A sinc-collocation method for Burgers Equation
by Timothy Scott Carlson

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Mathematics
Montana State University

Abstract:
Various aspects of the numerical solution to the viscous Burgers' equation via sinc functions are presented. Discretization in the temporal domain using a sinc function basis and a proof of convergence for the related first-order initial value problem is given. The temporal problem is posed on the half-line, but the treatment also includes a viable computational procedure for initial value problems on the entire real line. The novelty of the solution of this initial value problem is that the computed solution is globally defined. When the Reynolds number, a parameter of interest in Burgers' equation, is large, boundary layer effects arise. A procedure for the efficient choice of mesh size for these boundary layer problems which maintains the form of the discrete system is discussed. These temporal and spatial procedures are combined in a product discretization method for Burgers' equation.

A SINC–COLLOCATION METHOD FOR BURGERS' EQUATION

by

TIMOTHY SCOTT CARLSON

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Mathematics

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 1995

APPROVAL

of a thesis submitted by

TIMOTHY SCOTT CARLSON

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

April 19, 1995
Date

John Lund
John Lund
Chairperson, Graduate Committee

Approved for the Major Department

April 19, 1995
Date

John Lund
John Lund
Head, Mathematics

Approved for the College of Graduate Studies

5/1/95
Date

R. A. Brown
Robert Brown
Graduate Dean

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for extensive copying or reproduction of this thesis should be referred to University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute copies of the dissertation for sale in and from microform or electronic format, along with the right to reproduce and distribute my abstract in any format in whole or in part."

Signature 

Date 4/25/95

ACKNOWLEDGEMENTS

I would like to thank my parents, Norman and Miriam Carlson, for their love and support.

I would like to thank my advisor Dr. John Lund, not only for his mathematical guidance, but also his eloquent words of wisdom.

I would like to thank the members of my committee: Dr. Jack Dockery for all of his assistance, Dr. Ken Bowers who first introduced me to the sinc function, Dr. Curt Vogel who taught me everything I know about classical finite element and finite difference methods, and Dr. Gary Bogar who first introduced me to numerical analysis as an undergraduate.

I would like to thank Dr. Jeff Banfield who funded my final year of research through the Office of Naval Research under contract N–00014–89–J–1114.

I would like to dedicate this work to my wife Debbie for all her support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Various aspects of the numerical solution to the viscous Burgers' equation via sinc functions are presented. Discretization in the temporal domain using a sinc function basis and a proof of convergence for the related first–order initial value problem is given. The temporal problem is posed on the half–line, but the treatment also includes a viable computational procedure for initial value problems on the entire real line. The novelty of the solution of this initial value problem is that the computed solution is globally defined. When the Reynolds number, a parameter of interest in Burgers' equation, is large, boundary layer effects arise. A procedure for the efficient choice of mesh size for these boundary layer problems which maintains the form of the discrete system is discussed. These temporal and spatial procedures are combined in a product discretization method for Burgers' equation.

CHAPTER 1

Introduction

Burgers' equation

$$\begin{aligned} u_t(x,t) - \epsilon u_{xx}(x,t) + u(x,t)u_x(x,t) &= g(x,t), \quad a < x < b, \quad t > 0 \\ \alpha_1 u_x(a,t) - \alpha_0 u(a,t) &= 0, \quad t \geq 0 \\ \beta_1 u_x(b,t) + \beta_0 u(b,t) &= 0, \quad t \geq 0 \\ u(x,0) &= f(x), \quad a \leq x \leq b \end{aligned} \tag{1.1}$$

is a nonlinear parabolic partial differential equation that can be used as a prototype for the Navier–Stokes equations. In this work, a numerical method for solving (1.1) is discussed, developed, and implemented. The underlying idea in the numerical solution to (1.1) is based on the notion of a product method: the combination of a method to handle the spatial discretization along with a method to carry out the temporal discretization.

For the temporal discretization, fix $x = \hat{x}$ in (1.1) to obtain an initial value problem of the form

$$\begin{aligned} u'(t) &= F(t,u(t)), \quad t > 0 \\ u(0) &= f(\hat{x}) \end{aligned} \tag{1.2}$$

where

$$F(t,u(t)) = \epsilon u_{xx}(\hat{x},t) - u(\hat{x},t)u_x(\hat{x},t) + g(\hat{x},t) \quad .$$

In Chapter 2, a collocation procedure for

$$\begin{aligned} u'(t) &= f(t,u(t)), \quad t > 0 \\ u(0) &= 0 \end{aligned} \tag{1.3}$$

is developed. The linear transformation

$$v(t) = u(t) - \exp(-t)f(\hat{x})$$

can be used to transform (1.2) into the form (1.3). The work in Chapter 2 builds an algorithm based on sinc functions that defines a global numerical solution to (1.3), and a convergence proof for the method is given. This global numerical solution is in sharp contrast to the well known finite difference and finite element procedures for (1.3).

