DELFT UNIVERSITY OF TECHNOLOGY

REPORT 10-14

On the Convergence of GMRES with Invariant-Subspace Deflation

M.C. Yeung, J.M. Tang, and C. Vuik

ISSN 1389-6520

Reports of the Delft Institute of Applied Mathematics

Delft 2010

On the Convergence of GMRES with Invariant-Subspace Deflation

M.C. Yeung *, J.M. Tang †, C. Vuik ‡

June 23, 2010

Abstract

We consider the solution of large and sparse linear systems of equations by GMRES. Due to the appearance of unfavorable eigenvalues in the spectrum of the coefficient matrix, the convergence of GMRES may hamper. To overcome this, a deflated variant of GMRES can be used, which treats those unfavorable eigenvalues effectively. In the literature, several deflated GMRES variants are applied successfully to various problems, while a theoretical justification is often lacking. In contrast to deflated CG, the convergence of deflated GMRES seems to be harder to analyze and to understand.

This paper presents some new theoretical insights into deflated GMRES based on A-invariant deflation subspaces. Fundamental results regarding the convergence of deflated GMRES are proved in order to show the effectiveness and robustness of this method. Numerical experiments are provided to illustrate the theoretical results and to show some further properties of deflated GMRES. Consequently, practical variants of deflated GMRES from the literature can be better understood based on the results presented in this paper.

*University of Wyoming, Department of Mathematics, 1000 East University Avenue Laramie, WY 82071, USA (myeung@uwyo.edu), This research was supported by Flittie Sabbatical Augmentation Award, University of Wyoming.

†University of Minnesota, Department of Computer Science and Engineering, 200 Union Street S.E, Minneapolis, MN 55455, USA (jtang@cs.umn.edu), Work supported in part by DOE under grant DE-FG 08ER 25841.

‡Delft University of Technology, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft Institute of Applied Mathematics, P.O. Box 5031, 2600 GA Delft, The Netherlands, (c.vuik@tudelft.nl)

1 Introduction

This paper is devoted to the solution of large and sparse linear systems of the form

$$Ax = b, \tag{1}$$

where $A \in \mathbb{C}^{N\times N}$ and $b \in \mathbb{C}^N$ are a given nonsingular coefficient matrix and right-hand side, respectively. Eq. (1) arises in many practical applications, and much attention in the literature is paid to its efficient solvers.

Popular methods to solve (1) are Krylov-subspace methods, such as CG [12], LSQR [25], GCR [8], GMRES [29], and Bi-CGSTAB [35]. For a comprehensive overview of those methods, we refer to [28]. Recently, some promising Krylov-subspace methods have appeared, such as ML(k)BiCGSTAB [38] and IDR(s) [31].

Krylov-subspace methods based on a symmetric positive-definite matrix A, such as CG, are analyzed in, e.g., [20, 19, 14, 13, 34]. Convergence bounds of CG can be given in terms of the spectral properties of A. Krylov-subpace methods applied to general matrices are usually harder to analyze, and mostly restricted to methods with an optimality property, such as GMRES, see, e.g., [8, 11, 18, 29, 36]. Convergence bounds of GMRES can be provided in terms of the spectral properties of A and the condition of its eigenvector matrix.

A Krylov-subspace method is usually combined with preconditioning in order to obtain a fast convergence of the iterative process, see [28] for a discussion of various preconditioning techniques. For relatively simple problems, a preconditioned Krylov-subspace method is proved to be a successful method. However, for more ill-conditioned problems, the convergence of preconditioned Krylov-subspace methods may deteriorate, due to the appearance of unfavorable eigenvalues in the spectrum of the preconditioned coefficient matrix. A cure for this problem is to project the corresponding eigenvector components out of the residuals, so that the effect of those eigenvalues on the convergence is eliminated. This technique is known as second-level preconditioning, which includes deflation, coarse-grid correction, augmentation, and recycling techniques. In most of these second-level preconditioning approaches, spectral information is projected out of the error. In this paper, we focus on *deflation* methods [7, 21, 24], and note that some results presented here may be generalized to other methods.

As exact eigenvectors are usually not available and hard to compute, approximations of them are often adopted in the deflation approach. This strategy works surprisingly well

for many problems. For CG-like methods with deflation, this is illustrated in, e.g., [10, 17, 22, 23, 30, 37]. It can be shown that the use of approximated eigenvectors or even arbitrary vectors as deflation vectors accelerates the CG convergence [22]. For a discussion on GMRES-like methods with deflation, we refer to, e.g., [2, 3, 4, 5, 6, 9, 15, 16, 26, 27, 32]. In contrast to deflated variants of CG, a convergence analysis for deflated variants of GMRES is more complicated to perform. This is partially caused by the fact that deflation seems not always be effective, for example if A is nonnormal. In this case, convergence bounds on (deflated) GMRES do not necessarily predict the actual convergence. In addition, restarted and truncated variants rather than the original GMRES method are frequently used in practice, whose convergence is usually not easy to predict. It has been shown in the literature that deflated GMRES based on well-approximated eigenvectors, such as Ritz vectors, is effective in several applications. The choice for approximated eigenvectors as deflation vectors is usually justified by numerical experiments, while an appropriate theoretical justification is often lacking. It even seems that the convergence theory on deflated GMRES based on exact eigenvectors is hardly explored, while it is fundamental to understand the performance of advanced deflation approaches. Most convergence theory on deflated GMRES and their advanced variants are solely based on showing that practical deflation vectors span an approximate invariant subspace that is sufficiently close to the exact invariant subspace, under the assumption that deflation based on this exact invariant subspace is effective. To the best of our knowledge, there are however no theoretical results known in the literature that justisfy the latter assumption.

Here, we focus on solving Eq. (1) by deflated GMRES, also denoted by D-GMRES. A straightforward implementation of D-GMRES is adopted in which the deflation vectors span an exact invariant subspace of A and no restart or truncation of the iterative procedure is used. The aim of this paper is to provide some fundamental but new convergence analysis of D-GMRES. Major results include that D-GMRES does not break down and always converges faster than GMRES. Those insights can be used to get a better understanding of the performance of practical deflated GMRES methods in which more general deflation vectors and restarted/truncated variants are applied.

The outline of the rest of this paper is as follows. In Section 2, we define the considered problem and review some convergence results for GMRES. Section 3 is devoted to describe the D-GMRES and compare the spectrum of GMRES and D-GMRES. In Section 4, we prove that D-GMRES can be interpreted as GMRES applied to solve a smaller linear system. As a consequence of the latter result, it can be shown that D-GMRES always

converges to the correct solution for any initial guess, see Section 4. The (convergence of) residuals of D-GMRES are further analyzed in Section 5. We prove that the norm of the D-GMRES residuals is always smaller than that of the GMRES residual during the whole iterative process. In addition, we show that increasing the dimension of the A-invariant subspace leads to a faster convergence. Finally, we generalize some theoretical bounds on the GMRES residuals to D-GMRES residuals. Section 6 is devoted to the application of D-GMRES for solving real linear systems, and show how all computations can be kept in real arithmetic. Numerical experiments are presented in Section 7 in order to illustrate the theory and to provide some additional insights into D-GMRES. We end up with the conclusions in Section 8.

2 Problem Definition

The linear system (1) is considered, where we assume that $A \in \mathbb{C}^{N\times N}$ is nonsingular and $b \in \mathbb{C}^N$ throughout this paper. *Generalized Minimal Residual method* (GMRES) [29] is chosen to solve (1). Some results on GMRES are reviewed below.

Let an initial guess $x_0 \in \mathbb{C}^N$ be given along with its residual $r_0 \equiv b - Ax_0$. Then, GMRES recursively constructs an approximate solution, x_m, such that

$$x_m \in x_0 + \mathcal{K}_m(A, r_0) \equiv x_0 + \text{span}\{r_0, Ar_0, \ldots, A^{m-1}r_0\}, \tag{2}$$

and

$$\|r_m\|_2 = \min_{\xi\in x_0+\mathcal{K}_m(A,r_0)} \|b - A\xi\|_2, \tag{3}$$

where $r_m \equiv b - Ax_m$ and $m = 1, 2, \ldots$. We call $\mathcal{K}_m(A, r_0)$ the m-th *searching subspace*, as GMRES searches for an approximate solution in the affine subspace $x_0 + \mathcal{K}_m(A, r_0)$ at iteration m. Because of (2), x_m can be written as

$$x_m = x_0 + p_{m-1}(A)r_0,$$

for some $p_{m-1} \in \mathcal{P}_{m-1}$, where $\mathcal{P}_{m-1}$ denotes the set of all the polynomials with degree at most $m - 1$. In addition, we have

$$\begin{aligned}\|r_m\|_2 &= \|b - Ax_m\|_2 = \|(I - Ap_{m-1}(A))r_0\|_2 \\ &= \min_{p\in\mathcal{P}_{m-1}} \|(I - Ap(A))r_0\|_2 = \min_{p\in\mathcal{P}_m, p(0)=1} \|p(A)r_0\|_2.\end{aligned}$$

On the other hand, we can apply the Arnoldi process [1], starting with $u_1 = r_0/\|r_0\|_2$, to generate an orthonormal basis $\{u_1, u_2, \ldots, u_{m+1}\}$ of $\mathcal{K}_{m+1}(A, r_0)$ and an upper Hessenberg matrix $H_{m+1,m} \in \mathbb{C}^{(m+1)\times m}$ such that they satisfy

$$AU_m = U_{m+1}H_{m+1,m}, \tag{4}$$

where $U_j \equiv [u_1, u_2, \ldots, u_j]$. Then, Eq. (2) can be rewritten as $x_m = x_0 + U_m\eta$, for some $\eta \in \mathbb{C}^m$. Therefore, (3) becomes

$$\begin{aligned} \|r_m\|_2 &= \|b - A(x_0 + U_m\eta)\|_2 = \min_{\theta\in\mathbb{C}^m} \|b - A(x_0 + U_m\theta)\|_2 \\ &= \min_{\theta\in\mathbb{C}^m} \|r_0 - AU_m\theta\|_2 = \min_{\theta\in\mathbb{C}^m} \|\|r_0\|_2 u_1 - U_{m+1}H_{m+1,m}\theta\|_2 \\ &= \min_{\theta\in\mathbb{C}^m} \|\|r_0\|_2 e_1 - H_{m+1,m}\theta\|_2, \end{aligned} \tag{5}$$

where e_1 is the first column of the $(m+1)\times(m+1)$ identity matrix, I.

Remark 2.1 *In the Arnoldi process, if we assume that all entries in the lower sub-diagonal of $H_{m+1,m}$ are nonnegative, then U_m and $H_{m+1,m}$ are uniquely determined by A and u_1.*

If A is diagonalizable, an upper bound on $\|r_m\|_2$ is provided by the following result.

Proposition 2.2 *(see [28, Prop. 6.32]) Assume that A can be decomposed as*

$$A = V\Lambda V^{-1}, \tag{6}$$

with Λ being the diagonal matrix of eigenvalues. Then,

$$\|r_m\|_2 \le \kappa_2(V)\epsilon_m\|r_0\|_2,$$

where $\kappa_2(V) = \|V\|_2\|V^{-1}\|_2$ and $\epsilon_m = \min_{p\in\mathcal{P}_m, p(0)=1} \max_{\lambda\in\sigma(A)} |p(\lambda)|$.

