Preconditioning for Toeplitz-Related Systems

Sean Hon

St Hugh's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2018

To my family.

Acknowledgements

I would like to express my gratitude to my supervisor Andy Wathen for his constant encouragement, guidance, and support throughout my ups and downs over the years at Oxford. Not only did Andy introduce an intriguing research topic to me but also shaped my mathematical taste as well as my attitude towards doing research. It has been my greatest fortune during my DPhil to have the freedom that Andy has provided to nurture my ability of doing research independently.

I also thank Andrew Thompson for the useful advice he has given to me, which has greatly improved my work; and Coralia Cartis for being an examiner of my internal exams.

Among the many others who have importantly influenced me are Hadrien Montanelli, Rodrigo Mendoza Smith, and Kuan Xu whom I am thankful for their significant support and advice. The fruitful discussion with Jennifer Pestana also provided a crucial insight in relation to the development of my research.

For his deep support and inspiration, I am greatly indebted to Stefano Serra-Capizzano whom I also thank for giving me the opportunity to collaborate with, but more particularly for guiding me in the research of theory of Toeplitz sequences.

Last but not least, I am profoundly grateful to the Croucher Foundation of Hong Kong for the generous support that allows me to pursue my own research without any burden.

Abstract

This thesis concerns preconditioning for Toeplitz-related systems. Specifically, we consider functions of Toeplitz matrices, i.e. $h(T_n)\mathbf{x}_n = \mathbf{b}_n$, where $h(z)$ is an analytic function and $T_n \in \mathbb{C}^{n\times n}$ is a Toeplitz matrix.

We propose the absolute value circulant matrix $|h(C_n)|$ as a preconditioner for $h(T_n)$, where $C_n \in \mathbb{C}^{n\times n}$ is the optimal circulant preconditioner, the superoptimal circulant preconditioner, or Strang's circulant preconditioner derived from T_n, and show that $|h(C_n)|^{-1}h(T_n)$ has clustered spectra that account for the effectiveness of such preconditioner.

When $h(T_n)$ is a real matrix, we can first premultiply it by the anti-identity matrix $Y_n \in \mathbb{R}^{n\times n}$ to obtain a (real) symmetric matrix $Y_n h(T_n)$ without normalizing the original matrix. To ensure $|h(C_n)|$ is an effective preconditioner for $Y_n h(T_n)$, we show that $|h(C_n)|^{-1}Y_n h(T_n)$ has clustered spectra around ± 1. As $Y_n h(T_n)$ is symmetric yet possibly indefinite, we can use the minimal residual method for the corresponding linear system with guaranteed convergence that depends only on its eigenvalues.

We further show that the ideas of symmetrization and absolute value preconditioning for Toeplitz systems can be extended to the block Toeplitz matrix case. An application on time-stepping methods for evolutionary ordinary/partial differential equation problems is also discussed.

Numerical results are given to demonstrate the effectiveness of our proposed preconditioners.

Contents

List of Algorithms

List of Figures

List of Tables

Chapter 1

Introduction

In this thesis, we investigate the applicability of circulant preconditioners for Toeplitz-related systems, in particular, considering systems defined by functions of Toeplitz matrices, i.e. $h(T_n)\mathbf{x}_n = \mathbf{b}_n$, where $h(z)$ is an analytic function and $T_n \in \mathbb{C}^{n\times n}$ is a Toeplitz matrix. We note that $h(T_n)$ is not a Toeplitz matrix in general. However, when $h(z) = z$, the system reduces to the usual Toeplitz system, i.e. $T_n\mathbf{x}_n = \mathbf{b}_n$. In the following sections, we provide the research aims and the thesis outline.

1.1 Aims of research

Circulant preconditioners for functions of Toeplitz matrices have been recently of interest in the literature for their crucial applications. For example, the Toeplitz matrix exponential e^{T_n} arises from the discretization of integro-differential equations with a shift-invariant kernel. Solving these equations is often required in areas like option pricing [58, 137]. Related work on computing the exponential of block Toeplitz matrices arising in approximations of Markovian fluid queues can also be found in [11]. As for the matrix sine and cosine functions, an example application which arises in finite element semidiscretization of the wave equation is solving the following system of second order differential equations [80]

$$\frac{d^2}{dt^2}y + T_n^2 y = 0, \qquad y(0) = y_0, \quad y'(0) = y_0',$$

whose solution is given by

$$y(t) = \cos(tT_n)y_0 + T_n^{-1}\sin(tT_n)y_0'.$$

Jin, Zhao, and Tam [99] are the first to propose using optimal circulant preconditioners for functions of matrices. The authors provided some properties of such preconditioners and numerically demonstrated their effectiveness for certain functions of Toeplitz matrices, including e^z, $\sin z$, $\cos z$, $\ln(1+z)$, and $(1-z)^{-1}(1+z^2)$. Later, Bai, Jin, and Tao [7] proposed the use of superoptimal circulant preconditioners in the same context and illustrated their success in a series of numerical experiments.

We present in this thesis a collection of theoretic results that explain the effectiveness of circulant preconditioners for functions of Toeplitz matrices, and their corresponding numerical results that support our findings. In addition, we propose the use of absolute value circulant preconditioners for functions of Toeplitz matrices. An extension of our results to block Toeplitz systems is also provided. Each main chapter contains new contributions to numerical analysis in the context of preconditioning for structured systems.

1.2 Thesis outline

This thesis is structured as follows.

We first review in Chapter 2 the background on Toeplitz matrices before presenting our main results. Different aspects of Toeplitz matrices, including the asymptotic spectral distribution, direct and iterative solvers, and the related preconditioning techniques, are discussed. Also, we introduce a number of key concepts, such as absolute value circulant matrices and functions of matrices, and notation that are crucial for developing our results in later chapters.

In Chapter 3, we summarize several common Krylov subspace methods that we use throughout this thesis. A pseudocode of these methods and their convergence results relevant to our work on Toeplitz-related systems are provided.

Chapter 4 provides our main results on preconditioning for functions of Toeplitz matrices. In particular, we consider $h(T_n)\mathbf{x}_n = \mathbf{b}_n$, where $h(z)$ is an analytic function and T_n is the Toeplitz matrix generated by a continuous complex-valued function f defined on $[-\pi, \pi]$. We show that the absolute value circulant matrix $|h(c(T_n))|$ is an effective preconditioner for $h(T_n)$, provided that $c(T_n)$ is the optimal circulant preconditioner for T_n. Numerical tests are performed to support our results.

In Chapter 5, considering the Toeplitz matrices generated by functions in the Wiener class, we show that our results provided in Chapter 4 also apply to superoptimal circulant preconditioners and Strang's circulant preconditioners. Numerical results are provided to demonstrate the effectiveness of these circulant preconditioners.

As a simple application, we present in Chapter 6 time-stepping methods for evolutionary ordinary differential equation problems to illustrate how our results can be used. A further generalization of our work to the block Toeplitz all-at-once systems arising from time-dependent partial differential equation problems is also briefly discussed.

In Chapter 7, we comment on an extension of our work to block Toeplitz systems. We show that both block Toeplitz matrices with Toeplitz blocks and block Toeplitz matrices with commuting Hermitian blocks can be symmetrized by a simple permutation matrix. The corresponding absolute value block circulant preconditioners for these symmetrized block Toeplitz matrices are also given. Numerical results are provided to illustrate the success of our proposed preconditioners.

Chapter 8 concludes this thesis and provides a few directions for future research. We in particular present some preliminary results concerning the asymptotic spectral distribution of symmetrized Toeplitz matrices.

Chapter 2

Toeplitz matrices

Before discussing our proposed research, as Toeplitz matrices are of utmost importance to our work, we give the background on such matrices and the existing methods in the literature for solving the linear systems that involve them.

Throughout this thesis, we assume that the given *Toeplitz matrix* $T_n \in \mathbb{C}^{n\times n}$ is associated with the *spectral symbol/generating function* f via its Fourier series

$$S[f] = \sum_{k=-\infty}^{\infty} a_k e^{\mathbf{i}kx}$$

defined on $[-\pi, \pi]$. We have

$$T_n = \begin{bmatrix} a_0 & a_{-1} & \cdots & a_{-(n-2)} & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \ddots & a_{-(n-2)} \\ \vdots & a_1 & a_0 & \ddots & \vdots \\ a_{n-2} & \ddots & \ddots & \ddots & a_{-1} \\ a_{n-1} & a_{n-2} & \cdots & a_1 & a_0 \end{bmatrix},$$

where

$$a_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-\mathbf{i}kx}\, dx, \qquad k = 0, \pm 1, \pm 2, \ldots,$$

are the Fourier coefficients of f. Namely, the entries of a Toeplitz matrix are constant along its diagonals.

Several important properties of T_n associated with f are listed as follows [118, 25].

- If f is complex-valued, T_n is non-Hermitian for all n.
- If f is real-valued, T_n is Hermitian for all n.
- If f is real-valued and positive, T_n is Hermitian positive definite for all n.
- If f is real-valued and even, T_n is (real) symmetric for all n.

Such a matrix is named after Otto Toeplitz for his work on bilinear forms related to Laurent series [153] and we refer to [73] for details.

2.1 Spectral distribution of Toeplitz matrices

In this section, we discuss the asymptotic distribution of singular values and eigenvalues of Toeplitz matrices.

Let $L^p([-\pi,\pi])$, $1 \leq p < \infty$, denote the space of all (equivalence classes of) Lebesgue integrable functions f defined on $[-\pi,\pi]$ equipped with the norm

$$\|f\|_p = \left(\int_{-\pi}^{\pi} |f(x)|^p \, dx \right)^{\frac{1}{p}} < \infty$$

and let $\|f\|_\infty$ be the essential supremum norm.

Definition 2.1.1 *[172] Let $f(x) \in L^1([-\pi,\pi])$. A sequence $\{\lambda_k^{(n)}\}$ is said to be distributed as $f(x)$ if for any continuous function $F(x)$ with compact support*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} F(\lambda_k^{(n)}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(f(x))\, dx.$$

We write

$$\lambda_i^{(n)} \sim f.$$

The singular value and spectral distribution of Toeplitz matrices has been of interest over the past few decades. The earliest result on the eigenvalue distribution of Toeplitz matrices was established by Szegő [73], namely the eigenvalues of the Toeplitz matrix $T_n[f]$ generated by a real-valued $f \in L^\infty([-\pi,\pi])$ are asymptotically distributed as f. Considering the same class of functions, Avram and Parter [3, 128] showed that the singular values of $T_n[f]$ are distributed as $|f|$. Tyrtyshnikov [160, 158] later generalized the result for $T_n[f]$ generated by $f \in L^2([-\pi,\pi])$. Zamarashkin and Tyrtyshnikov [172] further weakened the requirement on f and showed that the same result holds for $f \in L^1([-\pi,\pi])$. Based on an approximation class sequence approach, Garoni, Serra-Capizzano, and Vassalos [68] recently provided the same theorem for $f \in L^1([-\pi,\pi])$ in the framework of the newly developed theory of Generalized Locally Toeplitz (GLT) sequences. The GLT theory aims to study and compute the spectral symbol of matrices arising from discretising differential operators, and we refer to [67] for its latest development. The spectral distribution of block Toeplitz matrices has also been studied for example by Tyrtyshnikov and Zamarashkin [162], and Tilli [151, 152, 115, 142].

We provide the generalized Szegő theorem in the following.

Theorem 2.1.1 *[73, 3, 128, 160, 158, 172, 151] Let $f \in L^1([-\pi,\pi])$. Let $T_n[f] \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f. Then,*

$$\sigma(T_n[f]) \sim |f|.$$

If, moreover, f is real-valued, then

$$\lambda(T_n[f]) \sim f.$$

The following simple example illustrates the point: we consider

$$T_n[f] = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

whose generating function is $f(x) = -e^{\mathbf{i}x} + 2 - e^{-\mathbf{i}x} = 2 - 2\cos x$. In Figure 2.1, we see that the eigenvalues of $T_n[f]$ are distributed as $f(x) = 2 - 2\cos x$ on $[-\pi,\pi]$.

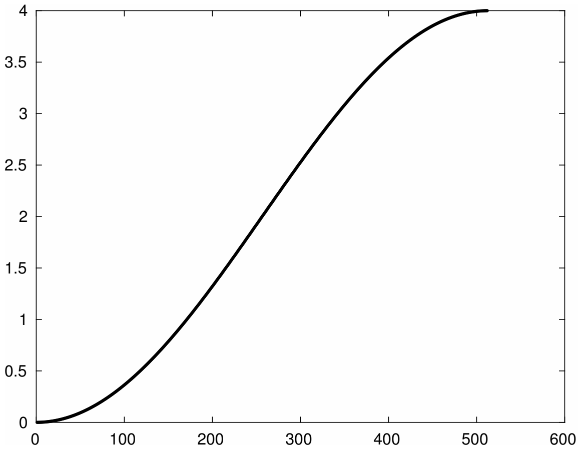


Figure 2.1: Spectral distribution of $T_n[f]$ with $f(x) = 2 - 2\cos x$ at $n = 512$.

Having knowledge about the asymptotic spectra of Toeplitz matrices is crucial to developing efficient solvers for them, such as preconditioned Krylov subspace methods which will be discussed in details in Chapter 3.

2.2 Toeplitz solvers

Toeplitz matrices are ubiquitous: they arise in many different fields of mathematics and scientific computing, including numerical ordinary/partial differential equations, image restoration, numerical solution of convolution integral equations, queueing networks, and option pricing, to name just a few. For more applications related to Toeplitz matrices, see for example [25, 118]. These applications often involve solving Toeplitz-related systems with a large dimension, so it is of high importance to develop efficient solvers that provide fast computation. Due to the wide-ranging applicability of Toeplitz systems and the computational consideration of applications, an extensive amount of work has been devoted to developing fast algorithms, known as *Toeplitz solvers*, for solving such systems.

2.2.1 Direct methods

Developing fast direct methods was primarily the focus of early research on Toeplitz solvers. Direct solvers are said to be *fast* if they are of about $\mathcal{O}(n^2)$ complexity. In 1946, Levinson [108] was the first to propose a fast direct solver, which was a significant improvement over methods such as Gaussian elimination, that require $\mathcal{O}(n^3)$ operations. Levinson-type variants [155, 173, 174] with similar complexity were later developed. Similarly, Schur-type fast solvers were proposed by Bareiss [8] and Rissanen [134]. Displacement equation-type fast solvers were also proposed for example by Heinig [77], Gohberg, Kailath, and Olshevsky [69], and Gu [74].

Superfast direct solvers, which are defined to be direct solvers that require less than $\mathcal{O}(n^2)$ costs, have then been developed since the 1980s. Various superfast Toeplitz solvers with only $\mathcal{O}(n \log^2 n)$ complexity can be found in [116, 1, 13, 16, 53]. However, Bunch indicated in [18] that the Schur solvers and the Levinson solvers are weakly stable in some cases, but both could be highly unstable in the case of indefinite and nonsymmetric Toeplitz matrices. The author identified that if T_n has a singular or ill-conditioned principal submatrix, a breakdown will occur, which could lead to numerical instabilities in these algorithms. Some methods were therefore developed to avoid breakdowns [54, 150]. In particular, T. Chan and Hansen [41, 42] proposed the look-ahead Levinson algorithm to enhance the numerical stability for solving Toeplitz systems. The basic idea of the algorithm is that it will look ahead to the next well-conditioned leading principal submatrix if a singular or a nonsingular ill-conditioned submatrix occurs. However, this look-ahead strategy increases the complexity to an overall $\mathcal{O}(n^3)$ operation, as it requires condition number estimates for all leading principal submatrices. Several extensions of look-ahead Toeplitz solvers were also proposed in [65, 66]. However, reliable look-ahead strategies are difficult to design and the resulting algorithms may not be fast. Some other superfast solvers [163, 45, 170] can also be found in the literature.

2.2.2 Iterative methods

Other than direct solvers, iterative methods can also be used to solve Toeplitz systems. In fact, for a large class of Toeplitz systems, Krylov subspace methods like the conjugate gradient method can achieve overall $\mathcal{O}(n \log n)$ complexity. Since a Toeplitz matrix-vector product is required at each iteration of a typical Krylov subspace method, it is crucial to compute this product in a fast manner in order to reduce the overall complexity.

One way to compute such Toeplitz matrix-vector multiplication is to use Fast Fourier Transforms (FFT) [51, 166]. We first embed T_n into a $2n \times 2n$ circulant matrix, namely

$$\begin{bmatrix} T_n & B_n \\ B_n & T_n \end{bmatrix} \begin{bmatrix} \mathbf{d}_n \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} T_n \mathbf{d}_n \\ B_n \mathbf{d}_n \end{bmatrix},$$

where

$$B_n = \begin{bmatrix} 0 & a_{n-1} & \cdots & a_2 & a_1 \\ a_{-(n-1)} & 0 & a_{n-1} & & a_2 \\ \vdots & a_{-n+1} & 0 & \ddots & \vdots \\ a_{-2} & & \ddots & \ddots & a_{n-1} \\ a_{-1} & a_{-2} & \cdots & a_{-(n-1)} & 0 \end{bmatrix}.$$

As will be explained in Section 2.3.1, the product $T_n \mathbf{d}_n$ for any vector $\mathbf{d}_n$ can be computed by several FFTs in $\mathcal{O}(2n \log 2n) = \mathcal{O}(n \log n)$ operations.

Due to the reduction in complexity and the ability to solve a large class of (e.g. indefinite or non-Hermitian) Toeplitz systems for which direct solvers could be unstable [118], iterative solvers are competitive alternatives to solving Toeplitz systems.

2.2.2.1 Preconditioning for Toeplitz systems

As will be discussed in Chapter 3, the convergence rate of Krylov subspace methods, such as the conjugate gradient method, depends on eigenvalues. Besides, the eigenvalues of Toeplitz matrices are distributed as their generating function and hence are not usually clustered as mentioned in Section 2.1. Therefore, the number of iterations required for a Krylov subspace method to converge is in general large. Due to this property of Toeplitz matrices, preconditioning is desired to improve complexity.

Suppose originally we want to solve an $n \times n$ nonsingular system

$$A_n \mathbf{x}_n = \mathbf{b}_n.$$

We can achieve the same goal by solving the following system instead

$$P_n^{-1} A_n \mathbf{x}_n = P_n^{-1} \mathbf{b}_n,$$

where a *preconditioner* P_n is any $n \times n$ nonsingular matrix.

Provided that we solve the linear system using a Krylov subspace method, the convergence rate will depend on $P_n^{-1}A_n$ instead of A_n. In order to accelerate the convergence, the preconditioner P_n should be well chosen such that the following criteria are satisfied [4, 25, 92, 118, 47, 9, 168]:

1. For any vector $\mathbf{d}_n$, the product $P_n^{-1}\mathbf{d}_n = \mathbf{v}_n$ can be computed efficiently or the system $P_n\mathbf{v}_n = \mathbf{d}_n$ can be solved efficiently.

2. The spectrum of $P_n^{-1}A_n$ is clustered and/or $P_n^{-1}A_n$ is well-conditioned compared to A_n.

In the next section, we will explain why circulant matrices as preconditioners for Toeplitz systems can be effective by illustrating that they satisfy the abovementioned criteria under certain assumptions.

2.3 Circulant matrices

A *circulant* matrix $C_n \in \mathbb{C}^{n\times n}$ is defined by

$$C_n = \begin{bmatrix} c_0 & c_{n-1} & \cdots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & & c_2 \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_{n-2} & & \ddots & \ddots & c_{n-1} \\ c_{n-1} & c_{n-2} & \cdots & c_1 & c_0 \end{bmatrix}.$$

Note that C_n itself is a Toeplitz matrix, and every row of C_n is a right cyclic shift of the row above it.

2.3.1 Diagonalization of circulant matrices

Theorem 2.3.1 *[52, Theorem 3.2.2 and Theorem 3.2.3] Let $C_n \in \mathbb{C}^{n\times n}$ be a circulant matrix. Then, C_n is given by*

$$C_n = F_n^* \Lambda_n F_n,$$

where $F_n \in \mathbb{C}^{n\times n}$ is the Fourier matrix of which the entries are given by $[F_n]_{jk} = \frac{1}{\sqrt{n}}e^{2\pi \mathbf{i} jk/n}$, $j,k = 0,1,\ldots,n-1$, and $\Lambda_n \in \mathbb{C}^{n\times n}$ is the diagonal matrix in the eigendecomposition of C_n.

Due to the diagonalization of circulant matrices $C_n = F_n^* \Lambda_n F_n$ (see Section 3.2 in [52]), one can easily show that for any vector $\mathbf{d}_n$ the product $C_n^{-1}\mathbf{d}_n$ (or $C_n\mathbf{d}_n$) can be efficiently computed as follows.

Since the first column of F_n is $\frac{1}{\sqrt{n}}\mathbf{1}_n$, where $\mathbf{1}_n = (1, 1, 1 \ldots, 1)^T \in \mathbb{R}^n$, we have

$$F_n C_n \boldsymbol{e}_1 = \frac{1}{\sqrt{n}} \Lambda_n \mathbf{1}_n,$$

where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^n$. Therefore, the diagonal matrix Λ_n can be computed in $\mathcal{O}(n \log n)$ operations by taking an FFT of the first column of C_n. Since $C_n^{-1} = F_n^* \Lambda_n^{-1} F_n$, $C_n^{-1}\mathbf{d}_n$ for any vector $\mathbf{d}_n$ can then be computed by several FFTs in $\mathcal{O}(n \log n)$ operations once Λ_n is obtained. Therefore, C_n has the potential to be an efficient preconditioner due to this low-complexity computation.

2.3.2 Absolute value circulant matrices

As circulant matrices are diagonalizable, we can define their corresponding absolute value matrix [167].

Definition 2.3.1 *[129] Let $C_n \in \mathbb{C}^{n\times n}$ be a circulant matrix. The* absolute value circulant matrix $|C_n| \in \mathbb{C}^{n\times n}$ *of C_n is defined by*

$$\begin{aligned} |C_n| &= (C_n^* C_n)^{1/2} \\ &= (C_n C_n^*)^{1/2} \\ &= F_n^* |\Lambda_n| F_n, \end{aligned}$$

where $F_n \in \mathbb{C}^{n\times n}$ is the Fourier matrix and $|\Lambda_n| \in \mathbb{R}^{n\times n}$ is the diagonal matrix in the eigendecomposition of C_n with all entries replaced by their magnitude.

By definition, $|C_n|$ is Hermitian positive definite provided that C_n is nonsingular. Also, since $|C_n|$ itself is a circulant matrix, $|C_n|^{-1}\mathbf{d}_n$ for any vector $\mathbf{d}_n$ can be computed by several FFTs in $\mathcal{O}(n \log n)$ operations. Due to these desired properties, $|C_n|$ can be used as a preconditioner in cooperation with Krylov subspace methods.

2.3.3 Circulant matrices as preconditioners

Strang [146] and Olkin [123] are the first to independently propose using circulant matrices as preconditioners for Toeplitz matrices. Numerical results in [148] showed that such preconditioners are effective for solving a wide range of Toeplitz systems. Theoretical results that guarantee fast convergence with circulant preconditioners were later given by R. Chan and Strang [34]. Other circulant preconditioners, such as optimal circulant preconditioners by T. Chan [40], Huckle's preconditioners [86], and superoptimal preconditioners by Tyrtyshnikov [157], were developed for the Toeplitz matrices generated by certain real-valued and positive functions f. The restriction on f was later relaxed for example in [29, 141, 55] to allow f to have zeros.

Work has also been done on preconditioning for Hermitian indefinite Toeplitz systems [33, 121], non-Hermitian Toeplitz systems [88], and nonsymmetric Toeplitz systems [130]. Tyrtyshnikov, Yeremin, and Zamarashkin [161] proposed improved circulant preconditioners and showed that the eigenvalues of the preconditioned matrices were clustered around unity for the Toeplitz matrix generated by sparsely vanishing functions f in $L^2([-\pi,\pi])$, i.e. the zero set of f is of measure zero. Di Benedetto and Serra Capizzano [56], and Oseledets and Tyrtyshnikov [125] later provided a unifying way of constructing circulant preconditioners. For references on the development of preconditioning for Toeplitz matrices, we refer to [118, 25, 124].

We first provide the following definition in relation to the clustered spectra around ±1 produced by absolute value circulant preconditioners, which is relevant to our main results.

Definition 2.3.2 *[129, Definition 4.5] A sequence of matrices $\{H_n\}_{n=1}^{\infty}$ is said to have clustered spectra around ± 1 if for any $\epsilon > 0$ there exist positive integers M and N such that for all $n > N$, at most M eigenvalues λ of H_n are such that $|\lambda - 1| > \epsilon$ and $|\lambda + 1| > \epsilon$.*

Remark A sequence of matrices having clustered spectra around unity can be defined in a similar way.

In what follows, we provide three commonly used circulant preconditioners. Given a Toeplitz matrix $T_n \in \mathbb{C}^{n\times n}$

$$T_n = \begin{bmatrix} a_0 & a_{-1} & \cdots & a_{-(n-2)} & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & & a_{-(n-2)} \\ \vdots & a_1 & a_0 & \ddots & \vdots \\ a_{n-2} & & \ddots & \ddots & a_{-1} \\ a_{n-1} & a_{n-2} & \cdots & a_1 & a_0 \end{bmatrix},$$

we provide the following circulant preconditioners for T_n.

2.3.3.1 Optimal circulant preconditioners

We let

$$\mathcal{M}_{F_n} = \{F_n^* \Lambda_n F_n \mid \Lambda_n \text{ is any } n \times n \text{ diagonal matrix}\}$$

be the set of all circulant matrices [52]. The *optimal circulant preconditioner* [40] $c(T_n) \in \mathbb{C}^{n \times n}$ for T_n is defined to be the minimizer of

$$\|T_n - C_n\|_F$$

over all $C_n \in \mathcal{M}_{F_n}$, where $\| \cdot \|_F$ is the Frobenius norm.

The entries of $c(T_n)$ can be explicitly obtained:

$$c(T_n) = \begin{bmatrix} c_0 & c_{n-1} & \cdots & c_2 & c_1 \\ c_1 & c_0 & c_{n-1} & & c_2 \\ \vdots & c_1 & c_0 & \ddots & \vdots \\ c_{n-2} & & \ddots & \ddots & c_{n-1} \\ c_{n-1} & c_{n-2} & \cdots & c_1 & c_0 \end{bmatrix},$$

where

$$c_k = \begin{cases} \frac{(n-k)a_k + ka_{k-n}}{n} & 0 \le k < n \\ c_{n+k} & 0 < -k < n \end{cases}.$$

Remark In fact, optimal circulant preconditioners can be defined for general square matrices. However, we only focus on Toeplitz matrices in this thesis.

Let $\delta(A_n) \in \mathbb{C}^{n \times n}$ denote the diagonal matrix whose diagonal is equal to the diagonal of $A_n \in \mathbb{C}^{n \times n}$. The following theorem provides some important properties of $c(T_n)$.

Theorem 2.3.2 *[25, 26] Let $T_n \in \mathbb{C}^{n \times n}$ be a Toeplitz matrix and let $c(T_n) \in \mathbb{C}^{n \times n}$ be the optimal circulant preconditioner for T_n. Then, the followings hold:*

(i) $c(T_n)$ is uniquely determined by T_n and is given by

$$c(T_n) = F_n^* \delta(F_n T_n F_n^*) F_n.$$

(ii) We have

$$\sigma_{max}\big(c(T_n)\big) \le \sigma_{max}(T_n),$$

where $\sigma_{max}(A_n)$ denotes the largest singular value of A_n.

(iii) If T_n is Hermitian, then $c(T_n)$ is also Hermitian. Furthermore, we have

$$\lambda_{min}(T_n) \le \lambda_{min}\big(c(T_n)\big) \le \lambda_{max}\big(c(T_n)\big) \le \lambda_{max}(T_n),$$

where $\lambda_{min}(A_n)$ and $\lambda_{max}(A_n)$ denote the smallest and largest eigenvalue of A_n, respectively. In particular, if T_n is positive definite, then so is $c(T_n)$.

Considering the Toeplitz matrices generated by positive functions in the Wiener class, i.e. T_n whose entries satisfy

$$\sum_{k=-\infty}^{\infty} |a_k| < \infty,$$

R. Chan proved that $c(T_n)^{-1}T_n$ has clustered spectra around unity for sufficiently large n.

Theorem 2.3.3 *[21, 22] Let f be a positive function in the Wiener class. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal preconditioner for T_n. Then, $c(T_n)^{-1}T_n$ has clustered spectra around* 1 *for sufficiently large n.*

R. Chan and Yeung later extended the result to positive functions in $\mathcal{C}[-\pi,\pi]$, the Banach space of continuous complex-valued functions defined on $[-\pi,\pi]$ equipped with the supremum norm $\|\cdot\|_\infty$.

Theorem 2.3.4 *[36] Let $f \in \mathcal{C}[-\pi,\pi]$ be a positive function. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal preconditioner for T_n. Then, $c(T_n)^{-1}T_n$ has clustered spectra around* 1 *for sufficiently large n.*

Optimal preconditioners have undergone many extensions and generalizations. Their properties, such as the stability in numerical ordinary differential equation problems, have been studied for example by Jin in [20, 97, 99, 19, 98, 48], and related preconditioners like modified optimal preconditioners can be found in [111].

2.3.3.2 Superoptimal circulant preconditioners

Similar to $c(T_n)$, the *superoptimal circulant preconditioner* [157] $t(T_n) \in \mathbb{C}^{n\times n}$ for T_n is defined to be the minimizer of

$$\|I_n - C_n^{-1}T_n\|_F$$

over all $C_n \in \mathcal{M}_{F_n}$

The following theorem provides a relation between $t(T_n)$ and $c(T_n)$.

Remark Superoptimal circulant preconditioners are also defined for general square matrices.

Theorem 2.3.5 *[26, Theorem 4] Let $T_n \in \mathbb{C}^{n\times n}$ be a Toeplitz matrix and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. If both T_n and $c(T_n)$ are nonsingular, then the superoptimal circulant preconditioner $t(T_n) \in \mathbb{C}^{n\times n}$ for T_n exists and is given by*

$$t(T_n) = c(T_n^*)^{-1}c(T_nT_n^*).$$

Considering the effectiveness of $t(T_n)$ for T_n, R. Chan, Jin, and Yeung showed the following result.

Theorem 2.3.6 *[27, Theorem 5] Let f be a positive function in the Wiener class. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $t(T_n) \in \mathbb{C}^{n\times n}$ be the superoptimal circulant preconditioner for T_n. Then, $t(T_n)^{-1}T_n$ has clustered spectra around* 1 *for sufficiently large n.*

Like optimal circulant preconditioners, superoptimal circulant preconditioners have been generalized over the years for example in [46, 7, 49, 96].

2.3.3.3 Strang's circulant preconditioners

Strang's circulant preconditioner $s(T_n) \in \mathbb{C}^{n\times n}$ [146] for T_n is defined by

$$s(T_n) = \begin{bmatrix} s_0 & s_{n-1} & \cdots & s_2 & s_1 \\ s_1 & s_0 & s_{n-1} & & s_2 \\ \vdots & s_1 & s_0 & \ddots & \vdots \\ s_{n-2} & & \ddots & \ddots & s_{n-1} \\ s_{n-1} & s_{n-2} & \cdots & s_1 & s_0 \end{bmatrix},$$

where

$$s_k = \begin{cases} a_k & 0 \le k \le \lfloor n/2 \rfloor \\ a_{k-n} & \lfloor n/2 \rfloor < k < n \\ s_{n+k} & 0 < -k < n \end{cases}$$

with $\lfloor n/2 \rfloor$ denotes the greatest integer $m \le n/2$. Namely, $s(T_n)$ copies the central diagonals of T_n and wraps them around to form a circulant matrix.

Remark Strang's circulant preconditioners can only be defined for Toeplitz matrices by the definition given. For other matrices, R. Chan, Ng, and Plemmons [30] proposed the generalized Strang type circulant preconditioners.

For the Toeplitz matrix generated by positive functions in the Wiener class, Strang's circulant preconditioners are effective in the sense that the eigenvalues of the preconditioned matrices are clustered.

Theorem 2.3.7 *[34] Let f be an even positive function in the Wiener class. Let $T_n \in \mathbb{R}^{n\times n}$ be the Toeplitz matrix generated by f and let $s(T_n) \in \mathbb{R}^{n\times n}$ be Strang's circulant preconditioner for T_n. Then, $s(T_n)^{-1}T_n$ has clustered spectra around 1 for sufficiently large n.*

Using a purely linear algebra approach, R. Chan later extended the result via the following theorem.

Theorem 2.3.8 *[21] Let f is a positive function in the Wiener class. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $s(T_n) \in \mathbb{C}^{n\times n}$ be Strang's circulant preconditioner for T_n. Then, $s(T_n)^{-1}T_n$ has clustered spectra around 1 for sufficiently large n.*

With more knowledge about f, descriptive convergence rates of the conjugate gradient method for $s(T_n)^{-1}T_n$ can be found for example in [34, 37].

2.3.3.4 Other circulant preconditioners

There are also other circulant preconditioners proposed for Toeplitz matrices over years in the literature.

For the Toeplitz matrices generated by certain positive functions, i.e. positive functions whose Fourier coefficients satisfy

$$\sum_{k=0}^{\infty} |k||a_k|^2 < \infty,$$

Huckle's preconditioners developed in [86] were shown to be able to give clustered spectra.

Circulant preconditioners constructed by embedding the Toeplitz matrix into a $2n$-by-$2n$ circulant matrix have also been used. R. Chan's preconditioners [21] and the preconditioners constructed by Ku and Kuo [103] belong to this category.

$\{\omega\}$-circulant preconditioners $C_n(\omega) \in \mathbb{C}^{n\times n}$ were proposed by Potts and Steidl [131, 132], namely $C_n(\omega) = \Omega_n^* F_n^* \Lambda_n F_n \Omega_n$, where $\Lambda_n \in \mathbb{C}^{n\times n}$ is the diagonal matrix in the eigendecomposition of $C_n(\omega)$, $F_n \in \mathbb{C}^{n\times n}$ is the Fourier matrix, and

$$\Omega_n = \operatorname{diag}(1, e^{w_n\mathbf{i}}, e^{2w_n\mathbf{i}}, \ldots, e^{(n-1)w_n\mathbf{i}}) \in \mathbb{C}^{n\times n}$$

is a unitary matrix with $w_n \in [-\pi, -\pi + 2\pi/n)$.

Best circulant preconditioners by R. Chan, Yip, and Ng [39, 32] were proposed to solve ill-conditioned Toeplitz systems. For more about other circulant preconditioners, we refer to [118, 25].

2.4 Noncirculant preconditioners

Other than circulant preconditioners, noncirculant preconditioners have also been used for Toeplitz systems.

Optimal transform-based preconditioners were proposed, due to the desired property that they can be diagonalized by a fast transform. Optimal sine transform-based preconditioners [10, 31, 14, 87], optimal cosine transform-based preconditioners [23, 119, 100], and optimal Hartley transform-based preconditioners [12, 89] belong to this kind.

Band-Toeplitz matrices have also been used as preconditioners for certain ill-conditioned Toeplitz systems, for example see [28, 35, 139, 140, 76, 110].

Other preconditioners were also proposed, such as approximate inverse-free preconditioners [169]. For more about noncirculant preconditioners, we refer to [118, 25].

2.5 Functions of Toeplitz matrices

In this section, we review a few definitions and several properties of matrix functions for developing our main results on functions of Toeplitz matrices.

Throughout this thesis, we assume that the given function $h(z)$ is analytic with radius of convergence r. It is therefore sufficient to consider the following representation of matrix functions via the Taylor series expansion of $h(z)$. Also, without loss of generality, we choose α in the series representation of $h(z)$ to be zero in order to simplify notation.

Theorem 2.5.1 *[79, Theorem 4.7] Suppose $h(z)$ has a Taylor series expansion*

$$h(z) = \sum_{k=0}^{\infty} a_k (z-\alpha)^k,$$

where $a_k = \frac{h^{(k)}(\alpha)}{k!}$, with radius of convergence r. If $A_n \in \mathbb{C}^{n\times n}$, then $h(A_n) \in \mathbb{C}^{n\times n}$ is defined and is given by

$$h(A_n) = \sum_{k=0}^{\infty} a_k (A_n - \alpha I_n)^k$$

if and only if each of the distinct eigenvalues $\lambda_1, \ldots, \lambda_s$ of A_n satisfies one of the conditions

(a) $|\lambda_i - \alpha| < r$,

(b) $|\lambda_i - \alpha| = r$ and the series for $h^{(n_i-1)}(\lambda)$, where n_i is the index of λ_i, is convergent at the point $\lambda = \lambda_i$, $i = 1, \ldots, s$.

As trigonometric matrix functions of Toeplitz matrices will be discussed in Chapter 4, we provide their definition in the following.

Definition 2.5.1 *[79] For any $A_n \in \mathbb{C}^{n\times n}$,*

$$\begin{aligned} e^{A_n} &= I_n + A_n + \frac{1}{2!}A_n^2 + \frac{1}{3!}A_n^3 + \cdots, \\ \sin A_n &= A_n - \frac{1}{3!}A_n^3 + \frac{1}{5!}A_n^5 - \frac{1}{7!}A_n^7 + \cdots, \\ \cos A_n &= I_n - \frac{1}{2!}A_n^2 + \frac{1}{4!}A_n^4 - \frac{1}{6!}A_n^6 + \cdots, \\ \sinh A_n &= A_n + \frac{1}{3!}A_n^3 + \frac{1}{5!}A_n^5 + \frac{1}{7!}A_n^7 + \cdots, \end{aligned}$$

and

$$\cosh A_n = I_n + \frac{1}{2!}A_n^2 + \frac{1}{4!}A_n^4 + \frac{1}{6!}A_n^6 + \cdots.$$

We also provide the following important theorems concerning matrix functions that will be used to prove our main results.

Theorem 2.5.2 *[79, Theorem 1.18] Let $h(z)$ be analytic on an open subset $\Omega \subseteq \mathbb{C}$ such that each connected component of Ω is closed under conjugation. Consider the corresponding matrix function $h(z)$ on its natural domain in $\mathbb{C}^{n\times n}$, i.e. the set $\mathcal{D} = \{A_n \in \mathbb{C}^{n\times n} : \Lambda(A_n) \subseteq \Omega\}$. Then, the followings are equivalent:*

(a) $h(A_n^*) = h(A_n)^*$ for all $A_n \in \mathcal{D}$.

(b) $h(\overline{A_n}) = \overline{h(A_n)}$ for all $A_n \in \mathcal{D}$.

(c) $h(\mathbb{R}^{n\times n} \cap \mathcal{D}) \subseteq \mathbb{R}^{n\times n}$.

(d) $h(\mathbb{R} \cap \Omega) \subseteq \mathbb{R}$.

Theorem 2.5.3 *[79, Theorem 4.8] Suppose $h(z)$ has the Taylor series expansion*

$$h(z) = \sum_{k=0}^{\infty} a_k (z-\alpha)^k,$$

where $a_k = \frac{h^{(k)}(\alpha)}{k!}$, with radius of convergence r. If $A_n \in \mathbb{C}^{n\times n}$ with $\rho(A_n - \alpha I_n) < r$, then for any matrix norm $\|\cdot\|$

$$\|h(A_n) - \sum_{k=0}^{K-1} a_k (A_n - \alpha I_n)^k\| \le \frac{1}{K!} \max_{0\le t\le 1} \|(A_n - \alpha I_n)^K h^{(K)}(\alpha I_n + t(A_n - \alpha I_n))\|.$$

The following lemma shows that $Y_n h(T_n)$ is (real) symmetric when T_n is a real Toeplitz matrix, which will be used in Chapter 4.

Theorem 2.5.4 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let $Y_n \in \mathbb{R}^{n \times n}$ be the anti-identity matrix. If $A_n \in \mathbb{R}^{n \times n}$ with $\rho(A_n) < r$ is (real) persymmetric, i.e. $Y_n A_n = A_n^T Y_n$, then $h(A_n)$ is also (real) persymmetric.*

Proof We start by showing that A_n^k is persymmetric for any nonnegative integer k:

$$\begin{aligned}
Y_n A^k &= Y_n A_n A_n^{k-1} \\
&= A_n^T Y_n A_n^{k-1} \\
&= A_n^T Y_n A_n A_n^{k-2} \\
&= (A_n^T)^2 Y_n A_n^{k-2} \\
&\vdots \\
&= (A_n^T)^k Y_n \\
&= (A_n^k)^T Y_n.
\end{aligned}$$

Since $h(z)$ is analytic on $|z| < r$, it has the following Taylor series representation:

$$h(z) = \sum_{k=0}^{\infty} a_k z^k$$

with radius of convergence r. By Theorem 2.5.1 and the assumption that $\rho(A_n) < r$, we have

$$h(A_n) = \sum_{k=0}^{\infty} a_k A_n^k.$$

Thus,

$$\begin{aligned}
Y_n h(A_n) &= Y_n \lim_{K\to\infty} \sum_{k=0}^{K} a_k A_n^k \\
&= \lim_{K\to\infty} \sum_{k=0}^{K} a_k Y_n A_n^k \\
&= \lim_{K\to\infty} \sum_{k=0}^{K} a_k (A_n^k)^T Y_n \\
&= \lim_{K\to\infty} (\sum_{k=0}^{K} a_k (A_n^k))^T Y_n \\
&= h(A_n)^T Y_n.
\end{aligned}$$

Moreover, $h(A_n)$ is a real matrix when A_n is a real matrix by Lemma 2.5.2 (c). The result follows. ∎

Chapter 3

Krylov subspace methods

Throughout this thesis, we focus on Krylov subspace methods as Toeplitz iterative solvers. In this chapter, we briefly review the following three commonly used methods: the conjugate gradient method, the minimal residual method, and the generalized minimal residual method for solving a linear system $A_n\mathbf{x} = \mathbf{b}_n$. For each of these methods, we provide a pseudocode and the convergence results related to preconditioning for Toeplitz matrix functions.

Given an initial guess $\mathbf{x}_n^{(0)}$ and the corresponding initial residual $\mathbf{r}_n^{(0)} = \mathbf{b}_n - A_n\mathbf{x}_n^{(0)}$, we define the k-th Krylov subspace

$$\mathcal{K}_k(A_n, \mathbf{r}_n^{(0)}) = \operatorname{span}\{\mathbf{r}_n^{(0)}, A_n\mathbf{r}_n^{(0)}, A_n^2\mathbf{r}_n^{(0)}, \ldots, A_n^{k-1}\mathbf{r}_n^{(0)}\}, \qquad k = 1, 2, \ldots.$$

Krylov subspace methods for finding $\mathbf{x}_n \in \mathbb{C}^{n\times n}$ satisfying $A_n\mathbf{x}_n = \mathbf{b}_n$ compute iterates $\mathbf{x}_n^{(k)}$ for which

$$\mathbf{x}_n^{(k)} - \mathbf{x}_n^{(0)} \in \mathcal{K}_k(A_n, \mathbf{r}_n^{(0)}), \qquad k = 1, 2, \ldots,$$

from some initial guess $\mathbf{x}_n^{(0)}$, and thus require one matrix-vector product computation at each iteration. The vector $\mathbf{r}_n^{(k)} = \mathbf{b}_n - A_n\mathbf{x}_n^{(k)}$, $k = 0, 1, 2, \ldots$, is the k-th residual. If $\mathbf{x}_n^{(0)} = \mathbf{0}$, then

$$\mathbf{x}_n^{(k)} \in \mathcal{K}_k(A_n, \mathbf{b}_n), \qquad k = 1, 2, \ldots.$$

Thus, the iterates and residuals of any Krylov subspace method satisfy

$$\mathbf{x}_n^{(k)} - \mathbf{x}_n^{(0)} = \sum_{i=0}^{k-1} \alpha_i A_n^i \mathbf{r}_n^{(0)}$$

for some coefficients $\alpha_i, i = 0, 1, 2, \cdots, k-1$. Hence,

$$\mathbf{x}_n^{(k)} = \mathbf{x}_n^{(0)} + q(A_n)\mathbf{r}_n^{(0)}, \tag{3.1}$$

where $q(z)$ is the polynomial of degree $k-1$ given by

$$q(z) = \sum_{i=0}^{k-1} \alpha_i z^i.$$

Premultiplying (3.1) by A_n and then subtracting it from $\mathbf{b}_n$, we obtain

$$\mathbf{b}_n - A_n \mathbf{x}_n^{(k)} = \mathbf{b}_n - A_n \mathbf{x}_n^{(0)} - A_n q(A_n) \mathbf{r}_n^{(0)},$$

and therefore

$$\begin{aligned} \mathbf{r}_n^{(k)} &= \mathbf{r}_n^{(0)} - A_n q(A_n) \mathbf{r}_n^{(0)} \\ &= p(A_n) \mathbf{r}_n^{(0)}, \end{aligned} \tag{3.2}$$

where

$$\begin{aligned} p(z) &= 1 - z \sum_{i=0}^{k-1} \alpha_i z^i \\ &= 1 - \sum_{i=1}^{k} \alpha_{i-1} z^i \end{aligned}$$

is a polynomial of degree k that satisfies $p(0) = 1$.

