Analysis of GMRES for Low-Rank and Small-Norm Perturbations of the Identity Matrix

Arielle K. Carr^{1,*}, **Eric de Sturler**^{2,**}, and **Mark Embree**^{2,***}

[1] Department of Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania, USA
[2] Department of Mathematics, Virginia Tech, Blacksburg, Virginia, USA

In many applications, linear systems arise where the coefficient matrix takes the special form $\mathbf{I}+\mathbf{K}+\mathbf{E}$, where $\mathbf{I}$ is the identity matrix of dimension n, $\operatorname{rank}(\mathbf{K}) = p \ll n$, and $\|\mathbf{E}\| \le \epsilon < 1$. GMRES convergence rates for linear systems with coefficient matrices of the forms $\mathbf{I}+\mathbf{K}$ and $\mathbf{I}+\mathbf{E}$ are guaranteed by well-known theory, but only relatively weak convergence bounds specific to matrices of the form $\mathbf{I}+\mathbf{K}+\mathbf{E}$ currently exist. In this paper, we explore the convergence properties of linear systems with such coefficient matrices by considering the pseudospectrum of $\mathbf{I}+\mathbf{K}$. We derive a bound for the GMRES residual in terms of ϵ when approximately solving the linear system $(\mathbf{I}+\mathbf{K}+\mathbf{E})\mathbf{x} = \mathbf{b}$ and identify the eigenvalues of $\mathbf{I}+\mathbf{K}$ that are sensitive to perturbation. In particular, while a clustered spectrum away from the origin is often a good indicator of fast GMRES convergence, that convergence may be slow when some of those eigenvalues are ill-conditioned. We show there can be at most $2p$ eigenvalues of $\mathbf{I}+\mathbf{K}$ that are sensitive to small perturbations. We present numerical results when using GMRES to solve a sequence of linear systems of the form $(\mathbf{I}+\mathbf{K}_j+\mathbf{E}_j)\mathbf{x}_j = \mathbf{b}_j$ that arise from the application of Broyden's method to solve a nonlinear partial differential equation.

1 Introduction

In this paper, we consider the convergence of iterative solvers for linear systems with coefficient matrices of the form

$$\mathbf{I}+\mathbf{K}+\mathbf{E} \in \mathbb{C}^{n\times n}, \tag{1}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{K}$ is *low rank* (i.e., $\operatorname{rank}(\mathbf{K}) = p \ll n$), and $\mathbf{E}$ is *small-in-norm* (i.e., $\|\mathbf{E}\|_2 = \epsilon < 1$). Sequences of matrices of the special form (1) arise in many applications. As an example, we consider a preconditioned nonlinear partial differential equation (PDE) solved using Broyden's method [1, 2]. Linear systems of this type also arise in methods where a highly effective initial preconditioner gets updated using a sequence of rank-one or low-rank updates [3, 4]. We study the convergence of GMRES [5] applied to linear systems of the form $(\mathbf{I}+\mathbf{K}+\mathbf{E})\mathbf{x} = \mathbf{b}$. At iteration m, GMRES approximates the solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ with the estimate $\mathbf{x}_m \in \mathcal{K}^m(\mathbf{A};\mathbf{r}_0) = \operatorname{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{m-1}\mathbf{r}_0\}$ that minimizes the 2-norm of the residual $\mathbf{r}_m = \mathbf{b} - \mathbf{A}\mathbf{x}_m$. To start, consider the following two well-known convergence bounds with $\mathbf{I}$, $\mathbf{K}$, and $\mathbf{E}$ defined as above (for the proofs, see [6, 7] and [8], respectively).

Theorem 1.1 *Let* $\mathbf{A} = \mathbf{I}+\mathbf{E}$*, Then* $\|\mathbf{r}_m\|_2 \le \epsilon^m \|\mathbf{r}_0\|_2$.

Theorem 1.2 *Let* $\mathbf{A} = \mathbf{I}+\mathbf{K}$. *Then GMRES will converge in at most* $p+1$ *iterations:* $\|\mathbf{r}_{p+1}\|_2 = 0$.

In practice, n is often very large. Given a prescribed convergence tolerance $\tau > 0$, we consider convergence to be fast when $\|\mathbf{r}_m\|_2 \le \tau$ for $m \ll n$. Clearly, the bound in Theorem 1.1 guarantees faster convergence for smaller ϵ, though ϵ need not be very small to ensure rapid GMRES convergence (consider, for example, $\epsilon = 1/3$). Further, as we are interested in cases where $p \ll n$, Theorem 1.2 guarantees very fast convergence. To our knowledge, only relatively weak bounds are currently known for the convergence of GMRES when solving linear systems with coefficient matrices of the form (1). In [9], for $\operatorname{rank}(\mathbf{K}) = p$, it is shown that there exists $\mathcal{C} > 0$ such that (for GMRES) $\|\mathbf{r}_{(p+1)M}\|_2 \le \mathcal{C}^M \|\mathbf{E}\|_2^M \|\mathbf{r}_0\|_2$, for $M > 0$. In particular, one can write the (scaled) minimal polynomial for $\mathbf{I}+\mathbf{K}$ in the form $\mu(z) = 1 + \sum_{k=1}^{p+1} \beta_k z^k$, and take

$$\mathcal{C} = \sum_{k=1}^{q+1} k|\beta_k| \Big\|\mathbf{I}+\mathbf{K}\Big\|_2^{k-1} + O\left(\Big\|\mathbf{E}\Big\|_2^2\right). \tag{2}$$

Further, the empirical observation is made that when $\|\mathbf{E}\|_2 \ll \tau$, GMRES tends to converge in $p+1$ iterations [9]. However, this condition severely limits the size of the perturbation to $\mathbf{I}+\mathbf{K}$ we can consider.