An upper bound for ϵ_m can be derived based on Chebyshev polynomials and ellipses in which the spectrum of A is contained. The result is given in Corollary 2.3.

Corollary 2.3 *(see [28, Cor. 6.3.3]) Suppose that A has a spectral decomposition (6). Let $E(c, d, a)$ denote the ellipse with center $c \in \mathbb{R}$, focal distance $d \ge 0$, and semi-major axis $a \ge 0$. Let C_m be the Chebyshev polynomial of degree m. If all eigenvalues of A are located*

in $E(c,d,a)$ that excludes the origin of the complex plane, then

$$\|r_m\|_2 \leq \kappa_2(V)\frac{C_m(\frac{a}{d})}{C_m(\frac{c}{d})}\|r_0\|_2 \approx \kappa_2(V)\delta^m\|r_0\|_2, \tag{7}$$

where $\delta = \frac{a+\sqrt{a^2-d^2}}{c+\sqrt{c^2-d^2}}$.

The upper bound in (7) contains two factors: the condition number of the eigenvector matrix V, $\kappa_2(V)$, and the scalar δ determined by the distribution of the eigenvalues. If A is nearly normal and has a spectrum $\sigma(A)$ which is clustered around 1, we obtain $\kappa_2(V) \approx 1$ and $\delta < 1$. In this case, $\|r_m\|_2$ decays in a rate of power δ^m, resulting in a fast convergence of GMRES.

In practice, GMRES is generally applied to solve a preconditioned system rather than to solve (1) directly. For example, the right-preconditioned system

$$AM^{-1}y = b \quad \text{and} \quad y = Mx,$$

can be solved by GMRES. The preconditioner $M \in \mathbb{C}^{N\times N}$ is usually a good approximation of A, namely $AM^{-1} \approx I$, so that we can assume without loss of generality that the eigenvalues of A in (1) are clustered around 1 with a few possible outliers, and avoid the explicit use of M in the remainder of this paper.

Since the ellipse $E(c,d,a)$ in Corollary 2.3 is required to include all eigenvalues of A, the outlying eigenvalues may keep the ellipse large, implying a large δ. To reduce the δ in (7), we therefore wish to remove these outlying eigenvalues from $\sigma(A)$. Any procedure of doing so is known as *deflation*. GMRES in combination with deflation is called *deflated GMRES* (D-GMRES).

In this paper, we focus on deflated GMRES based on the following. Suppose $\{v_1,\ldots,v_k\}$ is a set of vectors to be deflated that span a basis of some A-invariant subspace. For example, those vectors can be eigenvectors of A. Then, in D-GMRES, the m-th searching subspace $\mathcal{K}_m(A,r_0)$ is projected onto a subspace that is orthogonal to $v_1,\ldots,v_k$. Therefore, approximate solutions are found that are orthogonal to $v_1,\ldots,v_k$. As a consequence, these (eigen)vectors have no effect on the solution process, and the iterative process could be improved significantly.

Remark 2.4 *There are several other variants known of deflation that is applied to GMRES. For example, an approach referred to as Augmented GMRES proceeds as follows.*

Add eigenvectors $v_1, \ldots, v_k$ to the m-th searching subspace $\mathcal{K}_m(A, r_0)$ to form an augmented subspace $\mathcal{K}_m^{aug}$. Then, find $x_m \in x_0 + \mathcal{K}_m^{aug}$ with the property $\|r_m\|_2 = \min_{\xi \in x_0 + \mathcal{K}_m^{aug}} \|b - A\xi\|_2$, or, equivalently, $r_m \in (A\mathcal{K}_m^{aug})^\perp$. Note that the subspace $(A\,\mathcal{K}_m^{aug})^\perp$ is orthogonal to $v_1, \ldots, v_k$, so that it is mathematically equivalent to deflated GMRES as described above.

3 Deflated GMRES

In this section, we describe the deflated GMRES in more detail.

Consider the solution of system (1) and suppose x^* is its exact solution. Suppose that a so-called deflation-subspace matrix $Z = [z_1, \ldots, z_k] \in \mathbb{C}^{N \times k}$ is given, whose columns are linearly independent. Define the two projectors

$$P \equiv I - AZ(Z^H AZ)^{-1} Z^H \quad \text{and} \quad \widetilde{P} \equiv I - Z(Z^H AZ)^{-1} Z^H A, \tag{8}$$

where $Z^H AZ$ is assumed to be invertible. It is straightforward to verify that $P^2 = P$, $\widetilde{P}^2 = \widetilde{P}$ and $PA = A\widetilde{P}$. Using $\widetilde{P}$, we split x^* into two parts:

$$x^* = (I - \widetilde{P})x^* + \widetilde{P}x^* \equiv x_1^* + x_2^*.$$

For x_1^*, we have

$$x_1^* = (I - \widetilde{P})x^* = Z(Z^H AZ)^{-1} Z^H Ax^* = Z(Z^H AZ)^{-1} Z^H b.$$

For x_2^*, we obtain

$$x_2^* = A^{-1} Pb,$$

since $Ax_2^* = A\widetilde{P}x^* = PAx^* = Pb$. Now, if $x^\#$ is a solution of the singular system

$$PAx = Pb, \tag{9}$$

then

$$A\widetilde{P}x^\# = Pb \;\; \Leftrightarrow \;\; \widetilde{P}x^\# = A^{-1}Pb = x_2^*.$$

Based on the above observations, a deflated GMRES algorithm is given in Algorithm 1.

Unless stated otherwise, the following assumption holds in the remainder of this paper.

Assumption 3.1 *The columns of Z form a basis for an A-invariant subspace.*

Algorithm 1: Deflated GMRES

1 Choose Z
2 Compute $x_1^* = Z(Z^H AZ)^{-1} Z^H b$
3 Solve $PAx = Pb$ by GMRES to obtain a solution $x^{\#}$
4 Compute $x_2^* = \widetilde{P} x^{\#}$
5 Determine $x^* = x_1^* + x_2^*$

For example, if $A = VJV^{-1}$ is a Jordan canonical form of A with $V = [v_1, \ldots, v_N]$, then $\{v_1, \ldots, v_k\}$ with $1 \le k \le N$ can be used for Z, as it is a basis for the A-invariant subspace $\text{span}\{v_1, \ldots, v_k\}$. Furthermore, except for the results in Section 5.3, all theoretical results presented next also hold for A that is not necessarily diagonalizable. If A is diagonalizable, a spectral decomposition exists and eigenvectors can be used as columns for Z.

Suppose that A is diagonalizable, and set $Z = [v_1, \ldots, v_k]$, whose columns are eigenvectors of A associated with eigenvalues $\lambda_1, \ldots, \lambda_k$, respectively. Then, the spectrum $\sigma(PA)$ would contain the same eigenvalues of A except $\lambda_1, \ldots, \lambda_k$. This is a consequence of Proposition 3.2, which is given below.

When GMRES is applied to solve (9), the m-th searching subspace is $\mathcal{K}_m(PA, Pr_0)$. Since $PAP = PA$ follows from Lemma 4.1, we have $\mathcal{K}_m(PA, Pr_0) = P\mathcal{K}_m(A, r_0)$. Therefore, $\mathcal{K}_m(PA, Pr_0)$ is the projection (induced by the projector P) of $\mathcal{K}_m(A, r_0)$ onto the full Krylov subspace $\mathcal{K}(PA, Pr_0) \equiv \text{span}\{(PA)^k (Pr_0) \,|\, k = 0, 1, 2, \ldots\}$, so that this full Krylov subspace is orthogonal to $v_1, \ldots, v_k$.

Proposition 3.2 *Suppose that $A \in \mathbb{C}^{N\times N}$ is nonsingular and $Z \in \mathbb{C}^{N\times k}$ has columns that form a basis for some A-invariant subspace, $\mathcal{Z}$. Choose $W \in \mathbb{C}^{N\times(N-k)}$ such that its columns form a basis for $\mathcal{Z}^{\perp}$.*[1] *Then,*

(a) A can be decomposed as

$$A = [Z, W] \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} [Z, W]^{-1}, \tag{10}$$

for some $B_{11} \in \mathbb{C}^{k\times k}$, $B_{12} \in \mathbb{C}^{k\times(N-k)}$ and $B_{22} \in \mathbb{C}^{(N-k)\times(N-k)}$;

(b) $Z^H AZ$ is nonsingular;

[1] W is an auxiliary quantity in this and the following proofs. In an application of deflated GMRES, we do not need to compute W explicitly, while Z should be provided to form the deflation matrix P.

(c) $PA = [Z, W] \begin{bmatrix} 0 & 0 \\ 0 & B_{22} \end{bmatrix} [Z, W]^{-1}$.

Proof. Part (a) is obvious. To prove Parts (b) and (c), we rewrite (10) as

$$A[Z, W] = [Z, W] \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix},$$

which yields

$$AZ = ZB_{11} \quad \text{and} \quad AW = ZB_{12} + WB_{22}. \tag{11}$$

Since $\det(A) = \det(B_{11}) \det(B_{22})$ holds due to (10) and A is nonsingular, B_{11} and B_{22} are both nonsingular. Therefore, $Z^H AZ = Z^H(ZB_{11}) = (Z^H Z)B_{11}$. The nonsingularity of $Z^H AZ$ follows from those of $Z^H Z$ and B_{11}.

By noting that $Z^H W = 0$, we have

$$\begin{aligned} PAW &= [I - AZ(Z^H AZ)^{-1} Z^H]AW = AW - AZ(Z^H AZ)^{-1} Z^H AW \\ &= AW - AZ(Z^H AZ)^{-1} Z^H (ZB_{12} + WB_{22}) = AW - AZ(Z^H AZ)^{-1}(Z^H Z)B_{12} \\ &= AW - (ZB_{11})(Z^H ZB_{11})^{-1}(Z^H Z)B_{12} = AW - ZB_{12} = WB_{22}. \end{aligned}$$

Moreover, $PAZ = [I - AZ(Z^H AZ)^{-1} Z^H]AZ = 0$. Hence,

$$PA[Z, W] = [0, WB_{22}] = [Z, W] \begin{bmatrix} 0 & 0 \\ 0 & B_{22} \end{bmatrix}.$$

■

Corollary 3.3 *Under the assumptions of Proposition 3.2,* $\sigma(A) = \sigma(B_{11}) \cup \sigma(B_{22})$ *and* $\sigma(PA) = \{0, \ldots, 0\} \cup \sigma(B_{22})$ *hold.*

Thus, $\sigma(B_{11})$ is deflated from $\sigma(A)$, when the projector P (with Z consisting of vectors that form a basis for an A-invariant subspace) is applied to A. Moreover, it is not required to assume $Z^H AZ$ to be nonsingular in D-GMRES, as it is automatically satisfied for A-invariant subspace deflation.

4 Solution Equivalence of Linear Systems

In this section, we show that solving the deflated and singular system (9) is equivalent to solving a smaller and nonsingular system by GMRES. Based on this result, we prove that

GMRES finds a solution of (9) starting with an arbitrary initial guess.

