12-2022

GMRES Convergence of Block Preconditioners for Nonsymmetric Matrices

Miguel A. Mascorro
The University of Texas Rio Grande Valley

Recommended Citation

Mascorro, Miguel A., "GMRES Convergence of Block Preconditioners for Nonsymmetric Matrices" (2022). *Theses and Dissertations*. 1163.
https://scholarworks.utrgv.edu/etd/1163

GMRES CONVERGENCE OF BLOCK PRECONDITIONERS FOR NONSYMMETRIC MATRICES

A Thesis

by

MIGUEL A. MASCORRO

Submitted in Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE

Major Subject: Applied Mathematics

The University of Texas Rio Grande Valley

December 2022

GMRES CONVERGENCE OF BLOCK PRECONDITIONERS FOR NONSYMMETRIC MATRICES

A Thesis
by
MIGUEL A. MASCORRO

COMMITTEE MEMBERS

Dr. Josef Sifuentes
Chair of Committee

Cristina Villalobos
Committee Member

Andras Balogh
Committee Member

Mrinal Roychowdhury
Committee Member

December 2022

ABSTRACT

Mascorro, Miguel A., GMRES Convergence of Block Preconditionersfor Nonsymmetric Matrices. Master of Science (MS), December, 2022, 42 pp., 15 figures, references, 17 titles.

GMRES is an iterative method for solving linear systems that minimizes the residual over the *k*-dimensional Krylov subspace at iteration k. Murphy, Golub and Wathen in [11] show that saddle point type matrices can be preconditioned so that GMRES converges in two or three steps. Ipsen in [10] extends this work to matrixes where the (2,2) block is nonzero. However, the three step convergence result no longer holds in this case. In this thesis we investigate how many more steps are needed for convergence as a function of the size of that (2,2) block.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

In this thesis we are going to consider matrices of the form

$$\mathcal{A}_0 = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \quad \text{and} \quad \mathcal{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{1.1}$$

where the former is often referred to as saddle point matrices or Karush-Kuhn-Tucker (KKT) matrices [1][7][12] and where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{m\times n}$. We will also assume throughout that both $\mathcal{A}_0$ and $\mathcal{A}$ are invertible.

For such matrices we are going to explore solving the system

$$\mathcal{A}x = b. \tag{1.2}$$

where $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$, $b \in \mathbb{R}^{n+m}$, and $\mathcal{A}$ and A are both invertible. However, often the size of $\mathcal{A}$ is too large for Gaussian elimination to be practical. So, other numerical methods have to be used to solve the problem in a more reasonable time frame. In particular, we focus on working with the GMRES algorithm. In many instances, some ideal preconditioner $\mathcal{P}$ for a linear system (1.2) results in a coefficient matrix $\mathcal{A}\mathcal{P}^{-1}$ for which the GMRES algorithm converges exactly in some small number of iterations. In fact, in a paper from 2000, Murphy, Golub, and Wathen [11] proposed to precondition the KKT matrix

$$\mathcal{A}_0 = \begin{bmatrix} A & B^* \\ C & 0 \end{bmatrix} \tag{1.3}$$

with the block matrix

$$\mathcal{P}_{\pm} = \begin{bmatrix} A & B^* \\ 0 & \pm CA^{-1}B^* \end{bmatrix} \tag{1.4}$$

Here, $A \in \mathbb{R}^{n\times n}$, and $B, C \in \mathbb{R}^{n\times m}$ and it is assumed both A and $\mathcal{A}_0$ are invertible. Further, this gives the right preconditioned system

$$\mathcal{A}_0\mathcal{P}_{\pm}^{-1} = \begin{bmatrix} I & 0 \\ CA^{-1} & \mp I \end{bmatrix} \tag{1.5}$$

the degree-2 minimal polynomial $(z-1)(z+1)$ for $\mathcal{A}_0\mathcal{P}_{+}^{-1}$ and $(z-1)^2$ for $\mathcal{A}_0\mathcal{P}_{-}^{-1}$, making GMRES converge in two steps. Murphy, Golub, and Wathen [11] also consider the block-diagonal preconditioner

$$\mathcal{P}_{\varphi} = \begin{bmatrix} A & 0 \\ 0 & CA^{-1}B^* \end{bmatrix} \tag{1.6}$$

giving the preconditioned matrices $\mathcal{A}_0\mathcal{P}_{\varphi}^{-1}$ and $\mathcal{P}_{\varphi}^{-1}\mathcal{A}_0$ the same three eigenvalues:

$$\lambda_1 = 1, \quad \lambda_2 = \frac{1+\sqrt{5}}{2}, \quad \lambda_3 = \frac{1-\sqrt{5}}{2} \tag{1.7}$$

Since the eigenvalues are not defective, when applied to the preconditioned linear system $\mathcal{P}_{\varphi}^{-1}\mathcal{A}_0 x = \mathcal{P}_{\varphi}^{-1}b$ or $\mathcal{A}_0\mathcal{P}_{\varphi}^{-1}y = b$ with $\mathcal{A}_0$ nonsingular, the GMRES algorithm converges in no more than three steps [17].

Ipsen [10] extends this idea to generalize the preconditioners (1.4) and (1.6) to the generic block matrix

$$\mathcal{A} = \begin{bmatrix} A & B^* \\ C & D \end{bmatrix} \tag{1.8}$$

via the Schur complement $S := D - CA^{-1}B^*$. If (1.8) is preconditioned by

$$\mathcal{P}_\pm = \begin{bmatrix} A & B^* \\ 0 & \pm S \end{bmatrix}, \tag{1.9}$$

then

$$\mathcal{A}\mathcal{P}_\pm^{-1} = \begin{bmatrix} I & 0 \\ CA^{-1} & \pm I \end{bmatrix}, \tag{1.10}$$

and $\mathcal{P}_\pm^{-1}\mathcal{A}$ and $\mathcal{A}\mathcal{P}_\pm^{-1}$ have the minimal polynomial $(z-1)(z \mp 1)$. As such, the GMRES algorithm still converges in at most two steps. However, if (1.8) is preconditioned by

$$\mathcal{P} = \begin{bmatrix} A & 0 \\ 0 & -S \end{bmatrix}, \tag{1.11}$$

then the right-preconditioned matrix is [10]

$$\mathcal{A}\mathcal{P}^{-1} = \begin{bmatrix} I & -B^*S^{-1} \\ CA^{-1} & -DS^{-1} \end{bmatrix}. \tag{1.12}$$

In this case, unfortunately, the GMRES algorithm convergence in three steps does not hold. So, how many more steps does it take for GMRES to converge?

1.1 KKT Matrices

These types of matrices can arise from a quadratic optimization problem coupled with linear constraints:

$$\text{min or max } f(x) = \frac{1}{2}x^TQx + b^Tx + c$$
$$\text{such that } g(x) := C^Tx = d$$

where $Q \in \mathbb{R}^{n\times n}$, $x, b \in \mathbb{R}^n$, $C \in \mathbb{R}^{n\times m}$ is full rank, and n is often much larger than m. Note that

$$\nabla f(x) = Qx + b \quad \text{and} \quad \nabla g(x) = C$$

So, our first order conditions and constraints are satisfied if

$$Qx + b = C\lambda$$
$$C^T x = d$$

or

$$Qx + C(-\lambda) = -b$$
$$C^T x = d$$

This set of equations are an example of Karush-Kuhn-Tucker (KKT) [1] first order conditions and can be succintly written as

$$\begin{bmatrix} Q & C \\ C^T & 0 \end{bmatrix} \begin{bmatrix} x \\ -\lambda \end{bmatrix} = \begin{bmatrix} -b \\ d \end{bmatrix}$$

These types of matrices also arise from the Stokes equations modeling fluid flow at low Reynold numbers [1][7]. However, sometimes we may be interested in matrices of the form

$$\mathcal{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where $A \in \mathbb{R}^{n\times n}$,$B \in \mathbb{R}^{n\times m}$,$C \in \mathbb{R}^{m\times n}$, and $D \in \mathbb{R}^{m\times m}$ is a non-zero matrix.

1.2 GMRES

GMRES is an iterative method that aims to approximate the solution to (1.2) by minimizing the residual over a finite-dimensional subspace [13]. In particular, GMRES at iteration k solves

$$\min_{x \in x_0 + \mathcal{K}_k(A,r_0)} \|b - Ax\| \tag{1.13}$$

where the k-th Krylov subspace is defined as

$$\mathcal{K}_k(A,r_0) = \text{span}\{r_0, Ar_0, A^2 r_0, \dots, A^{k-1} r_0\}.$$

Here, $r_k = b - Ax_k$ is the residual, and x_k is the k-th iterate.

Notice then that $x = x_0 + y$ for $y \in \mathcal{K}_k(A,r_0)$ and so we also have that

$$b - Ax = b - A(x_0 + y) = b - Ax_0 - Ay = r_0 - Ay.$$

Therefore, we equivalently have that GMRES at iteration k

$$\min_{y \in \mathcal{K}_k(A,r_0)} \|r_0 - Ay\| \tag{1.14}$$

where $x = x_0 + y$.