3.1 Generalized minimal residual method

One of the most widely used iterative methods for non-Hermitian problems is the generalized minimum residual (GMRES) method developed by Saad and Schultz [136]. A typical GMRES implementation is given in Algorithm 1. The idea of GMRES is to find the k-th iterate $\mathbf{x}_n^{(k)}$ that minimizes

$$\|\mathbf{b}_n - A_n \mathbf{x}_n^{(k)}\|_2$$

over $\mathbf{x}_n^{(0)} + \mathcal{K}_k(A_n, \mathbf{r}_n^{(0)})$.

GMRES is based on the Arnoldi method [2] given in Algorithm 2, which uses a modified Gram-Schmidt process to construct an orthogonal basis

$$\{\mathbf{q}_n^{(1)}, \mathbf{q}_n^{(2)}, \ldots, \mathbf{q}_n^{(k)}\}$$

for $\mathcal{K}_k(A_n, \mathbf{r}_n^{(0)})$. We can view this method as a computation of projections onto successive Krylov subspaces in matrix form. Let

$$Q_k = [\mathbf{q}_n^{(1)} \ \mathbf{q}_n^{(2)} \ \ldots \ \mathbf{q}_n^{(k)}] \in \mathbb{C}^{n \times k}$$

Algorithm 1 Generalized minimal residual method [61, Algorithm 7.1]

```
Choose $\mathbf{x}_n^{(0)}$, compute $\mathbf{r}_n^{(0)} = \mathbf{b}_n - A_n\mathbf{x}_n^{(0)}$, set $\mathbf{q}_n^{(1)} = \mathbf{r}_n^{(0)}/\|\mathbf{r}_n^{(0)}\|_2$
for $k = 1$ until convergence do
    Do step $k$ of the Arnoldi method
    Update the QR factorization of $H_{k+1,k}$ in (3.4)
    Solve $\mathbf{y}_k = \arg\min \| \|\mathbf{r}_n^{(0)}\|_2 \boldsymbol{e}_1 - H_{k+1,k}\mathbf{y}_k\|_2$
    Set $\mathbf{x}_n^{(k)} = \mathbf{x}_n^{(0)} + Q_k\mathbf{y}_k$
    <Test for convergence>
end for
```

Algorithm 2 Arnoldi method [61]

```
Choose $\mathbf{q}_n^{(1)} \in \mathbb{C}^n$ with $\|\mathbf{q}_n^{(1)}\|_2 = 1$
for $k = 1, 2, \dots$ do
    $\mathbf{w}_n = A_n\mathbf{q}_n^{(k)}$
    for $j = 1, 2, \dots, k$ do
        $h_{j,k} = \langle \mathbf{q}_n^{(j)}, \mathbf{w}_n \rangle$
        $\mathbf{w}_n = \mathbf{w}_n - h_{j,k}\mathbf{q}_n^{(j)}$
    end for
    $h_{k+1,k} = \|\mathbf{w}_n\|_2$
    $\mathbf{q}_n^{(k+1)} = \mathbf{w}_n / h_{k+1,k}$
end for
```

and let

$$H_k = \begin{bmatrix} h_{11} & \cdots & \cdots & h_{1k} \\ h_{21} & h_{22} & & \vdots \\ & \ddots & \ddots & \vdots \\ & & h_{k,k-1} & h_{k,k} \end{bmatrix} \in \mathbb{C}^{k\times k} \tag{3.3}$$

be an upper Hessenberg matrix. Defining

$$H_{k+1,k} = \begin{bmatrix} H_k \\ h_{k+1,k}\boldsymbol{e}_k^T \end{bmatrix} \in \mathbb{C}^{(k+1)\times k}, \tag{3.4}$$

we have

$$\begin{aligned} A_nQ_k &= Q_kH_k + h_{k+1,k}\mathbf{q}_n^{(k+1)}\boldsymbol{e}_k^T \\ &= Q_{k+1}H_{k+1,k}, \qquad k = 1, 2, \dots, \end{aligned}$$

where $\boldsymbol{e}_k = (0, \dots, 0, 1)^T$ is the k-th coordinate k-vector. Choosing $\mathbf{q}_n^{(1)} = \frac{\mathbf{r}_n^{(0)}}{\|\mathbf{r}_n^{(0)}\|_2}$, the k-th iterate $\mathbf{x}_n^{(k)}$ corresponds to

$$\mathbf{x}_n^{(k)} = \mathbf{x}_n^{(0)} + Q_k\mathbf{y}_k$$

for some $\mathbf{y}_k \in \mathbb{C}^k$.

We can restate GMRES as solving the following least squares problem

$$\begin{aligned}
\min_{\mathbf{x}_n^{(k)} \in\ \mathbf{x}_n^{(0)} + \mathcal{K}_k(A_n, \mathbf{r}_n^{(0)})} \|\mathbf{b}_n - A_n \mathbf{x}_n^{(k)}\|_2 &= \min_{\mathbf{y}_k \in \mathbb{C}^k} \|\mathbf{r}_n^{(0)} - A_n Q_k \mathbf{y}_k\|_2 \\
&= \min_{\mathbf{y}_k \in \mathbb{C}^k} \|\mathbf{r}_n^{(0)} - Q_{k+1} H_{k+1,k} \mathbf{y}_k\|_2 \\
&= \min_{\mathbf{y}_k \in \mathbb{C}^k} \left\| Q_{k+1}(\|\mathbf{r}_n^{(0)}\|_2 \boldsymbol{e}_1 - H_{k+1,k}\mathbf{y}_k) \right\|_2 \\
&= \min_{\mathbf{y}_k \in \mathbb{C}^k} \left\| \|\mathbf{r}_n^{(0)}\|_2 \boldsymbol{e}_1 - H_{k+1,k}\mathbf{y}_k \right\|_2,
\end{aligned}$$

where $\boldsymbol{e}_1 = (1, , 0 \ldots, 0)^T$ is the first coordinate $(k+1)$-vector. We can then solve this least squares problem using the QR factorization of the Hessenberg matrix $H_{k+1,k}$, which can be achieved by one additional Givens rotation for each k since $H_{k+1,k}$ is built up by adding the last column for each k.

Assuming the (full) QR factorization of $H_{k,k-1}$ is

$$V_k^* H_{k,k-1} = R_k,$$

where $R_k \in \mathbb{C}^{k,k-1}$ is upper triangular, we have

$$\begin{aligned}
\begin{bmatrix} V_k^* & \\ & 1 \end{bmatrix} H_{k,k-1} &= \begin{bmatrix} V_k^* & \\ & 1 \end{bmatrix} \begin{bmatrix} H_{k,k-1} & \mathbf{h}_k \\ & h_{k+1,k} \end{bmatrix} \\
&= \begin{bmatrix} R_k & V_k^* \mathbf{h}_k \\ & h_{k+1,k} \end{bmatrix}, \qquad (3.5)
\end{aligned}$$

where $\mathbf{h}_k = [h_{1,k}\ h_{2,k}\ \ldots\ h_{k,k}]^T$. In other words, premultiplying (3.5) by

$$G_{k+1} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & c_{k+1} & s_{k+1} \\ & & & -\bar{s}_{k+1} & c_{k+1} \end{bmatrix} \in \mathbb{C}^{k+1,k+1},$$

where $c_{k+1} = \cos\theta_{k+1}$ and $s_{k+1} = \sin\theta_{k+1}$ represent the appropriate rotation to zero out the entry $h_{k+1,k}$, can render the QR factorization of $H_{k+1,k}$ as $H_{k+1,k} = V_{k+1}R_{k+1}$. Namely,

$$\underbrace{G_{k+1} \begin{bmatrix} V_k^* & \\ & 1 \end{bmatrix}}_{V_{k+1}^*} H_{k+1,k} = \underbrace{G_{k+1} \begin{bmatrix} R_k & V_k^* \mathbf{h}_k \\ & h_{k+1,k} \end{bmatrix}}_{R_{k+1}}.$$

The full GMRES algorithm given requires increasing storage and operations as k increases. If the number of iterations needed to solve a linear system is large, this creates significant demands of storage and computation. In such case, one can for example use GMRES(l) [136] (restarted GMRES), which simply restarts every l steps and uses the newest iterate as the initial guess for the next GMRES cycle.

3.1.1 Convergence of GMRES

Let Π_k be the set of polynomials of degree at most k and let $\Lambda(A_n)$ be the set of all eigenvalues of the matrix $A_n \in \mathbb{C}^{n\times n}$.

By (3.2), we have the following theorem that relates the convergence of GMRES with minimal polynomials.

Theorem 3.1.1 *[61] Let $A_n \in \mathbb{C}^{n\times n}$ and let $\mathbf{r}_n^{(k)}$ be the k-th residual of GMRES. Then,*

$$\frac{\|\mathbf{r}_n^{(k)}\|_2}{\|\mathbf{r}_n^{(0)}\|_2} \le \min_{p_k \in \Pi_k, p(0)=1} \|p_k(A_n)\|_2.$$

As a consequence, we have the following result.

Corollary 3.1.2 *[61] Let $A_n \in \mathbb{C}^{n\times n}$. Then, GMRES will terminate with the solution within n iterations in exact arithmetic.*

We have now a guarantee that GMRES will converge within a number of iterations equal to the dimension of the system. In fact, we can improve this termination bound if we have information about the minimal polynomial of the matrix.

Let $\kappa(B_n) = \|B_n\|_2 \|B_n^{-1}\|_2$ denote the *condition number* of $B_n \in \mathbb{C}^{n\times n}$.

Theorem 3.1.3 *[61, Theorem 7.1] Let $A_n \in \mathbb{C}^{n\times n}$ and let $\mathbf{r}_n^{(k)}$ be the k-th residual of GMRES. If A_n is diagonalizable, i.e. $A_n = V_n \Lambda_n V_n^{-1}$ where Λ_n is the diagonal matrix containing the eigenvalues of A_n and V_n is the matrix whose columns are the eigenvectors, then*

$$\frac{\|\mathbf{r}_n^{(k)}\|_2}{\|\mathbf{r}_n^{(0)}\|_2} \le \kappa(V_n) \min_{p_k \in \Pi_k, p(0)=1} \max_{\lambda_i \in \Lambda(A_n)} \|p_k(\lambda_i)\|_2.$$

Proof By Theorem 3.1.1 and the assumption that $A_n = V_n \Lambda_n V_n^{-1}$,

$$\begin{aligned}
\frac{\|\mathbf{r}_n^{(k)}\|_2}{\|\mathbf{r}_n^{(0)}\|_2} &\le \min_{p_k \in \Pi_k, p(0)=1} \|p_k(V_n \Lambda_n V_n^{-1})\|_2 \\
&\le \underbrace{\|V_n\|_2 \|V_n^{-1}\|_2}_{\kappa(V_n)} \min_{p_k \in \Pi_k, p(0)=1} \|p_k(\Lambda_n)\|_2 \\
&\le \kappa(V_n) \min_{p_k \in \Pi_k, p(0)=1} \max_{\lambda_i \in \Lambda(A_n)} \|p_k(\lambda_i)\|_2. \qquad \blacksquare
\end{aligned}$$

From Theorem 3.1.3, we immediately have the following corollary.

Corollary 3.1.4 *[61] Let $A_n \in \mathbb{C}^{n\times n}$. If A_n is diagonalizable with k distinct eigenvalues, then GMRES will terminate with the solution within k iterations in exact arithmetic.*

Even though, by Theorem 3.1.3, we have a convergence bound that depends on $\kappa(V_n)$ and a term depending on the spectrum of A_n, knowledge about $\kappa(V_n)$ is usually not available and therefore this result is often not easily applicable. Besides, the eigenvalues of A_n alone cannot fully describe the convergence of GMRES. In fact, Greenbaum, Pták, and Strakoš in [72] provided the following negative result in this regard. Given any n eigenvalues and any nonincreasing convergence curve terminating at or before the n-th iteration, for any given $\mathbf{b}_n$ there exists a matrix $A_n \in \mathbb{C}^{n\times n}$ with those eigenvalues and an initial guess $\mathbf{x}_n^{(0)}$ such that GMRES will give such a convergence curve. More negative results can be found for example in [59, 60].

Preconditioning [135] can be incorporated in the GMRES method provided in Algorithm 1. The application of a left preconditioner P_n requires only minor modifications to GMRES and a right preconditioner is also readily applied; see for example [61]. In the left PGMRES method, the initial residual is replaced by the preconditioned residual $\mathbf{r}_n^{(0)} = P_n^{-1}(\mathbf{b}_n - A_n\mathbf{x}_n^{(0)})$ and the vector $\mathbf{w}_n$ becomes $\mathbf{w}_n = P_n^{-1}A_n\mathbf{r}_n^{(k)}$ at each step of the Arnoldi method (Algorithm 2).

3.2 Minimal residual method

The minimal residual (MINRES) method developed by Paige and Saunders [126] can be regarded as a special case of GMRES when the matrix is Hermitian. A pseudocode of the algorithm is presented in Algorithm 3 for reference. Similar to GMRES, MINRES finds the k-th iterate $\mathbf{x}_n^{(k)}$ that minimizes

$$\|\mathbf{b}_n - A_n\mathbf{x}_n^{(k)}\|_2$$

over $\mathbf{x}_n^{(0)} + \mathcal{K}_k(A_n, \mathbf{r}_n^{(0)})$.

However, unlike GMRES the Arnoldi method used in the previous section can be simplified to a 3-term recurrence known as the Lanczos method [105, 106] for Hermitian matrices to generate an orthonormal basis for the Krylov subspace. We can write the Lanczos method in matrix form: recalling (3.3), the upper Hessenberg

Algorithm 3 Minimal residual method [61, Algorithm 2.4]

```
Set $\mathbf{r}_n^{(0)} = \mathbf{w}_n^{(0)} = \mathbf{w}_n^{(1)} = \mathbf{0}$
Choose $\mathbf{x}_n^{(0)}$, compute $\mathbf{r}_n^{(1)} = \mathbf{b}_n - A_n\mathbf{x}_n^{(0)}$, set $\gamma_1 = \|\mathbf{r}_n^{(1)}\|_2$
Set $\eta_1 = \gamma_1$, $s_0 = s_1 = 0$, $c_0 = c_1 = 1$
for $k = 1$ until convergence do
    $\mathbf{r}_n^{(k)} = \mathbf{r}_n^{(k)}/\gamma_k$
    $\delta_k = \left\langle A_n\mathbf{r}_n^{(k)}, \mathbf{r}_n^{(k)}\right\rangle$
    $\mathbf{r}_n^{(k+1)} = A_n\mathbf{r}_n^{(k)} - \delta_k\mathbf{r}_n^{(k)} - \gamma_k\mathbf{r}_n^{(k-1)}$ (the Lanczos method)
    $\gamma_{k+1} = \|\mathbf{r}_n^{(k+1)}\|_2$
    $\alpha_0 = c_k\delta_k - c_{k-1}s_k\gamma_k$ (update the QR factorization)
    $\alpha_1 = \sqrt{\alpha_0^2 + \gamma_{k+1}^2}$
    $\alpha_2 = s_k\delta_k + c_{k-1}c_k\gamma_k$
    $\alpha_3 = s_{k-1}\gamma_k$
    $c_{k+1} = \alpha_0/\alpha_1$; $s_{k+1} = \gamma_{k+1}/\alpha_1$ (Givens rotation)
    $\mathbf{w}_n^{(k+1)} = (\mathbf{r}_n^{(k)} - \alpha_3\mathbf{w}_n^{(k-1)} - \alpha_2\mathbf{w}_n^{(k)})/\alpha_1$
    $\mathbf{x}_n^{(k)} = \mathbf{x}_n^{(k-1)} + c_{k+1}\eta\mathbf{x}_n^{(k+1)}$
    $\eta = -s_{k+1}\eta$
    <Test for convergence>
end for
```

matrix H_k in this case is simplified to be a Hermitian tridiagonal matrix of recurrence coefficients denoted by $\widehat{H}_k \in \mathbb{C}^{k\times k}$, namely

$$\widehat{H}_k = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \alpha_3 & \ddots & \\ & & \ddots & \ddots & \beta_{k-1} \\ & & & \beta_{k-1} & \alpha_k \end{bmatrix}$$

where $\alpha_i = h_{ii}$ and $\beta_i = h_{i+1,i} = h_{i,i+1}$, $i = 1, \ldots, k-1$.

3.2.1 Convergence of MINRES

By Theorem 3.1.3 and the fact that Hermitian matrices are unitarily diagonalizable, we have the following theorem.

Theorem 3.2.1 *[61] Let $A_n \in \mathbb{C}^{n\times n}$ be Hermitian and let $\mathbf{r}_n^{(k)}$ be the k-th residual of MINRES. Then,*

$$\frac{\|\mathbf{r}_n^{(k)}\|_2}{\|\mathbf{r}_n^{(0)}\|_2} \le \min_{p_k\in\Pi_k, p(0)=1} \max_{\lambda_i\in\Lambda(A_n)} \|p_k(\lambda_i)\|_2.$$

By Theorem 3.2.1, we know that MINRES will perform well in exact arithmetic when the eigenvalues of A_n are clustered. With more knowledge about the spectrum of A_n, a more descriptive error bound can be found via the following theorem.

Theorem 3.2.2 *[171, Theorem 3.1] Let $A_n \in \mathbb{C}^{n\times n}$ be Hermitian and let $\mathbf{r}_n^{(k)}$ be the k-th residual of MINRES. If the eigenvalues of A_n are ordered such that*

$$\lambda_1 \le \cdots \le \lambda_p \le -b_1 \le \lambda_{p+1} \le \cdots \le \lambda_{n_1-q} \le -b_2 \le \lambda_{n_1-q+1} \le \cdots \le \lambda_{n_1} < 0$$

and

$$0 < \lambda_{n_1+1} \le \cdots \le \lambda_{n_1+m} \le b_3 \le \lambda_{n_1+m+1} \le \cdots \le \lambda_{n-l} \le b_4 \le \lambda_{n-l+1} \le \cdots \le \lambda_n,$$

where $b_1, b_2, b_3,$ and b_4 are positive constants with $b_1-b_2 = b_4-b_3$, then for $k \ge p+q+m+l$

$$\frac{\|\mathbf{r}_n^{(k)}\|_2}{\|\mathbf{r}_n^{(0)}\|_2} \le 2\left(\frac{\sqrt{b_1b_4}-\sqrt{b_2b_3}}{\sqrt{b_1b_4}+\sqrt{b_2b_3}}\right)^{[(k-p-q-m-l)/2]} \cdot \max_{\lambda\in[-b_1,-b_2]\cup[b_3,b_4]} |P(\lambda)|,$$

where

$$P(\lambda) := \prod_{j=1}^{p}\left(\frac{\lambda-\lambda_j}{\lambda_j}\right)\cdot \prod_{j=n_1-q+1}^{n_1}\left(\frac{\lambda_j-\lambda}{\lambda_j}\right)\cdot \prod_{j=n_1+1}^{n_1+m}\left(\frac{\lambda-\lambda_j}{\lambda_j}\right)\cdot \prod_{j=n-l+1}^{n}\left(\frac{\lambda_j-\lambda}{\lambda_j}\right).$$

A corollary concerns the convergence rate for A_n with clustered spectra around ±1, which is relevant to our main results on Toeplitz matrix functions, is given in the following.

Corollary 3.2.3 *[171, Theorem 3.2] Let $A_n \in \mathbb{C}^{n\times n}$ be Hermitian, let $\mathbf{r}_n^{(k)}$ be the k-th residual of MINRES, and let $0 < \epsilon < 1$. If the eigenvalues of A_n are ordered such that*

$$\lambda_1 \le \cdots \le \lambda_p \le -1-\epsilon \le \lambda_{p+1} \le \cdots \le \lambda_{n_1-q} \le -1+\epsilon \le \lambda_{n_1-q+1} \le \cdots \le \lambda_{n_1} \le -\delta < 0$$

and

$$0 < \delta \le \lambda_{n_1+1} \le \cdots \le \lambda_{n_1+m} \le 1-\epsilon \le \lambda_{n_1+m+1} \le \cdots \le \lambda_{n-l} \le 1+\epsilon \le \lambda_{n-l+1} \le \cdots \le \lambda_n,$$

where $\delta > 0$ is a constant independent of n, then

$$\frac{\|\mathbf{r}_n^{(k)}\|_2}{\|\mathbf{r}_n^{(0)}\|_2} \le 2B\epsilon^{[(k-p-q-m-l)/2]},$$

where $k \ge p+q+m+l$ and

$$B := \max\left\{\left(\frac{1+\epsilon}{\delta}\right)^q\cdot\left(\frac{2}{\delta}\right)^m\cdot 2^l,\ 2^p\cdot\left(\frac{2}{\delta}\right)^q\cdot\left(\frac{1+\epsilon}{\delta}\right)^m\right\}.$$

If a Hermitian positive definite preconditioner P_n is employed, the above theorems hold with A_n replaced by $P_n^{-1}A_n$. Since $P_n^{1/2}P_n^{-1}A_nP_n^{-1/2} = P_n^{-1/2}A_nP_n^{-1/2}$ when P_n is Hermitian positive definite, $P_n^{-1}A_n$ is similar to the Hermitian matrix $P_n^{-1/2}A_nP_n^{-1/2}$. Hence, it suffices to consider the preconditioned matrix $P_n^{-1}A_n$, which we want to have a clustered spectrum. This method differs only slightly from Algorithm 3: $\mathbf{r}_n^{(1)}$ is replaced by $\mathbf{r}_n^{(1)} = P_n^{-1}(\mathbf{b}_n - A_n\mathbf{x}_n^{(0)})$ and, at each iteration,

$$\delta_k = \left\langle P_n^{-1}A_n\mathbf{r}_n^{(k)}, \mathbf{r}_n^{(k)}\right\rangle$$

and

$$\mathbf{r}_n^{(k+1)} = P_n^{-1}A_n\mathbf{r}_n^{(k)} - \delta_k\mathbf{r}_n^{(k)}\gamma_k\mathbf{r}_n^{(k-1)}.$$

3.3 Conjugate gradient method

The conjugate gradient (CG) method was invented by Hestenes and Stiefel [78] for solving Hermitian positive definite systems $A_n\mathbf{x}_n = \mathbf{b}_n$. The k-th iterate $\mathbf{x}_n^{(k)}$ of CG minimizes the error

$$\| \underbrace{\mathbf{x}_n - \mathbf{x}_n^{(k)}}_{\mathbf{e}_n^{(k)}} \|_{A_n}$$

over $\mathbf{x}_n^{(0)} + \mathcal{K}_k(A_n, \mathbf{r}_n^{(0)})$, where $\|\mathbf{u}_n\|_{A_n} = \sqrt{\mathbf{u}_n^* A_n \mathbf{u}_n}$. Equivalently, the k-th iterate $\mathbf{x}_n^{(k)}$ of CG minimizes the functional

$$\phi(\mathbf{x}_n) = \frac{1}{2}\mathbf{x}_n^* A_n \mathbf{x}_n - \mathbf{x}_n^* \mathbf{b}_n$$

over $\mathbf{x}_n^{(0)} + \mathcal{K}_k(A_n, \mathbf{r}_n^{(0)})$. A typical CG implementation is given in Algorithm 4, and CG requires only one matrix-vector product with A_n at each iteration to compute $\mathbf{x}_n^{(k)}$.

Algorithm 4 Conjugate gradient method [61, Algorithm 2.1]

```
Choose $\mathbf{x}_n^{(0)}$, compute $\mathbf{r}_n^{(0)} = \mathbf{b}_n - A_n\mathbf{x}_n^{(0)}$, and set $\mathbf{p}_n^{(0)} = \mathbf{r}_n^{(0)}$
for $k = 0$ until convergence do
    $\alpha_k = \langle \mathbf{r}_n^{(k)}, \mathbf{r}_n^{(k)} \rangle / \langle A_n\mathbf{p}_n^{(k)}, \mathbf{p}_n^{(k)} \rangle$
    $\mathbf{x}_n^{(k+1)} = \mathbf{x}_n^{(k)} + \alpha_k\mathbf{p}_n^{(k)}$
    $\mathbf{r}_n^{(k+1)} = \mathbf{r}_n^{(k)} - \alpha_k A_n\mathbf{p}_n^{(k)}$
    <Test for convergence>
    $\beta_k = \langle \mathbf{r}_n^{(k+1)}, \mathbf{r}_n^{(k+1)} \rangle / \langle \mathbf{r}_n^{(k)}, \mathbf{r}_n^{(k)} \rangle$
    $\mathbf{p}_n^{(k+1)} = \mathbf{r}_n^{(k+1)} + \beta_k\mathbf{p}_n^{(k)}$
end for
```

3.3.1 Convergence of CG

We also provide several convergence theorems of CG that are relevant to preconditioning for Toeplitz-related systems.

Theorem 3.3.1 *[61] Let $A_n \in \mathbb{C}^{n\times n}$ be Hermitian positive definite and let $\mathbf{e}_n^{(k)}$ be the error of the k-th CG iterate. Then,*

$$\frac{\|\mathbf{e}_n^{(k)}\|_{A_n}}{\|\mathbf{e}_n^{(0)}\|_{A_n}} \leq \min_{p\in\Pi_k, p(0)=1} \max_{\lambda\in\Lambda(A_n)} |p(\lambda)|.$$

If $k = n$, one can choose the n-th degree polynomial that passes through all the eigenvalues of A_n with $p(0) = 1$ and obtain the following corollary.

Corollary 3.3.2 *[61] Let $A_n \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. Then, CG will terminate with the solution within n iterations in exact arithmetic.*

Using Theorem 3.3.1, one can choose p to be the k-th degree Chebyshev polynomial and obtain the following theorem.

Theorem 3.3.3 *[61] Let $A_n \in \mathbb{C}^{n\times n}$ be Hermitian positive definite and let $\mathbf{e}_n^{(k)}$ be the error of the k-th CG iterate. Then,*

$$\frac{\|\mathbf{e}_n^{(k)}\|_{A_n}}{\|\mathbf{e}_n^{(0)}\|_{A_n}} \leq 2\left(\frac{\sqrt{\kappa(A_n)}-1}{\sqrt{\kappa(A_n)}+1}\right)^k.$$

By Theorem 3.3.1 and 3.3.3, we can expect that CG will perform well in exact arithmetic when the eigenvalues of A_n are clustered. Similar to MINRES, with more knowledge about the spectrum of A_n, a more descriptive error bound can be found for example in [4] and via the following theorem.

Theorem 3.3.4 *[118, Theorem 2.3] Let $A_n \in \mathbb{C}^{n\times n}$ be Hermitian positive definite and let $\mathbf{e}_n^{(k)}$ be the error of the k-th CG iterate. If the eigenvalues of A_n are ordered such that*

$$0 < \lambda_1 \leq \cdots \leq \lambda_i \leq b_1 \leq \lambda_{i+1} \leq \cdots \leq \lambda_{n-j} \leq b_2 \leq \lambda_{n-j+1} \leq \cdots \leq \lambda_n,$$

then for $k \geq i + j$

$$\frac{\|\mathbf{e}_n^{(k)}\|_{A_n}}{\|\mathbf{e}_n^{(0)}\|_{A_n}} \leq 2\left(\frac{b-1}{b+1}\right)^{k-i-j} \cdot \max_{\lambda\in[b_1,b_2]} \left\{\prod_{l=1}^{i}\left(\frac{\lambda-\lambda_l}{\lambda_l}\right)\prod_{l=n-j+1}^{n}\left(\frac{\lambda_l-\lambda}{\lambda_l}\right)\right\},$$

where $b = \sqrt{b_2/b_1} \geq 1$.

The following corollary concerns the convergence rate for A_n with clustered spectra around 1. As we will see in Chapter 4, this corollary applies to Toeplitz matrix functions with clustered spectra.

Corollary 3.3.5 *Let $A_n \in \mathbb{C}^{n\times n}$ be Hermitian positive definite, let $\mathbf{e}_n^{(k)}$ be the error of the k-th CG iterate, and let $0 < \epsilon < 1$. If the eigenvalues of A_n are ordered such that*

$$0 < \delta < \lambda_1 \leq \cdots \leq \lambda_i \leq 1-\epsilon \leq \lambda_{i+1} \leq \cdots \leq \lambda_{n-j} \leq 1+\epsilon \leq \lambda_{n-j+1} \leq \cdots \leq \lambda_n,$$

then

$$\frac{\|\mathbf{e}_n^{(k)}\|_{A_n}}{\|\mathbf{e}_n^{(0)}\|_{A_n}} \leq 2\left(\frac{1+\epsilon}{\delta}\right)^i \epsilon^{k-i-j},$$

where $k \geq i + j$.

A Hermitian positive definite preconditioner P_n can be easily incorporated with CG to achieve such a clustered spectrum of eigenvalues. In Algorithm 5, a preconditioned conjugate gradient algorithm for $P_n^{-1}A_n$ is given. Note that $P_n^{-1}A_n$ is similar to the Hermitian positive definite matrix $P_n^{-1/2}A_nP_n^{-1/2}$, provided both A_n and P_n are Hermitian positive definite. With A_n replaced by $P_n^{-1}A_n$, the abovementioned theorems concerning the convergence rate of CG hold.

Algorithm 5 Preconditioned conjugate gradient method [61, Algorithm 2.2]

```
Choose $\mathbf{x}_n^{(0)}$, compute $\hat{\mathbf{r}}_n^{(0)} = P_n^{-1}(\mathbf{b}_n - A_n\mathbf{x}_n^{(0)})$, set $\mathbf{p}_n^{(0)} = \hat{\mathbf{r}}_n^{(0)}$
for $k = 0$ until convergence do
    $\alpha_k = \langle \hat{\mathbf{r}}_n^{(k)}, \hat{\mathbf{r}}_n^{(k)} \rangle / \langle P_n^{-1}A_n\mathbf{p}_n^{(k)}, \mathbf{p}_n^{(k)} \rangle$
    $\mathbf{x}_n^{(k+1)} = \mathbf{x}_n^{(k)} + \alpha_k\mathbf{p}_n^{(k)}$
    $\hat{\mathbf{r}}_n^{(k+1)} = \hat{\mathbf{r}}_n^{(k)} - \alpha_k P_n^{-1}A_n\mathbf{p}_n^{(k)}$
    <Test for convergence>
    $\beta_k = \langle \hat{\mathbf{r}}_n^{(k+1)}, \hat{\mathbf{r}}_n^{(k+1)} \rangle / \langle \hat{\mathbf{r}}_n^{(k)}, \hat{\mathbf{r}}_n^{(k)} \rangle$
    $\mathbf{p}_n^{(k+1)} = \hat{\mathbf{r}}_n^{(k+1)} + \beta_k\mathbf{p}_n^{(k)}$
end for
```

Other well-known iterative methods such as multigrid methods [17, 156], domain decomposition methods [43, 154], the biconjugate gradient stabilized (BiCGStab) method [164], BiCGStab(l) [144], the quasi-minimal residual (QMR) method [64], and induced dimension reduction (IDR) methods [145] can also be used. However, as we do not implement these methods in this thesis, we refer the readers to [109, 165, 124, 168, 61, 71, 114, 75, 63, 70, 5] for more detailed discussions about iterative solvers.

3.4 Conclusions

We have discussed several typical Krylov subspace methods, including CG, MINRES, and GMRES. Their algorithm and the convergence results relevant to Toeplitz matrix functions have also been provided. In the numerical tests provided in the subsequent chapters, these methods will be used to demonstrate the effectiveness of our proposed preconditioners.

Chapter 4

Optimal preconditioners for functions of Toeplitz matrices[1]

Instead of directly dealing with a $n \times n$ (real) nonsymmetric Toeplitz system $T_n \mathbf{x}_n = \mathbf{b}_n$, where

$$T_n = \begin{bmatrix} a_0 & a_{-1} & \cdots & a_{-n+2} & a_{-n+1} \\ a_1 & a_0 & a_{-1} & & a_{-n+2} \\ \vdots & a_1 & a_0 & \ddots & \vdots \\ a_{n-2} & & \ddots & \ddots & a_{-1} \\ a_{n-1} & a_{n-2} & \cdots & a_1 & a_0 \end{bmatrix} \in \mathbb{R}^{n \times n},$$

Pestana and Wathen suggested in [129, 112] that one can premultiply it by the anti-identity matrix Y_n, defined as

$$Y_n = \begin{bmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{bmatrix} \in \mathbb{R}^{n \times n},$$

to obtain the symmetric system $Y_n T_n \mathbf{x}_n = Y_n \mathbf{b}_n$ (i.e. a Hankel system), where

$$Y_n T_n = \begin{bmatrix} a_{n-1} & a_{n-2} & \cdots & a_1 & a_0 \\ a_{n-2} & & \cdot^{\cdot^{\cdot}} & \cdot^{\cdot^{\cdot}} & a_{-1} \\ \vdots & a_1 & a_0 & \cdot^{\cdot^{\cdot}} & \vdots \\ a_1 & a_0 & a_{-1} & & a_{-n+2} \\ a_0 & a_{-1} & \cdots & a_{-n+2} & a_{-n+1} \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Using a suitable absolute value circulant matrix $|C_n|$ as a preconditioner, they further proved that the eigenvalues of the preconditioned matrix $|C_n|^{-1} Y_n T_n$ are clustered around ±1.

[1] This chapter is adapted from two papers [81, 84]. My main contribution to the paper in collabration with Andy Wathen is proving the theorems as well as providing the numerical results therein.

In this chapter, we show that Pestana and Wathen's idea of using Y_n as a reordering device and $|C_n|$ as a preconditioner can also be applied to systems defined by analytic functions of Toeplitz matrices, i.e. systems of the form $h(T_n)\mathbf{x}_n = \mathbf{b}_n$, where $h(z)$ is an analytic function and T_n is the Toeplitz matrix generated by a continuous complex-valued function f defined on $[-\pi,\pi]$. Based on these ideas, we show that one can solve the symmetrized system $Y_n h(T_n)\mathbf{x}_n = Y_n\mathbf{b}_n$ using MINRES with guaranteed convergence that depends only the eigenvalues of $Y_n h(T_n)$.

In particular, we show that the eigenvalues of $|h(c(T_n))|^{-1}Y_n h(T_n)$ are clustered around ±1 under certain assumptions, where $c(T_n)$ is the optimal circulant preconditioner for T_n. As for a general non-Hermitian $h(T_n)$, we also provide similar results for its normal equation system with the preconditioner $|h(c(T_n))|$.

Given a circulant matrix C_n, we remark that $|h(C_n)|$ and $h(C_n)$ are circulant matrices. By the diagonalization of circulant matrices $C_n = F_n^*\Lambda_n F_n$, we have $|h(C_n)| = F_n^*|h(\Lambda_n)|F_n$ and $h(C_n) = F_n^* h(\Lambda_n)F_n$. Therefore, for any vector $\mathbf{d}_n$ the products $|h(C_n)|^{-1}\mathbf{d}_n$ and $h(C_n)^{-1}\mathbf{d}_n$ can be efficiently computed by several FFTs in $\mathcal{O}(n\log n)$ operations.

It must be noted however that fast matrix vector multiplication with the matrix $h(T_n)$ is not readily achieved by circulant embedding, in contrast to the simplest case $h(z) = z$, though sparsity may still help. Indeed, for e^{T_n} the matrix-vector product can be computed efficiently for example by a fast algorithm in [107, 102].

4.1 Preliminaries on $c(T_n)$ and T_n

We first provide some lemmas concerning Toeplitz matrices.

Lemma 4.1.1 *[38, Lemmas 1 and 3] Let $f \in \mathcal{C}[-\pi,\pi]$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. Then,*

$$\begin{aligned} \|T_n\|_2 &\leq 2\|f\|_\infty \\ \textit{and} \quad \|c(T_n)\|_2 &\leq 2\|f\|_\infty, \qquad n = 1,2,\ldots. \end{aligned}$$

Lemma 4.1.1 states that the 2-norm of the circulant matrix and that of the Toeplitz matrix generated by f are bounded by a constant independent of n.

Theorem 4.1.2 *[38, Theorem 1] Let $f \in \mathcal{C}[-\pi,\pi]$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for*

T_n. Then, for all $\epsilon > 0$ there exist positive integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$

$$c(T_n) - T_n = V_n + W_n,$$

where

$$\mathit{rank}(V_n) \leq 2M^{(\epsilon)}$$

and

$$\|W_n\|_2 \leq \epsilon.$$

Proof This proof is adapted from Theorem 1 in [38].

Suppose $f \in \mathcal{C}[-\pi, \pi]$. Let $T_n[f] \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $C_n[f] \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for $T_n[f]$.

By the Weierstrass approximation theorem, for any $\epsilon > 0$ there exists a trigonometric polynomial

$$p_M(x) = \sum_{k=-M}^{M} \rho_k e^{\mathbf{i}kx}$$

such that

$$\|f - p_M\|_\infty \leq \epsilon.$$

For all $n > 2M$, we decompose

$$\begin{aligned} C_n[f] - T_n[f] &= C_n[f] - C_n[p_M] + C_n[p_M] - T_n[p_M] + T_n[p_M] - T_n[f] \\ &= C_n[f - p_M] - T_n[f - p_M] + C_n[p_M] - T_n[p_M] \\ &= C_n[f - p_M] - T_n[f - p_M] + V_n + \overline{W}_n, \end{aligned}$$

where both

$$V_n = \begin{bmatrix} & & & \frac{n-M}{n}\rho_M & \cdots & \frac{n-1}{n}\rho_1 \\ & & & & \ddots & \vdots \\ & & & & & \frac{n-M}{n}\rho_M \\ \frac{n-M}{n}\rho_{-M} & & & & & \\ \vdots & \ddots & & & & \\ \frac{n-1}{n}\rho_{-1} & \cdots & \frac{n-M}{n}\rho_{-M} & & & \end{bmatrix} \tag{4.1}$$

and

$$\overline{W}_n = -\begin{bmatrix} 0 & \frac{1}{n}\rho_{-1} & \cdots & \frac{M}{n}\rho_{-M} & & & & \\ \frac{1}{n}\rho_1 & \ddots & \ddots & \ddots & \ddots & & & \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ \frac{M}{n}\rho_M & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \frac{M}{n}\rho_{-M} \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & \ddots & \ddots & \ddots & \ddots & \frac{1}{n}\rho_{-1} \\ & & & & \frac{M}{n}\rho_M & \cdots & \frac{1}{n}\rho_1 & 0 \end{bmatrix}$$

are Toeplitz matrices.

We can see from (4.1) that

$$\operatorname{rank}(V_n) \leq 2M.$$

By Lemma 4.1.1, we have

$$\begin{aligned}\|C_n[f-p_M]-T_n[f-p_M]\|_2 &\leq \|C_n[f-p_M]\|_2+\|T_n[f-p_M]\|_2\\ &\leq 2\|f-p_M\|_\infty+2\|f-p_M\|_\infty\\ &\leq 4\epsilon. \end{aligned} \tag{4.2}$$

We now estimate $\|\overline{W}_n\|_2$. For all $|k| \leq M$, we note that

$$\begin{aligned}|\rho_k| &= \left|\frac{1}{2\pi}\int_{-\pi}^{\pi} p_M(x)e^{-\mathbf{i}kx}\,dx\right|\\ &\leq \left|\frac{1}{2\pi}\int_{-\pi}^{\pi}(p_M(x)-f(x))e^{-\mathbf{i}kx}\,dx\right|+\left|\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-\mathbf{i}kx}\,dx\right|\\ &\leq \|f-p_M\|_\infty+\|f\|_\infty\\ &\leq \epsilon+\|f\|_\infty.\end{aligned}$$

By the definition of W_n, we have

$$\begin{aligned}\|\overline{W}_n\|_\infty &= \|\overline{W}_n\|_1\\ &= \frac{M}{n}|\rho_{-M}|+\cdots+\frac{2}{n}|\rho_{-2}|+\frac{1}{n}|\rho_{-1}|\\ &\quad +\frac{1}{n}|\rho_1|+\frac{2}{n}|\rho_2|+\cdots+\frac{M}{n}|\rho_M|\\ &\leq \frac{2}{n}(1+2+\cdots+M)(\epsilon+\|f\|_\infty)\\ &= \frac{1}{n}M(1+M)(\epsilon+\|f\|_\infty).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\|\overline{W}_n\|_2 &\leq (\|\overline{W}_n\|_\infty\|\overline{W}_n\|_1)^{1/2}\\ &\leq \frac{1}{n}M(M+1)(\epsilon+\|f\|_\infty).\end{aligned}$$

Thus, if we pick

$$\begin{aligned}N_M^{(\epsilon)} &:= \max\left\{M(M+1)(1+\frac{\|f\|_\infty}{\epsilon}), 2M\right\}\\ &= M(M+1)(1+\frac{\|f\|_\infty}{\epsilon}),\end{aligned}$$

we get

$$\|\overline{W}_n\|_2 \leq \epsilon. \tag{4.3}$$

We therefore conclude that for all $n \geq N_M^{(\epsilon)}$, there exist positive integers $M > 0$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$

$$\underbrace{C_n[f]}_{c(T_n)} - \underbrace{T_n[f]}_{T_n} = V_n + \underbrace{C_n[f - p_M] - T_n[f - p_M] + \overline{W}_n}_{W_n},$$

where

$$\text{rank}(V_n) \leq 2 \underbrace{M}_{M^{(\epsilon)}}$$

and, by (4.2) and (4.3),

$$\begin{aligned} \|W_n\|_2 &= \|C_n[f - p_M] - T_n[f - p_M] - \overline{W}_n\|_2 \\ &\leq \|C_n[f - p_M] - T_n[f - p_M]\|_2 + \|\overline{W}_n\|_2 \\ &\leq 5\epsilon. \end{aligned}$$

The result follows. ∎

Theorem 4.1.2 indicates that the difference between the circulant matrix and the Toeplitz matrix generated by f can be decomposed into the sum of a low rank matrix and a small norm matrix for sufficiently large n. In the next section, this theorem will be used to prove that a similar decomposition holds for Toeplitz matrix functions.

4.2 Main results

In this section, we show that the preconditioned matrix

$$|h(c(T_n))|^{-1} h(T_n)$$

can be decomposed into the sum of a unitary matrix, a low rank matrix, and a small norm matrix when n is sufficiently large under certain assumptions.

Without loss of generality, we assume that $h(z)$ is represented by the following Taylor series:

$$h(z) = \sum_{k=0}^{\infty} a_k z^k.$$

Theorem 4.2.1 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let $f \in \mathcal{C}[-\pi, \pi]$ with $2\|f\|_\infty < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. Then, for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$h(c(T_n)) - h(T_n) = R_n + E_n,$$

where

$$\mathit{rank}(R_n) \le 2M^{(\epsilon)}$$

and

$$\|E_n\|_2 \le \epsilon.$$

Proof Since $h(z)$ is analytic on $|z| < r$ by assumption, we can write it as the following Taylor series representation:

$$h(z) = \sum_{k=0}^{\infty} a_k z^k$$

with radius of convergence

$$r = \left(\lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| \right)^{-1}.$$

By the assumption that $2\|f\|_\infty < r$ and Lemma 4.1.1, we have

$$\begin{aligned} r &> 2\|f\|_\infty \\ &> \|T_n\|_2 \\ &\ge \rho(T_n) \\ &= \max_j |\lambda_j(T_n)| \\ &\ge |\lambda_j(T_n)|, \qquad j = 1, 2, \ldots, n, \end{aligned}$$

where $\rho(T_n)$ denotes the spectral radius of T_n and $\lambda_j(T_n)$ denotes the j-th eigenvalue of T_n. Therefore, by Lemma 2.5.1, we have

$$h(T_n) = \sum_{k=0}^{\infty} a_k T_n^k.$$

Similarly, we have

$$h(c(T_n)) = \sum_{k=0}^{\infty} a_k c(T_n)^k.$$

We now decompose

$$\begin{aligned} & h(c(T_n)) - h(T_n) \\ = \ & \underbrace{h(c(T_n)) - \sum_{k=0}^{K} a_k c(T_n)^k}_{\Delta_n^{(1)}} + \underbrace{\sum_{k=0}^{K} a_k c(T_n)^k - \sum_{k=0}^{K} a_k T_n^k}_{\Phi_n} + \underbrace{\sum_{k=0}^{K} a_k T_n^k - h(T_n)}_{\Delta_n^{(2)}}. \end{aligned}$$

We first obtain an upper bound for $\|\Delta_n^{(1)} + \Delta_n^{(2)}\|_2$. Using Theorem 2.5.3, we have

$$\begin{aligned}
& \|\Delta_n^{(1)} + \Delta_n^{(2)}\|_2 \\
\leq\ & \|\Delta_n^{(1)}\|_2 + \|\Delta_n^{(2)}\|_2 \\
=\ & \left\|h(c(T_n)) - \sum_{k=0}^{K} a_k c(T_n)^k\right\|_2 + \left\|h(T_n) - \sum_{k=0}^{K} a_k T_n^k\right\|_2 \\
\leq\ & \frac{\max_{0\leq t\leq 1} \|c(T_n)^{K+1} h^{(k+1)}(tc(T_n))\|_2}{(K+1)!} + \frac{\max_{0\leq t\leq 1} \|T_n^{K+1} h^{(k+1)}(tT_n)\|_2}{(K+1)!} \\
\leq\ & \frac{\|c(T_n)^{K+1}\|_2}{(K+1)!} \max_{0\leq t\leq 1} \|h^{(k+1)}(tc(T_n))\|_2 + \frac{\|T_n^{K+1}\|_2}{(K+1)!} \max_{0\leq t\leq 1} \|h^{(k+1)}(tT_n)\|_2.
\end{aligned}$$

By Lemma 4.1.1,

$$\begin{aligned}
\max_{0\leq t\leq 1} \|h^{(k+1)}(tT_n)\|_2 &= \max_{0\leq t\leq 1} \left\| \sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} a_{K+k+1} (tT_n)^k \right\|_2 \\
&\leq \max_{0\leq t\leq 1} \sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| \|tT_n\|_2^k \\
&\leq \sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| \|T_n\|_2^k \\
&\leq \sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| (2\|f\|_\infty)^k.
\end{aligned}$$

We now show that $\sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| (2\|f\|_\infty)^k$ is a convergent series using the ratio test. By the assumption that $2\|f\|_\infty < r = (\lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right|)^{-1}$, we have

$$\begin{aligned}
\lim_{k\to\infty} \left|\frac{a_{K+k+2}}{a_{K+k+1}}\right| \left(\frac{K+k+2}{k+1}\right)(2\|f\|_\infty) &= \lim_{k\to\infty} \left|\frac{a_{k+1}}{a_k}\right| (2\|f\|_\infty) \\
&< \left(\frac{1}{r}\right)(2\|f\|_\infty)
\end{aligned}$$

which is a constant less than 1 because $2\|f\|_\infty < r$.

