MATHEMATICS OF COMPUTATION
Volume 66, Number 217, January 1997, Pages 215–235
S 0025-5718(97)00789-8

# A SINC–COLLOCATION METHOD FOR INITIAL VALUE PROBLEMS

TIMOTHY S. CARLSON, JACK DOCKERY, AND JOHN LUND

ABSTRACT. A collocation procedure is developed for the initial value problem $u'(t) = f(t, u(t))$, $u(0) = 0$, using the globally defined sinc basis functions. It is shown that this sinc procedure converges to the solution at an exponential rate, i.e., $\mathcal{O}(M^2 \exp(-\kappa\sqrt{M}))$ where $\kappa > 0$ and $2M$ basis functions are used in the expansion. Problems on the domains $\mathbb{R} = (-\infty, \infty)$ and $\mathbb{R}^+ = (0, \infty)$ are used to illustrate the implementation and accuracy of the procedure.

## 1. INTRODUCTION

In this paper a collocation procedure for the numerical solution of the initial value problem

$$\frac{du(t)}{dt} = f(t, u(t)), \quad u(a) = 0 \ , \tag{1.1}$$

is developed. A global approximation of the solution of (1.1), which is valid for $t \in [a, b)$, is obtained from the sinc functions. These functions are derived from the entire function

$$\operatorname{sinc}(z) \equiv \begin{cases} \dfrac{\sin(\pi z)}{\pi z}, & z \neq 0 \ , \\ 1, & z = 0, \end{cases}$$

by translations. For each integer $j$ and the mesh size $h$, the sinc basis functions are defined on $\mathbb{R}$ by

$$S_j(x) \equiv \begin{cases} \dfrac{\sin\left[(\frac{\pi}{h})(x - jh)\right]}{\left[(\frac{\pi}{h})(x - jh)\right]}, & x \neq jh \ , \\ 1, & x = jh \ . \end{cases} \tag{1.2}$$

The sinc functions form an interpolatory set of functions, i.e.,

$$S_j(kh) = \delta_{jk}^{(0)} = \begin{cases} 1 & \text{if } j = k \ , \\ 0 & \text{if } j \neq k \ . \end{cases} \tag{1.3}$$

Received by the editor February 27, 1995 and, in revised form, November 2, 1995 and January 26, 1996.

1991 *Mathematics Subject Classification.* Primary 65L05, 65L60.

The first author was supported in part by the Office of Naval Research under contract ONR-00014-89-J-1114.

The second author was supported in part by the National Science Foundation grants OSR-93-50-546 and DMS-94-04-160.

Since the basis functions are defined on the whole real line, a convenient starting point is the construction of an approximation to the solution of the problem

$$\frac{du(x)}{dx} = f(x, u(x)), \qquad \lim_{x \to -\infty} u(x) = 0 \ . \tag{1.4}$$

From (1.2), the basis functions satisfy the limit condition in (1.4), so that the assumed approximate solution

$$w_m(x) = \sum_{j=-M}^{M-1} w_j S_j(x) \ , \quad m = 2M \ , \tag{1.5}$$

has the same property. The most direct method for the determination of the error requires the additional assumption

$$\lim_{x \to \infty} u(x) = 0 \ , \tag{1.6}$$

since the assumed approximate solution (1.5) has this behavior. For this introductory material it is assumed that the solution of (1.4) satisfies (1.6). This assumption, (1.6), will be removed in §2 with the introduction of an auxiliary basis function in the expansion (1.5).

A collocation scheme is defined by substituting (1.5) into (1.4) and evaluating the result at $x_k = kh, k = -M, \dots, M-1$. This gives the equation

$$\frac{1}{h} I_m^1 \vec{w} = -\vec{f}(\vec{x}, \vec{w}) \ , \text{ where } \ \vec{f}(\vec{x}, \vec{w}) = \begin{pmatrix} f(x_{-M}, w_{-M}) \\ f(x_{-M+1}, w_{-M+1}) \\ \vdots \\ f(x_{M-1}, w_{M-1}) \end{pmatrix} , \tag{1.7}$$

$\vec{x} = [x_{-M}, \dots, x_{M-1}]^t$, and $\vec{w} = [w_{-M}, \dots, w_{M-1}]^t$. The coefficient matrix in (1.7) is obtained from the explicit values for the derivative of the sinc basis functions at the nodes,

$$\left. \frac{dS_j(x)}{dx} \right|_{x = x_k = kh} = \frac{1}{h} \delta_{jk}^{(1)} = \frac{1}{h} \begin{cases} 0 & \text{if } j = k \ , \\ \dfrac{(-1)^{k-j}}{k-j} & \text{if } j \neq k \ . \end{cases} \tag{1.8}$$

Collecting the numbers $\delta_{jk}^{(1)}, -M \leq j, k \leq M-1$, leads to the definition of the $m \times m$ skew-symmetric coefficient matrix in (1.7):

$$I_m^1 = \begin{bmatrix} 0 & -1 & \frac{1}{2} & -\frac{1}{3} & \cdots & -\frac{1}{2M-1} \\ 1 & 0 & -1 & \frac{1}{2} & \cdots & \frac{1}{2M-2} \\ -\frac{1}{2} & 1 & 0 & -1 & \cdots & -\frac{1}{2M-3} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ -\frac{1}{2M-2} & \frac{1}{2M-3} & \cdots & 1 & 0 & -1 \\ \frac{1}{2M-1} & -\frac{1}{2M-2} & \frac{1}{2M-3} & \cdots & 1 & 0 \end{bmatrix}_{m \times m} \tag{1.9}$$

The procedure then is to solve the system (1.7) for the $m \times 1$ vector of coefficients $\vec{w}$ in (1.5). The discrete system in (1.7) can also be obtained via a Sinc–Galerkin procedure as outlined in [5, pp. 136-138]. Furthermore, the sinc discretization of differential equations, whether by Galerkin or collocation procedures, has been addressed by a number of authors. In particular, Sinc-Collocation procedures for the eigenvalue problem have been addressed in [6, 3], and for the two-point boundary value problem in [8, 1] and [9]. These procedures, as well as an extensive summary of properties of sinc approximation, can be found in [10].

It is shown in §2 that if the function $f(x, u(x))$ is continuously differentiable and $u(x)$, the solution of (1.4), is sinc approximable then there exists a unique solution $\vec{w}$ to (2.7) so that

$$\|\vec{u} - \vec{w}\| \leq KM^2 \exp(-\kappa\sqrt{M}) \ , \tag{1.10}$$

where $\vec{u} = [u(x_{-M}), \dots, u(x_{M-1})]^t$. Furthermore, the error between the approximation defined by (1.5) and the solution $u(x)$ to (1.4) satisfies

$$\|u - w_m\| \leq \widehat{K}M^2 \exp(-\kappa\sqrt{M}) \ , \tag{1.11}$$

where $K, \widehat{K}$ and $\kappa$ are positive constants. The notation $\|\cdot\|$ denotes the discrete or continuous two-norm. The proof of the estimate (1.11) depends on, among other things, the spectrum of $I_m^1$, and in turn, on the Toeplitz structure of $I_m^1$. This spectral study is also carried out in §2.

In the case that $f(s, u) = g(s)$, a connection with the method of Stenger, [11], can be developed by integrating (1.4) from $-\infty$ to a node $x_k$, giving

$$u(x_k) = \int_{-\infty}^{x_k} g(s)\, ds \ .$$

If the $w_j$ in (1.5) are replaced by $g(jh)$ and the resulting sinc expansion of $g(s)$ is substituted in the right-hand side, then the approximation

$$u(x_k) \approx \widehat{u}(x_k) = \sum_{j=-M}^{M-1} g(jh) \int_{-\infty}^{x_k} S_j(s)\, ds$$

results. Letting $k$ vary from $-M$ to $M-1$ gives the matrix equation

$$\vec{\widehat{u}} = h I_m^{(-1)} \vec{g} \ , \tag{1.12}$$

where the entries in the matrix $I_m^{(-1)}$ are defined by the integrals

$$\delta_{jk}^{(-1)} = \frac{1}{h} \int_{-\infty}^{x_k} S_j(s)\, ds \ .$$

It is shown in [10, p. 175] that the error in approximating $u(x)$ by (1.5), where the coefficients are the components of $\vec{\widehat{u}}$, satisfies the exponential convergence rate (1.10). Within this exponential accuracy these two methods are the same. This is numerically illustrated in each of the examples in §2.

The convergence proof which gives the order statement in (1.10) also applies to problems on an interval $[a, b)$ via the method of conformal mapping. The case of the mapping $x = \phi(t) = \ln(t)$, $t \in (0, \infty)$, is addressed in §3. The main motivation for restricting to the half–line is for implementation in the numerical solution of parabolic partial differential equations, where the convergence to an asymptotic state may be at an algebraic rate.

