Les Cahiers du GERAD ISSN: 0711–2440

Deflation-based spectral preconditioning for the conjugate gradient method on perturbed systems

Y. Diouane, S. Gürol, O. Mouhtal, D. Orban

G–2026–70

October 2026

La collection *Les Cahiers du GERAD* est constituée des travaux de recherche menés par nos membres. La plupart de ces documents de travail a été soumis à des revues avec comité de révision. Lorsqu'un document est accepté et publié, le pdf original est retiré si c'est nécessaire et un lien vers l'article publié est ajouté.

Citation suggérée : Y. Diouane, S. Gürol, O. Mouhtal, D. Orban (Octobre 2026). Deflation-based spectral preconditioning for the conjugate gradient method on perturbed systems, Rapport technique, Les Cahiers du GERAD G– 2026–70, GERAD, HEC Montréal, Canada.

Avant de citer ce rapport technique, veuillez visiter notre site Web (https://www.gerad.ca/fr/papers/G-2026-70) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.

The series *Les Cahiers du GERAD* consists of working papers carried out by our members. Most of these pre-prints have been submitted to peer-reviewed journals. When accepted and published, if necessary, the original pdf is removed and a link to the published article is added.

Suggested citation: Y. Diouane, S. Gürol, O. Mouhtal, D. Orban (October 2026). Deflation-based spectral preconditioning for the conjugate gradient method on perturbed systems, Technical report, Les Cahiers du GERAD G–2026–70, GERAD, HEC Montréal, Canada.

Before citing this technical report, please visit our website (https://www.gerad.ca/en/papers/G-2026-70) to update your reference data, if it has been published in a scientific journal.

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2026
– Bibliothèque et Archives Canada, 2026

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2026
– Library and Archives Canada, 2026

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Deflation-based spectral preconditioning for the conjugate gradient method on perturbed systems

Youssef Diouane [a, b]

Selime Gürol [c]

Oussama Mouhtal [a, b, c]

Dominique Orban [a, b]

[a] *Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Montréal (Qc), Canada, H3T 1J4*

[b] *GERAD, Montréal (Qc), Canada, H3T 1J4*

[c] *CERFACS / CECI CNRS UMR 5318, 31100 Toulouse, France*, Toulouse, France

youssef.diouane@polymtl.ca
oussama-2.mouhtal@polymtl.ca
dominique.orban@polymtl.ca
gurol@cerfacs.fr

October 2026
Les Cahiers du GERAD
G–2026–70

Copyright © 2026 Diouane, Gürol, Mouhtal, Orban

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs. Les auteurs conservent leur droit d'auteur et leurs droits moraux sur leurs publications et les utilisateurs s'engagent à reconnaître et respecter les exigences légales associées à ces droits. Ainsi, les utilisateurs:

- Peuvent télécharger et imprimer une copie de toute publication du portail public aux fins d'étude ou de recherche privée;
- Ne peuvent pas distribuer le matériel ou l'utiliser pour une activité à but lucratif ou pour un gain commercial;
- Peuvent distribuer gratuitement l'URL identifiant la publication.

Si vous pensez que ce document enfreint le droit d'auteur, contactez-nous en fournissant des détails. Nous supprimerons immédiatement l'accès au travail et enquêterons sur votre demande.

The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*. Copyright and moral rights for the publications are retained by the authors and the users must commit themselves to recognize and abide the legal requirements associated with these rights. Thus, users:

- May download and print one copy of any publication from the public portal for the purpose of private study or research;
- May not further distribute the material or use it for any profit-making activity or commercial gain;
- May freely distribute the URL identifying the publication.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Abstract : We consider sequences of large symmetric positive-definite linear systems in which spectral information from a reference matrix A is reused to accelerate the solution of a nearby system with perturbed matrix $A(t) = A + tE$, where E is symmetric, $\|E\|_F = 1$, and $t \geq 0$ is the perturbation size. We compare deflated conjugate gradient (deflated CG) with preconditioned CG (PCG) equipped with a scaled spectral preconditioner constructed from the same spectral information of A. First, we derive an energy-norm error characterization of deflated CG. We then use perturbation analysis to show that, after one additional PCG iteration, the squared PCG error is bounded, up to terms of order t^2, by an explicit multiple of the squared deflated CG error. This result clarifies the role of the scaling parameter in the preconditioner. Experiments show that the scaled spectral preconditioner recovers much of the convergence improvement provided by deflation while requiring fewer products with $A(t)$.

Keywords : Krylov methods; conjugate gradient; PCG; deflated CG; spectral preconditioner

Acknowledgements: This research was partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants held by Diouane and Orban. Gürol and Mouhtal were partially funded by the French national programme LEFE/INSU.

1 Introduction

We consider sequences of large symmetric positive-definite (SPD) linear systems solved by the conjugate gradient (CG) method [15]. Such sequences arise in many applications, such as variational data assimilation [8], where consecutive systems are often closely related. Information generated while solving one system can then be recycled to accelerate subsequent solves; see, for example, [12, 22, 25, 28].

We consider a reference matrix A together with a nearby matrix $A(t) = A + tE$, where E is symmetric, $\|E\|_F = 1$, and t is sufficiently small for $A(t)$ to remain positive definite. Our goal is to reuse spectral information from A to accelerate CG applied to the perturbed system $A(t)x = b$. This is particularly attractive in large-scale matrix-free settings, where constructing a preconditioner may be prohibitively expensive and only a limited number of matrix–vector products can be afforded.

One important approach for exploiting recycled information is deflated CG. Given a deflation subspace $\mathcal{W} = \mathrm{Range}(W)$, deflated CG splits the solution into a component in $\mathcal{W}$ and a complementary component computed by CG on a projected system [9, 21, 24]. Tang et al. [27] mention that, for any deflation subspace, the effective condition number of the deflated system does not exceed that of the original system. A widely used choice for the deflation subspace is a set of approximate eigenvectors obtained from previous Krylov iterations through Ritz or harmonic Ritz projections [22, 24, 25, 29]. In the nonsymmetric setting, SVD-based strategies have also been proposed for recycling Krylov subspaces, in particular for GMRES [1]. Problem-dependent alternatives include coarse spaces arising from domain decomposition and multilevel methods [27]. These approaches seek to construct a low-dimensional subspace that captures the components responsible for slow convergence at moderate additional cost.

Approaches related to deflation are limited-memory preconditioners (LMPs) [13], augmentation methods [10], balancing strategies [20], and low-rank updates [3]. When spectral information is used in the preconditioner, we refer to such methods as spectral preconditioners [5, 11, 13]. For broader overviews of preconditioning, see [2, 23, 30].

The effect of using approximate spectral information has therefore received considerable attention. The sensitivity of deflated CG to perturbations of the deflation subspace is studied in [17], while Giraud and Gratton [11] analyze how errors in approximate eigenpairs affect the spectrum of the preconditioned matrix. These results characterize the influence of inexact spectral information on the projected or preconditioned spectrum. Although deflation can significantly improve CG convergence, its construction may become costly when the coefficient matrix changes. For a deflation basis $W \in \mathbb{R}^{n\times k}$, one must form and solve a system with the matrix $W^\top A(t)W$, which requires additional products with $A(t)$. This problem generally needs to be solved accurately, since the convergence of deflated CG can be sensitive to errors in its solution [20]. To mitigate this sensitivity, related approaches avoid mapping the selected eigenvalues to zero, e.g., Erlangga and Nabben [7] shift selected eigenvalues to a nonzero part of the spectrum and use the resulting preconditioner within GMRES.

In [6], the authors introduced a scaled spectral preconditioner designed to reduce the energy-norm error after a prescribed number of preconditioned CG (PCG) iterations. The preconditioner is constructed from the eigenvalues and eigenvectors of A and maps the selected eigenvalues to a common value θ. Experiments showed that an appropriate choice of the scaling parameter can reproduce much of the convergence acceleration obtained with deflation while requiring a lower initialization cost. In particular, this behavior was observed in a data assimilation problem when approximate eigenpairs were recycled. However, the theoretical analysis in [6] assumes that the spectral information used to construct the preconditioner is exact and is associated with the matrix of the system being solved. It therefore does not explain why similar behavior persists when spectral information from a nearby matrix is reused.

Our main objective in this work is to establish this connection theoretically when spectral information obtained from A is reused for the perturbed system $A(t)x = b$. We focus on eigenpairs associated with the largest eigenvalues of A, although the analysis extends to any selected subset of eigenpairs.

We compare the energy-norm errors generated by deflated CG and by PCG equipped with the scaled spectral preconditioner. Using first-order perturbation analysis, we show that, after one additional PCG iteration, the squared energy-norm error of PCG is bounded by the squared energy-norm error of deflated CG multiplied by an explicit factor, plus higher-order terms in t. The resulting bound clarifies how the scaling parameter θ should be chosen relative to the nondeflated spectrum and provides a theoretical explanation for the similar convergence behavior observed numerically. We complement the theoretical analysis with numerical experiments on synthetic linear systems.

The paper is organized as follows. In Section 2, we review deflated CG and derive a characterization of its energy-norm error. We formulate the perturbed problem, develop the first-order analysis, establish the comparison between PCG and deflated CG, and discuss practical choices of the scaling parameter in Section 3. We report the numerical experiments in Section 4 and present our conclusions in Section 5.

Notation: The spectral norm of a matrix X is denoted $\|\|_2$, its spectral radius $\rho(X)$, and its range $\mathcal{R}(X)$. For a symmetric matrix $X \in \mathbb{R}^{n\times n}$, we denote its eigenvalues $\lambda_1(X), \ldots, \lambda_n(X)$ and the corresponding eigenvectors $s_1(X), \ldots, s_n(X)$. Throughout the paper, $A \in \mathbb{R}^{n\times n}$ is SPD. When there is no ambiguity, its ith eigenvalue and corresponding eigenvector are denoted λ_i and s_i, respectively. For $k \in \{1, \ldots, n-1\}$, we define $S_k = [s_1, \ldots, s_k] \in \mathbb{R}^{n\times k}$ and $\overline{S}_k = [s_{k+1}, \ldots, s_n] \in \mathbb{R}^{n\times(n-k)}$. Thus, S_k contains the eigenvectors associated with the k largest eigenvalues of A, while the columns of $\overline{S}_k$ span the orthogonal complement of $\mathcal{R}(S_k)$. For $x \in \mathbb{R}^n$, the A-norm, also called the *energy norm*, is defined by $\|x\|_A = \sqrt{x^\top Ax}$. We denote $\mathbb{P}_\ell(0)$ the set of polynomials of degree at most ℓ with value 1 at zero.

In what follows, all results are derived under the assumption of exact arithmetic.

2 Deflated conjugate gradient

Let $W \in \mathbb{R}^{n\times k}$ have full column rank. Deflated CG [9, 24] decomposes the solution of $Ax = b$ into a component in the deflation subspace $\mathcal{R}(W)$ and a complementary component computed by CG on a projected system. Let $V = AW$ and define the projection $P = I_n - V\big(W^\top V\big)^{-1}W^\top$. Deflated CG then generates iterates of the form [24]

$$x_\ell^{\mathrm{D}} \;=\; W\big(W^\top V\big)^{-1}W^\top b \;+\; P^\top z_\ell, \tag{1}$$

where z_ℓ are the iterates of CG applied to the projected system

$$\widehat{A}z \;=\; \widehat{b}, \tag{2}$$

with $\widehat{A} = PA$, $\widehat{b} = Pb$, and arbitrary starting vector z_0. The projected system (2) can be solved by CG [18], since $\widehat{A}$ is symmetric positive semidefinite, satisfies $\dim(\mathcal{R}(\widehat{A})) = n - k$ [24], and $\widehat{b} \in \mathcal{R}(\widehat{A})$. Consider the eigen-decomposition

$$\widehat{A} \;=\; [Q_k\ \overline{Q}_k]\begin{bmatrix}0 & 0\\ 0 & \overline{\Delta}_k\end{bmatrix}[Q_k\ \overline{Q}_k]^\top,$$

where $Q_k \in \mathbb{R}^{n\times k}$ spans the nullspace of $\widehat{A}$, $\overline{\Delta}_k \in \mathbb{R}^{(n-k)\times(n-k)}$ is diagonal of nonzero eigenvalues of $\widehat{A}$, and $\overline{Q}_k \in \mathbb{R}^{n\times(n-k)}$ the corresponding eigenvectors. In the next lemma, we show that solving (2) is equivalent to solving the reduced system

$$\overline{\Delta}_k\widetilde{y} \;=\; \widetilde{b}, \qquad \text{where} \qquad \widetilde{b} = \overline{Q}_k^\top\widehat{b}. \tag{3}$$