Therefore, by the ratio test,

$$\sum_{k=0}^{\infty} \frac{(K+k+1)!}{k!} |a_{K+k+1}| (2\|f\|_\infty)^k =: m_{(2\|f\|_\infty)}$$

is convergent. We conclude that $m_{(2\|f\|_\infty)}$ is independent of n.

Hence,

$$\max_{0\leq t\leq 1} \|h^{(k+1)}(tT_n)\|_2 \leq m_{(2\|f\|_\infty)}.$$

Similarly,

$$\max_{0\leq t\leq 1} \|h^{(k+1)}(tc(T_n))\|_2 \leq m_{(2\|f\|_\infty)}.$$

Therefore,

$$\begin{aligned}
&\|\Delta_n^{(1)} + \Delta_n^{(2)}\|_2 \\
\le\ & \frac{\|c(T_n)^{K+1}\|_2}{(K+1)!} \max_{0\le t\le 1} \|h^{(k+1)}(tc(T_n))\|_2 + \frac{\|T_n^{K+1}\|_2}{(K+1)!} \max_{0\le t\le 1} \|h^{(k+1)}(tT_n)\|_2 \\
\le\ & \frac{\|c(T_n)\|_2^{K+1}}{(K+1)!} m_{(2\|f\|_\infty)} + \frac{\|T_n\|_2^{K+1}}{(K+1)!} m_{(2\|f\|_\infty)} \\
\le\ & \frac{(2\|f\|_\infty)^{K+1}}{(K+1)!} m_{(2\|f\|_\infty)} + \frac{(2\|f\|_\infty)^{K+1}}{(K+1)!} m_{(2\|f\|_\infty)} \\
=\ & \frac{(2\|f\|_\infty)^{K+1}}{(K+1)!} (2m_{(2\|f\|_\infty)}) =: \epsilon_K
\end{aligned}$$

which converges to zero as K goes to infinity. Therefore, for a given $\epsilon_K > 0$, there exists an integer K such that for all $k > K$,

$$\|\Delta_n^{(1)} + \Delta_n^{(2)}\|_2 \le \epsilon_K \le \epsilon. \tag{4.4}$$

We next show that Φ_n can be decomposed into the sum of a fixed rank matrix and a small norm matrix. First, by Theorem 4.1.2, for all $\epsilon > 0$ there exist integers N_1 and $M_1 > 0$ such that for all $n > N_1$,

$$c(T_n) - T_n = V_n + W_n,$$

where

$$V_n = \begin{bmatrix} & & & & \Diamond & \cdots & \Diamond \\ & & & & & \ddots & \vdots \\ & & & & & & \Diamond \\ \Diamond & & & & & & \\ \vdots & \ddots & & & & & \\ \Diamond & \cdots & \Diamond & & & & \end{bmatrix}$$

with diamonds representing nonzero entries by (4.1),

$$\operatorname{rank}(V_n) \le 2M_1,$$

and

$$\|W_n\|_2 \le \epsilon.$$

We then decompose Φ_n into

$$\begin{aligned}
\Phi_n &= \sum_{k=0}^{K} a_k c(T_n)^k - \sum_{k=0}^{K} a_k T_n^k \\
&= \sum_{k=1}^{K} a_k (c(T_n)^k - T_n^k) \\
&= \sum_{k=1}^{K} a_k (\sum_{j=0}^{k-1} c(T_n)^j (c(T_n) - T_n) T_n^{k-1-j}) \\
&= \sum_{k=1}^{K} a_k (\sum_{j=0}^{k-1} c(T_n)^j (V_n + W_n) T_n^{k-1-j}) \\
&= \underbrace{\sum_{k=1}^{K} a_k (\sum_{j=0}^{k-1} c(T_n)^j V_n T_n^{k-1-j})}_{R_n} + \underbrace{\sum_{k=1}^{K} a_k (\sum_{j=0}^{k-1} c(T_n)^j W_n T_n^{k-1-j})}_{\Delta_n^{(3)}}.
\end{aligned}$$

Using Lemma 4.1.1, we can estimate the norm of $\Delta_n^{(3)}$:

$$\begin{aligned}
\|\Delta_n^{(3)}\|_2 &= \Big\| \sum_{k=1}^{K} a_k \sum_{j=0}^{k-1} c(T_n)^j W_n T_n^{k-1-j} \Big\|_2 \\
&\leq \sum_{k=1}^{K} |a_k| \Big\| \sum_{j=0}^{k-1} c(T_n)^j W_n T_n^{k-1-j} \Big\|_2 \\
&\leq \|W_n\|_2 \sum_{k=1}^{K} |a_k| \sum_{j=0}^{k-1} \|c(T_n)\|_2^j \|T_n\|_2^{k-1-j} \\
&\leq \|W_n\|_2 \sum_{k=1}^{K} |a_k| \sum_{j=0}^{k-1} (2\|f\|_\infty)^j (2\|f\|_\infty)^{k-1-j} \\
&= \|W_n\|_2 \underbrace{\sum_{k=1}^{K} |a_k| \sum_{j=0}^{k-1} (2\|f\|_\infty)^{k-1}}_{m_K} \\
&\leq m_K \epsilon, \qquad (4.5)
\end{aligned}$$

where m_K is a constant independent of n.

We now estimate the rank of R_n by investigating its sparsity structure. Simple computations similar to those in the proof of [120, Lemma 3.11] yield the following provided that n is sufficiently large

$$c(T_n)^\alpha V_n T_n^\beta \approx \begin{bmatrix} \Diamond & \cdots & \Diamond & & \Diamond & \cdots & \Diamond \\ \vdots & \Diamond & \vdots & & \vdots & \Diamond & \vdots \\ \Diamond & \cdots & \Diamond & & \Diamond & \cdots & \Diamond \\ & & & & & & \\ & & & & & & \\ \Diamond & \cdots & \Diamond & & \Diamond & \cdots & \Diamond \\ \vdots & \Diamond & \vdots & & \vdots & \Diamond & \vdots \\ \Diamond & \cdots & \Diamond & & \Diamond & \cdots & \Diamond \end{bmatrix},$$

where the rhombuses represent the nonzero entries and appear only in the four $(\alpha+1)M_1$ by $(\beta+1)M_1$ blocks in the corners, provided that n is larger than $2\max(\alpha+1,\beta+1)M_1$. Since the rank of

$$R_n = \sum_{k=1}^{K} a_i(\sum_{j=0}^{k-1} c(T_n)^j V_n T_n^{k-1-j})$$

is determined by that of $\sum_{j=0}^{K-1} c(T_n)^j V_n T_n^{K-1-j}$, which is a block matrix with only four nonzero KM_1 by KM_1 blocks in its corners, it follows that

$$\text{rank} R_n \le 2 \underbrace{KM_1}_{M^{(\epsilon)}}$$

if we assume $n > 2M^{(\epsilon)}$.

Hence, we pick

$$N_M^{(\epsilon)} := \max\{N_1, 2M^{(\epsilon)}\}$$

and, combining (4.4) and (4.5), it follows that for all $n > N_M^{(\epsilon)}$

$$\| \underbrace{\Delta_n^{(1)} + \Delta_n^{(2)} + \Delta_n^{(3)}}_{E_n} \|_2 \le (m_K + 1)\epsilon.$$

The result follows. ∎

From Theorem 4.2.1, we can derive the following corollaries on the clustered spectra of $|h(c(T_n))|^{-1}h(T_n)$, depending the positive definiteness of $h(T_n)$.

Corollary 4.2.2 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let $f \in \mathcal{C}[-\pi,\pi]$ with $2\|f\|_\infty < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. If $\|h(c(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$|h(c(T_n))|^{-1}h(T_n) = Q_n + \widetilde{R}_n + \widetilde{E}_n,$$

where Q_n is unitary,

$$rank(\widetilde{R}_n) \le 2M^{(\epsilon)},$$

and

$$\|\widetilde{E}_n\|_2 \le \epsilon.$$

Proof As $h(c(T_n))$ is a circulant matrix, we write $h(c(T_n)) = U_n^* h(\Lambda_n) U_n$ where $h(\Lambda_n)$ is the diagonal matrix in the eigendecomposition of $h(c(T_n))$. We then have

$$\begin{aligned} |h(c(T_n))| &= F_n^* |h(\Lambda_n)| F_n \\ &= F_n^* h(\Lambda_n) F_n (\underbrace{F_n^* \widetilde{h(\Lambda_n)} F_n}_{Q_n})^{-1} \\ &= h(c(T_n)) Q_n^{-1}, \end{aligned} \tag{4.6}$$

where $\widetilde{h(\Lambda_n)} = \text{diag}\big(\frac{h(\Lambda_i)}{|h(\Lambda_i)|}\big)$, and Q_n is unitary.

By Theorem 4.2.1, for all $\epsilon > 0$, there exist positive integers N_M and M such that for all $n > N_M$

$$h(c(T_n)) - h(T_n) = R_n + E_n,$$

where

$$\text{rank}(R_n) \le 2M$$

and

$$\|E_n\|_2 \le \epsilon.$$

We have

$$\begin{aligned} h(c(T_n))^{-1} h(T_n) &= I_n + h(c(T_n))^{-1}(h(T_n) - h(c(T_n))) \\ &= I_n + h(c(T_n))^{-1}(-R_n) + h(c(T_n))^{-1}(-E_n). \end{aligned}$$

Further using (4.6),

$$\begin{aligned} |h(c(T_n))|^{-1} h(T_n) &= Q_n h(c(T_n))^{-1} h(T_n) \\ &= Q_n + \underbrace{Q_n h(c(T_n))^{-1}(-R_n)}_{\widetilde{R}_n} + \underbrace{Q_n h(c(T_n))^{-1}(-E_n)}_{\widetilde{E}_n}. \end{aligned}$$

Since Q_n is unitary, we know

$$\begin{aligned} \text{rank}(\widetilde{R}_n) &= \text{rank}(Q_n h(c(T_n))^{-1} R_n) \\ &= \text{rank}(R_n) \\ &\le 2M \end{aligned}$$

and

$$\begin{aligned} \|\widetilde{E}_n\|_2 &= \|Q_n h(c(T_n))^{-1}E_n\|_2 \\ &= \|h(c(T_n))^{-1}E_n\|_2 \\ &\leq m_0\epsilon, \end{aligned}$$

where m_0 is a constant independent of n, resulting from the uniform boundedness of $\|h(c(T_n))^{-1}\|_2$. The result follows. ∎

Assuming f is a real-valued, we know by Theorem 2.5.2 (a) that $h(T_n)$ is Hermitian when T_n is Hermitian. Similarly, $h(c(T_n))$ is Hermitian when $c(T_n)$ is Hermitian. We can now show that the eigenvalues of $|h(c(T_n))|^{-1}h(T_n)$ are clustered around ±1 using Corollary 4.2.2. Note however that both $\widetilde{R}_n$ and $\widetilde{E}_n$ in the corollary are non-Hermitian in general. Besides, one must deal with the unitary matrix Q_n instead of the usual identity matrix in the matrix decomposition. Therefore, Cauchy's interlacing theorem that was used for example in [25] to show clustered spectra does not straightforwardly apply. Nevertheless, we are still able to show the clustered spectra of our concerned preconditioned matrices via a simple trick.

Corollary 4.2.3 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let $f \in \mathcal{C}[-\pi,\pi]$ with $2\|f\|_\infty < r$ be real-valued. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. If $\|h(c(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then $|h(c(T_n))|^{-1}h(T_n)$ has clustered spectra around ±1 for sufficiently large n.*

Proof By Corollary 4.2.2, for all $\epsilon > 0$ there exist positive integers N and M such that for all $n > N$

$$\begin{aligned} &\underbrace{|h(c(T_n))|^{-1/2}h(T_n)|h(c(T_n))|^{-1/2}}_{H_n} \\ =\ &\underbrace{|h(c(T_n))|^{1/2}Q_n|h(c(T_n))|^{-1/2}}_{\mathcal{Q}_n} + \underbrace{|h(c(T_n))|^{1/2}\widetilde{R}_n|h(c(T_n))|^{-1/2}}_{\mathcal{R}_n} + \underbrace{|h(c(T_n))|^{1/2}\widetilde{E}_n|h(c(T_n))|^{-1/2}}_{\mathcal{E}_n}, \end{aligned}$$

where $\mathcal{Q}_n$ is similar to Q_n,

$$\operatorname{rank}(\mathcal{R}_n) \leq 2M,$$

and

$$\|\mathcal{E}_n\|_2 \leq \epsilon.$$

We introduce the following matrix decomposition

$$\underbrace{\begin{bmatrix} & H_n \\ H_n^* & \end{bmatrix}}_{\mathcal{H}} = \underbrace{\begin{bmatrix} & \mathcal{Q}_n \\ \mathcal{Q}_n^* & \end{bmatrix}}_{\mathcal{Q}} + \underbrace{\begin{bmatrix} & \mathcal{R}_n \\ \mathcal{R}_n^* & \end{bmatrix}}_{\mathcal{R}} + \underbrace{\begin{bmatrix} & \mathcal{E}_n \\ \mathcal{E}_n^* & \end{bmatrix}}_{\mathcal{E}},$$

where $\mathcal{Q}$ has only eigenvalues ± 1,

$$\text{rank}(\mathcal{R}) \leq 4M,$$

and

$$\text{rank}(\mathcal{E}) \leq \epsilon.$$

Note that all $\mathcal{H}$, $\mathcal{Q}$, $\mathcal{R}$, and $\mathcal{E}$ are Hermitian. By [15, Corollary 3] (Theorem B.2.3), we know that there are at most $2(4M) = 8M$ eigenvalues of $\mathcal{H}$ that are not around ± 1. Thus, $\mathcal{H}$ has clustered spectra around ± 1 by Definition 2.3.2. As the eigenvalues of $\mathcal{H}$ are the same as the singular values of H_n up to ± sign, the singular values of H_n are clustered around 1. Consequently, as H_n is Hermitian and is similar to $|h(c(T_n))|^{-1}h(T_n)$, we conclude that $|h(c(T_n))|^{-1}h(T_n)$ has clustered spectra around ± 1. ∎

As a consequence of Corollary 4.2.3, we have the following two cases.

(i) If $h(T_n)$ is Hermitian indefinite, MINRES together with a Hermitian positive definite preconditioner $|h(c(T_n))|$ should be used [168, Section 5].

(ii) If both $h(T_n)$ and $h(c(T_n))$ are Hermitian positive definite, i.e. $|h(c(T_n))|$ reduces to $h(c(T_n))$, CG can be applied and the eigenvalues of the preconditioned matrix are clustered around unity in this case.

We will illustrate these two cases in the next section where the trigonometric functions e^z, $\sin z$, and $\cos z$ are under consideration.

Remark If both $h(T_n)$ and $h(c(T_n))$ are Hermitian negative definite, the situation is similar to Case (ii) since we can premultiply the matrices by $-I_n$ to make them positive definite.

In the general case where $h(T_n)$ is non-Hermitian, we consider its normal equations system via the follow corollaries.

Corollary 4.2.4 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let $f \in \mathcal{C}[-\pi, \pi]$ with $2\|f\|_\infty < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. If*

$\||h(c(T_n))|^{-1}\|_2$ *is uniformly bounded with respect to* n*, then for all* $\epsilon > 0$ *there exist integers* $M^{(\epsilon)}$ *and* $N_M^{(\epsilon)}$ *such that for all* $n > N_M^{(\epsilon)}$

$$[|h(c(T_n))|^{-1}h(T_n)]^*|h(c(T_n))|^{-1}h(T_n) = I_n + \overline{R}_n + \overline{E}_n,$$

where

$$rank(\overline{R}_n) \leq 4M^{(\epsilon)}$$

and

$$\|\overline{E}_n\|_2 \leq \epsilon.$$

Proof By Corollary 4.2.2, we know that for all $\epsilon > 0$ there exist positive integers N_M and M such that for all $n > N_M$

$$|h(c(T_n))|^{-1}h(T_n) = Q_n + \widehat{R}_n + \widehat{E}_n,$$

where Q_n is unitary,

$$\text{rank}(\widehat{R}_n) \leq 2M,$$

and

$$\|\widehat{E}_n\|_2 \leq \epsilon.$$

We then have

$$\begin{aligned} & [|h(c(T_n))|^{-1}h(T_n)]^*|h(c(T_n))|^{-1}h(T_n) \\ = & (Q_n + \widehat{R}_n + \widehat{E}_n)^*(Q_n + \widehat{R}_n + \widehat{E}_n) \\ = & \underbrace{Q_n^* Q_n}_{I_n} + \underbrace{\widehat{R}_n^*(I_n + \widehat{R}_n + \widehat{E}_n) + (I_n + \widehat{E}_n^*)\widehat{R}_n}_{\overline{R}_n} \\ & + \underbrace{\widehat{E}_n + \widehat{E}_n^* + \widehat{E}_n^*\widehat{E}_n}_{\overline{E}_n} \\ = & I_n + \overline{R}_n + \overline{E}_n. \end{aligned}$$

It immediately follows that

$$\text{rank}(\overline{R}_n) \leq 4M$$

and

$$\|\overline{E}_n\|_2 \leq \epsilon^2 + 2\epsilon. \quad \blacksquare$$

Since $[|h(c(T_n))|^{-1}h(T_n)]^*|h(c(T_n))|^{-1}h(T_n)$ in Corollary 4.2.4 is Hermitian positive definite, by Cauchy's interlacing theorem we know that its eigenvalues of are mostly close to 1 when n is sufficiently large. In this case, CG can be used.

Corollary 4.2.5 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let $f \in \mathcal{C}[-\pi, \pi]$ with $2\|f\|_\infty < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. If $\|h(c(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then*

$$[|h(c(T_n))|^{-1}h(T_n)]^*|h(c(T_n))|^{-1}h(T_n)$$

has clustered spectra around 1 for sufficiently large n.

In the special case where $h(T_n)$ is (real) nonsymmetric, we show in the following corollaries that it is not necessary to normalize the original matrix in order to have clustered spectra around ±1. Namely, we can symmetrize $h(T_n)$ by premultiplying it with Y_n and then precondition the symmetric matrix $Y_n h(T_n)$ with $|h(c(T_n))|$. As in Case (i) before, we can employ MINRES in this special case.

Corollary 4.2.6 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let $f \in \mathcal{C}[-\pi, \pi]$ have real Fourier coefficients and suppose $2\|f\|_\infty < r$. Let $T_n \in \mathbb{R}^{n\times n}$ be the real Toeplitz matrix generated by f, let $c(T_n) \in \mathbb{R}^{n\times n}$ be the real optimal circulant preconditioner for T_n, and let $Y_n \in \mathbb{R}^{n\times n}$ be the anti-identity matrix. If $\||h(c(T_n))|^{-1}\|_2$ is uniformly bounded with respect to n, then for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$|h(c(T_n))|^{-1}Y_n h(T_n) = Q_n + \widehat{R}_n + \widehat{E}_n,$$

where $Q_n \in \mathbb{R}^{n\times n}$ is orthogonal and symmetric,

$$\mathit{rank}(\widehat{R}_n) \le 2M^{(\epsilon)},$$

and

$$\|\widehat{E}_n\|_2 \le \epsilon.$$

Proof As $h(c(T_n))$ is a circulant matrix we write $h(c(T_n)) = F_n^* h(\Lambda_n) F_n$ where $h(\Lambda_n)$ is the diagonal matrix in the eigendecomposition of $h(c(T_n))$. We then have

$$\begin{aligned} |h(c(T_n))| &= F_n^*|h(\Lambda_n)|F_n \\ &= F_n^* h(\Lambda_n) F_n (\underbrace{F_n^* \widetilde{h(\Lambda_n)} F_n}_{\widetilde{C}_n})^{-1} \\ &= h(c(T_n))\widetilde{C}_n^{-1}, \end{aligned} \tag{4.7}$$

where $\widetilde{h(\Lambda_n)} = \mathrm{diag}(\frac{h(\Lambda_n)_j}{|h(\Lambda_n)_j|})$ with $h(\Lambda_n)_j$ being the j-th eigenvalue of $h(\Lambda_n)$.

By Theorem 4.2.1, we know that for all $\epsilon > 0$, there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$

$$h(c(T_n)) - h(T_n) = R_n + E_n,$$

where

$$\operatorname{rank}(R_n) \leq 2M^{(\epsilon)}$$

and

$$\|E_n\|_2 \leq \epsilon.$$

By (4.7), we have

$$\begin{aligned} |h(c(T_n))|^{-1} Y_n h(T_n) &= Y_n |h(c(T_n))|^{-1} h(T_n) \\ &= Y_n |h(c(T_n))|^{-1} (h(c(T_n)) - R_n - E_n) \\ &= \underbrace{Y_n \widetilde{C}_n}_{Q_n} + \underbrace{Y_n |h(c(T_n))|^{-1}(-R_n)}_{\widehat{R}_n} + \underbrace{Y_n |h(c(T_n))|^{-1}(-E_n)}_{\widehat{E}_n}. \end{aligned}$$

Since $\widetilde{C}_n$ itself is a real Toeplitz matrix, $Q_n = Y_n \widetilde{C}_n$ is symmetric. Thus, we can show that Q_n is orthogonal:

$$\begin{aligned} Q_n^T Q_n &= (Y_n \widetilde{C}_n)^T (Y_n \widetilde{C}_n) \\ &= \widetilde{C}_n^T (Y_n^T Y_n) \widetilde{C}_n \\ &= \widetilde{C}_n^T \widetilde{C}_n \\ &= (F_n^* \widetilde{h(\Lambda_n)} F_n)^T F_n^* \widetilde{h(\Lambda_n)} F_n \\ &= F_n^* \underbrace{|\widetilde{h(\Lambda_n)}|^2}_{I_n} F_n \\ &= I_n. \end{aligned}$$

We have

$$\begin{aligned} \operatorname{rank}(\widehat{R}_n) &= \operatorname{rank}(Y_n |h(c(T_n))|^{-1} R_n) \\ &= \operatorname{rank}(R_n) \\ &\leq 2M^{(\epsilon)} \end{aligned}$$

and

$$\begin{aligned} \|\widehat{E}_n\|_2 &= \|Y_n |h(c(T_n))|^{-1} E_n\|_2 \\ &\leq \|h(c(T_n))^{-1} E_n\|_2 \\ &\leq m_0 \epsilon, \end{aligned}$$

where m_0 is a constant independent of n, resulting from the uniform boundedness assumption on $\|h(c(T_n))^{-1}\|_2$. The result follows. ▮

Corollary 4.2.7 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let $f \in \mathcal{C}[-\pi, \pi]$ have real Fourier coefficients and suppose $2\|f\|_\infty < r$. Let $T_n \in \mathbb{R}^{n\times n}$ be the real Toeplitz matrix generated by f, let $c(T_n) \in \mathbb{R}^{n\times n}$ be the real optimal circulant preconditioner for T_n, and let $Y_n \in \mathbb{R}^{n\times n}$ be the anti-identity matrix. If $\|h(c(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then $|h(c(T_n))|^{-1}Y_nh(T_n)$ has clustered spectra around ± 1 for sufficiently large n.*

4.3 Trigonometric functions of Toeplitz matrices

In this section, we particularly consider the case where the Toeplitz matrix is Hermitian and $h(z) = e^z$, $\sin z$, or $\cos z$ to provide concrete examples of our results.

Note that the convergence radius of these functions is infinity. In other words, the condition that $2\|f\|_\infty < r = \infty$ is readily satisfied.

4.3.1 Preconditioning for e^{T_n}

For $h(z) = e^z$, the assumption that $\||h(c(T_n))|^{-1}\|_2 = \||e^{c(T_n)}|^{-1}\|_2$ is uniformly bounded with respect to n holds and can be shown easily via the following lemmas.

Lemma 4.3.1 *[79, Theorem 10.2] For $A_n, B_n \in \mathbb{C}^{n\times n}$,*

$$e^{(A_n+B_n)t} = e^{A_nt}e^{B_nt}$$

for all t if and only if $A_nB_n = B_nA_n$.

Lemma 4.3.2 *Let $f \in \mathcal{C}[-\pi, \pi]$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. Then,*

$$\|(e^{c(T_n)})^{-1}\|_2 \leq e^{2\|f\|_\infty}, \qquad n = 1, 2, \ldots.$$

Proof By Lemma 4.3.1

$$\begin{aligned} e^{c(T_n)}e^{-c(T_n)} &= e^{c(T_n)-c(T_n)} \\ &= I_n, \end{aligned}$$

$e^{-c(T_n)}$ is the inverse of $e^{c(T_n)}$. Further using Lemma 4.1.1, we have

$$\begin{aligned} \|(e^{c(T_n)})^{-1}\|_2 &= \|e^{-c(T_n)}\|_2 \\ &\leq e^{\|c_n[f]\|_2} \\ &\leq e^{2\|f\|_\infty}. \quad ▮ \end{aligned}$$

Assuming f is real-valued, e^{T_n} is Hermitian for all n and we have the following corollary.

Corollary 4.3.3 *Let $f \in \mathcal{C}[-\pi, \pi]$ be real-valued. Let $T_n \in \mathbb{C}^{n\times n}$ be the Hermitian Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. Then, for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$(e^{c(T_n)})^{-1}e^{T_n} = I_n + \widetilde{R}_n + \widetilde{E}_n,$$

where

$$\mathit{rank}(\widetilde{R}_n) \le 2M^{(\epsilon)}$$

and

$$\|\widetilde{E}_n\|_2 \le \epsilon.$$

Proof Since $c(T_n)$ (or T_n) is Hermitian when f is real-valued, we can write $c(T_n) = Z_n^* D_n Z_n$ where D_n is the diagonal matrix with real eigenvalues d_i being the eigenvalues of $c(T_n)$ and Z_n is a unitary matrix. Consequently, $e^{c(T_n)} = Z_n^* e^{D_n} Z_n$ is positive definite as its eigenvalues are all of the form $e^{d_i} > 0$. Thus,

$$\begin{aligned} |e^{c(T_n)}| &= Z_n^* |e^{D_n}| Z_n \\ &= Z_n^* e^{D_n} Z_n \\ &= e^{c(T_n)}. \end{aligned}$$

By Lemma 4.3.2 that $\|(e^{c(T_n)})^{-1}\|_2$ is uniformly bounded with respect to n and Corollary 4.2.2, we know that for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$

$$(e^{c(T_n)})^{-1}e^{T_n} = I_n + \widetilde{R}_n + \widetilde{E}_n,$$

where

$$\mathrm{rank}(\widetilde{R}_n) \le 2M^{(\epsilon)}$$

and

$$\|\widetilde{E}_n\|_2 \le \epsilon. \quad \blacksquare$$

Since both e^{T_n} and $e^{c(T_n)}$ in this case are Hermitian positive definite, CG can be used.

Corollary 4.3.4 *Let $f \in \mathcal{C}[-\pi, \pi]$ be real-valued. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. Then, $(e^{c(T_n)})^{-1}e^{T_n}$ has clustered spectra around 1 for sufficiently large n.*

4.3.2 Preconditioning for $\sin T_n$ and $\cos T_n$

We provide similar results for $h(z) = \sin z$ (or $\cos z$) in this subsection.

Unlike e^z, $\|(\sin c(T_n))^{-1}\|_2$ is unbounded in general. From

$$\|(\sin c(T_n))^{-1}\|_2 = \max_i \left|\frac{1}{\sin \lambda_i}\right|,$$

we know that $\|(\sin c(T_n))^{-1}\|_2$ could be arbitrarily large since $\sin \lambda_i$ could be close to zero, where λ_i is the i-th eigenvalue of $c(T_n)$. Therefore, the uniform boundedness of $\|(\sin c(T_n))^{-1}\|_2$ is required.

Consider now the case where $\sin T_n$ is Hermitian. Unlike the case with the matrix exponential, we cannot use CG for $\sin T_n$ since it is not positive definite in general. By the diagonalization of $\sin T_n = Z_n^*(\sin D_n)Z_n$, where D_n is the diagonal matrix with real eigenvalues d_i being the eigenvalues of T_n, we see that its eigenvalues are all of the form $-1 \le \sin d_i \le 1$. MINRES together with the Hermitian positive definite preconditioner $|\sin c(T_n)|$ should be used.

Corollary 4.3.5 *Let $f \in \mathcal{C}[-\pi, \pi]$ be real-valued. Let $T_n \in \mathbb{C}^{n\times n}$ be the Hermitian Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. If $\|(\sin c(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$|\sin c(T_n)|^{-1} \sin T_n = Q_n + \widetilde{R}_n + \widetilde{E}_n,$$

where Q_n is Hermitian and unitary,

$$\mathit{rank}(\widetilde{R}_n) \le 2M^{(\epsilon)},$$

and

$$\|\widetilde{E}_n\|_2 \le \epsilon.$$

Corollary 4.3.6 *Let $f \in \mathcal{C}[-\pi, \pi]$ be real-valued. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n. If $\|(\sin c(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then $|\sin c(T_n)|^{-1} \sin T_n$ has clustered spectra around ±1 for sufficiently large n.*

With $\sin z$ replaced by $\cos z$, Corollaries 4.3.5 and 4.3.6 still hold.

4.4 Numerical results

In this section, we demonstrate the effectiveness of our proposed preconditioners using CG, MINRES, and GMRES.

Throughout all numerical tests, e^{T_n} is computed by the MATLAB R2016b built-in function **expm** while $\sin T_n$, $\cos T_n$, and other matrix functions are computed by **funm**. The vector $\mathbf{b}_n$ is generated by **ones(n,1)** and the initial guess is the zero vector. Also, we use **pcg** to solve Hermitian positive definite systems. For Hermitian indefinite systems, we use **minres**. As a comparison, GMRES is also used and is executed by **gmres**. The stopping criterion used is

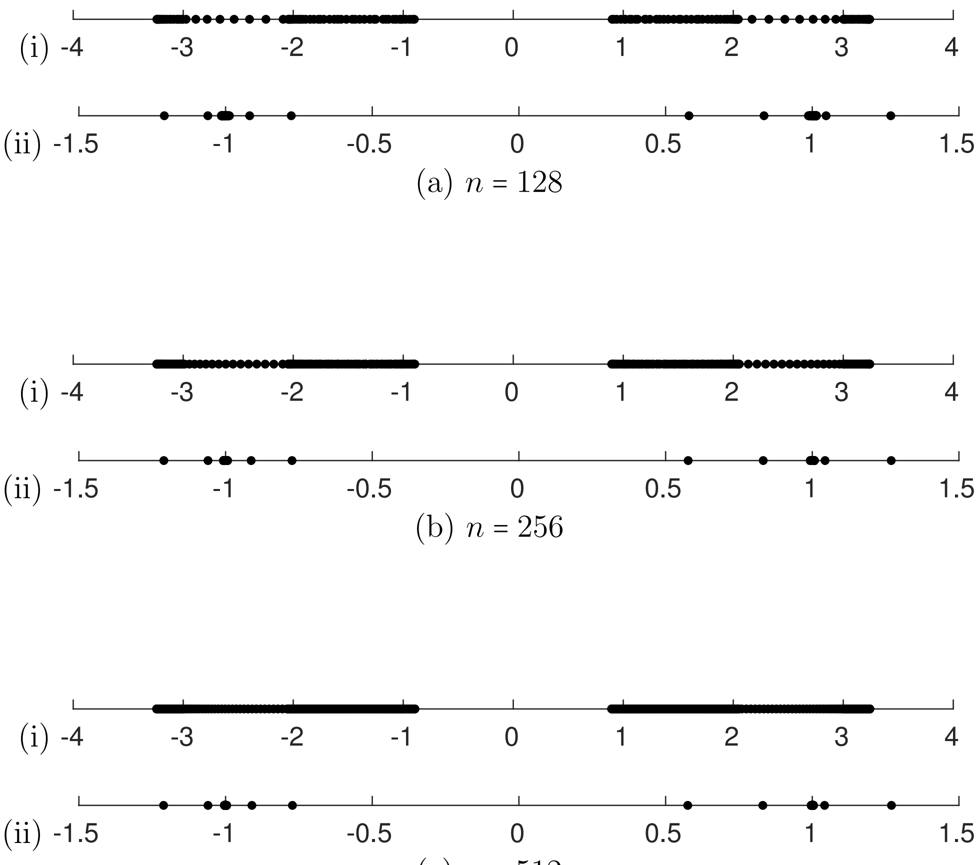
$$\frac{\|\mathbf{r}_n^{(j)}\|_2}{\|\mathbf{b}_n\|_2} < 10^{-7},$$

where $\mathbf{r}_n^{(j)}$ is the residual vector after j iterations.

Example 4.1. We first consider the following Grcar matrix

$$T_n = \begin{bmatrix} 1 & 1 & 1 & 1 & & \\ -1 & \ddots & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & \ddots & 1 \\ & & \ddots & \ddots & \ddots & 1 \\ & & & \ddots & \ddots & 1 \\ & & & & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{4.8}$$

Table 4.1 (a) and (b) show the numerical results for the system using MINRES and GMRES, respectively. The preconditioner appears effective for speeding up the convergence rate for both Krylov subspace methods. Figure 4.1 shows the spectra of Y_nT_n before and after the preconditioner $|c(T_n)|$ is applied at different n. We observe that the spectra are highly clustered around ±1.



(a) n = 128

(b) n = 256

(c) n = 512

Figure 4.1: Spectra of Y_nT_n with T_n given in Example 4.1 at different n (i) with no preconditioner or (ii) with the preconditioner $|c(T_n)|$.

Table 4.1: Numbers of iterations with (a) MINRES for $Y_n T_n$ and (b) GMRES for T_n with T_n given in Example 4.1.

(a)

n	with no preconditioner	with $\lvert c(T_n)\rvert$
128	49	13
256	49	12
512	49	11
1024	47	11

(b)

n	with no preconditioner	with $c(T_n)$
128	94	6
256	158	6
512	218	6
1024	213	5

Example 4.2. We consider the Toeplitz matrix polynomial $h(T_n) \in \mathbb{R}^{n\times n}$, where $h(z) = z^2 + z + 1$ and T_n is the Grcar matrix given by (4.8).

Table 4.2 (a) and (b) show the numbers of iterations for the system. In Figure 4.2 (a) and (b), we show the spectra of $Y_n h(T_n)$ before or after the preconditioner $|h(c(T_n))|$ is applied at different n. Again, we observe a speed-up in convergence and the clusters of eigenvalues around ±1.

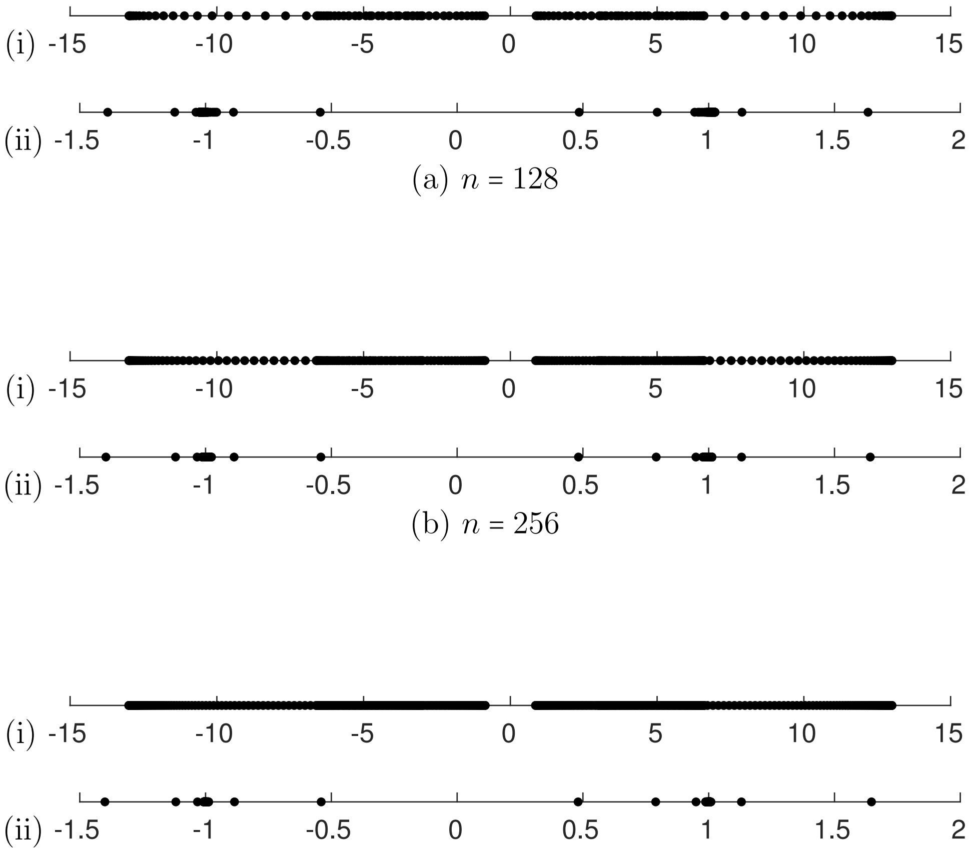
Table 4.2: Numbers of iterations with (a) MINRES for $Y_n h(T_n)$ and (b) GMRES for $h(T_n)$ with $h(T_n)$ given in Example 4.2.

(a)

n	with no preconditioner	with $\lvert h(c(T_n))\rvert$
128	144	16
256	167	15
512	194	14
1024	190	13

(b)

n	with no preconditioner	with $h(c(T_n))$
128	128	9
256	256	8
512	512	8
1024	1024	7



(a) n = 128

(b) n = 256

(c) n = 512

Figure 4.2: Spectra of $Y_n h(T_n)$ with $h(T_n)$ given in Example 4.2 at different n (i) with no preconditioner or (ii) with the preconditioner $|h(c(T_n))|$.

Example 4.3. In this example, the Toeplitz hyperbolic sine function $\sinh T_n \in \mathbb{R}^{n\times n}$ is considered, where T_n is again the Grcar matrix given by (4.8). Note that both $\sinh T_n$ and $Y_n \sinh T_n$ become highly ill-conditioned when $n = 512$. Hence, we only show the numerical results up to $n = 256$.

Table 4.3 (a) and (b) show the numbers of iterations for the system. The convergence rate appears accelerated with our proposed preconditioners. Figure 4.3 shows the expected clusters of eigenvalues when the preconditioner $|\sinh c(T_n)|$ is applied.

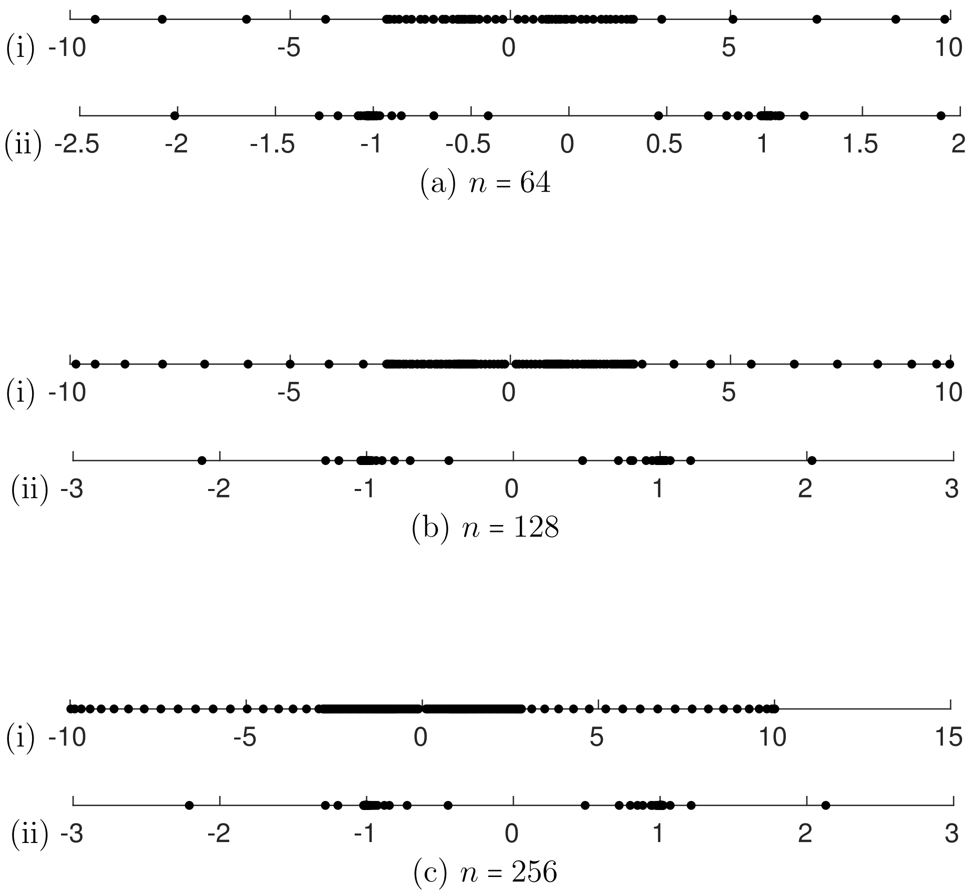
Table 4.3: Numbers of iterations with (a) MINRES for $Y_n \sinh T_n$ and (b) GMRES for $\sinh T_n$ with T_n given in Example 4.3.

(a)

n	with no preconditioner	with $\lvert\sinh c(T_n)\rvert$
32	39	21
64	106	24
128	172	22
256	391	20

(b)

n	with no preconditioner	with $\sinh c(T_n)$
32	32	14
64	64	14
128	124	13
256	240	11



(a) $n = 64$

(b) $n = 128$

(c) $n = 256$

Figure 4.3: Spectra of $Y_n \sinh T_n$ with T_n given in Example 4.3 at different n (i) with no preconditioner or (ii) with the preconditioner $|\sinh c(T_n)|$.

Example 4.4. The next example is the nonsymmetric Toeplitz matrix generated by $f(x) = e^{\mathbf{i}x} + 2e^{-\mathbf{i}x}$, namely

$$T_n = \begin{bmatrix} & 2 & & \\ 1 & & \ddots & \\ & \ddots & & 2 \\ & & 1 & \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{4.9}$$

We remark that both T_n and Y_nT_n in this case are highly ill-conditioned. For example, we have $\kappa(T_{128}) = \kappa(Y_{128}T_{128}) = 3.6884 \times 10^{19}$. Therefore, we consider the numerical results up to $n = 64$.

Table 4.4 shows the numbers of iterations for the system. Note that both MINRES and GMRES fail to converge when $n = 64$ due to the large condition number. In Figure 4.4, we nevertheless still observe the highly clusters of eigenvalues around ± 1 up to $n = 512$.

