# On Sinc Discretization and Banded Preconditioning for Linear Third-Order Ordinary Differential Equations [*]

Zhong-Zhi Bai

*State Key Laboratory of Scientific/Engineering Computing*
*Institute of Computational Mathematics and Scientific/Engineering Computing*
*Academy of Mathematics and Systems Science*
*Chinese Academy of Sciences, P.O. Box 2719, Beijing 100190, P.R. China*
*Email: bzz@lsec.cc.ac.cn*

Raymond H. Chan

*Department of Mathematics*
*The Chinese University of Hong Kong*
*Shatin, Hong Kong*
*Email: rchan@math.cuhk.edu.hk*

Zhi-Ru Ren

*State Key Laboratory of Scientific/Engineering Computing*
*Institute of Computational Mathematics and Scientific/Engineering Computing*
*Academy of Mathematics and Systems Science*
*Chinese Academy of Sciences, P.O. Box 2719, Beijing 100190, P.R. China*
*Email: renzr@lsec.cc.ac.cn*

April 22, 2010

**Abstract**

Some draining or coating fluid-flow problems and problems concerning the flow of thin films of viscous fluid with a free surface can be described by third-order ordinary differential equations. In this paper, we solve the boundary value problems of such equations by sinc discretization and prove that the discrete solutions converge to the true solutions of the ordinary differential equations exponentially. The discrete solution is determined by a linear system with the coefficient matrix being a combination of Toeplitz and diagonal matrices. The system can be effectively solved by Krylov subspace iteration methods such as GMRES preconditioned by banded matrices. We demonstrate that the eigenvalues of the preconditioned matrix are uniformly bounded within a rectangle on the complex plane independent of the size of the linear system. Numerical examples are given to illustrate the effective performance of our method.

[*] Supported by The National Basic Research Program (No. 2005CB321702), The China Outstanding Young Scientist Foundation (No. 10525102), Hong Kong Research Grants Council Grant (No. 400508), and CUHK DAG 2060257.

**Keywords:** third-order ordinary differential equation, sinc-collocation discretization, sinc-Galerkin discretization, convergence analysis, banded preconditioning, Krylov subspace methods

**AMS(MOS) Subject Classifications:** 65F08, 65F10, 65F15, 65L60, 65T10; CR: G1.3.

# 1 Introduction

We consider the numerical solution for the two-point boundary value problem of linear third-order ordinary differential equation (ODE):

$$\begin{cases} Ly(x) := y^{'''}(x) + \mu_2(x)y^{''}(x) + \mu_1(x)y^{'}(x) + \mu_0(x)y(x) = \sigma(x), \\ y(a) = 0, \quad y(b) = 0, \quad y^{'}(a) = 0, \quad a < x < b, \end{cases} \tag{1.1}$$

where $\mu_j(x)$ $(j = 0, 1, 2)$ and $\sigma(x)$ are known bounded functions, and $a$ and $b$ are given real numbers. This class of problems arises from many practical applications such as draining or coating fluid-flow problems [26, 29, 31, 32] and problems concerning the flow of thin films of viscous fluid with a free surface [10, 11, 12, 14]. A remarkable feature of this class of ODEs is that its highest term is of order three, which makes the coefficient matrices of the correspondingly resulted linear system be strongly nonsymmetric and highly ill-conditioned and, hence, causes much difficulty in solving it numerically.

We first use the sinc-collocation and the sinc-Galerkin methods to discrete the ODE (1.1). The sinc function used is

$$\text{sinc}(t) = \frac{\sin(\pi t)}{\pi t}, \quad -\infty < t < \infty,$$

and the set of basis functions adopted are

$$S(j, h)(t) := \frac{\sin[\pi(t - jh)/h]}{\pi(t - jh)/h}, \quad -\infty < t < \infty, \quad j \in \mathbb{Z}, \tag{1.2}$$

where $h$ is the step-size and $\mathbb{Z}$ denotes the set of all integers [27]. The points $t_j = jh$, $j \in \mathbb{Z}$, are called the sinc grid-points.

Using the sinc discretizations on (1.1), we can obtain an $n$-by-$n$ system of linear equations of the form $\mathbf{Aw} = \mathbf{p}$. Theoretically, we demonstrate that the discrete solution determined by the linear system converges exponentially to the true solution of the continuous problem when the step-size $h$ tends to zero. We will see that the coefficient matrix $\mathbf{A}$ is a combination of Toeplitz and diagonal matrices. Hence, a straightforward application of the Gaussian elimination will result in an algorithm of $\mathcal{O}(n^3)$ complexity. In fact, for $n$-by-$n$ Toeplitz linear systems, fast direct solvers of complexity $\mathcal{O}(n \log^2 n)$ have been developed; see, for instance [1]. However, there does not exist fast direct solver for Toeplitz-plus-diagonal linear systems yet, since the displacement rank of a Toeplitz-plus-diagonal matrix can take any value between 0 and $n$; see [18]. Therefore, fast direct Toeplitz solvers that are based on small displacement-rank are not applicable to the Toeplitz-plus-diagonal linear systems.

However, it is known [9, 24] that for any $n$-dimensional vector $\mathbf{q}$, the matrix-vector product $\mathbf{Aq}$ can be computed in $\mathcal{O}(n \log n)$ operations. Thus Krylov subspace iteration methods can be employed to solve the linear system $\mathbf{Aw} = \mathbf{p}$ economically; see, e.g., [15, 16, 17]. In order to accelerate the convergence speeds of Krylov subspace methods, we need to construct an efficient and effective preconditioner for the matrix $\mathbf{A}$. In this paper, we use a banded matrix with a fixed bandwidth as a preconditioning matrix for the matrix $\mathbf{A}$; see [6, 5, 7, 23, 24]. We show that the eigenvalues of the preconditioned matrix are tightly bounded within a rectangle on the complex plane. Numerical results show that the new discretization scheme is accurate and the banded preconditioner is effective in accelerating the convergence property of GMRES.

The remainder of the paper is outlined as follows. In Section 2, we discretize the ODE (1.1) by both sinc-collocation and sinc-Galerkin methods. A combination of these two methods leads to the linear system $\mathbf{Aw} = \mathbf{p}$. In Section 3, we estimate the error between the discrete and the continuous solutions of (1.1). A banded preconditioning matrix for the coefficient matrix $\mathbf{A}$ is constructed and the spectral distribution about the corresponding preconditioned matrix is estimated in Section 4. Several numerical examples are given in Section 5 to show the effectiveness of our new approach. Finally, in Section 6, we end this paper with a few concluding remarks.

## 2 Sinc Discretization Methods

Let $\mathcal{D}$ be a simply-connected domain having boundary $\partial\mathcal{D}$. Let $a$ and $b$ denote two distinct points of $\partial\mathcal{D}$, and $t = \phi(z)$ denote a conformal mapping of $\mathcal{D}$ onto a strip region $\mathcal{D}_d$ such that $\phi(a) = -\infty$ and $\phi(b) = \infty$, where $\mathcal{D}_d := \{t \in \mathbb{C} : |\mathrm{Im}(t)| < d\}$. Conversely, $z = \psi(t) := \phi^{-1}(t)$ maps $\mathcal{D}_d$ onto $\mathcal{D}$ with a boundary $\partial\mathcal{D}$ on which the points $a$ and $b$ lie. Here and in the sequel, we use $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ to denote the real and the imaginary parts of a complex number, and we will write a function $f(x)$ simply as $f$ if no confusion arises.

In this section, we discretize the ODE (1.1) by both sinc-collocation and sinc-Galerkin methods. To this end, we approximate the exact solution $y(x)$ of (1.1) by the function

$$y_N(x) = \frac{1}{\phi'(x)} \sum_{j=-N}^{N} w_j S(j,h) \circ \phi(x), \tag{2.1}$$

where $\phi(x)$ is a conformal mapping from $\mathcal{D}$ to $\mathcal{D}_d$, $\{S(j,h)\}_{j\in\mathbb{Z}_N}$ are the sinc-basis functions in (1.2), and $\{w_j\}_{j\in\mathbb{Z}_N}$ are the unknown coefficients to be found. Here, we have used the notation $\mathbb{Z}_N = \{-N, -N+1, \ldots, N\}$.

### 2.1 Sinc-Collocation Method

For the sinc-collocation method, the unknown coefficients $\{w_j\}_{j=-N}^{N}$ in (2.1) are determined by the collocation technique, which uses the sinc grid-points as the collocation points. More precisely, we impose the conditions

$$Ly_N(x_k) = \sigma(x_k), \quad k \in \mathbb{Z}_N, \tag{2.2}$$

with $x_k = \psi(kh) = \phi^{-1}(kh)$ and $L$ being the operator in (1.1). After substituting $y_N(x)$ in (2.1) into (2.2) and multiplying $h^3/(\phi')^2$ to both sides, we obtain a system of linear equations with respect to $\{w_j\}_{j=-N}^{N}$ as follows:

$$\begin{aligned}
&\sum_{j=-N}^{N}\Bigg\{\delta_{jk}^{(3)}+h\frac{\mu_2}{\phi'}\,\delta_{jk}^{(2)}+h^2\left[\frac{2}{\phi'}\left(\frac{1}{\phi'}\right)''-\left(\left(\frac{1}{\phi'}\right)'\right)^2+\frac{\mu_2}{\phi'}\left(\frac{1}{\phi'}\right)'+\frac{\mu_1}{(\phi')^2}\right]\delta_{jk}^{(1)}\\
&\quad+h^3\frac{1}{(\phi')^2}\left[\left(\frac{1}{\phi'}\right)'''+\mu_2\left(\frac{1}{\phi'}\right)''+\mu_1\left(\frac{1}{\phi'}\right)'+\frac{\mu_0}{\phi'}\right]\delta_{jk}^{(0)}\Bigg\}(x_k)\cdot w_j\\
&\quad=h^3\frac{\sigma}{(\phi')^2}(x_k),\quad k\in\mathbb{Z}_N,
\end{aligned}\tag{2.3}$$

where

$$\delta_{jk}^{(m)}:=h^m\frac{\mathrm{d}^m}{\mathrm{d}\phi^m}[S(j,h)\circ\phi(x)]\big|_{x=x_k},\quad m=0,1,2,3.\tag{2.4}$$

More concretely, it holds that

$$\delta_{jk}^{(0)}=\begin{cases}1, & j=k,\\ 0, & j\neq k,\end{cases}\qquad \delta_{jk}^{(1)}=\begin{cases}0, & j=k,\\ \frac{(-1)^{k-j}}{k-j}, & j\neq k,\end{cases}$$

$$\delta_{jk}^{(2)}=\begin{cases}-\frac{\pi^2}{3}, & j=k,\\ \frac{(-1)^{k-j}(-2)}{(k-j)^2}, & j\neq k,\end{cases}\qquad \delta_{jk}^{(3)}=\begin{cases}0, & j=k,\\ \frac{(-1)^{k-j}[6-(k-j)^2\pi^2]}{(k-j)^3}, & j\neq k.\end{cases}$$

Noting that

$$\delta_{jk}^{(0)}=\delta_{kj}^{(0)},\quad \delta_{jk}^{(1)}=-\delta_{kj}^{(1)},\quad \delta_{jk}^{(2)}=\delta_{kj}^{(2)},\quad \delta_{jk}^{(3)}=-\delta_{kj}^{(3)},\quad j,k\in\mathbb{Z}_N,$$

we may rewrite the system of linear equations (2.3) in the form

$$\begin{aligned}
&\sum_{j=-N}^{N}\Bigg\{-\delta_{kj}^{(3)}+h\frac{\mu_2}{\phi'}\,\delta_{kj}^{(2)}-h^2\left[\frac{2}{\phi'}\left(\frac{1}{\phi'}\right)''-\left(\left(\frac{1}{\phi'}\right)'\right)^2+\frac{\mu_2}{\phi'}\left(\frac{1}{\phi'}\right)'+\frac{\mu_1}{(\phi')^2}\right]\delta_{kj}^{(1)}\\
&\quad+h^3\frac{1}{(\phi')^2}\left[\left(\frac{1}{\phi'}\right)'''+\mu_2\left(\frac{1}{\phi'}\right)''+\mu_1\left(\frac{1}{\phi'}\right)'+\frac{\mu_0}{\phi'}\right]\delta_{kj}^{(0)}\Bigg\}(x_k)\cdot w_j\\
&\quad=h^3\frac{\sigma}{(\phi')^2}(x_k),\quad k\in\mathbb{Z}_N.
\end{aligned}\tag{2.5}$$

In order to rewrite the system of linear equations (2.5) into a matrix-vector form, we let $\mathbf{T}^{(m)}$ $(m=0,1)$ be the matrices whose $(j,k)$th element is $\delta_{jk}^{(m)}$, and $\mathbf{T}^{(m)}$ $(m=2,3)$ be the matrices whose $(j,k)$th element is $-\delta_{jk}^{(m)}$. Denote by $n=2N+1$. Then $\mathbf{T}^{(0)}$ is the identity matrix and

$\mathbf{T}^{(m)}$ $(m = 1, 2, 3)$ are $n \times n$ Toeplitz matrices given by

$$\mathbf{T}^{(1)} = \begin{bmatrix} 0 & -1 & \frac{1}{2} & \dots & \frac{(-1)^{n-1}}{n-1} \\ 1 & 0 & \ddots & \ddots & \vdots \\ -\frac{1}{2} & 1 & \ddots & -1 & \frac{1}{2} \\ \vdots & \ddots & \ddots & 0 & -1 \\ -\frac{(-1)^{n-1}}{n-1} & \dots & -\frac{1}{2} & 1 & 0 \end{bmatrix}, \tag{2.6}$$

$$\mathbf{T}^{(2)} = \begin{bmatrix} \frac{\pi^2}{3} & -2 & \frac{2}{2^2} & \dots & \frac{(-1)^{n-1}2}{(n-1)^2} \\ -2 & \frac{\pi^2}{3} & \ddots & \ddots & \vdots \\ \frac{2}{2^2} & -2 & \ddots & -2 & \frac{2}{2^2} \\ \vdots & \ddots & \ddots & \frac{\pi^2}{3} & -2 \\ \frac{(-1)^{n-1}2}{(n-1)^2} & \dots & \frac{2}{2^2} & -2 & \frac{\pi^2}{3} \end{bmatrix} \tag{2.7}$$

and

$$\mathbf{T}^{(3)} = \begin{bmatrix} 0 & 6-\pi^2 & \frac{-(6-2^2\pi^2)}{2^3} & \dots & \frac{-(-1)^{n-1}[6-(n-1)^2\pi^2]}{(n-1)^3} \\ -(6-\pi^2) & 0 & \ddots & \ddots & \vdots \\ \frac{6-2^2\pi^2}{2^3} & -(6-\pi^2) & \ddots & 6-\pi^2 & \frac{-(6-2^2\pi^2)}{2^3} \\ \vdots & \ddots & \ddots & 0 & 6-\pi^2 \\ \frac{(-1)^{n-1}[6-(n-1)^2\pi^2]}{(n-1)^3} & \dots & \frac{6-2^2\pi^2}{2^3} & -(6-\pi^2) & 0 \end{bmatrix}. \tag{2.8}$$