Let the columns $\{z_1, \ldots, z_k\}$ of Z be extended to a basis $\{z_1, \ldots, z_k, z_{k+1}, \ldots, z_N\}$ of $\mathbb{C}^N$. Then, perform a QR factorization on the matrix $[z_1, \ldots, z_k, z_{k+1}, \ldots, z_N] \equiv [Z, \widetilde{Z}]$ as follows:

$$[Z, \widetilde{Z}] = QR \equiv [Q_1, Q_2] \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}, \tag{12}$$

where $Q_1 \in \mathbb{C}^{N \times k}$ and $R_{11} \in \mathbb{C}^{k \times k}$. This yields

$$Z = Q_1 R_{11} \quad \text{and} \quad \widetilde{Z} = Q_1 R_{12} + Q_2 R_{22}. \tag{13}$$

Lemma 4.1 now shows that the deflation matrix, P, and deflated coefficient matrix, PA, can be written in terms of Q_2 and B_{22}.

Lemma 4.1 *Set $W = Q_2$ in Proposition 3.2. Then, matrices P and PA in the proposition become*

$$P = Q_2 Q_2^H \quad \textit{and} \quad PA = Q_2 B_{22} Q_2^H,$$

and the equality $PAP = PA$ holds.

Proof. First, by using (11) and (13), we have

$$\begin{aligned} P &= I - AZ(Z^H AZ)^{-1} Z^H = I - ZB_{11}(Z^H Z B_{11})^{-1} Z^H = I - Z(Z^H Z)^{-1} Z^H \\ &= I - Q_1 R_{11} (R_{11}^H R_{11})^{-1} (Q_1 R_{11})^H = I - Q_1 Q_1^H = QQ^H - Q_1 Q_1^H \\ &= [Q_1, Q_2][Q_1, Q_2]^H - Q_1 Q_1^H = Q_2 Q_2^H. \end{aligned}$$

For the equation of PA, we first note that

$$\begin{aligned} A &= [Z, Q_2] \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} [Z, Q_2]^{-1} = [Q_1 R_{11}, Q_2] \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} [Q_1 R_{11}, Q_2]^{-1} \\ &= [Q_1, Q_2] \begin{bmatrix} R_{11} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{bmatrix} \begin{bmatrix} R_{11} & 0 \\ 0 & I \end{bmatrix}^{-1} [Q_1, Q_2]^{-1} \\ &= [Q_1, Q_2] \begin{bmatrix} R_{11} B_{11} R_{11}^{-1} & R_{11} B_{12} \\ 0 & B_{22} \end{bmatrix} [Q_1, Q_2]^{-1}. \end{aligned} \tag{14}$$

Hence,

$$\begin{aligned} PA &= Q_2 Q_2^H [Q_1, Q_2] \begin{bmatrix} R_{11} B_{11} R_{11}^{-1} & R_{11} B_{12} \\ 0 & B_{22} \end{bmatrix} [Q_1, Q_2]^H \\ &= Q_2 [0, I] \begin{bmatrix} R_{11} B_{11} R_{11}^{-1} & R_{11} B_{12} \\ 0 & B_{22} \end{bmatrix} [Q_1, Q_2]^H \\ &= Q_2 \begin{bmatrix} 0 & B_{22} \end{bmatrix} [Q_1, Q_2]^H = Q_2 B_{22} Q_2^H. \end{aligned}$$

Finally, we obtain

$$PAP = (Q_2 B_{22} Q_2^H)(Q_2 Q_2^H) = Q_2 B_{22} Q_2^H = PA.$$

■

Next, we consider the solution of the systems

$$PAx^{(1)} = Pb \quad \text{and} \quad B_{22} x^{(2)} = Q_2^H b \tag{15}$$

by GMRES. The superscripts (1) and (2) are used to distinguish similar quantities involved in the two solution processes. According to Lemma 4.1, the first equation of (15) can be written as $Q_2 B_{22} Q_2^H x^{(1)} = Q_2 Q_2^H b$.

Let $x_0^{(1)} \in \mathbb{C}^N$ be an initial guess of the first system of (15). The corresponding residual is

$$r_0^{(1)} = Pb - PAx_0^{(1)} = Q_2 Q_2^H b - Q_2 B_{22} Q_2^H x_0^{(1)}.$$

For the second system of (15), we choose $x_0^{(2)} = Q_2^H x_0^{(1)} \in \mathbb{C}^{N-k}$ as its initial guess.

Let us assume that $r_0^{(1)} \neq 0$. Then, $r_0^{(2)} \neq 0$ since $r_0^{(1)} = Q_2 r_0^{(2)}$. In the Arnoldi process associated with the first system of (15), we set $u_1^{(1)} = r_0^{(1)} / \|r_0^{(1)}\|_2$. Then, Eq. (4) becomes

$$PAU_m^{(1)} = U_{m+1}^{(1)} H_{m+1,m}^{(1)}. \tag{16}$$

The approximate solution at the m-th iteration is $x_m^{(1)} = x_0^{(1)} + U_m^{(1)} \eta_m^{(1)}$, where $\eta_m^{(1)}$ minimizes (cf. Eq. (5))

$$\|r_m^{(1)}\|_2 = \| \|r_0^{(1)}\|_2 e_1 - H_{m+1,m}^{(1)} \eta_m^{(1)} \|_2 = \min_{\theta \in \mathbb{C}^m} \| \|r_0^{(1)}\|_2 e_1 - H_{m+1,m}^{(1)} \theta \|_2.$$

On the other hand, for the Arnoldi process associated with the second equation of (15),

we set $u_1^{(2)} = r_0^{(2)}/\|r_0^{(2)}\|_2$. Then (4) becomes

$$B_{22}U_m^{(2)} = U_{m+1}^{(2)}H_{m+1,m}^{(2)}, \tag{17}$$

and the approximate solution at the m-th iteration is

$$x_m^{(2)} = x_0^{(2)} + U_m^{(2)}\eta_m^{(2)},$$

where $\eta_m^{(2)}$ minimizes (cf. Eq. (5))

$$\|r_m^{(2)}\|_2 = \|\|r_0^{(2)}\|_2 e_1 - H_{m+1,m}^{(2)}\eta_m^{(2)}\|_2 = \min_{\theta\in\mathbb{C}^m} \|\|r_0^{(2)}\|_2 e_1 - H_{m+1,m}^{(2)}\theta\|_2.$$

We now rewrite (17) as follows:

$$\begin{array}{rlcl} & B_{22}(Q_2^HQ_2)U_m^{(2)} & = & U_{m+1}^{(2)}H_{m+1,m}^{(2)} \\ \Rightarrow & (Q_2B_{22}Q_2^H)(Q_2U_m^{(2)}) & = & (Q_2U_{m+1}^{(2)})H_{m+1,m}^{(2)} \\ \Rightarrow & PA(Q_2U_m^{(2)}) & = & (Q_2U_{m+1}^{(2)})H_{m+1,m}^{(2)}. \end{array} \tag{18}$$

Since $r_0^{(1)} = Q_2r_0^{(2)}$, the first column of $Q_2U_m^{(2)}$ is

$$Q_2u_1^{(2)} = \frac{Q_2r_0^{(2)}}{\|r_0^{(2)}\|_2} = \frac{Q_2r_0^{(2)}}{\|Q_2r_0^{(2)}\|_2} = \frac{r_0^{(1)}}{\|r_0^{(1)}\|_2} = u_1^{(1)}.$$

Moreover,

$$(Q_2U_m^{(2)})^H(Q_2U_m^{(2)}) = (U_m^{(2)})^HQ_2^HQ_2U_m^{(2)} = (U_m^{(2)})^HU_m^{(2)} = I.$$

Thus, the last equation of (18) is an Arnoldi process applied to PA with starting vector $u_1^{(1)}$. Recall that Eq. (16) is also an Arnoldi process applied to PA with the same starting vector $u_1^{(1)}$. Hence, we have

$$H_{m+1,m}^{(1)} = H_{m+1,m}^{(2)} \quad \text{and} \quad Q_2U_m^{(2)} = U_m^{(1)}, \tag{19}$$

by the uniqueness of the Arnoldi process (see Remark 2.1). Therefore,

$$\begin{array}{rl} \|r_m^{(1)}\|_2 & = \min_{\theta\in\mathbb{C}^m} \|\|r_0^{(1)}\|_2 e_1 - H_{m+1,m}^{(1)}\theta\|_2 \\ & = \min_{\theta\in\mathbb{C}^m} \|\|r_0^{(2)}\|_2 e_1 - H_{m+1,m}^{(2)}\theta\|_2 \\ & = \|r_m^{(2)}\|_2. \end{array} \tag{20}$$

Moreover, since B_{22} is nonsingular, $\text{rank}(H^{(2)}_{m+1,m}) = m$. Therefore, $\text{rank}(H^{(1)}_{m+1,m}) = m$ follows from (19), and

$$\begin{aligned}\eta^{(1)}_m &= [(H^{(1)}_{m+1,m})^H H^{(1)}_{m+1,m}]^{-1}(H^{(1)}_{m+1,m})^H(\|r^{(1)}_0\|_2 e_1)\\ &= [(H^{(2)}_{m+1,m})^H H^{(2)}_{m+1,m}]^{-1}(H^{(2)}_{m+1,m})^H(\|r^{(2)}_0\|_2 e_1)\\ &= \eta^{(2)}_m.\end{aligned}$$

Hence,

$$x^{(2)}_m = x^{(2)}_0 + U^{(2)}_m \eta^{(2)}_m = Q_2^H x^{(1)}_0 + Q_2^H U^{(1)}_m \eta^{(1)}_m = Q_2^H x^{(1)}_m. \tag{21}$$

The above results considering the solution equivalences are summarized in Lemma 4.3.

Remark 4.2 *In the above discussion, we have implicitly assumed that all the entries in the lower sub-diagonals of $H^{(1)}_{m+1,m}$ and $H^{(2)}_{m+1,m}$ are nonnegative. This assumption, however, is not essential for the truth of (20) and (21), namely, (20) and (21) still hold for GMRES built on a general Arnoldi process.*

Lemma 4.3 *Let the two systems in (15) be solved by GMRES with initial guesses $x^{(1)}_0$ and $x^{(2)}_0 = Q_2^H x^{(1)}_0$, respectively. Then, the approximate solutions at the m-th iteration, $x^{(1)}_m$ and $x^{(2)}_m$, satisfy*

$$\|r^{(1)}_m\|_2 = \|r^{(2)}_m\|_2 \quad \textit{and} \quad x^{(2)}_m = Q_2^H x^{(1)}_m.$$

Subsequently, Lemma 4.3 can be used to prove the following theorem. Other side results can be found in Section 5.3.