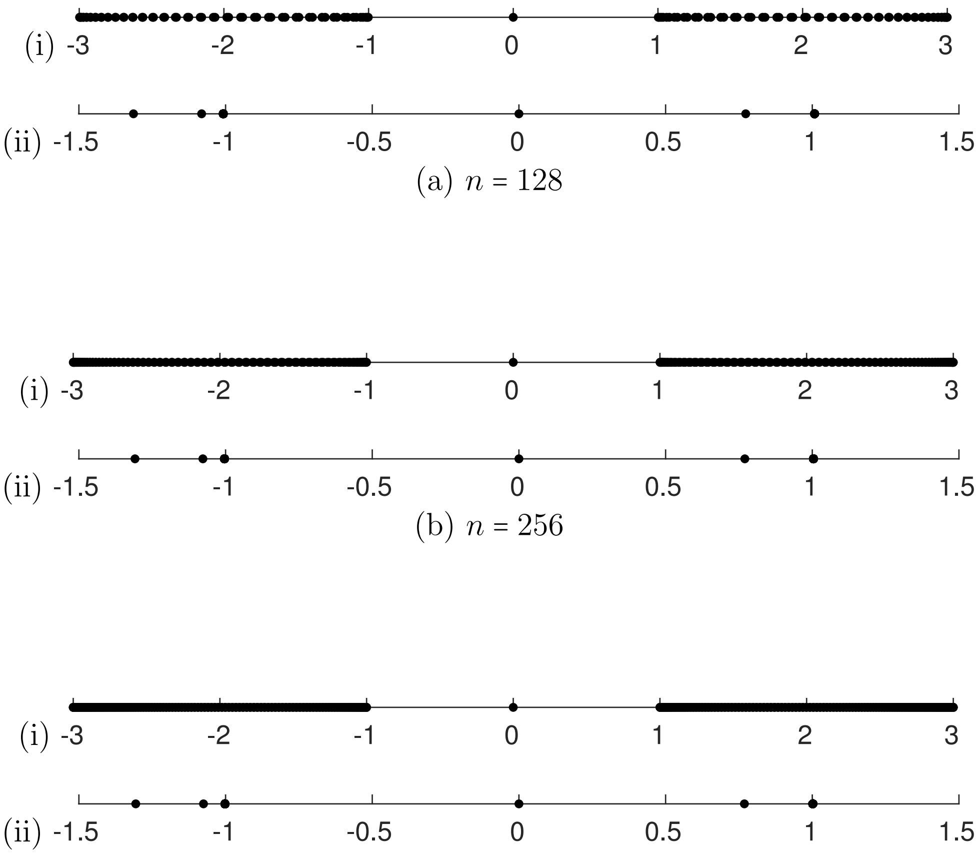
Table 4.4: Numbers of iterations with (a) MINRES for Y_nT_n and (b) GMRES for T_n with T_n given in Example 4.4.

(a)

n	with no preconditioner	with $\lvert c(T_n)\rvert$
8	8	5
16	16	5
32	32	5
64	no convergence	no convergence

(b)

n	with no preconditioner	with $c(T_n)$
8	8	3
16	10	3
32	32	3
64	no convergence	4



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 4.4: Spectra of Y_nT_n with T_n given in Example 4.4 at different n (i) with no preconditioner or (ii) with the preconditioner $|c(T_n)|$.

Example 4.5. We consider the matrix exponential e^{T_n} with the symmetric positive definite T_n generated by $f(x) = x^2$. Note that the symmetrization device Y_n is not used since e^{T_n} is already symmetric in this example.

Table 4.5 shows the numbers of iterations required with CG and GMRES for e^{T_n}. It is apparent that the proposed preconditioners are effective for speeding up the rate of convergence of CG. In Figure 4.5, we further show the spectra of the preconditioned matrices at different n. We observe that the highly clustered spectra independent of n. In Figure 4.5 (i) and (ii), the contrast between the spectra of the matrices is shown. In Figure 4.5 (iii), we show the zoom-in spectrum of (ii) and observe that the eigenvalues are highly clustered around 1. Due to the highly clustered eigenvalues, a fast convergence rate for CG is expected (see for example Section 3.3.1 or [6]).

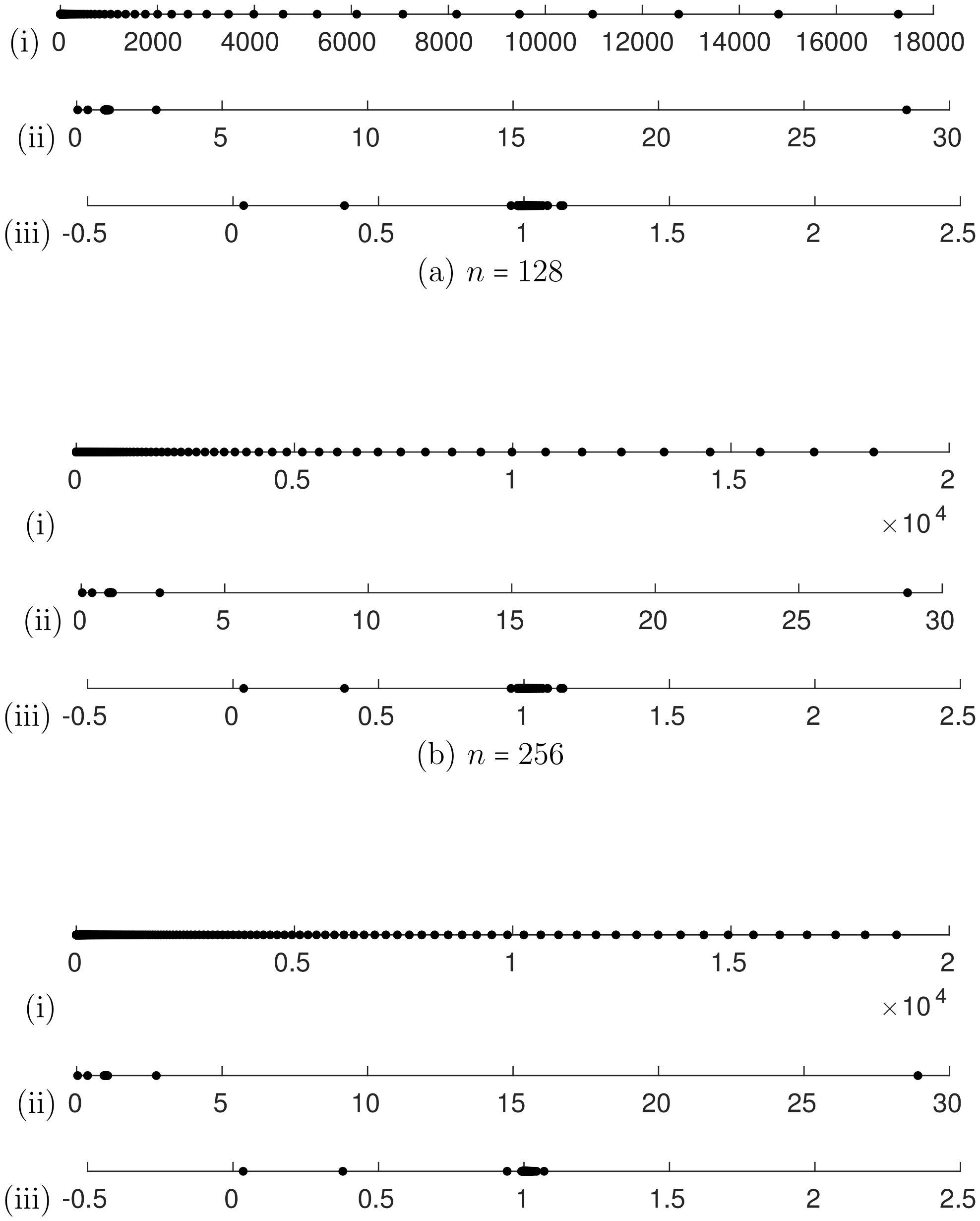
Table 4.5: Numbers of iterations with (a) CG and (b) GMRES for e^{T_n}, with T_n given in Example 4.5.

(a)

n	with no preconditioner	with $e^{c(T_n)}$
128	257	9
256	450	8
512	657	8
1024	833	8

(b)

n	with no preconditioner	with $e^{c(T_n)}$
128	76	6
256	127	5
512	205	5
1024	320	5



(c) n = 512

Figure 4.5: Spectra of e^{T_n} with T_n given in Example 4.5 at different n (i) with no preconditioner or (ii) with the preconditioner $e^{c(T_n)}$. (iii) Zoom-in spectrum of (ii).

Example 4.6. Next, we consider the symmetric indefinite Toeplitz matrix T_n generated by $f(x) = x^2 - \pi$.

Table 4.6 shows the numbers of iterations for the system. In Figure 4.6 (a) and (b), we also show the spectra of T_n before and after the preconditioner $|c(T_n)|$ is used. We again observe rapid convergence and the expected clusters around ±1.

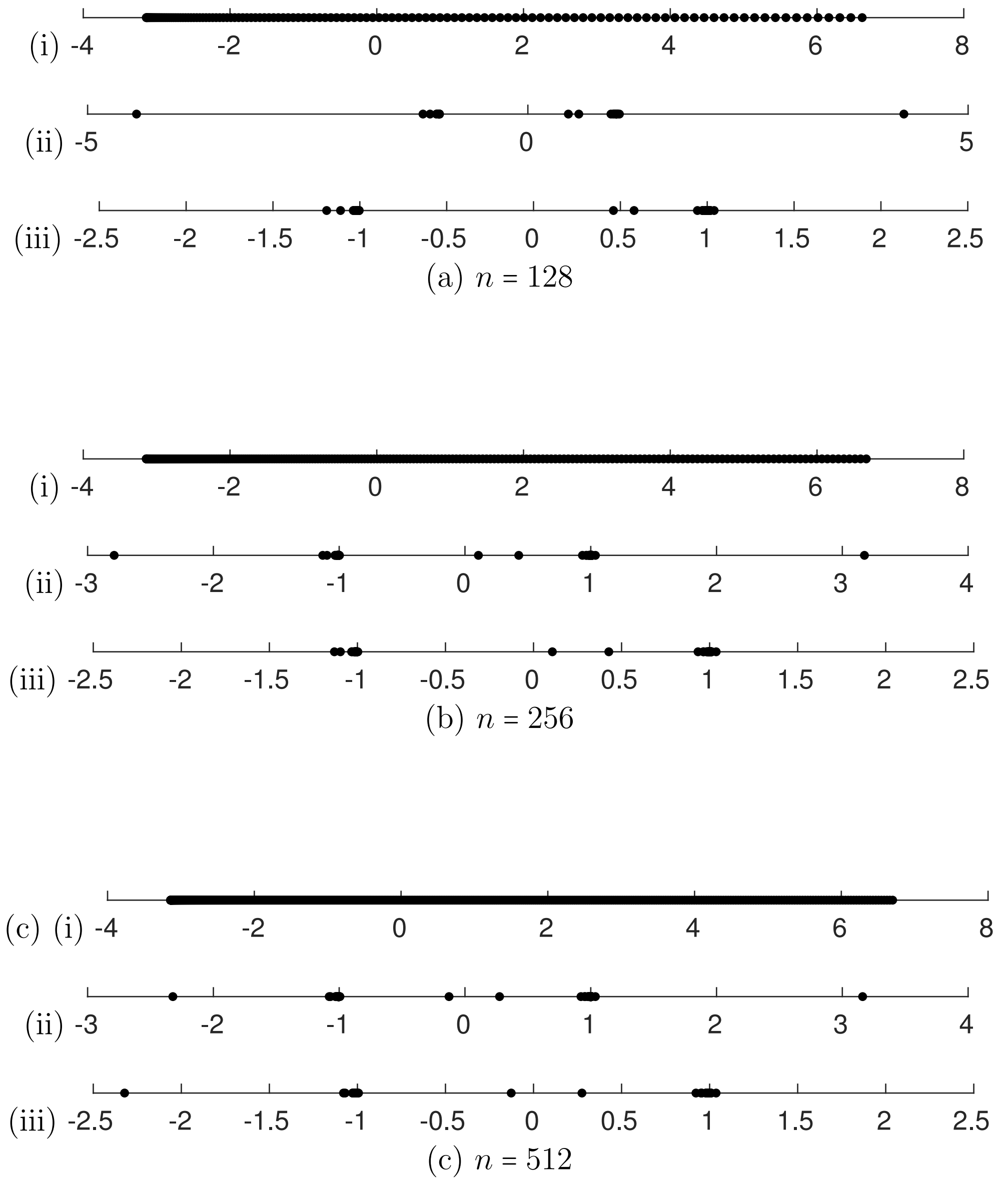
Table 4.6: Numbers of iterations with (a) MINRES and (b) GMRES for T_n given in Example 4.6.

(a)

n	with no preconditioner	with $\lvert c(T_n)\rvert$
128	76	11
256	159	11
512	321	10
1024	648	10

(b)

n	with no preconditioner	with $c(T_n)$
128	74	6
256	150	7
512	299	6
1024	595	6



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 4.6: Spectra of T_n given in Example 4.6 at different n (i) with no preconditioner or (ii) with the preconditioner $|c(T_n)|$. (iii) Zoom-in spectrum of (ii).

Example 4.7. We also consider $\cos T_n$ with the same symmetric indefinite T_n considered in Example 4.6.

Table 4.7 shows the numbers of iterations for the system. Again, we observe a significant reduce in the iteration counts for both MINRES and GMRES with the proposed preconditioners. In Figure 4.7, we also show the spectra of $|\cos c(T_n)|^{-1}\cos T_n$ at different n and observe the expected clusters around ±1.

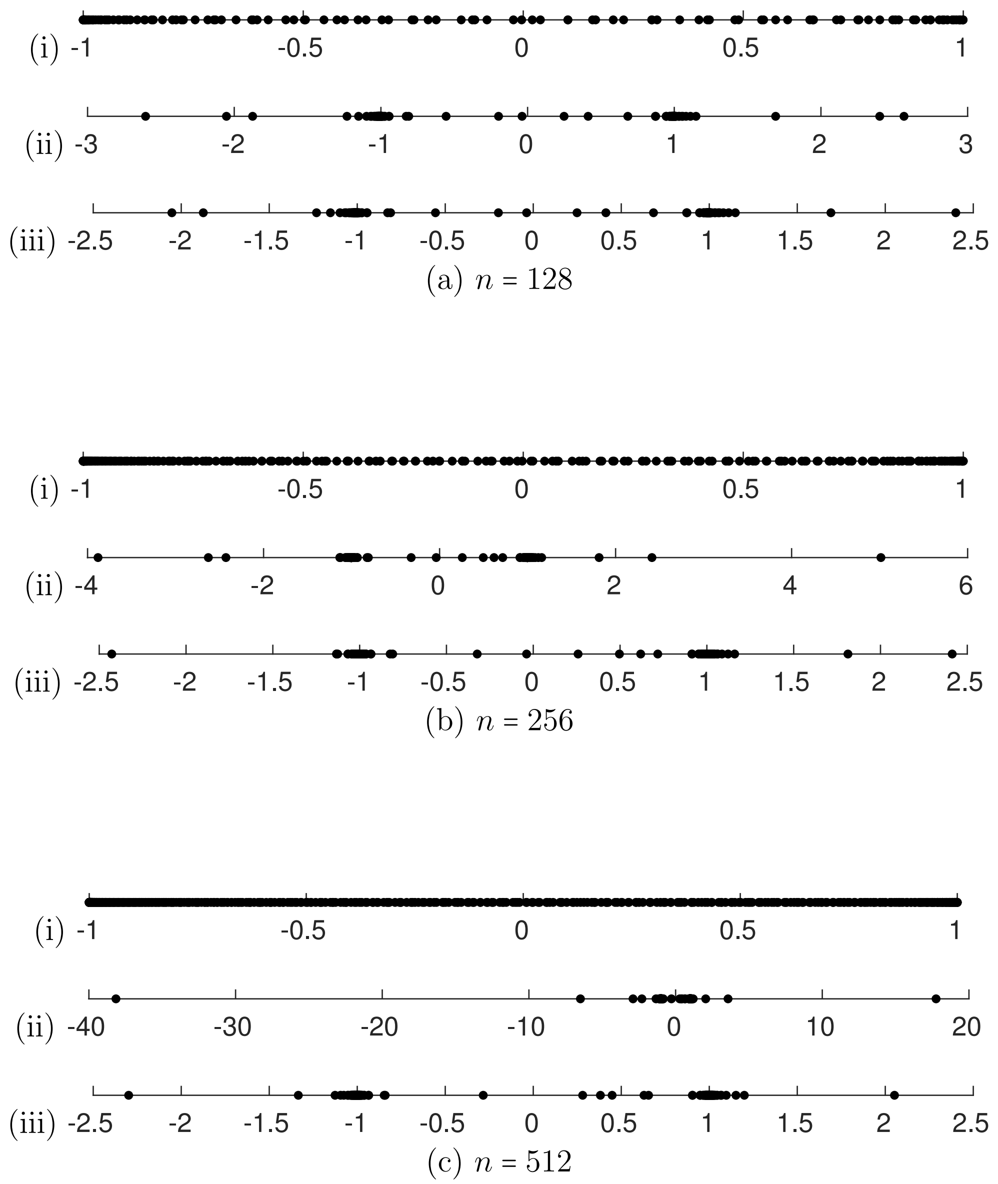
Table 4.7: Numbers of iterations with (a) MINRES and (b) GMRES for $\cos T_n$ with T_n given in Example 4.7.

(a)

n	with no preconditioner	with $\lvert\cos c(T_n)\rvert$
128	58	24
256	110	24
512	212	24
1024	417	22

(b)

n	with no preconditioner	with $\cos c(T_n)$
128	58	12
256	110	13
512	212	11
1024	417	11



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 4.7: Spectra of $\cos T_n$ with T_n given in Example 4.7 at different n (i) with no preconditioner or (ii) with the preconditioner $|\cos c(T_n)|$. (iii) Zoom-in spectrum of (ii).

Example 4.8. We now consider the complex Toeplitz matrix T_n whose entries are given by

$$a_k = \begin{cases} (1+|k|)^{-1.1} + \mathbf{i}(1+|k|)^{-1.1} & k \neq 0 \\ 0 & k = 0 \end{cases}. \tag{4.10}$$

As T_n is non-Hermitian, we resort to its normal equations system. The matrix function considered in this example is the matrix exponential.

Table 4.8 (a) and (b) show the numbers of iterations for the system using CG and GMRES, respectively. Again, we observe that the preconditioners are effective for speeding up the rate of convergence. In Figure 4.8, the highly clustered spectra around 1 are observed.

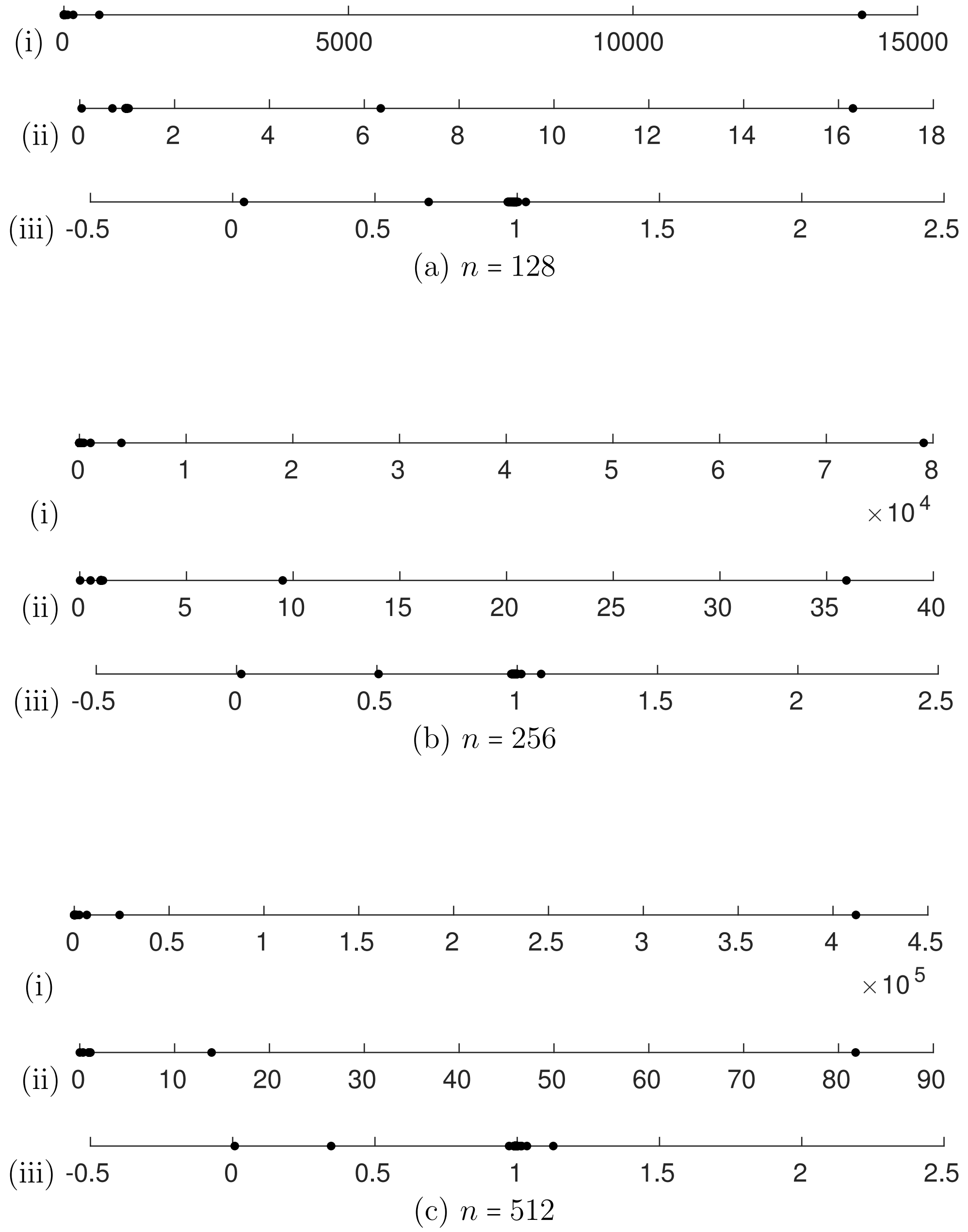
Table 4.8: Numbers of iterations with (a) CG for $(e^{T_n})^* e^{T_n}$ and (b) GMRES for e^{T_n} with T_n given Example 4.8.

(a)

n	with no preconditioner	with $e^{c(T_n)}$
128	24	7
256	43	7
512	76	8
1024	155	9

(b)

n	with no preconditioner	with $e^{c(T_n)}$
128	19	6
256	26	6
512	38	7
1024	55	7



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 4.8: Spectra of $(e^{T_n})^* e^{T_n}$ with T_n given in Example 4.8 at different n (i) with no preconditioner or (ii) with the preconditioner $e^{c(T_n)}$. (iii) Zoom-in spectrum of (ii).

Example 4.9. This example concerns of $\sin T_n$ with T_n given by (4.10). Table 4.9 shows the iteration counts with CG and GMRES for the system. The numbers of iterations are reduced significantly with our proposed preconditioners. Figure 4.9 shows the expected clusters of eigenvalues around 1.

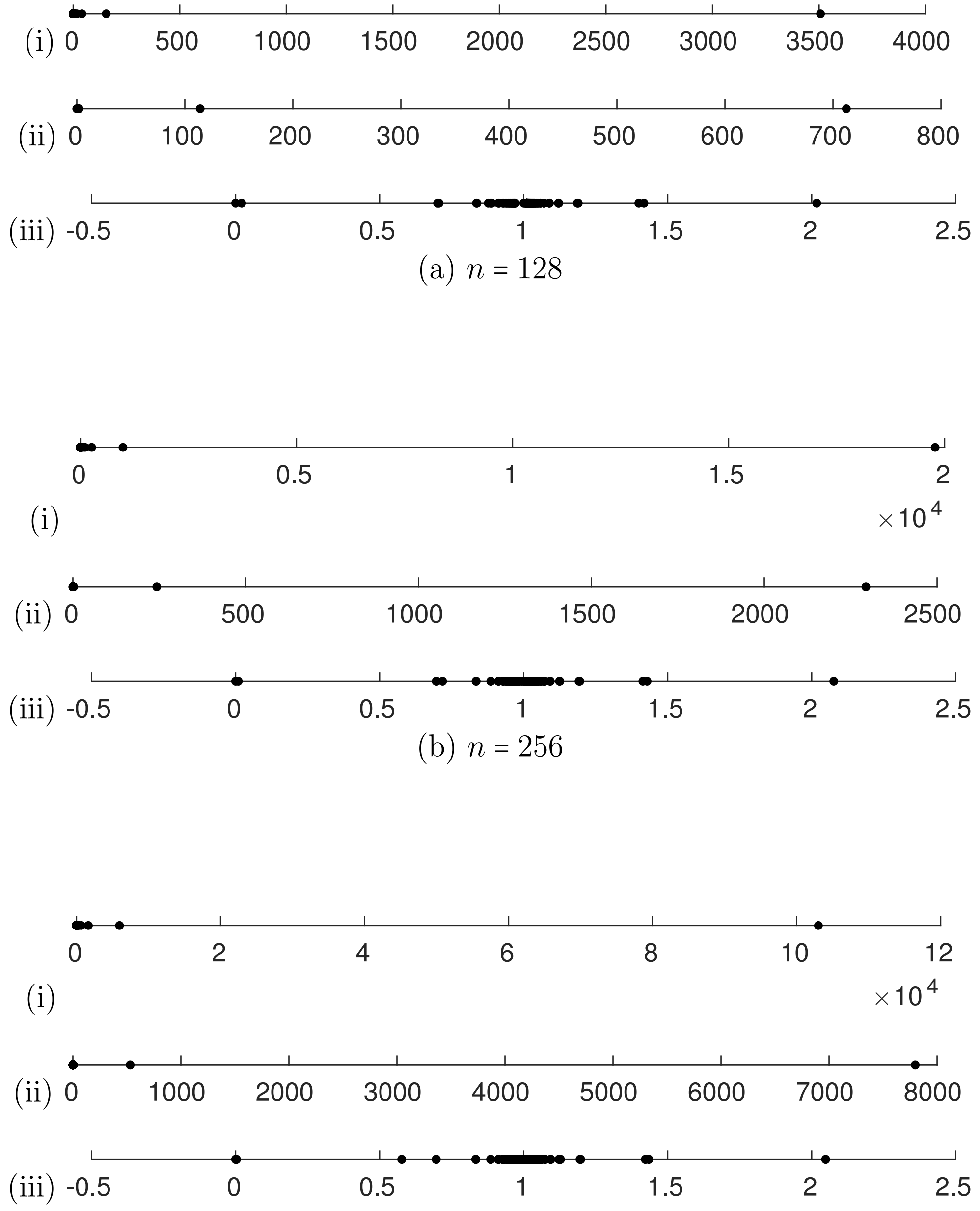
Table 4.9: Numbers of iterations with (a) CG for $(\sin T_n)^* \sin T_n$ and (b) GMRES for $\sin T_n$ with T_n given in Example 4.9.

(a)

n	with no preconditioner	with $\sin c(T_n)$
128	151	18
256	308	26
512	1104	22
1024	1429	27

(b)

n	with no preconditioner	with $\sin c(T_n)$
128	48	11
256	96	13
512	180	11
1024	353	12



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 4.9: Spectra of $(\sin T_n)^* \sin T_n$ with T_n given in Example 4.9 at different n (i) with no preconditioner or (ii) with the preconditioner $\sin c(T_n)$. (iii) Zoom-in spectrum of (ii).

Example 4.10. Finally, we consider the hyperbolic cosine function. Namely, $\cosh T_n$ with T_n given by (4.10).

Table 4.10 shows the iteration counts for the system. In Figure 4.10 (a) and (b), we show the spectra before and after the preconditioner $\cosh c(T_n)$ is applied. In the zoom-in spectrum shown in Figure 4.10 (c), we observe that the eigenvalues cluster around 1.

Table 4.10: Numbers of iterations with (a) CG for $(\cosh T_n)^* \cosh T_n$ and (b) GMRES for $\cosh T_n$ with T_n given in Example 4.10.

(a)

n	with no preconditioner	with $\cosh c(T_n)$
128	10	6
256	19	7
512	41	8
1024	75	9

(b)

n	with no preconditioner	with $\cosh c(T_n)$
128	12	6
256	19	6
512	29	7
1024	43	7

Remark Among the matrices tested in the numerical examples, several of them are ill-conditioned and yet the proposed circulant preconditioning technique still appears to be working moderately well in those cases. However, those are just particular examples in which the condition number is not extremely large. In fact, our proposed circulant preconditioner fails to work for the severely ill-conditioned matrix in Example 4.4 when the dimension increases, which is in line with the well-known results such as [139, 140] that circulant preconditioners are not optimal for ill-conditioned Toeplitz matrices. As will be discussed in Section 8.1.4, band-Toeplitz preconditioners are preferred in the ill-conditioned case.

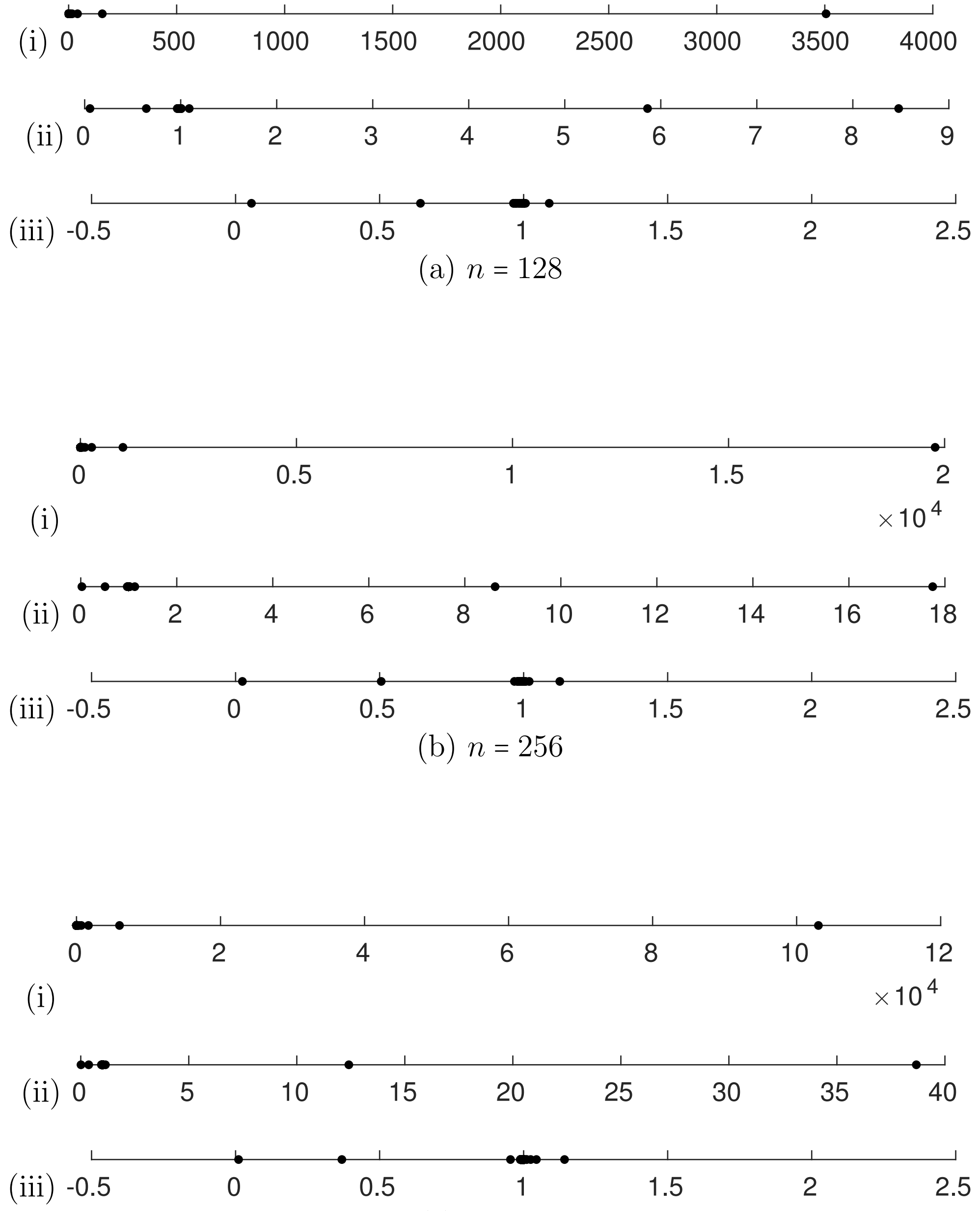
4.5 Conclusions

We have proposed the use of $|h(c(T_n))|$ as a preconditioner for $h(T_n)$, where T_n is generated by $f \in \mathcal{C}[-\pi, \pi]$ and $c(T_n)$ is the optimal circulant preconditioner derived from T_n. Also, we have provided several theorems that account for the effectiveness of $|h(c(T_n))|$.

The decomposition of the obtained preconditioned matrices is often a key to explain the effectiveness of a circulant type preconditioner. Much work has been devoted to showing such decomposition holds in the literature, see for example in [95, 93]. Even for nonsymmetric matrices [120, 127] for which CG and MINRES cannot be straightforwardly applied, such decomposition is still a heuristic indicator to design effective preconditioners. We remark that for nonsymmetric matrices error bounds for Krylov subspace methods do not depend on eigenvalues in general, which is a stark contrast to our case with the symmetrized matrix $Y_nh(T_n)$.

Moreover, we have given a series of numerical examples concerning different $h(z)$ and different T_n. In each of those examples, we have observed significant reductions in numbers of iterations needed for convergence and the expected clusters of eigenvalues.

Even though in many examples GMRES requires fewer iterations than MINRES, it is not necessarily a reduction in work recalling from Chapter 3 that the cost of GMRES increases for every iteration whereas MINRES has a constant cost per iteration. The rapid convergence guarantees established here for MINRES are an indication of the success of our preconditioning approach; it is not perhaps surprising that related preconditioning strategies such as those we have used with GMRES are also successful.



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 4.10: Spectra of $(\cosh T_n)^* \cosh T_n$ with T_n given in Example 4.10 at different n (i) with no preconditioner or (ii) with the preconditioner $\cosh c(T_n)$. (iii) Zoom-in spectrum of (ii).

Chapter 5

Superoptimal preconditioners for functions of Toeplitz matrices[1]

In this chapter, we show that superoptimal circulant preconditioners and Strang's circulant preconditioners are also effective for systems defined by analytic functions $h(z)$ of Toeplitz matrices.

Specifically, we show that $|h(t(T_n))|^{-1}h(T_n)$ can be decomposed into the sum of a Hermitian unitary matrix, a low rank matrix, and a small norm matrix, provided that T_n is the Toeplitz matrix generated by a positive function in the Wiener class and $t(T_n)$ is the superoptimal circulant preconditioner for T_n. Moreover, we show that Strang's circulant preconditioner $s(T_n)$ for T_n also satisfies such a decomposition. As a result, $|h(t(T_n))|^{-1}h(T_n)$ and $|h(s(T_n))|^{-1}h(T_n)$ have clustered spectra around ± 1. Numerical examples are given to support our results. As a comparison, $|h(c(T_n))|$ is also used in the numerical tests, where $c(T_n)$ is the optimal circulant preconditioner used in the previous chapter.

Note that while it only takes $\mathcal{O}(n)$ operations to construct $s(T_n)$ and $c(T_n)$ (see Section 2.3.3.3), $t(T_n)$ requires $\mathcal{O}(n\log n)$ operations to be built using a fast algorithm proposed in [157].

Throughout this chapter, we assume that f is a positive function in the Wiener class, namely

$$\sum_{k=-\infty}^{\infty} |a_k| < \infty,$$

where a_k, $k = 0, \pm 1, \pm 2, \ldots$, are the Fourier coefficients of f.

Recalling Theorem 2.1.1, the eigenvalues of the corresponding Toeplitz matrix T_n must be positive, i.e. T_n is Hermitian positive definite for all n. By Theorem 2.5.2, $h(T_n)$ is also Hermitian positive definite for all n. Hence, neither the normal

[1] This chapter is adapted from [82].

equations matrix of $h(T_n)$ nor the symmetrization matrix Y_n are considered in this chapter.

5.1 Preliminaries on T_n, $t(T_n)$, and $s(T_n)$

We will require the following lemmas in order to show our main results.

Lemma 5.1.1 *[27, Lemma 1] Let f be a positive function in the Wiener class. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f, let $t(T_n) \in \mathbb{C}^{n\times n}$ be the superoptimal circulant preconditioner for T_n, and let $c(T_n^2) \in \mathbb{C}^{n\times n}$ be the optimal circulant preconditioner for T_n^2. Then,*

$$\begin{array}{rcl} \|T_n\|_2 & \leq & f_{\max}, \\ \|t(T_n)\|_2 & \leq & \dfrac{f_{\max}^2}{f_{\min}}, \\ \textit{and} \quad \|c(T_n^2)^{-1}\|_2 & \leq & \dfrac{1}{f_{\min}^2}, \qquad n = 1, 2, \ldots. \end{array}$$

Lemma 5.1.2 *[27, Corollary] Let f be a positive function in the Wiener class. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $c(T_n) \in \mathbb{C}^{n\times n}$ (or $c(T_n^2)$) be the optimal circulant preconditioner for T_n (or T_n^2). Then,*

$$\lim_{n\to\infty} \|c(T_n)^2 - c(T_n^2)\|_2 = 0.$$

Theorem 5.1.3 *[27, Theorem 5] Let f be a positive function in the Wiener class. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $t(T_n) \in \mathbb{C}^{n\times n}$ be the superoptimal circulant preconditioner for T_n. Then, for all $\epsilon > 0$ there exist integers N and $M > 0$ such that for all $n > N$*

$$t(T_n) - T_n = V_n + W_n,$$

where

$$\mathit{rank}(V_n) \leq 2M$$

and

$$\|W_n\|_2 \leq \epsilon.$$

Proof With $t(T_n) = c(T_n)^{-1}c(T_n^2)$ by Theorem 2.3.5, we first decompose

$$t(T_n) - T_n = c(T_n) - T_n - t(T_n)c(T_n^2)^{-1}[c(T_n)^2 - c(T_n^2)],$$

where $c(T_n)$ (or $c(T_n^2)$) is the optimal circulant matrix for T_n (or T_n^2). By Theorem 4.1.2, we know that for all $\epsilon > 0$ there exist integers N_1 and $M > 0$ such that for all $n > N_1$

$$t(T_n) - T_n = V_n + \overline{W_n} - t(T_n)c(T_n^2)^{-1}[c(T_n)^2 - c(T_n^2)],$$

where

$$\text{rank}(V_n) \le 2M$$

and

$$\|\overline{W_n}\|_2 \le \epsilon.$$

By Lemma 5.1.1 and 5.1.2, we know that there exits a positive integer N_2 such that for all $n > N_2$

$$\begin{aligned} \|t(T_n)c(T_n^2)^{-1}[c(T_n)^2 - c(T_n^2)]\|_2 &\le \|t(T_n)\|_2\|c(T_n^2)^{-1}\|_2\|c(T_n)^2 - c(T_n^2)\|_2 \\ &\le (\frac{f_{max}^2}{f_{min}})(\frac{1}{f_{min}^2})\epsilon \\ &\le \frac{f_{max}^2}{f_{min}^3}\epsilon. \end{aligned}$$

Hence, we pick $N := \max\{N_1, N_2\}$ and for all $n > N$

$$\|\underbrace{\overline{W_n} + t(T_n)c(T_n^2)^{-1}[c(T_n)^2 - c(T_n^2)]}_{W_n}\|_2 \le (1 + \frac{f_{max}^2}{f_{min}^3})\epsilon.$$

The result follows. ▮

Similarly, we require the following results for Strang's preconditioners.

Lemma 5.1.4 *[21, Theorem 1] Let f be a positive function in the Wiener class. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $s(T_n) \in \mathbb{C}^{n\times n}$ be Strang's circulant preconditioner for T_n. Then,*

$$\|s(T_n)\|_2 \le f_{\max}.$$

Theorem 5.1.5 *[21, Theorem 2] Let f be a positive function in the Wiener class. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $s(T_n) \in \mathbb{C}^{n\times n}$ be Strang's circulant preconditioner for T_n. Then, for all $\epsilon > 0$ there exist integers N and $M > 0$ such that for all $n > N$*

$$s(T_n) - T_n = V_n + W_n,$$

where

$$\mathit{rank}(V_n) \le 2M$$

and

$$\|W_n\|_2 \le \epsilon.$$

5.2 Main results

In this section, we show that the preconditioned matrix

$$|h(t(T_n))|^{-1}h(T_n) \text{ (or } |h(s(T_n))|^{-1}h(T_n))$$

can be decomposed into the sum of a Hermitian unitary matrix, a low rank matrix, and a small norm matrix for sufficiently large n under certain conditions. As a result, the preconditioned matrix has clustered spectra around ±1.

As in Chapter 4 before, without loss of generality, we assume that $h(z)$ is represented by the following Taylor series:

$$h(z) = \sum_{k=0}^{\infty} a_k z^k.$$

Theorem 5.2.1 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let f be a positive function in the Wiener class and suppose $\frac{f_{max}^2}{f_{min}} < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $t(T_n) \in \mathbb{C}^{n\times n}$ be the superoptimal circulant preconditioner for T_n. Then, for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$h(t(T_n)) - h(T_n) = R_n + E_n$$

where

$$\mathit{rank}(R_n) \leq 2M^{(\epsilon)}$$

and

$$\|E_n\|_2 \leq \epsilon.$$

Proof The proof of this theorem echoes that of Theorem 4.2.1 with slight modifications: mainly replacing Lemma 4.1.1 and Theorem 4.1.2 by Lemma 5.1.1 and Theorem 5.1.3, respectively.

Corollary 5.2.2 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let f be a positive function in the Wiener class and suppose $\frac{f_{max}^2}{f_{min}} < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $t(T_n) \in \mathbb{C}^{n\times n}$ be the superoptimal circulant preconditioner for T_n. If $\|h(t(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$|h(t(T_n))|^{-1}h(T_n) = Q_n + \widehat{R}_n + \widehat{E}_n,$$

where Q_n is Hermitian and unitary,

$$rank(\widehat{R}_n) \le 2M^{(\epsilon)},$$

and

$$\|\widehat{E}_n\|_2 \le \epsilon.$$

Corollary 5.2.3 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let f be a positive function in the Wiener class and suppose $\frac{f_{max}^2}{f_{min}} < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $t(T_n) \in \mathbb{C}^{n\times n}$ be the super-optimal circulant preconditioner for T_n. If $\|h(t(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then $|h(t(T_n))|^{-1}h(T_n)$ has clustered spectra around ±1 for sufficiently large n.*

By Lemma 5.1.4 and Theorem 5.1.5, we can show similar results for Strang's preconditioners.

Theorem 5.2.4 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let f be a positive function in the Wiener class and suppose $f_{\max} < r$. Let T_n be the Toeplitz matrix generated by f and let $s(T_n)$ be Strang's circulant preconditioner for T_n. Then, for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$h(s(T_n)) - h(T_n) = R_n + E_n,$$

where

$$rank(R_n) \le 2M^{(\epsilon)}$$

and

$$\|E_n\|_2 \le \epsilon.$$

Proof Similar to showing Theorem 5.2.1, the proof of this theorem can be done with Lemma 5.1.1 and Theorem 5.1.3 replaced by Lemma 5.1.4 and Theorem 5.1.5, respectively.

Corollary 5.2.5 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let f be a positive function in the Wiener class and suppose $f_{\max} < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $s(T_n) \in \mathbb{C}^{n\times n}$ be Strang's circulant preconditioner for T_n. If $\|h(s(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then for all $\epsilon > 0$ there exist integers $M^{(\epsilon)}$ and $N_M^{(\epsilon)}$ such that for all $n > N_M^{(\epsilon)}$*

$$|h(s(T_n))|^{-1}h(T_n) = Q_n + \widehat{R}_n + \widehat{E}_n,$$

where Q_n is Hermitian and unitary,

$$rank(\widehat{R}_n) \le 2M^{(\epsilon)},$$

and

$$\|\widehat{E}_n\|_2 \le \epsilon.$$

Corollary 5.2.6 *Suppose $h(z)$ is an analytic function on $|z| < r$ with radius of convergence r. Let f be a positive function in the Wiener class and suppose $f_{\max} < r$. Let $T_n \in \mathbb{C}^{n\times n}$ be the Toeplitz matrix generated by f and let $s(T_n) \in \mathbb{C}^{n\times n}$ be Strang's circulant preconditioner for T_n. If $\|h(s(T_n))^{-1}\|_2$ is uniformly bounded with respect to n, then $|h(s(T_n))|^{-1}h(T_n)$ has clustered spectra around ±1 for sufficiently large n.*

5.3 Numerical results

In this section, we demonstrate the effectiveness of $|h(t(T_n))|$, $|h(s(T_n))|$, and $|h(c(T_n))|$ for $h(T_n)\mathbf{x}_n = \mathbf{b}_n$ using CG, MINRES, and GMRES. The settings such as convergence criterion are the same as those used in Chapter 4.