Lemma 1. *Let $\{z_\ell\}$ be the iterates generated by CG applied to (2) with an initial vector z_0, and let $\{\widetilde{y}_\ell\}$ be the iterates generated by CG applied to (3) with initial vector $\widetilde{y}_0 = \overline{Q}_k^\top z_0$. Let $\widetilde{r}_\ell$ and $\widetilde{p}_\ell$ be the residuals and search directions when solving (3) with CG, and $\widehat{r}_\ell$ and $\widehat{p}_\ell$ be the residuals and search directions when solving (2) with CG. Then, for every $\ell \in \{0, \ldots, n-k\}$ such that $\widetilde{y}_\ell$ exists,*

$$\widetilde{r}_\ell = \overline{Q}_k^\top\widehat{r}_\ell \quad \textit{and} \quad \widehat{r}_\ell \in \mathcal{R}(\widehat{A}) = \mathcal{R}(\overline{Q}_k), \tag{4}$$

$$\widetilde{p}_\ell = \overline{Q}_k^\top \widehat{p}_\ell \quad \textit{and} \quad \widehat{p}_\ell \in \mathcal{R}(\widehat{A}), \tag{5}$$

$$\widetilde{y}_\ell = \overline{Q}_k^\top z_\ell, \tag{6}$$

$$Q_k Q_k^\top z_\ell = Q_k Q_k^\top z_0. \tag{7}$$

Proof. We proceed by induction. For $\ell = 0$,

$$\widetilde{r}_0 = \overline{Q}_k^\top \widehat{b} - \overline{\Delta}_k \widetilde{y}_0 = \overline{Q}_k^\top (\widehat{b} - \overline{Q}_k \overline{\Delta}_k \overline{Q}_k^\top z_0) = \overline{Q}_k^\top (\widehat{b} - \widehat{A} z_0) = \overline{Q}_k^\top \widehat{r}_0.$$

Because the projected system is consistent, $\widehat{b} \in \mathcal{R}(\widehat{A})$; hence $\widehat{r}_0 \in \mathcal{R}(\widehat{A})$. Furthermore, $\widetilde{p}_0 = \widetilde{r}_0$ and $\widehat{p}_0 = \widehat{r}_0$, so $\widetilde{p}_0 = \overline{Q}_k^\top \widehat{p}_0$. The identities (6) and (7) hold at $\ell = 0$ by definition. Assume now that (4)–(7) hold at iteration $\ell < n - k$ and CG has not terminated. The CG update for the linear system (3) gives

$$\widetilde{y}_{\ell+1} = \widetilde{y}_\ell + \frac{\widetilde{r}_\ell^\top \widetilde{r}_\ell}{\widetilde{p}_\ell^\top \overline{\Delta}_k \widetilde{p}_\ell} \widetilde{p}_\ell = \overline{Q}_k^\top z_\ell + \frac{\widehat{r}_\ell^\top \overline{Q}_k \overline{Q}_k^\top \widehat{r}_\ell}{\widehat{p}_\ell^\top \overline{Q}_k \overline{\Delta}_k \overline{Q}_k^\top \widehat{p}_\ell} \overline{Q}_k^\top \widehat{p}_\ell.$$

Since $\widehat{r}_\ell \in \mathcal{R}(\overline{Q}_k)$, then $\overline{Q}_k \overline{Q}_k^\top \widehat{r}_\ell = \widehat{r}_\ell$. We get $\widetilde{y}_{\ell+1} = \overline{Q}_k^\top \left(z_\ell + \frac{\widehat{r}_\ell^\top \widehat{r}_\ell}{\widehat{p}_\ell^\top \widehat{A} \widehat{p}_\ell} \widehat{p}_\ell \right) = \overline{Q}_k^\top z_{\ell+1}$. Using this identity, the reduced residual satisfies

$$\widetilde{r}_{\ell+1} = \overline{Q}_k^\top \widehat{b} - \overline{\Delta}_k \widetilde{y}_{\ell+1} = \overline{Q}_k^\top \left(\widehat{b} - \overline{Q}_k \overline{\Delta}_k \overline{Q}_k^\top z_{\ell+1} \right) = \overline{Q}_k^\top (\widehat{b} - \widehat{A} z_{\ell+1}) = \overline{Q}_k^\top \widehat{r}_{\ell+1},$$

where $\widehat{r}_{\ell+1} \in \mathcal{R}(\widehat{A})$. For the search direction, from the CG recurrence, we get

$$\begin{aligned} \widetilde{p}_{\ell+1} = \widetilde{r}_{\ell+1} + \frac{\widetilde{r}_{\ell+1}^\top \widetilde{r}_{\ell+1}}{\widetilde{r}_\ell^\top \widetilde{r}_\ell} \widetilde{p}_\ell &= \overline{Q}_k^\top \widehat{r}_{\ell+1} + \frac{\widehat{r}_{\ell+1}^\top \overline{Q}_k \overline{Q}_k^\top \widehat{r}_{\ell+1}}{\widehat{r}_\ell^\top \overline{Q}_k \overline{Q}_k^\top \widehat{r}_\ell} \overline{Q}_k^\top \widehat{p}_\ell \\ &= \overline{Q}_k^\top \left(\widehat{r}_{\ell+1} + \frac{\widehat{r}_{\ell+1}^\top \widehat{r}_{\ell+1}}{\widehat{r}_\ell^\top \widehat{r}_\ell} \widehat{p}_\ell \right) = \overline{Q}_k^\top \widehat{p}_{\ell+1}. \end{aligned}$$

Since $\widehat{p}_{\ell+1}$ is a linear combination of $\widehat{p}_\ell$ and $\widehat{r}_{\ell+1}$, it also belongs to $\mathcal{R}(\widehat{A})$. Finally, $Q_k Q_k^\top \widehat{p}_\ell = 0$, so the null-space component of $z_{\ell+1}$ remains equal to that of z_0. This proves (7) and completes the induction. □

In exact arithmetic, CG finds the exact solution $\widetilde{y}^*$ of (3) in at most $n - k$ iterations. From (6) and (7), the corresponding final iterate for (2) is

$$z^* = Q_k Q_k^\top z_0 + \overline{Q}_k \overline{\Delta}_k^{-1} \overline{Q}_k^\top \widehat{b}. \tag{8}$$

The first term in (8) is the component of the initial vector z_0 in the null space of $\widehat{A}$. By (7), in exact arithmetic, CG leaves it unchanged. The second term is the uniquely determined component in $\mathcal{R}(\widehat{A})$ that solves the projected system (2). For completeness, we now express the energy norm of the deflated CG error.

Proposition 1. *Let $1 \leq \ell \leq n-k$. Assuming that the ℓ-th deflated CG iterate x_ℓ^D exists, the corresponding error satisfies*

$$\left\| x^* - x_\ell^D \right\|_A^2 = \min_{p \in \mathbb{P}_\ell(0)} \| p(\overline{\Delta}_k) \left(\widetilde{y}^* - \widetilde{y}_0 \right) \|_{\overline{\Delta}_k}^2 = \min_{p \in \mathbb{P}_\ell(0)} \| p(\overline{\Delta}_k) \overline{Q}_k^\top \left(z^* - z_0 \right) \|_{\overline{\Delta}_k}^2,$$

where x^* represents the exact solution of $Ax = b$ and $\widetilde{y}_0 = \overline{Q}_k^\top z_0$.

Proof. Applying (1) to both x_ℓ^{D} and the exact solution x^* (with z_ℓ replaced by z^*), and using $PAP^\top = PA = \overline{Q}_k \overline{\Delta}_k \overline{Q}_k^\top$ (see [9]) gives

$$\left\| x^* - x_\ell^{\mathrm{D}} \right\|_A^2 = \left\| P^\top \left(z^* - z_\ell \right) \right\|_A^2 = \left(z^* - z_\ell \right)^\top PAP^\top \left(z^* - z_\ell \right)$$

$$
\begin{aligned}
&= \left(z^* - z_\ell\right)^\top PA\left(z^* - z_\ell\right) \\
&= \left(z^* - z_\ell\right)^\top \overline{Q}_k \overline{\Delta}_k \overline{Q}_k^\top \left(z^* - z_\ell\right) \\
&= \left\| \overline{Q}_k^\top (z^* - z_\ell) \right\|_{\overline{\Delta}_k}^2 = \left\| \widetilde{y}^* - \widetilde{y}_\ell \right\|_{\overline{\Delta}_k}^2 .
\end{aligned}
$$

The CG minimization property for (3) now gives

$$
\left\| \widetilde{y}^* - \widetilde{y}_\ell \right\|_{\overline{\Delta}_k}^2 = \min_{p \in \mathbb{P}_\ell(0)} \| p(\overline{\Delta}_k) \left(\widetilde{y}^* - \widetilde{y}_0 \right) \|_{\overline{\Delta}_k}^2 = \min_{p \in \mathbb{P}_\ell(0)} \| p(\overline{\Delta}_k) \overline{Q}_k^\top \left(z^* - z_0 \right) \|_{\overline{\Delta}_k}^2 . \qquad \square
$$

3 Deflated CG and PCG for a perturbed system

Let $A(t) = A + tE$, where $t \geq 0$ and E is symmetric with $\|E\|_F = 1$. Since A is SPD, $A(t)$ is SPD for sufficiently small values of t. In this section, we compare the scaled spectral preconditioner in [6] with deflated CG when solving the perturbed linear system

$$
A(t)x = b. \tag{9}
$$

Let $1 \leq k \leq n-1$. For deflated CG, we use the subspace $W = S_k = [s_1, \ldots, s_k]$, formed by the eigenvectors associated with the k largest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ of A. Set $V(t) = A(t)S_k$ and $P(t) = I_n - V(t)\left(S_k^\top V(t)\right)^{-1} S_k^\top$. The projected system is

$$
\widehat{A}(t)z = \widehat{b}(t), \tag{10}
$$

where $\widehat{A}(t) := P(t)A(t)$ and $\widehat{b}(t) := P(t)b$. The ℓ-th iterate of deflated CG is given by

$$
x_\ell^{[\mathrm{D}]}(t) = S_k \big(S_k^\top V(t)\big)^{-1} S_k^\top b + P(t)^\top z_\ell(t), \tag{11}
$$

where $z_\ell(t)$ is the iterate of CG when solving (10).

Let $\theta > 0$ and $\Lambda_k := \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$, the scaled spectral preconditioner is defined by $U^\theta = I_n + S_k\left(\sqrt{\theta}\,\Lambda_k^{-1/2} - I_k\right) S_k^\top$. The associated preconditioned system is

$$
U^\theta A(t) U^\theta y = U^\theta b, \qquad x = U^\theta y. \tag{12}
$$

Similarly, the ℓth iterate of PCG is $x_\ell^{[P]}(t) = U^\theta y_\ell(t)$, where $y_\ell(t)$ is the iterate of CG when solving (12) starting from y_0.

3.1 Perturbation analysis

In this section, we introduce the perturbation results to quantify how eigenvalues and eigenvectors change with t and to control the gap terms in our subsequent comparison of the energy-norm errors of PCG and deflated CG. We first recall the first-order perturbation theory for simple eigenvalues and eigenvectors of symmetric matrices. These classical results (see, e.g., [26]) describe how the eigenpairs $\{\lambda_i(t), s_i(t)\}$ of $A(t)$ evolve with small t. For $i = 1, \ldots, n$ and if λ_i is simple, then

$$
\lambda_i(t) = \lambda_i + t s_i^\top E s_i + O(t^2), \quad s_i(t) = s_i + t \underline{S}_i (\lambda_i I_{n-1} - \underline{\Lambda}_i)^{-1} \underline{S}_i^\top E s_i + O(t^2), \tag{13}
$$

where $\underline{\Lambda}_i = \mathrm{diag}(\lambda_j)_{j \neq i}$, and $\underline{S}_i \in \mathbb{R}^{n \times (n-1)}$ is obtained by removing the ith column from the eigenvector matrix of A. Second-order remainder estimates for eigenpairs are known; see, e.g., [19, Ch. II]. For our analysis, we provide a short self-contained proof for symmetric matrices. The resulting bound involves only concrete constants and an explicit range of validity for t, both expressed in terms of the spectral gap, without relying on contour integration.

