace. When taking matrix-vector multiplications of M into account, as Mq can be computed in advance, we only need to compute two M-orthogonalizations, i.e., $\overline{x}_k$ and z_{k+1}, and one M matrix-vector multiplication for residual vector r_k, while LOPCG only needs one matrix-vector multiplication. Since the main cost comes from the preconditioned linear systems and matrix-vector multiplications of A, the cost of EPIC and LOPCG are essentially the same.

5.1. Convergence analysis of EPIC. Similarly to Theorem 4.1, we can establish the convergence of EPIC as follows.

THEOREM 5.1. *Given a positive definite matrix pair* (A, M) *with a simple smallest eigenvalue* λ_1, *i.e.,* $0 < \lambda_1 < \lambda_2$, *assume that the initial vector* $z_0 = x_0 \in \mathcal{X}$ *of EPIC is chosen such that*

$$0 \leq \mathrm{Rq}(x_0) - \lambda_1 \leq \frac{1}{\max\{8\kappa_P, 2\kappa_P(1+\tau_P\kappa_P)^2\}} (\mathrm{Rq}(q) - \lambda_1),$$

where $\mathrm{Rq}(q)$ *satisfies the condition* (2.23), *and* q *is the vector in* (2.2) *to define the auxiliary problem* (2.9), $\kappa_P = L_P/\mu_P$, μ_P *and* L_P *are convexity parameters of* $\phi(y)$ *defined in Theorem* 2.2. *If the step size* τ_P *satisfies* $0 < \tau_P \leq \kappa_P^{-1/2}$, *then the Rayleigh quotient sequence of* x_k *generated by EPIC (Algorithm* 1*) satisfies*

$$\mathrm{Rq}(x_k) \leq \mathrm{Rq}(x_{k-1}) \leq \cdots \leq \mathrm{Rq}(x_0) \tag{5.12}$$

and

$$\mathrm{Rq}(x_k) - \lambda_1 \leq 2(1-\tau_P)^k\big(\mathrm{Rq}(x_0) - \lambda_1\big). \tag{5.13}$$

The proof of Theorem 5.1 is analogous to the proof of Theorem 4.1. The only difference is replacing the standard inner-product by P inner-product.

Now let us discuss the quantification of the quality of preconditioner P and copreconditioner T. First, from practical viewpoints, the linear system $Tx = b$ should be easy to solve. From theoretical viewpoints, based on the convergence of EPIC in (5.15), the ratio ι_ξ should be close to 1. Since

$$\iota_\xi := \frac{\xi_{\max}}{\xi_{\min}} = \frac{\lambda_{\max}(P^{-1}B)}{\lambda_{\min}(P^{-1}B)} \leq \frac{\nu_{\max}}{\nu_{\min}} = \frac{\lambda_{\max}(T^{-1}A)}{\lambda_{\min}(T^{-1}A)} := \iota_\nu, \tag{5.14}$$

the copreconditioner T should be chosen as a good spectral approximation of A, i.e., ι_ν is close to 1. As a by-product, a good preconditioner P (therefore, the copreconditioner T) enlarges the permissible region for the choice of q, since the requirement of $\mathrm{Rq}(q)$ in (2.12) is

$$\lambda_1 \leq \mathrm{Rq}(q) < \lambda_1 + \frac{\lambda_2 - \lambda_1}{2+\chi_P}, \quad \text{where} \quad \chi_P = \frac{8\lambda_2}{\lambda_1}\Big(\frac{\lambda_2+\lambda_1}{2(\lambda_2-\lambda_1)}\Big)^{1/2}\iota_\xi.$$

According to (5.14), we know $\iota_\xi \leq \iota_\nu$. Thus, when T is a good preconditioner for A, i.e., ι_ν is close to 1, the parameter χ_P will be significantly contracted, and the permissible region for $\mathrm{Rq}(q)$ is enlarged. However, even taking $\iota_\xi = \iota_\nu = 1$, the region is still very limited, which is even smaller than the condition $\mathrm{Rq}(x_0) < \lambda_2$ from Knyazev and Neymeyr [15].

To end this section, let us compare the convergence rate of EPIC with other methods. Neglecting the term $\mathcal{O}(\mathrm{Rq}(q) - \lambda_1)$ of μ_P and L_P in Corollary 2.1, the rate of convergence of EPIC is

$$\mathrm{Rq}(x_k) - \lambda_1 \leq 2(1-\sqrt{\eta_\nu})^k\big(\mathrm{Rq}(x_0)-\lambda_1\big), \quad \text{where} \quad \eta_\nu = \frac{1-\lambda_1/\lambda_n}{\iota_\nu(1-\lambda_1/\lambda_2)}. \tag{5.15}$$

Clearly, the bound (5.15) is better than the following sharp estimation for the preconditioned inverse iteration in [2][1]:

$$\frac{\mathrm{Rq}(x_{k+1})-\lambda_1}{\lambda_2 - \mathrm{Rq}(x_{k+1})} \leq (1-\eta_\nu)^2 \frac{\mathrm{Rq}(x_k)-\lambda_1}{\lambda_2-\mathrm{Rq}(x_k)},$$

since the exponent of η_ν is 1/2 rather than 1. For LOPCG, Knyazev conjectured the following rate of convergence in [13, eq. (5.5)], which is essentially same as our result (5.15):

$$\frac{\mathrm{Rq}(x_{k+1})-\lambda_1}{\lambda_2 - \mathrm{Rq}(x_{k+1})} \leq \Big(1 - \frac{2\sqrt{\eta_\nu}}{1+\sqrt{\eta_\nu}}\Big)^2 \frac{\mathrm{Rq}(x_k)-\lambda_1}{\lambda_2-\mathrm{Rq}(x_k)}. \tag{5.16}$$

To the best of our knowledge, a proof of upper bound (5.16) is elusive so far.