If the time domain is the half-line, the sinc basis functions in (1.2) are replaced by

$$S_j \circ \phi(t) \equiv \frac{\sin[(\pi/h)(\phi(t) - jh)]}{[(\pi/h)(\phi(t) - jh)]} \; . \tag{1.13}$$

With this alteration, the approximation procedure is the same: assume an approximate solution of (1.1) of the form

$$w_m(t) = \sum_{j=-M}^{M-1} w_j S_j \circ \phi(t) \; , \quad m = 2M \; , \tag{1.14}$$

substitute (1.14) into (1.1) and collocate at the nodes $t_k = \phi^{-1}(x_k), k = -M, -M+1, \ldots, M-1$. This leads to the equation

$$\frac{1}{h} I_m^1 \vec{w} = -\mathcal{D}\left(\frac{1}{\phi'}\right) \vec{f}(\vec{t}, \vec{w}) \; , \tag{1.15}$$

where, given a function $g(t)$ defined on the nodes $t_k, k = -M, \ldots, M-1$, the notation $\mathcal{D}(g)$ denotes a $2M \times 2M$ diagonal matrix with the $k^{\text{th}}$ diagonal entry given by $g(t_k)$. One of the implementation conveniences of this sinc procedure is that the only alteration to the numerical procedure given by (1.7) is the introduction of a diagonal matrix on the right-hand side of (1.15). This procedure has the same rate of convergence as the procedure for the real line. Another convenience in the implementation of the method is that, in the case of using Newton's method, the Jacobian update is simply a diagonal matrix evaluation. This method is implemented and illustrated in §3.

The method of this paper is not limited to the scalar initial value problem (1.1). Indeed, for the initial value problem

$$\mathbf{y}'(t) = \mathbf{f}(t, \mathbf{y}) \; , \quad \mathbf{y}(t) \in \mathbb{R}^n \; , \quad t > 0,$$

the development leading to (1.15) gives the $n$ systems

$$\frac{1}{h} I_m^1 \vec{w}_j = -\mathcal{D}\left(\frac{1}{\phi'}\right) \vec{f}_j(\vec{t}, \vec{w}_1, \ldots \vec{w}_n), \qquad j = 1, 2, \ldots, n \; , \tag{1.16}$$

where $f_j$ denotes the $j^{\text{th}}$ component of $\mathbf{f}$. As in the previous paragraph, the implementation of Newton's method for (1.16) is simplified, owing to the diagonal matrix evaluations in the Jacobian update.

## 2. Collocation on $\mathbb{R}^1$

In this section the convergence rate given in (1.10) is obtained for the problem

$$u'(x) = f(x, u(x)), \quad \lim_{x \to -\infty} u(x) = 0 \; . \tag{2.1}$$

The space of functions where the sinc approximant given by (1.5) yields an exponential discretization error is given in the following definition.

**Definition 2.1.** The function $u$ is in the space of $\mathcal{H}^2(\mathcal{D}_d)$, where

$$\mathcal{D}_d = \{z = x + iy : 0 < |y| < d\}$$

if $u$ is analytic in $\mathcal{D}_d$ and satisfies

$$\int_{-d}^{d} |u(x + iy)| dy = \mathcal{O}(|x|^\gamma) \; , \quad x \to \pm\infty \; , \quad 0 \le \gamma < 1 \; ,$$

and

$$\mathcal{N}^2(u,\mathcal{D}_d) \equiv \lim_{y\to d^-} \left(\int_{-\infty}^{\infty} |u(x+iy)|^2 dx\right)^{1/2} + \left(\int_{-\infty}^{\infty} |u(x-iy)|^2 dx\right)^{1/2} < \infty \, .$$

There are many properties of the sinc expansion of functions in the class $\mathcal{H}^2(\mathcal{D}_d)$. A complete development is found in the text [10]. For the present paper the following interpolation and quadrature theorems play a key role.

**Theorem 2.2. Interpolation.** *Assume that $u \in \mathcal{H}^2(\mathcal{D}_d)$; then for all $z \in \mathcal{D}_d$*

$$\begin{aligned} E(u,h)(z) &\equiv u(z) - \sum_{k=-\infty}^{\infty} u(kh)S_k(z) \\ &= \frac{\sin(\pi z/h)}{2\pi i} \int_{-\infty}^{\infty} \left\{ \frac{u(s-id^-)}{(s-z-id^-)\sin(\pi(s-id^-)/h)} \right. \\ &\qquad \left. - \frac{u(s+id^-)}{(s-z+id^-)\sin(\pi(s+id^-)/h)} \right\} ds \end{aligned} \tag{2.2}$$

*and*

$$\|E(u,h)\| \le \frac{\mathcal{N}^2(u,\mathcal{D}_d)}{\sinh(\pi d/h)} = \mathcal{O}\left(e^{-\pi d/h}\right) \, . \tag{2.3}$$

**Corollary 2.3.** *Assume that $u \in \mathcal{H}^2(\mathcal{D}_d)$ and there are positive constants $\alpha$ and $K_1$ such that*

$$|u(x)| \le K_1 \exp(-\alpha|x|) \, , \quad x \in \mathbb{R} \, . \tag{2.4}$$

*If the mesh selection*

$$h = \sqrt{\frac{\pi d}{\alpha(M-1)}} \tag{2.5}$$

*is made in the finite sinc interpolant*

$$u_m(x) = \sum_{j=-M}^{M-1} u(x_j)S_j(x) \tag{2.6}$$

*to $u(x)$, then the error is bounded by*

$$\|u - u_m\| \le K_2 M^{1/2} \exp\left(-\sqrt{\pi d \alpha (M-1)}\right) \, . \tag{2.7}$$

**Theorem 2.4. Quadrature.** *Assume that $u \in \mathcal{H}^2(\mathcal{D}_d)$ is integrable; then*

$$\begin{aligned} \eta &\equiv \int_{-\infty}^{\infty} E(u,h)(x)\, dx = \int_{-\infty}^{\infty} u(x)\, dx - h \sum_{k=-\infty}^{\infty} u(kh) \\ &= \frac{e^{-\pi d/h}}{2i} \int_{-\infty}^{\infty} \left\{ \frac{u(s+id^-)e^{i\pi s/h}}{\sin(\pi(s+id^-)/h)} - \frac{u(s-id^-)e^{-i\pi s/h}}{\sin(\pi(s-id^-)/h)} \right\} ds \, . \end{aligned}$$

*Furthermore,*

$$|\eta| \le \frac{\mathcal{N}^2(u,\mathcal{D})e^{-\pi d/h}}{2\sinh(\pi d/h)} = \mathcal{O}\left(e^{-2\pi d/h}\right) . \tag{2.8}$$

Upon differentiating (2.2) one obtains the identity

$$\begin{aligned} u'(x) - \sum_{j=-M}^{M-1} u(jh)S_j'(x) &= \sum_{|j|>M} u(jh)S_j'(x) + u(Mh)S_M'(x) \\ &+ \frac{d}{dx}\left[\frac{\sin(\pi x/h)}{2\pi i}\int_{-\infty}^{\infty} \frac{u(s-id^-)}{(s-x-id^-)\sin(\pi(s-id^-)/h)}\right. \\ &\left. - \frac{u(s+id^-)}{(s-x+id^-)\sin(\pi(s+id^-)/h)}ds\right] , \end{aligned} \tag{2.9}$$

where the two terms on the right-hand side are called the truncation and the discretization errors, respectively. If the function $u(x)$ lies in $\mathcal{H}^2(\mathcal{D}_d)$, then it is shown in [7, Eq. # 4.3] that

$$\begin{aligned} &\left|\frac{d}{dx}\left[\frac{\sin(\pi x/h)}{2\pi i}\int_{-\infty}^{\infty} \frac{u(s-id^-)}{(s-x-id^-)\sin(\pi(s-id^-)/h)}\right.\right. \\ &\qquad \left.\left. - \frac{u(s+id^-)}{(s-x+id^-)\sin(\pi(s+id^-)/h)}ds\right]\right| \\ &\le \frac{K_3}{h}\exp(-\pi d/h) . \end{aligned} \tag{2.10}$$

A short calculation gives the bound

$$|S_j'(x)| = \left|\frac{dS_j(x)}{dx}\right| \le \frac{\pi}{2h}, \quad x \in \mathbb{R} .$$

Combining this inequality with (2.4) gives the following bound on the truncation error:

$$\begin{aligned} \left|\sum_{|j|>M} u(jh)S_j'(x) + u(Mh)S_M'(x)\right| &\le \sum_{|j|>M} \left|u(jh)S_j'(x)\right| + |u(Mh)S_M'(x)| \\ &\le \frac{\pi}{h}\sum_{j=M}^{\infty} |u(jh)| \\ &\le \frac{\pi K_1}{h}\sum_{j=M}^{\infty} |\exp(-\alpha jh)| \\ &= \frac{\pi K_1}{h}\left(\frac{\exp(-\alpha h)}{1-\exp(-\alpha h)}\right)\exp(-\alpha(M-1)h) \\ &\le \frac{\pi K_1}{\alpha h^2}\exp(-\alpha(M-1)h) \\ &\le \frac{K_4}{h^2}\exp(-\alpha(M-1)h) , \end{aligned} \tag{2.11}$$

where the inequality

$$\frac{\exp(-\alpha h)}{1-\exp(-\alpha h)} \le \frac{1}{\alpha h}$$

yields the first inequality in the last line of (2.11).