Proposition 2. *Let $i \in \{1, \dots, n\}$, and assume that λ_i is simple. Let $g_i := \min_{j \neq i} |\lambda_i - \lambda_j| > 0$ be the spectral gap around λ_i. Then, for every $t \in [0, g_i/(4\|E\|_2)]$, $A(t)$ admits a simple eigenpair $\{\lambda_i(t), s_i(t)\}$ such that*

$$\left|\lambda_i(t) - \lambda_i - t\, s_i^\top E s_i\right| \leq \frac{4\|E\|_2^2}{g_i}\, t^2, \quad \left\|s_i(t) - s_i - t\underline{S}_i(\lambda_i I_{n-1} - \underline{\Lambda}_i)^{-1}\underline{S}_i^\top E s_i\right\|_2 \leq \frac{12\|E\|_2^2}{g_i^2}\, t^2.$$

Proof. We first establish an a priori bound. By Weyl's inequality for symmetric matrices (see, e.g., [4, 16]), $|\lambda_i(t) - \lambda_i| \leq t\,\|E\|_2$ for every t. Hence, for $0 \leq t \leq g_i/(2\|E\|_2)$, $|\lambda_i(t) - \lambda_i| \leq g_i/2$. Since, for every $j \neq i$, $|\lambda_i - \lambda_j| \geq g_i$, this gives

$$|\lambda_i(t) - \lambda_j| \geq |\lambda_i - \lambda_j| - |\lambda_i(t) - \lambda_i| \geq \frac{g_i}{2}. \tag{14}$$

Consequently, (14) implies that, for $0 \leq t \leq g_i/(2\|E\|_2)$,

$$\left\|(\lambda_i(t) I_{n-1} - \underline{\Lambda}_i)^{-1}\right\|_2 \leq \frac{2}{g_i}. \tag{15}$$

Let $0 \leq t \leq g_i/(4\|E\|_2)$, we next show that $\lambda_i(t)$ remains simple. For $j \neq i$,

$$|\lambda_i(t) - \lambda_j(t)| \geq |\lambda_i - \lambda_j| - |\lambda_i(t) - \lambda_i| - |\lambda_j(t) - \lambda_j| \geq |\lambda_i - \lambda_j| - 2t\|E\|_2 \geq g_i - 2t\|E\|_2 \geq \frac{g_i}{2},$$

where we have used in the last inequality that $0 \leq t \leq g_i/(4\|E\|_2)$. Thus $\lambda_i(t)$ is simple. We now derive exact equations for the perturbed eigenpair. Let $s_i(t)$ be a unit eigenvector associated with $\lambda_i(t)$ and write $s_i(t) = a(t)s_i + \underline{S}_i d(t)$. Its eigenvalue equation is

$$(A + tE)s_i(t) = \lambda_i(t)s_i(t). \tag{16}$$

Multiplying (16) by $\underline{S}_i^\top$ and using $\underline{S}_i^\top A s_i = 0$, $\underline{S}_i^\top A \underline{S}_i = \underline{\Lambda}_i$, and $\underline{S}_i^\top s_i = 0$, gives

$$\big(\lambda_i(t) I_{n-1} - \underline{\Lambda}_i\big) d(t) = t\, \underline{S}_i^\top E s_i(t). \tag{17}$$

The matrix in (17) is invertible by (15), and hence

$$d(t) = t\big(\lambda_i(t) I_{n-1} - \underline{\Lambda}_i\big)^{-1} \underline{S}_i^\top E s_i(t). \tag{18}$$

Because $\underline{S}_i$ has orthonormal columns and $\|s_i(t)\|_2 = 1$, taking norms in (18) and applying (15) yields,

$$\|d(t)\|_2 \leq t\frac{2}{g_i}\|\underline{S}_i^\top E s_i(t)\|_2 \leq \frac{2t\|E\|_2}{g_i}. \tag{19}$$

Since $a(t)^2 + \|d(t)\|_2^2 = 1$, combining (19) and $\frac{2t\|E\|_2}{g_i} \leq \frac{1}{2}$, gives $|a(t)| \geq \sqrt{3}/2 > 1/2$.

The eigenvalue and eigenvector gap can be bounded as follows. Multiplying (16) by $s_i^\top$ gives

$$a(t)\big(\lambda_i(t) - \lambda_i - t\, s_i^\top E s_i\big) = t\, s_i^\top E \underline{S}_i d(t). \tag{20}$$

Because $|a(t)| > 1/2$, combining (20) and (19) gives,

$$\left|\lambda_i(t) - \lambda_i - t\, s_i^\top E s_i\right| \leq \frac{1}{|a(t)|} t\|E\|_2\|d(t)\|_2 \leq 2t\|E\|_2\|d(t)\|_2 \leq \frac{4\|E\|_2^2}{g_i}\, t^2.$$

On the other hand,

$$\left\|s_i(t) - s_i - t\underline{S}_i R_0 \underline{S}_i^\top E s_i\right\|_2 \leq |a(t) - 1| + \left\|d(t) - tR_0\underline{S}_i^\top E s_i\right\|_2, \tag{21}$$

where $R_0 := \left(\lambda_i I_{n-1} - \underline{\Lambda}_i\right)^{-1}$. We bound the two terms on the right-hand side of (21) separately. Choose the sign of $s_i(t)$ so that $a(t) > 0$. Since $a(t)^2 + \|d(t)\|_2^2 = 1$,

$$0 \leq 1 - a(t) = \frac{\|d(t)\|_2^2}{1 + a(t)} \leq \|d(t)\|_2^2 \leq \frac{4t^2\|E\|_2^2}{g_i^2}. \tag{22}$$

Define $R(t) := \left(\lambda_i(t) I_{n-1} - \underline{\Lambda}_i\right)^{-1}$. Using the expression of $s_i(t)$ in (18) gives

$$d(t) = tR(t)\left(a(t)\underline{S}_i^\top E s_i + \underline{S}_i^\top E \underline{S}_i d(t)\right). \tag{23}$$

Subtracting $tR_0\underline{S}_i^\top E s_i$ from (23) yields

$$d(t) - tR_0\underline{S}_i^\top E s_i = t\big(R(t) - R_0\big)\underline{S}_i^\top E s_i + tR(t)\left((a(t) - 1)\underline{S}_i^\top E s_i + \underline{S}_i^\top E \underline{S}_i d(t)\right). \tag{24}$$

The resolvent identity gives $R(t) - R_0 = \big(\lambda_i - \lambda_i(t)\big)R(t)R_0$. Since $|\lambda_i - \lambda_i(t)| \leq t\|E\|_2$, $\|R(t)\|_2 \leq 2/g_i$ and $\|R_0\|_2 = 1/g_i$, we thus deduce,

$$\|R(t) - R_0\| \leq \frac{2t\|E\|_2}{g_i^2}. \tag{25}$$

Using the triangular inequality in (24), and (25), (19) and (22), we obtain

$$\begin{aligned}\left\|d(t) - tR_0\underline{S}_i^\top E s_i\right\|_2 &\leq t\|R(t) - R_0\|_2\|E\|_2 + t\|R(t)\|_2\left(|a(t) - 1|\|E\|_2 + \|E\|_2\|d(t)\|_2\right) \\ &\leq \frac{2t^2\|E\|_2^2}{g_i^2} + \frac{8t^3\|E\|_2^3}{g_i^3} + \frac{4t^2\|E\|_2^2}{g_i^2} \leq \frac{8t^2\|E\|_2^2}{g_i^2}.\end{aligned}$$

where we have used $\|\underline{S}_i^\top E s_i\| \leq \|E\|_2$, $\|\underline{S}_i^\top E \underline{S}_i\| \leq \|E\|_2$, and the last inequality uses $t\|E\|_2/g_i \leq 1/4$. Combining this together with to (21), (22) and (3.1) completes the proof. □

To keep the analysis simple, in the remainder of the paper, we make the following assumption on the matrix A.

Assumption 3.1. The eigenvalues of the matrix A are all simple.

3.2 Deflated CG for the perturbed system

In this section, we analyze deflated CG on the perturbed system (9).

Let $1 \leq k \leq n-1$. For sufficiently small $t \geq 0$, let $(\widehat{\lambda}_i(t), \widehat{s}_i(t))$, $i = k+1, \ldots, n$, denote the nonzero eigenpairs of $\widehat{A}(t) = P(t)A(t)$, where the eigenvectors are chosen orthonormal and the pairs are indexed by continuation from the eigenpairs (λ_i, s_i) of A at $t = 0$. Collecting them, we define

$$\overline{Q}_k(t) = [\widehat{s}_{k+1}(t), \ldots, \widehat{s}_n(t)] \quad \text{and} \quad \overline{\Delta}_k(t) = \operatorname{diag}(\widehat{\lambda}_{k+1}(t), \ldots, \widehat{\lambda}_n(t)),$$

so that $\widehat{A}(t) = \overline{Q}_k(t)\,\overline{\Delta}_k(t)\,\overline{Q}_k(t)^\top$. The next theorem gives an explicit expression for the energy-norm error of deflated CG, based on Proposition 1 and Lemma 1.

Theorem 1. *Let $1 \leq \ell < n - k$ and let z_0 be any starting vector. Assume that the deflated CG iterate $x_\ell^{[D]}(t)$ is well defined for sufficiently small $t \geq 0$. Then,*

$$\left\|x^*(t) - x_\ell^{[D]}(t)\right\|_{A(t)}^2 = \sum_{i=k+1}^{n} \frac{\widetilde{\zeta}_i(t)^2}{\widehat{\lambda}_i(t)} p_\ell^{(t)}(\widehat{\lambda}_i(t))^2,$$

where $\widetilde{\zeta}_i(t) := \widehat{s}_i(t)^\top\left(\widehat{b}(t) - \widehat{A}(t)z_0\right)$, $x^*(t)$ is the exact solution of (9), and

$$\{p_\ell^{(t)}\} = \arg\min_{p \in \mathbb{P}_\ell(0)} \|p(\overline{\Delta}_k(t))\overline{Q}_k(t)^\top(z^*(t) - z_0)\|_{\overline{\Delta}_k(t)}^2.$$

Proof. By Proposition 1 and Lemma 1, applied to (10), we have

$$\begin{aligned}\left\|x^*(t) - x_\ell^{[\mathrm{D}]}(t)\right\|_{A(t)}^2 &= \min_{p\in\mathbb{P}_\ell(0)} \left\|p(\overline{\Delta}_k(t))\overline{Q}_k(t)^\top \left(z^*(t) - z_0\right)\right\|_{\overline{\Delta}_k(t)}^2 \\ &= \left\|p_\ell^{(t)}(\overline{\Delta}_k(t))\overline{Q}_k(t)^\top \left(z^*(t) - z_0\right)\right\|_{\overline{\Delta}_k(t)}^2,\end{aligned} \tag{26}$$

where $z^*(t)$ denotes a solution of (10). The component of $z^*(t)$ in $\mathcal{R}(\overline{Q}_k(t))$ satisfies $\overline{Q}_k(t)^\top z^*(t) = \overline{\Delta}_k(t)^{-1}\overline{Q}_k(t)^\top \widehat{b}(t)$. Therefore,

$$\overline{Q}_k(t)^\top \left(z^*(t) - z_0\right) = \overline{\Delta}_k(t)^{-1}\overline{Q}_k(t)^\top \widehat{b}(t) - \overline{Q}_k(t)^\top z_0 = \overline{\Delta}_k(t)^{-1}\overline{Q}_k(t)^\top (\widehat{b}(t) - \widehat{A}(t)z_0).$$

where we used $\widehat{A}(t) = \overline{Q}_k(t)\overline{\Delta}_k(t)\overline{Q}_k(t)^\top$. Substituting this identity into (26),

$$\begin{aligned}\left\|x^*(t) - x_\ell^{[\mathrm{D}]}(t)\right\|_{A(t)}^2 &= \left\|p_\ell^{(t)}(\overline{\Delta}_k(t))\overline{\Delta}_k(t)^{-1}\overline{Q}_k(t)^\top (\widehat{b}(t) - \widehat{A}(t)z_0)\right\|_{\overline{\Delta}_k(t)}^2 \\ &= \sum_{i=k+1}^{n} \frac{\widetilde{\zeta}_i(t)^2}{\widehat{\lambda}_i(t)} p_\ell^{(t)}(\widehat{\lambda}_i(t))^2.\end{aligned}$$

□

We now present the asymptotic expansion of $\widehat{A}(t) = P(t)A(t)$, along with its associated nonzero eigenvalues and corresponding eigenvectors, as well as $\widehat{b}(t) = P(t)b$.

Lemma 2. *Under Assumption 3.1, for sufficiently small $t \geq 0$,*

$$\begin{aligned}&\widehat{A}(t) = \overline{S}_k\overline{\Lambda}_k\overline{S}_k^\top + t\overline{S}_k\overline{S}_k^\top E\overline{S}_k\overline{S}_k^\top + O(t^2), \\ &\widehat{\lambda}_i(t) = \lambda_i(t) + O(t^2), && \text{for} \quad i = k+1,\dots,n, \\ &\widehat{s}_i(t) = s_i + t\underline{S}_i D_i \underline{S}_i^\top E s_i + O(t^2), && \text{for} \quad i = k+1,\dots,n, \\ &\widehat{b}(t) = \overline{S}_k\overline{S}_k^\top b - t\overline{S}_k\overline{S}_k^\top E S_k\Lambda_k^{-1}S_k^\top b + O(t^2),\end{aligned}$$

where $\overline{\Lambda}_k := \mathrm{diag}(\lambda_{k+1},\dots,\lambda_n)$ and D_i is a diagonal matrix such that

$$[D_i]_{jj} = \begin{cases} 0, & \text{for} \quad j = 1,\dots,k, \\ 1/(\lambda_i - \lambda_j), & \text{for} \quad j = k+1,\dots,n : j \neq i. \end{cases}$$

Proof. We first expand the projector $P(t)$. Since $S_k^\top V(t) = S_k^\top(A+tE)S_k = \Lambda_k + tS_k^\top E S_k$, a first-order expansion of the inverse gives

$$(S_k^\top V(t))^{-1} = \Lambda_k^{-1} - t\Lambda_k^{-1}S_k^\top E S_k\Lambda_k^{-1} + O(t^2). \tag{27}$$

Using this expansion in the definition of $P(t)$ yields

$$\begin{aligned}P(t) &= I_n - V(t)(S_k^\top V(t))^{-1}S_k^\top \\ &= I_n - (S_k\Lambda_k + tES_k)(\Lambda_k^{-1} - t\Lambda_k^{-1}S_k^\top E S_k\Lambda_k^{-1})S_k^\top + O(t^2) \\ &= I_n - S_kS_k^\top + tS_kS_k^\top E S_k\Lambda_k^{-1}S_k^\top - tES_k\Lambda_k^{-1}S_k^\top + O(t^2) \\ &= \overline{S}_k\overline{S}_k^\top - t(I_n - S_kS_k^\top)ES_k\Lambda_k^{-1}S_k^\top + O(t^2).\end{aligned}$$

Hence, $P(t) = \overline{S}_k\overline{S}_k^\top - t\overline{S}_k\overline{S}_k^\top E S_k\Lambda_k^{-1}S_k^\top + O(t^2)$. We now use the expression of $P(t)$ to expand $\widehat{A}(t)$ and $\widehat{b}(t)$,

$$\begin{aligned}\widehat{A}(t) = P(t)A(t) &= \overline{S}_k\overline{S}_k^\top(A+tE) - t\overline{S}_k\overline{S}_k^\top E S_k\Lambda_k^{-1}S_k^\top(A+tE) + O(t^2) \\ &= \overline{S}_k\overline{\Lambda}_k\overline{S}_k^\top + t\overline{S}_k\overline{S}_k^\top E(I_n - S_kS_k^\top) + O(t^2) \\ &= \overline{S}_k\overline{\Lambda}_k\overline{S}_k^\top + t\overline{S}_k\overline{S}_k^\top E\overline{S}_k\overline{S}_k^\top + O(t^2).\end{aligned}$$

Similarly, $\widehat{b}(t) = P(t)b = \overline{S}_k\overline{S}_k^\top b - t\overline{S}_k\overline{S}_k^\top E S_k\Lambda_k^{-1}S_k^\top b + O(t^2)$.