In practice, we want to find an orthonormal basis for $\mathcal{K}_k(A,r_0)$. Suppose $\{v_1, v_2, \dots, v_k\}$ is an orthonormal basis for $\mathcal{K}_k(A,r_0)$. The orthonormal basis is constructed by applying the modified Gram-Schmidt to the basis $v_1, Av_2, Av_3, Av_4, Av_5, \dots$

If $v_j \in \mathcal{K}_j$, then

$$\begin{aligned} v_j &= \sum_{l=0}^{j-1} c_l A^l r_0 \\ Av_j &= \sum_{l=0}^{j-1} c_l A^{l+1} r_0 \in \mathcal{K}_{j+1} \end{aligned}$$

So at step $k+1$ we project Av_k onto a space orthogonal to $\mathcal{K}_k$ and then normalize as follows:

$$\tilde{v}_{k+1} = (I - v_k v_k^*) \cdots (I - v_1 v_1^*) A v_k \tag{1.15a}$$

$$v_{k+1} = \frac{\tilde{v}_{k+1}}{\|\tilde{v}_{k+1}\|} \tag{1.15b}$$

Then,

$$H_k := V_k^* A V_k$$

is a $k \times k$ upper Hessenberg matrix since

$$(H_k)_{j\ell} = v_j^* A v_\ell$$

and $Av_\ell \in \mathcal{K}_{\ell+1}$, thus making $(H_k)_{j\ell} = 0$ if $j > \ell + 1$.

We also know that $\|\tilde{v}_{k+1}\| = h_{k+1,k}$ and thus from the equations in (1.15) we have

$$\begin{aligned} \|\tilde{v}_{k+1}\| v_{k+1} = h_{k+1,k} v_{k+1} &= (I - V_k V_k^*) A V_k e_k \\ &= A V_k e_k - V_k H_k e_k \end{aligned}$$

Thus,

$$A v_k = V_k H_k e_k + h_{k+1,k} v_{k+1}$$

and since this equality holds for indeces 1 through k, we have that

$$\begin{aligned}
\begin{bmatrix} | & | & & | \\ Av_1 & Av_2 & \cdots & Av_k \\ | & | & & | \end{bmatrix} &= \begin{bmatrix} | & | & & | \\ V_1 H_1 e_1 & V_2 H_2 e_2 & \cdots & V_k H_k e_k \\ | & | & & | \end{bmatrix} + \begin{bmatrix} | & | & & | \\ h_{2,1} v_1 & h_{3,2} v_2 & \cdots & h_{k+1,k} v_{k+1} \\ | & | & & | \end{bmatrix} \\
\begin{bmatrix} | & | & & | \\ Av_1 & Av_2 & \cdots & Av_k \\ | & | & & | \end{bmatrix} &= \begin{bmatrix} | & | & & | \\ h_{1,1} v_1 + h_{2,1} v_2 & h_{1,2} v_1 + h_{2,2} v_2 + h_{3,2} v_3 & \cdots & V_k H_k e_k^T + h_{k+1,k} v_{k+1} \\ | & | & & | \end{bmatrix} \\
AV_k &= V_k H_k + \begin{bmatrix} | & & | & | \\ 0 & \cdots & 0 & h_{k+1,k} v_{k+1} \\ | & & | & | \end{bmatrix} \\
AV_k &= V_k H_k + h_{k+1,k} v_{k+1} e_k^T \qquad (1.16)
\end{aligned}$$

where we can rewrite the second equation as the third since H_k is an upper Hessenberg matrix. Next, recall that the product of a matrix multiplication is the sum of matrices that are products of the columns of the left matrix and rows of the right matrix. So, equation (1.16) can be rewritten as

$$\begin{aligned}
AV_K &= v_1 h_1 + v_2 h_2 + \cdots + v_k h_k + v_{k+1} h_{k+1,k} e_k^T \\
AV_k &= V_{k+1} \tilde{H}_k \qquad (1.17)
\end{aligned}$$

where h_j represent the rows of H_k and

$$\tilde{H}_k = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,k} \\ h_{2,1} & h_{2,2} & & \vdots \\ & h_{3,2} & \ddots & \vdots \\ & & \ddots & h_{k,k} \\ & & & h_{k+1,k} \end{bmatrix} = \begin{bmatrix} H_k \\ h_{k+1,k}e_k^T \end{bmatrix}$$

is a $(k+1) \times k$ matrix.

Now, note that

$$v_1 = V_k e_1 = V_{k+1} e_1 = \frac{r_0}{\|r_0\|} \tag{1.18}$$

and since the span of the columns of V_k are the Krylov subspace, we have the following equivalence

$$\min_{x \in \mathcal{K}_k(A, r_0)} \|r_0 - Ax\| \iff \min_{c \in \mathbb{C}^k} \|r_0 - AV_k c\|$$

Thus, from equations (1.17) and (1.18), we have

$$\min_{c \in \mathbb{C}^k} \|r_0 - AV_k c\| \iff \min_{c \in \mathbb{C}^k} \| \|r_0\| V_{k+1} e_1 - V_{k+1} \tilde{H}_k c\|$$

and since the columns of V_{k+1} are orthonormal, $\|V_{k+1} z\| = \|z\|$ for any $z \in \mathbb{C}^k$ and we have

$$\begin{aligned} \min_{c \in \mathbb{C}^k} \| \|r_0\| V_{k+1} e_1 - V_{k+1} \tilde{H}_k c\| &\iff \min_{c \in \mathbb{C}^k} \|V_{k+1}(\|r_0\| e_1 - \tilde{H}_k c)\| \\ &\iff \min_{c \in \mathbb{C}^k} \| \|r_0\| e_1 - \tilde{H}_k c\| \end{aligned}$$

and thus at every step, we are solving a $(k+1) \times k$ Hessenberg least squares problem. Specifically, after specifying an initial guess x_0 and some tolerance t, we are using algorithm 1 [14] to solve the problem.

Algorithm 1 GMRES

```
Input: $A, b, t, x_0$
Output: $x_k$
  $r_0 = b - Ax_0$
  $v_1 = r_0 / \|r_0\|$
  for $k = 1, 2, 3, \ldots$ do
      $q = Av_k$
      for $j = 1$ to $k$ do
          $h_{j,k} = v_j^* q$
          $q = q - h_{j,k} v_j$
      end for
      $h_{k+1,k} = \|q\|$
      $v_{k+1} = q / h_{k+1,k}$
      Find $c$ to minimize $\| \|r_0\| e_1 - \tilde{H}_k c\| \quad (= \|r_k\|)$
      $x_k = V_k c$
      if $\| \|r_0\| e_1 - \tilde{H}_k c\| / \|r_0\| < t$ then return $x_k$
      end if
  end for
```

1.3 GMRES Analysis

Interestingly, GMRES can be thought of as a polynomial optimization problem. Consider $y \in \mathcal{K}_k(A, r_0)$, then

$$\begin{aligned} y &= c_0 r_0 + c_1 A r_0 + c_2 A^2 r_0 + \ldots + c_{k-1} A^{k-1} r_0 \\ &= (c_0 + c_1 A + c_2 A^2 + \ldots + c_{k-1} A^{k-1}) r_0 \\ &= q(A) r_0, \text{ where } q \text{ is a polynomial of order} \leq k - 1 \end{aligned}$$

and if y_k solves (1.14), then

$$\begin{aligned} r_k = r_0 - Ay_k = r_0 - Aq(A)r_0 &= (I - Aq(A))r_0 \\ &= (I - A(c_0 + c_1 A + c_2 A^2 + \ldots + c_{k-1} A^{k-1}))r_0 \\ &= (I - c_0 A - c_1 A^2 - c_2 A^3 - \ldots + c_{k-1} A^k) r_0 \\ &= p(A) r_0 \end{aligned}$$

where p is a polynomial of order k and $p(0) = I$ or 1.

Thus, GMRES is equivalent to

$$\min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \|p(A)r_0\|$$

at iteration k and we can bound the relative residual as

$$\|r_k\| = \min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \|p(A)r_0\| \leq \min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \|p(A)\|\|r_0\|$$

and thus

$$\frac{\|r_k\|}{\|r_0\|} \leq \min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \|p(A)\|. \tag{1.19}$$

To bound the relative residual then, we shift our focus to what is sometimes referred to as the *Ideal GMRES problem* seen above [8][15].

1.3.1 Some Known Bounds

The minimal polynomial of A provides some insight into GMRES convergence. Here, the minimal polynomial $q_A(z)$ of a matrix A is the unique monic polynomial of least degree that annihilates A [9].

Theorem 1. *Let $A \in \mathbb{C}^{n\times n}$ be invertible, d be the order of the minimal polynomial q_A of A, and r_k be the k-th residual vector produced by the GMRES algorithm in 1. Then, GMRES converges in at most d steps.*

Proof. Note that $q_A(\lambda) = 0$ if and only if λ is an eigenvalue of A [9]. So, for a nonsingular matrix A, $q_A(0) \neq 0$. That is, for a nonsingular matrix A, we can always construct a polynomial $\tilde{p}(z)$ such that $\tilde{p}(0) = 1$ if we let

$$\tilde{p}(z) := \frac{q_A(z)}{q_A(0)}. \tag{1.20}$$

Then, since $\tilde{p}$ is also of order d,

$$\frac{\|r_d\|}{\|r_0\|} \leq \min_{\substack{p \in \mathcal{P}_d \\ p(0)=1}} \|p(A)\| \leq \|\tilde{p}(A)\| = 0. \tag{1.21}$$

Hence, GMRES must converge in at most d steps. □

Furthermore, the interpretation of the GMRES algorithm as a polynomial optimization problem allows us to understand its convergence via spectral sets of the matrix A. First note that when A is normal, $A = U\Lambda U^*$ for unitary U and a diagonal matrix Λ consisting of the eigenvalues of A, the following is true [15]:

$$\min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \|p(A)\| = \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \|Up(\Lambda)U^*\| = \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \max_{\lambda \in \sigma(A)} |p(\lambda)| \tag{1.22}$$

where $\sigma(A)$ is the set of eigenvalues of A. With this in mind, the first bound [15] we present for the relative residual using a spectral set of A is the following:

