FAST ITERATIVE METHODS FOR SINC SYSTEMS

MICHAEL K. NG * AND DANIEL POTTS †

Abstract. We consider linear systems of equations arising from the Sinc method of boundary value problems which are typically nonsymmetric and dense. For the solutions of these systems we propose Krylov subspace methods with banded preconditioners. We prove that our preconditioners are invertible and discuss the convergence behavior of the conjugate gradient method for the normal equations (CGNE). In particular, we show that the solution of an n-by-n discrete Sinc system arising from the model problem can be obtained in $\mathcal{O}(n \log^2 n)$ operations by using the preconditioned CGNE method. Numerical results are given to illustrate the effectiveness of our fast iterative solvers.

Key words. Sinc method, Toeplitz matrices, Krylov subspace methods, preconditioners, banded matrices

AMS subject classifications. 65F10, 65F15, 65T10

1. Introduction.

In the Sinc–Galerkin method, the basis functions are derived from the Whittaker cardinal (sinc) function

$$\operatorname{sinc}(x) := \begin{cases} \dfrac{\sin(\pi x)}{\pi x}, & x \in \mathbb{R} \setminus 0 \\ 1, & x = 0. \end{cases}$$

and its translates

$$s(k,h)(x) := \operatorname{sinc}\left(\frac{x-kh}{h}\right), \quad (x \in \mathbb{R},\ k \in \mathbb{Z},\ h > 0).$$

The globally supported basis functions can be transformed via a composition with a suitable conformal map to any connected subset of the real line. This basis has been proved useful in the numerical analysis of a number of problems [17, 23, 24].

We seek an approximate solution of the linear two–point boundary value problem

$$\begin{gathered} \mathcal{L}u = u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \quad a < x < b, \\ u(a) = u(b) = 0. \end{gathered} \tag{1.1}$$

We approximate u by

$$u_{M+N+1}(x) = \sum_{k=-M}^{N} u_k s(k,h) \circ \phi(x), \tag{1.2}$$

where $\phi(z)$ is a conformal map of a simply connected domain $\mathcal{S}$ with boundary points $a \neq b$ onto

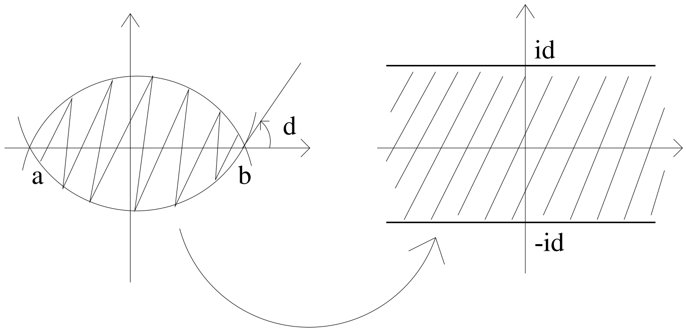
$$\mathcal{S}_d = \{z : z = x + \mathrm{i}y, \quad |y| < d, \quad d > 0\}. \tag{1.3}$$

---

* Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: mng@maths.hku.hk. Research supported in part by RGC Grant Nos. HKU 7147/99P and 7132/00P. This work described in this paper was also supported by a grant from the German Academic Exchange Services and the Research Grants Council of the Hong Kong Joint Research Scheme(Project No. G-HK020/00).

† Medical University of Lübeck, Institute of Mathematics, Wallstr. 40, D-23560 Lübeck. E-mail: potts@math.mu-luebeck.de. Research supported in part by the Hong Kong-German Joint Research Collaboration Grant from the Deutscher Akademischer Austauschdienst and the Hong Kong Research Grants Council.

FIG. 1.1. *The conformal map* $\phi(z) = \log\left(\frac{z-a}{b-z}\right)$



such that $\phi(a) = -\infty$ and $\phi(b) = \infty$. In Figure 1.1 (see for instance Lund and Bowers [17, p.118], and Stenger [24, pp.67–68]), we give an example of such conformal map. The simply connected domain $\mathcal{S}$ is the eye-shaped region

$$\left\{ z : \left| \arg \left(\frac{z-a}{b-z} \right) \right| < d \right\}$$

and the conformal map is given by

$$\phi(z) = \log \left(\frac{z-a}{b-z} \right).$$

Other conformal maps can also be found in [17, 23]. The general Galerkin method enables us to determine $\{u_k\}_{k=-M}^{N}$ by solving the linear system of equations

$$\langle \mathcal{L} u_{M+N+1} - f, s(k,h) \circ \phi \rangle = 0, \quad -M \le k \le N, \tag{1.4}$$

where the inner product is defined by

$$\langle f, g \rangle := \int_a^b f(x) g(x) w(x) \mathrm{d}x.$$

Here w plays the role of a weight function. For the case of second order problems, it is convenient to take $w(x) = \frac{1}{\phi'(x)}$, see [17, p.116]. The most distinctive feature of the Sinc basis is the resulting exponential convergence rate of the error. Moreover, the convergence rate maintains when the solution of the boundary value problem has boundary singularities.

The approximate explicit expressions for the inner products in (1.4) have been thoroughly treated in [17, 23]. The resulting discrete Sinc–Galerkin matrix coupling with collocation (see [24, pp. 465]) is given by the dense matrix

$$\boldsymbol{A} = \boldsymbol{T}_2 + \boldsymbol{D}_1 \boldsymbol{T}_1 + \boldsymbol{T}_1 \boldsymbol{D}_1 + \boldsymbol{D}_2, \quad (\boldsymbol{A} \in \mathbb{R}^{n \times n}), \tag{1.5}$$

where $\boldsymbol{T}_2$ is a symmetric Toeplitz matrix, $\boldsymbol{T}_1$ is a skew–symmetric Toeplitz matrix, and $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are diagonal matrices. Here $n = M + N + 1$. A straightforward application of the Gaussian elimination method will result in an algorithm, which takes $\mathcal{O}(n^3)$ arithmetical operations.

For n-by-n Toeplitz systems, fast and superfast direct solvers requiring $\mathcal{O}(n^2)$ and $\mathcal{O}(n\log^2 n)$ arithmetical operations respectively have been developed, see for instance Levinson [15] and Ammar and Gragg [1]. However, there exist no fast direct solvers for solving the system in (1.5). This is mainly because the displacement rank of the coefficient matrix can take any value between 0 and n. Hence fast Toeplitz solvers that are based on low displacement rank of matrices cannot be applied. The details of displacement ranks can be found in [14].

However, we note that given any n–vector $\boldsymbol{q}$, the matrix–vector product $\boldsymbol{A}\boldsymbol{q}$ can be computed in $\mathcal{O}(n\log n)$ operations. In fact, $\boldsymbol{T}_l\boldsymbol{q}$ ($l \in \{1,2\}$) can be obtained by using fast trigonometric transforms, see e.g., [11, 21]. Since $\boldsymbol{D}_l$ is a diagonal matrix, the product $\boldsymbol{D}_l\boldsymbol{q}$ ($l \in \{1,2\}$) can be computed in $\mathcal{O}(n)$ operations. Thus Krylov subspace methods, which are based on matrix–vector products, can be employed for solving Sinc systems. Since $\boldsymbol{A}$ is nonsymmetric, we suggest to solve the equations

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}, \tag{1.6}$$

by conjugate gradient type methods like GMRES [22, p.158], BiCGSTAB [22, p.217] or the conjugate gradient method for the normal equations (CGNE) [22, p.238].

One way to speed up the convergence rate of CGNE is to precondition the coefficient matrix. Instead of solving the original system $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$, we solve the preconditioned system

$$(\boldsymbol{M}^{-1}\boldsymbol{A})\boldsymbol{u} = \boldsymbol{M}^{-1}\boldsymbol{f}. \tag{1.7}$$

We note that the convergence rate of the CGNE method depends on the singular values of the preconditioned matrix [5, 28]. The matrix $\boldsymbol{M}$, called a preconditioner to the matrix $\boldsymbol{A}$, should be chosen with two criteria in mind: $\boldsymbol{M}\boldsymbol{r} = \boldsymbol{d}$ is easy to solve for any vector $\boldsymbol{d}$; the spectrum of $(\boldsymbol{M}^{-1}\boldsymbol{A})(\boldsymbol{M}^{-1}\boldsymbol{A})^{\mathrm{T}}$ is uniformly bounded and well–separated from the origin compared to that of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$.