We remark that the generating functions of $\mathbf{T}^{(1)}$, $\mathbf{T}^{(2)}$ and $\mathbf{T}^{(3)}$ are $\imath\theta$, $\theta^2$ and $\imath\theta^3$, respectively; see [27]. It follows that the system of linear equations (2.5) can be written as

$$\mathbf{A}_C\mathbf{w} = \mathbf{p},$$

where

$$\mathbf{A}_C = \mathbf{T}^{(3)} + \mathbf{D}_C^{(2)}\mathbf{T}^{(2)} + \mathbf{D}_C^{(1)}\mathbf{T}^{(1)} + \mathbf{D}_C^{(0)} \in \mathbb{R}^{n\times n}, \tag{2.9}$$

$$\mathbf{w} = (w_{-N}, w_{-N+1}, \dots, w_N)^T \in \mathbb{R}^n \tag{2.10}$$

and

$$\mathbf{p} = h^3 \left( \frac{\sigma}{(\phi')^2}(x_{-N}), \frac{\sigma}{(\phi')^2}(x_{-N+1}), \dots, \frac{\sigma}{(\phi')^2}(x_N) \right)^T \in \mathbb{R}^n. \tag{2.11}$$

In addition,

$$\mathbf{D}_C^{(i)} := \mathrm{diag}(g_C^{(i)}(x_{-N}), g_C^{(i)}(x_{-N+1}), \dots, g_C^{(i)}(x_N)), \quad i = 0, 1, 2,$$

are diagonal matrices, with

$$g_C^{(2)} = -h\frac{\mu_2}{\phi'},$$

$$g_C^{(1)} = -h^2 \left[ \frac{2}{\phi'}\left(\frac{1}{\phi'}\right)'' - \left(\left(\frac{1}{\phi'}\right)'\right)^2 + \frac{\mu_2}{\phi'}\left(\frac{1}{\phi'}\right)' + \frac{\mu_1}{(\phi')^2} \right]$$

and

$$g_C^{(0)} = h^3 \frac{1}{(\phi')^2}\left[\left(\frac{1}{\phi'}\right)''' + \mu_2\left(\frac{1}{\phi'}\right)'' + \mu_1\left(\frac{1}{\phi'}\right)' + \frac{\mu_0}{\phi'}\right].$$

## 2.2 Sinc-Galerkin Method

For the sinc-Galerkin method, the unknown coefficients $\{w_j\}_{j=-N}^{N}$ in (2.1) are determined by orthogonalizing the residual $Ly_N(x) - \sigma(x)$ with the functions $\{S(k,h)\circ\phi(x)\}_{k=-N}^{N}$, where $L$ is the operator in (1.1). This yields the discretized system

$$\langle Ly_N - \sigma,\, S(k,h)\circ\phi\rangle = 0, \quad k \in \mathbb{Z}_N, \tag{2.12}$$

where $\langle\cdot,\cdot\rangle$ represents the inner product defined by

$$\langle f,\, g\rangle = \int_a^b \frac{f(x)g(x)}{\phi'(x)}\,\mathrm{d}x,$$

with $\phi(x)$ being a conformal mapping from $\mathcal{D}$ to $\mathcal{D}_d$ (see Section 2). By integrating (2.12) by part, and using Corollary 4.2.15 in [27], we obtain a system of linear equations with respect to $\{w_j\}_{j=-N}^{N}$ as follows:

$$\begin{aligned}\sum_{j=-N}^{N}\Bigg\{ & -\delta_{kj}^{(3)} + h\delta_{kj}^{(2)}\frac{\mu_2}{\phi'} - h^2\delta_{kj}^{(1)}\left[\frac{2}{\phi'}\left(\frac{1}{\phi'}\right)'' - \left(\left(\frac{1}{\phi'}\right)'\right)^2 - \frac{2\mu_2'}{(\phi')^2} - \frac{\mu_2}{\phi'}\left(\frac{1}{\phi'}\right)' + \frac{\mu_1}{(\phi')^2}\right] \\ & - h^3\delta_{kj}^{(0)}\frac{1}{(\phi')^2}\left[\left(\frac{1}{\phi'}\right)''' - \left(\frac{\mu_2}{\phi'}\right)'' + \left(\frac{\mu_1}{\phi'}\right)' - \frac{\mu_0}{\phi'}\right]\Bigg\}(x_j)\cdot w_j \\ & = h^3\frac{\sigma}{(\phi')^2}(x_k), \quad k \in \mathbb{Z}_N,\end{aligned} \tag{2.13}$$

where $\delta_{jk}^{(m)}$ $(j,k \in \mathbb{Z}_N;\ m = 0,1,2,3)$ are the same as in (2.4).

The system of linear equations (2.13) can be rewritten in the matrix-vector form

$$\mathbf{A}_G\mathbf{w} = \mathbf{p},$$

where

$$\mathbf{A}_G = \mathbf{T}^{(3)} + \mathbf{T}^{(2)}\mathbf{D}_G^{(2)} + \mathbf{T}^{(1)}\mathbf{D}_G^{(1)} + \mathbf{D}_G^{(0)} \in \mathbb{R}^{n\times n}, \tag{2.14}$$

$\mathbf{T}^{(m)}$ $(m = 1,2,3)$ are the Toeplitz matrices defined in (2.6)–(2.8), and $\mathbf{w}$ and $\mathbf{p}$ are the unknown and the right-hand side vectors defined in (2.10) and (2.11), respectively. In addition,

$$\mathbf{D}_G^{(i)} := \mathrm{diag}(g_G^{(i)}(x_{-N}), g_G^{(i)}(x_{-N+1}), \ldots, g_G^{(i)}(x_N)), \quad i = 0,1,2,$$

are diagonal matrices, with

$$\begin{aligned} g_G^{(2)} &= -h\frac{\mu_2}{\phi'}, \\ g_G^{(1)} &= -h^2\left[\frac{2}{\phi'}\left(\frac{1}{\phi'}\right)'' - \left(\left(\frac{1}{\phi'}\right)'\right)^2 - \frac{2\mu_2'}{(\phi')^2} - \frac{\mu_2}{\phi'}\left(\frac{1}{\phi'}\right)' + \frac{\mu_1}{(\phi')^2}\right]\end{aligned}$$

and

$$g_G^{(0)} = -h^3 \frac{1}{(\phi')^2} \left[ \left(\frac{1}{\phi'}\right)''' - \left(\frac{\mu_2}{\phi'}\right)'' + \left(\frac{\mu_1}{\phi'}\right)' - \frac{\mu_0}{\phi'} \right].$$

## 2.3 A Combination of Sinc-Collocation and Sinc-Galerkin Methods

A symmetric or positive definite system of linear equations[1] often possesses preferable algebraic and numerical properties. Moreover, there are many economical and fast direct and iterative methods, with plenty of error analysis and convergence theory, for solving symmetric or positive definite systems of linear equations; see [2, 3, 4, 13]. Therefore, one should try to construct a discretized linear system for (1.1) such that its coefficient matrix is as symmetrical or positive definite as possible, if it itself is not so. To this end, we average the sinc-collocation matrix $\mathbf{A}_C$ in (2.9) and the sinc-Galerkin matrix $\mathbf{A}_G$ in (2.14) to obtain the system of linear equations

$$\mathbf{A}\mathbf{w} = \mathbf{p}, \tag{2.15}$$

where

$$\begin{aligned}\mathbf{A} &= \frac{1}{2}(\mathbf{A}_C + \mathbf{A}_G)\\ &= \mathbf{T}^{(3)} + \frac{1}{2}(\mathbf{D}^{(2)}\mathbf{T}^{(2)} + \mathbf{T}^{(2)}\mathbf{D}^{(2)}) + \frac{1}{2}(\mathbf{D}^{(1)}\mathbf{T}^{(1)} + \mathbf{T}^{(1)}\mathbf{D}^{(1)})\\ &\quad + \frac{1}{2}(\mathbf{D}_s^{(1)}\mathbf{T}^{(1)} - \mathbf{T}^{(1)}\mathbf{D}_s^{(1)}) + \mathbf{D}^{(0)},\end{aligned} \tag{2.16}$$

with

$$\begin{aligned}\mathbf{D}^{(i)} &= \frac{1}{2}(\mathbf{D}_C^{(i)} + \mathbf{D}_G^{(i)}) := \mathrm{diag}(g^{(i)}(x_{-N}), g^{(i)}(x_{-N+1}), \ldots, g^{(i)}(x_N)), \quad i = 0, 1, 2,\\ \mathbf{D}_s^{(1)} &= \frac{1}{2}(\mathbf{D}_C^{(1)} - \mathbf{D}_G^{(1)}) := \mathrm{diag}(g_s^{(1)}(x_{-N}), g_s^{(1)}(x_{-N+1}), \ldots, g_s^{(1)}(x_N)),\end{aligned} \tag{2.17}$$

and $\mathbf{w}$ and $\mathbf{p}$ are defined as in (2.10). In addition,

$$\begin{aligned}g^{(2)} &= -h\frac{\mu_2}{\phi'},\\ g^{(1)} &= -h^2 \left[ \frac{2}{\phi'}\left(\frac{1}{\phi'}\right)'' - \left(\left(\frac{1}{\phi'}\right)'\right)^2 - \frac{\mu_2'}{(\phi')^2} + \frac{\mu_1}{(\phi')^2} \right],\\ g^{(0)} &= \frac{h^3}{2}\frac{1}{(\phi')^2} \left[ \mu_2\left(\frac{1}{\phi'}\right)'' + \left(\frac{\mu_2}{\phi'}\right)'' + \mu_1\left(\frac{1}{\phi'}\right)' - \left(\frac{\mu_1}{\phi'}\right)' + \frac{2\mu_0}{\phi'} \right]\end{aligned}$$

and

$$g_s^{(1)} = -h^2 \frac{1}{\phi'}\left(\frac{\mu_2}{\phi'}\right)'.$$

[1] A complex system of linear equations is called positive definite if the Hermitian part of its coefficient matrix is positive definite; see, e.g., [3].

Evidently, the matrix $\mathbf{A}$ in (2.16) is more symmetrically structured than either of the matrices $\mathbf{A}_C$ and $\mathbf{A}_G$ in (2.9) and (2.14), respectively. For example, instead of having $\mathbf{D}_C^{(2)}\mathbf{T}^{(2)}$ in (2.9) or $\mathbf{T}^{(2)}\mathbf{D}_G^{(2)}$ in (2.14), we now have $\frac{1}{2}(\mathbf{D}^{(2)}\mathbf{T}^{(2)}+\mathbf{T}^{(2)}\mathbf{D}^{(2)})$ in (2.16) which is symmetric. Moreover, when the matrix $\frac{1}{2}(\mathbf{D}^{(2)}\mathbf{T}^{(2)}+\mathbf{T}^{(2)}\mathbf{D}^{(2)})+\mathbf{D}^{(0)}$ is symmetric positive definite, the matrix $\mathbf{A}$ is almost positive definite provided that $\mathbf{D}_s^{(1)}\mathbf{T}^{(1)}-\mathbf{T}^{(1)}\mathbf{D}_s^{(1)}$ is small, or $\mathbf{D}_s^{(1)}$ and $\mathbf{T}^{(1)}$ are nearly commutative. In particular, if $\mathbf{D}_s^{(1)}$ and $\mathbf{T}^{(1)}$ commute exactly, then the matrix $\mathbf{A}$ is positive definite.

The following lemma, originally proved in [20, 28] and recently re-stated in [6], describes the eigenvalue distributions of the Toeplitz matrices $\mathbf{T}^{(i)}$ $(i=1,2,3)$; see also [27]. The results follow directly from the fact that the generating functions of $\mathbf{T}^{(1)}$, $\mathbf{T}^{(2)}$ and $\mathbf{T}^{(3)}$ are $\imath\theta$, $\theta^2$ and $\imath\theta^3$, respectively.

**Lemma 2.1** *[20, 28] Let $\mathbf{T}^{(i)}$ $(i=1,2,3)$ be the Toeplitz matrices defined in (2.6)–(2.8). Then*

(i) $\mathbf{T}^{(1)}$ *is a skew-symmetric matrix and its eigenvalues $\{\imath\lambda_j^{(1)}\}_{j=-N}^{N}$ satisfy $\lambda_j^{(1)}\in[-\pi,\pi]$;*

(ii) $\mathbf{T}^{(2)}$ *is a symmetric positive-definite matrix and its eigenvalues $\{\lambda_j^{(2)}\}_{j=-N}^{N}$ satisfy $\lambda_j^{(2)}\in[4\sin^2(\frac{\pi}{4(N+1)}),\pi^2]$;*

(iii) $\mathbf{T}^{(3)}$ *is a skew-symmetric matrix and its eigenvalues $\{\imath\lambda_j^{(3)}\}_{j=-N}^{N}$ satisfy $\lambda_j^{(3)}\in[-\pi^3,\pi^3]$.*

Hereafter, we use $(\cdot)^*$ to denote the conjugate transpose of either a vector or a matrix. For a square matrix $\mathbf{X}$, we represent by $\mathcal{H}(\mathbf{X})$ and $\mathcal{S}(\mathbf{X})$, respectively, its Hermitian and skew-Hermitian parts, and $\lambda(\mathbf{X})$ its spectral set. In particular, when $\mathbf{X}$ is Hermitian or real symmetric, we use $\lambda_{\max}(\mathbf{X})$ and $\lambda_{\min}(\mathbf{X})$ to represent its largest and smallest eigenvalues. In addition, $\mathbf{I}$ is used to denote the identity matrix of suitable dimension.