Theorem 2. *Suppose that A is invertible and diagonalizable, and that $A = V\Lambda V^{-1}$ where Λ is a diagonal matrix consisting of the eigenvalues of A, then*

$$\frac{\|r_k\|}{\|r_0\|} \leq \|V\|\|V^{-1}\| \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \max_{\lambda \in \sigma(A)} |p(\lambda)|. \tag{1.23}$$

Proof. Recall that we know (1.19), and since we also have that

$$\begin{aligned}
\min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \|p(A)\| &= \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \|p(V\Lambda V^{-1})\| \\
&\leq \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \|V\|\|p(\Lambda)\|\|V^{-1}\| \\
&= \|V\|\|V^{-1}\| \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \max_{\lambda \in \sigma(A)} |p(\lambda)|
\end{aligned}$$

where the last equality comes from (1.22), the desired inequality is proven. □

When $\|V\|\|V^{-1}\|$ is large or infinite, the ε-pseudospectra of A provides an alternative approach that can be more descriptive [15]. This ε-pseudospectra of A is denoted $\sigma_\varepsilon(A)$ and defined as

$$\begin{aligned}\sigma_\varepsilon(A) &= \{z \in \mathbb{C} : \exists v \in \mathbb{C}^n \text{ such that } v^*v = 1, \|Av - zv\| < \varepsilon\} &(1.24)\\ &= \{z \in \mathbb{C} : z \in \sigma(A+E) \text{ for some } \|E\| < \varepsilon\} &(1.25)\\ &= \{z \in \mathbb{C} : \left\|(zI - A)^{-1}\right\| > 1/\varepsilon\}. &(1.26)\end{aligned}$$

We use the convention that $\|M^{-1}\| = \infty$ when M is singular. Trefethen and Embree use a Dunford-Taylor integral to arrive at the bound we know for GMRES [15] that uses the pseudospectra of A:

Theorem 3. *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and r_k be the k-th residual vector produced by the GMRES algorithm in 1. Then,*

$$\frac{\|r_k\|}{\|r_0\|} \leq \frac{\mathcal{L}_\varepsilon}{2\pi\varepsilon} \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \max_{z\in\partial\sigma_\varepsilon(A)} |p(z)| \tag{1.27}$$

for $\varepsilon > 0$, where $\mathcal{L}_\varepsilon$ is the arc length of $\partial\sigma_\varepsilon(A)$.

Proof. Let $\Gamma := \partial\sigma_\varepsilon(A)$. Since $\sigma(A)$, which are all the poles of $(zI - A)^{-1}$, are contained in $\sigma_\varepsilon(A)$, then

$$\begin{aligned}\|p(A)\| &= \left\|\frac{1}{2\pi i}\int_\Gamma p(z)(zI - A)^{-1}dz\right\|\\ &\leq \frac{1}{2\pi}\int_\Gamma |p(z)| \left\|(zI - A)^{-1}\right\| |dz|\\ &\leq \frac{1}{2\pi}\max_{z\in\partial\sigma_\varepsilon(A)} |p(z)| \frac{1}{\varepsilon}\mathcal{L}_\varepsilon\\ &= \frac{\mathcal{L}_\varepsilon}{2\pi\varepsilon}\max_{z\in\partial\sigma_\varepsilon(A)} |p(z)|\end{aligned}$$

by the definition of pseudospectra, and the desired inequality follows from (1.19). □

Interestingly, we can also bound the GMRES algorithm using the field of values of a matrix, also known as the numerical range [4][5]. In a paper by Michel Crouzeix [4], he proves the following bound for matrix polynomials:

Theorem 4. *For any matrix $A \in \mathbb{C}^{n\times n}$ and any polynomial $p : \mathbb{C} \to \mathbb{C}$,*

$$\|p(A)\| \le 11.08 \sup_{z\in W(A)} |p(z)| \tag{1.28}$$

where

$$W(A) = \{x^*Ax : x^*x = 1, x \in \mathbb{C}^n\}. \tag{1.29}$$

Knowing this, another bound for GMRES is given by

$$\frac{\|r_k\|}{\|r_0\|} \le 11.08 \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \sup_{z\in W(A)} |p(z)| \tag{1.30}$$

via (1.19).

Crouzeix and Palencia expanded upon this work in a 2017 paper [5] where they develop a tighter bound:

Theorem 5. *For any matrix $A \in \mathbb{C}^{n\times n}$ and any polynomial $p : \mathbb{C} \to \mathbb{C}$,*

$$\|p(A)\| \le \left(1+\sqrt{2}\right) \sup_{z\in W(A)} |p(z)|. \tag{1.31}$$

With this, we have yet another bound for the relative residual:

$$\frac{\|r_k\|}{\|r_0\|} \le \left(1+\sqrt{2}\right) \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \sup_{z\in W(A)} |p(z)| \tag{1.32}$$

once again via (1.19).

In a 1999 seminal paper [6], Bernard and Francois Delyon showed that for a smooth, bounded, convex domain $\Omega \subset \mathbb{C}$, there exists a best constant C_Ω such that for all rational functions f, there holds

$$\|f(A)\| \leq C_\Omega \sup_{z \in \Omega} |f(z)| \tag{1.33}$$

whenever A is a bounded linear operator in a complex Hilbert space $(H, \langle , \rangle, \| \|)$ whose numerical range

$$W(A) := \{\langle Av, v\rangle : v \in H, \|v\| = 1\}$$

satisfies $\overline{W(A)} \subset \Omega$. Though it has been shown in [4] that $\mathcal{Q} := \sup_\Omega C_\Omega$ is $2 \leq \mathcal{Q} \leq 11.08$ and more recently in [5] that $2 \leq \mathcal{Q} \leq 1 + \sqrt{2}$, Crouzeix conjectures [3][4][5] that $\mathcal{Q} = 2$, but it remains an open problem.

CHAPTER II

CONDITIONING AND PRECONDITIONING

This condition number of a linear system $Ax = b$ defined by

$$\kappa(A) = \|A\| \|A^{-1}\|$$

and it measures the sensitivity of the solution to small perturbations in the input data A. We say a problem is well-conditioned when it has a low condition number, and it is ill-conditioned if it has a high condition number. That is, an ill-conditioned problem is more susceptible to large changes in the answer even when there is a small change in the inputs.

However, the idea of preconditioning a system has less to do with it's condition number and more to do with the rate of convergence of the iterative method applied to it. So, if we wish to solve the $m \times m$ nonsingular system

$$Ax = b, \tag{2.1}$$

notice that for any nonsingular $m \times m$ matrix M, the systems

$$M^{-1}Ax = M^{-1}b \tag{2.2}$$

$$\text{and } AM^{-1}Mx = b \tag{2.3}$$

have the same solution. However, if we solve (2.2) iteratively, the convergence will depend on the properties of $M^{-1}A$ rather than those of A. Thus, if the preconditioner M is well chosen, convergence for (2.2) might be much faster than that of (2.1).

Of course, since it must be possible to compute the operation represented by the product $M^{-1}A$ efficiently, we do not explicitly construct the inverse M^{-1}, rather we construct the solution of systems of equations of the form

$$My = c.$$

Additionally, since how well a preconditioner performs depends on the problem. We say a preconditioner M is good if $M^{-1}A$ is not too far from normal and its eigenvalues are clustered [14]. However, since it is hard to discern when a particular preconditioner will be better than another, Wathen [16] when he says

> “There is, of course, no such concept as a best preconditioner: the only two candidates for this would be $P = I$, for which the preconditioning takes no time at all, and $P = A$, for which only one iteration would be required for solution by any iterative method. However, every practitioner knows when they have a good preconditioner which enables feasible computation and solution of problems. In this sense, preconditioning will always be an art rather than a science.”

CHAPTER III

CAN THE FIELD OF VALUES BOUND BE USED IF THE FIELD OF VALUES CONTAINS ZERO?

Recall from Section 1.3.1 that for GMRES applied to any system $Ax = b$ where $A \in \mathbb{C}^{n \times n}$ is nonsingular and $b \in \mathbb{C}^n$,

$$\frac{\|r_k\|}{\|r_0\|} = \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \|p(A)\| \leq C \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \max_{z \in W(A)} |p(z)| \tag{3.1}$$

where $C = 1 + \sqrt{2}$, r_k is the relative residual at iteration k, and $\mathcal{P}_k$ is the set of polynomials of degree k or less. Note that the above follows from [5]

$$\|p(A)\| \leq C \max_{z \in W(A)} |p(z)|. \tag{3.2}$$

We would like to be able to use this bound to describe GMRES convergence, but is it possible to use this bound when $0 \in W(A)$?

Well, suppose $0 \in W(A)$. Then, if we choose $\tilde{p}(z) = 1$, $\tilde{p} \in \mathcal{P}_k$, $\tilde{p}(0) = 1$, and we have

$$\min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \max_{z \in W(A)} |p(z)| \leq \max_{z \in W(A)} |\tilde{p}(z)| \leq 1 \tag{3.3}$$

Furthermore, for any polynomial p such that $p(0) = 1$,

$$1 \leq \max_{z \in W(A)} |p(z)|$$

and thus

$$1 \leq \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \max_{z\in W(A)} |p(z)| \tag{3.4}$$

Hence, by (3.3) and (3.4), we have

$$\frac{\|r_k\|}{\|r_0\|} = \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \|p(A)\| \leq C \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \max_{z\in W(A)} |p(z)| = C \tag{3.5}$$

and we obtain a bound that is not very useful.