In [19], we have considered the symmetric Sinc–Galerkin method [16] for discretization of the second–order self–adjoint boundary value problem. In this case, the Sinc–Galerkin matrix $\boldsymbol{A}$ is the sum of a symmetric Toeplitz matrix and a diagonal matrix. We have used banded matrices $\boldsymbol{R}$ with band–widths independent of the size of the matrix as preconditioners. We have shown that they give rise to the fast convergence of the preconditioned conjugate gradient (PCG) method. In particular, we proved that the spectra of $\boldsymbol{R}^{-1}\boldsymbol{A}$ are uniformly bounded from above and below by positive constants independent of the size of the matrix. The banded system $\boldsymbol{R}\boldsymbol{r} = \boldsymbol{d}$ can be solved in $\mathcal{O}(n)$ operations, where n is the size of the matrix. Therefore the cost of each PCG iteration is of $\mathcal{O}(n\log n)$ operations. It follows that the solution of $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}$ can be obtained in $\mathcal{O}(n\log n)$ operations. However, these preconditioners cannot be applied to nonsymmetric Sinc systems.

The main aim of this paper is to propose other banded preconditioners $\boldsymbol{B}$ for $\boldsymbol{A}$, given by (1.5). We show that the singular values of the preconditioned Sinc matrix arising from the model problem are uniformly bounded except for at most a finite number of outliers. Using this result, we show that the CGNE method applied to (1.7) converges at most in $\mathcal{O}(\log n)$ iteration steps. Hence the method requires $\mathcal{O}(n\log^2 n)$ operations.

The outline of this paper is as follows: In §2, we study some properties of the discrete Sinc system. In §3, we introduce our preconditioners. The convergence analysis of the CGNE method is given in §4. Numerical results are presented in §5 to illustrate the effectiveness of our method. Furthermore we compare the CGNE method with Krylov subspace methods like GMRES or BiCGSTAB which do not require the translation of (1.7) to the normal equations. Finally §6 contains some concluding remarks.

2. Properties of Discrete Systems. Let $\mathcal{S}$ be a simply connected domain in the complex plane with boundary points $a \neq b$. Let ϕ be a conformal mapping of $\mathcal{S}$ onto the strip $\mathcal{S}_d$ defined by (1.3) such that $\phi(a) = -\infty$ and $\phi(b) = \infty$. For $1 \leq k \leq \infty$, let $\mathcal{H}^k(\mathcal{S})$ denote the family of all functions f that are analytic in $\mathcal{S}$ and fulfill

$$\begin{cases} \left(\int_{\partial\mathcal{S}} |f(z)|^k \mathrm{d}z\right)^{1/k} < \infty, & 1 \leq k < \infty, \\ \sup_{z\in\mathcal{S}} |f(z)| < \infty, & k = \infty. \end{cases}$$

Corresponding to the number α, let $\mathcal{L}_\alpha(\mathcal{S})$ denote the family of all analytic functions on $\mathcal{S}$ for which there exists a constant C such that

$$|f(z)| \leq C \, \frac{|\mathrm{e}^{\phi(z)}|^\alpha}{(1+|\mathrm{e}^{\phi(z)}|)^{2\alpha}} \quad \text{for all} \quad z \in \mathcal{S} \, .$$

To study the convergence of the Sinc-Galerkin method for differential problems, assumptions on the functions ϕ, p and q are required.

Assumption (A1): (see [24, pp. 467, 469]) *Assume for the differential equation* (1.1) *that p/ϕ', $(p/\phi')/\phi'$, $q/(\phi')^2$, $(1/\phi')'$, and $(1/\phi')''/\phi'$ are real valued, belong to $\mathcal{H}^\infty(S)$ and that problem* (1.1) *has a unique solution $u \in \mathcal{L}_\alpha(\mathcal{S})$.*

Assumption (A2): (see [24, p. 478]) *Assume for the differential equation* (1.1) *that*

$$\mathrm{Re}\left(\frac{1}{\phi'(x)}\left(\frac{1}{\phi'(x)}\right)'' - \frac{1}{\phi'(x)}\left(\frac{p(x)}{\phi'(x)}\right)' + \frac{2q(x)}{(\phi'(x))^2}\right) \leq 0, \quad \text{for } a < x < b.$$

The following theorem about the approximate solution was given in [24].

Theorem 2.1. [24, Theorem 7.2.6] *Let Assumption* **(A1)** *and* **(A2)** *be satisfied. Let*

$$\begin{aligned} \boldsymbol{A}_n^{(g)} := \boldsymbol{T}_n[g_2] + h\boldsymbol{T}_n[g_1]\boldsymbol{D}_n\left[\frac{-\phi''}{(\phi')^2} - \frac{p}{\phi'}\right] + \\ h^2\boldsymbol{D}_n\left[\frac{1}{\phi'}\left(\frac{1}{\phi'}\right)'' - \frac{1}{\phi'}\left(\frac{p}{\phi'}\right)' + \frac{q}{(\phi')^2}\right], \end{aligned} \tag{2.1}$$

$$\boldsymbol{A}_n^{(c)} := \boldsymbol{T}_n[g_2] + h\boldsymbol{D}_n\left[\frac{-\phi''}{(\phi')^2} - \frac{p}{\phi'}\right]\boldsymbol{T}_n[g_1] + h^2\boldsymbol{D}_n\left[\frac{q}{(\phi')^2}\right] \tag{2.2}$$

and

$$\boldsymbol{A}_n := \frac{1}{2}\left(\boldsymbol{A}_n^{(g)} + \boldsymbol{A}_n^{(c)}\right). \tag{2.3}$$

Here $\boldsymbol{T}_n[g_\ell]$ ($\ell \in \{1,2\}$) denotes the n-by-n Toeplitz matrix with the (j,k)th entry given by the $(j-k)$th Fourier coefficient of the function,

$$g_\ell(\theta) = (\mathrm{i}\theta)^\ell, \quad \forall\theta \in [-\pi,\pi]\,, \tag{2.4}$$

$\boldsymbol{D}_n[\psi]$ is an n-by-n diagonal matrix given by

$$\boldsymbol{D}_n[\psi] = \text{ diag } [\psi(x_{-M}), \cdots, \psi(x_0), \cdots, \psi(x_N)],$$

with $x_k = \phi^{-1}(kh)$ for $k = 0, \pm 1, \pm 2, \ldots$. If the vector $\boldsymbol{u} = (u_{-M}, \cdots, u_N)^T$ denotes the exact solution of the system of equations

$$\boldsymbol{A}_n \boldsymbol{u} = h^2 \boldsymbol{D}_n \left[\frac{1}{(\phi')^2} \right] \boldsymbol{f}, \tag{2.5}$$

where $\boldsymbol{f} = [f(x_{-M}), \cdots, f(x_N)]^T$, then

$$|u(x) - u_n(x)| \leq C n^{1/2} \mathrm{e}^{-(\pi d \alpha n)^{1/2}}, \quad \text{for } a < x < b. \tag{2.6}$$

THEOREM 2.2. [24, Lemma 7.2.5] *Let Assumptions* **(A1)** *and* **(A2)** *be satisfied. Let $\boldsymbol{A}_n^{(g)}$, $\boldsymbol{A}_n^{(c)}$ and $\boldsymbol{A}_n$ be defined as in* (2.1), (2.2) *and* (2.3) *respectively. Then the following hold true:*
(i) *There exists a constant c_1 independent of n such that*

$$\|\boldsymbol{A}_n^{(g)}\|_2, \ \|\boldsymbol{A}_n^{(c)}\|_2, \ \|\boldsymbol{A}_n\|_2 \leq \pi^2 (1 + \frac{c_1}{\sqrt{n}}).$$

(ii) *There exists a constant c_2 independent of n such that*

$$\|(\boldsymbol{A}_n^{(g)})^{-1}\|_2, \ \|(\boldsymbol{A}_n^{(c)})^{-1}\|_2, \ \|\boldsymbol{A}_n^{-1}\|_2 \leq \frac{4n^2}{\pi^2}(1 + \frac{c_2}{n}).$$

In particular, the condition number $\kappa(\boldsymbol{A}_n \boldsymbol{A}_n^T)$ of $\boldsymbol{A}_n \boldsymbol{A}_n^T$ satisfies

$$\kappa(\boldsymbol{A}_n \boldsymbol{A}_n^T) \leq 4n^2 (1 + \frac{c_1}{\sqrt{n}})(1 + \frac{c_2}{n}).$$

Since $\kappa(\boldsymbol{A}_n \boldsymbol{A}_n^{\mathrm{T}}) = \mathcal{O}(n^2)$, the convergence of the CGNE method might be very slow with increasing n, see for instance Theorem 4.1 in Section 4. In the next section, we introduce the banded preconditioner to precondition the Sinc coefficient matrix in order to speed up the convergence rate of the CGNE method.

