Received: 8 September 2022 | Revised: 5 January 2023 | Accepted: 26 February 2023

DOI: 10.1002/nla.2498

RESEARCH ARTICLE

WILEY

Convergence analysis of a block preconditioned steepest descent eigensolver with implicit deflation

Ming Zhou[1] | **Zhaojun Bai**[2] | **Yunfeng Cai**[3] | **Klaus Neymeyr**[1]

[1]Department of Mathematics, University of Rostock, Rostock, Mecklenburg-Vorpommern, Germany

[2]Department of Computer Science and Department of Mathematics, University of California, Davis, Davis, California, USA

[3]Cognitive Computing Lab, Baidu Research, Beijing, China

Correspondence
Ming Zhou, Universität Rostock, Institut für Mathematik, Ulmenstraße 69, 18055 Rostock, Germany.
Email: ming.zhou@uni-rostock.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 463329614; National Science Foundation, Grant/Award Number: DMS-1913364

Abstract

Gradient-type iterative methods for solving Hermitian eigenvalue problems can be accelerated by using preconditioning and deflation techniques. A preconditioned steepest descent iteration with implicit deflation (PSD-id) is one of such methods. The convergence behavior of the PSD-id is recently investigated based on the pioneering work of Samokish on the preconditioned steepest descent method (PSD). The resulting non-asymptotic estimates indicate a superlinear convergence of the PSD-id under strong assumptions on the initial guess. The present paper utilizes an alternative convergence analysis of the PSD by Neymeyr under much weaker assumptions. We embed Neymeyr's approach into the analysis of the PSD-id using a restricted formulation of the PSD-id. More importantly, we extend the new convergence analysis of the PSD-id to a practically preferred block version of the PSD-id, or BPSD-id, and show the cluster robustness of the BPSD-id. Numerical examples are provided to validate the theoretical estimates.

KEYWORDS

block eigensolvers, gradient iterations, Rayleigh quotient

1 | INTRODUCTION

Determining the smallest eigenvalues and the associated eigenfunctions of a self-adjoint elliptic partial differential operator is a common task in many application areas. The computational costs are usually high due to large dimensions of discretized problems. By considering an equivalent minimization problem for the Rayleigh quotient, one can develop suitable eigensolvers based on gradient iterations. The performance can be significantly improved by preconditioning techniques for better descent directions,[1] and by a blockwise implementation for computing clustered eigenvalues.[2,3] A further acceleration is enabled by modifying the preconditioners with certain shifts after deflation. These methodological improvements are particularly necessary for ill-conditioned eigenvalue problems arising from applications, such as the partition-of-unity finite element method for solving the Kohn-Sham equation in electronic structure calculations, where the target eigenvalues are not well separated from the rest of the spectrum; see Reference 4 and references therein.

In this article, we consider the generalized Hermitian definite matrix eigenvalue problem

$$Hu = \lambda Su, \tag{1}$$

where $H, S \in \mathbb{C}^{n\times n}$ are Hermitian, and S is positive definite. The eigenvalues of (H, S) are all real. We denote them by $\lambda_1, \ldots, \lambda_n$ and arrange them as $\lambda_1 \le \cdots \le \lambda_n$. We aim at computing the first m eigenvalues together with the associated eigenvectors. Typically, $m \ll n$.

When $m = 1$, the task can be restated as minimizing the Rayleigh quotient

$$\rho : \mathbb{C}^n \setminus \{0\} \to \mathbb{R}, \quad \rho(z) = \frac{z^* Hz}{z^* Sz}, \tag{2}$$

where the superscript * denotes conjugate transpose. The preconditioned steepest descent iteration (PSD)

$$z^{(\ell+1)} = z^{(\ell)} - \omega^{(\ell)} K r^{(\ell)}, \tag{3}$$

is applicable to the minimization problem of (2). Therein $K \in \mathbb{C}^{n\times n}$ is a Hermitian preconditioner, and the residual $r^{(\ell)} = Hz^{(\ell)} - \rho(z^{(\ell)})Sz^{(\ell)}$ is collinear with the gradient vector $\nabla\rho(z^{(\ell)})$ of the Rayleigh quotient $\rho(z^{(\ell)})$. The descent direction $-Kr^{(\ell)}$ generalizes $-\nabla\rho(z^{(\ell)})$, and is expected to enable an acceleration. An optimal stepsize $\omega^{(\ell)}$ can be determined by the Rayleigh-Ritz procedure applied on the subspace $\operatorname{span}\{z^{(\ell)}, Kr^{(\ell)}\}$.

The choice of K significantly affects the convergence rate of the PSD. Setting K equal to the $n \times n$ identity matrix I leads to a slow and mesh-dependent convergence for discretized operator eigenvalue problems.[5] A shift-and-invert preconditioner $K = (H - \sigma S)^{-1}$ with $\sigma < \lambda_1$ is much more efficient. According to the analyses of gradient eigensolvers in References 6 and 7, one can derive the single-step convergence factor

$$\kappa = \left(\frac{\lambda_1 - \sigma}{\lambda_2 - \sigma}\right)\left(\frac{\lambda_n - \lambda_2}{\lambda_n - \lambda_1}\right),$$

for approximate eigenvectors (with respect to tangent values of error angles) and additionally $\kappa^2/(2-\kappa)^2$ for approximate eigenvalues in the final phase (with respect to relative positions between λ_1 and λ_2). These convergence factors are bounded away from 1 when $\lambda_1 \ll \lambda_2$ and can be improved by empirically increasing σ.

In practice, the shift-and-invert preconditioner K is often implemented approximately, for example, by iteratively solving linear systems of the form

$$\tilde{H}p = r \quad \text{with} \quad \tilde{H} = H - \sigma S. \tag{4}$$

Thus it is more important to analyze a preconditioner $K \approx \tilde{H}^{-1}$. The analysis from Reference 8 begins with a convergence argument for (4) with respect to the error propagation matrix $I - K\tilde{H}$. By using a quality parameter ε defined in the condition

$$\|I - K\tilde{H}\|_{\tilde{H}} \le \varepsilon < 1, \tag{5}$$

the eigenvalue convergence factor $\kappa^2/(2-\kappa)^2$ for the special case $K = \tilde{H}^{-1}$ (where $\varepsilon = 0$) is generalized as

$$(\kappa + \varepsilon(2-\kappa))^2/((2-\kappa) + \varepsilon\kappa)^2. \tag{6}$$

As we can see, the convergence rate of the PSD can be deteriorated by $\varepsilon \approx 1$ or $\kappa \approx 1$. One can reduce ε by using a proper linear system solver. However, reducing κ is difficult if λ_1 and λ_2 are tightly clustered. In this case, the PSD needs to be modified, for example, by a blockwise implementation or certain subspace extensions. The resulting iterations also provide approximations of further eigenvalues.

For $m > 1$, computing the m smallest eigenvalues of the matrix pair (H, S) corresponds to the trace minimization of Z^*HZ among all S-orthonormal matrices $Z \in \mathbb{C}^{n\times m}$; compare Reference 9. A typical approach is the block version of the PSD, BPSD in short. Therein each iterate $Z^{(\ell)}$ is an S-orthonormal Ritz basis matrix, and $\Theta^{(\ell)} = (Z^{(\ell)})^* HZ^{(\ell)}$ is a diagonal matrix whose diagonal consists of the Ritz values in $\operatorname{span}\{Z^{(\ell)}\}$. The next iterate $Z^{(\ell+1)}$ is determined by applying the Rayleigh-Ritz procedure to the trial subspace

$$\operatorname{span}\{Z^{(\ell)}, KR^{(\ell)}\} \quad \text{with} \quad R^{(\ell)} = HZ^{(\ell)} - SZ^{(\ell)}\Theta^{(\ell)}.$$

The columns of $Z^{(\ell+1)}$ are S-orthonormal Ritz vectors associated with the m smallest Ritz values in span$\{Z^{(\ell)}, KR^{(\ell)}\}$. The convergence behavior of the BPSD with respect to individual Ritz values has been analyzed in Reference 10. The convergence factor for the ith Ritz value in the final phase is given by (6) with the generalized parameter

$$\kappa = \left(\frac{\lambda_i - \sigma}{\lambda_{i+1} - \sigma}\right)\left(\frac{\lambda_n - \lambda_{i+1}}{\lambda_n - \lambda_i}\right).$$

An alternative convergence factor presented in Reference 11 concerns the non-optimized version $Z^{(\ell+1)} = Z^{(\ell)} - KR^{(\ell)}$ of the BPSD, and depends on the ratio $(\lambda_i - \sigma)/(\lambda_{m+1} - \sigma)$ instead of $(\lambda_i - \sigma)/(\lambda_{i+1} - \sigma)$. This reflects the cluster robustness of the BPSD for $\lambda_i \ll \lambda_{m+1}$ which can be ensured by enlarging the block size m.

Deflation is required for block iterative eigensolvers if the number of target eigenvalues exceeds the block size. In the (B)PSD, once the residual of an approximate eigenvector is sufficiently small, one can store it in a basis matrix U consisting of accepted S-orthonormal approximate eigenvectors. The further iterates are S-orthogonalized against U so that they converge toward eigenvectors associated with the next eigenvalues. Such an orthogonalization is called implicit deflation; see Section 6.2.3 in Reference 12.

Recently, a combination of the PSD with implicit deflation (PSD-id) has been investigated in Reference 4. A remarkable feature of the PSD-id is that the preconditioner is variable depending on the current approximate eigenvalue, somewhat similarly to the Jacobi-Davidson method or an inexact Rayleigh quotient iteration.[13] This accelerates the convergence in the final phase in comparison to a fixed preconditioner.

The convergence behavior of the PSD-id is analyzed in Reference 4 based on the pioneering work of Samokish on the PSD.[1] In particular, Theorem 3.2 in Reference 4 extends a non-asymptotic reformulation of the classical estimate (10) in Reference 1 presented by Ovtchinnikov; see Theorem 2.1 in Reference 14. Precisely, a convergence rate bound for approximating an interior eigenvalue λ_i ($> \lambda_{i-1}$) with the PSD-id and a variable preconditioner K is derived by using the matrix $M = Q^*(H - \lambda_i S)Q$, where Q is an S-orthogonal projector onto the invariant subspace associated with the eigenvalues $\lambda_i, \ldots, \lambda_n$. The resulting estimate contains an essential parameter

$$q = \frac{\beta - \alpha}{\beta + \alpha},$$

where α and β are the smallest positive and largest eigenvalues of the matrix product KM. A corresponding asymptotic estimate then uses q^2 as the convergence factor. It is possible to reduce q^2 by modifying K with proper shifts for a superlinear convergence.

A drawback of Theorem 3.2 in Reference 4 is its technical assumption on parameters associated with K and M. The assumption is rather restricted as the current approximate eigenvalue has to be sufficiently close to λ_i. More significantly, it would be extremely difficult, if possible, to generalize the analysis to a block version of the PSD-id, or BPSD-id, which is the algorithm used in practice.

In this article, we first recast the PSD-id as the PSD applied to a restricted eigenvalue problem, and then utilize the approaches proposed in Reference 8 to provide a convergence analysis of the PSD-id. More importantly, we are able to present a convergence analysis of the BPSD-id by extending the convergence analysis of the BPSD in Reference 10. The resulting sharp single-step estimates of the PSD-id and the BPSD-id are presented in Theorems 2 and 3. Further estimates in Sections 3.3 and 4 deal with larger shifts and clustered eigenvalues. Numerical examples are provided to verify the sharpness of convergence estimates. These theoretical results and numerical experiments advance our in-depth understanding of the convergence behaviors of the PSD-id and the BPSD-id.

For ease of references, we restate the estimates of the PSD and the BPSD in References 8 and 10 for a standard Hermitian eigenvalue problem with notation that are compatible with a restricted formulation of the PSD-id and the BPSD-id in Section 2.

Theorem 1 (References 8,10). *Consider a Hermitian positive definite matrix $A \in \mathbb{C}^{d\times d}$ together with its eigenvalues $\zeta_1 \le \cdots \le \zeta_d$ and a Hermitian positive definite preconditioner $T \in \mathbb{C}^{d\times d}$ satisfying*

$$\|I - TA\|_A \le \varepsilon < 1. \tag{7}$$

If $\eta = (x^*Ax)/(x^*x)$ for a nonzero vector $x \in \mathbb{C}^d$ is located in an eigenvalue interval (ζ_j, ζ_{j+1}), then it holds for the smallest Ritz value η' of A in span $\{x, T(Ax - \eta x)\}$ that

$$\frac{\eta' - \zeta_j}{\zeta_{j+1} - \eta'} \leq \left(\frac{\kappa + \varepsilon(2-\kappa)}{(2-\kappa) + \varepsilon\kappa} \right)^2 \frac{\eta - \zeta_j}{\zeta_{j+1} - \eta} \quad with \quad \kappa = \frac{\zeta_j(\zeta_d - \zeta_{j+1})}{\zeta_{j+1}(\zeta_d - \zeta_j)}. \tag{8}$$

The equality in (8) is attainable in the limit case $\eta \to \zeta_j$ in an invariant subspace associated with the eigenvalues ζ_j, ζ_{j+1} and ζ_d.

Consider further a c-dimensional subspace $\mathcal{X} \subset \mathbb{C}^d$ and orthonormal Ritz vectors $x_1, \ldots, x_c$ associated with the Ritz values $\eta_1 \leq \cdots \leq \eta_c$ of A induced by $\mathcal{X}$. Denote by $\eta'_1 \leq \cdots \leq \eta'_c$ the c smallest Ritz values of A induced by span $\{X, T(AX - XD)\}$ with $X = [x_1, \ldots, x_c]$ and $D = diag(\eta_1, \ldots, \eta_c)$. For each $t \in \{1, \ldots, c\}$, if η_t is located in an eigenvalue interval (ζ_j, ζ_{j+1}), then it holds that

$$\frac{\eta'_t - \zeta_j}{\zeta_{j+1} - \eta'_t} \leq \left(\frac{\kappa + \varepsilon(2-\kappa)}{(2-\kappa) + \varepsilon\kappa} \right)^2 \frac{\eta_t - \zeta_j}{\zeta_{j+1} - \eta_t} \quad with \quad \kappa = \frac{\zeta_j(\zeta_d - \zeta_{j+1})}{\zeta_{j+1}(\zeta_d - \zeta_j)}. \tag{9}$$

The equality in (9) is attainable in the limit case $\eta_t \to \zeta_j$ in an invariant subspace associated with the eigenvalues ζ_j, ζ_{j+1} and ζ_d.

The remaining part of this article is organized as follows. In Section 2, we introduce the algorithmic structure of the (B)PSD-id, and present a restricted formulation as the starting point of our analysis. In Section 3, the estimates of the (B)PSD in Theorem 1 are applied to certain representations of the (B)PSD-id in an invariant subspace in order to derive sharp single-step estimates. Multi-step estimates are presented in Section 4 for the cluster robustness of the BPSD-id. The theoretical convergence estimates of the BPSD-id are demonstrated by numerical experiments in Section 5.

2 | PSD-ID AND BPSD-ID ALGORITHMS AND RESTRICTED FORMULATIONS

2.1 | PSD-id and BPSD-id algorithms

In this section, we reformulate the original algorithm of the PSD-id from Reference 4 in Algorithm 1. Therein we drop the step indices of the iterates since they are not needed in the derivation of our new estimates. A few remarks of Algorithm 1 are in order.

- Line 1: The S-orthogonalization against the $i-1$ accepted approximate eigenvectors is made once at initialization and then satisfied automatically in the following while-loop.
- Line 2: The stopping criterion uses the norm $\|\cdot\|_{S^{-1}}$. It implies that there exists an eigenvalue λ of (H, S) which fulfills

$$|\lambda - \rho(z)| \leq \|r\|_{S^{-1}} / \|z\|_S,$$

  compare Theorem 15.9.1 in Reference 15.
- Line 3: The preconditioner K is constructed to be effectively positive definite for ensuring the convergence; see Definition 1. In practice, $p = Kr$ can be implemented by approximate solution of the linear system $(H - \sigma S)p = r$. The shift σ will be discussed in the convergence analysis and the numerical experiments. In principle we set σ slightly smaller than the ith smallest eigenvalue λ_i, or equal to the current approximate eigenvalue $\rho(z)$ if it is close to λ_i.