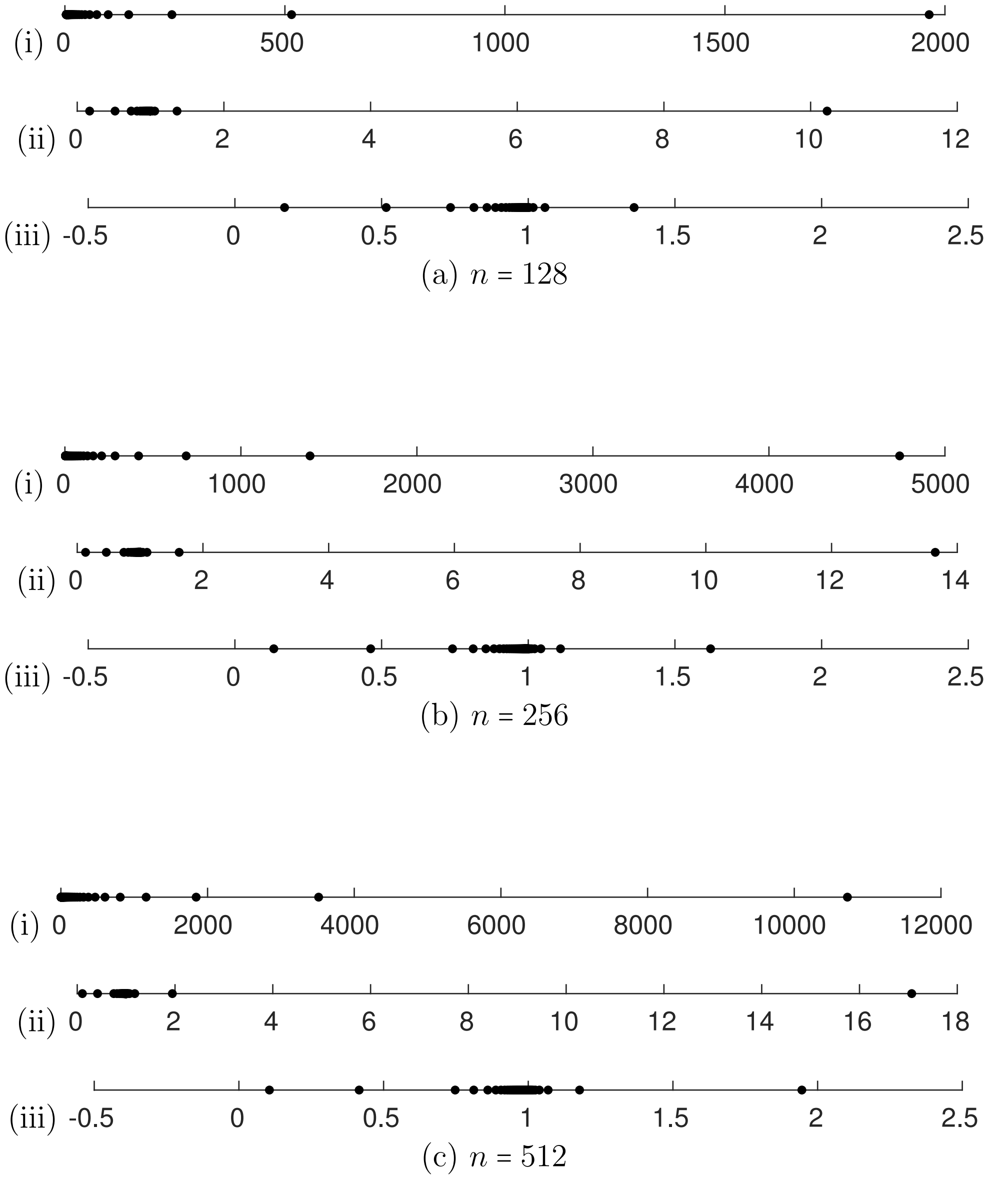
In all numerical tests, we consider the Hermitian Toeplitz matrix T_n generated by

$$f(x) = 2\sum_{k=0}^{\infty} \frac{\sin(kx) + \cos(kx)}{(1+k)^{1.1}}$$

in the Wiener class, which is analysed in [27]. The entries of T_n are given by

$$a_k = \begin{cases} \frac{1+\mathbf{i}}{(1+k)^{1.1}} & k > 0 \\ 2 & k = 0 \\ \bar{a}_{-k} & k < 0 \end{cases}.$$

Example 5.1. We first consider e^{T_n}. Table 5.1 shows the numbers of iterations needed for e^{T_n} with or without the preconditioners. It is observed that the proposed preconditioners are efficient for speeding up the convergence rate of CG. In Figure 5.1 (i) and (ii), the contrast between the spectrum of the preconditioned matrix with $e^{t(T_n)}$ is shown at different n. In Figure 5.1 (iii), we see that the eigenvalues of the preconditioned matrices are highly clustered around 1. A fast convergence rate of CG is expected due to the clustered eigenvalues. Figure 5.2 and 5.3 show similar observations for e^{T_n} when preconditioned by $e^{s(T_n)}$ and $e^{c(T_n)}$ respectively.

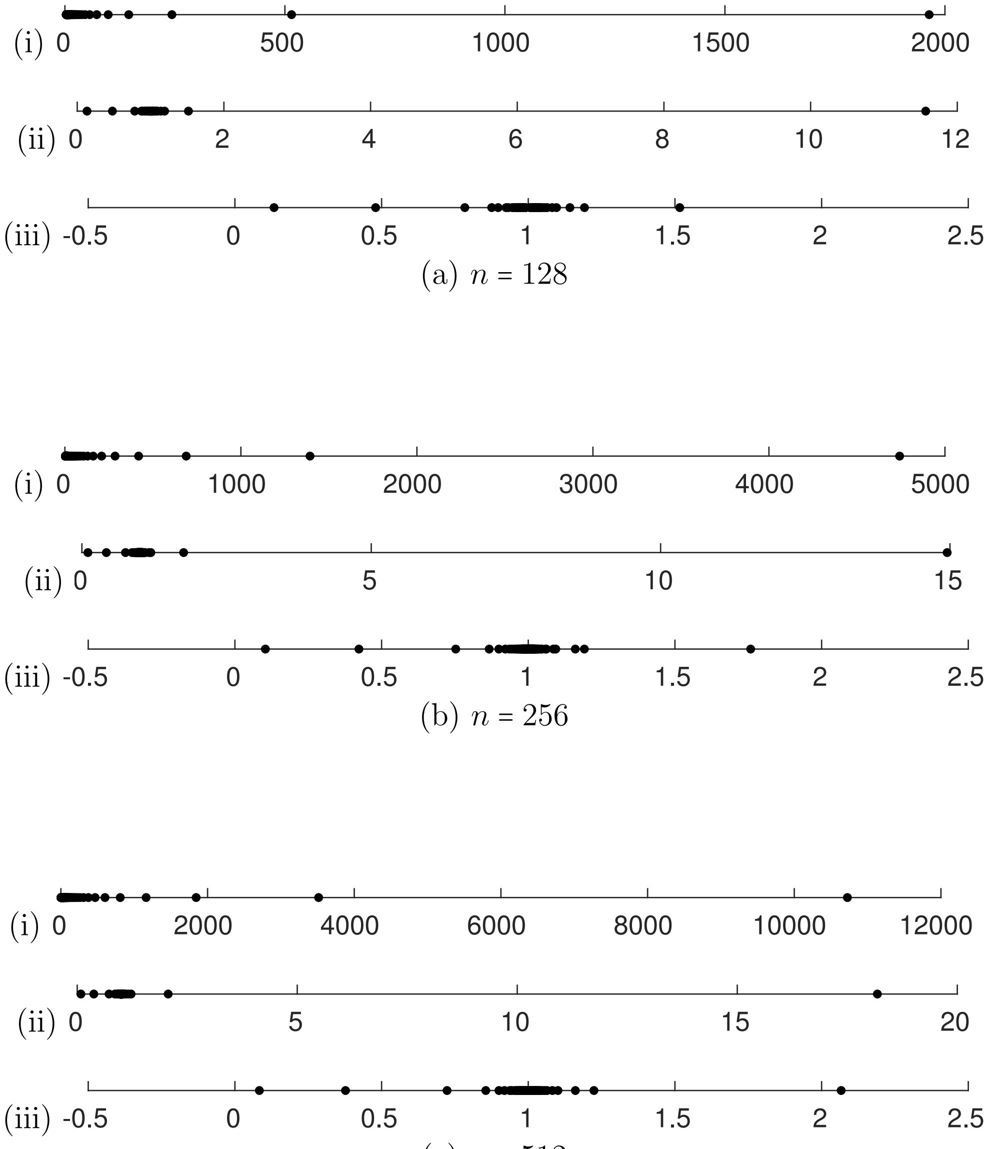


(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.1: Spectra of e^{T_n} given in Example 5.1 at different n (i) with no preconditioner or (ii) with the preconditioner $e^{t(T_n)}$. (iii) Zoom-in spectrum of (ii).

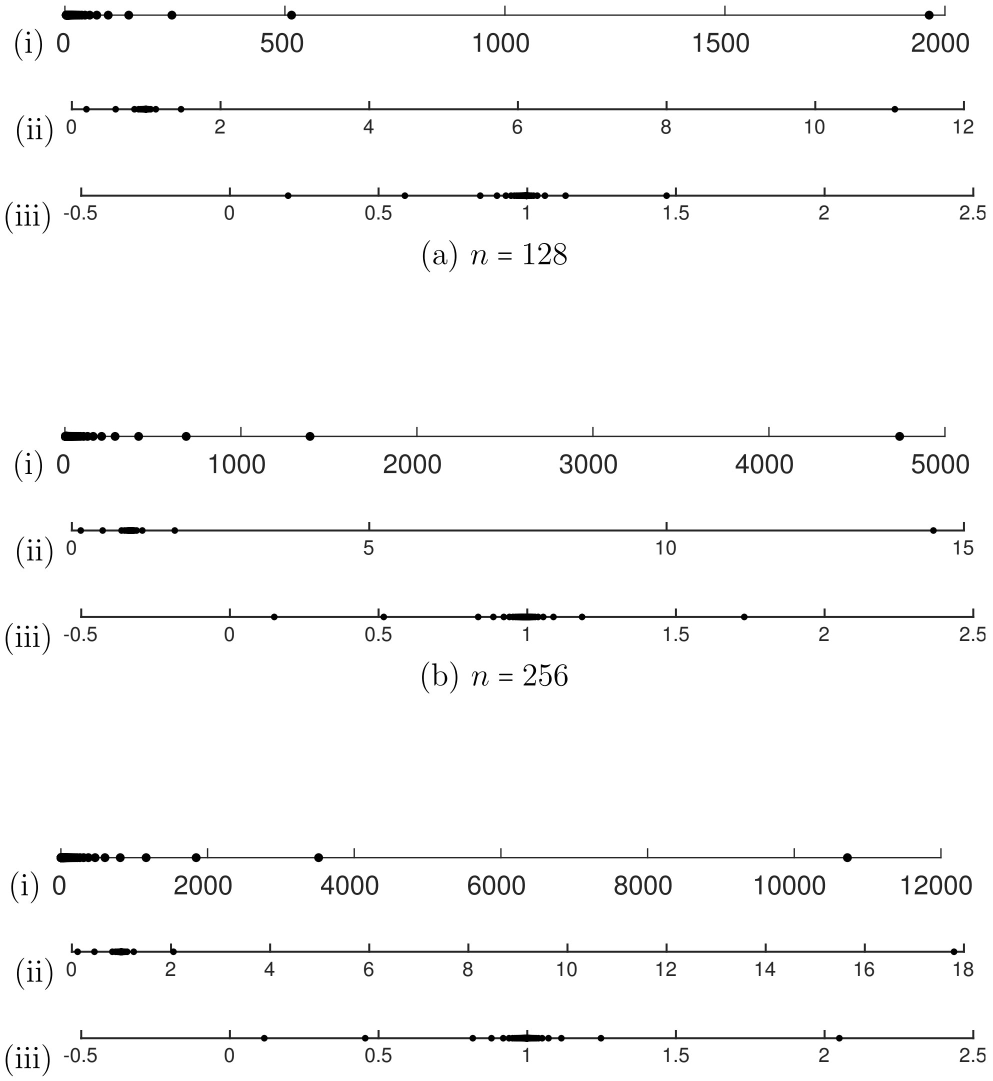


(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.2: Spectra of e^{T_n} given in Example 5.1 at different n (i) with no preconditioner or (ii) with the preconditioner $e^{s(T_n)}$. (iii) Zoom-in spectrum of (ii).



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.3: Spectra of e^{T_n} given in Example 5.1 at different n (i) with no preconditioner or (ii) with the preconditioner $e^{c(T_n)}$. (iii) Zoom-in spectrum of (ii).

Table 5.1: Numbers of iterations with (a) CG or (b) GMRES for e^{T_n} given in Example 5.1.

(a)

n	with no preconditioner	with $e^{t(T_n)}$	with $e^{s(T_n)}$	with $e^{c(T_n)}$
128	34	11	11	10
256	53	11	11	11
512	79	11	12	11
1024	121	12	13	12

(b)

n	with no preconditioner	with $e^{t(T_n)}$	with $e^{s(T_n)}$	with $e^{c(T_n)}$
128	26	11	11	10
256	35	11	12	11
512	46	12	13	11
1024	62	12	13	12

Example 5.2. Table 5.2 shows the numerical results using MINRES and GMRES for $\cos T_n$. Again, we observe that the preconditioners $|\cos t(T_n)|$, $|\cos s(T_n)|$, and $|\cos c(T_n)|$ appear effective for $\cos T_n$. In Figure 5.4, we further show the spectrum of $|\cos t(T_n)|^{-1} \cos T_n$ with different n. We conclude that the highly clustered eigenvalues around ±1 seem independent of n. In Figure 5.5 and 5.6, we also observe the expected clusters of eigenvalues when the system is preconditioned by $|\cos s(T_n)|$ and $|\cos c(T_n)|$ respectively.

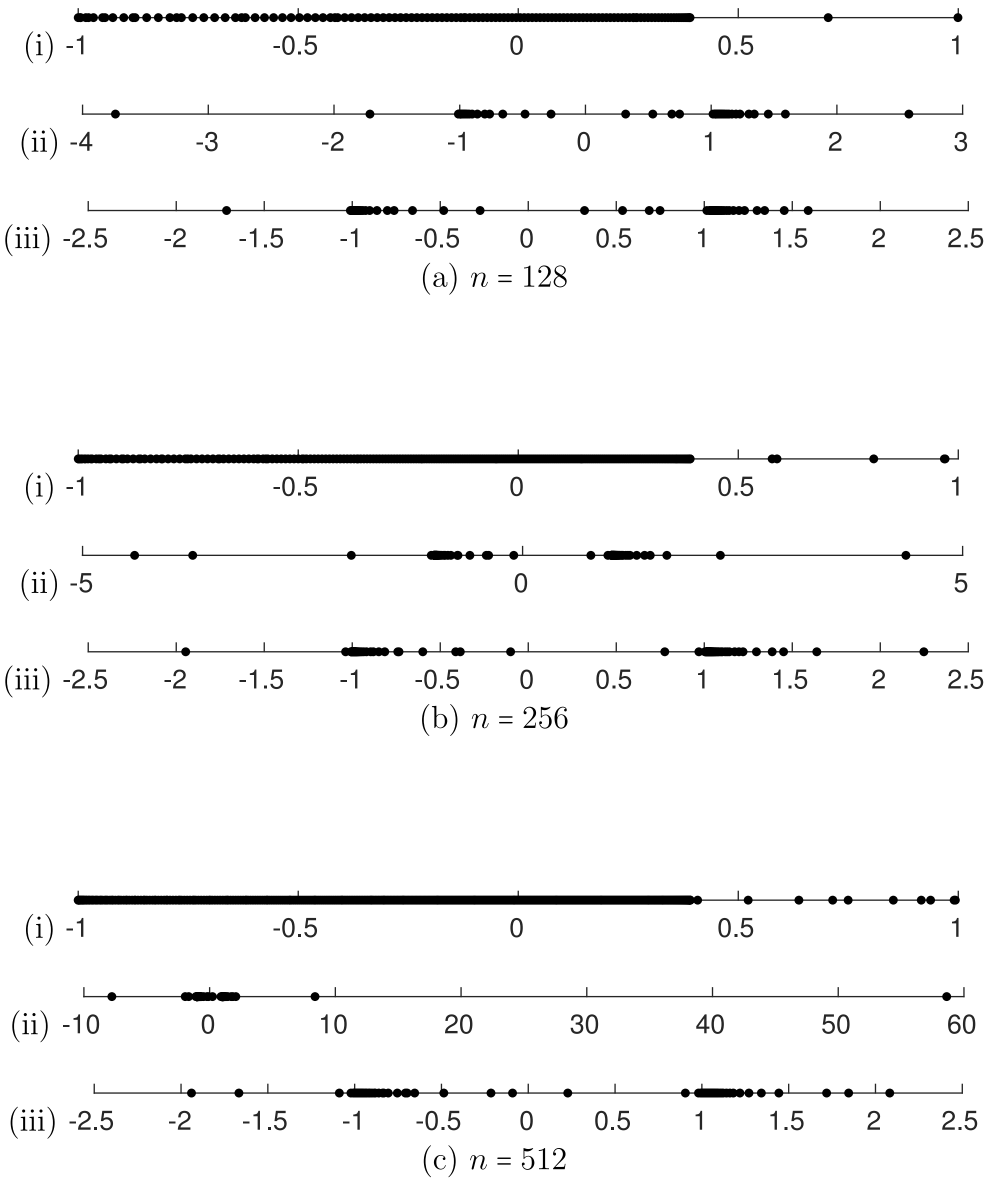
Table 5.2: Numbers of iterations with (a) MINRES or (b) GMRES for $\cos T_n$ given in Example 5.2.

(a)

n	with no preconditioner	with $\lvert\cos t(T_n)\rvert$	with $\lvert\cos s(T_n)\rvert$	with $\lvert\cos c(T_n)\rvert$
128	178	29	42	23
256	412	32	50	36
512	952	49	46	31
1024	2152	47	48	35

(b)

n	with no preconditioner	with $\cos t(T_n)$	with $\cos s(T_n)$	with $\lvert\cos c(T_n)\rvert$
128	128	18	21	15
256	256	18	20	16
512	512	21	24	18
1024	1024	21	24	18

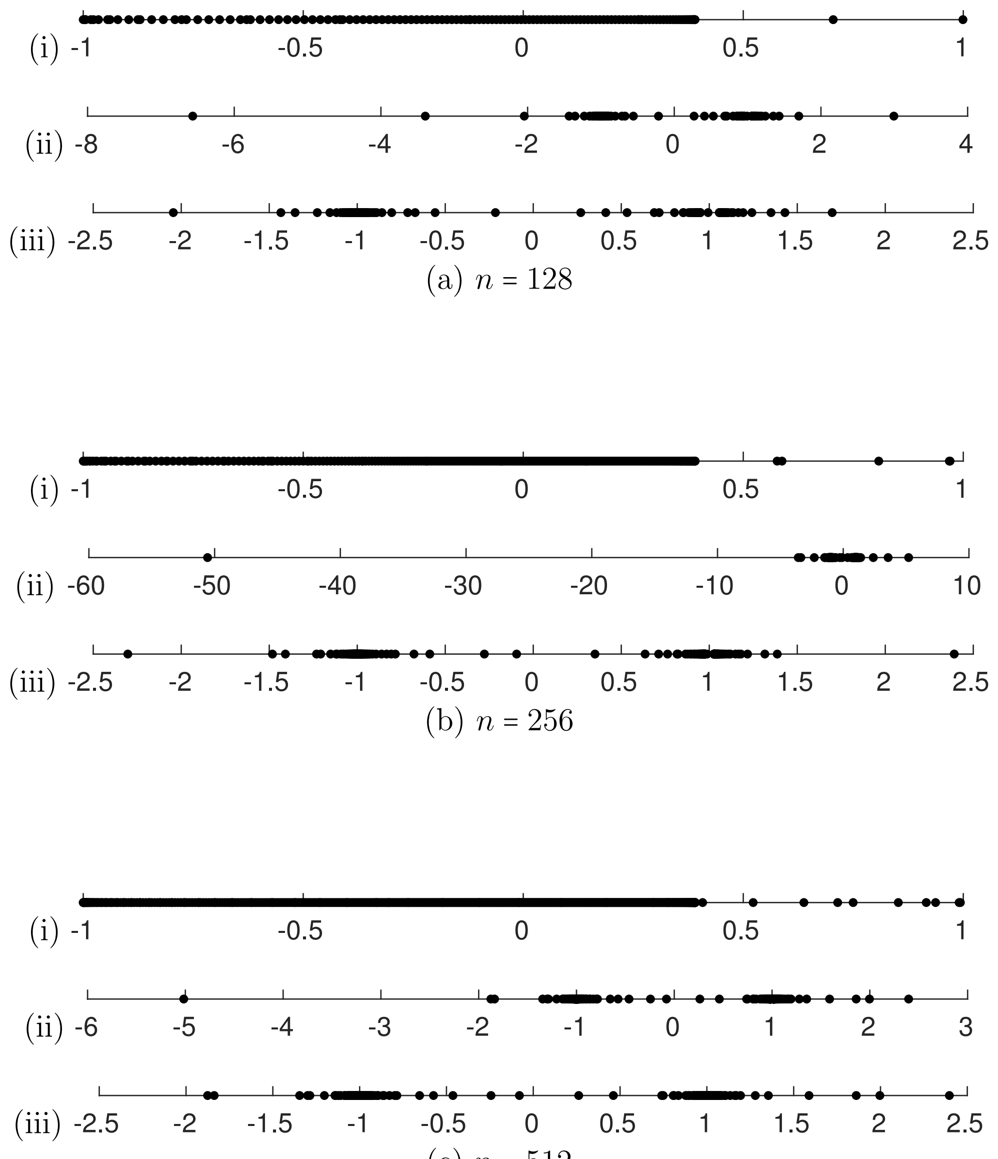


(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.4: Spectra of $\cos T_n$ given in Example 5.2 at different n (i) with no preconditioner or (ii) with the preconditioner $|\cos t(T_n)|$. (iii) Zoom-in spectrum of (ii).

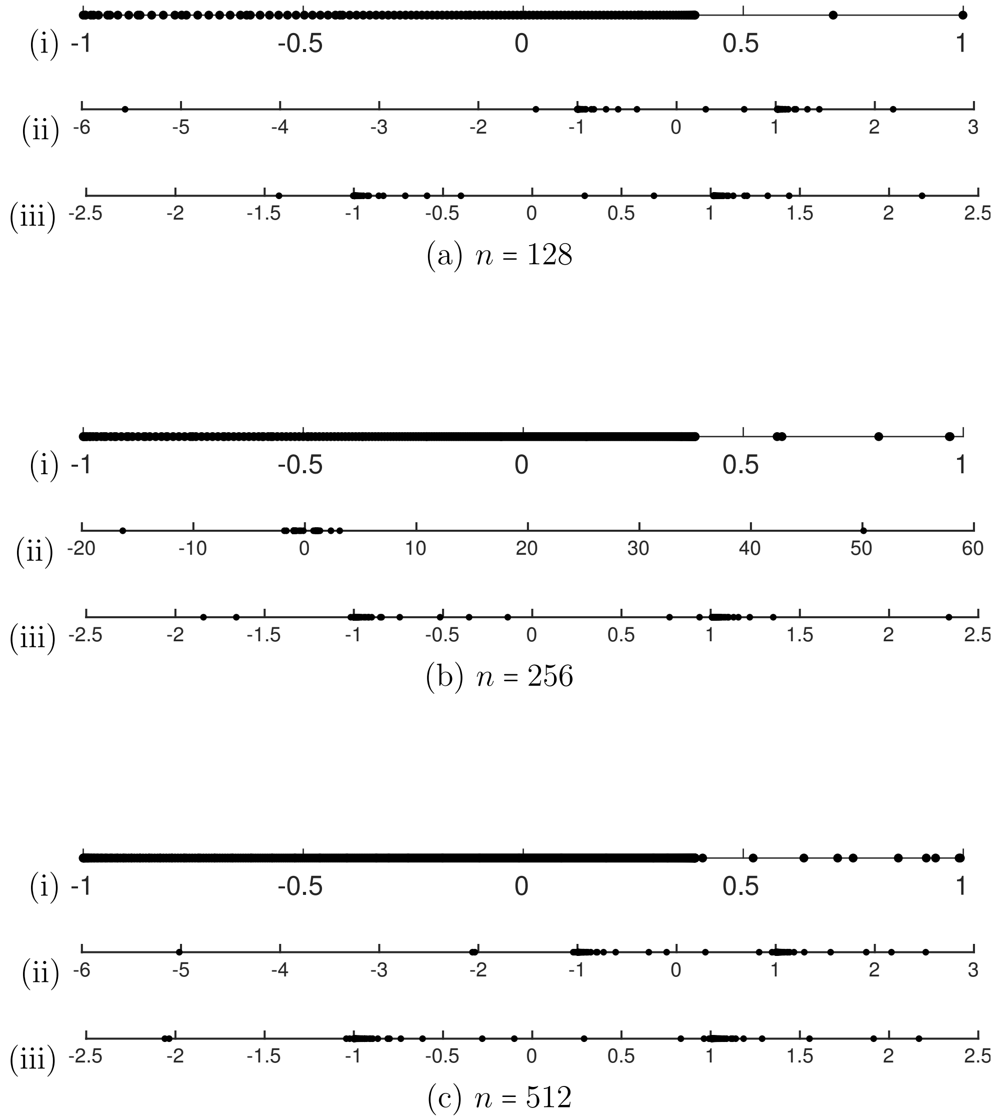


(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.5: Spectra of $\cos T_n$ given in Example 5.2 at different n (i) with no preconditioner or (ii) with the preconditioner $|\cos s(T_n)|$. (iii) Zoom-in spectrum of (ii).



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.6: Spectra of $\cos T_n$ given in Example 5.2 at different n (i) with no preconditioner or (ii) with the preconditioner $|\cos c(T_n)|$. (iii) Zoom-in spectrum of (ii).

Example 5.3. We next consider the system defined by the hyperbolic sine function. Table 5.3 shows the numerical results using CG for $\sinh T_n$. The numbers of iterations needed are reduced significantly with the proposed preconditioners. In Figures 5.7 - 5.9, we observe the clusters around 1 when the system is preconditioned by our proposed preconditioners.

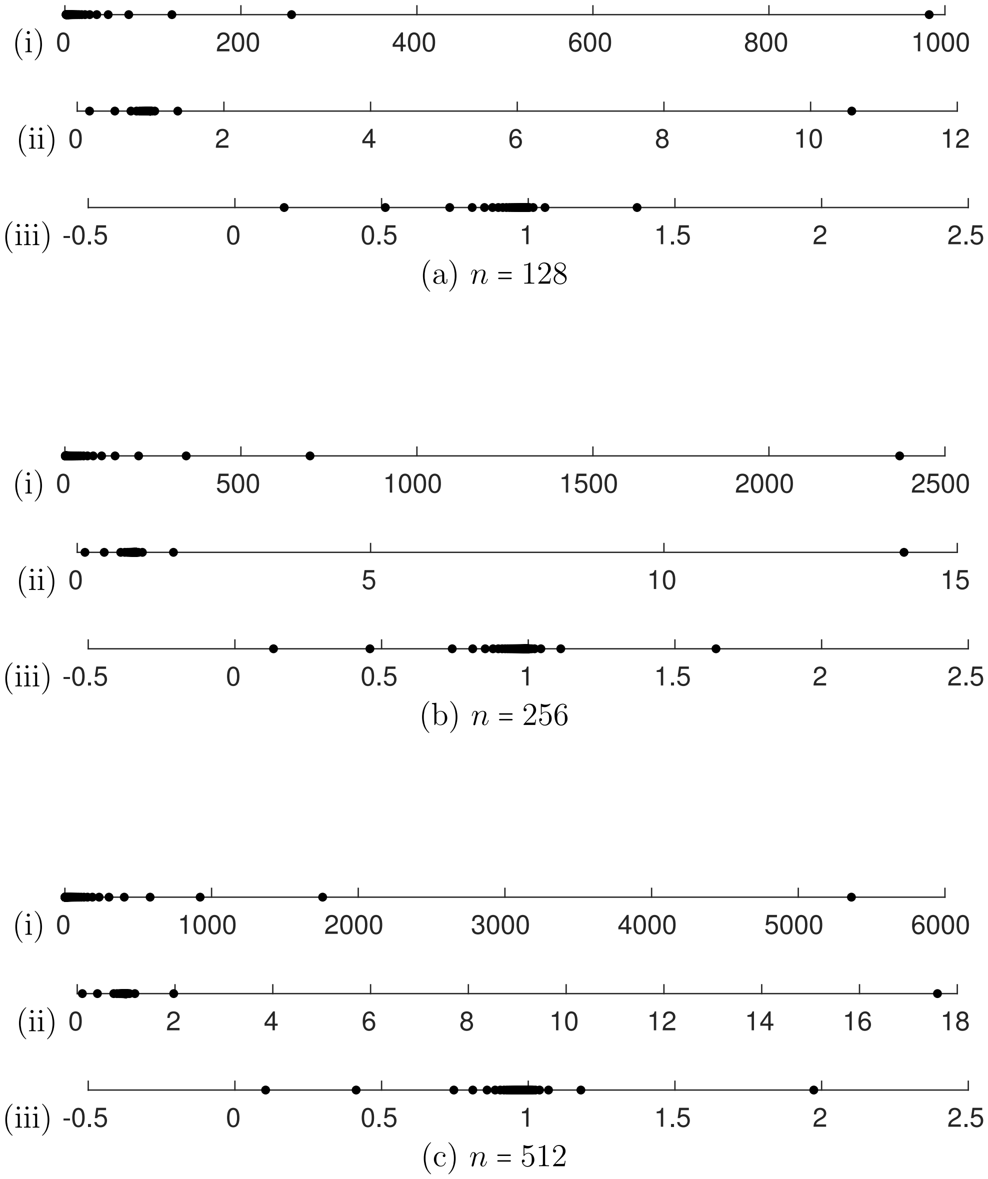
Table 5.3: Numbers of iterations with (a) CG or (b) GMRES for $\sinh T_n$ given in Example 5.3.

(a)

n	with no preconditioner	with $\sinh t(T_n)$	with $\sinh s(T_n)$	with $\sinh c(T_n)$
128	38	11	11	9
256	57	11	12	11
512	82	11	12	11
1024	129	12	13	12

(b)

n	with no preconditioner	with $\sinh t(T_n)$	with $\sinh s(T_n)$	with $\sinh c(T_n)$
128	27	11	11	10
256	36	11	12	11
512	47	12	13	11
1024	63	12	13	12

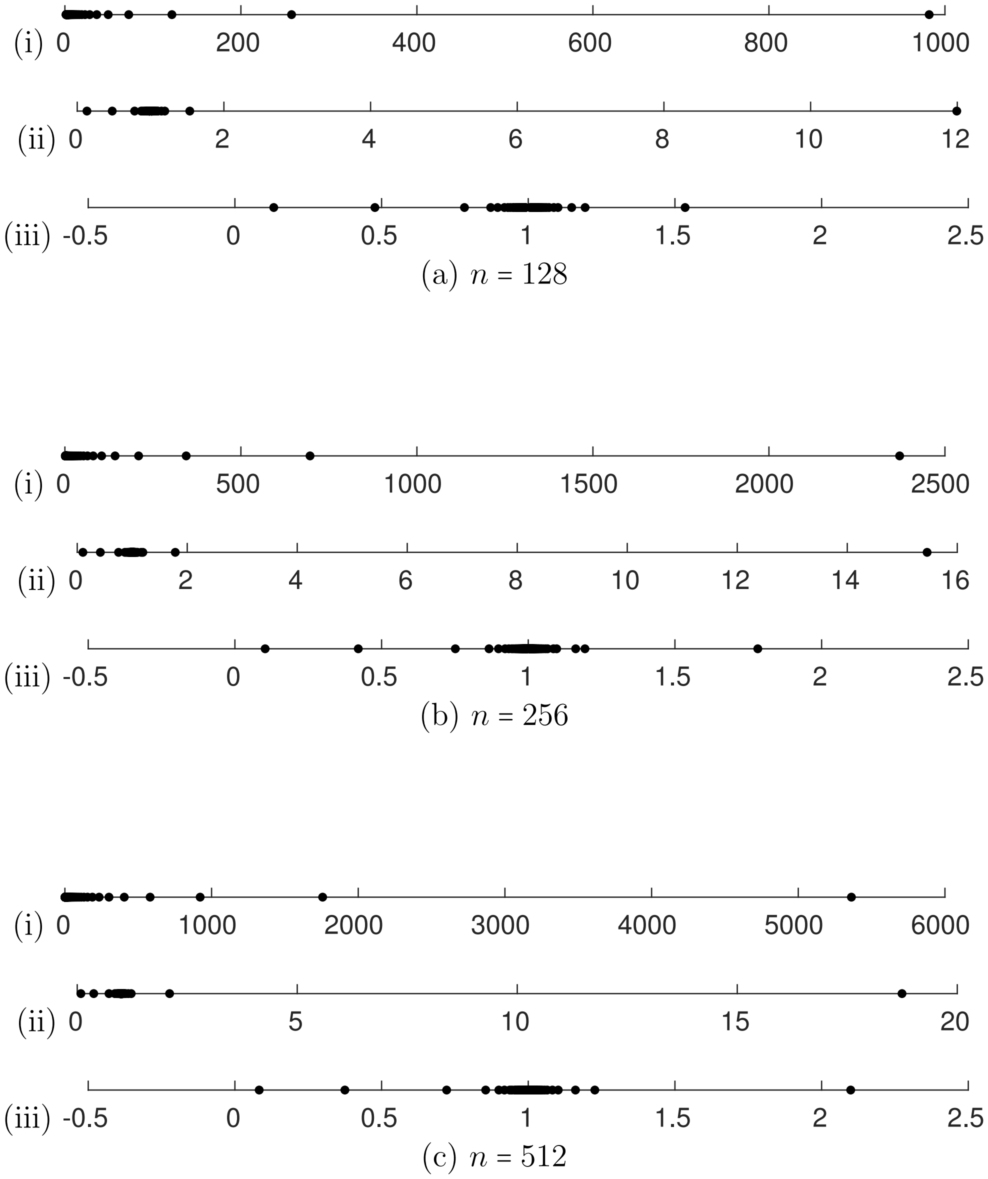


(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.7: Spectra of $\sinh T_n$ given in Example 5.3 at different n (i) with no preconditioner or (ii) with the preconditioner $\sinh t(T_n)$. (iii) Zoom-in spectrum of (ii).

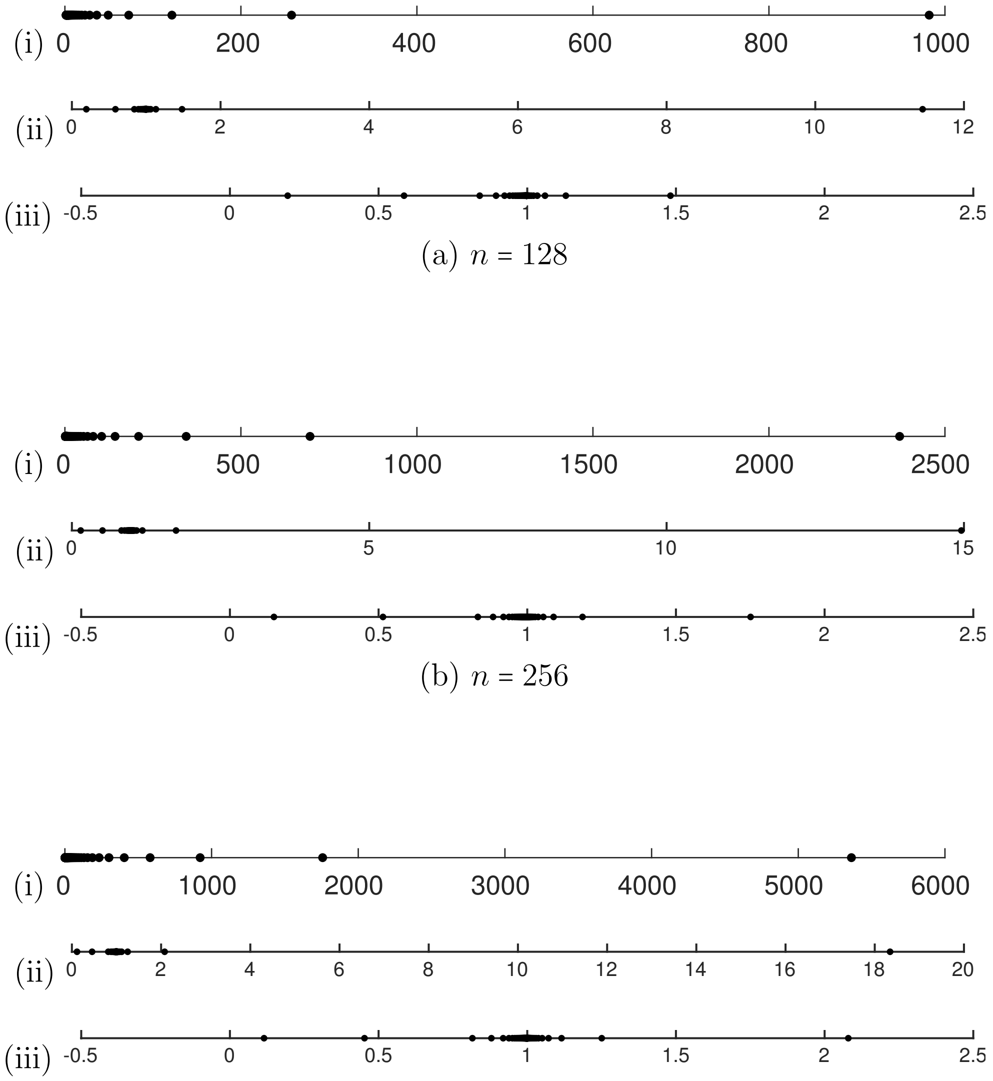


(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.8: Spectra of $\sinh T_n$ given in Example 5.3 at different n (i) with no preconditioner or (ii) with the preconditioner $\sinh s(T_n)$. (iii) Zoom-in spectrum of (ii).



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.9: Spectra of $\sinh T_n$ given in Example 5.3 at different n (i) with no preconditioner or (ii) with the preconditioner $\sinh c(T_n)$. (iii) Zoom-in spectrum of (ii).

Example 5.4. Lastly, we consider the Toeplitz matrix polynomial

$$h(T_n) = T_n^3 + T_n^2 + T_n + I_n.$$

Table 5.4 shows the numerical results for $h(T_n)$. We conclude that the preconditioner $h(t(T_n))$ and $h(s(T_n))$ appear effective for $h(T_n)$. In Figures 5.10 - 5.12, we again observe the clusters of eigenvalues around 1 with our proposed preconditioners.

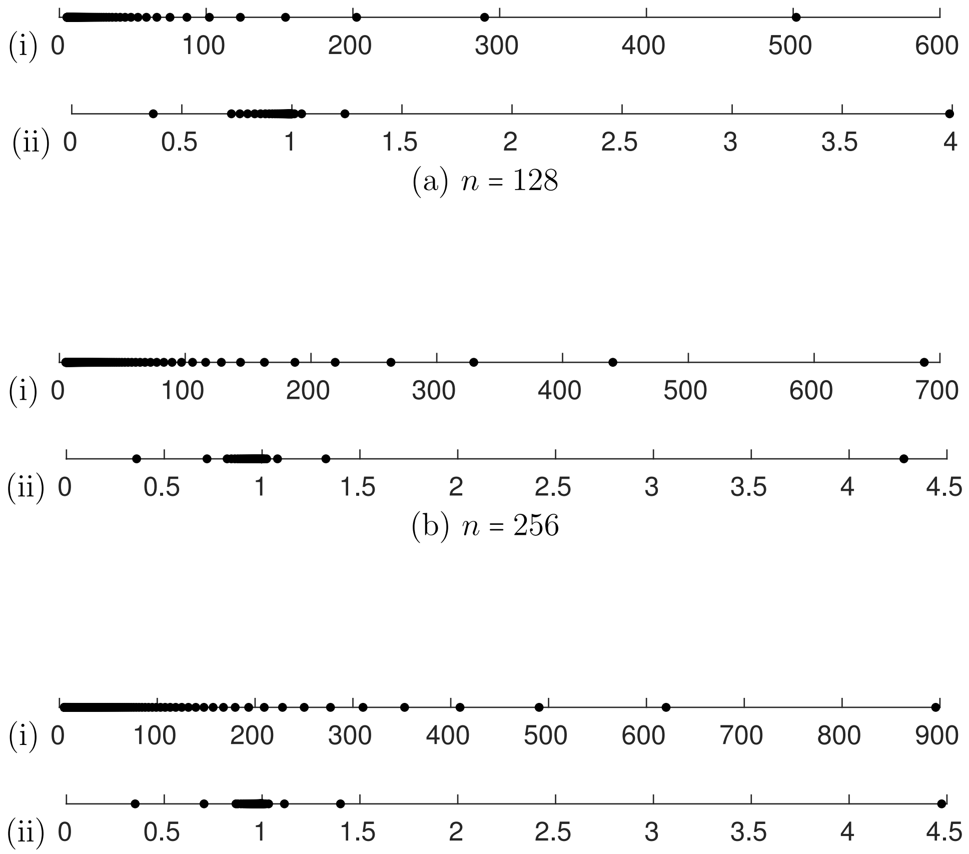
Table 5.4: Numbers of iterations with (a) CG or (b) GMRES for $h(T_n)$ given in Example 5.4.

(a)

n	with no preconditioner	with $h(t(T_n))$	with $h(s(T_n))$	with $h(c(T_n))$
128	32	9	9	9
256	40	9	9	8
512	50	9	9	9
1024	63	9	9	9

(b)

n	with no preconditioner	with $h(t(T_n))$	with $h(s(T_n))$	with $h(c(T_n))$
128	27	10	10	9
256	35	10	10	9
512	43	9	10	9
1024	51	10	10	9

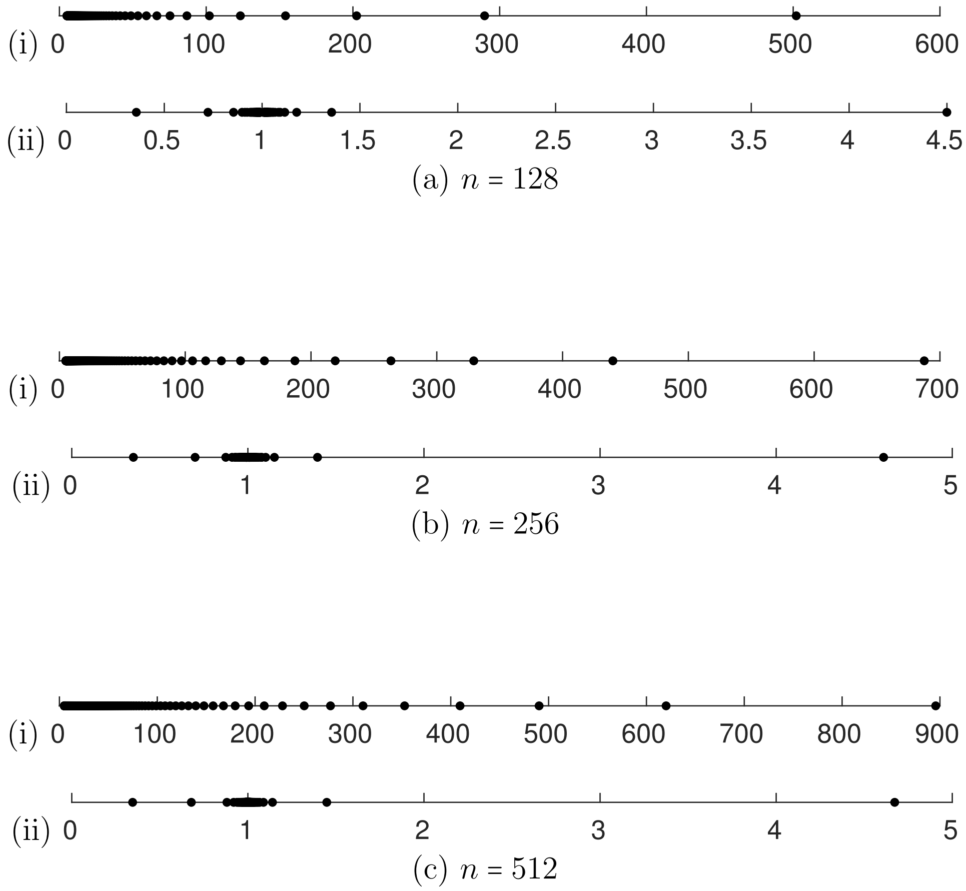


(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.10: Spectra of $h(T_n)$ given in Example 5.4 at different n (i) with no preconditioner or (ii) with the preconditioner $h(t(T_n))$.