Regarding the nonzero eigenvalues of $\widehat{A}(t)$, the matrix $\widehat{A}(t) = P(t)A(t)$, is analytic for all sufficiently small t. Standard perturbation theory therefore applies to each simple nonzero eigenvalue of $\widehat{A}(0)$. In particular, the $O(t^2)$ term in the matrix expansion above does not affect the first-order eigenvalue and eigenvector coefficients. The unperturbed matrix $\widehat{A}(0) = \overline{S}_k\Lambda_k\overline{S}_k^\top$ has, by Assumption 3.1, $n-k$ simple eigenpairs (λ_i, s_i) for $i = k+1,\dots,n$. Applying the first-order perturbation (13) to $\widehat{A}(t)$ gives, for $i = k+1,\dots,n$,

$$\widehat{\lambda}_i(t) = \lambda_i + ts_i^\top\overline{S}_k\overline{S}_k^\top E\overline{S}_k\overline{S}_k^\top s_i + O(t^2) = \lambda_i + ts_i^\top Es_i + O(t^2) = \lambda_i(t) + O(t^2),$$

where we used $\overline{S}_k\overline{S}_k^\top s_i = s_i$. Let $W_i = \operatorname{diag}(\underbrace{0,\dots,0}_{k \text{ times}}, \lambda_{k+1},\dots,\lambda_{i-1},\lambda_{i+1},\dots,\lambda_n)$, then the same first-order perturbation formula for the eigenvectors gives

$$\widehat{s}_i(t) = s_i + t\underline{S}_i(\lambda_i I_{n-1} - W_i)^{-1}\underline{S}_i^\top\overline{S}_k\overline{S}_k^\top E\overline{S}_k\overline{S}_k^\top s_i + O(t^2) = s_i + t\underline{S}_i D_i\underline{S}_i^\top Es_i + O(t^2). \qquad \square$$

3.3 Perturbation analysis of the preconditioned matrix

The perturbation expansions for the eigenvectors of $U^\theta A(t)U^\theta$ involve the factors $1/(\lambda_i - \theta)$ for $i = k+1,\dots,n$, so obtaining estimates that are uniform in the scaling parameter requires θ to stay a fixed positive distance away from the nondeflated eigenvalues. We formalize this through the following set.

Definition 1. Let γ_0 be the smallest spectral gap $\gamma_0 := \min_{1\le i\le n-1}(\lambda_i - \lambda_{i+1})$ which is positive by Assumption 3.1. For a fixed $k \in \{1,\dots,n-1\}$ and $\delta > 0$, define

$$\Theta_\delta := \left\{\theta \in [\lambda_n, \lambda_k] \mid \min_{i=k+1,\dots,n} |\theta - \lambda_i| \ge \delta\right\}. \tag{28}$$

Note that, since the eigenvalues of A are ordered decreasingly,

$$\min_{i=k+1,\dots,n} |\lambda_k - \lambda_i| = \lambda_k - \lambda_{k+1} \ge \gamma_0.$$

Thus, $\lambda_k \in \Theta_{\gamma_0}$. Decreasing δ relaxes the separation condition, so $\Theta_{\gamma_0} \subseteq \Theta_\delta$ whenever $0 < \delta \le \gamma_0$. Hence, $\lambda_k \in \Theta_\delta$ and Θ_δ is nonempty. We henceforth fix $0 < \delta \le \gamma_0$ and choose $\theta \in \Theta_\delta$, ensuring a distance of at least δ from $\lambda_{k+1},\dots,\lambda_n$.

Throughout the remainder of the paper, an estimate $R(t,\theta) = O(t^2)$ stated uniformly for $\theta \in \Theta_\delta$ means that there exist constants $C > 0$ and $\hat{t} > 0$, both independent of θ, such that

$$\|R(t,\theta)\| \le C\,t^2, \quad \text{for every } \theta \in \Theta_\delta \text{ and } 0 \le t \le \hat{t}.$$

We now turn to the asymptotic analysis of the eigenvalues and eigenvectors of the preconditioned matrix $U^\theta A(t)U^\theta$. For conciseness, set

$$A^\theta(t) := U^\theta A(t)U^\theta, \qquad \lambda_i^\theta(t) := \lambda_i(A^\theta(t)), \qquad s_i^\theta(t) := s_i(A^\theta(t)).$$

In this analysis, eigenvalue indices denote local perturbation branches rather than decreasing numerical order: indices $1,\dots,k$ label the branch issued from the eigenvalue θ, while each index $i = k+1,\dots,n$ labels the branch issued from λ_i.

Lemma 3. *Let Assumption 3.1 hold and fix $\delta \in (0,\gamma_0]$. Let $1 \le k \le n-1$. For sufficiently small $t \ge 0$, the eigenpairs of the preconditioned matrix indexed by $i = k+1,\dots,n$ admit the following expansions, uniformly in $\theta \in \Theta_\delta$,*

$$\lambda_i^\theta(t) = \lambda_i(t) + O(t^2), \qquad s_i^\theta(t) = s_i + t\underline{S}_i T_i^\theta\underline{S}_i^\top Es_i + O(t^2),$$

where T_i^θ *is an* $(n-1)\times(n-1)$ *diagonal matrix such that*

$$[T_i^\theta]_{jj} = \begin{cases} \sqrt{\theta}/(\sqrt{\lambda_j}(\lambda_i-\theta)), & \text{for} \quad j=1,\dots,k \\ 1/(\lambda_i-\lambda_j), & \text{for} \quad j=k+1,\dots,n : j\neq i. \end{cases}$$

Proof. Since $A^\theta(t) = U^\theta A U^\theta + tU^\theta E U^\theta$, the first-order perturbation (13) gives

$$\lambda_i^\theta(t) = \lambda_i^\theta(0) + ts_i^\top U^\theta E U^\theta s_i + O(t^2) = \lambda_i + ts_i^\top E s_i + O(t^2) = \lambda_i(t) + O(t^2),$$

where we used $U^\theta s_i = s_i$ and $\lambda_i^\theta(0) = \lambda_i$ for $i = k+1,\dots,n$. Applying the eigenvector expansion gives,

$$s_i^\theta(t) = s_i + t\underline{S}_i \left(\lambda_i I_{n-1} - \Omega_i\right)^{-1} \underline{S}_i^\top U^\theta E U^\theta s_i + O(t^2), \tag{29}$$

where $\Omega_i = \operatorname{diag}(\underbrace{\theta,\dots,\theta}_{k \text{ times}}, \lambda_{k+1},\dots,\lambda_{i-1},\lambda_{i+1},\dots,\lambda_n)$. Using $U^\theta s_i = s_i$ and $U^\theta \underline{S}_i = \underline{S}_i$ $\operatorname{diag}(\frac{\sqrt{\theta}}{\sqrt{\lambda_1}},\dots,\frac{\sqrt{\theta}}{\sqrt{\lambda_k}}, \underbrace{1,\dots,1}_{n-k-1 \text{ times}})$, (29) becomes $s_i^\theta(t) = s_i + t\underline{S}_i T_i^\theta \underline{S}_i^\top E s_i + O(t^2)$.

For every $i > k$ and every $\theta \in \Theta_\delta$, the distance from λ_i to the remaining spectrum of $U^\theta A U^\theta$ is at least δ. Moreover, $\|U^\theta E U^\theta\|_2 \le \|U^\theta\|_2^2 \|E\|_2 \le 1$. Hence Proposition 2, applied with the common gap δ, provides a common perturbation radius and θ-independent constants in the eigenvalue and eigenvector remainders. □

3.4 Comparison between PCG and deflated CG

This section presents the main comparison between the energy-norm errors of PCG and deflated CG. To connect the two error representations, we first compare their residual coefficients and then establish spectral bounds used in the main comparison theorem.

Throughout this section, we use the following assumptions.

Assumption 3.2. The initial guess z_0 when solving (10) is chosen such that $S_k^\top r_0 = 0$, with $r_0 = b - Az_0$. A possible choice is $z_0 = S_k \Lambda_k^{-1} S_k^\top b$.

Assumption 3.3. The initial guess for the preconditioned system (12) is chosen such that $z_0 = U^\theta y_0$.

Lemma 4. *Let Assumption 3.1 hold and fix* $\delta \in (0, \gamma_0]$. *Let* $1 \le k \le n-1$. *For* $i \in \{k+1,\dots,n\}$, *define*

$$\widetilde{\zeta}_i(t) := \widehat{s}_i(t)^\top \left(\widehat{b}(t) - \widehat{A}(t) z_0\right) \quad \text{and} \quad \widetilde{\beta}_i^\theta(t) := s_i^\theta(t)^\top \left(U^\theta b - A^\theta(t) y_0\right).$$

Then, under Assumptions 3.2 and 3.3, for a $t \ge 0$ *sufficiently small,* $\widetilde{\zeta}_i(t)^2 = \widetilde{\beta}_i^\theta(t)^2 + O(t^2)$ *uniformly in* $\theta \in \Theta_\delta$.

Proof. Using Lemma 2,

$$\begin{aligned}
\widetilde{\zeta}_i(t) &= \widehat{s}_i(t)^\top (\widehat{b}(t) - \widehat{A}(t) z_0) \\
&= (s_i + t\underline{S}_i D_i \underline{S}_i^\top E s_i)^\top (\overline{S}_k \overline{S}_k^\top r_0 - t\overline{S}_k \overline{S}_k^\top E S_k \Lambda_k^{-1} S_k^\top b - t\overline{S}_k \overline{S}_k^\top E \overline{S}_k \overline{S}_k^\top z_0) + O(t^2) \\
&= s_i^\top r_0 + t\left(s_i^\top E \underline{S}_i D_i \underline{S}_i^\top r_0 - s_i^\top E S_k \Lambda_k^{-1} S_k^\top b - s_i^\top E \overline{S}_k \overline{S}_k^\top z_0\right) + O(t^2) \\
&= s_i^\top r_0 + t\left(s_i^\top E \underline{S}_i D_i \underline{S}_i^\top r_0 - s_i^\top E S_k S_k^\top x^* - s_i^\top E \overline{S}_k \overline{S}_k^\top z_0\right) + O(t^2).
\end{aligned}$$

Set $\eta_i := s_i^\top r_0$. Then, squaring the expression of $\widetilde{\zeta}_i(t)$ gives

$$\widetilde{\zeta}_i(t)^2 = \eta_i^2 + 2t\eta_i \left(s_i^\top E \underline{S}_i D_i \underline{S}_i^\top r_0 - s_i^\top E S_k S_k^\top x^* - s_i^\top E \overline{S}_k \overline{S}_k^\top z_0\right) + O(t^2). \tag{30}$$

Let $\widehat{T}_i^\theta$ be an $(n-1)\times(n-1)$ diagonal matrix such that

$$[\widehat{T}_i^\theta]_{jj} = \begin{cases} \theta/(\lambda_j(\lambda_i-\theta)), & \text{for} \quad j=1,\dots,k, \\ 1/(\lambda_i-\lambda_j), & \text{for} \quad j=k+1,\dots,n : j\neq i. \end{cases}$$

Then, by Assumption 3.3, the preconditioned initial residual satisfies

$$U^\theta b - A^\theta(t)y_0 = U^\theta\big(b-A(t)z_0\big) = U^\theta r_0 - tU^\theta E z_0.$$