Theorem 4.4 *Suppose $A \in \mathbb{C}^{N\times N}$ is nonsingular and the deflation matrix $Z \in \mathbb{C}^{N\times k}$ is chosen such that its columns form a basis of an A-invariant subspace. If GMRES is applied to solve the singular system (9), then GMRES always finds a solution to (9) starting with any initial guess.*

Proof. Let $x^{(1)}_0$ be an arbitrary initial guess to the first equation of (15) and set $x^{(2)}_0 = Q_2^H x^{(1)}_0$ for the second equation of (15) . Since the second equation is a nonsingular system, GMRES finds its solution at some iteration t for $t \le N - k$, namely, $r^{(2)}_t = 0$. Therefore, $x^{(1)}_t$ is a solution of the first equation due to Lemma 4.3. ∎

Thus, Theorem 4.4 guarantees that deflated GMRES converges to the solution. This is not a trivial result, see the following remark.

Remark 4.5 *The singular system (9) is always consistent, so there exists at least one solution. However, the consistency of a singular system is generally not a sufficient condition for a Krylov-subspace method to converge to a solution of the system. For example, consider the system*

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

which is clearly consistent. If we select the initial guess $x_0 = [1, 0]^T$*, then the affine subspace (2) would contain no solution of the system for any* m*. Therefore, a Krylov-subspace method would fail to find a solution starting with this initial guess.*

5 Bounds on Deflated GMRES Residuals

This section presents some results on bounds of residuals computed by deflated GMRES. In Section 5.1, those residual bounds are used to show that D-GMRES converges faster than GMRES. Subsequently, residual bounds are provided in Section 5.2 showing that the convergence of D-GMRES is accelerated if the deflation subspace is extended. Finally, we generalize Proposition 2.2 and Corollary 2.3 to D-GMRES in Section 5.3.

5.1 Comparison of GMRES and Deflated GMRES

We consider the solution of Eqs. (1) and (9) by GMRES, and denote the corresponding residuals by r_m and r_m^D, respectively. Let $x_0 \in \mathbb{C}^N$ be an initial guess of the two solution processes. Then, for any scalars $c_1, c_2, \ldots, c_m$, we have

$$Pr_0 + \sum_{k=1}^{m} c_k (PA)^k Pr_0 = Pr_0 + \sum_{k=1}^{m} c_k PA^k r_0 = P\left(r_0 + \sum_{k=1}^{m} c_k A^k r_0 \right), \tag{22}$$

where we have used the fact that $PAP = PA$ due to Lemma 4.1. Let $p_m(\lambda) = 1 + \sum_{k=1}^{m} c_k \lambda^k$. Then, (22) becomes

$$p_m(PA) r_0^D = p_m(PA) Pr_0 = P p_m(A) r_0.$$

Therefore,

$$\begin{aligned} \|r_m^D\|_2 &= \min_{p\in\mathcal{P}_m, p(0)=1} \|p(PA)r_0^D\|_2 \le \|p_m(PA)r_0^D\|_2 = \|Pp_m(A)r_0\|_2 \\ &= \|Q_2Q_2^H p_m(A)r_0\|_2 = \|Q_2^H p_m(A)r_0\|_2. \end{aligned} \tag{23}$$

Moreover, since

$$\begin{aligned} p_m(A)r_0 &= QQ^H p_m(A)r_0 = (Q_1Q_1^H + Q_2Q_2^H)p_m(A)r_0 \\ &= Q_1Q_1^H p_m(A)r_0 + Q_2Q_2^H p_m(A)r_0 \equiv d_1 + d_2, \end{aligned}$$

and $d_1^H d_2 = 0$, we have

$$\|p_m(A)r_0\|_2^2 = \|d_1\|_2^2 + \|d_2\|_2^2 = \|Q_1^H p_m(A)r_0\|_2^2 + \|Q_2^H p_m(A)r_0\|_2^2.$$

Thus, (23) can be written as

$$\|r_m^D\|_2^2 \le \|Q_2^H p_m(A)r_0\|_2^2 = \|p_m(A)r_0\|_2^2 - \|Q_1^H p_m(A)r_0\|_2^2. \tag{24}$$

Note that (24) holds for any $p_m \in \mathcal{P}_m$ with $p_m(0) = 1$. If we choose p_m such that $p_m(A)r_0 = r_m$, then (24) yields

$$\|r_m^D\|_2^2 \le \|Q_2^H r_m\|_2^2 = \|r_m\|_2^2 - \|Q_1^H r_m\|_2^2. \tag{25}$$

Therefore, we have proved the following theorem.

Theorem 5.1 *Suppose $A \in \mathbb{C}^{N\times N}$ is nonsingular and $Z \in \mathbb{C}^{N\times k}$ is a deflation-subspace matrix whose columns form a basis of some A-invariant subspace. Let GMRES be used to solve (1) and (9), where r_m and r_m^D denote the corresponding m-th residual, respectively. Then, starting with the same initial guess, r_m and r_m^D obey (25). In particular, we obtain $\|r_m^D\|_2 \le \|r_m\|_2$ for all $m = 1, 2, \ldots$.*

We note that, for a general choice of Z, Theorem 5.1 is not necessarily true, see the experiment in Section 7.1.3. The theorem only holds for an A-invariant deflation subspace; in this case, deflated GMRES always converges faster than GMRES. This result is quite strong by regarding the fact that spectral properties of the (deflated) coefficient matrix do not necessarily predict the GMRES convergence, see, e.g., [11].

Remark 5.2 *For any polynomial method starting with the same initial guess x_0 and ap-*

plied to solve (1), the inequality (24) indicates that the residual, denoted by r_m^{pol}, computed by the method at the m-th iteration satisfies

$$\|r_m^D\|_2^2 \leq \|Q_2^H r_m^{pol}\|_2^2 = \|r_m^{pol}\|_2^2 - \|Q_1^H r_m^{pol}\|_2^2.$$

In particular,

$$\|r_m^D\|_2^2 \leq \|Q_2^H r_m^{Bi\text{-}CG}\|_2^2 = \|r_m^{Bi\text{-}CG}\|_2^2 - \|Q_1^H r_m^{Bi\text{-}CG}\|_2^2,$$

and

$$\|r_m^D\|_2^2 \leq \|Q_2^H r_{m/2}^{Bi\text{-}CGSTAB}\|_2^2 = \|r_{m/2}^{Bi\text{-}CGSTAB}\|_2^2 - \|Q_1^H r_{m/2}^{Bi\text{-}CGSTAB}\|_2^2,$$

where $r_m^{Bi\text{-}CG}$ is the Bi-CG residual, and $r_{m/2}^{Bi\text{-}CGSTAB}$ is the Bi-CGSTAB residual at iteration $m/2$. In other words, Theorem 5.1 can be easily generalized to Bi-CG and Bi-CGSTAB.

5.2 Deflated GMRES for Different Deflation Subspaces

The argument that leads to (25) also implies a relation between different deflation processes.

Suppose $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are two A-invariant subspaces with $\mathcal{Z}_1 \subseteq \mathcal{Z}_2$. Let $Z_1 = [z_1^{(1)}, \ldots, z_{k_1}^{(1)}]$ be a basis of $\mathcal{Z}_1$ and $Z_2 = [z_1^{(2)}, \ldots, z_{k_2}^{(2)}]$ a basis of $\mathcal{Z}_2$. Form the projectors

$$P^{(1)} = I - AZ_1(Z_1^H AZ_1)^{-1}Z_1^H \quad \text{and} \quad P^{(2)} = I - AZ_2(Z_2^H AZ_2)^{-1}Z_2^H,$$

and consider the solution of

$$P^{(1)}Ax = P^{(1)}b \quad \text{and} \quad P^{(2)}Ax = P^{(2)}b, \tag{26}$$

by GMRES.

Lemma 5.3 shows that the projector $P = I - AZ(Z^H AZ)^{-1}Z^H$ is independent of the choice of a basis of an A-invariant subspace.

Lemma 5.3 *If $\mathcal{Z}_1 = \mathcal{Z}_2$, then $P^{(1)} = P^{(2)}$.*

Proof. Let $\mathcal{Z} = \mathcal{Z}_1 = \mathcal{Z}_2$ and $k = k_1 = k_2$. Then, $\{z_1^{(1)}, \ldots, z_k^{(1)}\}$ and $\{z_1^{(2)}, \ldots, z_k^{(2)}\}$ are two bases of the same space $\mathcal{Z}$, and, therefore, there exists a nonsingular matrix $B \in \mathbb{C}^{k\times k}$ such that $Z_1 = Z_2 B$. Thus,

$$\begin{aligned} P^{(1)} &= I - AZ_1(Z_1^H AZ_1)^{-1}Z_1^H = I - A(Z_2B)\left((Z_2B)^H A(Z_2B)\right)^{-1}(Z_2B)^H \\ &= I - AZ_2(Z_2^H AZ_2)^{-1}Z_2^H = P^{(2)}. \end{aligned}$$

■

We now pick a basis $\{z_1, \ldots, z_{k_1}, \ldots, z_{k_2}, \ldots, z_N\}$ of $\mathbb{C}^N$ with $\mathcal{Z}_1 = \text{span}\{z_1, \ldots, z_{k_1}\}$ and $\mathcal{Z}_2 = \text{span}\{z_1, \ldots, z_{k_2}\}$. Because of Lemma 5.3, it can be assumed without loss of generality that $z_i^{(1)} = z_i$ for $i = 1, 2, \ldots, k_1$ and $z_i^{(2)} = z_i$ for $i = 1, 2, \ldots, k_2$. We perform a QR factorization on the basis matrix:

$$[z_1, \ldots, z_{k_1}, \ldots, z_{k_2}, \ldots, z_N] = QR. \tag{27}$$

Lemma 4.1 then indicates that

$$P^{(1)} = Q_2^{(1)}(Q_2^{(1)})^H \quad \text{and} \quad P^{(2)} = Q_2^{(2)}(Q_2^{(2)})^H,$$

where $Q_2^{(1)}$ and $Q_2^{(2)}$ are the matrices of the last $N-k_1$ and $N-k_2$ columns of Q, respectively. From this, it can be seen that $P^{(2)} = P^{(2)}P^{(1)}$.

Let $c_1, \ldots, c_m$ be any scalars. Similar to (22), we obtain

$$\begin{aligned}
& P^{(2)}r_0 + \sum_{k=1}^{m} c_k (P^{(2)}A)^k P^{(2)} r_0 = P^{(2)}r_0 + \sum_{k=1}^{m} c_k P^{(2)} A^k r_0 \\
= \;& P^{(2)}\left(r_0 + \sum_{k=1}^{m} c_k A^k r_0 \right) = P^{(2)}P^{(1)}\left(r_0 + \sum_{k=1}^{m} c_k A^k r_0 \right) \\
= \;& P^{(2)}\left(P^{(1)}r_0 + \sum_{k=1}^{m} c_k P^{(1)} A^k r_0 \right) = P^{(2)}\left(P^{(1)}r_0 + \sum_{k=1}^{m} c_k (P^{(1)}A)^k P^{(1)} r_0 \right).
\end{aligned}$$

Therefore,

$$p_m(P^{(2)}A)P^{(2)}r_0 = P^{(2)}p_m(P^{(1)}A)P^{(1)}r_0, \tag{28}$$

where $p_m(\lambda) = 1 + \sum_{k=1}^{m} c_k \lambda^k$.