Algorithm 2 describes the BPSD-id. Therein the usage of a suitable block size $\tilde{k}$ can overcome the possible convergence stagnation of the PSD-id for clustered eigenvalues; see Section 4. A few remarks of Algorithm 2 are in order.

- Line 4: The block residual R actually consists of the residuals of individual Ritz vectors. The first k columns of R are considered in the stopping criterion.
- There are two implementations with different choices of the block size for computing the smallest eigenvalues of (H, S) and the associated eigenvectors, beginning with a random initial guess $W \in \mathbb{C}^{n \times m}$ where m is larger than the number of target eigenvalues. A straightforward implementation with a fixed block size $\tilde{k} < m$ has the form

$$\text{BPSD-id}\left(W(:, 1 : i-1), W(:, i : i-1+\tilde{k}), k\right),$$

Algorithm 1. $z = \text{PSD-id}(U, z)$

Input: S-orthonormal matrix $U \in \mathbb{C}^{n\times(i-1)}$ whose columns are accepted approximate eigenvectors associated with the $i-1$ smallest eigenvalues; initial guess $z \in \mathbb{C}^n$.

Output: approximate eigenvector z associated with the ith smallest eigenvalue.

S-orthogonalize z against U: $z = z - UU^*Sz$; $z = z/\|z\|_S$; $r = Hz - \rho(z)Sz$;
while $\|r\|_{S^{-1}}$ not sufficiently small **do**
  compute a preconditioned residual $p = Kr$;
  update z by an S-normalized Ritz vector associated with the ith smallest Ritz value in the subspace $\text{span}\{U, z, p\}$;
  $r = Hz - \rho(z)Sz$;
end while

that is, each outer step only treats $\tilde{k}$ columns of W. The leading index i of these columns is initially 1 and will be updated together with $W(:, 1:i-1)$ by using already converged columns. An alternative implementation uses the block size $\tilde{k} = m - i + 1$ depending on the current index i. Thus W is entirely modified as early as in the first outer step. The latter columns of W can provide more accurate initial data for the inner loop in the next outer step.

Algorithm 2. $Z = \text{BPSD-id}(U, Z, k)$

Input: S-orthonormal matrix $U \in \mathbb{C}^{n\times(i-1)}$ whose columns are accepted approximate eigenvectors associated with the $i-1$ smallest eigenvalues; initial guess $Z \in \mathbb{C}^{n\times\tilde{k}}$, $\tilde{k}$ is the block size.

Output: matrix $Z \in \mathbb{C}^{n\times k}(k \le \tilde{k})$ consisting of approximate eigenvectors associated with the ith to the $(i-1+k)$th smallest eigenvalues.

S-orthogonalize Z against U;
update Z by S-orthonormal Ritz vectors in $\text{span}\{Z\}$;
$R = HZ - SZ(Z^*HZ)$;
while $\|R(:, 1:k)\|_{S^{-1}}$ not sufficiently small **do**
  compute a block preconditioned residual $P = KR$;
  update Z by S-orthonormal Ritz vectors associated with the ith to the $(i-1+\tilde{k})$th smallest Ritz values in the subspace $\text{span}\{U, Z, P\}$;
  $R = HZ - SZ(Z^*HZ)$;
end while
$Z = Z(:, 1:k)$;

2.2 | Restricted formulations of the PSD-id and the BPSD-id

As in Reference 4, we assume by ignoring sufficiently small numerical errors that the columns $u_1, \ldots, u_{i-1}$ of the matrix U in the PSD-id (Algorithm 1) and the BPSD-id (Algorithm 2) are exact S-orthonormal eigenvectors associated with the $i-1$ smallest eigenvalues $\lambda_1 \le \cdots \le \lambda_{i-1}$ of (H, S), and the iterate z or Z after the first S-orthogonalization against U has full rank.

As the starting point of our new convergence analysis, we represent the PSD-id by the PSD applied to a restricted eigenvalue problem. Let us first extend U by $V = [u_i, \ldots, u_n]$ as an S-orthonormal basis of $\mathbb{C}^n$ where $u_i, \ldots, u_n$ are eigenvectors associated with the remaining eigenvalues $\lambda_i \le \cdots \le \lambda_n$, that is,

$$HV = SV\Lambda \quad \text{with} \quad \Lambda = \text{diag}(\lambda_i, \ldots, \lambda_n) \quad \text{and} \quad \text{span}\{V\} = \text{span}\{U\}^{\perp_S}. \tag{10}$$

For the PSD-id, an arbitrary $\tilde{z} \in \text{span}\{V\} \setminus \{0\}$ can be represented by

$$\tilde{z} = V\tilde{c} \quad \text{with} \quad \tilde{c} = V^*S\tilde{z}. \tag{11}$$

By using the identity (10), a relation between the Rayleigh quotient $\rho(\cdot)$ of (H, S) defined in (2) and the restricted Rayleigh quotient of Λ:

$$\tilde{\rho} : \mathbb{C}^{n-i+1} \setminus \{0\} \to \mathbb{R}, \quad \tilde{\rho}(w) = \frac{w^*\Lambda w}{w^* w}, \tag{12}$$

is given by

$$\rho(\tilde{z}) = \rho(V\tilde{c}) = \frac{\tilde{c}^* V^* H V \tilde{c}}{\tilde{c}^* V^* S V \tilde{c}} \overset{(10)}{=} \frac{\tilde{c}^* V^* S V \Lambda \tilde{c}}{\tilde{c}^* \tilde{c}} = \frac{\tilde{c}^* \Lambda \tilde{c}}{\tilde{c}^* \tilde{c}} = \tilde{\rho}(\tilde{c}). \tag{13}$$

Consequently, the target eigenvalue of the PSD-id (Algorithm 1) can be interpreted by

$$\min_{\tilde{z} \in \mathbb{C}^n \setminus \{0\},\ U^* S \tilde{z} = 0} \rho(\tilde{z}) = \min_{\tilde{z} \in \mathrm{span}\{V\} \setminus \{0\}} \rho(\tilde{z}) = \min_{\tilde{c} \in \mathbb{C}^{n-i+1} \setminus \{0\}} \tilde{\rho}(\tilde{c}).$$

It implies that the PSD-id for computing the ith smallest eigenvalue of (H, S) is equivalent to the PSD for computing the smallest eigenvalue of Λ. The following lemma presents such relationship in detail.

Lemma 1. *Denote by z and z' two successive iterates of the PSD-id.*

(i) If z belongs to span$\{V\} \setminus \{0\}$, then also z'.
(ii) Let $c = V^* S z$ and $c' = V^* S z'$ be the coefficient vectors of z and z' with respect to the representation (11). Then c' is a minimizer of $\tilde{\rho}(\cdot)$ in span$\{c, \tilde{K}\tilde{r}\}$ with $\tilde{K} = V^* S K S V$ and $\tilde{r} = \Lambda c - \tilde{\rho}(c)c$.

Proof. (i) If z belongs to span$\{V\} \setminus \{0\}$, then the dimension of the trial subspace span$\{U, z, p\}$ is at least i. This verifies the existence of the ith smallest Ritz value in span$\{U, z, p\}$ and the existence of the next iterate z' which is an associated Ritz vector. Moreover, the columns of U are eigenvectors associated with the $i-1$ smallest eigenvalues and automatically Ritz vectors associated with the $i-1$ smallest Ritz values in span$\{U, z, p\}$. Thus z' is S-orthogonal to span$\{U\}$ and belongs to span$\{V\} \setminus \{0\}$.

(ii) By using the representation $z = Vc$ and the projector $Q = VV^*S$ onto span$\{V\}$, we get

$$\mathrm{span}\{U, z, p\} = \mathrm{span}\{U\} \oplus \mathrm{span}\{Vc, Qp\},$$

and

$$\begin{aligned} Qp = (VV^*S)Kr = VV^*SK\,(Hz - \rho(z)Sz) &\overset{(13)}{=} VV^*SK\,(HVc - \tilde{\rho}(c)SVc) \\ &\overset{(10)}{=} VV^*SK\,(SV\Lambda c - \tilde{\rho}(c)SVc) = VV^*SKSV\,(\Lambda c - \tilde{\rho}(c)c) = V\tilde{K}\tilde{r}. \end{aligned}$$

Therefore

$$\mathrm{span}\{U, z, p\} = \mathrm{span}\{U\} \oplus \mathrm{span}\{Vc, V\tilde{K}\tilde{r}\} = \mathrm{span}\{U\} \oplus V \cdot \mathrm{span}\{c, \tilde{K}\tilde{r}\}.$$

Recall that the columns of U are Ritz vectors associated with the $i-1$ smallest Ritz values in span$\{U, z, p\}$, the ith smallest Ritz value is just the smallest Ritz value in $V \cdot$ span$\{c, \tilde{K}\tilde{r}\}$. The associated Ritz vector z' is thus a minimizer of $\rho(\cdot)$ therein. The relation (13) ensures that minimizing $\rho(\cdot)$ in the subspace $V \cdot$ span$\{c, \tilde{K}\tilde{r}\}$ is equivalent to minimizing $\tilde{\rho}(\cdot)$ in the "coefficient subspace" span$\{c, \tilde{K}\tilde{r}\}$. Consequently, the coefficient vector c' of z' is a minimizer of $\tilde{\rho}(\cdot)$ in span$\{c, \tilde{K}\tilde{r}\}$. ∎

Lemma 1 indicates that c' is a Ritz vector associated with the smallest Ritz value of Λ in span$\{c, \tilde{K}\tilde{r}\}$. Thus c and c' can be regarded as two successive iterates of a PSD iteration for minimizing $\tilde{\rho}(\cdot)$ with $\tilde{K}$ as preconditioner. Consequently, we can analyze the PSD-id in terms of the PSD iteration with successive iterates c and c'.

Now let us represent the BPSD-id by the BPSD applied to a restricted eigenvalue problem. For an arbitrary matrix $\tilde{Z} \in \mathbb{C}^{n \times l}$ having full rank and satisfying span$\{\tilde{Z}\} \subseteq$ span$\{V\}$, one can define $\tilde{C} = V^* S \tilde{Z}$ so that $\tilde{Z} = V\tilde{C}$. Based on the

relations

$$\tilde{Z}^*H\tilde{Z} = \tilde{C}^*V^*HV\tilde{C} \overset{(10)}{=} \tilde{C}^*V^*SV\Lambda\tilde{C} = \tilde{C}^*\Lambda\tilde{C} \quad \text{and} \quad \tilde{Z}^*S\tilde{Z} = \tilde{C}^*V^*SV\tilde{C} = \tilde{C}^*\tilde{C}, \tag{14}$$

the Ritz values of (H, S) in $\text{span}\{\tilde{Z}\}$ coincide with those of Λ in $\text{span}\{\tilde{C}\}$. The respective Ritz vectors can be converted by multiplications with V or V^*S analogously to (11).

Lemma 2. *Denote by Z and Z' two successive iterates of the BPSD-id.*

(i) If Z has full rank and all its $\tilde{k}$ columns belong to $\text{span}\{V\}$, then also Z'.
(ii) Define the coefficient matrices $C = V^*SZ$ and $C' = V^*SZ'$. Then the columns of C are orthonormal Ritz vectors of Λ in $\text{span}\{C\}$, that is, $\Theta = C^*\Lambda C$ is a diagonal matrix whose diagonal entries are the corresponding Ritz values. Furthermore, the columns of C' are orthonormal Ritz vectors associated with the $\tilde{k}$ smallest Ritz values of Λ in $\text{span}\{C, \tilde{K}\tilde{R}\}$ for $\tilde{K} = V^*SKSV$ and $\tilde{R} = \Lambda C - C\Theta$.

Proof. (i) The dimension of $\text{span}\{U, Z, P\}$ is at least $i - 1 + \tilde{k}$ so that Z' is constructed by S-orthonormal Ritz vectors associated with the ith to the $(i - 1 + \tilde{k})$th smallest Ritz values of (H, S) in $\text{span}\{U, Z, P\}$, and has full rank. These Ritz vectors (columns of Z') are S-orthogonal to those associated with the $i - 1$ smallest eigenvalues and thus belong to the S-orthogonal complement of $\text{span}\{U\}$, that is, $\text{span}\{V\}$.

(ii) The statement for C is simply based on (14) applied to Z and C together with the fact that the columns of Z are S-orthonormal Ritz vectors of (H, S) in $\text{span}\{Z\}$. In order to verify the statement for C', the relation

$$\text{span}\{U, Z, P\} = \text{span}\{U\} \oplus \text{span}\{Z, QP\}$$

with the projector $Q = VV^*S$ onto $\text{span}\{V\}$ indicates that the columns of Z' are S-orthonormal Ritz vectors associated with the $\tilde{k}$ smallest Ritz values of (H, S) in $\text{span}\{Z, QP\}$. Moreover, by using $\Theta = C^*\Lambda C = Z^*HZ$, it holds that

$$\begin{aligned} QP &= (VV^*S)KR = VV^*SK\,(HZ - SZ\Theta) = VV^*SK\,(HVC - SVC\Theta) \\ &= VV^*SK\,(SV\Lambda C - SVC\Theta) = VV^*SKSV\,(\Lambda C - C\Theta) = V\tilde{K}\tilde{R}, \end{aligned}$$

so that $\text{span}\{Z, QP\} = V \cdot \text{span}\{C, \tilde{K}\tilde{R}\}$. Applying (14) to $\text{span}\{Z, QP\}$ and $\text{span}\{C, \tilde{K}\tilde{R}\}$ completes the verification. ∎

By Lemma 2, we can analyze the convergence behavior of the BPSD-id in terms of the BPSD iteration with successive iterates C and C'.

3 | SHARP SINGLE-STEP ESTIMATES

In this section, we first analyze the convergence behavior of the PSD-id and the BPSD-id. Section 3.1 presents an alternative convergence analysis of the PSD-id (Algorithm 1) in comparison to Reference 4. The estimate (8) of the PSD with weaker assumptions is applied to the restricted formulation of the PSD-id introduced in Lemma 1. This results in a sharp single-step estimate of the PSD-id. In Section 3.2, the convergence of the BPSD-id (Algorithm 2) is investigated by using the estimate (9) together with Lemma 2. In Section 3.3, we discuss an extension of the main results under the notion of so-called larger shifts.

In preparation for the main analysis in this section, we characterize the preconditioner K for the PSD-id and the BPSD-id with respect to its restricted form $\tilde{K}$ arising from Lemmas 1 and 2.

Definition 1 (Definition 2.1 in Reference 4). A preconditioner K is called *effectively positive definite*, if $\tilde{K} = V^*SKSV \in \mathbb{C}^{(n-i+1)\times(n-i+1)}$ is positive definite, where V is defined in (10). $\tilde{K}$ is called an *effective form* of K.

A typical example of effectively positive definite preconditioners is the shift-invert preconditioner $K = (H - \sigma S)^{-1}$ where the shift σ is smaller than λ_i, and not an eigenvalue of (H, S). In this case, the corresponding effective form $\tilde{K}$ is actually a diagonal matrix since by (10),

$$\tilde{K} = V^*SKSV = V^*S(H - \sigma S)^{-1}SV = V^*SV(\Lambda - \sigma\tilde{I})^{-1} = (\Lambda - \sigma\tilde{I})^{-1}, \tag{15}$$

where $\tilde{I} = I_{n-i+1}$. $\tilde{K}$ is positive definite since $(\lambda_j - \sigma)^{-1} > 0$ for each $j \geq i$ due to $\sigma < \lambda_i$.

3.1 | Sharp single-step estimate of the PSD-id

We first provide a simple proof on the monotonicity of the approximate eigenvalues, which has been proven in a cumbersome way; see Proposition 2.2 in Reference 4.