Recently, a provable accelerated eigensolver with preconditioning named RAP is proposed in [32]. The RAP achieves an acceleration similar to (5.15), but the analysis is different. For RAP, due to the operations on a manifold, preconditioning will significantly change problems since the objective function is modified. The theoretical guarantee of acceleration involves additional terms, beyond η_ν, related to the preconditioner T. In contrast, EPIC benefits from the natural incorporation of preconditioning, facilitated by the subtle structure of implicit convexity and the transformation between the eigenvalue problem and the auxiliary problem.

[1]The result in [2] is slightly different, where there is no λ_n in η_ν.

6. Numerical experiments. In this section, numerical results are presented to support our theoretical analysis in the previous sections. In the first example, we examine the sharpness of the convergence rate $1-\kappa_P^{-1/2}$ of EPIC in Theorem 5.1. Our focus will be on the exponent $-1/2$ in $\kappa_P^{-1/2}$. In the second example, we select a set of matrix pairs from SuiteSparse matrix collection to compare convergence behaviors of EPIC and LOPCG.

6.1. Sharpness of the estimated convergence rate of EPIC. Following the setting in [13, sect. 6], let

$$A = \mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \quad \text{and} \quad M = I,$$

where $\lambda_i = \omega^{i-1}$ for some $\omega > 1$. Then the relative spectral gap of (A, M) is

$$\frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1} = \frac{\omega^{n-1} - 1}{\omega - 1} \geq \omega^{n-2}.$$

When $\omega > 1$, the gap grows exponentially and the eigenvalue problem is ill-conditioned. Let the preconditioner P be given by $P = Q_\perp^\mathsf{T} T Q_\perp$, where q is an approximation of the eigenvector u_1, the matrix $[q, Q_\perp]$ is M-orthonormal, and the copreconditioner T is constructed as

$$T = A^{1/2} S^{-1} D^{-1} S A^{1/2},$$

where S and S^{-1} are the discrete sine transformation matrix and its inverse, which can be implemented by the MATLAB built-in functions `dst` and `idst`, respectively, and

$$D = \mathtt{Diag}\big(\mathtt{logspace}(0, \mathtt{log10}(\iota_\nu), \mathtt{n})\big),$$

where $\iota_\nu > 1$ is a parameter.

By variational characterizations of eigenvalues, we know that

$$\begin{aligned} \nu_{\min} &:= \lambda_{\min}(T^{-1}A) = 1 \leq \lambda_{\min}(Q_\perp^\mathsf{T} A Q_\perp, Q_\perp^\mathsf{T} T Q_\perp) = \lambda_{\min}(P^{-1}B) := \xi_{\min}, \\ \nu_{\max} &:= \lambda_{\max}(T^{-1}A) = \iota_\nu \geq \lambda_{\max}(Q_\perp^\mathsf{T} A Q_\perp, Q_\perp^\mathsf{T} T Q_\perp) = \lambda_{\max}(P^{-1}B) := \xi_{\max}. \end{aligned}$$

Consequently, by Corollary 2.1, up to the first order of $\mathrm{Rq}(q) - \lambda_1$, parameters μ_P and L_P for convexity of function ϕ in Theorem 2.2 are

$$\mu_P = 2\xi_{\min}\Big(1 - \frac{\lambda_1}{\lambda_2}\Big) \geq \frac{2(\omega-1)}{\omega} + \mathcal{O}(\mathrm{Rq}(q) - \lambda_1), \tag{6.1a}$$

$$L_P = 2\xi_{\max}\Big(1 - \frac{\lambda_1}{\lambda_n}\Big) \leq \frac{2\iota_\nu(\omega^{n-1} - 1)}{\omega^{n-1}} + \mathcal{O}(\mathrm{Rq}(q) - \lambda_1). \tag{6.1b}$$

Then, the condition number of the auxiliary function ϕ in P inner-product is

$$\kappa_P = \frac{L_P}{\mu_P} \leq \frac{\omega^{n-1} - 1}{\omega^{n-2}(\omega - 1)}\iota_\nu + \mathcal{O}(\mathrm{Rq}(q) - \lambda_1) = \frac{\omega \iota_\nu}{\omega - 1} + \mathcal{O}(\mathrm{Rq}(q) - \lambda_1).$$

For fixed ω and n, neglecting the term $\mathcal{O}(\mathrm{Rq}(q) - \lambda_1)$, the condition number κ_P is bounded by

$$\iota_\nu \leq \kappa_P \leq \left(\frac{\omega}{\omega - 1}\right)\iota_\nu. \tag{6.2}$$

Therefore, we can modify ι_ν for different condition numbers κ_P of convex function ϕ.

Let $\epsilon_k = \mathrm{Rq}(x_k) - \lambda_1$, where $\{x_k\}$ are iterates of EPIC. By the rate of convergence of EPIC in (5.13), with initial value x_0 and stopping criteria $\epsilon_k \le \epsilon_*$, we have

$$\ln\Big(\frac{\epsilon_*}{2\epsilon_0}\Big) \le m_P \cdot \ln(1-\tau_P), \tag{6.3}$$

where m_P is the iteration number of EPIC to the convergence. Therefore, when the step size is chosen as $\tau_P = \kappa_P^{-1/2}$, by the first-order expansion of $\ln(1-\tau_P)$ and (6.2), we have

$$-\ln(1-\tau_P) \approx \tau_P = \kappa_P^{-1/2} \approx \iota_\nu^{-1/2}. \tag{6.4}$$

By combining (6.3) and (6.4), the relationship between ι_ν and m_P should be

$$m_P \le \frac{\ln\big(\frac{\epsilon_*}{2\epsilon_0}\big)}{-\ln(1-\tau_P)} \le C \cdot \iota_\nu^{1/2} \cdot \ln\Big(\frac{\epsilon_*}{2\epsilon_0}\Big), \tag{6.5}$$

where C is an absolute constant from the approximation (6.4). With (6.5), we expect iteration numbers of EPIC, and LOPCG based on its conjectured rate, will be at the order of $\iota_\nu^{1/2}$.