## 3 Convergence Analysis

In this section, we show that the approximate solution $y_N(x)$ given in (2.1) converges exponentially to the true solution $y(x)$ of the ODE (1.1) as $N$ tends to infinity. Similar to the treatments of the second-order and the fourth-order ODEs [27, 22, 25], the arguments here also proceed in the following three steps:

(i) estimate the Euclidean norm of $\mathbf{A}\tilde{\mathbf{y}}-\mathbf{p}$, where $\tilde{\mathbf{y}}$ is an $n$-dimensional real vector defined by

$$\tilde{\mathbf{y}}=(\tilde{y}(x_{-N}),\tilde{y}(x_{-N+1}),\ldots,\tilde{y}(x_N))^T, \tag{3.1}$$

with $\tilde{y}(x):=y(x)\phi^{'}(x)$;

(ii) derive an upper bound for the Euclidean norm of the matrix $\mathbf{A}^{-1}$; and

(iii) demonstrate the boundedness of the error $|y(x)-y_N(x)|$.

In order to precisely describe the convergence, we introduce two necessary functional spaces $\mathbb{L}_\alpha(\mathcal{D})$ and $\mathbb{H}^\infty(\mathcal{D})$: the space $\mathbb{L}_\alpha(\mathcal{D})$ is the set of all analytic functions $F$ in $\mathcal{D}$ such that

$$|F(z)| \leq \frac{c|e^{\phi(z)}|^\alpha}{(1+|e^{\phi(z)}|)^{2\alpha}}$$

for all $z \in \mathcal{D}$, where $c$ and $\alpha$ are positive constants, and $\phi : \mathcal{D} \to \mathcal{D}_d$ is a conformal mapping; while the space $\mathbb{H}^\infty(\mathcal{D})$ is the space of analytic functions in $\mathcal{D}$ equipped with the maximum norm.

We first estimate an upper bound for $\|\mathbf{A}\tilde{\mathbf{y}} - \mathbf{p}\|_2$.

**Lemma 3.1** *Assume that the ODE (1.1) has a unique solution $y := y(x) \in \mathbb{L}_\alpha(\mathcal{D})$. Let $\mathbf{A}_C$, $\mathbf{A}_G$, $\mathbf{A}$, $\tilde{\mathbf{y}}$ and $\mathbf{p}$ be defined as in (2.9), (2.14), (2.16), (3.1) and (2.11), respectively.*

(i) *If $\mu_2/\phi'$, $\mu_1/(\phi')^2$, $(1/\phi')'$ and $(1/\phi')''/\phi'$ belong to $\mathbb{H}^\infty(\mathcal{D})$, $\sigma/(\phi')^2$ is in $\mathbb{L}_\alpha(\mathcal{D})$, and $\tilde{y}(x) := y(x)\phi'(x)$ belongs to $\mathbb{L}_\alpha(\mathcal{D})$, then there exists a constant $c_1$, independent of $N$, such that*
$$\|\mathbf{A}_C\tilde{\mathbf{y}} - \mathbf{p}\|_2 \leq c_1 N^{1/2} e^{-(\pi d\alpha N)^{1/2}}.$$

(ii) *If the conditions in* (i) *are satisfied, and $\mu_2'/(\phi')^2$, $(\mu_2/\phi')''/(\phi')^2$, $(\mu_1/\phi')'/(\phi')^2$, $\mu_0/(\phi')^3$ and $(1/\phi')'''/(\phi')^2$ belong to $\mathbb{H}^\infty(\mathcal{D})$, then there exists a constant $c_1'$, independent of $N$, such that*
$$\|\mathbf{A}_G\tilde{\mathbf{y}} - \mathbf{p}\|_2 \leq c_1' N^{1/2} e^{-(\pi d\alpha N)^{1/2}}.$$

*It then follows immediately from* (i) and (ii) that

$$\|\mathbf{A}\tilde{\mathbf{y}} - \mathbf{p}\|_2 \leq \frac{1}{2}(c_1 + c_1') N^{1/2} e^{-(\pi d\alpha N)^{1/2}}. \tag{3.2}$$

*Proof.* See Appendix for the proof of (i) and (ii). □

We now derive an upper bound for the Euclidean norm of $\mathbf{A}^{-1}$, say, $\|\mathbf{A}^{-1}\|_2$, where the matrix $\mathbf{A}$ is defined by (2.16).

**Lemma 3.2** *Let $\mathbf{A}$ be defined as in (2.16) and $\mathbf{D}^{(0)}$ be defined as in (2.17). Assume that $\mathbf{D}^{(0)}$ is a positive-definite diagonal matrix, and $\mu_2(x) = \xi\phi'(x)$ with $\xi$ being a negative constant. Then there exists a constant $c_2$, independent of $N$, such that*

$$\|\mathbf{A}^{-1}\|_2 \leq \frac{4N^2}{d^{(2)}\pi^2}(1 + c_2 N^{-1}) \tag{3.3}$$

*holds for a sufficiently large $N$, with $d^{(2)} := -h\xi > 0$.*

*Proof.* Let $\delta_i$ $(i = 1, 2, \ldots, n)$ be the singular values of the matrix $\mathbf{A}$ satisfying $\delta_i \leq \delta_{i+1}$, and $\lambda_i(\cdot)$ $(i = 1, 2, \ldots, n)$ be the eigenvalues of the corresponding Hermitian matrix ordered as $\lambda_i(\cdot) \leq \lambda_{i+1}(\cdot)$. By making use of Lemma 2.1, in accordance with the assumptions and [21] we have

$$\begin{aligned}\delta_1 &\geq \min_{1\leq i\leq n}\left|\lambda_i\left(\frac{\mathbf{A}+\mathbf{A}^*}{2}\right)\right| = \min_{1\leq i\leq n}\left|\lambda_i(d^{(2)}\mathbf{T}^{(2)} + \mathbf{D}^{(0)})\right| \\ &\geq d^{(2)}\min_{1\leq i\leq n}\left|\lambda_i(\mathbf{T}^{(2)})\right| = 4d^{(2)}\sin^2\left(\frac{\pi}{4(N+1)}\right).\end{aligned}$$

This readily leads to the estimate in (3.3). □

A bound for the maximum norm of the function $y(x) - y_N(x)$ is described in the following theorem.

**Theorem 3.1** *Let $y$ be the exact solution of the ODE (1.1) and $y_N$ be its sinc approximation of the form (2.1). Then, under the assumptions of Lemmas 3.1 and 3.2, there exists a constant $c$, independent of $N$, such that*

$$\sup_{x\in\phi^{-1}((-\infty,\infty))} |y(x) - y_N(x)| \leq cN^{5/2}e^{-(\pi d\alpha N)^{1/2}}, \tag{3.4}$$

*holds for a sufficiently large $N$.*

*Proof.* Define the function

$$\zeta_N(x) = \frac{1}{\phi'(x)} \sum_{j=-N}^{N} y(x_j)\phi'(x_j)S(j,h)\circ\phi(x).$$

Then by making use of the triangular inequality we have

$$|y(x) - y_N(x)| \leq |y(x) - \zeta_N(x)| + |\zeta_N(x) - y_N(x)|. \tag{3.5}$$

Since $\tilde{y} \in \mathbb{L}_\alpha(\mathcal{D})$, from [27] we know that there exists a constant $c_3$, independent of $N$, such that

$$\sup_{x\in\phi^{-1}((-\infty,\infty))} |y(x) - \zeta_N(x)| \leq c_3N^{1/2}e^{-(\pi d\alpha N)^{1/2}}. \tag{3.6}$$

The second term in the right-hand side of (3.5) satisfies

$$\begin{aligned}
|\zeta_N(x) - y_N(x)| &= \left| \frac{1}{\phi'(x)} \sum_{j=-N}^{N} [\tilde{y}(x_j) - w_j]S(j,h)\circ\phi(x) \right| \\
&\leq \sum_{j=-N}^{N} |\tilde{y}(x_j) - w_j| \left| \frac{S(j,h)\circ\phi(x)}{\phi'(x)} \right| \\
&\leq \left( \sum_{j=-N}^{N} |\tilde{y}(x_j) - w_j|^2 \right)^{1/2} \left( \sum_{j=-N}^{N} \left| \frac{S(j,h)\circ\phi(x)}{\phi'(x)} \right|^2 \right)^{1/2}.
\end{aligned}$$

Because $x \in \phi^{-1}((-\infty,\infty))$, the summation $\sum_{j=-\infty}^{\infty} \left| \frac{S(j,h)\circ\phi(x)}{\phi'(x)} \right|^2$ is bounded by a constant. Hence, we further get

$$|\zeta_N(x) - y_N(x)| \leq c_3' \left( \sum_{j=-N}^{N} |\tilde{y}(x_j) - w_j|^2 \right)^{1/2} = c_3' \|\tilde{\mathbf{y}} - \mathbf{w}\|_2,$$

where $\mathbf{w}$ defined in (2.10) is the exact solution of the linear system (2.15). By (3.2) and (3.3), we can obtain

$$\|\tilde{\mathbf{y}} - \mathbf{w}\|_2 = \|\mathbf{A}^{-1}(\mathbf{A}\tilde{\mathbf{y}} - \mathbf{p})\|_2 \leq \|\mathbf{A}^{-1}\|_2 \|\mathbf{A}\tilde{\mathbf{y}} - \mathbf{p}\|_2 \leq c_3'' N^{5/2}e^{-(\pi d\alpha N)^{1/2}}, \tag{3.7}$$

where $c_3''$ is a constant independent of $N$. Now the estimate (3.4) follows immediately by substituting (3.6) and (3.7) into (3.5). □

**Remark 3.1** *The assumptions imposed on the matrix $\mathbf{D}^{(0)}$ and the coefficient $\mu_2(x)$ of the ODE (1.1) in Lemma 3.2 are stronger than necessary for guaranteeing the validity of the conclusion in Theorem 3.1, as there exist examples that the assumptions in Lemma 3.2 is violated, but $y_N(x)$ still converges to $y(x)$; see Example 5.2.*

# 4 Banded Preconditioning

In this section, we discuss how to solve the system of linear equations (2.15) efficiently by Krylov subspace iteration methods such as GMRES. The crucial point here is to construct an efficient and effective preconditioner $\mathbf{P}$ for the coefficient matrix $\mathbf{A}$ defined in (2.16). We propose to use a banded preconditioner $\mathbf{P}$ and we prove that the eigenvalues of the preconditioned matrix $\mathbf{P}^{-1}\mathbf{A}$ are uniformly bounded within a rectangle on the complex plane, which is independent of the size of the linear system.

## 4.1 Construction of the Banded Preconditioners

The banded preconditioning matrix $\mathbf{P}$ is constructed by considering the special structure of the matrix $\mathbf{A}$. In [23] and [24], the authors proposed to use banded matrices as preconditioners for Toeplitz matrices $\mathbf{T}^{(1)}$ and $\mathbf{T}^{(2)}$, which possess satisfactory theoretical properties and are computational efficient. Following the approach, we construct and study the following banded preconditioner for the matrix $\mathbf{A}$ defined in (2.16):

$$\begin{aligned}\mathbf{P} = \mathbf{B}^{(3)} &+ \frac{1}{2}(\mathbf{D}^{(2)}\mathbf{B}^{(2)} + \mathbf{B}^{(2)}\mathbf{D}^{(2)}) + \frac{1}{2}(\mathbf{D}^{(1)}\mathbf{B}^{(1)} + \mathbf{B}^{(1)}\mathbf{D}^{(1)}) \\ &+ \frac{1}{2}(\mathbf{D}_s^{(1)}\mathbf{B}^{(1)} - \mathbf{B}^{(1)}\mathbf{D}_s^{(1)}) + \mathbf{D}^{(0)}, \end{aligned} \tag{4.1}$$

where

$$\mathbf{B}^{(1)} = \text{tridiag}\left[\frac{1}{2}, 0, -\frac{1}{2}\right] \quad \text{and} \quad \mathbf{B}^{(2)} = \text{tridiag}[-1, 2, -1] \tag{4.2}$$

are tridiagonal matrices approximating the Toeplitz matrices $\mathbf{T}^{(1)}$ and $\mathbf{T}^{(2)}$, respectively, and

$$\mathbf{B}^{(3)} = \text{pentadiag}\left[-\frac{1}{2}, 1, 0, -1, \frac{1}{2}\right] \tag{4.3}$$

is a penta-diagonal matrix approximating the Toeplitz matrix $\mathbf{T}^{(3)}$. We remark that the generating functions of $\mathbf{B}^{(1)}$, $\mathbf{B}^{(2)}$ and $\mathbf{B}^{(3)}$ are $\imath \sin\theta$, $2 - 2\cos\theta$ and $\imath \sin\theta(2 - 2\cos\theta)$, respectively; see [6, 5, 7]. We also remark that the preconditioner $\mathbf{P}$ is, in whole, a penta-diagonal matrix, and hence can be inverted fast.

First of all, we estimate bounds for the eigenvalues of the banded matrices $\mathbf{B}^{(i)}$ $(i = 1, 2, 3)$.

**Lemma 4.1** *Let the banded matrices $\mathbf{B}^{(i)}$ $(i = 1, 2, 3)$ be defined as in (4.2)–(4.3). Then*

(i) $\mathbf{B}^{(1)}$ *is a skew-symmetric matrix and its eigenvalues $\{\imath\lambda_j^{(1)}\}_{j=-N}^{N}$ satisfy*

$$\lambda_j^{(1)} \in \left[-\cos\left(\frac{\pi}{2(N+1)}\right), \cos\left(\frac{\pi}{2(N+1)}\right)\right];$$

(ii) $\mathbf{B}^{(2)}$ *is a symmetric positive-definite matrix and its eigenvalues* $\{\lambda_j^{(2)}\}_{j=-N}^{N}$ *satisfy*

$$\lambda_j^{(2)} \in \left[4\sin^2\left(\frac{\pi}{4(N+1)}\right), 4\cos^2\left(\frac{\pi}{4(N+1)}\right)\right];$$

(iii) $\mathbf{B}^{(3)}$ *is a skew-symmetric matrix and its eigenvalues* $\{\imath\lambda_j^{(3)}\}_{j=-N}^{N}$ *satisfy*

$$\lambda_j^{(3)} \in \left(-\frac{3\sqrt{3}}{2}, \frac{3\sqrt{3}}{2}\right).$$

*Proof.* The results of (i) and (ii) can be found in [3]; see also [6, 5, 7]. Hence, we only need to demonstrate the validity of (iii).