The initial value problem (2.1) gives $u'(x_k) = f(x_k, u(x_k))$, so that, evaluating (2.9) at the nodes, and using the inequalities in (2.10) and (2.11), one can show that the $k^{\text{th}}$ component of

$$N_m(\vec{u}) = \frac{1}{h} I_m^1 \vec{u} + \vec{f}(\vec{x}, \vec{u}) \tag{2.12}$$

is bounded by

$$\begin{aligned} |N_m(u_k)| &\le \frac{K_3}{h} \exp(-\pi d/h) + \frac{K_4}{h^2} \exp -(\alpha(M-1)h) \\ &\le \left[\widehat{K}_3 \sqrt{M} + \widehat{K}_4 M\right] \exp\left(-\sqrt{\pi d \alpha (M-1)}\right) \ , \end{aligned}$$

where the mesh selection $h$ in (2.5) was substituted in the first inequality to obtain the second inequality. Therefore, the vector $N_m(\vec{u})$, in the two-norm, is bounded by

$$\begin{aligned} \|N_m(\vec{u})\| &= \left( \sum_{k=-M}^{M-1} |N_m(u_k)|^2 \right)^{1/2} \\ &\le \sqrt{2M} \max_{-M \le k \le M-1} |N_m(u_k)| \\ &\le K_5 M^{3/2} \exp\left(-\sqrt{\pi d \alpha (M-1)}\right) \ . \end{aligned} \tag{2.13}$$

From (2.12) and the inequality (2.13), an estimate of the error in the approximation requires a bound on the norm of the inverse of the matrix $I_m^1, m = 2M$. It has been numerically shown that this matrix is invertible for all $M \le 250$ and the sixth column of Table 2 in Example 2.10 numerically supports this invertibility. These numerics, as well as analytic evidence supporting the invertibility of $I_{2M}^1$, motivates the assumption:

$$\|(I_m^1)^{-1}\| \le m = 2M, \quad M \ge 1 \ . \tag{2.14}$$

**Theorem 2.5.** *Assume that the function $u$ is in $\mathcal{H}^2(\mathcal{D}_d)$ and satisfies* (2.4). *Further, assume that the function $f(x,u)$ is continuously differentiable and that $f_u = \partial f / \partial u$ is Lipschitz continuous with Lipschitz constant $K_L$ and that* (2.14) *holds. Then in a sufficiently small ball about $u(x)$ the function*

$$w_m(x) = \sum_{j=-M}^{M-1} w_j S_j(x) \ , \tag{2.15}$$

*where the coefficients are determined by solving the equation*

$$N_m(\vec{w}) \equiv \frac{1}{h} I_m^1 \vec{w} + \vec{f}(\vec{x}, \vec{w}) = \vec{0} \ , \tag{2.16}$$

*satisfies*

$$\|w_m - u\| \le K_6 M^2 \exp\left(-\sqrt{\pi d \alpha (M-1)}\right) \ . \tag{2.17}$$

If $\vec{u} = (u(x_{-M}), \dots, u(x_{M-1}))^t$ is the vector of coefficients in the sinc expansion (2.6), then the equality of function and vector norms,

$$\|w_m - u_m\| = \|\vec{w} - \vec{u}\| \ ,$$

follows from the orthogonality of the sinc basis,

$$\int_{-\infty}^{\infty} S_j(x)S_k(x) = 0, \quad j \neq k \, .$$

Hence, the triangle inequality takes the form

$$\begin{aligned} \|w_m - u\| &\le \|w_m - u_m\| + \|u_m - u\| \\ &= \|\vec{w} - \vec{u}\| + \|u_m - u\| \\ &\le \|\vec{w} - \vec{u}\| + K_2 M^{1/2} \exp\left(-\sqrt{\pi d\alpha(M-1)}\right) \, , \end{aligned} \tag{2.18}$$

where the last inequality follows from (2.7). It remains to bound the error in the coefficients $\|\vec{w} - \vec{u}\|$, and this is addressed in the following two lemmas.

**Lemma 2.6.** *Assume that the function $u$ is in $\mathcal{H}^2(\mathcal{D}_d)$ and satisfies* (2.4). *Further, assume that the function $f(x,u)$ is continuously differentiable and that $f_u = \partial f/\partial u$ is Lipschitz continuous with Lipschitz constant $K_L$ and that* (2.14). *Then in a sufficiently small ball about $\vec{u}$ there is a unique solution $\vec{w}$ to* (2.16) *which satisfies the inequality*

$$\|\vec{w} - \vec{u}\| \le K_5 M^2 \exp\left(-\sqrt{\pi d\alpha(M-1)}\right) \, . \tag{2.19}$$

The idea of the proof is to use the Contraction Mapping Principle. This argument requires an estimate on the norm of the inverse of the matrix

$$L_m(\vec{u}) \equiv \frac{1}{h} I_m^1 + \mathcal{D}(\vec{f}_u(\vec{x}, \vec{u})) \, , \tag{2.20}$$

which, in turn, depends on the norm of the inverse of the matrix $I_m^1$. The assumed invertibility of the matrix $I_m^1$, $m = 2M$, and the estimate in (2.14) will be motivated following a discussion of the Toeplitz structure of the matrix $I_m^1$. Assuming the estimate in (2.14), the following lemma is needed in the proof of Lemma 2.6.

**Lemma 2.7.** *Let $ie_1$ be the purely imaginary eigenvalue of $I_m^1$, $m = 2M$, with smallest positive imaginary part $e_1$. Let $\mathcal{D}$ be an arbitrary $m \times m$ real diagonal matrix. Then*

$$\|(I_m^1 + \mathcal{D})^{-1}\| \le \frac{1}{e_1} = \|(I_m^1)^{-1}\| \, . \tag{2.21}$$

*Proof.* Since $I_m^1$ has real entries and is skew-symmetric, its eigenvalues are purely imaginary. To see the first inequality, let $\vec{v}$ be a unit eigenvector of $I_m^1$ corresponding to the eigenvalue $ie_1$. For an arbitrary unit vector $\vec{z} \in \mathbb{C}^{2M}$ one has

$$\begin{aligned} \|I_m^1 + \mathcal{D}\|^2 &\equiv \max_{\|\vec{z}\|^2=1} \left((I_m^1 + \mathcal{D})\vec{z}, (I_m^1 + \mathcal{D})\vec{z}\right) \\ &\ge ((I_m^1 + \mathcal{D})\vec{v}, (I_m^1 + \mathcal{D})\vec{v}) \\ &= (ie_1\vec{v} + \mathcal{D}\vec{v}, ie_1\vec{v} + \mathcal{D}\vec{v}) \\ &= (ie_1\vec{v} + \mathcal{D}\vec{v})^*(ie_1\vec{v} + \mathcal{D}\vec{v}) \\ &= |e_1|^2\vec{v}^{\,*}\vec{v} + (ie_1\vec{v})^*\mathcal{D}\vec{v} + ie_1\vec{v}^{\,*}\mathcal{D}^*\vec{v} + \vec{v}^*\mathcal{D}^*\mathcal{D}\vec{v} \\ &= |e_1|^2 + [(ie_1)^* + ie_1]\vec{v}^{\,*}\mathcal{D}\vec{v} + \vec{v}^{\,*}\mathcal{D}^2\vec{v} \ge |e_1|^2 \, , \end{aligned}$$

since the entries in $\mathcal{D}$ and $e_1$ are real. This implies (2.21) which completes the proof of Lemma 2.7. □