For numerical examples, we set $n = 512$ and $\omega^{n-1} = 10^{10}$. The resulting eigenvalue problem is highly ill-conditioned since the relative spectral gap is as large as

$$\frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1} = \frac{\omega^{n-1} - 1}{\omega - 1} \ge \omega^{n-2} \ge 10^9.$$

The vector q for the auxiliary problem is constructed as

$$q = \eta[1, (\omega-1)^2, \ldots, (\omega-1)^{2n-2}]^\mathsf{T},$$

where η ensures $\|q\| = 1$. In this case, the vector q is super close to the eigenvector u_1 since $\mathrm{Rq}(q) - \lambda_1 \approx 2 \times 10^{-7}$. The initial vector of the EPIC is $x_0 = q$, the step size τ_P is set as $\tau_P = \kappa_P^{-1/2}$, where $\kappa_P = L_P/\mu_P$, and parameters μ_P and L_P are selected by dropping the first-order term of $\mathrm{Rq}(q) - \lambda_1$ in (6.1). The stopping criteria are set to when relative errors of approximate eigenvalues are less than 10^{-14}, i.e., $\epsilon_* = 10^{-14}\lambda_1$.

The number of iterations of EPIC and LOPCG depicted in Table 1 are for parameters $\iota_\nu = (10k)^2$ with $k = 1, 2, \ldots, 12$. These parameters represent the effective condition numbers of auxiliary functions ϕ in the P-inner product due the bounds (6.2). The data in Table 1 validate the theoretical linear relationship between $\iota_\nu^{1/2}$, and thus the condition number κ_P, and iteration numbers m_P in (6.5) for EPIC. Meanwhile, we observe that such a linear relationship also holds for LOPCG.

The discrepancy in the actual number of iterations of EPIC and LOPCG is due to the suboptimal choice of the parameters μ_P and L_P of EPIC. It is a subject of future study to develop an adaptive strategy to adjust these parameters to minimize the number of the iteration of EPIC.

TABLE 1

Iteration numbers of EPIC and LOPCG with respect to the square roots of condition numbers κ_P.

$\iota_\nu^{1/2}$ ($\approx \kappa_P^{1/2}$)	10	20	30	40	50	60	70	80	90	100	110	120
EPIC	170	330	476	618	759	929	1074	1217	1351	1481	1612	1744
LOPCG	78	142	201	257	312	365	416	467	518	566	615	664

Table 2
Iteration numbers and elapsed times.

Matrices (A, M)	n	LOPCG	Time (s)	EPIC	Time (s)
2cubes_sphere	101492	82	1.8985	62	**1.6748**
boneS01	127224	×	×	516	**27.0412**
Dubcova3	146689	217	7.7819	150	**6.2790**
finan512	74752	74	0.9762	52	**0.7907**
G2_circuit	150102	18	**0.4572**	22	0.6808
(bcsstk09,bcsstm09)	1083	52	**0.1154**	52	0.1209
(bcsstk21,bcsstm21)	3600	97	**0.2354**	95	0.2487
(Kuu,Muu)	7102	49	**0.2667**	51	0.3166

6.2. Test matrices from the SuiteSparse matrix collection. In this part, we compare numerical behaviors of EPIC and LOPCG with a set of test matrices (A, M) from the SuiteSparse matrix collection [6].

The vector q is chosen as a random Gaussian vector with normalization. For both methods, initial vectors are set to $x_0 = q$. Since the choice of q will affect the behavior of EPIC, and the probability of a random Gaussian vector satisfying the condition in Theorem 2.2 is extremely low, a restart strategy will be applied to EPIC. Specifically, when $|x_k^{\mathsf{T}} M q| < 0.5$, we restart EPIC with $q = x_k$. Actually, such a restart scheme will significantly improve the behavior of EPIC in our experiments.

For the copreconditioner T, we employ the aggregation-based algebraic multigrid preconditioner [24]. Differently from previous experiments, less attention will be paid to the choice of μ and L in EPIC. We just set $\mu = L = 6$ for all test matrices.

The stopping criteria of EPIC and LOPCG are chosen as when the relative errors of approximate eigenvalue are less than 10^{-8}, i.e., $\mathrm{Rq}(x_k) - \lambda_1 \leq 10^{-8}\lambda_1$, where λ_1 is computed from the MATLAB built-in function `eigs`.

Numerical results are depicted in Table 2 and Figure 2. We can see that convergence histories of EPIC and LOPCG are very close, for both the Rayleigh quotient and the components in u_1. In terms of elapsed times per iteration, EPIC is slightly longer than LOPCG. We observe that the restart of EPIC only happens in the very early stages.

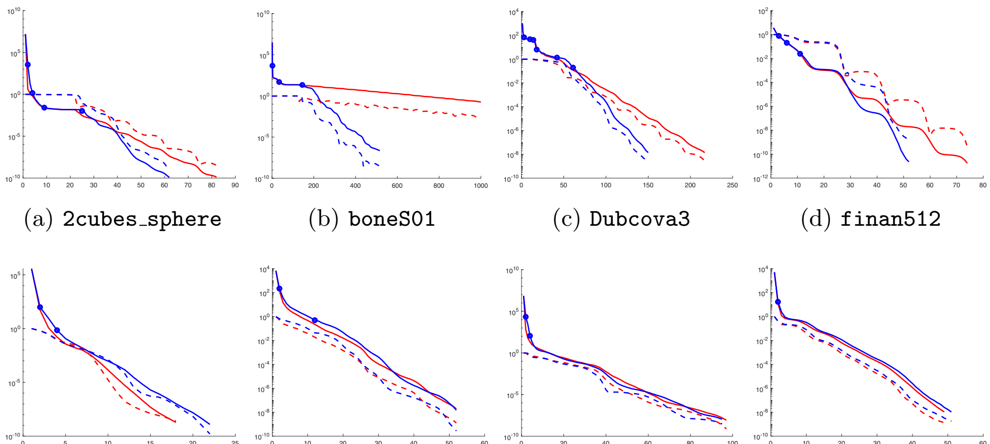
For the test matrix `boneS01`, LOPCG does not converge in 1000 iterations. In this case, EPIC outperforms LOPCG significantly. It is observed from Figure 2(b) that the LOPCG could converge linearly. In the last 600 iterations, the convergence rate is roughly 0.9949, which aligns closely with the theoretical result for acceleration as