Subsequently, let r_m^{D1} and r_m^{D2} denote the residuals computed at the m-th iteration of GMRES, applied to solve the first and the second system in (26), respectively. Then, (28) implies

$$\begin{aligned}
\|r_m^{D2}\|_2 &= \min_{p \in \mathcal{P}_m, p(0)=1} \|p(P^{(2)}A)r_0^{D2}\|_2 \leq \|p_m(P^{(2)}A)r_0^{D2}\|_2 \\
&= \|P^{(2)}p_m(P^{(1)}A)r_0^{D1}\|_2 = \|(Q_2^{(2)})^H p_m(P^{(1)}A)r_0^{D1}\|_2.
\end{aligned}$$

Starting with the above inequality and using the same argument that leads to (24), we have

$$\|r_m^{D2}\|_2^2 \leq \|p_m(P^{(1)}A)r_0^{D1}\|_2^2 - \|(Q_1^{(2)})^H p_m(P^{(1)}A)r_0^{D1}\|_2^2,$$

where $Q_1^{(2)}$ is the matrix consisting of the first k_2 columns of the Q in (27). Then, by choosing p_m so that $p_m(P^{(1)}A)r_0^{D1} = r_m^{D1}$, this yields

$$\|r_m^{D2}\|_2^2 \leq \|r_m^{D1}\|_2^2 - \|(Q_1^{(2)})^H r_m^{D1}\|_2^2. \tag{29}$$

Thus, we have proved Theorem 5.4 (cf. Theorem 5.1), which states that extending the deflation subspace results in a faster convergence of Deflated GMRES.

Theorem 5.4 *Suppose $A \in \mathbb{C}^{N\times N}$ is nonsingular and $\mathcal{Z}_1, \mathcal{Z}_2$ are two A-invariant subspaces with $\mathcal{Z}_1 \subseteq \mathcal{Z}_2$. Let $Z_1 \in \mathbb{C}^{N\times k_1}$ and $Z_2 \in \mathbb{C}^{N\times k_2}$ be chosen so that their columns form a basis of $\mathcal{Z}_1$ and a basis of $\mathcal{Z}_2$, respectively. Suppose GMRES is used to solve the systems in (26). Then, by starting with the same initial guess, the residuals r_m^{D1} and r_m^{D2} satisfy (29). In particular, $\|r_m^{D2}\|_2 \leq \|r_m^{D1}\|_2$ holds for all $m = 1, 2, \ldots$.*

5.3 Generalization of GMRES Results

We aim at extending Proposition 2.2 and Corollary 2.3 from GMRES to deflated GMRES. To do so, we need the following assumption, which holds throughout this subsection.

Assumption 5.5 *We assume that the nonsingular $A \in \mathbb{C}^{N\times N}$ has a spectral decomposition (6) with $V = [v_1, \ldots, v_N]$ and $\Lambda = diag\{\lambda_1, \ldots, \lambda_N\}$. Moreover, we choose $Z = [v_1, \ldots, v_k] \in \mathbb{C}^{N\times k}$, perform a QR factorization on the matrix $V = [v_1, \ldots, v_k, v_{k+1}, \ldots, v_N] \equiv [Z, \widetilde{Z}]$ as in (12), and set $W = Q_2$.*

Thus, the columns of the matrix Z form a basis of the A-invariant subspace spanned by the eigenvectors $\{v_1, \ldots, v_k\}$.

Lemma 5.6 *Under Assumption 5.5, the matrix B_{22} from Proposition 3.2 satisfies*

$$B_{22} = R_{22}\Lambda_2 R_{22}^{-1},$$

where $\Lambda_2 = diag\{\lambda_{k+1}, \ldots, \lambda_N\}$, and matrices B_{22} and R_{22} are the same as in Eqs. (10) and (12), respectively.

Proof. Right-multiplying (14) by $[Q_1, Q_2]$ yields $AQ_2 = Q_1R_{11}B_{12} + Q_2B_{22}$. This implies

$$B_{22} = Q_2^H Q_2 B_{22} = Q_2^H(AQ_2 - Q_1R_{11}B_{12}) = Q_2^H AQ_2. \tag{30}$$

On the other hand, combining (11) and (13) yields

$$\begin{aligned} AQ_2R_{22} &= A(\widetilde{Z} - Q_1R_{12}) = A\widetilde{Z} - AQ_1R_{12} = A\widetilde{Z} - AZR_{11}^{-1}R_{12} \\ &= \widetilde{Z}\Lambda_2 - ZB_{11}R_{11}^{-1}R_{12} = (Q_1R_{12} + Q_2R_{22})\Lambda_2 - (Q_1R_{11})B_{11}R_{11}^{-1}R_{12}, \end{aligned} \tag{31}$$

where we have used

$$A\widetilde{Z} = A[v_{k+1}, \ldots, v_N] = [v_{k+1}, \ldots, v_N]\Lambda_2 = \widetilde{Z}\Lambda_2.$$

Left-multiplying (31) by Q_2^H leads to

$$Q_2^H AQ_2R_{22} = R_{22}\Lambda_2. \tag{32}$$

By combining Eqs. (30) and (32), the lemma follows immediately. ■

In Section 4, we have proved the solution equivalence (20) and (21), when GMRES is used to solve the two systems in (15). Subsequently, by applying Proposition 2.2 and Lemma 5.6 to the second equation of (15), we obtain the following result.

Theorem 5.7 *Suppose that GMRES is used to solve Eq. (9). Then, under Assumption 5.5, r_m^D obeys*

$$\|r_m^D\|_2 \leq \kappa_2(R_{22})\epsilon_m^{(2)}\|r_0^D\|_2, \tag{33}$$

where $\epsilon_m^{(2)} = \min\limits_{p\in\mathcal{P}_m, p(0)=1} \max\limits_{\lambda\in\{\lambda_{k+1},\ldots,\lambda_N\}} |p(\lambda)|$.

Similarly, by applying Corollary 2.3 and Lemma 5.6 to the second equation of (15), we obtain the following result.

Corollary 5.8 *Under the assumptions of Theorem 5.7, if all the eigenvalues $\lambda_{k+1}, \ldots, \lambda_N$ of A are located in an ellipse, $E(c, d, a)$, that excludes the origin of the complex plane, then*

$$\|r_m^D\|_2 \leq \kappa_2(R_{22})\frac{C_m(\frac{a}{d})}{C_m(\frac{c}{d})}\|r_0^D\|_2 \approx \kappa_2(R_{22})\delta^m\|r_0^D\|_2, \tag{34}$$

where $\delta = \frac{a+\sqrt{a^2-d^2}}{c+\sqrt{c^2-d^2}}$.

Thus, the upper bound of the residual norm of deflated GMRES is determined by the condition number of R_{22} (rather than V), and the scalar δ determined by the distribution

of the undeflated eigenvalues. Furthermore, we note that the ellipse $E(c, d, a)$ as used in Corollary 5.8 is smaller than that of Corollary 2.3 for the same A. Additionally, we notice that it depends on the exact eigenvalue distribution of A which eigenvalues of A can be deflated best. This is further illustrated in Section 7.

6 Solving Real Linear Systems by Deflated GMRES

The results presented in Sections 4, 5.1, and 5.2 remain unchanged in real arithmetic. In this subsection, we focus on the results of Section 5.3, and show how to choose Z from complex eigenvectors such that all computations remain in real arithmetic while the effectiveness of deflation does not change.

Suppose A has a spectral decomposition $A = \widetilde{V}\widetilde{\Lambda}\widetilde{V}^{-1}$, where $\widetilde{V}$ and $\widetilde{\Lambda}$ are $N \times N$ complex matrices. Since A is real, its eigenvalues and eigenvectors appear in conjugate pairs. Therefore, with appropriate permutations, we can express $\widetilde{\Lambda}$ and $\widetilde{V}$ as follows:

$$\widetilde{\Lambda} = \mathrm{diag}\{\lambda_1, \bar{\lambda}_1, \ldots, \lambda_l, \bar{\lambda}_l, \mu_1, \ldots, \mu_{N-2l}\},$$

and

$$\widetilde{V} = [v_1, \bar{v}_1, \ldots, v_l, \bar{v}_l, w_1, \ldots, w_{N-2l}],$$

where $\mu_1, \ldots, \mu_{N-2l}$ and $w_1, \ldots, w_{N-2l}$ are real, and the overbar denotes complex conjugation.

Now, suppose that we want to remove

$$\lambda_1, \bar{\lambda}_1, \ldots, \lambda_i, \bar{\lambda}_i, \mu_1, \ldots, \mu_j, \quad 2i + j = k, \tag{35}$$

from $\sigma(A)$. Further permutations allow us to rearrange these eigenvalues and the associated eigenvectors such that they come first in $\widetilde{\Lambda}$ and $\widetilde{V}$, i.e.,

$$\widetilde{\Lambda} = \mathrm{diag}\{\lambda_1, \bar{\lambda}_1, \ldots, \lambda_i, \bar{\lambda}_i, \mu_1, \ldots, \mu_j, \lambda_{i+1}, \bar{\lambda}_{i+1}, \ldots, \lambda_l, \bar{\lambda}_l, \mu_{j+1}, \ldots, \mu_{N-2l}\},$$

and

$$\widetilde{V} = [v_1, \bar{v}_1, \ldots, v_i, \bar{v}_i, w_1, \ldots, w_j, v_{i+1}, \bar{v}_{i+1}, \ldots, v_l, \bar{v}_l, w_{j+1}, \ldots, w_{N-2l}].$$

In order to remove the desired eigenvalues (35) while keeping all the quantities real, we want to choose an appropriate deflation-subspace matrix $Z \in \mathbb{R}^{N\times k}$ in (8). To that end,

we set

$$K_t = \begin{bmatrix} \lambda_t^{\text{real}} & \lambda_t^{\text{imag}} \\ -\lambda_t^{\text{imag}} & \lambda_t^{\text{real}} \end{bmatrix}, \quad F = \begin{bmatrix} 1 & \sqrt{-1} \\ \sqrt{-1} & 1 \end{bmatrix}, \quad G = \frac{1}{2}\begin{bmatrix} 1 & -\sqrt{-1} \\ 1 & \sqrt{-1} \end{bmatrix},$$

where the superscripts 'real' and 'imag' denote the real and imaginary parts of a complex quantity, respectively. It is then straightforward to verify that

$$K_t = F \begin{bmatrix} \lambda_t & 0 \\ 0 & \bar{\lambda}_t \end{bmatrix} F^{-1}, \quad [v_t^{\text{real}}, v_t^{\text{imag}}] = [v_t, \bar{v}_t]G.$$

Therefore, if we set

$$\begin{aligned}
\Lambda &= \text{diag}\{K_1, \ldots, K_i, \mu_1, \ldots, \mu_j, K_{i+1}, \ldots, K_l, \mu_{j+1}, \ldots, \mu_{N-2l}\} \equiv \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}; \\
V &= [v_1^{\text{real}}, v_1^{\text{imag}}, \ldots, v_i^{\text{real}}, v_i^{\text{imag}}, w_1, \ldots, w_j, v_{i+1}^{\text{real}}, v_{i+1}^{\text{imag}}, \ldots, v_l^{\text{real}}, v_l^{\text{imag}}, w_{j+1}, \ldots, w_{N-2l}] \\
&\equiv [V_1, V_2]; \\
\Gamma &= \text{diag}\{\underbrace{F, \ldots, F}_{i}, \underbrace{1, \ldots, 1}_{j}, \underbrace{F, \ldots, F}_{l-i}, \underbrace{1, \ldots, 1}_{N-2l-j}\} \equiv \begin{bmatrix} \Gamma_1 & 0 \\ 0 & \Gamma_2 \end{bmatrix}; \\
\Omega &= \text{diag}\{\underbrace{G, \ldots, G}_{i}, \underbrace{1, \ldots, 1}_{j}, \underbrace{G, \ldots, G}_{l-i}, \underbrace{1, \ldots, 1}_{N-2l-j}\},
\end{aligned}$$

where $\Lambda_1 \in \mathbb{R}^{k\times k}$, $V_1 \in \mathbb{R}^{N\times k}$ and $\Gamma_1 \in \mathbb{C}^{k\times k}$, we obtain

$$\Lambda = \Gamma\widetilde{\Lambda}\Gamma^{-1}, \quad V = \widetilde{V}\Omega,$$

and

$$A = \widetilde{V}\widetilde{\Lambda}\widetilde{V}^{-1} = (V\Omega^{-1})(\Gamma^{-1}\Lambda\Gamma)(V\Omega^{-1})^{-1} = V(\Gamma\Omega)^{-1}\Lambda(\Gamma\Omega)V^{-1} = V\Lambda V^{-1}, \tag{36}$$

since $(\Gamma\Omega)^{-1}\Lambda(\Gamma\Omega) = \Lambda$.