*Proof of Lemma* 2.6. Let $B_r(\vec{u})$ denote a ball of radius $r$ in $\mathbb{R}^{2M}$ about $\vec{u}$. Consider the fixed point problem

$$\begin{aligned} \vec{w} &= F_m(\vec{w}) \ , \\ F_m(\vec{w}) &\equiv \vec{w} - L_m^{-1}[\vec{u}]N_m(\vec{w}) \ . \end{aligned}$$

Lemma 2.7 shows that the function $L_m^{-1}[\vec{u}]$ in (2.20) exists and is bounded by

$$\|L_m^{-1}[\vec{u}]\| = \left\| \left[ \frac{1}{h} I_m^1 + \mathcal{D}(\vec{f}_u(\vec{x},\vec{u})) \right]^{-1} \right\| \le h(2M) \le K_6\sqrt{M} \ , \tag{2.22}$$

where the mesh size in (2.5) yields the last inequality. It follows that a fixed point of $F_m$ gives a solution of (2.16). Let $\vec{v} \in B_r(\vec{u})$; then the calculation

$$\begin{aligned} &\|F_m(\vec{v}) - \vec{u}\| = \|\vec{v} - \vec{u} - L_m^{-1}[\vec{u}]N_m(\vec{v})\| \\ &\quad = \|\vec{v} - \vec{u} - L_m^{-1}[\vec{u}] \left[ N_m(\vec{u}) + \left( \int_0^1 \frac{dN_m(t\vec{v} + (1-t)\vec{u})}{dt} dt \right) \right] \| \\ &\quad = \|\vec{v} - \vec{u} - L_m^{-1}[\vec{u}] \left[ N_m(\vec{u}) + \left( \int_0^1 D_{\vec{u}}(N_m)(t\vec{v} + (1-t)\vec{u}) dt \right) (\vec{v} - \vec{u}) \right] \| \\ &\quad \le \|L_m^{-1}[\vec{u}]N_m(\vec{u})\| \\ &\qquad + \|L_m^{-1}[\vec{u}] \left[ \int_0^1 \{ L_m[\vec{u}] - D_{\vec{u}}(N_m)(t\vec{v} + (1-t)\vec{u}) \} \, dt \right] (\vec{v} - \vec{u})\| \end{aligned} \tag{2.23}$$

follows from the Taylor polynomial for the function $N_m$ and the triangle inequality. The first term following the inequality sign in (2.23) can be bounded by the product of the right-hand sides of (2.13) and (2.22).

Consider bounding the second term following the inequality sign on the right-hand side of (2.23). Using the assumed Lipschitz continuity of $f_u$ yields

$$\begin{aligned} &\|L_m^{-1}[\vec{u}] \left[ \int_0^1 \{ L_m[\vec{u}] - D_{\vec{u}}(N_m)(t\vec{v} + (1-t)\vec{u}) \} \, dt \right] (\vec{v} - \vec{u})\| \\ &\quad = \|L_m^{-1}[\vec{u}] \left[ \int_0^1 \mathcal{D}\left( \vec{f}_u(\vec{x},\vec{u}) - \vec{f}_u(\vec{x}, t\vec{v} + (1-t)\vec{u}) \right) dt \right] (\vec{v} - \vec{u})\| \\ &\quad \le \|L_m^{-1}[\vec{u}]\|_2 K_L r^2 \ . \end{aligned} \tag{2.24}$$

Substituting (2.24) in the right-hand side of (2.23) leads to the inequality

$$\begin{aligned} \|F_m(\vec{v}) - \vec{u}\| &\le \|L_m^{-1}[\vec{u}]\| \left( \|N_m(\vec{u})\| + K_L r^2 \right) \\ &\le K_6\sqrt{M} \left[ \left( K_5 M^{3/2} \right) \exp\left( -\sqrt{\pi d\alpha(M-1)} \right) + K_L r^2 \right] \\ &\le K_7 M^2 \exp\left( -\sqrt{\pi d\alpha(M-1)} \right) + \sqrt{M}\hat{K}_L r^2 \ , \end{aligned}$$

where (2.13) and (2.22) yield the second inequality. The quadratic inequality

$$K_7 M^2 \exp\left( -\sqrt{\pi d\alpha(M-1)} \right) + \sqrt{M}\hat{K}_L r^2 < r$$

is satisfied for all $r \in (r_0, r_1)$, where

$$r_0 = \mathcal{O}(M^2 \exp\left( -\sqrt{\pi d\alpha(M-1)} \right)) < r_1 = \mathcal{O}(\frac{1}{\sqrt{M}}) \ , \tag{2.25}$$

since $M^2 \exp\left(-\sqrt{\pi d\alpha(M-1)}\right) \to 0$ as $M \to \infty$. This shows that $F_m$ maps $B_r(\vec{u})$ into itself.

Next it is shown that on $B_r(\vec{u})$, for $r$ sufficiently small, $F_m$ is a contraction mapping. Let $\vec{w}, \vec{v} \in B_r(\vec{u})$; then

(2.26)

$$\begin{aligned}
\|F_m(\vec{v}) - F_m(\vec{w})\| &= \|\vec{v} - \vec{w} - L_m^{-1}[\vec{u}]\,(N_m(\vec{v}) - N_m(\vec{w}))\,\| \\
&= \|L_m^{-1}[\vec{u}]\,[L_m[\vec{u}](\vec{v} - \vec{w}) - (N_m(\vec{v}) - N_m(\vec{w}))]\,\| \\
&\leq \|L_m^{-1}[\vec{u}]\|\;\|\mathcal{D}\left\{\vec{f}_u(\vec{x},\vec{u})\right\}(\vec{v}-\vec{w}) - [\vec{f}(\vec{x},\vec{v}) - \vec{f}(\vec{x},\vec{w})]\| \\
&= \|L_m^{-1}[\vec{u}]\|\;\|\int_0^1 \mathcal{D}\left\{\vec{f}_u(\vec{x}, t\vec{u} + (1-t)\vec{u}) - \vec{f}_u(\vec{x}, t\vec{v} + (1-t)\vec{w})\right\} dt\,(\vec{v}-\vec{w})\,\| \\
&\leq 2rK_L\|L_m^{-1}[\vec{u}]\|\;\|\vec{v} - \vec{w}\|\,,
\end{aligned}$$

where $K_L$ is a Lipschitz constant for $f_u$. By choosing

$$r = \frac{r_+}{2K_L} < \frac{1}{2K_7\sqrt{M}}$$

it follows from the inequalities in (2.26) and (2.22) that $F_m$ is a contractive map on $B_r(\vec{u})$, so that $F_m$ has a unique fixed point. Furthermore, from (2.25), one can choose

$$r = \mathcal{O}\left(M^2 \exp\left(-\sqrt{\alpha\pi d(M-1)}\right)\right)\,.$$

This completes the proof of Lemma 2.6. □

In order to provide support for the assumed invertibility of the matrix $I_m^1, m = 2M$, in (1.9) it is convenient to use the theorem of Toeplitz [4, p. 63].

**Theorem 2.8. Toeplitz.** *Denote the Fourier coefficients of the real-valued function $f \in L(-\pi,\pi)$ by*

$$f_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\exp(-inx)\,dx, \qquad n = 0, \pm 1, \pm 2, \dots\,,$$

*and define the $m \times m$ Toeplitz matrix of the function $f$ by*

$$(2.27) \qquad C_m(f) \equiv \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{m-1} \\ f_{-1} & f_0 & f_1 & \cdots & f_{m-2} \\ f_{-2} & f_{-1} & f_0 & \ddots & f_{m-3} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ f_{-m+2} & \cdots & f_{-1} & f_0 & f_1 \\ f_{-m+1} & \cdots & f_{-2} & f_{-1} & f_0 \end{bmatrix}_{m\times m}$$

*Denote the real eigenvalues of the Hermitian matrix $C_m(f)$ in increasing order by $\{e_j^m\}_{j=1}^m$. If the function $f$ has a minimum $\mathcal{M}_l$ and maximum $\mathcal{M}_u$ on $[-\pi,\pi]$, then for every $m$,*

$$\mathcal{M}_l \leq e_1^m \leq e_2^m \leq \cdots \leq e_m^m \leq \mathcal{M}_u\,.$$