$$1 - \sqrt{\eta_\nu} \approx 1 - \left(\frac{1 - \lambda_1/\lambda_n}{\iota_\nu (1 - \lambda_1/\lambda_2)} \right)^{1/2} \approx 0.9945.$$

This experiment highlights that, beyond acceleration, EPIC may discover advantageous pathways for achieving faster convergence.

7. Concluding remarks. We introduce the concept of implicit convexity of symmetric eigenvalue problems. A symmetric EPIC with provable acceleration is proposed. Numerical results verify the theoretical rate of convergence of EPIC, and show similar rates of convergence of EPIC and LOPCG for a set of test matrices from applications.

There are two research directions for future work. One is how to develop a parameter-free variant similar to LOPCG, and the other one is the development of a block version of EPIC.



(a) `2cubes_sphere` (b) `boneS01` (c) `Dubcova3` (d) `finan512`

(e) `G2_circuit` (f) `bcsst(k/m)09` (g) `bcsst(k/m)21` (h) `(K/M)uu`

FIG. 2. *Convergence history of LOPCG (red) and EPIC (blue). The x-axis is the iterations number. Solid lines are relative errors of approximate smallest eigenvalues, and dashed lines are* $1-|x_k^\mathsf{T} M u_1|$*, where* x_k *and* u_1 *are both* M*-normalized. Restart points are marked by a circle. Note: color appears only in the online article.*

Appendix A. Proof of Lemma 2.1. For item 1: For any $x_1, x_2 \in \mathcal{S}_q^{n-1}$, if $\psi(x_1) = \psi(x_2)$, we have

$$Q_\perp^\mathsf{T} M\Big(\frac{x_1}{q^\mathsf{T} M x_1} - \frac{x_2}{q^\mathsf{T} M x_2}\Big) = 0.$$

By the M-orthogonality of q and $Q_\perp$, there exists $\alpha \in \mathbb{R}$ such that

$$\frac{x_1}{q^\mathsf{T} M x_1} - \frac{x_2}{q^\mathsf{T} M x_2} = \alpha q.$$

Multiplying $q^\mathsf{T} M$ on the left of both sides in the above equation, we have $\alpha = 0$, i.e.,

$$x_1 = \frac{q^\mathsf{T} M x_1}{q^\mathsf{T} M x_2} x_2.$$

Then $x_1 = x_2$ is obtained by $q^\mathsf{T} M x > 0$ and $\|x\|_M = 1$ for all $x \in \mathcal{S}_q^{n-1}$.

For $\psi^\dagger$, if $\psi^\dagger(y_1) = \psi^\dagger(y_2)$, we have

$$Q_\perp\Big(\frac{y_1}{\|Q_\perp y_1 + q\|} - \frac{y_2}{\|Q_\perp y_2 + q\|}\Big) = \Big(\frac{1}{\|Q_\perp y_2 + q\|} - \frac{1}{\|Q_\perp y_1 + q\|}\Big) q.$$

Using the M-orthogonality of q and $Q_\perp$, we know that $y_1 = y_2$.

For item 2, by direct computation, for any $x \in \mathcal{S}_q^{n-1}$,

$$\psi^\dagger\big(\psi(x)\big) = \frac{\dfrac{Q_\perp Q_\perp^\mathsf{T} M x}{q^\mathsf{T} M x} + q}{\Big\|\dfrac{Q_\perp Q_\perp^\mathsf{T} M x}{q^\mathsf{T} M x} + q\Big\|_M} = \frac{\dfrac{x - qq^\mathsf{T} M x}{q^\mathsf{T} M x} + q}{\Big\|\dfrac{x - qq^\mathsf{T} M x}{q^\mathsf{T} M x} + q\Big\|_M} = x,$$

because of $q^\mathsf{T} M x > 0$ and $qq^\mathsf{T} M + Q_\perp Q_\perp^\mathsf{T} M = I$.

For item 3, for any $y \in \mathbb{R}^{n-1}$, by $\psi^\dagger\big(\psi(x)\big) = x$, we know

$$\psi^\dagger\Big(\psi\big(\psi^\dagger(y)\big)\Big) = \psi^\dagger(y).$$

Then $\psi\big(\psi^\dagger(y)\big) = y$ is obtained by ψ is an injection.