3. Banded Preconditioners. Recall that the coefficient matrix $\boldsymbol{A}_n$ in (2.3) is the sum of Toeplitz–times–diagonal matrices and diagonal matrices. There are many "good" preconditioners for the individual parts. For instance, the diagonal matrix system can be solved easily. For Toeplitz systems, circulant preconditioners have been proved to be successful choices, see the recent survey paper by Chan and Ng [3]. However, we remark that circulant preconditioners do not work for Toeplitz–plus–banded systems. Even T. Chan's circulant preconditioner [6] which is well–defined for non–Toeplitz matrices, will – while defined for $\boldsymbol{A}_n$ – not work well when $\boldsymbol{D}_n[\cdot]$ are not identity matrices, see numerical results in [4]. If we approximate $\boldsymbol{T}_n[g_\ell]$ in (2.3) by a circulant preconditioner $\boldsymbol{C}_n[g_\ell]$, then

$$\boldsymbol{C}_n[g_2] + \frac{h}{2}(\boldsymbol{D}_n^I \boldsymbol{C}_n[g_1] + \boldsymbol{C}_n[g_1]\boldsymbol{D}_n^I) + \frac{h^2}{2}\, \boldsymbol{D}_n^{II},$$

where

$$\boldsymbol{D}_n^I := \boldsymbol{D}_n \left[\frac{-\phi''}{(\phi')^2} - \frac{p}{\phi'} \right] \quad \text{and} \quad \boldsymbol{D}_n^{II} := \boldsymbol{D}_n \left[\frac{1}{\phi'} \left(\frac{1}{\phi'} \right)'' - \frac{1}{\phi'} \left(\frac{p}{\phi'} \right)' + \frac{2q}{(\phi')^2} \right]$$

can be expected to be "good" approximation to $\boldsymbol{A}_n$. Unfortunately, the resulting circulant–type matrix system *cannot* be solved easily in general. Hence, this approach of constructing preconditioner for $\boldsymbol{A}_n$ cannot work in most situations. In this paper, we consider a preconditioner which is easily invertible.

In [19], we have proposed to use banded matrices as preconditioners for symmetric Sinc–Galerkin systems. Following this approach, we introduce our preconditioners $\boldsymbol{B}_n$ by

$$\boldsymbol{B}_n := \boldsymbol{P}_n^{II} + \frac{h}{2}(\boldsymbol{D}_n^I \boldsymbol{P}_n^I + \boldsymbol{P}_n^I \boldsymbol{D}_n^I) + \frac{h^2}{2}\boldsymbol{D}_n^{II}, \tag{3.1}$$

where $\boldsymbol{P}_n^{II}$ and $\boldsymbol{P}_n^I$ are the banded Toeplitz matrices:

$$\boldsymbol{P}_n^{II} := \boldsymbol{T}_n(p_2) \;=\; \text{tridiag}\,[1,-2,1] \quad \text{and} \quad \boldsymbol{P}_n^{I} := \boldsymbol{T}_n(p_1) \;=\; \text{tridiag}\,[-\frac{1}{2},0,\frac{1}{2}]$$

with generating functions of $\boldsymbol{P}_n^I$ and $\boldsymbol{P}_n^{II}$ given by

$$p_1(\theta) := \mathrm{i}\sin\theta \quad \text{and} \quad p_2(\theta) := -2 + 2\cos\theta, \quad \forall\theta\in[-\pi,\pi], \tag{3.2}$$

respectively.

We note that the preconditioner $\boldsymbol{B}_n$ is just an n-by-n tridiagonal matrix. It follows that the system $\boldsymbol{B}_n\boldsymbol{r} = \boldsymbol{d}$ can be solved by using any efficient tridiagonal solver in $\mathcal{O}(n)$ operations.

The symmetric and skew–symmetric parts of $\boldsymbol{B}_n$ are given by

$$\boldsymbol{B}_n^{(h)} := \boldsymbol{P}_n^{II} + \frac{h^2}{2}\boldsymbol{D}_n^{II} \quad \text{and} \quad \boldsymbol{B}_n^{(s)} := \frac{h}{2}(\boldsymbol{D}_n^I \boldsymbol{P}_n^I + \boldsymbol{P}_n^I \boldsymbol{D}_n^I)\,,$$

respectively. Moreover, we have by the theorem of Bendixson [25, p. 418] that

$$\lambda_{\min}(\boldsymbol{B}_n^{(h)}) \leq \ \mathrm{Re}[\lambda(\boldsymbol{B}_n)] \leq \lambda_{\max}(\boldsymbol{B}_n^{(h)})$$

and

$$\lambda_{\min}(\frac{1}{\mathrm{i}}\boldsymbol{B}_n^{(s)}) \leq \ \mathrm{Im}[\lambda(\boldsymbol{B}_n)] \leq \lambda_{\max}(\frac{1}{\mathrm{i}}\boldsymbol{B}_n^{(s)}),$$

where $\lambda(\boldsymbol{B})$ denotes the eigenvalues of the matrix $\boldsymbol{B}$.

LEMMA 3.1. *Let Assumption* **(A2)** *be satisfied. Further let*

$$d_2 := \min_{x\in\phi^{-1}(\mathbb{R})}\left\{\frac{1}{\phi'(x)}\left(\frac{1}{\phi'(x)}\right)'' - \frac{1}{\phi'(x)}\left(\frac{p(x)}{\phi'(x)}\right)' + \frac{2q(x)}{(\phi'(x))^2}\right\}$$

and

$$d_3 := \max_{x\in\phi^{-1}(\mathbb{R})}\left\{\frac{1}{\phi'(x)}\left(\frac{1}{\phi'(x)}\right)'' - \frac{1}{\phi'(x)}\left(\frac{p(x)}{\phi'(x)}\right)' + \frac{2q(x)}{(\phi'(x))^2}\right\}.$$

Then we have

$$\boldsymbol{P}_n^{II} + \frac{d_2 h^2}{2}\boldsymbol{I}_n \leq \boldsymbol{B}_n^{(h)} \leq \boldsymbol{P}_n^{II} + \frac{d_3 h^2}{2}\boldsymbol{I}_n.$$

In particular, the preconditioners $\boldsymbol{B}_n$ are nonsingular for all n.
Proof: The assertion follows from **(A2)** and the fact that the matrices $\boldsymbol{B}_n^{(h)}$ are negative definite. ■

Remark: In [24, p. 481], Stenger has shown that the approximate solution for $u_{M+N+1}(x)$ in (1.2) can also be obtained by solving the linear systems involving the coefficient matrices $\boldsymbol{A}_n^{(g)}$ and $\boldsymbol{A}_n^{(c)}$ given in (2.1) and (2.2), respectively. We note that we can also develop similar banded preconditioners

$$\boldsymbol{B}_n^{(g)} := \boldsymbol{P}_n^{II} + h\,\boldsymbol{P}_n^{I}\boldsymbol{D}_n^{I} + h^2\boldsymbol{D}_n\left[\frac{1}{\phi'}\left(\frac{1}{\phi'}\right)'' - \frac{1}{\phi'}\left(\frac{p}{\phi'}\right)' + \frac{q}{(\phi')^2}\right]$$

and

$$\boldsymbol{B}_n^{(c)} := \boldsymbol{P}_n^{II} + h\,\boldsymbol{D}_n^{I}\boldsymbol{P}_n^{I} + h^2\,\boldsymbol{D}_n\left[\frac{q}{(\phi')^2}\right]$$

for the matrices $\boldsymbol{A}_n^{(g)}$ and $\boldsymbol{A}_n^{(c)}$, respectively. Numerical tests show that these preconditioners work similarly well as the preconditioner $\boldsymbol{B}_n$ for $\boldsymbol{A}_n$. However, we remark that the convergence analysis for these preconditioned systems $(\boldsymbol{B}_n^{(g)})^{-1}\boldsymbol{A}_n^{(g)}$ and $(\boldsymbol{B}_n^{(c)})^{-1}\boldsymbol{A}_n^{(c)}$ is still an open problem. □

