2007

GMRES ON A TRIDIAGONAL TOEPLITZ LINEAR SYSTEM

Wei Zhang
University of Kentucky, wzhang@ms.uky.edu

ABSTRACT OF DISSERTATION

Wei Zhang

The Graduate School

University of Kentucky

2007

GMRES ON A TRIDIAGONAL TOEPLITZ LINEAR SYSTEM

ABSTRACT OF DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements of the degree of Doctor of Philosophy in the
College of Arts and Sciences at the University of Kentucky

By

Wei Zhang

Lexington, Kentucky

Director: Dr. Ren-Cang Li, Dr. Qiang Ye, Department of Mathematics

Lexington, Kentucky

2007

ABSTRACT OF DISSERTATION

GMRES ON A TRIDIAGONAL TOEPLITZ LINEAR SYSTEM

The Generalized Minimal Residual method (GMRES) is often used to solve a nonsymmetric linear system $Ax = b$. But its convergence analysis is a rather difficult task in general. A commonly used approach is to diagonalize $A = X\Lambda X^{-1}$ and then separate the study of GMRES convergence behavior into optimizing the condition number of X and a polynomial minimization problem over A's spectrum. This artificial separation could greatly overestimate GMRES residuals and likely yields error bounds that are too far from the actual ones. On the other hand, considering the effects of both A's spectrum and the conditioning of X at the same time poses a difficult challenge, perhaps impossible to deal with in general but only possible for certain particular linear systems. This thesis will do so for a (nonsymmetric) tridiagonal Toeplitz system. Sharp error bounds on and sometimes exact expressions for residuals are obtained. These expressions and/or bounds are in terms of the three parameters that define A and Chebyshev polynomials of the first kind or the second kind.

KEYWORDS: GMRES, rate of convergence, tridiagonal Toeplitz matrix, linear system, Chebyshev Polynomials

Wei Zhang

23 July 2007

GMRES ON A TRIDIAGONAL TOEPLITZ LINEAR SYSTEM

By

Wei Zhang

Dr. Ren-Cang Li, Dr. Qiang Ye
Director of Dissertation

Dr. Serge Ochanine
Director of Graduate Studies

23 July 2007

DISSERTATION

Wei Zhang

The Graduate School

University of Kentucky

2007

GMRES ON A TRIDIAGONAL TOEPLITZ LINEAR SYSTEM

DISSERTATION

A dissertation submitted in partial fulfillment of the
requirements of the degree of Doctor of Philosophy in the
College of Arts and Sciences at the University of Kentucky

By

Wei Zhang

Lexington, Kentucky

Director: Dr. Ren-Cang Li, Dr. Qiang Ye, Department of Mathematics

Lexington, Kentucky

2007

ACKNOWLEDGMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. I would like to thank Dr. Zhongwei Shen, Dr. David Johnson, Dr. Robert Molzon, Dr. Russell Brown, Dr. Peter Hislop and Dr. Serge Ochanine from Department of Mathematics of University of Kentucky, and Dr. Yuming Zhang from Department of Electrical Engineering of University of Kentucky, for their continuous support and encouragement since I joined the mathematics program.

I am deeply indebted to my advisor Dr. Ren-Cang Li and Dr. Qiang Ye from Department of Mathematics of University of Kentucky, who helped and encouraged me in all the time of research for and writing of this thesis and I would like to give my special thanks to my parents whose support enabled me to complete this work.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Iteration methods are often used to solve large sparse systems of linear equations. The Generalized Minimal Residual (GMRES) method [32, 14] is such an algorithm and is often used for solving a non-symmetric linear system

$$Ax = b, \tag{1.1.1}$$

where A is an $N \times N$ nonsingular matrix, and b is a vector with dimension N.

The basic idea is to seek approximate solutions, which minimize the residual norm, within the Krylov subspaces. Specifically, the kth approximation, x_k, is sought so that the kth residual, $r_k = b - Ax_k$, satisfies [32] (without loss of generality, we take initially $x_0 = 0$ and thus $r_0 = b$)

$$\|r_k\|_2 = \min_{y \in \mathcal{K}_k} \|b - Ay\|_2,$$

where the *kth Krylov subspace* $\mathcal{K}_k \equiv \mathcal{K}_k(A, b)$ of A on b is defined as

$$\mathcal{K}_k \equiv \mathcal{K}_k(A, b) \overset{\text{def}}{=} \mathsf{span}\{b, Ab, \ldots, A^{k-1}b\}, \tag{1.1.2}$$

and generic norm $\|\cdot\|_2$ is the usual ℓ_2 norm of a vector or the spectral norm of a matrix.

According to R. W. Freund, N. M. Nachtigal [12] and M. Embree [10], the residual norms by other algorithms, such as QMR [12, 13] and BiCGSSTAB [39, 20], are somehow related to the GMRES residual norm. Hence, understanding convergence for GMRES helps us to study convergence of other algorithms.

Roughly speaking, there are three kinds of convergence bounds for GMRES:

1. The bound based on eigenvalues with the eigenvector condition number [8, 32];
2. The bound based on the field of values [6, 7];
3. The bound based on pseudospectra [37, 36].

In this thesis we are most interested in the first kind. It starts by diagonalizing $A = X\Lambda X^{-1}$ and then separating the study of GMRES convergence behavior into optimizing the condition number of X and a polynomial minimization problem over A's spectrum. This artificial separation could greatly overestimate GMRES residuals and likely yields error bounds that are too far from the actual ones. On the other hand, considering the effects of both A's spectrum and the conditioning of X at the same time poses a difficult challenge, perhaps impossible to deal with in general but possible for certain particular linear systems.

1.2 Objective

This thesis is concerned with the convergence analysis of GMRES on a linear system $Ax = b$ whose coefficient matrix A is a (nonsymmetric) tridiagonal Toeplitz coefficient matrix

$$A = \begin{pmatrix} \lambda & \mu & & \\ \nu & \ddots & \ddots & \\ & \ddots & \ddots & \mu \\ & & \nu & \lambda \end{pmatrix},$$

where λ, μ, ν are assumed nonzero and possibly complex. Linear systems as such have been studied quite extensively in the past. For the nonsymmetric case, i.e., $\mu \neq \nu$ as we are interested in here, most up-to-date and detailed studies are due to Liesen and Strakoš [26] and Ernst [11].

Motivated to better understand the convergence behavior of GMRES on a convection-diffusion model problem [27], Liesen and Strakoš and Ernst established various bounds on residual ratios. Most results in Liesen and Strakoš [26] are of a qualitative nature, intended to explain GMRES convergence behaviors for such linear systems. In particular, Liesen and Strakoš showed that GMRES for tiny $|\mu|$ behaves much like GMRES after setting μ to 0. Instead of eigenvalue information, Ernst used the field of values to assess the convergence rate. Our object here is to analyze the kth residual for the GMRES on tridiagonal Toeplitz A directly and arrive at simple quantitative results.

The remainder of this thesis is organized as follows.

Chapter 2 reviews necessary material about GMRES method and Chebyshev polynomials. Section 2.1 reviews projection methods and some of their properties. Section 2.2 introduces Krylov subspaces. Based on projection methods and Krylov subspaces, Section 2.3 explains the Arnoldi process and GMRES algorithm. Section 2.4 introduces Chebyshev polynomials of the first kind and the second kind.

Chapter 3 covers the rate of convergence analysis of GMRES Method on a tridiagonal Toeplitz linear system by applying the Chebyshev polynomial of the first kind. Section 3.1 gives the calculation of the kth residual of GMRES method. Section 3.2 analyzes the norm of the residual based on rectangular Vandermonde matrices and the Chebyshev polynomial of the first kind. Section 3.3 follows the calculation of the residual in Section 3.2, and shows an estimation of the upper bound of the kth residual in a general case, which is given by Theorem 3.3.1. Section 3.3 finishes the proof of Theorem 3.3.1 by analyzing the decomposition of the residual. Some numerical examples are also given in Section 3.3. Section 3.4 lists the estimation result with some special right-hand sides: e_1, e_N, or $b_{(1)}e_1 + b_{(N)}e_N$. The estimation comes from Theorem 3.3.1 and gives a tight upper bound. Section 3.5 estimates what the worst convergence speed could be.

Chapter 4 applies Chebyshev Polynomials of the second kind to estimate the kth residual, r_k, of GMRES method on a tridiagonal Teoplitz linear system. Section 4.1 calculates the residual by applying rectangular Vandermonde matrices and the Chebyshev polynomial of the second kind. The computation is similar to those in Chapter 3, and similar bounds are obtained. Section 4.2 follows the computation in the previous section, and gives an estimation of residuals of GMRES with a general right-hand side. The residuals with special right-hand side: $b = e_1$ or $b = e_N$ are calculated exactly in Section 4.3. Some of the complicated computations, needed in Section 4.2, are presented in Section 4.4. Chapter 5 presents our concluding remarks.

1.3 Notation

Throughout this thesis, $\mathbb{K}^{n\times m}$ is the set of all $n \times m$ matrices with entries in $\mathbb{K}$, where $\mathbb{K}$ is $\mathbb{C}$ (the set of complex numbers) or $\mathbb{R}$ (the set of real numbers), $\mathbb{K}^n = \mathbb{K}^{n\times 1}$, and $\mathbb{K} = \mathbb{K}^1$. I_n (or simply I if its dimension is clear from the context) is the $n \times n$ identity matrix, and e_j is its jth column. The superscript "$\cdot^*$" takes conjugate transpose while "$\cdot^T$" takes transpose only. The smallest singular value of X is denoted by $\sigma_{\min}(X)$.

We shall also adopt MATLAB-like convention to access the entries of vectors and matrices. The set of integers from i to j inclusive is $i : j$. For vector u and matrix X, $u_{(j)}$ is u's jth entry, $X_{(i,j)}$ is X's (i,j)th entry, $\operatorname{diag}(u)$ is the diagonal matrix with $(\operatorname{diag}(u))_{(j,j)} = u_{(j)}$; X's submatrices $X_{(k:\ell,i:j)}$, $X_{(k:\ell,:)}$, and $X_{(:,i:j)}$ consists of intersections of row k to row ℓ and column i to column j, row k to row ℓ and all columns, and all rows and column i to column j, respectively. Finally $\|\cdot\|_p$ $(1 \leq p \leq \infty)$ is the ℓ_p norm of a vector or the ℓ_p operator norm of a matrix, defined as

$$\|u\|_p = \left(\sum_j |u_{(j)}|^p \right)^{1/p}, \quad \|X\|_p = \max_{\|u\|_p = 1} \|Xu\|_p.$$

$\lfloor \alpha \rfloor$ be the largest integer that is no bigger than α, and $\lceil \alpha \rceil$ the smallest integer that is no less than α.

Throughout this thesis, exact arithmetic is assumed, A is N-by-N, and k is GMRES iteration index. Since in exact arithmetic GMRES computes the exact solution in at most N steps, $r_N = 0$. For this reason, we restrict $k < N$ at all times. This restriction is needed to interpret our later results concerning (worst) asymptotic speed in terms of certain limits of $\|r_k\|^{1/k}$ as $k \to \infty$.

Chapter 2

Preliminary

In this chapter, we briefly present the projection method, the Krylov subspace method, the GMRES method, and Chebyshev polynomials. The projection methods and some of their properties are presented in Section 2.1. In Section 2.2, the Krylov subspace method is introduced as a special case of the Projection method. Section 2.3 introduces the GMRES method, including Arnoldi process and the general GMRES algorithm. Finally Section 2.4 presents Chebyshev polynomials of the first kind and the second kind.

2.1 Projection Method

2.1.1 Basic Idea

A projection method [21, 15] is to find an approximation to the solution of a linear system from a subspace, and is widely used for solving large linear systems.

Without loss of generality, we take initially $x_0 = 0$. A general projection method for solving the linear system

$$Ax = b$$

is a method which seeks an approximate solution x_m from a subspace $\mathcal{J}_m$ of dimension m by imposing the Petrov-Galekin condition:

$$b - Ax_m \perp \mathcal{L}_m,$$

i.e.,

$$\langle b - Ax_m, w \rangle = 0, \forall w \in \mathcal{L}_m, \tag{2.1.1}$$

where $\mathcal{L}_m$ is another subspace of dimension m, and $\langle \cdot, \cdot \rangle$ is the inner product.

The subspace $\mathcal{J}_m$ is the search subspace, which contains the approximate solution x_m. $\mathcal{L}_m$ is called the subspace of constraints, i.e. the constraints in the Petrov-Galekin condition. $\mathcal{J}_m$ and $\mathcal{L}_m$ can be the same subspace, which results in an orthogonal projection method. If they are different, it is called an oblique projection method.

Let $\mathcal{J}_m$ have a basis $\{v_1, v_2, \ldots, v_m\}$, and $V = [v_1, v_2, \ldots, v_m]$ be an $n \times m$ matrix. Let $\mathcal{L}_m$ have a basis $\{w_1, w_2, \ldots, w_m\}$, and $W = [w_1, w_2, \ldots, w_m]$ be an $n \times m$ matrix. Then let $x_m = Vy$, where y is a vector with dimension m. According to (2.1.1), we have

$$W^T AVy = W^T b. \tag{2.1.2}$$

If $W^T AV$ is nonsingular, then x_m can be written as

$$x_m = Vy = V(W^T AV)^{-1} W^T b. \tag{2.1.3}$$

The projection method can be presented by Algorithm 2.1.1 [29].

Algorithm 2.1.1 Projection Method

```
repeat
    Select a pair of subspace $\mathcal{J}_m$ and $\mathcal{L}_m$;
    Choose bases $V = [v_1, v_2, \ldots, v_m]$ and $W = [w_1, w_2, \ldots, w_m]$ for $\mathcal{J}_m$ and $\mathcal{L}_m$;
    $r := b - Ax$;
    $y := (W^T AV)^{-1} W^T r$;
    $x := x + Vy$;
until convergence
```

2.1.2 Properties

Obviously, Algorithm 2.1.1 works only if $W^T AV$ is nonsingular. A special case, where $W^T AV$ is nonsingular, is presented by [29, Proposition 5.1].

Proposition 2.1.1. *If A is nonsingular and $\mathcal{L}_m = A\mathcal{J}_m$, then the matrix W^TAV is nonsingular for any basis matrices V and W of $\mathcal{J}_m$ and $\mathcal{L}_m$, respectively.*

Proof: Since $\mathcal{L}_m = A\mathcal{J}_m$, then we have

$$W = AVG,$$

where G is a nonsingular $m \times m$ matrix. Hence

$$W^TAV = G^T(AV)^TAV.$$

Since A is a nonsingular $N \times N$ matrix, and V is a basis of $\mathcal{J}_m$ with dimension $N \times m$, then $N \times m$ matrix AV is of full column rank and $(AV)^TAV$ is nonsingular. ∎.

[29, Proposition 5.1] also presents another particular case as follows,

Proposition 2.1.2. *If A is positive definite and $\mathcal{L}_m = \mathcal{J}_m$, then the matrix W^TAV is nonsingular for any basis matrices V and W of $\mathcal{J}_m$ and $\mathcal{L}_m$, respectively.*

Proof: Since $\mathcal{L}_m = \mathcal{J}_m$, then

$$W = VG,$$

where G is a nonsingular $m \times m$ matrix. Hence

$$W^TAV = G^TV^TAV.$$

Since A is positive definite, V^TAV is also positive definite. Hence $W^TAV = G^TV^TAV$ is nonsingular. ∎.

According to the above propositions, $\mathcal{L}_m$ is chosen as $\mathcal{J}_m$ or $A\mathcal{J}_m$ very often. Let $\mathcal{L}_m = A\mathcal{J}_m$. The projection method minimizes 2-norm of the residual $r = b - Ax$ [29, Proposition 5.3].

Proposition 2.1.3. *Let A be an arbitrary square matrix, $\mathcal{L}_m = A\mathcal{J}_m$. Given an initial solution $x_0 = 0$, a vector $\tilde{x}$ is the result of an projection method onto $\mathcal{J}_m$, orthogonally to $\mathcal{L}_m$ if and only if it minimizes the 2-norm of the residual vector $b - A\tilde{x}$ over $x \in \mathcal{J}_m$, i.e.*

$$\|b - A\tilde{x}\|_2 = \min_{x \in \mathcal{K}_m} \|b - Ax\|_2.$$

Proof: If $\tilde{x}$ is the minimizer of $\|b - Ax\|_2$, it is necessary and sufficient that $b - A\tilde{x}$ is orthogonal to all vectors of the form $v = Ay$, where $y \in \mathcal{J}_m$. Since $\mathcal{L}_m = A\mathcal{J}_m$, $v = Ay \in \mathcal{L}_m$. Hence $\tilde{x}$ is the minimizer of $\|b - Ax\|_2$, if and only if,

$$\langle b - A\tilde{x}, v\rangle = 0, \forall v \in \mathcal{L}_m,$$

i.e. the Petrov-Galerkin condition is satisfied, and $\tilde{x}$ is the approximated solution. ∎

Proposition 2.1.3 is applied in GMRES method. For $\mathcal{L}_m = \mathcal{J}_m$, there is a similar property, which is used in this thesis and omitted here.

2.2 Krylov Subspace

A Krylov subspace method [29, 31] is a projection method for which the subspace $\mathcal{J}_m$ is the Krylov subspace

$$\mathcal{J}_m = \mathcal{K}_m(A, b) = \mathsf{span}\{b, Ab, A^2b, \ldots, A^{m-1}b\}.$$

In this thesis, $\mathcal{K}_m(A, b)$ will be denoted by $\mathcal{K}_m$.

The approximations obtained from a Krylov subspace method are of the form

$$A^{-1}b \approx x_m = q_{m-1}(A)b,$$

where q_{m-1} is a certain polynomial of degree $(m-1)$.

The dimension of the subspace increases by one at each step[1] of the iterative process. The subspace $\mathcal{K}_m$ is the subspace of all vectors in $\mathbb{R}^N$ which can be written as $x = p(A)b$, where p is a polynomial of degree not exceeding $m - 1$.

2.3 GMRES method

The Generalized Minimum Residual Method(GMRES) [32, 14] is a projection method based on taking $\mathcal{J}_m = \mathcal{K}_m$ and $\mathcal{L}_m = A\mathcal{K}_m$, in which $\mathcal{K}_m$ is the m-th Krylov subspace with $v_1 = r_0/\|r_0\|_2$. The GMRES method minimizes the residual norm over all vectors in the Krylov subspace $\mathcal{K}_m$.

2.3.1 Arnoldi Algorithm

Arnoldi's process [2, 30] is applied to build an orthogonal basis of the Krylov subspace $\mathcal{K}_m$. In exact arithmetic, the Arnoldi algorithm is described as Algorithm 2.3.1.

Algorithm 2.3.1 Arnoldi algorithm

```
Choose a vector $v_1$ of norm 1;
for $j$ = 1,2,...,$m$ do
    Compute $h_{ij} = \langle Av_j, v_i \rangle$ for $i$ = 1,2,...,$j$;
    Compute $w_j = Av_j - \sum_{i=1}^{j} h_{ij} v_i$;
    $h_{j+1,j} = \|w_j\|_2$;
    if $h_{j+1,j} = 0$ then stop;
    $v_{j+1} = w_j / h_{j+1,j}$;
end for
```

According to Algorithm 2.3.1, $\{v_1, v_2, ...v_m\}$ forms an orthogonal basis of the Krylov subspace $\mathcal{K}_m$. At each step, the algorithm multiplies the previous vector v_j by A and then orthonormalizes the resulting vector w_j against all previous v_i's by a standard Gram-Schmidt procedure. If w_j vanishes, then the process stops. If w_j does not vanish, then another vector v_{j+1} is obtained, and the algorithm produces a bigger Krylov subspace $\mathcal{K}_{j+1}$.

[1] If $A^m b \in \mathcal{K}_m$, then the dimension of $\mathcal{K}_{m+1}$ is equal to the dimension of $\mathcal{K}_m$.

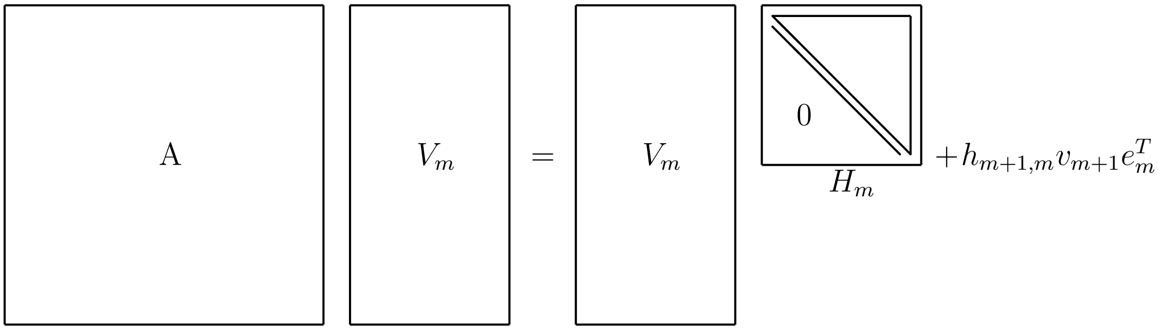

Figure 2.3.1: $AV_m = V_m H_m + h_{m+1,m}v_{m+1}e_m^T$.

According to Algorithm 2.3.1, the following property is obtained [29].

Proposition 2.3.1. *Let $V_m = [v_1, v_2, \ldots, v_m]$, where $\{v_1, \ldots, v_m\}$ are obtained from Algorithm 2.3.1, and $\bar{H}_m$ be the $(m+1) \times m$ Hessenberg matrix whose nonzero entries h_{ij} are defined by Algorithm 2.3.1, and H_m obtained from $\bar{H}_m$ by deleting the last row. Then*

$$\begin{aligned} AV_m &= V_{m+1}\bar{H}_m \qquad (2.3.4) \\ &= V_m H_m + h_{m+1,m}v_{m+1}e_m^T, \qquad (2.3.5) \end{aligned}$$

and hence

$$V_m^T A V_m = H_m. \qquad (2.3.6)$$

The proof just follows from Algorithm 2.3.1, and is omitted here. Obviously, the relation (2.3.4) and (2.3.5) are equivalent, the relation (2.3.6) is obtained by multiplying both sides of (2.3.5) by V_m^T.

The relation (2.3.5) is represented in Figure 2.3.1.

In practice, the Modified Gram-Schmidt [34] or the Householder Arnoldi [40] is used instead of the standard Gram-Schmidt algorithm. The Modified Gram-Schmidt results in Algorithm 2.3.2 [34].

Algorithm 2.3.2 Arnoldi-Modified Gram-Schmidt

```
Choose a vector $v_1$ of norm 1;
for $j = 1,2,\ldots,m$ do
  Compute $w_j = Av_j$;
  for $i = 1,2,\ldots,j$ do
    $h_{ij} = \langle w_j, v_i \rangle$;
    $w_j = w_j - h_{ij}v_i$;
  end for
  $h_{j+1,j} = \|w_j\|_2$. If $h_{j+1,j} = 0$ then stop;
  $v_{j+1} = w_j/h_{j+1,j}$;
end for
```

The Arnoldi-Modified Gram-Schmidt and Algorithm 2.3.1 are mathematically equivalent in exact arithmetic. If we consider the round-off in practice, Algorithm 2.3.2 is more reliable. If even Algorithm 2.3.2 is inadequate, then double orthogonalization or the Housholder Arnoldi can be used.

The Arnoldi-Modified Gram-schmidt algorithm is applied in the GMRES algorithm in the next subsection.

2.3.2 GMRES Algorithm

By using the Arnoldi-Modified Gram-schmidt algorithm [34] to build an orthogonal basis of the Krylov subspace $\mathcal{K}_m$, the GMRES method seek an approximation minimizing the residual norm over all vectors in the Krylov subspace $\mathcal{K}_m$.

Any vector x in the Krylov subspace $\mathcal{K}_m$ can be written as:

$$x = V_m y,$$

where y is an m-vector. Then

$$\|b - Ax\|_2 = \|b - AV_m y\|_2.$$

Furthermore, we have

$$\begin{aligned} b - Ax &= b - AV_m y \\ &= \beta v_1 - V_{m+1}\bar{H}_m y \\ &= V_{m+1}(\beta e_1 - \bar{H}_m y), \end{aligned}$$

where $\beta = \|b\|_2$. Since the column-vectors of V_{m+1} are othonormal, then

$$\|b - Ax\|_2 = \|\beta e_1 - \bar{H}_m y\|_2. \tag{2.3.7}$$

The GMRES approximation gives the unique vector in $\mathcal{K}_m$, which minimizes $\|b - Ax\|_2$. According to (2.3.7), this approximation can be obtained by seeking an optimal y_m that minimizes $\|\beta e_1 - \bar{H}_m y\|_2$, hence the approximation is $x_m = V_m y_m$. The y_m is easier to compute since it is the solution of an $(m+1) \times m$ least-squares problem where m is not big generally. The algorithm is described as Algorithm 2.3.3 [32].

Algorithm 2.3.3 GMRES

```
 1: Compute $r_0 = b, \beta = \|r_0\|_2$, and $v_1 = r_0/\beta$;
 2: Define the $(m+1) \times m$ matrix $\bar{H}_m = \{h_{ij}\}_{1 \le i \le m+1, 1 \le j \le m}$;
 3: for $j$ = 1,2,...,$m$ do
 4:   Compute $w_j = Av_j$;
 5:   for $i$ = 1,2,...,$j$ do
 6:     $h_{ij} = \langle w_j, v_i \rangle$;
 7:     $w_j = w_j - h_{ij} v_i$;
 8:   end for
 9:   $h_{j+1,j} = \|w_j\|_2$. If $h_{j+1,j} = 0$ set $m = j$ and go to 12;
10:   $v_{j+1} = w_j / h_{j+1,j}$;
11: end for
12: Compute $y_m$, the minimizer of $\|\beta e_1 - \bar{H}_m y\|_2$, and $x_m = V_m y_m$.
```

Algorithm 2.3.3 calculates the GMRES approximation as:

$$x_m = V_m y_m, \tag{2.3.8}$$

where

$$y_m = \text{argmin}_y \|\beta e_1 - \bar{H}_m y\|_2. \tag{2.3.9}$$

Similar to Algorithm 2.3.1, Algorithm 2.3.3 will stop if w_j vanishes, i.e., when $h_{j+1,j} = 0$ at some step j. If the algorithm stops in this way, that means the residual vector is zero, hence the approximation of GMRES will be the exact solution. The following proposition is Proposition 2 in [32].

Proposition 2.3.2. *The solution x_j produced by GMRES at step j is exact if and only if the following four equivalent conditions hold:*

1. *The algorithm 2.3.3 breaks down at step j.*

2. $v_{j+1} = 0$.

3. $h_{j+1,j} = 0$.

4. *The degree of the minimal polynomial of A on the initial residual vector r_0 is equal to j, where the minimal polynomial of A on r_0 is the monic polynomial of least degree such that $p(A)r_0 = 0$.*

The breakdown results in the exact solution. Because the degree of the minimal polynomial of v_1 can not exceed N for an N-dimensional problem, GMRES terminates in at most N steps [32].

We also can illustrate the above property by Figure 2.3.2, which is obtained by letting $h_{j+1,j} = 0$ in Figure 2.3.1.

2.3.3 Practical Implementation of GMRES

In order to solve the least-squares problem $\min\|\beta e_1 - \bar{H}_m y\|_2$ in the last step of Algorithm 2.3.3, the Hessenberg matrix is transformed into upper triangular form by using

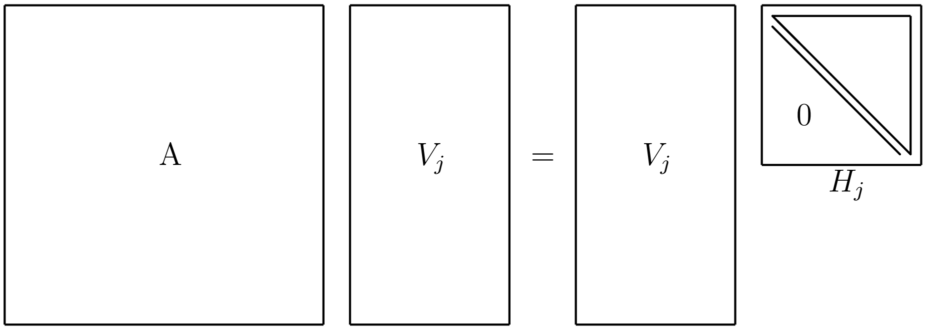

Figure 2.3.2: $AV_j = V_j H_j + 0$.

Givens rotations [16, 32]. Define the rotation matrices as

$$Q_i = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 1 & & & & & \\ & & & c_i & s_i & & & \\ & & & -s_i & c_i & & & \\ & & & & & 1 & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{pmatrix}, \tag{2.3.10}$$

where $c_i^2 + s_i^2 = 1$. Given $\bar{H}_m$ as an $(m+1) \times m$ matrix, Q_i's are $(m+1) \times (m+1)$ matrices for $1 \leq i \leq m$.