Appendix B. Proof of Lemma 2.3. For the first bound (2.16), let us begin with an upper bound of the angle between x_1 and $x_2 \in \mathcal{X}$. Suppose x_j for $j = 1$ and 2 admit decomposition on the basis of the M-orthonormal eigenvectors u_i:

$$x_j = \sum_{i=1}^{n} c_{i,j} u_i \quad \text{and} \quad \sum_{i=1}^{n} c_{i,j}^2 = 1. \tag{B.1}$$

Since $x_j \in \mathcal{X}$, we have

$$\mathrm{Rq}(q) \geq \mathrm{Rq}(x_j) = \sum_{i=1}^{n} c_{i,j}^2 \lambda_i \geq c_{1,j}^2 \lambda_1 + (1 - c_{1,j}^2)\lambda_2.$$

Combining it with the assumption $\mathrm{Rq}(q) < (\lambda_1 + \lambda_2)/2$, we know that

$$c_{1,j}^2 \geq \frac{\lambda_2 - \mathrm{Rq}(q)}{\lambda_2 - \lambda_1} > \frac{1}{2}, \tag{B.2}$$

which means for any $x \in \mathcal{X}$, $x^\mathsf{T} M u_1 \neq 0$. Since $\mathcal{X}$ is connected on the hemisphere $\mathcal{S}_q^{n-1}$ and $u_1 \in \mathcal{X}$, we have $c_{1,j} > 0$. By the simple fact

$$\begin{aligned} \sum_{i=2}^{n} c_{i,1} c_{i,2} &= \frac{1}{2} \sum_{i=2}^{n} \big((c_{i,1} + c_{i,2})^2 - (c_{i,1}^2 + c_{i,2}^2) \big) \\ &\geq -\frac{1}{2} \sum_{i=2}^{n} (c_{i,1}^2 + c_{i,2}^2) = \frac{1}{2}(c_{1,1}^2 + c_{1,2}^2) - 1, \end{aligned}$$

we have the following upper bound on the angle between x_1 and $x_2 \in \mathcal{X}$:

$$\begin{aligned} x_1^\mathsf{T} M x_2 &= \sum_{i=1}^{n} c_{i,1} c_{i,2} \geq c_{1,1} c_{1,2} + \frac{1}{2}(c_{1,1}^2 + c_{1,2}^2) - 1 \\ &= \frac{(c_{1,1} + c_{1,2})^2}{2} - 1 \geq 1 - \frac{2(\mathrm{Rq}(q) - \lambda_1)}{\lambda_2 - \lambda_1}, \end{aligned} \tag{B.3}$$

where (B.2) is used in the last inequality. Note that

$$q^\mathsf{T} M \psi^\dagger(y) = \frac{q^\mathsf{T} M (Q_\perp y + q)}{\|Q_\perp y + q\|_M} = \frac{1}{\sqrt{y^\mathsf{T} y + 1}}$$

holds for all $y \in \mathbb{R}^{n-1}$. Taking $x_1 = q$ and $x_2 = \psi^\dagger(y)$ in (B.3), we have

$$\frac{1}{y^\mathsf{T} y + 1} = (q^\mathsf{T} M x)^2 \geq \Big(1 - \frac{2(\mathrm{Rq}(q) - \lambda_1)}{\lambda_2 - \lambda_1} \Big)^2,$$

which is the first desired bound (2.16).

For the second bound (2.17), it is sufficient to show

$$|s^\mathsf{T} \nabla\phi(y) y^\mathsf{T} s + s^\mathsf{T} y \big(\nabla\phi(y)\big)^\mathsf{T} s| \leq 2|s^\mathsf{T} \nabla\phi(y)||s^\mathsf{T} y| \leq \chi_g \|s\|_B^2$$

holds for any $s \in \mathbb{R}^{n-1}$. We will prove

$$|s^\mathsf{T} \nabla\phi(y)| \leq 2\|s\|_B \left(\frac{\mathrm{Rq}(q)(\mathrm{Rq}(q) - \lambda_1)}{\lambda_1(1 + \|y\|^2)} \right)^{1/2}, \tag{B.4}$$

$$|s^\mathsf{T} y| \leq \frac{\|s\|_B \|y\|}{\sqrt{\lambda_1}}. \tag{B.5}$$

First, consider the bound (B.4). The gradients of $\mathrm{Rq}(x)$ and $\phi(y)$ are easily computed as follows:

$$\text{(B.6)}\qquad \begin{aligned}\nabla \mathrm{Rq}(x) &= 2\big(Ax - \mathrm{Rq}(x)Mx\big),\\ \nabla\phi(y) &= \frac{2}{\|y\|^2+1}\big(By - \phi(y)y + Q_\perp^\mathsf{T} Aq\big).\end{aligned}$$

Let $x = \psi^\dagger(y)$, and noting that $B = Q_\perp^\mathsf{T} A Q_\perp$, $\mathrm{Rq}(x) = \phi(y)$, $Q_\perp Q_\perp^\mathsf{T} M + qq^\mathsf{T} M = I$, and

$$1 + \|y\|^2 = 1 + \frac{x^\mathsf{T} M Q_\perp Q_\perp^\mathsf{T} Mx}{(q^\mathsf{T} Mx)^2} = \frac{x^\mathsf{T} Mx}{(q^\mathsf{T} Mx)^2} = \frac{1}{(q^\mathsf{T} Mx)^2},$$

we have

$$\begin{aligned}\nabla\phi(y) &= \frac{2}{\|y\|^2+1}\left(\frac{Q_\perp^\mathsf{T} A Q_\perp Q_\perp^\mathsf{T} Mx}{q^\mathsf{T} Mx} - \frac{\mathrm{Rq}(x) Q_\perp^\mathsf{T} Mx}{q^\mathsf{T} Mx} + Q_\perp^\mathsf{T} Aq\right)\\ &= \frac{2}{\sqrt{1+\|y\|^2}}\big(Q_\perp^\mathsf{T} A(Q_\perp Q_\perp^\mathsf{T} M + qq^\mathsf{T} M)x - \mathrm{Rq}(x) Q_\perp^\mathsf{T} Mx\big)\\ &= \frac{2}{\sqrt{1+\|y\|^2}}\big(Q_\perp^\mathsf{T} Ax - \mathrm{Rq}(x) Q_\perp^\mathsf{T} Mx\big) = \frac{Q_\perp^\mathsf{T}\nabla\mathrm{Rq}(x)}{\sqrt{1+\|y\|^2}}.\end{aligned}$$