Equation (36) is a spectral decomposition of A in $\mathbb{R}^{N\times N}$. By setting $Z = V_1$, we obtain the result that $\sigma(PA)$ does not contain the eigenvalues (35) according to Proposition 3.2.

Moreover, the same procedures of derivation in real arithmetic can be used to derive the results as presented in Sections 4 and 5. All results remain true with *real* quantities $B_{22}, R_{22}, \Lambda_2, r_0^D$ and r_m^D, except that $\kappa_2(R_{22})$ in both (33) and (34) should be replaced by $\kappa_2(R_{22}\Gamma_2)$. This is because (33) and (34) were obtained by applying Proposition 2.2 and

Corollary 2.3 to the second system of (15), and we now have

$$B_{22} = R_{22}\Lambda_2 R_{22}^{-1} = (R_{22}\Gamma_2)\widetilde{\Lambda}_2(R_{22}\Gamma_2)^{-1},$$

where $\widetilde{\Lambda}_2$ is the $(N-k)\times(N-k)$ lower-diagonal block of $\widetilde{\Lambda}$.

7 Numerical Experiments

To illustrate the theoretical results as presented in the previous sections, we perform some numerical experiments in which the performance of GMRES and deflated GMRES (D-GMRES) is tested. The computations are carried out in MATLAB 7.4.0 on a sequential LINUX machine (Dell Precision T5500 with a Quad-core Intel Xeon 5500 series processor and 4 GB memory).

Our main application is a variant of the 2-D convection-diffusion-reaction equation, i.e.,

$$-u_{xx} - u_{yy} + \alpha(u_x + u_y) - \beta u = f,$$

with homogeneous Dirichlet boundary conditions and

$$x, y \in [0,1]^2, \quad u = u(x,y), \quad f = f(x,y) = 1 + \sin(\pi x)\sin(\pi y), \quad \alpha \geq 0, \quad \beta \in \mathbb{C}.$$

The equation is discretized by a standard second-order finite-difference scheme on a uniform Cartesian grid, where central discretization is used for the first-order derivatives, and N_x and N_y grid points are chosen in the x- and y-direction, respectively. The resulting matrix is $A \in \mathbb{C}^{N\times N}$ with $N = N_x N_y$, and is non-Hermitian and unsymmetric if $\alpha, \beta \neq 0$. In the experiments, we fix $N_x = N_y = 20$, while parameters α and β are varied to control the symmetry and the definiteness of A, respectively.

We consider the linear system $Ax = b$ for various classes of matrices:

(a) real and positive-definite A (i.e., $\beta = 0$);

(b) real and indefinite A (i.e., a real $\beta > 0$);

(c) complex and indefinite A (i.e., a complex β with a positive real and imaginary part).

For both GMRES and D-GMRES, we choose a random initial guess, no preconditioner, and a relative termination tolerance of 10^{-8}. In the experiments, the eigenvalues and

eigenvectors of A are explicitly computed via the MATLAB command `[V,D]=eig(A)`. Matrix V is the eigenvector-matrix. In most experiments, we can write $V = [Z, \widetilde{Z}]$, where $Z \in \mathbb{C}^{N\times k}$ is the deflation-subspace matrix containing k eigenvectors of A. The value of k is varied in the experiments. Notice that, for the sake of convenience, eigenvectors are not divided into their real and imaginary parts for the deflation vectors as described in Section 6. Furthermore, we measure the exact residuals (rather than the GMRES-generated residuals) in the experiments, i.e., $r_m = b - Ax_m$ for $m = 1, 2 \ldots$ are analyzed.

We emphasize that the results of the experiments are mainly meant to illustrate the theoretical results rather than to come up with an efficient solver. Therefore, we do not present results in terms of computational cost, but in terms of number of iterations, residual convergence and spectral plots. We note that, for large and realistic problems, a good preconditioner is required to reduce the number of iterations, approximations of eigenvectors should be used, and attention should be paid to an efficient implementation of D-GMRES, see, e.g., [5, 26].

7.1 Experiments with a Real and Positive-Definite Matrix

In the first experiment, we set $\beta = 0$ and vary the value of α. The resulting matrix, A, is real and positive definite.

7.1.1 Deflation of the Smallest Eigenvalues

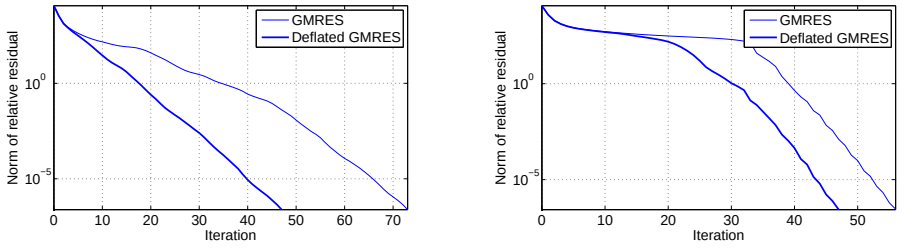
We perform the experiment with $k = 10$ deflation vectors, which are equal to eigenvectors associated with eigenvalues of A that are the smallest in magnitude. The results can be found in Table 1, Figure 1 and Figure 2. In the table, we measure the condition numbers of A, V and Z^HAZ, which are denoted by $\kappa(A)$, $\kappa(V)$ and $\kappa(Z^HAZ)$, respectively. Moreover, the quantity $\frac{||A-A^T||_2}{||A||_2}$ measures the symmetry of A. In addition, the quantity $\frac{||Z^T\widetilde{Z}||_2}{||\widetilde{Z}||_2}$ measures the orthogonality of Z with respect to $\widetilde{Z}$. Finally, we can perform a QR factorization on V as in (12). The condition number of block R_{22} of the matrix R is measured, which gives some insights into the residuals, see Theorem 5.7.

We can see in Figure 2 that the imaginary parts of the eigenvalues of A become larger for an increasing α. Moreover, as can be observed in Table 1, a larger α also yields the following:

- A becomes better conditioned, while $\kappa(V)$ and $\kappa(Z^HAZ)$ becomes worse conditioned;

α	$\kappa(A)$	$\kappa(V)$	$\kappa(Z^H AZ)$	$\kappa(R_{22})$	$\frac{\|Z^T\widetilde{Z}\|_2}{\|Z\|_2}$	$\frac{\|A-A^T\|_2}{\|A\|_2}$	GMRES	D-GMRES	Benefit
0	1.8E2	1.0	8.3	1.0	0.0	0.00	73	47	36%
1	1.8E2	27	11	1.6	0.3	0.02	72	47	35%
5	1.5E2	1.2E3	2.8E2	8.6	1.3	0.12	72	49	32%
10	1.1E2	2.3E4	1.4E4	60	2.6	0.24	68	50	26%
20	68	8.4E8	9.5E6	3.1E3	5.5	0.47	57	48	16%

Table 1: Results for $k = 10$, $\beta = 0$, and various values of α. The matrix Z consists of eigenvectors associated with the smallest eigenvalues of A. The columns associated with 'GMRES' and 'D-GMRES' present the numbers of required iterations for convergence. The 'Benefit' denotes the improvement of using D-GMRES instead of GMRES in terms of iterations.

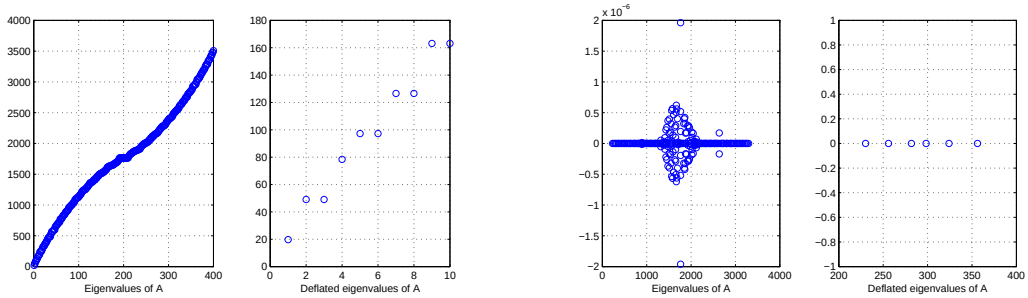


(a) $\alpha = 0$. (b) $\alpha = 20$.

Figure 1: Residual plots of GMRES and D-GMRES for $k = 10$, $\beta = 0$, and various values of α. Deflation vectors are eigenvectors corresponding to the smallest eigenvalues of A in magnitude.

- $\kappa(R_{22})$ grows, so the residuals as given in Eqs. (33) and (34) are bounded by larger values;
- the space spanned by the columns of Z becomes less orthogonal to that of $\widetilde{Z}$;
- A becomes more unsymmetric;
- GMRES requires fewer iterations;
- the benefit of D-GMRES decreases compared to GMRES.

Those observations are in agreement with Theorem 5.7 and Corollary 5.8. Due to an increasing $\kappa(R_{22})$, the bound in both Eqs. (33) and (34) increases correspondingly. Therefore, the improvement of D-GMRES with respect to GMRES becomes less significant as the symmetry of A decreases.



(a) $\alpha = 0$ (eigenvalues are depicted in the real plane).

(b) $\alpha = 20$ (eigenvalues are depicted in the complex plane). Note that some eigenvalues are symmetric about the x-axis, as the complex ones appear in conjugate pairs.

Figure 2: Spectral properties of GMRES and D-GMRES for $k = 10$, $\beta = 0$, and various values of α. Deflation vectors are eigenvectors corresponding to the smallest eigenvalues of A in magnitude.

