On Preconditioned Iterative Methods for Burgers Equations

Bai, Zhong Zhi; Huang, Yu Mei; Ng, Michael K.

Published in:
SIAM Journal on Scientific Computing

DOI:
10.1137/060649124

Published: 15/02/2007

Document Version:
Publisher's PDF, also known as Version of record

SIAM J. SCI. COMPUT.
Vol. 29, No. 1, pp. 415–439

ON PRECONDITIONED ITERATIVE METHODS FOR BURGERS EQUATIONS*

ZHONG-ZHI BAI†, YU-MEI HUANG‡, AND MICHAEL K. NG‡

Abstract. We study the Newton method and a fixed-point method for solving the system of nonlinear equations arising from the Sinc–Galerkin discretization of the Burgers equations. In each step of the Newton method or the fixed-point method, a structured subsystem of linear equations is obtained and needs to be solved numerically. In this paper, preconditioning techniques are applied to solve such linear subsystems. The bounds for eigenvalues of the preconditioned matrices are derived and numerical examples are given to illustrate the effectiveness of the proposed methods. We also find that a combination of the Newton/fixed-point iteration with the preconditioned GMRES method is quite efficient for the Sinc–Galerkin discretization of the Burgers equations.

Key words. Burgers equation, Sinc–Galerkin discretization, Toeplitz-like matrices, preconditioners, GMRES method

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

DOI. 10.1137/060649124

1. Introduction.

We consider an iterative solution of the system of nonlinear equations

$$\mathbf{F}(\mathbf{u}) \equiv B\mathbf{u} + C\Psi(\mathbf{u}) - \mathbf{b} = 0, \tag{1.1}$$

where B and C are n-by-n matrices, $\mathbf{b}$ is a given n-vector, and $\Psi : \mathbb{R}^n \to \mathbb{R}^n$, with

$$\Psi(\mathbf{u}) = (\psi_1(u_1), \psi_2(u_2), \ldots, \psi_n(u_n))^T \quad \text{and} \quad \mathbf{u} = (u_1, u_2, \ldots, u_n)^T, \tag{1.2}$$

is a continuous diagonal mapping defined on the open ball

$$\mathcal{U}_\delta := \{u \in \mathbb{R}^n \mid \|\mathbf{u}\| < \delta\}.$$

See [19] for general background and applications about the system of mildly nonlinear equations.

Nonlinear systems such as (1.1)–(1.2) may also arise from the Sinc–Galerkin discretization of the Burgers equation [15]:

$$\begin{cases} \mathcal{P}^{(2)}u(x,t) \equiv \dfrac{\partial u}{\partial t}(x,t) + u(x,t)\dfrac{\partial u}{\partial x}(x,t) - \varepsilon\dfrac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), & a < x < b,\ t \geq 0, \\ u(a,t) = \gamma(t) \quad \text{and} \quad u(b,t) = \delta(t), & t \geq 0, \\ u(x,0) = g(x), & a \leq x \leq b. \end{cases} \tag{1.3}$$

*Received by the editors January 5, 2006; accepted for publication (in revised form) September 19, 2006; published electronically February 15, 2007.
http://www.siam.org/journals/sisc/29-1/64912.html

†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 100080, The People's Republic of China (bzz@lsec.cc.ac.cn). This author's research was supported by The National Basic Research Program (2005CB321702), The China NNSF National Outstanding Young Scientist Foundation (10525102), and The National Natural Science Foundation (10471146), The People's Republic of China.

‡Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (ymhuang@math.hkbu.edu.hk, mng@math.hkbu.edu.hk). The research of the third author was supported in part by RGC grants 7046/03P, 7035/04P, and 7035/05P, and HKBU FRGs.

In the Sinc–Galerkin method, we approximate u by

$$\widehat{u} = \sum_{j=-m_t-1}^{n_t} \sum_{i=-m_x-1}^{n_x+1} u_{ij}\chi_i(x)\theta_j(t),$$

where the functions $\chi_i(x)$ and $\theta_j(t)$ are bases in the spatial space and the temporal space, respectively, with

$$\chi_i(x) = \begin{cases} \dfrac{b-x}{b-a}, & i = -m_x - 1, \\ \mathcal{S}(i, h_x) \circ \phi_x(x), & -m_x \le i \le n_x, \\ \dfrac{x-a}{b-a}, & i = n_x + 1, \end{cases}$$

and

$$\mathcal{S}(i, h_x) \circ \phi_x(x) := \operatorname{sinc}\left[\frac{\phi_x(x) - ih_x}{h_x}\right],$$

$$\theta_j(t) = \begin{cases} \dfrac{t+1}{t^2+1}, & j = -m_t - 1, \\ \mathcal{S}(j, h_t) \circ \phi_t(t), & -m_t \le j \le n_t; \end{cases}$$

see [15] for details. Here, $\phi_x(x)$ and $\phi_t(t)$ are the restrictions of the conformal mapping $\phi_z(z)$ onto the real intervals (a, b) and $(0, +\infty)$, respectively, with $\phi_z(z)$ a mapping from a simply connected domain $\mathcal{D}$ onto

$$\mathcal{D}_d := \{z \mid z = x + \imath y, |y| < d, d > 0\},$$

with $\imath$ the imaginary unit. We remark that the first and second derivatives of $\phi_z(z)$ with respect to the variable z will be denoted as $\phi_z^{'}(z)$ and $\phi_z^{''}(z)$, respectively.

The Galerkin method enables us to determine

$$u_{ij}, \quad -m_x - 1 \le i \le n_x + 1, \quad -m_t - 1 \le j \le n_t,$$

through solving the system of nonlinear equations

$$\langle \mathcal{P}^{(2)}\widehat{u} - f, \mathcal{S}_{kl} \rangle = 0,$$

where the inner product is defined as

$$\langle f, g \rangle = \int_0^{\infty} \int_a^b f(x,t)g(x,t)\omega_x(x)\omega_t(t)dxdt,$$

with $\omega_x(x)$ and $\omega_t(t)$ being two weighting functions with respect to the spatial and the temporal variables, respectively, and $\mathcal{S}_{kl} = \mathcal{S}(k, h_x)\mathcal{S}(l, h_t)$; see [15].

The most distinctive feature of the Sinc basis is that it can lead to exponential decreasing rate of the error. Moreover, the decreasing rate is maintained when the solution of the boundary value problem has boundary singularities.

In the form of (1.1)–(1.2), the resulting Sinc–Galerkin system can be described as follows; see [15] for more details. Now, the mapping Ψ is given by

$$\Psi(\mathbf{u}) = (u_1^2, u_2^2, \ldots, u_n^2)^T, \quad \text{with} \quad \mathbf{u} = (u_1, u_2, \ldots, u_n)^T, \tag{1.4}$$

the matrices B and C are given by

$$B = (\varepsilon(T_x^{(2)} + D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)} + D_x^{(2)})) \otimes Q_t + Q_x \otimes (D_t^{(3)}T_t^{(1)} + T_t^{(1)}D_t^{(3)} + D_t^{(4)}), \tag{1.5}$$

and

$$C = (D_x^{(3)}T_x^{(1)} + T_x^{(1)}D_x^{(3)} + D_x^{(4)}) \otimes Q_t, \tag{1.6}$$

where $T_z^{(i)}$ ($i = 1, 2$ and $z \in \{x, t\}$) are $(m_z + n_z + 1)$-by-$(m_z + n_z + 1)$ Toeplitz matrices, with

$$T_z^{(1)} = \begin{bmatrix} 0 & -1 & \frac{1}{2} & \cdots & \frac{(-1)^{m_z+n_z}}{m_z+n_z} \\ 1 & & & & \vdots \\ -\frac{1}{2} & & \ddots & & \frac{1}{2} \\ \vdots & & & & -1 \\ \frac{(-1)^{m_z+n_z}}{m_z+n_z} & \cdots & -\frac{1}{2} & 1 & 0 \end{bmatrix}, \tag{1.7}$$

$$T_z^{(2)} = \begin{bmatrix} \frac{\pi^2}{3} & -2 & \frac{2}{2^2} & \cdots & \frac{(-1)^{m_z+n_z}2}{(m_z+n_z)^2} \\ -2 & & & & \vdots \\ \frac{2}{2^2} & & \ddots & & \frac{2}{2^2} \\ \vdots & & & & -2 \\ \frac{(-1)^{m_z+n_z}2}{(m_z+n_z)^2} & \cdots & \frac{2}{2^2} & -2 & \frac{\pi^2}{3} \end{bmatrix}, \tag{1.8}$$

and $D_z^{(i)}$ and Q_z ($i = 1, 2, 3, 4$ and $z \in \{x, t\}$) are $(m_z + n_z + 1)$-by-$(m_z + n_z + 1)$ diagonal matrices, with

$$D_z^{(1)} = \frac{h_z}{2} \cdot \operatorname{diag}\left[\left\{-\frac{\phi_z''(z)}{(\phi_z'(z))^2} - \frac{2\omega_z'(z)}{\phi'(z)\omega_z(z)}\right\}_{z=-m_z}^{n_z}\right], \tag{1.9}$$

$$D_z^{(2)} = \frac{h_z^2}{2} \cdot \operatorname{diag}\left[\left\{-\frac{\omega_z''(z)}{(\phi_z'(z))^2\omega_z(z)}\right\}_{z=-m_z}^{n_z}\right], \tag{1.10}$$

$$D_z^{(3)} = \frac{h_z}{2} \cdot \operatorname{diag}\left[\{-\omega_z(z)\}_{z=-m_z}^{n_z}\right], \tag{1.11}$$

$$D_z^{(4)} = \frac{h_z^2}{2} \cdot \operatorname{diag}\left[\left\{-\frac{\omega_z'(z)}{\phi_z'(z)}\right\}_{z=-m_z}^{n_z}\right], \tag{1.12}$$

and

$$Q_z = \operatorname{diag}\left[\left\{\frac{\omega_z(z)}{\phi_z'(z)}\right\}_{z=-m_z}^{n_z}\right]. \tag{1.13}$$

There are two classes of methods for computing the solution of the system of nonlinear equations (1.1). One casts this problem as a general nonlinear system and uses a nonlinear solver such as the Newton method [15] to solve it, and the other considers this problem as a special system of nonlinear equations

$$B\mathbf{u} = \mathbf{b} - C\mathbf{u} \odot \mathbf{u},$$

where $\odot$ is the entrywise multiplication, and uses the fixed-point method to solve it. It has been shown in [15] that the spectral radius of the inverse of the coefficient matrix B is bounded above by $1/(4\varepsilon)$. Evidently, as ε decreases, the iterative method used to solve this nonlinear system deteriorates. See also [1, 2, 3, 4, 8, 9, 10, 21, 22] and references therein for sequential and parallel matrix splitting iteration methods for solving the system of nonlinear equations (1.1).