This paper is outlined as follows. In Section 2, we consider the pseudospectrum of $\mathbf{I}+\mathbf{K}$ and derive a straightforward bound for the GMRES residual in terms of ϵ. In Section 3, we identify those eigenvalues of $\mathbf{I}+\mathbf{K}$ that are potentially sensitive to perturbation (i.e., the introduction of $\mathbf{E}$). In Section 4, we study a sequence of linear systems with coefficient matrices of

[*] Corresponding author: arg318@lehigh.edu
[**] sturler@vt.edu
[***] embree@vt.edu

type (1) arising from the solution of a nonlinear PDE using Broyden's method and GMRES. Finally, in Section 5, we provide conclusions and future work. All plots of pseudospectra were computed using EigTool [10].

2 Pseudospectral GMRES Bounds for Perturbed Matrices

In our analysis, rather than directly considering coefficient matrices of the form (1), we examine how $\|\mathbf{E}\|_2$ affects the eigenvalues of $\mathbf{A} + \mathbf{E}$, where

$$\mathbf{A} = \mathbf{I} + \mathbf{K}. \tag{3}$$

Using spectral perturbation theory applied to the resolvents of a general coefficient matrix, $\mathbf{A}$, and the perturbation, $\mathbf{A} + \mathbf{E}$, it is shown in [11] that the norm of the GMRES residual when solving the unperturbed linear system (i.e., $\mathbf{Ax} = \mathbf{b}$) can only increase by at most $O(\epsilon)$, regardless of the magnitude of the change to the eigenvalues of $\mathbf{A}$, provided that $\|\mathbf{E}\|_2 < 1/\|\mathbf{A}^{-1}\|_2$ [11]. Here we specialize this result for the case when the coefficient matrix takes the particular form (3).

The δ-pseudospectrum of the matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ can be defined in the equivalent ways (see, e.g., [12, chap. 2]):

$$\sigma_\delta(\mathbf{A}) = \left\{z \in \mathbb{C} \mid \|(z\mathbf{I} - \mathbf{A})^{-1}\|_2 > 1/\delta\right\} = \{z \in \sigma(\mathbf{A} + \mathbf{E}) \text{ for some } \mathbf{E} \in \mathbb{C}^{n\times n} \text{ with } \|\mathbf{E}\|_2 < \delta\}, \tag{4}$$

where $\sigma(\cdot)$ denotes the spectrum (set of eigenvalues). Note that $\sigma(\mathbf{A})$ is contained in $\sigma_\delta(\mathbf{A})$ for all $\delta > 0$ and the boundary $\partial\sigma_\delta(\mathbf{A})$ encloses a region containing all eigenvalues of $\mathbf{A}$. When $\delta > \epsilon$, we can further conclude that $\partial\sigma_\delta(\mathbf{A})$ encloses the region containing all eigenvalues of $\mathbf{A} + \mathbf{E}$ [11]. Let $\delta_0 > 0$ denote the smallest value for which $\partial\sigma_{\delta_0}(\mathbf{A})$ passes through the origin; then $\sigma_{\delta_0}(\mathbf{A})$ is the largest δ-pseudospectrum of $\mathbf{A}$ that does not contain the origin. Substituting (3) into (4) gives

$$\sigma_\delta(\mathbf{I} + \mathbf{K}) = \left\{z \in \mathbb{C} \;\middle|\; \|((z-1)\mathbf{I} - \mathbf{K})^{-1}\|_2 > 1/\delta\right\},$$

and hence $\sigma_\delta(\mathbf{I} + \mathbf{K}) = 1 + \sigma_\delta(\mathbf{K})$. In Section 3 we identify precisely those eigenvalues of $\mathbf{I} + \mathbf{K}$ that are sensitive to small-in-norm perturbations.

Let $\boldsymbol{\rho}_m$ denote the residual at the mth iteration of (full) GMRES applied to $(\mathbf{I} + \mathbf{K})\mathbf{x} = \mathbf{b}$, and $\mathbf{r}_m$ denote the analogous quantity for GMRES applied to $(\mathbf{I} + \mathbf{K} + \mathbf{E})\mathbf{x} = \mathbf{b}$ (both with the same $\mathbf{x}_0$). Applying [11, Corollary 2.2] to $\mathbf{A} = \mathbf{I} + \mathbf{K}$ gives, for all $\delta > \epsilon = \|\mathbf{E}\|_2$,

$$\|\mathbf{r}_m\|_2 \le \|\boldsymbol{\rho}_m\|_2 + \epsilon\, C_m(\delta), \tag{5}$$

where

$$C_m(\delta) = \left(\frac{L_\delta \|\mathbf{b}\|_2}{\pi\delta^2}\right) \sup_{z\in\sigma_\delta(\mathbf{A})} |\psi_m(z)|. \tag{6}$$

Here L_δ denotes the arc length of $\partial\sigma_\delta(\mathbf{I} + \mathbf{K})$, and $\psi_m(z)$ denotes the residual polynomial at the mth iteration of GMRES applied to $(\mathbf{I} + \mathbf{K})\mathbf{x} = \mathbf{b}$. Theorem 1.2 guarantees that $\|\boldsymbol{\rho}_{p+1}\|_2 = 0$, and hence for any $m \ge p + 1$ the bound in (5) becomes

$$\|\mathbf{r}_m\|_2 < \epsilon\, C_m(\delta), \qquad m \ge p + 1. \tag{7}$$

This bound provides a concrete alternative to the asymptotic expression in (2), and formalizes an empirical observation made in [9]: When $\epsilon < \tau/C_{p+1}(\delta)$, we can guarantee our perturbed system will converge in $p + 1$ steps. Of course this last observation still limits the size of $\|\mathbf{E}\|_2$ we can consider. However, the analysis in (7) provides a mechanism for bounding GMRES convergence for $\mathbf{I} + \mathbf{K} + \mathbf{E}$ at iterations *beyond* $m = p + 1$. At iterations m beyond the point at which GMRES has exactly converged for the $\mathbf{I} + \mathbf{K}$ system, the residual polynomial $\psi_m(z)$ can be taken to be any polynomial for which $\psi_m(0) = 1$ and $\psi_m(\mathbf{I} + \mathbf{K})\boldsymbol{\rho}_0 = \mathbf{0}$ (an idea employed by Ymbert to analyze a different situation in which GMRES converges quickly for the unperturbed system [13]). Here we simply bound $\|\mathbf{r}_{(p+1)M}\|_2$ using $\psi_{(p+1)M}(z) = \psi_{p+1}(z)^M$, and the bound (7) gives convergence when $\epsilon < \tau/C_{(p+1)M}(\delta)$.