Furthermore, in Figure 1, we observe that all residuals of D-GMRES are smaller than those of GMRES for all iterations. This is in conformation with Theorem 5.1.

To conclude this subsection, we give the results of D-GMRES for a various number of deflation vectors, see Figure 3. As can be seen in this figure, D-GMRES converges faster as k grows. This statement is even stronger: the residuals for D-GMRES with s deflation vectors are equal or greater than those residuals for D-GMRES with more than s deflation vectors. This is in agreement with Theorem 5.4.

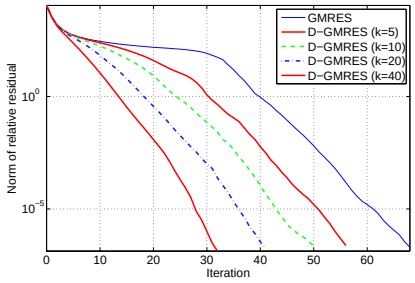


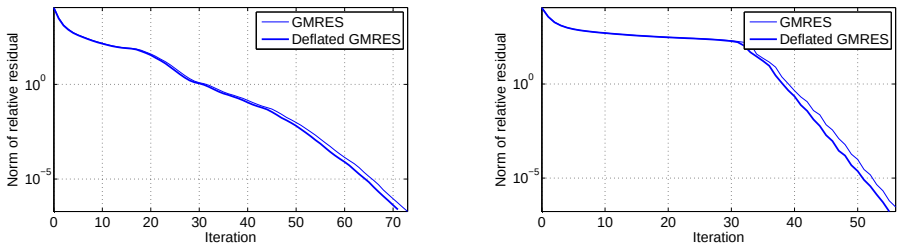
Figure 3: Residual plots of GMRES and D-GMRES for $\alpha = 10$, $\beta = 0$, and various values of k. Deflation vectors are eigenvectors corresponding to the smallest eigenvalues of A in magnitude.

7.1.2 Deflation of Other Eigenvalues

We test the performance of D-GMRES in which the largest or middle, instead of the smallest, eigenvalues in magnitude are deflated. The convergence results of this experiment can be found in Figures 4 and 5.

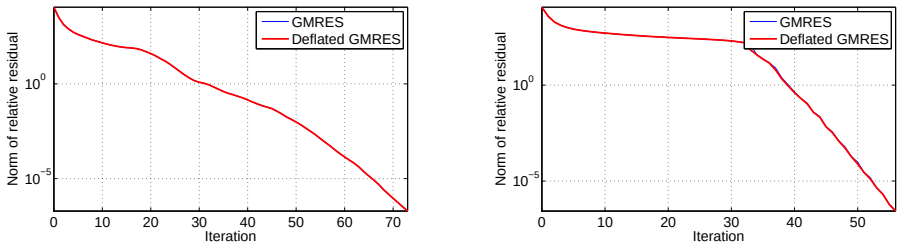
From the figures, we observe that GMRES and D-GMRES show a similar convergence behavior. Hence, deflating eigenvalues other than the smallest ones does not significantly improve the convergence of GMRES for this test case. Note that the residuals of D-GMRES are equal or slightly smaller than those of GMRES, in agreement of Theorem 5.1.

The fact that deflating the smallest eigenvalues of A has the best performance could be explained by examining the parameter δ as used in Eq. (34). Recall that $\delta = \frac{a+\sqrt{a^2-d^2}}{c+\sqrt{c^2-d^2}}$ where a, c, d are derived from the ellipse $E(c, d, a)$ in which all the eigenvalues of PA are located. If this $E(c, d, a)$ is chosen to be a circle with center c (on the x-axis) and radius r, then $\delta = \frac{r}{c}$. The eigenvalues of A in this experiment are all nearly real. Let $\lambda_1, \ldots, \lambda_N$ be the eigenvalues of A in an increasing order of magnitude. Then, we find that the circle centered at $c_1 = \frac{\lambda_{k+1}+\lambda_N}{2}$ with radius $r_1 = \frac{\lambda_N-\lambda_{k+1}}{2}$ contains all of the eigenvalues $\lambda_{k+1}, \ldots, \lambda_N$. Similarly, if the largest k eigenvalues are deleted, the circle centered at $c_2 = \frac{\lambda_1+\lambda_{N-k}}{2}$ with radius $r_2 = \frac{\lambda_{N-k}-\lambda_1}{2}$ contains the remaining eigenvalues $\lambda_1, \ldots, \lambda_{N-k}$. In this experiment, it happens that $\frac{r_1}{c_1} < \frac{r_2}{c_2}$, so the δ associated with deflating the smallest eigenvalues is smaller than the δ associated with deflated the largest eigenvalues. In addition, deflating interior eigenvalues of A does obviously not change δ. Hence, it can be motivated by inequality (34) that deflating the smallest eigenvalues results in a faster convergence of D-GMRES compared to deflating the middle or largest eigenvalues of A for this specific test case.



(a) $\alpha = 0$. (b) $\alpha = 20$.

Figure 4: Residual plots of GMRES and D-GMRES for $k = 10$, $\beta = 0$, and various values of α. Deflation vectors are eigenvectors of A corresponding to the largest eigenvalues in magnitude.



(a) $\alpha = 0$. (b) $\alpha = 20$.

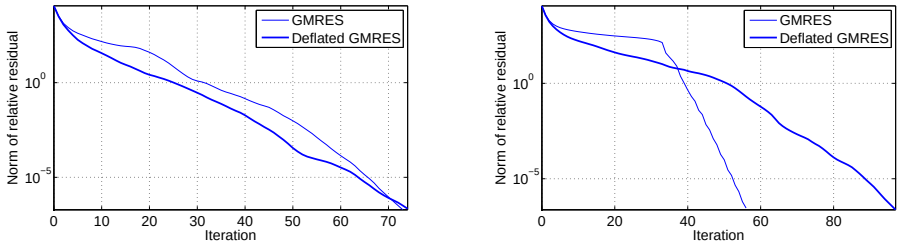
Figure 5: Residual plots of GMRES and D-GMRES for $k = 10$, $\beta = 0$, and various values of α. Deflation vectors are eigenvectors of A associated with the eigenvalues in the center of the spectrum.

7.1.3 Deflation with General Vectors

In the next experiment, we use random vectors as deflation vectors in D-GMRES. The theory, as presented in the previous sections, is not valid for this case, but it is interesting to see how D-GMRES performs. For *symmetric* matrices, it is known that deflation with general vectors would not harm the convergence of the iterative process, see [22, Section 2.3]. In the next experiment, we test if this is also the case for nonsymmetric matrices.

We take the same parameter set as above (i.e., $k = 10$, $\beta = 0$, and various values of α). The convergence results of GMRES and D-GMRES can be found in Figure 6.

From Figure 6, it can be observed that for a symmetric A, GMRES and D-GMRES show a similar convergence behavior, while D-GMRES is significantly slower than GMRES for a (strongly) unsymmetric A. Hence, the theory for symmetric matrices, as provided in [22], does not apply to unsymmetric matrices.



(a) $\alpha = 0$. (b) $\alpha = 20$.

Figure 6: Residual plots of GMRES and D-GMRES for $k = 10$, $\beta = 0$, and various values of α. Random vectors are used as deflation vectors in D-GMRES.

7.2 Experiment with a Real and Indefinite Matrix

In the next experiment, we fix all parameters in the problem ($\alpha = 10$, and $k = 20$), and set $\beta = 500$. Choosing a positive β implies a shift of the eigenvalues of A towards the left half-plane, resulting in a real and indefinite matrix A. The corresponding eigenvalues of A can be found in Figure 7. Different choices of deflation vectors are examined; we consider eigenvectors of A associated with

1. the smallest eigenvalues in absolute sense;
2. the eigenvalues with the largest negative real parts;
3. the eigenvalues with the largest positive real parts;
4. the largest eigenvalues in absolute sense.

For this specific test case, Choice 3 and 4 are the same. The results for GMRES and D-GMRES can be found in Table 2 and Figure 8.

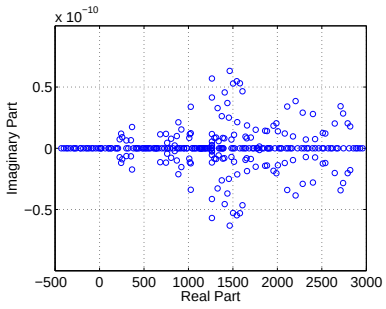


Figure 7: Eigenvalues of A corresponding to $\alpha = 10$ and $\beta = 500$.

Method	Iterations	Benefit
GMRES	176	–
D-GMRES with Choice 1	109	38.1%
D-GMRES with Choice 2	128	27.3%
D-GMRES with Choice 3 / 4	165	6.3%

Table 2: Results for GMRES and D-GMRES (with Choices 1, 2, 3, and 4 for the deflation vectors) for the test case of $\alpha = 10$, $\beta = 500$, and $k = 20$.

From the table and figure, we observe that D-GMRES based on eigenvectors associated with the smallest eigenvalues in absolute sense is the best choice. In this case, the deflated

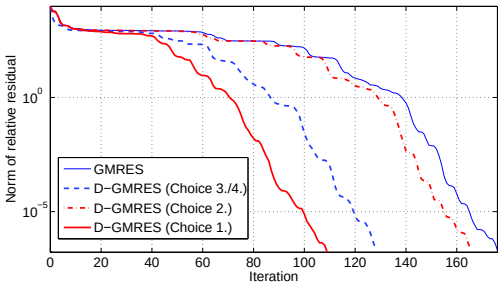


Figure 8: Residual plots for GMRES and D-GMRES (with Choices 1, 2, 3, and 4 for the deflation vectors) for the test case of $\alpha = 10$, $\beta = 500$, and $k = 20$.

matrix consists of two spectral clusters away from zero, and this seems to be more favorable than reducing the size of one spectral cluster, which happens when deflation vectors are taken to be eigenvectors corresponding to eigenvalues with the largest negative or positive real parts or are largest in absolute sense.

The effectiveness of deflating eigenvalues from the interior of the spectrum of A looks counterintuitive by considering Corollary 5.8. However, note that this corollary cannot be applied to motivate this result, as the eigenvalues of A do not lie in the same half plane, Therefore, one cannot find an ellipse $E(c, d, a)$ that contains all the desired eigenvalues while the origin is excluded.

7.3 Experiment with a Complex and Indefinite Matrix

In the next experiment, we take the same parameters as above ($\alpha = 10$ and $k = 20$), and set $\beta = 500(1 + i)$, so that A is a complex matrix. Then, the imaginary parts of the eigenvalues of A are approximately constant (around -500), while the real parts vary between -500 and 3000.

We again investigate the four different choices of deflation vectors, as done in Subsection 7.2. The results of the experiment can be found in Table 3 and Figure 9.

Method	Iterations	Benefit
GMRES	64	–
D-GMRES with Choice 1	61	4.7%
D-GMRES with Choice 2	51	20.3%
D-GMRES with Choice 3 /4	61	4.7%

Table 3: Results for GMRES and D-GMRES with different choices for the deflation vectors for the test case of a complex A with $\alpha = 10$, $\beta = 500(1 + i)$, and $k = 20$.