For both Newton and fixed-point iterations employed to solve the system of nonlinear equations (1.1), at each step we need to solve a subsystem of linear equations of the form

$$(B + CD)\mathbf{z} = \mathbf{r}, \tag{1.14}$$

where D is a diagonal matrix. A straightforward application of the Gaussian elimination to the linear system (1.14) will result in an algorithm of $\mathcal{O}(n^6)$ complexity, where n is the size of the matrix $T_z^{(i)}$ $(i = 1, 2)$, $D_z^{(i)}$ $(i = 1, 2, 3, 4)$, or Q_z, $z \in \{x, t\}$. For an n-by-n Toeplitz linear system, fast direct solvers of complexity $\mathcal{O}(n^2)$ have been developed; see, for instance, Levinson [14]. 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. Hence, fast direct Toeplitz solvers that are based on small displacement rank matrices are not applicable to the Toeplitz-plus-diagonal linear systems or to the block Toeplitz-like-plus-diagonal linear systems. For details about displacement ranks, we refer the reader to Kailath and Sayed [13].

However, we note that for any n^2-vector $\mathbf{q}$, the matrix-vector product $(B+CD)\mathbf{q}$ can be computed in $\mathcal{O}(n^2 \log n)$ operations. In fact, for $i = 1, 2$ and $z \in \{x, t\}$ the matrix-vector product $T_z^{(i)}\mathbf{q}$ can be obtained by first embedding $T_z^{(i)}$ into a $2n$-by-$2n$ circulant matrix and then using *fast Fourier transforms* (FFTs) to compute it [12]. Since for all $i \in \{1, 2, 3, 4\}$ and $z \in \{x, t\}$, $D_z^{(i)}$ is a diagonal matrix, the matrix-vector product $D_z^{(i)}\mathbf{q}$ can be computed in $\mathcal{O}(n^2)$ operations. Thus, iterative methods such as GMRES [20] can be employed to economically solve the linear system (1.14). Usually, in order to accelerate the convergence speed of GMRES, we need to precondition the linear system (1.14) by an approximate matrix with respect to the coefficient matrix $B + CD$. Therefore, to solve the original linear system, we turn to solving the preconditioned linear system instead; see [17, 7] and the references therein.

The outline of this paper is as follows. In section 2, we demonstrate some basic properties about the system of nonlinear equations (1.1). In section 3, we construct preconditioners for the coefficient matrix of the linear system (1.14); in addition, we study preconditioning properties of the preconditioned matrix and show that its eigenvalues are tightly and uniformly bounded in a rectangular on the complex plane independent of the size of the matrix. Several numerical examples are used to show the effectiveness of the proposed methods in section 4. Finally, in section 5, we end this paper with some concluding remarks.

2. Preliminary results. In this section, we give a few theoretical properties of the system of nonlinear equations (1.1).

THEOREM 2.1. *Let $B \in \mathbb{R}^{n\times n}$ be nonsingular, and $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ satisfy*

$$\|\Psi(\mathbf{u})\| \le \gamma\|\mathbf{u}\|^p \quad \forall \mathbf{u} \in \mathcal{U}_\delta,$$

where γ is a positive constant, $p \ge 2$ is a positive integer, $\|\cdot\|$ is any consistent matrix or vector norm, and $\mathcal{U}_\delta$ is the ball centered at the origin with the radius δ. Denote

$$\alpha = \|B^{-1}C\| \quad \textit{and} \quad \beta = \|B^{-1}b\|,$$

and assume that

$$\alpha\beta^{p-1}\gamma < (p-1)^{p-1}p^{-p}.$$

If $\mathbf{u}_* \in \mathcal{U}_\delta$ is a solution of the system of nonlinear equations (1.1), then $\mathbf{u}_* \in \mathcal{U}_{\delta_0}$, with δ_0 the smallest positive solution of the nonlinear equation

$$\alpha\gamma t^p - t + \beta = 0.$$

Proof. Define

$$g(t) = \alpha\gamma t^p - t + \beta$$

and

$$t_0 = (\alpha\gamma p)^{-\frac{1}{p-1}}.$$

Then we have

$$\begin{cases} g(0) = \beta > 0, & g(t_0) = (\alpha\gamma)^{-\frac{1}{p-1}}[(\alpha\beta^{p-1}\gamma)^{\frac{1}{p-1}} - (p-1)p^{-\frac{p}{p-1}}] < 0, \\ g'(t_0) = 0, & g'(t)(t-t_0) > 0 \quad \forall t \in (0,+\infty). \end{cases}$$

It follows that $g(t)$ has two positive roots; we denote the smallest one as δ_0. Therefore, $g(t) \ge 0$ for $t \in [0, \delta_0]$. Since $\mathbf{u}_* \in \mathcal{U}_\delta$ is a solution of the system of nonlinear equations (1.1), it holds that

$$B\mathbf{u}_* + C\Psi(\mathbf{u}_*) = \mathbf{b},$$

or equivalently,

$$\mathbf{u}_* + B^{-1}C\Psi(\mathbf{u}_*) = B^{-1}\mathbf{b}.$$

Therefore,

$$\|B^{-1}b\| \ge \|\mathbf{u}_*\| - \|B^{-1}C\|\|\Psi(\mathbf{u}_*)\| \ge \|\mathbf{u}_*\| - \gamma\|B^{-1}C\|\|\mathbf{u}_*\|^p,$$

i.e.,

$$\alpha\gamma\|\mathbf{u}_*\|^p - \|\mathbf{u}_*\| + \|B^{-1}b\| \ge 0.$$

We immediately know that this inequality holds when $\|\mathbf{u}_*\| \le \delta_0$ or $\mathbf{u}_* \in \mathcal{U}_{\delta_0}$. ∎

COROLLARY 2.2. *Let the assumptions of Theorem* 2.1 *be satisfied. If* $\Psi(\mathbf{u}) = (u_1^2, u_2^2, \ldots, u_n^2)^T$, *with* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T$, *then* $\gamma = 1$ *and*

$$\delta_0 = \frac{1}{2\alpha}\left[1 - \sqrt{1 - 4\alpha\beta}\right],$$

provided $\alpha\beta < \frac{1}{4}$.

THEOREM 2.3. *Let* $B \in \mathbb{R}^{n\times n}$ *be nonsingular. Denote* $\alpha = \|B^{-1}C\|$ *and* $\beta = \|B^{-1}b\|$, *where* $\|\cdot\|$ *is any consistent matrix or vector norm. Assume that* $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ *satisfies*

(i) *the bounded condition* $\|\Psi(\mathbf{u})\| \le \gamma\|\mathbf{u}\|^p \ \forall \mathbf{u} \in \mathcal{U}_\delta$, *where* $\mathcal{U}_\delta$ *is the ball centered at the origin with the radius* δ, γ *is a positive constant, and* $p \ge 2$ *a positive integer, such that*

$$\alpha\beta^{p-1}\gamma < (p-1)^{p-1}p^{-p};$$

(ii) *the Lipschitz condition* $\|\Psi(\mathbf{u}) - \Psi(\mathbf{v})\| \le \tau\|\mathbf{u} - \mathbf{v}\| \ \forall \mathbf{u}, \mathbf{v} \in \mathcal{U}_{\delta_0}$, *where* $\tau > 0$ *is the Lipschitz constant and* δ_0 *the smallest positive root of the nonlinear equation*

$$\alpha\gamma t^p - t + \beta = 0.$$

If $\alpha\tau < 1$, *then the system of nonlinear equations* (1.1) *has a unique solution* $\mathbf{u}_* \in \mathcal{U}_{\delta_0}$.

Proof. Define $\mathbf{G} : \mathcal{U}_{\delta_0} \subset \mathbb{R}^n \to \mathbb{R}^n$ as

$$\mathbf{G}(\mathbf{u}) = B^{-1}b - B^{-1}C\Psi(\mathbf{u}).$$

Then $\mathbf{G}$ is a contraction mapping defined on $\mathcal{U}_{\delta_0}$ as

$$\|\mathbf{G}(\mathbf{u}) - \mathbf{G}(\mathbf{v})\| = \|B^{-1}C[\Psi(\mathbf{u}) - \Psi(\mathbf{v})]\| \le \alpha\tau\|\mathbf{u} - \mathbf{v}\|.$$

Therefore, $\mathbf{G}$ has a unique fixed point $\mathbf{u}_*$ in $\mathcal{U}_{\delta_0}$. It then follows from Theorem 2.1 that the system of nonlinear equations (1.1) has a unique solution $\mathbf{u}_*$ in $\mathcal{U}_{\delta_0}$. ∎

COROLLARY 2.4. *Let the assumptions of Theorem* 2.3 *be satisfied. Assume* $\alpha\beta < \frac{1}{4}$. *If* $\Psi(\mathbf{u}) = (u_1^2, u_2^2, \ldots, u_n^2)^T$, *with* $\mathbf{u} = (u_1, u_2, \ldots, u_n)^T$, *then*

$$\alpha\tau \le 2\alpha\delta_0 = 1 - \sqrt{1 - 4\alpha\beta} < 1.$$

Therefore, the system of nonlinear equations

$$B\mathbf{u} + C\mathbf{u} \odot \mathbf{u} = \mathbf{b}$$

has a unique solution $\mathbf{u}_*$ *in* $\mathcal{U}_{\delta_0}$.

3. The preconditioners. In this section, we will solve the system of nonlinear equations

$$B\mathbf{u} + C\mathbf{u} \odot \mathbf{u} = \mathbf{b} \tag{3.1}$$

by the Newton or the fixed-point iteration method.

3.1. The Newton and fixed-point iterations. Let $\Omega \in \mathbb{R}^{n\times n}$ be a diagonal matrix such that $B + C\Omega$ is nonsingular. Then we consider the following Newton iteration for solving the system of nonlinear equations (3.1):

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} - (B + 2C\,\mathrm{diag}(\mathbf{u}^k))^{-1}(B\mathbf{u}^{(k)} + C\mathbf{u}^{(k)} \odot \mathbf{u}^{(k)} - \mathbf{b}), \quad k = 0, 1, 2, \ldots.$$

If $\mathbf{u}_* \in \mathcal{U}_{\delta_0}$ is a solution of the nonlinear system (3.1), then it holds that $\{\mathbf{u}^{(k)}\} \subset \mathcal{U}_{\delta_0}$ and $\lim_{k\to\infty} \mathbf{u}^{(k)} = \mathbf{u}_*$, with the convergence speed being at least quadratic, provided $\mathbf{u}^{(0)}$ is in a neighborhood of $\mathbf{u}_*$.

The fixed-point iteration for the system of nonlinear equations (3.1) is as follows:

$$(B + C\Omega)\mathbf{u}^{(k+1)} = C[\Omega\mathbf{u}^{(k)} - \mathbf{u}^{(k)} \odot \mathbf{u}^{(k)}] + \mathbf{b}, \quad k = 0, 1, 2, \ldots,$$

or equivalently,

$$\mathbf{u}^{(k+1)} = (B + C\Omega)^{-1}C[\Omega\mathbf{u}^{(k)} - \mathbf{u}^{(k)} \odot \mathbf{u}^{(k)}] + (B + C\Omega)^{-1}\mathbf{b}, \quad k = 0, 1, 2, \ldots.$$

If $\mathbf{u}_* \in \mathcal{U}_{\delta_0}$ is a solution of the nonlinear system (3.1), then it holds that $\{\mathbf{u}^{(k)}\} \subset \mathcal{U}_{\delta_0}$ and $\lim_{k\to\infty} \mathbf{u}^{(k)} = \mathbf{u}_*$, provided $\mathbf{u}^{(0)}$ is in a neighborhood of $\mathbf{u}_*$ and

$$\rho((B + C\Omega)^{-1}C(\Omega - 2\,\mathrm{diag}(\mathbf{u}_*))) < 1,$$

where $\mathrm{diag}(\mathbf{u}_*)$ is the diagonal matrix with its ith diagonal entries being $[\mathbf{u}_*]_i$, and $\rho(\cdot)$ denotes the spectral radius of the corresponding matrix. See [19] for details.