The idea can be best explained for $m = 4$. We have

$$\bar{H}_4 = \begin{pmatrix} h_{11} & h_{12} & h_{13} & h_{14} \\ h_{21} & h_{22} & h_{23} & h_{24} \\ & h_{32} & h_{33} & h_{34} \\ & & h_{43} & h_{44} \\ & & & h_{54} \end{pmatrix}, \quad \bar{\gamma} = \begin{pmatrix} \beta \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{2.3.11}$$

Multiply $\bar{H}_4$ and $\bar{\gamma}$ by

$$Q_1 = \begin{pmatrix} c_1 & s_1 & & & \\ -s_1 & c_1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix},$$

where

$$c_1 = \frac{h_{11}}{\sqrt{h_{11}^2 + h_{21}^2}},$$
$$s_1 = \frac{h_{21}}{\sqrt{h_{11}^2 + h_{21}^2}}.$$

Then we have

$$\bar{H}_4^{(1)} = \begin{pmatrix} h_{11}^{(1)} & h_{12}^{(1)} & h_{13}^{(1)} & h_{14}^{(1)} \\ & h_{22}^{(1)} & h_{23}^{(1)} & h_{24}^{(1)} \\ & h_{32} & h_{33} & h_{34} \\ & & h_{43} & h_{44} \\ & & & h_{54} \end{pmatrix}, \quad \bar{\gamma}^{(1)} = \begin{pmatrix} c_1\beta \\ -s_1\beta \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{2.3.12}$$

Now h_{21} has been eliminated, and we can multiply Q_2 to eliminate h_{32}, and continue the elimination process until the upper triangular $H_4^{(4)}$ is obtained,

$$\bar{R}_4 = \bar{H}_4^{(4)} = \begin{pmatrix} h_{11}^{(4)} & h_{12}^{(4)} & h_{13}^{(4)} & h_{14}^{(4)} \\ & h_{22}^{(4)} & h_{23}^{(4)} & h_{24}^{(4)} \\ & & h_{33}^{(4)} & h_{34}^{(4)} \\ & & & h_{44}^{(4)} \\ & & & 0 \end{pmatrix}, \quad \bar{\gamma}^{(4)} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \gamma_5 \end{pmatrix} = \begin{pmatrix} c_1\beta \\ -c_2 s_1\beta \\ c_3 s_2 s_1\beta \\ -c_4 s_3 s_2 s_1\beta \\ s_4 s_3 s_2 s_1\beta \end{pmatrix}, \tag{2.3.13}$$

where c_i and s_i are defined as

$$c_i = \frac{h_{ii}^{(i-1)}}{\sqrt{(h_{ii}^{(i-1)})^2 + h_{i+1,i}^2}}, \tag{2.3.14}$$

$$s_i = \frac{h_{i+1,i}}{\sqrt{(h_{ii}^{(i-1)})^2 + h_{i+1,i}^2}}. \tag{2.3.15}$$

Let $\bar{Q} = Q_4 Q_3 Q_2 Q_1$, then

$$\bar{R}_4 = \bar{Q}\bar{H}_4, \tag{2.3.16}$$

$$\bar{\gamma}^{(4)} = \bar{Q}\bar{\gamma} = \bar{Q}\beta e_1, \tag{2.3.17}$$

and

$$\begin{aligned} \bar{\gamma}^{(4)} - \bar{R}_4 y &= \bar{Q}\beta e_1 - \bar{Q}\bar{H}_4 y & (2.3.18) \\ &= \bar{Q}(\beta e_1 - \bar{H}_4 y). & (2.3.19) \end{aligned}$$

Since $\bar{Q}$ is unitary,

$$\min \|\bar{\gamma}^{(4)} - \bar{R}_4 y\|_2 = \min \|\beta e_1 - \bar{H}_4 y\|_2. \tag{2.3.20}$$

The solution to $\min \|\bar{\gamma}^{(4)} - \bar{R}_4 y\|_2$ can be calculated by solving the upper triangular system,

$$\begin{pmatrix} h_{11}^{(4)} & h_{12}^{(4)} & h_{13}^{(4)} & h_{14}^{(4)} \\ & h_{22}^{(4)} & h_{23}^{(4)} & h_{24}^{(4)} \\ & & h_{33}^{(4)} & h_{34}^{(4)} \\ & & & h_{44}^{(4)} \end{pmatrix} y = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix}. \tag{2.3.21}$$

Moreover, $\min \|\bar{\gamma}^{(4)} - \bar{R}_4 y\|_2 = |\gamma_5|$.

2.4 Chebyshev Polynomials

Chebyshev polynomials [28, 1], named after Pafnuty Chebyshev, are a sequence of orthogonal polynomials which are related to de Moivre's formula and which are easily defined recursively. In this thesis, we calculate Chebyshev polynomials of the first kind which are denoted by $T_n(x)$ and Chebyshev polynomials of the second kind which are denoted by $U_n(x)$.

2.4.1 Chebyshev Polynomials of the First Kind

The Chebyshev polynomials of the first kind [28] are a set of orthogonal polynomials defined as the solutions to the Chebyshev differential equation. They are used as an approximation to a least squares fit. They are also intimately connected with trigonometric multiple-angle formulas. They are normalized such that $T_n(1) = 1$.

The Chebyshev polynomials of the first kind can be obtained from the generating function

$$g(t, x) = \frac{1 - xt}{1 - 2xt + t^2} = \sum_{n=0}^{\infty} T_n(x) t^n,$$

for $|x| < 1$ and $|t| < 1$.

The Chebyshev polynomials of the first kind are:

$$\begin{aligned}
T_0(x) &= 1, \\
T_1(x) &= x, \\
T_2(x) &= 2x^2 - 1, \\
\dots, & \\
T_{n+2}(x) &= 2T_{n+1}x - T_n(x).
\end{aligned}$$

The Chebyshev polynomials of the first kind can also be defined through the identity

$$T_n(\cos\theta) = \cos(n\theta).$$

Let $x = \cos(\theta)$. Then

$$T_n(x) = \cos(n\theta) = \cos(n \arccos(x)).$$

The sequence of Chebyshev polynomials $T_n(x)$ composes a sequence of orthogonal polynomial. Two different polynomials $T_n(x)$, $T_m(x)$ in the sequence are orthogonal to each other in the following sense

$$\int_{-1}^{1} T_n(x)\, T_m(x) \frac{dx}{\sqrt{1-x^2}} = \begin{cases} 0, & \text{if} \quad m \neq n, \\ \pi, & \text{if} \quad m = n = 0, \\ \pi/2, & \text{if} \quad m = n \neq 0. \end{cases} \tag{2.4.22}$$

2.4.2 Chebyshev Polynomials of the Second Kind

The Chebyshev polynomials of the second kind [28] are denoted $U_n(x)$ and are defined by a different generating function

$$g(t,x) = \frac{1}{1 - 2xt + t^2} = \sum_{n=0}^{\infty} U_n(x) t^n,$$

for $|x| < 1$ and $|t| < 1$.

The Chebyshev polynomials of the second kind are:

$$\begin{aligned}
U_0(x) &= 1, \\
U_1(x) &= 2x, \\
U_2(x) &= 4x^2 - 1, \\
\ldots, & \\
U_{n+2}(x) &= 2U_{n+1}x - U_n(x).
\end{aligned}$$

Letting $x = \cos(\theta)$ allows the Chebyshev polynomials of the second kind to be written as

$$U_n(x) = \frac{\sin[(n+1)\theta)]}{\sin(\theta)}.$$

The sequence of $U_n(x)$ also composes a orthogonal polynomial sequence. Two different polynomials $U_n(x)$, $U_m(x)$ in the sequence are orthogonal to each other in the following sense

$$\int_{-1}^{1} U_n(x)\, U_m(x) \sqrt{1-x^2} dx = \left\{ \begin{array}{rll} 0, & \text{if} & m \neq n, \\ \pi/2, & \text{if} & m = n. \end{array} \right. \tag{2.4.23}$$

Chapter 3

Convergence Analysis Using Chebyshev Polynomials of the First Kind

In this and next chapter, we will present our main result of the convergence analysis of GMRES, i.e. estimation of the residuals of a tridiagonal Toeplitz matrix. Section 3.1 introduces how the residuals are calculated. Section 3.2 applies Chebyshev polynomial to calculate the residuals. In Section 3.3, the upper bounds of estimated residuals in general case are given, and numerical examples are provided to show how sharp our error bounds are. More error bounds of special cases are obtained in Section 3.4 and 3.5. In this chapter, a general case means a linear system with a general right hand side; special cases are ones with special right hand sides, such as $b = e_1$, $b = e_N$.

3.1 Residual Formulation for a Diagonalizable Linear System

The result in this section applies to any $Ax = b$ with diagonalizable A. Suppose

$$A = X\Lambda X^{-1}, \tag{3.1.1}$$

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_N), \tag{3.1.2}$$

and X is an $N \times N$ nonsingular matrix.

Recall we assumed, without loss of generality, the initial approximation $x_0 = 0$ and thus the initial residual $r_0 = b - Ax_0 = b$. The kth GMRES residual is

$$\begin{aligned} \|r_k\|_2 &= \min_{p_k(0)=1} \|p_k(A)b\|_2 \\ &= \min_{p_k(0)=1} \|Xp_k(\Lambda)X^{-1}b\|_2 \\ &= \min_{u_{(1)}=1} \|YV_{k+1,N}^T u\|_2, \end{aligned} \tag{3.1.3}$$

where

$$Y = X\,\mathrm{diag}(X^{-1}b), \tag{3.1.4}$$

$V_{k+1,N}$ is the $(k+1) \times N$ rectangular Vandermonde matrix

$$V_{k+1,N} \overset{\text{def}}{=} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_N \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^k & \lambda_2^k & \cdots & \lambda_N^k \end{pmatrix}, \tag{3.1.5}$$

having nodes $\{\lambda_j\}_{j=1}^N$[41, Lemma 2.1] and the coefficients of $p_k(A)$ forms a vector u with $u_{(1)} = 1$.

3.2 Residual Reformulation Using Chebyshev Polynomials of the First Kind

An $N \times N$ tridiagonal Toeplitz A takes this form

$$A = \begin{pmatrix} \lambda & \mu & & \\ \nu & \ddots & \ddots & \\ & \ddots & \ddots & \mu \\ & & \nu & \lambda \end{pmatrix} \in \mathbb{C}^{N\times N}. \tag{3.2.1}$$

Given parameters ν, λ, and μ are nonzero, A is diagonalizable. In fact [33, pp.113-115] (see also [26]),

$$A = X\Lambda X^{-1}, \tag{3.2.2}$$

where

$$
\begin{aligned}
\Lambda &= \operatorname{diag}(\lambda_1, \ldots, \lambda_N), && (3.2.3)\\
\lambda_j &= \lambda - 2\sqrt{\mu\nu}\, t_j, && (3.2.4)\\
t_j &= \cos\theta_j, \quad \theta_j = \frac{j\pi}{N+1}, && (3.2.5)\\
X &= \Omega Z, && (3.2.6)\\
\Omega &= \operatorname{diag}(\xi^0, \xi^{-1}, \ldots, \xi^{-N+1}), \quad \xi = -\frac{\sqrt{\mu\nu}}{\nu}, && (3.2.7)\\
Z_{(:,j)} &= \sqrt{\frac{2}{N+1}}\,(\sin j\theta_1, \ldots, \sin j\theta_N)^T. && (3.2.8)
\end{aligned}
$$

It can be verified that $Z^T Z = I_N$. So A is normal if $|\xi| = 1$, i.e., $|\mu| = |\nu| > 0$. By (3.2.4), we have

$$\lambda_j = \omega(t_j - \tau), \quad 1 \le j \le N, \tag{3.2.9}$$

where

$$\omega = -2\sqrt{\mu\nu}, \tag{3.2.10}$$

$$\tau = \frac{\lambda}{2\sqrt{\mu\nu}}. \tag{3.2.11}$$

Any branch of $\sqrt{\mu\nu}$, once picked and fixed, is a valid choice in this thesis.

According to (3.1.3), we have

$$
\begin{aligned}
\|r_k\|_2 &= \min_{u_{(1)}=1} \|Y\, V_{k+1,N}^T u\|_2\\
&\le \|Y\|_2 \min_{u_{(1)}=1} \|V_{k+1,N}^T u\|_2\\
&\le \|X\|_2 \max_i |(X^{-1}b)_{(i)}| \min_{u_{(1)}=1} \|V_{k+1,N}^T u\|_2\\
&\le \|X\|_2 \|X^{-1}b\|_2 \min_{u_{(1)}=1} \|V_{k+1,N}^T u\|_2\\
&\le \|X\|_2 \|X^{-1}\|_2 \|b\|_2 \min_{u_{(1)}=1} \|V_{k+1,N}^T u\|_2\\
&\le \kappa(X)\|b\|_2 \min_{p_k(0)=1} \max_i |p_k(\lambda_i)|, && (3.2.12)
\end{aligned}
$$

where p_k is a polynomial of degree no higher than k. Thus, together with $r_0 = b$, they imply

$$\|r_k\|_2/\|r_0\|_2 \le \kappa(X) \min_{p_k(0)=1} \max_i |p_k(\lambda_i)|. \tag{3.2.13}$$

Inequality (3.2.13) is often the starting point in existing quantitative analysis on GMRES convergence [18, Page 54], as it seems that there is no easy way to do otherwise. It simplifies the analysis by separating the study of GMRES convergence behavior into optimizing the condition number of X and a polynomial minimization problem over A's spectrum, but it could potentially overestimate GMRES residuals. This is partly because, as observed by Liesen and Strakoš [26], possible cancelations of huge components in X and/or X^{-1} were artificially ignored for the sake of the convergence analysis. However, for tridiagonal Toeplitz matrix A the rich structure (the Chebyshev polynomial) allows us to estimate differently, namely starting with (3.1.3) directly.

We define the mth *Translated Chebyshev Polynomial of the first kind* in z of degree m as

$$\begin{aligned} \mathcal{T}_m(z;\omega,\tau) &\overset{\text{def}}{=} \mathcal{T}_m(z/\omega+\tau) && (3.2.14)\\ &= a_{mm}z^m + a_{m-1\,m}z^{m-1} + \cdots + a_{1m}z + a_{0m}, && (3.2.15)\end{aligned}$$

where the coefficients $a_{jm} \equiv a_{jm}(\omega,\tau)$ are functions of ω and τ, and forms an upper triangular R_{m+1} in terms of ω and τ as

$$R_{m+1} \equiv R_{m+1}(\omega,\tau) \overset{\text{def}}{=} \begin{pmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0\,m} \\ & a_{11} & a_{12} & \cdots & a_{1\,m} \\ & & a_{22} & \cdots & a_{2\,m} \\ & & & \ddots & \vdots \\ & & & & a_{m\,m} \end{pmatrix}, \tag{3.2.16}$$

i.e., the $(m+1)$th column consists of the coefficients of $T_m(\lambda;\omega,\tau)$. Set

$$\boldsymbol{T}_N \overset{\text{def}}{=} \begin{pmatrix} T_0(t_1) & T_0(t_2) & \cdots & T_0(t_N) \\ T_1(t_1) & T_1(t_2) & \cdots & T_1(t_N) \\ \vdots & \vdots & & \vdots \\ T_{N-1}(t_1) & T_{N-1}(t_2) & \cdots & T_{N-1}(t_N) \end{pmatrix}, \tag{3.2.17}$$

and $V_N = V_{N,N}$ for short. Then

$$V_N^T R_N = \boldsymbol{T}_N^T. \tag{3.2.18}$$

Equation (3.2.18) yields $V_N^T = \boldsymbol{T}_N^T R_N^{-1}$. Extracting the first $k+1$ columns from both sides of $V_N^T = \boldsymbol{T}_N^T R_N^{-1}$ yields

$$V_{k+1,N}^T = (\boldsymbol{T}_N^T)_{(:,1:k+1)}\, R_{k+1}^{-1}. \tag{3.2.19}$$

Now notice $Y = X\ \mathrm{diag}(X^{-1}b)$ and $X = \Omega Z$ with Z in (3.2.6) being real and orthogonal to get

$$\begin{aligned} Y\, V_{k+1,N}^T &= \Omega Z\ \mathrm{diag}(Z^T\Omega^{-1}b)\, (\boldsymbol{T}_N^T)_{(:,1:k+1)} R_{k+1}^{-1} \\ &= \Omega M_{(:,1:k+1)} R_{k+1}^{-1} && (3.2.20) \\ &= \Omega M_{(:,1:k+1)} \Omega_{k+1}^{-1}\, \Omega_{k+1} R_{k+1}^{-1}. && (3.2.21) \end{aligned}$$

where $\Omega_{k+1} = \Omega_{(1:k+1,1:k+1)}$, the $(k+1)$th leading submatrix of Ω,

$$M = Z\ \mathrm{diag}(Z^T\Omega^{-1}b)\, \boldsymbol{T}_N^T. \tag{3.2.22}$$

Now we can estimate GMRES residual

$$\|r_k\|_2 = \min_{u_{(1)}=1} \|Y\, V_{k+1,N}^T u\|_2 = \min_{u_{(1)}=1} \|\Omega M_{(:,1:k+1)} \Omega_{k+1}^{-1}\, \Omega_{k+1} R_{k+1}^{-1} u\|_2. \tag{3.2.23}$$

Then

$$\sigma_{\min}(\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}) \le \frac{\|r_k\|_2}{\min_{u_{(1)}=1} \|\Omega_{k+1} R_{k+1}^{-1} u\|_2} \le \|\Omega M_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2. \tag{3.2.24}$$

Hence, the upper bound and lower bound of the residual $\|r_k\|_2$ could be estimated.

3.3 Estimation of Residual in General Case

Set

$$\zeta = \min\left\{|\xi|, \frac{1}{|\xi|}\right\}, \tag{3.3.1}$$

$$\rho = \max\left\{\left|\tau + \sqrt{\tau^2-1}\right|, \left|\tau - \sqrt{\tau^2-1}\right|\right\}. \tag{3.3.2}$$

Note $\rho \geq 1$ always because $(\tau + \sqrt{\tau^2-1})(\tau - \sqrt{\tau^2-1}) = 1$. In particular if $\lambda \in \mathbb{R}$, $\mu < 0$ and $\nu > 0$, then $\rho = |\tau| + \sqrt{|\tau|^2+1}$.

Recall Chebyshev polynomials of the first kind

$$\begin{aligned} T_m(t) &= \cos(m \arccos t), && \text{for } |t| \leq 1, && (3.3.3)\\ &= \frac{1}{2}\left(t+\sqrt{t^2-1}\right)^m + \frac{1}{2}\left(t-\sqrt{t^2-1}\right)^m, && \text{for } |t| \geq 1. && (3.3.4) \end{aligned}$$

Define

$$\Phi_{k+1}^{(+)}(\tau,\xi) \overset{\text{def}}{=} \sum_{j=0}^{k}{}' \, |\xi|^{2j} \, |T_j(\tau)|^2, \tag{3.3.5}$$

$$\Phi_{k+1}^{(-)}(\tau,\xi) \overset{\text{def}}{=} \sum_{j=0}^{k}{}' \, |\xi|^{-2j} \, |T_j(\tau)|^2, \tag{3.3.6}$$

$$\Phi_{k+1}(\tau,\xi) \overset{\text{def}}{=} \sum_{j=0}^{k}{}' \, \zeta^{2j} \, |T_j(\tau)|^2 \equiv \min\left\{\Phi_{k+1}^{(+)}(\tau,\xi), \Phi_{k+1}^{(-)}(\tau,\xi)\right\}, \tag{3.3.7}$$

where $\sum_j'$ means the first term is halved. Obviously, if $|\xi| \leq 1$, then

$$\Phi_{k+1}(\tau,\xi) = \Phi_{k+1}^{(+)}(\tau,\xi);$$

otherwise,

$$\Phi_{k+1}(\tau,\xi) = \Phi_{k+1}^{(-)}(\tau,\xi).$$

3.3.1 Residual with General Right-hand Sides

Given a tridiagonal Toeplitz with a general right hand side, we have

Theorem 3.3.1. *For $Ax = b$, where A is tridiagonal Toeplitz as in (3.2.1) with nonzero (real or complex) parameters ν, λ, and μ. Then the kth GMRES residual r_k satisfies for $1 \leq k < N$*

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \sqrt{k+1}\left[\frac{1}{2} + \Phi_{k+1}(\tau, \xi)\right]^{-1/2}. \tag{3.3.8}$$

The proof of Theorem 3.3.1 involves a complicated computation. We will prove the theorem for the case $|\xi| \leq 1$ first.

Recall the computation in the previous section, the proof follows the inequality (3.2.24)

$$\sigma_{\min}(\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}) \leq \frac{\|r_k\|_2}{\min_{u_{(1)}=1} \|\Omega_{k+1} R_{k+1}^{-1} u\|_2} \leq \|\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}\|_2.$$

Rewrite the second inequality in (3.2.24):

$$\|r_k\|_2 \leq \min_{u_{(1)}=1} \|\Omega_{k+1} R_{k+1}^{-1} u\|_2 \|\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}\|_2. \tag{3.3.9}$$

This is our foundation to prove Theorem 3.3.1. There are two quantities to deal with

$$\min_{u_{(1)}=1} \|\Omega_{k+1} R_{k+1}^{-1} u\|_2 \quad \text{and} \quad \|\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}\|_2. \tag{3.3.10}$$

For the first part, the following lemma was proven in [23, 24], and also implied by the proof of [25, Theorem 2.1]. See also [22].

Lemma 3.3.1. *If W has full column rank, then*

$$\min_{u_{(1)}=1} \|Wu\|_2 = \left[e_1^T (W^* W)^{-1} e_1\right]^{-1/2}. \tag{3.3.11}$$

In particular if W is nonsingular, $\min_{u_{(1)}=1} \|Wu\|_2 = \|W^{-*}e_1\|_2^{-1}$.

Proof: Set $v = Wu$. Since W has full column rank N, its Moore-Penrose pseudo-inverse is $W^\dagger = (W^*W)^{-1}W^*$[35], and thus $u = W^\dagger v$. This gives a one-one and onto mapping between $u \in \mathbb{C}^m$ and the column space $v \in \mathsf{span}(W)$, where $\mathsf{span}(W) = \text{span}\{W(:$

$, 1), W(:, 2), \ldots, W(:, N)\}$. Then

$$\min_{|u_{(1)}|=1} \|Wu\|_2 = \min_u \frac{\|Wu\|_2}{|u_{(1)}|} = \min_{v \in \mathsf{span}(W)} \frac{\|v\|_2}{|e_1^T W^\dagger v|} \geq \min_v \frac{\|v\|_2}{|e_1^T W^\dagger v|} = \|e_1^T W^\dagger\|_2^{-1}, \tag{3.3.12}$$

where the last min is achieved at

$$v_{\text{opt}} = (e_1^T W^\dagger)^* = W(W^*W)^{-1}e_1 \in \text{span}(W),$$

which implies the "$\geq$" in (3.3.12) is an equality, and

$$u_{\text{opt}} = \frac{W^\dagger v_{\text{opt}}}{e_1^T W^\dagger v_{\text{opt}}}.$$

Finally

$$\|e_1^T W^\dagger\|_2 = \sqrt{e_1^T W^\dagger (W^\dagger)^* e_1} = \sqrt{e_1^T (W^*W)^{-1} e_1}.$$

∎

The above proof is taken from [23].

By this lemma, we have (note $a_{00} = 1$)

$$\min_{u_{(1)}=1} \|\Omega_{k+1} R_{k+1}^{-1} u\|_2 = \|\Omega_{k+1}^{-*} R_{k+1}^* e_1\|_2^{-1} = \left[\frac{1}{2} + \Phi_{k+1}^{(+)}(\tau, \xi)\right]^{-1/2}. \tag{3.3.13}$$

This gives the first quantity in (3.3.10).

Rewrite Theorem3.3.1 as

$$\begin{aligned} \|r_k\|_2 &\leq \|r_0\|_2 \sqrt{k+1} \left[\frac{1}{2} + \Phi_{k+1}(\tau, \xi)\right]^{-1/2} \\ &\leq \|b\|_2 \sqrt{k+1} \left[\frac{1}{2} + \Phi_{k+1}(\tau, \xi)\right]^{-1/2} \\ &\leq \left[\frac{1}{2} + \Phi_{k+1}(\tau, \xi)\right]^{-1/2} \|b\|_2 \sqrt{k+1}. \end{aligned}$$

If we can show

$$\|\Omega M_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2 \leq \|b\|_2 \sqrt{k+1}, \tag{3.3.14}$$

then according to inequality (3.3.9) and (3.3.13), Theorem 3.3.1 is proved. We prove the inequality (3.3.14) in the next subsection and finish the proof of Theorem 3.3.1.

3.3.2 The Second Part of the Proof

Now let's investigate $\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}$ and prove the inequality (3.3.14).

It can be seen that $\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1} = (\Omega M \Omega^{-1})_{(:,1:k+1)}$ since Ω is diagonal. To compute $\Omega M \Omega^{-1}$, we shall investigate M in (3.2.22) first.

$$\begin{aligned} M &= Z\ \mathrm{diag}(Z^T\Omega^{-1}b)\,\boldsymbol{T}_N^T \\ &= \sum_{\ell=1}^{N} Z\ \mathrm{diag}(Z\Omega^{-1}b_{(\ell)}e_\ell)\,\boldsymbol{T}_N^T \\ &= \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1} Z\ \mathrm{diag}(Ze_\ell)\,\boldsymbol{T}_N^T \\ &= \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1} Z\ \mathrm{diag}(Z_{(:,\ell)})\,\boldsymbol{T}_N^T \\ &= \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1} M_\ell, \end{aligned} \tag{3.3.15}$$

where $M_\ell = Z\ \mathrm{diag}(Z_{(:,\ell)})\,\boldsymbol{T}_N^T$ and $b_{(\ell)}$ is the ℓth element of the right hand side b.

It is not easy to obtain M directly, but M_ℓ can be calculated in Lemma 3.3.3. In Lemma 3.3.2 and in the proof of Lemma 3.3.3 below, without causing notational conflict, we will temporarily use k as a running index, as opposed to the rest of the paper where k is reserved for GMRES step index.

Lemma 3.3.2. *For $\theta_j = \frac{j}{N+1}\pi$ and integer ℓ,*

$$\sum_{k=1}^{N} \cos \ell\theta_k = \begin{cases} N, & \text{if } \ell = 2m(N+1) \text{ for some integer } m, \\ -1, & \text{if } \ell \text{ is even, but } \ell \neq 2m(N+1) \text{ for any integer } m, \\ 0, & \text{if } \ell \text{ is odd.} \end{cases} \tag{3.3.16}$$

Proof: If $\ell = 2m(N+1)$ for some integer m, then $\ell\theta_k = 2mk\pi$ and thus $\cos \ell\theta_k = 1$. Assume that $\ell \neq 2m(N+1)$ for any integer m. Set $\Omega = \ell\pi/(N+1)$. We have [17, p.30]

$$\sum_{k=1}^{N} \cos \ell\theta_k = \sum_{k=1}^{N} \cos k\Omega = \cos \frac{N+1}{2}\Omega \times \frac{\sin \frac{N\Omega}{2}}{\sin \frac{\Omega}{2}}.$$

Now notice $\cos\frac{N+1}{2}\Omega = \cos\frac{\ell}{2}\pi = 0$ for odd ℓ and $(-1)^{\ell/2}$ for even ℓ, and $\sin\frac{N\Omega}{2} = \sin(\frac{\ell}{2}\pi - \Omega) = -(-1)^{\ell/2}\sin\Omega$ for even ℓ to conclude the proof. ∎

Lemma 3.3.2 is Lemma 3.1 in [23].

Lemma 3.3.3. *Let $M_\ell \overset{\text{def}}{=} Z\,\text{diag}(Z_{(:,\ell)})\boldsymbol{T}_N^T$ for $1 \le \ell \le N$. Then the entries of M_ℓ are zeros, except at those positions (i,j), graphically forming four straight lines:*

$$\begin{array}{ll} \text{(a)} & i+j=\ell+1, \\ \text{(b)} & i-j=\ell-1, \\ \text{(c)} & j-i=\ell+1, \\ \text{(d)} & i+j=2(N+1)-\ell+1. \end{array} \tag{3.3.17}$$

$(M_\ell)_{(i,j)} = 1/2$ for (a) and (b), except at their intersection $(\ell,1)$ for which $(M_\ell)_{(\ell,1)} = 1$. $(M_\ell)_{(i,j)} = -1/2$ for (c) and (d). Notice no valid entries for (c) if $\ell \ge N-1$ and no valid entries for (d) if $\ell \le 2$.