As a simple example, consider a matrix of the form (3), with $n = 25$, rank($\mathbf{K}$) $= 2$, and three eigenvalues: $\lambda_1 = 1$ of multiplicity 23, and simple eigenvalues $\lambda_2 \approx 1.25$ and $\lambda_3 \approx 12.25$. Figure 1 shows pseudospectra of this $\mathbf{I} + \mathbf{K}$, and Figure 2 shows $\|\mathbf{r}_m\|_2$, $\|\boldsymbol{\rho}_m\|_2$ (for 100 random perturbations $\mathbf{E}$), and the upper bound (7) on $\|\mathbf{r}_m\|_2$ for different δ. We took perturbations of sizes $\epsilon = 10^{-3.5}$ and $\epsilon = 10^{-5}$, and used several values of $\delta \in (\epsilon, 1/\|\mathbf{A}^{-1}\|_2)$. The right-hand side $\mathbf{b}$ is fixed throughout, generated by MATLAB's `randn` command and normalized to be a unit vector. The unperturbed system converges to tolerance $\tau = 10^{-10}$ in three iterations, as guaranteed by Theorem 1.2. The perturbation causes a slight delay, with GMRES converging to $\tau = 10^{-10}$ in four iterations for $\|\mathbf{E}\|_2 = 10^{-5}$, and five iterations for $\|\mathbf{E}\|_2 = 10^{-3.5}$.

We estimate the bound (7) for $\delta \le 10^{-2}$ by computing the three components of $\sigma_\delta(\mathbf{I} + \mathbf{K})$ in EigTool [10] (reducing the axes to appropriate scale as δ decreases). From this data MATLAB's contour plotting routine approximates the boundary $\partial\sigma_\delta(\mathbf{I} + \mathbf{K})$ as discrete points that determine three closed curves, from which we estimate the boundary length L_δ. The sup term in (7) is estimated by taking the maximum value of $|\psi_m(z)|$ for z among the contour points that approximate $\partial\sigma_\delta(\mathbf{I}+\mathbf{K})$.

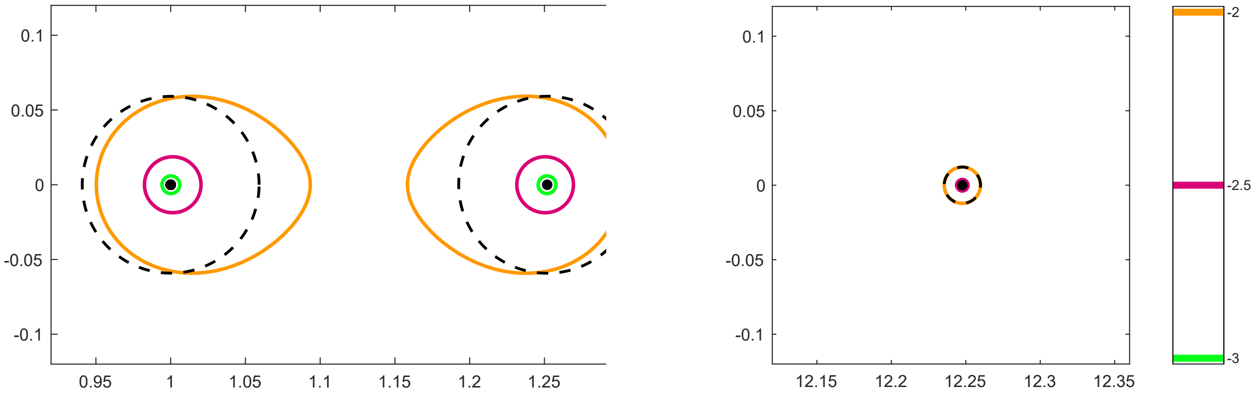


Fig. 1: Boundaries of $\sigma_\delta(A)$ for $\delta = 10^{-2}, 10^{-2.5}, 10^{-3}$ for an example with $n = 25$ and $\mathrm{rank}(\mathbf{K}) = 2$. (The color bar on the right shows $\log_{10}(\delta)$.) The black dots show the three eigenvalues of $\mathbf{I}+\mathbf{K}$. The dashed black curves show the asymptotic sets $\{\lambda_j + \kappa_j \delta \mathrm{e}^{\mathrm{i}\theta} : \theta \in [0, 2\pi)\}$