3.1. The Model Problem. In this subsection, we consider some model Sinc–Galerkin matrices and analyze the spectra of these preconditioned matrices. By using the Bendixson theorem again, we obtain that symmetric and skew–symmetric parts of $\boldsymbol{A}_n$ are given by

$$\boldsymbol{A}_n^{(h)} := \boldsymbol{T}_n[g_2] + \frac{h^2}{2}\boldsymbol{D}_n^{II} \quad \text{and} \quad \boldsymbol{A}_n^{(s)} := \frac{h}{2}\left(\boldsymbol{D}_n^{I}\boldsymbol{T}_n[g_1] + \boldsymbol{T}_n[g_1]\boldsymbol{D}_n^{I}\right),$$

respectively, and that

$$\lambda_{\min}(\boldsymbol{A}_n^{(h)}) \leq \operatorname{Re}[\lambda(\boldsymbol{A}_n)] \leq \lambda_{\max}(\boldsymbol{A}_n^{(h)})$$

and

$$\lambda_{\min}\left(\frac{1}{\mathrm{i}}\boldsymbol{A}_n^{(s)}\right) \leq \operatorname{Im}[\lambda(\boldsymbol{A}_n)] \leq \lambda_{\max}\left(\frac{1}{\mathrm{i}}\boldsymbol{A}_n^{(s)}\right).$$

Let

$$d_1 := \max_{x\in\phi^{-1}(\mathbb{R})}\left\{\left|\frac{-\phi''(x)}{(\phi'(x))^2} - \frac{p(x)}{\phi'(x)}\right|\right\}. \tag{3.3}$$

Then we have

$$-\lambda_{\max}\left(\frac{d_1 h}{\mathrm{i}}\boldsymbol{T}_n[g_1]\right) \leq \lambda_{\min}\left(\frac{1}{\mathrm{i}}\boldsymbol{A}_n^{(s)}\right) \leq \lambda_{\max}\left(\frac{1}{\mathrm{i}}\boldsymbol{A}_n^{(s)}\right) \leq \lambda_{\max}\left(\frac{d_1 h}{\mathrm{i}}\boldsymbol{T}_n[g_1]\right).$$

For the symmetric part of $\boldsymbol{A}_n$, we find

$$\boldsymbol{T}_n[g_2] + \frac{d_2 h^2}{2}\boldsymbol{I}_n \leq \boldsymbol{A}_n^{(h)} \leq \boldsymbol{T}_n[g_2] + \frac{d_3 h^2}{2}\boldsymbol{I}_n,$$

where d_2 and d_3 are defined as in Lemma 3.1. In particular, we have

$$\lambda_{\min}\left(\boldsymbol{T}_n[g_2]+\frac{d_2h^2}{2}\boldsymbol{I}_n\right)\le\lambda_{\min}\left(\boldsymbol{A}_n^{(h)}\right)$$
$$\le\lambda_{\max}\left(\boldsymbol{A}_n^{(h)}\right)\ \le\ \lambda_{\max}\left(\boldsymbol{T}_n[g_2]+\frac{d_2h^2}{2}\boldsymbol{I}_n\right).$$

The spectrum of the matrix $\boldsymbol{A}_n$ is contained in the box

$$\left[\lambda_{\min}\left(\boldsymbol{T}_n[g_2]+\frac{d_2h^2}{2}\boldsymbol{I}_n\right),\lambda_{\max}\left(\boldsymbol{T}_n[g_2]+\frac{d_3h^2}{2}\boldsymbol{I}_n\right)\right]\times$$
$$\left[-\lambda_{\max}\left(\frac{d_1h}{\mathrm{i}}\boldsymbol{T}_n[g_1]\right),\lambda_{\max}\left(\frac{d_1h}{\mathrm{i}}\boldsymbol{T}_n[g_1]\right)\right]$$

in the complex plane. This suggests us to analyze the banded preconditioners for the following model Sinc–Galerkin matrices

$$\boldsymbol{T}_n[g_2]+h\gamma_1\boldsymbol{T}_n[g_1]+h^2\,\gamma_2\boldsymbol{I}_n \text{ with } \gamma_1\in\{\pm d_1\} \text{ and } \gamma_2\in\{d_2/2,d_3/2\}\,. \tag{3.4}$$

If the corresponding banded matrices are good preconditioners of these model Sinc–Galerkin matrices, then we expect that $\boldsymbol{B}_n$ will be a good preconditioner for $\boldsymbol{A}_n$. Numerical results in §5 will show that our banded preconditioners give rise to fast convergence of the iterative method.

3.2. Spectra of the Preconditioned Matrices for the Model Problem. We note that the model problem matrices in (3.4) are Toeplitz matrices. Therefore, we analyze the spectra of their corresponding preconditioned matrices by using their generating functions. We first establish the following lemma.

LEMMA 3.2. *Let $c_1\in\mathbb{R}$, c_2 be a negative number and h be a positive number. Let $g_1(\theta),g_2(\theta),p_1(\theta)$ and $p_2(\theta)$ be defined as in (2.4) and (3.2). If*

$$r(\theta)=\frac{hc_1g_1(\theta)+g_2(\theta)+h^2c_2}{hc_1p_1(\theta)+p_2(\theta)+h^2c_2},\quad\forall\theta\in[-\pi,\pi]\,,$$

then

$$1\le\ \mathrm{Re}(r(\theta))<\frac{3\pi^3}{8}+\frac{\pi^2}{16}h^2c_1^2,\quad\forall\theta\in[-\pi,\pi]\,, \tag{3.5}$$

and

$$-\frac{h|c_1|\pi}{4-h^2c_2}\le\ \mathrm{Im}(r(\theta))\le\frac{h|c_1|\pi}{4-h^2c_2},\quad\forall\theta\in[-\pi,\pi],$$

where

$$r(\theta)\ =\ \mathrm{Re}(r(\theta))+i\,\mathrm{Im}(r(\theta))\,. \tag{3.6}$$

Proof: We have

$$\mathrm{Re}(r(\theta))=\frac{(-\theta^2+h^2c_2)(-2+2\cos\theta+h^2c_2)+h^2c_1^2\theta\sin\theta}{|hc_1p_1(\theta)+p_2(\theta)+h^2c_2|^2}$$

and

$$\mathrm{Im}(r(\theta))=\frac{hc_1\theta(-2+2\cos\theta+h^2c_2)-hc_1\sin\theta(-\theta^2+h^2c_2)}{|hc_1p_1(\theta)+p_2(\theta)+h^2c_2|^2}\,.$$

Let us start with the real part. First we see that $\mathrm{Re}(r) - 1$ is nonnegative because

$$\begin{aligned}&\frac{(-\theta^2 + h^2c_2)(-2+2\cos\theta + h^2c_2) + (hc_1)^2\theta\sin\theta}{|hc_1p_1(\theta)+p_2(\theta)+h^2c_2|^2} - 1\\ =\ & \frac{h^2c_1^2(\theta-\sin\theta)\sin\theta + (-2+2\cos\theta+h^2c_2)(-\theta^2+2-2\cos\theta)}{|hc_1p_1(\theta)+p_2(\theta)+h^2c_2|^2}\end{aligned} \tag{3.7}$$

and both functions $(\theta - \sin\theta)\sin\theta$ and $(-2+2\cos\theta+h^2c_2)(-\theta^2+2-2\cos\theta)$ are nonnegative on $[-\pi,\pi]$.
Since

$$\mathrm{Re}(r(\theta)) = \frac{(-\theta^2+h^2c_2)(-2+2\cos\theta+h^2c_2) + h^2c_1^2\theta\sin\theta}{(2-2\cos\theta-h^2c_2)^2 + h^2c_1^2\sin^2\theta} \tag{3.8}$$

we get with

$$\frac{2}{\pi}\theta^2 \le 2-2\cos\theta \le \theta^2, \quad \left(0\le\theta\le\frac{\pi}{2}\right) \quad \text{and} \quad \frac{2}{\pi}\theta \le \sin\theta \le \theta, \quad \left(0\le\theta\le\frac{\pi}{2}\right)$$

that

$$\begin{aligned}\mathrm{Re}(r(\theta)) &\le \frac{(\theta^2-h^2c_2)(\theta^2-h^2c_2) + h^2\gamma_1^2\theta^2}{(\frac{2}{\pi}\theta^2-h^2c_2)^2 + h^2\gamma_1^2\left(\frac{2}{\pi}\theta\right)^2}\\ &\le \max\left\{\left(\frac{\theta^2-h^2c_2}{\frac{2}{\pi}\theta^2-h^2c_2}\right)^2, \quad \frac{h^2\gamma_1^2\theta^2}{h^2\gamma_1^2\frac{4}{\pi^2}\theta^2}\right\}\\ &\le \max\left\{\left(\max\left\{\frac{\pi}{2},1\right\}\right)^2, \quad \frac{\pi^2}{4}\right\} \ = \ \frac{\pi^2}{4}, \qquad \left(0\le\theta\le\frac{\pi}{2}\right).\end{aligned}$$