Proof: For $1 \le i,\, j \le N$,

$$\begin{aligned} 2(N+1)\cdot(M_\ell)_{(i,j)} &= 4\sum_{k=1}^{N}\sin k\theta_i\,\sin\ell\theta_k\,\cos(j-1)\theta_k \\ &= 4\sum_{k=1}^{N}\sin i\theta_k\,\sin\ell\theta_k\,\cos(j-1)\theta_k \\ &= 2\sum_{k=1}^{N}[\cos(i-\ell)\theta_k - \cos(i+\ell)\theta_k]\cos(j-1)\theta_k \\ &= \sum_{k=1}^{N}[\cos(i+j-\ell-1)\theta_k + \cos(i-j-\ell+1)\theta_k \\ &\qquad - \cos(i+j+\ell-1)\theta_k - \cos(i-j+\ell+1)\theta_k]\,. \end{aligned}$$

Since all

$$\begin{aligned} i_1 &= i+j-\ell-1, \\ i_2 &= i-j-\ell+1, \\ i_3 &= i+j+\ell-1, \\ i_4 &= i-j+\ell+1, \end{aligned}$$

are either even or odd at the same time, Lemma 3.3.2 implies $(M_\ell)_{(i,j)} = 0$ unless one of them takes the form $2m(N+1)$ for some integer m. We now investigate all possible situations as such, keeping in mind that $1 \le i,\ j,\ \ell \le N$.

1. $i_1 = i + j - \ell - 1 = 2m(N+1)$. This happens if and only if $m = 0$, and thus $i + j = \ell + 1$. Then

   $$i_2 = -2j + 2,\ i_3 = 2\ell,\ i_4 = -2j + 2\ell + 2.$$

   They are all even. i_3 and i_4 do not take the form $2m(N+1)$ for some integers m. This is obvious for i_3, while $i_4 = 2m(N+1)$ implies $m = 0$ and $j = \ell + 1$, and thus $i = 0$ which cannot happen. However if $i_2 = 2m(N+1)$, then $m = 0$ and $j = 1$, and thus $i = \ell$.

   So Lemma 3.3.2 implies $(M_\ell)_{(i,j)} = 1/2$ for $i+j = \ell+1$ and $i \neq \ell$, while $(M_\ell)_{(\ell,1)} = 1$.

2. $i_2 = i - j - \ell + 1 = 2m(N+1)$. This happens if and only if $m = 0$, and thus $i - j = \ell - 1$. Then

   $$i_1 = 2j - 2,\ i_3 = 2j + 2\ell - 2,\ i_4 = 2\ell.$$

   They are all even. i_3 and i_4 do not take the form $2m(N+1)$ for some integers m. This is obvious for i_4, while $i_3 = 2m(N+1)$ implies $m = 1$ and $j = N + 2 - \ell$, and thus $i = N + 1$ which cannot happen. However if $i_1 = 2m(N+1)$, then $m = 0$ and thus $j = 1$ and $i = \ell$ which has already been considered in Item 1.

   So Lemma 3.3.2 implies $(M_\ell)_{(i,j)} = 1/2$ for $i-j = \ell-1$ and $i \neq \ell$, while $(M_\ell)_{(\ell,1)} = 1$.

3. $i_3 = i + j + \ell - 1 = 2m(N+1)$. This happens if and only if $m = 1$, and thus $i + j = 2(N+1) - \ell + 1$. Then

   $$i_1 = 2(N+1) - 2\ell,\ i_2 = 2(N+1) - 2j - 2\ell + 2,\ i_4 = 2(N+1) - 2j + 2.$$

They are all even. i_1 and i_2 do not take the form $2m(N+1)$ for some integers m. This is obvious for i_1, while $i_2 = 2m(N+1)$ implies $m = 0$ and $j = N+2-\ell$, and thus $i = N+1$ which cannot happen. However if $i_4 = 2m(N+1)$, then $m = 1$ and thus $j = 1$ and $i = 2(N+1)-\ell$ which is bigger than $N+2$ and not possible.

So Lemma 3.3.2 implies $(M_\ell)_{(i,j)} = -1/2$ for $i+j = 2(N+1)-\ell+1$.

4. $i_4 = i - j + \ell + 1 = 2m(N+1)$. This happens if and only if $m = 0$, and thus $j - i = \ell + 1$. Then

$$i_1 = 2j - 2\ell - 2,\ i_2 = -2\ell,\ i_3 = 2j - 2.$$

They are all even, and do not take the form $2m(N+1)$ for some integers m. This is obvious for i_2. $i_1 = 2m(N+1)$ implies $m = 0$ and $j = \ell+1$, and thus $i = 0$ which cannot happen. $i_3 = 2m(N+1)$ implies $m = 0$ and thus $j = 1$ and $i = -\ell$ which cannot happen either.

So Lemma 3.3.2 implies $(M_\ell)_{(i,j)} = -1/2$ for $j - i = \ell + 1$.

This completes the proof. ∎

Figure 3.3.1 gives the example for M_ℓ with $N = 8$ for $L = 1, 2, 3, 4$. The structure of M_ℓ, i.e. the nonzero entries, is showed in each graph. The entries on the red lines have the value $\frac{1}{2}$ except on the first column with $M_\ell(\ell, 1) = 1$; the entries on the black lines have the value $-\frac{1}{2}$. The lines are labeled with a, b, c, d according to Lemma 3.3.3.

Figure 3.3.2 gives the example for M_ℓ with $N = 8$ for $L = 5, 6, 7, 8$ similarly.

Now we know M_ℓ. We still need to find out $\Omega M \Omega^{-1}$. Let us examine it for $N = 5$ in

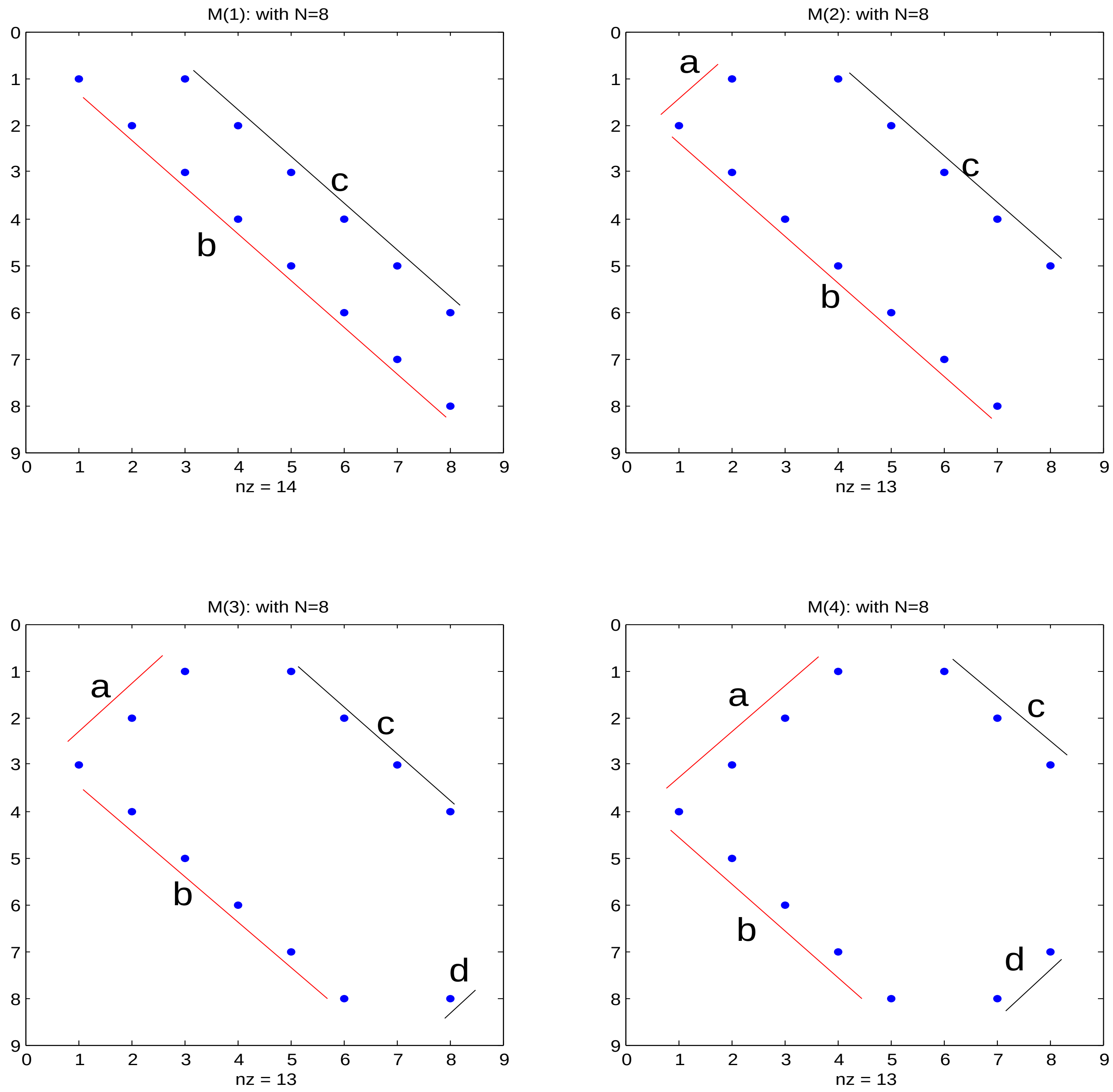

Figure 3.3.1: The structure of M_ℓ with $N = 8$ for $L = 1, 2, 3, 4$.

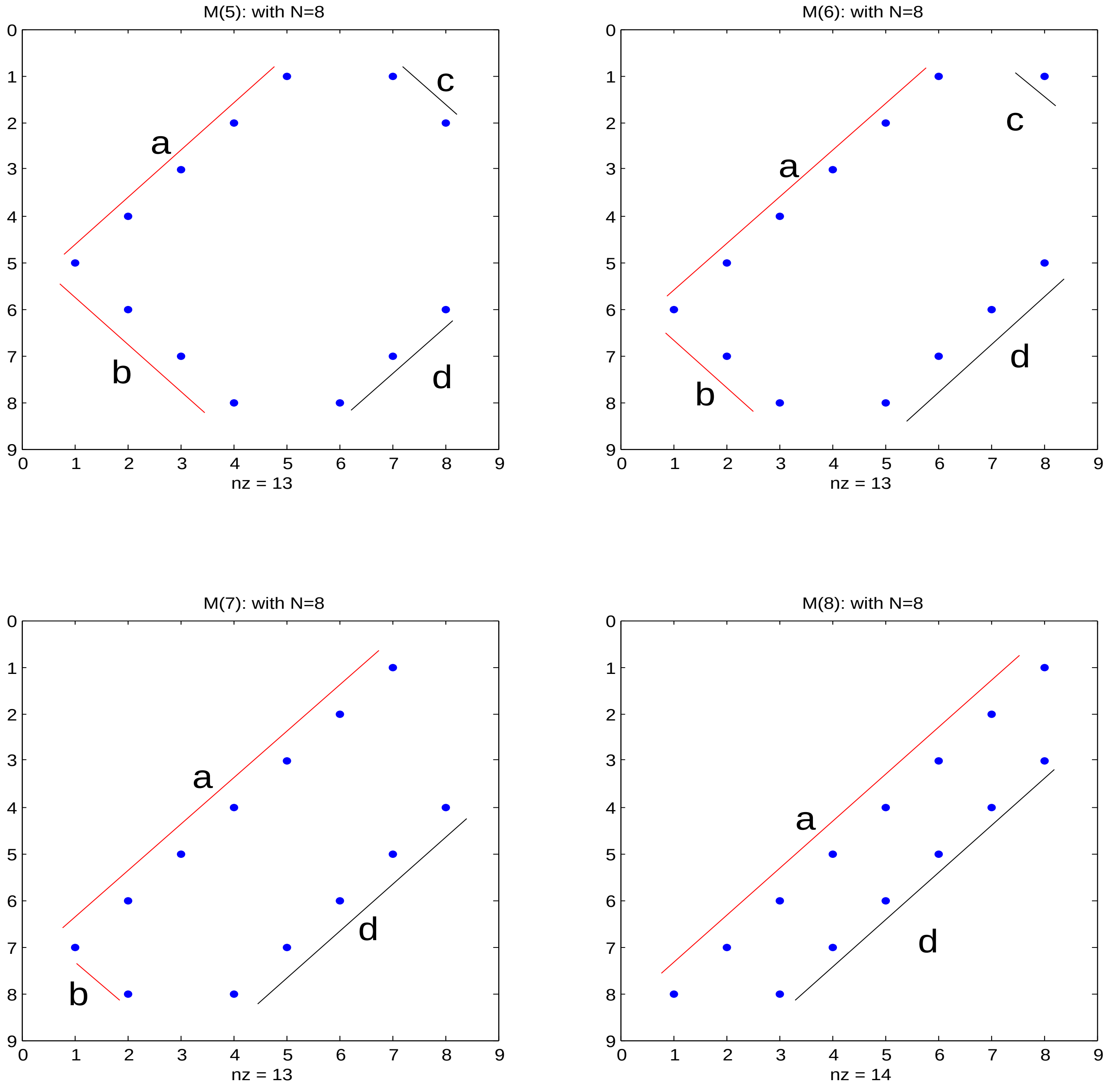

Figure 3.3.2: The structure of M_ℓ with $N = 8$ for $L = 5, 6, 7, 8$.

order to get some idea about what it may look like. $\Omega M \Omega^{-1}$ for $N = 5$ is

$$\begin{pmatrix} b_{(1)} & \frac{1}{2}\,\xi^2\,b_{(2)} & -\frac{1}{2}\,\xi^2\,b_{(1)} + \frac{1}{2}\,\xi^4\,b_{(3)} & -\frac{1}{2}\,\xi^4\,b_{(2)} + \frac{1}{2}\,\xi^6\,b_{(4)} & -\frac{1}{2}\,\xi^6\,b_{(3)} + \frac{1}{2}\,\xi^8\,b_{(5)} \\ b_{(2)} & \frac{1}{2}\,b_{(1)} + \frac{1}{2}\,b_{(3)}\,\xi^2 & \frac{1}{2}\,b_{(4)}\,\xi^4 & -\frac{1}{2}\,\xi^2\,b_{(1)} + \frac{1}{2}\,\xi^6\,b_{(5)} & -\frac{1}{2}\,\xi^4\,b_{(2)} \\ b_{(3)} & \frac{1}{2}\,b_{(2)} + \frac{1}{2}\,\xi^2\,b_{(4)} & \frac{1}{2}\,b_{(1)} + \frac{1}{2}\,b_{(5)}\,\xi^4 & 0 & -\frac{1}{2}\,\xi^2\,b_{(1)} - \frac{1}{2}\,\xi^6\,b_{(5)} \\ b_{(4)} & \frac{1}{2}\,b_{(3)} + \frac{1}{2}\,b_{(5)}\,\xi^2 & \frac{1}{2}\,b_{(2)} & \frac{1}{2}\,b_{(1)} - \frac{1}{2}\,b_{(5)}\,\xi^4 & -\frac{1}{2}\,b_{(4)}\,\xi^4 \\ b_{(5)} & \frac{1}{2}\,b_{(4)} & \frac{1}{2}\,b_{(3)} - \frac{1}{2}\,b_{(5)}\,\xi^2 & \frac{1}{2}\,b_{(2)} - \frac{1}{2}\,\xi^2\,b_{(4)} & \frac{1}{2}\,b_{(1)} - \frac{1}{2}\,b_{(3)}\,\xi^2 \end{pmatrix}.$$

We observe that for $N = 5$, the entries of $\Omega M \Omega^{-1}$ are polynomials in ξ with at most two terms. This turns out to be true for all N.

Lemma 3.3.4. *The following statements hold.*

1. *The first column of $\Omega M \Omega^{-1}$ is b. Entries in every other columns taking one of the three forms: $(b_{(n_1)}\xi^{m_1} + b_{(n_2)}\xi^{m_2})/2$ with $n_1 \neq n_2$, $b_{(n_1)}\xi^{m_1}/2$, and 0, where $1 \leq n_1,\, n_2 \leq N$ and $m_i \geq 0$ are nonnegative integer.*

2. *In each given column of $\Omega M \Omega^{-1}$, any particular entry of b appears at most twice.*

As the consequence, we have $\|\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}\|_2 \leq \sqrt{k+1}\;\|b\|_2$, if $|\xi| \leq 1$.

Proof: Notice $M = \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1}M_\ell$ and consider M's (i,j)th entry which comes from the contributions from all M_ℓ. But not all of M_ℓ contribute as most of them are zero at the position. Precisely, with the help of Lemma 3.3.3, those M_ℓ that contribute nontrivially to the (i,j)th position are the following ones subject to satisfying the given inequalities.

(a) if $1 \leq i + j - 1 \leq N$ or equivalently $i + j \leq N + 1$, M_{i+j-1} gives a $1/2$.

(b) if $1 \leq i - j + 1 \leq N$ or equivalently $i \geq j$, M_{i-j+1} gives a $1/2$.

(c) if $1 \leq j - i - 1 \leq N$ or equivalently $j \geq i + 2$, M_{j-i-1} gives a $-1/2$.

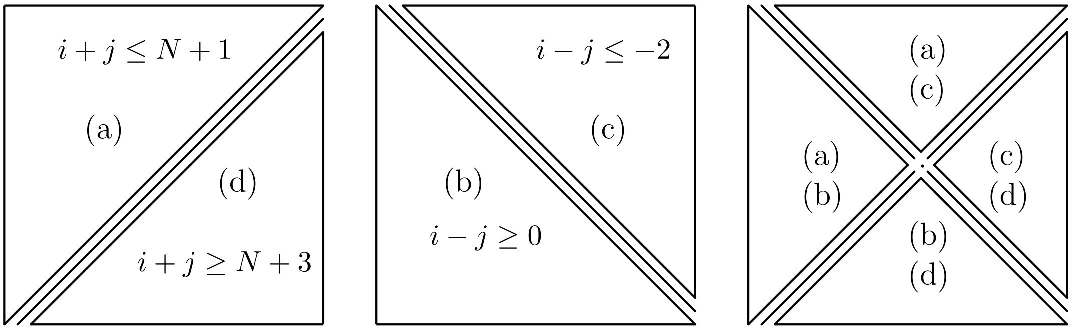

Figure 3.3.3: Computation of $M_{(i,j)}$.

(d) if $1\leq 2(N+1)-(i+j)+1\leq N$ or equivalently $i+j\geq N+3$, $M_{2(N+1)-(i+j)+1}$ gives a $-1/2$.

These inequalities, effectively 4 of them, are described in Figure 3.3.3. The left graph shows the regions of entries as divided by inequalities in (a) and (d); the middle shows Regions of entries as divided by inequalities in (b) and (c). These four regions divided entries of M into nine possible regions showed as the right graph of Figure 3.3.3. We shall examine each region one by one. Recall

$$(\Omega M\Omega^{-1})_{(i,j)}=\xi^{-i+1}M_{(i,j)}\xi^{j-1}=\xi^{j-i}M_{(i,j)},$$

and let

$$\begin{aligned}\gamma_a &= \frac{1}{2}b_{(i+j-1)}\xi^{2j-2},\\ \gamma_b &= \frac{1}{2}b_{(i-j+1)},\\ \gamma_c &= -\frac{1}{2}b_{(j-i-1)}\xi^{2(j-i-1)},\\ \gamma_d &= -\frac{1}{2}b_{(2(N+1)-(i+j)+1)}\xi^{2(N+1)-(i+j)}.\end{aligned}$$

Each entry in the 9 possible regions in the rightmost plot of Figure 3.3.3 is as follows.

1. (a) and (b): $(\Omega M\Omega^{-1})_{(i,j)}=\gamma_a+\gamma_b$.

2. (a) and (c): $(\Omega M\Omega^{-1})_{(i,j)}=\gamma_a+\gamma_c$.

3. (b) and (d): $(\Omega M \Omega^{-1})_{(i,j)} = \gamma_b + \gamma_d$.

4. (c) and (d): $(\Omega M \Omega^{-1})_{(i,j)} = \gamma_c + \gamma_d$.

5. (a) and $i - j = -1$: $(\Omega M \Omega^{-1})_{(i,j)} = \gamma_a$.

6. (b) and $i + j = N + 2$: $(\Omega M \Omega^{-1})_{(i,j)} = \gamma_b$.

7. (c) and $i + j = N + 2$: $(\Omega M \Omega^{-1})_{(i,j)} = \gamma_c$.

8. (d) and $i - j = -1$: $(\Omega M \Omega^{-1})_{(i,j)} = \gamma_d$.

9. $i - j = -1$ and $i + j = N + 2$: $(\Omega M \Omega^{-1})_{(i,j)} = 0$. In this case, $i = (N + 1)/2$ and $j = (N + 3)/2$. So there is only one such entry when N is odd, and none when N is even.

With this profile on the entries of $\Omega M \Omega^{-1}$, we have Item 1 of the lemma immediately. Item 2 is the consequence of $M = \sum_{\ell=1}^{N} b_{(\ell)} \xi^{\ell-1} M_\ell$ and Lemma 3.3.3 which implies that there are at most two nonzero entries in each column of M_ℓ.

As the consequence of Item 1 and Item 2, each column of $\Omega M \Omega^{-1}$ can be expressed as the sum of two vectors w and v such that $\|w\|_2$, $\|v\|_2 \le \|b\|_2/2$ when $|\xi| \le 1$, and thus $\|(\Omega M \Omega^{-1})_{(:,j)}\|_2 \le \|b\|_2$ for all $1 \le j \le N$. Therefore

$$\|\Omega M_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2 \le \sqrt{\sum_{j=1}^{k+1} \|(\Omega M \Omega^{-1})_{(:,j)}\|_2^2} \le \sqrt{k+1}\; \|b\|_2,$$

as expected. ∎

We can prove Theorem 3.3.1 now.

Proof: First we can prove

$$\|r_k\|_2 \le \|b\|_2 \sqrt{k+1} \left[\frac{1}{2} + \Phi_{k+1}^{(+)}(\tau, \xi) \right]^{-1/2}, \quad \text{for } |\xi| \le 1. \tag{3.3.18}$$

Assume $|\xi| \leq 1$. Inequality (3.3.18) is the consequence of estimation of residuals (3.2.24)

$$\sigma_{\min}(\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}) \leq \frac{\|r_k\|_2}{\min_{u_{(1)}=1} \|\Omega_{k+1} R_{k+1}^{-1} u\|_2} \leq \|\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}\|_2,$$

the equation(3.3.13)

$$\min_{u_{(1)}=1} \|\Omega_{k+1} R_{k+1}^{-1} u\|_2 = \|\Omega_{k+1}^{-*} R_{k+1}^{*} e_1\|_2^{-1} = \left[\frac{1}{2} + \Phi_{k+1}^{(+)}(\tau, \xi)\right]^{-1/2},$$

and Lemma 3.3.4

$$\|\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1}\|_2 \leq \sqrt{k+1}\ \|b\|_2, \quad \text{if}|\xi| \leq 1.$$

For the other case, $|\xi| > 1$, the proof can be turned into this case as follows. Let $\Pi = (e_N, \ldots, e_2, e_1) \in \mathbb{R}^{N\times N}$ be the permutation matrix. Notice $\Pi^T A \Pi = A^T$ and thus $Ax = b$ is equivalent to

$$A^T \Pi^T x = (\Pi^T A \Pi)(\Pi^T x) = \Pi^T b. \tag{3.3.19}$$

Note $\mathcal{K}_k(A^T, \Pi^T b) = \mathcal{K}_k(\Pi^T A \Pi, \Pi^T b) = \Pi^T\, \mathcal{K}_k(A, b)$, and

$$\begin{aligned} \|r_k\|_2 = \min_{y\in\mathcal{K}_k(A,b)} \|b - Ay\|_2 &= \min_{\Pi^T y \in \Pi^T \mathcal{K}_k(A,b)} \|\Pi^T(b - A\Pi\,\Pi^T y)\|_2 \\ &= \min_{w\in\mathcal{K}_k(A^T,\Pi^T b)} \|\Pi^T b - A^T w\|_2. \end{aligned}$$

Since (3.3.18) is proven true, then for $|\xi| > 1$ we have

$$\|r_k\|_2 \leq \|\Pi^T b\|_2 \sqrt{k+1} \left[\frac{1}{2} + \Phi_{k+1}^{(-)}(\tau, \xi)\right]^{-1/2},$$

because the ξ for A^T is the reciprocal of the one for A. ∎

REMARK **3.3.1.** The leftmost inequality in (3.2.24) gives a lower bound on $\|r_k\|_2$ in terms of $\sigma_{\min}(\Omega M_{(:,1:k+1)}\Omega_{k+1}^{-1})$ which, however, is hard to bound from below because it can be as small as zero, unless we know more about b such as $b = e_1$ or e_N as in Theorems 3.4.1 and 3.4.2.

The upper bounds can also be calculated according to the inequality (3.2.13).

Theorem 3.3.2. *For $Ax = b$, where A is tridiagonal Toeplitz as in (3.2.1) with nonzero (real or complex) parameters ν, λ, and μ. Then the kth GMRES residual r_k satisfies for $1 \le k < N$*

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \kappa(X) \left[\sum_{j=0}^{k} |T_j(\tau)|^2 \right]^{-1/2}. \tag{3.3.20}$$

Proof: According to (3.2.13),

$$\|r_k\|_2 / \|r_0\|_2 \le \kappa(X) \min_{p_k(0)=1} \max_i |p_k(\lambda_i)|. \tag{3.3.21}$$

Now we calculate $\min_{p_k(0)=1} \max_i |p_k(\lambda_i)|$ according to (3.2.18)

$$\begin{aligned} \min_{p_k(0)=1} \max_i |p_k(\lambda_i)| &= \min_{u_{(1)}=1} \|V_{k+1,N}^T u\|_2 \\ &= \min_{u_{(1)}=1} \|(\boldsymbol{T}_N^T)_{(:,1:k+1)} R_{k+1}^{-1} u\|_2. \end{aligned} \tag{3.3.22}$$

Apply Lemma 3.3.1, we have

$$\begin{aligned} \min_{p_k(0)=1} \max_i |p_k(\lambda_i)| &= \min_{u_{(1)}=1} \|(\boldsymbol{T}_N^T)_{(:,1:k+1)} R_{k+1}^{-1} u\|_2 \\ &= \left[e_1^T (((\boldsymbol{T}_N^T)_{(:,1:k+1)} R_{k+1}^{-1})^* (\boldsymbol{T}_N^T)_{(:,1:k+1)} R_{k+1}^{-1})^{-1} e_1 \right]^{-1/2} \\ &= \left[e_1^T (R_{k+1}^{-*} R_{k+1}^{-1})^{-1} e_1 \right]^{-1/2} \\ &= \left[e_1^T (R_{k+1} R_{k+1}^{*}) e_1 \right]^{-1/2} \\ &= \left[\sum_{j=0}^{k} |T_j(\tau)|^2 \right]^{-1/2}. \end{aligned} \tag{3.3.23}$$

Now (3.2.13) can be written as (3.3.20). ∎

3.3.3 Numerical Examples

In all numerical examples in this paper, we always take $N = 50$ and $\lambda = 1$, and choose different $|\tau|$ and $|\xi|$, then calculate μ and ν as following:

$$|\mu| = \frac{|\xi|}{2|\tau|}, \quad \mu = \pm|\mu|, \quad \text{and} \quad \nu = |\nu| = \frac{1}{2|\tau\xi|}.$$

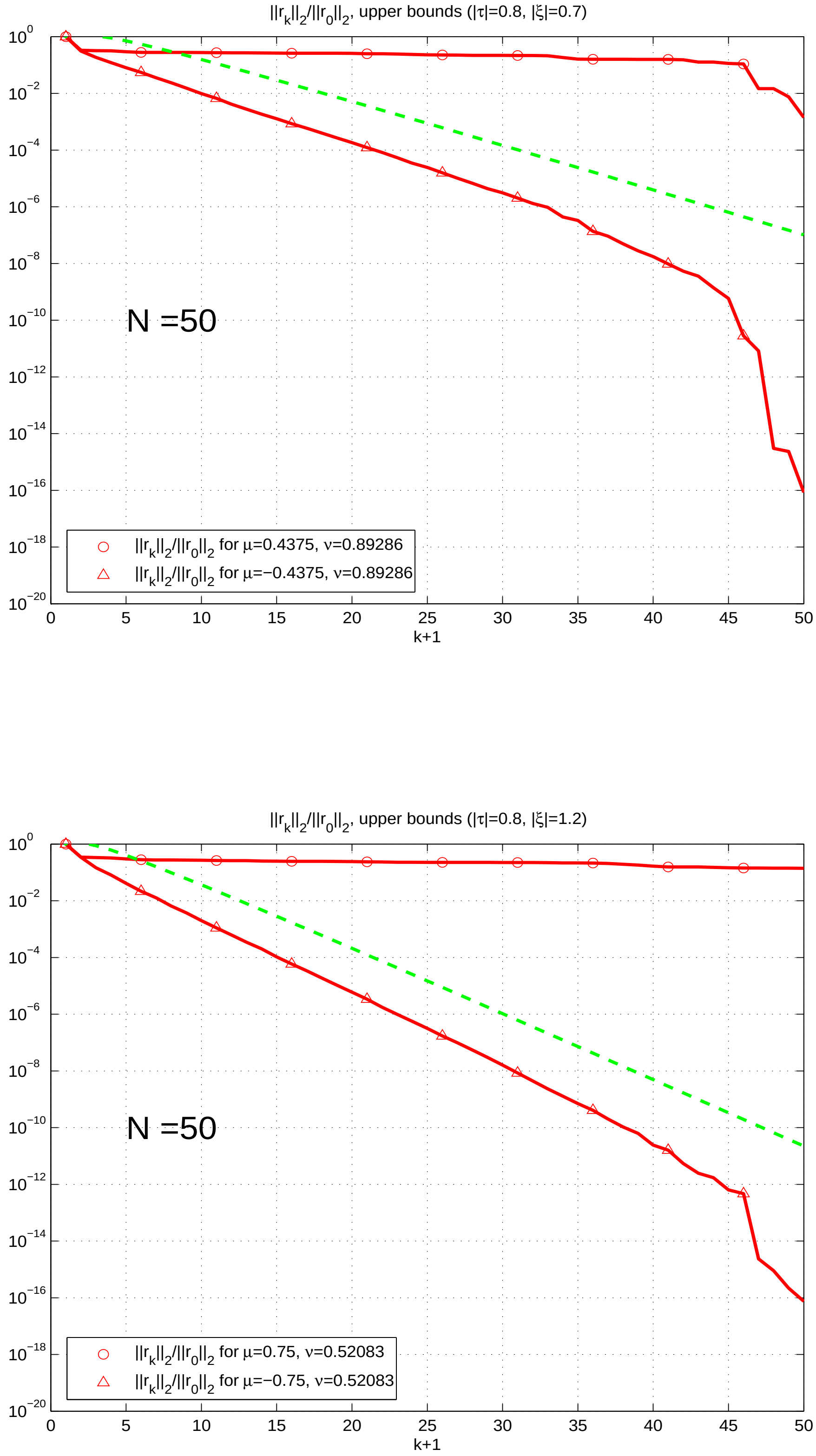

Figure 3.3.4: GMRES residuals for random b with $|\tau| = 0.8$, and the upper bounds by Theorem 3.3.1.