3.2. Construction of the preconditioners. For both fixed-point and Newton iterations, at each step we need to solve a subsystem of linear equations of the form

$$A\mathbf{z} \equiv (B + CD)\mathbf{z} = \mathbf{r}. \tag{3.2}$$

This kind of linear system may be efficiently solved by the preconditioned GMRES method. The key point here is constructing a high-quality preconditioner M for the coefficient matrix A by making use of its concrete properties and special structure.

Consider the system of nonlinear equations (1.1) with the function $\Psi(\mathbf{u})$ as given in (1.4) and the matrices B and C as given in (1.5) and (1.6), respectively, where $T_z^{(i)}$ and $D_z^{(i)}$ ($i = 1, 2, 3, 4$ and $z \in \{x, t\}$), and Q_z ($z \in \{x, t\}$) are defined as in (1.7)–(1.13). Let $\Omega = \mathrm{diag}(\omega_{\mathbf{i}})$ be a positive definite matrix such that $D := I \otimes \Omega$ is an approximation to the Jacobian of $\Psi(\mathbf{u})$, i.e., $D\mathbf{u} \approx \mathbf{\Psi}(\mathbf{u})$. Then the target matrix under consideration is

$$\begin{aligned} A &= B + CD \\ &= \varepsilon(T_x^{(2)} + D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)} + D_x^{(2)}) \otimes Q_t \\ &\quad + Q_x \otimes (D_t^{(3)}T_t^{(1)} + T_t^{(1)}D_t^{(3)} + D_t^{(4)}) \\ &\quad + (D_x^{(3)}T_x^{(1)} + T_x^{(1)}D_x^{(3)} + D_x^{(4)}) \otimes (Q_t\Omega). \end{aligned} \tag{3.3}$$

By utilizing the special structure of the matrix A, we can construct its preconditioner M as

$$\begin{aligned} M &= \hat{B} + \hat{C}D \\ &= \varepsilon(B_x^{(2)} + D_x^{(1)}B_x^{(1)} + B_x^{(1)}D_x^{(1)} + D_x^{(2)}) \otimes Q_t \\ &\quad + Q_x \otimes (D_t^{(3)}B_t^{(1)} + B_t^{(1)}D_t^{(3)} + D_t^{(4)}) \\ &\quad + (D_x^{(3)}B_x^{(1)} + B_x^{(1)}D_x^{(3)} + D_x^{(4)}) \otimes (Q_t\Omega), \end{aligned} \tag{3.4}$$

where

$$\hat{B} = (\varepsilon(B_x^{(2)} + D_x^{(1)}B_x^{(1)} + B_x^{(1)}D_x^{(1)} + D_x^{(2)})) \otimes Q_t + Q_x \otimes (D_t^{(3)}B_t^{(1)} + B_t^{(1)}D_t^{(3)} + D_t^{(4)})$$

and

$$\hat{C} = (D_x^{(3)}B_x^{(1)} + B_x^{(1)}D_x^{(3)} + D_x^{(4)}) \otimes Q_t,$$

and, for $z \in \{x, t\}$,

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

are tridiagonal approximations to $T_z^{(1)}$ and $T_z^{(2)}$, respectively.

We remark that the preconditioner M is a block tridiagonal matrix, and is usually of mild size because, compared with the finite-difference system, the Sinc–Galerkin system need not be very large in order to achieve the same discretization accuracy [16, 18]. Therefore, for any given vector $\mathbf{r}$, the generalized residual equation $M\mathbf{w} = \mathbf{r}$ involved in the preconditioned GMRES iteration method can be solved in $\mathcal{O}(N_x N_t^2)$ or $\mathcal{O}(N_x^2 N_t)$ operations by using a variety of linear solvers such as the sparse direct methods, where $N_z = m_z + n_z + 1$, with $z \in \{x, t\}$; see also [6, 5] and the references therein.

3.3. Analysis of the preconditioning matrix. In this subsection, we will demonstrate the positive definiteness of the matrix A defined in (3.3) and its preconditioning matrix M defined in (3.4). In addition, we will derive precise bounds for the eigenvalues of the preconditioned matrix $M^{-1}A$ by making use of the generalized Bendixson theorem established in [7].

To this end, in what follows we use $(\cdot)^*$ to denote the conjugate transpose of either a vector or a square matrix. For a given square matrix X, we use $\mathcal{H}(X)$ and $\mathcal{S}(X)$ to denote, respectively, its Hermitian and skew-Hermitian parts [6], and use $\lambda(X)$ to denote its spectral set. In particular, when X is Hermitian or real symmetric, we use $\lambda_{\max}(X)$ to represent its largest eigenvalue.

THEOREM 3.1. *Assume that $D_z^{(i)}$ $(i = 2, 4$ and $z \in \{x, t\})$ are positive semidefinite diagonal matrices. Then, both $\mathcal{H}(A)$ and $\mathcal{H}(M)$ are symmetric positive definite matrices.*[1] *Hence, A and M are positive definite and, thus, are nonsingular.*

Proof. The Hermitian and the skew-Hermitian parts of A and M are

$$\mathcal{H}(A) = \frac{1}{2}(A + A^*) = \varepsilon(T_x^{(2)} + D_x^{(2)}) \otimes Q_t + Q_x \otimes D_t^{(4)} + D_x^{(4)} \otimes (Q_t\Omega),$$

$$\mathcal{S}(A) = \frac{1}{2}(A - A^*) = \varepsilon(D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)}) \otimes Q_t + Q_x \otimes (D_t^{(3)}T_t^{(1)} + T_t^{(1)}D_t^{(3)}) + (D_x^{(3)}T_x^{(1)} + T_x^{(1)}D_x^{(3)}) \otimes (Q_t\Omega),$$

[1] A matrix is positive definite if its Hermitian part is positive definite. Note that a positive definite matrix is not necessarily Hermitian; see [6, 5].

and

$$\begin{aligned}\mathcal{H}(M) &= \frac{1}{2}(M + M^*)\\ &= \varepsilon(B_x^{(2)} + D_x^{(2)}) \otimes Q_t + Q_x \otimes D_t^{(4)} + D_x^{(4)} \otimes (Q_t\Omega),\end{aligned}$$

$$\begin{aligned}\mathcal{S}(M) &= \frac{1}{2}(M - M^*)\\ &= \varepsilon(D_x^{(1)}B_x^{(1)} + B_x^{(1)}D_x^{(1)}) \otimes Q_t + Q_x \otimes (D_t^{(3)}B_t^{(1)} + B_t^{(1)}D_t^{(3)})\\ &\quad +(D_x^{(3)}B_x^{(1)} + B_x^{(1)}D_x^{(3)}) \otimes (Q_t\Omega).\end{aligned}$$

Because the diagonal matrices $D_z^{(i)}$ ($i = 2, 4$ and $z \in \{x, t\}$) are positive semidefinite, and the Toeplitz matrices $T_x^{(2)}$ and the diagonal matrices Q_z ($z \in \{x, t\}$) are symmetric positive definite [12], we know that $\mathcal{H}(A)$ is symmetric positive definite. Therefore, A is a positive definite matrix and, thus, is nonsingular.

By applying the same arguments to the preconditioning matrix M, we can immediately show that M is positive definite and nonsingular, too. ∎

The following *generalized Bendixson theorem*, established in [7], is essential for us to derive a rectangular domain for bounding the eigenvalues of the preconditioned matrix $M^{-1}A$.

THEOREM 3.2 ([7, Theorem 2.4]). *Let $A, M \in C^{n\times n}$ be n-by-n complex matrices and, for all $v \in C^n \setminus \{0\}$, $v^*\mathcal{H}(A)v \neq 0$ and $v^*\mathcal{H}(M)v \neq 0$ hold. Let the functions $h(v)$, $f_A(v)$, and $f_M(v)$ be defined as*

$$h(v) = \frac{v^*\mathcal{H}(A)v}{v^*\mathcal{H}(M)v}, \quad f_A(v) = \frac{1}{\imath} \cdot \frac{v^*\mathcal{S}(A)v}{v^*\mathcal{H}(A)v}, \quad \text{and} \quad f_M(v) = \frac{1}{\imath} \cdot \frac{v^*\mathcal{S}(M)v}{v^*\mathcal{H}(M)v}.$$

Assume that there exist positive constants γ_1 and γ_2 such that

$$\gamma_1 \le h(v) \le \gamma_2 \quad \forall\ v \in C^n \setminus \{0\},$$

and nonnegative constants η and μ such that

$$-\eta \le f_A(v) \le \eta \quad \text{and} \quad -\mu \le f_M(v) \le \mu \ \ \forall\ v \in C^n \setminus \{0\}.$$

Then, when $\eta\mu \le 1$, we have

$$\begin{cases} \frac{(1-\eta\mu)\gamma_1}{1+\mu^2} \le \ \ \mathrm{Re}(\lambda(M^{-1}A)) \le (1+\eta\mu)\gamma_2,\\ -(\eta+\mu)\gamma_2 \le \ \mathrm{Im}(\lambda(M^{-1}A)) \le (\eta+\mu)\gamma_2. \end{cases}$$

Here, $\mathrm{Re}(\cdot)$ *and* $\mathrm{Im}(\cdot)$ *represent the real and imaginary parts of the corresponding complex, respectively.*

In [7] the following estimates about the bounds of the function $h(v)$ have been derived.

LEMMA 3.3 (see [7, Lemma 4.2]). *Assume that $D_z^{(i)}(i = 2, 4$ and $z \in \{x, t\})$ are positive semidefinite diagonal matrices. Then*

$$1 \le \frac{v^*\mathcal{H}(A)v}{v^*\mathcal{H}(M)v} \le \frac{\pi^2}{4} \quad \forall v \in \mathbb{C}^n \setminus \{0\}.$$

To obtain the bounds for the functions $f_A(v)$ and $f_M(v)$, we need to establish the following estimates.