*Further, if $C_m(g)$ is the Toeplitz matrix of the real-valued function $g \in L(-\pi,\pi)$ and $g(x) \leq f(x)$, then*

$$(2.28) \qquad c_j^m \leq e_j^m, \qquad j = 1, 2, \dots, m,$$

*where $\{c_j^m\}_{j=1}^m$ are the eigenvalues of $C_m(g)$.*

The role of the Toeplitz theorem in the present development is as follows. The Fourier coefficients of the function $f(x) = x$ are

$$\begin{aligned} f_n &= \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\exp(-inx)\,dx \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} x\exp(-inx)\,dx \\ &= \begin{cases} 0 & \text{if } n = 0\ , \\ \frac{i}{n}\cos(n\pi) & \text{if } n \neq 0, \end{cases} \\ &= i\delta_n^{(1)} = i\begin{cases} 0 & \text{if } n = 0\ , \\ \frac{(-1)^n}{n} & \text{if } n \neq 0\ , \end{cases} \end{aligned}$$

so that upon comparing these coefficients with the entries of the matrix $I_m^1, m = 2M$, in (1.9) shows that the Toeplitz matrix $C_{2M}(f) = iI_{2M}^1$. The eigenvalues of the real skew–symmetric matrix $I_{2M}^1$ occur in conjugate pairs $\{\pm ie_p^m\}_{p=1}^M$ and the nonnegative real numbers, $e_p^m$, satisfy the inequality

$$-\pi \le -e_M^m \le \cdots \le -e_1^m \le e_1^m \le \cdots \le e_M^m \le \pi\ . \tag{2.29}$$

To see that zero is not in the above list, consider the function

$$g(x) = \sin(x) = -\frac{e^{-ix}}{2i} + \frac{e^{ix}}{2i},$$

whose Fourier coefficients are given by $g_{\pm 1} = \pm\frac{1}{2i}$, and $g_n = 0$ if $n \neq \pm 1$, so that the Toeplitz matrix $C_m(g)$ is given by the matrix

$$C_m(g) \equiv \frac{1}{2i}\begin{bmatrix} 0 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & -1 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{m\times m} \tag{2.30}$$

The eigenvalues of the real skew–symmetric matrix $iC_m(g)$ also occur in conjugate pairs $\{\pm ic_p^m\}_{p=1}^M, m = 2M$, and the real numbers, $c_p^m$, are given by the explicit formula

$$c_p^m = \cos\left(\frac{[M-p+1]\pi}{2M+1}\right), \quad p = 1, 2, \ldots, M\ ,$$

and are ordered by

$$0 < c_1^m \le c_2^m \le \cdots \le c_M^m < 1\ . \tag{2.31}$$

The inequality $g(x) = \sin(x) \le f(x) = x$ is satisfied on the interval $[0, \pi]$. From this one would expect that

$$c_j^m \le e_j^m, \quad j = 1, ..., M\ , \tag{2.32}$$

as in Theorem 2.8, i.e., the positive part of the spectrum is monotonically ordered.

However, a proof of this last statement for general odd functions is not possible as the following example shows. Consider the $4 \times 4$ Toeplitz matrix generated by the function $-f(x) = -x$:

$$C_4(-f) = -iI_4^1 = i \begin{bmatrix} 0 & 1 & -\frac{1}{2} & \frac{1}{3} \\ -1 & 0 & 1 & -\frac{1}{2} \\ \frac{1}{2} & -1 & 0 & 1 \\ -\frac{1}{3} & \frac{1}{2} & -1 & 0 \end{bmatrix}_{4\times 4} ,$$

and the $4 \times 4$ Toeplitz matrix generated by $h(x) = \sin(x) + \sin(2x)$:

$$C_4(h) = \frac{1}{2i} \begin{bmatrix} 0 & -1 & -1 & 0 \\ 1 & 0 & -1 & -1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}_{4\times 4} .$$

The nonnegative eigenvalues of $C_4(-f)$ are $c_1^4 = \sqrt{13}/6$ and $c_2^4 = \sqrt{13}/2$ whereas the nonnegative eigenvalues $C_4(h)$ are $e_1^4 = 0$ and $e_2^4 = \sqrt{5}/2$ even though the inequality $-x \leq \sin(x) + \sin(2x)$ holds on $[0, \pi]$.

It has been shown that the inequality (2.32) holds for $g(x) = \sin(x)$ and $f(x) = x$ for all corresponding matrices up to size $500 \times 500$. Assuming this holds, it follows that

$$\min_{j=1,\dots,M} e_j^m = e_1^m \geq c_1^m = \cos\left(\frac{M\pi}{2M+1}\right) \geq \frac{1}{2M} . \tag{2.33}$$

Hence, combining (2.21) and (2.33), one finds

$$\|(I_{2M}^1)^{-1}\| = \frac{1}{e_1^m} \leq \frac{1}{c_1^m} .$$

In view of the upper bound in (2.29) for the eigenvalues of the matrix $I_{2M}^1$ the spectral condition number of this matrix is

$$\kappa(I_{2M}^1) = \|I_{2M}^1\| \, \|(I_{2M}^1)^{-1}\| = \frac{e_M^m}{e_1^m} \leq \frac{\pi}{\cos\left(\frac{M\pi}{2M+1}\right)} .$$

The following example clearly exposes the various parameter selections yielding the mesh selection $h$ in (2.5) and also illustrates the close connection of this method with the method found in [10, §7.1].

**Example 2.9.** The function

$$u(z) = \frac{1}{\cosh(\pi z)} , \tag{2.34}$$

is analytic in a strip of width one (the pole closest to $\mathbb{R}^1$ of $u(z)$ occurs at $z = \frac{\pm i}{2}$), so that the domain of analyticity of this function is $\mathcal{D}_{\frac{1}{2}}$. Further, this function satisfies the inequality (2.4) with $K_1 = 2$ and $\alpha = \pi$ and is the unique solution to the problem

$$u'(x) = -\pi \sinh(\pi x)[u(x)]^2, \quad \lim_{x \to -\infty} u(x) = 0 . \tag{2.35}$$

The function in (2.34) satisfies the auxiliary assumption $\lim_{x\to\infty} u(x) = 0$ so that the Theorem 2.5 applies. Hence, setting $d = 1/2$ and $\alpha = \pi$ leads to the mesh size

$h = \sqrt{1/(2M)}$. The coefficients $\{w_j\}_{j=-M}^{M-1}$ in (2.15) are obtained by solving the system

$$\frac{1}{h} I_m^1 \vec{w} = -\vec{g}, \quad \text{where } g(x) = \frac{\pi \sinh(\pi x)}{\cosh^2(\pi x)} , \tag{2.36}$$

and the second column in Table 1 displays the error between the solution at the nodes and the coefficients,

$$ERR(M) = \|\vec{u} - \vec{w}\| , \tag{2.37}$$

which, owing to the factor $M^2$ in (2.19) and the inequality in (2.18), represents the dominant error contribution to $\|u - w_m\|$ .

TABLE 1. Errors in the computed solution of (2.35)

| $M$ | $ERR(M)$ | $\widehat{ERR}(M)$ | $\mathcal{R}$ |
|---|---|---|---|
| 4 | 7.9514e-02 | 6.5952e-02 | 8.29e-01 |
| 8 | 1.6165e-02 | 1.3654e-02 | 8.45e-01 |
| 16 | 1.6267e-03 | 1.4151e-03 | 8.70e-01 |
| 32 | 5.6978e-05 | 5.1164e-05 | 8.98e-01 |
| 64 | 4.3819e-07 | 4.0450e-07 | 9.23e-01 |
| 128 | 3.9179e-10 | 3.6964e-10 | 9.43e-01 |

Instead of the matrix inversion in (2.36), an alternative procedure [10, §7.1] begins by rewriting (2.36) as the indefinite integral

$$u(x_k) = \int_{-\infty}^{x_k} g(x)dx \ , \quad k = -M, -M+1, \ldots , M-1 \ . \tag{2.38}$$

Now replace $g(x)$ in (2.38) by the finite sinc expansion

$$g(x) \approx \sum_{j=-M}^{M-1} g(jh) S_j(x) \ , \tag{2.39}$$

integrate this expression from $-\infty$ to the node $x_k$, and define the approximation

$$\widehat{w}_k = \sum_{j=-M}^{M-1} g(jh) \int_{-\infty}^{kh} S_j(x)dx = \sum_{j=-M}^{M-1} h\delta_{jk}^{(-1)} g(jh) \ . \tag{2.40}$$

The numbers $\delta_{jk}^{(-1)}$ take the form

$$\delta_{jk}^{(-1)} = \frac{1}{h} \int_{-\infty}^{kh} \frac{\sin\left[(\frac{\pi}{h})(x - jh)\right]}{\left[(\frac{\pi}{h})(x - jh)\right]} dx = \int_{-\infty}^{k-j} \frac{\sin(\pi y)}{\pi y} dy \ . \tag{2.41}$$

It is shown in [10, p. 719] that the error in approximating $u(x_k)$ by the expression in (2.40) is, in supremum norm, of order $\mathcal{O}\left(M^{1/2} \exp\left(-\sqrt{\pi d \alpha (M-1)}\right)\right)$. Hence, to exponential order, this is the same as the bound on the right-hand side of (2.17). By letting $k = -M, \ldots , 0, \ldots , M-1$, the equations in (2.40) admit the matrix form

$$\vec{\widehat{w}} = h \left[\delta_{jk}^{(-1)}\right]_{m \times m} \vec{g} \ \equiv \ h \ I_m^{(-1)} \ \vec{g} \ . \tag{2.42}$$

These numbers are calculated and the error $\widehat{ERR}(M) = \|\vec{w} - \vec{u}\|$ is compared with the error $ERR(M)$ in (2.37). The ratios

$$\mathcal{R} = \frac{\widehat{ERR}(M)}{ERR(M)} \tag{2.43}$$

are displayed in the final column of Table 1.