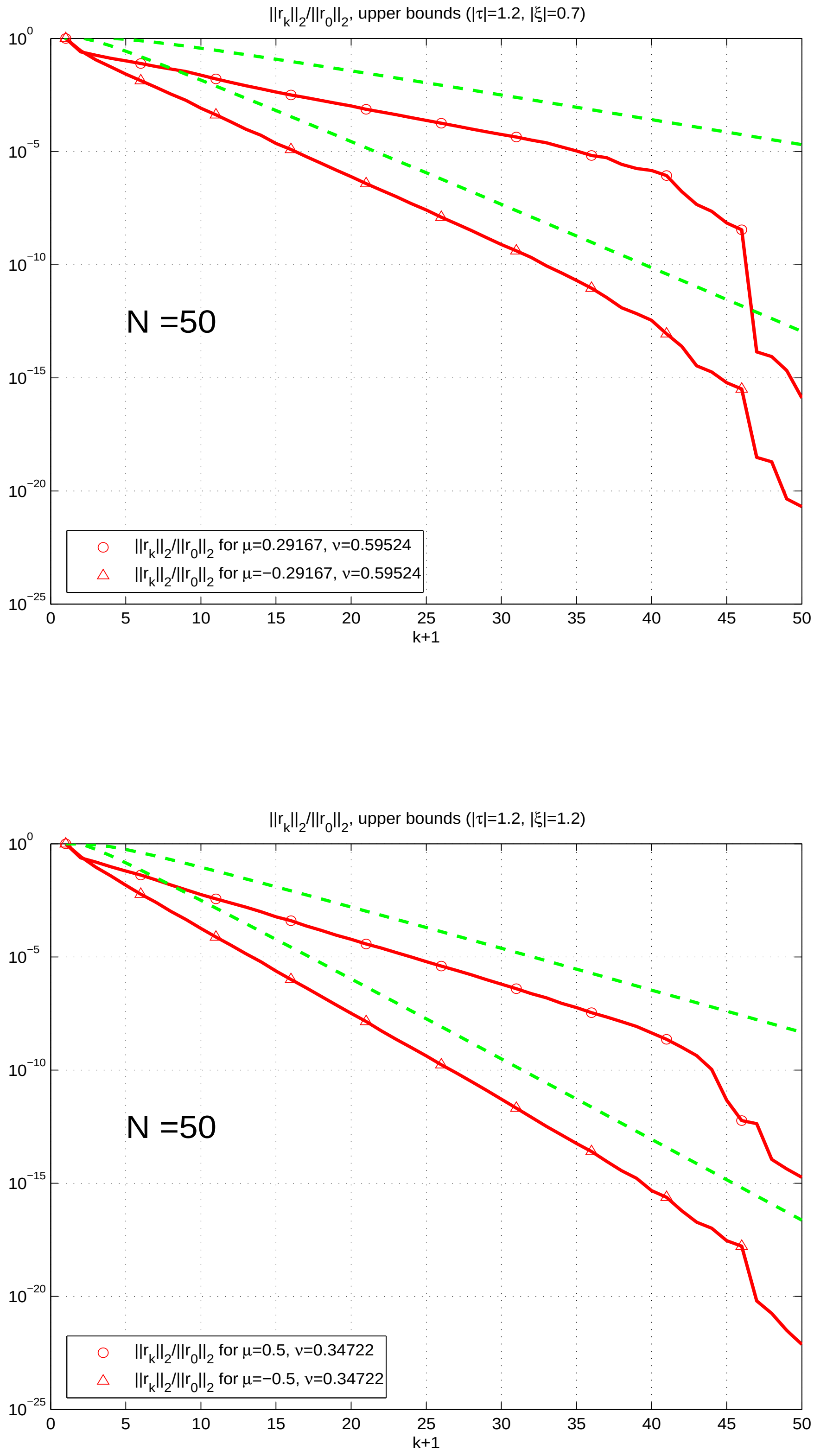

Figure 3.3.5: GMRES residuals for random b with $|\tau| = 1.2$, and the upper bounds by Theorem 3.3.1.

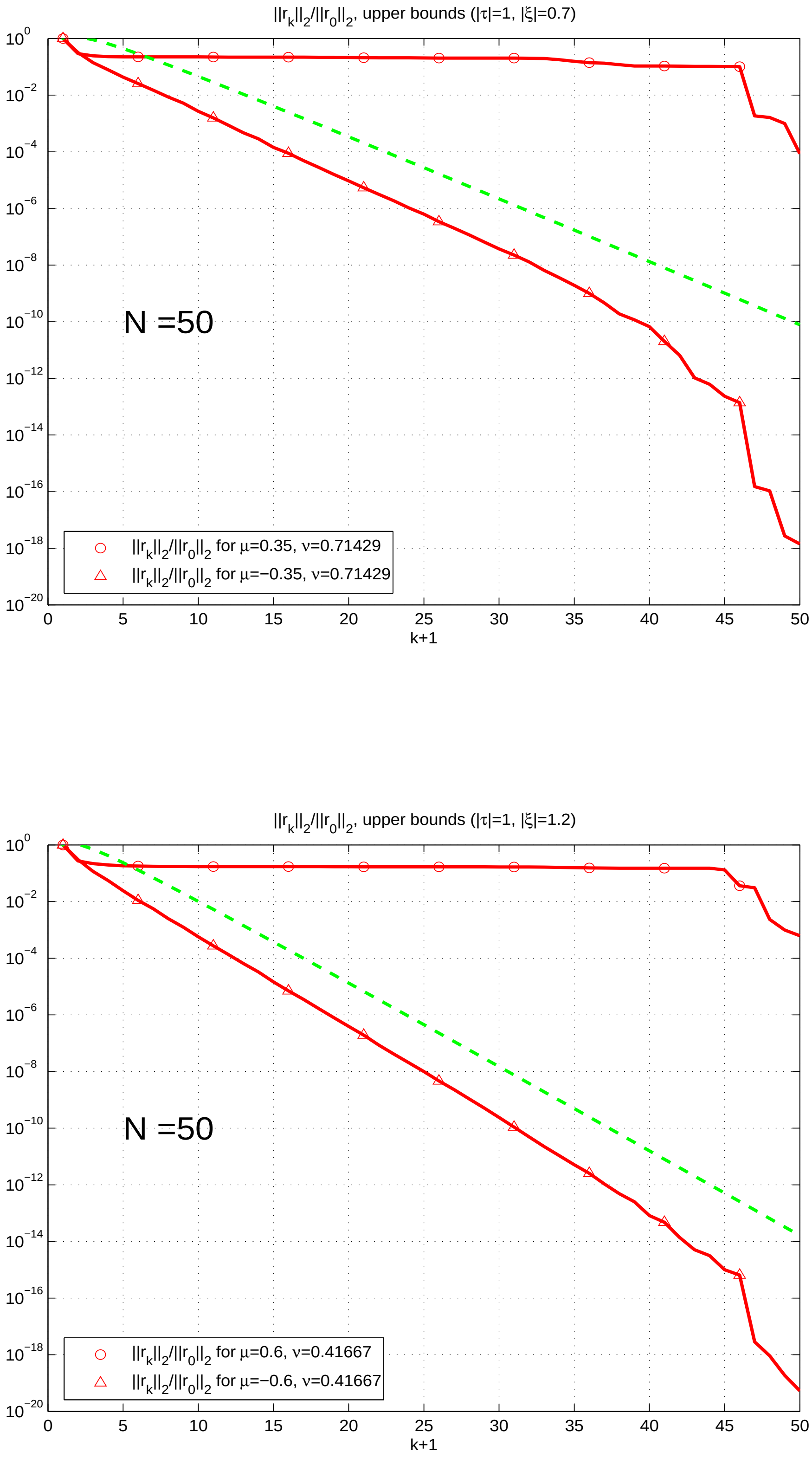

Figure 3.3.6: GMRES residuals for random b with $|\tau| = 1$, and the upper bounds by Theorem 3.3.1.

Thus μ, $\nu \in \mathbb{R}$, and in fact $\nu > 0$ always. When $\mu > 0$, $\xi = -\sqrt{\mu/\nu} < 0$ and $\tau = 1/(2\sqrt{\mu\nu}) > 0$, but when $\mu < 0$, both $\xi = -\iota\sqrt{|\mu/\nu|}$ and $\tau = -\iota/(2\sqrt{|\mu\nu|})$ are imaginary, where $\iota = \sqrt{-1}$ is the imaginary unit.

Figures 3.3.4, 3.3.5, 3.3.6 plot residual histories for examples of GMRES with each of b's entries being uniformly random in $[-1, 1]$, and their upper bounds (dashed lines) by Theorem 3.3.1. The parameters, τ and ξ are set in table 3.3.1.

Table 3.3.1: Parameters Setting

	$\lvert\tau\rvert$	$\lvert\xi\rvert$	
Figure 3.3.4	0.8	0.7	1.2
Figure 3.3.5	1.2	0.7	1.2
Figure 3.3.6	1	0.7	1.2

Each figure has two graphs with different ξ, 0.7 or 1.2. Each graph has two cases for a real τ and a pure imaginary τ. In each graph, the plot for $\mu > 0$, i.e. τ is real, is above that for $\mu < 0$, i.e. τ is purely imaginary. In other words, this indicates that GMRES converges much faster for $\mu < 0$ than for $\mu > 0$ in each of the plots. There is a simple explanation for this: the eigenvalues of A (see (3.2.9) below) are further away from the origin for a pure imaginary τ than for a real τ for any fixed $|\tau|$.

All plots indicate that our upper bounds are tight, except for the last few steps. Note that the upper bounds for the case $\mu > 0$ are visually indistinguishable from the horizontal line 10^0 when $\tau \leq 1$, suggesting slow convergence.

Figures 3.3.7, 3.3.8, 3.3.9 plot residual histories for examples of GMRES with each of b's entries being uniformly random in $[-1, 1]$, and their upper bounds (dashed lines) by Theorem 3.3.2. In each figure, the upper bounds are much bigger than the residuals.

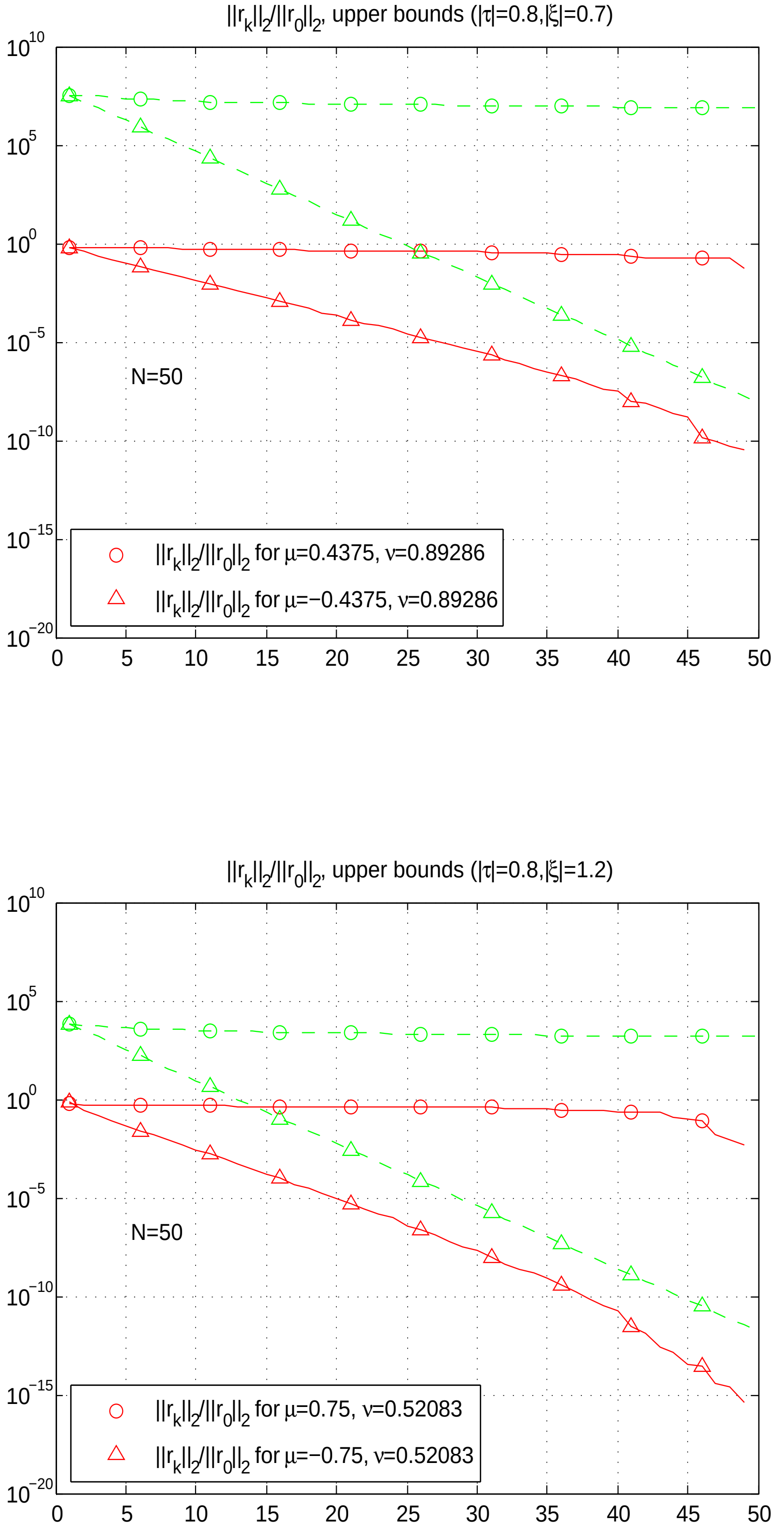

Figure 3.3.7: GMRES residuals for random b with $|\tau| = 0.8$, and the upper bounds by Theorem 3.3.2.

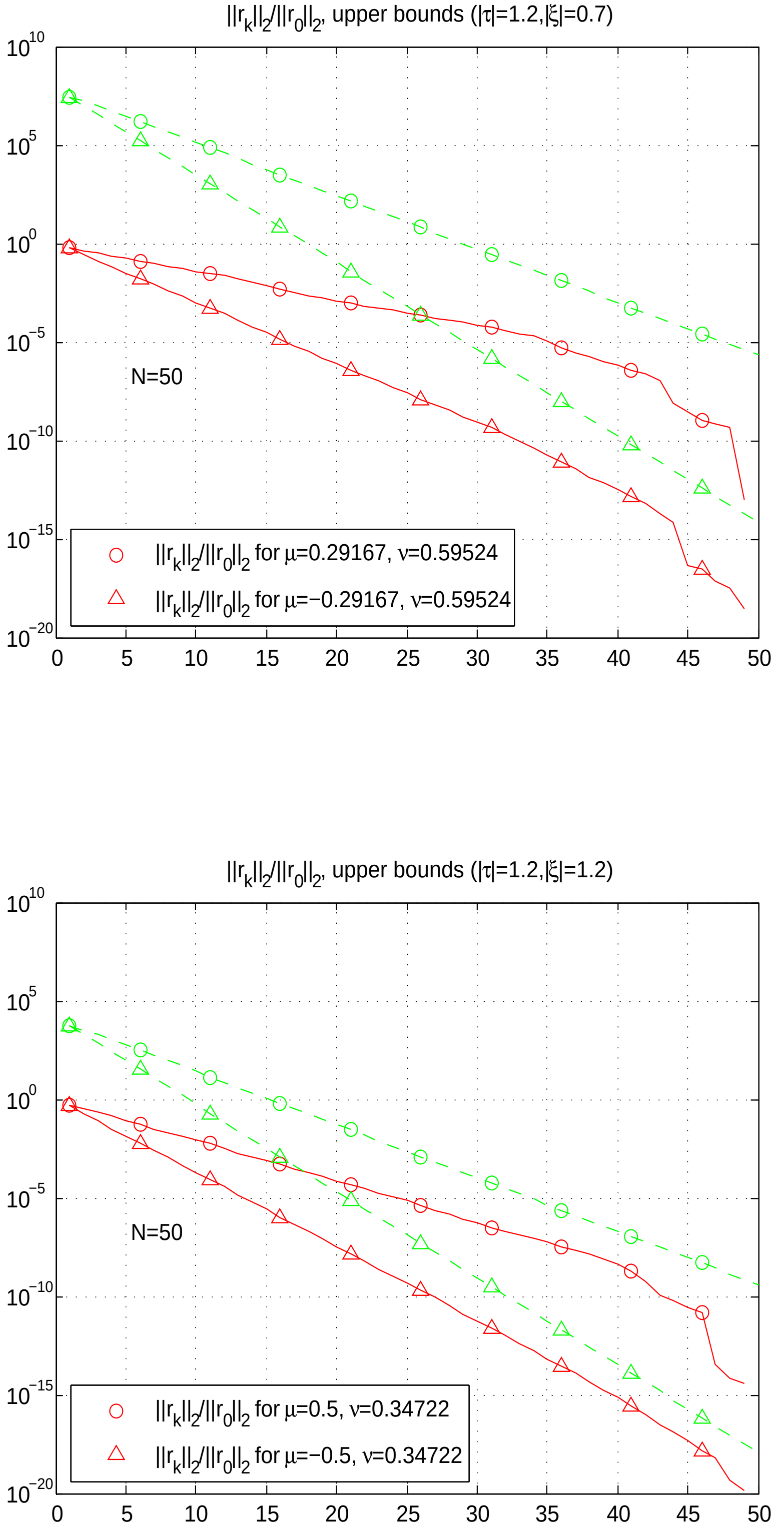

Figure 3.3.8: GMRES residuals for random b with $|\tau| = 1.2$, and the upper bounds by Theorem 3.3.2.

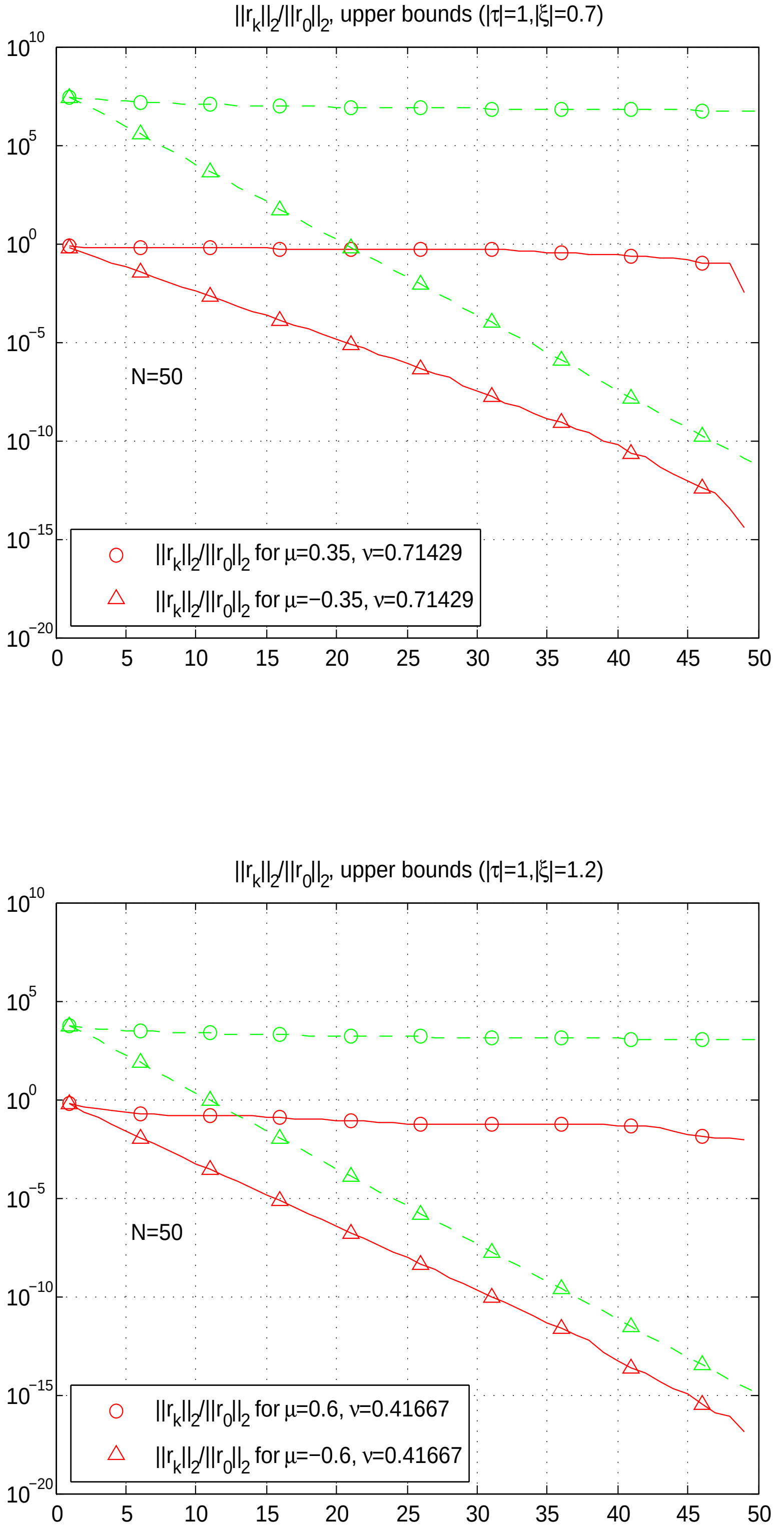

Figure 3.3.9: GMRES residuals for random b with $|\tau| = 1$, and the upper bounds by Theorem 3.3.2.

3.4 Special Right-hand Sides: $\boldsymbol{b = e_1, b = e_N}$

We now consider three special right-hand sides: $b = e_1$, e_N or $b_{(1)}e_1 + b_{(N)}e_N$. In particular, we also can find out the lower bound, and show that the upper bound in Theorem 3.3.1 is within a factor about $(k+1)$ of the true residual for $b = e_1$ or e_N, depending on $|\xi| \leq 1$ or $|\xi| \geq 1$, respectively.

3.4.1 Right-hand Sides $\boldsymbol{b = e_1}$

For the right-hand sides: $b = e_1$, we have the following theorem .

Theorem 3.4.1. *If $b = e_1$, then the kth GMRES residual r_k satisfies for $1 \leq k < N$*

$$\frac{1}{2}\left[\sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1}|\xi|^{2j}\right]^{-1}\left[\Phi_{k+1}^{(+)}(\tau,\xi)-\frac{1}{4}\right]^{-1/2} \leq \frac{\|r_k\|_2}{\|r_0\|_2} \leq \frac{1}{2}(1+|\xi|^2)\left[\Phi_{k+1}^{(+)}(\tau,\xi)-\frac{1}{4}\right]^{-1/2}. \tag{3.4.1}$$

In particular,

$$\frac{1}{2\lceil\frac{k+1}{2}\rceil}\left[\Phi_{k+1}^{(+)}(\tau,\xi)-\frac{1}{4}\right]^{-1/2} \leq \frac{\|r_k\|_2}{\|r_0\|_2} \leq \left[\Phi_{k+1}^{(+)}(\tau,\xi)-\frac{1}{4}\right]^{-1/2}, \qquad \textit{for } |\xi| \leq 1. \tag{3.4.2}$$

Proof: If $b = e_1$, then $M = M_1$ is upper triangular. More specifically

$$M = M_1 = \begin{pmatrix} 1 & 0 & -1/2 & & \\ & 1/2 & 0 & \ddots & \\ & & 1/2 & & -1/2 \\ & & & \ddots & 0 \\ & & & & 1/2 \end{pmatrix}, \tag{3.4.3}$$

and, by (3.2.20),

$$YV_{k+1,N} = \begin{pmatrix} \Omega_{k+1}\widetilde{M}R_{k+1}^{-1} \\ 0 \end{pmatrix} = \begin{pmatrix} \Omega_{k+1}\widetilde{M}\Omega_{k+1}^{-1}D^{-1} \times D\Omega_{k+1}R_{k+1}^{-1} \\ 0 \end{pmatrix},$$

where $\widetilde{M} = M_{(1:k+1,1:k+1)}$ and $D = \mathrm{diag}(2, 1, 1, \ldots, 1)$. Therefore

$$\sigma_{\min}(\Omega_{k+1}\widetilde{M}\Omega_{k+1}^{-1}D^{-1}) \leq \frac{\min_{u_{(1)}=1}\|Y\,V_{k+1,N}^T u\|_2}{\min_{u_{(1)}=1}\|D\Omega_{k+1}R_{k+1}^{-1}u\|_2} \leq \|\Omega_{k+1}\widetilde{M}\Omega_{k+1}^{-1}D^{-1}\|_2. \tag{3.4.4}$$

Let $P_{k+1} = (e_1, e_3, \ldots, e_2, e_4, \ldots) \in \mathbb{R}^{(k+1)\times(k+1)}$. It can be seen that

$$P_{k+1}^T(\Omega_{k+1}\widetilde{M}\Omega_{k+1}^{-1}D^{-1})P_{k+1} = \frac{1}{2}\begin{pmatrix} E_1 & \\ & E_2 \end{pmatrix},$$

where $E_1 \in \mathbb{R}^{\lceil\frac{k+1}{2}\rceil\times\lceil\frac{k+1}{2}\rceil}$, $E_2 \in \mathbb{R}^{\lfloor\frac{k+1}{2}\rfloor\times\lfloor\frac{k+1}{2}\rfloor}$, and

$$E_i = \begin{pmatrix} 1 & -\xi^2 & & \\ & 1 & \ddots & \\ & & \ddots & -\xi^2 \\ & & & 1 \end{pmatrix}, \ E_i^{-1} = \begin{pmatrix} 1 & \xi^2 & \cdots & \xi^{2(m-1)} \\ & 1 & \ddots & \vdots \\ & & \ddots & \xi^2 \\ & & & 1 \end{pmatrix}.$$

Hence $\|E_i\|_2 \le \sqrt{\|E_i\|_1\|E_i\|_\infty} = 1 + |\xi|^2$. Therefore

$$\|\Omega_{k+1}\widetilde{M}\Omega_{k+1}^{-1}D^{-1}\|_2 = \frac{1}{2}\max\{\|E_1\|_2, \|E_2\|_2\} \le \frac{1}{2}(1+|\xi|^2).$$

Similarly use $\|E_i^{-1}\|_2 \le \sqrt{\|E_i^{-1}\|_1\|E_i^{-1}\|_\infty}$ to get

$$\|E_1^{-1}\|_2 \le \sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1}|\xi|^{2j}, \quad \|E_2^{-1}\|_2 \le \sum_{j=0}^{\lfloor\frac{k+1}{2}\rfloor-1}|\xi|^{2j}.$$

Therefore

$$\begin{aligned}\sigma_{\min}(\Omega_{k+1}\widetilde{M}\Omega_{k+1}^{-1}D^{-1}) &= \frac{1}{2}\min\{\sigma_{\min}(E_1), \sigma_{\min}(E_2)\} \\ &= \frac{1}{2}\min\{\|E_1^{-1}\|_2^{-1}, \|E_2^{-1}\|_2^{-1}\} \\ &\ge \frac{1}{2}\left[\sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1}|\xi|^{2j}\right]^{-1}.\end{aligned}$$

Finally, by Lemma 3.3.1, we have

$$\min_{u_{(1)}=1}\|D\Omega_{k+1}R_{k+1}^{-1}u\|_2 = \|D^{-*}\Omega_{k+1}^{-*}R_{k+1}^{*}e_1\|_2^{-1} = \left[\Phi_{k+1}^{(+)}(\tau,\xi) - \frac{1}{4}\right]^{-1/2}.$$

This, together with (3.4.4), lead to (3.4.1). ∎

3.4.2 Numerical Examples for $b = e_1$

The numerical examples are presented in Figures 3.4.1, 3.4.2,and 3.4.3. Figure 3.4.1 shows GMRES residuals for $b = e_1$ with $\tau = 0.8$, and their bounds according to Theorem 3.4.1; Figure 3.4.2 shows GMRES residuals for $b = e_1$ with $\tau = 1.2$; and Figure 3.4.3 with $\tau = 1$.