LEMMA 3.4. *Assume that* $D_z^{(i)}(i=2,4$ *and* $z\in\{x,t\}), Q_z(z\in\{x,t\})$, *and* Ω *are positive definite diagonal matrices. For* $z\in\{x,t\}$, *let* $N_z=m_z+n_z+1$, *and define*

$$d_z^{(1)}=\max_{1\le j\le N_z}\{|[D_z^{(1)}]_{jj}|,\ |[D_z^{(3)}]_{jj}|\},\quad d_z^{(2)}=\min_{1\le j\le N_z}\{|[D_z^{(2)}]_{jj}|,\ |[D_z^{(4)}]_{jj}|\},$$

$$\mu^{(1)}=\max_{z\in\{x,t\}}\left\{\frac{2\pi d_z^{(1)}}{\sqrt{(\pi^2+d_z^{(2)})d_z^{(2)}}},\quad \frac{2\pi d_z^{(1)}}{d_z^{(2)}}\right\},$$

and

$$\mu^{(2)}=\max_{z\in\{x,t\}}\left\{d_z^{(1)}\left(\sqrt{1+\frac{4}{d_z^{(2)}}}-1\right),\quad \frac{2d_z^{(1)}}{d_z^{(2)}}\right\}.$$

Then, for all $v\in\mathbb{C}^n\setminus\{0\}$, *it holds that*

$$\left|\frac{v^*[(D_x^{(1)}T_x^{(1)}+T_x^{(1)}D_x^{(1)})\otimes Q_t]v}{v^*[(T_x^{(2)}+D_x^{(2)})\otimes Q_t]v}\right|\le\mu^{(1)},\tag{3.5}$$

$$\left|\frac{v^*[(D_x^{(1)}B_x^{(1)}+B_x^{(1)}D_x^{(1)})\otimes Q_t]v}{v^*[(B_x^{(2)}+D_x^{(2)})\otimes Q_t]v}\right|\le\mu^{(2)},\tag{3.6}$$

$$\left|\frac{v^*[Q_x\otimes(D_t^{(3)}T_t^{(1)}+T_t^{(1)}D_t^{(3)})]v}{v^*[Q_x\otimes D_t^{(4)}]v}\right|\le\mu^{(1)},\tag{3.7}$$

$$\left|\frac{v^*[Q_x\otimes(D_t^{(3)}B_t^{(1)}+B_t^{(1)}D_t^{(3)})]v}{v^*[Q_x\otimes D_t^{(4)}]v}\right|\le\mu^{(2)},\tag{3.8}$$

$$\left|\frac{v^*[(D_x^{(3)}T_x^{(1)}+T_x^{(1)}D_x^{(3)})\otimes(Q_t\Omega)]v}{v^*[D_x^{(4)}\otimes(Q_t\Omega)]v}\right|\le\mu^{(1)},\tag{3.9}$$

and

$$\left|\frac{v^*[(D_x^{(3)}B_x^{(1)}+B_x^{(1)}D_x^{(3)})\otimes(Q_t\Omega)]v}{v^*[D_x^{(4)}\otimes(Q_t\Omega)]v}\right|\le\mu^{(2)}.\tag{3.10}$$

Proof. We first prove (3.5) and (3.6). Because $D_x^{(1)}T_x^{(1)}+T_x^{(1)}D_x^{(1)}$ is a skew-symmetric matrix and $T_x^{(2)}+D_x^{(2)}$ is a symmetric positive definite matrix, for all

$v \neq 0$, by direct computations we can obtain

$$\begin{aligned}
&\left| \frac{v^*[(D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)}) \otimes Q_t]v}{v^*[(T_x^{(2)} + D_x^{(2)}) \otimes Q_t]v} \right| \\
&\quad \leq \max_{v \neq 0} \left\{ \left| \frac{\frac{1}{\imath} v^*[(D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)}) \otimes Q_t]v}{v^*[(T_x^{(2)} + D_x^{(2)}) \otimes Q_t]v} \right| \right\} \\
&\quad \leq \left| \lambda_{\max} \left(\frac{1}{\imath} (T_x^{(2)} + D_x^{(2)})^{-1/2} (D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)}) (T_x^{(2)} + D_x^{(2)})^{-1/2} \right) \right| \\
&\quad \leq \left\| (T_x^{(2)} + D_x^{(2)})^{-1/2} (D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)}) (T_x^{(2)} + D_x^{(2)})^{-1/2} \right\|_2 \\
&\quad \leq \left\| (T_x^{(2)} + D_x^{(2)})^{-1/2} D_x^{(1)}T_x^{(1)} (T_x^{(2)} + D_x^{(2)})^{-1/2} \right\|_2 \\
&\qquad + \left\| (T_x^{(2)} + D_x^{(2)})^{-1/2} T_x^{(1)}D_x^{(1)} (T_x^{(2)} + D_x^{(2)})^{-1/2} \right\|_2 \\
&\quad \leq \left\| (T_x^{(2)} + D_x^{(2)})^{-1/2} (T_x^{(2)} + d_x^{(2)}I)^{1/2} \right\|_2 \cdot \left\| (T_x^{(2)} + d_x^{(2)}I)^{-1/2} \right\|_2 \\
&\qquad \cdot \left\| D_x^{(1)} \right\|_2 \cdot \left\| (T_x^{(2)} + d_x^{(2)}I)^{1/2} (T_x^{(2)} + D_x^{(2)})^{-1/2} \right\|_2 \\
&\qquad \cdot \left[\left\| T_x^{(1)} (T_x^{(2)} + d_x^{(2)}I)^{-1/2} \right\|_2 + \left\| (T_x^{(2)} + d_x^{(2)}I)^{-1/2} T_x^{(1)} \right\|_2 \right] \\
&\quad \leq \left\| (T_x^{(2)} + d_x^{(2)}I)^{-1/2} \right\|_2 \cdot \left\| D_x^{(1)} \right\|_2 \\
&\qquad \cdot \left[\left\| T_x^{(1)} (T_x^{(2)} + d_x^{(2)}I)^{-1/2} \right\|_2 + \left\| (T_x^{(2)} + d_x^{(2)}I)^{-1/2} T_x^{(1)} \right\|_2 \right]. \qquad (3.11)
\end{aligned}$$

Here, we have applied the fact

$$\begin{aligned}
&\left\| (T_x^{(2)} + D_x^{(2)})^{-1/2} (T_x^{(2)} + d_x^{(2)}I)^{1/2} \right\|_2^2 \\
&\quad = \left\| (T_x^{(2)} + d_x^{(2)}I)^{1/2} (T_x^{(2)} + D_x^{(2)})^{-1/2} \right\|_2^2 \\
&\quad = \lambda_{\max} \left((T_x^{(2)} + D_x^{(2)})^{-1/2} (T_x^{(2)} + d_x^{(2)}I) (T_x^{(2)} + D_x^{(2)})^{-1/2} \right) \\
&\quad = \max_{w \neq 0} \left\{ \frac{w^*(T_x^{(2)} + D_x^{(2)})^{-1/2} (T_x^{(2)} + d_x^{(2)}I) (T_x^{(2)} + D_x^{(2)})^{-1/2} w}{w^*w} \right\} \\
&\quad = \max_{w \neq 0} \left\{ \frac{w^*(T_x^{(2)} + d_x^{(2)}I)w}{w^*(T_x^{(2)} + D_x^{(2)})w} \right\} \\
&\quad \leq \max \left\{ 1, \quad d_x^{(2)} \cdot \max_{w \neq 0} \frac{w^*w}{w^*D_x^{(2)}w} \right\} \\
&\quad \leq \ 1.
\end{aligned}$$

Now, we estimate the terms $\|(T_x^{(2)} + d_x^{(2)}I)^{-1/2}\|_2$, $\|T_x^{(1)}(T_x^{(2)} + d_x^{(2)}I)^{-1/2}\|_2$ and $\|(T_x^{(2)} + d_x^{(2)}I)^{-1/2}T_x^{(1)}\|_2$ involved in (3.11).

As the generating function of the Toeplitz matrix $T_x^{(2)}$ is θ^2, we have

$$\begin{aligned}\left\|(T_x^{(2)}+d_x^{(2)}I)^{-1/2}\right\|_2^2 &= \lambda_{\max}\left((T_x^{(2)}+d_x^{(2)}I)^{-1}\right)\\ &= \max_{w\neq 0}\left\{\frac{w^*(T_x^{(2)}+d_x^{(2)}I)^{-1}w}{w^*w}\right\}\\ &< \max_{-\pi\le\theta\le\pi}\left\{\frac{1}{\theta^2+d_x^{(2)}}\right\}\\ &= \frac{1}{d_x^{(2)}}.\end{aligned}$$

Therefore,

$$\left\|(T_x^{(2)}+d_x^{(2)}I)^{-1/2}\right\|_2 < \frac{1}{\sqrt{d_x^{(2)}}}.$$

In addition, we have

$$\begin{aligned}&\left\|T_x^{(1)}(T_x^{(2)}+d_x^{(2)}I)^{-1/2}\right\|_2^2\\ &= \left\|(T_x^{(2)}+d_x^{(2)}I)^{-1/2}T_x^{(1)}\right\|_2^2\\ &= \lambda_{\max}\left((T_x^{(2)}+d_x^{(2)}I)^{-1/2}T_x^{(1)}(T_x^{(1)})^*(T_x^{(2)}+d_x^{(2)}I)^{-1/2}\right)\\ &= \max_{w\neq 0}\left\{\frac{w^*(T_x^{(2)}+d_x^{(2)}I)^{-1/2}T_x^{(1)}(T_x^{(1)})^*(T_x^{(2)}+d_x^{(2)}I)^{-1/2}w}{w^*w}\right\}\\ &= \max_{w\neq 0}\left\{\frac{w^*T_x^{(1)}(T_x^{(1)})^*w}{w^*(T_x^{(2)}+d_x^{(2)}I)w}\right\}.\end{aligned}$$

By making use of Theorems 3.1 and 3.3 in [11], we know that for any $\epsilon>0$ there exist a positive semidefinite matrix R_x of fixed rank and a matrix E_x of small norm such that $\|E_x\|<d_x^{(2)}\epsilon$, and

$$T_x^{(1)}(T_x^{(1)})^*+R_x+E_x=T_x^{(3)},$$

where $T_x^{(3)}$ is the Toeplitz matrix generated by the positive function θ^2. We remark that the generating function of $T_x^{(1)}$ is $\imath\theta$ and $T_x^{(1)}$ is a skew-symmetric matrix. Because

$$\frac{w^*R_xw}{w^*(T_x^{(2)}+d_x^{(2)}I)w}\ge 0 \quad\text{and}\quad \left|\frac{w^*E_xw}{w^*(T_x^{(2)}+d_x^{(2)}I)w}\right|\le\epsilon \quad \forall w\neq 0,$$

it follows from the above matrix decomposition that

$$\max_{w\neq 0}\left\{\frac{w^*T_x^{(1)}(T_x^{(1)})^*w}{w^*(T_x^{(2)}+d_x^{(2)}I)w}\right\} < \max_{w\neq 0}\left\{\frac{w^*T_x^{(3)}w}{w^*(T_x^{(2)}+d_x^{(2)}I)w}\right\}+\epsilon.$$

Since ϵ can be arbitrarily small, it holds that

$$\begin{aligned}\max_{w\neq 0}\left\{\frac{w^*T_x^{(1)}(T_x^{(1)})^*w}{w^*(T_x^{(2)}+d_x^{(2)}I)w}\right\} &\le \max_{w\neq 0}\left\{\frac{w^*T_x^{(3)}w}{w^*(T_x^{(2)}+d_x^{(2)}I)w}\right\}\\ &< \max_{-\pi\le\theta\le\pi}\left\{\frac{\theta^2}{\theta^2+d_x^{(2)}}\right\}\\ &= \frac{\pi^2}{\pi^2+d_x^{(2)}}. \end{aligned}\tag{3.12}$$

Therefore,

$$\left\|T_x^{(1)}(T_x^{(2)}+d_x^{(2)}I)^{-1/2}\right\|_2 = \left\|(T_x^{(2)}+d_x^{(2)}I)^{-1/2}T_x^{(1)}\right\|_2 < \frac{\pi}{\sqrt{\pi^2+d_x^{(2)}}}.$$

Based upon (3.11) we immediately have

$$\begin{aligned}\left|\frac{v^*[(D_x^{(1)}T_x^{(1)}+T_x^{(1)}D_x^{(1)})\otimes Q_t]v}{v^*[(T_x^{(2)}+D_x^{(2)})\otimes Q_t]v}\right| &\le \frac{1}{\sqrt{d_x^{(2)}}}\cdot d_x^{(1)}\cdot\frac{2\pi}{\sqrt{\pi^2+d_x^{(2)}}}\\ &= \frac{2\pi d_x^{(1)}}{\sqrt{(\pi^2+d_x^{(2)})d_x^{(2)}}}\\ &\le \mu^{(1)}.\end{aligned}$$

This shows that (3.5) holds true.