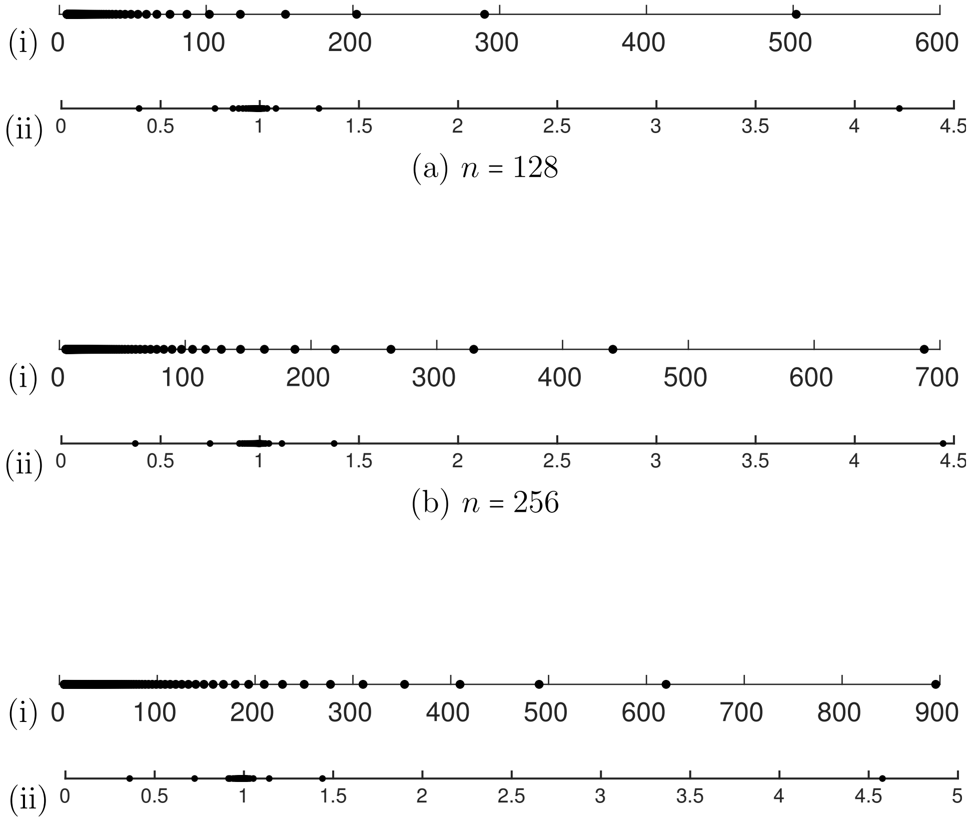


(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.11: Spectra of $h(T_n)$ given in Example 5.4 at different n (i) with no preconditioner or (ii) with the preconditioner $h(s(T_n))$.



(a) $n = 128$

(b) $n = 256$

(c) $n = 512$

Figure 5.12: Spectra of $h(T_n)$ given in Example 5.4 at different n (i) with no preconditioner or (ii) with the preconditioner $h(c(T_n))$.

5.4 Conclusions

In this chapter, considering superoptimal circulant preconditioners and Strang's preconditioners, we have provided several theorems that explain their effectiveness for $h(T_n)$ with T_n generated by a positive function in the Wiener class. A number of numerical examples concerning different functions $h(z)$ have been given to support our results. In each example, we have observed significant improvements in convergence and the expected clusters of eigenvalues around ± 1. Also, while all proposed preconditioners appear effective, $h(c(T_n))$ works best among them. This finding is in accordance with the results given in [149] in which $c(T_n)$ is shown to be superior compared with $t(T_n)$ and $s(T_n)$ for the Toeplitz matrices generated by continuous functions.

Chapter 6

Application to time-stepping methods for ordinary differential equations[1]

In this chapter, we illustrate that the symmetrization technique and absolute value preconditioning can be employed on time-stepping problems for ordinary differential equations (ODEs) as simple applications.

In particular, we describe one simple and frequently arising situation: banded (real) nonsymmetric Toeplitz matrices, where we can guarantee rapid convergence of the appropriate iterative method by manipulating the problem into a symmetric one without recourse to the normal equations. Moreover, considering simple bidiagonal nonsymmetric Toeplitz matrices, we show that the minimal polynomial of the matrices when preconditioned by Strang's circulant preconditioner is quadratic. As a result, Krylov subspace methods like GMRES will converge in two steps. The result that we establish in this setting is the guaranteed convergence of an iterative method for an all-at-once formulation in a number of iterations independent of the number of time-steps.

6.1 Theta method

We consider only a scalar linear ordinary differential equation,

$$\frac{\mathrm{d}y}{\mathrm{d}t} = ay + f, \qquad y(0) = y_0,$$

[1] This chapter is adapted from the book chapter with Eleanor McDonald, Jennifer Pestana, and Andy Wathen [112]. My main contribution to this work is showing Theorem 6.1.1.

on the time interval $[0, T]$. For simplicity, we consider only the simple two-level θ-method, which gives

$$\frac{\mathrm{y}^{(k+1)} - \mathrm{y}^{(k)}}{\tau} = a\theta \mathrm{y}^{(k+1)} + a(1-\theta)\mathrm{y}^{(k)} + f^{(k)}, \qquad \mathrm{y}^{(0)} = y_0,$$

where $\theta \in [0, 1]$ and τ is the constant time step with $n\tau = T$. Euler's method corresponds to the choice $\theta = 1$ and the trapezoidal rule to $\theta = 1/2$.

The discrete equations to be solved are

$$(1 - a\theta\tau)\mathrm{y}^{(k+1)} = (1 + a(1-\theta)\tau)\mathrm{y}^{(k)} + \tau f^{(k)}, \qquad k = 0, 1, 2, \dots, n-1, \tag{6.1}$$

with $\mathrm{y}^{(0)} = y_0$.

A usual approach is solving the equations (6.1) sequentially for $k = 0, 1, 2, \dots, n-1$, which is exactly forward substitution for the all-at-once system

$$T_n \underbrace{\begin{bmatrix} \mathrm{y}^{(1)} \\ \mathrm{y}^{(2)} \\ \mathrm{y}^{(3)} \\ \vdots \\ \mathrm{y}^{(n)} \end{bmatrix}}_{\mathbf{y}_n} = \underbrace{\begin{bmatrix} \tau f^{(0)} + (1 + a(1-\theta)\tau)y_0 \\ \tau f^{(1)} \\ \tau f^{(2)} \\ \vdots \\ \tau f^{(n-1)} \end{bmatrix}}_{\mathbf{f}_n}$$

where

$$T_n = \begin{bmatrix} 1 - a\theta\tau & & & \\ -1 - a(1-\theta)\tau & \ddots & & \\ & \ddots & \ddots & \\ & & -1 - a(1-\theta)\tau & 1 - a\theta\tau \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{6.2}$$

However, we can see that the coefficient matrix T_n in the all-at-once system is a nonsymmetric Toeplitz matrix, provided that a and τ are real numbers. Hence, the symmetrization technique given in the previous chapters applies.

Thus, we have

$$Y_n T_n = \begin{bmatrix} & & -1 - a(1-\theta)\tau & 1 - a\theta\tau \\ & \iddots & \iddots & \\ -1 - a(1-\theta)\tau & \iddots & & \\ 1 - a\theta\tau & & & \end{bmatrix} \in \mathbb{R}^{n\times n}$$

and Strang's absolute value circulant preconditioner $|S_n|$ for T_n is given by

$$|S_n| = \left| \begin{bmatrix} 1 - a\theta\tau & & & -1 - a(1-\theta)\tau \\ -1 - a(1-\theta)\tau & \ddots & & \\ & \ddots & \ddots & \\ & & -1 - a(1-\theta)\tau & 1 - a\theta\tau \end{bmatrix} \right| \in \mathbb{R}^{n\times n}.$$

Remark It is interesting to note that the condition number of T_n in (6.2) depends on the coefficients a, θ, and τ in general. Although the singular values of T_n are distributed as the norm of its generating function, the condition number of T_n is not the mainly determined by $|f|$. In fact, even if $|f|$ does not have any zero, the corresponding Toeplitz matrix can still be ill-conditioned or even singular. The following example illustrates this observation. Choosing $a = 1/\tau$ and $\theta = 1$, we have the bidiagonal Toeplitz matrix

$$T_n[f] = \begin{bmatrix} 0 & & & \\ -1 & 0 & & \\ & \ddots & \ddots & \\ & & -1 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n}$$

generated by $f(x) = -e^{\mathbf{i}x}$. Clearly, $f(x)$ has no zeros since $|f(x)| = 1$ and yet $T_n[f]$ is singular.

Example 6.1. In Tables 6.1 and 6.2, we show the numerical results for $Y_nT_n\mathbf{y}_n = Y_n\mathbf{f}_n$ using MINRES with Strang's absolute value circulant preconditioner $|S_n|$ and observe that MINRES converges in 4 iterations independently of n. The parameters for the presented results are $a = -0.3, \tau = 0.2$, and $\theta = 0.8$; similar behaviour has been observed for many other sets of parameters. The eigenvalues of the preconditioned matrix for this problem are shown in Table 6.2.

Table 6.1: Condition numbers $\kappa(T_n)$ for a time-dependent linear ODE using the θ-method for T_n given by (6.2) and the numbers of iterations with MINRES and the preconditioner $|S_n|$ for Y_nT_n.

n	$\kappa(T_n)$	Iterations
10	10.474	4
100	30.852	4
1000	33.887	4

In fact, for such a simple bidiagonal Toeplitz matrix T_n, we can show that the minimal polynomial of $S_n^{-1}T_n$ is quadratic with Strang's preconditioner S_n. Therefore, this is a rare situation in which it is possible to show that GMRES must terminate with the solution in 2 iterations.

Theorem 6.1.1 *Suppose $\alpha \neq 0$ and $\beta \neq \alpha \in \mathbb{C}$. Assuming $n \geq 2$, if*

$$T_n = \begin{bmatrix} \alpha & & & \\ \beta & \alpha & & \\ & \ddots & \ddots & \\ & & \beta & \alpha \end{bmatrix} \in \mathbb{C}^{n\times n}$$

Table 6.2: Eigenvalues of the preconditioned system (to 4 decimal places). Repeated eigenvalues are shown in brackets with the number of repeated eigenvalues indicated.

n	Eigenvalues of $\lvert S_n\rvert^{-1}Y_nT_n$
10	$\{-0.7206, (-1 \times 4), (1 \times 4), 3.1155\}$
100	$\{-0.4975, (-1 \times 49), (1 \times 49), 2.0157\}$
1000	$\{-0.4966, (-1 \times 499), (1 \times 499), 2.0139\}$

is preconditioned by

$$S_n = \begin{bmatrix} \alpha & & & \beta \\ \beta & \ddots & & \\ & \ddots & \ddots & \\ & & \beta & \alpha \end{bmatrix} \in \mathbb{C}^{n \times n},$$

then the minimal polynomial of $S_n^{-1}T_n$ is quadratic.

Proof Simple calculations yield

$$B_n = S_n^{-1}T_n = \begin{bmatrix} 1 & & & \frac{-\alpha^{n-1}\beta}{\det S_n} \\ & \ddots & & \frac{\alpha^{n-2}\beta^2}{\det S_n} \\ & & \ddots & \vdots \\ & & 1 & \frac{(-1)^{n-1}\alpha\beta^{n-1}}{\det S_n} \\ & & & \frac{\alpha^n}{\det S_n} \end{bmatrix} \in \mathbb{R}^{n \times n},$$

where

$$\det S_n = \begin{cases} \alpha^n + \beta^n & \text{if } n \text{ is odd} \\ \alpha^n - \beta^n & \text{if } n \text{ is even} \end{cases}.$$

We can now easily show that B_n satisfies

$$(B_n - I_n)(B_n - \frac{\alpha^n}{\det S_n} I_n) = 0.$$

Since $(B_n - I_n) \neq 0$ and $(B_n - \frac{\alpha^n}{\det S_n} I_n) \neq 0$, $(B_n - I_n)(B_n - \frac{\alpha^n}{\det S_n} I_n)$ is the minimal polynomial of B_n. ▌

Since the minimal polynomial of the preconditioned coefficient matrix in this case is quadratic, we know that the Krylov subspace is of dimension 2. Because of its minimization property, GMRES termination must occur within 2 iterations. We note that MINRES for Y_nT_n takes 4 iterations (Table 6.1).

6.2 Multistep method

In order to examine a slightly more complex system where the minimal polynomial is not as trivial as before, we examine also a 2-step backward differentiation formula (BDF2) time stepping method. We remark that the generalization to the multistep method is straightforward.

We now require two initial conditions y_{-1} and y_0. For the BDF2 method, we have

$$\frac{\mathrm{y}^{(k+1)} - \frac{4}{3}\mathrm{y}^{(k)} + \frac{1}{3}\mathrm{y}^{(k-1)}}{\tau} = \tfrac{2}{3}a\mathrm{y}^{(k+1)} + \tfrac{2}{3}f^{(k+1)}, \qquad \mathrm{y}^{(0)} = y_0, \quad \mathrm{y}^{(-1)} = y_{-1},$$

where τ is the constant time step with $n\tau = T$. The discrete equations to be solved are

$$(1 - \tfrac{2}{3}a\tau)\mathrm{y}^{(k+1)} = \tfrac{4}{3}\mathrm{y}^{(k)} - \tfrac{1}{3}\mathrm{y}^{(k-1)} + \tfrac{2}{3}\tau f^{(k+1)}, \qquad k = 0, 1, 2, \ldots, n-1,$$

with $\mathrm{y}^{(0)} = y_0$ and $\mathrm{y}^{(-1)} = y_{-1}$. The corresponding all-at-once system is

$$T_n \underbrace{\begin{bmatrix} \mathrm{y}^{(1)} \\ \mathrm{y}^{(2)} \\ \mathrm{y}^{(3)} \\ \vdots \\ \mathrm{y}^{(n)} \end{bmatrix}}_{\mathbf{y}_n} = \underbrace{\begin{bmatrix} \frac{2}{3}\tau f^{(1)} + \frac{4}{3}y_0 - \frac{1}{3}y_{-1} \\ \frac{2}{3}\tau f^{(2)} - \frac{1}{3}y_0 \\ \frac{2}{3}\tau f^{(3)} \\ \vdots \\ \frac{2}{3}\tau f^{(n)} \end{bmatrix}}_{\mathbf{f}_n},$$

where

$$T_n = \begin{bmatrix} 1 - \frac{2}{3}a\tau & & & & \\ -\frac{4}{3} & \ddots & & & \\ \frac{1}{3} & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{3} & -\frac{4}{3} & 1 - \frac{2}{3}a\tau \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{6.3}$$

The coefficient matrix T_n in (6.3) has an additional subdiagonal but is still a nonsymmetric Toeplitz matrix if a and τ are real numbers. Thus, we have

$$Y_n T_n = \begin{bmatrix} & & \frac{1}{3} & -\frac{4}{3} & 1 - \frac{2}{3}a\tau \\ & & \iddots & \iddots & \iddots \\ \frac{1}{3} & \iddots & \iddots & & \\ -\frac{4}{3} & \iddots & & & \\ 1 - \frac{2}{3}a\tau & & & & \end{bmatrix} \in \mathbb{R}^{n \times n}$$

and Strang's absolute value circulant preconditioner $|S_n|$ for T_n is given by

$$|S_n| = \left| \begin{bmatrix} 1 - \frac{2}{3}a\tau & & & \frac{1}{3} & -\frac{4}{3} \\ -\frac{4}{3} & \ddots & & & \frac{1}{3} \\ \frac{1}{3} & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{3} & -\frac{4}{3} & 1 - \frac{2}{3}a\tau \end{bmatrix} \right| \in \mathbb{R}^{n \times n}.$$

Example 6.2. Applying MINRES to solve the system $Y_n T_n \mathbf{y}_n = Y_n \mathbf{f}_n$ with $|S_n|$ as a preconditioner and a random initial guess, from the results in Table 6.3 we see convergence in 6 iterations independently of n. The parameter values for the presented results are again set as $a = -0.3$ and $\tau = 0.2$. As we have used implicit time-stepping, we have no restrictions on the value of τ to maintain stability. Also, as implied by Theorem 6.1.1, it is the lower diagonal Toeplitz structure of T_n that ensures the number of unique eigenvalues of the preconditioned matrix so it is expected that other parameter values behave in the same manner for the symmetrized system. The eigenvalues of the preconditioned matrix in this case are shown in Table 6.4.

Table 6.3: Condition numbers $\kappa(T_n)$ for a time-dependent linear ODE using the BDF2 method for T_n given by (6.3) and the numbers of iteration with MINRES and the preconditioner $|S_n|$ for $Y_n T_n$.

n	$\kappa(A_n)$	Iterations
10	29.33	6
100	67.49	6
1000	67.67	6

Table 6.4: Eigenvalues of the preconditioned system (to 4 decimal places). Repeated eigenvalues are shown in brackets with the number of repeated eigenvalues indicated.

n	Eigenvalues of $\lvert S_n\rvert^{-1} Y_n T_n$
10	$\{-1.0442, (-1 \times 3), -0.6781, 0.9219, (1 \times 3), 3.3921\}$
100	$\{-1.0610, (-1 \times 48), -0.4410, 0.9424, (1 \times 48), 2.2736\}$
1000	$\{-1.0610, (-1 \times 498), -0.4401, 0.9425, (1 \times 498), 2.2720\}$

6.3 Conclusions

We have shown that for simple nonsymmetric Toeplitz matrices with real entries the use of a simple trick gives symmetry so that convergence estimates for MINRES that are based only on eigenvalues rigorously apply.

It is further observed how this preconditioning can be applied in the context of time-stepping problems for ODEs and that convergence is achieved in a small number of iterations independent of the number of time-steps.

This approach for time-dependent problems may not be advantageous for such a simple problem considered here because MINRES requires matrix-vector products with T_n and Y_n as well as solution of a system with $|S_n|$ at each iteration. However, its potential for time-dependent problems for partial differential equations (PDEs) is more interesting. In [113], McDonald, Pestana, and Wathen employed the same techniques for time-dependent evolutionary PDEs problems, namely they symmetrized the all-at-once nonsymmetric block Toeplitz system resulting from discretising the PDE using a permutation matrix and then proposed an absolute value block circulant preconditioner for such symmetrized system. Rapid convergence that depends only on eigenvalues can then be guaranteed. We will briefly discuss how these techniques can be applied to block Toeplitz matrices in the next chapter and refer the readers to [113] for more discussions on time-dependent PDEs problems.

Chapter 7

Extension to block Toeplitz systems

In this chapter, we present our preliminary results on how the symmetrization technique and absolute value preconditioning described in Chapter 4 can be generalized to block Toeplitz systems.

A block Toeplitz matrix $T_{(n,m)} \in \mathbb{C}^{nm \times nm}$ is given by

$$T_{(n,m)} = \begin{bmatrix} A_{(0)} & A_{(-1)} & \cdots & A_{(-(n-1))} \\ A_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{(-1)} \\ A_{(n-1)} & \cdots & A_{(1)} & A_{(0)} \end{bmatrix}$$

with the blocks $A_{(k)} \in \mathbb{C}^{m \times m}$, $|k| \le n-1$. We consider in particular the following two types of block Toeplitz matrices: real block Toeplitz matrices with Toeplitz blocks (BTTB) and block Toeplitz matrices with commuting Hermitian blocks (BTHB).

We propose two simple permutation matrices that can symmetrize BTTB matrices and BTHB matrices, respectively. Then, we introduce their corresponding absolute value block circulant matrix that can be used as a preconditioner for these block Toeplitz systems. As a result, Krylov subspace methods like MINRES can be employed with guaranteed convergence that depends only on eigenvalues. Numerical tests are performed to support our results.

We remark that our results can be further extended to the multilevel Toeplitz case. However, we restrict our focus to block Toeplitz matrices in this chapter for the ubiquitous applications involving them.

7.1 Block Toeplitz matrices with Toeplitz blocks

We first consider block Toeplitz matrices with Toeplitz blocks (BTTB) in this section. Let $T_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be a BTTB matrix, namely

$$T_{(n,m)} = \begin{bmatrix} T_{(0)} & T_{(-1)} & \cdots & T_{(-(n-1))} \\ T_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_{(-1)} \\ T_{(n-1)} & \cdots & T_{(1)} & T_{(0)} \end{bmatrix},$$

where the blocks $T_{(k)} \in \mathbb{C}^{m\times m}$, $|k| \le n-1$, are Toeplitz matrices.

Like Toeplitz systems, BTTB systems have many crucial applications in numerical differential equations, networks, and image processing (see [25, 118, 92] for more examples).

We assume that the entries of the given BTTB matrix $T_{(n,m)}$ are denoted by

$$[T_{(n,m)}]_{p,q;r,s} = a_{p-q}^{(r-s)}$$

for $1 \le p, q \le m$ and $1 \le r, s \le n$. Also, $T_{(n,m)}$ is associated with its generating function $f(x,y)$ via the Fourier series

$$\sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_k^{(j)} e^{\mathbf{i}(jx+ky)}$$

defined on $[-\pi,\pi]\times[-\pi,\pi]$, where the Fourier coefficients $a_k^{(j)}$ are given by

$$a_k^{(j)} = \frac{1}{(2\pi)^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x,y)e^{-\mathbf{i}(jx+ky)}\,dxdy, \qquad j,k = 0, \pm1, \pm2, \ldots.$$

Several important properties of $T_{(n,m)}$ associated with f for any n and m are listed as follows.

- If f is real-valued, $T_{(n,m)}$ is Hermitian, i.e. $a_k^{(j)} = \bar{a}_{-k}^{(-j)}$.
- If f is real-valued with $f(x,y) = f(-x,-y)$, $T_{(n,m)}$ is symmetric, i.e. $a_k^{(j)} = a_{-k}^{(-j)}$.
- If f is real-valued with $f(x,y) = f(|x|,|y|)$, $T_{(n,m)}$ satisfies $a_k^{(j)} = a_{|k|}^{(|j|)}$.

For more properties of BTTB matrices, we refer to Chapter 5 of [25].

We now show that a real BTTB matrix $T_{(n,m)} \in \mathbb{R}^{nm\times nm}$ can be symmetrized by the permutation matrix

$$\begin{aligned} Y_{(n,m)} &= Y_n \otimes Y_m \in \mathbb{R}^{nm\times nm} \\ &= \begin{bmatrix} & & Y_m \\ & \iddots & \\ Y_m & & \end{bmatrix}. \end{aligned}$$

Theorem 7.1.1 *Let $T_{(n,m)} \in \mathbb{R}^{nm\times nm}$ be a real BTTB matrix, namely*

$$T_{(n,m)} = \begin{bmatrix} T_{(0)} & T_{(-1)} & \cdots & T_{(-(n-1))} \\ T_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_{(-1)} \\ T_{(n-1)} & \cdots & T_{(1)} & T_{(0)} \end{bmatrix},$$

where the blocks $T_{(k)} \in \mathbb{R}^{m\times m}$, $|k| \le n-1$, are real Toeplitz matrices and let $Y_{(n,m)} = Y_n \otimes Y_m \in \mathbb{R}^{nm\times nm}$. Then,

$$Y_{(n,m)}T_{(n,m)} = T_{(n,m)}^T Y_{(n,m)}.$$

Proof As the blocks $T_{(k)} \in \mathbb{R}^{m\times m}$, $|k| \le n-1$, are real Toeplitz matrices,

$$Y_m T_{(k)} = T_{(k)}^T Y_m, \qquad |k| \le n-1.$$

Thus,

$$\begin{aligned}
(Y_{(n,m)}T_{(n,m)})^T &= (\begin{bmatrix} & & & Y_m \\ & & \iddots & \\ & \iddots & & \\ Y_m & & & \end{bmatrix}\begin{bmatrix} T_{(0)} & T_{(-1)} & \cdots & T_{(-(n-1))} \\ T_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_{(-1)} \\ T_{(n-1)} & \cdots & T_{(1)} & T_{(0)} \end{bmatrix})^T \\
&= \begin{bmatrix} Y_mT_{(n-1)} & \cdots & Y_mT_{(1)} & Y_mT_{(0)} \\ \vdots & \iddots & \iddots & Y_mT_{(-1)} \\ Y_mT_{(1)} & \iddots & \iddots & \vdots \\ Y_mT_{(0)} & Y_mT_{(-1)} & \cdots & Y_mT_{(-(n-1))} \end{bmatrix}^T \\
&= \begin{bmatrix} (Y_mT_{(n-1)})^T & \cdots & (Y_mA_{(1)})^T & (Y_mA_{(0)})^T \\ \vdots & \iddots & \iddots & (Y_mT_{(-1)})^T \\ (Y_mT_{(1)})^T & \iddots & \iddots & \vdots \\ (Y_mT_{(0)})^T & (Y_mT_{(-1)})^T & \cdots & (Y_mA_{(-(n-1))}))^T \end{bmatrix} \\
&= \begin{bmatrix} Y_mT_{(n-1)} & \cdots & Y_mT_{(1)} & Y_mT_{(0)} \\ \vdots & \iddots & \iddots & Y_mT_{(-1)} \\ Y_mT_{(1)} & \iddots & \iddots & \vdots \\ Y_mT_{(0)} & Y_mT_{(-1)} & \cdots & Y_mT_{(-(n-1))} \end{bmatrix} \\
&= Y_{(n,m)}T_{(nm)}.
\end{aligned}$$

As $Y_{(n,m)}$ is symmetric,

$$\begin{aligned}
Y_{(n,m)}T_{(n,m)} &= (Y_{(n,m)}T_{(n,m)})^T \\
&= T_{(n,m)}^T Y_{(n,m)}^T \\
&= T_{(n,m)}^T Y_{(n,m)}. \quad \blacksquare
\end{aligned}$$

7.1.1 Block circulant matrices with circulant blocks

A block circulant matrix with circulant blocks (BCCB) $C_{(n,m)} \in \mathbb{C}^{nm\times nm}$ is given by

$$C_{(n,m)} = \begin{bmatrix} C_{(0)} & C_{(n-1)} & \cdots & C_{(1)} \\ C_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & C_{(n-1)} \\ C_{(n-1)} & \cdots & C_{(1)} & C_{(0)} \end{bmatrix},$$

where the blocks $C_{(k)} \in \mathbb{C}^{m\times m}$, $k \le n-1$, are circulant matrices.

Theorem 7.1.2 *[52, Theorem 5.8.1] Let $C_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be a BCCB matrix. Then, $C_{(n,m)}$ is given by*

$$C_{(n,m)} = (F_n \otimes F_m)^* \Lambda_{(n,m)} (F_n \otimes F_m),$$

where $F_n \in \mathbb{C}^{n\times n}$ is the Fourier matrix and $\Lambda_{(n,m)} \in \mathbb{C}^{nm\times nm}$ is the diagonal matrix in the eigendecomposition of $C_{(n,m)}$.

7.1.1.1 Diagonalization of BCCB matrices

Due to the diagonalization of BCCB matrices, one can easily show that for any vector $\mathbf{d}_{nm}$ the product $C^{-1}_{(n,m)}\mathbf{d}_{nm}$ (or $C_{(n,m)}\mathbf{d}_{nm}$) can be efficiently computed as follows.

Since the first column of $F_n \otimes F_m$ is $\frac{1}{\sqrt{nm}}\mathbf{1}_{nm}$, where $\mathbf{1}_{nm} = (1,1,1\ldots,1)^T \in \mathbb{R}^{nm}$, we have

$$(F_n \otimes F_m) C_{(n,m)} \boldsymbol{e}_1 = \frac{1}{\sqrt{nm}} \Lambda_{(n,m)} \mathbf{1}_{nm},$$

where $\boldsymbol{e}_1 = (1,0,\ldots,0)^T \in \mathbb{R}^{nm}$. Therefore, the diagonal matrix $\Lambda_{(n,m)}$ can be computed in $\mathcal{O}(nm \log mn)$ operations by taking a two dimensional FFT of the first column of $C_{(n,m)}$. Since $C^{-1}_{(n,m)} = (F_n \otimes F_m)^* \Lambda^{-1}_{(n,m)} (F_n \otimes F_m)$, $C^{-1}_{(n,m)}\mathbf{d}_{nm}$ for any vector $\mathbf{d}_{nm}$ can then be computed by several two dimensional FFTs in $\mathcal{O}(nm \log nm)$ operations once $\Lambda_{(n,m)}$ is obtained.

7.1.1.2 Absolute value BCCB matrices

The absolute BCCB matrices are readily well-defined, since BCCB matrices are diagonalizable by $F_n \otimes F_m$.

Definition 7.1.1 *Let $C_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be a BCCB matrix. The* absolute value BCCB matrix $|C_{(n,m)}| \in \mathbb{C}^{nm\times nm}$ *of $C_{(n,m)}$ is defined by*

$$\begin{aligned} |C_{(n,m)}| &= (C^*_{(n,m)} C_{(n,m)})^{1/2} \\ &= (C_{(n,m)} C^*_{(n,m)})^{1/2} \\ &= (F_n \otimes F_m)^* |\Lambda_{(n,m)}| (F_n \otimes F_m), \end{aligned}$$

where $F_n \in \mathbb{C}^{n\times n}$ is the Fourier matrix and $|\Lambda_{(n,m)}| \in \mathbb{R}^{nm\times nm}$ is the diagonal matrix in the eigendecomposition of $C_{(n,m)}$ with all entries replaced by their magnitude.

Remark $|C_{(n,m)}|$ is Hermitian positive definite by definition, provided that $C_{(n,m)}$ is nonsingular.

7.1.1.3 Optimal BCCB preconditioners

We let

$$\mathcal{M}_{F_n\otimes F_m} = \big\{(F_n \otimes F_m)^*\Lambda_{(n,m)}(F_n \otimes F_m) \mid \Lambda_{(n,m)} \text{ is any } nm\text{-by-}nm \text{ diagonal matrix}\big\}$$

be the set of all BCCB matrices (Section 5.8 of [52]). The *optimal BCCB preconditioner* [44] $c(T_{(n,m)}) \in \mathbb{C}^{nm\times nm}$ for $T_{(n,m)}$ is defined to be the minimizer of

$$\|T_{(n,m)} - C_{(n,m)}\|_F$$

over all $C_{(n,m)} \in \mathcal{M}_{F_n\otimes F_m}$.

Remark Optimal BCCB preconditioners are defined for general square matrices.

Let $\delta(A_{(n,m)}) \in \mathbb{C}^{nm\times nm}$ denote the diagonal matrix whose diagonal is equal to the diagonal of $A_{(n,m)}$. The following theorem provides some important properties of $c(T_{(n,m)})$.

Theorem 7.1.3 *[25, Theorem 5.6] Let $T_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be a BTTB matrix and let $c(T_{(n,m)}) \in \mathbb{C}^{nm\times nm}$ be the optimal BCCB preconditioner for $T_{(n,m)}$. Then, the followings hold:*

(i) $c(T_{(n,m)})$ is uniquely determined by $T_{(n,m)}$ and is given by

$$(F_n \otimes F_m)^*\delta\big((F_n \otimes F_m)T_{(n,m)}(F_n \otimes F_m)^*\big)(F_n \otimes F_m).$$

(ii) We have

$$\sigma_{max}\big(c(T_{(n,m)})\big) \le \sigma_{max}(T_{(n,m)}).$$

(iii) If $T_{(n,m)}$ is Hermitian, then $c(T_{(n,m)})$ is also Hermitian. Furthermore, we have

$$\lambda_{min}(T_{(n,m)}) \le \lambda_{min}\big(c(T_{(n,m)})\big) \le \lambda_{max}\big(c(T_{(n,m)})\big) \le \lambda_{max}(T_{(n,m)}).$$

7.1.2 Preconditioning for BTTB systems

Many preconditioners have been used for block Toeplitz systems, including block diagonal preconditioners and Schur complement preconditioners [50, 122]. For BTTB systems, band BTTB preconditioners were proposed by Serra-Capizzano [138, 57], Ng [117], and Jin [91]. BCCB matrices have also been used as preconditioners for BTTB systems for example by R. Chan and Jin [24], Ku and Kuo [104], and Jin [94, 90]. For a large class of BTTB matrices, e.g. the BTTB matrices generated by sparsely vanishing functions [159], optimal BCCB preconditioners for example are shown to be effective. A key to their success is the fact that the difference between a BTTB matrix and its corresponding optimal BCCB preconditioner can be decomposed into the sum of a low rank matrix and a small norm matrix for sufficiently large n and m.

Given a real BTTB matrix $T_{(n,m)} \in \mathbb{R}^{nm \times nm}$, we first symmetrize it using $Y_{(n,m)}$ and propose using an absolute value BCCB matrix $|C_{(n,m)}| \in \mathbb{R}^{nm \times nm}$ as a preconditioner for $Y_{(n,m)}T_{(n,m)}$. The following theorem, which generalizes Proposition 4.1 in [129] to the two dimensional block matrix case, accounts for the effectiveness of $|C_{(n,m)}|$.

Theorem 7.1.4 *Let $T_{(n,m)} \in \mathbb{R}^{nm \times nm}$ be a real BTTB matrix, let $C_{(n,m)} \in \mathbb{R}^{nm \times nm}$ be a real BCCB matrix, and let $Y_{(n,m)} = Y_n \otimes Y_m \in \mathbb{R}^{nm \times nm}$. Suppose for all $\epsilon > 0$ there exist positive integers N and M such that for all $n > N$ and all $m > M$*

$$C^{-1}_{(n,m)}T_{(n,m)} = I_{(n,m)} + \widetilde{R}_{(n,m)} + \widetilde{E}_{(n,m)},$$

where

$$rank(\widetilde{R}_{(n,m)}) \leq \mathcal{O}(n) + \mathcal{O}(m)$$

and

$$\|\widetilde{E}_{(n,m)}\|_2 \leq \epsilon.$$

Then,

$$|C_{(n,m)}|^{-1}Y_{(n,m)}T_{(n,m)} = Q_{(n,m)} + \widehat{R}_{(n,m)} + \widehat{E}_{(n,m)},$$

where $Q_{(n,m)}$ is symmetric and orthogonal,

$$rank(\widehat{R}_{(n,m)}) \leq \mathcal{O}(n) + \mathcal{O}(m),$$

and

$$\|\widehat{E}_{(n,m)}\|_2 \leq \epsilon.$$

Proof The proof echoes to that of Corollary 4.2.6. ∎

By setting the block dimension $m = 1$, we recover Proposition 4.1 of [129] in the following.

Corollary 7.1.5 *[129, Proposition 4.1] Let $T_n \in \mathbb{R}^{n\times n}$ be a real Toeplitz matrix, let $C_n \in \mathbb{R}^{n\times n}$ be a real circulant matrix, and let $Y_n \in \mathbb{R}^{n\times n}$ be the anti-identity matrix. Suppose for all $\epsilon > 0$ there exist positive integers N such that for all $n > N$*

$$C_n^{-1}T_n = I_n + \widetilde{R}_n + \widetilde{E}_n,$$

where

$$\mathit{rank}(\widetilde{R}_n) \le \mathcal{O}(n)$$

and

$$\|\widetilde{E}_n\|_2 \le \epsilon.$$

Then,

$$|C_n|^{-1}Y_nT_n = Q_n + \widehat{R}_n + \widehat{E}_n,$$

where Q_n is symmetric and orthogonal,

$$\mathit{rank}(\widehat{R}_n) \le \mathcal{O}(n),$$

and

$$\|\widehat{E}_{(n,m)}\|_2 \le \epsilon.$$

7.1.3 Symmetrizing other BTTB matrices

In the special case where the BTTB matrices are symmetric at the block level, we show that they can be symmetrized by another permutation matrix in addition to $Y_{(n,m)} = Y_n \otimes Y_m$. Namely, considering a real BTTB matrix $T_{(n,m)} \in \mathbb{R}^{nm\times nm}$ such that

$$T_{(n,m)} = \begin{bmatrix} T_{(0)} & T_{(1)} & \cdots & T_{(n-1)} \\ T_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_{(1)} \\ T_{(n-1)} & \cdots & T_{(1)} & T_{(0)} \end{bmatrix}$$

where the blocks $T_{(k)} \in \mathbb{R}^{m\times m}$, $k \le n-1$, are real Toeplitz matrices, we can symmetrize it using the permutation matrix

$$\begin{aligned} \mathscr{Y}_{(n,m)} &= I_n \otimes Y_m \in \mathbb{R}^{nm\times nm} \\ &= \begin{bmatrix} Y_m & & \\ & \ddots & \\ & & Y_m \end{bmatrix}. \end{aligned}$$

Theorem 7.1.6 *Let $T_{(n,m)} \in \mathbb{R}^{nm\times nm}$ be a real BTTB matrix such that*

$$T_{(n,m)} = \begin{bmatrix} T_{(0)} & T_{(1)} & \cdots & T_{(n-1)} \\ T_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_{(1)} \\ T_{(n-1)} & \cdots & T_{(1)} & T_{(0)} \end{bmatrix},$$

where the blocks $T_{(k)} \in \mathbb{R}^{m\times m}$, $k \le n-1$, are Toeplitz matrices and $\mathscr{Y}_{(n,m)} = I_n \otimes Y_m \in \mathbb{R}^{nm\times nm}$. Then,

$$\mathscr{Y}_{(n,m)} T_{(n,m)} = T^T_{(n,m)} \mathscr{Y}_{(n,m)}.$$

Proof As the blocks $T_{(k)} \in \mathbb{R}^{m\times m}$, $k \le n-1$, are real Toeplitz matrices,

$$Y_m T_{(k)} = T^T_{(k)} Y_m, \qquad k \le n-1.$$

Thus,

$$\begin{aligned}
(\mathscr{Y}_{(n,m)} T_{(n,m)})^T &= (\begin{bmatrix} Y_m & & & \\ & \ddots & & \\ & & \ddots & \\ & & & Y_m \end{bmatrix} \begin{bmatrix} T_{(0)} & T_{(1)} & \cdots & T_{(n-1)} \\ T_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & T_{(1)} \\ T_{(n-1)} & \cdots & T_{(1)} & T_{(0)} \end{bmatrix})^T \\
&= \begin{bmatrix} Y_m T_{(0)} & Y_m T_{(1)} & \cdots & Y_m T_{(n-1)} \\ Y_m T_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & Y_m T_{(1)} \\ Y_m T_{(n-1)} & \cdots & Y_m T_{(1)} & Y_m T_{(0)} \end{bmatrix}^T \\
&= \begin{bmatrix} (Y_m T_{(0)})^T & (Y_m T_{(1)})^T & \cdots & (Y_m T_{(n-1)})^T \\ (Y_m T_{(1)})^T & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & (Y_m T_{(1)})^T \\ (Y_m T_{(n-1)})^T & \cdots & (Y_m T_{(1)})^T & (Y_m T_{(0)})^T \end{bmatrix} \\
&= \begin{bmatrix} Y_m T_{(0)} & Y_m T_{(1)} & \cdots & Y_m T_{(n-1)} \\ Y_m T_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & Y_m T_{(1)} \\ Y_m T_{(n-1)} & \cdots & Y_m T_{(1)} & Y_m T_{(0)} \end{bmatrix} \\
&= \mathscr{Y}_{(n,m)} T_{(nm)}.
\end{aligned}$$

As $\mathscr{Y}_{(n,m)}$ is symmetric,

$$\begin{aligned}
\mathscr{Y}_{(n,m)} T_{(n,m)} &= (\mathscr{Y}_{(n,m)} T_{(n,m)})^T \\
&= T^T_{(n,m)} \mathscr{Y}^T_{(n,m)} \\
&= T^T_{(n,m)} \mathscr{Y}_{(n,m)}. \quad \blacksquare
\end{aligned}$$

We emphasize that $Y_{(n,m)} = Y_n \otimes Y_m$ is sufficient to symmetrize a general real BTTB matrix, including the one mentioned above. Hence, we only focus on using $Y_{(n,m)}$ in this chapter.

7.2 Block Toeplitz matrices with commuting Hermitian blocks

In this section, we consider block Toeplitz matrices with commuting Hermitian blocks (BTHB). Let $\mathcal{T}_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be a BTHB matrix, namely

$$\mathcal{T}_{(n,m)} = \begin{bmatrix} A_{(0)} & A_{(-1)} & \cdots & A_{(-(n-1))} \\ A_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{(-1)} \\ A_{(n-1)} & \cdots & A_{(1)} & A_{(0)} \end{bmatrix},$$

where the blocks $A_{(k)} \in \mathbb{C}^{m\times m}$, $|k| \leq n-1$, commute and are Hermitian (hence simultaneously diagonalizable). Such matrices can be found for example in evolutionary PDE problems [113]. Similar to the previous section, we propose here a way to symmetrize a BTHB matrices and their corresponding absolute value preconditioner.

We first show that one can obtain a Hermitian matrix by premultiplying $\mathcal{T}_{(n,m)}$ by the permutation matrix

$$\begin{aligned} \mathcal{Y}_{(n,m)} &= Y_n \otimes I_m \in \mathbb{R}^{nm\times nm} \\ &= \begin{bmatrix} & & I_m \\ & \iddots & \\ I_m & & \end{bmatrix}. \end{aligned}$$

Theorem 7.2.1 *Let $\mathcal{T}_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be a BTHB matrix, namely*

$$\mathcal{T}_{(n,m)} = \begin{bmatrix} A_{(0)} & A_{(-1)} & \cdots & A_{(-(n-1))} \\ A_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{(-1)} \\ A_{(n-1)} & \cdots & A_{(1)} & A_{(0)} \end{bmatrix},$$

where the block $A_{(k)} \in \mathbb{C}^{m\times m}$ satisfies $A^*_{(k)} = A_{(k)}$, $|k| \leq n-1$, and $\mathcal{Y}_{(n,m)} = Y_m \otimes I_n \in \mathbb{R}^{nm\times nm}$. Then,

$$\mathcal{Y}_{(n,m)}\mathcal{T}_{(n,m)} = \mathcal{T}^*_{(n,m)}\mathcal{Y}_{(n,m)}.$$

Proof As the blocks $A_{(k)} \in \mathbb{C}^{m\times m}$, $|k| \leq n-1$, are Hermitian matrices,

$$\begin{aligned}
(\mathcal{Y}_{(n,m)}\mathcal{T}_{(n,m)})^* &= (\begin{bmatrix} & & & I_m \\ & & \iddots & \\ & \iddots & & \\ I_m & & & \end{bmatrix}\begin{bmatrix} A_{(0)} & A_{(-1)} & \cdots & A_{(-(n-1))} \\ A_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{(-1)} \\ A_{(n-1)} & \cdots & A_{(1)} & A_{(0)} \end{bmatrix})^* \\
&= \begin{bmatrix} A_{(n-1)} & \cdots & A_{(1)} & A_{(0)} \\ \vdots & \iddots & \iddots & A_{(-1)} \\ A_{(1)} & \iddots & \iddots & \vdots \\ A_{(0)} & A_{(-1)} & \cdots & A_{(-(n-1))} \end{bmatrix}^* \\
&= \begin{bmatrix} A^*_{(n-1)} & \cdots & A^*_{(1)} & A^*_{(0)} \\ \vdots & \iddots & \iddots & A^*_{(-1)} \\ A^*_{(1)} & \iddots & \iddots & \vdots \\ A^*_{(0)} & A^*_{(-1)} & \cdots & A^*_{(-(n-1))} \end{bmatrix} \\
&= \begin{bmatrix} A_{(n-1)} & \cdots & A_{(1)} & A_{(0)} \\ \vdots & \iddots & \iddots & A_{(-1)} \\ A_{(1)} & \iddots & \iddots & \vdots \\ A_{(0)} & A_{(-1)} & \cdots & A_{(-(n-1))} \end{bmatrix} \\
&= \mathcal{Y}_{(n,m)}\mathcal{T}_{(n,m)}.
\end{aligned}$$

As $\mathcal{Y}_{(n,m)}$ is symmetric,

$$\begin{aligned}
\mathcal{Y}_{(n,m)}\mathcal{T}_{(n,m)} &= (\mathcal{Y}_{(n,m)}\mathcal{T}_{(n,m)})^* \\
&= \mathcal{T}^*_{(n,m)}\mathcal{Y}^*_{(n,m)} \\
&= \mathcal{T}^*_{(n,m)}\mathcal{Y}_{(n,m)}.
\end{aligned}$$

The result follows. ▮

7.2.1 Block circulant matrices with commuting Hermitian blocks

A block circulant matrix with commuting Hermitian blocks (BCHB) $\mathcal{C}_{(n,m)} \in \mathbb{C}^{nm\times nm}$ is given by

$$\mathcal{C}_{(n,m)} = \begin{bmatrix} A_{(0)} & A_{(n-1)} & \cdots & A_{(1)} \\ A_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{(n-1)} \\ A_{(n-1)} & \cdots & A_{(1)} & A_{(0)} \end{bmatrix},$$

where the blocks $A_{(k)} \in \mathbb{C}^{m\times m}$, $k \leq n-1$, commute and are Hermitian.

7.2.1.1 Diagonalization of BCHB matrices

Theorem 7.2.2 *Let $\mathcal{C}_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be a BCHB matrix. Then, $\mathcal{C}_{(n,m)}$ is given by*

$$\mathcal{C}_{(n,m)} = (F_n \otimes U_m)^* \Upsilon_{(n,m)} (F_n \otimes U_m),$$

where $F_n \in \mathbb{C}^{n\times n}$ is the Fourier matrix, $U_m \in \mathbb{C}^{m\times m}$ is a unitary matrix in the eigendecomposition of the blocks of $\mathcal{C}_{(n,m)}$, and $\Upsilon_{(n,m)} \in \mathbb{C}^{nm\times nm}$ is the diagonal matrix in the eigendecomposition of $\mathcal{C}_{(n,m)}$.