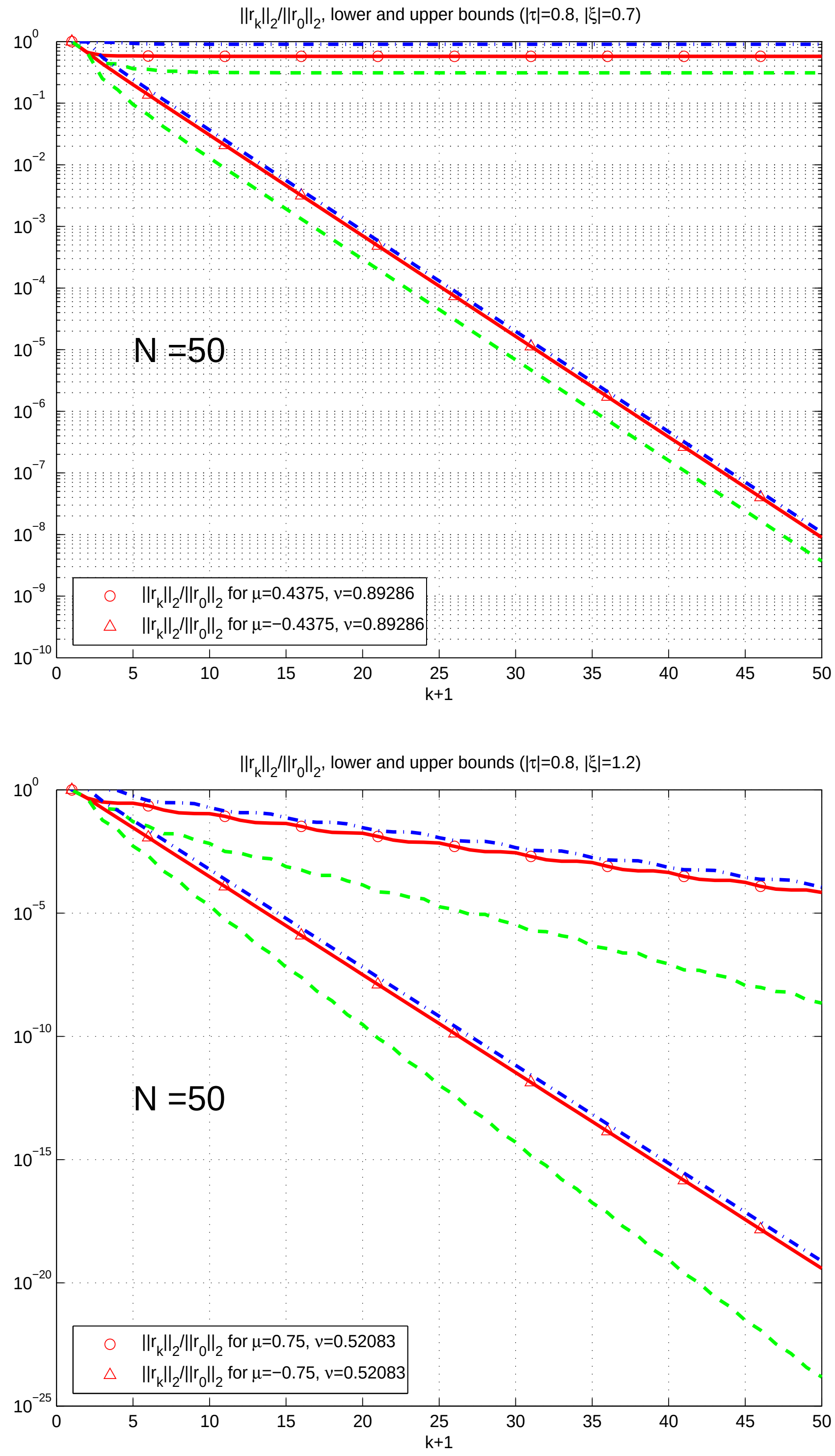

Figure 3.4.1: GMRES residuals for $b = e_1$, $|\tau| = 0.8$, and the bounds by Theorem 3.4.1.

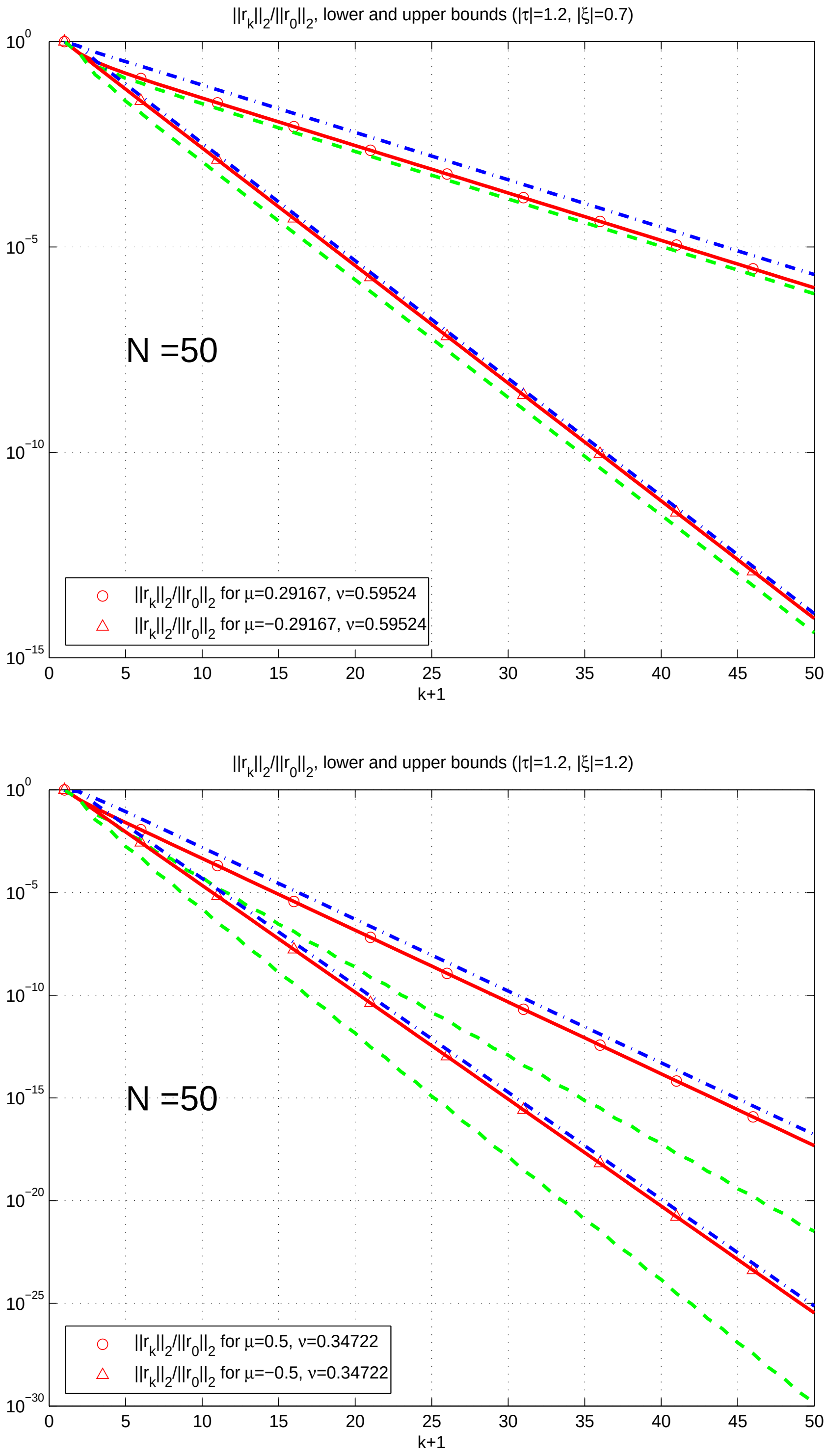

Figure 3.4.2: GMRES residuals for $b = e_1$, $|\tau| = 1.2$, and the bounds by Theorem 3.4.1.

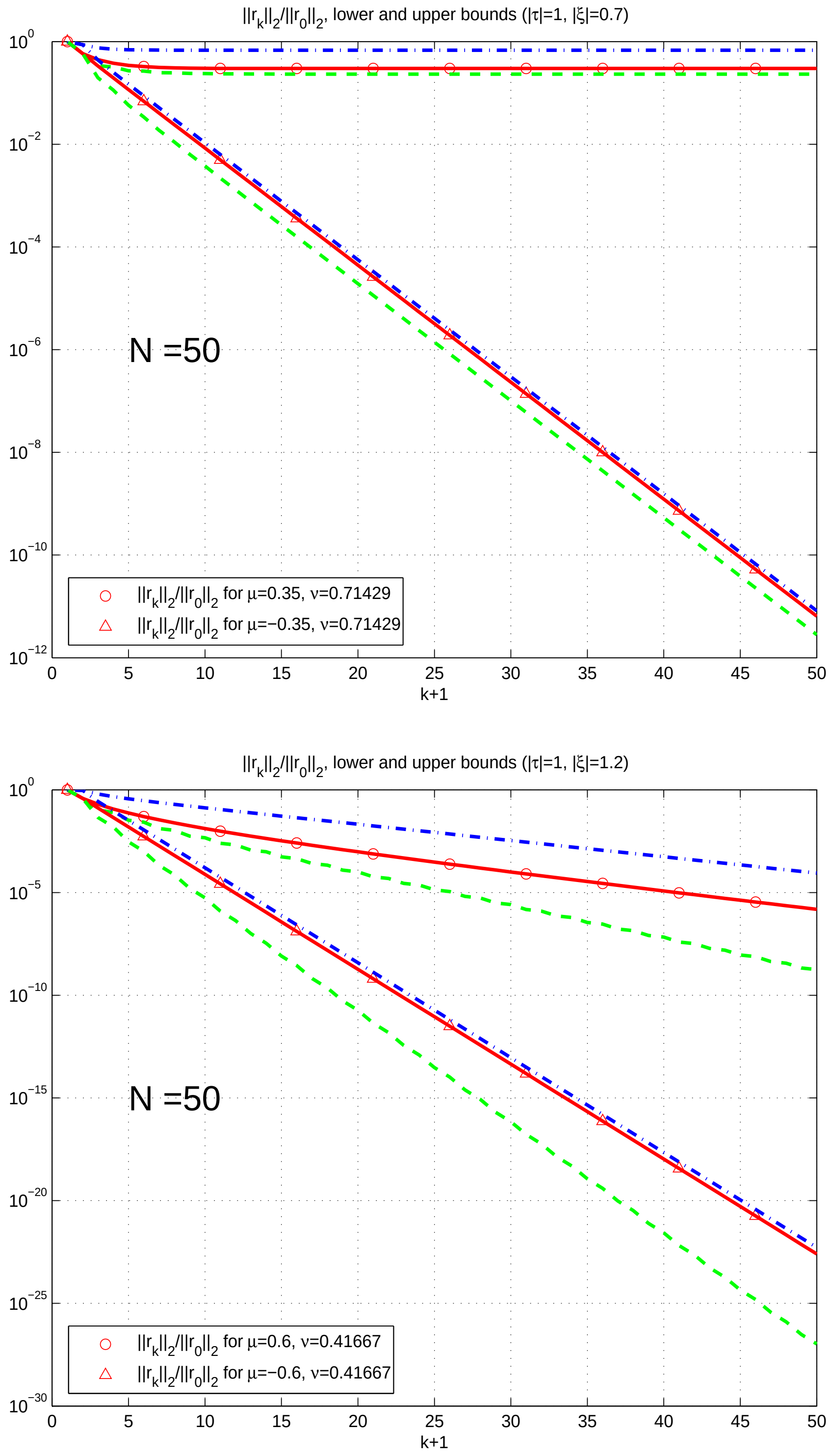

Figure 3.4.3: GMRES residuals for $b = e_1$, $|\tau| = 1$, and the bounds by Theorem 3.4.1.

The upper bounds are blue dashed lines, the lower bounds are green dashed lines. The figures show that GMRES residuals for $b = e_1$ are sandwiched by the lower and upper bounds by Theorem 3.4.1. The upper bounds are very close to the GMRES residuals, when either $|\xi| \leq 1$ or $|\xi| > 1$; the lower bounds are also good when $|\xi| > 1$, but not as good as the cases of $|\xi| \leq 1$.

The upper bound and the lower bound in (3.4.2) differ by a factor roughly $(k+1)$, and thus they are rather sharp; if $|\xi| \leq 1$, the bounds in (3.4.1) are even sharper. The upper bound according to Theorem 3.4.1 is within a factor about $(k+1)$ of the true residual for $b = e_1$.

3.4.3 Right-hand Sides $b = e_N$

For $b = e_N$, we have a similar theorem.

Theorem 3.4.2. *In* Theorem 3.3.1*, if $b = e_N$, then the kth GMRES residual r_k satisfies for $1 \leq k < N$*

$$\frac{1}{2}\left[\sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1} |\xi|^{-2j}\right]^{-1}\left[\Omega_{k+1}^{(-)}(\tau,\xi)-\frac{1}{4}\right]^{-1/2} \leq \frac{\|r_k\|_2}{\|r_0\|_2} \leq \frac{1}{2}(1+|\xi|^{-2})\left[\Omega_{k+1}^{(-)}(\tau,\xi)-\frac{1}{4}\right]^{-1/2}. \tag{3.4.5}$$

In particular,

$$\frac{1}{2\lceil\frac{k+1}{2}\rceil}\left[\Omega_{k+1}^{(-)}(\tau,\xi)-\frac{1}{4}\right]^{-1/2} \leq \frac{\|r_k\|_2}{\|r_0\|_2} \leq \left[\Omega_{k+1}^{(-)}(\tau,\xi)-\frac{1}{4}\right]^{-1/2}, \qquad \textit{for } |\xi| \geq 1. \tag{3.4.6}$$

Proof: As in the proof of Theorem 3.3.1, by applying Theorem 3.4.1 to the permuted system (3.3.19), we get Theorem 3.4.2 for $b = e_N$. ∎

The upper bound and the lower bound in (3.4.6) differ by a factor roughly $(k+1)$, and thus they are rather sharp; so are the bounds in (3.4.5) for $|\xi| \geq 1$.

The results of numerical examples for $b = e_N$ are very similar to these for $b = e_1$ in the previous subsection, and are not presented again. The upper bound and the lower bound in (3.4.6) differ by a factor roughly $(k+1)$, and thus they are rather sharp; if $|\xi| > 1$, the bounds in (3.4.5) are even sharper.

Our numerical examples indicate that the upper bounds are rather good regardless of the magnitude of $|\xi|$ for both cases $b = e_1$ and $b = e_N$.

3.4.4 Right-hand Sides $b_{(1)}e_1 + b_{(N)}e_N$

Given Theorems 3.4.1 and 3.4.2, it would not be unreasonable to expect that the upper bound in Theorem 3.3.1 would be sharp for very large or tiny $|\xi|$ within a factor possibly about at most $(k+1)^{3/2}$ for right-hand side b with $b_{(i)} = 0$ for $2 \le i \le N-1$ and $|b_{(1)}| = |b_{(N)}| > 0$. The following theorem indeed confirms this but only for $k \le N/2$. Our numerical examples even support that the lower bounds by (3.4.7) would be good for $k > N/2$ (see Figure 3.4.4), too, but we do not have a way to mathematically justify it yet.

Theorem 3.4.3. *In* Theorem 3.3.1, *if* $b_{(i)} = 0$ *for* $2 \le i \le N-1$, *then the* k*th GMRES residual* r_k *satisfies*

$$\frac{\|r_k\|_2}{\|r_0\|_2} \ge \frac{\min\limits_{i\in\{1,N\}} |b_{(i)}|}{2\chi\,\|r_0\|_2}\left[\Phi_{k+1}(\tau,\xi) - \frac{1}{4}\right]^{-1/2}, \quad \textit{for } 1 \le k \le N/2, \tag{3.4.7}$$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \sqrt{3}\left[\frac{1}{2} + \Phi_{k+1}(\tau,\xi)\right]^{-1/2}, \tag{3.4.8}$$

where

$$1 < \chi = \sum_{j=0}^{\lceil\frac{k+1}{2}\rceil - 1} \zeta^{2j} \le \left\lceil \frac{k+1}{2} \right\rceil.$$

Proof: Now $b = b_{(1)}e_1 + b_{(N)}e_N$. Notice the form of M_1 in (3.4.3), and that M_N is M_1 after its rows reordered from the last to the first. For the case $M = b_{(1)}M_1 + \xi^{N-1}b_{(N)}M_N$,

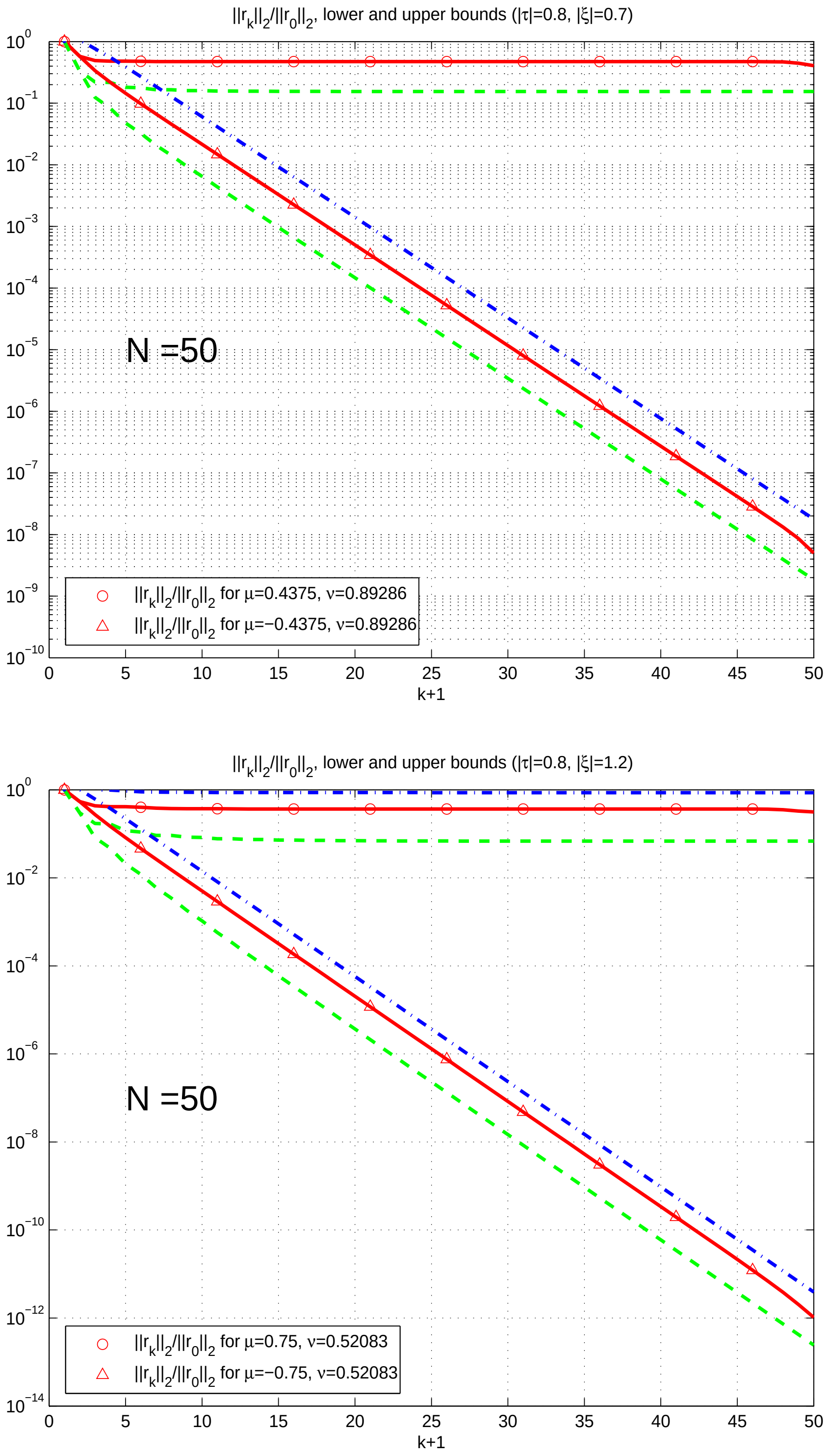

Figure 3.4.4: GMRES residuals for $b = e_1 + e_N$, $|\tau| = 0.8$, and the bounds by Theorem 3.4.3.

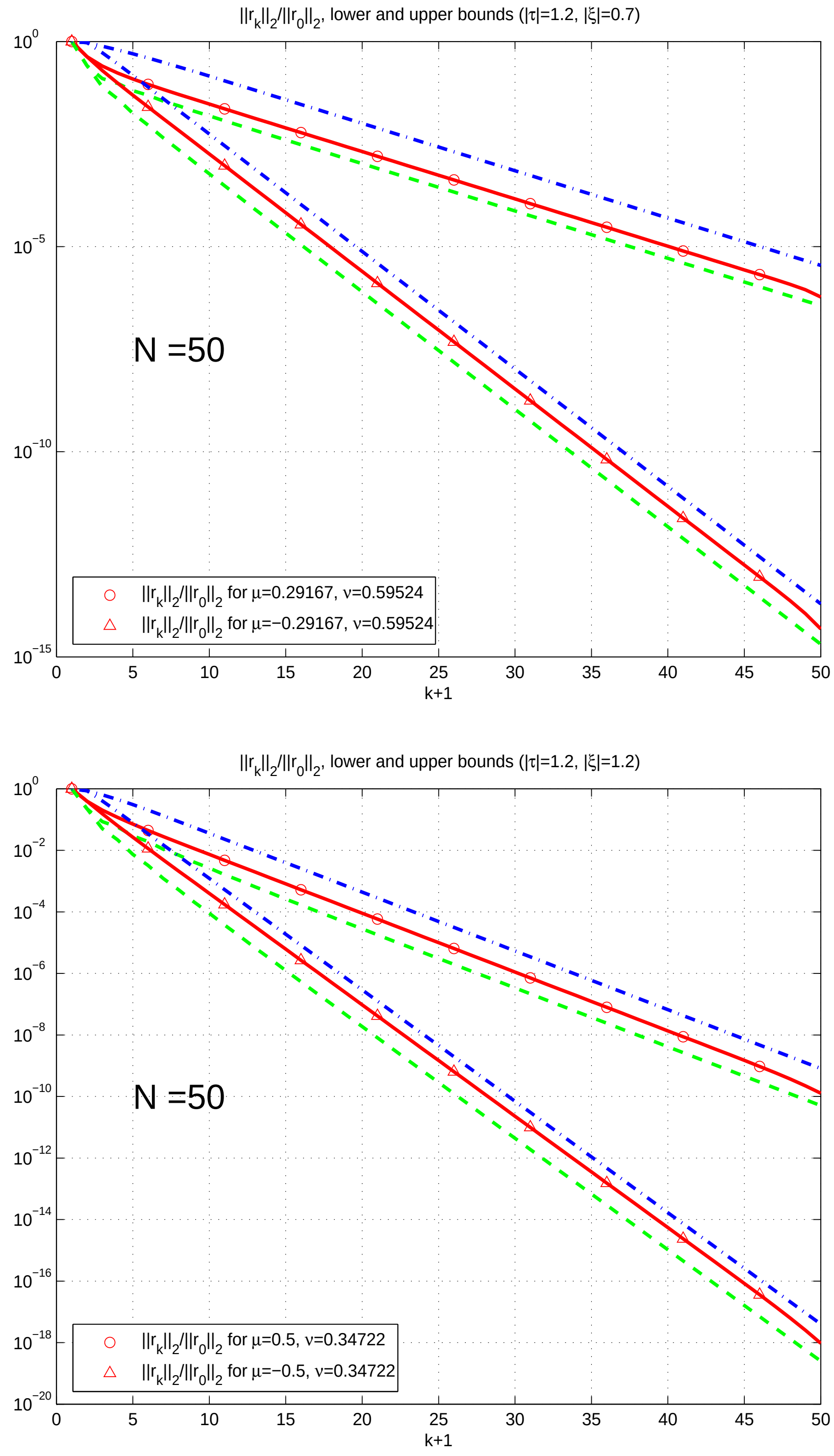

Figure 3.4.5: GMRES residuals for $b = e_1 + e_N$, $|\tau| = 1.2$, and the bounds by Theorem 3.4.3.

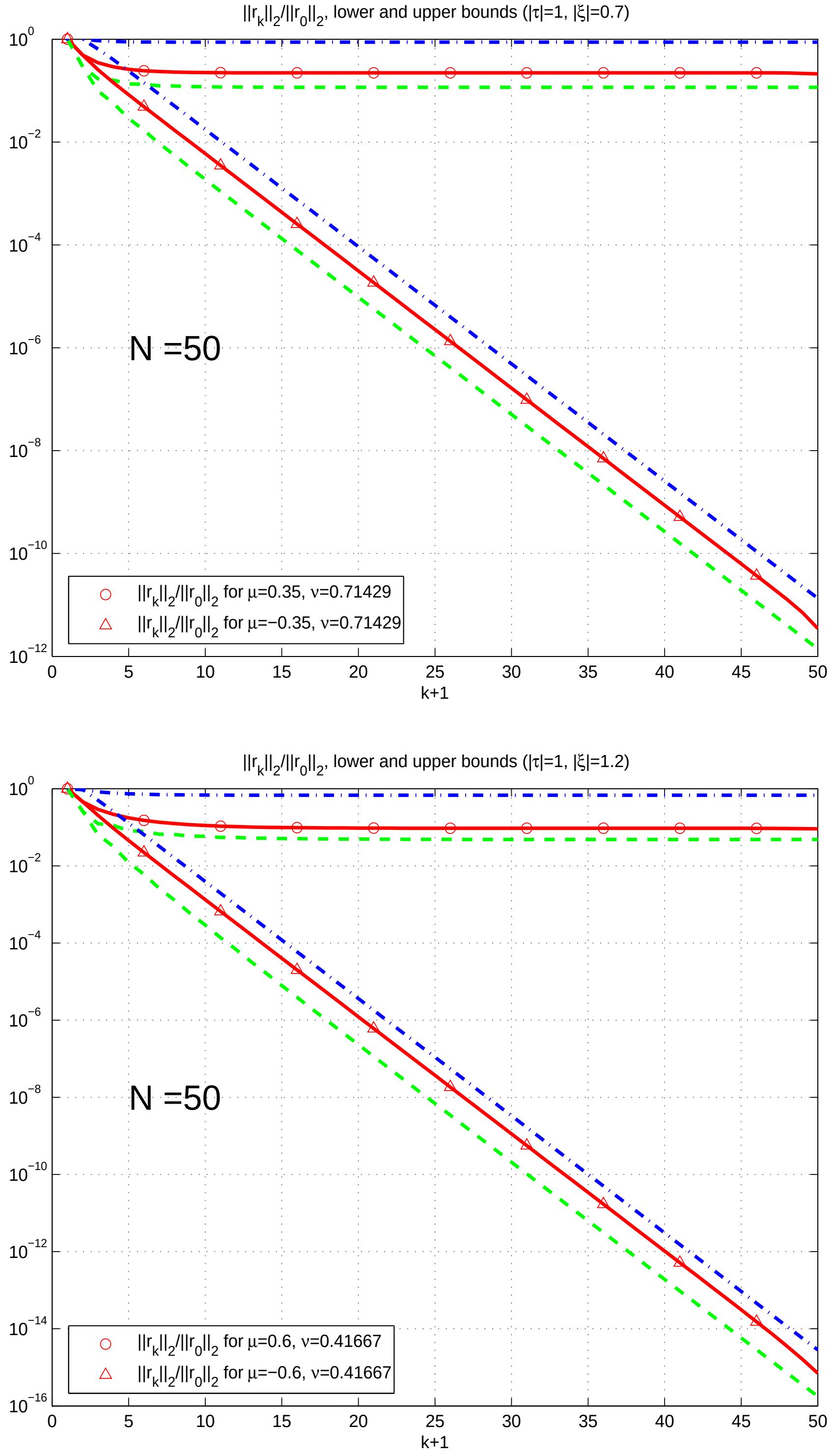

Figure 3.4.6: GMRES residuals for $b = e_1 + e_N$, $|\tau| = 1$, and the bounds by Theorem 3.4.3.

and also Lemma 3.3.4 implies that only positive powers of ξ appear in the entries of $\Omega M \Omega^{-1}$. Therefore when $|\xi| \leq 1$,

$$\begin{aligned} \|\Omega M_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2 &\leq \|\Omega M \Omega^{-1}\|_2 \\ &\leq |b_{(1)}| \, \| \, |M_1| \, \|_2 + |b_{(N)}| |\xi|^{N-1} \, \| \, |M_N| \, \|_2 \\ &\leq |b_{(1)}| \sqrt{3/2} + |b_{(N)}| \sqrt{3/2} \\ &\leq \sqrt{3} \|b\|_2, \end{aligned} \tag{3.4.9}$$

where $|M_\ell|$ takes entrywise absolute value, and we have used

$$\| \, |M_N| \, \|_2 = \| \, |M_1| \, \|_2 \leq \sqrt{\|M_1\|_1 \|M_1\|_\infty} = \sqrt{3/2}.$$

Inequality (3.4.8) for $|\xi| \leq 1$ is the consequence of (3.2.24), (3.3.13), and (3.4.9). Inequality (3.4.8) for $|\xi| \geq 1$ follows from itself for $|\xi| \leq 1$ applied to the permuted system (3.3.19).