Because the generating function of $\mathbf{B}^{(3)}$ is

$$f(\theta) = \imath(2 - 2\cos\theta)\sin\theta \equiv \imath\widetilde{f}(\theta),\ \theta \in [-\pi, \pi],$$

with $\widetilde{f}(\theta) = (2 - 2\cos\theta)\sin\theta$, from [17, Theorem 1.10] we have

$$\min_{-\pi\le\theta\le\pi} \widetilde{f}(\theta) \le \min \mathrm{Im}(\lambda(\mathbf{B}^{(3)})) \le \max \mathrm{Im}(\lambda(\mathbf{B}^{(3)})) \le \max_{-\pi\le\theta\le\pi} \widetilde{f}(\theta).$$

By directly calculating the minimum and the maximum values of $\widetilde{f}(\theta)$ and from [17, Theorem 1.11], we can obtain (iii). □

Hereafter in this section, we consider the coefficient matrix $\mathbf{A}$ when $\mu_2(x) = \xi\phi'(x)$, with $\xi$ a negative constant. It turns out that $\mathbf{D}_s^{(1)}$ is a zero matrix and $\mathbf{D}^{(2)} = d^{(2)}\mathbf{I}$, with $d^{(2)} := -h\xi > 0$. Thus, the coefficient matrix $\mathbf{A}$ defined in (2.16) is reduced to

$$\mathbf{A} = \mathbf{T}^{(3)} + d^{(2)}\mathbf{T}^{(2)} + \frac{1}{2}(\mathbf{D}^{(1)}\mathbf{T}^{(1)} + \mathbf{T}^{(1)}\mathbf{D}^{(1)}) + \mathbf{D}^{(0)} \tag{4.4}$$

and the banded preconditioner $\mathbf{P}$ defined in (4.1) is then given by

$$\mathbf{P} = \mathbf{B}^{(3)} + d^{(2)}\mathbf{B}^{(2)} + \frac{1}{2}(\mathbf{D}^{(1)}\mathbf{B}^{(1)} + \mathbf{B}^{(1)}\mathbf{D}^{(1)}) + \mathbf{D}^{(0)}. \tag{4.5}$$

The following theorem shows that both matrices $\mathbf{A}$ and $\mathbf{P}$ are positive definite.

**Theorem 4.1** *Assume that* $\mathbf{D}^{(0)}$ *defined in (2.17) is a positive-definite diagonal matrix. Then both* $\mathcal{H}(\mathbf{A})$ *and* $\mathcal{H}(\mathbf{P})$ *are symmetric positive-definite matrices. Hence,* $\mathbf{A}$ *and* $\mathbf{P}$ *are positive definite and, thus, nonsingular.*

*Proof.* Evidently, the Hermitian and the skew-Hermitian parts of $\mathbf{A}$ and $\mathbf{P}$ are given as follows:

$$\begin{aligned}
\mathcal{H}(\mathbf{A}) &= \frac{1}{2}(\mathbf{A} + \mathbf{A}^*) = d^{(2)}\mathbf{T}^{(2)} + \mathbf{D}^{(0)},\\
\mathcal{S}(\mathbf{A}) &= \frac{1}{2}(\mathbf{A} - \mathbf{A}^*) = \mathbf{T}^{(3)} + \frac{1}{2}\left(\mathbf{D}^{(1)}\mathbf{T}^{(1)} + \mathbf{T}^{(1)}\mathbf{D}^{(1)}\right),\\
\mathcal{H}(\mathbf{P}) &= \frac{1}{2}(\mathbf{P} + \mathbf{P}^*) = d^{(2)}\mathbf{B}^{(2)} + \mathbf{D}^{(0)},\\
\mathcal{S}(\mathbf{P}) &= \frac{1}{2}(\mathbf{P} - \mathbf{P}^*) = \mathbf{B}^{(3)} + \frac{1}{2}\left(\mathbf{D}^{(1)}\mathbf{B}^{(1)} + \mathbf{B}^{(1)}\mathbf{D}^{(1)}\right).
\end{aligned}$$

Because $d^{(2)} > 0$, the diagonal matrix $\mathbf{D}^{(0)}$ is positive definite and the Toeplitz matrix $\mathbf{T}^{(2)}$ is symmetric positive definite (see Lemma 2.1), we know that $\mathcal{H}(\mathbf{A})$ is symmetric positive definite. Therefore, $\mathbf{A}$ is a positive definite matrix and is, thus, nonsingular.

By applying the same arguments to the preconditioning matrix $\mathbf{P}$, with Lemma 4.1 we can immediately show that $\mathbf{P}$ is positive definite and nonsingular, too. □

## 4.2 Several Preliminary Lemmas

In this section, we establish several lemmas that are indispensable for estimating eigenvalue bounds for the preconditioned matrix $\mathbf{P}^{-1}\mathbf{A}$. The generalized Bendixson theorem, established in [7], is an essential tool for deriving a rectangular domain that bounds the eigenvalues of the preconditioned matrix $\mathbf{P}^{-1}\mathbf{A}$.

**Theorem 4.2** *[7, Theorem 2.4] Let $\mathbf{A}, \mathbf{P} \in \mathbb{C}^{n\times n}$ be complex matrices such that, for all $\mathbf{x} \in \mathbb{C}^n\backslash\{0\}$, $\mathbf{x}^*\mathcal{H}(\mathbf{A})\mathbf{x} \neq 0$ and $\mathbf{x}^*\mathcal{H}(\mathbf{P})\mathbf{x} \neq 0$. Let the functions $h(\mathbf{x})$, $f_A(\mathbf{x})$ and $f_P(\mathbf{x})$ be defined as*

$$h(\mathbf{x}) = \frac{\mathbf{x}^*\mathcal{H}(\mathbf{A})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{P})\mathbf{x}}, \quad f_A(\mathbf{x}) = \frac{1}{\imath}\frac{\mathbf{x}^*\mathcal{S}(\mathbf{A})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{A})\mathbf{x}} \quad \text{and} \quad f_P(\mathbf{x}) = \frac{1}{\imath}\frac{\mathbf{x}^*\mathcal{S}(\mathbf{P})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{P})\mathbf{x}}.$$

*Assume that there exist positive constants $\tau_1$ and $\tau_2$ such that*

$$\tau_1 \leq h(\mathbf{x}) \leq \tau_2, \quad \forall \mathbf{x} \in \mathbb{C}^n\backslash\{0\},$$

*and nonnegative constants $\eta$ and $\mu$ such that*

$$-\eta \leq f_A(\mathbf{x}) \leq \eta \quad \text{and} \quad -\mu \leq f_P(\mathbf{x}) \leq \mu, \quad \forall \mathbf{x} \in \mathbb{C}^n\backslash\{0\}.$$

*Then, when $\eta\mu < 1$, we have*

$$\begin{cases} \dfrac{(1-\eta\mu)\tau_1}{1+\mu^2} \leq \mathrm{Re}(\lambda(\mathbf{P}^{-1}\mathbf{A})) \leq (1+\eta\mu)\tau_2, \\ -(\eta+\mu)\tau_2 \leq \mathrm{Im}(\lambda(\mathbf{P}^{-1}\mathbf{A})) \leq (\eta+\mu)\tau_2. \end{cases}$$

In order to use Theorem 4.2 to bound the eigenvalues of $\mathbf{P}^{-1}\mathbf{A}$, we need bounds for the generalized Rayleigh quotients with respect to the Hermitian and the skew-Hermitian parts of $\mathbf{A}$ and $\mathbf{P}$. To this end, we need to review and establish the following lemmas.

**Lemma 4.2** *[7] Let $\mathbf{T}^{(2)}$ and $\mathbf{B}^{(2)}$ be the matrices defined in (2.7) and (4.2), respectively. Then it holds that*

$$1 < \frac{\mathbf{x}^*\mathbf{T}^{(2)}\mathbf{x}}{\mathbf{x}^*\mathbf{B}^{(2)}\mathbf{x}} < \frac{\pi^2}{4}, \quad \forall \mathbf{x} \in \mathbb{C}^n\backslash\{0\}.$$

**Lemma 4.3** *Let $\mathbf{T}^{(i)}$ $(i = 1, 3)$ be the Toeplitz matrices defined in (2.6) and (2.8). Denote by $f_i(\theta)$ the generating function of $\mathbf{T}^{(i)}$ and define*

$$d^{(0)} = \min_{1\leq\ell\leq n}\{|[\mathbf{D}^{(0)}]_{\ell\ell}|\}. \tag{4.6}$$

*Then, for all* $\mathbf{x} \in \mathbb{C}^n \backslash \{0\}$*, it holds that*

$$\max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^* \mathbf{T}^{(i)} (\mathbf{T}^{(i)})^* \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{T}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right\} < \max_{-\pi \leq \theta \leq \pi} \left\{ \frac{|f_i(\theta)|^2}{d^{(2)}\theta^2 + d^{(0)}} \right\}, \quad i = 1, 3.$$

*Proof.* From Lemma 2.1 we know that $\mathbf{T}^{(i)}$ $(i = 1, 3)$ are skew-symmetric Toeplitz matrices and their generating functions are in the Wiener class. By making use of Theorems 3.1 and 3.3 in [8], we know that for any $\epsilon > 0$ there exist positive semidefinite matrices $\mathbf{R}_i$ of fixed ranks and matrices $\mathbf{E}_i$ of small norms such that $\|\mathbf{E}_i\|_2 < d^{(0)}\epsilon$ and

$$\mathbf{T}^{(i)} (\mathbf{T}^{(i)})^* + \mathbf{R}_i + \mathbf{E}_i = \widehat{\mathbf{T}}^{(i)},$$

where $\widehat{\mathbf{T}}^{(i)}$ are the Toeplitz matrices generated by the positive functions $|f_i(\theta)|^2$. Because $\mathbf{T}^{(2)}$ is positive definite, we have

$$\frac{\mathbf{x}^* \mathbf{R}_i \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{T}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \geq 0 \quad \text{and} \quad \left| \frac{\mathbf{x}^* \mathbf{E}_i \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{T}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right| \leq \epsilon, \quad \forall \mathbf{x} \neq 0.$$

It then follows from the above matrix decompositions that

$$\max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^* \mathbf{T}^{(i)} (\mathbf{T}^{(i)})^* \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{T}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right\} < \max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^* \widehat{\mathbf{T}}^{(i)} \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{T}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right\} + \epsilon.$$

Since $\epsilon$ is arbitrary, this inequality readily implies

$$\max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^* \mathbf{T}^{(i)} (\mathbf{T}^{(i)})^* \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{T}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right\} \leq \max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^* \widehat{\mathbf{T}}^{(i)} \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{T}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right\} < \max_{-\pi \leq \theta \leq \pi} \left\{ \frac{|f_i(\theta)|^2}{d^{(2)}\theta^2 + d^{(0)}} \right\}.$$

□

**Lemma 4.4** *Let* $\mathbf{B}^{(i)}$ $(i = 1, 3)$ *be the matrices defined in (4.2)–(4.3) and denote by* $g_i(\theta)$ *the generating function of* $\mathbf{B}^{(i)}$*. Then, for all* $\mathbf{x} \in \mathbb{C}^n \backslash \{0\}$*, it holds that*

$$\max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^* \mathbf{B}^{(i)} (\mathbf{B}^{(i)})^* \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{B}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right\} < \max_{-\pi \leq \theta \leq \pi} \left\{ \frac{|g_i(\theta)|^2}{d^{(2)}(2 - 2\cos\theta) + d^{(0)}} \right\}, \quad i = 1, 3,$$

*where* $d^{(0)}$ *is defined as in (4.6).*

*Proof.* From Lemma 4.1 we know that $\mathbf{B}^{(i)}$ $(i = 1, 3)$ are skew-symmetric banded Toeplitz matrices. By making use of [8, Theorem 3.1] again we know that there exist positive semidefinite matrices $\mathbf{F}_i$ of fixed ranks such that $\mathbf{B}^{(i)}(\mathbf{B}^{(i)})^* + \mathbf{F}_i = \widehat{\mathbf{B}}^{(i)}$, where $\widehat{\mathbf{B}}^{(i)}$ are the Toeplitz matrices generated by the positive functions $|g_i(\theta)|^2$. Because $\mathbf{B}^{(2)}$ is positive definite, we have

$$\frac{\mathbf{x}^* \mathbf{F}_i \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{B}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \geq 0, \quad \forall \mathbf{x} \neq 0.$$

It then follows from the above matrix decompositions that

$$\max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^* \mathbf{B}^{(i)} (\mathbf{B}^{(i)})^* \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{B}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right\} \leq \max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^* \widehat{\mathbf{B}}^{(i)} \mathbf{x}}{\mathbf{x}^* (d^{(2)} \mathbf{B}^{(2)} + d^{(0)} \mathbf{I}) \mathbf{x}} \right\}$$
$$< \max_{-\pi \leq \theta \leq \pi} \left\{ \frac{|g_i(\theta)|^2}{d^{(2)}(2 - 2\cos\theta) + d^{(0)}} \right\}.$$

□

**Lemma 4.5** *Assume that* $\mathbf{D}^{(0)}$ *defined in (2.17) is a positive-definite diagonal matrix. Define*

$$d^{(1)} = \max_{1\le\ell\le n}\{|[\mathbf{D}^{(1)}]_{\ell\ell}|\}. \tag{4.7}$$

*Then, for all* $\mathbf{x}\in\mathbb{C}^n\backslash\{0\}$*, it holds that*

$$1 < \left|\frac{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right| < \frac{\pi^2}{4}, \tag{4.8}$$

$$\left|\frac{\mathbf{x}^*\mathbf{T}^{(3)}\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right| < \frac{\pi^3}{\sqrt{d^{(0)}(d^{(2)}\pi^2+d^{(0)})}}, \tag{4.9}$$

$$\left|\frac{\mathbf{x}^*\mathbf{B}^{(3)}\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right| < \frac{2(\sqrt{4d^{(2)}+d^{(0)}}-\sqrt{d^{(0)}})}{d^{(2)}\sqrt{d^{(0)}}}, \tag{4.10}$$

$$\left|\frac{\mathbf{x}^*(\mathbf{D}^{(1)}\mathbf{T}^{(1)}+\mathbf{T}^{(1)}\mathbf{D}^{(1)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right| < \frac{2d^{(1)}\pi}{\sqrt{d^{(0)}(d^{(2)}\pi^2+d^{(0)})}}, \tag{4.11}$$

$$\left|\frac{\mathbf{x}^*(\mathbf{D}^{(1)}\mathbf{B}^{(1)}+\mathbf{B}^{(1)}\mathbf{D}^{(1)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right| < \frac{d^{(1)}(\sqrt{4d^{(2)}+d^{(0)}}-\sqrt{d^{(0)}})}{d^{(2)}\sqrt{d^{(0)}}}. \tag{4.12}$$

*Proof.* We first demonstrate the estimate (4.8). Because the diagonal matrix $\mathbf{D}^{(0)}$ is positive definite, from Lemma 4.2 we know that

$$\frac{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}} \le \max_{\mathbf{x}\ne0}\left\{\frac{\mathbf{x}^*\mathbf{T}^{(2)}\mathbf{x}}{\mathbf{x}^*\mathbf{B}^{(2)}\mathbf{x}},\, 1\right\} < \frac{\pi^2}{4}$$

holds for any $\mathbf{x}\in\mathbb{C}^n\backslash\{0\}$. By similar argument we can get the lower bound in (4.8).