Then by the Cauchy–Schwarz inequality,

$$\text{(B.7)}\qquad |s^\mathsf{T}\nabla\phi(y)| = \frac{\big|(Q_\perp s)^\mathsf{T}\nabla\mathrm{Rq}(x)\big|}{\sqrt{1+\|y\|^2}} \le \frac{\|Q_\perp s\|_A \|\nabla\mathrm{Rq}(x)\|_{A^{-1}}}{\sqrt{1+\|y\|^2}} = \frac{\|s\|_B\|\nabla\mathrm{Rq}(x)\|_{A^{-1}}}{\sqrt{1+\|y\|^2}}.$$

Let $\rho = \mathrm{Rq}(x) = \phi(y)$, and assuming $x = \sum_{i=1}^n c_i u_i$ like (B.1), we know that

$$\|\nabla\mathrm{Rq}(x)\|_{A^{-1}}^2 = 4\sum_{i=1}^n \frac{c_i^2(\lambda_i - \rho)^2}{\lambda_i}.$$

Since $x \in \mathcal{X}$, we have $\rho \le \mathrm{Rq}(q) < \lambda_2$; then

$$\sum_{i=1}^n \frac{c_i^2(\lambda_i-\rho)^2}{\lambda_i} \le \frac{c_1^2(\rho-\lambda_1)^2}{\lambda_1} + \sum_{i=2}^n c_i^2(\lambda_i - \rho) = \frac{c_1^2\rho(\rho-\lambda_1)}{\lambda_1} \le \frac{\rho(\rho-\lambda_1)}{\lambda_1},$$

where the equation is based on the fact $\sum_{i=1}^n c_i^2 = 1$ and $\sum_{i=1}^n c_i^2\lambda_i = \rho$. Combining these two relationships above, we know

$$\text{(B.8)}\qquad \|\nabla\mathrm{Rq}(x)\|_{A^{-1}}^2 \le \frac{4\rho(\rho-\lambda_1)}{\lambda_1}.$$

Then (B.4) is proved by $\rho \le \mathrm{Rq}(q)$, (B.7), and (B.8).

For the bound (B.5), by the Cauchy–Schwarz inequality and the Courant–Fischer minimax theorem, we have

$$s^\mathsf{T} y = (Q_\perp s)^\mathsf{T} M Q_\perp y \le \|Q_\perp s\|_A \|MQ_\perp s\|_{A^{-1}} \le \frac{\|s\|_B\|MQ_\perp y\|_{M^{-1}}}{\sqrt{\lambda_1}} = \frac{\|s\|_B\|y\|}{\sqrt{\lambda_1}}.$$

Then the bound (2.17) is proved by the assumption $\mathrm{Rq}(q) < (\lambda_1+\lambda_2)/2$, (B.4), and (B.5), and the first result in (2.16).

Appendix C. Proof of Corollary 2.2. In Theorem 2.1, it has been proved that

$$\mathcal{Y} = \{y \in \mathbb{R}^{n-1} \mid \lambda_1 \le \phi(y) \le \mathrm{Rq}(q)\}.$$

By $\nabla\phi(y_*) = \mathbf{0}$, estimation (2.21) is directly obtained from Theorem 2.2.

Now if condition (2.22) is satisfied, let $\mathcal{D}$ be a set for y satisfying (2.22) as

$$\mathcal{D} := \left\{ y \,\middle|\, \|y - y_*\|_P^2 \le \frac{2(\mathrm{Rq}(q) - \lambda_1)}{L_P} \right\}.$$

We will show that $\mathcal{D} \subset \mathcal{Y}$. Otherwise, there exists a $\widehat{y}_1 \in \mathcal{D}$ but $\phi(\widehat{y}_1) > \mathrm{Rq}(q)$. Note that $\phi(y_*) \le \mathrm{Rq}(q)$; by the intermediate value theorem and convexity of $\mathcal{D}$, there exists a $\widehat{y}_2 \in \mathcal{D}$ such that $\phi(\widehat{y}_2) = \mathrm{Rq}(q)$ and

$$\|\widehat{y}_2 - y_*\|_P^2 < \|\widehat{y}_1 - y_*\|_P^2 \le \frac{2(\mathrm{Rq}(q) - \lambda_1)}{L_P},$$

where the last inequality uses the fact $\widehat{y}_1 \in \mathcal{D}$. Noticing that $\widehat{y}_2 \in \mathcal{Y}$ due to $\phi(\widehat{y}_2) = \mathrm{Rq}(q)$, we can obtain

$$\phi(\widehat{y}_2) - \lambda_1 \le \frac{L_P}{2}\|\widehat{y}_2 - y_*\|_P^2 < \mathrm{Rq}(q) - \lambda_1$$

by (2.21), which is in contradiction to $\phi(\widehat{y}_2) = \mathrm{Rq}(q)$.

Appendix D. Proof of Theorem 4.1. The monotonically decreasing property of the Rayleigh quotient sequence $\mathrm{Rq}(x_k)$ in (4.7) is a direct consequence of the local optimization problem (4.5).

For the convergence of the Rayleigh quotient sequence $\mathrm{Rq}(x_k)$ in (4.8), since EIC is equivalent to applying the LONAG (4.1) for auxiliary problem (2.9), the convergence of EIC can be concluded by verifying that the assumption (3.9) of Theorem 3.3 is satisfied if the initial vector x_0 is chosen to satisfy (4.6). Therefore, for the rest of the proof, we need to show that

(i) if the initial vector x_0 of EIC is chosen to satisfy (4.6), then the assumption (3.9) of Theorem 3.2 holds, i.e.,

$$\{y \mid \phi(y) \le \phi(y_0)\} \subset \mathcal{B}_{R_1} \subset \mathcal{B}_{R_2} \subset \mathcal{Y},$$

where $R_1 = (2\mathcal{L}_0/\mu)^{1/2}$ and $R_2 = \max\{2R_1, (1+\tau\kappa)R_1\}$.