Lemma 3. *Denote by z and z' two successive iterates of the PSD-id (Algorithm 1) where $z \in span\{V\} \setminus \{0\}$. Let the preconditioner K be effectively positive definite. If z is not an eigenvector, then $\rho(z') < \rho(z)$.*

Proof. We use the coefficient vectors c and c' defined in Lemma 1. Since c' is a minimizer of $\tilde{\rho}(\cdot)$ in $\text{span}\{c, \tilde{K}\tilde{r}\}$, we get $\tilde{\rho}(c') \leq \tilde{\rho}(c)$. Therein the equality does not hold, since otherwise c would also be a minimizer of $\tilde{\rho}(\cdot)$ and thus a Ritz vector in $\text{span}\{c, \tilde{K}\tilde{r}\}$. Then the residual $\tilde{r} = \Lambda c - \tilde{\rho}(c)c$ would be orthogonal to $\text{span}\{c, \tilde{K}\tilde{r}\}$ so that $\tilde{r}^*\tilde{K}\tilde{r} = 0$. Subsequently, the positive definiteness of the restricted form $\tilde{K}$ of K leads to $\tilde{r} = 0$ and

$$0 = SV\tilde{r} = SV\,(\Lambda c - \tilde{\rho}(c)c) \overset{(10)}{=} HVc - \tilde{\rho}(c)SVc \overset{(13)}{=} Hz - \rho(z)Sz,$$

that is, z would be an eigenvector. Thus $\tilde{\rho}(c') < \tilde{\rho}(c)$ holds and implies $\rho(z') < \rho(z)$ by (13). ∎

The following lemma provides a quantitative measure on the quality of an effectively positive definite preconditioner.

Lemma 4. *Consider an effectively positive definite preconditioner K, its restricted form $\tilde{K} = V^*SKSV$ and the diagonal matrix $\Lambda_\nu = \Lambda - \nu\tilde{I}$, where Λ is from (10), $\tilde{I} = I_{n-i+1}$, and ν is a parameter such that $\nu < \lambda_i$. Denote by α and β the smallest and largest eigenvalues of $\tilde{K}\Lambda_\nu$. Then $\beta \geq \alpha > 0$, and*

$$\left\|\tilde{I} - \omega\tilde{K}\Lambda_\nu\right\|_{\Lambda_\nu} \leq \varepsilon < 1, \tag{16}$$

where $\omega = 2/(\beta + \alpha)$ and $\varepsilon = (\beta - \alpha)/(\beta + \alpha)$.

Proof. The matrices $\tilde{K}$ and Λ_ν are evidently Hermitian positive definite so that their square root matrices are available. By using $\Lambda_\nu^{1/2}$, the matrix $\tilde{K}\Lambda_\nu$ is similar to $\widehat{K} = \Lambda_\nu^{1/2}\tilde{K}\Lambda_\nu^{1/2}$ which is Hermitian positive definite. This shows the positiveness of all eigenvalues of $\tilde{K}\Lambda_\nu$ and $\widehat{K}$. Moreover, the norm $\left\|\tilde{I} - \omega\tilde{K}\Lambda_\nu\right\|_{\Lambda_\nu} = \left\|\tilde{I} - \omega\widehat{K}\right\|_2$ is actually the maximum of $|1 - \omega\lambda|$ among all eigenvalues λ of $\widehat{K}$. Then the quality condition (16) is verified by

$$0 < \alpha \leq \lambda \leq \beta \quad \Rightarrow \quad \frac{2\alpha}{\beta + \alpha} \leq \omega\lambda \leq \frac{2\beta}{\beta + \alpha} \quad \Rightarrow \quad \frac{\alpha - \beta}{\beta + \alpha} \leq 1 - \omega\lambda \leq \frac{\beta - \alpha}{\beta + \alpha}.$$

∎

The diagonal matrix Λ_ν in Lemma 4 is motivated by the special case $K = (H - \sigma S)^{-1}$ with a shift $\sigma < \lambda_i$. Therein the restricted form $\tilde{K}$ coincides with $(\Lambda - \sigma\tilde{I})^{-1}$ as shown in (15). Then K can be viewed as an optimal preconditioner by setting $\nu = \sigma$ for which $\tilde{K}\Lambda_\nu = \tilde{I}$ holds so that $\alpha = \beta = 1 = \omega$, and $\varepsilon = 0$. Consequently, a natural choice of ν is $\nu = \sigma$ if we construct K by approximating $(H - \sigma S)^{-1}$. However, for evaluating an arbitrary effectively positive definite preconditioner K, the parameter ν is not necessarily related to a shift from a certain algorithm.

The quantity ω is determined by α and β and serves to scale the restricted form $\tilde{K}$ so that (16) is compatible with the quality condition (7) from Theorem 1 where T corresponds to $\omega\tilde{K}$.

The following lemma interprets the coefficient vectors from Lemma 1 as iterates of a PSD iteration for a shifted matrix.

Lemma 5. *With the diagonal matrix Λ_ν from Lemma 4 and the corresponding Rayleigh quotient*

$$\tilde{\rho}_\nu : \mathbb{C}^{n-i+1} \setminus \{0\} \to \mathbb{R}, \quad \tilde{\rho}_\nu(w) = \frac{w^*\Lambda_\nu w}{w^* w}, \tag{17}$$

the coefficient vector c' of z' for the PSD-id (Algorithm 1) is a minimizer of $\tilde{\rho}_\nu(\cdot)$ in span$\{c, \tilde{K}\tilde{r}_\nu\}$, where $\tilde{K} = V^*SKSV$ and $\tilde{r}_\nu = \Lambda_\nu c - \tilde{\rho}_\nu(c)c$.

Proof. For an arbitrary $\tilde{c} \in \mathbb{C}^{n-i+1} \setminus \{0\}$, it holds that

$$\tilde{\rho}_\nu(\tilde{c}) = \frac{\tilde{c}^*\Lambda_\nu\tilde{c}}{\tilde{c}^*\tilde{c}} = \frac{\tilde{c}^*\Lambda\tilde{c} - \nu\tilde{c}^*\tilde{c}}{\tilde{c}^*\tilde{c}} = \tilde{\rho}(\tilde{c}) - \nu.$$

Thus minimizing $\tilde{\rho}_\nu(\cdot)$ is equivalent to minimizing $\tilde{\rho}(\cdot)$. Moreover, the relation

$$\tilde{r}_\nu = \Lambda_\nu c - \tilde{\rho}_\nu(c)c = \left(\Lambda - \nu\tilde{I}\right)c - (\tilde{\rho}(c) - \nu)\,c = \Lambda c - \tilde{\rho}(c)c = \tilde{r},$$

implies span$\{c, \tilde{K}\tilde{r}_\nu\}$ = span$\{c, \tilde{K}\tilde{r}\}$ so that the statement for c' from Lemma 1 is directly reformulated in terms of $\tilde{\rho}_\nu(\cdot)$ and $\tilde{r}_\nu$. ∎

By Lemmas 4 and 5, the following theorem shows that by a reverse transformation, the PSD estimate (8) in Theorem 1 leads to a sharp single-step estimate of the PSD-id based on relations of Rayleigh quotients in (13).

Theorem 2. *Denote by z and z' two successive iterates of the PSD-id (Algorithm 1) where $z \in$ span$\{V\} \setminus \{0\}$. Let the preconditioner K be effectively positive definite with the quality parameter ε defined in (16) for any $\nu < \lambda_i$.*

If $\rho(z) \in (\lambda_j, \lambda_{j+1})$ for certain $j \geq i$, then

$$\frac{\rho(z') - \lambda_j}{\lambda_{j+1} - \rho(z')} \leq \left(\frac{\kappa + \varepsilon(2-\kappa)}{(2-\kappa) + \varepsilon\kappa}\right)^2 \frac{\rho(z) - \lambda_j}{\lambda_{j+1} - \rho(z)} \tag{18}$$

with

$$\kappa = \left(\frac{\lambda_j - \nu}{\lambda_{j+1} - \nu}\right)\left(\frac{\lambda_n - \lambda_{j+1}}{\lambda_n - \lambda_j}\right).$$

The equality in (18) is attainable in the limit case $\rho(z) \to \lambda_j$ in an invariant subspace associated with the eigenvalues λ_j, λ_{j+1} and λ_n.

Proof. We use coefficient vectors c and c' introduced in Lemma 1. According to Lemma 5, c' is a minimizer of $\tilde{\rho}_\nu(\cdot)$ in span$\{c, \tilde{K}\tilde{r}_\nu\}$ concerning the matrix Λ_ν.

By applying Theorem 1 to

$$A \to \Lambda_\nu, \quad x \to c, \quad T \to \omega\tilde{K}, \quad \eta \to \tilde{\rho}_\nu(c), \quad \eta' \to \tilde{\rho}_\nu(c'),$$

and substituting the eigenvalues, the estimate (8) is specified as

$$\frac{\tilde{\rho}_\nu(c') - (\lambda_j - \nu)}{(\lambda_{j+1} - \nu) - \tilde{\rho}_\nu(c')} \leq \left(\frac{\kappa + \varepsilon(2-\kappa)}{(2-\kappa) + \varepsilon\kappa}\right)^2 \frac{\tilde{\rho}_\nu(c) - (\lambda_j - \nu)}{(\lambda_{j+1} - \nu) - \tilde{\rho}_\nu(c)}$$

with

$$\kappa = \frac{(\lambda_j - \nu)\left((\lambda_n - \nu) - (\lambda_{j+1} - \nu)\right)}{(\lambda_{j+1} - \nu)\left((\lambda_n - \nu) - (\lambda_j - \nu)\right)} = \left(\frac{\lambda_j - \nu}{\lambda_{j+1} - \nu}\right)\left(\frac{\lambda_n - \lambda_{j+1}}{\lambda_n - \lambda_j}\right).$$

Therein the terms $\tilde{\rho}_\nu(c)$ and $\tilde{\rho}_\nu(c')$ coincide with $\rho(z)-\nu$ and $\rho(z')-\nu$ due to the relation

$$\tilde{\rho}_\nu(\tilde{c}) = \tilde{\rho}(\tilde{c}) - \nu \overset{(13)}{=} \rho(\tilde{z}) - \nu.$$

Thus (18) is shown. Furthermore, the sharpness statement in Theorem 1 is specified for the limit case $\tilde{\rho}_\nu(c) \to \lambda_j - \nu$ and the matrix Λ_ν. The corresponding invariant subspace is associated with the eigenvalues $\lambda_j - \nu$, $\lambda_{j+1} - \nu$ and $\lambda_n - \nu$. Denoting this subspace by $\widetilde{C}$, then $\widetilde{\mathcal{Z}} = V\widetilde{C}$ is an invariant subspace of (H, S) associated with the eigenvalues λ_j, λ_{j+1} and λ_n due to (10), and the limit case $\tilde{\rho}_\nu(c) \to \lambda_j - \nu$ is converted into $\rho(z) \to \lambda_j$. ∎

Remark 1. The assumption $\rho(z) \in (\lambda_j, \lambda_{j+1})$ in Theorem 2 does not cover the case that $\rho(z)$ is equal to λ_j or λ_{j+1}. Applying Lemma 3 to for example, $\rho(z) = \lambda_j$ provides the following supplement: if z is an eigenvector, then the iteration is simply terminated; otherwise Lemma 3 ensures that $\rho(z)$ is smaller in the next step and can match the assumption $\rho(z) \in (\lambda_j, \lambda_{j+1})$ for a smaller index j so that Theorem 2 is applicable. In summary, $\rho(z)$ either converges to an eigenvalue λ_j with $j > i$ or reaches the interval $(\lambda_i, \lambda_{i+1})$ in the final phase. In the latter (and usual) case, two possible phenomena can be interpreted by the ratio $(\lambda_i - \nu)/(\lambda_{i+1} - \nu)$ from the convergence bound: the convergence rate is deteriorated for $\lambda_i \approx \lambda_{i+1}$; for well-separated λ_i and λ_{i+1}, a fast convergence can be obtained by (proper approximations of) the shift-invert preconditioner $K = (H - \sigma S)^{-1}$ with $\sigma \approx \lambda_i$ since $\kappa \to 0$ for $\nu = \sigma \to \lambda_i$. In Section 5 in Reference 4, it is shown that an efficient shift σ can be chosen from the interval $(\lambda_{i-1}, \lambda_i)$, for example, by initially setting σ slightly larger than the computed λ_{i-1} and then enlarging it with a weighted mean of λ_{i-1} and the current approximation of λ_i. The sharpness statement in Theorem 2 can be illustrated similarly to Figure 4.5 in Reference 10. One can notice that the numerical maxima of the single-step convergence rate of the PSD-id form a monotonically decreasing function $g(\theta)$ against $\theta = \rho(z) \in (\lambda_j, \lambda_{j+1})$ for various j, and the limit $\lim_{\theta \to \lambda_j} g(\theta)$ coincides with the convergence factor in (18). Moreover, $g(\theta)$ for arbitrary $\theta \in (\lambda_j, \lambda_{j+1})$ can be calculated based on certain lengthy implicit representations; compare Page 3204 in Reference 8. Determining an explicit form of $g(\theta)$ is still challenging.

Remark 2. In comparison to Theorem 3.2 in Reference 4, the current approximate eigenvalue $\rho(z)$ in Theorem 2 is located in an arbitrary eigenvalue interval so that the statement is much more flexible. Moreover, the bound in (18) has a simpler form where only one technical term is used, namely the quality parameter ε of preconditioning. With a dynamic shift σ approximating λ_i from below, the parameter κ in Theorem 2 with $\nu = \sigma$ can be close to zero in the final phase and indicates a superlinear convergence. The limit case $\kappa \to 0$ corresponds to an optimal shift-invert preconditioner which allows a one-step convergence in the sense that computing the preconditioned residual $p = Kr$ by solving the linear system $(H - \sigma S)p = r$ for $\sigma \approx \lambda_i$, that is, for almost singular $H - \sigma S$, already gives a good approximate eigenvector associated with λ_i. The convergence factor in (18) is close to ε^2 for $\kappa \to 0$ and thus mainly depends on the solution quality of the linear system. However, this approach is usually not optimal with respect to the total computational time. In addition, Theorem 4.1 in Reference 4 can be improved by Theorem 2 with the convergence factor $\kappa^2/(2-\kappa)^2$ arising from the special case $\varepsilon = 0$, that is, $K = (H - \sigma S)^{-1}$.

3.2 | Sharp single-step estimate of the BPSD-id

Let us now analyze the evolution of Ritz values of the BPSD-id within two successive subspace iterates. We first interpret the coefficient matrices of the BPSD-id from Lemma 2 as iterates of a BPSD iteration for the shifted matrix Λ_ν introduced in Lemma 4.

Lemma 6. *Denote by Z and Z' two successive iterates of the BPSD-id (Algorithm 2) where Z has full rank and all its $\tilde{k}$ columns belong to span$\{V\}$. With the diagonal matrix Λ_ν from Lemma 4, the columns of the coefficient matrix $C = V^*SZ$ are orthonormal Ritz vectors of Λ_ν in span$\{C\}$. Moreover, by using the corresponding Ritz value matrix $\Theta_\nu = C^*\Lambda_\nu C$, the columns of the coefficient matrix $C' = V^*SZ'$ are orthonormal Ritz vectors associated with the $\tilde{k}$ smallest Ritz values of Λ_ν in span$\{C, \tilde{K}\tilde{R}_\nu\}$ for $\tilde{K} = V^*SKSV$ and $\tilde{R}_\nu = \Lambda_\nu C - C\Theta_\nu$.*

Proof. The statements are verified by using Lemma 2 and the transformations

$$\Theta_\nu = C^*\Lambda_\nu C = C^*\left(\Lambda - \nu\tilde{I}\right)C = \Theta - \nu I_{\tilde{k}},$$

and the fact that

$$\tilde{R}_\nu = \left(\Lambda - \nu\tilde{I}\right)C - C\left(\Theta - \nu I_{\tilde{k}}\right) = \Lambda C - C\Theta = \tilde{R},$$

implies that

$$\operatorname{span}\{C, \tilde{K}\tilde{R}_\nu\} = \operatorname{span}\{C, \tilde{K}\tilde{R}\}.$$

■

The following lemma shows a strict reduction of Ritz values concerning the coefficient subspaces $\operatorname{span}\{C\}$ and $\operatorname{span}\{C'\}$.