Using this identity, the expression for $\widehat{T}_i^\theta$, and Lemma 3, we obtain

$$\begin{aligned}
\widetilde{\beta}_i^\theta(t) &= s_i^\theta(t)^\top\big(U^\theta b - A^\theta(t)y_0\big) \\
&= \Big(s_i + t\underline{S}_i T_i^\theta \underline{S}_i^\top E s_i + O(t^2)\Big)^\top \Big(U^\theta r_0 - tU^\theta E z_0\Big) \\
&= s_i^\top r_0 + t\Big(s_i^\top E\underline{S}_i\widehat{T}_i^\theta\underline{S}_i^\top r_0 - s_i^\top E z_0\Big) + O(t^2) \\
&= s_i^\top r_0 + t\Big(s_i^\top E\underline{S}_i\widehat{T}_i^\theta\underline{S}_i^\top r_0 - s_i^\top E S_k S_k^\top z_0 - s_i^\top E\overline{S}_k\overline{S}_k^\top z_0\Big) + O(t^2).
\end{aligned}$$

Squaring this expression gives

$$\widetilde{\beta}_i^\theta(t)^2 = \eta_i^2 + 2t\eta_i\Big(s_i^\top E\underline{S}_i\widehat{T}_i^\theta\underline{S}_i^\top r_0 - s_i^\top E S_k S_k^\top z_0 - s_i^\top E\overline{S}_k\overline{S}_k^\top z_0\Big) + O(t^2). \tag{31}$$

From (30) and (31),

$$\begin{aligned}
\widetilde{\beta}_i^\theta(t)^2 - \widetilde{\zeta}_i(t)^2 &= 2t\eta_i\Big(s_i^\top E\underline{S}_i(\widehat{T}_i^\theta - D_i)\underline{S}_i^\top r_0 + s_i^\top E S_k S_k^\top(x^* - z_0)\Big) + O(t^2) \\
&= 2t\eta_i\Big(s_i^\top E\underline{S}_i(\widehat{T}_i^\theta - D_i)\underline{S}_i^\top r_0 + s_i^\top E S_k S_k^\top A^{-1} r_0\Big) + O(t^2) \\
&= 2t\eta_i\Big(s_i^\top E S_k\widehat{D}_i^\theta S_k^\top r_0 + s_i^\top E S_k\Lambda_k^{-1} S_k^\top r_0\Big) + O(t^2) \\
&= 2t\eta_i\Big(s_i^\top E S_k\Big(\widehat{D}_i^\theta + \Lambda_k^{-1}\Big) S_k^\top r_0\Big) + O(t^2),
\end{aligned}$$

where we used $\underline{S}_i(\widehat{T}_i^\theta - D_i)\underline{S}_i^\top = S_k\widehat{D}_i^\theta S_k^\top$, with $\widehat{D}_i^\theta := \operatorname{diag}\Big(\frac{\theta}{\lambda_j(\lambda_i-\theta)}\Big)_{j=1,\dots,k}$. By Assumption 3.2, $S_k^\top r_0 = 0$, so the entire coefficient of t vanishes. It remains to verify that the resulting $O(t^2)$ estimate is uniform in θ. In fact, first, Definition 1 gives $\lambda_n \le \theta \le \lambda_k$ and $|\lambda_i - \theta| \ge \delta$ for $i = k+1,\dots,n$. Consequently, for any $j \le k$, $\left|\frac{\theta}{\lambda_j(\lambda_i-\theta)}\right| \le \frac{1}{\delta}$, and $\left|\frac{\sqrt{\theta}}{\sqrt{\lambda_j}(\lambda_i-\theta)}\right| \le \frac{1}{\delta}$, because $\theta \le \lambda_k \le \lambda_j$. Thus, the matrices $\widehat{T}_i^\theta$ and T_i^θ are uniformly bounded for any $\theta \in \Theta_\delta$. Second, $\|U^\theta\|_2 \le 1$ uniformly. Third, the second-order eigenvector remainder in Lemma 3 is uniform over Θ_δ. All factors produced when the two expansions are multiplied and squared are therefore bounded independently of θ. Moreover, by Proposition 2 and the argument in Lemma 3, the threshold underlying the eigenvector remainder is independent of θ; hence the threshold for the resulting $O(t^2)$ estimate is independent of θ. □

The following lemma is useful for the proof of the main result.

Lemma 5. *Let Assumption 3.1 hold and fix $\delta \in (0, \gamma_0]$. Let $1 \le k \le n-1$ and $t_0 > 0$, such that $t_0\|E\|_2 < \lambda_n$. Then for every $0 \le t \le t_0$:*

(i) $[\lambda_n - t_0\|E\|_2, \lambda_{k+1} + t_0\|E\|_2]$ *contains the nonzero spectrum of $\widehat{A}(t)$;*

(ii) $[\lambda_n - t_0\|E\|_2, \lambda_k + t_0\|E\|_2]$ *contains the spectrum of $A^\theta(t)$, for every $\theta \in \Theta_\delta$.*

Proof. First, let us present the projected matrix as a Schur complement. Since $[S_k, \overline{S}_k]$ is an orthonormal basis of $\mathbb{R}^n$, write $A(t)$ in this basis as $M_{11} := S_k^\top A(t) S_k$, $M_{12} := S_k^\top A(t)\overline{S}_k$, and

$M_{22} := \overline{S}_k^\top A(t)\overline{S}_k$. Using $A(t)S_k = S_k M_{11} + \overline{S}_k M_{12}^\top$ and $S_k^\top A(t) = M_{11}S_k^\top + M_{12}\overline{S}_k^\top$, the matrix $\widehat{A}(t) = A(t) - A(t)S_k M_{11}^{-1} S_k^\top A(t)$ expands to

$$\widehat{A}(t) = A(t) - \big(S_k M_{11} + \overline{S}_k M_{12}^\top\big) M_{11}^{-1} \big(M_{11} S_k^\top + M_{12}\overline{S}_k^\top\big) = \overline{S}_k \big(M_{22} - M_{12}^\top M_{11}^{-1} M_{12}\big)\overline{S}_k^\top.$$

Since $\overline{S}_k$ has orthonormal columns, the nonzero spectrum of $\widehat{A}(t)$ therefore coincides with the spectrum of $M_{22} - M_{12}^\top M_{11}^{-1} M_{12}$, and it suffices to prove

$$\overline{\Lambda}_k - t_0\|E\|_2 I_{n-k} \preceq M_{22} - M_{12}^\top M_{11}^{-1} M_{12} \preceq \overline{\Lambda}_k + t_0\|E\|_2 I_{n-k}. \tag{32}$$

We use the following characterization of the Schur complement: For any $v \in \mathbb{R}^{n-k}$,

$$v^\top (M_{22} - M_{12}^\top M_{11}^{-1} M_{12}) v = \min_{z\in\mathbb{R}^k} (S_k z + \overline{S}_k v)^\top A(t)(S_k z + \overline{S}_k v).$$

Indeed, expanding the quadratic form gives

$$(S_k z + \overline{S}_k v)^\top A(t)(S_k z + \overline{S}_k v) = z^\top M_{11} z + 2 z^\top M_{12} v + v^\top M_{22} v, \tag{33}$$

and, since M_{11} is SPD, the minimum of (33) with respect to z is attained at $z = -M_{11}^{-1} M_{12} v$; substituting back gives $v^\top (M_{22} - M_{12}^\top M_{11}^{-1} M_{12}) v$, as claimed.

We now prove the two bounds in (32). Since $0 \le t \le t_0$, we have $A - t_0\|E\|_2 I_n \preceq A(t) \preceq A + t_0\|E\|_2 I_n$. For the lower bound, this and the characterization above give, for any $v \in \mathbb{R}^{n-k}$,

$$\begin{aligned} v^\top (M_{22} - M_{12}^\top M_{11}^{-1} M_{12}) v &\ge \min_{z\in\mathbb{R}^k} (S_k z + \overline{S}_k v)^\top (A - t_0\|E\|_2 I_n)(S_k z + \overline{S}_k v) \\ &= \min_{z\in\mathbb{R}^k} \Big\{ z^\top (\Lambda_k - t_0\|E\|_2 I_k) z + v^\top (\overline{\Lambda}_k - t_0\|E\|_2 I_{n-k}) v \Big\} \\ &= v^\top (\overline{\Lambda}_k - t_0\|E\|_2 I_{n-k}) v, \end{aligned}$$

where we used $S_k^\top A \overline{S}_k = 0$ and $S_k^\top \overline{S}_k = 0$, and where the minimum is attained at $z = 0$ because $\Lambda_k - t_0\|E\|_2 I_k \succ 0$. For the upper bound, evaluating (33) at $z = 0$ and using $A(t) \preceq A + t_0\|E\|_2 I_n$ gives, for any $v \in \mathbb{R}^{n-k}$,

$$v^\top (M_{22} - M_{12}^\top M_{11}^{-1} M_{12}) v \le v^\top \overline{S}_k^\top A(t) \overline{S}_k v \le v^\top (\overline{\Lambda}_k + t_0\|E\|_2 I_{n-k}) v,$$

where we used $\overline{S}_k^\top A \overline{S}_k = \overline{\Lambda}_k$. Together, the two bounds establish (32), and hence that the nonzero spectrum of $\widehat{A}(t)$ lies in $[\lambda_n - t_0\|E\|_2, \lambda_{k+1} + t_0\|E\|_2]$.

For $\theta \in \Theta_\delta$, the eigenvalues of $A^\theta(0)$ lie in $[\lambda_n, \lambda_k]$, and $\|U^\theta\|_2 \le 1$. Hence

$$\|A^\theta(t) - A^\theta(0)\|_2 = t\,\|U^\theta E U^\theta\|_2 \le t\,\|E\|_2 \le t_0\|E\|_2.$$

By Weyl's inequality we prove (ii). □

We now state our main result: a second-order comparison of the energy-norm errors of PCG with U^θ and deflated CG, with an explicit remainder.

Theorem 2. *Let Assumption 3.1 hold and fix $\delta \in (0, \gamma_0]$. Consider $k \in \{1, \dots, n-1\}$ and $\theta \in \Theta_\delta$. Let z_0 and y_0 be as in Assumptions 3.2 and 3.3. Then, there exist $t^* \in (0, \lambda_n/\|E\|_2)$, $\gamma_1 > 0$, and $\gamma_2 > 0$ such that, for every $1 \le \ell \le n-k$ and $t \in [0, t^*]$ such that $x_{\ell+1}^{[P]}(t) = U^\theta y_{\ell+1}$ and $x_\ell^{[\mathrm{D}]}(t)$ exist,*

$$\Big\| x^*(t) - x_{\ell+1}^{[P]}(t) \Big\|_{A(t)}^2 \le \frac{\alpha(t,\theta)^2}{\theta^2} \Big\| x^*(t) - x_\ell^{[\mathrm{D}]}(t) \Big\|_{A(t)}^2 + \Big(\gamma_1 q_k^{2\ell} + \gamma_2 \ell q_k^{2\ell-1}\Big) t^2, \tag{34}$$

where $\alpha(t,\theta) := \max\limits_{k+1\le i\le n} \Big|\widehat{\lambda}_i(t) - \theta\Big|$ and $q_k := \frac{\lambda_k + t^*\|E\|_2}{\lambda_n - t^*\|E\|_2} - 1$.

Proof. The proof proceeds in three steps: (1) we split the PCG error into a clustered and a nonclustered contribution using an admissible test polynomial; (2) we bound the clustered contribution; and (3) we bound the nonclustered contribution using Lemma 4. Let $t_0 > 0$ be chosen as in Lemma 5. A value $t^* > 0$ that is uniform in k and in $\theta \in \Theta_\delta$ is obtained by taking the minimum of t_0 and the finitely many positive thresholds associated, for $1 \le k \le n-1$, with Lemmas 2 to 4.

Fix $1 \le \ell \le n-k$ and $t \in [0, t^*]$ for which both iterates exist. For $i = 1, \dots, n$, set $\widetilde{\beta}_i^\theta(t) := s_i^\theta(t)^\top \big(U^\theta b - A^\theta(t) y_0\big)$ and let $y^*(t)$ be the exact solution of $A^\theta(t) y = U^\theta b$. Then, $x^*(t) = U^\theta y^*(t)$ and $\left\| x^*(t) - U^\theta y \right\|_{A(t)}^2 = \left\| y^*(t) - y \right\|_{A^\theta(t)}^2$ for any y. The standard CG minimization property therefore gives

$$\left\| x^*(t) - x_{\ell+1}^{[P]}(t) \right\|_{A(t)}^2 = \min_{q \in \mathbb{P}_{\ell+1}(0)} \sum_{i=1}^{n} \frac{\widetilde{\beta}_i^\theta(t)^2}{\lambda_i^\theta(t)} q\big(\lambda_i^\theta(t)\big)^2. \tag{35}$$

Consider $q_{\ell+1}^{(t)}(\lambda) = (1 - \lambda/\theta)\, p_\ell^{(t)}(\lambda) \in \mathbb{P}_{\ell+1}(0)$. Using this polynomial in (35) and separating the sum over the eigenvalues of $A^\theta(t)$ into the clustered indices $i = 1, \dots, k$ (close to θ) and the nonclustered indices $i = k+1, \dots, n$ gives

$$\left\| x^*(t) - x_{\ell+1}^{[P]}(t) \right\|_{A(t)}^2 \le T_{\mathrm{c}}^\theta(t) + T_{\mathrm{n}}^\theta(t), \tag{36}$$

where $T_{\mathrm{c}}^\theta(t) = \sum_{i=1}^{k} \widetilde{\beta}_i^\theta(t)^2 f_{\ell,t}^\theta(\lambda_i^\theta(t))$ and $T_{\mathrm{n}}^\theta(t) = \sum_{i=k+1}^{n} \widetilde{\beta}_i^\theta(t)^2 f_{\ell,t}^\theta(\lambda_i^\theta(t))$ with $f_{\ell,t}^\theta(\lambda) := \frac{1}{\lambda}\left(1 - \frac{\lambda}{\theta}\right)^2 p_\ell^{(t)}(\lambda)^2$. For conciseness, define $\underline{\mu} := \lambda_n - t^* \|E\|_2$ and $\overline{\mu}_j := \lambda_j + t^* \|E\|_2$ for $j \in \{k, k+1\}$.