However, a paper by Greenbaum and Choi [2] from 2015 tells us that we can, in fact, use the field of values bound when $0 \in W(A)$. To do this, we want to use the roots of A. Notice that if we let $B = A^{1/\ell}$, $B^{\ell} = A$, and if $p(A)$ is a polynomial of order k, then $p(A) = p(B^{\ell}) = q(B)$ where q is a polynomial of order ℓk and has only powers that are multiples of ℓ. Thus, we have

$$\begin{aligned} \|p(A)\| = \|p(B^{\ell})\| = \|q(B)\| &\leq C \max_{z\in W(B)} |q(z)| \\ &= C \max_{z\in W(A^{1/\ell})} |p(z^{\ell})| \\ &= C \max_{z\in (W(A^{1/\ell}))^{\ell}} |p(z)| \end{aligned}$$

In particular, notice that if $(W(A^{1/\ell}))^{\ell}$ can be contained in a ball centered at c with radius R, then we may be able to use this idea to give us a practical bound for the GMRES algorithm. Also of note is the fact that Greenbaum and Choi prove in [2] that for any nonsingular matrix A and any positive integer ℓ, $\lim_{\ell\to\infty}[W(A^{1/\ell})]^{\ell} = \exp[W(\log A)]$ where $\log A$ is defined using the same branch cut used to define the ℓ-th root of A. Using the inequality above, we have that

$$\frac{\|r_k\|}{\|r_0\|} \leq C \min_{\substack{p\in\mathcal{P}_k \\ p(0)=1}} \max_{z\in (W(A^{1/\ell}))^{\ell}} |p(z)| \tag{3.6}$$

and if we choose $\tilde{p}(z) = \left(1 - \frac{z}{c}\right)^k$, then $\tilde{p} \in \mathcal{P}_k$, $\tilde{p}(0) = 1$, and

$$\begin{aligned}\tilde{p}(c + Re^{i\theta}) &= \left(1 - \frac{c + Re^{i\theta}}{c}\right)^k \\ &= \left(1 - 1 - \frac{Re^{i\theta}}{c}\right)^k \\ &= \frac{R^k e^{ik\theta}}{c^k}.\end{aligned}$$

Thus, for any $z \in \partial B_R(c)$

$$|\tilde{p}(z)| = \left(\frac{R}{|c|}\right)^k$$

and if $(W(A^{1/\ell}))^\ell \subset B_R(c)$ where $B_R(c)$ is a ball of radius R centered at c, we have

$$\begin{aligned}C \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \max_{z \in (W(A^{1/\ell}))^\ell} |p(z)| &\leq C \min_{\substack{p \in \mathcal{P}_k \\ p(0)=1}} \max_{z \in \partial B_R(c)} |p(z)| && (3.7) \\ &\leq C \max_{z \in \partial B_R(c)} |\tilde{p}(z)| && (3.8) \\ &= C \left(\frac{R}{|c|}\right)^k. && (3.9)\end{aligned}$$

Hence, we hope to use a ball with $R < |c|$ containing $(W(A^{1/\ell}))^\ell$ but not 0 to bound the rate of convergence of the GMRES algorithm since otherwise this bound is not useful.

Unfortunately, in our tests, we could not find any such ball for multiple values of ℓ. Figure 3.1 shows the tests for matrices $\mathcal{A}\mathcal{P}^{-1} \in \mathbb{R}^{55 \times 55}$ for $\mathcal{A}$ of the form $\mathcal{A} = I + 1.5\mathcal{B}$ where I is the 55×55 identity matrix and $\mathcal{B} \in \mathbb{R}^{55 \times 55}$ is a random normalized matrix, and for $\mathcal{P}$ as described in (1.11). Figure 3.2 shows the test for matrices $\mathcal{A}_0\mathcal{P}_+^{-1} \in \mathbb{R}^{55 \times 55}$ for $\mathcal{A}$ as defined in (1.3) and $\mathcal{P}_+$ as defined in (1.4) and where $A \in \mathbb{R}^{50 \times 50}$ is an orthogonal random matrix, $B \in \mathbb{R}^{5 \times 50}$ is a random matrix, and $C \in \mathbb{R}^{5 \times 50}$ is a random normalized matrix. In both of these figures, the eigenvalues of the base matrix are shown as red dots and the origin is shown as a blue dot.

In fact, these figures provide some intuition for the following theorem.

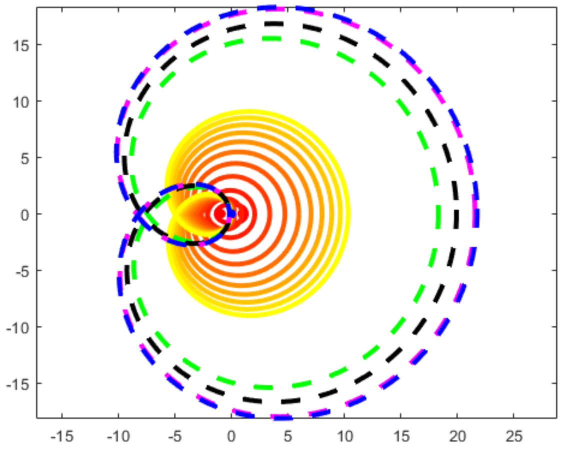

Figure 3.1: $(W((\mathcal{A}\mathcal{P}^{-1})^{1/\ell}))^{\ell}$ for values of $\ell = 1,2,\ldots,10$ where the line with the deepest red is $\ell = 1$, the one with the brightest yellow is $\ell = 10$, the green dashed line is $\ell = 50$, the black dashed line is $\ell = 100$, the magenta dashed line is $\ell = 1000$, and the blue dashed line is $\exp[W(\log(\mathcal{A}\mathcal{P}^{-1}))]$.

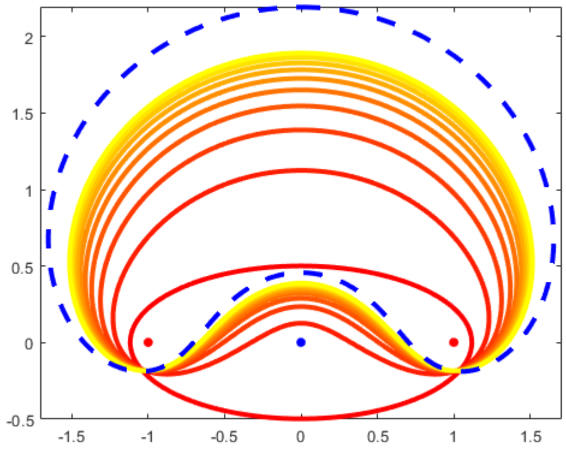

Figure 3.2: $(W((\mathcal{A}_0\mathcal{P}_+^{-1})^{1/\ell}))^{\ell}$ for values of $\ell = 1,2,\ldots,10$ where the line with the deepest red is $\ell = 1$, the one with the brightest yellow is $\ell = 10$, and the blue dashed line is $\exp[W(\log(\mathcal{A}_0\mathcal{P}_+^{-1}))]$.

Theorem 6. *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $A^{1/\ell}$ be the matrix such that $(A^{1/\ell})^{\ell} = A$. Then, if the eigenvalues of A cannot be contained by a ball $B_R(c)$ of radius R centered at c such that $0 \notin B_R(c)$, then there does not exist such a ball to contain $(W(A^{1/\ell}))^{\ell}$ but not 0.*

Proof. Notice that the eigenvalues of A must be contained in $(W(A^{1/\ell}))^{\ell}$ for any ℓ. Indeed, suppose $\lambda_{A^{1/\ell}}$ is an eigenvalue of $A^{1/\ell}$ and λ_A is an eigenvalue of A, then for some nonzero vector x

$$\lambda_A x = Ax = (A^{1/\ell})^\ell x = \lambda_{A^{1/\ell}}^\ell x.$$

Thus, $\lambda_{A^{1/\ell}} = \lambda_A^{1/\ell}$. Then, since $\lambda_A^{1/\ell} \in W(A^{1/\ell})$, $\lambda_A \in (W(A^{1/\ell}))^\ell$.

Hence, if there does not exist a ball that can contain the eigenvalues of A but not 0, there does not exist a ball to contain $(W(A^{1/\ell}))^\ell$ but not 0. □

CHAPTER IV

RELATIVE RESIDUAL BOUND

We are attempting to answer the question of how many more steps does it take for the GMRES algorithm to converge when applied to the preconditioned linear system $\mathcal{A}\mathcal{P}^{-1}x = b$ given that the $(2{,}2)$ block in $\mathcal{A}$, D, is nonzero. To do this, we can think of $\mathcal{A}\mathcal{P}^{-1}$ as a perturbation of the preconditioned matrix $\mathcal{A}_0\mathcal{P}^{-1}$. To put it in more general terms, if we have a fixed perturbation $E = \mathcal{A}\mathcal{P}^{-1} - \mathcal{A}_0\mathcal{P}^{-1}$, and if

$$\|r_k\| = \min_{\substack{p\in\mathcal{P}_k\\ p(0)=1}} \|p(\mathcal{A}_0\mathcal{P}^{-1})b\| \quad \text{and} \quad \|\rho_k\| = \min_{\substack{p\in\mathcal{P}_n\\ p(0)=1}} \|\phi(\mathcal{A}_0\mathcal{P}^{-1}+E)b\|$$

are the residuals produced by GMRES applied to

$$\mathcal{A}_0\mathcal{P}^{-1}x = b \quad \text{and} \quad (\mathcal{A}_0\mathcal{P}^{-1}+E)x = b,$$

respectively, how far does the residual ρ_k lag behind r_k? To answer this question, we need to find the norm of the difference between these residuals, but a bound can also be enlightening.