(ii) From the decrease (3.13) of the discrete Lyapounov function $\mathcal{L}_k$ of Theorem 3.2, we show the convergence of the Rayleigh quotient sequence $\{\mathrm{Rq}(x_k)\}$ as in (4.8).

For item (i), by Proposition 2.1, we know that $\mathrm{Rq}(x_0) = \phi(y_0)$. Therefore we need to show that if

$$\phi(y_0) - \lambda_1 \le \frac{1}{\max\{8\kappa, 2\kappa(1+\tau\kappa)^2\}}\big(\mathrm{Rq}(q) - \lambda_1\big), \tag{D.1}$$

then the assumption (3.9) of Theorem 3.2 holds. Let us first show that

$$\{y \mid \phi(y) \le \phi(y_0)\} \subset \mathcal{B}_{R_1}. \tag{D.2}$$

In fact, by Theorem 2.1 and $\mathrm{Rq}(x_0) \le \mathrm{Rq}(q)$, we have

$$\{y \mid \phi(y) \le \phi(y_0)\} = \{y \mid \phi(y) \le \mathrm{Rq}(x_0)\} \subset \{y \mid \phi(y) \le \mathrm{Rq}(q)\} = \mathcal{Y}.$$

Furthermore, for any y satisfying $\phi(y) \leq \phi(y_0)$, by the convexity of ϕ on $\mathcal{Y}$, the first-order characterization (1.3), and $\nabla\phi(y_*) = \mathbf{0}$, we have

$$\|y - y_*\|^2 \leq \frac{2}{\mu}\big(\phi(y) - \phi(y_*)\big) \leq \frac{2}{\mu}\big(\phi(y_0) - \phi(y_*)\big) \leq \frac{2\mathcal{L}_0}{\mu}, \tag{D.3}$$

which means $y \in \mathcal{B}_{R_1}$. Therefore, (D.2) is proved.

For the other two relationships, i.e.,

$$\mathcal{B}_{R_1} \subset \mathcal{B}_{R_2} \quad \text{and} \quad \mathcal{B}_{R_2} \subset \mathcal{Y}, \tag{D.4}$$

the first one comes from $R_1 \leq R_2$. For the second one, noting that

$$\mathcal{L}_0 = \phi(y_0) - \phi(y_*) + \frac{\mu}{2}\|y_0 - y_*\|^2 \leq 2\big(\phi(y_0) - \phi(y_*)\big) = 2(\mathrm{Rq}(x_0) - \lambda_1),$$

we can obtain $\mathcal{B}_{R_2} \subset \mathcal{Y}$ by

$$R_2^2 = \max\{8, 2(1+\tau\kappa)^2\}\frac{\mathcal{L}_0}{\mu} \leq \max\{16, 4(1+\tau\kappa)^2\}\frac{\mathrm{Rq}(x_0) - \lambda_1}{\mu} \leq \frac{2(\mathrm{Rq}(q) - \lambda_1)}{L}$$

and Corollary 2.2. Combining (D.2) and (D.4), we conclude (3.9).

For item (ii), since the assumption (3.9) of Theorem 3.2 holds when $\mathrm{Rq}(x_0)$ satisfies (4.6), we have the convergence of LONAG in (3.15) as

$$\phi(y_k) - \lambda_1 \leq (1-\tau)^k \mathcal{L}_0.$$

Combining it with Proposition 2.1, we have

$$\mathrm{Rq}(x_k) - \lambda_1 = \phi(y_k) - \phi(y_*) \leq 2(1-\tau)^k(\mathrm{Rq}(x_0) - \lambda_1),$$

which is the result (4.8).

Acknowledgments. The authors would like to thank Long Chen and Daniel Kressner for fruitful discussions. They would also like to thank the anonymous referees for their constructive comments and suggestions on the early version of the manuscript, in particular for the structure of Theorem 2.2 and the simplified proof of the inequality (B.3). Part of this work was performed when the first author Shao was at School of Mathematical Sciences, Fudan University.

REFERENCES

[1] F. Alimisis and B. Vandereycken, *Geodesic convexity of the symmetric eigenvalue problem and convergence of steepest descent*, J. Optim. Theory Appl., 203 (2024), pp. 920–959, https://doi.org/10.1007/s10957-024-02538-8.

[2] M. E. Argentati, A. V. Knyazev, K. Neymeyr, E. E. Ovtchinnikov, and M. Zhou, *Convergence theory for preconditioned eigenvalue solvers in a nutshell*, Found. Comput. Math., 17 (2017), pp. 713–727, https://doi.org/10.1007/s10208-015-9297-1.

[3] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*, Software Environ. Tools 11, SIAM, Philadelphia, 2000, https://doi.org/10.1137/1.9780898719581.

[4] M. Benzi, *Preconditioning techniques for large linear systems: A survey*, J. Comput. Phys., 182 (2002), pp. 418–477, https://doi.org/10.1006/jcph.2002.7176.

[5] W. Chen, N. Shao, and X. Xu, *A locally optimal preconditioned Newton-Schur method for symmetric elliptic eigenvalue problems*, Math. Comp., 92 (2023), pp. 2655–2684, https://doi.org/10.1090/mcom/3860.

[6] T. A. Davis and Y. Hu, *The University of Florida sparse matrix collection*, ACM Trans. Math. Software, 38 (2011), 1, https://doi.org/10.1145/2049662.2049663.

[7] J. A. Duersch, M. Shao, C. Yang, and M. Gu, *A robust and efficient implementation of LOBPCG*, SIAM J. Sci. Comput., 40 (2018), pp. C655–C676, https://doi.org/10.1137/17M1129830.

[8] E. G. D'yakonov, *Optimization in Solving Elliptic Problems*, CRC Press, Boca Raton, FL, 1996, https://doi.org/10.1201/9781351075213.

[9] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353, https://doi.org/10.1137/S0895479895290954.

[10] G. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., The Johns Hopkins University Press, Baltimore, MD, 2013.