We first bound $T_{\mathrm{c}}^\theta(t)$. Let $p_\ell^{(t)}(\lambda) = \prod_{j=1}^{\ell} \left(1 - \frac{\lambda}{\xi_j(t)}\right)$, where $\xi_1(t), \dots, \xi_\ell(t)$ are the Ritz values. Since these Ritz values lie in $[\underline{\mu}, \overline{\mu}_{k+1}]$, for every $\lambda \in [\underline{\mu}, \overline{\mu}_k]$,

$$\left| p_\ell^{(t)}(\lambda) \right| \le \max_{\substack{\lambda \in [\underline{\mu}, \overline{\mu}_k] \\ \xi \in [\underline{\mu}, \overline{\mu}_{k+1}]}} \left| 1 - \frac{\lambda}{\xi} \right|^\ell = \left(\frac{\lambda_k + t^* \|E\|_2}{\lambda_n - t^* \|E\|_2} - 1 \right)^\ell = q_k^\ell, \tag{37}$$

The equality follows from the fact that the maximum is attained at $\lambda = \overline{\mu}_k$ and $\xi = \underline{\mu}$ (using $\underline{\mu} > 0$ and $\overline{\mu}_k \ge \overline{\mu}_{k+1}$). For $i \le k$, Weyl's inequality gives $|\lambda_i^\theta(t) - \theta| \le t$; since $\theta \ge \lambda_n$, this yields $\left| 1 - \frac{\lambda_i^\theta(t)}{\theta} \right|^2 \le \frac{t^2}{\theta^2} \le \frac{t^2}{\lambda_n^2}$. Using (37), gives

$$\begin{aligned} T_{\mathrm{c}}^\theta(t) \le \left(\sum_{i=1}^{k} \frac{\widetilde{\beta}_i^\theta(t)^2}{\lambda_i^\theta(t)} \right) q_k^{2\ell} t^2 / \lambda_n^2 &\le \frac{\left\| U^\theta \big(b - A(t) z_0\big) \right\|_2^2}{\underline{\mu}\, \lambda_n^2} q_k^{2\ell} t^2 \\ &\le \frac{\max_{0 \le s \le t^*} \|b - A(s) z_0\|_2^2}{\underline{\mu}\, \lambda_n^2} q_k^{2\ell} t^2 =: c_{1,k} q_k^{2\ell} t^2. \end{aligned}$$

Here, we used $\lambda_i^\theta(t) \ge \underline{\mu}$ (Lemma 5) and $\|U^\theta\|_2 \le 1$. The constant $c_{1,k}$ depends on k because the initial guess z_0 depends on k. Since $k \in \{1, \dots, n-1\}$, define $c_1 := \max_{1 \le j \le n-1} c_{1,j}$. Then

$$T_{\mathrm{c}}^\theta(t) \le c_1 q_k^{2\ell} t^2. \tag{38}$$

We next bound $T_{\mathrm{n}}^\theta(t)$. Differentiating $p_\ell^{(t)}$ and using $\xi_j(t) \ge \underline{\mu}$ gives, for $\lambda \in [\underline{\mu}, \overline{\mu}_k]$,

$$\left| \partial_\lambda p_\ell^{(t)}(\lambda) \right| = \left| -\sum_{j=1}^{\ell} \frac{1}{\xi_j(t)} \prod_{\substack{h=1 \\ h \ne j}}^{\ell} \left(1 - \frac{\lambda}{\xi_h(t)}\right) \right| \le \frac{\ell}{\underline{\mu}} q_k^{\ell-1}. \tag{39}$$

Write $f^{\theta}_{\ell,t}(\lambda) = g^{\theta}(\lambda) p^{(t)}_{\ell}(\lambda)^2$ with $g^{\theta}(\lambda) = \lambda^{-1}(1-\lambda/\theta)^2$. For $\lambda \in [\underline{\mu}, \overline{\mu}_k]$,

$$|g^{\theta}(\lambda)| = \frac{1}{\lambda}(1-\frac{\lambda}{\theta})^2 \le \frac{1}{\underline{\mu}}(1+\frac{\lambda_1 + t^*\|E\|_2}{\lambda_n})^2 =: c_2, \ |\partial_\lambda g^{\theta}(\lambda)| = \left|-\frac{1}{\lambda^2}+\frac{1}{\theta^2}\right| \le \frac{1}{\underline{\mu}^2}+\frac{1}{\lambda_n^2} =: c_2'.$$

Combining these estimates with (37) and (39) yields $c_2, c_3, c_4 > 0$, independent of ℓ and of $\theta \in \Theta_\delta$, such that the following bounds hold for every $\lambda \in [\underline{\mu}, \overline{\mu}_k]$:

$$|f^{\theta}_{\ell,t}(\lambda)| \le c_2 q_k^{2\ell}, \qquad |\partial_\lambda f^{\theta}_{\ell,t}(\lambda)| \le c_3 q_k^{2\ell} + c_4 \ell q_k^{2\ell-1}. \tag{40}$$

For each $i = k+1, \dots, n$

$$\begin{aligned} \widetilde{\beta}^{\theta}_i(t)^2 f^{\theta}_{\ell,t}(\lambda^{\theta}_i(t)) = \widetilde{\zeta}_i(t)^2 f^{\theta}_{\ell,t}(\widehat{\lambda}_i(t)) + \big(\widetilde{\beta}^{\theta}_i(t)^2 - \widetilde{\zeta}_i(t)^2\big) f^{\theta}_{\ell,t}(\lambda^{\theta}_i(t)) \\ + \widetilde{\zeta}_i(t)^2 \big(f^{\theta}_{\ell,t}(\lambda^{\theta}_i(t)) - f^{\theta}_{\ell,t}(\widehat{\lambda}_i(t))\big). \end{aligned} \tag{41}$$

Lemmas 3 and 4 provide constants $c_{\beta,k,i}, c_{\lambda,k,i} \ge 0$ (independent of θ) such that

$$\left|\widetilde{\beta}^{\theta}_i(t)^2 - \widetilde{\zeta}_i(t)^2\right| \le c_{\beta,k,i} t^2, \qquad \left|\lambda^{\theta}_i(t) - \widehat{\lambda}_i(t)\right| \le c_{\lambda,k,i} t^2, \qquad i = k+1, \dots, n.$$

Setting $c_\beta := \max\limits_{1\le j\le n-1, j+1\le l\le n} c_{\beta,j,l}$ and $c_\lambda := \max\limits_{1\le j\le n-1, j+1\le l\le n} c_{\lambda,j,l}$, yields

$$\left|\widetilde{\beta}^{\theta}_i(t)^2 - \widetilde{\zeta}_i(t)^2\right| \le c_\beta t^2, \qquad \left|\lambda^{\theta}_i(t) - \widehat{\lambda}_i(t)\right| \le c_\lambda t^2, \qquad i = k+1, \dots, n. \tag{42}$$

Using (40), the sum of the second terms in (41) satisfies

$$\sum_{i=k+1}^{n} \left|\widetilde{\beta}^{\theta}_i(t)^2 - \widetilde{\zeta}_i(t)^2\right| \left|f^{\theta}_{\ell,t}(\lambda^{\theta}_i(t))\right| \le n c_\beta c_2 q_k^{2\ell} t^2. \tag{43}$$

Set $c_\zeta := \max\limits_{1\le j\le n-1} \max_{0\le s\le t^*} \sum_{i=j+1}^{n} \widetilde{\zeta}_i(s)^2$. (40) and the mean value theorem yield

$$\begin{aligned} \sum_{i=k+1}^{n} \widetilde{\zeta}_i(t)^2 \left|f^{\theta}_{\ell,t}(\lambda^{\theta}_i(t)) - f^{\theta}_{\ell,t}(\widehat{\lambda}_i(t))\right| &\le c_\zeta \max_{k+1\le i\le n} \left|f^{\theta}_{\ell,t}(\lambda^{\theta}_i(t)) - f^{\theta}_{\ell,t}(\widehat{\lambda}_i(t))\right| \\ &\le c_\zeta \sup_{\lambda\in[\underline{\mu},\overline{\mu}_k]} |\partial_\lambda f^{\theta}_{\ell,t}(\lambda)| \max_{k+1\le i\le n} \left|\lambda^{\theta}_i(t) - \widehat{\lambda}_i(t)\right| \\ &\le c_\zeta c_\lambda \left(c_3 q_k^{2\ell} + c_4 \ell q_k^{2\ell-1}\right) t^2. \end{aligned} \tag{44}$$

Summing (41) over $i = k+1, \dots, n$, and combining (43)–(44), gives

$$T^{\theta}_{\mathrm{n}}(t) \le \sum_{i=k+1}^{n} \widetilde{\zeta}_i(t)^2 f^{\theta}_{\ell,t}(\widehat{\lambda}_i(t)) + \left(c_5 q_k^{2\ell} + c_6 \ell q_k^{2\ell-1}\right) t^2,$$

where $c_5 := n c_\beta c_2 + c_\zeta c_\lambda c_3$ and $c_6 := c_\zeta c_\lambda c_4$. Using $\left|1 - \widehat{\lambda}_i(t)/\theta\right|^2 \le \alpha(t,\theta)^2/\theta^2$ and the deflated CG error representation in Theorem 1, we obtain

$$T^{\theta}_{\mathrm{n}}(t) \le \frac{\alpha(t,\theta)^2}{\theta^2} \left\|x^*(t) - x^{[\mathrm{D}]}_{\ell}(t)\right\|^2_{A(t)} + \left(c_5 q_k^{2\ell} + c_6 \ell q_k^{2\ell-1}\right) t^2. \tag{45}$$

The result follows from (36), (38), and (45) with $\gamma_1 := c_1 + c_5$ and $\gamma_2 := c_6$. □

Theorem 2 provides an explicit second-order comparison between PCG with U^{θ} and deflated CG. After one additional PCG iteration, the squared energy-norm error is bounded by the squared deflated

CG error multiplied by $(\alpha(t,\theta)/\theta)^2$, plus a remainder of order t^2. The quantity $\alpha(t,\theta)$ is the largest distance from θ to the nonzero eigenvalues of $\widehat{A}(t)$; by Lemma 2, these eigenvalues agree with the corresponding eigenvalues of $A(t)$ up to $O(t^2)$. Thus, if $\alpha(t,\theta)/\theta < 1$, the leading term is contractive.

The explicit factor $\gamma_1 q_k^{2\ell} + \gamma_2 \ell q_k^{2\ell-1}$ describes how the perturbation remainder depends on the iteration index. When $q_k > 1$, this upper bound can grow rapidly with ℓ and can be pessimistic; when $0 < q_k < 1$, its behavior is different and it may decrease as ℓ grows. This concerns only the quality of the perturbation estimate and does not imply that either iterative method has increasing error. Increasing k removes more eigenvalues from the nondeflated spectrum and may reduce q_k.

Following Diouane et al. [6], the next proposition introduces two scaling choices and shows that both yield a contractive leading factor for sufficiently small perturbations.

Proposition 3. *Suppose the assumptions of Theorem 2 hold.*

(i) *Then, for sufficiently small $t \geq 0$, the bound (34) holds with $\theta = \lambda_k$ and its leading factor is contractive, i.e., $\alpha(t,\lambda_k)/\lambda_k < 1$.*

(ii) *Suppose, in addition, that λ_n is known, and let $\theta_m := (\lambda_k + \lambda_n)/2$ with $\theta_m \notin \{\lambda_{k+1}, \ldots, \lambda_n\}$. Then, for sufficiently small $t \geq 0$, the bound (34) also holds with $\theta = \theta_m$, and its leading factor is strictly smaller than the one in* (i)*, i.e., $\alpha(t,\theta_m)/\theta_m < \alpha(t,\lambda_k)/\lambda_k < 1$.*

Proof. Indeed, since $\lambda_k - \lambda_i \geq \lambda_k - \lambda_{k+1} \geq \gamma_0$ for every $i > k$, we have $\lambda_k \in \Theta_{\gamma_0}$. Let $t^* > 0$ be the threshold supplied by Theorem 2 when it is applied with $\delta = \gamma_0$, i.e., for any $t \leq t^*$, the bound (34) holds with $\theta = \lambda_k$. At $t = 0$, the nonzero eigenvalues of $\widehat{A}(0)$ are $\lambda_{k+1}, \ldots, \lambda_n$, and hence $\alpha(0,\lambda_k)/\lambda_k = (\lambda_k - \lambda_n)/\lambda_k < 1$. Since the eigenvalues of $\widehat{A}(t)$ depend continuously on t, and hence so does $t \mapsto \alpha(t,\lambda_k)$. Therefore, there exists $\bar{t} > 0$ such that $\alpha(t,\lambda_k)/\lambda_k < 1$ for every $t \in [0,\bar{t}]$. Hence, by taking $\tau := \min\{t^*, \bar{t}\}$ proves (i).