Analogously to (3.11), for all $v\neq 0$ we can obtain

$$\begin{aligned}&\left|\frac{v^*[(D_x^{(1)}B_x^{(1)}+B_x^{(1)}D_x^{(1)})\otimes Q_t]v}{v^*[(B_x^{(2)}+D_x^{(2)})\otimes Q_t]v}\right|\\ &\quad\le \left\|(B_x^{(2)}+d_x^{(2)}I)^{-1/2}\right\|_2\cdot\left\|D_x^{(1)}\right\|_2\\ &\quad\quad\cdot\left[\left\|B_x^{(1)}(B_x^{(2)}+d_x^{(2)}I)^{-1/2}\right\|_2+\left\|(B_x^{(2)}+d_x^{(2)}I)^{-1/2}B_x^{(1)}\right\|_2\right].\end{aligned}\tag{3.13}$$

Here, we have applied the fact

$$\left\|(B_x^{(2)}+D_x^{(2)})^{-1/2}(B_x^{(2)}+d_x^{(2)}I)^{1/2}\right\|_2^2 = \left\|(B_x^{(2)}+d_x^{(2)}I)^{1/2}(B_x^{(2)}+D_x^{(2)})^{-1/2}\right\|_2^2 \le 1.$$

Now, we estimate the terms $\|(B_x^{(2)}+d_x^{(2)}I)^{-1/2}\|_2$, $\|B_x^{(1)}(B_x^{(2)}+d_x^{(2)}I)^{-1/2}\|_2$, and $\|(B_x^{(2)}+d_x^{(2)}I)^{-1/2}B_x^{(1)}\|_2$ involved in (3.13).

As the generating function of the Toeplitz matrix $B_x^{(2)}$ is $2-2\cos\theta$, we have

$$\begin{aligned}\left\|(B_x^{(2)}+d_x^{(2)}I)^{-1/2}\right\|_2^2 &= \lambda_{\max}\left((B_x^{(2)}+d_x^{(2)}I)^{-1}\right)\\ &= \max_{w\neq 0}\left\{\frac{w^*(B_x^{(2)}+d_x^{(2)}I)^{-1}w}{w^*w}\right\}\\ &< \max_{-\pi\le\theta\le\pi}\left\{\frac{1}{2(1-\cos\theta)+d_x^{(2)}}\right\}\\ &= \frac{1}{d_x^{(2)}}.\end{aligned}$$

Therefore,

$$\left\|(B_x^{(2)} + d_x^{(2)} I)^{-1/2}\right\|_2 < \frac{1}{\sqrt{d_x^{(2)}}}.$$

By making use of Theorem 3.1 in [11] again, we know that there exists a positive semidefinite matrix R_x of fixed rank such that

$$B_x^{(1)}(B_x^{(1)})^* + R_x = B_x^{(3)},$$

where $B_x^{(3)}$ is the Toeplitz matrix generated by the positive function $\sin^2\theta$. We remark that the generating function of $B_x^{(1)}$ is $\imath \sin\theta$ and $B_x^{(1)}$ is a skew-symmetric matrix. Because

$$\frac{w^* R_x w}{w^*(B_x^{(2)} + d_x^{(2)} I)w} \geq 0 \quad \forall w \neq 0,$$

it follows from the above matrix decomposition that

$$\begin{aligned}
&\left\|B_x^{(1)}(B_x^{(2)} + d_x^{(2)} I)^{-1/2}\right\|_2^2 \\
&\quad = \left\|(B_x^{(2)} + d_x^{(2)} I)^{-1/2} B_x^{(1)}\right\|_2^2 \\
&\quad = \lambda_{\max}\left((B_x^{(2)} + d_x^{(2)} I)^{-1/2} B_x^{(1)}(B_x^{(1)})^*(B_x^{(2)} + d_x^{(2)} I)^{-1/2}\right) \\
&\quad = \max_{w\neq 0}\left\{\frac{w^*(B_x^{(2)} + d_x^{(2)} I)^{-1/2} B_x^{(1)}(B_x^{(1)})^*(B_x^{(2)} + d_x^{(2)} I)^{-1/2} w}{w^* w}\right\} \\
&\quad = \max_{w\neq 0}\left\{\frac{w^* B_x^{(1)}(B_x^{(1)})^* w}{w^*(B_x^{(2)} + d_x^{(2)} I)w}\right\} \\
&\quad \leq \max_{w\neq 0}\left\{\frac{w^* B_x^{(3)} w}{w^*(B_x^{(2)} + d_x^{(2)} I)w}\right\} \\
&\quad < \max_{-\pi\leq\theta\leq\pi}\left\{\frac{\sin^2\theta}{2(1-\cos\theta) + d_x^{(2)}}\right\} \\
&\quad = 1 + \frac{d_x^{(2)}}{2} - \frac{1}{2}\sqrt{(d_x^{(2)} + 4) d_x^{(2)}} \\
&\quad = \frac{1}{4}\left(\sqrt{d_x^{(2)} + 4} - \sqrt{d_x^{(2)}}\right)^2. \qquad (3.14)
\end{aligned}$$

Therefore,

$$\left\|B_x^{(1)}(B_x^{(2)} + d_x^{(2)} I)^{-1/2}\right\|_2 = \left\|(B_x^{(2)} + d_x^{(2)} I)^{-1/2} B_x^{(1)}\right\|_2 < \frac{1}{2}\left(\sqrt{d_x^{(2)} + 4} - \sqrt{d_x^{(2)}}\right).$$

Based upon (3.13) we immediately have

$$\left|\frac{v^*[(D_x^{(1)}B_x^{(1)}+B_x^{(1)}D_x^{(1)})\otimes Q_t]v}{v^*[(B_x^{(2)}+D_x^{(2)})\otimes Q_t]v}\right| \le \frac{1}{\sqrt{d_x^{(2)}}}\cdot d_x^{(1)}\cdot\left(\sqrt{d_x^{(2)}+4}-\sqrt{d_x^{(2)}}\right)$$
$$= d_x^{(1)}\cdot\left(\sqrt{1+\frac{4}{d_x^{(2)}}}-1\right)$$
$$\le \mu^{(2)}.$$

This shows the validity of (3.6).

We remark that the last bound in (3.14) is obtained by using the maximum value of the one-variable function

$$\varphi(\theta)=\frac{\sin^2\theta}{2(1-\cos\theta)+d_x^{(2)}}$$

attained at the stationary point

$$\theta_*=\arccos\left(\frac{2+d_x^{(2)}-\sqrt{(d_x^{(2)})^2+4d_x^{(2)}}}{2}\right).$$

Now, we are going to prove (3.7)–(3.10). Analogously, for all $v\neq 0$, by straightforward computations we obtain

$$\left|\frac{v^*[Q_x\otimes(D_t^{(3)}T_t^{(1)}+T_t^{(1)}D_t^{(3)})]v}{v^*[Q_x\otimes D_t^{(4)}]v}\right|$$
$$\le \max_{v\neq 0}\left\{\left|\frac{\frac{1}{\imath}v^*[Q_x\otimes(D_t^{(3)}T_t^{(1)}+T_t^{(1)}D_t^{(3)})]v}{v^*[Q_x\otimes D_t^{(4)}]v}\right|\right\}$$
$$\le \left|\lambda_{\max}\left(\frac{1}{\imath}(D_t^{(4)})^{-\frac{1}{2}}(D_t^{(3)}T_t^{(1)}+T_t^{(1)}D_t^{(3)})(D_t^{(4)})^{-\frac{1}{2}}\right)\right|$$
$$\le \left\|(D_t^{(4)})^{-\frac{1}{2}}(D_t^{(3)}T_t^{(1)}+T_t^{(1)}D_t^{(3)})(D_t^{(4)})^{-\frac{1}{2}}\right\|_2$$
$$\le \left\|(D_t^{(4)})^{-\frac{1}{2}}(d_t^{(2)}I)^{\frac{1}{2}}\right\|_2\left\|(d_t^{(2)}I)^{-\frac{1}{2}}\right\|_2\left\|D_t^{(3)}\right\|_2\left\|(d_t^{(2)}I)^{\frac{1}{2}}(D_t^{(4)})^{-\frac{1}{2}}\right\|_2$$
$$\cdot\left[\left\|T_t^{(1)}(d_t^{(2)}I)^{-\frac{1}{2}}\right\|_2+\left\|(d_t^{(2)}I)^{-\frac{1}{2}}T_t^{(1)}\right\|_2\right]. \tag{3.15}$$

It is straightforward that

$$\left\|(D_t^{(4)})^{-\frac{1}{2}}(d_t^{(2)}I)^{\frac{1}{2}}\right\|_2\le 1 \quad\text{and}\quad \left\|(d_t^{(2)}I)^{\frac{1}{2}}(D_t^{(4)})^{-\frac{1}{2}}\right\|_2\le 1 \tag{3.16}$$

hold true, and it follows from

$$\begin{aligned}\left\|T_t^{(1)}(d_t^{(2)}I)^{-\frac{1}{2}}\right\|_2^2 &= \left\|(d_t^{(2)}I)^{-\frac{1}{2}}T_t^{(1)}\right\|_2^2\\ &\le \max_{v\neq 0}\left\{\frac{v^*T_t^{(1)}(T_t^{(1)})^*v}{v^*(d_t^{(2)}I)v}\right\}\\ &\le \max_{-\pi\le\theta\le\pi}\left\{\frac{\theta^2}{d_t^{(2)}}\right\}\\ &= \frac{\pi^2}{d_t^{(2)}}\end{aligned}$$

that

$$\left\|T_t^{(1)}(d_t^{(2)}I)^{-\frac{1}{2}}\right\|_2 = \left\|(d_t^{(2)}I)^{-\frac{1}{2}}T_t^{(1)}\right\|_2 < \frac{\pi}{\sqrt{d_t^{(2)}}};$$

see (3.12). By substituting these bounds into (3.15) we know that (3.7) holds true.

Similar arguments can lead to the estimate

$$\left|\frac{v^*[(D_x^{(3)}T_x^{(1)}+T_x^{(1)}D_x^{(3)})\otimes(Q_t\Omega)]v}{v^*[D_x^{(4)}\otimes(Q_t\Omega)]v}\right| \le \frac{2\pi d_x^{(1)}}{d_x^{(2)}} \le \mu^{(1)},$$

which is exactly (3.9).