The development up to this point has assumed that the solution of the initial value problem (2.1) vanishes at infinity. This limit assumption is removed by appending an auxiliary basis function to the sinc expansion in (2.15). Define the basis function

$$\omega(x) = \frac{e^x}{e^x + e^{-x}}$$

and form the augmented approximate sinc solution

$$w^a(x) = w_{m-1}(x) + c_\infty \omega(x) = \sum_{j=-M}^{M-2} c_j S_j(x) + c_\infty \omega(x) \ . \tag{2.44}$$

The additional basis function $\omega(x)$ satisfies

$$\lim_{x \to \pm\infty} \omega(x) = \lim_{x \to \pm\infty} \frac{e^x}{e^x + e^{-x}} = \begin{cases} 1, & x \to \infty \ , \\ 0, & x \to -\infty \end{cases}$$

and is included in the expansion to allow nonzero boundary values of $u$, $u(\infty) = u_\infty$. The change of variable

$$v(x) = u(x) - u_\infty \omega(x) \tag{2.45}$$

transforms the problem

$$u'(x) = f(x, u(x)), \qquad \lim_{x \to -\infty} u(x) = 0 \ , \tag{2.46}$$

to the problem

$$v'(x) = f(x, v(x) + u_\infty \omega(x)) - u_\infty \omega'(x), \qquad \lim_{x \to \pm\infty} v(x) = 0 \ . \tag{2.47}$$

If $u_\infty$ is known, then the method defined by (2.16) determines the $\{w_j\}_{j=-M}^{M-1}$ in the expansion

$$w_m(x) = \sum_{j=-M}^{M-1} w_j S_j(x)$$

and the result of Theorem 2.5 applies to the approximation of $v(x)$ in (2.47) by $w_m(x)$. However, if $u_\infty$ is unknown, one approach which preserves the error of Theorem 2.5 is to replace this unknown by $c_\infty$ in (2.45) and use the Quadrature Theorem 2.4 to write

$$\begin{aligned} v(\infty) = 0 &= \int_{-\infty}^{\infty} [f(x, v(x) + u_\infty \omega(x)) - u_\infty \omega(x)]\, ds \\ &\approx \int_{-\infty}^{\infty} [f(x, w_{m-1}(x) + c_\infty \omega(x)) - c_\infty \omega(x)]\, ds \\ &\approx h \sum_{k=-M}^{M-2} [f(x_k, c_k + c_\infty \omega(x_k)) - c_\infty \omega(x_k)] \ . \end{aligned}$$

Add this equation to the solution procedure to obtain the approximate value for $c_\infty$. Since the error in the quadrature theorem is the square of the error of interpolation, this introduces no more error than the error in the method defined by (2.16).

Incorporating the above side condition in the approximate method to determine the coefficients in (2.44) is less convenient to implement than the following approach. Directly substitute the augmented approximate sinc solution (2.44) into the differential equation (2.46) and collocate this expansion at the $m = 2M$ nodes $x_k, k = -M, \ldots, 0, \ldots M-1$. This leads to the bordered matrix system

$$A\vec{c} = \left[\frac{1}{h} I^1_{m \times m-1} \mid -\vec{\omega}'\right] \vec{c} = -\vec{f}(\vec{x}, T_\omega \vec{c}) \ . \tag{2.48}$$

In (2.48) the vector $\vec{c} = [c_{-M}, \ldots, c_0, \ldots c_{M-2}, c_\infty]^{\ t}$ contains the coefficients in (2.44), and the approximate values to the solution

$$\vec{w}^{\ a} = [w^a_{-M}, \ldots, w^a_0, \ldots w^a_{M-2}, w^a_\infty]^{\ t}$$

are obtained from the transformation

$$\vec{w}^{\ a} = T_\omega \vec{c} \ , \tag{2.49}$$

where the matrix $T_\omega$ is defined by

$$T_\omega = \begin{bmatrix} 1 & 0 & \cdots & 0 & \omega_{-M} \\ 0 & 1 & \cdots & 0 & \omega_{-M+1} \\ 0 & 0 & \ddots 1 & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & \omega_{M-2} \\ 0 & 0 & \cdots & 0 & \omega_{M-1} \end{bmatrix} . \tag{2.50}$$

Since the matrix $T_\omega$ has the explicit inverse

$$T_\omega^{-1} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -\dfrac{\omega_{-M}}{\omega_{M-1}} \\ 0 & 1 & \cdots & 0 & -\dfrac{\omega_{-M+1}}{\omega_{M-1}} \\ 0 & 0 & 1 & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & -\dfrac{\omega_{M-2}}{\omega_{M-1}} \\ 0 & 0 & \cdots & 0 & \dfrac{1}{\omega_{M-1}} \end{bmatrix} , \tag{2.51}$$

one may regard either the vector $\vec{c}$ or $\vec{w}^{\ a}$ as the unknown in (2.49).

The system in (2.48) is solved for the coefficients by applying Newton's method to the function

$$N_m(\vec{c}) = A\vec{c} + \vec{f}(\vec{x}, T_\omega \vec{c}) \ . \tag{2.52}$$

If the matrix $A$ satisfies the conclusion of Lemma 2.7, then Theorem 2.5 applies to the function $N_m$ so that the rate of convergence of the present method is also given by (2.17). Although an argument implying the validity of Lemma 2.7 for the matrix $A$ does not seem to be an immediate corollary of the argument implying its validity for $I^1_m$, the numerical results displayed in the next example provide compelling evidence for the validity of Lemma 2.7 with $I^1_m$ replaced by the matrix $A$ in (2.52).

**Example 2.10.** In this example the function

$$u(x) = \frac{\exp(x)}{\exp(x) + 1}$$

is a solution to

$$u'(x) = -[u(x)]^2 + g(x) \ , \quad \lim_{x \to -\infty} u(x) = 0 \ , \tag{2.53}$$

provided $g(x) = u(x)$. The coefficients $\vec{c}$ in the approximation $w^a(x)$ are found by solving (2.52), which takes the form

$$N_m(\vec{c}) = A\vec{c} + \mathcal{D}\left((T_\omega \vec{c}\,)^{\,2}\right) - \vec{g} = \vec{0} \ .$$

The matrix $\mathcal{D}\left((T_\omega \vec{c}\,)^{\,2}\right)$ is the diagonal matrix whose $k^{\text{th}}$ diagonal entry is given by the square of the $k^{\text{th}}$ component of the vector $T_\omega \vec{c}$. This system is solved by Newton's method, and the number of iterations $n$ used in the calculations is recorded in Table 2. As in the last example, the error of the method,

$$ERR(M) = \|\vec{u} - \vec{w}^a\| \ , \tag{2.54}$$

is displayed in the third column of Table 2. The method of [10] discussed in the previous example was also applied and the ratio of the errors defined in (2.43) are recorded in the fifth column of Table 2.

TABLE 2. Errors in the computed solution of (2.51)

| M | $n$ | $ERR(M)$ | $\widehat{ERR}(M)$ | $\mathcal{R}$ | $R((I_m^1)^{-1})$ | $R(A^{-1})$ |
|---|---|---|---|---|---|---|
| 4 | 6 | 1.2284e-01 | 1.1243e-02 | 9.15e-02 | 5.19e-01 | 6.71e-01 |
| 8 | 6 | 2.5326e-02 | 3.0833e-03 | 1.21e-01 | 5.13e-01 | 6.02e-01 |
| 16 | 7 | 2.6765e-03 | 3.7675e-04 | 1.41e-01 | 5.09e-01 | 5.51e-01 |
| 32 | 8 | 9.7673e-05 | 1.5030e-05 | 1.54e-01 | 5.06e-01 | 5.23e-01 |
| 64 | 9 | 7.7053e-07 | 1.2637e-07 | 1.64e-01 | 5.03e-01 | 5.11e-01 |
| 128 | 10 | 6.9836e-10 | 1.2016e-10 | 1.72e-01 | 5.02e-01 | 5.05e-01 |

To amplify the remarks preceding the opening of this example, the final two columns in Table 2 compare the ratios

$$R((I_m^1)^{-1}) = \frac{\|(I_m^1)^{-1}\|}{2M} \quad \text{and} \quad R(A^{-1}) = \frac{\|A^{-1}\|}{2M} \ .$$

For this example the rank-one change from the matrix $I_m^1$ to $A$ has not, in magnitude, altered the norm in any significant manner. Indeed, since the matrix $A$ in (2.52) is independent of the problem (it only depends on the choice of $\omega(x)$), this comparison remains the same for other initial value problems.