To prove (ii), suppose that λ_n is known and that $\theta_m \notin \{\lambda_{k+1}, \ldots, \lambda_n\}$. Set $\delta := \min\left\{\gamma_0, \min_{i=k+1,\ldots,n} |\theta_m - \lambda_i|\right\}$. Then, $0 < \delta \leq \gamma_0$ and $\theta_m \in \Theta_\delta$. Let $t_m^* > 0$ be the threshold supplied by Theorem 2 for this δ, i.e., for any $t \leq t_m^*$, the bound (34) holds with $\theta = \theta_m$. At $t = 0$,

$$\frac{\alpha(0,\theta_m)}{\theta_m} = \frac{\lambda_k - \lambda_n}{\lambda_k + \lambda_n} < \frac{\lambda_k - \lambda_n}{\lambda_k} = \frac{\alpha(0,\lambda_k)}{\lambda_k} < 1. \tag{46}$$

By continuity, there exists $\bar{t}_m > 0$ such that the strict inequalities in (46) remain valid for every $t \in [0,\bar{t}_m]$. Taking $\tau_m := \min\{\tau, t_m^*, \bar{t}_m\}$ proves (ii). □

We note that the choice $\theta = \theta_m$ is particularly practical in applications such as variational data assimilation, where λ_n is often known a priori.

4 Numerical experiments

4.1 Computational complexity: PCG versus deflated CG

Assume that $S_k = [s_1, \ldots, s_k]$ and $\Lambda_k = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ are available. Both deflated CG and PCG with U^θ use CG as an iterative solver; the main computational difference lies in the cost of applying the deflation projector versus applying the scaled spectral preconditioner.

Deflated CG requires the projector $P(t) = I_n - V_k(t)\left(S_k^\top V_k(t)\right)^{-1} S_k^\top$, where $V_k(t) = A(t)S_k$. Constructing this projector requires k products with $A(t)$ to form $V_k(t)$, $\mathcal{O}(nk^2)$ operations to assemble $S_k^\top V_k(t)$, and $\mathcal{O}(k^3)$ operations to factorize it. Once these quantities are available, applying $P(t)$ requires products with $S_k^\top$ and $V_k(t)$ and the solution of a $k \times k$ linear system. Deflated CG stores S_k, $V_k(t)$, and the dense matrix $S_k^\top V_k(t)$. In contrast, PCG with the scaled spectral preconditioner

U^θ stores only S_k, the associated eigenvalues, and the scalar parameter θ. Applying U^θ requires one multiplication by each of $S_k^\top$ and S_k, together with a diagonal scaling.

4.2 Synthetic data experiments

We consider the linear system $A(t)x = (A + tE)x = b$, where $A \in \mathbb{R}^{n\times n}$ is diagonal with entries

$$\lambda_i = \lambda_n + \frac{n-i}{n-1}\left[\psi + (\lambda_1 - \lambda_n - \psi)\rho^{i-1}\right], \qquad i = 1, \dots, n, \tag{47}$$

where $0 < \rho < 1$ controls the exponential decay of the spectrum and $0 < \psi < \lambda_1 - \lambda_n$ ensures that the smallest consecutive eigenvalue gap satisfies $\gamma_0 \geq \psi/(n-1) > 0$ in exact arithmetic. Setting $\psi = 0$ recovers the spectrum used in [14]. In what follows, we fix $n = 4 \times 10^3$, $\lambda_1 = 10^6$, $\lambda_n = 1$, and $\psi = 1$. Throughout this section, E is a fixed randomly generated symmetric matrix satisfying $\|E\|_F = 1$, and $b = [1, \dots, 1]^\top/\sqrt{n}$. The initial guesses satisfy Assumptions 3.2 and 3.3.

The numerical study is organized in two parts. First, we compare the PCG error with the leading term in Theorem 2 and discuss the effect of the omitted remainder. Second, we compare the energy-norm convergence of CG, PCG using U^θ, and deflated CG for the same family of test matrices. All methods are run with full reorthogonalization. In the supplementary material, we present additional numerical results for a variational data assimilation problem.

4.2.1 Comparison with the theoretical bound

We assess the leading term in the upper bound of Theorem 2 by evaluating $\left\|x^*(t) - x_{\ell+1}^{[P]}(t)\right\|_{A(t)}^2$ and $(\alpha(t,\theta)^2/\theta^2)\left\|x^*(t) - x_\ell^{[\mathrm{D}]}(t)\right\|_{A(t)}^2$, as functions of the perturbation size t for different ℓ and ρ. We set $k = 60$ and choose θ_m as scaling. For this choice, $\alpha(t,\theta_m)/\theta_m \approx 0.99 < 1$ over the tested range of t. Table 1 reports, for $\rho \in \{0.9, 0.94\}$, the gap γ_0, the ratio λ_k/λ_n governing q_k, and the separation of θ_m from the nondeflated spectrum.

Table 1: Diagnostic quantities: the minimum eigenvalue gap γ_0, the ratio λ_k/λ_n governing q_k, and the separation $\min_{i>k}|\theta_m - \lambda_i|$ of the midrange scaling parameter from the nondeflated spectrum.

	γ_0	λ_k/λ_n	$\min_{i=k+1,\dots,n}\lvert\theta_m - \lambda_i\rvert$
$\rho = 0.9$	2.5×10^{-4}	1.969×10^3	4.387×10^1
$\rho = 0.94$	2.5×10^{-4}	2.559×10^4	1.254×10^2

From Table 1, Assumption 3.1 is satisfied; choosing $\delta = \gamma_0$ ensures that $\theta_m \in \Theta_\delta$. The behavior of the omitted remainder in Theorem 2 is governed by q_k, whose size is controlled mainly by the ratio λ_k/λ_n. From Table 1, the ratio $\lambda_k/\lambda_n > 10^3$ for both cases; consequently, the remainder $\gamma_1 q_k^{2\ell} + \gamma_2 \ell q_k^{2\ell-1}$ becomes large for large values of ℓ.

For $\rho = 0.9$ and $\ell = 20, 30$, the leading term of the right hand side of (34) lies above the PCG error over the full range of perturbations. For $\ell = 40$, it falls below the PCG error, although the discrepancy decreases for small t. For $\rho = 0.94$, the leading term lies above the PCG error for $\ell = 40$ over the full range of perturbations. For $\ell = 50$, it falls below the PCG error for t close to 1, and for $\ell = 60$ once t exceeds 10^{-6}. This behavior does not contradict Theorem 2: the theorem bounds the PCG error by the leading term *plus* an $O(t^2)$ remainder, which can be large here because of q_k. Its influence becomes apparent at later iterations, when the energy-norm errors are small. Thus, as predicted by Theorem 2, the accuracy of the leading-order approximation depends strongly on ℓ.

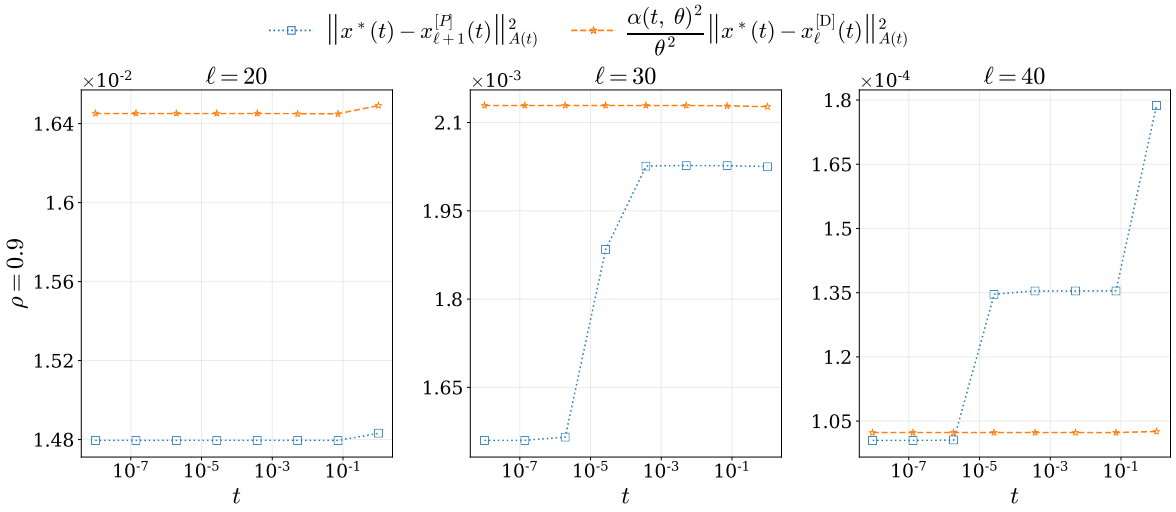

Figure 1: The squared PCG error (blue) and the leading term of the bound in Theorem 2 (orange) for $\rho = 0.9$ and $\ell \in \{20, 30, 40\}$.

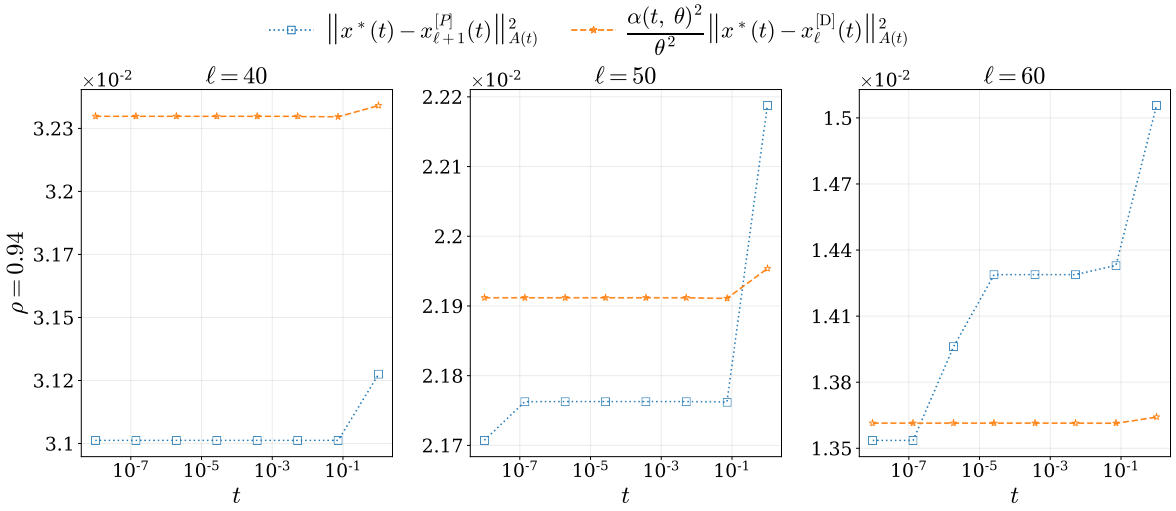

Figure 2: The squared PCG error (blue) and the leading term of the bound in Theorem 2 (orange) for $\rho = 0.94$ and $\ell \in \{40, 50, 60\}$.

4.2.2 Comparison of energy-norm errors

These numerical experiments complement the preceding comparison of the theoretical bound by examining the convergence behavior of the different solvers. We consider $k \in \{60, 80\}$, choose $\theta \in \{\theta_m, \lambda_k\}$, and apply the strategies listed in the table below for $t \in \{1, 10^{-3}, 10^{-6}\}$. Table 2 reports the methods used in the numerical experiments.

During the first iterations, the perturbation size t has little effect on the difference between deflated CG and $\mathbf{PCG}(U^{\theta_m})$. Their convergence curves separate mainly near the end of the iteration history, once the error is small. For both $\rho = 0.9$ and $\rho = 0.94$, $\mathbf{PCG}(U^{\theta_m})$ substantially improves upon CG and closely follows deflated CG at moderate accuracy. The difference is most visible at high accuracy and for the largest perturbation, $t = 1$; it narrows for $t = 10^{-3}$ and 10^{-6}. Increasing k from 60 to 80 reduces the gap between $\mathbf{PCG}(U^{\theta_m})$ and deflated CG, especially during the final iterations. Increasing

k removes more of the largest eigenvalues and decreases λ_k. Consequently, the terms $q_k^{2\ell}$ and $\ell q_k^{2\ell-1}$ in the perturbation remainder are reduced.

Table 2: Description of methods used in the numerical experiments.