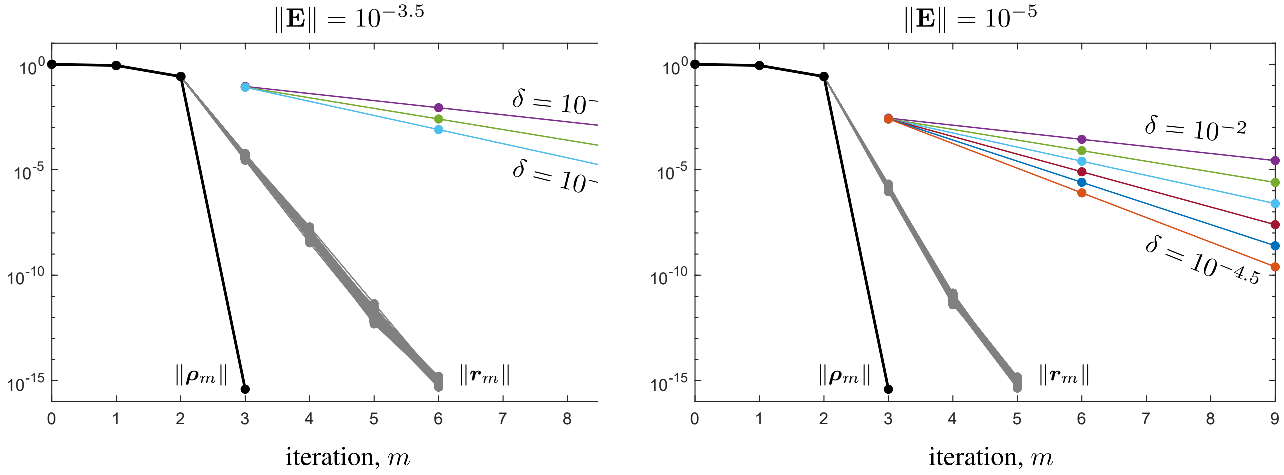


Fig. 2: GMRES convergence for $\mathbf{I} + \mathbf{K}$ (black line) and for $\mathbf{I} + \mathbf{K} + \mathbf{E}$ (gray lines, for 100 random perturbations $\mathbf{E}$). The colored lines show the upper bound (7) for iterations $m = 3$, 6, and 9, for $\log_{10}(\delta) = -2, -2.5, -3$ for $\|\mathbf{E}\| = 10^{-3.5}$ (left) and $\log_{10}(\delta) = -2, -2.5, \ldots, -4.5$ for $\|\mathbf{E}\| = 10^{-5}$ (right).

This process results in the estimated bounds shown in Figure 2. For reference, $C_3(10^{-3}) \approx 256.7$, $C_6(10^{-3}) \approx 2.52$, and $C_9(10^{-3}) \approx 0.025$. These values are multiplied by $\epsilon = \|\mathbf{E}\|_2$, so, for example, when $\epsilon = 10^{-5}$ and $\delta = 10^{-3}$, (7) ensures convergence to $\|\mathbf{r}_3\|_2 < 3 \cdot 10^{-3}$, $\|\mathbf{r}_6\|_2 < 3 \cdot 10^{-5}$, and $\|\mathbf{r}_9\|_2 < 3 \cdot 10^{-7}$. While Figure 2 suggests that these bounds can be somewhat pessimistic, they will capture eventual convergence. Note that while δ must be chosen such that $\delta > \epsilon$, (6) itself does not depend on ϵ. Thus very small perturbations of (3) will give small values of $\epsilon\, C_m$.

Following [13] and assuming $\mathbf{I} + \mathbf{K}$ is diagonalizable, we can approximate the bounds in terms of the *condition numbers of the eigenvalues* (i.e., the norms of the spectral projectors associated with each distinct eigenvalue; for simple eigenvalues, the secant of the angle between left and right eigenvectors); see, e.g., [12, 14]. Let $\kappa_j \geq 1$ denote the condition number of λ_j, $j = 1, \ldots, p + 1$. Then the asymptotic behavior of $(z\mathbf{I} - (\mathbf{I} + \mathbf{K}))^{-1}$ as $z \to \lambda_j$ implies that $\sigma_\delta(\mathbf{I} + \mathbf{K})$ behaves, as $\delta \to 0$, like the union of $p + 1$ disks, each centered at an eigenvalue λ_j and having radius $\kappa_j \delta$; see [12, chap. 52]. (For the example in Figure 1, we compare these asymptotic sets with the true pseudospectra for $\delta = 10^{-2}$.) For small δ, we can thus estimate $L_\delta = 2\pi\delta(\kappa_1 + \cdots + \kappa_{p+1})$ and hence, as $\delta \to 0$,

$$\|\mathbf{r}_m\|_2 < \epsilon\, C_m(\delta) = \epsilon \left(\frac{L_\delta \|\mathbf{b}\|_2}{\pi \delta^2} \right) \sup_{z \in \sigma_\delta(\mathbf{A})} |\psi_m(z)| \approx \frac{2\,\epsilon\,(\kappa_1 + \cdots + \kappa_{p+1})}{\delta} \max_{\substack{j=1,\ldots,p+1 \\ \theta \in [0,2\pi)}} |\psi_m(\lambda_j + \kappa_j \delta \mathrm{e}^{\mathrm{i}\theta})|.$$

Next, we seek more insight into the conditioning of the eigenvalues of $\mathbf{I} + \mathbf{K}$.

3 Sources of Sensitive Eigenvalues

While a clustered spectrum away from the origin is often a good indicator of fast GMRES convergence, convergence may be slow when some of those eigenvalues are ill-conditioned. In the last section we saw that we can instead focus on the δ-pseudospectrum of $\sigma_\delta(\mathbf{I}+\mathbf{K})$. In fact, there can only be at most $2p$ eigenvalues of $\mathbf{I}+\mathbf{K}$ that are sensitive to the introduction of $\mathbf{E}$, which we formalize now. Let $R(\cdot)$ denote the range of an $n \times n$ matrix and $N(\cdot)$ denote the null space.

First, consider the (full) singular value decomposition (SVD) of $\mathbf{K}$ (with $\mathbf{\Sigma}_1$ nonsingular), written as

$$\mathbf{K} = \left[\begin{array}{c|c} \mathbf{U}_1 & \mathbf{U}_2 \end{array}\right] \left[\begin{array}{c|c} \mathbf{\Sigma}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right] \left[\begin{array}{c} \mathbf{V}_1^* \\ \hline \mathbf{V}_2^* \end{array}\right] = \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^*. \tag{8}$$

Given $\mathbf{A}$ of the form (3) and the SVD of $\mathbf{K}$ as in (8), we can study the equivalent eigenvalue problems

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x} \quad \Longleftrightarrow \quad \mathbf{x} + \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{x} = \lambda \mathbf{x} \quad \Longleftrightarrow \quad \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{x} = (\lambda - 1)\mathbf{x}. \tag{9}$$

We characterize the eigenpairs $(\lambda, \mathbf{x})$ to identify the source of potentially ill-conditioned eigenvalues. We begin by defining the eigenvectors corresponding to eigenvalue $\lambda = 1$.

Theorem 3.1 $(1, \mathbf{x})$ *is an eigenpair of* $\mathbf{A} = \mathbf{I} + \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^*$ *if and only if* $\mathbf{x} \in R(\mathbf{V}_2)$.