Lemma 7. *Let the preconditioner K of the BPSD-id be effectively positive definite. Following Lemma 6, denote by $\varphi_1 \le \cdots \le \varphi_{\tilde{k}}$ and $\varphi'_1 \le \cdots \le \varphi'_{\tilde{k}}$ the Ritz values of Λ_ν in $\operatorname{span}\{C\}$ and $\operatorname{span}\{C'\}$, respectively. If $\operatorname{span}\{C\}$ contains no eigenvectors of Λ_ν, then $\varphi'_t < \varphi_t$ holds for each $t \in \{1, \ldots, \tilde{k}\}$.*

Proof. Let us apply Lemma 4 to K, and define the auxiliary matrix

$$\tilde{C} = C - \omega\tilde{K}\tilde{R}_\nu.$$

Then $\tilde{C}$ has full rank since otherwise there would exist a $g \in \mathbb{C}^{\tilde{k}} \setminus \{0\}$ satisfying $0 = \tilde{C}g = Cg - \omega\tilde{K}(\Lambda_\nu C - C\Theta_\nu)g$, that is,

$$\Lambda_\nu^{-1}C\Theta_\nu g = (\tilde{I} - \omega\tilde{K}\Lambda_\nu)(\Lambda_\nu^{-1}C\Theta_\nu g - Cg).$$

Moreover, the condition $\nu < \lambda_i$ from Lemma 4 ensures that $\Lambda_\nu^{-1}C\Theta_\nu$ has full rank so that $\Lambda_\nu^{-1}C\Theta_\nu g \neq 0$. Applying (16) implies

$$\|\Lambda_\nu^{-1}C\Theta_\nu g\|_{\Lambda_\nu} \le \|\tilde{I} - \omega\tilde{K}\Lambda_\nu\|_{\Lambda_\nu}\|\Lambda_\nu^{-1}C\Theta_\nu g - Cg\|_{\Lambda_\nu} < \|\Lambda_\nu^{-1}C\Theta_\nu g - Cg\|_{\Lambda_\nu}, \tag{19}$$

where $\|\Lambda_\nu^{-1}C\Theta_\nu g - Cg\|_{\Lambda_\nu} = 0$ is excluded due to the first inequality and $\Lambda_\nu^{-1}C\Theta_\nu g \neq 0$. However, the orthogonality

$$(Cg)^*\Lambda_\nu(\Lambda_\nu^{-1}C\Theta_\nu g - Cg) = g^*C^*C\Theta_\nu g - g^*C^*\Lambda_\nu Cg = 0,$$

leads to

$$\|\Lambda_\nu^{-1}C\Theta_\nu g\|_{\Lambda_\nu}^2 = \|Cg\|_{\Lambda_\nu}^2 + \|\Lambda_\nu^{-1}C\Theta_\nu g - Cg\|_{\Lambda_\nu}^2 \ge \|\Lambda_\nu^{-1}C\Theta_\nu g - Cg\|_{\Lambda_\nu}^2,$$

which contradicts (19). Thus $\tilde{C}$ has full rank.

Consequently, there are $\tilde{k}$ Ritz values in $\operatorname{span}\{\tilde{C}\}$. We denote them by $\tilde{\varphi}_1 \le \cdots \le \tilde{\varphi}_{\tilde{k}}$. Then $\varphi'_t \le \tilde{\varphi}_t$ holds for each $t \in \{1, \ldots, \tilde{k}\}$ due to $\operatorname{span}\{\tilde{C}\} \subseteq \operatorname{span}\{C, \tilde{K}\tilde{R}_\nu\}$ and the Courant-Fischer principles. For proving $\varphi'_t < \varphi_t$, it remains to be shown

$$\tilde{\varphi}_t < \varphi_t. \tag{20}$$

For this purpose, we use a submatrix E_t of $I_{\tilde{k}}$ such that the columns of CE_t are Ritz vectors associated with the Ritz values $\varphi_1 \le \cdots \le \varphi_t$. Then φ_t is the largest Ritz value in $\operatorname{span}\{CE_t\}$. Correspondingly, we consider the largest Ritz value $\widehat{\varphi}_t$ in $\operatorname{span}\{\tilde{C}E_t\}$. The relation $\operatorname{span}\{\tilde{C}E_t\} \subseteq \operatorname{span}\{\tilde{C}\}$ ensures $\tilde{\varphi}_t \le \widehat{\varphi}_t$. In addition, we use a

Ritz vector $\widehat{c} = \tilde{C}E_t\widehat{g}$ associated with $\widehat{\varphi}_t$ and the auxiliary vector

$$c = C\Theta_\nu E_t\widehat{g} = CE_t\Theta_{\nu,t}\widehat{g} \quad \text{with} \quad \Theta_{\nu,t} = \operatorname{diag}(\varphi_1, \ldots, \varphi_t).$$

Then by using the definitions of $\tilde{C}$ and $\tilde{R}_\nu$, we have

$$\Lambda_\nu^{-1}c - \widehat{c} = \Lambda_\nu^{-1}C\Theta_\nu E_t\widehat{g} - \left(C - \omega\tilde{K}(\Lambda_\nu C - C\Theta_\nu)\right)E_t\widehat{g} = (\tilde{I} - \omega\tilde{K}\Lambda_\nu)(\Lambda_\nu^{-1}c - CE_t\widehat{g}),$$

so that

$$\|\Lambda_\nu^{-1}c - \widehat{c}\|_{\Lambda_\nu} \overset{(16)}{\leq} \varepsilon\|\Lambda_\nu^{-1}c - CE_t\widehat{g}\|_{\Lambda_\nu}.$$

Therein $\|\Lambda_\nu^{-1}c - CE_t\widehat{g}\|_{\Lambda_\nu}$ further fulfills

$$\|\Lambda_\nu^{-1}c - CE_t\widehat{g}\|_{\Lambda_\nu} \leq \|\Lambda_\nu^{-1}c - \varphi^{-1}c\|_{\Lambda_\nu},$$

for $\varphi = \tilde{\rho}_\nu(c)$ with (17) since the orthogonality

$$(CE_t\tilde{g})^*\Lambda_\nu(\Lambda_\nu^{-1}c - CE_t\widehat{g}) = \tilde{g}^*E_t^*C^*C\Theta_\nu E_t\widehat{g} - \tilde{g}^*E_t^*C^*\Lambda_\nu CE_t\widehat{g} = 0,$$

for $\tilde{g} = \widehat{g} - \varphi^{-1}\Theta_{\nu,t}\widehat{g}$ leads to

$$\begin{aligned}\|\Lambda_\nu^{-1}c - CE_t\widehat{g}\|_{\Lambda_\nu} &\leq \left(\|\Lambda_\nu^{-1}c - CE_t\widehat{g}\|_{\Lambda_\nu}^2 + \|CE_t\tilde{g}\|_{\Lambda_\nu}^2\right)^{1/2} \\ &= \|\Lambda_\nu^{-1}c - CE_t(\varphi^{-1}\Theta_{\nu,t}\widehat{g})\|_{\Lambda_\nu} = \|\Lambda_\nu^{-1}c - \varphi^{-1}c\|_{\Lambda_\nu}.\end{aligned}$$

If $\operatorname{span}\{C\}$ contains no eigenvectors of Λ_ν, then the difference $\Lambda_\nu^{-1}c - \varphi^{-1}c$ is nonzero since otherwise c would be an eigenvector of Λ_ν. Consequently, it holds that

$$\|\Lambda_\nu^{-1}c - \widehat{c}\|_{\Lambda_\nu}^2 < \|\Lambda_\nu^{-1}c - \varphi^{-1}c\|_{\Lambda_\nu}^2.$$

By the definition of the norm $\|\cdot\|_{\Lambda_\nu}$, we have

$$\widehat{c}^*\Lambda_\nu\widehat{c} - c^*\widehat{c} - \widehat{c}^*c < \varphi^{-2}c^*\Lambda_\nu c - 2\varphi^{-1}c^*c = -\varphi^{-1}c^*c.$$

Multiplying both sides by φ^{-1} yields

$$\varphi^{-1}\widehat{c}^*\Lambda_\nu\widehat{c} < \varphi^{-1}(c^*\widehat{c} + \widehat{c}^*c) - \varphi^{-2}c^*c = \widehat{c}^*\widehat{c} - \|\widehat{c} - \varphi^{-1}c\|_2^2 \leq \widehat{c}^*\widehat{c},$$

so that $\tilde{\rho}_\nu(\widehat{c}) = (\widehat{c}^*\Lambda_\nu\widehat{c})/(\widehat{c}^*\widehat{c}) < \varphi$. Thus the inequality (20) is verified by

$$\tilde{\varphi}_t \leq \widehat{\varphi}_t = \tilde{\rho}_\nu(\widehat{c}) < \varphi = \tilde{\rho}_\nu(c) = \tilde{\rho}_\nu(CE_t\Theta_{\nu,t}\widehat{g}) \leq \varphi_t.$$

■

Lemma 7 can easily be adapted to successive iterates of the BPSD-id. Therein the relation (14) applied to C, Z and C', Z' (introduced in Lemma 6) shows that $\varphi_t + \nu$ and $\varphi'_t + \nu$ are Ritz values of (H, S) in $\operatorname{span}\{Z\}$ and $\operatorname{span}\{Z'\}$, respectively. Therefore a strict reduction of Ritz values occurs if $\operatorname{span}\{Z\}$ contains no eigenvectors of (H, S). This actually generalizes Lemma 3 to the BPSD-id where the proof is however not a direct generalization of that of Lemma 3.

The following theorem includes sharp single-step estimates on the convergence of the BPSD-id.

Theorem 3. *Denote by Z and Z' two successive iterates of the BPSD-id (Algorithm 2) where Z has full rank and all its $\tilde{k}$ columns belong to $\operatorname{span}\{V\}$. Let the preconditioner K be effectively positive definite with the quality parameter ε defined in (16). Consider the Ritz values $\theta_1 \leq \cdots \leq \theta_{\tilde{k}}$ and $\theta'_1 \leq \cdots \leq \theta'_{\tilde{k}}$ in $\operatorname{span}\{Z\}$ and $\operatorname{span}\{Z'\}$, respectively.*

If $\theta_t \in (\lambda_j, \lambda_{j+1})$ for $t \in \{1, \dots, \tilde{k}\}$ and certain $j \geq i-1+t$, then

$$\frac{\theta_t' - \lambda_j}{\lambda_{j+1} - \theta_t'} \leq \left(\frac{\kappa + \varepsilon(2-\kappa)}{(2-\kappa) + \varepsilon\kappa}\right)^2 \frac{\theta_t - \lambda_j}{\lambda_{j+1} - \theta_t} \tag{21}$$

with

$$\kappa = \left(\frac{\lambda_j - \nu}{\lambda_{j+1} - \nu}\right)\left(\frac{\lambda_n - \lambda_{j+1}}{\lambda_n - \lambda_j}\right).$$

The equality in (21) is attainable in the limit case $\theta_t \to \lambda_j$ in an invariant subspace associated with the eigenvalues λ_j, λ_{j+1} and λ_n.

Proof. We use the coefficient matrices C and C' of Z and Z' introduced in Lemma 2. Following the relation (14), Lemmas 6 and 7, we apply Theorem 1 to

$$A \to \Lambda_\nu, \quad X \to C, \quad T \to \omega\tilde{K}, \quad \eta_t \to \varphi_t = \theta_t - \nu, \quad \eta_t' \to \varphi_t' = \theta_t' - \nu.$$

This results in

$$\frac{(\theta_t' - \nu) - (\lambda_j - \nu)}{(\lambda_{j+1} - \nu) - (\theta_t' - \nu)} \leq \left(\frac{\kappa + \varepsilon(2-\kappa)}{(2-\kappa) + \varepsilon\kappa}\right)^2 \frac{(\theta_t - \nu) - (\lambda_j - \nu)}{(\lambda_{j+1} - \nu) - (\theta_t - \nu)}$$

with

$$\kappa = \frac{(\lambda_j - \nu)\left((\lambda_n - \nu) - (\lambda_{j+1} - \nu)\right)}{(\lambda_{j+1} - \nu)\left((\lambda_n - \nu) - (\lambda_j - \nu)\right)} = \left(\frac{\lambda_j - \nu}{\lambda_{j+1} - \nu}\right)\left(\frac{\lambda_n - \lambda_{j+1}}{\lambda_n - \lambda_j}\right),$$

and further yields the estimate (21) including the sharpness statement. ∎

Similarly to Remark 1, we can additionally discuss the case $\theta_t = \lambda_j$ based on Lemma 7. This yields

$$\theta_t' - \nu = \varphi_t' < \varphi_t = \theta_t - \nu \quad \Rightarrow \quad \theta_t' < \theta_t.$$

so that the limit of the Ritz value θ_t is either λ_{i-1+t} or some λ_j with $j > i-1+t$.

We note that the convergence factor in (21) is only meaningful for well-separated eigenvalues and cannot predict the so-called cluster robustness of block iterations. Section 4 serves to fill this theoretical gap.

3.3 | Larger shifts

The estimates in Theorems 2 and 3 are concerned with effectively positive definite preconditioners such as shift-invert preconditioners of the form $K = (H - \sigma S)^{-1}$ as well as their approximations. Setting $\sigma < \lambda_i$ easily ensures the effectively positive definiteness of K; see (15). However, the practical implementation of the PSD-id in Section 5 in Reference 4 uses the shift $\sigma = \rho(z)$ after $\rho(z)$ is sufficiently close to λ_i, that is, σ is larger than λ_i. Nevertheless, the convergence analysis therein is still based on some estimates which actually treat the case that σ is slightly smaller than λ_i. Thus the resulting bounds can only be applied in an asymptotic way. This section presents a direct analysis together with an extension to the BPSD-id.

We begin with the condition $\lambda_i < \sigma < (\lambda_i + \lambda_{i+1})/2$ in the PSD-id and the exact shift-invert preconditioner $K = (H - \sigma S)^{-1}$.

Theorem 4. *Let z be the current iterate of the PSD-id (Algorithm 1), and z' the next iterate. Assume that the columns $u_1, \dots, u_{i-1}$ of U are S-orthonormal eigenvectors associated with the $i-1$ smallest eigenvalues*

$\lambda_1 \le \cdots \le \lambda_{i-1}$, define $V = [u_i, \ldots, u_n]$ such that $[U, V] = [u_1, \ldots, u_n]$ is an S-orthonormal eigenbasis associated with $\lambda_1 \le \cdots \le \lambda_n$, and consider the case $K = (H - \sigma S)^{-1}$ for a shift $\sigma \in (\lambda_i, (\lambda_i + \lambda_{i+1})/2)$.

If $\rho(z) \ne \sigma$ and $\rho(z) \in (\lambda_i, \lambda_{i+1})$, then

$$\frac{\rho(z') - \lambda_i}{\lambda_{i+1} - \rho(z')} \le \left(\frac{\kappa}{2 - \kappa}\right)^2 \frac{\rho(z) - \lambda_i}{\lambda_{i+1} - \rho(z)} \tag{22}$$

with

$$\kappa = \left(\frac{\lambda_i - \sigma}{\lambda_{i+1} - \sigma}\right)\left(\frac{\lambda_n - \lambda_{i+1}}{\lambda_n - \lambda_i}\right).$$

Proof. The condition $\rho(z) \ne \sigma$ excludes a stagnation caused by

$$p = (H - \rho(z)S)^{-1}(Hz - \rho(z)Sz) = z \quad \Rightarrow \quad \operatorname{span}\{U, z, p\} = \operatorname{span}\{U, z\}.$$

For proving (22), we use a restricted formulation of the considered iteration with respect to $z = Vc$ and $z' = Vc'$, namely,

$$c' = \mathrm{RR}\left(\operatorname{span}\{c, (\Lambda - \sigma\tilde{I})^{-1}c\}\right), \tag{23}$$

where the Rayleigh-Ritz procedure $\mathrm{RR}(\cdot)$ extracts an orthonormal Ritz vector of Λ associated with the smallest Ritz value; compare Lemma 1. Evidently, (23) is an acceleration of

$$\widehat{c} = (\Lambda - \sigma\tilde{I})^{-1}c + \beta c,$$

for arbitrary $\beta \in \mathbb{R}$. By using the auxiliary matrix $A = -\Lambda$ together with its eigenvalues $\alpha_1 \ge \cdots \ge \alpha_{n-i+1}$ and the corresponding Rayleigh quotient $\alpha(\cdot)$, we get

$$\alpha_j = -\lambda_{i-1+j}, \quad \alpha(c) = -\tilde{\rho}(c) = -\rho(z) > -\lambda_{i+1} = \alpha_2, \quad \alpha(\widehat{c}) = -\tilde{\rho}(\widehat{c}),$$

$$\text{and} \quad \widehat{c} = f(A)c \quad \text{with} \quad f(\eta) = (-\eta - \sigma)^{-1} + \beta.$$