Proof Given a block circulant matrix $\mathcal{C}_{(n,m)} \in \mathbb{C}^{nm\times nm}$, it can be written as

$$\begin{aligned}
\mathcal{C}_{(n,m)} &= \begin{bmatrix} A_{(0)} & A_{(n-1)} & \cdots & A_{(1)} \\ A_{(1)} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & A_{(n-1)} \\ A_{(n-1)} & \cdots & A_{(1)} & A_{(0)} \end{bmatrix} \\
&= \sum_{k=0}^{n-1} \Pi_n^{-k} \otimes A_{(k)},
\end{aligned}$$

where

$$\Pi_n = \begin{bmatrix} & 1 & & \\ & & \ddots & \\ & & & 1 \\ 1 & & & \end{bmatrix} \in \mathbb{R}^{n\times n}$$

is the *fundamental circulant matrix.*

As the Hermitian blocks $A_{(k)} \in \mathbb{C}^{m\times m}$, $k \leq n-1$, commute by assumption, there exists a unitary matrix $U_m \in \mathbb{C}^{m\times m}$ such that $A_{(k)} = U_m^* \Lambda_{(k)} U_m$, $k \leq n-1$. Using the eigendecomposition $\Pi_n = F_n^* \Omega_n F_n$, where $F_n \in \mathbb{C}^{n\times n}$ is the Fourier matrix, we write $\mathcal{C}_{(n,m)}$ as

$$\begin{aligned}
\mathcal{C}_{(n,m)} &= \sum_{k=0}^{n-1} \Pi_n^{-k} \otimes A_{(k)} \\
&= \sum_{k=0}^{n-1} (F_n^* \Omega_n^{-k} F_n) \otimes (U_m^* \Lambda_{(k)} U_m) \\
&= (F_n \otimes U_m)^* \underbrace{(\sum_{k=0}^{n-1} \Omega_n^{-k} \otimes \Lambda_{(k)})}_{\Upsilon_{(n,m)}} (F_n \otimes U_m),
\end{aligned}$$

where $F_n \otimes U_m$ is unitary and $\Upsilon_{(n,m)}$ is the diagonal matrix in the eigendecomposition of $\mathcal{C}_{(n,m)}$. ▮

7.2.1.2 Absolute value BCHB matrices

As block circulant matrices do not commute in general, we require an additional property that all the blocks are simultaneously diagonalizable in order to define their absolute value block circulant matrix for our purposes.

Definition 7.2.1 *Let* $\mathcal{C}_{(n,m)} \in \mathbb{C}^{n\times n}$ *be a BCHB matrix. The* absolute value BCHB matrix $|\mathcal{C}_{(n,m)}| \in \mathbb{C}^{nm\times nm}$ *of* $\mathcal{C}_{(n,m)}$ *is defined by*

$$\begin{aligned} |\mathcal{C}_{(n,m)}| &= (\mathcal{C}^*_{(n,m)}\mathcal{C}_{(n,m)})^{1/2} \\ &= (\mathcal{C}_{(n,m)}\mathcal{C}^*_{(n,m)})^{1/2} \\ &= (F_n \otimes U_m)^*|\Upsilon_{(n,m)}|(F_n \otimes U_m), \end{aligned}$$

where $F_n \in \mathbb{C}^{n\times n}$ *is the Fourier matrix,* $U_m \in \mathbb{C}^{m\times m}$ *is the unitary matrix in the eigendecomposition of the blocks of* $\mathcal{C}_{(n,m)}$*, and* $|\Upsilon_{(n,m)}| \in \mathbb{R}^{nm\times nm}$ *is the diagonal matrix in the eigendecomposition of* $\mathcal{C}_{(n,m)}$ *with all entries replaced by their magnitude.*

Remark $|\mathcal{C}_{(n,m)}|$ is Hermitian positive definite by definition, provided that the blocks $A_{(k)}$, $k \le n-1$, are nonsingular.

7.2.1.3 Optimal block circulant preconditioners

For $A_{(n,m)} \in \mathbb{C}^{nm\times nm}$ partitioned as

$$A_{(n,m)} = \begin{bmatrix} A_{0,0} & A_{0,1} & \cdots & A_{0,n-1} \\ A_{1,0} & A_{1,1} & \cdots & A_{1,n-1} \\ \vdots & \ddots & \ddots & \vdots \\ A_{n-1,0} & A_{n-1,1} & \cdots & A_{n-1,n-1} \end{bmatrix},$$

where the blocks $A_{k,k} \in \mathbb{C}^{m\times m}$, $k \le n-1$, we define $\widetilde{\delta}(A_{(n,m)}) \in \mathbb{C}^{nm\times nm}$ as

$$\widetilde{\delta}(A_{(n,m)}) = \begin{bmatrix} A_{0,0} & & & \\ & A_{1,1} & & \\ & & \ddots & \\ & & & A_{n-1,n-1} \end{bmatrix}. \qquad (7.1)$$

Let $\widetilde{\mathcal{D}}_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be the set of all matrices of the form given by (7.1) and let

$$\widetilde{\mathcal{M}}_{F_n} = \{(F_n \otimes I_m)^*\widetilde{\Lambda}_{(n,m)}(F_n \otimes I_m) \mid \widetilde{\Lambda}_{(n,m)} \in \widetilde{\mathcal{D}}_{(n,m)}\}.$$

The *optimal block circulant preconditioner* [44] $\widetilde{c}(\mathcal{T}_{(n,m)}) \in \mathbb{C}^{nm\times nm}$ for $\mathcal{T}_{(n,m)}$ is defined to be the minimizer of

$$\|\mathcal{T}_{(n,m)} - \mathcal{W}_{(n,m)}\|_F$$

over all $\mathcal{W}_{(n,m)} \in \widetilde{\mathcal{M}}_{F_n}$.

Remark Optimal block circulant preconditioners are defined for general square matrices.

The following theorem provides some important properties of $\widetilde{c}(T_{(n,m)})$.

Theorem 7.2.3 *[25, Theorem 5.3] Let $\mathcal{T}_{(n,m)} \in \mathbb{C}^{nm \times nm}$ be a BTHB matrix and let $\widetilde{c}(\mathcal{T}_{(n,m)}) \in \mathbb{C}^{nm \times nm}$ be the optimal block circulant preconditioner for $\mathcal{T}_{(n,m)}$. Then, the followings hold:*

(i) $\widetilde{c}(\mathcal{T}_{(n,m)})$ is uniquely determined by $\mathcal{T}_{(n,m)}$ and is given by

$$(F_n \otimes I_m)^* \widetilde{\delta}\big((F_n \otimes I_m)\mathcal{T}_{(n,m)}(F_n \otimes I_m)^*\big)(F_n \otimes I_m).$$

(ii) We have

$$\sigma_{max}\big(\widetilde{c}(\mathcal{T}_{(n,m)})\big) \leq \sigma_{max}(\mathcal{T}_{(n,m)}).$$

(iii) If $\mathcal{T}_{(n,m)}$ is Hermitian, then $\widetilde{c}(\mathcal{T}_{(n,m)})$ is also Hermitian. Furthermore, we have

$$\lambda_{min}(\mathcal{T}_{(n,m)}) \leq \lambda_{min}\big(\widetilde{c}(\mathcal{T}_{(n,m)})\big) \leq \lambda_{max}\big(\widetilde{c}(\mathcal{T}_{(n,m)})\big) \leq \lambda_{max}(\mathcal{T}_{(n,m)}).$$

7.2.2 Preconditioning for BTHB systems

Like BTTB matrices, we have the following theorem for the symmetrized BTHB matrices on the effectiveness of absolute value BCHB preconditioners.

Theorem 7.2.4 *Let $\mathcal{T}_{(n,m)} \in \mathbb{C}^{nm \times nm}$ be a BTHB matrix, let $\mathcal{C}_{(n,m)} \in \mathbb{C}^{nm \times nm}$ be a BCHB matrix, and let $\mathcal{Y}_{(n,m)} = Y_n \otimes I_m \in \mathbb{R}^{nm \times nm}$. Suppose for all $\epsilon > 0$ there exist a positive integer N such that for all $n > N$ and all $m > 0$*

$$\mathcal{C}_{(n,m)}^{-1}\mathcal{T}_{(n,m)} = I_{(n,m)} + \widetilde{R}_{(n,m)} + \widetilde{E}_{(n,m)},$$

where

$$rank(\widetilde{R}_{(n,m)}) \leq \mathcal{O}(n)$$

and

$$\|\widetilde{E}_{(n,m)}\|_2 \leq \epsilon.$$

Then,

$$|\mathcal{C}_{(n,m)}|^{-1}\mathcal{Y}_{(n,m)}\mathcal{T}_{(n,m)} = Q_{(n,m)} + \widehat{R}_{(n,m)} + \widehat{E}_{(n,m)},$$

where $Q_{(n,m)}$ is Hermitian and unitary,

$$rank(\widehat{R}_{(n,m)}) \leq \mathcal{O}(n),$$

and

$$\|\widehat{E}_{(n,m)}\|_2 \leq \epsilon.$$

Proof The proof echoes to that of Theorem 7.1.4. ▮

Remark Note that $\text{rank}(\widetilde{R}_{(n,m)})$ is assumed to be bounded by $\mathcal{O}(n)$ instead of $\mathcal{O}(n)+\mathcal{O}(m)$ as in Theorem 7.1.4, since $\mathcal{C}_{(n,m)}$ approximates the blocks of $\mathcal{T}_{(n,m)}$ exactly. Similar results involving $\widetilde{c}(\mathcal{T}_{(n,m)})$ that satisfy such assumptions can be found for example in Theorem 5.8 and Corollary 5.9 of [25].

7.3 Numerical results

In this section, we demonstrate the effectiveness of $|C_{(n,m)}|$ and $|\mathcal{C}_{(n,m)}|$ for the system $T_{(n,m)}\mathbf{x}_{nm} = \mathbf{b}_{nm}$ and $\mathcal{T}_{(n,m)}\mathbf{x}_{nm} = \mathbf{b}_{nm}$, respectively, using MINRES and GMRES. The settings such as convergence criterion are the same as those used in Chapter 4.

Example 7.1. We first consider the symmetric indefinite BTTB matrix $T_{(n,m)} = T_n \otimes T_m \in \mathbb{R}^{nm\times nm}$, where $T_n \in \mathbb{R}^{n\times n}$ is generated by $f_1(x) = x^2 - \pi$ and $T_m \in \mathbb{R}^{m\times m}$ is generated by $f_2(x) = 2 + 2\cos x$. The preconditioner we used is the optimal BCCB matrix $|c(T_n \otimes T_m)|$.

Table 7.1 (a) and (b) show the numbers of iterations with MINRES and GMRES for the system, respectively. We observe that a reduction in iterations when the preconditioner is applied. Figure 7.1 shows that the clusters of eigenvalues around ±1 appear as n and m get large.

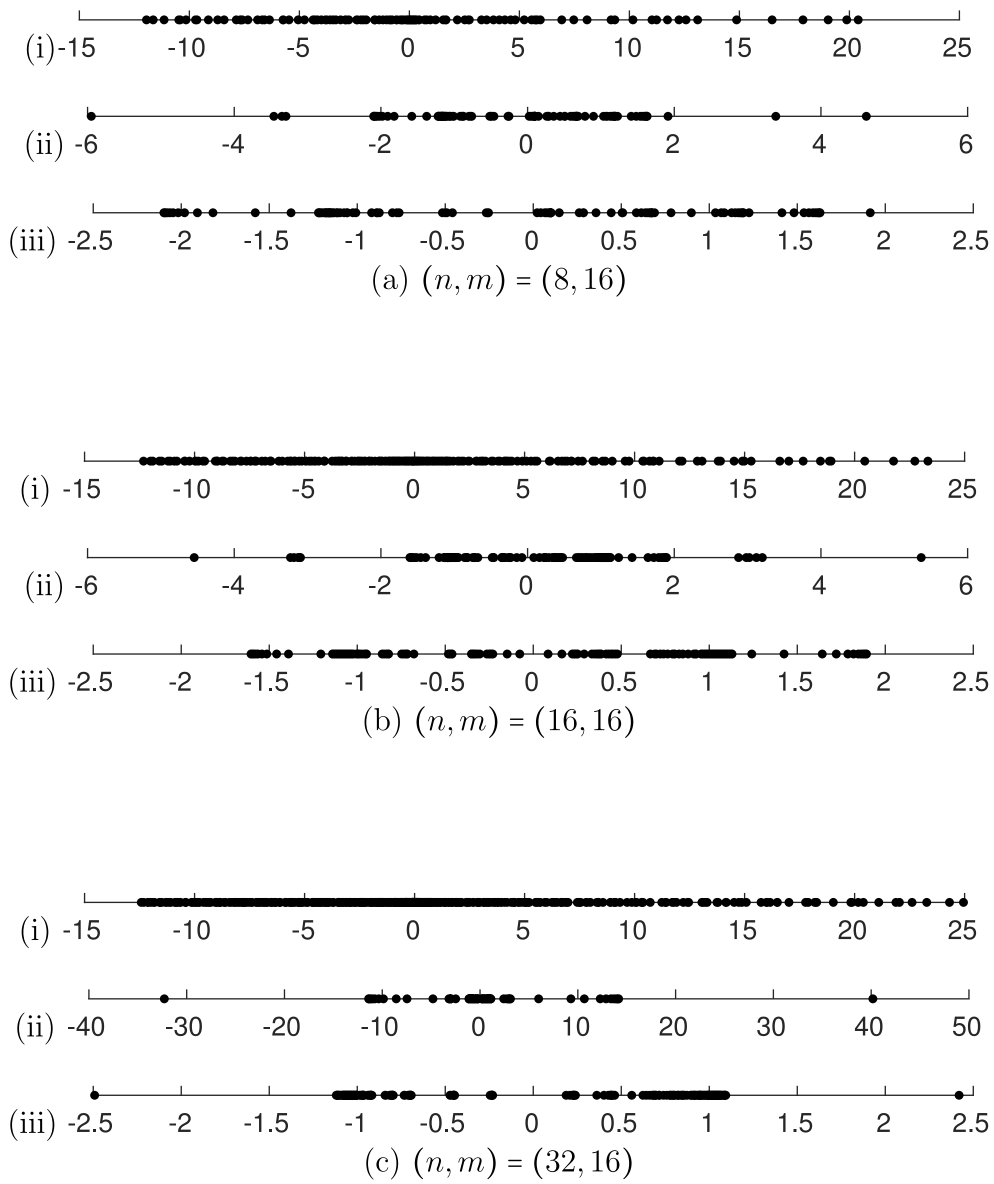
Table 7.1: Numbers of iterations with (a) MINRES and (b) GMRES for $T_{(n,m)}$ given in Example 7.1.

(a)

n	m	nm	with no preconditioner	with $\lvert c(T_n \otimes T_m)\rvert$
16	16	256	175	55
16	32	512	671	76
32	16	512	508	93
32	32	1024	1846	142

(b)

n	m	nm	with no preconditioner	with $c(T_n \otimes T_m)$
16	16	256	143	14
16	32	512	373	20
32	16	512	330	36
32	32	1024	789	51



(a) $(n, m) = (8, 16)$

(b) $(n, m) = (16, 16)$

(c) $(n, m) = (32, 16)$

Figure 7.1: Spectra of $T_{(n,m)}$ given in Example 7.1 at different (n, m) (i) with no preconditioner or (ii) with the preconditioner $|c(T_n \otimes T_m)|$. (iii) Zoom-in spectrum of (ii).

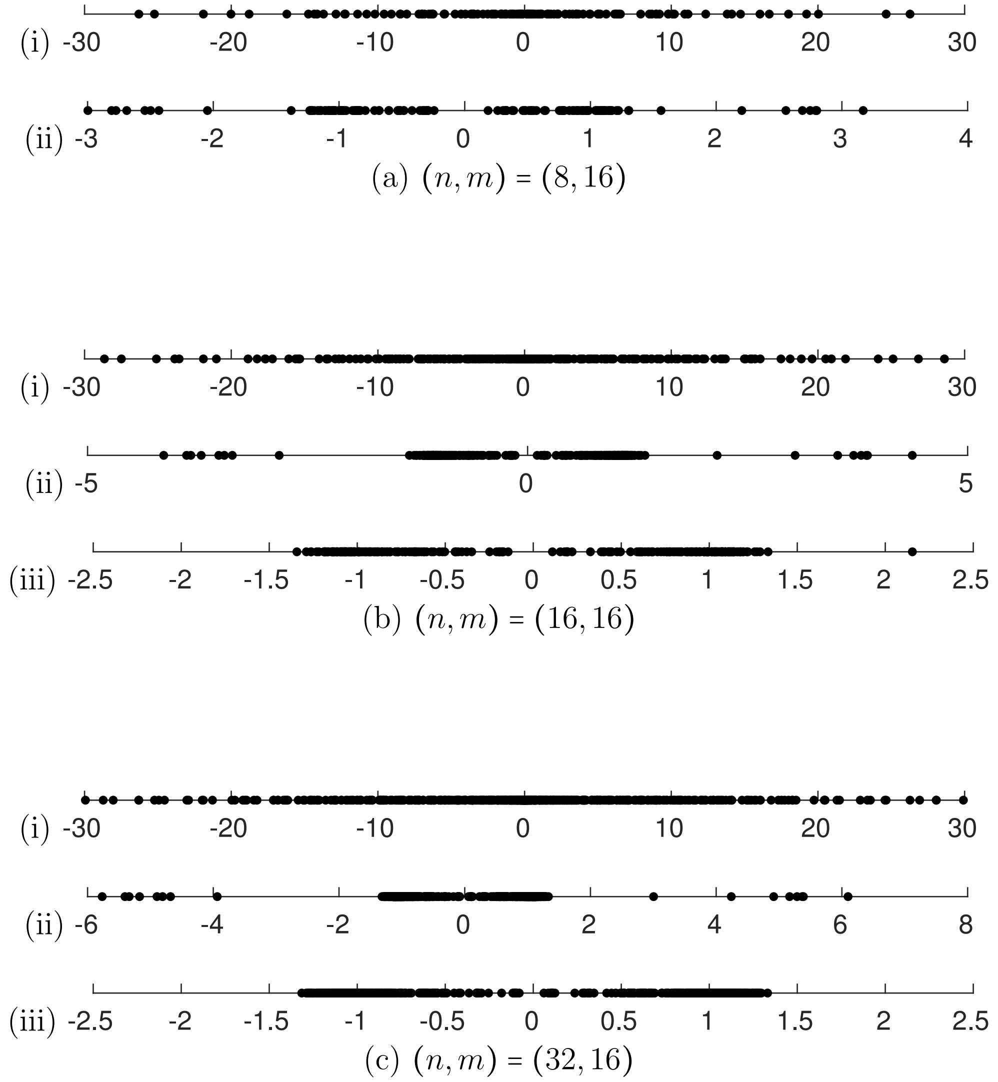
Example 7.2. We consider the nonsymmetric BTTB matrix $T_{(n,m)} = T_n \otimes T_m \in \mathbb{R}^{nm \times nm}$, where $T_n \in \mathbb{R}^{n \times n}$ is generated by $f(x) = x^2$ and $T_m \in \mathbb{R}^{m \times m}$ is the Grcar matrix given in (4.8). The preconditioner used here is the optimal BCCB matrix $|c(T_n \otimes T_m)|$.

Table 7.2 shows the numbers of iterations for the system and Figure 7.2 shows its spectrum at different (n, m). We again observe improved convergence when the preconditioner is applied. The eigenvalues of the preconditioned matrix are getting close to ±1 as the dimensions grow, even though they are not highly clustered. We will explain the loosely clusters of eigenvalues in the conclusion of this chapter.

Table 7.2: Numbers of iterations with (a) MINRES for $Y_{(n,m)}T_{(n,m)}$ and (b) GMRES for $T_{(n,m)}$ given in Example 7.2.

	n	m	nm	with no preconditioner	with $\lvert c(T_n \otimes T_m)\rvert$
	16	16	256	877	128
(a)	16	32	512	3308	148
	32	16	512	3848	199
	32	32	1024	14454	236

	n	m	nm	with no preconditioner	with $c(T_n \otimes T_m)$
	16	16	256	244	45
(b)	16	32	512	497	47
	32	16	512	500	57
	32	32	1024	1008	66



(a) $(n, m) = (8, 16)$

(b) $(n, m) = (16, 16)$

(c) $(n, m) = (32, 16)$

Figure 7.2: Spectra of $Y_{(n,m)}T_{(n,m)}$ with $T_{(n,m)}$ given in Example 7.2 at different (n, m) (i) with no preconditioner or (ii) with the preconditioner $|c(T_n \otimes T_m)|$. (iii) Zoom-in spectrum of (ii).

Example 7.3. We now consider a simple symmetric indefinite BTHB matrix, namely the block Toeplitz matrix with diagonal blocks $\mathcal{T}_{(n,m)} = T_n \otimes D_m$, where $T_n \in \mathbb{R}^{n\times n}$ is generated by $f(x) = x^2 - \pi$ and $D_m \in \mathbb{R}^{m\times m}$ is the diagonal matrix

$$D_m = \begin{bmatrix} 1 & & & & \\ & 1+\frac{100}{m} & & & \\ & & 1+2\cdot\frac{100}{m} & & \\ & & & \ddots & \\ & & & & 1+(m-1)\cdot\frac{100}{m} \end{bmatrix}. \tag{7.2}$$

As all diagonal matrices commute, D_m satisfies the requirement of being commuting Hermitian. Also, the preconditioner used is the optimal block circulant matrix $|\widetilde{c}(\mathcal{T}_{(n,m)})| = |c(T_n) \otimes D_m|$, where $c(T_n) \in \mathbb{R}^{n\times n}$ is the optimal circulant preconditioner for T_n.

Table 7.3 shows the iterations counts for MINRES and GMRES. Again, we see that the convergence is significantly improved with the proposed preconditioner. Figure 7.3 shows the expected clusters of eigenvalues around ±1 at different (n, m).

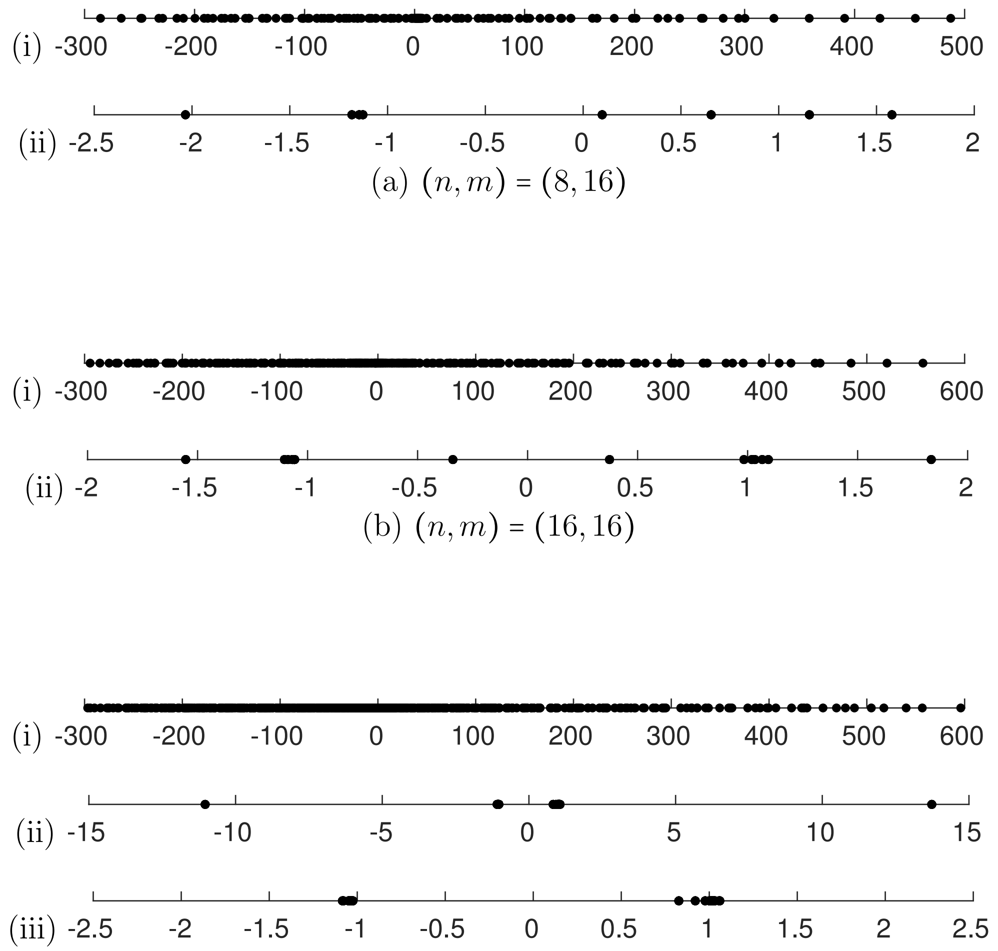
Table 7.3: Numbers of iterations with (a) MINRES and (b) GMRES for $\mathcal{T}_{(n,m)}$ given in Example 7.3 .

(a)

n	m	nm	with no preconditioner	with $\lvert c(T_n) \otimes D_m\rvert$
16	16	256	431	8
16	32	512	975	8
32	16	512	1192	11
32	32	1024	2403	11

(b)

n	m	nm	with no preconditioner	with $c(T_n) \otimes D_m$
16	16	256	220	6
16	32	512	438	6
32	16	512	451	6
32	32	1024	887	6



(a) $(n, m) = (8, 16)$

(b) $(n, m) = (16, 16)$

(c) $(n, m) = (32, 16)$

Figure 7.3: Spectra of $\mathcal{T}_{(n,m)}$ given in Example 7.3 at different (n, m) (i) with no preconditioner or (ii) with the preconditioner $|c(T_n) \otimes D_m|$. (iii) Zoom-in spectrum of (ii).

Example 7.4. We consider a nonsymmetric BTHB matrix - the block Toeplitz matrix with diagonal blocks $\mathcal{T}_{(n,m)} = T_n \otimes D_m$, where $T_n \in \mathbb{R}^{n \times n}$ is the Grcar matrix given by (4.8) and $D_m \in \mathbb{R}^{m \times m}$ is the diagonal matrix given by (7.2). Note that $\mathcal{T}_{(n,m)}$ in this case can be symmetrized by $\mathcal{Y}_{(n,m)}$, and the preconditioner used is the absolute value optimal block circulant matrix $|\widetilde{c}(\mathcal{T}_{(n,m)})| = |c(T_n) \otimes D_m|$.

Both Table 7.4 and Figure 7.4 show that our proposed preconditioner appears effective.

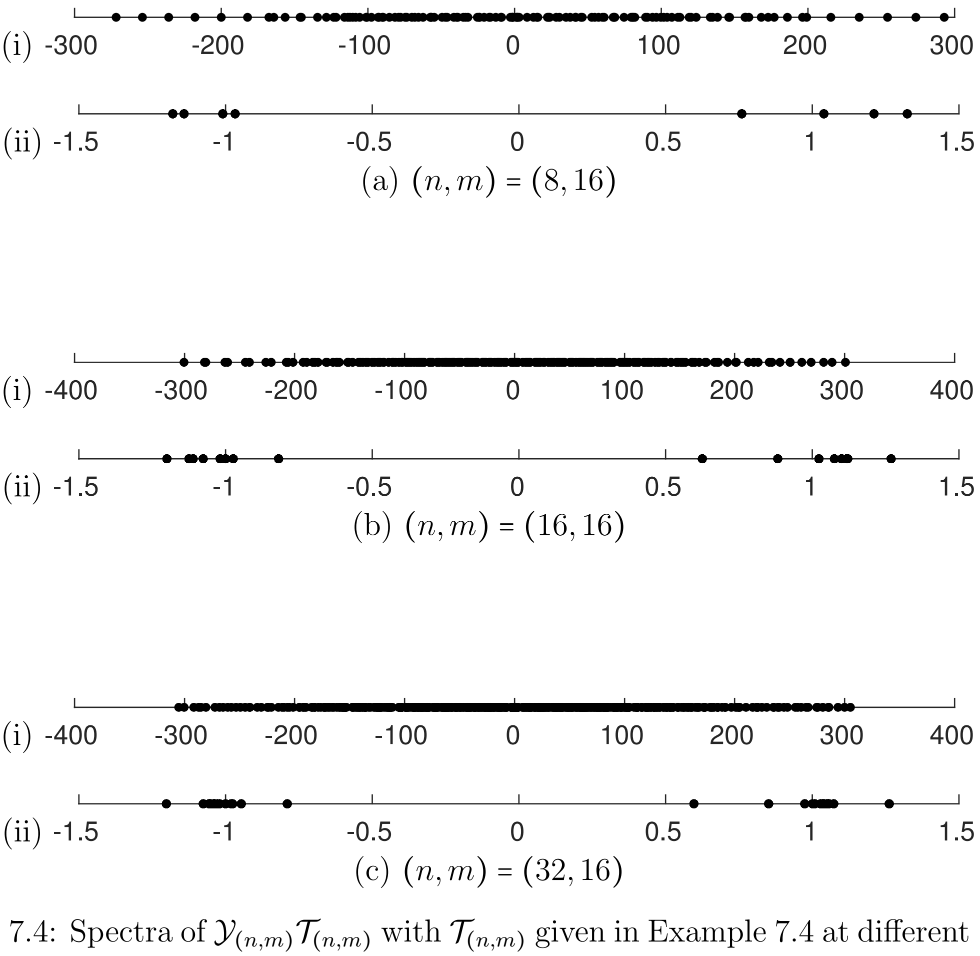
Table 7.4: Numbers of iterations with (a) MINRES for $\mathcal{Y}_{(n,m)}\mathcal{T}_{(n,m)}$ and (b) GMRES for $\mathcal{T}_{(n,m)}$ given in Example 7.4.

(a)

n	m	nm	with no preconditioner	with $\lvert c(T_n) \otimes D_m \rvert$
16	16	256	551	14
16	32	512	1145	14
32	16	512	1569	15
32	32	1024	2511	15

(b)

n	m	nm	with no preconditioner	with $c(T_n) \otimes D_m$
16	16	256	256	9
16	32	512	507	9
32	16	512	508	8
32	32	1024	950	8



(a) $(n, m) = (8, 16)$

(b) $(n, m) = (16, 16)$

(c) $(n, m) = (32, 16)$

Figure 7.4: Spectra of $\mathcal{Y}_{(n,m)}\mathcal{T}_{(n,m)}$ with $\mathcal{T}_{(n,m)}$ given in Example 7.4 at different (n, m) (i) with no preconditioner or (ii) with the preconditioner $|c(T_n) \otimes D_m|$.

7.4 Conclusions

We have discussed how the symmetrization technique and absolute value preconditioning can be extended to block Toeplitz matrices. Considering two special types of block Toeplitz matrices, we proposed two simple permutation matrices that can symmetrize them. The corresponding absolute value block circulant preconditioners were also provided. When the block dimension is equal to unity, our results recover those for the usual Toeplitz matrices.

Moreover, numerical examples have been given to illustrate our results. Even though our preconditioners are able to improve convergence, negative results were in fact shown by Serra-Capizzano and Tyrtyshnikov [143] that circulant type preconditioners are not optimal for multilevel Toeplitz systems. Namely, the eigenvalues of the preconditioned matrices are not tightly clustered. As a result, constructing other effective preconditioners like band block Toeplitz preconditioners will be a direction for future research.

Chapter 8

Conclusions

This thesis has aimed to extend the use of absolute value circulant preconditioners in the context of functions of Toeplitz matrices.

Chapter 2 covered not only the background on Toeplitz matrices but also the direct solvers and iterative solvers for them. Other aspects such as the asymptotic spectra and the circulant-based preconditioning techniques for Toeplitz matrices were also provided, together with the three commonly used circulant preconditioners, including optimal circulant preconditioners, superoptimal circulant preconditioners, and Strang's circulant preconditioners.

We then discussed the following Krylov subspace methods: CG, MINRES, and GMRES as well as the related convergence results concerning Toeplitz systems in Chapter 3.

In Chapters 4 and 5, we provided our main results on preconditioning for functions of Toeplitz matrices, proposing the use of absolute value circulant preconditioners for systems defined by analytic functions of Toeplitz matrices. A number of theorems that account for the effectiveness of optimal circulant preconditioners, superoptimal circulant preconditioners, and Strang's circulant preconditioners in this context were also given, with numerical examples to support our results.

Later in Chapter 6, we explored how our results on the symmetrization technique and absolute value preconditioning can be applied in time-stepping methods for ODE/PDE problems.

In Chapter 7, we showed our preliminary results on preconditioning for block Toeplitz systems. Considering two kinds of block Toeplitz matrices: block Toeplitz matrices with Toeplitz blocks (BTTB) and block Toeplitz matrices with commuting Hermitian blocks (BTHB), we demonstrated that they can be symmetrized by a permutation matrix. We also defined the related absolute value block circulant preconditioners and illustrated their effectiveness via a number of numerical tests.

8.1 Future work

In the course of my research, I noticed the following directions for future research.

8.1.1 Asymptotic spectral distribution of symmetrized Toeplitz matrices[1]

In the previous chapters, we showed that for sufficiently large n we have the following matrix decomposition

$$Y_n T_n = Y_n C_n - Y_n R_n - Y_n E_n,$$

where $Y_n R_n$ is a low rank matrix and $Y_n E_n$ is a small norm matrix.

From such decomposition, we can see that the singular values of $Y_n T_n[f]$ with $f \in L^1([-\pi, \pi])$ are in fact distributed as $|f|$, since premultiplying $T_n[f]$ by an unitary matrix (Y_n) does not alter its singular values. In other words, as $Y_n T_n[f]$ is symmetric, the eigenvalues of $\{Y_n T_n[f]\}_n$ are essentially distributed as $|f|$ up to ± sign.

The following simple example illustrates my observation: we consider

$$T_n[f] = \begin{bmatrix} 2 & & & \\ 1 & 2 & & \\ & \ddots & \ddots & \\ & & 1 & 2 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

generated by $f(x) = 2 + e^{\mathbf{i}x}$ and $n = 512$. In Figure 8.1 (a), we see that the singular values of $Y_n T_n[f]$ are distributed as $|f(x)| = \sqrt{5 + 4\cos x}$. In Figure 8.1 (b), we observe that the eigenvalues of $Y_n T_n[f]$ are distributed essentially as $\pm|f|$. Such bidiagonal matrix arises for example in time stepping methods for ODEs, as discussed in Chapter 6.

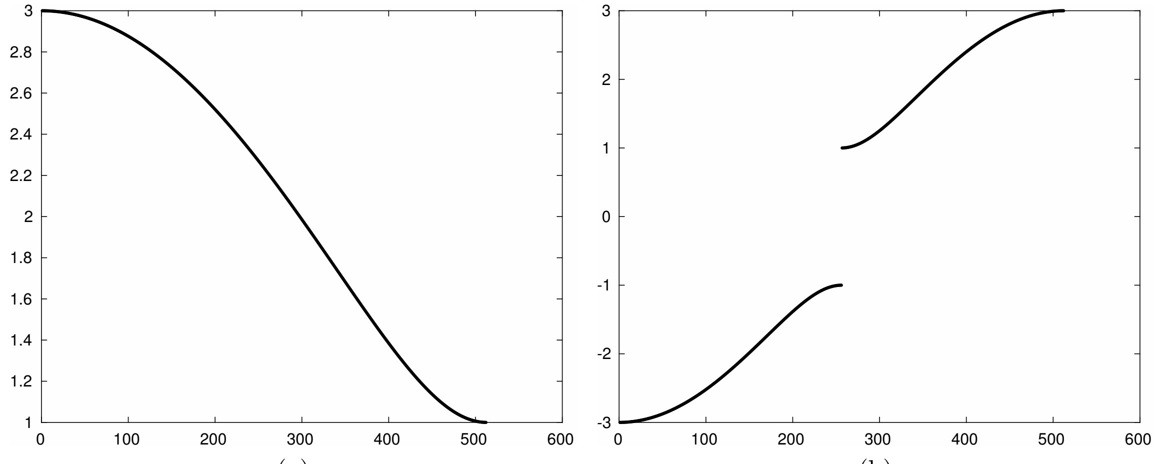
The following theorem provides a relationship with $Y_n T_n$ and $Y_n C_n$ that could help explain my observation.

Theorem 8.1.1 *Let $C_n \in \mathbb{R}^{n \times n}$ be a circulant matrix and let $Y_n \in \mathbb{R}^{n \times n}$ be the anti-identity matrix. Then,*

$$Y_n C_n = U_n^* \Sigma_n |\Lambda_n| U_n,$$

where $U_n \in \mathbb{C}^{n \times n}$ is a unitary matrix and $\Sigma_n |\Lambda_n| \in \mathbb{R}^{n \times n}$ is the diagonal matrix in the eigendecomposition of $Y_n C_n$, having only eigenvalues $\pm|\lambda_j|$ with λ_j being the j-th eigenvalue of C_n.

[1] This subsection is partially adapted from [83, 62], which is joint work mainly with Mohammad Ayman Mursaleen and Stefano Serra-Capizzano.



(a)

(b)

Figure 8.1: (a) Singular value and (b) spectral distribution of $Y_nT_n[f]$ with $f(x) = 2 + e^{\mathbf{i}x}$ at $n = 512$.

Proof Recalling the definition of absolute circulant matrices $|C_n| = F_n^*|\Lambda_n|F_n \in \mathbb{R}^{n\times n}$, we have

$$\begin{aligned} Y_nC_n &= Y_nF_n^*\Lambda_nF_n \\ &= Y_n \underbrace{F_n^*\widetilde{\Lambda}_nF_n}_{\widetilde{C}_n} F_n^*|\Lambda_n|F_n \\ &= \underbrace{Y_n\widetilde{C}_n}_{Q_n} |C_n| \\ &= Q_n|C_n|, \end{aligned}$$

where $\widetilde{\Lambda}_n \in \mathbb{C}^{n\times n}$ is the diagonal matrix having the sign of eigenvalues of $\Lambda_n \in \mathbb{C}^{n\times n}$ as its eigenvalues. Note that $|C_n|$ is symmetric by Definition 2.3.1.

Moreover, $Y_n|C_n|$ is symmetric as $|C_n|$ itself is a real Toeplitz matrix, i.e.

$$\begin{aligned} Y_n|C_n| &= (Y_n|C_n|)^T \\ &= |C_n|^TY_n^T \\ &= |C_n|Y_n. \end{aligned}$$

Similarly, we know that $Q_n = Y_n\widetilde{C}_n$ is also symmetric.

As any circulant matrices commute, we have

$$\begin{aligned} Q_n|C_n| &= Y_n\widetilde{C}_n|C_n| \\ &= Y_n|C_n|\widetilde{C}_n \\ &= |C_n|Y_n\widetilde{C}_n \\ &= |C_n|Q_n, \end{aligned}$$

namely Q_n and $|C_n|$ commute. Therefore, there exists a unitary matrix $U_n \in \mathbb{C}^{n\times n}$ such that both $U_n Q_n U_n^*$ and $U_n|C_n|U_n^*$ are diagonal.

Furthermore, Q_n is orthogonal as

$$\begin{aligned}
Q_n^T Q_n &= (Y_n\widetilde{C}_n)^T(Y_n\widetilde{C}_n) \\
&= \widetilde{C}_n^T Y_n^T Y_n \widetilde{C}_n \\
&= \widetilde{C}_n^T \widetilde{C}_n \\
&= (F_n^*\widetilde{\Lambda}_n F_n)^*(F_n^*\widetilde{\Lambda}_n F_n) \\
&= F_n^* \underbrace{|\widetilde{\Lambda}_n|^2}_{I_n} F_n \\
&= I_n.
\end{aligned}$$

With Q_n being both symmetric and orthogonal, we have $Q_n = U_n^*\Sigma_n U_n$ where Σ_n, having only eigenvalues ±1, is the diagonal matrix of the eigenvalue decomposition of Q_n.

As a result,

$$\begin{aligned}
Y_n C_n &= Q_n|C_n| \\
&= U_n^*\Sigma_n|\Lambda_n|U_n,
\end{aligned}$$

where $\Sigma_n|\Lambda_n|$ is the diagonal matrix in the eigenvalue decomposition of Y_nC_n, having only eigenvalues $\pm|\lambda_j|$ with λ_j being the j-th eigenvalue of C_n. The eigenvalues of Y_nC_n are therefore all of the form $\pm|\lambda_j|$ while its singular values are of $|\lambda_j|$. ∎

In other words, for sufficiently large n, we have

$$\begin{aligned}
Y_n T_n &= Y_n C_n - Y_n R_n - Y_n E_n \\
&= U_n^*\Sigma_n|\Lambda_n|U_n - Y_n R_n - Y_n E_n.
\end{aligned}$$

Assuming $\widetilde{C}_n^{1/2}$ exists and is real, we have Σ_n is similar to Y_n (i.e. $\Sigma_n = (\widetilde{C}_n^{1/2}U_n)^* Y_n \widetilde{C}_n^{1/2} U_n)$. Thus, there are roughly half of the eigenvalues of Y_nC_n are positive/negative provided that $|\Lambda_n|$ has only nonzero entries. The same result also holds for Y_nT_n, since both Y_nC_n and Y_nT_n have the same asymptotic spectral distribution (see Theorem 6.1 in [160]). As a result, Y_nT_n is always indefinite for sufficiently large n.

Like nonsymmetric Toeplitz matrices, the asymptotic spectral distribution of block Toeplitz matrices can be investigated in a similar fashion.

Theorem 8.1.2 *Let $C_{(n,m)} \in \mathbb{R}^{nm\times nm}$ be a block circulant matrix with circulant blocks and let $Y_{(n,m)} = Y_m \otimes Y_n \in \mathbb{R}^{nm\times nm}$. Then,*

$$Y_{(n,m)}C_{(n,m)} = U^*_{(n,m)}\Sigma_{(n,m)}|\Lambda_{(n,m)}|U_{(n,m)},$$

where $U_{(n,m)} \in \mathbb{C}^{nm\times nm}$ is a unitary matrix and $\Sigma_{(n,m)}|\Lambda_{(n,m)}| \in \mathbb{R}^{nm\times nm}$ is the diagonal matrix in the eigendecomposition of $Y_{(n,m)}C_{(n,m)}$, having only eigenvalues $\pm|\lambda_j|$ with λ_j being the j-th eigenvalue of $C_{(n,m)}$.

Theorem 8.1.3 *Let $\mathcal{C}_{(n,m)} \in \mathbb{C}^{nm\times nm}$ be a block circulant matrix with commuting Hermitian blocks $A_{(k)} \in \mathbb{C}^{m\times m}$, $|k| \le n-1$, and let $\mathcal{Y}_{(n,m)} = Y_m \otimes I_n \in \mathbb{R}^{nm\times nm}$. Then,*

$$\mathcal{Y}_{(n,m)}\mathcal{C}_{(n,m)} = U^*_{(n,m)}\Sigma_{(n,m)}|\Upsilon_{(n,m)}|U_{(n,m)},$$

where $U_{(n,m)} \in \mathbb{C}^{nm\times nm}$ is a unitary matrix and $\Sigma_{(n,m)}|\Upsilon_{(n,m)}| \in \mathbb{R}^{nm\times nm}$ is the diagonal matrix in the eigendecomposition of $\mathcal{Y}_{(n,m)}\mathcal{C}_{(n,m)}$, having only eigenvalues $\pm|\lambda_j|$ with λ_j being the j-th eigenvalue of $\mathcal{C}_{(n,m)}$.

8.1.2 Preconditioning for block Toeplitz systems

As discussed in Chapter 7, both the symmetrization technique and absolute value preconditioners can be extended to block Toeplitz systems. The BTHB systems mentioned have certain applications for example in [113] on time-dependent PDEs problems. We remark that the all-at-once BTHB matrix $T_{(n,m)}$ examined in [113] is of a relatively simple block lower triangular structure, i.e.

$$\mathcal{T}_{(n,m)} = \begin{bmatrix} A_{(0)} & & & & & \\ A_{(1)} & \ddots & & & & \\ \vdots & \ddots & \ddots & & & \\ A_{(p-1)} & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & A_{(p-1)} & \cdots & A_{(1)} & A_{(0)} \end{bmatrix}.$$

More complicated block Toeplitz systems could be found in other applications.

Hence, further developing absolute value preconditioning techniques for other block Toeplitz systems is a direction for future research.

8.1.3 Matrix-vector multiplication $h(T_n)\mathbf{d}_n$

We recall that a matrix-vector multiplication is required at each iteration of Krylov subspace methods. Although it was shown $\mathcal{O}(n\log n)$ complexity is achievable for certain Toeplitz matrix exponential e^{T_n} [107, 102], the multiplication $h(T_n)\mathbf{d}_n$ for any vector $\mathbf{d}_n$, as mentioned in Chapter 4, is not readily computed efficiently in general.

Therefore, developing efficient numerical methods for computing $h(T_n)\mathbf{d}_n$ will be future work. If such efficient computation is possible, fast iterative solvers for functions of Toeplitz matrices can be achieved.

8.1.4 Preconditioning for nonsymmetric ill-conditioned Toeplitz systems

For symmetric ill-conditioned Toeplitz systems, band-Toeplitz preconditioners [28, 35, 139, 140] have been shown to be effective as discussed in Section 2.4. A direction for future research will be developing band-Toeplitz preconditioners for nonsymmetric ill-conditioned Toeplitz systems. A possible approach is to first symmetrize a nonsymmetric Toeplitz matrix with the anti-identity matrix Y_n and then design an effective band-Toeplitz preconditioner that can give clustered spectra at ±1 for the symmetrized matrix.