[11] A. Knyazev and A. Shorokhodov, *On exact estimates of the convergence rate of the steepest ascent method in the symmetric eigenvalue problem*, Linear Algebra Appl., 154 (1991), pp. 245–257, https://doi.org/10.1016/0024-3795(91)90379-B.

[12] A. V. Knyazev, *Preconditioned eigensolvers - An oxymoron?*, Electron. Trans. Numer. Anal., 7 (1998), pp. 104–123.

[13] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541, https://doi.org/10.1137/S1064827500366124.

[14] A. V. Knyazev, M. E. Argentati, I. Lashuk, and E. Ovtchinnikov, *Block locally optimal preconditioned eigenvalue xolvers (BLOPEX) in Hypre and PETSc*, SIAM J. Sci. Comput., 29 (2007), pp. 2224–2239, https://doi.org/10.1137/060661624.

[15] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration* III*: A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114, https://doi.org/10.1016/S0024-3795(01)00461-X.

[16] A. V. Knyazev and A. L. Skorokhodov, *Preconditioned gradient-type iterative methods in a subspace for partial generalized symmetric eigenvalue problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1226–1239, https://doi.org/10.1137/0731064.

[17] H. Luo and L. Chen, *From differential equation solvers to accelerated first-order methods for convex optimization*, Math. Program., 195 (2022), pp. 735–781, https://doi.org/10.1007/s10107-021-01713-3.

[18] M. Muehlebach and M. Jordan, *Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives*, J. Mach. Learn. Res., 22 (2021), pp. 1–50, http://jmlr.org/papers/v22/20-207.html.

[19] Y. Nesterov, *A method for solving the convex programming problem with convergence rate* $\mathcal{O}(1/k^2)$, Sov. Math. Dokl., 269 (1983), pp. 543–547.

[20] Y. Nesterov, *Lectures on Convex Optimization*, Springer Nature, Cham, Switzerland, 2018, https://doi.org/10.1007/978-3-319-91578-4.

[21] K. Neymeyr, *A geometric convergence theory for the preconditioned steepest descent iteration*, SIAM J. Numer. Anal., 50 (2012), pp. 3188–3207, https://doi.org/10.1137/11084488X.

[22] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 2006, https://doi.org/10.1007/978-0-387-40065-5.

[23] Y. Notay, *Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44, https://doi.org/10.1002/nla.246.

[24] Y. Notay, *An aggregation-based algebraic multigrid method*, Electron. Trans. Numer. Anal., 37 (2010), pp. 123–146.

[25] E. E. Ovtchinnikov, *Sharp convergence estimates for the preconditioned steepest descent method for Hermitian eigenvalue problems*, SIAM J. Numer. Anal., 43 (2006), pp. 2668–2689, https://doi.org/10.1137/040620643.

[26] B. O'donoghue and E. Candès, *Adaptive restart for accelerated gradient schemes*, Found. Comput. Math., 15 (2015), pp. 715–732, https://doi.org/10.1007/s10208-013-9150-3.

[27] J.-H. Park, A. J. Salgado, and S. M. Wise, *Preconditioned accelerated gradient descent methods for locally Lipschitz smooth objectives with applications to the solution of nonlinear PDEs*, J. Sci. Comput., 89 (2021), pp. 1–37, https://doi.org/10.1007/s10915-021-01615-8.

[28] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics Appl. Math. 20, SIAM, Philadelphia, 1998, https://doi.org/10.1137/1.9781611971163.

[29] B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phys., 4 (1964), pp. 1–17, https://doi.org/10.1016/0041-5553(64)90137-5.

[30] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, rev. ed., Classic Appl. Math. 66, SIAM, Philadelphia, 2011, https://doi.org/10.1137/1.9781611970739.

[31] B. Samokish, *The steepest descent method for an eigenvalue problem with semi-bounded operators*, Izv. Vyssh. Uchebn. Zaved. Mat., 5 (1958), pp. 105–114.
[32] N. Shao and W. Chen, *Riemannian Acceleration with Preconditioning for Symmetric Eigenvalue Problems*, preprint, 2309.05143, 2023, https://doi.org/10.48550/arXiv.2309.05143.
[33] B. Shi, S. S. Du, M. Jordan, and W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Math. Program., 195 (2022), pp. 79–148, https://doi.org/10.1007/s10107-021-01681-8.
[34] A. Stathopoulos, *Nearly optimal preconditioned methods for Hermitian eigenproblems under limited memory. Part* I*: Seeking one eigenvalue*, SIAM J. Sci. Comput., 29 (2007), pp. 481–514, https://doi.org/10.1137/050631574.
[35] G. W. Stewart, *On the numerical analysis of oblique projectors*, SIAM J. Matrix Anal. Appl., 32 (2011), pp. 309–348, https://doi.org/10.1137/100792093.
[36] W. Su, S. Boyd, and E. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, in Advances in Neural Information Processing Systems, Vol. 27, Curran Associates, Red Hook, NY, 2014, https://proceedings.neurips.cc/paper/2014/file/f09696910bdd874a99cd74c8f05b5c44-Paper.pdf.
[37] A. J. Wathen, *Preconditioning*, Acta Numer., 24 (2015), pp. 329–376, https://doi.org/10.1017/S0962492915000021.
[38] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965, https://doi.org/10.1017/S0013091500012104.
[39] M. Zhou, Z. Bai, Y. Cai, and K. Neymeyr, *Convergence analysis of a block preconditioned steepest descent eigensolver with implicit deflation*, Numer. Linear Algebra Appl., 30 (2023), e2498, https://doi.org/10.1002/nla.2498.
[40] M. Zhou and K. Neymeyr, *Cluster robust estimates for block gradient-type eigensolvers*, Math. Comp., 88 (2019), pp. 2737–2765, https://doi.org/10.1090/mcom/3446.