Subsequently, we choose $\beta = -\frac{1}{2}\left((\lambda_{i+1} - \sigma)^{-1} + (\lambda_n - \sigma)^{-1}\right)$ so that

$$|f(\alpha_1)| > |f(\alpha_2)| \ge |f(\alpha_j)| \quad \forall\, j \in \{2, \ldots, n - i + 1\},$$

holds (by elementary comparison). Then $\widehat{c} = f(A)c$ can be analyzed as the power method for a matrix function; see Section 1.1 in Reference 16. In particular, the estimate (1.9) in Reference 16 is applicable due to $\alpha(c) > \alpha_2$, and implies

$$\frac{\alpha_1 - \alpha(\widehat{c})}{\alpha(\widehat{c}) - \alpha_2} \le \left(\frac{|f(\alpha_2)|}{|f(\alpha_1)|}\right)^2 \frac{\alpha_1 - \alpha(c)}{\alpha(c) - \alpha_2},$$

which is equivalent to

$$\frac{\tilde{\rho}(\widehat{c}) - \lambda_i}{\lambda_{i+1} - \tilde{\rho}(\widehat{c})} \le \left(\frac{(\lambda_{i+1} - \sigma)^{-1} + \beta}{(\lambda_i - \sigma)^{-1} + \beta}\right)^2 \frac{\tilde{\rho}(c) - \lambda_i}{\lambda_{i+1} - \tilde{\rho}(c)}.$$

This leads to (22) by inserting β and using $\tilde{\rho}(\widehat{c}) \ge \tilde{\rho}(c') = \rho(z')$, $\tilde{\rho}(c) = \rho(z)$. ∎

Theorem 4 relaxes the condition $\sigma < \lambda_i$. A reasonable interval for selecting the shift is $(\lambda_{i-1}, (\lambda_i + \lambda_{i+1})/2)$ where the lower bound can be determined by the previous outer step and the upper bound can be detected by the Rayleigh quotient

and the residual; compare Section 5.2 in Reference 17. The excluded case $\rho(z) = \sigma$ in Theorem 4 can be treated by solving the restricted linear system

$$\tilde{Q}^*(H - \sigma S)\tilde{Q}p = r, \quad p \in \operatorname{span}\{U, z\}^{\perp_S} \tag{24}$$

with the S-orthogonal projector $\tilde{Q}$ onto $\operatorname{span}\{U, z\}^{\perp_S}$. Then the subspace $\operatorname{span}\{U, z, p\}$ can be shown to contain the vector $(H - \sigma S)^{-1}Sz$ as in Section 4 in Reference 13 so that the restricted formulation (23) is applicable. Therefore the estimate (22) with $j = i$ additionally holds for $\rho(z) = \sigma$. Moreover, the special β used for deriving (22) is determined by minimizing $|f(\alpha_2)|/|f(\alpha_1)|$ among $\beta \in \mathbb{R}$. Thus setting $\beta = 0$ yields a less accurate estimate than (22), namely,

$$\frac{\rho(z') - \lambda_i}{\lambda_{i+1} - \rho(z')} \leq \left(\frac{\rho(z) - \lambda_i}{\lambda_{i+1} - \rho(z)}\right)^3. \tag{25}$$

In this sense, (22) indicates a "supercubic" convergence.

For the PSD-id with inexact shift-invert preconditioners which are not effectively positive definite, an alternative of the quality parameter ε from (16) is required so that (22) or (25) can be generalized as an estimate like (18). In particular, for generalizing (25), an appropriate parameter can be constructed with respect to the restricted linear system (24). Therein the restriction of $H - \sigma S$ to $\operatorname{span}\{U, z\}^{\perp_S}$ is positive definite; compare Lemma 3.1 in Reference 13 or the following simpler verification: for an arbitrary $w \in \operatorname{span}\{U, z\}^{\perp_S}$ with $\|w\|_S = 1$, the matrix $W = [U, z, w]$ fulfills $W^*SW = I_{i+1}$ so that the Ritz values of (H, S) in $\operatorname{span}\{W\}$ are eigenvalues of W^*HW. The trace of W^*HW reads

$$\tau = \rho(u_1) + \cdots + \rho(u_{i-1}) + \rho(z) + \rho(w) = \lambda_1 + \cdots + \lambda_{i-1} + \rho(z) + \rho(w).$$

Moreover, the Courant-Fischer principles ensure that τ is at least $\lambda_1 + \cdots + \lambda_{i+1}$. Thus

$$\begin{aligned} &\rho(z) + \rho(w) \geq \lambda_i + \lambda_{i+1} \quad \Rightarrow \quad \rho(w) \geq \lambda_i + \lambda_{i+1} - \sigma \\ &\Rightarrow \quad w^*(H - \sigma S)w = \rho(w) - \sigma \geq \lambda_i + \lambda_{i+1} - 2\sigma > 0. \end{aligned}$$

Correspondingly, the restricted formulation of PSD-id within $\operatorname{span}\{V\}$ includes the restriction of $\Lambda - \sigma\tilde{I}$ to $\operatorname{span}\{c\}^{\perp}$ which is also positive definite. Then the accuracy of the approximate solution p of (24) or its coefficient vector V^*Sp can be interpreted by a vector norm induced by the restriction of $H - \sigma S$ or $\Lambda - \sigma\tilde{I}$; compare Theorem 3.2 in Reference 13. A sufficiently accurate p is characterized by a quality parameter $\varepsilon \in [0, 1)$ so that the estimate (25) can be generalized as

$$\frac{\rho(z') - \lambda_i}{\lambda_{i+1} - \rho(z')} \leq \left(\frac{\rho(z) - \lambda_i + \varepsilon\,(\lambda_{i+1} - \rho(z))}{\lambda_{i+1} - \rho(z) + \varepsilon\,(\rho(z) - \lambda_i)}\right)^2 \frac{\rho(z) - \lambda_i}{\lambda_{i+1} - \rho(z)}. \tag{26}$$

Now we turn to the BPSD-id and analyze its convergence for the exact shift-inverse preconditioning.

Theorem 5. *Let Z be the current iterate of the BPSD-id (Algorithm 2), and Z' the next iterate. Assume that the columns $u_1, \ldots, u_{i-1}$ of U are S-orthonormal eigenvectors associated with the $i-1$ smallest eigenvalues $\lambda_1 \leq \cdots \leq \lambda_{i-1}$, define $V = [u_i, \ldots, u_n]$ such that $[U, V] = [u_1, \ldots, u_n]$ is an S-orthonormal eigenbasis associated with $\lambda_1 \leq \cdots \leq \lambda_n$, and consider the case $K = (H - \sigma S)^{-1}$ for certain $\sigma \in (\lambda_i, (\lambda_i + \lambda_{i+1})/2)$. Denote by $\theta_1 \leq \cdots \leq \theta_{\tilde{k}}$ and $\theta'_1 \leq \cdots \leq \theta'_{\tilde{k}}$ the Ritz values in $\operatorname{span}\{Z\}$ and $\operatorname{span}\{Z'\}$, respectively.*

If $\theta_1 \neq \sigma$ and $\theta_1 \in (\lambda_i, \lambda_{i+1})$, then

$$\frac{\theta'_1 - \lambda_i}{\lambda_{i+1} - \theta'_1} \leq \left(\frac{\kappa}{2 - \kappa}\right)^2 \frac{\theta_1 - \lambda_i}{\lambda_{i+1} - \theta_1} \tag{27}$$

with

$$\kappa = \left(\frac{\lambda_i - \sigma}{\lambda_{i+1} - \sigma}\right)\left(\frac{\lambda_n - \lambda_{i+1}}{\lambda_n - \lambda_i}\right).$$

If $\theta_t \in (\lambda_j, \lambda_{j+1})$ *for* $t \in \{2, \ldots, \tilde{k}\}$ *and* $j = i - 1 + t$, *then*

$$\frac{\theta_t' - \lambda_j}{\lambda_{j+1} - \theta_t'} \le \left(\frac{\kappa}{2 - \kappa}\right)^2 \frac{\theta_t - \lambda_j}{\lambda_{j+1} - \theta_t} \tag{28}$$

with

$$\kappa = \left(\frac{\lambda_j - \sigma}{\lambda_{j+1} - \sigma}\right)\left(\frac{\lambda_n - \lambda_{j+1}}{\lambda_n - \lambda_j}\right).$$

Proof. By applying the PSD-id to a Ritz vector z_1 in $\mathrm{span}\{Z\}$ associated with θ_1, the next iterate z_1' is contained in $\mathrm{span}\{Z'\}$. Then the estimate (22) for $z = z_1$ and $z' = z_1'$ leads to (27) due to $\rho(z_1') \ge \theta_1'$ and $\rho(z_1) = \theta_1$.

The derivation of (28) for $t = \tilde{k}$ is analogous to the proof of Theorem 4 by using the generalization (2.22) of (1.9) in Reference 16 concerning the block form of an abstract power method. Moreover, for $t \in \{2, \ldots, \tilde{k} - 1\}$, we can adapt this derivation to the BPSD-id applied to the subset $\mathcal{Z}_t = \mathrm{span}\{z_1, \ldots, z_t\}$ of $\mathrm{span}\{Z\}$ spanned by Ritz vectors associated with $\theta_1, \ldots, \theta_t$. This yields an intermediate estimate for the largest (tth smallest) Ritz value $\tilde{\theta}_t$ in the next iterate $\mathcal{Z}_t'$ of $\mathcal{Z}_t$. Subsequently, (28) is proved by considering that $\mathcal{Z}_t'$ is a subset of $\mathrm{span}\{Z'\}$ so that $\tilde{\theta}_t \ge \theta_t'$. ∎

For the BPSD-id with inexact shift-invert preconditioners, we particularly consider that the current θ_1 fulfills $\theta_1 < (\lambda_i + \lambda_{i+1})/2$ and is used as the next shift σ. Then the estimates (25) and (26) can be adapted to θ_1. It is also remarkable that by using the positive definiteness of the restriction of $H - \sigma S$, Theorem 3 can be modified for discussing the deviation between the Ritz values in the subspace iterates and the Ritz values in the larger subspace $\mathrm{span}\{U, Z(:,1)\}^{\perp_S}$. The latter ones get closer to the eigenvalues $\lambda_{i+1}, \ldots, \lambda_n$ for the decreasing θ_1 toward λ_i. Then the corresponding estimates turn into those in Theorem 3 with the index update $i \leftarrow i + 1$.

4 | MULTI-STEP ESTIMATES ON THE CLUSTER ROBUSTNESS

A well-known feature of block eigensolvers is the cluster robustness, that is, fast convergence toward clustered eigenvalues can be guaranteed by sufficiently large block sizes. A classical estimate of the subspace iteration has been presented by Parlett in Section 14.4 in Reference 15. Therein the block inverse iteration (also called inverse subspace iteration) $\mathrm{span}\{X^{(\ell+1)}\} = \mathrm{span}\{A^{-1}X^{(\ell)}\}$ for a real symmetric positive definite matrix A with eigenvalues $\alpha_1 \le \cdots \le \alpha_n$ is investigated. The convergence is measured by the angle between an eigenvector and the current subspace. The resulting bound contains the term $(\alpha_j/\alpha_{m+1})^k$ for k steps with the block size m. A Ritz value estimate with the same ratio α_j/α_{m+1} can be derived based on the analysis of the block form of an abstract power method, see Section 2.2 in Reference 16.

These two estimates can be modified for the subspace iteration implemented with implicit deflation and the exact shift-invert preconditioner $K = (H - \sigma S)^{-1}$ for $\sigma < \lambda_i$. Therein the current Ritz basis matrix $Z \in \mathbb{C}^{n \times \tilde{k}}$ is updated by S-orthonormal Ritz vectors associated with the ith to the $(i - 1 + \tilde{k})$th smallest Ritz values in the subspace $\mathrm{span}\{U, (H - \sigma S)^{-1}SZ\}$. By using a transformation with the representation $Z = VC$ and the S-orthogonal projector $Q = VV^*S$, we have

$$Q(H - \sigma S)^{-1}SZ = VV^*S(H - \sigma S)^{-1}SVC \overset{(15)}{=} VV^*S\left(V(\Lambda - \sigma\tilde{I})^{-1}\right)C = V(\Lambda - \sigma\tilde{I})^{-1}C.$$

Consequently, we have

$$\mathrm{span}\{U, (H - \sigma S)^{-1}SZ\} = \mathrm{span}\{U\} \oplus V \cdot \mathrm{span}\{(\Lambda - \sigma\tilde{I})^{-1}C\}.$$

Thus the restricted iteration within $\mathrm{span}\{V\}$ reads

$$\mathrm{span}\{C'\} = \mathrm{span}\{(\Lambda - \sigma\tilde{I})^{-1}C\}, \tag{29}$$

that is, the block inverse iteration for the diagonal matrix $(\Lambda - \sigma\tilde{I})^{-1}$. Then applying the analysis from References 15 and 16 gives the ratio $(\lambda_{i-1+t} - \sigma)/(\lambda_{i+\tilde{k}} - \sigma)$.

Furthermore, the restricted form of the BPSD-id (Algorithm 2) has the trial subspace $\mathrm{span}\{C, (\Lambda - \sigma\tilde{I})^{-1}C\}$ which is a superset of $\mathrm{span}\{C'\}$ from (29). Thus Ritz value estimates for (29) lead to indirect estimates for BPSD-id.

A more accurate estimate of this kind for BPSD-id can be shown by using another auxiliary iteration whose restricted form reads

$$\mathrm{span}\{C'\} = \mathrm{span}\{(\Lambda - \sigma\tilde{I})^{-1}C + \beta C\}. \tag{30}$$

The accuracy can benefit from the choice of $\beta \in \mathbb{R}$.

The following theorem predicts the cluster robustness of the BPSD-id.

Theorem 6. *Consider the BPSD-id (Algorithm 2) with $K = (H - \sigma S)^{-1}$ and $\sigma < \lambda_i$. Denote by $\theta_1^{(\ell)} \le \cdots \le \theta_{\tilde{k}}^{(\ell)}$ the Ritz values in the ℓth subspace $\mathrm{span}\{Z^{(\ell)}\}$. Then*

$$\frac{\theta_t^{(\ell)} - \lambda_{i-1+t}}{\lambda_n - \theta_t^{(\ell)}} \le \left(\frac{\kappa}{2-\kappa}\right)^{2\ell} \frac{\lambda_{i-1+t} - \sigma}{\lambda_n - \sigma}\tau \tag{31}$$

with

$$\kappa = \left(\frac{\lambda_{i-1+t} - \sigma}{\lambda_{i+\tilde{k}} - \sigma}\right)\left(\frac{\lambda_n - \lambda_{i+\tilde{k}}}{\lambda_n - \lambda_{i-1+t}}\right),$$

and a constant $\tau > 0$ depending on the initial subspace $\mathrm{span}\{Z^{(0)}\}$.

Proof. Following Lemma 6, we observe the accompanying iteration

$$\mathrm{span}\{C^{(\ell+1)}\} = \mathrm{RR}_{\Lambda_\sigma, \tilde{k}, \uparrow}\left(\mathrm{span}\{C^{(\ell)}, \Lambda_\sigma^{-1}C^{(\ell)}\}\right) \tag{32}$$

with the coefficient matrix $C^{(\ell)} = V^*SZ^{(\ell)}$ where the Rayleigh-Ritz procedure $\mathrm{RR}_{\Lambda_\sigma, \tilde{k}, \uparrow}(\cdot)$ extracts orthonormal Ritz vectors of $\Lambda_\sigma = \mathrm{diag}(\lambda_i - \sigma, \ldots, \lambda_n - \sigma)$ associated with the $\tilde{k}$ smallest Ritz values, that is, $\theta_1^{(\ell+1)} - \sigma, \ldots, \theta_{\tilde{k}}^{(\ell+1)} - \sigma$ are Ritz values of Λ_σ in $\mathrm{span}\{C^{(\ell+1)}\}$, and coincide with the $\tilde{k}$ smallest Ritz values of Λ_σ in $\mathrm{span}\{C^{(\ell)}, \Lambda_\sigma^{-1}C^{(\ell)}\}$.