Proof. Let $(1, \mathbf{x})$ be an eigenpair of $\mathbf{A} = \mathbf{I} + \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^*$. Then, $\mathbf{x} + \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{x} = \mathbf{x}$ implies $\mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{x} = \mathbf{0}$, and clearly $\mathbf{x} \in N(\mathbf{V}_1^*) = R(\mathbf{V}_2)$. Now assume $\mathbf{0} \neq \mathbf{x} \in R(\mathbf{V}_2) = N(\mathbf{V}_1^*)$. Then $\mathbf{V}_1^* \mathbf{x} = \mathbf{0}$ and (9) implies $\lambda = 1$. □

When $\lambda \neq 1$, (9) implies that the corresponding eigenvector $\mathbf{x} \in R(\mathbf{U}_1)$. In this case, we need only analyze the eigenpairs of the small matrix $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1$, which we make precise now.

Theorem 3.2 $\mathbf{x} = \mathbf{U}_1 \boldsymbol{\xi}$ *is an eigenvector of* $\mathbf{A} = \mathbf{I} + \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^*$ *if and only if* $\boldsymbol{\xi}$ *is an eigenvector of* $\mathbf{\Sigma_1 V_1^* U_1}$.

Proof. Let $\mathbf{x} = \mathbf{U}_1 \boldsymbol{\xi}$ be an eigenvector of $\mathbf{I} + \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^*$. Then $\mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1 \boldsymbol{\xi} = \gamma \mathbf{U}_1 \boldsymbol{\xi}$, with $\gamma = \lambda - 1$. Left multiplying by $\mathbf{U}_1^*$ gives $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1 \boldsymbol{\xi} = \gamma \boldsymbol{\xi}$, and so $\boldsymbol{\xi}$ is an eigenvector of $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1$ with eigenvalue $\gamma = \lambda - 1$. Now assume $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1 \boldsymbol{\xi} = \gamma \boldsymbol{\xi}$ for $\boldsymbol{\xi} \neq \mathbf{0}$. Then $\mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* (\mathbf{U}_1 \boldsymbol{\xi}) = \gamma (\mathbf{U}_1 \boldsymbol{\xi})$, which is equivalent to (9) with $\lambda - 1 = \gamma$ and $\mathbf{x} = \mathbf{U}_1 \boldsymbol{\xi}$. □

Suppose $\boldsymbol{\xi} \neq \mathbf{0}$ is an eigenvector of $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1$ corresponding to eigenvalue γ. First, consider the case that $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1$ is diagonalizable. We need to distinguish between $\gamma = 0$ and $\gamma \neq 0$. If $\gamma \neq 0$, $\mathbf{U}_1 \boldsymbol{\xi}$ is an eigenvector of $\mathbf{A}$ with eigenvalue $\lambda = \gamma + 1$. When $\gamma = 0$, $\lambda = 1$ and $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1 \boldsymbol{\xi} = \mathbf{0}$ implies $\mathbf{V}_1^* \mathbf{U}_1 \boldsymbol{\xi} = \mathbf{0}$, and hence $\mathbf{U}_1 \boldsymbol{\xi} \in R(\mathbf{V}_2)$ is not a 'new' eigenvector of $\mathbf{A}$. In this case, $\mathbf{V}_1 \mathbf{\Sigma}_1^{-1} \boldsymbol{\xi}$ is a generalized eigenvector of order two with defective eigenvalue $\lambda = 1$:

$$(\mathbf{A} - \mathbf{I})^2 \mathbf{V}_1 \mathbf{\Sigma}_1^{-1} \boldsymbol{\xi} = (\mathbf{A} - \mathbf{I}) \mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{V}_1 \mathbf{\Sigma}_1^{-1} \boldsymbol{\xi} = (\mathbf{A} - \mathbf{I})(\mathbf{U}_1 \boldsymbol{\xi}) = \mathbf{0},$$

with $\mathbf{U}_1 \boldsymbol{\xi} \neq \mathbf{0}$. Such defective eigenvalues are sensitive to perturbations [15]. Let ℓ denote the multiplicity of $\gamma = 0$. Then we have ℓ eigenvectors $\mathbf{U}_1 \boldsymbol{\xi} \in R(\mathbf{V}_2)$ and ℓ generalized eigenvectors of order two as above, $p - \ell$ eigenvectors of type $\mathbf{U}_1 \boldsymbol{\xi}$ (with $\gamma \neq 0$), in addition to $n - p - \ell$ further eigenvectors in $R(\mathbf{V}_2)$ for $\lambda = 1$.