On the other hand, we have for $\frac{\pi}{2}\le\theta\le\pi$ that

$$\frac{4}{\pi}\theta \le 2-2\cos\theta \le \frac{3}{2}\theta \quad \text{and} \quad \sin\theta \le \theta - \frac{1}{\pi^2}\theta^3$$

and further by (3.8)

$$\begin{aligned}\mathrm{Re}(r(\theta)) &\le \frac{(\theta^2-h^2c_2)\left(\frac{3}{2}\theta-h^2c_2\right) + h^2\gamma_1^2\left(\theta^2-\frac{1}{\pi^2}\theta^4\right)}{\left(\frac{4}{\pi}\theta-h^2c_2\right)^2}\\ &\le \frac{(\theta^2-h^2c_2)\left(\frac{3}{2}\theta-h^2c_2\right)}{\left(\frac{4}{\pi}\theta-h^2c_2\right)^2} + h^2c_1^2\frac{\theta^2-\frac{1}{\pi^2}\theta^4}{\left(\frac{4}{\pi}\theta\right)^2}\\ &\le \frac{(\pi^2-h^2c_2)\left(\frac{3}{2}\pi-h^2c_2\right)}{(2-h^2c_2)^2} + \frac{\pi^2}{16}h^2\gamma_1^2\\ &\le \frac{\pi^2}{2}\cdot\frac{3}{4}\pi + \frac{\pi^2}{16}h^2c_1^2 \ = \ \frac{3\pi^3}{8} + \frac{\pi^2}{16}h^2c_1^2, \qquad \left(\frac{\pi}{2}\le\theta\le\pi\right).\end{aligned}$$

Since $\mathrm{Re}(r)$ is even, (3.5) follows. Furthermore, we have

$$\mathrm{Im}(r(\theta)) = \frac{hc_1(-2\theta+2\theta\cos\theta+\theta h^2c_2+\theta^2\sin\theta-h^2c_2\sin\theta)}{4-8\cos\theta-4h^2c_2+4\cos^2\theta+4h^2c_2\cos\theta+h^4c_2^2+h^2c_1^2-h^2c_1^2\cos^2\theta}. \tag{3.9}$$

By using Taylor series of $\cos\theta$ and $\sin\theta$ we get for $c_1 > 0$ that the numerator of the right hand side of (3.9) is less than $\frac{h^3 c_1 c_2}{6}\theta$ but the denominator is bigger than $h^4 c_2^2 + h^2(c_1^2 - 2c_2)\theta^2$. Hence the maximum and minimum values of $\mathrm{Im}(r(\theta))$ are attained at $\theta = \pi$ and $\theta = -\pi$. The result follows by noting that

$$\mathrm{Im}(r(-\pi)) = -\mathrm{Im}(r(\pi)) = \frac{hc_1\pi}{4 - h^2 c_2}\,.$$

■

The next lemma follows immediately from the close relationship between the spectrum of a Toeplitz matrix and its generating function [9].

LEMMA 3.3. *Let γ_1 and γ_2 be defined as in* (3.4). *Then we have*

$$1 \leq \ \lambda(\boldsymbol{T}_n[\mathrm{Re}(r)]) < \frac{3\pi^3}{8} + \frac{\pi^2}{16}h^2\gamma_1^2,$$

and

$$-\frac{h|\gamma_1|\pi}{4 - h^2\gamma_2} \leq \ \lambda(\boldsymbol{T}_n[\mathrm{Im}(r)]) \leq \frac{h|\gamma_1|\pi}{4 - h^2\gamma_2}, \quad \forall\theta \in [-\pi, \pi].$$

Next we prove the following lemma.

LEMMA 3.4. *Let Assumptions* **(A1)** *and* **(A2)** *be satisfied. Then, for all n,*

$$\boldsymbol{T}_n[g_2] + h\gamma_1\boldsymbol{T}_n[g_1] + h^2\gamma_2\boldsymbol{I}_n = (\boldsymbol{P}_n^{II} + h\gamma_1\boldsymbol{P}_n^{I} + h^2\gamma_2\boldsymbol{I}_n)\boldsymbol{T}_n[r] + \boldsymbol{L}_n, \tag{3.10}$$

where $\boldsymbol{L}_n$ has only nonzero entries in the first and last columns.

Proof: The result can be derived by noting that $\boldsymbol{P}_n^{II} + h\gamma_1\boldsymbol{P}_n^{I} + h^2\gamma_2\boldsymbol{I}_n$ is a tridiagonal Toeplitz matrix. ■

With Lemma 3.4, we have that the spectra of the preconditioned matrices are also essentially bounded.

THEOREM 3.5. *Let Assumptions* **(A1)** *and* **(A2)** *be satisfied. Then at most 8 eigenvalues of*

$$(\boldsymbol{P}_n^{II} + h\gamma_1\boldsymbol{P}_n^{I} + h^2\gamma_2\boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1\boldsymbol{T}_n[g_1] + h^2\gamma_2\boldsymbol{I}_n) \tag{3.11}$$

are outside the box

$$\left[1, \frac{3\pi^3}{8} + \frac{\pi^2}{16}h^2\gamma_1^2\right] \times \left[-\frac{h|\gamma_1|\pi}{4 - h^2\gamma_2}, \frac{h|\gamma_1|\pi}{4 - h^2\gamma_2}\right]$$

in the complex plane.

Proof: Since the matrix $\boldsymbol{P}_n^{II} + h\gamma_1\boldsymbol{P}_n^{I} + h^2\gamma_2\boldsymbol{I}_n$ is nonsingular, we obtain from (3.10) that

$$(\boldsymbol{P}_n^{II} + h\gamma_1\boldsymbol{P}_n^{I} + h^2\gamma_2\boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1\boldsymbol{T}_n[g_1] + h^2\gamma_2\boldsymbol{I}_n) = \boldsymbol{T}_n[r] + \tilde{\boldsymbol{L}}_n, \tag{3.12}$$

where $\tilde{\boldsymbol{L}}_n = (\boldsymbol{P}_n^{II} + h\gamma_1\boldsymbol{P}_n^{I} + h^2\gamma_2\boldsymbol{I}_n)^{-1}\boldsymbol{L}_n$ and the rank of $\tilde{\boldsymbol{L}}_n$ is at most 2. Let λ be an eigenvalue of the preconditioned matrix in (3.11). Then we get by Bendixson's theorem that

$$\lambda_{\min}\left(\boldsymbol{T}_n[\mathrm{Re}(r)] + \frac{\tilde{\boldsymbol{L}}_n + \tilde{\boldsymbol{L}}_n^{\mathrm{T}}}{2}\right) \leq \ \mathrm{Re}(\lambda) \leq \lambda_{\max}\left(\boldsymbol{T}_n[\mathrm{Re}(r)] + \frac{\tilde{\boldsymbol{L}}_n + \tilde{\boldsymbol{L}}_n^{\mathrm{T}}}{2}\right),$$

where $\mathrm{Re}(r)$ and $\mathrm{Im}(r)$ are defined as in (3.6). Since

$$\mathrm{rank}\left(\frac{\tilde{\boldsymbol{L}}_n + \tilde{\boldsymbol{L}}_n^{\mathrm{T}}}{2}\right) = 4,$$

by using Weyl's theorem [12, Theorem 4.3.1], at most 4 eigenvalues of $\boldsymbol{T}_n[\mathrm{Re}(r)] + (\tilde{\boldsymbol{L}}_n + \tilde{\boldsymbol{L}}_n^*)/2$ are not contained in the interval $[\min \mathrm{Re}(r(\theta)), \max \mathrm{Re}(r(\theta))]$. Similarly, we prove that at most 4 eigenvalues of $\boldsymbol{T}_n[\mathrm{Im}(r)] + (\tilde{\boldsymbol{L}}_n - \tilde{\boldsymbol{L}}_n^{\mathrm{T}})/2$ are not contained in the interval i $[\min \mathrm{Im}(r(\theta)), \max \mathrm{Im}(r(\theta))]$. Now the assertion follows from Lemma 3.3. ∎

We remark that it is well-known that the knowledge of the eigenvalues alone is not sufficient to estimate the convergence rate of GMRES, see for instance [8, 18]. As a matter of fact, it still remains an open problem to describe the convergence of GMRES in terms of some simple characteristic properties of the coefficient matrix. Even though we show in Theorem 3.5 that the eigenvalues of the preconditioned matrices are contained in a bounded region except for a finite number of outliers, we cannot provide a tight convergence bound of GMRES. However, we expect that GMRES may converge very fast when we apply GMRES to solve these preconditioned systems. Our numerical results in §5 will show that GMRES indeed converges very fast.

Next we consider the singular values distribution of the preconditioned matrix. This will be useful to estimate the number of iterations required for convergence of the CGNE method.