Now, we are going to verify the validity of (4.9) and (4.10). Because $\mathbf{T}^{(3)}$ is a skew-symmetric matrix and $\mathbf{T}^{(2)}$ is a symmetric positive definite matrix, for any $\mathbf{x}\in\mathbb{C}^n\backslash\{0\}$, by direct computations we can obtain the following estimates:

$$\begin{aligned}
&\left|\frac{\mathbf{x}^*\mathbf{T}^{(3)}\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right| \\
\le\;& \max_{\mathbf{x}\ne0}\left\{\left|\frac{\mathbf{x}^*(\imath\mathbf{T}^{(3)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right|\right\} \\
=\;& \max\left\{\left|\lambda\left((d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}(\imath\mathbf{T}^{(3)})(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\right)\right|\right\} \\
=\;& \left\|(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\mathbf{T}^{(3)}(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\right\|_2 \\
\le\;& \left\|(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{1/2}\right\|_2 \\
&\cdot\left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2\cdot\left\|\mathbf{T}^{(3)}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2 \\
&\cdot\left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{1/2}(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\right\|_2 \\
\le\;& \left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2\cdot\left\|\mathbf{T}^{(3)}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2.
\end{aligned} \tag{4.13}$$

Here, we have applied the facts

$$\left\|(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{1/2}\right\|_2=\left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{1/2}(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\right\|_2$$

and

$$\begin{aligned}&\left\|(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{1/2}\right\|_2^2\\ =\ &\lambda_{\max}\left((d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\right)\\ =\ &\max_{\mathbf{x}\neq 0}\left\{\frac{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right\}\leq\max\left\{1,\quad\max_{\mathbf{x}\neq 0}\frac{d^{(0)}\mathbf{x}^*\mathbf{x}}{\mathbf{x}^*\mathbf{D}^{(0)}\mathbf{x}}\right\}\leq 1.\end{aligned}$$

We further estimate the two terms on the right-hand side of the inequality (4.13). As the generating function of the Toeplitz matrix $\mathbf{T}^{(2)}$ is $\theta^2$, we have

$$\begin{aligned}&\left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2^2\\ =\ &\lambda_{\max}\left((d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1}\right)\\ =\ &\max_{\mathbf{x}\neq 0}\left\{\frac{\mathbf{x}^*\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})\mathbf{x}}\right\}<\max_{-\pi\leq\theta\leq\pi}\left\{\frac{1}{d^{(2)}\theta^2+d^{(0)}}\right\}=\frac{1}{d^{(0)}}.\end{aligned}$$

Therefore, it holds that

$$\left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2<\frac{1}{\sqrt{d^{(0)}}}. \tag{4.14}$$

In addition, recalling that

$$\begin{aligned}\left\|\mathbf{T}^{(3)}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2^2&=\lambda_{\max}\left((d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{T}^{(3)}(\mathbf{T}^{(3)})^*(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right)\\ &=\max_{\mathbf{x}\neq 0}\left\{\frac{\mathbf{x}^*\mathbf{T}^{(3)}(\mathbf{T}^{(3)})^*\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})\mathbf{x}}\right\},\end{aligned}$$

since the generating function of $\mathbf{T}^{(3)}$ is $\imath\theta^3$, from Lemma 4.3 we have

$$\max_{\mathbf{x}\neq 0}\left\{\frac{\mathbf{x}^*\mathbf{T}^{(3)}(\mathbf{T}^{(3)})^*\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})\mathbf{x}}\right\}<\max_{-\pi\leq\theta\leq\pi}\left\{\frac{\theta^6}{d^{(2)}\theta^2+d^{(0)}}\right\}=\frac{\pi^6}{d^{(2)}\pi^2+d^{(0)}}.$$

Therefore, it holds that

$$\left\|\mathbf{T}^{(3)}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2<\frac{\pi^3}{\sqrt{d^{(2)}\pi^2+d^{(0)}}}. \tag{4.15}$$

By substituting (4.14) and (4.15) into (4.13), we immediately obtain

$$\left|\frac{\mathbf{x}^*\mathbf{T}^{(3)}\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right|<\frac{\pi^3}{\sqrt{d^{(0)}(d^{(2)}\pi^2+d^{(0)})}},$$

which is exactly the estimate in (4.9).

Analogous to the derivation of (4.13), for any $\mathbf{x} \neq 0$ we can obtain

$$\left| \frac{\mathbf{x}^*\mathbf{B}^{(3)}\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)} + \mathbf{D}^{(0)})\mathbf{x}} \right| \leq \left\| (d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1/2} \right\|_2 \cdot \left\| \mathbf{B}^{(3)}(d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1/2} \right\|_2. \tag{4.16}$$

Here we have applied the fact

$$\left\| (d^{(2)}\mathbf{B}^{(2)} + \mathbf{D}^{(0)})^{-1/2}(d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{1/2} \right\|_2 = \left\| (d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{1/2}(d^{(2)}\mathbf{B}^{(2)} + \mathbf{D}^{(0)})^{-1/2} \right\|_2 \leq 1.$$

We further estimate the two terms on the right-hand side of the inequality (4.16). As the generating function of the Toeplitz matrix $\mathbf{B}^{(2)}$ is $(2 - 2\cos\theta)$, we have

$$\begin{aligned} & \left\| (d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1/2} \right\|_2^2 \\ = \; & \lambda_{\max}\left( (d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1} \right) \\ = \; & \max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^*\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})\mathbf{x}} \right\} < \max_{-\pi\leq\theta\leq\pi} \left\{ \frac{1}{d^{(2)}(2 - 2\cos\theta) + d^{(0)}} \right\} = \frac{1}{d^{(0)}}. \end{aligned}$$

Therefore, it holds that

$$\left\| (d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1/2} \right\|_2 < \frac{1}{\sqrt{d^{(0)}}}. \tag{4.17}$$

In addition, by recalling that the generating function of the Toeplitz matrix $\mathbf{B}^{(3)}$ is $\imath(2 - 2\cos\theta)\sin\theta$, we can obtain from Lemma 4.4 that

$$\begin{aligned} \left\| \mathbf{B}^{(3)}(d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1/2} \right\|_2^2 &= \lambda_{\max}\left( (d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1/2}\mathbf{B}^{(3)}(\mathbf{B}^{(3)})^*(d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1/2} \right) \\ &= \max_{\mathbf{x}\neq 0} \left\{ \frac{\mathbf{x}^*\mathbf{B}^{(3)}(\mathbf{B}^{(3)})^*\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})\mathbf{x}} \right\} \\ &< \max_{-\pi\leq\theta\leq\pi} \left\{ \frac{(2 - 2\cos\theta)^2\sin^2\theta}{d^{(2)}(2 - 2\cos\theta) + d^{(0)}} \right\} \\ &\leq \frac{4\left(\sqrt{4d^{(2)} + d^{(0)}} - \sqrt{d^{(0)}}\right)^2}{(d^{(2)})^2}. \end{aligned}$$

Therefore, it holds that

$$\left\| \mathbf{B}^{(3)}(d^{(2)}\mathbf{B}^{(2)} + d^{(0)}\mathbf{I})^{-1/2} \right\|_2 < \frac{2(\sqrt{4d^{(2)} + d^{(0)}} - \sqrt{d^{(0)}})}{d^{(2)}}. \tag{4.18}$$

By substituting (4.17) and (4.18) into (4.16), we immediately obtain the estimate in (4.10).

Finally, we demonstrate the estimates (4.11) and (4.12). Because $\mathbf{D}^{(1)}\mathbf{T}^{(1)}+\mathbf{T}^{(1)}\mathbf{D}^{(1)}$ is a skew-symmetric matrix, for all $\mathbf{x}\neq 0$ we have

$$\begin{aligned}
&\left|\frac{\mathbf{x}^*(\mathbf{D}^{(1)}\mathbf{T}^{(1)}+\mathbf{T}^{(1)}\mathbf{D}^{(1)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right|\\
\leq\ & \max_{\mathbf{x}\neq 0}\left\{\left|\frac{\mathbf{x}^*\imath(\mathbf{D}^{(1)}\mathbf{T}^{(1)}+\mathbf{T}^{(1)}\mathbf{D}^{(1)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right|\right\}\\
=\ & \left\|(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\imath(\mathbf{D}^{(1)}\mathbf{T}^{(1)}+\mathbf{T}^{(1)}\mathbf{D}^{(1)})(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\right\|_2\\
\leq\ & \left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2\cdot\left\|\mathbf{D}^{(1)}\right\|_2\\
&\quad\cdot\left\{\left\|\mathbf{T}^{(1)}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2+\left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{T}^{(1)}\right\|_2\right\}.
\end{aligned}\tag{4.19}$$

Noticing that the generating function of the Toeplitz matrix $\mathbf{T}^{(1)}$ is $\imath\theta$, from Lemma 4.3 we can obtain

$$\left\|\mathbf{T}^{(1)}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2=\left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{T}^{(1)}\right\|_2$$

and

$$\begin{aligned}
\left\|\mathbf{T}^{(1)}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2^2 &= \lambda_{\max}\left((d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{T}^{(1)}(\mathbf{T}^{(1)})^*(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right)\\
&= \max_{\mathbf{x}\neq 0}\left\{\frac{\mathbf{x}^*\mathbf{T}^{(1)}(\mathbf{T}^{(1)})^*\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})\mathbf{x}}\right\}\\
&< \max_{-\pi\leq\theta\leq\pi}\left\{\frac{\theta^2}{d^{(2)}\theta^2+d^{(0)}}\right\}=\frac{\pi^2}{d^{(2)}\pi^2+d^{(0)}}.
\end{aligned}$$

Therefore, it holds that

$$\left\|\mathbf{T}^{(1)}(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2=\left\|(d^{(2)}\mathbf{T}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{T}^{(1)}\right\|_2<\frac{\pi}{\sqrt{d^{(2)}\pi^2+d^{(0)}}}.\tag{4.20}$$

By substituting (4.14) and (4.20) into (4.19), we immediately obtain the inequality in (4.11).

Analogous to the derivation of (4.19), by noticing that $\mathbf{D}^{(1)}\mathbf{B}^{(1)}+\mathbf{B}^{(1)}\mathbf{D}^{(1)}$ is a skew-symmetric matrix, we have for all $\mathbf{x}\neq 0$ that

$$\begin{aligned}
&\left|\frac{\mathbf{x}^*(\mathbf{D}^{(1)}\mathbf{B}^{(1)}+\mathbf{B}^{(1)}\mathbf{D}^{(1)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right|\\
\leq\ & \max_{\mathbf{x}\neq 0}\left\{\left|\frac{\mathbf{x}^*\imath(\mathbf{D}^{(1)}\mathbf{B}^{(1)}+\mathbf{B}^{(1)}\mathbf{D}^{(1)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}\right|\right\}\\
=\ & \left\|(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})^{-1/2}(\mathbf{D}^{(1)}\mathbf{B}^{(1)}+\mathbf{B}^{(1)}\mathbf{D}^{(1)})(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})^{-1/2}\right\|_2\\
\leq\ & \left\|(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2\cdot\left\|\mathbf{D}^{(1)}\right\|_2\\
&\quad\cdot\left\{\left\|\mathbf{B}^{(1)}(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2+\left\|(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{B}^{(1)}\right\|_2\right\}.
\end{aligned}\tag{4.21}$$

Because the generating function of $\mathbf{B}^{(1)}$ is $\imath \sin\theta$, from Lemma 4.4 we know

$$\left\|\mathbf{B}^{(1)}(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2 = \left\|(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{B}^{(1)}\right\|_2$$

and

$$\begin{aligned}\left\|\mathbf{B}^{(1)}(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2^2 &= \lambda_{\max}\left((d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{B}^{(1)}(\mathbf{B}^{(1)})^*(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right)\\ &= \max_{\mathbf{x}\neq 0}\left\{\frac{\mathbf{x}^*\mathbf{B}^{(1)}(\mathbf{B}^{(1)})^*\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})\mathbf{x}}\right\}\\ &< \max_{-\pi\leq\theta\leq\pi}\left\{\frac{\sin^2\theta}{d^{(2)}(2-2\cos\theta)+d^{(0)}}\right\}\\ &= \frac{\left(\sqrt{4d^{(2)}+d^{(0)}}-\sqrt{d^{(0)}}\right)^2}{4(d^{(2)})^2}.\end{aligned}$$

Therefore, it holds that

$$\left\|\mathbf{B}^{(1)}(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\right\|_2 = \left\|(d^{(2)}\mathbf{B}^{(2)}+d^{(0)}\mathbf{I})^{-1/2}\mathbf{B}^{(1)}\right\|_2 < \frac{\sqrt{4d^{(2)}+d^{(0)}}-\sqrt{d^{(0)}}}{2d^{(2)}}. \quad (4.22)$$

By substituting (4.17) and (4.22) into (4.21), we immediately obtain the estimate in (4.12). □

## 4.3 Analysis of the Preconditioned Matrix

In this section, we derive eigenvalue bounds for the preconditioned matrix $\mathbf{P}^{-1}\mathbf{A}$. To this end, we first estimate bounds for the function $h(\mathbf{x})$ defined in Theorem 4.2.