In addition, $\sigma < \lambda_i$ ensures that Λ_σ is positive definite. By using its square root matrix $\Lambda_\sigma^{1/2}$, we define $E^{(\ell)} = \Lambda_\sigma^{1/2}C^{(\ell)}$ and $A = \Lambda_\sigma^{-1}$ so that

$$\mathrm{span}\{E^{(\ell)}, AE^{(\ell)}\} = \mathrm{span}\{\Lambda_\sigma^{1/2}C^{(\ell)}, \Lambda_\sigma^{-1/2}C^{(\ell)}\} = \Lambda_\sigma^{1/2}\mathrm{span}\{C^{(\ell)}, \Lambda_\sigma^{-1}C^{(\ell)}\}.$$

Then the iteration (32) is equivalent to

$$\mathrm{span}\{E^{(\ell+1)}\} = \mathrm{RR}_{A, \tilde{k}, \downarrow}\left(\mathrm{span}\{E^{(\ell)}, AE^{(\ell)}\}\right), \tag{33}$$

where the Rayleigh-Ritz procedure $\mathrm{RR}_{A, \tilde{k}, \downarrow}(\cdot)$ extracts A-orthonormal Ritz vectors of A associated with the $\tilde{k}$ largest Ritz values. For explaining this equivalence, we can denote by C a basis matrix of $\mathrm{span}\{C^{(\ell)}, \Lambda_\sigma^{-1}C^{(\ell)}\}$ consisting of orthonormal Ritz vectors of Λ_σ. Then C^*C is an identity matrix, and $C^*\Lambda_\sigma C$ is a diagonal matrix containing Ritz values of Λ_σ. The corresponding $E = \Lambda_\sigma^{1/2}C$ fulfills

$$E^*E = C^*\Lambda_\sigma C, \quad E^*AE = C^*C,$$

so that its columns are A-orthonormal Ritz vectors of A in $\mathrm{span}\{E\} = \Lambda_\sigma^{1/2}\mathrm{span}\{C\}$. Moreover, the concerned Ritz values of A are reciprocals of those of Λ_σ.

We further denote by $\alpha_1 \geq \cdots \geq \alpha_{n-i+1}$ the eigenvalues of A, and by $\psi_1^{(\ell)} \geq \cdots \geq \psi_{\tilde{k}}^{(\ell)}$ the Ritz values in the ℓth subspace from (33). Then

$$\alpha_t = (\lambda_{i-1+t} - \sigma)^{-1}, \quad \psi_t^{(\ell)} = \left(\theta_t^{(\ell)} - \sigma\right)^{-1},$$

so that the estimate (31) is equivalent to

$$\frac{\alpha_t - \psi_t^{(\ell)}}{\psi_t^{(\ell)} - \alpha_{n-i+1}} \leq \left(\frac{\kappa}{2-\kappa}\right)^{2\ell} \tau \tag{34}$$

with

$$\kappa = \frac{\alpha_{\tilde{k}+1} - \alpha_{n-i+1}}{\alpha_t - \alpha_{n-i+1}}.$$

Finally, we derive (34) via the simplified version

$$\operatorname{span}\{E^{(\ell+1)}\} = \operatorname{span}\{AE^{(\ell)} + \beta E^{(\ell)}\},$$

of (33). Therein $\operatorname{span}\{E^{(\ell)}\}$ can be represented by $f(A)\operatorname{span}\{E^{(0)}\}$ with $f(\eta) = \eta + \beta$. This corresponds to the block form of an abstract power method so that its convergence behavior can be analyzed as in Section 2.2 in Reference 16. By using $\beta = -\frac{1}{2}(\alpha_{t+1} + \alpha_{n-i+1})$, we get

$$|f(\alpha_1)| \geq \cdots \geq |f(\alpha_t)| \geq |f(\alpha_j)| \quad \forall\, j \in \{t+1, \ldots, n-i+1\}.$$

Then the estimate (2.20) in Reference 16 implies (34) where the constant τ is given by the tangent square of the angle between $\operatorname{span}\{E^{(0)}\}$ and the invariant subspace of A associated with the $\tilde{k}$ largest eigenvalues. This angle depends on $\operatorname{span}\{Z^{(0)}\}$ since $E^{(0)} = \Lambda_\sigma^{1/2} V^* S Z^{(0)}$. ∎

Theorem 6 predicts the cluster robustness of the BPSD-id since it indicates that the convergence rate increases with the Ritz value index t and the parameter κ can be bounded away from 1 for a sufficiently large block size $\tilde{k}$.

It is desirable to extend Theorem 6 to effectively positive definite preconditioners where $(\kappa + \varepsilon(2-\kappa))^{2\ell}/((2-\kappa) + \varepsilon\kappa)^{2\ell}$ is a suitable convergence factor with the quality parameter ε defined in (16). Toward this aim, the analysis of the cluster robustness of a preconditioned subspace iteration in Reference 11 can be adapted to the restricted formulation of the BPSD-id and yields an estimate of the form

$$\frac{\theta_t^{(\ell)} - \lambda_{i-1+t}}{\lambda_{i+\tilde{k}} - \theta_t^{(\ell)}} \leq \left(\varepsilon + (1-\varepsilon)\frac{\lambda_{i-1+t} - \nu}{\lambda_{i+\tilde{k}} - \nu}\right)^{2\ell} \frac{\theta_{\tilde{k}}^{(0)} - \lambda_{i-1+t}}{\lambda_{i+\tilde{k}} - \theta_{\tilde{k}}^{(0)}}, \tag{35}$$

for an arbitrary $\nu < \lambda_i$. Therein the quality parameter ε is related to ν and certain auxiliary vectors, but close to that from an error propagation formulation such as (16).

A recent approach in Reference 18 for analyzing the cluster robustness of the BPSD is also compatible with the restricted formulation of the BPSD-id. The resulting estimate reads

$$\frac{\theta_t^{(\ell)} - \lambda_{i-1+t}}{\lambda_{i+\tilde{k}} - \theta_t^{(\ell)}} \leq \left(\frac{\kappa + \varepsilon(2-\kappa)}{(2-\kappa) + \varepsilon\kappa}\right)^{2\ell} \frac{\theta_{\tilde{k}}^{(0)} - \lambda_{i-1+t}}{\lambda_{i+\tilde{k}} - \theta_{\tilde{k}}^{(0)}} \tag{36}$$

with

$$\kappa = \left(\frac{\lambda_{i-1+t} - \nu}{\lambda_{i+\tilde{k}} - \nu}\right)\left(\frac{\lambda_n - \lambda_{i+\tilde{k}}}{\lambda_n - \lambda_{i-1+t}}\right).$$

In the case $\varepsilon = 0$ and $\nu = \sigma$, the convergence factor in (36) coincides with that in (31).

An alternative of the multi-step estimates (35) and (36) follows from the single-step estimate (21) in Theorem 3 for $j = i - 1 + t$, namely,

$$\frac{\theta_t^{(\ell)} - \lambda_{i-1+t}}{\lambda_{i+t} - \theta_t^{(\ell)}} \le \left(\frac{\kappa + \varepsilon(2-\kappa)}{(2-\kappa) + \varepsilon\kappa}\right)^{2\ell} \frac{\theta_t^{(0)} - \lambda_{i-1+t}}{\lambda_{i+t} - \theta_t^{(0)}} \tag{37}$$

with

$$\kappa = \left(\frac{\lambda_{i-1+t} - \nu}{\lambda_{i+t} - \nu}\right)\left(\frac{\lambda_n - \lambda_{i+t}}{\lambda_n - \lambda_{i-1+t}}\right).$$

The estimate (37) cannot predict the cluster robustness, but can provide better bounds for the first steps in comparison to (35) and (36).

5 | NUMERICAL EXPERIMENTS

In this section, we illustrate main convergence estimates of the PSD-id and the BPSD-id presented in the previous sections by several numerical examples. Example 1 with non-clustered targets eigenvalues demonstrates single-step estimates for the BPSD-id in Section 3.2. In Example 2, we implement the BPSD-id for a large-scale eigenvalue problem and compare three effectively positive definite preconditioners. In Example 3, we revisit Example 5.2 in Reference 4 where the eigenvalue problem is derived by the partition-of-unity finite element for a self-consistent pseudopotential density functional calculation. We refine the implementation of the PSD-id therein (by accelerating inner steps) and interpret a cubic convergence for dynamic larger shifts by estimates in Section 3.3. Example 4 with clustered targets eigenvalues demonstrates multi-step estimates for the BPSD-id in Section 4.

Example 1. We consider the Laplacian eigenvalue problem $-\Delta u = \lambda u$ on the rectangle $[0, 1.5] \times [0, 1]$ with two slits $\{0.5\} \times [0.45, 0.55]$ and $\{1\} \times [0.45, 0.55]$; see Figure 1. The boundary condition on the rectangle boundary and the slits is simply $u = 0$. By using the five-point star discretization with the mesh size $h = 45/3600$, we get the eigenvalue problem (1) of order $n = 9383$ and $S = I$. The seven smallest eigenvalues of (H, S) are well separated:

$$\lambda_1 \approx 27.07834, \quad \lambda_2 \approx 38.24327, \quad \lambda_3 \approx 45.24858,$$

$$\lambda_4 \approx 49.32646, \quad \lambda_5 \approx 58.36810, \quad \lambda_6 \approx 78.91626, \quad \lambda_7 \approx 89.70648.$$

We compute the six smallest eigenvalues by the BPSD-id (Algorithm 2) with $\{k, \tilde{k}\} = \{1, 2\}, \{2, 3\}, \{3, 4\}$. Recall that k is the number of wanted eigenvalues in each run and $\tilde{k}$ is the block size. For instance, if $\{k, \tilde{k}\} = \{2, 3\}$, then the BPSD-id with block size 3 computes 2 eigenvalues in each of three successive runs.

We first consider fixed shifts and use incomplete matrix factorizations for generating the preconditioner K, namely,

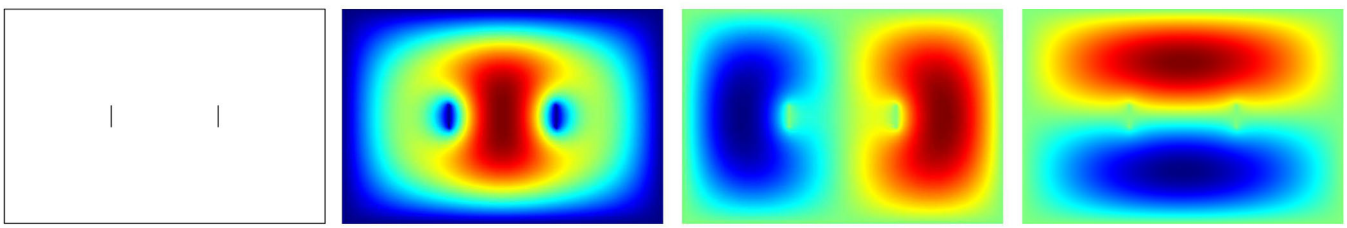

FIGURE 1 Laplacian eigenvalue problem in Example 1. The two slits on the domain clearly influence the shapes of eigenfunctions associated with the three smallest operator eigenvalues.

```
ichol(H-sigma*S,struct('type','ict','droptol',3e-5))
```

with the shift $\sigma = 20$ for the first run (i.e., for $i = 1$) and

```
ilu(H-sigma*S,struct('type','crout','milu','row','droptol',3e-5))
```

with the shift $\sigma = \lambda_{i-1}$ for further runs.

Figure 2 depicts the convergence of the residual norm $\|R(:, 1:k)\|_{S^{-1}}$. We run 1000 random initial subspaces and show the slowest convergence. We observe that the residual norm decreases monotonically for different choices of $\{k, \tilde{k}\}$.

The convergence of the BPSD-id in terms of the Ritz value error $\theta_t - \lambda_t$ for $t \in \{1, \dots, 6\}$ is depicted in Figure 3, where the indices of Ritz values are permutated to match the target eigenvalues. We observe that Theorem 3 provides sharp error bounds in dotted curves. For evaluating these bounds, we determine the quality parameter ε by using Lemma 4 with $\nu = \sigma$ and the fact that the nonzero eigenvalues of $\tilde{K}\Lambda_\sigma = (V^*SKSV)V^*(H - \sigma S)V$ coincide with those of $K(SVV^*)(H - \sigma S)VV^*S$. Then the extremal eigenvalues of $\tilde{K}\Lambda_\sigma$ can be obtained by `eigs` applied to a subroutine for matrix-vector multiplications where $(VV^*S)x$ is computed by $x - U(U^*(Sx))$, and $(SVV^*)y$ by $y - S(U(U^*y))$. Once ε is determined, we calculate the right-hand side of the bound (21) in Theorem 3 with $\nu = \sigma$ and precomputed eigenvalues. Thereafter the (absolute) error bound of the corresponding Ritz value is obtained by solving an elementary equation. The values of ε concerning the construction of the preconditioner K by `ichol` ($\sigma = 20$) or `ilu` ($\sigma = \lambda_{i-1}$) with `droptol` $\in$ {`2e-5`, `3e-5`, `6e-5`} are listed in Table 1. For each σ, the value of ε increases with `droptol`. The case $\varepsilon \approx 1$ occurs for `droptol` > `1e-4`. Then the convergence factor in (21) is also close to 1 and reflects a slow convergence. Lemma 4 thus enables controlling the preconditioning quality in the BPSD-id and extends the approach for the PSD mentioned in (5) or its equivalent form (1.6) in Reference 8. The application also covers preconditioners generated by multigrid techniques and approximately solving linear systems.

We note that the sharpness statement for the estimate (21) cannot easily be observed for random iterates as shown in Figure 3. The attainability of the associated equality within certain low-dimensional invariant subspaces can however be illustrated similarly to Figure 4.5 in Reference 10.

Let us consider refining the preconditioner K by a dynamic shift σ. With the index i of the smallest target eigenvalue in the current run, we estimate the ratio $\eta = (\theta_i - \lambda_i)/(\lambda_{i+1} - \theta_i)$ roughly by $(\theta_{i,\mathrm{old}} - \theta_i)/(\theta_{i+1} - \theta_i)$ as suggested for the PSD-id in Section 5 in Reference 4. If η and the residual norm $\|R(:, 1:k)\|_{S^{-1}}$ are both smaller than the threshold 0.1, we update σ by $\sigma \leftarrow (\sigma + \theta_i)/2$ and refine the preconditioner K by

```
ilu(H-sigma*S,struct('type','crout','milu','row','droptol',max(eta,1e-12))).
```

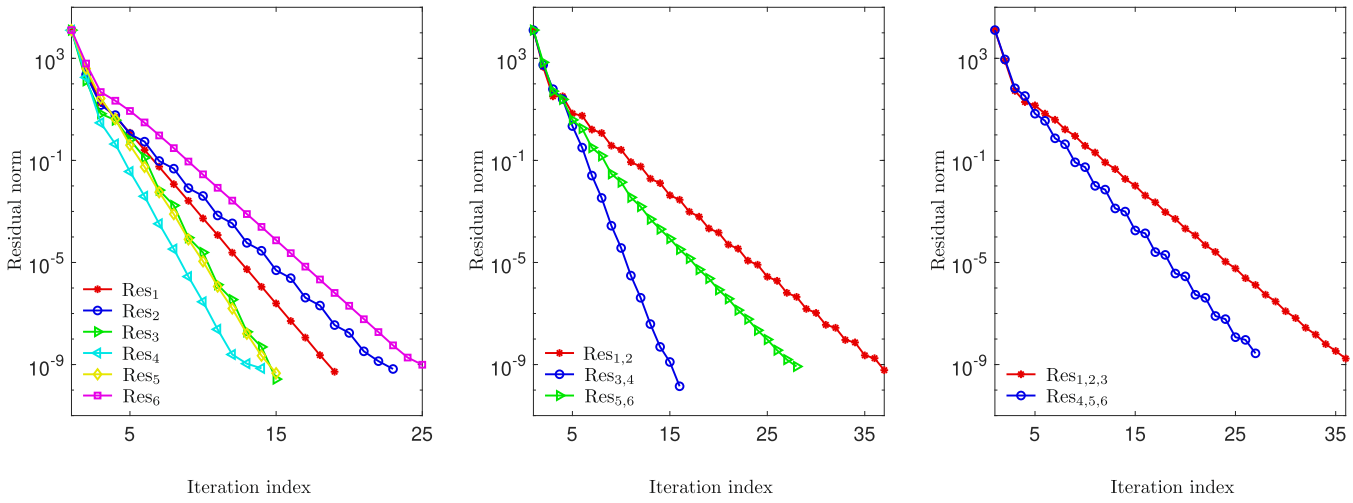

FIGURE 2 Convergence of the BPSD-id (Algorithm 2) in terms of the residual norm $\|R(:, 1:k)\|_{S^{-1}}$ for computing the six smallest eigenvalues in Example 1. The pair $\{k, \tilde{k}\}$ is $\{1, 2\}$ (left), $\{2, 3\}$ (center) and $\{3, 4\}$ (right).