With Lemma 3.3 and Lemma 3.4, we have our main theorem which states that the spectra of the preconditioned normal equations matrices are essentially bounded.

THEOREM 3.6. *Let Assumptions* **(A1)** *and* **(A2)** *be satisfied. Then there exist $\beta \geq 1$ independent of n, such that at most 6 singular values of*

$$(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)$$

are outside the interval $[1, \beta]$.

Proof: By Lemma 3.4, we obtain

$$\begin{aligned}
&[(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)] \cdot \\
&[(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)]^{\mathrm{T}} \\
= {} & \boldsymbol{T}_n[r]\boldsymbol{T}_n[r]^{\mathrm{T}} + \hat{\boldsymbol{L}}_n,
\end{aligned}$$

where $\hat{\boldsymbol{L}}_n$ is Hermitian and $\mathrm{rank}(\hat{\boldsymbol{L}}_n) = 6$. By using the Courant–Fischer theorem about the inequalities between individual singular values of $\boldsymbol{T}_n[r]$ and eigenvalues of its Hermitian part [13, p.151], we have

$$\sigma_{\min}(\boldsymbol{T}_n[r]) \geq \lambda_{\min}(\boldsymbol{T}_n[\mathrm{Re}(r)]) \geq 1\,.$$

Here $\sigma(\cdot)$ denotes the singular values of a matrix. By using Lemma 3.3, we get

$$\begin{aligned}
\sigma_{\max}(\boldsymbol{T}_n[r]) \;\leq\; \|\boldsymbol{T}_n[r]\|_2 \;\leq\; 2\,\|r\|_\infty &\leq 2\sqrt{\left(\frac{3\pi^3}{8} + \frac{\pi^2}{16}h^2\gamma_1^2\right)^2 + \left(\frac{h|\gamma_1|\pi}{4 - h^2\gamma_2}\right)^2} \\
&\leq 2\sqrt{\left(\frac{3\pi^3}{8} + \frac{\pi^2}{16}\gamma_1^2\right)^2 + \left(\frac{|\gamma_1|\pi}{4}\right)^2} := \beta.
\end{aligned} \tag{3.13}$$

Hence the result follows. ■

4. Convergence Analysis of CGNE. An important practical aspect of solving boundary value problem (1.1) is the efficient solution of the resulting linear system

$$\boldsymbol{C}_n \boldsymbol{x} = \boldsymbol{B}_n^{-1} \boldsymbol{b} = \tilde{\boldsymbol{b}}, \tag{4.1}$$

with $\boldsymbol{C}_n = \boldsymbol{B}_n^{-1} \boldsymbol{A}_n$.

CGNE for solving the linear system (4.1) amounts to applying CG to the system $\boldsymbol{C}_n \boldsymbol{C}_n^{\mathrm{T}} \boldsymbol{y} = \tilde{\boldsymbol{b}}$ under the change of variables $\boldsymbol{x} = \boldsymbol{C}_n^{\mathrm{T}} \boldsymbol{y}$, see [8, p.105]. We note that the convergence rate of the CGNE method depends on the singular values of the preconditioned matrix. Since the singular values of the preconditioned Sinc matrix arising from the model problem are uniformly bounded except for at most a finite number of outliers (cf. Theorem 3.6), we will show that the convergence rate of the preconditioned conjugate gradient method for the normal equations will converge in at most $\mathcal{O}(\log n)$ steps. We begin by noting the following error estimate of the conjugate gradient method for the normal equations; see [28].

THEOREM 4.1. *Let $\boldsymbol{x}$ be the solution to $\boldsymbol{C}_n \boldsymbol{x} = \tilde{\boldsymbol{b}}$ and $\boldsymbol{x}^{(j)}$ be the j-th iterate of CGNE applied to the system $\boldsymbol{C}_n \boldsymbol{C}_n^T \boldsymbol{y} = \tilde{\boldsymbol{b}}$ under the change of variables $\boldsymbol{x} = \boldsymbol{C}_n^T \boldsymbol{y}$. If the eigenvalues $\{\delta_k\}$ of $\boldsymbol{C}_n \boldsymbol{C}_n^T$ are such that*

$$0 < \delta_1 \le \cdots \le \delta_p \le b_1 \le \delta_{p+1} \le \cdots \le \delta_{n-q} \le b_2 \le \delta_{n-q+1} \le \cdots \le \delta_n,$$

then

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(j)}||_2}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_2} \le 2 \left(\frac{b-1}{b+1} \right)^{j-p-q} \cdot \max_{\delta \in [b_1, b_2]} \left\{ \prod_{k=1}^{p} \left(\frac{\delta - \delta_k}{\delta_k} \right) \prod_{k=n-q+1}^{n} \left(\frac{\delta_k - \delta}{\delta_k} \right) \right\}, \tag{4.2}$$

for $j \ge p + q$. Here $b \equiv (b_2/b_1)^{\frac{1}{2}} \ge 1$.

We can derive (4.2) by passing linear polynomials through the outlying eigenvalues δ_k for $1 \le k \le p$ and $n - q + 1 \le k \le n$, and using a $(j - p - q)$th degree Chebyshev polynomial to minimize the error in the interval $[b_1, b_2]$. Since we always have

$$0 \le \frac{\delta_k - \delta}{\delta_k} \le 1, \quad n - q + 1 \le k \le n$$

for $\delta \in [b_1, b_2]$, (4.2) can be simplified to

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(j)}||_2}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_2} \le 2 \left(\frac{b-1}{b+1} \right)^{j-p-q} \cdot \max_{\delta \in [b_1, b_2]} \left\{ \prod_{k=1}^{p} \left(\frac{\delta - \delta_k}{\delta_k} \right) \right\}. \tag{4.3}$$

For the preconditioned system, the iteration matrix $\boldsymbol{C}_n$ is given by

$$\boldsymbol{C}_n = (\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2 \gamma_2 \boldsymbol{I}_n)^{-1} (\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2 \gamma_2 \boldsymbol{I}_n).$$

Theorem 3.6 implies that we can choose $b_1 = 1$ and $b_2 = \beta$ in (3.13). Then, p and q are constants that are independent of n. In order to use (4.3), we need a lower bound for δ_k, $1 \le k \le p$. We note that

$$\begin{aligned}
&||(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2 \gamma_2 \boldsymbol{I}_n)^{-1} (\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2 \gamma_2 \boldsymbol{I}_n)||_2 \\
&\le ||\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2 \gamma_2 \boldsymbol{I}_n||_2^{-1} ||\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2 \gamma_2 \boldsymbol{I}_n||_2 \cdot \\
&\quad \cdot \kappa(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2 \gamma_2 \boldsymbol{I}_n),
\end{aligned}$$

and there exists a constant $c_3 > 0$ independent of n such that

$$\|\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n\|_2 \le c_3 := 4 + \gamma_1\pi + \gamma_2.$$

Therefore, it remains to show that there exists $c_4 > 0$ independent of n such that

$$\left\|\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n\right\|_2 \ge c_4. \tag{4.4}$$

But this follows from the fact that

$$\|\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n\|_2 \ge \|\boldsymbol{T}_n[g_2] + h^2\gamma_2 \boldsymbol{I}_n\|_2 - \|h\gamma_1 \boldsymbol{T}_n[g_1]\|_2.$$

We remark that the singular values of $\boldsymbol{T}_n[g_2]$ and $\boldsymbol{T}_n[g_1]$ are distributed as $|g_2| = \theta^2$ and $|g_1| = |\theta|$ respectively (see [20, 27]). Therefore, for sufficiently small h, we have the inequality stated in (4.4). It follows by Theorem 2.2 that

$$\begin{aligned}
\delta_k &\ge \min_{\ell} \delta_\ell \\
&= \left\|(\boldsymbol{T}_n[g_2] + h\gamma_1 \boldsymbol{T}_n[g_1] + h^2\gamma_2 \boldsymbol{I}_n)^{-1}(\boldsymbol{P}_n^{II} + h\gamma_1 \boldsymbol{P}_n^{I} + h^2\gamma_2 \boldsymbol{I}_n)\right\|_2^{-2} \\
&\ge (\frac{c_4}{c_3})^2 16n^4(1+\frac{c_1}{\sqrt{n}})^2(1+\frac{c_2}{n})^2 \;=\; cn^{-4},
\end{aligned}$$

for $1 \le k \le n$, where c is a positive constant. Thus, for $1 \le k \le p$ and $\delta \in [1, \beta]$, we have that

$$0 \le \frac{\delta - \delta_k}{\delta_k} \le cn^4.$$

Hence, (4.2) becomes

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(j)}||_2}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_2} < c^p n^{4p} \left(\frac{b-1}{b+1}\right)^{j-p-q}.$$

Given arbitrary tolerance $\epsilon > 0$, an upper bound for the number of iterations required to make

$$\frac{||\boldsymbol{x} - \boldsymbol{x}^{(j_0)}||_2}{||\boldsymbol{x} - \boldsymbol{x}^{(0)}||_2} < \epsilon$$

is therefore given by

$$j_0 \equiv p + q - \frac{p\log c + 4p\log n - \log\epsilon}{\log\left(\frac{b-1}{b+1}\right)} = \mathcal{O}(\log n).$$

Since each CGNE iteration requires $\mathcal{O}(n\log n)$ operations, the total cost of CGNE is at most $\mathcal{O}(n\log^2 n)$ arithmetical operations.