Fortunately, in a paper from 2017, Ymbert, Embree, and Sifuentes [17] show that for any $\delta > \epsilon := \|E\|$,

$$\frac{\|\rho_k\|}{\|b\|} \le \frac{\|r_k\|}{\|b\|} + \left(\frac{\epsilon}{\delta-\epsilon}\right)\left(\frac{L_\delta}{2\pi\delta}\right) \sup_{z\in\partial\sigma_\delta(\mathcal{A}_0\mathcal{P}^{-1})} |p_k(z)| \tag{4.1}$$

where L_δ is the length of the boundary $\partial\sigma_\delta(\mathcal{A}_0\mathcal{P}^{-1})$ of the δ-pseudospectrum of $\mathcal{A}_0\mathcal{P}^{-1}$

and p_k is some polynomial of degree k or less satisfying $p_k(0) = 1$ for which the GMRES residual vector at step k takes the form

$$r_k = p_k(\mathcal{A}_0\mathcal{P}^{-1})b.$$

Further, since $\mathcal{A}_0\mathcal{P}^{-1}$ has a low-degree polynomial for any of the preconditioners described before, GMRES will converge exactly in just a few steps. Let d denote the degree of this minimal polynomial. Then, if $k \geq d$, $r_k = 0$ and we can write [17] $r_k = p_k(\mathcal{A}_0\mathcal{P}^{-1})b$ for any polynomial of the form

$$p_k(z) = \alpha(z)q(z)$$

where α is the degree d polynomial that annihilates $\mathcal{A}_0\mathcal{P}^{-1}$ and q is a polynomial of degree $k - d$ or less, with $\alpha(0) = q(0) = 1$. As such, when $k \geq d$, (4.1) bounds the residual remaining for the perturbed preconditioned problem [17]:

$$\frac{\|\rho_k\|}{\|b\|} \leq \left(\frac{\epsilon}{\delta - \epsilon}\right)\left(\frac{L_\delta}{2\pi\delta}\right) \min_{\substack{deg(q) \leq k-d \\ q(0)=1}} \sup_{z \in \partial\sigma_\delta(\mathcal{A}_0\mathcal{P}^{-1})} |\alpha(z)||q(z)| \tag{4.2}$$

Ymbert, Embree, and Sifuentes [17] expand upon this further by assembling a bound for the preconditioner (1.6) via estimates of the pseudospectra of $\mathcal{A}_0\mathcal{P}^{-1}$ using the spectral projectors Π_j for its eigenvalues λ_j. Specifically, if we let $\|\Pi_j\| \leq \kappa_j$, then

$$\sigma_\delta(\mathcal{A}_0\mathcal{P}^{-1}) \subseteq S_\delta := \bigcup_{j=1}^{3} \{z \in \mathbb{C} : |z - \lambda_j| \leq 3\delta\kappa_j\}$$

and if $\|\mathcal{A}\mathcal{P}^{-1} - \mathcal{A}_0\mathcal{P}^{-1}\| \leq \epsilon' < \delta$, for $k \geq 3$

$$\frac{\|\rho_k\|}{\|b\|} \leq 3\left(\frac{\epsilon'}{\delta - \epsilon'}\right)(\kappa_1 + \kappa_2 + \kappa_3) \sup_{z \in \partial S_\delta} |p_k(z)| \tag{4.3}$$

where

$$p_k(z) = (1-z)^{d_1}\left(1-\frac{z}{\phi}\right)^{d_2}\left(1-\frac{z}{1-\phi}\right)^{d_3}$$

and $d_1 + d_2 + d_3 = k$ are chosen to minimize $|p_k(z)|$ over ∂S_δ. We can apply this bound to our prpblem for own bound on the relative residual:

Theorem 7. *Let $\mathcal{A} \in \mathbb{C}^{(n+m)\times(n+m)}$ be the 2×2 block matrix defined in (1.8), $\mathcal{A}_0 \in \mathbb{C}^{(n+m)\times(n+m)}$ be the matrix defined in (1.3), $\mathcal{P}$ be the preconditioner defined in (1.11), Π_j be the spectral projectors of $\mathcal{A}_0\mathcal{P}^{-1}$, and ρ_k be the residual at iteration k of GMRES applied to the system $\mathcal{A}\mathcal{P}^{-1}x = b$. If $\|\Pi_j\| \leq \kappa_j$ and $\varepsilon := \|\mathcal{A}\mathcal{P}^{-1} - \mathcal{A}_0\mathcal{P}^{-1}\| < \delta$, then*

$$\frac{\|\rho_k\|}{\|b\|} \leq 3\left(\frac{\varepsilon}{\delta-\varepsilon}\right)(\kappa_1+\kappa_2+\kappa_3)\sup_{z\in\partial S_\delta}|p_k(z)| \tag{4.4}$$

Proof. The proof follows immediately from the inequality (4.3). □

Thus, if we can bound $\|\mathcal{A}\mathcal{P}^{-1} - \mathcal{A}_0\mathcal{P}^{-1}\|$, then we can bound the relative residual for GMRES applied to the inexact preconditioned problem.

4.1 A Bound for the Difference Between the Perturbed System and the Unperturbed One

We have the preconditioned systems

$$\mathcal{A}\mathcal{P}^{-1} = \begin{pmatrix} I & -B^*S^{-1} \\ CA^{-1} & -DS^{-1} \end{pmatrix}$$

and

$$\mathcal{A}_0\mathcal{P}^{-1} = \begin{pmatrix} I & -B^*S_0^{-1} \\ CA^{-1} & 0 \end{pmatrix}$$

where

$$\mathcal{A} = \begin{pmatrix} A & B^* \\ C & D \end{pmatrix},$$

$\mathcal{A}\mathcal{P}^{-1}$ is the preconditioned matrix with eigenvalues 1, the golden ratio, and its conjugate, and

$$S_0 = -CA^{-1}B^*$$

$$S = D - CA^{-1}B^*$$

Theorem 8. *Let $\mathcal{A} \in \mathbb{C}^{(n+m)\times(n+m)}$ be the 2×2 block system defined in (1.8), $\mathcal{P}$ be the preconditioner defined in (1.11), $\delta' := \|D\|$, and $\gamma := \|S_0^{-1}\|$. If $\|DS_0^{-1}\| < 1$, then*

$$\varepsilon := \left\|\mathcal{A}\mathcal{P}^{-1} - \mathcal{A}_0\mathcal{P}^{-1}\right\| \leq \frac{\delta'\gamma\sqrt{\|B^*\|^2\gamma^2 + 1}}{1 - \delta'\gamma}.$$

Proof. We have

$$\begin{aligned}
\|\mathcal{A}\mathcal{P}^{-1} - \mathcal{A}_0\mathcal{P}^{-1}\| &= \left\| \begin{pmatrix} I & -B^*S^{-1} \\ CA^{-1} & -DS^{-1} \end{pmatrix} - \begin{pmatrix} I & -B^*S_0^{-1} \\ CA^{-1} & 0 \end{pmatrix} \right\| \\
&= \left\| \begin{pmatrix} 0 & -B^*S^{-1} + B^*S_0^{-1} \\ 0 & -DS^{-1} \end{pmatrix} \right\| \\
&\leq \sqrt{\left\|-B^*S^{-1} + B^*S_0^{-1}\right\|^2 + \|-DS^{-1}\|^2} \\
&= \sqrt{\left\|B^*S_0^{-1} - B^*S^{-1}\right\|^2 + \|DS^{-1}\|^2}
\end{aligned}$$

So, notice

$$\begin{aligned}
B^*S_0^{-1} - B^*S^{-1} &= B^*S_0^{-1} - B^*(D + S_0)^{-1} \\
&= B^*(S_0^{-1} - (D + S_0)^{-1}) \\
&= B^*(S_0^{-1} - ((DS_0^{-1} + I)S_0)^{-1}) \\
&= B^*(S_0^{-1} - S_0^{-1}(I + DS_0^{-1})^{-1}) \\
&= B^*S_0^{-1}(I - (I + DS_0^{-1})^{-1})
\end{aligned}$$

Since $\left\| DS_0^{-1} \right\| < 1$, then

$$(I + DS_0^{-1})^{-1} = \sum_{j=0}^{\infty} (-DS_0^{-1})^j$$

Now,

$$\begin{aligned}\left\| I - (I + DS_0^{-1})^{-1} \right\| &= \left\| \sum_{j=1}^{\infty} (-DS_0^{-1})^j \right\| \\ &\leq \sum_{j=1}^{\infty} (\delta' \gamma)^j \\ &= \frac{1}{1 - \delta'\gamma} - \frac{1 - \delta'\gamma}{1 - \delta'\gamma} \\ &= \frac{\delta'\gamma}{1 - \delta'\gamma}\end{aligned}$$

Therefore,

$$\left\| B^* S_0^{-1} - B^* S^{-1} \right\| \leq \left\| B^* S_0^{-1} \right\| \left\| I - (I + DS_0^{-1})^{-1} \right\| \leq \frac{\|B^*\| \left\| S_0^{-1} \right\| \delta'\gamma}{1 - \delta'\gamma} = \frac{\|B^*\| \delta'\gamma^2}{1 - \delta'\gamma}$$

Additionally,

$$\begin{aligned}\left\| DS^{-1} \right\| &= \left\| D(D + S_0)^{-1} \right\| \\ &= \left\| D(S_0(S_0^{-1}D + I))^{-1} \right\| \\ &= \left\| D(I + S_0^{-1}D)^{-1} S_0^{-1} \right\| \\ &\leq \|D\| \left\| (I + S_0^{-1}D)^{-1} \right\| \left\| S_0^{-1} \right\| \\ &\leq \delta'\gamma \frac{1}{1 - \delta'\gamma} \\ &= \frac{\delta'\gamma}{1 - \delta'\gamma}\end{aligned}$$

Hence,

$$\sqrt{\left\|B^*S_0^{-1} - B^*S^{-1}\right\|^2 + \|DS^{-1}\|^2} \leq \sqrt{\frac{\|B^*\|^2 \delta'^2\gamma^4}{(1-\delta'\gamma)^2} + \frac{\delta'^2\gamma^2}{(1-\delta'\gamma)^2}}$$
$$= \frac{\delta'\gamma\sqrt{\|B^*\|^2\gamma^2+1}}{1-\delta'\gamma}$$

That is,

$$\left\|\mathcal{A}\mathcal{P}^{-1} - \mathcal{A}_0\mathcal{P}^{-1}\right\| \leq \frac{\delta'\gamma\sqrt{\|B^*\|^2\gamma^2+1}}{1-\delta'\gamma}.$$

□

4.2 The Bound

Finally then, we can construct a more specific bound for the relative residual of GMRES applied to the problem preconditioned by (1.11).