**Lemma 4.6** *Assume that $\mathbf{D}^{(0)}$ defined in (2.17) is a positive-definite diagonal matrix. Let $\mathbf{A}$ and $\mathbf{P}$ be defined in (4.4) and (4.5), respectively. Then it holds that*

$$1 < \frac{\mathbf{x}^*\mathcal{H}(\mathbf{A})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{P})\mathbf{x}} < \frac{\pi^2}{4}, \quad \forall \mathbf{x}\in\mathbb{C}^n\backslash\{0\}.$$

*Proof.* From Lemma 4.5, for all $\mathbf{x}\neq 0$ we have

$$\frac{\mathbf{x}^*\mathcal{H}(\mathbf{A})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{P})\mathbf{x}} = \frac{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{B}^{(2)}+\mathbf{D}^{(0)})\mathbf{x}} < \frac{\pi^2}{4}.$$

Similarly, we can obtain the lower bound in Lemma 4.6. □

For bounds about the functions $f_A(\mathbf{x})$ and $f_P(\mathbf{x})$ defined in Theorem 4.2, we have the estimates described in the following lemma.

**Lemma 4.7** *Assume that $\mathbf{D}^{(0)}$ defined in (2.17) is a positive-definite diagonal matrix. Let $\mathbf{A}$ and $\mathbf{P}$ be defined in (4.4) and (4.5), respectively. Denote by*

$$\eta = \frac{\pi(\pi^2+d^{(1)})}{\sqrt{d^{(0)}(d^{(2)}\pi^2+d^{(0)})}}, \quad \mu = \frac{(4+d^{(1)})(\sqrt{4d^{(2)}+d^{(0)}}-\sqrt{d^{(0)}})}{2d^{(2)}\sqrt{d^{(0)}}}, \quad (4.23)$$

*where $d^{(0)}$ and $d^{(1)}$ are defined in (4.6) and (4.7), respectively . Then, for all $\mathbf{x} \in \mathbb{C}^n \backslash \{0\}$, it holds that*

$$\left|\frac{\mathbf{x}^*\mathcal{S}(\mathbf{A})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{A})\mathbf{x}}\right| \leq \eta \quad \text{and} \quad \left|\frac{\mathbf{x}^*\mathcal{S}(\mathbf{P})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{P})\mathbf{x}}\right| \leq \mu.$$

*Proof.* By making use of Lemma 4.5, with straightforward computations we have

$$\begin{aligned}\left|\frac{\mathbf{x}^*\mathcal{S}(\mathbf{A})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{A})\mathbf{x}}\right| &= \frac{1}{2}\left|\frac{\mathbf{x}^*(2\mathbf{T}^{(3)} + \mathbf{D}^{(1)}\mathbf{T}^{(1)} + \mathbf{T}^{(1)}\mathbf{D}^{(1)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)} + \mathbf{D}^{(0)})\mathbf{x}}\right| \\ &\leq \left|\frac{\mathbf{x}^*\mathbf{T}^{(3)}\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)} + \mathbf{D}^{(0)})\mathbf{x}}\right| + \frac{1}{2}\left|\frac{\mathbf{x}^*(\mathbf{D}^{(1)}\mathbf{T}^{(1)} + \mathbf{T}^{(1)}\mathbf{D}^{(1)})\mathbf{x}}{\mathbf{x}^*(d^{(2)}\mathbf{T}^{(2)} + \mathbf{D}^{(0)})\mathbf{x}}\right| \leq \eta.\end{aligned}$$

Similarly, we can get

$$\left|\frac{\mathbf{x}^*\mathcal{S}(\mathbf{P})\mathbf{x}}{\mathbf{x}^*\mathcal{H}(\mathbf{P})\mathbf{x}}\right| \leq \mu.$$

□

Based on Theorem 4.2 and Lemmas 4.6 and 4.7, we can readily obtain a domain that bounds the eigenvalues of the preconditioned matrix $\mathbf{P}^{-1}\mathbf{A}$.

**Theorem 4.3** *Assume that $\mathbf{D}^{(0)}$ defined in (2.17) is a positive-definite diagonal matrix. Let $\mathbf{A}$ and $\mathbf{P}$ be defined in (4.4) and (4.5), respectively. Then it holds that*

$$\frac{1-\eta\mu}{1+\mu^2} \leq \mathrm{Re}(\lambda(\mathbf{P}^{-1}\mathbf{A})) \leq \frac{\pi^2(1+\eta\mu)}{4}, \quad \textit{for} \quad \mu\eta < 1,$$

*and*

$$-\frac{\pi^2(\eta+\mu)}{4} \leq \mathrm{Im}(\lambda(\mathbf{P}^{-1}\mathbf{A})) \leq \frac{\pi^2(\eta+\mu)}{4},$$

*where $\eta$ and $\mu$ are defined in (4.23).*

By employing Theorem 4.3 we can immediately obtain a theoretical estimate about the asymptotic convergence rate of the preconditioned GMRES method, with the preconditioner $\mathbf{P}$ being defined in (4.5), for solving the system of linear equations (2.15). For details, we refer to [6, 7, 30].

When using Theorem 4.3, we should suitably scale the ODE (1.1) and appropriately choose the conformal mapping $\phi(x)$ such that $\mu\eta < 1$, so that correct and accurate estimates about the eigenvalue bounds may be obtained. For example, if we take $\phi(x) = \nu \ln(x/(1-x))$ in Example 5.1, with $\nu > 0$ a scaling factor, then corresponding to different mesh sizes $h = \pi/\sqrt{2N}$ we can obtain the following computed and estimated eigenvalue bounds about the the preconditioned matrix $\mathbf{P}^{-1}\mathbf{A}$:

(i) $N = 8$ and $\nu = 10^4$, the computed eigenvalues are bounded in the rectangle $[1.0000, 1.0057] \times [-0.0975, 0.0975]$, while the estimated eigenvalues are bounded in the rectangle $[0.8332, 2.7925] \times [-2.0928, 2.0928]$;

(ii) $N = 16$ and $\nu = 10^5$, the computed eigenvalues are bounded in the rectangle $[1.0000, 1.0283] \times [-0.2323, 0.2323]$, while the estimated eigenvalues are bounded in the rectangle $[0.4520, 3.6495] \times [-3.9906, 3.9906]$;

(iii) $N = 32$ and $\nu = 10^7$, the computed eigenvalues are bounded in the rectangle $[1.0000, 1.0118]\times[-0.1643, 0.1643]$, while the estimated eigenvalues are bounded in the rectangle $[0.8105, 2.8391]\times[-2.2383, 2.2383]$.

Clearly, these estimated rectangles contain sharply the computed eigenvalues of the preconditioned matrices.

## 5 Numerical Examples

In this section, we examine the accuracy of the sinc discretization and test the effectiveness of the proposed banded preconditioner. To this end, we apply GMRES and BiCGSTAB, incorporated with the banded preconditioner $\mathbf{P}$ defined in (4.1), to the system of linear equations (2.15) obtained from the sinc discretization of the ODE (1.1).

The two examples of the ODEs used in our numerical performance are given below.

**Example 5.1** *The third-order ODE*

$$\begin{cases} y^{'''}(x) - \dfrac{1}{x(1-x)}y^{''}(x) - \dfrac{1}{x^2}y^{'}(x) + \dfrac{1}{x^3}y(x) = 21x + 4 - \dfrac{3}{x} - \dfrac{2}{1-x}, \\ y(0) = 0, \quad y(1) = 0, \quad y^{'}(0) = 0, \end{cases}$$

*with the exact solution being* $y(x) = x^2(1-x)^2$.

In Example 5.1, the conformal mapping $\phi(x)$ is chosen as $\phi(x) = \ln(x/(1-x))$ and the mesh size is set to be the optimal one $h = \pi/\sqrt{2N}$ since $\tilde{y} = y\phi^{'} \in \mathbb{L}_\alpha(\mathcal{D})$. With calculations we can verify that this problem satisfies all assumptions in Theorems 3.1 and 4.3.

**Example 5.2** *The third-order ODE*

$$\begin{cases} y^{'''}(x) - y^{''}(x) - y^{'}(x) + y(x) = \sigma(x), \\ y(0) = 0, \quad y(1) = 0, \quad y^{'}(0) = 0, \\ \sigma(x) = (\pi^2 + 1)\sin(\pi x) - (\pi^3 + \pi)\cos(\pi x) + \pi(x^2 - 3x - 1), \end{cases}$$

*with the exact solution being* $y(x) = \sin(\pi x) + \pi(x^2 - x)$.

In Example 5.2, the conformal mapping $\phi(x)$ is chosen as $\phi(x) = \ln(x/(1-x))$ and the mesh size is set to be the optimal one $h = \pi/\sqrt{2N}$. Note that this problem does not satisfy some of the assumptions, e.g., $\mu_2(x) = \xi\phi^{'}(x)$, in Theorem 3.1.

Both test examples are ODEs of homogeneous boundary values and with known solutions, which make it easy to verify the accuracy of both discrete and computed solutions.

In our performance, both GMRES and BiCGSTAB are applied to the preconditioned linear system

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{w} = \mathbf{P}^{-1}\mathbf{p},$$

where $\mathbf{P}$ represents the banded preconditioner given in (4.1) and $\mathbf{p}$ stands for the vector given in (2.11). We remark that at each step GMRES uses only one matrix-vector product, whereas

Table 5.1: Numerical Results for Example 5.1

| | | GMRES | | BiCGSTAB | |
|---|---|---|---|---|---|
| $N$ | $E_s(h)$ | $\mathbf{I}_{\text{iter}}$ | $\mathbf{P}_{\text{iter}}$ | $\mathbf{I}_{\text{iter}}$ | $\mathbf{P}_{\text{iter}}$ |
| 8 | 3.26e-05 | 17 | 14 | 32 | 10 |
| 16 | 2.16e-06 | 33 | 19 | 156 | 15 |
| 32 | 3.66e-08 | 65 | 26 | $**$ | 24 |
| 64 | 1.20e-10 | 129 | 35 | $**$ | 38 |
| 128 | 3.91e-14 | 257 | 46 | $**$ | 64 |
| 256 | 1.41e-14 | 513 | 60 | $**$ | 106 |

Table 5.2: Numerical Results for Example 5.2

| | | GMRES | | BiCGSTAB | |
|---|---|---|---|---|---|
| $N$ | $E_s(h)$ | $\mathbf{I}_{\text{iter}}$ | $\mathbf{P}_{\text{iter}}$ | $\mathbf{I}_{\text{iter}}$ | $\mathbf{P}_{\text{iter}}$ |
| 8 | 2.06e-04 | 17 | 14 | 174 | 13 |
| 16 | 7.42e-06 | 33 | 19 | $**$ | 18 |
| 32 | 9.57e-08 | 65 | 27 | $**$ | 27 |
| 64 | 2.93e-10 | 129 | 35 | $**$ | 43 |
| 128 | 1.95e-13 | 257 | 45 | $**$ | 86 |
| 256 | 2.11e-13 | 513 | 58 | $**$ | 100 |

BiCGSTAB uses two matrix-vector products. All codes are written in MATLAB 7.04 and all experiments are done on a personal computer with 0.98G memory. In addition, the initial guess is taken to be zero and the iteration process is terminated once the current residual $\mathbf{r}^{(j)}$ satisfies

$$\frac{\|\mathbf{r}^{(j)}\|_2}{\|\mathbf{r}^{(0)}\|_2} \leq 10^{-6}.$$

Tables 5.1 and 5.2 list the numbers of iteration steps and the errors $E_s(h)$ between the numerical approximate solution $y_N(x)$ and the true solution $y(x)$ at the sinc points. More precisely, the error $E_s(h)$ is defined as

$$E_s(h) = \max_{-N\leq j\leq N} |y(x_j) - y_N(x_j)|,$$

where the coefficients $\{w_j\}_{j=-N}^{N}$ in $y_N(x_j)$ are solved by the direct method $\mathbf{w} = \mathbf{A} \setminus \mathbf{p}$ with MATLAB. In these two tables, we use "$**$" to indicate that the iteration method does not converge within 1000 iterations, "$\mathbf{I}$" to represent the iteration method with no preconditioner, and "$\mathbf{P}$" to denote the iteration method with the banded preconditioner $\mathbf{P}$. In addition, "$\mathbf{P}_{\text{iter}}$" and "$\mathbf{I}_{\text{iter}}$" denote their numbers of iterations required for convergence.

From Tables 5.1 and 5.2, we see that if no preconditioner is used, GMRES converges very slowly and the number of iteration steps increase approximately like $2N$. BiCGSTAB converges more slowly than GMRES for Examples 5.1 and 5.2, and it even fails to solve the ODEs when $N > 16$. However, when the banded preconditioner $\mathbf{P}$ is used, the preconditioned GMRES and the preconditioned BiCGSTAB can successfully compute satisfactory approximations to the exact solutions of Examples 5.1 and 5.2, and both methods converge in less iteration steps. Hence, for these two examples the banded preconditioner $\mathbf{P}$ is very effective in accelerating the convergence rates of GMRES and BiCGSTAB. Moreover, for both examples we observe that the error function $E_s(h)$ reduces exponentially when $N$ is growing. We note also from Table 5.2 that the discrete accuracy of the sinc method is high and the convergence speeds of the preconditioned GMRES and BiCGSTAB methods are fast even though the assumptions in Theorem 3.1 are violated by Example 5.2.

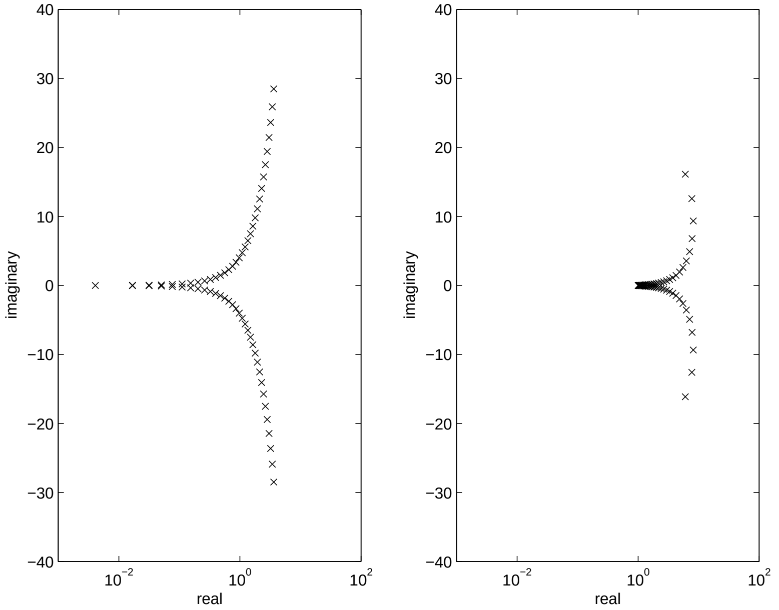

Figure 5.1: Spectra of $\mathbf{A}$ (left) and $\mathbf{P}^{-1}\mathbf{A}$ (right) for Example 5.1 with $N$=32.