5. Numerical Results. In this section, we test our banded preconditioners on a SGI O2 workstation. All experiments are performed in MATLAB with a machine precision of 10^{-16}.

Our problems have homogeneous Dirichlet boundary conditions and known solutions. We apply GMRES, BiCGSTAB and CGNE methods to

$$\boldsymbol{B}_n^{-1}\boldsymbol{A}_n\boldsymbol{x} = \boldsymbol{B}_n^{-1}\boldsymbol{b}.$$

Here $\boldsymbol{B}_n$ represents the banded preconditioner (3.1). The iterative method started with the zero vector and the vector $\boldsymbol{b}$ is given by (2.5).

Tables 5.1 – 5.4 list the number of matrix-vector products of $\boldsymbol{A}_n$ or $\boldsymbol{A}_n^{\mathrm{T}}$ required until the residual norms produced by the different iterative method satisfied $\|\boldsymbol{r}^{(j)}\|_2/\|\boldsymbol{r}^{(0)}\|_2 < 10^{-7}$. The symbol $*$ denotes that the method stopped without converging to the desired tolerance in 1000 iteration steps. We remark that GMRES uses one matrix-vector product per step, and BiCGSTAB and CGNE use two matrix-vector products per step. Note that the preconditioned systems need in addition the solution of $\boldsymbol{B}_n\boldsymbol{x} = \boldsymbol{y}$ or $\boldsymbol{B}_n^{\mathrm{T}}\boldsymbol{x} = \boldsymbol{y}$. But since $\boldsymbol{B}_n$ is a tridiagonal matrix we compute the solution quickly by a permuted back-substitution algorithm as implemented in MATLAB. In the tables, the symbol $\boldsymbol{I}_n$ means that the system is solved without using a preconditioner.

In the tables, we also determine the error between the numerical approximation and the true solution at the sinc points defined as follows:

$$E := \sqrt{\sum_{k=-M}^{N} |u_k - u(x_k)|^2}.$$

Here we obtain this error by determining $\{u_k\}_{k=-M}^{N}$, where we solve the system (2.5) by a direct method.

In the numerical tests, we consider the following examples:

Example 5.1. (see [17, p. 119]) The discretization of

$$\begin{gathered} u''(x) + \tfrac{1}{6x}u'(x) - \tfrac{1}{x^2}u(x) = -\tfrac{19}{6}\sqrt{x} \quad (x \in (0,1))\,, \\ u(0) = u(1) = 0, \end{gathered}$$

which has solution $u(x) = x^{3/2}(1-x)$, is given by (2.3) with

$$\boldsymbol{D}_n^I \;=\; \boldsymbol{D}_n\left[\frac{-\phi''}{(\phi')^2} - \frac{p}{\phi'}\right] = \boldsymbol{D}_n\left[\frac{5-11x}{6}\right]$$

and

$$\boldsymbol{D}_n^{II} \;=\; \boldsymbol{D}_n\left[\frac{1}{\phi'}\left(\frac{1}{\phi'}\right)'' - \frac{1}{\phi'}\left(\frac{p}{\phi'}\right)' + \frac{2q}{(\phi')^2}\right] = \boldsymbol{D}_n\left[\frac{(x-1)(12-x)}{6}\right]\,.$$

We choose the conformal map $\phi(z) = \log\left(\frac{z}{1-z}\right)$ and as in [17, p. 119] $M = 2^l, N = \frac{3M}{2} - 1$ and $h = \frac{\pi}{\sqrt{3M}}$. This problem has a regular singular point at $x = 0$. $\square$

Example 5.2. (see [17, p. 126]) The discretization for the problem on $(0,\infty)$ given by

$$\begin{gathered} u''(x) - \tfrac{x}{x^2+1}u'(x) - \tfrac{1}{x^2+1}u(x) = \tfrac{2x(x^2-4)}{(x^2+1)^3} \quad (x \in (0,\infty))\,, \\ u(0) = \lim_{x\to\infty} u(x) = 0, \end{gathered}$$

which has solution $u(x) = \frac{x}{x^2+1}$, takes the form (2.3) with

$$\boldsymbol{D}_n^I = \boldsymbol{D}_n\left[\frac{2x^2+1}{x^2+1}\right] \quad \text{and} \quad \boldsymbol{D}_n^{II} = \boldsymbol{D}_n\left[\frac{-2x^4}{(x^2+1)^2}\right].$$

TABLE 5.1
Results for Example 5.1.

		CGNE		GMRES		BiCGSTAB	
n	E	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$
10	4.50e-03	24	12	11	8	21	10
20	8.48e-04	56	26	21	9	39	9
40	5.92e-05	154	28	40	8	67	9
80	1.05e-06	492	26	72	6	117	6
160	2.77e-09	1696	24	107	4	181	4
320	5.08e-13	*	24	153	3	261	3

TABLE 5.2
Results for Example 5.2.

		CGNE		GMRES		BiCGSTAB	
n	E	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$
8	3.14e-02	18	18	9	9	17	12
16	4.01e-03	40	28	17	12	35	14
32	3.55e-04	106	32	33	13	75	14
64	1.37e-05	312	32	64	12	138	12
128	1.18e-07	1020	30	125	10	250	10
256	1.15e-10	*	28	213	7	437	7
512	5.07e-14	*	26	373	5	921	5

We choose the conformal map $\phi(z) = \log(z)$ and as in [17, p. 126] $M = 2^l, N = M - 1$ and $h = \frac{\pi}{\sqrt{2M}}$. □

For Example 5.1 and Example 5.2, Assumptions **(A1)** and **(A2)** are fulfilled. In Figure 5.1 we plot the singular values of $\boldsymbol{A}_n$ and of the preconditioned matrix $\boldsymbol{B}_n^{-1}\boldsymbol{A}_n$. We see, that except some outliers the singular values of $\boldsymbol{B}_n^{-1}\boldsymbol{A}_n$ lie in an fixed interval independent of n. For CGNE, our numerical results confirm our expected theoretical results, that the number of CGNE iterations is of order $\mathcal{O}(\log n)$.

We note that BiCGSTAB and GMRES use different Krylov subspaces [8, p.90] and therefore we cannot compare their iteration results directly. However, we observe in the tables that GMRES and BiCGSTAB converge very fast. These numerical results illustrate the effectiveness of our proposed preconditioners.

In the following examples we apply the banded preconditioner to precondition the Sinc coefficient matrix when Assumption **(A2)** is not fulfilled.

EXAMPLE 5.3. (see [2, 7]) We consider the convection problem

$$\begin{aligned} u''(x) - \kappa u'(x) &= f(x) \quad (x \in (0,1)), \\ u(0) = u(1) &= 0. \end{aligned} \tag{5.1}$$

The solution of (5.1) is difficult to compute for large values κ. We compute the

FIG. 5.1. *Singular values of* $\boldsymbol{A}_n$ *(left) and of* $\boldsymbol{B}_n^{-1}\boldsymbol{A}_n$ *(right) for* $n \in \{10, 20, 40, 80, 160, 320\}$ *given in Example* 5.1.

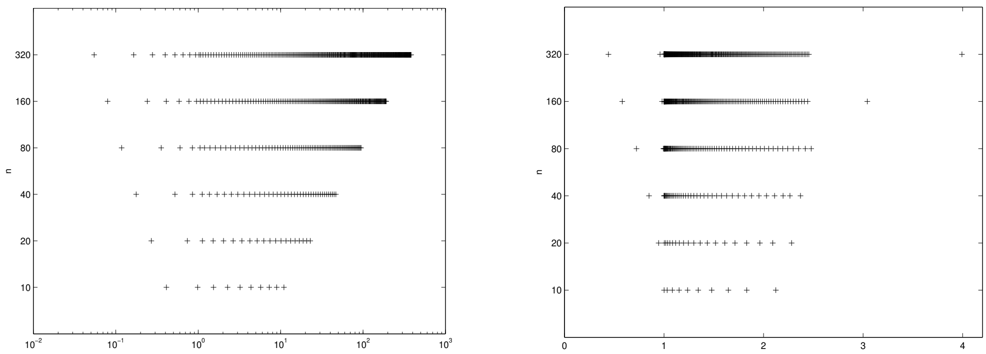


TABLE 5.3
Results for Example 5.3 *with* $\kappa = 100$.