To prove (3.4.7), we use the lines of arguments in the proof for Theorem 3.4.1 and notice that for $1 \leq k \leq N/2$

$$Y V_{k+1,N} = \begin{array}{c} \\ {\scriptstyle k} \\ {\scriptstyle N-2k} \\ {\scriptstyle k} \end{array} \begin{array}{c} {\scriptstyle k} \\ \begin{pmatrix} W_1 \\ 0 \\ W_2 \end{pmatrix} \end{array}.$$

It can be seen from the proof for Theorem 3.4.1 that

$$\begin{aligned} \min_{u_{(1)}=1} \|W_1 u\|_2 &\geq \frac{|b_{(1)}|}{2} \left[\sum_{j=0}^{\lceil \frac{k+1}{2} \rceil - 1} |\xi|^{2j} \right]^{-1} \left[\Omega_{k+1}^{(+)}(\tau, \xi) - \frac{1}{4} \right]^{-1/2}, \\ \min_{u_{(1)}=1} \|W_2 u\|_2 &\geq \frac{|b_{(N)}|}{2} \left[\sum_{j=0}^{\lceil \frac{k+1}{2} \rceil - 1} |\xi|^{-2j} \right]^{-1} \left[\Omega_{k+1}^{(-)}(\tau, \xi) - \frac{1}{4} \right]^{-1/2}. \end{aligned}$$

Finally use

$$\min_{u_{(1)}=1} \|Y V_{k+1,N} u\|_2 \geq \max \left\{ \min_{u_{(1)}=1} \|W_1 u\|_2, \min_{u_{(1)}=1} \|W_2 u\|_2 \right\}$$

to complete the proof. ∎

Figuress 3.4.4, 3.4.5 and 3.4.6 plot residual histories for several examples of GMRES with $b = e_1 + e_N$. According to the graphs, GMRES residuals for $b = e_1 + e_N$, are sandwiched by their lower and upper bounds by (3.4.7) and (3.4.8). Strictly speaking, (3.4.7) is only proved for $k \leq N/2$, but it seems to be very good even for $k > N/2$ as well. We also ran GMRES for $b = e_1 - e_N$ and obtained residual history that is very much the same. Finally we have the following theorem about the asymptotic speeds of $\|r_k\|_2$ for $b = e_1$ and $b = e_N$.

Theorem 3.4.4. *Assume the conditions of* Theorem 3.3.1 *hold.*

1. *Let $b = e_1$. If $\rho > 1$, then*

$$\begin{aligned} \min\{(|\xi|^2\rho)^{-1}, (|\xi|\rho)^{-1}, 1\} &\le \liminf_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \\ &\le \limsup_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \le \min\{(|\xi|\rho)^{-1}, 1\}. \end{aligned} \tag{3.4.10}$$

If $\rho = 1$ (which happens when and only when $\tau \in [-1, 1]$), then

$$\min\{|\xi|^{-1}, 1\} \times \eta \le \liminf_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \le \limsup_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \le \eta, \tag{3.4.11}$$

where $\eta = \limsup_{k\to\infty} \left[1/4 + \sum_{j=1}^{k} |\xi|^{2j} (\cos j\theta)^2\right]^{-1/(2k)}$ *and* $\theta = \arccos\tau$. *Regardless of $\rho > 1$ or $\rho = 1$,*

$$\lim_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} = \min\{(|\xi|\rho)^{-1}, 1\} \quad \textit{for } |\xi| \le 1. \tag{3.4.12}$$

2. *Let $b = e_N$. If $\rho > 1$, then*

$$\begin{aligned} \min\{(|\xi|^{-2}\rho)^{-1}, (|\xi|^{-1}\rho)^{-1}, 1\} &\le \liminf_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \\ &\le \limsup_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \le \min\{(|\xi|^{-1}\rho)^{-1}, 1\}. \end{aligned} \tag{3.4.13}$$

If $\rho = 1$ (which happens when and only when $\tau \in [-1, 1]$), then

$$\min\{|\xi|, 1\} \times \eta \le \liminf_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \le \limsup_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \le \eta, \tag{3.4.14}$$

where $\eta = \limsup_{k\to\infty} \left[1/4 + \sum_{j=1}^{k} |\xi|^{-2j} (\cos j\theta)^2\right]^{-1/(2k)}$ *and* $\theta = \arccos\tau$. *Regardless of $\rho > 1$ or $\rho = 1$,*

$$\lim_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} = \min\{(|\xi|^{-1}\rho)^{-1}, 1\} \quad \textit{for } |\xi| \ge 1. \tag{3.4.15}$$

Proof: We note that

$$\limsup_{k\to\infty} \left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \le 1, \quad \limsup_{k\to\infty} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \le 1$$

for any b because $\|r_k\|_2$ is nonincreasing.

Suppose $b = e_1$. Consider first $\rho > 1$. Then $|T_j(\tau)| \sim \frac{1}{2}\rho^j$, and thus

$$\Phi_{k+1}^{(+)}(\tau,\xi) - \frac{1}{4} \sim \frac{1}{4}\sum_{j=0}^{k}(|\xi|\rho)^{2j} = \frac{1}{4}\cdot\frac{(|\xi|\rho)^{2(k+1)}-1}{(|\xi|\rho)^2-1}. \tag{3.4.16}$$

If $|\xi|\rho > 1$, then (3.4.16) and Theorem 3.4.1 imply

$$\begin{aligned}
\limsup_{k\to\infty}\left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} &\le \lim_{k\to\infty}\left[\frac{1}{2}(1+|\xi|^2)\right]^{1/k}\left[\Omega_{k+1}^{(+)}(\tau,\xi)-\frac{1}{4}\right]^{-1/(2k)} &(3.4.17)\\
&= (|\xi|\rho)^{-1},\\
\liminf_{k\to\infty}\left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} &\ge \lim_{k\to\infty}\frac{1}{2^{1/k}}\left[\sum_{j=0}^{\lceil\frac{k+1}{2}\rceil-1}|\xi|^{2j}\right]^{-1/k}\left[\Omega_{k+1}^{(+)}(\tau,\xi)-\frac{1}{4}\right]^{-1/(2k)} &(3.4.18)\\
&= \begin{cases}(|\xi|\rho)^{-1}, & \text{for } |\xi|\le 1,\\ (|\xi|^2\rho)^{-1}, & \text{for } |\xi|>1.\end{cases}
\end{aligned}$$

They together give (3.4.10) for the case $|\xi|\rho > 1$. If $|\xi|\rho \le 1$, then must $|\xi| < 1$ and $\min\{(|\xi|\rho)^{-1}, 1\} = 1$, $\min\{(|\xi|^2\rho)^{-1}, (|\xi|\rho)^{-1}, 1\} = 1$, and

$$\liminf_{k\to\infty}\left[\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \ge 1$$

by (3.4.18) because $\Phi_{k+1}^{(+)}(\tau,\xi) - \frac{1}{4}$ is approximately bounded by $(k+1)/4$ by (3.4.16). So (3.4.10) holds for the case $|\xi|\rho \le 1$, too. Now consider $\rho = 1$. Then $\tau + \sqrt{\tau^2-1} = e^{\iota\theta}$ for some $0 \le \theta \le \pi$, where $\iota = \sqrt{-1}$ is the imaginary unit. Thus $\tau \in [-1, 1]$ and in fact

$$2\tau = (\tau + \sqrt{\tau^2-1}) + (\tau - \sqrt{\tau^2-1}) = 2\cos\theta, \quad T_j(\tau) = \cos j\theta.$$

Therefore $\Phi_{k+1}^{(+)}(\tau,\xi) - \frac{1}{4} \sim \frac{1}{4} + \sum_{j=1}^{k}|\xi|^{2j}(\cos j\theta)^2$ which implies

$$\lim_{k\to\infty}\left[\Phi_{k+1}^{(+)}(\tau,\xi) - \frac{1}{4}\right]^{-1/(2k)} = \eta.$$

Inequalities (3.4.17) and (3.4.18) remain valid and yield (3.4.11). Finally regardless of $\rho > 1$ or $\rho = 1$, if $|\xi| \le 1$, then all leftmost sides and rightmost sides in (3.4.10) and

(3.4.11) are equal to $\min\{(|\xi|\rho)^{-1}, 1\}$. This proves (3.4.12). The proof for the case $b = e_1$ is done.

The case for $b = e_N$ can be dealt with by applying the results for $b = e_1$ to the permuted system (3.3.19). ∎

Remark 3.4.1. As we commented before, our numerical examples indicate that the upper bounds in Theorems 3.4.1 and 3.4.2 are rather accurate regardless of the magnitude of $|\xi|$ for both cases $b = e_1$ and $b = e_N$ (see Figure 3.4.1) and the lower bound in Theorem 3.4.3 is also accurate regardless of whether $k \leq N/2$ or not (see Figure 3.4.4). This leads us to conjecture that the following equations would hold.

$$\lim_{k\to\infty} \|r_k\|_2^{1/k} = \min\{(|\xi|\rho)^{-1}, 1\} \qquad \text{for } b = e_1, \tag{3.4.19}$$

$$\lim_{k\to\infty} \|r_k\|_2^{1/k} = \min\{(|\xi|^{-1}\rho)^{-1}, 1\} \quad \text{for } b = e_N, \tag{3.4.20}$$

where no constraint is assumed between k and N, except $k < N$ as usual.

3.5 Worst Convergence Speed

Furthermore, Theorem 3.5.1 tells the worst asymptotic speed for $\|r_k\|_2$.

Theorem 3.5.1. *Under the conditions of* Theorem 3.3.1,

$$\lim_{k\to\infty} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} = \lim_{k\to\infty} \left[\max_{r_0\in\{e_1,e_N\}} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} = \min\{(\zeta\rho)^{-1}, 1\}. \tag{3.5.1}$$

Proof: Note again that $\limsup_{k\to\infty} \left(\sup_{r_0} \|r_k\|_2/\|r_0\|_2\right)^{1/k} \leq 1$.

First we prove

$$\limsup_{k\to\infty} \left[\max_{r_0\in\{e_1,e_N\}} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \leq \limsup_{k\to\infty} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \leq \min\{(\zeta\rho)^{-1}, 1\}. \tag{3.5.2}$$

The first inequality is obvious because $\{e_1, e_N\} \in \{r_0\}$. We now prove the second one. If $\rho = 1$, then $\min\{(\zeta\rho)^{-1}, 1\} = 1$ because $\zeta^{-1} \geq 1$; no proof is needed. If $\rho > 1$, then $|T_j(\tau)| \sim \frac{1}{2}\rho^j$, and thus (3.4.16). Now if $\zeta\rho > 1$, then (3.4.16) and Theorem 3.3.1 imply $\limsup_{k\to\infty} \left(\sup_{r_0} \|r_k\|_2/\|r_0\|_2\right)^{1/k} \leq (\zeta\rho)^{-1}$ which also holds if $\zeta\rho \leq 1$ because then $(\zeta\rho)^{-1} \geq 1$.

Next we prove

$$\liminf_{k\to\infty} \left[\max_{r_0\in\{e_1,e_N\}} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \geq \min\{(\zeta\rho)^{-1}, 1\}. \tag{3.5.3}$$

If $|\xi| \leq 1$, then $\zeta = |\xi|$ and thus

$$\liminf_{k\to\infty} \left[\max_{r_0\in\{e_1,e_N\}} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} \geq \liminf_{k\to\infty} \left[\max_{r_0=e_1} \frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} = \min\{(\zeta\rho)^{-1}, 1\},$$

by (3.4.12) in Theorem 3.4.4. This is (3.5.3) for $|\xi| \leq 1$. For the case $|\xi| \geq 1$, we also have (3.5.3) similarly by (3.4.15). The proof is completed by combining (3.5.2) and (3.5.3). ∎

Chapter 4

Convergence Analysis Using Chebyshev Polynomials of the Second Kind

In this chapter, we use Chebyshev polynomials of the second kind to analyze the residuals of GMRES on tridiagonal Toeplitz. Section 4.1 introduces how to evaluate the residuals by applying Chebyshev polynomials of the second kind. Section 4.2 derives the upper bound of the residuals. Section 4.3 analyzes the residuals on special cases $b = e_1$ or $b = e_N$. Some of the complicated computations, needed in Section 4.2, are presented in Section 4.4.

4.1 Residual Reformulation Using Chebyshev Polynomials of the Second Kind

As in Section 3.2, for a nonsymmetric linear system (3.2.1), the kth GMRES residual can be written as

$$\|r_k\|_2 = \min_{u_{(1)}=1} \|Y\, V_{k+1,N}^T u\|_2. \tag{4.1.1}$$

In Section 3.2, the Chebyshev polynomial of the first kind is used to calculate the residual; here the kth residual is calculated through applying the Chebyshev polynomial of the second kind. We define the mth *Translated Chebyshev Polynomial of the second kind* in z of degree m as

$$\begin{aligned} U_m(z;\omega,\tau) &\overset{\text{def}}{=} U_m(z/\omega+\tau) && (4.1.2)\\ &= \tilde{a}_{mm} z^m + \tilde{a}_{m-1\,m} z^{m-1} + \cdots + \tilde{a}_{1m} z + \tilde{a}_{0m}, && (4.1.3)\end{aligned}$$

where $\tilde{a}_{jm} \equiv \tilde{a}_{jm}(\omega, \tau)$ are functions of ω and τ, and upper triangular $\tilde{R}_m \in \mathbb{C}^{m\times m}$, a matrix-valued function in ω and τ, too, as

$$\tilde{R}_{m+1} \equiv \tilde{R}_{m+1}(\omega,\tau) \overset{\text{def}}{=} \begin{pmatrix} \tilde{a}_{00} & \tilde{a}_{01} & \tilde{a}_{02} & \cdots & \tilde{a}_{0\,m} \\ & \tilde{a}_{11} & \tilde{a}_{12} & \cdots & \tilde{a}_{1\,m} \\ & & \tilde{a}_{22} & \cdots & \tilde{a}_{2\,m} \\ & & & \ddots & \vdots \\ & & & & \tilde{a}_{m\,m} \end{pmatrix}, \tag{4.1.4}$$

i.e., the jth column consists of the coefficients of $U_{j-1}(z;\omega,\tau)$. Set

$$\mathbf{U}_N \overset{\text{def}}{=} \begin{pmatrix} U_0(t_1) & U_0(t_2) & \cdots & U_0(t_N) \\ U_1(t_1) & U_1(t_2) & \cdots & U_1(t_N) \\ \vdots & \vdots & & \vdots \\ U_{N-1}(t_1) & U_{N-1}(t_2) & \cdots & U_{N-1}(t_N) \end{pmatrix} \tag{4.1.5}$$

and again $V_N = V_{N,N}$ for short. Then

$$V_N^T \tilde{R}_N = \mathbf{U}_N^T. \tag{4.1.6}$$

Equation (4.1.6) yields $V_N^T = \mathbf{U}_N^T \tilde{R}_N^{-1}$. Extracting the first $k+1$ columns from both sides of $V_N^T = \mathbf{U}_N^T \tilde{R}_N^{-1}$ yields

$$V_{k+1,N}^T = \mathbf{U}_{k+1,N}^T\, \tilde{R}_{k+1}^{-1}, \tag{4.1.7}$$

where $\mathbf{U}_{k+1,N} = (\mathbf{U}_N)_{(1:k+1,:)}$.

Now we can estimate the kth GMRES residual

$$\|r_k\|_2 = \min_{u_{(1)}=1} \|Y\, V_{k+1,N}^T u\|_2$$

for $Ax = b$, where $Y = X\,\mathrm{diag}(X^{-1}b)$. Now notice $Y = X\ \mathrm{diag}(X^{-1}b)$ and $X = \Omega Z$ with Z in (3.2.6) being real and orthogonal to get

$$\begin{aligned} Y\, V_{k+1,N}^T &= \Omega Z\ \mathrm{diag}(Z^T\Omega^{-1}b)\,(\mathbf{U}_N^T)_{(:,1:k+1)}\tilde{R}_{k+1}^{-1} \\ &= \Omega \tilde{M}_{(:,1:k+1)}\tilde{R}_{k+1}^{-1} & (4.1.8) \\ &= \Omega \tilde{M}_{(:,1:k+1)}\Omega_{k+1}^{-1}\,\Omega_{k+1}\tilde{R}_{k+1}^{-1}. & (4.1.9) \end{aligned}$$

where $\Omega_{k+1} = \Omega_{(1:k+1,1:k+1)}$, the $(k+1)$th leading submatrix of Ω,

$$\tilde{M} = Z \ \mathrm{diag}(Z^T \Omega^{-1} b) \, \mathbf{U}_N^T. \tag{4.1.10}$$

It follows from (4.1.1) and (4.1.9) that

$$\sigma_{\min}(\Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}) \leq \frac{\|r_k\|_2}{\min_{u_{(1)}=1} \|\Omega_{k+1} \tilde{R}_{k+1}^{-1} u\|_2} \leq \|\Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2. \tag{4.1.11}$$

Hence, the upper bound and lower bound of the residual $\|r_k\|_2$ could be estimated.

4.2 Estimation of Residual in General Case

Similarly to the previous chapter, we also define

$$\begin{aligned}
\zeta &\overset{\text{def}}{=} \min\left\{|\xi|, \frac{1}{|\xi|}\right\}, \\
\tilde{\Phi}_{k+1}^{(+)}(\tau, \xi) &\overset{\text{def}}{=} {\sum_{j=0}^{k}}' \, |\xi|^{2j} \left|U_j(\tau)\right|^2, \\
\tilde{\Phi}_{k+1}^{(-)}(\tau, \xi) &\overset{\text{def}}{=} {\sum_{j=0}^{k}}' \, |\xi|^{-2j} \left|U_j(\tau)\right|^2, \\
\tilde{\Phi}_{k+1}(\tau, \xi) &\overset{\text{def}}{=} {\sum_{j=0}^{k}}' \, \zeta^{2j} \left|U_j(\tau)\right|^2 \equiv \min\left\{\tilde{\Phi}_{k+1}^{(+)}(\tau, \xi), \tilde{\Phi}_{k+1}^{(-)}(\tau, \xi)\right\},
\end{aligned}$$

where $\sum_j'$ means the first term is halved. In the following part of this section, we try to calculate the upper bound of the residuals.

Based on inequality (4.1.11), we have

$$\|r_k\|_2 \leq \min_{u_{(1)}=1} \|\Omega_{k+1} \tilde{R}_{k+1}^{-1} u\|_2 \ \ \|\Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2, \tag{4.2.1}$$

and according to inequality (3.3.13), we have

$$\min_{u_{(1)}=1} \|\Omega_{k+1} \tilde{R}_{k+1}^{-1} u\|_2 = \|\Omega_{k+1}^{-*} \tilde{R}_{k+1}^{*} e_1\|_2^{-1} = \left[\frac{1}{2} + \Phi_{k+1}^{(+)}(\tau, \xi)\right]^{-1/2}. \tag{4.2.2}$$

Hence, if we can estimate the second part $\|\Omega\tilde{M}_{(:,1:k+1)}\Omega_{k+1}^{-1}\|_2$, then the upper bound is obtained.

Similarly, we analyze $\tilde{M}$ to calculate the second part. Recall Chebyshev polynomials of the second kind:

$$\begin{aligned} U_{m-1}(t) &= \frac{\sin(m \arccos t)}{\sin(\arccos t)} && \text{for } |t| \le 1, && (4.2.3) \\ &= \frac{\left(t+\sqrt{t^2-1}\right)^m + \left(t-\sqrt{t^2-1}\right)^m}{2\sqrt{t^2-1}} && \text{for } |t| \ge 1, && (4.2.4) \end{aligned}$$

and the definition of $\mathbf{U}_N$ in (4.1.5) and $t_j = \cos\theta_j, \theta_j = \frac{j\pi}{N+1}$ in (3.2.5). We rewrite $\mathbf{U}_N$ as

$$\mathbf{U}_N \overset{\text{def}}{=} \begin{pmatrix} \frac{\sin(\theta_1)}{\sin(\theta_1)} & \frac{\sin(\theta_2)}{\sin(\theta_2)} & \cdots & \frac{\sin(\theta_N)}{\sin(\theta_N)} \\ \frac{\sin(2\theta_1)}{\sin(\theta_1)} & \frac{\sin(2\theta_2)}{\sin(\theta_2)} & \cdots & \frac{\sin(2\theta_N)}{\sin(\theta_N)} \\ \vdots & \vdots & & \vdots \\ \frac{\sin(N\theta_1)}{\sin(\theta_1)} & \frac{\sin(N\theta_2)}{\sin(\theta_2)} & \cdots & \frac{\sin(N\theta_N)}{\sin(\theta_N)} \end{pmatrix}. \tag{4.2.5}$$

Then we rewrite $\tilde{M}$ as:

$$\begin{aligned} \tilde{M} &= Z \text{ diag}(Z^T\Omega^{-1}b)\, \mathbf{U}_N^T \\ &= \sum_{\ell=1}^{N} Z \text{ diag}(Z\Omega^{-1}b_{(\ell)}e_\ell)\, \mathbf{U}_N^T \\ &= \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1} Z \text{ diag}(Ze_\ell)\, \mathbf{U}_N^T \\ &= \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1} Z \text{ diag}(Z_{(:,\ell)})\, \mathbf{U}_N^T \\ &= \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1} \tilde{M}_\ell, \end{aligned} \tag{4.2.6}$$

where $\tilde{M}_\ell = Z \text{ diag}(Z_{(:,\ell)})\, \mathbf{U}_N^T$. Now recall $Z_{(:,\ell)} = \sqrt{\frac{2}{N+1}}(\sin \ell\theta_1, \ldots, \sin \ell\theta_N)^T$, then $\tilde{M}_\ell$

can be written as

$$\begin{aligned}
\tilde{M_\ell} &= \sqrt{\frac{2}{N+1}} \begin{pmatrix} \sin(\theta_1) & \sin(2\theta_1) & \cdots & \sin(N\theta_1) \\ \sin(\theta_2) & \sin(2\theta_2) & \cdots & \sin(N\theta_2) \\ \vdots & \vdots & & \vdots \\ \sin(\theta_N) & \sin(2\theta_N) & \cdots & \sin(N\theta_N) \end{pmatrix} \\
& \times\sqrt{\frac{2}{N+1}} \begin{pmatrix} \sin(\ell\theta_1) & & & \\ & \sin(\ell\theta_2) & & \\ & & \ddots & \\ & & & \sin(\ell\theta_N) \end{pmatrix} \\
& \times \begin{pmatrix} \frac{\sin(\theta_1)}{\sin(\theta_1)} & \frac{\sin(2\theta_1)}{\sin(\theta_1)} & \cdots & \frac{\sin(N\theta_1)}{\sin(\theta_1)} \\ \frac{\sin(\theta_2)}{\sin(\theta_2)} & \frac{\sin(2\theta_2)}{\sin(\theta_2)} & \cdots & \frac{\sin(N\theta_2)}{\sin(\theta_2)} \\ \vdots & \vdots & & \vdots \\ \frac{\sin(\theta_N)}{\sin(\theta_N)} & \frac{\sin(2\theta_N)}{\sin(\theta_N)} & \cdots & \frac{\sin(N\theta_N)}{\sin(\theta_N)} \end{pmatrix} \\
&= Z \begin{pmatrix} \frac{\sin(\ell\theta_1)}{\sin(\theta_1)} & & & \\ & \frac{\sin(\ell\theta_2)}{\sin(\theta_2)} & & \\ & & \ddots & \\ & & & \frac{\sin(\ell\theta_N)}{\sin(\theta_N)} \end{pmatrix} Z.
\end{aligned}$$

Since $Z = Z^T$, we also can write

$$\tilde{M_\ell} = Z\, D_\ell\, Z^T, \tag{4.2.7}$$

where

$$D_\ell = \begin{pmatrix} \frac{\sin(\ell\theta_1)}{\sin(\theta_1)} & & & \\ & \frac{\sin(\ell\theta_2)}{\sin(\theta_2)} & & \\ & & \ddots & \\ & & & \frac{\sin(\ell\theta_N))}{\sin(\theta_N)} \end{pmatrix}. \tag{4.2.8}$$

4.2.1 Residual with General Right-hand Sides

Given a tridiagonal Toeplitz with a general right-hand side, we have

Theorem 4.2.1. *For $Ax = b$, where A is tridiagonal Toeplitz as in (3.2.1) with nonzero (real or complex) parameters ν, λ, and μ. Then given $1 \le k \le N$, the kth GMRES residual r_k satisfies*

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \sqrt{k+1}\psi(k,\zeta)\left[\frac{1}{2} + \tilde{\Phi}_{k+1}(\tau,\xi)\right]^{-1/2}, \tag{4.2.9}$$

where

$$\psi(k,\zeta) = \frac{1-|\zeta|^{2\min\left\{k+1,\lfloor\frac{N+1}{2}\rfloor\right\}}}{1-|\zeta|^2}, \quad \zeta = \min\left\{|\xi|, \frac{1}{|\xi|}\right\}. \tag{4.2.10}$$

Specially, for $|\xi| \le 1$,

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \sqrt{k+1}\psi(k,\zeta)\left[\frac{1}{2} + \tilde{\Phi}^{(+)}_{k+1}(\tau,\xi)\right]^{-1/2}, \tag{4.2.11}$$

and for $|\xi| > 1$,

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \sqrt{k+1}\psi(k,\zeta)\left[\frac{1}{2} + \tilde{\Phi}^{(-)}_{k+1}(\tau,\xi)\right]^{-1/2}. \tag{4.2.12}$$

The proof relies on (4.2.1)

$$\|r_k\|_2 \le \min_{u_{(1)}=1} \|\Omega_{k+1}\tilde{R}^{-1}_{k+1}u\|_2 \;\; \|\Omega\tilde{M}_{(:,1:k+1)}\Omega^{-1}_{k+1}\|_2, \tag{4.2.13}$$

and according to inequality (3.3.13), we have

$$\min_{u_{(1)}=1} \|\Omega_{k+1}\tilde{R}^{-1}_{k+1}u\|_2 = \|\Omega^{-*}_{k+1}\tilde{R}^{*}_{k+1}e_1\|_2^{-1} = \left[\frac{1}{2} + \tilde{\Phi}^{(+)}_{k+1}(\tau,\xi)\right]^{-1/2}. \tag{4.2.14}$$

Now we need to analyze $\Omega\tilde{M}_{(:,1:k+1)}\Omega^{-1}_{k+1}$ to finish the proof.

According to (4.2.7), we can calculate $\tilde{M}_\ell$, and find some properties of $b_\ell\xi^{\ell-1}\Omega\tilde{M}_\ell\Omega^{-1}$, which is a part of $\Omega\tilde{M}_\ell\Omega^{-1} = \sum_{\ell=1}^{N} b_\ell\xi^{\ell-1}\;\Omega\tilde{M}_\ell\Omega^{-1}$.