Theorem 9. *Let $\mathcal{A} \in \mathbb{C}^{(n+m)\times(n+m)}$ be the 2×2 block matrix defined in (1.8), $\mathcal{A}_0 \in \mathbb{C}^{(n+m)\times(n+m)}$ be the matrix defined in (1.3), $\mathcal{P}$ be the preconditioner defined in (1.11), $\delta' := \|D\|$, $\gamma := \|S_0^{-1}\|$, Π_j be the spectral projectors of $\mathcal{A}_0\mathcal{P}^{-1}$, and ρ_k be the residual at iteration k of GMRES applied to the system $\mathcal{A}\mathcal{P}^{-1}x = b$. If $\|DS_0^{-1}\| < 1$, $\|\Pi_j\| \leq \kappa_j$, and $\frac{\delta'\gamma\sqrt{\|B^*\|^2\gamma^2+1}}{1-\delta'\gamma} < \delta$, then*

$$\frac{\|\rho_k\|}{\|b\|} \leq 3\left(\frac{\delta'\gamma\sqrt{\|B^*\|^2\gamma^2+1}}{\delta - \delta\delta'\gamma - \delta'\gamma\sqrt{\|B^*\|^2\gamma^2+1}}\right)(\kappa_1+\kappa_2+\kappa_3)\sup_{z\in\partial S_\delta}|p_k(z)| \tag{4.5}$$

Proof. The proof follows from theorem 7. □

4.3 An Important Result

Theorem 10. *Suppose $\mathcal{A}_0$ and $\mathcal{P}$ defined in (1.3) and (1.11) are both invertible. If $m < n$, the right and left invariant subspaces $\mathfrak{R}_1$ and $\mathcal{L}_1$ of $\mathcal{A}\mathcal{P}^{-1}$ for $\mathcal{A}$ defined in (1.8) and associated with $\lambda_1 = 1$*

have dimension $n - m$:

$$\mathcal{R}_1 = \left\{ \begin{bmatrix} Az \\ 0 \end{bmatrix} : z \in \operatorname{Ker}(C) \right\}, \quad \mathcal{L}_1 = \left\{ \begin{bmatrix} z \\ 0 \end{bmatrix} : z \in \operatorname{Ker}(B) \right\}. \tag{4.6}$$

Proof. Since $\mathcal{A}_0$ is invertible, B and C must have full row rank. Recall that the Ipsen preconditioned matrix is

$$\mathcal{A}\mathcal{P}^{-1} = \begin{bmatrix} I & -B^*S^{-1} \\ CA^{-1} & -DS^{-1} \end{bmatrix}$$

where $S = D - CA^{-1}B^*$.

To compute $\mathcal{R}_1$, note that a right eigenvector $u = [x^T y^T]^T$ where $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^m$, associated with $\lambda_1 = 1$ satisfies

$$\begin{aligned} x - B^*S^{-1}y &= x \\ CA^{-1} - DS^{-1}y &= y \end{aligned}$$

Since B has full row rank and S is invertible, the first equation implies $y = 0$. That reduces the second equation to the condition $A^{-1}x \in \operatorname{Ker}(C)$, establishing the form for $\mathcal{R}_1$ in (4.6). The computation for $\mathcal{L}_1$ is similar: the left eigenvector $v = [r^T s^T]^T$ where $r \in \mathbb{C}^n$ and $s \in \mathbb{C}^m$ satisfies $(\mathcal{A}\mathcal{P}^{-1})^*$, that is,

$$\begin{aligned} r + A^{-*}C^*s &= r \\ -S^{-*}Br - S^{-*}Ds &= s \end{aligned}$$

Since C has full row rank and A is invertible, the first equation implies $s = 0$. That reduces the second equation to the condition $r \in \operatorname{Ker}(B)$, and thus we have the formula for $\mathcal{L}_1$ in (4.6). Since $B, C \in \mathbb{C}^{m \times n}$ have full row rank and $m \leq n$, $\dim(\operatorname{Ker}(B)) = \dim(\operatorname{Ker}(C)) = n - m$, establishing the dimensions of $\mathcal{R}_1$ and $\mathcal{L}_1$. □

Theorem 11. *Suppose $\mathcal{A}_0$ and $\mathcal{P}$ defined in (1.3) and (1.11) are both invertible. If $m < n$, then the GMRES algorithm applied to the system $\mathcal{A}\mathcal{P}^{-1}x = b$ for $\mathcal{A}$ defined in (1.8) converges in $2m + 1$ steps or less.*

Proof. Since $\mathcal{A}\mathcal{P}^{-1}$ is invertible, theorem 1 tells us that GMRES converges in an amount of steps that is at most equivalent to the degree of the minimal polynomial of $\mathcal{A}\mathcal{P}^{-1}$.

So, to find the minimal polynomial of $\mathcal{A}\mathcal{P}^{-1}$, consider its Jordan canonical form. Since $\mathcal{A}_0$ and $\mathcal{P}$ are invertible and $m < n$, then from theorem 10 we know that the geometric multiplicity of the eigenvalue $\lambda_1 = 1$ of $\mathcal{A}\mathcal{P}^{-1}$ is $n - m$. Let the algebraic multiplicity of $\lambda_1 = 1$ be r. Then $r \geq n - m$, and since the Jordan canonical form will have a 1 just above the its main diagonal corresponding to each eigenvalue with a missing eigenvector, the Jordan canonical form of $\mathcal{A}\mathcal{P}^{-1}$ can be written

$$J = T^{-1}AT = \begin{bmatrix} I_{n-m-1} & & \\ & J_1 & \\ & & J_2 \end{bmatrix} \tag{4.7}$$

where T is the matrix containing all eigenvectors and generalized eigenvectors corresponding to the eigenvalues of $\mathcal{A}\mathcal{P}^{-1}$, I_{n-m-1} is the $(n-m-1) \times (n-m-1)$ identity matrix,

$$J_1 = \begin{bmatrix} 1 & 1 & & \\ & 1 & \ddots & \\ & & \ddots & 1 \\ & & & 1 \end{bmatrix} \in \mathbb{C}^{(r-(n-m)+1)\times(r-(n-m)+1)}, \tag{4.8}$$

and $J_2 \in \mathbb{C}^{(n+m-r)\times(n+m-r)}$ is the matrix containing all other Jordan blocks corresponding to the eigenvalues of $\mathcal{A}\mathcal{P}^{-1}$. Note that if the eigenvalue $\lambda_1 = 1$ is nondefective, $J_1 = 1$ and $J_2 \in \mathbb{C}^{2m\times 2m}$.

The minimal polynomial p_J of such a matrix is

$$p_J(z) = (1-z)(1-z)^{r-(n-m)} q_{J_2}(z) = (1-z)^{r-(n-m)+1} q_{J_2}(z) \tag{4.9}$$

where q_{J_2} is the minimal polynomial of J_2 which must have order $n+m-r$ or less. Indeed, notice

$$p_J(z) = T p_J(J) T^{-1}$$

and

$$\begin{aligned}
p_J(J) &= (I-J)^{r-(n-m)+1} q_{J_2}(J) \\
&= \begin{bmatrix} 0 & & \\ & (I-J_1)^{r-(n-m)+1} & \\ & & (I-J_2)^{r-(n-m)+1} \end{bmatrix} \begin{bmatrix} q_{J_2}(I) & & \\ & q_{J_2}(J_1) & \\ & & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & & \\ & 0 & \\ & & (I-J_2)^{r-(n-m)+1} \end{bmatrix} \begin{bmatrix} q_{J_2}(I) & & \\ & q_{J_2}(J_1) & \\ & & 0 \end{bmatrix} \\
&= 0
\end{aligned}$$

since $r-(n-m)+1$ is the size of $I-J_1$ and

$$(I-J_1)^{r-(n-m)+1} = \begin{bmatrix} 0 & -1 & & \\ & 0 & \ddots & \\ & & \ddots & -1 \\ & & & 0 \end{bmatrix}^{r-(n-m)+1} = 0.$$

Thus, since the degree of q_{J_2} is at most $n + m - r$,

$$\begin{aligned} \deg(p_J) &\leq r - (n - m) + 1 + n + m - r \\ &= 2m + 1. \end{aligned}$$