## 3. COLLOCATION ON $\mathbb{R}^+$

The procedure and the proof of convergence in the last section applies to the problem

$$u'(t) = f(t, u(t)), \quad u(0) = 0 \ , \tag{3.1}$$

via the method of conformal mapping. Specifically, the map

$$z = \phi(w) = \ln(w), \quad w = e^z \ ,$$

is a conformal equivalence of the strip $D_d$ in Definition 2.1 onto the wedge

$$\mathcal{D}_W = \{w \in \mathbb{C} : w = re^{i\theta},\ \ |\theta| < d \leq \pi/2\}\ . \tag{3.2}$$

The analogue of the space $\mathcal{H}^2(\mathcal{D}_d)$ for this domain is contained in the following definition.

**Definition 3.1.** The function $u(z)$ is in the space $\mathcal{H}^2(\mathcal{D}_W)$ if $u$ is analytic in $\mathcal{D}_W$ and satisfies

$$\int_{-d}^{d} |F(re^{i\theta})| r\, d\theta = \mathcal{O}(|\ln(r)|^a), \qquad r \to 0^+, \infty, \quad 0 \leq a < 1\ ,$$

and

$$\lim_{\gamma \to \partial D_W} \int_\gamma |F(w)dw| = \lim_{\substack{r\to 0^+\\ R\to\infty}} \int_r^R |F(\rho e^{id})| d\rho < \infty.$$

A sinc approximate solution of (3.1) takes the form

$$w_m(t) = \sum_{j=-M}^{M-1} w_j S_j \circ \phi(t), \qquad m = 2M\ , \tag{3.3}$$

where the basis functions for the half-line are defined by the composition

$$S_j \circ \phi(t) \equiv \frac{\sin[(\pi/h)\phi(t) - jh]}{[(\pi/h)\phi(t) - jh]}\ . \tag{3.4}$$

With this alteration, the derivation of the approximation procedure is the same as it was in §2. Substitute $w_m$ into (3.1) and evaluate at the $m = 2M$ sinc nodes $\phi^{-1}(x_k) \equiv t_k = \exp(kh), k = -M, \ldots, M-1$, to arrive at the discrete system

$$\frac{1}{h} I_m^1 \vec{w} = -\mathcal{D}\left(\frac{1}{\phi'}\right) \vec{f}(\vec{t}, \vec{w})\ . \tag{3.5}$$

As mentioned in the Introduction, the only difference between this matrix equation and the one presented in (2.16) of §2 is the multiplicative diagonal matrix $\mathcal{D}(\frac{1}{\phi'})$.

The importance of the class of analytic functions in Definition 3.1 lies in the fact that if $\phi'(w)u(w) \in \mathcal{H}^2(\mathcal{D}_W)$, and if there are positive constants $\alpha$ and $K_1$ so that

$$|u(t)| \leq K_1 \frac{t^\alpha}{(1+t)^{2\alpha}}, \qquad t > 0\ , \tag{3.6}$$

then the sinc interpolant to $u(t)$ also satisfies (2.7) and (2.9). Since $u'(t_k) = f(t_k, u(t_k))$, it again follows that the error in the $k^{\text{th}}$ component of the function

$$N_m(\vec{u}) = \frac{1}{h} I_m^1 \vec{u} + \mathcal{D}\left(\frac{1}{\phi'}\right) \vec{f}(\vec{x}, \vec{u})$$

is bounded by

$$|N_m(u_k)| \leq \frac{K_3}{h} \exp(-\pi d/h) + \frac{K_4}{h^2} \exp\left(-\alpha(M-1)h\right)\ . \tag{3.7}$$

Finally, the mesh selection

$$h = \sqrt{\frac{\pi d}{\alpha(M-1)}}\ ,$$

when substituted into the right-hand side of (3.7), leads to the bound in (2.13) for $||N_m(\vec{u})||$ in (3.7).

**Theorem 3.2.** *Assume that the function $\phi'(w)u(w)$ is in $\mathcal{H}^2(\mathcal{D}_W)$ and satisfies (3.6). Further, assume that the function $f(t,u)$ is continuously differentiable and that $f_u = \partial f/\partial u$ is Lipschitz continuous with Lipschitz constant $K_L$. Then in a sufficiently small ball about $u(t)$ there is a unique vector $\vec{w}$ which provides the coefficients for $w_m(t)$ in (3.3) and*

$$\|w_m - u\| \leq KM^2 \exp\left(\sqrt{\pi d\alpha(M-1)}\right) . \tag{3.8}$$

The proof of this again follows from Lemma 2.6 and Lemma 2.7, both of which remain valid with the stated assumptions and owing to the fact that the coefficient matrix in (3.5) remains the same as in §2.

Since the functions $S_j \circ \phi(t)$ have the property $\lim_{t\to\infty} S_j \circ \phi(t) = 0$, the assumed approximate $w_m(t)$ in (3.3) has the same property so that the method can only be expected to approximate initial value problems with this added assumption. This limit assumption is removed by appending an auxiliary basis function to the sinc expansion in (3.3) much as in the last section, and is discussed in the next example.

**Example 3.3.** Let $\gamma$ be a real parameter in the family of initial value problems

$$u'(t) = f(t) = (1-\gamma t)\exp(-t) \ , \quad t > 0 \ , \quad u(0) = 0 \ . \tag{3.9}$$

The solution is given by

$$u(t) = 1 - \exp(-t) + \gamma\left(\exp(-t) + t\exp(-t) - 1\right)$$

and satisfies

$$\lim_{t\to\infty} u(t) = u_\infty = 1 - \gamma \ .$$

This example serves to illustrate that the procedure not only tracks a nonzero limit value ($\gamma \neq 1$), but also that the method still tracks a zero steady state ($\gamma = 1$).

As discussed in the lines following (2.47), add the additional basis function

$$\omega(t) = \frac{t}{t+1} \tag{3.10}$$

to the sinc approximate (3.3) to obtain the augmented sinc approximant

$$w^a(t) = \sum_{j=-M}^{M-2} c_j S_j \circ \phi(t) + c_\infty \omega(t) \ . \tag{3.11}$$

Substitute (3.11) into (3.1) and evaluate this result at the sinc nodes $t_k = \exp(kh)$, $k = -M, -M+1, \ldots, M-1$. This yields the matrix system

$$A\vec{c} = -h\mathcal{D}\left(\frac{1}{\phi'}\right)\vec{f} \ , \tag{3.12}$$

where

$$A = \left[\frac{1}{h} I^{(1)}_{m\times m-1} \Big| -\frac{\vec{\omega}'}{\vec{\phi}'}\right] \ . \tag{3.13}$$

The approximation to the solution $w^a(t_k) = c_k + c_\infty \omega(t_k) \approx u(t_k)$ is obtained from the transformation $\vec{w}^a = T_\omega \vec{c}$. The coefficients $c_k, k = -M, \ldots, M-2$, and $c_\infty$, are assembled in the $m \times 1$ vector $\vec{c}$, and the matrix

$$T_\omega = [I_{m\times m-1} | \ \vec{\omega}]$$

is the same as in (2.50) with $\omega$ replaced by (3.10). It is important that the system (3.12) calculates the limit value when $\gamma = 1$, namely zero. For purposes of illustration, the system (3.5), without the augmented basis function has also been computed and the results of solving that system for the coefficients in (3.3) are given in Table 3 as well.

TABLE 3. Errors in the augmented and non–augmented approximation for the solution of (3.9) with $\gamma = 1$

| $M$ | $\|\vec{w}^a - \vec{u}\|$ | $\|\vec{w} - \vec{u}\|$ |
|---|---|---|
| 4 | 1.4419e-01 | 8.1682e-02 |
| 8 | 3.1887e-02 | 1.7142e-02 |
| 16 | 6.4556e-03 | 3.2712e-03 |
| 32 | 3.4783e-05 | 2.9180e-05 |
| 64 | 2.3802e-06 | 1.2030e-06 |
| 128 | 2.0902e-09 | 1.0572e-09 |

If the bound on the inverse of $A$ in (3.13) satisfies the conclusion of Lemma 2.7, then the results displayed in the above table are not specific to this example.