Method	Description
CG	CG applied to $A(t)x = b$, starting with $x_0 = z_0$
Deflated CG	CG applied to the projected system $P(t)A(t)z = P(t)b$
PCG$(U^{\theta_{\mathrm{m}}})$	CG applied to $U^\theta A(t)U^\theta y = U^\theta b$, with $\theta = \theta_m$ and $x = U^\theta y$
PCG(U^{λ_k})	CG applied to $U^\theta A(t)U^\theta y = U^\theta b$, with $\theta = \lambda_k$ and $x = U^\theta y$

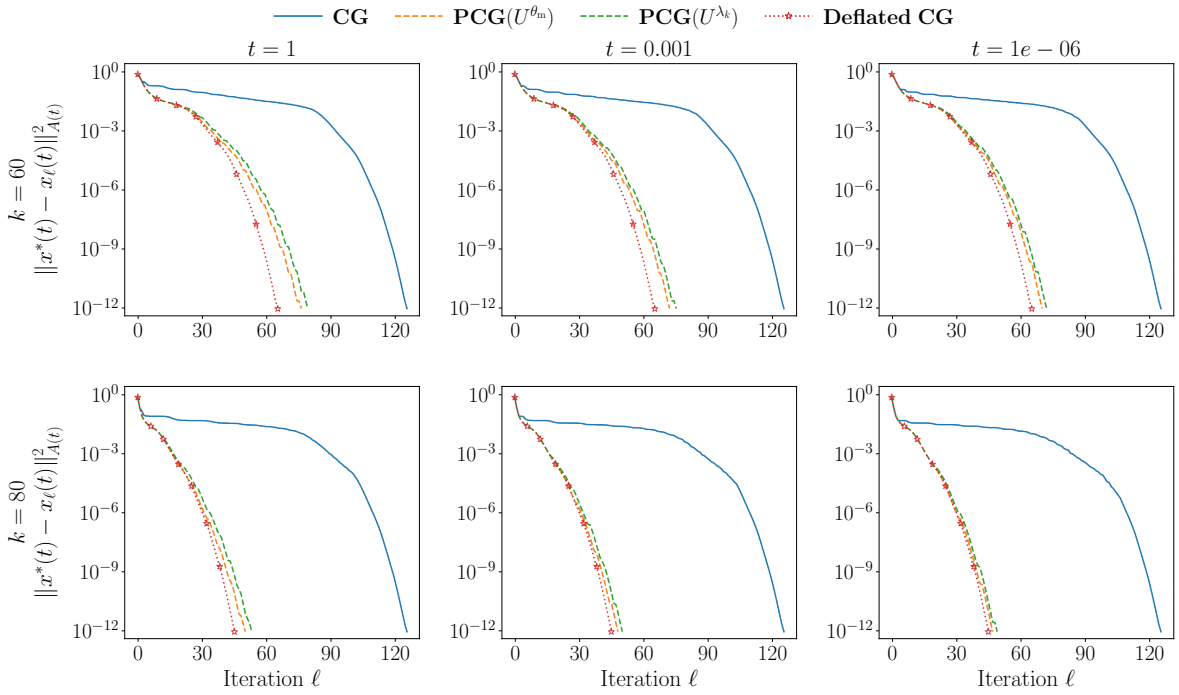

Figure 3: Squared energy-norm convergence for $\rho = 0.9$. The top and bottom rows correspond to $k = 60$ and $k = 80$, respectively; from left to right, the columns correspond to $t = 1$, 10^{-3}, and 10^{-6}.

Thus, t primarily affects the number of late iterations, rather than the initial convergence rate. PCG with θ_m retains most of the benefit of deflation but may require a few additional iterations at high accuracy. This cost can be modest relative to constructing the deflation projector, which requires k matrix–vector products with $A(t)$. The preconditioned approach is therefore especially attractive when these products are expensive. The same conclusion holds for **PCG**(U^{λ_k}), although **PCG**(U^{θ_m}) performs better, consistently with Proposition 3.

5 Conclusion

We analyzed sequences of SPD linear systems in which spectral information from a reference matrix A with simple eigenvalues is reused for a perturbed matrix $A(t) = A + tE$. Our comparison considered deflated CG and PCG equipped with a scaled spectral preconditioner constructed from the same selected eigenvectors of A.

An energy-norm characterization of deflated CG, combined with first-order perturbation analysis, yields an asymptotic comparison of the two methods. After one additional PCG iteration, the squared PCG error is bounded, up to higher-order perturbation terms, by an explicit multiple of the squared

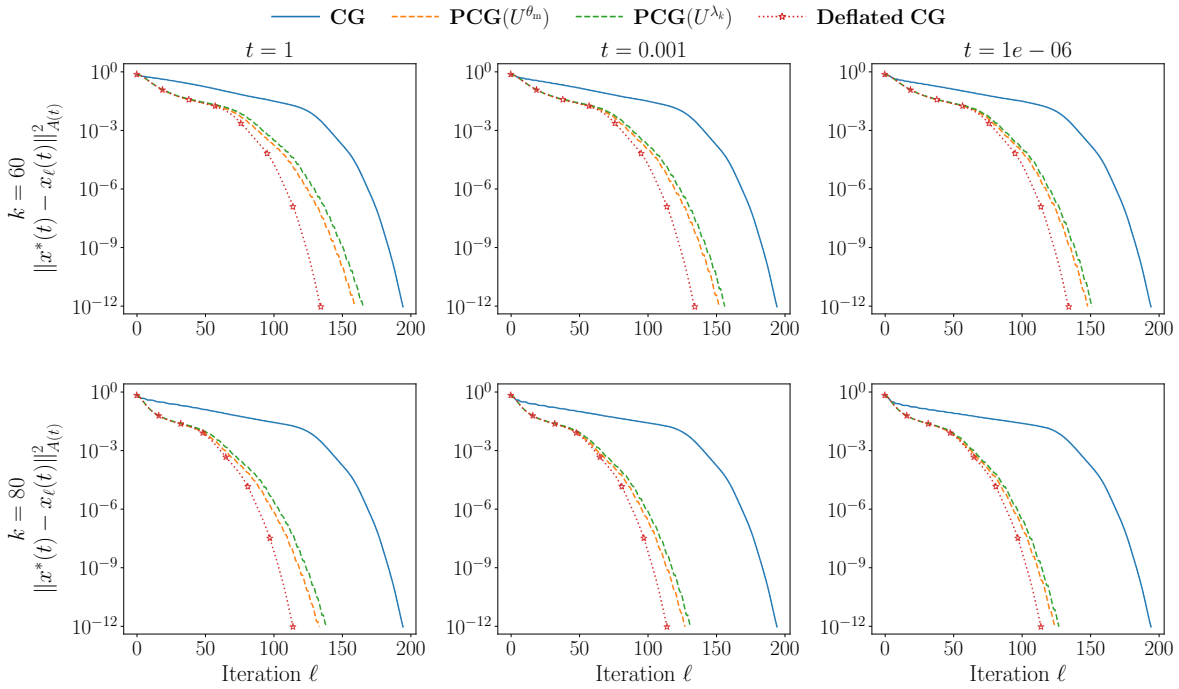

Figure 4: Same as Figure 3 for $\rho = 0.94$.

deflated CG error. The analysis also motivates the scaling parameter θ_m. Numerical experiments show that the scaled spectral preconditioner closely reproduces the convergence of deflated CG while avoiding the construction of the deflation projector and its associated products with the perturbed matrix. PCG with this preconditioner is therefore a promising alternative to deflation for sequences of closely related large-scale systems.

An important direction for future work is to extend the perturbation analysis to SPD matrices with multiple eigenvalues. We also fix the perturbation direction E and vary only its magnitude t; allowing E to vary and deriving bounds over a class of symmetric perturbations would yield a more general analysis.

References

[1] H. Al Daas, L. Grigori, P. Hénon, and P. Ricoux. Recycling Krylov subspaces and truncating deflation subspaces for solving sequence of linear systems. ACM Trans. Math. Software, 47(2):1–30, 2021. cited on p. 1

[2] M. Benzi. Preconditioning techniques for large linear systems: A survey. J. Comput. Phys., 182(2):418–477, 2002. cited on p. 1

[3] L. Bergamaschi. A survey of low-rank updates of preconditioners for sequences of symmetric linear systems. Algorithms, 13(4):100, 2020. cited on p. 1

[4] R. Bhatia. Matrix Analysis, volume 169 of Graduate Texts in Mathematics. Springer, New York, 1997. cited on p. 5

[5] B. Carpentieri, I. S. Duff, and L. Giraud. A class of spectral two-level preconditioners. SIAM J. Sci. Comput., 25(2):749–765, 2003. cited on p. 1

[6] Y. Diouane, S. Gürol, O. Mouhtal, and D. Orban. A spectral preconditioner for the conjugate gradient method with iteration budget. SIAM J. Matrix Anal. Appl., 2026. To appear. cited on pp. 1, 4, and 14

[7] Y. A. Erlangga and R. Nabben. Multilevel projection-based nested Krylov iteration for boundary value problems. SIAM J. Sci. Comput., 30(3):1572–1595, 2008. cited on p. 1

[8] M. Fisher, J. Nocedal, Y. Trémolet, and S. J. Wright. Data assimilation in weather forecasting: a case study in PDE-constrained optimization. Optim. Eng., 10(3):409–426, 2009. cited on p. 1

[9] J. Frank and C. Vuik. On the construction of deflation-based preconditioners. SIAM J. Sci. Comput., 23 (2):442–462, 2001. cited on pp. 1, 2, and 3

[10] A. Gaul, M. H. Gutknecht, J. Liesen, and R. Nabben. A framework for deflated and augmented Krylov subspace methods. SIAM J. Matrix Anal. Appl., 34(2):495–518, 2013. cited on p. 1

[11] L. Giraud and S. Gratton. On the sensitivity of some spectral preconditioners. SIAM J. Matrix Anal. Appl., 27(4):1089–1105, 2006. cited on p. 1

[12] P. Gosselet, C. Rey, and J. Pebrel. Total and selective reuse of Krylov subspaces for the resolution of sequences of nonlinear structural problems. International Journal for Numerical Methods in Engineering, 94(1):60–83, 2013. cited on p. 1

[13] S. Gratton, A. Sartenaer, and J. Tshimanga. On a class of limited memory preconditioners for large scale linear systems with multiple right-hand sides. SIAM J. Optim., 21(3):912–935, 2011. cited on p. 1

[14] A. Greenbaum and Z. Strakoš. Predicting the behavior of finite precision Lanczos and conjugate gradient computations. SIAM J. Matrix Anal. Appl., 13(1):121–137, 1992. cited on p. 15

[15] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. J. Res. Natl. Bur. Stand., 49:409–435, 1952. cited on p. 1

[16] R. A. Horn and C. R. Johnson. Matrix Analysis. Cambridge University Press, Cambridge, 2nd edition, 2012. cited on p. 5

[17] K. Kahl and H. Rittich. The deflated conjugate gradient method: Convergence, perturbation and accuracy. Linear Algebra Appl., 515:111–129, 2017. cited on p. 1

[18] W. J. Kammerer and M. Z. Nashed. On the convergence of the conjugate gradient method for singular linear operator equations. SIAM J. Numer. Anal., 9(1):165–181, 1972. cited on p. 2

[19] T. Kato. Perturbation Theory for Linear Operators. Classics in Mathematics. Springer, 1995. cited on p. 4

[20] R. Nabben and C. Vuik. A comparison of deflation and the balancing preconditioner. SIAM J. Sci. Comput., 27(5):1742–1759, 2006. cited on p. 1

[21] R. A. Nicolaides. Deflation of conjugate gradients with applications to boundary value problems. SIAM J. Numer. Anal., 24(2):355–365, 1987. cited on p. 1

[22] M. L. Parks, E. de Sturler, G. Mackey, D. D. Johnson, and S. Maiti. Recycling Krylov subspaces for sequences of linear systems. SIAM J. Sci. Comput., 28(5):1651–1674, 2006. cited on p. 1

[23] J. W. Pearson and J. Pestana. Preconditioners for Krylov subspace methods: An overview. GAMM-Mitt., 43(4):e202000015, 2020. cited on p. 1

[24] Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc'h. A deflated version of the conjugate gradient algorithm. SIAM J. Sci. Comput., 21(5):1909–1926, 2000. cited on pp. 1 and 2

[25] A. Stathopoulos and K. Orginos. Computing and deflating eigenvalues while solving multiple right-hand side linear systems with an application to quantum chromodynamics. SIAM J. Sci. Comput., 32(1):439–462, 2010. cited on p. 1

[26] G. W. Stewart. Matrix Algorithms: Volume II: Eigensystems. SIAM, 2001. cited on p. 4

[27] J. M. Tang, R. Nabben, C. Vuik, and Y. A. Erlangga. Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods. J. Sci. Comput., 39(3):340–370, 2009. cited on p. 1

[28] J. Tshimanga, S. Gratton, A. T. Weaver, and A. Sartenaer. Limited-memory preconditioners, with application to incremental four-dimensional variational data assimilation. Q. J. R. Meteorol. Soc., 134 (632):751–769, 2008. cited on p. 1

[29] N. Venkovic, P. Mycek, L. Giraud, and O. Le Maître. Comparative study of harmonic and Rayleigh-Ritz procedures with applications to deflated conjugate gradients. PhD thesis, CERFACS, 2020. cited on p. 1

[30] A. J. Wathen. Preconditioning. Acta Numer., 24:329–376, 2015. cited on p. 1