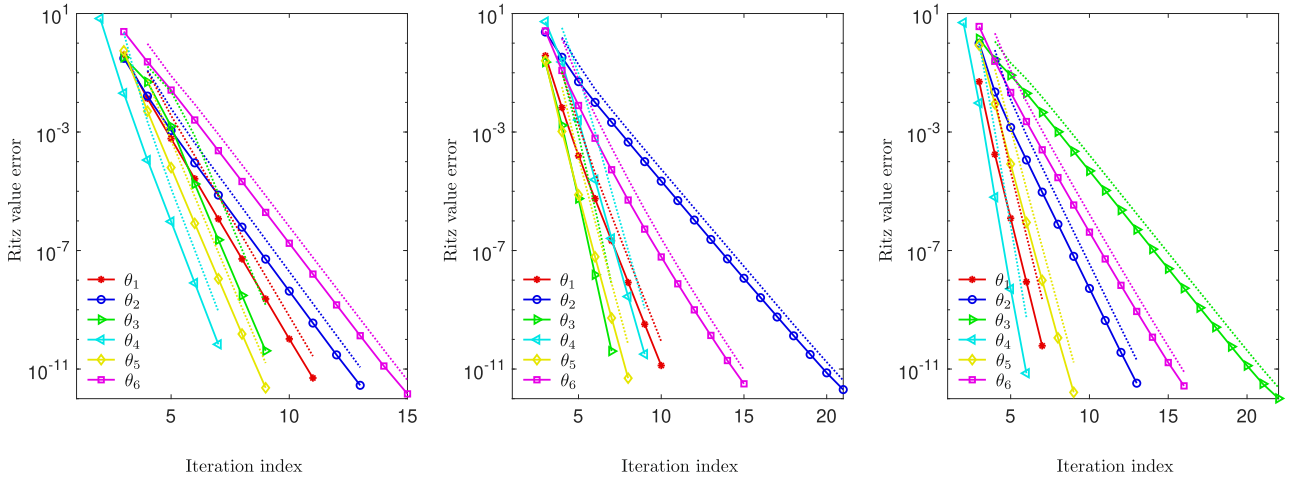

FIGURE 3 Convergence of the BPSD-id (Algorithm 2) in terms of the Ritz value error $\theta_t - \lambda_t$ for computing the six smallest eigenvalues in Example 1 with fixed shifts. The error bounds in dotted curves are determined by Theorem 3.

TABLE 1 Quality parameter ϵ of the preconditioner K in Example 1.

droptol	$\sigma = 20$	$\sigma = \lambda_1$	$\sigma = \lambda_2$	$\sigma = \lambda_3$	$\sigma = \lambda_4$	$\sigma = \lambda_5$
2e-5	0.2391	0.1386	0.1848	0.0830	0.1672	0.0618
3e-5	0.3106	0.2384	0.3293	0.1567	0.1918	0.1224
6e-5	0.4605	0.5432	0.9126	0.3808	0.2890	0.3147

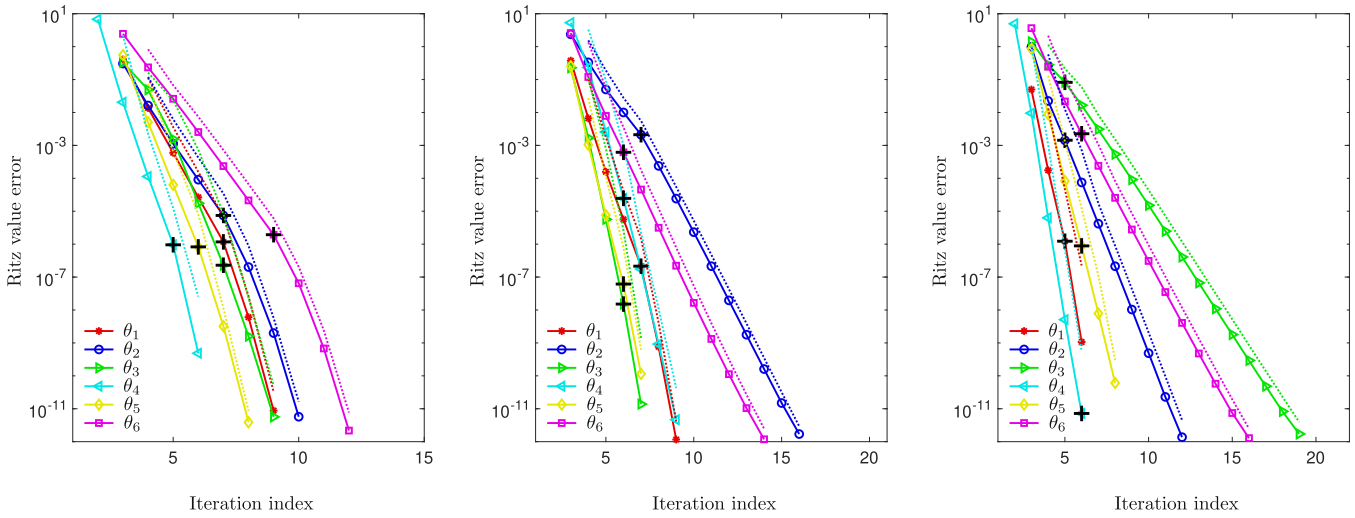

FIGURE 4 Convergence of the BPSD-id (Algorithm 2) in terms of the Ritz value error $\theta_t - \lambda_t$ for computing the six smallest eigenvalues in Example 1 with dynamic shifts.

We mark the first refinement by "+." As shown in Figure 4, this modification leads to an acceleration. The improvement is evident for the $\{k, \tilde{k}\} = \{1, 2\}$ (left subfigure). Furthermore, Theorem 3 produces proper bounds in dotted curves which are built similarly as in Figure 3 but with a dynamic shift σ in various steps. Therein the quality parameter ϵ can be reduced to `1e-9`.

Example 2. For discussing the performance of the BPSD-id in large-scale problems, let us consider a matrix pair arising from an adaptive finite element discretization of the Laplacian eigenvalue problem on

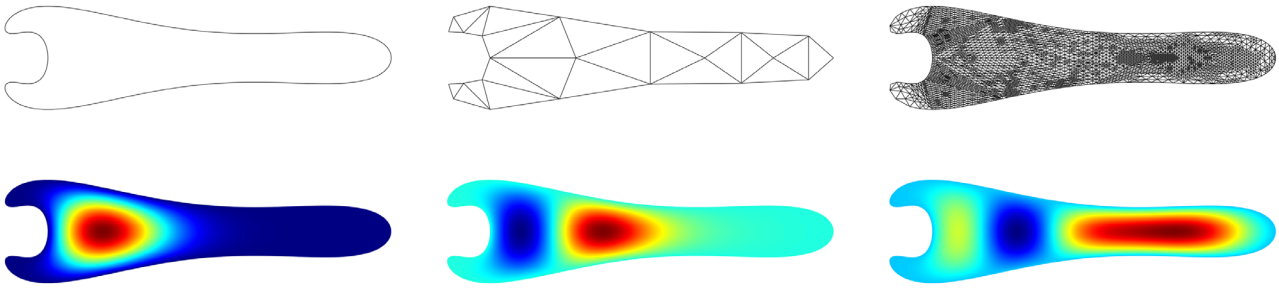

FIGURE 5 Laplacian eigenvalue problem in Example 2. The first row displays the domain, the initial grid and an adaptively refined grid. The second row illustrates approximate eigenfunctions associated with the three smallest operator eigenvalues. The corresponding residuals are used for the grid refinement.

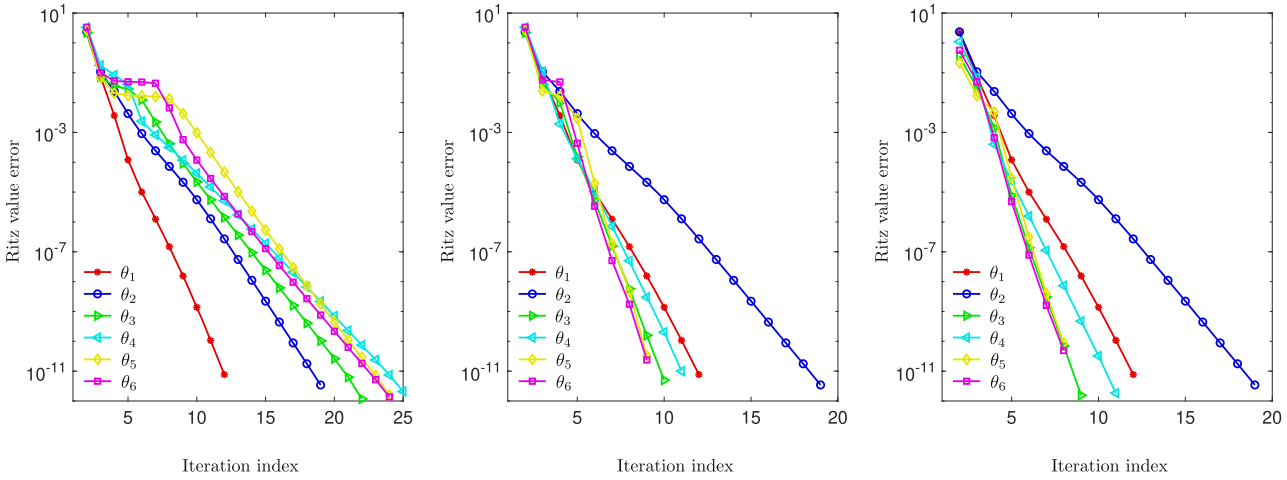

FIGURE 6 Convergence of the BPSD-id (Algorithm 2) in terms of the Ritz value error $\theta_t - \lambda_t$ for computing the six smallest eigenvalues in Example 2. Left: Using `ichol` for generating the preconditioner with a constant shift. Center: Modifying the preconditioner by MINRES with the tolerance 0.1. Right: Modifying the preconditioner by MINRES with the tolerance 0.01.

a wrench-shaped domain with homogeneous Dirichlet boundary conditions; see Figure 5. The boundary is defined by

$$\{20\cos(t) + 10\cos(2t),\ 5\sin(t) + \sin(5t);\ t \in [0, 2\pi)\}\,.$$

Similarly to Appendix in Reference 11, matrix eigenvalue problems are generated successively by an adaptive finite element discretization. The refinement is controlled by residuals of approximate eigenfunctions associated with the three smallest operator eigenvalues.

We consider the matrix pair (H, S) from the 29th grid of the discretization with $n = 1{,}618{,}797$ degrees of freedom. The seven smallest eigenvalues of (H, S) are well separated and located in the interval $(0.1418824, 0.4562653)$. Similarly to Example 1, we compute the six smallest eigenvalues $\lambda_1, \ldots, \lambda_6$ by 3 successive runs of the BPSD-id (Algorithm 2) with $\{k, \tilde{k}\} = \{2, 4\}$.

Figure 6 illustrates the convergence of the BPSD-id in terms of the Ritz value error $\theta_t - \lambda_t$ for $t \in \{1, \ldots, 6\}$. Therein we compare three effectively positive definite preconditioners.

In the left subfigure, the preconditioner K is generated by the incomplete matrix factorization

```
ichol(H-sigma*S,struct('type','ict','droptol',1e-6))
```

with the shift $\sigma = 0.1 < \lambda_1$. We use this K for each run, since generating K by `ilu` with $\sigma = \lambda_{i-1}$ is too costly. Consequently, more outer steps are required in the second and third runs (curves for θ_3, θ_4 and θ_5, θ_6) than in the first run (curves for θ_1, θ_2).

In the central subfigure, we modify K for the second and third runs where approximations of $(H-\sigma S)^{-1}$ for $\sigma=\lambda_{i-1}$ are constructed by MINRES with the above `ichol` factorization and the tolerance 0.1. This clearly reduces the number of required outer steps. However, the computational time for these two runs reads 176 s, longer than 105 s measured in the left subfigure.

The same modification with the tolerance 0.01, as illustrated in the right subfigure, does not lead to a further significant reduction of step numbers, but increases the computational time to 208 s. The values of the corresponding quality parameter are given in Table 2 (columns $\sigma=\lambda_2$ and $\sigma=\lambda_4$). In addition, modifications with dynamic shifts are also less efficient with respect to the total computational time. Thus a simply applicable preconditioner with fixed shifts is occasionally more appropriate.

Example 3. To verify the sharpness of the estimates with larger shifts discussed in Section 3.3, we use the matrix pencil (H,S) of order $n=5336$ derived from partition-of-unity finite element method for quantum-mechanical materials calculation; see Example 5.2 in Reference 4. The matrix H is given by a rank-l modification of an $n\times n$ sparse matrix $\widehat{H}$ with $l=26$. Both H and S are ill-conditioned and their condition numbers are $\mathcal{O}(10^{10})$. Furthermore, H and S share a common near-nullspace span$\{V\}$ of dimension 1000 such that $\|HV\|=\|SV\|=\mathcal{O}(10^{-4})$. This is considered as an extremely ill-conditioned eigenvalue problem.

For the BPSD-id, the explicit form of H is dense so that constructing K by incomplete matrix factorizations is not efficient. Thus MINRES with preconditioner `chol(S)` is used for computing the preconditioned residual Kr with $K\approx(H-\sigma S)^{-1}$. The stopping criterion of MINRES uses the residual norm $\psi=\|r\|_2/(\|Hz\|_2+\|\rho(z)Sz\|_2)$ instead of $\|r\|_{S^{-1}}$. The initial σ is -1, which is smaller than the smallest eigenvalue $\lambda_1\approx-0.8888$. The tolerance of MINRES is 0.1 for $\sigma=-1$. If ψ and an estimated value of $(\rho(z)-\lambda_i)\,/\,(\lambda_{i+1}-\rho(z))$ are sufficiently small, then σ is set equal to $\rho(z)$.

We begin with the PSD-id (Algorithm 1) implemented as a weakened form of the BPSD-id where $r=R(:,1)$ is used for constructing the trial subspace. This corresponds to an acceleration of the PSD-id so that Theorem 4 is still applicable and the estimate (22) indicates a cubic convergence for larger shifts from the interval $(\lambda_i,(\lambda_i+\lambda_{i+1})/2)$.

We note that the inner steps (MINRES) in the above implementation can significantly be accelerated by using $\widehat{H}+3S$ instead of S in `chol(S)`. In other words, $\widehat{H}+3S$ gives a better approximation of $H-\sigma S$ in comparison to S. In addition, a substantial acceleration of the outer steps in the third and fourth runs is enabled by using λ_{i-1} as the initial σ. This improvement can be observed by comparing the reduction of the residual norm ψ in Figure 7 (left) with that in Figure 5.2 in Reference 4.

Furthermore, Figure 7 (center) depicts the reduction of ψ for fixed shifts, that is, $\sigma=-1$ for the first run and $\sigma=\lambda_{i-1}$ for further runs. The tolerance of MINRES is constantly 0.1. In comparison to the version with dynamic shifts, only the second run is considerably slowed down with respect to the outer steps. The total computational time is however only slightly increased. In Figure 7 (right), a hybrid version is implemented where dynamic shifts are used if the iteration index for the first possible switch is larger than 6. Then dynamic shifts are enabled in the second run, and the total computational time is reduced in comparison to the previous two versions.

Since our estimates are formulated for Ritz value errors, we observe in Figure 8 the error $\theta_t-\lambda_t$ for the three versions mentioned above. Therein θ_t denotes the approximate eigenvalue in the tth run. The convergence rate can be made cubic by using dynamic shifts as predicted by Theorem 4. For fixed shifts, the linear convergence predicted by Theorem 2 can be observed for $t\in\{1,3,4\}$. The error for $t=2$ indicates various convergence rates so that Theorem 2 would give overestimation for several iteration steps. This phenomenon however reflects the fact that the implemented PSD-id uses an augmented trial subspace.

TABLE 2 Quality parameter of the MINRES preconditioner in Example 2.