Appendix A

Notation

A.1 Vector/Function spaces

- $\mathbb{C}$: The complex plane
- $\mathbb{C}^n$: The space of n-dimensional complex vectors
- $\mathbb{C}^{m\times n}$: The space of $m \times n$ complex matrices
- $\mathbb{R}$: The real plane
- $\mathbb{R}^n$: The space of n-dimensional real vectors
- $\mathbb{R}^{m\times n}$: The space of $m \times n$ real matrices
- $\mathcal{C}[-\pi,\pi]$: The Banach space of continuous complex-valued functions defined on $[-\pi,\pi]$ with the supremum norm $\|\cdot\|_\infty$
- $L^p[-\pi,\pi]$, $1 \le p < \infty$: The Banach space of functions f equipped with the norm

$$\|f\|_p = \left(\int_{-\pi}^{\pi} |f(x)|^p \, dx\right)^{\frac{1}{p}} < \infty$$

- $L^\infty[-\pi,\pi]$: The Banach space of functions f equipped with the essential supremum norm

A.2 Krylov subspaces

- $\mathcal{K}_k(A_n, \mathbf{r}_n^{(0)})$: The Krylov subspace of dimension n of $A_n \in \mathbb{C}^{n\times n}$ and an initial residual $\mathbf{r}_n^{(0)}$
- Π_k: The set of polynomials of degree at most k
- CG: The conjugate gradient method
- GMRES: The generalized minimal residual method
- MINRES: The minimal residual method

A.3 Linear algebra

- $\mathrm{diag}(a_1, a_2, \dots)$: A diagonal matrix having $a_1, a_2, \dots$ as its entries
- I_n: The $n \times n$ identity matrix
- F_n: The $n \times n$ Fourier matrix
- Y_n: The $n \times n$ anti-identity matrix
- T_n: An $n \times n$ Toeplitz matrix
- C_n: An $n \times n$ circulant matrix
- $|C_n|$: The $n \times n$ absolute value circulant matrix of $C_n \in \mathbb{C}^{n\times n}$
- $c(T_n)$: The $n \times n$ optimal circulant preconditioner for $T_n \in \mathbb{C}^{n\times n}$
- $t(T_n)$: The $n \times n$ superoptimal circulant preconditioner for $T_n \in \mathbb{C}^{n\times n}$
- $s(T_n)$: Strang's $n \times n$ circulant preconditioner for $T_n \in \mathbb{C}^{n\times n}$
- $T_{(n,m)}$: An $nm \times nm$ block Toeplitz matrix with Toeplitz blocks
- $C_{(n,m)}$: An $nm \times nm$ circulant matrix with circulant blocks (BCCB)
- $|C_{(n,m)}|$: The $nm \times nm$ absolute value BCCB matrix of $C_{(n,m)} \in \mathbb{C}^{nm\times nm}$
- $c(T_{(n,m)})$: The $nm \times nm$ optimal BCCB preconditioner for $T_{(n,m)} \in \mathbb{C}^{nm\times nm}$
- $\mathcal{T}_{(n,m)}$: An $nm \times nm$ block Toeplitz matrix with commuting Hermitian blocks

- $\mathcal{C}_{(n,m)}$: An $nm \times nm$ block circulant matrix with commuting Hermitian blocks (BCHB)
- $|\mathcal{C}_{(n,m)}|$: The $nm \times nm$ absolute value BCHB matrix of $\mathcal{C}_{(n,m)} \in \mathbb{C}^{nm\times nm}$
- $\widetilde{c}(\mathcal{T}_{(n,m)})$: The $nm \times nm$ optimal block circulant preconditioner for $\mathcal{T}_{(n,m)} \in \mathbb{C}^{nm\times nm}$
- $\delta(A_n)$: The diagonal matrix whose diagonal is equal to the diagonal of the matrix $A_n \in \mathbb{C}^{n\times n}$
- $\Lambda(A_n)$: The spectrum of the matrix $A_n \in \mathbb{C}^{n\times n}$
- $\lambda_{\max}(A_n)$: The largest eigenvalue of the matrix $A_n \in \mathbb{C}^{n\times n}$
- $\lambda_{\min}(A_n)$: The smallest eigenvalue of the matrix $A_n \in \mathbb{C}^{n\times n}$
- $\sigma_{\max}(A_n)$: The largest singular value of the matrix $A_n \in \mathbb{C}^{n\times n}$
- $\sigma_{\min}(A_n)$: The smallest singular value of the matrix $A_n \in \mathbb{C}^{n\times n}$
- $\kappa(A_n)$: The condition number of the matrix $A_n \in \mathbb{C}^{n\times n}$
- $\rho(A_n)$: The spectral radius of the matrix $A_n \in \mathbb{C}^{n\times n}$
- $\mathrm{rank}(A)$: The rank of the matrix $A \in \mathbb{C}^{m\times n}$
- $\mathrm{span}(\mathbf{v}_1, \mathbf{v}_2, \dots)$: The span of the vectors $\mathbf{v}_1, \mathbf{v}_2, \dots$
- $A \otimes B$: The Kronecker product of the matrices A and B
- A^T: The transpose of the matrix $A \in \mathbb{C}^{m\times n}$
- A^*: The conjugate transpose of the matrix $A \in \mathbb{C}^{m\times n}$
- A_n^{-1}: The inverse of the matrix $A_n \in \mathbb{C}^{n\times n}$

A.4 Complex analysis

- $h(z)$: An analytic function

Appendix B

Linear algebra results

B.1 Matrix norms

A function $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$ is called a *matrix norm* [85] if for all $A_n, B_n \in \mathbb{C}^{n\times n}$ it satisfies the following axioms:

- $\|A_n\| \geq 0$ (Nonnegative)
- $\|A_n\| = 0$ if and only if $A_n = 0$ (Positive)
- $\|cA_n\| = |c|\|A_n\|$ for all complex scalars c (Homogeneous)
- $\|A_n + B_n\| \leq \|A_n\| + \|B_n\|$ (Triangle inequality)
- $\|A_nB_n\| \leq \|A_n\|\|B_n\|$ (Submultiplicative).

Theorem B.1.1 *[85] Let $A \in \mathbb{C}^{m\times n}$ and let $\|\cdot\|$ be a vector norm on $\mathbb{C}^n$. We define $\|\cdot\|$ on $\mathbb{C}^{m\times n}$ by*

$$\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|.$$

For example, we have the following matrix norms. In each case, we take $A = [a_{ij}] \in \mathbb{C}^{m\times n}$.

The *maximum column sum matrix norm*

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|.$$

The *maximum row sum matrix norm*

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{i=1}^{n} |a_{ij}|.$$

The *spectral norm*

$$\|A\|_2 = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^*A\}.$$

An important matrix norm that is not induced by a vector norm is the Frobenius norm.

Definition B.1.1 *[85] Let $A \in \mathbb{C}^{m\times n}$. The Frobenius norm of A is defined by*

$$\|A\|_F = \left(\sum_{j=1}^{n}\sum_{i=1}^{m} |a_{ij}|^2\right)^{\frac{1}{2}}.$$

The above matrix norms satisfy the following properties:

- $\|A\|_2 \le \|A\|_F \le \sqrt{n}\|A\|_2$
- $\frac{1}{\sqrt{m}}\|A\|_1 \le \|A\|_2 \le \sqrt{n}\|A\|_1$
- $\frac{1}{\sqrt{n}}\|A\|_\infty \le \|A\|_2 \le \sqrt{m}\|A\|_\infty$
- $\|A\|_2 \le \sqrt{\|A\|_1\|A\|_\infty}$.

Definition B.1.2 *[85] A matrix norm $\|\cdot\|$ on $\mathbb{C}^{m\times n}$ is said to be unitarily invariant if*

$$\|UAV\| = \|A\|$$

for all $A \in \mathbb{C}^{m\times n}$ and for all unitary matrices $U \in \mathbb{C}^{m\times m}, V \in \mathbb{C}^{n\times n}$.

Remark Both the Frobenius norm and the spectral norm are unitarily invariant.

Definition B.1.3 *[85] Let $A_n \in \mathbb{C}^{n\times n}$. The condition number $\kappa(A_n)$ of A_n is defined by*

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

Definition B.1.4 *[85] Let $A_n \in \mathbb{C}^{n\times n}$. The spectral radius $\rho(A_n)$ of A_n is defined by*

$$\rho(A_n) = \max\{|\lambda| : \lambda \textit{ is an eigenvalue of } A_n\}.$$

Theorem B.1.2 *[85] Let $A_n \in \mathbb{C}^{n\times n}$. If $\|\cdot\|$ is any matrix norm, then*

$$\rho(A_n) \le \|A_n\|.$$

B.2 Singular values and eigenvalues

Theorem B.2.1 *[85, Cauchy's interlacing theorem] Let $A_n, B_n \in \mathbb{C}^{n\times n}$ be Hermitian and let the eigenvalues of $A_n, B_n, A_n + B_n$ be arranged in an increasing order. Then, for every pair of integers j, k such that $1 \le j, k \le n$ and $j + k \ge n + 1$*

$$\lambda_{j+k-n}(A_n + B_n) \le \lambda_j(A_n) + \lambda_k(B_n)$$

and for every pair of integers j, k such that $1 \le j, k \le n$ and $j + k \le n + 1$

$$\lambda_j(A_n) + \lambda_k(B_n) \le \lambda_{j+k-1}(A_n + B_n).$$

Theorem B.2.2 *[15, Proposition 2] Let $\sigma \in \mathbb{R}, w \in \mathbb{C}^n$ with $\|w\| = 1$ and define $H := w\sigma w^*$. Let $X \in \mathbb{R}^{n\times n}$ be diagonal with exactly t distinct eigenvalues denoted by λ_i $(1 \le i \le t)$. Write P_i for the orthogonal projection onto the eigenspace χ_i corresponding to λ_i and define θ_i by*

$$\|P_i w\| = \cos\theta_i \qquad \textit{where } \theta_i \textit{ is chosen in } [-\frac{\pi}{2}, \frac{\pi}{2}].$$

Define yet the $t \times t$ matrix

$$Y = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_t \end{bmatrix} + y\sigma y^*, \qquad \textit{where } y = \begin{bmatrix} \cos\theta_1 \\ \vdots \\ \cos\theta_t \end{bmatrix}.$$

Then, the eigenvalues of $X + H$ are the eigenvalues of X, of which single copies of multiple eigenvalues are replaced by the eigenvalues of Y.

Theorem B.2.3 *[15, Corollary 3] If in the conditions of Theorem B.2.2 H is a Hermitian rank-k matrix (for $1 \le k \le n$), then the eigenvalues of $X + H$ are the eigenvalues of X of which at most k copies of each different eigenvalue have been perturbed. Moreover, at most kt of the total number of eigenvalues are perturbed.*

B.3 Kronecker products

Definition B.3.1 *The Kronecker product of $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$ is defined by*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{C}^{mp\times nq}.$$

The Kronecker product satisfies the following properties:

- $A \otimes (B + C) = A \otimes B + A \otimes C$
- $(A + B) \otimes C = A \otimes C + A \otimes B$
- $(kA) \otimes B = A \otimes (kB) = k(A \otimes B)$
- $(A \otimes B) \otimes C = A \otimes (B \otimes C)$
- $(A \otimes B)^* = A^* \otimes B^*$,

where A, B, and C are matrices and k is a scalar.

Moreover, if A, B, C, and D are matrices of appropriate size that can form the matrix products AC and BD, we have the mixed-product property

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD).$$

B.4 Fourier matrices

Definition B.4.1 *The Fourier matrix $F_n \in \mathbb{C}^{n\times n}$ is defined by*

$$F_n = \frac{1}{\sqrt{n}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^{2\cdot 2} & \cdots & \omega_n^{2\cdot(n-1)} \\ \vdots & \cdots & \cdots & \cdots & \cdots \\ 1 & \omega_n^{n-1} & \omega_n^{(n-1)\cdot 2} & \cdots & \omega_n^{(n-1)\cdot(n-1)} \end{bmatrix},$$

where $\omega_n = e^{2\pi \mathbf{i}/n}$.

Remark Note that F_n is unitary.

B.5 Fast Fourier transforms

In this section, we provide an algebraic approach to drive a fast Fourier transform (FFT), namely an algorithm to compute the matrix-vector product $F_n \mathbf{d}_n$ efficiently.

The key idea is to factorize

$$\overline{F}_n = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_n & \omega_n^2 & \cdots & \omega_n^{n-1} \\ 1 & \omega_n^2 & \omega_n^{2\cdot 2} & \cdots & \omega_n^{2\cdot(n-1)} \\ \vdots & \cdots & \cdots & \cdots & \cdots \\ 1 & \omega_n^{n-1} & \omega_n^{(n-1)\cdot 2} & \cdots & \omega_n^{(n-1)\cdot(n-1)} \end{bmatrix} \in \mathbb{C}^{n\times n},$$

to reveal its relation with $\overline{F}_{n/2}$. Assuming n is even, simple calculations yield

$$\overline{F}_n = \begin{bmatrix} I_{n/2} & D_{n/2} \\ I_{n/2} & -D_{n/2} \end{bmatrix} \begin{bmatrix} \overline{F}_{n/2} & \\ & \overline{F}_{n/2} \end{bmatrix} \Pi_n,$$

where $D_{n/2} \in \mathbb{C}^{n/2 \times n/2}$ is the diagonal matrix with entries $(1, \omega_n, \omega_n^2, \ldots, \omega_n^{(n/2)-1})$ and $\Pi_n \in \mathbb{R}^{n \times n}$ is the permutation matrix that separates the incoming vector into its even and odd parts.

A simple example demonstrates this idea. When $n = 4$, we have

$$\begin{aligned} \overline{F}_4 &= \begin{bmatrix} I_2 & D_2 \\ I_2 & -D_2 \end{bmatrix} \begin{bmatrix} \overline{F}_2 & \\ & \overline{F}_2 \end{bmatrix} \Pi_4 \\ &= \begin{bmatrix} 1 & & 1 & \\ & 1 & & \mathbf{i} \\ 1 & & -1 & \\ & 1 & & -\mathbf{i} \end{bmatrix} \begin{bmatrix} 1 & 1 & & \\ 1 & \mathbf{i}^2 & & \\ & & 1 & 1 \\ & & 1 & \mathbf{i}^2 \end{bmatrix} \begin{bmatrix} 1 & & & \\ & & 1 & \\ & 1 & & \\ & & & 1 \end{bmatrix}. \end{aligned}$$

Therefore, the original transform with $\overline{F}_n$ reduces to two similar transforms with $\overline{F}_{n/2}$, which involve $n/2$ multiplications and n additions/subtractions. Consequently, we repeat the same process and count the number of work needed recursively. The complexity of the final work count turns out to be $\mathcal{O}(nd) = \mathcal{O}(n \log n)$ provided that $n = 2^d$. Therefore, the matrix-vector product $F_n \mathbf{d}_n$ can be computed in $\mathcal{O}(n \log n)$ operations. For FFTs with an arbitrary positive integer n and other related subjects, we refer to [166, 147, 124].

B.6 Commuting matrices

Definition B.6.1 *[85] Let $A_n, B_n \in \mathbb{C}^{n \times n}$. We call A_n and B_n simultaneously diagonalizable if there is a single similarity matrix $S_n \in \mathbb{C}^{n \times n}$ such that both $S_n^{-1} A_n S_n$ and $S_n^{-1} B_n S_n$ are diagonal.*

Theorem B.6.1 *[85] Let $A_n, B_n \in \mathbb{C}^{n \times n}$ and let $\mathfrak{F}$ be a given family of Hermitian matrices. There exists a unitary matrix U_n such that $U_n A_n U_n^*$ is diagonal for all $A_n \in \mathfrak{F}$ if and only if $A_n B_n = B_n A_n$ for all $A_n, B_n \in \mathfrak{F}$.*

Appendix C

Approximation theory results

The following selected results from approximation theory are used in this thesis.

C.1 Polynomial approximations

Definition C.1.1 *[101] A trigonometric polynomial on $[-\pi,\pi]$ is an expression of the form*

$$p(x) = \sum_{k=-M}^{M} a_k e^{\mathbf{i}kx}.$$

The numbers k are called the frequencies of $p(x)$ and the largest integer k such that $|a_k| + |a_{-k}| \neq 0$ is called the degree of $p(x)$.

Theorem C.1.1 *[133, Weierstrass approximation theorem] For any $f \in \mathcal{C}[-\pi,\pi]$ and for any $\epsilon > 0$, there exists a polynomial on $[-\pi,\pi]$ such that*

$$\|f(x) - p(x)\|_\infty \le \epsilon$$

for all $x \in [-\pi,\pi]$.

C.2 Fourier series

Definition C.2.1 *[101] The Fourier series $S[f]$ of a function $f \in L^1([-\pi,\pi])$ is the trigonometric series*

$$S[f] = \sum_{k=-\infty}^{\infty} a_k e^{\mathbf{i}kx},$$

where the Fourier coefficients

$$a_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x) e^{-\mathbf{i}kx}\, dt, \qquad k = 0, \pm 1, \pm 2, \ldots.$$

Bibliography

[1] G. Ammar and W. Gragg. Superfast solution of real positive definite Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 9(1):61–76, 1988.

[2] W. Arnoldi. The principle of minimized iteration in the solution of the matrix eigenvalue problem. *Quarterly of Applied Mathematics*, 9:17–29, 1951.

[3] F. Avram. On bilinear forms in Gaussian random variables and Toeplitz matrices. *Probability Theory and Related Fields*, 79(1):37–45, 1988.

[4] O. Axelsson. *Iterative solution methods.* Cambridge University Press, Cambridge, 1994.

[5] O. Axelsson and V. Barker. *Finite element solution of boundary value problems*, volume 35 of *Classics in Applied Mathematics.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[6] O. Axelsson and G. Lindskog. On the rate of convergence of the preconditioned conjugate gradient method. *Numerische Mathematik*, 48:499–524, 1986.

[7] Z. Bai, X. Jin, and T. Yao. Superoptimal preconditioners for functions of matrices. *Numerical Mathematics: Theory, Methods and Applications*, 8(4):515–529, 2015.

[8] E. Bareiss. Numerical solution of linear equations with Toeplitz and vector Toeplitz matrices. *Numerische Mathematik*, 13(5):404–424, 1969.

[9] M. Benzi. Preconditioning techniques for large linear systems: a survey. *Journal of Computational Physics*, 182(2):418–477, 2002.

[10] D. Bini and F. Benedetto. A new preconditioner for the parallel solution of positive definite Toeplitz systems. In *Proceedings of the Second Annual ACM Symposium on Parallel Algorithms and Architectures*, SPAA '90, pages 220–223. ACM, 1990.

[11] D. Bini, S. Dendievel, G. Latouche, and B. Meini. Computing the exponential of large block-triangular block-Toeplitz matrices encountered in fluid queues. *Linear Algebra and its Applications*, 502(Supplement C):387–419, 2016.

[12] D. Bini and P. Favati. On a matrix algebra related to the discrete Hartley transform. *SIAM Journal on Matrix Analysis and Applications*, 14(2):500–507, 1993.

[13] R. Bitmead and B. Anderson. Asymptotically fast solution of Toeplitz and related systems of linear equations. *Linear Algebra and its Applications*, 34:103–116, 1980.

[14] E. Boman and I. Koltracht. Fast transform based preconditioners for Toeplitz equations. *SIAM Journal on Matrix Analysis and Applications*, 16(2):628–645, 1995.

[15] J. Brandts and R. Reis da Silva. Computable eigenvalue bounds for rank-k perturbations. *Linear Algebra and its Applications*, 432(12):3100–3116, 2010.

[16] R. Brent, F. Gustavson, and D. Yun. Fast solution of Toeplitz systems of equations and computation of Padé approximants. *Journal of Algorithms*, 1(3):259–295, 1980.

[17] W. Briggs, V. Henson, and S. McCormick. *A multigrid tutorial.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2000.

[18] J. Bunch. Stability of methods for solving Toeplitz systems of equations. *SIAM Journal on Scientific and Statistical Computing*, 6(2):349–364, 1985.

[19] M. Cai and X. Jin. A note on T. Chan's preconditioner. *Linear Algebra and its Applications*, 376:283–290, 2004.

[20] M. Cai, X. Jin, and Y. Wei. A generalization of T. Chan's preconditioner. *Linear Algebra and its Applications*, 407:11–18, 2005.

[21] R. Chan. Circulant preconditioners for Hermitian Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 10(4):542–550, 1989.

[22] R. Chan. The spectrum of a family of circulant preconditioned Toeplitz systems. *SIAM Journal on Numerical Analysis*, 26(2):503–506, 1989.

[23] R. Chan, T. Chan, and C. Wong. Cosine transform based preconditioners for total variation deblurring. *IEEE Transactions on Image Processing*, 8(10):1472–1478, 1999.

[24] R. Chan and X. Jin. A family of block preconditioners for block systems. *SIAM Journal on Scientific and Statistical Computing*, 13(5):1218–1235, 1992.

[25] R. Chan and X. Jin. *An introduction to iterative Toeplitz solvers*, volume 5 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

[26] R. Chan, X. Jin, and M. Yeung. The circulant operator in the Banach algebra of matrices. *Linear Algebra and its Applications*, 149:41–53, 1991.

[27] R. Chan, X. Jin, and M. Yeung. The spectra of super-optimal circulant preconditioned Toeplitz systems. *SIAM Journal on Numerical Analysis*, 28(3):871–879, 1991.

[28] R. Chan and K. Ng. Toeplitz preconditioners for Hermitian Toeplitz systems. *Linear Algebra and its Applications*, 190:181–208, 1993.

[29] R. Chan and M. Ng. Conjugate gradient methods for Toeplitz systems. *SIAM Review*, 38(3):427–482, 1996.

[30] R. Chan, M. Ng, and R. Plemmons. Generalization of Strang's preconditioner with applications to Toeplitz least squares problems. *Numerical Linear Algebra with Applications*, 3(1):45–64, 1996.

[31] R. Chan, M. Ng, and C. Wong. Sine transform based preconditioners for symmetric Toeplitz systems. *Linear Algebra and its Applications*, 232:237–259, 1996.

[32] R. Chan, M. Ng, and A. Yip. The best circulant preconditioners for Hermitian Toeplitz systems II: the multiple-zero case. *Numerische Mathematik*, 92(1):17–40, 2002.

[33] R. Chan, D. Potts, and G. Steidl. Preconditioners for nondefinite Hermitian Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 22(3):647–665, 2001.

[34] R. Chan and G. Strang. Toeplitz equations by conjugate gradients with circulant preconditioner. *SIAM Journal on Scientific and Statistical Computing*, 10(1):104–119, 1989.

[35] R. Chan and P. Tang. Fast band-Toeplitz preconditioners for Hermitian Toeplitz systems. *SIAM Journal on Scientific Computing*, 15(1):164–171, 1994.

[36] R. Chan and M. Yeung. Circulant preconditioners for Toeplitz matrices with positive continuous generating functions. *Mathematics of Computation*, 58(197):233–240, 1992.

[37] R. Chan and M. Yeung. Jackson's theorem and circulant preconditioned Toeplitz systems. *Journal of Approximation Theory*, 70(2):191–205, 1992.

[38] R. Chan and M. Yeung. Circulant preconditioners for complex Toeplitz matrices. *SIAM Journal on Numerical Analysis*, 30(4):1193–1207, 1993.

[39] R. Chan, A. Yip, and M. Ng. The best circulant preconditioners for Hermitian Toeplitz systems. *SIAM Journal on Numerical Analysis*, 38(3):876–896, 2000.

[40] T. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM Journal on Scientific and Statistical Computing*, 9(4):766–771, 1988.

[41] T. Chan and P. Hansen. A look-ahead Levinson algorithm for general Toeplitz systems. *IEEE Transactions on Signal Processing*, 40(5):1079–1090, 1992.

[42] T. Chan and P. Hansen. A look-ahead Levinson algorithm for indefinite Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 13(2):490–506, 1992.

[43] T. Chan and T. Mathew. Domain decomposition algorithms. *Acta Numerica*, 3:61–143, 1994.

[44] T. Chan and J. Olkin. Circulant preconditioners for Toeplitz-block matrices. *Numerical Algorithms*, 6(1):89–101, 1994.

[45] S. Chandrasekaran, M. Gu, X. Sun, J. Xia, and J. Zhu. A superfast algorithm for Toeplitz systems of linear equations. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1247–1266, 2008.

[46] J. Chen and X. Jin. The generalized superoptimal preconditioner. *Linear Algebra and its Applications*, 432(1):203–217, 2010.

[47] K. Chen. *Matrix preconditioning techniques and applications*, volume 19 of *Cambridge Monographs on Applied and Computational Mathematics.* Cambridge University Press, Cambridge, 2005.

[48] C. Cheng and X. Jin. Some stability properties of T. Chan's preconditioner. *Linear Algebra and its Applications*, 395:361–365, 2005.

[49] C. Cheng, X. Jin, S. Vong, and W. Wang. A note on spectra of optimal and superoptimal preconditioned matrices. *Linear Algebra and its Applications*, 422(2):482–485, 2007.

[50] W. Ching, M. Ng, and Y. Wen. Block diagonal and Schur complement preconditioners for block-Toeplitz systems with small size blocks. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1101–1119, 2008.

[51] J. Cooley and J. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19:297–301, 1965.

[52] P. Davis. *Circulant matrices.* John Wiley & Sons, New York-Chichester-Brisbane, 1979.

[53] F. de Hoog. A new algorithm for solving Toeplitz systems of equations. *Linear Algebra and its Applications*, 88:123–138, 1987.

[54] P. Delsarte, Y. Genin, and Y. Kamp. A generalization of the Levinson algorithm for Hermitian Toeplitz matrices with any rank profile. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(4):964–971, 1985.

[55] F. Di Benedetto. Analysis of preconditioning techniques for ill-conditioned Toeplitz matrices. *SIAM Journal on Scientific Computing*, 16(3):682–697, 1995.

[56] F. Di Benedetto and S. Serra Capizzano. A unifying approach to abstract matrix algebra preconditioning. *Numerische Mathematik*, 82(1):57–90, 1999.

[57] M. Donatelli, C. Garoni, M. Mazza, S. Serra-Capizzano, and D. Sesana. Spectral behavior of preconditioned non-Hermitian multilevel block Toeplitz matrices with matrix-valued symbol. *Applied Mathematics and Computation*, 245:158–173, 2014.

[58] D. Duffy. *Finite Difference Methods in Financial Engineering: A Partial Differential Equation Approach.* The Wiley Finance Series. Wiley, 2013.

[59] J. Duintjer Tebbens and G. Meurant. Any Ritz value behavior is possible for Arnoldi and for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 33(3):958–978, 2012.

[60] J. Duintjer Tebbens and G. Meurant. Prescribing the behavior of early terminating GMRES and Arnoldi iterations. *Numerical Algorithms*, 65(1):69–90, 2014.

[61] H. Elman, D. Silvester, and A. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, second edition, 2014.

[62] P. Ferrari, I. Furci, S. Hon, M. Ayman Mursaleen, and S. Serra-Capizzano. The eigenvalue distribution of special 2-by-2 block matrix sequences, with applications to the case of symmetrized Toeplitz structures. *ArXiv e-prints*, 2018.

[63] B. Fischer. *Polynomial based iteration methods for symmetric linear systems*, volume 68 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2011.

[64] R. Freund and N. Nachtigal. QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numerische Mathematik*, 60(3):315–339, 1991.

[65] R. Freund and H. Zha. Formally biorthogonal polynomials and a look-ahead Levinson algorithm for general Toeplitz systems. *Linear Algebra and its Applications*, 188:255–303, 1993.

[66] W. Freund. A look-ahead Bareiss algorithm for general Toeplitz matrices. *Numerische Mathematik*, 68(1):35–69, 1994.

[67] C. Garoni and S. Serra Capizzano. *Generalized locally Toeplitz sequences: theory and applications. Vol. I.* Springer, Cham, 2017.

[68] C. Garoni, S. Serra Capizzano, and P. Vassalos. A general tool for determining the asymptotic spectral distribution of Hermitian matrix-sequences. *Operators and Matrices*, 9(3):549–561, 2015.

[69] I. Gohberg, T. Kailath, and V. Olshevsky. Fast Gaussian elimination with partial pivoting for matrices with displacement structure. *Mathematics of Computation*, 64(212):1557–1576, 1995.

[70] G. Golub and C. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, fourth edition, 2013.

[71] A. Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[72] A. Greenbaum, V. Pták, and Z. Strakoš. Any nonincreasing convergence curve is possible for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 17(3):465–469, 1996.

[73] U. Grenander and G. Szegő. *Toeplitz forms and their applications*. Chelsea Publishing Co., New York, second edition, 1984.

[74] M. Gu. Stable and efficient algorithms for structured systems of linear equations. *SIAM Journal on Matrix Analysis and Applications*, 19(2):279–306, 1998.

[75] W. Hackbusch. *Iterative solution of large sparse systems of equations*, volume 95 of *Applied Mathematical Sciences*. Springer, Cham, second edition, 2016.

[76] M. Hanke and J. Nagy. Toeplitz approximate inverse preconditioner for banded Toeplitz matrices. *Numerical Algorithms*, 7(2-4):183–199, 1994.

[77] G. Heinig. Inversion of generalized Cauchy matrices and other classes of structured matrices. In Adam Bojanczyk and George Cybenko, editors, *Linear Algebra for Signal Processing*, pages 63–81, New York, NY, 1995. Springer, New York.

[78] M. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49:409–436, 1952.

[79] N. Higham. *Functions of matrices*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

[80] N. Higham and P. Kandolf. Computing the action of trigonometric and hyperbolic matrix functions. *SIAM Journal on Scientific Computing*, 39(2):A613–A627, 2017.

[81] S. Hon. Optimal preconditioners for systems defined by functions of Toeplitz matrices. *Linear Algebra and its Applications*, 548:148–171, 2018.

[82] S. Hon. Circulant preconditioners for functions of Hermitian Toeplitz matrices. *Journal of Computational and Applied Mathematics*, 352:328–340, 2019.

[83] S. Hon, M. Ayman Mursaleen, and S. Serra-Capizzano. A note on the spectral distribution of symmetrized Toeplitz sequences. *ArXiv e-prints*, 2018.

[84] S. Hon and A. Wathen. Circulant preconditioners for analytic functions of Toeplitz matrices. *Numerical Algorithms*, 79(4):1211–1230, 2018.

[85] R. Horn and C. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, second edition, 2013.

[86] T. Huckle. Circulant and skewcirculant matrices for solving Toeplitz matrix problems. *SIAM Journal on Matrix Analysis and Applications*, 13(3):767–777, 1992.

[87] T. Huckle. Fast transforms for tridiagonal linear equations. *BIT Numerical Mathematics*, 34(1):99–112, 1994.

[88] T. Huckle, S. Serra Capizzano, and C. Tablino-Possio. Preconditioning strategies for non-Hermitian Toeplitz linear systems. *Numerical Linear Algebra with Applications*, 12(2-3):211–220, 2005.

[89] X. Jin. Hartley preconditioners for Toeplitz systems generated by positive continuous functions. *BIT Numerical Mathematics*, 34(3):367–371, 1994.

[90] X. Jin. A note on preconditioned block Toeplitz matrices. *SIAM Journal on Scientific Computing*, 16(4):951–955, 1995.

[91] X. Jin. Band Toeplitz preconditioners for block Toeplitz systems. *Journal of Computational and Applied Mathematics*, 70(2):225–230, 1996.

[92] X. Jin. *Developments and applications of block Toeplitz iterative solvers*, volume 2 of *Combinatorics and Computer Science*. Kluwer Academic Publishers Group, Dordrecht; Science Press Beijing, Beijing, 2002.

[93] X. Jin, S. Lei, and Y. Wei. Circulant preconditioners for solving differential equations with multidelays. *Computers & Mathematics with Applications*, 47(8):1429–1436, 2004.

[94] X. Jin and F. Lin. Block preconditioners with circulant blocks for general linear systems. *Computers & Mathematics with Applications*, 58(7):1309–1319, 2009.

[95] X. Jin, V. Sin, and L. Song. Circulant-block preconditioners for solving ordinary differential equations. *Applied Mathematics and Computation*, 140(2):409–418, 2003.

[96] X. Jin and Y. Wei. A short note on singular values of optimal and superoptimal preconditioned matrices. *International Journal of Computer Mathematics*, 84(8):1261–1263, 2007.

[97] X. Jin and Y. Wei. A survey and some extensions of T. Chan's preconditioner. *Linear Algebra and its Applications*, 428(2):403–412, 2008.

[98] X. Jin, Y. Wei, and W. Xu. A stability property of T. Chan's preconditioner. *SIAM Journal on Matrix Analysis and Applications*, 25(3):627–629, 2003.

[99] X. Jin, Z. Zhao, and S. Tam. Optimal preconditioners for functions of matrices. *Linear Algebra and its Applications*, 457:224–243, 2014.

[100] T. Kailath and V. Olshevsky. Displacement structure approach to discrete-trigonometric-transform based preconditioners of G. Strang type and of T. Chan type. *Calcolo*, 33(3-4):191–208 (1998), 1996.

[101] Y. Katznelson. *An introduction to harmonic analysis*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, third edition, 2004.

[102] D. Kressner and R. Luce. Fast computation of the matrix exponential for a Toeplitz matrix. *SIAM Journal on Matrix Analysis and Applications*, 39(1):23–47, 2018.

[103] T. Ku and C. Kuo. Design and analysis of Toeplitz preconditioners. *IEEE Transactions on Signal Processing*, 40(1):129–141, 1992.

[104] T. Ku and C. Kuo. On the spectrum of a family of preconditioned block Toeplitz matrices. *SIAM Journal on Scientific and Statistical Computing*, 13(4):948–966, 1992.

[105] C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45:255–282, 1950.

[106] C. Lanczos. Solution of systems of linear equations by minimized-iterations. *Journal of Research of the National Bureau of Standards*, 49:33–53, 1952.

[107] S. Lee, H. Pang, and H. Sun. Shift-invert Arnoldi approximation to the Toeplitz matrix exponential. *SIAM Journal on Scientific Computing*, 32(2):774–792, 2010.

[108] N. Levinson. The Wiener (root mean square) error criterion in filter design and prediction. *Journal of Mathematics and Physics*, 25(1-4):261–278, 1946.

[109] J. Liesen and Z. Strakoš. *Krylov subspace methods.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[110] F. Lin and M. Ng. Inverse product Toeplitz preconditioners for non-Hermitian Toeplitz systems. *Numerical Algorithms*, 54(2):279–295, 2010.

[111] X. Lv, T. Huang, and Z. Ren. A modified T. Chan's preconditioner for Toeplitz systems. *Computers & Mathematics with Applications*, 58(4):693–699, 2009.

[112] E. McDonald, S. Hon, J. Pestana, and A. Wathen. *Preconditioning for Nonsymmetry and Time-Dependence*, pages 81–91. Springer International Publishing, 2017.

[113] E. McDonald, J. Pestana, and A. Wathen. Preconditioning and iterative solution of all-at-once systems for evolutionary partial differential equations. *SIAM Journal on Scientific Computing*, 40(2):A1012–A1033, 2018.

[114] G. Meurant. *Computer solution of large linear systems*, volume 28 of *Studies in Mathematics and its Applications.* North-Holland Publishing Co., Amsterdam, 1999.

[115] M. Miranda and P. Tilli. Asymptotic spectra of Hermitian block Toeplitz matrices and preconditioning results. *SIAM Journal on Matrix Analysis and Applications*, 21(3):867–881, 2000.

[116] M. Morf. *Fast Algorithms for Multivariable Systems.* PhD thesis, Stanford University, 1974.

[117] M. Ng. Band preconditioners for block-Toeplitz-Toeplitz-block systems. *Linear Algebra and its Applications*, 259:307–327, 1997.

[118] M. Ng. *Iterative methods for Toeplitz systems.* Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2004.

[119] M. Ng, R. Chan, T. Chan, and A. Yip. Cosine transform preconditioners for high resolution image reconstruction. *Linear Algebra and its Applications*, 316(1):89–104, 2000.

[120] M. Ng and J. Pan. Approximate inverse circulant-plus-diagonal preconditioners for Toeplitz-plus-diagonal matrices. *SIAM Journal on Scientific Computing*, 32(3):1442–1464, 2010.

[121] M. Ng and D. Potts. Circulant preconditioners for indefinite Toeplitz systems. *BIT Numerical Mathematics*, 41(5):1079–1088, 2001.

[122] B. Ning, D. Zhao, and H. Li. Improved schur complement preconditioners for block-Toeplitz systems with small size blocks. *Journal of Computational and Applied Mathematics*, 311:655–663, 2017.

[123] J. Olkin. *Linear and Nonlinear Deconvolution Problems (Optimization).* PhD thesis, Rice University, 1986.

[124] M. Olshanskii and E. Tyrtyshnikov. *Iterative methods for linear systems.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2014.

[125] I. Oseledets and E. Tyrtyshnikov. A unifying approach to the construction of circulant preconditioners. *Linear Algebra and its Applications*, 418(2):435–449, 2006.

[126] C. Paige and M. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[127] J. Pan, R. Ke, M. Ng, and H. Sun. Preconditioning techniques for diagonal-times-Toeplitz matrices in fractional diffusion equations. *SIAM Journal on Scientific Computing*, 36(6):A2698–A2719, 2014.

[128] S. Parter. On the distribution of the singular values of Toeplitz matrices. *Linear Algebra and its Applications*, 80:115–130, 1986.

[129] J. Pestana and A. Wathen. A preconditioned MINRES method for nonsymmetric Toeplitz matrices. *SIAM Journal on Matrix Analysis and Applications*, 36(1):273–288, 2015.

[130] D. Potts and G. Steidl. Optimal trigonometric preconditioners for nonsymmetric Toeplitz systems. *Linear Algebra and its Applications*, 281(1):265–292, 1998.

[131] D. Potts and G. Steidl. Preconditioners for ill-conditioned Toeplitz matrices. *BIT Numerical Mathematics*, 39(3):513–533, 1999.

[132] D. Potts and G. Steidl. Preconditioners for ill-conditioned Toeplitz systems constructed from positive kernels. *SIAM Journal on Scientific Computing*, 22(5):1741–1761, 2001.

[133] M. Powell. *Approximation theory and methods.* Cambridge University Press, Cambridge-New York, 1981.

[134] J. Rissanen. Algorithms for triangular decomposition of block Hankel and Toeplitz matrices with application to factoring positive matrix polynomials. *Mathematics of Computation*, 27(121):147–154, 1973.

[135] Y. Saad. A flexible inner-outer preconditioned GMRES algorithm. *SIAM Journal on Scientific Computing*, 14(2):461–469, 1993.

[136] Y. Saad and M. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[137] E. Sachs and A. Strauss. Efficient solution of a partial integro-differential equation in finance. *Applied Numerical Mathematics*, 58(11):1687–1703, 2008.

[138] S. Serra. Preconditioning strategies for asymptotically ill-conditioned block Toeplitz systems. *BIT Numerical Mathematics*, 34(4):579–594, 1994.

[139] S. Serra. Optimal, quasi-optimal and superlinear band-Toeplitz preconditioners for asymptotically ill-conditioned positive definite Toeplitz systems. *Mathematics of Computation*, 66(218):651–665, 1997.

[140] S. Serra Capizzano. Toeplitz preconditioners constructed from linear approximation processes. *SIAM Journal on Matrix Analysis and Applications*, 20(2):446–465, 1998.

[141] S. Serra Capizzano. Superlinear PCG methods for symmetric Toeplitz systems. *Mathematics of Computation*, 68:793–803, 1999.

[142] S. Serra Capizzano and P. Tilli. Extreme singular values and eigenvalues of non-Hermitian block Toeplitz matrices. *Journal of Computational and Applied Mathematics*, 108(1):113–130, 1999.

[143] S. Serra Capizzano and E. Tyrtyshnikov. Any circulant-like preconditioner for multilevel matrices is not superlinear. *SIAM Journal on Matrix Analysis and Applications*, 21(2):431–439, 2000.

[144] G. Sleijpen and D. Fokkema. BiCGstab(l) for linear equations involving unsymmetric matrices with complex spectrum. *Electronic Transactions on Numerical Analysis*, 1(Sept.):11–32, 1993.

[145] P. Sonneveld and M. van Gijzen. IDR(s): A family of simple and fast algorithms for solving large nonsymmetric systems of linear equations. *SIAM Journal on Scientific Computing*, 31(2):1035–1062, 2009.

[146] G. Strang. A proposal for Toeplitz matrix calculations. *Studies in Applied Mathematics*, 74(2):171–176, 1986.

[147] G. Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, MA, fourth edition, 2009.

[148] G. Strang and A. Edelman. The Toeplitz-circulant eigenvalue problem $Ax = \lambda Cx$. In *Oakland conference on partial differential equations and applied mathematics (Rochester, Mich., 1986)*, volume 154 of *Pitman Res. Notes Math. Ser.*, pages 109–117. Longman Sci. Tech., Harlow, 1987.

[149] V. Strela and E. Tyrtyshnikov. Which circulant preconditioner is better? *Mathematics of Computation*, 65(213):137–150, 1996.

[150] D. Sweet. The use of pivoting to improve the numerical performance of algorithms for Toeplitz matrices. *SIAM Journal on Matrix Analysis and Applications*, 14(2):468–493, 1993.

[151] P. Tilli. A note on the spectral distribution of Toeplitz matrices. *Linear and Multilinear Algebra*, 45(2-3):147–159, 1998.

[152] P. Tilli. Singular values and eigenvalues of non-Hermitian block Toeplitz matrices. *Linear Algebra and its Applications*, 272(1):59–89, 1998.

[153] O. Toeplitz. Zur Theorie der quadratischen und bilinearen Formen von unendlichvielen Veränderlichen. *Mathematische Annalen*, 70(3):351–376, 1911.

[154] A. Toselli and O. Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.

[155] W. Trench. An algorithm for the inversion of finite Toeplitz matrices. *Journal of the Society for Industrial and Applied Mathematics*, 12(3):515–522, 1964.

[156] U. Trottenberg, C. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, Inc., San Diego, CA, 2001.

[157] E. Tyrtyshnikov. Optimal and superoptimal circulant preconditioners. *SIAM Journal on Matrix Analysis and Applications*, 13(2):459–473, 1992.

[158] E. Tyrtyshnikov. New theorems on the distribution of eigenvalues and singular values of multilevel Toeplitz matrices. *Doklady Akademii Nauk*, 333(3):300–303, 1993.

[159] E. Tyrtyshnikov. Circulant preconditioners with unbounded inverses. *Linear Algebra and its Applications*, 216:1–23, 1995.

[160] E. Tyrtyshnikov. A unifying approach to some old and new theorems on distribution and clustering. *Linear Algebra and its Applications*, 232:1–43, 1996.

[161] E. Tyrtyshnikov, A. Yeremin, and N. Zamarashkin. Clusters, preconditioners, convergence. *Linear Algebra and its Applications*, 263:25–48, 1997.

[162] E. Tyrtyshnikov and N. Zamarashkin. Spectra of multilevel Toeplitz matrices: advanced theory via simple matrix relationships. *Linear Algebra and its Applications*, 270(1):15–27, 1998.

[163] M. Van Barel, G. Heinig, and P. Kravanja. A stabilized superfast solver for nonsymmetric Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 23(2):494–510, 2001.

[164] H. van der Vorst. Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 13(2):631–644, 1992.

[165] H. van der Vorst. *Iterative Krylov methods for large linear systems*, volume 13 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2009.

[166] C. Van Loan. *Computational frameworks for the fast Fourier transform*, volume 10 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[167] E. Vecharynski and A. Knyazev. Absolute value preconditioning for symmetric indefinite linear systems. *SIAM Journal on Scientific Computing*, 35(2):A696–A718, 2013.

[168] A. Wathen. Preconditioning. *Acta Numerica*, 24:329–376, 2015.

[169] Y. Wen, W. Ching, and M. Ng. Approximate inverse-free preconditioners for Toeplitz matrices. *Applied Mathematics and Computation*, 217(16):6856–6867, 2011.

[170] J. Xia, Y. Xi, and M. Gu. A superfast structured solver for Toeplitz linear systems via randomized sampling. *SIAM Journal on Matrix Analysis and Applications*, 33(3):837–858, 2012.

[171] Z. Xie, X. Jin, and Z. Zhao. A convergence analysis of the MINRES method for some Hermitian indefinite systems. *East Asian Journal on Applied Mathematics*, 7(4):827–836, 2017.

[172] N. Zamarashkin and E. Tyrtyshnikov. Distribution of the eigenvalues and singular numbers of Toeplitz matrices under weakened requirements on the generating function. *Matematicheskiĭ Sbornik*, 188(8):83–92, 1997.

[173] S. Zohar. Toeplitz matrix inversion: the algorithm of W. F. Trench. *Journal of the Association for Computing Machinery*, 16(4):592–601, 1969.

[174] S. Zohar. The solution of a Toeplitz set of linear equations. *Journal of the Association for Computing Machinery*, 21(2):272–276, 1974.