Hence, GMRES converges in at most $2m + 1$ steps. □

CHAPTER V

NUMERICAL EXAMPLES

In this chapter, we apply the bounds developed in Chapter IV to matrices $\mathcal{A}$ of the form described in (1.8) that have been preconditioned with $\mathcal{P}$ as defined in (1.11). In figures 5.1-5.8, these matrices are randomly generated with $A \in \mathbb{R}^{50\times 50}$, $B, C \in \mathbb{R}^{50\times 5}$, and $D \in \mathbb{R}^{5\times 5}$, the solid black line is the relative residual produced by GMRES applied to the system $\mathcal{A}_0\mathcal{P}^{-1}x = b$, and the dashed line is the residual produced by GMRES applied to the perturbed preconditioned system $\mathcal{A}\mathcal{P}^{-1}x = b$. In all of the examples where bound (4.4) or (4.5) is used, $\kappa_j = \|\Pi_j\|$ and

$$\mu = \frac{\delta'\gamma\sqrt{\|B^*\|^2\gamma^2+1}}{1-\delta'\gamma}.$$

Figures 5.1 and 5.2 demonstrate the bounds (4.4) and (4.5), respectively, on the same randomly generated matrix. Figures 5.3 and 5.4 demonstrate the bounds (4.4) and (4.5), respectively, on the same randomly generated matrix. Figures 5.5 and 5.6 demonstrate the bounds (4.4) and (4.5), respectively, on the same randomly generated matrix. Figures 5.7 and 5.8 demonstrate the bounds (4.4) and (4.5), respectively, on the same randomly generated matrix.

In figures 5.9-5.13, the dashed lines are the residuals produced by GMRES applied to the preconditioned system $\mathcal{A}\mathcal{P}^{-1}x = b$ for different values of n and m and the red line is the vertical line at the $2m+1$ iteration. That is, these figures are demonstrating the bound on GMRES convergence as described in theorem 11. In figures 5.12-5.13, $\mathcal{A} \in \mathbb{R}^{1000\times 1000}$ and δ' ranges from 10^{-10} to 10.

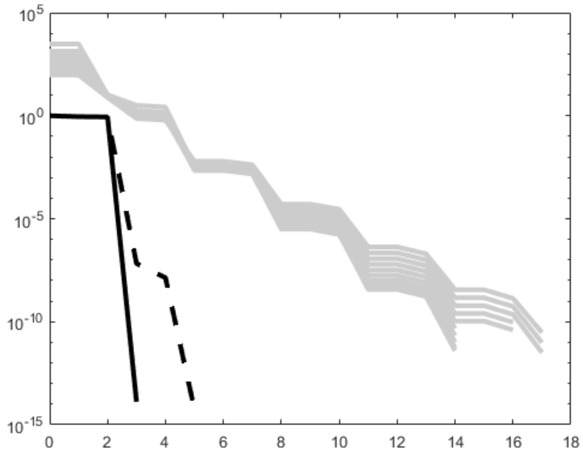

Figure 5.1: Here, $\varepsilon = 1.80 \times 10^{-8}$, $\delta' = 10^{-8}$, and the gray lines are the bounds using (4.3) using $\delta = \varepsilon 10^j$ for $j = 0.1, 0.2, \ldots 1$.

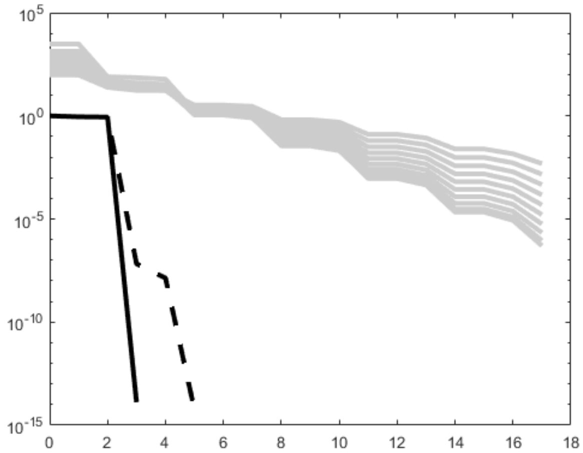

Figure 5.2: Here, $\varepsilon = 1.80 \times 10^{-8}$, $\delta' = 10^{-8}$, and the gray lines are the bounds using (4.3) using $\delta = \mu 10^j$ for $j = 0.1, 0.2, \ldots 1$.

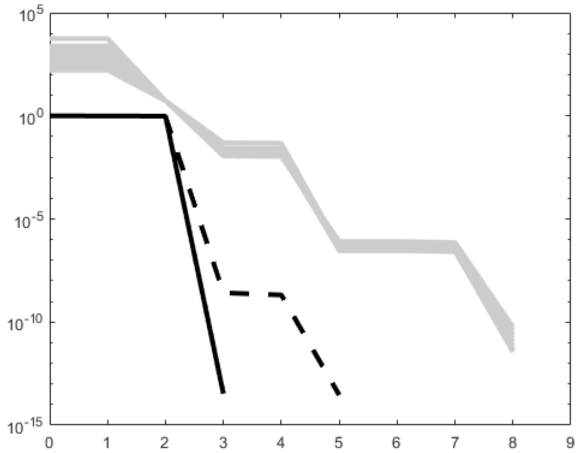

Figure 5.3: Here, $\varepsilon = 9.53 \times 10^{-9}$, $\delta' = 10^{-9}$, and the gray lines are the bounds using (4.3) using $\delta = \varepsilon 10^{j}$ for $j = 0.1, 0.2, \ldots 1$.

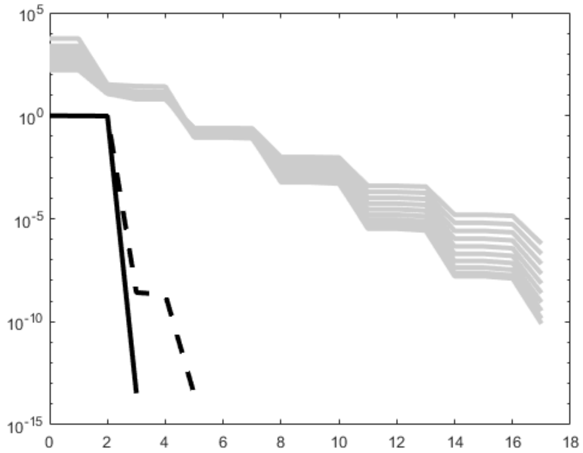

Figure 5.4: Here, $\varepsilon = 9.53 \times 10^{-9}$, $\delta' = 10^{-9}$, and the gray lines are the bounds using (4.3) using $\delta = \mu 10^{j}$ for $j = 0.1, 0.2, \ldots 1$.

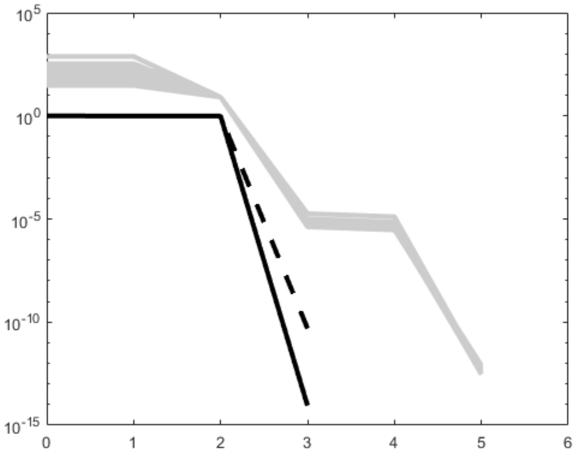

Figure 5.5: Here, $\varepsilon = 1.52 \times 10^{-10}$, $\delta' = 10^{-10}$, and the gray lines are the bounds using (4.3) using $\delta = \varepsilon 10^j$ for $j = 0.1, 0.2, \dots 1$.

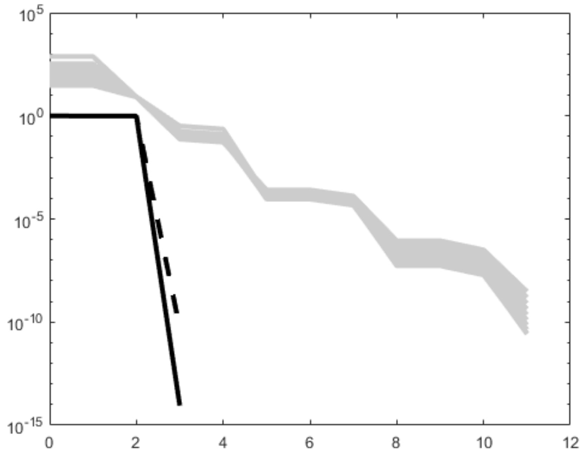

Figure 5.6: Here, $\varepsilon = 1.52 \times 10^{-10}$, $\delta' = 10^{-10}$, and the gray lines are the bounds using (4.3) using $\delta = \mu 10^j$ for $j = 0.1, 0.2, \dots 1$.