It turns out that $\tilde{M}_\ell$'s have the following forms, whose theoretical justifications are very complicated and thus are postponed to Section 4.4.

$$
\begin{aligned}
\tilde{M}_1 &= \begin{pmatrix}
1 & & & & & \\
 & 1 & & & & \\
 & & 1 & & & \\
 & & & 1 & & \\
 & & & & \ddots & \\
 & & & & & 1
\end{pmatrix}, \\
\tilde{M}_2 &= \begin{pmatrix}
 & 1 & & & & \\
1 & & 1 & & & \\
 & 1 & & 1 & & \\
 & & 1 & & \ddots & \\
 & & & \ddots & & 1 \\
 & & & & 1 &
\end{pmatrix}, \\
\tilde{M}_3 &= \begin{pmatrix}
 & & 1 & & & \\
 & 1 & & 1 & & \\
1 & & 1 & & \ddots & \\
 & 1 & & \ddots & & 1 \\
 & & \ddots & & 1 & \\
 & & & 1 & &
\end{pmatrix}, \\
\vdots \quad & \quad \vdots \\
\tilde{M}_{N-1} &= \begin{pmatrix}
 & & & & & 1 & \\
 & & & & 1 & & 1 \\
 & & & \cdot & & 1 & \\
 & & \cdot & & \cdot & & \\
 & \cdot & & \cdot & & & \\
1 & & \cdot & & & & \\
 & 1 & & & & &
\end{pmatrix}, \\
\tilde{M}_N &= \begin{pmatrix}
 & & & & & & 1 \\
 & & & & & 1 & \\
 & & & & \cdot & & \\
 & & & \cdot & & & \\
 & & \cdot & & & & \\
 & 1 & & & & & \\
1 & & & & & &
\end{pmatrix}.
\end{aligned}
\qquad (4.2.15)
$$

Now for $b_{(\ell)}\xi^{\ell}\Omega\tilde{M}_{\ell}\Omega^{-1}$, we have the following result

$$\begin{aligned}
b_{(1)}\Omega\tilde{M}_1\Omega^{-1} &= b_{(1)}\begin{pmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}, \\
b_{(2)}\xi^{1}\Omega\tilde{M}_2\Omega^{-1} &= b_{(2)}\begin{pmatrix} & \xi^2 & & & & \\ 1 & & \xi^2 & & & \\ & 1 & & \xi^2 & & \\ & & 1 & & \ddots & \\ & & & \ddots & & \xi^2 \\ & & & & 1 & \end{pmatrix}, \\
b_{(3)}\xi^{2}\Omega\tilde{M}_3\Omega^{-1} &= b_{(3)}\begin{pmatrix} & & \xi^4 & & & \\ & \xi^2 & & \xi^4 & & \\ 1 & & \xi^2 & & \ddots & \\ & 1 & & \ddots & & \xi^4 \\ & & \ddots & & \xi^2 & \\ & & & 1 & & \end{pmatrix}, \\
&\vdots \qquad \vdots \\
b_{(N-1)}\xi^{N-2}\Omega\tilde{M}_{N-1}\Omega^{-1} &= b_{(N-1)}\begin{pmatrix} & & & & & \xi^{2N-4} & \\ & & & & \xi^{2N-6} & & \xi^{2N-4} \\ & & & . & & \xi^{2N-6} & \\ & & . & & . & & \\ & . & & . & & & \\ 1 & & . & & & & \\ & 1 & & & & & \end{pmatrix}, \\
b_{(N)}\xi^{N-1}\Omega\tilde{M}_N\Omega^{-1} &= b_{(N)}\begin{pmatrix} & & & & & & \xi^{2N-2} \\ & & & & & \xi^{2N-4} & \\ & & & & . & & \\ & & & . & & & \\ & & . & & & & \\ & \xi^2 & & & & & \\ 1 & & & & & & \end{pmatrix},
\end{aligned} \tag{4.2.16}$$

and $\Omega\tilde{M}\Omega^{-1} = \sum_{\ell=1}^{N} b_{(\ell)}\xi^{\ell-1}\Omega\tilde{M}_{\ell}\Omega^{-1}$.

As the result of (4.2.16), if $k+1 \le \lfloor \frac{N+1}{2} \rfloor$, the $(k+1)$th column of $\Omega\tilde{M}\Omega^{-1}$ can be expressed as the sum of $(k+1)$ vectors $\{v_1, \xi^2 v_2, \ldots, \xi^{2k} v_{k+1}\}$, where $\|v_i\|_2 \le \|b\|_2$ for $i = 1, \ldots, k+1$. Then we have $\|(\Omega\tilde{M}\Omega^{-1})_{(:,k+1)}\|_2 \le \|b\|_2(1 + |\xi|^2 + \ldots + |\xi|^{2k})$. For

$|\xi| \leq 1$, if $j \geq \lfloor \frac{N+1}{2} \rfloor$, we still have [1]

$$\|(\Omega \tilde{M} \Omega^{-1})_{(:,j)}\|_2 \leq \|b\|_2 (1 + |\xi|^2 + \ldots + |\xi|^{2\lfloor \frac{N-1}{2} \rfloor}).$$

Therefore, when $|\xi| \leq 1$,

$$\|\Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2 \leq \sqrt{\sum_{j=1}^{k+1} \|(\Omega \tilde{M} \Omega^{-1})_{(:,j)}\|_2^2} \leq \sqrt{k+1}\ \|b\|_2 \psi(k,\zeta), \tag{4.2.17}$$

where $\psi(k,\zeta) = \frac{1-|\zeta|^{2\min\left\{k+1, \lfloor \frac{N+1}{2} \rfloor\right\}}}{1-|\zeta|^2}$, $\quad \zeta = \min\left\{|\xi|, \frac{1}{|\xi|}\right\}$.

Now we can prove Theorem 4.2.1.

Proof : First let us prove the second part of Theorem 4.2.1, i.e. for $|\xi| \leq 1$, the residual satisfied (4.2.11).

According to the inequality (4.2.1)

$$\|r_k\|_2 \leq \min_{u_{(1)}=1} \|\Omega_{k+1} \tilde{R}_{k+1}^{-1} u\|_2 \quad \|\Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2,$$

and the relation(3.3.13)

$$\min_{u_{(1)}=1} \|\Omega_{k+1} \tilde{R}_{k+1}^{-1} u\|_2 = \|\Omega_{k+1}^{-*} \tilde{R}_{k+1}^{*} e_1\|_2^{-1} = \left[\frac{1}{2} + \tilde{\Phi}_{k+1}^{(+)}(\tau,\xi)\right]^{-1/2}.$$

We have

$$\|r_k\|_2 \leq \left[\frac{1}{2} + \tilde{\Phi}_{k+1}^{(+)}(\tau,\xi)\right]^{-1/2} \quad \|\Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}\|_2.$$

Now consider the relation (4.2.17), then

$$\|r_k\|_2 \quad \leq \quad \left[\frac{1}{2} + \tilde{\Phi}_{k+1}^{(+)}(\tau,\xi)\right]^{-1/2} \sqrt{k+1}\ \|b\|_2\ \psi(k,\zeta). \tag{4.2.18}$$

Since $\|r_0\|_2 = \|b\|_2$, for $|\xi| \leq 1$, (4.2.11) is proved. For $|\xi| \geq 1$, just apply (4.2.11) to the permuted system (3.3.19). Theorem 4.2.1 is now proved. ∎

[1] Since $\tilde{M}$ is symmetric and also symmetric about antidiagonal, we have

$$\|(\Omega \tilde{M} \Omega^{-1})_{(:,j)}\|_2 \leq \|b\|_2 (1 + |\xi|^2 + \ldots + |\xi|^{2(N+1-j)}).$$

Note $(N+1-j) \leq \lfloor \frac{N-1}{2} \rfloor$ now. So we can write $\|(\Omega \tilde{M} \Omega^{-1})_{(:,j)}\|_2 \leq \|b\|_2 (1 + |\xi|^2 + \ldots + |\xi|^{2\lfloor \frac{N-1}{2} \rfloor})$.

4.2.2 Numerical Examples

Similar to Theorem 3.3.1, which gives

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \sqrt{k+1}\left[\frac{1}{2} + \Phi_{k+1}(\tau, \xi)\right]^{-1/2}, \tag{4.2.19}$$

Theorem 4.2.1 gives a bound with similar format,

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \sqrt{k+1}\psi(k, \zeta)\left[\frac{1}{2} + \tilde{\Phi}_{k+1}(\tau, \xi)\right]^{-1/2}.$$

We can compare these two upper bounds by evaluating the ratio

$$\frac{\left[\frac{1}{2} + \tilde{\Phi}^{(+)}_{k+1}(\tau, \xi)\right]^{-1/2}\psi(k, \zeta)}{\left[\frac{1}{2} + \Phi_{k+1}(\tau, \xi)\right]^{-1/2}}. \tag{4.2.20}$$

Numerical examples are showed in Figures 4.2.1, 4.2.2, 4.2.3. The parameters in Section 3.3.3 are used again. The upper bounds (blue solid lines with squares) are calculated according to Theorem 4.2.1, and the other upper bounds(green dashed lines) obtained by Theorem 3.3.1. These two upper bounds are very close to each other.

The ratio (4.2.20) is showed in Figure 4.2.4. The curve with circles is the ratio when $\mu > 0$; the lower plot with triangles presents the ratio when $\mu < 0$, i.e. τ is pure imaginary. The upper bounds, through applying Chebyshev polynomials of the second kind, have better performance when $\mu > 0$. The upper bounds, through applying Chebyshev polynomials of the first kind, are better when $\mu < 0$. Note in Figure 4.2.4, the bottom two graphs only contain one curve for the case $\mu < 0$; for the other case $\mu > 0$, since $\tau = 1$, the denominator in the Chebyshev polynomial of the second kind is zero.

4.3 Exact Residual for Special Right-hand Sides: $b = e_1$, $b = e_N$

Now let us consider some cases for special right-hand side. First, given the right hand side: $b = e_1$, we can calculate the exact residual through the analysis using Chebyshev polynomials of the second kind.

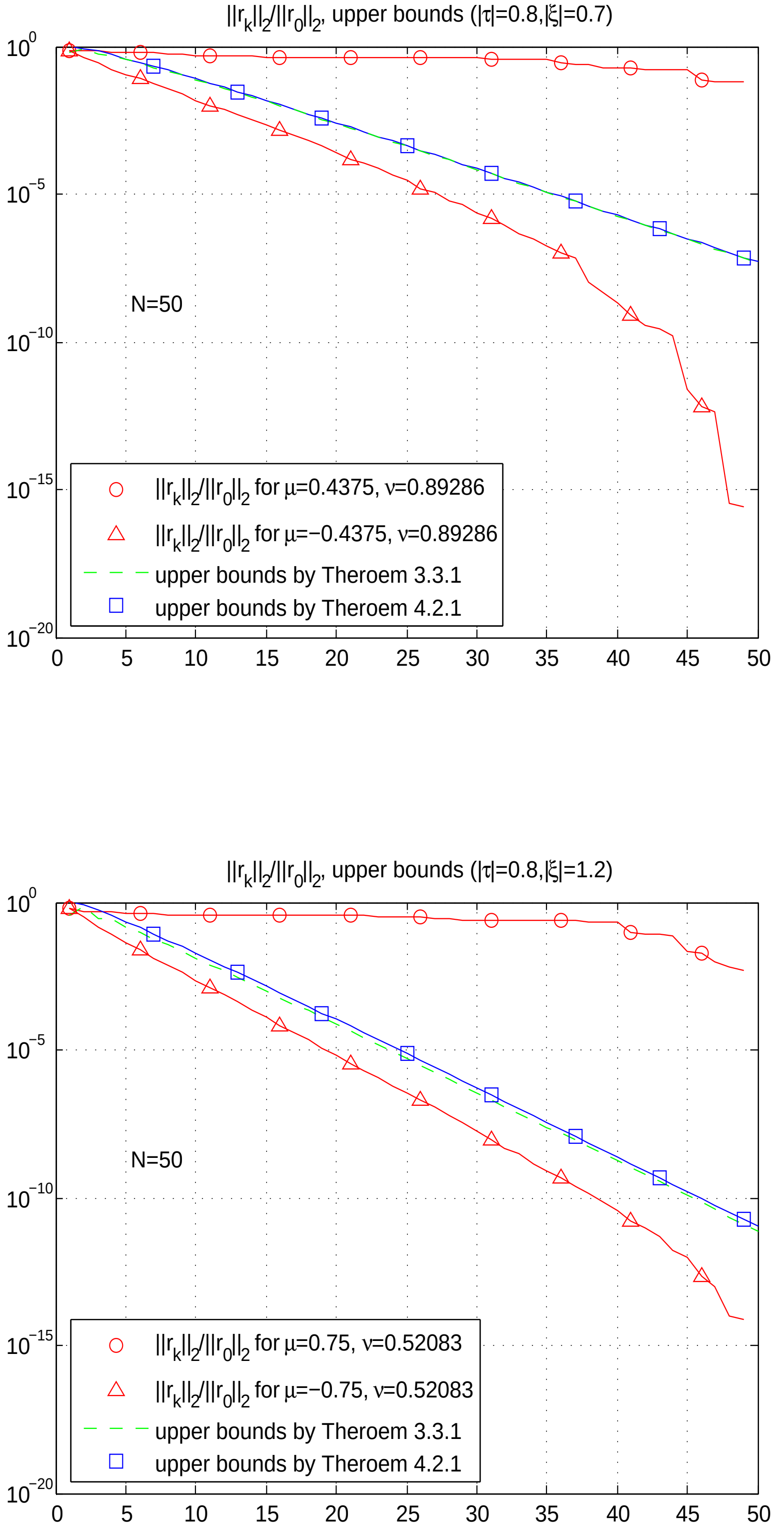

Figure 4.2.1: GMRES residuals for random b with $|\tau| = 0.8$, and the upper bounds by Theorem 3.3.1 and Theorem 4.2.1.

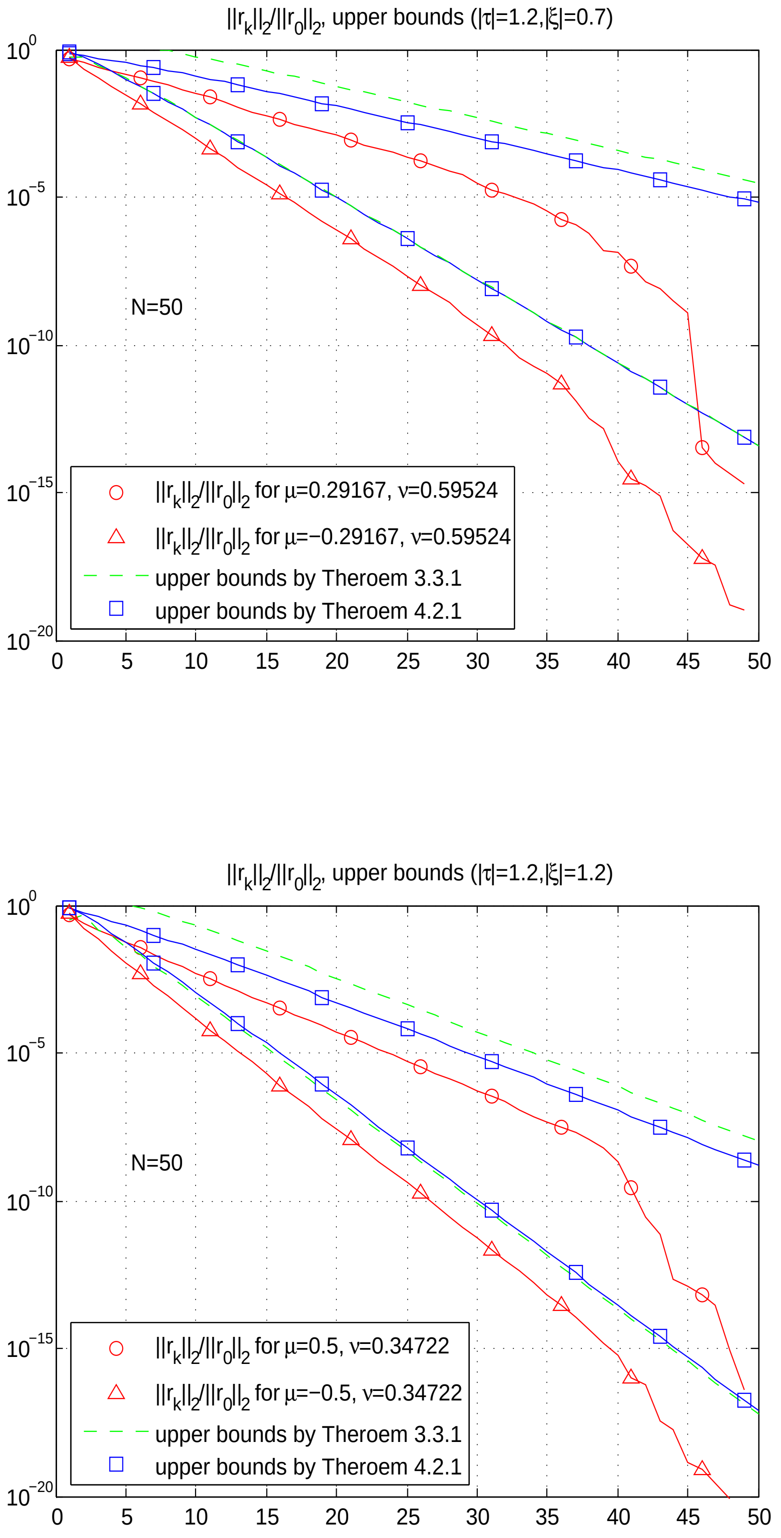

Figure 4.2.2: GMRES residuals for random b with $|\tau| = 1.2$, and the upper bounds by Theorem 3.3.1 and Theorem 4.2.1.

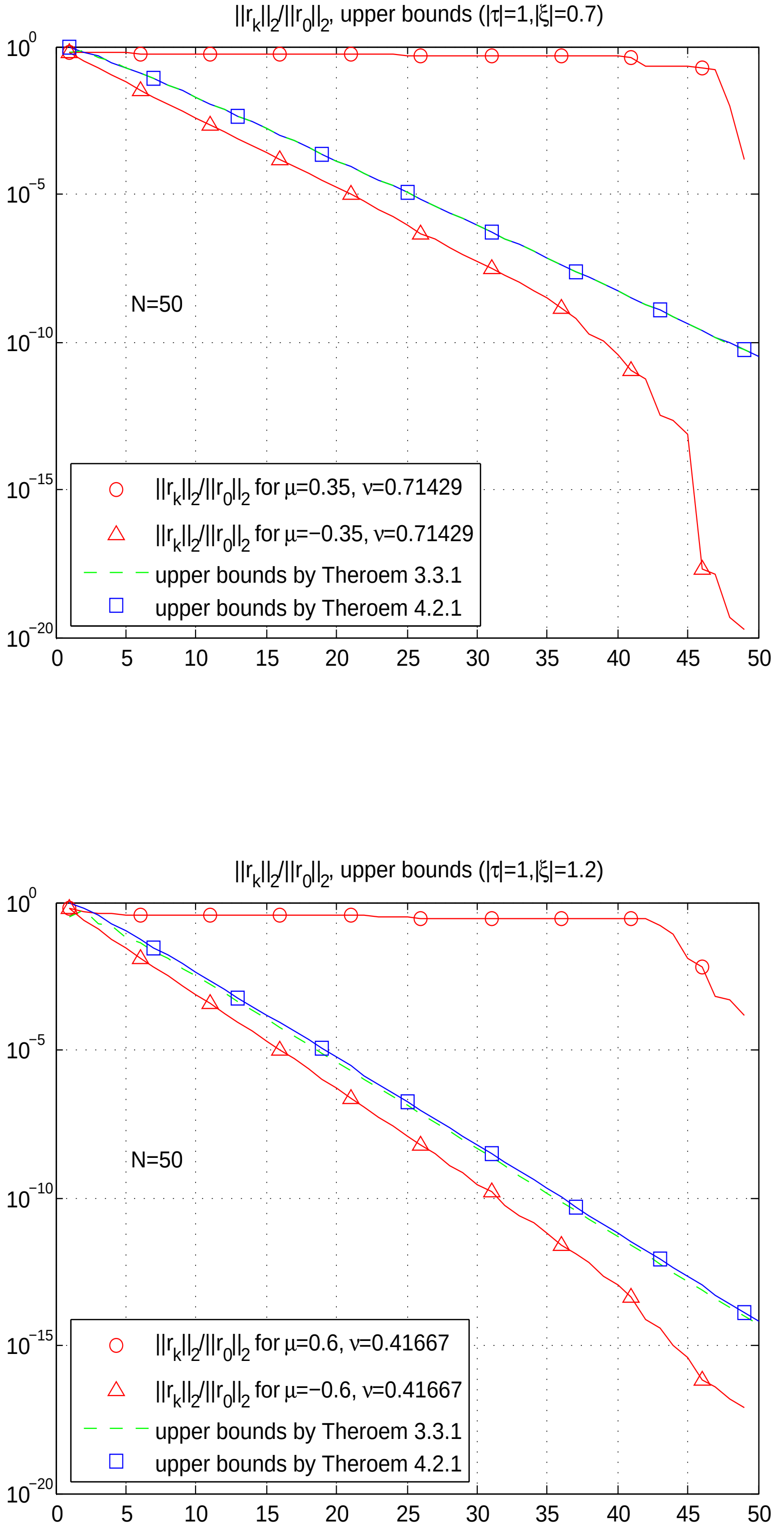

Figure 4.2.3: GMRES residuals for random b with $|\tau| = 1$, and the upper bounds by Theorem 3.3.1 and Theorem 4.2.1.

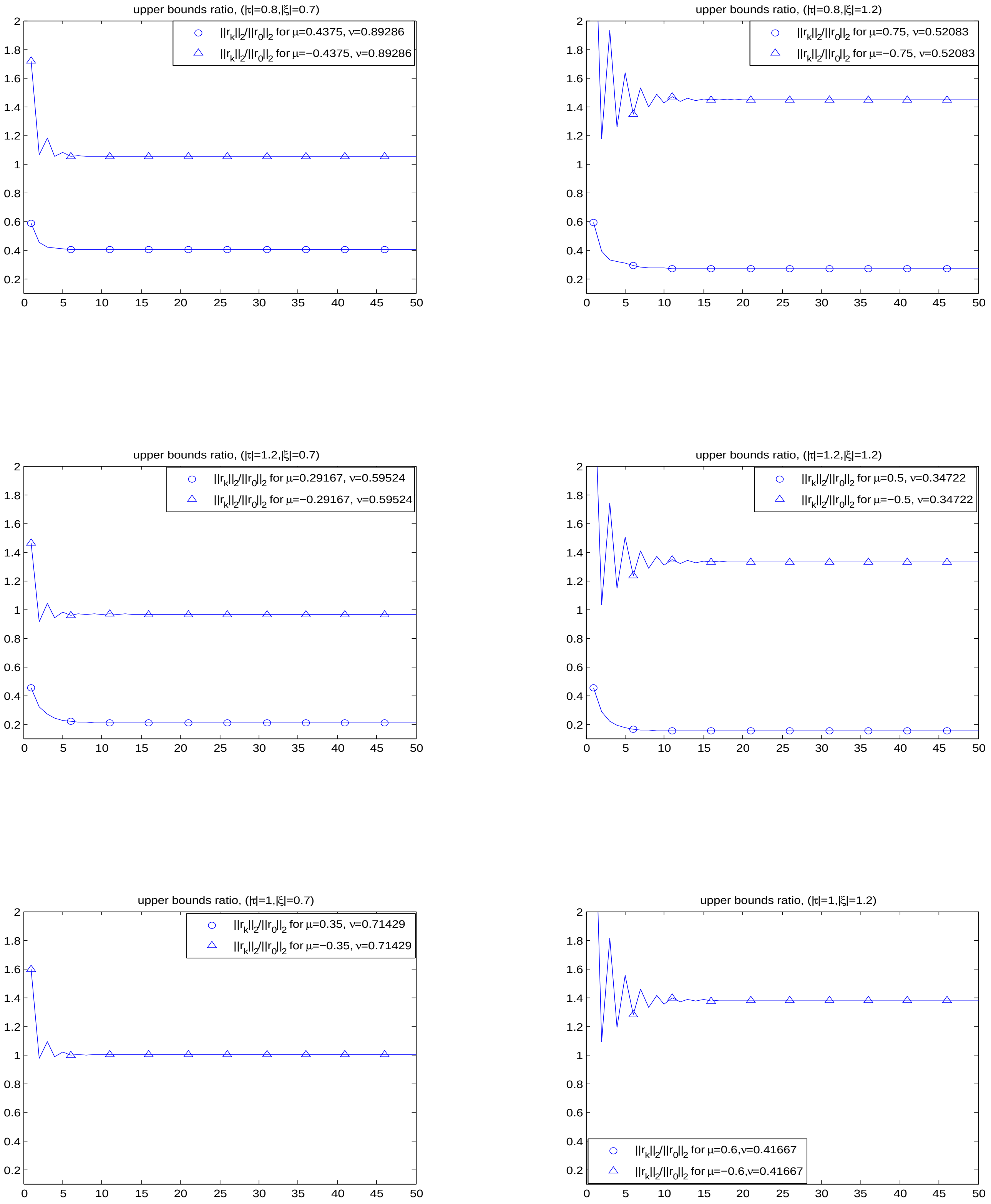

Figure 4.2.4: Upper bound ratios between the bounds obtained from Theorem 4.2.1 and those from Theorem 3.3.1.

The kth GMRES residual is

$$\|r_k\|_2 = \min_{u_{(1)}=1} \|Y\, V_{k+1,N}^T u\|_2,$$

and the equation (4.1.9) gives

$$Y\, V_{k+1,N}^T = \Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}\, \Omega_{k+1} \tilde{R}_{k+1}^{-1}, \tag{4.3.1}$$

then

$$\|r_k\|_2 = \min_{u_{(1)}=1} \|\Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}\, \Omega_{k+1} \tilde{R}_{k+1}^{-1}\|_2. \tag{4.3.2}$$

Recall (4.2.6)

$$\tilde{M} = \sum_{\ell=1}^{N} b_{(\ell)} \xi^{\ell-1} \tilde{M}_\ell,$$

Now, if $b = e_1$, i.e. only $b_{(1)} = 1$, then

$$\begin{aligned} \tilde{M} &= \tilde{M}_1 \\ &= Z\, D_1\, Z^T \\ &= I_N, \end{aligned}$$

since (4.2.7) gives $\tilde{M}_1 = Z\, D_1\, Z^T$, and $D_1 = I_N$. Hence, $\Omega \tilde{M} \Omega^{-1} = I_N$, and

$$\begin{aligned} \Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1} &= \left(\Omega \tilde{M} \Omega^{-1}\right)_{(:,1:k+1)} \\ &= (I_N)_{(:,1:k+1)}\,. \end{aligned} \tag{4.3.3}$$

Recall the relation (3.3.13), the kth GMRES residual is

$$\begin{aligned}
\|r_k\|_2 &= \min_{u_{(1)}=1} \|Y\, V_{k+1,N}^T u\|_2 \\
&= \min_{u_{(1)}=1} \|\Omega \tilde{M}_{(:,1:k+1)} \Omega_{k+1}^{-1}\, \Omega_{k+1} \tilde{R}_{k+1}^{-1} u\|_2 \\
&= \min_{u_{(1)}=1} \|\,(I_N)_{(:,1:k+1)}\ \Omega_{k+1} \tilde{R}_{k+1}^{-1} u\|_2 \\
&= \min_{u_{(1)}=1} \|\Omega_{k+1} \tilde{R}_{k+1}^{-1} u\|_2 \\
&= \left[\frac{1}{2} + \tilde{\Phi}_{k+1}^{(+)}(\tau,\xi)\right]^{-1/2}. \qquad (4.3.4)
\end{aligned}$$

The above calculation leads to Theorem 4.3.1.

Theorem 4.3.1. *If $b = e_1$, then the kth GMRES residual r_k satisfies for $1 \le k < N$*

$$\frac{\|r_k\|_2}{\|r_0\|_2} = \left[\frac{1}{2} + \tilde{\Phi}_{k+1}^{(+)}(\tau,\xi)\right]^{-1/2}. \qquad (4.3.5)$$

Note that $b = e_1$ and $\|r_0\|_2 = \|b\|_2 = 1$.

Now apply Theorem 4.3.1 to the permuted system (3.3.19) to get Theorem 4.3.2.