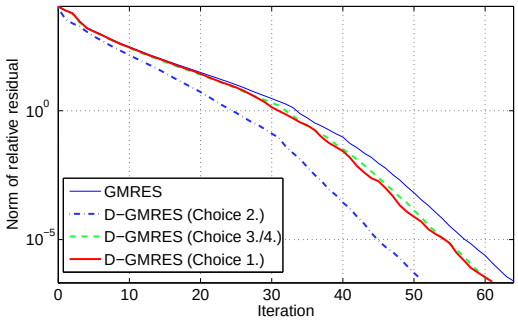


Figure 9: Residual plots for GMRES and D-GMRES (with Choices 1, 2, 3, and 4 for the deflation vectors) for the test case of $\alpha = 10$, $\beta = 500(1 + i)$, and $k = 20$.

From the table and figure, we can observe that D-GMRES based on eigenvectors associated with the eigenvalues with largest negative parts is the best choice. Note that this is a different observation compared to the case in which A is real, see Table 2. Hence, when A is indefinite, the choice of the eigenvalues to be deflated depends on the exact eigenvalue distribution of A.

Moreover, we note that a similar experiment as presented in Section 7.1.1 can be performed here, where the performance of D-GMRES is examined by varying the number of deflation vectors. As the results are similar to Figure 3 (and, therefore, obey Theorem 5.4), they are omitted here.

7.4 Stability of D-GMRES

In theory, D-GMRES always converges faster than GMRES when eigenvectors are used as deflation vectors, see Theorem 5.1. However, there are cases where D-GMRES is less effective than GMRES, as computations are done in finite precision. An example is presented below.

We consider the same problem setting as in the previous subsection (i.e., $\beta = 500(1+i)$ and $k = 20$), but we now use $\alpha = 20$. We apply D-GMRES where the deflation vectors are eigenvectors associated with the eigenvalues with the largest negative real parts. We show that D-GMRES fails for this specific test case, see Figure 10.

As can be observed in the figure, D-GMRES stagnates around 1E-4. This phenomenon may be explained by the fact that A is strongly nonnormal in this test case. Consider the

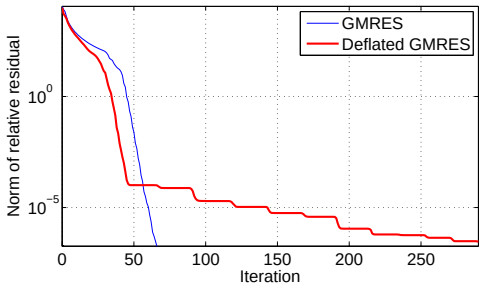


Figure 10: Residual plots of GMRES and D-GMRES for a complex A and parameters $\alpha = 20$, $\beta = 500(1+i)$, and $k = 20$.

condition numbers of the corresponding matrices A, V, and $Z^H AZ$:

$$\kappa(A) = 129, \quad \kappa(V) = 5.8\text{E}8, \quad \kappa(Z^H AZ) = 1.7\text{E}10.$$

As V is ill-conditioned, $Z^H AZ$ is also ill-conditioned. This results in the fact that the deflation matrix, as given in Eq. (8), cannot be constructed accurately in machine precision. Therefore, the projection of eigenvalues in D-GMRES can also not be performed accurately; eigenvalues are not projected to exactly zero, but close to zero. This problem could be resolved by stabilizing the deflation operator by projecting eigenvalues to a value in a cluster of the spectrum of A rather than to zero, so that perturbations of these deflated eigenvalues would not harm the convergence of the iterative process, see also [33].

8 Conclusions

In this paper, the convergence of Deflated GMRES is analyzed, where the deflation vectors span an A-invariant subspace. We deduce that the deflated eigenvalues are shifted to zero, whereas the other eigenvalues are unchanged. Then, by proving that deflated GMRES is equivalent to GMRES applied to solve a reduced linear system, we derive that deflated GMRES does not break down in exact arithmetic. Subsequently, we prove that the norm of the residuals of deflated GMRES are always below the norm of the residuals of GMRES. Hence, deflated GMRES always converges faster than GMRES. Furthermore, a monotonicity property of deflated GMRES is obtained: extending the deflation subspace leads to a faster convergence. In addition, bounds of residual norms that are valid for GMRES are generalized to deflated GMRES. Thereafter, for real-valued systems with possibly complex eigenvalues and eigenvectors, an analysis is provided to show how the computations in

deflated GMRES can be kept in real arithmetic, while the theoretical results still remain valid. Finally, numerical experiments are presented to illustrate the theoretical results. In addition, those experiments show that deflated GMRES might break down when the eigenvector-matrix is ill-conditioned, and deflated GMRES with general deflation vectors does not necessarily work well.

This paper provides some more fundamental insights into the theory of deflated GMRES. Future research should include the development of the theory for practical deflated GMRES variants: restarted and truncated GMRES with deflation can be examined, and deflation based on nearly A-invariant subspaces can be further explored.

References

[1] W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math.*, 9:17–29, 1951.

[2] J. Baglama, D. Calvetti, G. H. Golub, and L. Reichel. Adaptively preconditioned GMRES algorithms. *SIAM J. Sci. Comput.*, 20(1):243–269, 1999.

[3] K. Burrage, J. Erhel, B. Pohl, and A. Williams. A deflation technique for linear systems of equations. *SIAM J. Sci. Comput.*, 19:1245–1260, 1998.

[4] A. Chapman and Y. Saad. Deflated and augmented Krylov subspace techniques. *Numer. Linear Algebra Appl.*, 4(1):43–66, 1997.

[5] D. Darnell, R.B. Morgan, and W. Wilcox. Deflated GMRES for systems with multiple shifts and multiple right-hand sides. *Linear Algebra and its Applications*, 429:2415–2434, 2008.

[6] R.M. Dinkla. GMRES(m) with deflation applied to nonsymmetric systems arising from fluid mechanics problems. Master's thesis, Delft University of Technology, Delft, The Netherlands, 2009.

[7] Z. Dóstal. Conjugate gradient method with preconditioning by projector. *Int. J. Comput. Math.*, 23:315–323, 1988.

[8] S.C. Eisenstat, H.C. Elman, and M.H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 20 (2):345–357, 1983.

[9] J. Erhel, K. Burrage, and B. Pohl. Restarted GMRES preconditioned by deflation. *J. Comput. Appl. Math.*, 69(2):303–318, 1996.

[10] J. Frank and C. Vuik. On the construction of deflation-based preconditioners. *SIAM Journal on Scientific Computing*, 23:442–462, 2001.

[11] A. Greenbaum, V. Ptak, and Z. Strakos. Any nonincreasing convergence curve is possible for GMRES. *SIAM J. Matrix Anal. Appl.*, 17:465–469, 1996.

[12] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.*, 49:409–436, 1952.

[13] A. Jennings. Influence of the eigenvalue spectrum on the convergence rate of the conjugate gradient method. *IMA Journal*, 20:61–72, 1977.

[14] S. Kaniel. Estimates for some computational techniques in linear algebra. *Math. Comp.*, 20:369–378, 1966.

[15] S. A. Kharchenko and A. Yu. Yeremin. Eigenvalue translation based preconditioners for the GMRES(k) method. *Numer. Linear Algebra Appl.*, 2(1):51–77, 1995.

[16] M.E. Kilmer and E. de Sturler. Recycling subspace information for diffuse optical tomography. *SIAM J. Sci. Comput.*, 27(6):2140–2166, 2006.

[17] S.P. MacLachlan, J.M. Tang, and C. Vuik. Fast and robust solvers for pressure correction in bubbly flow problems. *Journal of Computational Physics*, 227:9742–9761, 2008.

[18] T.A. Manteuffel. The Tchebychew iteration for nonsymmetric linear systems. *Num. Math.*, 28:307–327, 1977.

[19] J.A. Meijerink and H.A. van der Vorst. An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Mathematics of Computation*, 31:148–162, 1977.

[20] G. Meinardus. Über eine Verallgemeinerung einer Ungleichung von L. V. Kantorowitsch. *Numer. Math.*, 5:14–23, 1963.

[21] R.B. Morgan. A restarted GMRES method augmented with eigenvectors. *SIAM J. Matrix Anal. Appl.*, 16:1154–1171, 1995.

[22] R. Nabben and C. Vuik. A comparison of Deflation and Coarse Grid Correction applied to porous media flow. *SIAM J. Numer. Anal.*, 42:1631–1647, 2004.

[23] R. Nabben and C. Vuik. A comparison of abstract versions of deflation, balancing and additive coarse grid correction preconditioners. *Numer. Linear Algebra Appl.*, 15:355–372, 2008.

[24] R.A. Nicolaides. Deflation of conjugate gradients with applications to boundary value problems. *SIAM J. Numer. Anal.*, 24:355–365, 1987.

[25] C.C. Paige and M.A. Saunders. LSQR : An algorithm for sparse linear equations and sparse least squares. *A.C.M. Trans. Math. Softw.*, 8:43–71, 1982.

[26] M.L. Parks, E. de Sturler, G. Mackey, D.D. Johnson, and S. Maiti. Recycling Krylov subspaces for sequences of linear systems. *SIAM J. Scientific Computing*, 28(5):1651–1674, 2006.

[27] Y. Saad. Analysis of augmented Krylov subspace methods. *SIAM J. Math. Anal. Appl.*, 18(2):435–449, 1997.

[28] Y. Saad. *Iterative methods for sparse linear systems, Second Edition*. SIAM, Philadelphia, 2003.

[29] Y. Saad and M.H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7:856–869, 1986.

[30] Y. Saad, M. Yeung, J. Ehrel, and F. Guyomarc'h. A deflated version of the conjugate gradient algorithm. *SIAM Journal on Scientific Computing*, 21:1909–1926, 2000.

[31] P. Sonneveld and M.B. van Gijzen. IDR(s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations. *SIAM Journal on Scientific Computing*, 31:1035–1062, 2008.

[32] E. de Sturler. Truncation strategies for optimal Krylov subspace methods. *SIAM J. Numer. Anal.*, 36(3):864–889, 1999.

[33] J.M. Tang, R. Nabben, C. Vuik, and Y.A. Erlangga. Comparison of two-level preconditioners derived from deflation, domain decomposition and multigrid methods. *Journal of Scientific Computing*, 39:340–370, 2009.

[34] A. van der Sluis and H.A. van der Vorst. The rate of convergence of conjugate gradients. *Numer. Math.*, 48:543–560, 1986.

[35] H.A. van der Vorst. Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for solution of non-symmetric linear systems. *SIAM J. Sci. Stat. Comp.*, 13:631–644, 1992.

[36] H.A. van der Vorst and C. Vuik. The superlinear convergence behaviour of GMRES. *J. Comp. Appl. Math.*, 48:327–341, 1993.

[37] C. Vuik, A. Segal, and J.A. Meijerink. An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts in the coefficients. *J. Comp. Phys.*, 152:385–403, 1999.

[38] M.C. Yeung and T.F. Chan. ML(k)BiCGSTAB: A Bi-CGSTAB variant based on multiple Lanczos starting vectors. *SIAM Journal on Scientific Computing*, 21:1263–1290, 1999.