Moreover, it follows from (3.14) and (3.16) that

$$\begin{aligned}&\left|\frac{v^*[Q_x\otimes(D_t^{(3)}B_t^{(1)}+B_t^{(1)}D_t^{(3)})]v}{v^*[Q_x\otimes D_t^{(4)}]v}\right|\\ &\le \left\|(D_t^{(4)})^{-\frac{1}{2}}(d_t^{(2)}I)^{\frac{1}{2}}\right\|_2\left\|(d_t^{(2)}I)^{-\frac{1}{2}}\right\|_2\left\|D_t^{(3)}\right\|_2\left\|(d_t^{(2)}I)^{\frac{1}{2}}(D_t^{(4)})^{-\frac{1}{2}}\right\|_2\\ &\quad\cdot\left[\left\|B_t^{(1)}(d_t^{(2)}I)^{-\frac{1}{2}}\right\|_2+\left\|(d_t^{(2)}I)^{-\frac{1}{2}}B_t^{(1)}\right\|_2\right]\\ &\le \frac{1}{\sqrt{d_t^{(2)}}}\cdot d_t^{(1)}\cdot 2\cdot\max_{-\pi\le\theta\le\pi}\left\{\frac{|\sin\theta|}{\sqrt{d_t^{(2)}}}\right\}\\ &\le \frac{2d_t^{(1)}}{d_t^{(2)}}\\ &\le \mu^{(2)}\end{aligned}$$

and

$$\left|\frac{v^*[(D_x^{(3)}B_x^{(1)}+B_x^{(1)}D_x^{(3)})\otimes(Q_t\Omega)]v}{v^*[D_x^{(4)}\otimes(Q_t\Omega)]v}\right| \le \frac{2d_x^{(1)}}{d_x^{(2)}} \le \mu^{(2)}$$

hold true, which are exactly (3.8) and (3.10), respectively. ∎

Lemma 3.4 immediately implies the following bounds about the functions $f_A(v)$ and $f_M(v)$ defined in Theorem 3.2.

THEOREM 3.5. *Assume that* $D_z^{(i)}(i = 2, 4$ *and* $z \in \{x, t\}), Q_z(z \in \{x, t\})$, *and* Ω *are positive definite diagonal matrices. Let the positive constants* $d_z^{(i)}(i = 1, 2$ *and* $z \in \{x, t\})$ *be defined as in Lemma* 3.4 *and the functions* $f_A(v)$ *and* $f_M(v)$ *be defined as in Theorem* 3.2. *Then it holds that*

$$|f_A(v)| \le \mu^{(1)} \quad \textit{and} \quad |f_M(v)| \le \mu^{(2)}.$$

Proof. Based on Lemmas 3.3 and 3.4, by straightforward computations we have

$$|f_A(v)| = \left|\frac{v^*\mathcal{S}(A)v}{v^*\mathcal{H}(A)v}\right| =$$

$$\left|\frac{v^*[\varepsilon(D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)}) \otimes Q_t + Q_x \otimes (D_t^{(3)}T_t^{(1)} + T_t^{(1)}D_t^{(3)}) + (D_x^{(3)}T_x^{(1)} + T_x^{(1)}D_x^{(3)}) \otimes (Q_t\Omega)]v}{v^*[\varepsilon(T_x^{(2)} + D_x^{(2)}) \otimes Q_t + Q_x \otimes D_t^{(4)} + D_x^{(4)} \otimes (Q_t\Omega)]v}\right|$$

$$\le \max_{v\neq 0}\left\{\left|\frac{v^*[\varepsilon(D_x^{(1)}T_x^{(1)} + T_x^{(1)}D_x^{(1)}) \otimes Q_t]v}{v^*[\varepsilon(T_x^{(2)} + D_x^{(2)}) \otimes Q_t]v}\right|, \left|\frac{v^*[Q_x \otimes (D_t^{(3)}T_t^{(1)} + T_t^{(1)}D_t^{(3)})]v}{v^*(Q_x \otimes D_t^{(4)})v}\right|,\right.$$

$$\left.\left|\frac{v^*[(D_x^{(3)}T_x^{(1)} + T_x^{(1)}D_x^{(3)}) \otimes (Q_t\Omega)]v}{v^*(D_x^{(4)} \otimes (Q_t\Omega))v}\right|\right\}$$

$$\le \mu^{(1)}$$

and, similarly,

$$|f_M(v)| = \left|\frac{v^*\mathcal{S}(M)v}{v^*\mathcal{H}(M)v}\right| \le \mu^{(2)}.$$

Here, we have used the inequality

$$\frac{\alpha_1 + \alpha_2 + \alpha_3}{\beta_1 + \beta_2 + \beta_3} \le \max_{1\le i\le 3}\left\{\frac{\alpha_i}{\beta_i}\right\},$$

with α_i and β_i, $i = 1, 2, 3$, being positive reals. ∎

By using Theorems 3.2 and 3.5, and Lemmas 3.3 and 3.4, we can straightforwardly obtain the main theorem of this paper.

THEOREM 3.6. *Assume that* $D_z^{(i)}(i = 2, 4$ *and* $z \in \{x, t\})$, $Q_z(z \in \{x, t\})$, *and* Ω *are positive definite diagonal matrices. Then it holds that*

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

and

$$-\frac{\pi^2(\mu^{(1)} + \mu^{(2)})}{4} \le \mathrm{Im}(\lambda(M^{-1}A)) \le \frac{\pi^2(\mu^{(1)} + \mu^{(2)})}{4}.$$

Based on Theorem 3.6, we can immediately obtain a theoretical estimate about the asymptotic convergence rate of the preconditioned GMRES method with the preconditioner M in (3.4) for solving the system of linear equations (3.2). Here, we should suitably scale the Burgers equation (1.3) and appropriately choose the weighting functions $\omega_x(x)$ and $\omega_t(t)$ and the conformal mappings $\phi_x(x)$ and $\phi_t(t)$ such that $\mu^{(1)}\mu^{(2)} < 1$. For details, we refer to [20, 7].

4. Numerical examples. In this section, we illustrate the effectiveness of the proposed preconditioner when it is used to precondition the linear subsystems involved in each step of the Newton or the fixed-point iteration for solving the system of nonlinear equations (1.1).

In our experiments, the initial guess is taken to be zero, and each iteration process is terminated when the current residual $r^{(j)}$ satisfies

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

The GMRES method is applied to solve the preconditioned linear subsystems of the form

$$M^{-1}A\mathbf{z} = M^{-1}\mathbf{r},$$

which forms the inner iteration process for solving the linear subsystems involved in each step of the Newton or the fixed-point method. Here, the stopping criterion for the preconditioned GMRES is that the relative reduction on the norm of the residual is less than 10^{-6}. Besides, all codes are written in MATLAB 7.01, and all experiments are tested on a PC with 0.99G memory.

The following two examples of the Burgers equation (1.3) are used to examine the numerical performance of our new preconditioner M defined in (3.4):

(i) The Burgers equation

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) + u(x,t)\dfrac{\partial u}{\partial x}(x,t) - \epsilon\dfrac{\partial^2 u}{\partial x^2}(x,t) \\ \quad = e^{-\pi^2 t}\sin(\pi x) \\ \qquad \cdot(1-\pi^2 t + \pi t^2 e^{-\pi^2 t}\cos(\pi x) + \epsilon\pi^2 t), & 0 < x < 1 \ \text{ and } \ t \geq 0, \\ u(0,t) = 0 \quad \text{and} \quad u(1,t) = 0, & t \geq 0, \\ u(x,0) = 0, & 0 \leq x \leq 1. \end{cases}$$

(ii) The Burgers equation

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) + u(x,t)\dfrac{\partial u}{\partial x}(x,t) - \epsilon\dfrac{\partial^2 u}{\partial x^2}(x,t) \\ \quad = e^{x-t}(1-t+t^2e^{x-t} - \epsilon t) \\ \qquad -e^{-t}(1-t)(1-x+ex) \\ \qquad -t^2e^{-2t}(e-1)(1-x+xe), & 0 < x < 1 \ \text{ and } \ t \geq 0, \\ u(0,t) = te^{-t} \quad \text{and} \quad u(1,t) = te^{1-t}, & t \geq 0, \\ u(x,0) = 0, & 0 \leq x \leq 1. \end{cases}$$

The conformal mappings are chosen as $\phi(z) = \ln(\frac{z}{1-z})$ and $\psi(z) = \ln(\sinh(z))$ so that their restrictions onto the real intervals $(0,1)$ and $(0,+\infty)$ are $\phi_x(x) := \phi(x) = \ln(\frac{x}{1-x})$ and $\phi_t(t) := \psi(t) = \ln(\sinh(t))$, which are used for the discretizations of the x and t variables, respectively. The weighting functions are chosen to be $\omega_x(x) = 1/\phi_x'(x)$ and $\omega_t(t) = 1/\phi_t'(t)$.

In the numerical tables, the symbol "I" means that no preconditioner is used when solving the linear subsystems involved in the nonlinear iterations, while "M" represents that the preconditioner M defined in (3.4) is used. We use "N" to denote the number of the Newton iteration steps, "F" the number of the fixed-point iteration steps, "G" the average number of GMRES iteration steps in each Newton or

fixed-point iteration, "CPU" the total computing times, "Se" the maximum absolute discretization error at the Sinc grid points, and "Ue" the maximum absolute discretization error on the corresponding uniform grid points, while we use "average Se" and "average Ue" to represent the average absolute errors at all the Sinc grid points and at all the uniform grid points, respectively. In addition, the symbol "$*$" is used to denote that the iteration does not satisfy the terminating criterion within 50 steps of the Newton or the fixed-point iteration, while "+" means that the inner iteration does not satisfy the GMRES terminating criterion within 1000 iteration steps.

With respect to different ϵ, Tables 4.1–4.6 list the numbers of iteration steps and the CPU times required for the convergence of the Newton iteration, and Tables 4.7–4.12 list those required for the convergence of the fixed-point iteration, respectively, when they are applied to solve the system of nonlinear equations (1.1) resulting from the Sinc–Galerkin discretization of example (i). Some errors for reflecting the accuracy of the computed solutions are also listed in these tables. From these tables, we can see that our new preconditioner can considerably improve the convergence properties of both iteration methods and greatly reduce the running times. Moreover, with increasing of the problem size m, the number of the Newton or the fixed-point iteration steps almost remains the same if the inner iteration solver GMRES is preconditioned by the new preconditioner. The GMRES cannot achieve the prescribed tolerance within 1000 iteration steps and, therefore, the Newton or the fixed-point iteration cannot achieve the prescribed tolerance within 50 iteration steps, if the inner iteration solver GMRES is employed without using a preconditioner. Therefore, the new preconditioning technique substantially improves the convergence behaviors of both Newton and fixed-point iterations and, consequently, leads to fast convergent methods for solving the Sinc–Galerkin nonlinear systems of the Burgers equations.