		CGNE		GMRES		BiCGSTAB	
n	E	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$
16	1.12e-01	48	34	17	13	61	19
32	2.07e-02	132	44	31	14	107	18
64	1.02e-03	420	44	52	13	206	18
128	9.77e-06	1408	38	95	12	347	16
256	1.06e-08	*	38	144	6	491	6
512	4.54e-13	*	30	206	4	890	4

solution for $f(x) = -\kappa$. The discretization is given by (2.3) with

$$\boldsymbol{D}_n^I = \boldsymbol{D}_n\,[1 - 2x + \kappa x(1-x)] \quad \text{and} \quad \boldsymbol{D}_n^{II} = \boldsymbol{D}_n\,[x(x-1)(2 + \kappa(2x-1))]\,.$$

We choose $\phi(z) = \log\left(\frac{z}{1-z}\right)$, $h = \frac{\pi}{\sqrt{2M}}$ and $N = 2^l, M = N - 1$. Note that Ernst [7] used a discretization based on the Galerkin finite element method and solved the resulting linear system by GMRES without a preconditioner. □

EXAMPLE 5.4. (see [2]) Consider the differential equation (for $\kappa > 0$) defined by

$$\begin{aligned} u'' - \tfrac{\kappa}{x} u'(x) &= -\kappa(\kappa+1)x^{\kappa-1} \quad (x \in (0,1))\,, \\ u(0) = u(1) &= 0. \end{aligned}$$

This problem has the difficulty represented by a regular singular point at $x = 0$ and a boundary layer at $x = 1$ when $\kappa \gg 0$. The linear system (1.7) takes the form (2.3) with

$$\boldsymbol{D}_n^I = \boldsymbol{D}_n\,[1 - 2x + \kappa(1-x)] \quad \text{and} \quad \boldsymbol{D}_n^{II} = \boldsymbol{D}_n\,[x(x-1)(2+\kappa)]\,.$$

We choose the conformal map $\phi(z) = \log\left(\frac{z}{1-z}\right)$ and $h = \frac{\pi}{\sqrt{2M}}$ and $N = 2^l, M = N-1$. □

TABLE 5.4
Results for Example 5.4 for $\kappa = 100$.

		CGNE		GMRES		BiCGSTAB	
n	E	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$	$\boldsymbol{I}_n$	$\boldsymbol{B}_n$
8	1.50e-01	18	20	9	9	29	17
16	1.06e-01	52	36	17	14	95	21
32	2.09e-02	160	56	33	17	517	29
64	1.04e-03	384	70	65	21	*	35
128	9.83e-06	*	92	129	52	*	42
256	1.02e-08	*	108	240	55	*	45
512	4.67e-13	*	102	430	6	*	12

6. Concluding Remarks. We remark that the accuracy of the computed solution depends only on the Galerkin method used in the discretization of the boundary value problem. However, the convergence rate of the discrete system and the costs per iteration of the iterative method depend on how we discretize the boundary value problem. It is advantageous to use the Sinc method to discretize the boundary value problem because the Sinc-Galerkin method for boundary value problems is convergent exponentially (see (2.6) and Tables 5.1–5.4). However, we require to solve n-by-n Sinc systems where their coefficient matrices are dense. A straightforward application of the Gaussian elimination method will result in an algorithm, which takes $\mathcal{O}(n^3)$ arithmetical operations. The main contribution of this paper is to propose banded preconditioners to precondition Sinc matrices and speed up the convergence rate of conjugate gradient type methods. The cost of our proposed method for Sinc systems is significantly less than the $\mathcal{O}(n^3)$ cost required by Gaussian elimination method for solving Sinc systems.

Finally, we remark that we can employ the finite difference or the finite element method to discrete the boundary value problem, and therefore banded system solvers can be used to solve the corresponding linear system in $\mathcal{O}(n)$ operations. However, in order to obtain a reasonably accurate solution, a small step-size has to be used in the finite difference or the finite element method and hence the dimension of the resulting matrix system will be very large compared with the size of the Sinc system [19].

Acknowledgment. Part of this work was done while the second author was visiting the Department of Mathematics of “The Chinese University of Hong Kong”. The authors would like to acknowledge R. Chan, B. Fischer and G. Steidl for numerous fruitful and enlightening discussions. We would also like to thank the referees for their valuable suggestions.

REFERENCES

[1] G. Ammar and W. Gragg, *Superfast Solution of Real Positive Definite Toeplitz Systems*, SIAM J. Matrix Appl. 9 (1988), pp. 61 – 76.
[2] T. S. Carlson J. Lund and K.L. Bowers, *A Sinc–Galerkin method for convection dominated transport*, in Computation and Control III, “Progress in Systems and Control Theory”,15, (1993), pp. 121 – 139.
[3] R. Chan and M. Ng, *Conjugate Gradient Methods for Solving Toeplitz Systems*, SIAM Review, 38 (1996), pp. 427 – 482.

[4] R. Chan and M. Ng, *Fast Iterative Solvers for Toeplitz-Plus-Band Systems*, SIAM J. Sci. Comput., 14 (1993), pp. 1013 – 1019.
[5] R. Chan and M. Yeung, Circulant preconditioners for complex Toeplitz Systems. SIAM *J. Numer. Anal.*, 30:1193 – 1207, 1993.
[6] T. Chan, *An Optimal Circulant Preconditioner for Toeplitz Systems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 766 – 771.
[7] O. Ernst, *Residual-minimizing Krylov Subspace Methods for Stabilized Discretizations of Convection Diffusion Equations*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1079–1101.
[8] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.
[9] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, Los Angeles, 1958.
[10] G. Golub and C. Van Loan, *Matrix Computations*, 2nd Ed., The Johns Hopkins University Press, Baltimore, 1989.
[11] G. Heinig and K. Rost, Representations of Toeplitz–plus–Hankel Matrices Using Trigonometric Transforms with Application to fast Matrix-vector Multiplication. *Linear Algebra Appl.*, 254 (1997), pp. 193 – 226.
[12] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.
[13] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.
[14] T. Kailath and A. Sayed, *Displacement Structure: Theory and Applications*, SIAM Review, 37 (1995), pp. 297 – 386.
[15] N. Levinson, *The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction*, J. Math. and Phys., 25 (1946), pp. 261 – 278.
[16] J. Lund, *Symmetrization of the Sinc-Galerkin Method for Boundary Value Problems*, Mathematics of Computation, 47, (1986), pp. 571 – 588.
[17] J. Lund and K. Bowers, *Sinc Methods for Quadrature and Differential Equations*, SIAM, 1992.
[18] N. Nachtigal, S. C. Reddy, and L. N. Trefethen, *How Fast are Nonsymmetric Matrix Iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778 – 795.
[19] M. Ng, *Fast Iterative Methods for Symmetric Sinc-Galerkin Systems*, IMA Journal of Numerical Analysis, 19 (1999), 357 – 373.
[20] S. V. Parter, On the distribution of singular values of Toeplitz matrices. *Linear Algebra Appl.*, 80:115 – 130, 1986.
[21] D. Potts and G. Steidl, Optimal Trigonometric Preconditioners for Nonsymmetric Toeplitz Systems. *Linear Algebra Appl.*, 281:265 – 292, 1998.
[22] Y. Saad, *Iterative Methods for Sparse Linear Systems.* PWS Publ., Boston, 1996.
[23] F. Stenger, *Numerical Methods Based on Whittaker Cardinal, Sinc Functions*, SIAM Review, 23 (1981), pp. 85–109.
[24] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer Series in Computational Mathematics, Springer-Verlag, 1993.
[25] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer-Verlag, 1993.
[26] G. Strang, *A Proposal for Toeplitz Matrix Calculations*, Stud. Appl. Math., 74 (1986), pp. 171 – 176.
[27] E. E. Tyrtyshnikov, A unifying approach to some old and new theorems on distribution and clustering. *Linear Algebra Appl.*, 232:1 – 43, 1996.
[28] H. Van der Vorst, Preconditioning by Incomplete Decomposition. *Ph. D thesis.*, Rijksuniversiteit Utrecht, 1982.