In the general case, the discretization of the problem (3.1) takes the form

$$A\vec{c} = -\mathcal{D}\left(\frac{1}{\phi'}\right)\vec{f}(\vec{t}, T_\omega\vec{c}) \ , \tag{3.14}$$

from which the coefficients in (3.11) are calculated and the approximation to the solution at the nodes is given by $w^a(t_k) = c_k + c_\infty \omega(t_k)$. In each of the following examples, Newton's method is applied to the function

$$N_m(\vec{c}) = A\vec{c} + \mathcal{D}\left(\frac{1}{\phi'}\right)\vec{f}(\vec{t}, T_\omega\vec{c}) \ . \tag{3.15}$$

The vector $\vec{c}^{\,0} = \vec{1}$ initializes the Newton iteration

$$\vec{c}^{\,n+1} = \vec{c}^{\,n} + \vec{\delta}^{\,n} \ , \tag{3.16}$$

where the update $\vec{\delta}^{\,n}$ is given by

$$-\mathcal{J}(N_m)(\vec{c}^{\,n})\vec{\delta}^{\,n} = N_m(\vec{c}^{n}) \tag{3.17}$$

and the Jacobian of (3.15) is

$$\mathcal{J}(N_m)(\vec{c}) = A + \mathcal{D}\left(\frac{1}{\phi'}\right)\mathcal{D}\left(\frac{\partial\vec{f}}{\partial u}(\vec{t}, T_\omega\vec{c})\right)T_\omega \ . \tag{3.18}$$

Note that besides the exponential rate of convergence given by (3.8), the computation involved for the Jacobian of the nonlinear system is straightforward. In fact, from (3.18), the update of the Jacobian is simply a diagonal evaluation.

**Example 3.4.** The initial value problem

$$u'(t) = f(t,u) = -\left(\frac{u^2 + 4u + 1}{2u + 4}\right) \ , \quad t > 0, \quad u(0) = 0 \ , \tag{3.19}$$

has the solution

$$u(t) = 2 - \sqrt{3 + \exp(-t)} \ ,$$

which tends to $2 - \sqrt{3}$ at the exponential rate

$$u(t) = (2 - \sqrt{3}) - \mathcal{O}(\exp(-t)) \quad \text{as} \quad t \to \infty\,. \tag{3.20}$$

The results in Table 4 display the number of Newton steps, $n$, in (3.16) and the two-norm error

$$ERR(M) = \|\vec{w}^{a} - \vec{u}\|\,.$$

TABLE 4. Error in the computed solution of (3.19)

| $M$ | $n$ | $ERR(M)$ |
|---|---|---|
| 4 | 4 | 2.2603e-03 |
| 8 | 5 | 2.9802e-03 |
| 16 | 5 | 2.6584e-04 |
| 32 | 5 | 7.6291e-06 |
| 64 | 6 | 4.2556e-08 |
| 128 | 6 | 2.0623e-12 |

A particularly useful application of the present procedure is to those initial value problems where the convergence to the asymptotic state is only of an algebraic rate. For example, an autonomous differential equation that has a nonhyperbolic rest point. The sinc approximation to such solutions also assumes algebraic decay at infinity, so that the convergence estimate in (3.8) is maintained. This is illustrated in the following example.

**Example 3.5.** For small positive parameters $\beta$ the problem

$$u'(t) = \beta(1 - u)^2, \quad t > 0, \quad u(0) = 0\,, \tag{3.21}$$

has the solution

$$u(t) = \frac{\beta t}{\beta t + 1} = 1 - \frac{1}{\beta t + 1}\,.$$

The asymptotic behavior

$$u(t) - 1 \sim \frac{1}{\beta t} \quad \text{as} \quad t \to \infty \tag{3.22}$$

shows the algebraic rate of approach to the asympototic state. In particular, for small $\beta$, this rate is is quite slow compared to the rate of approach in the previous example, given by (3.20).

In Table 5 the error in the calculated solution of (3.21) is displayed for several values of $\beta$. As one reads the table from left to right (decreasing $\beta$), there are fewer Newton steps computed to achieve the error, owing to the decreased accuracy in the computed solution. The reason for this decrease in accuracy can be traced to the truncation error, which is bounded by the second term on the right-hand side of (3.7). For $t$ large, the inequality in (3.6) implies

$$u(t) - 1 \sim K_1 \frac{1}{t^{\alpha}}\,.$$

As seen from (3.22), $K_1 \sim 1/\beta$. Hence, as $\beta$ is decreasing, the constant $K_1$ is increasing. In these cases of an algebraic rate of approach to the asymptotic state,

TABLE 5. Error in the computed solution of (3.21)

|  |  | $ERR(M)$ |  | $ERR(M)$ |  | $ERR(M)$ |
|---|---|---|---|---|---|---|
| $M$ | $n$ | $\beta = .1$ | $n$ | $\beta = .01$ | $n$ | $\beta = .001$ |
| 4 | 6 | 1.3231e-01 | 4 | 2.8747e-01 | 3 | 4.9698e-02 |
| 8 | 9 | 1.9510e-02 | 6 | 2.0021e-01 | 4 | 2.6669e-01 |
| 16 | 13 | 1.0601e-03 | 10 | 1.7213e-02 | 7 | 1.5763e-01 |
| 32 | 18 | 1.8684e-05 | 15 | 3.7626e-04 | 12 | 4.3506e-03 |
| 64 | 26 | 5.8273e-08 | 23 | 1.8770e-06 | 20 | 2.1567e-05 |
| 128 | 37 | 1.1437e-11 | 34 | 1.1200e-09 | 31 | 1.3027e-08 |

a simple change in the definition of the mesh selection (3.9) yields an accuracy bounded by $\exp(-(\delta\sqrt{M}))$ where $\delta \leq \alpha$. This alternative mesh selection, which defines a mesh reallocation, is also used, for example in boundary layer problems, and forms a portion of [2].

## ACKNOWLEDGMENT

The authors appreciate the clarifications suggested by the referee.

## REFERENCES

1. B. Bialecki, *Sinc-collocation methods for two-point boundary value problems*, IMA J. Numer. Anal. **11** (1991), 357–375. MR **92f:**65086
2. T. S. Carlson, *Sinc methods for Burgers' equation*, Ph.D. thesis, Montana State University, 1995.
3. N. Eggert, M. Jarratt, and J. Lund, *Sinc function computation of the eigenvalues of Sturm-Liouville problems*, J. Comput. Phys. **69** (1987), no. 1, 209–229. MR **89c:**65090
4. U. Grenander and G. Szegö, *Toeplitz forms and their applications*, 2nd ed., Chelsea Publishing Co., New York, 1984. MR **88b:**42031
5. J. Lund and K. L. Bowers, *Sinc methods for quadrature and differential equations*, SIAM, Philadelphia, 1992. MR **93i:**65004
6. J. Lund and B. V. Riley, *A sinc-collocation method for the computation of the eigenvalues of the radial Schrödinger equation*, IMA J. Numer. Anal. **4** (1984), 83–98. MR **86f:**65134
7. L. Lundin and F. Stenger, *Cardinal type approximations of a function and its derivatives*, SIAM J. Math. Anal. **10** (1979), 139–160. MR **81c:**41043
8. K. M. McArthur, *A collocative variation of the Sinc-Galerkin method for second order boundary value problems*, Computation and Control (K. Bowers and J. Lund, eds.), Birkhäuser, Boston, 1989, pp. 253–261. CMP 90:10
9. A. C. Morlet, *Convergence of the sinc method for a fourth-order ordinary differential equation with an application*, SIAM J. Numer. Anal. **32** (1995), 1475–1503. MR **96f:**65097
10. F. Stenger, *Numerical methods based on sinc and analytic functions*, Springer-Verlag, New York, 1993. MR **94k:**65003
11. F. Stenger, B. Barkey, and R. Vakili, *Sinc convolution approximate solution of Burgers' equation*, Computation and Control III (K. Bowers and J. Lund, eds.), Birkhäuser, Boston, 1993, pp. 341–354. CMP 94:04

SANTA FE INSTITUTE, 1399 HYDE PARK ROAD, SANTA FE, NEW MEXICO 87501
*E-mail address*: tim@santafe.edu

DEPARTMENT OF MATHEMATICS, MONTANA STATE UNIVERSITY, BOZEMAN, MONTANA 59717
*E-mail address*: umsfjdoc@math.montana.edu

DEPARTMENT OF MATHEMATICS, MONTANA STATE UNIVERSITY, BOZEMAN, MONTANA 59717
*E-mail address*: umsfjlun@math.montana.edu