TABLE 4.1
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 1.0$, *and the Newton method is used.*

	I			M						
m	N	G	CPU	N	G	Se	Average Se	Ue	Average Ue	CPU
4	2	76	0.25	2	13	1.3×10^{-3}	3.8×10^{-4}	2.2×10^{-3}	2.4×10^{-4}	0.14
8	2	268	2.78	2	20	1.1×10^{-3}	1.7×10^{-4}	5.1×10^{-4}	2.5×10^{-5}	0.28
16	2	924	90.9	2	33	3.9×10^{-4}	3.7×10^{-5}	1.6×10^{-4}	4.5×10^{-6}	1.36
32	*	+	—	2	55	3.6×10^{-5}	2.5×10^{-6}	1.4×10^{-5}	3.0×10^{-7}	10.8

TABLE 4.2
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.1$, *and the Newton method is used.*

	I			M						
m	N	G	CPU	N	G	Se	Average Se	Ue	Average Ue	CPU
4	2	77	0.30	2	20	9.3×10^{-3}	2.1×10^{-3}	9.4×10^{-3}	6.8×10^{-4}	0.16
8	2	265	2.63	2	32	2.2×10^{-3}	3.7×10^{-4}	1.8×10^{-3}	1.5×10^{-4}	0.41
16	3	932	119.9	3	47	2.5×10^{-4}	3.3×10^{-5}	2.0×10^{-4}	1.7×10^{-5}	2.83
32	*	+	—	3	79	3.6×10^{-5}	3.3×10^{-6}	3.1×10^{-5}	2.4×10^{-6}	21.3

TABLE 4.3
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.01$, *and the Newton method is used.*

	I			M						
m	N	G	CPU	N	G	Se	Average Se	Ue	Average Ue	CPU
4	3	78	0.28	3	25	7.1×10^{-3}	1.8×10^{-3}	6.7×10^{-3}	1.2×10^{-3}	0.22
8	3	271	3.78	3	40	5.1×10^{-4}	1.0×10^{-4}	3.6×10^{-4}	8.2×10^{-5}	0.66
16	3	942	124.4	3	59	5.6×10^{-5}	7.8×10^{-6}	4.3×10^{-5}	4.61×10^{-6}	3.53
32	*	+	—	3	103	3.1×10^{-6}	3.4×10^{-7}	2.4×10^{-6}	2.3×10^{-7}	27.0

TABLE 4.4
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.001$, *and the Newton method is used.*

	I			M						
m	N	G	CPU	N	G	Se	Average Se	Ue	Average Ue	CPU
4	3	75	0.39	3	34	2.6×10^{-3}	7.2×10^{-4}	2.7×10^{-3}	5.7×10^{-4}	0.25
8	3	275	4.06	3	57	4.2×10^{-4}	8.6×10^{-5}	2.7×10^{-4}	6.1×10^{-5}	0.94
16	3	939	124.1	2	80	5.6×10^{-5}	6.4×10^{-6}	2.4×10^{-5}	3.2×10^{-6}	3.14
32	*	+	—	2	132	3.1×10^{-6}	3.0×10^{-7}	1.4×10^{-6}	1.5×10^{-7}	23.0

TABLE 4.5
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.0001$, *and the Newton method is used.*

	I			M						
m	N	G	CPU	N	G	Se	Average Se	Ue	Average Ue	CPU
4	3	78	0.53	3	34	2.4×10^{-3}	8.1×10^{-4}	2.7×10^{-3}	5.6×10^{-4}	0.27
8	2	275	2.77	2	56	4.6×10^{-4}	6.9×10^{-5}	2.1×10^{-4}	5.6×10^{-5}	0.59
16	2	916	79.9	2	63	5.6×10^{-5}	4.8×10^{-6}	1.0×10^{-5}	2.9×10^{-6}	3.13
32	*	+	—	2	146	3.1×10^{-6}	2.3×10^{-7}	4.8×10^{-6}	1.2×10^{-7}	25.4

TABLE 4.6
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.00001$, *and the Newton method is used.*

	I			M						
m	N	G	CPU	N	G	Se	Average Se	Ue	Average Ue	CPU
4	3	77	0.39	3	28	2.4×10^{-3}	7.6×10^{-4}	2.6×10^{-3}	5.5×10^{-4}	0.22
8	3	281	6.50	3	64	8.0×10^{-4}	1.0×10^{-4}	3.9×10^{-4}	6.0×10^{-5}	1.02
16	2	944	85.3	2	91	7.0×10^{-5}	4.1×10^{-6}	1.2×10^{-5}	2.8×10^{-6}	3.67
32	*	+	—	2	151	3.1×10^{-6}	1.9×10^{-7}	2.3×10^{-7}	1.2×10^{-7}	26.6

TABLE 4.7
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 1.0$, *and the fixed-point method is used.*

	I			M						
m	F	G	CPU	F	G	Se	Average Se	Ue	Average Ue	CPU
4	3	51	0.16	3	8	1.3×10^{-3}	3.8×10^{-4}	2.2×10^{-3}	2.4×10^{-4}	0.05
8	3	178	2.31	2	10	1.1×10^{-3}	1.7×10^{-4}	5.1×10^{-4}	2.5×10^{-5}	0.11
16	2	444	35.3	2	16	4.0×10^{-4}	3.7×10^{-5}	1.6×10^{-4}	4.5×10^{-6}	0.61
32	*	+	—	2	27	3.6×10^{-5}	2.5×10^{-6}	1.4×10^{-5}	3.0×10^{-7}	4.31

TABLE 4.8
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.1$, *and the fixed-point method is used.*

	I			M						
m	F	G	CPU	F	G	Se	Average Se	Ue	Average Ue	CPU
4	4	58	0.23	3	13	9.3×10^{-3}	2.1×10^{-3}	9.4×10^{-3}	6.8×10^{-4}	0.05
8	3	114	2.28	3	21	2.2×10^{-3}	3.7×10^{-4}	1.8×10^{-3}	1.5×10^{-4}	0.31
16	3	577	66.9	3	32	2.5×10^{-4}	3.3×10^{-5}	2.0×10^{-4}	1.7×10^{-5}	1.83
32	*	+	—	3	55	3.6×10^{-5}	3.3×10^{-6}	3.1×10^{-5}	2.4×10^{-6}	12.8

Figures 4.1 and 4.3 depict the spectral distributions of the preconditioned matrix $M^{-1}A$ for examples (i) and (ii), respectively, corresponding to the Newton iteration, while Figure 4.5 depicts those for example (ii) corresponding to the fixed-point iteration, when $\epsilon = 1.0$. These figures clearly show that the matrices without preconditioning are very ill conditioned and, therefore, the corresponding GMRES method

TABLE 4.9
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.01$, *and the fixed-point method is used.*

	I			M						
m	F	G	CPU	F	G	Se	Average Se	Ue	Average Ue	CPU
4	4	59	0.20	4	19	7.1×10^{-3}	1.8×10^{-3}	6.7×10^{-3}	1.2×10^{-3}	0.13
8	4	211	5.19	4	32	5.1×10^{-4}	1.0×10^{-4}	3.6×10^{-4}	8.3×10^{-5}	0.64
16	3	653	102.9	3	40	5.6×10^{-5}	7.8×10^{-6}	4.3×10^{-5}	4.6×10^{-6}	2.28
32	*	+	—	3	70	3.1×10^{-6}	3.4×10^{-7}	2.4×10^{-6}	2.3×10^{-7}	16.6

TABLE 4.10
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.001$, *and the fixed-point method is used.*

	I			M						
m	F	G	CPU	F	G	Se	Average Se	Ue	Average Ue	CPU
4	4	58	0.39	4	26	2.6×10^{-3}	7.2×10^{-4}	2.7×10^{-3}	5.7×10^{-4}	0.20
8	3	185	2.55	3	37	4.2×10^{-4}	8.6×10^{-5}	2.7×10^{-4}	6.1×10^{-5}	0.53
16	3	633	86.1	2	41	5.6×10^{-5}	6.4×10^{-6}	2.4×10^{-5}	3.2×10^{-6}	1.55
32	*	+	—	2	68	3.1×10^{-6}	3.0×10^{-7}	1.4×10^{-6}	1.5×10^{-7}	11.0

TABLE 4.11
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.0001$, *and the fixed-point method is used.*

	I			M						
m	F	G	CPU	F	G	Se	Average Se	Ue	Average Ue	CPU
4	4	58	0.58	4	26	2.4×10^{-3}	8.1×10^{-4}	2.7×10^{-3}	5.6×10^{-4}	0.17
8	3	189	2.69	3	39	4.6×10^{-4}	6.9×10^{-5}	2.1×10^{-4}	5.6×10^{-5}	0.64
16	3	625	80.0	2	43	5.6×10^{-5}	4.8×10^{-6}	1.0×10^{-5}	2.9×10^{-6}	1.55
32	*	+	—	2	76	3.1×10^{-6}	2.3×10^{-7}	4.8×10^{-7}	1.2×10^{-7}	12.4

TABLE 4.12
Results for example (i): $h_x = h_t = \pi/\sqrt{3m}$, $\epsilon = 0.00001$, *and the fixed-point method is used.*

	I			M						
m	F	G	CPU	F	G	Se	Average Se	Ue	Average Ue	CPU
4	4	58	0.41	4	20	2.4×10^{-3}	7.6×10^{-4}	2.6×10^{-3}	5.5×10^{-4}	0.14
8	4	215	4.30	3	43	7.9×10^{-4}	1.0×10^{-4}	3.8×10^{-4}	6.0×10^{-5}	0.66
16	3	647	91.3	2	47	7.0×10^{-5}	4.1×10^{-6}	1.2×10^{-5}	2.8×10^{-6}	1.70
32	*	+	—	2	78	3.1×10^{-6}	1.9×10^{-7}	2.3×10^{-7}	1.2×10^{-7}	13.1

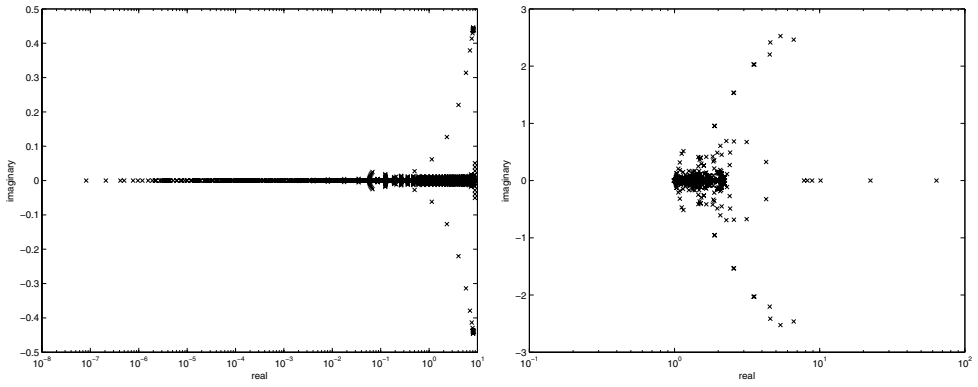

FIG. 4.1. *Spectra of the no-preconditioned (left) and preconditioned (right) matrices for example* (i), *with* $h_x = h_t = \pi/\sqrt{3m}$ *and* $m \equiv m_x = m_t = n_x = n_t = 16$, *when the Newton method is adopted.*