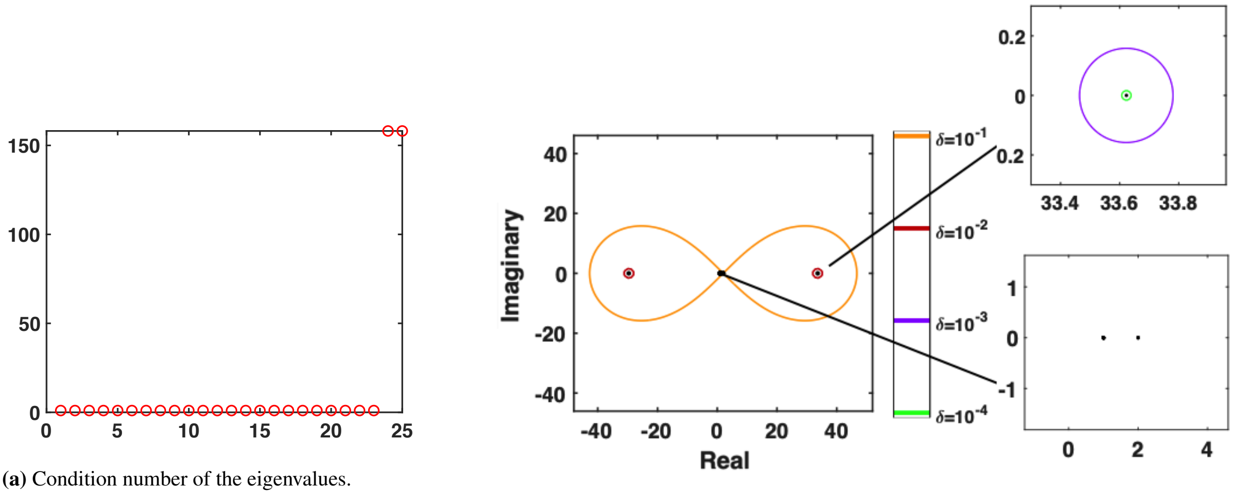
If $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1$ is not diagonalizable and $\gamma \neq 0$, then a generalized eigenvector $\boldsymbol{\xi}$ of order k corresponding to γ corresponds to a generalized eigenvector $\mathbf{U}_1 \boldsymbol{\xi}$ of the same order corresponding to $\lambda = \gamma + 1$ of $\mathbf{A}$. This follows directly from the observation that $(\mathbf{U}_1 \mathbf{\Sigma}_1 \mathbf{V}_1^* - \gamma \mathbf{I}_n) \mathbf{U}_1 = \mathbf{U}_1 (\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1 - \gamma \mathbf{I}_p)$ and hence $\mathbf{U}_1 (\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1 - \gamma \mathbf{I}_p)^k \boldsymbol{\xi} = \mathbf{0} \Leftrightarrow (\mathbf{A} - (\gamma + 1)\mathbf{I}_n)^k \mathbf{U}_1 \boldsymbol{\xi} = \mathbf{0}$. Again, defective eigenvalues are highly sensitive to perturbations [15].

As a result, we get up to p sensitive eigenvalues from small angles between the eigenvectors in $R(\mathbf{U}_1)$, defective eigenvalues corresponding to $\gamma = 0 \Leftrightarrow \lambda = 1$, or $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1$ being non-diagonalizable. Another up to p potentially sensitive eigenvalues arise from small angles between eigenvectors $\mathbf{U}_1 \boldsymbol{\xi}$ (with $\lambda \neq 1$) and $R(\mathbf{V}_2)$, as any vector $\mathbf{V}_2 \boldsymbol{\zeta}$ is an eigenvector (with $\lambda = 1$), and from small angles between eigenvectors $\mathbf{U}_1 \boldsymbol{\xi}$ and the vectors $\mathbf{V}_1 \mathbf{\Sigma}_1^{-1} \boldsymbol{\xi}$ corresponding to eigenvalues $\gamma = 0$.

If $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1$ is not diagonalizable and $\gamma = 0$, the results above show that a Jordan chain of lengths s for $\mathbf{\Sigma}_1 \mathbf{V}_1^* \mathbf{U}_1$ with eigenvalue $\gamma = 0$, $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_s$ (the subscript indicating the order) gives a Jordan chain of $\mathbf{A}$ (with eigenvalue $\lambda = 1$) of length $s + 1$ given by $\mathbf{U}_1 \boldsymbol{\xi}_1, \mathbf{V}_1 \mathbf{\Sigma}_1^{-1} \boldsymbol{\xi}_1, \ldots, \mathbf{V}_1 \mathbf{\Sigma}_1^{-1} \boldsymbol{\xi}_s$.

We have identified sources of sensitive eigenvalues for matrices taking the form (3). Next, we give two small examples, each highlighting how these sources of sensitivity can affect the convergence of GMRES.

Example 1. To visualize the effect of the sensitive eigenvalues of $\mathbf{\Sigma_1 V_1^* U_1}$ on GMRES convergence for matrices of type (3), we consider an example where $\mathbf{\Sigma_1 V_1^* U_1}$ has two ill-conditioned eigenvalues. We let the matrix size be $n = 25$ and $p = \mathrm{rank}(\mathbf{K}) = 5$. The sensitive eigenvalues of $\mathbf{K}$ are $\gamma_1 \approx 32.62$ and $\gamma_2 \approx -30.62$, with corresponding sensitive eigenvalues of $\mathbf{A}$, $\lambda_1 \approx 33.62$ and $\lambda_2 \approx -29.62$. We show in Figure 3(a) that the matrix has two sensitive eigenvalues (i.e., those shown with large condition number). Figure 3(b) shows pseudospectra of $\mathbf{A}$. We construct $\mathbf{E}_1$ and $\mathbf{E}_2$ such that $\|\mathbf{E}_1\|_2 = 10^{-3}$ and $\|\mathbf{E}_2\|_2 = 10^{-1}$ and run GMRES to solve the linear systems $\mathbf{A}_{1,2}\mathbf{x} = \mathbf{b}$, for $\mathbf{A}_1 = \mathbf{I} + \mathbf{K} + \mathbf{E}_1$ and $\mathbf{A}_2 = \mathbf{I} + \mathbf{K} + \mathbf{E}_2$. Here, we let $\mathbf{b}$ be a random vector generated by MATLAB's `rand` function. For the unperturbed system, GMRES converges in 4 iterations. For the system with coefficient matrix $\mathbf{A}_1$, GMRES converges in 7 iterations, and for $\mathbf{A}_2$, iterations further increase to 14. In fact, even for a very small perturbation $\mathbf{E}_3$ with $\|\mathbf{E}_3\|_2 = 10^{-6}$, GMRES converges in 6 iterations.