MINRES tol	$\sigma=\lambda_1$	$\sigma=\lambda_2$	$\sigma=\lambda_3$	$\sigma=\lambda_4$	$\sigma=\lambda_5$
0.01	0.2756	0.5157	0.7036	0.4972	0.7269
0.1	0.4947	0.6813	0.8707	0.7470	0.9265

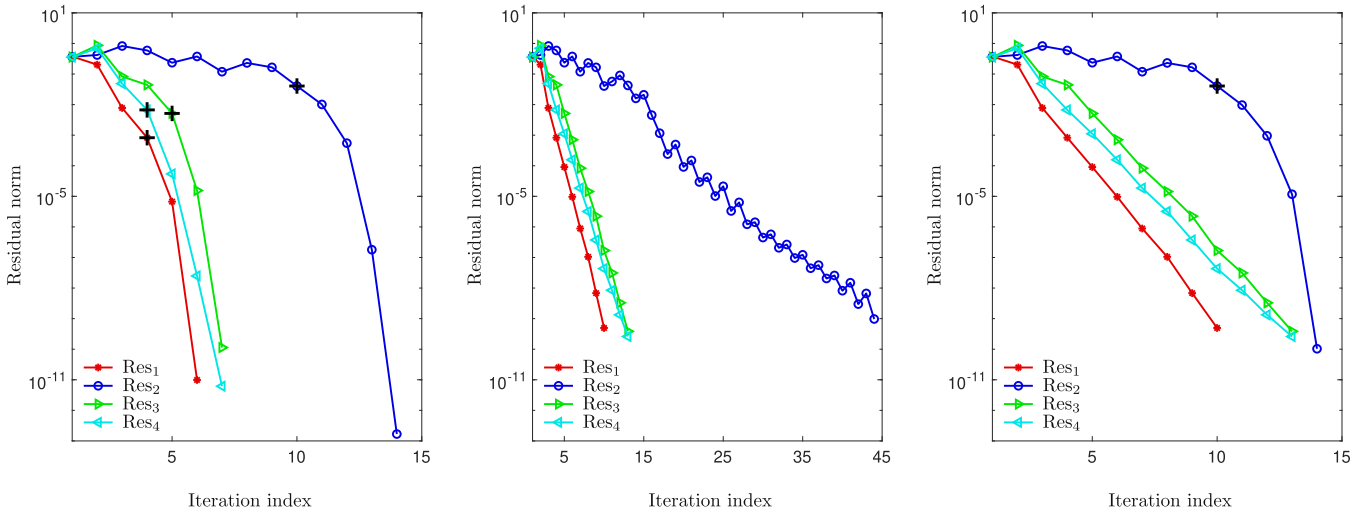

FIGURE 7 Convergence of a weakened form of BPSD-id using only one residual vector and an relative residual norm for computing the four smallest eigenvalues in Example 3. The preconditioner is constructed by MINRES with certain shifts. Left: Enabling dynamic shifts. The first dynamic step is marked by "+." Center: Using fixed shifts only. Right: Enabling dynamic shifts when the convergence in the first steps is slow.

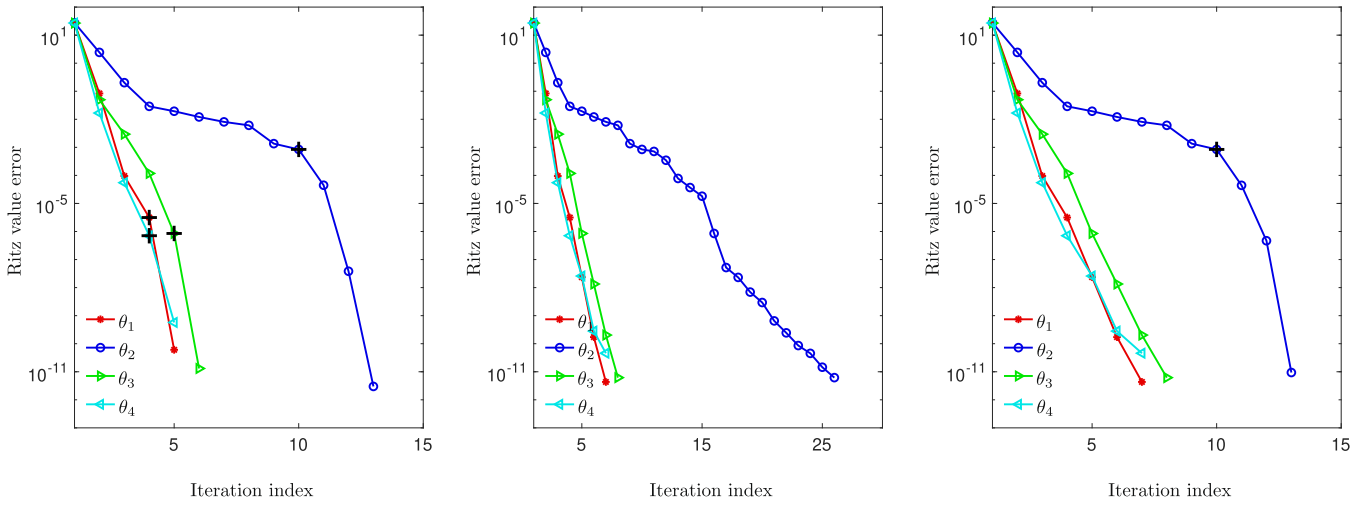

FIGURE 8 Convergence of a weakened form of BPSD-id in terms of the Ritz value error $\theta_t - \lambda_t$ in addition to Figure 7.

When implementing the standard BPSD-id with shifts from $(\lambda_i, (\lambda_i + \lambda_{i+1})/2)$ in Example 3, a cubic convergence can occur for the smallest Ritz value in each run. Further Ritz values only converge linearly. This difference matches the two estimates from Theorem 5.

Example 4. By using a block size which is larger than the cluster size, the BPSD-id can efficiently compute clustered eigenvalues. This fact has been analyzed in Theorem 6 for exact shift-invert preconditioning. The estimate (31) corresponds to a special form of (36) concerning effectively positive definite preconditioners. The estimate (36) can be derived by adapting the analysis from Reference 18 to the restricted formulation of the BPSD-id analogously to Section 3.2. In this example, we compare (36) with its counterparts (35) and (37) which are based on Reference 11 and the single-step estimate (21), respectively.

We modify the eigenvalue problem in Example 1 by setting larger slits $\{0.5\} \times [0.1, 0.9]$ and $\{1\} \times [0.1, 0.9]$ on the rectangle $[0, 1.5] \times [0, 1]$; see Figure 9. The mesh size $h = 1/80$ leads to $n = 9271$. The modified domain can be regarded as three small rectangles connected by narrow gates. For a sufficiently small gate width,

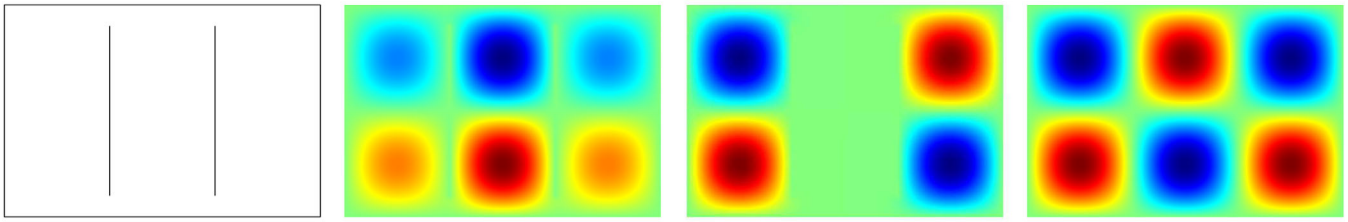

FIGURE 9 Laplacian eigenvalue problem in Example 4. The eigenfunctions associated with the fourth to sixth smallest operator eigenvalues are displayed.

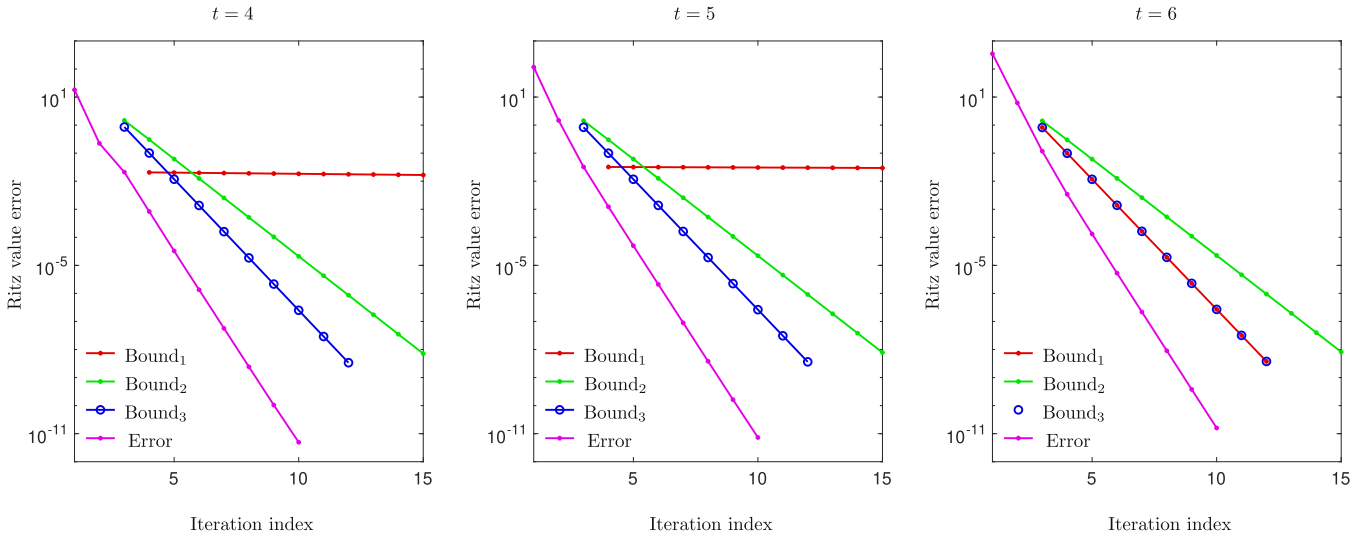

FIGURE 10 Convergence of the BPSD-id (Algorithm 2) in terms of the Ritz value error $\theta_t - \lambda_t$ for $t \in \{4, 5, 6\}$ in Example 4 with clustered eigenvalues. The three bounds are based on the multi-step estimates (37), (35), and (36).

the eigenvalue problem is almost split into three partial problems of the same size. Thus the eigenvalues are roughly copies of those from partial problems. As a result, there are two tight clusters among the seven smallest eigenvalues:

$$\lambda_1, \lambda_2, \lambda_3 \in (49.24886, 49.32647), \quad \lambda_4, \lambda_5, \lambda_6 \in (78.61283, 78.91626), \quad \lambda_7 \approx 127.5209.$$

Let us examine a run of the BPSD-id with $i = 4$ and $\tilde{k} = 3$, which means that the eigenvalues $\lambda_4, \lambda_5, \lambda_6$ are to be computed. The preconditioner K is generated by

```
ilu(H-sigma*S,struct('type','crout','milu','row','droptol',3e-5))
```

for $\sigma = \lambda_3$. We denote by Bound$_1$, Bound$_2$ and Bound$_3$ the bounds for the Ritz value error $\theta_t - \lambda_t$ (with simplified indices) which are determined on the basis of (37), (35), and (36), respectively. We compare these bounds with the error $\theta_t - \lambda_t$ in Figure 10 with subfigures for $t \in \{4, 5, 6\}$ by evaluating their numerical maxima concerning 1000 random initial subspaces.

We note that Bound$_2$ and Bound$_3$ are nearly invariant for t due to the eigenvalue cluster $\{\lambda_4, \lambda_5, \lambda_6\}$. Moreover, Bound$_3$ is evidently more accurate than Bound$_2$. For $t \in \{4, 5\}$, the clustered eigenvalues make Bound$_1$ nearly constant. For $t = 6$, the moderate gap between λ_6 and λ_7 leads to a meaningful Bound$_1$ which actually coincides with Bound$_3$.

6 | CONCLUSION

The limitation of the approach presented in Reference 4 for analyzing the convergence behavior of the PSD-id method is overcome by embedding concise bounds from References 8 and 10 concerning the PSD and BPSD methods. The new estimates are more flexible with weaker assumptions, natural description of preconditioning and extension to block iterations. Therein the preconditioners are assumed to be effectively positive definite and particularly include approximative shift-invert preconditioners where the shift is smaller than the target eigenvalue. In addition, the case of larger shifts is discussed based on the analysis of an abstract power method from Reference 16 and the analysis of an inexact Rayleigh quotient iteration from Reference 13. Furthermore, the cluster robustness of the BPSD-id is analyzed for exact shift-invert preconditioning analogously to an abstract block iteration from Reference 16. A more general analysis for effectively positive definite preconditioners is enabled by recent progress for the BPSD from Reference 18. Topics for further study include "implicit deflation" versions of the LOBPCG and various Davidson methods as well as practical settings of shifts and block sizes.

ACKNOWLEDGMENTS

We are grateful to the referees for their insightful comments that improve the presentation. Ming Zhou acknowledges the support by German Research Foundation (DFG), Project number 463329614. Zhaojun Bai acknowledges the support by U.S. National Science Foundation, Award number DMS-1913364. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no datasets were created or analyzed during the study.

ORCID

Ming Zhou https://orcid.org/0000-0001-7096-1649

REFERENCES

1. Samokish BA. The steepest descent method for an eigenvalue problem with semi-bounded operators. Izv Vyssh Uchebn Zaved Mat. 1958;5:105–14. (in Russian).
2. Bramble JH, Pasciak JE, Knyazev AV. A subspace preconditioning algorithm for eigenvector/eigenvalue computation. Adv Comput Math. 1996;6:159–89.
3. Ovtchinnikov EE. Cluster robustness of preconditioned gradient subspace iteration eigensolvers. Linear Algebra Appl. 2006;415:140–66.
4. Cai Y, Bai Z, Pask JE, Sukumar N. Convergence analysis of a locally accelerated preconditioned steepest descent method for Hermitian-definite generalized eigenvalue problems. J Comp Math. 2018;36:739–60.
5. Knyazev AV, Neymeyr K. Efficient solution of symmetric eigenvalue problems using multigrid preconditioners in the locally optimal block conjugate gradient method. Electron Trans Numer Anal. 2003;15:38–55.
6. Knyazev AV, Skorokhodov AL. On exact estimates of the convergence rate of the steepest ascent method in the symmetric eigenvalue problem. Linear Algebra Appl. 1991;154–156:245–57.
7. Neymeyr K, Ovtchinnikov EE, Zhou M. Convergence analysis of gradient iterations for the symmetric eigenvalue problem. SIAM J Matrix Anal Appl. 2011;32:443–56.
8. Neymeyr K. A geometric convergence theory for the preconditioned steepest descent iteration. SIAM J Numer Anal. 2012;50:3188–207.
9. Stewart GW, Sun J. Matrix perturbation theory. Cambridge, MA: Academic Press; 1990.
10. Neymeyr K, Zhou M. The block preconditioned steepest descent iteration for elliptic operator eigenvalue problems. Electron Trans Numer Anal. 2014;41:93–108.
11. Zhou M, Neymeyr K. Cluster robust estimates for block gradient-type eigensolvers. Math Comput. 2019;88:2737–65.
12. Saad Y. Numerical methods for large eigenvalue problems. Manchester: Manchester University Press; 1992.
13. Notay Y. Convergence analysis of inexact Rayleigh quotient iteration. SIAM J Matrix Anal Appl. 2003;24:627–44.
14. Ovtchinnikov EE. Sharp convergence estimates for the preconditioned steepest descent method for Hermitian eigenvalue problems. SIAM J Numer Anal. 2006;43:2668–89.
15. Parlett BN. The symmetric eigenvalue problem. Hoboken, NJ: Prentice-Hall; 1980 Reprinted as Classics in Applied Mathematics 20, SIAM; 1997.
16. Knyazev AV. Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem. Russ J Numer Anal Math Model. 1987;2:371–96.

17. Notay Y. Combination of Jacobi-Davidson and conjugate gradients for the partial symmetric eigenproblem. Numer Linear Algebra Appl. 2002;9:21–44.
18. Zhou M, Neymeyr K. Convergence rates of individual Ritz values in block preconditioned gradient-type eigensolvers. Technical report. 2022. Available from. https://arxiv.org/abs/2206.00585

How to cite this article: Zhou M, Bai Z, Cai Y, Neymeyr K. Convergence analysis of a block preconditioned steepest descent eigensolver with implicit deflation. Numer Linear Algebra Appl. 2023;e2498. https://doi.org/10.1002/nla.2498