Theorem 4.3.2. *If $b = e_N$, then the kth GMRES residual r_k satisfies for $1 \le k < N$*

$$\frac{\|r_k\|_2}{\|r_0\|_2} = \left[\frac{1}{2} + \tilde{\Phi}_{k+1}^{(-)}(\tau,\xi)\right]^{-1/2}. \qquad (4.3.6)$$

4.4 The Structure of $\tilde{M}_\ell$

According to (4.2.7), we have

Proposition 4.4.1. *Let $\tilde{M}_\ell \overset{\text{def}}{=} Z\ \mathrm{diag}(Z_{(:,\ell)})\mathbf{U}_N^T$ for $1 \le \ell \le N$. Then*

$$\tilde{M}_\ell(i,j) = \frac{2}{N+1} \sum_{k=1}^{N} \sin(i\theta_k)\sin(j\theta_k)\frac{\sin(\ell\theta_k)}{\sin(\theta_k)}, \qquad (4.4.7)$$

and $\tilde{M}_\ell$ is symmetric, i.e. $\tilde{M}_\ell(i,j) = \tilde{M}_\ell(j,i)$, and is also symmetric about the anti-diagonal, i.e.

$$\tilde{M}_\ell(i,j) = \tilde{M}_\ell(N+1-j, N+1-i).$$

Furthermore,

$$\tilde{M}_\ell(i,j) = \tilde{M}_{N+1-\ell}(i, N+1-j),$$

i.e., $\tilde{M}_{N+1-\ell}$ is $\tilde{M}_\ell$ after its columns reordered backwards from the last to the first.

Proof: The relation (4.4.7) follows the computation of $\tilde{M}_\ell$. According to (4.4.7), we have

$$\begin{aligned}
\tilde{M}_\ell(i,j) &= \frac{2}{N+1}\sum_{k=1}^{N} \sin(i\,\theta_k)\sin(j\,\theta_k)\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \frac{2}{N+1}\sum_{k=1}^{N} \sin(j\,\theta_k)\sin(i\,\theta_k)\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \tilde{M}_\ell(j,i),
\end{aligned}$$

i.e. $\tilde{M}_\ell$ is symmetric.

To prove $\tilde{M}_\ell$ is symmetric about the anti-diagonal, we need to show

$$\tilde{M}_\ell(i,j) = \tilde{M}_\ell(N+1-j, N+1-i).$$

According to (4.4.7), we have

$$\begin{aligned}
\tilde{M}_\ell(N+1-j, N+1-i) &= \frac{2}{N+1}\sum_{k=1}^{N} \sin((N+1-j)\,\theta_k)\sin((N+1-i)\,\theta_k)\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \frac{2}{N+1}\sum_{k=1}^{N} \sin(k\pi - j\,\theta_k)\sin(k\pi - i\,\theta_k)\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \frac{1}{N+1}\sum_{k=1}^{N} \left[\cos((i-j)\,\theta_k) - \cos(2k\pi - (i+j)\,\theta_k)\right]\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \frac{1}{N+1}\sum_{k=1}^{N} \left[\cos((i-j)\,\theta_k) - \cos((i+j)\,\theta_k)\right]\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \frac{2}{N+1}\sum_{k=1}^{N} \sin(i\,\theta_k)\sin(j\,\theta_k)\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \tilde{M}_\ell(i,j).
\end{aligned}$$

For the last part, we have

$$\begin{aligned}
\tilde{M}_{N+1-\ell}(i, N+1-j) &= \frac{2}{N+1}\sum_{k=1}^{N} \sin(i\,\theta_k)\sin((N+1-j)\,\theta_k)\frac{\sin((N+1-\ell)\theta_k)}{\sin(\theta_k)} \\
&= \frac{2}{N+1}\sum_{k=1}^{N} \sin(i\theta_k)\sin(k\pi - j\,\theta_k)\frac{\sin(k\pi - \ell\,\theta_k)}{\sin(\theta_k)} \\
&= \frac{1}{N+1}\sum_{k=1}^{N} [\cos((\ell-j)\,\theta_k) - \cos(2k\pi - (\ell+j)\,\theta_k)]\,\frac{\sin(i\,\theta_k)}{\sin(\theta_k)} \\
&= \frac{1}{N+1}\sum_{k=1}^{N} [\cos((\ell-j)\,\theta_k) - \cos((\ell+j)\,\theta_k)]\,\frac{\sin(i\,\theta_k)}{\sin(\theta_k)} \\
&= \frac{2}{N+1}\sum_{k=1}^{N} \sin(i\,\theta_k)\sin(j\,\theta_k)\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \tilde{M}_\ell(i,j).
\end{aligned}$$

■

According to Proposition 4.4.1, to compute $\tilde{M}_\ell$'s, it suffices to compute the first $\frac{N+1}{2}$ columns of $\tilde{M}_\ell$, for $\ell \le \frac{N+1}{2}$. This is done in Proposition 4.4.2 below.

Proposition 4.4.2. *Let* $\tilde{M}_\ell \stackrel{\text{def}}{=} Z\,\operatorname{diag}(Z_{(:,\ell)})\mathbf{U}_N^T$. *Then, for* $j \le \frac{N+1}{2}$ *and* $\ell \le \frac{N+1}{2}$, *the jth column of* $\tilde{M}_\ell$ *has at most* $\min\{j,\ell\}$ *nonzero entries. If* $j \le \ell$,

$$\begin{aligned}
\tilde{M}_\ell(\ell+j-1, j) &= 1, \\
\tilde{M}_\ell(\ell+j-3, j) &= 1, \\
\tilde{M}_\ell(\ell+j-5, j) &= 1, \\
&\vdots \\
\tilde{M}_\ell(\ell+j-(2j-1), j) &= \tilde{M}_\ell(\ell-j+1, j) = 1. \qquad (4.4.8)
\end{aligned}$$

If $\ell \leq j$,

$$\begin{aligned}
\tilde{M}_\ell(j+\ell-1,j) &= 1, \\
\tilde{M}_\ell(j+\ell-3,j) &= 1, \\
\tilde{M}_\ell(j+\ell-5,j) &= 1, \\
&\vdots \\
\tilde{M}_\ell(j+\ell-(2\ell-1),j) &= \tilde{M}_\ell(j-\ell+1,j) = 1. \qquad (4.4.9)
\end{aligned}$$

Proof: For $j \leq \ell$, the proposition can be proved by induction. When $j = 1$, the first column of $\tilde{M}_\ell$ can be calculated as

$$\begin{aligned}
\tilde{M}_\ell(i,1) &= \frac{2}{N+1}\sum_{k=1}^{N} \sin(i\theta_k)\sin(1\,\theta_k)\frac{\sin(\ell\theta_k)}{\sin(\theta_k)} \\
&= \frac{2}{N+1}\sum_{k=1}^{N} \sin(i\theta_k)\sin(\ell\theta_k).
\end{aligned}$$

Hence, $\tilde{M}_\ell(i,1) = 1$ only if $i = \ell$; otherwise, $\tilde{M}_\ell(i,1) = 0$. In other words, the nonzero entry of the first column is:

$$\tilde{M}_\ell(\ell+j-1,j) = \tilde{M}_\ell(\ell-j+1,j) = \tilde{M}_\ell(\ell,1) = 1.$$

Assume Proposition 4.4.2 is right for the jth column of $\tilde{M}_\ell$'s for $j \leq \ell$. We need to prove it is correct for the $(j+1)$th column of $\tilde{M}_\ell$'s for $j+1 \leq \ell$.

According to (4.4.7), we have

$$\begin{aligned}
\tilde{M}_\ell(i,j) &= \frac{2}{N+1}\sum_{k=1}^{N} \sin(i\,\theta_k)\sin(\ell\theta_k)\frac{\sin((j)\,\theta_k)}{\sin(\theta_k)} \\
&= \frac{1}{N+1}\sum_{k=1}^{N} \left(\cos((i-\ell)\,\theta_k) - \cos((i+\ell)\,\theta_k)\right)\frac{\sin(j\,\theta_k)}{\sin(\theta_k)}. \qquad (4.4.10)
\end{aligned}$$

Let's calculate $\tilde{M}_\ell(i, j+1)$ as

$$\begin{aligned}
\tilde{M}_\ell(i,j+1) &= \frac{2}{N+1}\sum_{k=1}^{N}\sin(i\,\theta_k)\sin(\ell\theta_k)\frac{\sin((j+1)\,\theta_k)}{\sin(\theta_k)} \\
&= \frac{1}{N+1}\sum_{k=1}^{N}\left(\cos((i-\ell)\,\theta_k)-\cos((i+\ell)\,\theta_k)\right)\frac{\sin((j+1)\,\theta_k)}{\sin(\theta_k)} \\
&= \frac{1}{N+1}\sum_{k=1}^{N}\left(\cos((i-\ell)\,\theta_k)-\cos((i+\ell)\,\theta_k)\right)\left(\cos(j\,\theta_k)+\cos(\theta_k)\frac{\sin(j\,\theta_k)}{\sin(\theta_k)}\right) \\
&= \Delta_1+\Delta_2,
\end{aligned}$$

where

$$\begin{aligned}
\Delta_1 &= \frac{1}{N+1}\sum_{k=1}^{N}\left(\cos((i-\ell)\,\theta_k)-\cos((i+\ell)\,\theta_k)\right)\cos(j\,\theta_k) \\
&= \frac{1/2}{N+1}\sum_{k=1}^{N}\left[\cos((i-\ell-j)\,\theta_k)+\cos((i-\ell+j)\,\theta_k)\right] \\
&\qquad -\left[\cos((i+\ell-j)\,\theta_k)+\cos((i+\ell+j)\,\theta_k)\right] \\
&= \frac{1/2}{N+1}\sum_{k=1}^{N}\left[\cos((i-\ell-j)\,\theta_k)-\cos((i+\ell+j)\,\theta_k)\right] \\
&\qquad +\left[\cos((i-\ell+j)\,\theta_k)-\cos((i+\ell-j)\,\theta_k)\right], \qquad (4.4.11)
\end{aligned}$$

and

$$\begin{aligned}
\Delta_2 &= \frac{1}{N+1}\sum_{k=1}^{N}\left(\cos((i-\ell)\,\theta_k)-\cos((i+\ell)\,\theta_k)\right)\cos(\theta_k)\frac{\sin(j\,\theta_k)}{\sin(\theta_k)} \\
&= \frac{1/2}{N+1}\sum_{k=1}^{N}\left[\cos((i-\ell-1)\,\theta_k)+\cos((i-\ell+1)\,\theta_k)\right]\frac{\sin(j\,\theta_k)}{\sin(\theta_k)} \\
&\qquad -\left[\cos((i+\ell-1)\,\theta_k)+\cos((i+\ell+1)\,\theta_k)\right]\frac{\sin(j\,\theta_k)}{\sin(\theta_k)} \\
&= \frac{1/2}{N+1}\sum_{k=1}^{N}\left[\cos((i-\ell-1)\,\theta_k)-\cos((i+\ell+1)\,\theta_k)\right]\frac{\sin(j\,\theta_k)}{\sin(\theta_k)} \\
&\qquad +\left[\cos((i-\ell+1)\,\theta_k)-\cos((i+\ell-1)\,\theta_k)\right]\frac{\sin(j\,\theta_k)}{\sin(\theta_k)}. \qquad (4.4.12)
\end{aligned}$$

According to Lemma 3.3.2, if $i = \ell + j$, i.e. $i = \ell + (j+1) - 1$, then

$$\sum_{k=1}^{N} \cos\left[(i - \ell - j)\theta_k\right] = N,$$
$$\sum_{k=1}^{N} -\cos\left[(i + \ell + j)\theta_k\right] = 1,$$
$$\sum_{k=1}^{N} \cos((i - \ell + j)\,\theta_k) - \cos((i + \ell - j)\,\theta_k) = 0.$$

Hence,

$$\Delta_1 = 1/2. \tag{4.4.13}$$

If $i = \ell - j$, i.e. $i = \ell - (j+1) + 1$, then

$$\sum_{k=1}^{N} \cos\left[(i - \ell + j)\theta_k\right] = N,$$
$$\sum_{k=1}^{N} -\cos\left[(i + \ell - j)\theta_k\right] = 1,$$
$$\sum_{k=1}^{N} \cos((i - \ell - j)\,\theta_k) - \cos((i + \ell + j)\,\theta_k) = 0.$$

So we also have

$$\Delta_1 = 1/2. \tag{4.4.14}$$

For other cases, we always have

$$\sum_{k=1}^{N} \cos((i - \ell - j)\,\theta_k) - \cos((i + \ell + j)\,\theta_k) = 0,$$
$$\sum_{k=1}^{N} \cos((i - \ell + j)\,\theta_k) - \cos((i + \ell - j)\,\theta_k) = 0.$$

It is because that $(i - \ell - j)$ and $(i + \ell + j)$ are even or odd at the same time, so are $(i - \ell + j)$ and $(i + \ell - j)$; and both $(i + \ell + j)$ and $(i + \ell - j)$ are less than $2(N+1)$ since $i \leq N, \ell \leq \frac{N+1}{2}$, and $j \leq \frac{N+1}{2}$. So the sum must be zero, i.e. $\Delta_1 = 0$.

According to the assumption and (4.4.10),

$$\Delta_2 = 1, \tag{4.4.15}$$

for

$$\begin{aligned} i &= \ell + (j+1) - 3, \\ &\vdots \\ i &= \ell - (j+1) + 3, \end{aligned}$$

and

$$\Delta_2 = 1/2, \tag{4.4.16}$$

for $i = \ell + (j+1) - 1$ or $i = \ell - (j+1) + 1$, otherwise Δ_2 is zero.

Hence for $j + 1 \leq \ell$, $(j+1)$th column of $\tilde{M}_\ell$ has $(j+1)$ nonzero entries, which are equal to 1.

So for $j \leq \ell$, the proposition is proved. For $\ell \leq j$, it can be proved by similarly repeating the above steps again. Proposition 4.4.2 is proved. ∎

According to Propostions 4.4.1 and 4.4.2, we have $\tilde{M}_\ell$ as in (4.2.15).

Chapter 5

Conclusions

There are a few GMRES error bounds with simplicity comparable to the well-known bound for the conjugate gradient method [4, 18, 29, 38]. In [7, Section 6], Eiermann and Ernst proved

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \left[1 - \gamma(A)\,\gamma(A^{-1})\right]^{k/2}, \tag{5.0.1}$$

where $\gamma(A) = \inf\{|z^*Az| \,:\, \|z\|_2 = 1\}$ is the distance from the origin to A's field of values. When A's Hermitian part, $H = (A + A^*)/2$, is positive definite, it yields a bound by Elman [9] (see also [8])

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le \left[1 - \left(\frac{1}{\|H^{-1}\|_2\|A\|_2}\right)^2\right]. \tag{5.0.2}$$

As observed in [3], this bound of Elman can be easily extended to cover the case when only $\gamma(A) > 0$

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le (\sin\theta)^k, \quad \theta = \arccos\frac{\gamma(A)}{\|A\|_2}. \tag{5.0.3}$$

Recently Beckermann, Goreinov, and Tyrtyshnikov [3] improved (5.0.3) to

$$\frac{\|r_k\|_2}{\|r_0\|_2} \le (2 + 2/\sqrt{3})(2+\delta)\delta^k, \quad \delta = 2\sin\frac{\theta}{4 - 2\theta/\pi}. \tag{5.0.4}$$

All three bounds (5.0.1), (5.0.3), and (5.0.4) yield meaningful estimates only when $\gamma(A) > 0$, i.e. A's field of values does not contain the origin.

However, in general, there is not much concrete quantitative results for the convergence rate of GMRES, based on limited information on A and/or b. In part, it is a very difficult problem, and such a result most likely does not exist, thanks to the negative result of Greenbaum, Pták, and Strakoš [19] which says that "*Any Nonincreasing Convergence Curve is Possible for GMRES.*" A commonly used approach, as a step towards getting a feel of how fast GMRES may be, is through assuming that A is diagonalizable to arrive at (3.2.13):

$$\|r_k\|_2/\|r_0\|_2 \leq \kappa(X) \min_{p_k(0)=1} \max_i |p_k(\lambda_i)|, \tag{5.0.5}$$

and then putting aside the effect of $\kappa(X)$ to study only the effect in the factor of the associated minimization problem. This approach does not always yield satisfactory results, especially when $\kappa(X) \gg 1$ which occurs when A is highly nonnormal. Getting a fairly accurate quantitative estimate for the convergence rate of GMRES for a highly nonnormal case is likely to be very difficult. Trefethen and Toh [37, 36] established residual bounds based on pseudospectra, which sometimes is more realistic than (5.0.5) but is very expensive to compute. In [5], Driscoll, Toh, and Trefethen provided a nice explanation on this matter.

Our analysis here on tridiagonal Toeplitz A represents one of few diagonalizable cases where one can analyze r_k directly to arrive at simple quantitative results.

Our first main contribution in this thesis is the following error bound (Theorem 3.3.1)

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \sqrt{k+1} \left[\sum_{j=0}^{k} \zeta^{2j} \, |T_j(\tau)|^2\right]^{-1/2}, \tag{5.0.6}$$

where $T_j(t)$ is the jth Chebyshev polynomial of the first kind, and

$$\xi = -\frac{\sqrt{\mu\nu}}{\nu}, \quad \tau = \frac{\lambda}{2\sqrt{\mu\nu}}, \quad \zeta = \min\{|\xi|, |\xi|^{-1}\}.$$

We also prove that this upper bound is nearly achieved by $b = e_1$ (the first column of the identity matrix) when $|\xi| \leq 1$ or by $b = e_N$ (the last column of the identity matrix) when

$|\xi| \geq 1$. By "nearly achieved," we mean it is within a factor, about at most $(k+1)^{3/2}$, of the exact relative residual.

Our second main contribution is about the worst asymptotic speed of $\|r_k\|_2$ among all possible r_0. It is proven that (Theorem 3.5.1)

$$\lim_{k\to\infty}\left[\sup_{r_0}\frac{\|r_k\|_2}{\|r_0\|_2}\right]^{1/k} = \min\left\{(\zeta\rho)^{-1}, 1\right\}, \tag{5.0.7}$$

where $\rho = \max\left\{\left|\tau+\sqrt{\tau^2-1}\right|, \left|\tau-\sqrt{\tau^2-1}\right|\right\}$. As a by-product, it says the worst asymptotic speed can be separated into the factor $\zeta^{-1} \geq 1$ contributed by A's departure from normality and the factor ρ^{-1} contributed by A's eigenvalue distribution. Take, for example, $\lambda = 0.5$, $\mu = -0.3$, and $\nu = 0.7$ which was used in [5, p.562] to explain the effect of nonnormality on GMRES convergence. We have $(\zeta\rho)^{-1} = 0.90672$, whereas in [5, p.562] it is implied $\|r_k\|_2/\|r_0\|_2 \leq (0.913)^k$ for $N = 50$, which is rather good, considering that $N = 50$ is rather small.

We also estimate error bounds using Chebyshev polynomials of the second kind. Theorem 4.2.1 gives an upper bound

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \sqrt{k+1}\psi(k,\xi)\left[\frac{1}{2}+\tilde{\Phi}_{k+1}(\tau,\xi)\right]^{-1/2}, \tag{5.0.8}$$

where

$$\psi(k,\zeta) = \frac{1-|\zeta|^{2\min\left\{k+1,\lfloor\frac{N+1}{2}\rfloor\right\}}}{1-|\zeta|^2}, \quad \zeta = \min\left\{|\xi|, \frac{1}{|\xi|}\right\}, \tag{5.0.9}$$

which is comparable to the bound by Theorem 3.3.1.

Ernst [11], in our notation, obtained the following inequality: *if A's field of values does not contain the origin, then*

$$\frac{\|r_k\|_2}{\|r_0\|_2} \leq \left(|\xi|^k + |\xi|^{-k}\right)\frac{\widetilde{\rho}^k}{1-\widetilde{\rho}^{2k}}, \tag{5.0.10}$$

where $\widetilde{\rho} = \max\left\{\left|\widetilde{\tau}+\sqrt{\widetilde{\tau}^2-1}\right|, \left|\widetilde{\tau}-\sqrt{\widetilde{\tau}^2-1}\right|\right\}$ *and* $\widetilde{\tau} = \left[\cos\frac{\pi}{N+1}\right]^{-1}\tau$. Our bound (5.0.6) is comparable to Ernst's bound for large N. This can be seen by noting that for N large enough, $\widetilde{\tau} \approx \tau$ and $\widetilde{\rho} \approx \rho$, and that $T_j(\tau) \approx \frac{1}{2}\rho^j$ when $\rho > 1$ and $|\zeta|^{-k} \le |\xi|^k + |\xi|^{-k} \le 2|\zeta|^{-k}$. Ernst's bound also leads to

$$\limsup_{k\to\infty} \left[\sup_{r_0} \frac{\|r_k\|_2}{\|r_0\|_2} \right]^{1/k} \le \min\left\{(\zeta\rho)^{-1}, 1\right\}. \tag{5.0.11}$$

In differentiating our contributions here from Ernst's, we use a different technique to arrive at (5.0.6) and (5.0.7). While our proof is not as elegant as Ernst's which was based on A's field of values (see also [6]), it allows us to establish both lower and upper bounds on relative residuals for special right-hand sides to conclude that our bound is nearly achieved. Also (5.0.7) is an equality while only an inequality (5.0.11) can be deduced from Ernst's bound and approach.

We also obtain residual bounds especially for right-hand sides $b = e_1$ and $b = e_N$ (Theorems 3.4.1 and 3.4.2). They suggest, besides the sharpness of (5.0.6), an interesting GMRES convergence behavior. For $b = e_1$, that $|\xi| > 1$ speeds up GMRES convergence, and in fact $\|r_k\|_2$ is roughly proportional to $|\xi|^{-k}$. So the bigger the $|\xi|$ is, the faster the convergence will be. Note as $|\xi|$ gets bigger, A gets further away from a normal matrix. Thus, loosely speaking, the nonnormality contributes to the convergence rate in the positive way. Nonetheless this does not contradict our usual perception that high nonnormality is bad for GMRES if the worst behavior of GMRES among all b is considered. This mystery can be best explained by looking at the extreme case: $|\xi| = \infty$, i.e., $\nu = 0$, for which $b = e_1$ is an eigenvector (and convergence occurs in just one step). In general for $\nu \ne 0$, as $|\xi|$ gets bigger and bigger, roughly speaking $b = e_1$ comes closer and closer to A's invariant subspaces of lower dimensions and consequently speedier convergence is witnessed. Similar comments apply to the case when $b = e_N$.

Applying Chebyshev polynomials of the second kind enables us to obatin the exact expressions of residuals for special right-hand sides $b = e_1$ and $b = e_N$ (Theorems 4.3.1 and 4.3.2).

Bibliography

[1] G. B. Arfken, H. J. Weber, *Mathematical methods for physicists*, San Diego : Harcourt/Academic Press, 2001.

[2] W. E. Arnoldi.,*The Principle of Minimized Itereation in the Solution of the Matrix Eigenvalue Problem*, Quart. Appl. Math., 9, 1951, pp. 17-29.

[3] B. Beckermann, S. A. Goreinov, and E. E. Tyrtyshnikov, *Some remarks on the Elman estimate for GMRES*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 772–778.

[4] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[5] T. A. Driscoll, K.-C. Toh, and L. N. Trefethen, *From potential theory to matrix iterations in six steps*, SIAM Rev., 40 (1998), pp. 547–578.

[6] M. Eiermann, *Fields of values and iterative methods*, Linear Algebra Appl., 180 (1993), pp. 167–197.

[7] M. Eiermann and O. G. Ernst, *Geometric aspects in the theory of Krylov subspace methods*, Acta Numer., 10 (2001), pp. 251–312.

[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 (1983), pp. 345–357.

[9] H. C. Elman, *Iterative Methods for Large, Sparse Nonsymmetric Systems of Linear Equations*, PhD thesis, Department of Computer Science, Yale University, 1982.

[10] Mark Embree, *How Descriptive Are GMRES Convergence Bounds?*, Oxford University Computing Laboratory Numerical Analysis Report 99/08, June 1999.

[11] O. G. Ernst, *Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1079–1101.

[12] R. W. Freund and N. M. Nachtigal, *QMR: A Quasi-Minimal Residual Method for non-Hermitian Linear Systems,* Number. Math., 60, 1991, pp. 315-339.

[13] R. W. Freund and T. Szeto, *A Quasi-Minimal Residual Method for non-Hermitian Linear Systems,* Tech. Rep. 91.26, Research Institute for Advanced Computer Science, NASA Ames Research Center, Moffett Filed, CA, Dec. 1991.

[14] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative Solution of Linear Systems*, Acta Numerica 1992, Cambridge U. Press, 1992.

[15] N. Gastinel, *Analyse Numérrique Linéaire*, Hermann, Paris, 1966.

[16] G. H. Golub, C. F. Van Loan ,*Matrix Computations, 3/e* , Baltimore: Johns Hopkins University Press, 1996.

[17] I. S. Gradshteyn and I. M. Ryzhik, *Table Of Integrals, Series, and Products*, Academic Press, New York, 1980. Corrected and Enlarged Edition prepared by A. Jeffrey, incorporated the fourth edition prepared by Yu. V. Geronimus and M. Yu. Tseytlin, translated from the Russian by Scripta Technica, Inc.

[18] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[19] A. Greenbaum, V. Pták, and Z. Strakoš, *Any nonincreasing convergence curve is possible for GMRES*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 465– 469.

[20] M. H. Gutknecht, *Variants of BICGSTAB for Matrices with Complex Spectrum*, Tech. Rep. 91-14, Interdisziplinäres Projektzentrum für Supercomputing, Eidgenössische Technische Hochschule, Zürich, Aug. 1991.

[21] A. S. Householder, *Thery of Matrices in Numerical Analysis*, Blaisdell Pub. Co., Johnson, CO, 1964.

[22] I. C. F. Ipsen, *Expressions and bounds for the GMRES residual*, BIT, 40 (2000), pp. 524–535.

[23] R.-C. Li, *Sharpness in rates of convergence for CG and symmetric Lanczos methods*, Technical Report 2005-01, Department of Mathematics, University of Kentucky, 2005. Avaliable at `http://www.ms.uky.edu/∼math/MAreport/`.

[24] ——, *On Meinardus' examples for the conjugate gradient method*, Math. Comp., (2006). to appear.

[25] J. Liesen, M. Rozloznı́k, and Z. Strakoš, *Least squares residuals and minimal residual methods*, SIAM J. Sci. Comput., 23 (2002), pp. 1503–1525.

[26] J. Liesen and Z. Strakoš, *Convergence of GMRES for tridiagonal Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 233–251.

[27] ——, *GMRES convergence analysis for a convection-diffusion model problem*, SIAM J. Sci. Comput., 26 (2005), pp. 1989–2009.

[28] T. J. Rivlin ,*Chebyshev polynomials : from approximation theory to algebra and number theory* , New York : J. Wiley, 1990.

[29] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, Philadelphia, 2nd ed., 2003.

[30] ——, *Variations on Arnoldi's Method for Computing Eigenelements of Large Unsymmetric Matrices*, Lin. Alg. Appl., 34, 1980, pp. 269-295.

[31] ——, *Krylov Subspace Methods for Solving Large Unsymmetric Linear Systems*, Math. Comput., 37, 1981, pp. 105-126.

[32] Y. Saad and M. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[33] G. D. Smith, *Numerical Solution of Partial Differential Equations*, Clarendon Press, Oxford, UK, 2nd ed., 1978.

[34] G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.

[35] G. W. Stewart and J. G. Sun,Matrix Pertubation Theory, Academic Press, Boston, 1990.

[36] K.-C. Toh and L. N. Trefethen, *Calculation of Pseudospectra by the Arnoldi Iteration*, SIAM J. Sci. Comp., Vol. 17, No. 1, 1996, pp. 1-15.

[37] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis 1991:Proceedings of the 14th Dundee Conference, June, 1991, D. F. Griffiths and G. A. Watson, eds., Research Notes in Mathematics Series, Longman Press, 1992.

[38] L. N. Trefethen and D. Bau, III, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[39] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of non-symmetric linear systems*, SIAM J. Sci. Comp., 12, 1992, pp. 631-644.

[40] H. F. Walker, *Implementation of the GMRES Method Using Householder Transformations*, SIAM J. Sci. Comp., 9, 1988, pp. 152-163.

[41] I. Zavorin, D. P. O'Leary, and H. Elman, *Complete stagnation of GMRES*, Linear Algebra Appl., 367 (2003), pp. 165–183.

Vita

1. Background.
   (a) Date of Birth: January 1st, 1975
   (b) Place of Birth: Jilin, China
2. Academic Degrees.
   (a) University of Kentucky, M.S. in Mathematics, 2003.
   (b) Tsinghua University, Beijing, China, M.S. in Mechanical Engineering, 2000.
   (c) Tsinghua University, Beijing, China, B.S. in Mechanical Engineering, 1997.
3. Professional Experience.
   (a) Research Associate in Department of Mathmatics, University of Texas at Arlington, 2006-2007.
   (b) Teaching Assistant in Department of Mathematics, University of Kentucky, 2001-2006.
4. Publications.
   (a) Ren-Cang Li, Wei Zhang, The rate of convergence of GMRES on a tridiagonal Toeplitz linear system (submitted).