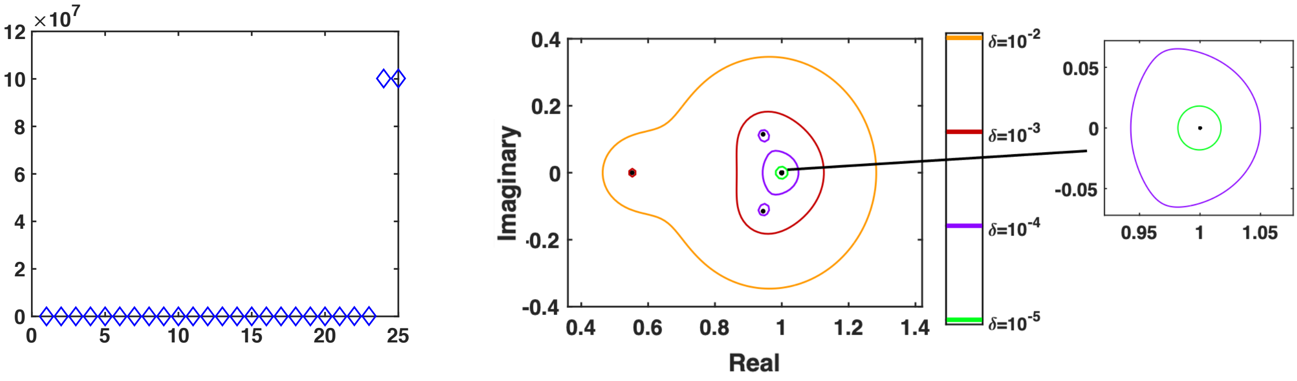


(a) Condition number of the eigenvalues.

(b) Pseudospectra of $\mathbf{I}+\mathbf{K}$. The color bar gives δ corresponding to $\sigma_\delta(\mathbf{I}+\mathbf{K})$.

Fig. 3: $\mathbf{I}+\mathbf{K}$ with dimension $n = 25$, rank$(\mathbf{K}) = 5$, and two ill-conditioned eigenvalues of $\mathbf{\Sigma}_1\mathbf{V}_1^*\mathbf{U}_1$.

Example 2. Next, we analyze a matrix with very small angles between $R(\mathbf{U}_1)$ and $R(\mathbf{V}_2)$. Specifically, we let the matrix size be $n = 25$ and let $p = \text{rank}(\mathbf{K}) = 5$. We construct $\mathbf{K}$ such that there are two small angles, $\theta_{1,2} = 10^{-5}$, between $R(\mathbf{U}_1)$ and $R(\mathbf{V}_2)$. We see in Figure 4(a) that the unperturbed matrix we construct has two sensitive eigenvalues. Figure 4(b) shows pseudospectra of $\mathbf{A}$. We choose $\mathbf{E}_1$ and $\mathbf{E}_2$ such that $\|\mathbf{E}_1\|_2 = 10^{-5}$ and $\|\mathbf{E}_2\|_2 = 10^{-2}$ and consider GMRES iterations for $\mathbf{A}_1 = \mathbf{I}+\mathbf{K}+\mathbf{E}_1$ and $\mathbf{A}_2 = \mathbf{I}+\mathbf{K}+\mathbf{E}_2$. For the unperturbed system, GMRES converges in 6 iterations. For the perturbed system corresponding to $\mathbf{A}_1$, iterations increase slightly to 7, and for $\mathbf{A}_2$ iterations increase to 10.



(a) Condition number of the eigenvalues.

(b) Pseudospectra of $\mathbf{I}+\mathbf{K}$. The color bar gives δ corresponding to $\sigma_\delta(\mathbf{I}+\mathbf{K})$.

Fig. 4: $\mathbf{I}+\mathbf{K}$ with dimension $n = 25$, rank$(\mathbf{K}) = 5$, and two small angles between $\mathbf{U}_1$ and $\mathbf{V}_2$.

4 Experimental Results

We consider the solution of the nonlinear PDE

$$\begin{aligned} -\nabla\cdot(\nabla u) + (1+u)(70u_x + 70u_y) &= f \quad \text{on} \quad (0,1)\times(0,1), \\ u &= 0 \quad \text{for} \quad x=0, x=1 \text{ or } y=0, y=1, \end{aligned} \tag{10}$$

on the unit square, where f is chosen so that the solution satisfies $u(x,y) = y\sin(\pi y)(1-x)\sin(\pi x)e^{4x}$. We use a finite difference discretization with second order central differences and mesh width $h = 1/201$. Discretization leads to the discrete system $\mathbf{F}(\mathbf{u}) = \mathbf{A}\mathbf{u} + (\mathbf{D}(\mathbf{u})+\mathbf{I})(70\mathbf{D}_x + 70\mathbf{D}_y)\mathbf{u} - \mathbf{f} = \mathbf{0}$, where $\mathbf{D}(\mathbf{x})$ denotes the diagonal matrix with the coefficients of $\mathbf{x}$ on the diagonal, $\mathbf{A}$ is the discretized diffusion operator, and $\mathbf{D}_x, \mathbf{D}_y$ are the convection operators in the x and y directions. Our initial solution is $\mathbf{u} = \mathbf{0}$, and hence the initial Jacobian is $\mathbf{J}_0 = \mathbf{A} + \mathbf{D}((70\mathbf{D}_x + 70\mathbf{D}_y)\mathbf{u}) + (\mathbf{I}+\mathbf{D}(\mathbf{u}))(70\mathbf{D}_x + 70\mathbf{D}_y)$.

In general, realistic systems are three dimensional and too large for a direct solver. Thus we consider a very good preconditioner, $\mathbf{P}$, but not an exact factorization (as, e.g., suggested in [2]). The high cost of a very good preconditioner is amortized over multiple linear solves. We use Broyden's method with GMRES for the preconditioned nonlinear system $\mathbf{P}\mathbf{F}(\mathbf{u}) = \mathbf{0}$ with initial approximation $\mathbf{u} = \mathbf{0}$, which results in the initial preconditioned Jacobian $\mathbf{B}_0 = \mathbf{P}\mathbf{J}_0 = \mathbf{I} + \mathbf{E}$, where $\|\mathbf{E}\|_2 = \varepsilon < 1$. This leads to a sequence of linear systems of the form $(\mathbf{I} + \mathbf{K} + \mathbf{E})\mathbf{x} = \mathbf{b}$, where the rank of $\mathbf{K}$ increases with each nonlinear iteration. We use ILUT [16] with drop tolerance 10^{-4}, which gives $\|\mathbf{E}\|_2 \approx 0.4$. In practice, we estimate $\|\mathbf{E}\|_2$ using information from GMRES in the first Broyden step. The line search uses 3-point parabolic interpolation and the Armijo condition [2]. Each Broyden step adds a rank-one update. So, at iteration p we solve a system of the form $\mathbf{I} + \mathbf{K} + \mathbf{E}$ with $\operatorname{rank}(\mathbf{K}) = p$.