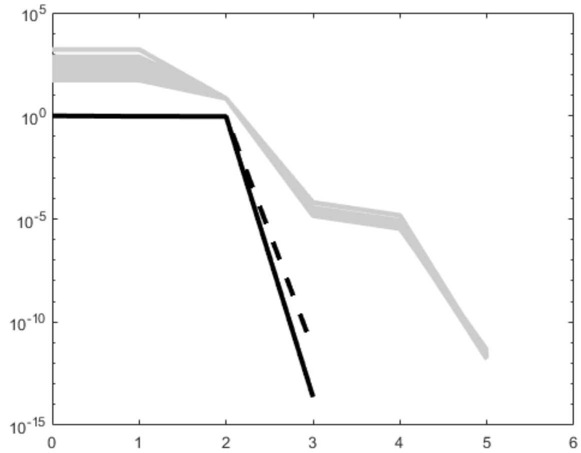

Figure 5.7: Here, $\varepsilon = 5.79 \times 10^{-11}$, $\delta' = 10^{-11}$, and the gray lines are the bounds using (4.3) using $\delta = \varepsilon 10^j$ for $j = 0.1, 0.2, \ldots 1$.

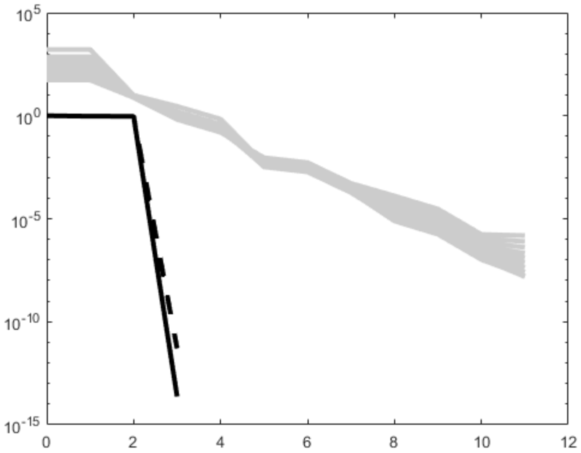

Figure 5.8: Here, $\varepsilon = 5.79 \times 10^{-11}$, $\delta' = 10^{-11}$, and the gray lines are the bounds using (4.3) using $\delta = \mu 10^j$ for $j = 0.1, 0.2, \ldots 1$.

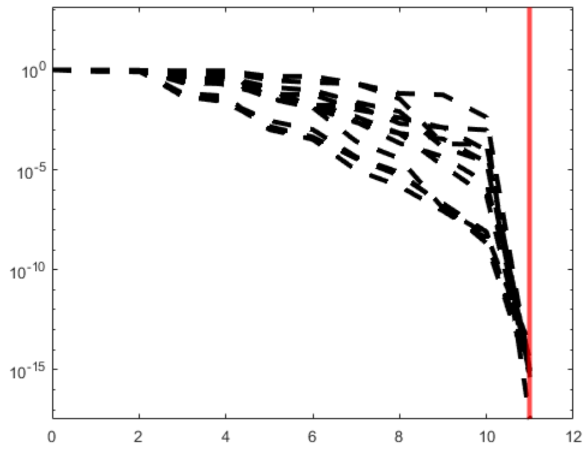

Figure 5.9: Here, $\delta' = 3.99$, $m = 5$, and $n = 5 \times 2^\ell$ for $\ell = 1, \cdots, 12$.

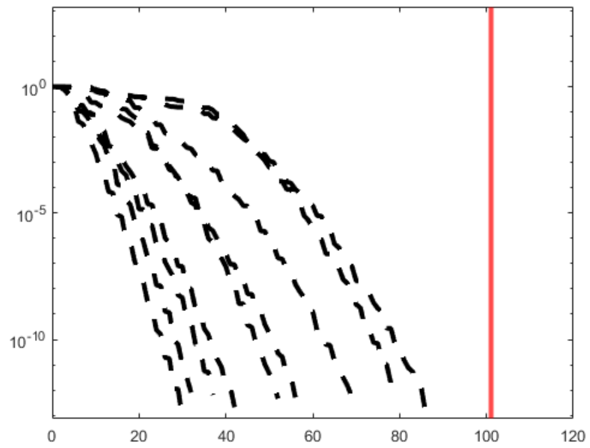

Figure 5.10: Here, $\delta' = 14.71$, $m = 50$, and $n = 5 \times 2^\ell$ for $\ell = 4, \cdots, 12$.

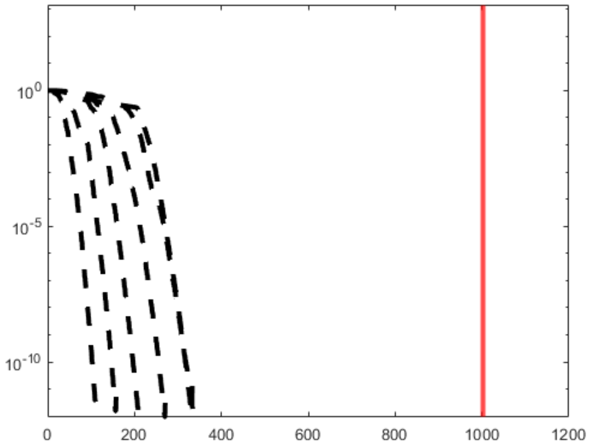

Figure 5.11: Here, $\delta' = 44.30$, $m = 500$, and $n = 5 \times 2^{\ell}$ for $\ell = 7, \cdots, 12$.

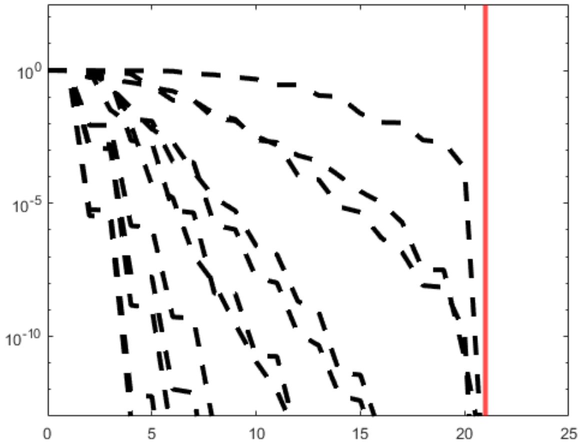

Figure 5.12: Here, $m = 10$, $n = 990$, and $\delta' = 10^{\ell}$ for $\ell = -10, \cdots, 1$.

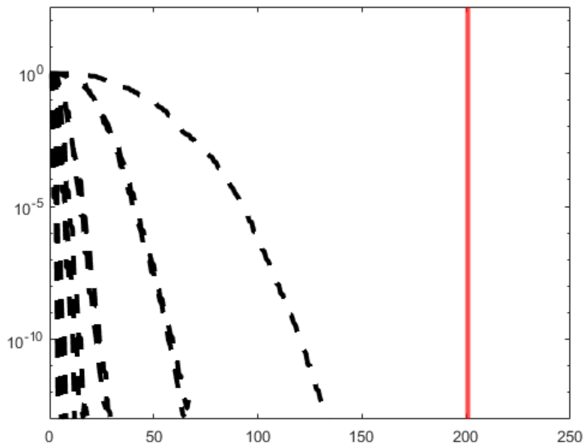

Figure 5.13: Here, $m = 100$, $n = 900$, and $\delta' = 10^{\ell}$ for $\ell = -10, \cdots, 1$.

REFERENCES

[1] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004, ch. 5, pp. 243–246.

[2] D. Choi and A. Greenbaum, *Roots of matrices in the study of gmres convergence and crouzeix's conjecture*, SIAM Journal on Matrix Analysis and Applications, 36 (2015), pp. 289–301.

[3] M. Crouzeix, *Bounds for analytical functions of matrices*, Integral Equations and Operator Theory, 48 (2004), pp. 461–477.

[4] M. Crouzeix, *Numerical range and functional calculus in hilbert space*, Journal of Functional Analysis, 244 (2007), pp. 668–690.

[5] M. Crouzeix and C. Palencia, *The numerical range as a spectral set*, 2017.

[6] B. Delyon and F. Delyon, *Generalization of von neumann's spectral sets and integral representation of operators*, Bulletin de la Société Mathématique de France, 127 (1999), pp. 25–41.

[7] H. C. Elman, A. Ramage, and D. J. Silvester, *Algorithm 866: Ifiss, a matlab toolbox for modelling incompressible flow*, ACM Transactions on Mathematical Software, 33 (2007).

[8] A. Greenbaum and L. N. Trefethen, *Gmres/cr and arnoldi/lanczos as matrix approximation problems*, SIAM Journal on Scientific Computing, 15 (1994), pp. 359–368.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge; New York, 2nd ed., 2013.

[10] I. C. F. Ipsen, *A note on preconditioning nonsymmetric matrices*, SIAM Journal on Scientific Computing, 23 (2001), pp. 1050–1051.

[11] M. F. Murphy, G. H. Golub, and A. J. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM Journal on Scientific Computing, 21 (2000), pp. 1969–1972.

[12] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Series in Operations Research and Financial Engineering, Springer, 2 ed., 2006.

[13] Y. Saad, *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, second ed., 2003.

[14] L. N. TREFETHEN AND D. BAU III, *Numerical Linear Algebra*, Society for Industrial and Applied Mathematics, 1997, pp. 313–314.

[15] L. N. TREFETHEN AND M. EMBREE, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, 2005, pp. 248–249.

[16] A. J. WATHEN, *Preconditioning*, Acta Numerica, 24 (2015), p. 329–376.

[17] G. YMBERT III, M. EMBREE, AND J. A. SIFUENTES, *Approximate Murphy-Golub-Wathen Preconditioning for Saddle Point Problems*, 2017.

BIOGRAPHICAL SKETCH

Miguel Mascorro is a student committed in pursuing a career in mathematics education. He completed his master's degree in Applied Mathematics at the University of Texas Rio Grande Valley in December 2022 and graduated with a Bachelor of Science in Applied Mathematics at the same university in December 2019. You can contact him via email at mmascorro1998@hotmail.com.