Figures 5.1-5.4 depict the distributions of the eigenvalues of the original matrix $\mathbf{A}$ and the preconditioned matrix $\mathbf{P}^{-1}\mathbf{A}$ for Examples 5.1 and 5.2. These figures clearly show that the original matrices are very ill conditioned and, therefore, the corresponding GMRES and BiCGSTAB methods may converge very slowly or even diverge. However, the preconditioned matrices have tightly clustered eigenvalues and, thus, are well conditioned. As a result, the corresponding preconditioned GMRES and BiCGSTAB methods converge considerably fast to the exact solutions of the linear systems.

## 6 Concluding Remarks

By discretizing a class of third-order ODEs with the sinc-collocation and the sinc-Galerkin methods, we have obtained the systems of linear equations with the coefficient matrices being combinations of Toeplitz and diagonal matrices. By making use of the special structures of the coefficient matrices, we have constructed and analyzed a class of banded preconditioners, which can considerably improve the numerical properties of the Krylov subspace iteration methods such as GMRES and BiCGSTAB. Both theoretical analyses and numerical implementations have shown that the sinc discretization scheme is accurate, the discrete solution is exponentially convergent, and the banded preconditioner is effective.

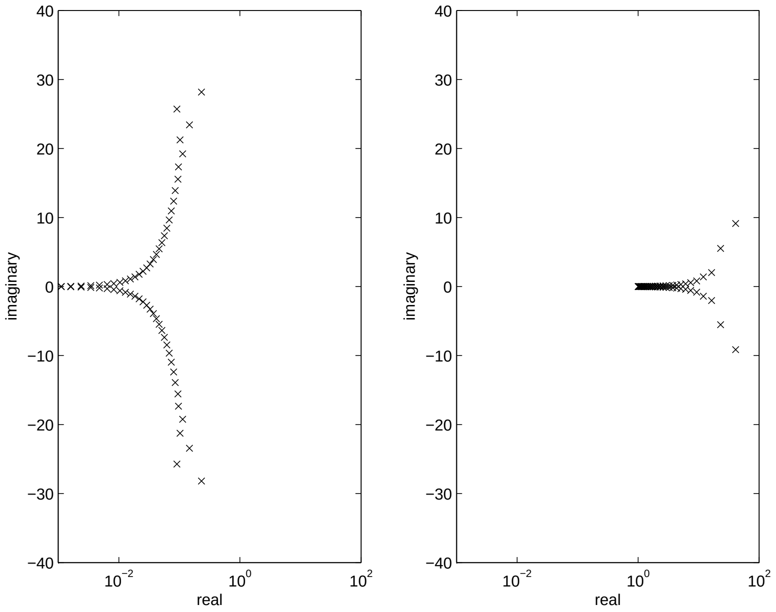

Figure 5.2: Spectra of $\mathbf{A}$ (left) and $\mathbf{P}^{-1}\mathbf{A}$ (right) for Example 5.2 with $N$=32.

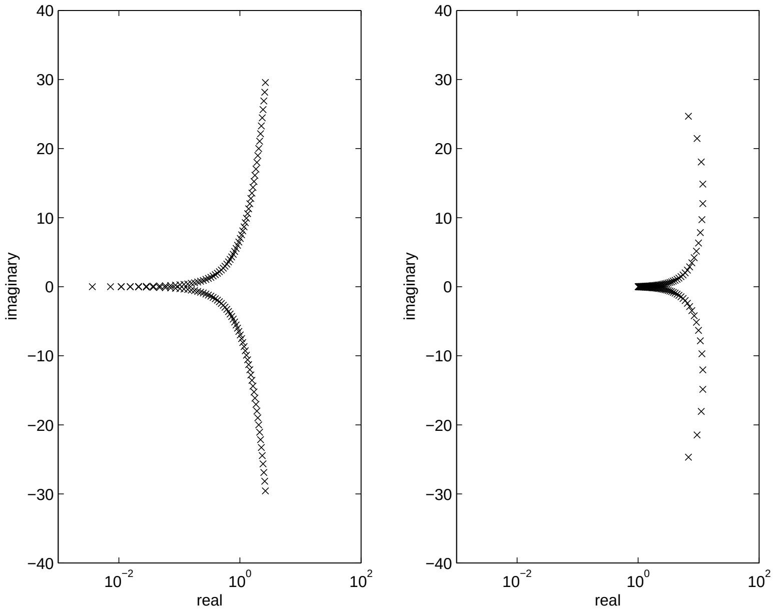

Figure 5.3: Spectra of $\mathbf{A}$ (left) and $\mathbf{P}^{-1}\mathbf{A}$ (right) for Example 5.1 with $N$=64.

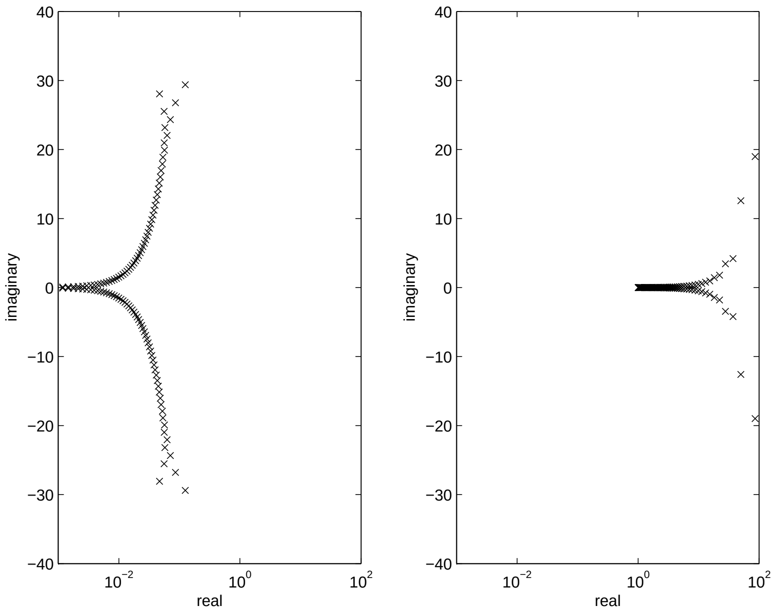

Figure 5.4: Spectra of $\mathbf{A}$ (left) and $\mathbf{P}^{-1}\mathbf{A}$ (right) for Example 5.2 with $N$=64.

A remarkable feature of this class of ODEs is that its highest order term is of order three, which makes the coefficient matrices of the resulting linear system strongly nonsymmetric and highly ill-conditioned. Hence, solving this class of linear systems should be a challengeable work. The Krylov subspace iteration methods, incorporated with the banded preconditioning matrices, provide a feasible approach for effectively tackling this class of linear systems.

The assumptions on the coefficients of the ODEs may limit the application scopes of our theories. They are, however, mainly due to algebraic difficulty in eigenvalue estimates, and may be removed by variable replacements in the ODEs. This will be a future research topic of both theoretical importance and practical value.

## Appendix: Proof of Lemma 3.1

**Part (i)**: To prove Lemma 3.1 (i), we need an error expression of the cardinal expansion of $\tilde{y}(x)$. For $m = 0, 1, 2, 3$ and $j \in \mathbb{Z}_N$, define $K_m(x,z)$ and $\omega_{m,j}(x)$ as

$$K_m(x,z) = \frac{1}{2\pi\imath[\phi'(x)]^{m-1}} \frac{\partial^m}{\partial x^m}\left(\frac{\sin[\pi\phi(x)/h]}{\phi'(x)[\phi(z)-\phi(x)]}\right),$$
$$\omega_{m,j}(x) = \frac{1}{[\phi'(x)]^{m-1}} \frac{\mathrm{d}^m}{\mathrm{d}x^m}\left(\frac{S(j,h)\circ\phi(x)}{\phi'(x)}\right).$$

Since $\tilde{y}(x) \in \mathbb{L}_\alpha(\mathcal{D})$, it follows that $\tilde{y}(x)\phi^{'}(x) \in \mathbb{H}^1(\mathcal{D})$. Hence, by making use of [20, Theorem 3.2], we know that the cardinal series expansion $\tilde{y}(x)$ has an error term

$$\tilde{y}(x) - \sum_{j=-\infty}^{\infty} \tilde{y}(x_j)\omega_{0,j}(x) = \int_{\partial\mathcal{D}} \frac{K_0(x,z)\tilde{y}(z)\phi^{'}(z)}{\sin[\pi\phi(z)/h]}\,\mathrm{d}z.$$

It then follows that,

$$\frac{\mathrm{d}^m y(x)}{\mathrm{d}x^m} - \sum_{j=-\infty}^{\infty} [\phi^{'}(x)]^{m-1} \omega_{m,j}(x)\tilde{y}(x_j) = \int_{\partial\mathcal{D}} \frac{[\phi^{'}(x)]^{m-1} K_m(x,z)}{\phi^{'}(x)\sin[\pi\phi(z)/h]} \tilde{y}(z)\phi^{'}(z)\,\mathrm{d}z.$$

We now estimate each component $v_k$ of the vector $\mathbf{A}_C\tilde{\mathbf{y}} - \mathbf{p}$. By replacing $w_j$ with $\tilde{y}(x_j)$ in (2.5) we obtain

$$\begin{aligned} v_k :&= [\mathbf{A}_C\tilde{\mathbf{y}} - \mathbf{p}]_k \\ &= h^3 \sum_{j=-N}^{N} \Big\{\omega_{3,j}(x_k) + \frac{\mu_2}{\phi^{'}}\omega_{2,j}(x_k) + \frac{\mu_1}{(\phi^{'})^2}\omega_{1,j}(x_k) + \frac{\mu_0}{(\phi^{'})^3}\omega_{0,j}(x_k)\Big\}\tilde{y}(x_j) \\ &\quad - h^3 \frac{\sigma}{(\phi^{'})^2}(x_k). \end{aligned} \tag{6.1}$$

Since $Ly(x) - \sigma(x) = 0$, from (6.1) we have

$$\begin{aligned} v_k &= [\mathbf{A}_C\tilde{\mathbf{y}} - \mathbf{p}]_k - h^3\left(\frac{Ly-\sigma}{(\phi^{'})^2}\right)_k \\ &= h^3 \sum_{j=-N}^{N} \Big\{\omega_{3,j}(x_k) + \frac{\mu_2}{\phi^{'}}\omega_{2,j}(x_k) + \frac{\mu_1}{(\phi^{'})^2}\omega_{1,j}(x_k)\Big\}\tilde{y}(x_j) \\ &\quad - \frac{h^3}{[\phi^{'}(x_k)]^2}[y^{'''}(x_k) + \mu_2(x_k)y^{''}(x_k) + \mu_1(x_k)y^{'}(x_k)] := v_k^{(1)} + v_k^{(2)}. \end{aligned}$$

Here we have split the summation into $\sum_{j=-N}^{N} = \sum_{j=-\infty}^{\infty} - \sum_{|j|>N}$, i.e.,

$$\begin{aligned} v_k^{(1)} &= h^3 \sum_{j=-\infty}^{\infty} \Big\{\omega_{3,j}(x_k) + \frac{\mu_2(x_k)}{[\phi^{'}(x_k)]}\omega_{2,j}(x_k) + \frac{\mu_1(x_k)}{[\phi^{'}(x_k)]^2}\omega_{1,j}(x_k)\Big\}\tilde{y}(x_j) \\ &\quad - \frac{h^3}{[\phi^{'}(x_k)]^2}[y^{'''}(x_k) + \mu_2(x_k)y^{''}(x_k) + \mu_1(x_k)y^{'}(x_k)] \\ &= -h^3 \int_{\partial\mathcal{D}} \left[K_3(x_k,z) + \frac{\mu_2(x_k)}{\phi^{'}(x_k)}K_2(x_k,z) + \frac{\mu_1(x_k)}{[\phi^{'}(x_k)]^2}K_1(x_k,z)\right] \frac{\phi^{'}(z)\tilde{y}(z)}{\sin[\pi\phi(z)/h]}\,\mathrm{d}z \end{aligned}$$

and

$$v_k^{(2)} = -h^3 \sum_{|j|>N} \Big\{\omega_{3,j}(x_k) + \frac{\mu_2(x_k)}{\phi^{'}(x_k)}\omega_{2,j}(x_k) + \frac{\mu_1(x_k)}{[\phi^{'}(x_k)]^2}\omega_{1,j}(x_k)\Big\}\tilde{y}(x_j).$$

In the expressions above, the explicit forms of $K_m(x,z)$, $m = 0,1,2,3,$ are as follows:

$$\begin{aligned} &K_0(x_k,z) = 0, \\ &K_1(x_k,z) = \frac{(-1)^k}{2\imath h[\phi(z)-kh]}, \\ &K_2(x_k,z) = \frac{(-1)^k}{2\imath h[\phi(z)-kh]^2}\left[2 + (\phi(z)-kh)\left(\frac{1}{\phi^{'}}\right)^{'}(x_k)\right], \\ &K_3(x_k,z) = \frac{(-1)^k}{2\imath h[\phi(z)-kh]^3}\left[6 + (\phi(z)-kh)^2\left(-\left(\frac{\pi}{h}\right)^2 + 2\frac{1}{\phi^{'}}\left(\frac{1}{\phi^{'}}\right)^{''} - \left(\left(\frac{1}{\phi^{'}}\right)^{'}\right)^2\right)(x_k)\right]. \end{aligned}$$