Next, we relate the observed convergence to the analysis provided above, for a realistic $\|\mathbf{E}\|_2 \approx 0.4$, obtained with an accurate preconditioner. Space limitations prevent us from providing all details, but, for this problem, the analysis shows that the eigenvalues introduced by the low-rank updates are not sensitive and the non-unit eigenvalues λ are very close to the $1+\gamma_j$ computed from $\mathbf{\Sigma}_1\mathbf{V}_1^*\mathbf{U}_1$, leading to a marginal increase in the number of GMRES iterations over the Broyden iteration.

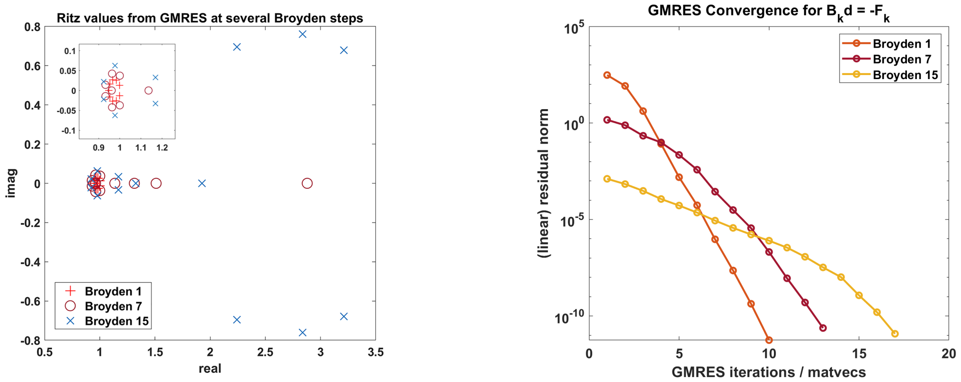


Fig. 5: The left figure shows the spectrum of $\mathbf{I} + \mathbf{E}$ (first Broyden step) and $\mathbf{I} + \mathbf{E} + \mathbf{K}$ at the 7th and the last Broyden step. The right figure shows the resulting GMRES convergence. Using the analysis above, we could estimate the spectrum and sensitivities from the SVD of $\mathbf{K}$.

5 Conclusions

In this paper, we characterize GMRES convergence for linear systems with coefficient matrices of the special form $\mathbf{I}+\mathbf{K}+\mathbf{E}$: low rank plus small-in-norm perturbations of the identity matrix. We define a bound for the GMRES residual in terms of the size of $\|\mathbf{E}\|_2$ by following the theoretical framework of [11]. We reveal the potentially sensitive eigenvalues of $\mathbf{I} + \mathbf{K}$ when introducing small perturbations (i.e., $\mathbf{E}$), showing that there are no more than $2p$ such eigenvalues, where p denotes the rank of $\mathbf{K}$.

Several interesting theoretical and practical extensions of the current paper are the focus of ongoing work, such as considering $((z-1)\mathbf{I} - \mathbf{K})^{-1}$ directly by way of the Sherman–Morrison–Woodbury formula, and when solving long sequences of linear systems of the form $(\mathbf{I} + \mathbf{K}_j + \mathbf{E}_j)\mathbf{x}_j = \mathbf{b}_j$. In particular, we will use this analysis more extensively to, e.g., determine when to recompute or update the preconditioner in the Broyden iteration when solving nonlinear PDEs using GMRES.

References

[1] C. Broyden, Mathematics of Computation **19**, 577–593 (1965).
[2] C. T. Kelley, Solving Nonlinear Equations with Newton's Method (SIAM, 2003).
[3] K. Ahuja, B. K. Clark, E. de Sturler, D. M. Ceperley, and J. Kim, SIAM Journal on Scientific Computing **33**(4), 1837–1859 (2011).
[4] L. Bergamaschi, R. Bru, A. Martínez, and M. Putti, Electronic Transactions in Numerical Analysis (ETNA) **23**, 76–87 (2006).
[5] Y. Saad and M. H. Schultz, SIAM Journal on Scientific and Statistical Computing **7**(3), 856–869 (1986).
[6] S. L. Campbell, I. C. Ipsen, C. T. Kelley, and C. D. Meyer, BIT Numerical Mathematics **36**(4), 664–675 (1996).
[7] N. Gmati and B. Philippe, Comments on the GMRES convergence for preconditioned systems, in: Large-Scale Scientific Computing, (Springer, 2008), pp. 40–51.
[8] R. Ehrig and P. Deuflhard, GMERR - an error minimizing variant of GMRES, Tech. Rep. SC-97-63, ZIB, Berlin, 1997.
[9] C. T. Kelley, I. G. Kevrekidis, and L. Qiao, Newton-Krylov solvers for time-steppers, preprint, arxiv.org/abs/math/0404374, 2004.
[10] T. G. Wright, EigTool, https://github.com/eigtool, 2002.
[11] J. A. Sifuentes, M. Embree, and R. B. Morgan, SIAM Journal on Matrix Analysis and Applications **34**(3), 1066–1088 (2013).
[12] L. N. Trefethen and M. Embree, Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators (Princeton University Press, Princeton, NJ, 2005).
[13] G. Ymbert III, Convergence bounds for approximate preconditioning, Master's thesis, Texas A&M University, Corpus Christi, 2011.
[14] G. H. Golub and C. F. Van Loan, Matrix Computations, fourth edition (Johns Hopkins University Press, Baltimore, 2012).
[15] J. Moro, J. V. Burke, and M. L. Overton, SIAM Journal on Matrix Analysis and Applications **18**, 793–817 (1997).
[16] Y. Saad, Numerical Linear Algebra with Applications **1**(4), 387–402 (1994).