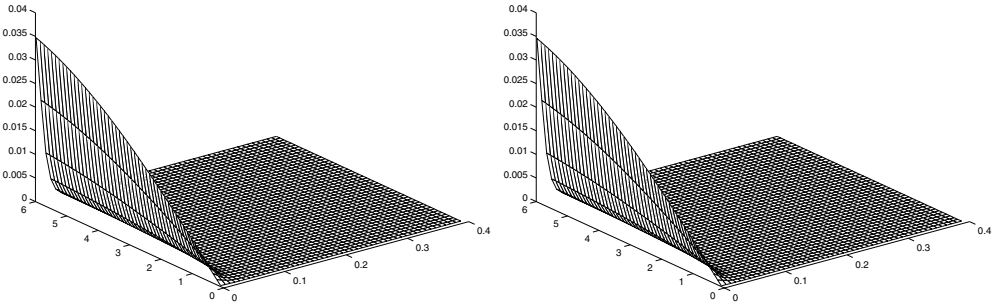

FIG. 4.2. *The exact (left) and computed (right) solutions of example* (i)*, with* $h_x = h_t = \pi/\sqrt{3m}$*,* $m \equiv m_x = m_t = n_x = n_t = 16$*, when the Newton method is adopted to obtain the computed solution.*

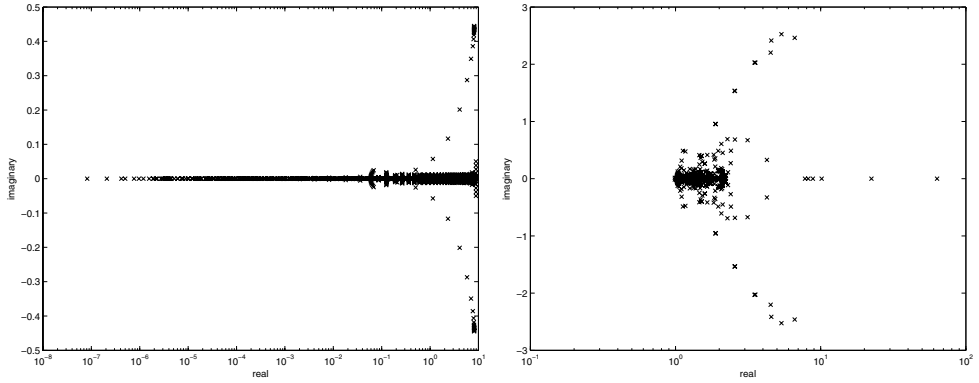

FIG. 4.3. *Spectra of the no-preconditioned (left) and preconditioned (right) matrices for example* (ii)*, with* $h_x = h_t = \pi/\sqrt{3m}$ *and* $m \equiv m_x = m_t = n_x = n_t = 16$*, when the Newton method is adopted.*

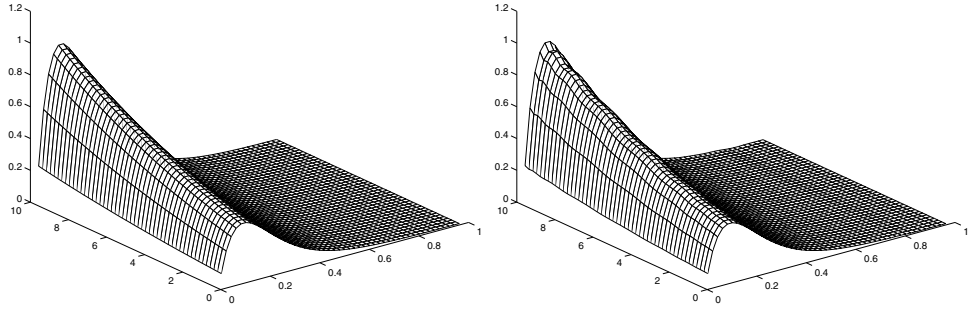

FIG. 4.4. *The exact (left) and computed (right) solutions of example* (ii)*, with* $h_x = h_t = \pi/\sqrt{3m}$*,* $m \equiv m_x = m_t = n_x = n_t = 16$*, when the Newton method is adopted to obtain the computed solution.*

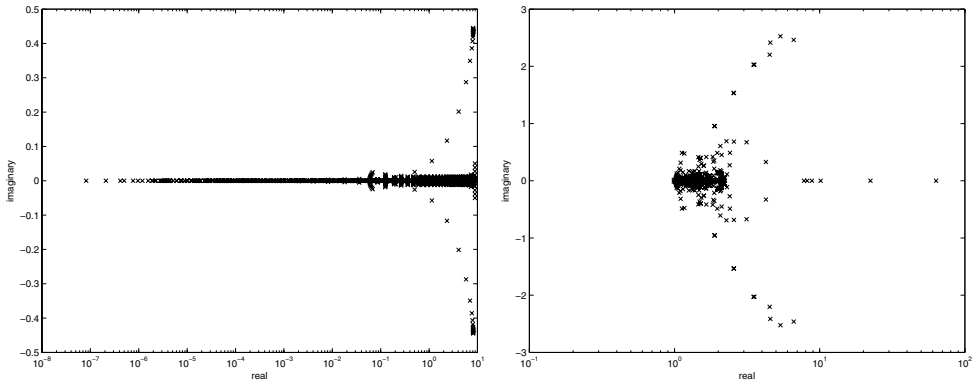

FIG. 4.5. *Spectra of the no-preconditioned (left) and preconditioned (right) matrices for example* (ii)*, with* $h_x = h_t = \pi/\sqrt{3m}$ *and* $m \equiv m_x = m_t = n_x = n_t = 16$*, when the fixed-point method is adopted.*

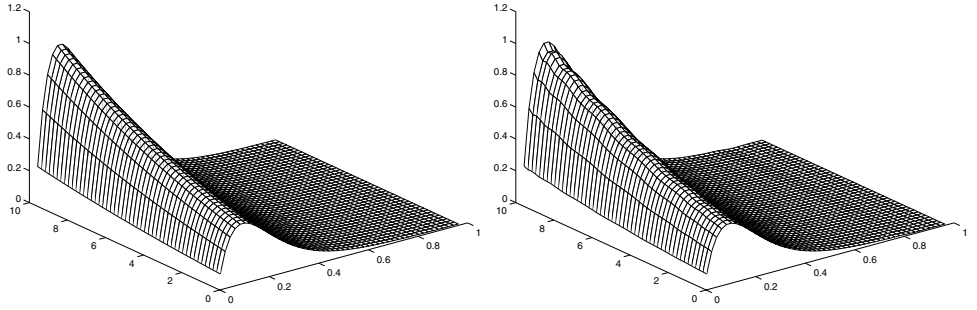

FIG. 4.6. *The exact (left) and computed (right) solutions of example* (i)*, with* $h_x = h_t = \pi/\sqrt{3m}$*,* $m \equiv m_x = m_t = n_x = n_t = 16$*, when the fixed-point method is adopted to obtain the computed solution.*

may be very slowly convergent or even divergent, while the matrices with preconditioning have tightly clustered eigenvalues and, thus, are well conditioned; as a result, the corresponding preconditioned GMRES method may converge quickly to the exact solutions of the subsystems of linear equations and, consequently, may lead to a fast convergent Newton or fixed-point method for solving the Sinc–Galerkin nonlinear systems of the Burgers equations.

In Figures 4.2, 4.4, and 4.6, we plot the exact and computed solutions of examples (i) and (ii) corresponding to the cases shown in the Figures 4.1, 4.3, and 4.5, respectively, where the computed solution is obtained by using the Newton or the fixed-point iteration method. Figures 4.2, 4.4, and 4.6 further show that our new methods can compute reasonably accurate results.

5. Concluding remarks. We have constructed a structured preconditioner that can efficiently improve the convergence property of the GMRES iteration employed to inexactly solve the subsystem of linear equations involved in each Newton or fixed-point iteration for solving the system of nonlinear equations resulting from the Sinc–Galerkin discretization of the Burgers equation. The bounds of the eigenvalues of the preconditioned matrix have been precisely estimated by making use of the generalized Bendixson theorem. Numerical experiments have shown the effectiveness of this new preconditioner.

REFERENCES

[1] Z.-Z. Bai, *Parallel multisplitting AOR method for solving a class of system of nonlinear algebraic equations*, Appl. Math. Mech., 16 (1995), pp. 675–682.
[2] Z.-Z. Bai, *Parallel nonlinear AOR method and its convergence*, Comput. Math. Appl., 31 (1996), pp. 21–31.
[3] Z.-Z. Bai, *A class of two-stage iterative methods for systems of weakly nonlinear equations,* Numer. Algorithms, 14 (1997), pp. 295–319.
[4] Z.-Z. Bai, *Parallel multisplitting two-stage iterative methods for large sparse systems of weakly nonlinear equations*, Numer. Algorithms, 15 (1997), pp. 347–372.
[5] 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), pp. 844–863.
[6] 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), pp. 603–626.
[7] Z.-Z. Bai and M.K. Ng, *Preconditioners for nonsymmetric block Toeplitz-like-plus-diagonal linear systems*, Numer. Math., 96 (2003), pp. 197–220.
[8] Z.-Z. Bai, J.-C. Sun, and D.-R. Wang, *A unified framework for the construction of various matrix multisplitting iterative methods for large sparse system of linear equations*, Comput. Math. Appl., 32 (1996), pp. 51–76.
[9] Z.-Z. Bai and D.-R. Wang, *Asynchronous multisplitting nonlinear Gauss-Seidel type method*, Appl. Math. J. Chinese Univ. Ser. B, 9 (1994), pp. 189–194.
[10] Z.-Z. Bai and D.-R. Wang, *Asynchronous parallel multisplitting nonlinear Gauss-Seidel iteration*, Appl. Math. J. Chinese Univ. Ser. B, 12 (1997), pp. 179–194.
[11] F. Di Benedetto, *Solution of Toeplitz normal equations by sine transform based preconditioning*, Linear Algebra Appl., 285 (1998), pp. 229–255.
[12] R.H. Chan and M.K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.
[13] T. Kailath and A.H. Sayed, *Displacement structure: Theory and applications*, SIAM Rev., 37 (1995), pp. 297–386.
[14] N. Levinson, *The Wiener RMS (root mean square) error criterion in filter design and prediction*, J. Math. Phys. Mass. Inst. Tech., 25 (1947), pp. 261–278.
[15] J. Lund and K.L. Bowers, *Sinc Methods for Quadrature and Differential Equations*, SIAM, Philadelphia, 1992.
[16] M.K. Ng, *Fast iterative methods for symmetric sinc-Galerkin systems*, IMA J. Numer. Anal., 19 (1999), pp. 357–373.
[17] M.K. Ng and Z.-Z. Bai, *A hybrid preconditioner of banded matrix approximation and alternating direction implicit iteration for symmetric Sinc-Galerkin linear systems*, Linear Algebra Appl., 366 (2003), pp. 317–335.
[18] M.K. Ng and D. Potts, *Fast iterative methods for sinc systems*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 581–598.
[19] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York and London, 1970.
[20] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[21] D.-R. Wang, Z.-Z. Bai, and D.J. Evans, *Asynchronous multisplitting relaxed iterations for weakly nonlinear systems*, Int. J. Comput. Math., 54 (1994), pp. 57–76.

[22] D.-R. Wang, Z.-Z. Bai, and D.J. Evans, *On the monotone convergence of multisplitting method for a class of system of weakly nonlinear equations*, Int. J. Comput. Math., 60 (1996), pp. 229–242.