Since $|\mathrm{Im}(t)| = d$ and $|t - kh| \geq d$ hold on $\partial\mathcal{D}_d$, we have $|\mathrm{Im}(\phi(z))| = d$ and $|\phi(z) - kh| \geq d$ on $\partial\mathcal{D}$. Using these facts, as well as the assumptions on the coefficients of the ODE (1.1) and on the mapping $\phi$, we obtain

$$\begin{aligned} & h^3 \left| K_3(x_k, z) + \frac{\mu_2(x_k)}{[\phi'(x_k)]} K_2(x_k, z) + \frac{\mu_1(x_k)}{[\phi'(x_k)]^2} K_1(x_k, z) \right| \\ \leq \; & h^3 \left\{ |K_3(x_k, z)| + \left| \frac{\mu_2(x_k)}{[\phi'(x_k)]} \right| |K_2(x_k, z)| + \left| \frac{\mu_1(x_k)}{[\phi'(x_k)]^2} \right| |K_1(x_k, z)| \right\} \\ \leq \; & \frac{c_4}{[(\mathrm{Re}(\phi(z)) - kh)^2 + d^2]^{1/2}}, \end{aligned}$$

with $c_4$ a constant depending on the bounds for the coefficients of the ODE (1.1), on the bounds for derivatives of the inverse of the mapping $\phi$, and on $d$. Therefore, it holds that

$$\|\mathbf{A}_C \tilde{\mathbf{y}} - \mathbf{p}\|_2 = \left( \sum_{k=-N}^{N} |v_k|^2 \right)^{1/2} \leq \left( \sum_{k=-N}^{N} |v_k^{(1)}|^2 \right)^{1/2} + \left( \sum_{k=-N}^{N} |v_k^{(2)}|^2 \right)^{1/2}. \tag{6.2}$$

The first term in the right-hand side of (6.2) satisfies

$$\begin{aligned} \sum_{k=-N}^{N} |v_k^{(1)}|^2 &\leq \sum_{k=-\infty}^{\infty} \left| \int_{\partial\mathcal{D}} \frac{c_4}{[(\mathrm{Re}(\phi(z)) - kh)^2 + d^2]^{1/2}} \frac{|\phi'(z)\tilde{y}(z)|}{|\sin[\pi\phi(z)/h]|} |\mathrm{d}z| \right|^2 \\ &\leq \sum_{k=-\infty}^{\infty} \frac{c_4'}{k^2h^2 + d^2} \left( \int_{\partial\mathcal{D}} \frac{|\phi'(z)\tilde{y}(z)\,\mathrm{d}z|}{|\sin[\pi\phi(z)/h]|} \right)^2 \leq \frac{c_4'' h^{-2}}{[\sinh(\pi d/h)]^2}. \end{aligned} \tag{6.3}$$

We remark that the first inequality in (6.3) comes from the fact that there exists a $k_0 \in \mathbb{Z}$ such that $k_0 h \leq \mathrm{Re}(\phi(z)) - kh \leq (k_0 + 1)h$, and the last inequality in (6.3) comes from the bound $\sinh[\pi d/h] \leq \sin[\pi\phi(z)/h]$ on $\partial\mathcal{D}$ and from the existence of the integral of $\tilde{y}\phi'$.

For the second term in the right-hand side in (6.2), by making use of the assumptions on the mapping $\phi$, on the coefficients of the ODE (1.1), and the expression for $\{\delta_{j,k}^{(m)}\}_{j,k=-N}^{N}$ ($m = 1, 2, 3$), we have

$$\begin{aligned} \sum_{k=-N}^{N} |v_k^{(2)}|^2 &= \sum_{k=-N}^{N} \left| h^3 \sum_{|j|>N} \left\{ \omega_{3,j}(x_k) + \frac{\mu_2(x_k)}{\phi'(x_k)} \omega_{2,j}(x_k) + \frac{\mu_1(x_k)}{[\phi'(x_k)]^2} \omega_{1,j}(x_k) \right\} \tilde{y}(x_j) \right|^2 \\ &= \sum_{k=-N}^{N} \left| \sum_{|j|>N} \left\{ \delta_{jk}^{(3)} + h \frac{\mu_2(x_k)}{\phi'(x_k)} \delta_{jk}^{(2)} + h^2 \widetilde{\phi}(x_k)\, \delta_{jk}^{(1)} \right\} \tilde{y}(x_j) \right|^2 \\ &\leq c_5' \sum_{k=-N}^{N} \left| \sum_{|j|>N} \gamma_{j,k} e^{-\alpha|j|h} \right|^2 \\ &\leq c_5' \sum_{|j|>N} \sum_{|\ell|>N} \sum_{k=-\infty}^{\infty} \gamma_{j,k}\gamma_{\ell,k} e^{-\alpha|j|h} e^{-\alpha|\ell|h} \leq \frac{c_5''}{h^2} e^{-2\alpha N h}, \end{aligned} \tag{6.4}$$

where

$$\widetilde{\phi} = \frac{2}{\phi'}\left(\frac{1}{\phi'}\right)'' - \left(\left(\frac{1}{\phi'}\right)'\right)^2 + \frac{\mu_2}{\phi'}\left(\frac{1}{\phi'}\right)' + \frac{\mu_1}{(\phi')^2}$$

and $\gamma_{j,k}$ is the maximum of $\{|\delta_{j,k}^{(m)}|\}$ $(m = 1, 2, 3)$. We remark that the first inequality in (6.4) is derived by considering the fact that $|\tilde{y}(x_j)|$ is bounded by an exponentially decaying factor.

By combining the bounds for $v_k^{(1)}$ and $v_k^{(2)}$ in (6.3) and (6.4), and replacing $h$ by its optimal choice $[\frac{\pi d}{\alpha N}]^{1/2}$, we finally obtain

$$\|\mathbf{A}_C\tilde{\mathbf{y}} - \mathbf{p}\|_2 = \left(\sum_{k=-N}^{N} |v_k|^2\right)^{1/2} \leq c_1 N^{1/2} e^{-(\pi d\alpha N)^{1/2}}. \tag{6.5}$$

**Part (ii)**: We select an arbitrary integer in the range $[-N, N]$ and simply write $S$ for $S(k,h)\circ\phi$. Then it holds that

$$\begin{aligned}
0 &= h^2 \int_a^b \frac{(Ly-\sigma)(x)S(x)}{\phi'(x)}\,\mathrm{d}x \\
&= h^2 \int_a^b \left\{\left[-\left(\frac{S}{\phi'}\right)'''(x) + \left(\frac{\mu_2 S}{\phi'}\right)''(x) - \left(\frac{\mu_1 S}{\phi'}\right)'(x) + \left(\frac{\mu_0 S}{\phi'}\right)(x)\right] y(x) - \frac{\sigma S}{\phi'}(x)\right\}\mathrm{d}x \\
&= v_k^{(1)} + v_k^{(2)} + v_k^{(3)},
\end{aligned}$$

where $v_k^{(1)}$ denotes the $k$th component of the vector $\mathbf{A}_G\tilde{\mathbf{y}} - \mathbf{p}$,

$$\begin{aligned}
v_k^{(2)} = \sum_{|j|>N} \Bigg\{ &-\delta_{kj}^{(3)} + h\delta_{kj}^{(2)}\frac{\mu_2}{\phi'} \\
&- h^2\delta_{kj}^{(1)}\left[2\frac{1}{\phi'}\left(\frac{1}{\phi'}\right)'' - \left(\left(\frac{1}{\phi'}\right)'\right)^2 - 2\frac{\mu_2'}{(\phi')^2} - \frac{\mu_2}{\phi'}\left(\frac{1}{\phi'}\right)' + \frac{\mu_1}{(\phi')^2}\right]\Bigg\}\tilde{y}(x_j)
\end{aligned}$$

and $v_k^{(3)}$ represents the error of infinite-point quadrature. The quadrature may be explicitly expressed by means of Theorem 4.2.1 in [27] as follows:

$$v_k^{(3)} = \frac{\imath}{2}\int_{\partial\mathcal{D}} \frac{\kappa(z,h)\left\{\left[-\left(\frac{S}{\phi'}\right)'''(x) + \left(\frac{\mu_2 S}{\phi'}\right)''(x) - \left(\frac{\mu_1 S}{\phi'}\right)'(x) + \left(\frac{\mu_0 S}{\phi'}\right)(x)\right] y(x) - \frac{\sigma S}{\phi'}(x)\right\}}{\sin[\pi\phi(z)/h]}\,\mathrm{d}z, \tag{6.6}$$

with

$$\kappa(z,h) = \exp\{(\imath\pi\phi(z)/h)\cdot\mathrm{sgn}(\mathrm{Im}(\phi(z)))\}$$

such that $|\kappa(z,h)| = e^{-\pi d/h}$ holds for $z \in \partial\mathcal{D}$, where $\mathrm{sgn}(x)$ is a sign function defined as

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases}$$

Again, we set $u(z) = \mathrm{Re}(\phi(z))$. Recall that if $z \in \partial\mathcal{D}$, then $|\phi(z) - kh| \geq d$. Under the assumptions in (ii), with the explicit expression of the numerator in (6.6), we see that there exists a constant $c_6$, independent of $h$, such that

$$\left|v_k^{(3)}\right| \leq c_6 e^{-\pi d/h} \int_{\partial\mathcal{D}} \frac{|\mathrm{d}z|}{[(u(z) - kh)^2 + d^2]^{1/2}}.$$

Analogous to the derivation of (6.3), we obtain

$$\left(\sum_{k=-N}^{N} |v_k^{(3)}|^2\right)^{1/2} \leq \left(\sum_{k\in\mathbb{Z}} |v_k^{(3)}|^2\right)^{1/2} \leq c_6 h^{-1/2} e^{-\pi d/h}.$$

Also, similar to the derivation of (6.4), we obtain

$$\left(\sum_{k=-N}^{N} |v_k^{(2)}|^2\right)^{1/2} \leq c_7 h^{-1} e^{-\alpha N h}.$$

It then follows that

$$\|\mathbf{A}_G\tilde{\mathbf{y}} - \mathbf{p}\|_2 \leq c_1^{'} N^{1/2} e^{-(\pi d\alpha N)^{1/2}}. \tag{6.7}$$

## References

[1] G. Ammar and W. Gragg, *Superfast solution of real positive definite Toeplitz systems*, SIAM J. Matrix Anal. Appl., 9 (1988), 61–76.

[2] Z.-Z. Bai, G.H. Golub, L.-Z. Lu and J.-F. Yin, *Block triangular and skew-Hermitian splitting methods for positive-definite linear systems*, SIAM J. Sci. Comput., 26 (2005), 844–863.

[3] Z.-Z. Bai, G.H. Golub and M.K. Ng, *Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear systems*, SIAM J. Matrix Anal. Appl., 24 (2003), 603–626.

[4] Z.-Z. Bai, G.H. Golub and M.K. Ng, *On successive-overrelaxation acceleration of the Hermitian and skew-Hermitian splitting iterations*, Numer. Linear Algebra Appl., 14 (2007), 319–335.

[5] Z.-Z. Bai, Y.-M. Huang and M.K. Ng, *On preconditioned iterative methods for Burgers equations*, SIAM J. Sci. Comput., 29 (2007), 415–439.

[6] Z.-Z. Bai, Y.-M. Huang and M.K. Ng, *On preconditioned iterative methods for certain time-dependent partial differential equations*, SIAM J. Numer. Anal., 47 (2009), 1019–1037.

[7] Z.-Z. Bai and M.K. Ng, *Preconditioners for nonsymmetric block Toeplitz-like-plus-diagonal linear systems*, Numer. Math., 96 (2003), 197–220.

[8] F.D. Benedetto, *Solution of Toeplitz normal equations by sine transform based preconditioning* , Linear Algebra Appl., 285 (1998), 229–255.

[9] R.H. Chan and M.K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), 427–482.

[10] B.R. Duffy and S.K. Wilson, *A third-order differential equation arising in thin-film flows and relevant to tanner's law*, Appl. Math. Lett., 10 (1997), 63–68.

[11] W.F. Ford, *A third-order differential equation*, SIAM Rev., 34 (1992), 121–122.

[12] G. Friz, *Über den dynamischen Randwinkel im Fall der vollständigen Benetzung*, Z. Angew. Phys., 19 (1965), 374–378.

[13] G.H. Golub and C.F. Van Loan, *Matrix Computations*, 3rd Edition, The Johns Hopkins University Press, Baltimore and London, 1996.

[14] F.A. Howes, *The asymptotic solution of a class of third-order boundary-value problems arising in the theory of thin film flows*, SIAM J. Appl. Math., 43 (1983), 993–1004.

[15] X.-Q. Jin, *A note on preconditioned block Toeplitz matrices*, SIAM J. Sci. Comput., 16 (1995), 951–955.

[16] X.-Q. Jin, *Band Toeplitz preconditioners for block Toeplitz systems*, J. Comput. Appl. Math., 70 (1996), 225–230.

[17] X.-Q. Jin, *Developments and Applications of Block Toeplitz Iterative Solvers*, Kluwer Academic Publishers Group, Dordrecht; Science Press, Beijing, 2002.

[18] T. Kailath and A.H. Sayed, *Displacement structure: theory and applications*, SIAM Rev., 37 (1995), 297–386.

[19] N. Levinson, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J. Math. Phys., 25 (1946), 261–278.

[20] J. Lund and K. Bowers, *Sinc Methods for Quadrature and Differential Equations*, SIAM, Philadelphia, 1992.

[21] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York, 1979.

[22] A.C. Morlet, *Convergence of the sinc method for a fourth-order ordinary differential equation with an application*, SIAM J. Numer. Anal., 32 (1995), 1475–1503.

[23] M.K. Ng, *Fast iterative methods for symmetric sinc-Galerkin systems*, IMA J. Numer. Anal., 19 (1999), 357–373.

[24] M.K. Ng and D. Potts, *Fast iterative methods for sinc systems*, SIAM J. Matrix Anal. Appl., 24 (2002), 581–598.

[25] A. Nurmuhammad, M. Muhammad, M. Mori and M. Sugihara, *Double exponential transformation in the sinc-collocation method for a boundary value problem with fourth-order ordinary differential equation*, J. Comput. Appl. Math., 182 (2005), 32–50.

[26] K.J. Ruschak, *Coating flows*, Ann. Rev. Fluid Mech., 17 (1985), 65–89.

[27] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer Ser. Comput. Math., Springer-Verlag, New York, 1993.

[28] R.C. Smith, G.A. Bogar, K.L. Bowers and J. Lund, *The sinc-Galerkin method for fourth-order differential equations*, SIAM J. Numer. Anal., 28 (1991), 760–788.

[29] R.P. Spiers, C.V. Subbaraman and W.L. Wilkinson, *Free coating of a Newtonian liquid onto a vertical surface*, Chem. Engrg. Sci., 29 (1974), 389–396.

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

[31] E.O. Tuck and L.W. Schwartz, *A numerical and asymptotic study of some third-order ordinary differential equations relevant to draining and coating flows*, SIAM Rev., 32 (1990), 453–469.

[32] S.D.R. Wilson, *The drag-out problem in film coating theory*, J. Engrg. Math., 16 (1982), 209–221.