Since the sinc function

$$\operatorname{sinc}(x) \equiv \begin{cases} \dfrac{\sin(\pi x)}{\pi x}, & x \neq 0 \\ 1, & x = 0 \end{cases} \tag{1.4}$$

is defined on the entire real line, a convenient starting point for the development of a collocation procedure for (1.3) is to consider

$$\begin{aligned} u'(x) &= f(x, u(x)), \quad -\infty < x < \infty \\ \lim_{x \to \infty} u(x) &= 0 \quad . \end{aligned} \tag{1.5}$$

The basis functions used throughout this work are derived from (1.4) by translation: for each integer j and a mesh size h the sinc basis functions are defined on $\mathbb{R}$ by

$$S_j(x) \equiv \begin{cases} \dfrac{\sin\left[(\frac{\pi}{h})(x - jh)\right]}{\left[(\frac{\pi}{h})(x - jh)\right]}, & x \neq jh \\ 1, & x = jh \end{cases} \quad . \tag{1.6}$$

If an approximate solution of the form

$$u_m(x) = \sum_{j=-M}^{M-1} c_j S_j(x), \quad m = 2M \tag{1.7}$$

is substituted into (1.5) then a collocation procedure is defined by evaluating the result at the nodes $x_k = kh$. This gives rise to the $m = 2M$ equations

$$\sum_{j=-M}^{M-1} c_j S_j'(x_k) = f(x, u_m(x_k)), \quad k = -M, \ldots, M-1 \tag{1.8}$$

whose solution c_j, $j = -M, \ldots, M-1$, defines the coefficients for the approximate solution (1.7). In Chapter 2 this system of equations is written in matrix form and a thorough discussion of the matrix equation, including a proof of convergence of (1.8) to the solution of (1.5), will be given. Fundamental to the convergence proof are the known spectral properties of Toeplitz matrices. A discussion of these properties is also included in Chapter 2. Having developed a method for (1.5), a conformal mapping is used to address the problem (1.3). This conformal mapping maintains the Toeplitz structure of the coefficient matrix and as a consequence the convergence proof need not be repeated. Examples are included which illustrate the proven convergence rate. Implementation issues arising from problems involving nonlinearities and nonzero steady states are addressed.

In Chapter 3 attention is turned to the spatial problem associated with (1.1), which is obtained by fixing $t = \hat{t}$. Doing this, one obtains a boundary value problem of the form

$$\begin{aligned} -u''(x) + p(x,u)u'(x) &= f(x), \quad a < x < b \\ \alpha_1 u'(a) - \alpha_0 u(a) &= 0, \\ \beta_1 u'(b) + \beta_0 u(b) &= 0 \quad . \end{aligned} \tag{1.9}$$

This nonlinear problem has received less attention both computationally and analytically than has the linear problem

$$\begin{aligned} -u''(x) + p(x)u'(x) + q(x)u(x) &= f(x), \quad a < x < b \\ \alpha_1 u'(a) - \alpha_0 u(a) &= 0, \\ \beta_1 u'(b) + \beta_0 u(b) &= 0 \quad . \end{aligned} \tag{1.10}$$

The use of sinc methods for differential equations originated with the work [19], which announced the Sinc–Galerkin method for boundary value problems. Since

that time a great deal of attention has been devoted to this spatial problem. A Sinc-collocation procedure was implicated in [19] and was outlined in the review paper [20]. This outline provided the motivation for the collocation method in [15] which addressed the eigenvalue computation for the radial Schrödinger equation. This work was expanded to include other Sturm–Louiville eigenvalue problems associated with (1.10) in [6]. The same discretization as found in [6] was studied for the boundary value problem (1.10) in [1]. These Sinc–collocation schemes and their relation to Sinc–Galerkin schemes were explicitly sorted out for (1.10) in [13]. In [21], Stenger shows his original Sinc–Galerkin scheme and the collocation scheme used in this thesis are the same in the sense that they converge at the same rate.

In Chapter 3, a brief review of Stenger's Sinc–Galerkin procedure and convergence theorem is given to identify the class of functions in which the sinc approximation can be expected to give an exponential convergence rate of the approximation to the true solution of (1.10). Although the discrete systems for the Sinc–Galerkin and Sinc–collocation methods are different, an example indicating the parameter selections for the methods shows that they are numerically equivalent. If the nonlinearity $p(x, u)$ in (1.9) is replaced by a constant κ, where κ is large, the performance of the numerical method deteriorates due to boundary layer effects. A review of the error terms associated with the method, as undertaken in [4], yields a mesh selection that allows one to maintain the accuracy despite the boundary layer. The nonlinearity in (1.9) adds yet another numerical difficulty as seen by the introduction of a Hadamard product in the resulting matrix system. A simple iterative scheme, as suggested in [13] and [21], naturally suggests itself as a solution method. It is numerically demonstrated that for moderately large values of κ, this iterative procedure breaks down and is abandoned in favor of Newton's method. The combination of the breakdown of the simple iterative procedure and the introduction of Newton's method motivates the

lengthy and important Example 3.5. The length of Example 3.5 is due to the entry of a Hadamard product into the discretization, and the importance lies in the discussion of the Jacobian calculation for Hadamard products, which is fundamental to the discretization of nonlinear problems. It is shown that Newton's method, combined with an alternative mesh selection, maintains the accuracy of the Sinc–collocation method for very large values of κ. In the last section of Chapter 3 the radiation boundary conditions are incorporated in (1.9) and the necessary modifications of the approximation procedure are developed and implemented.

The final chapter assembles the work of Chapter 2 and Chapter 3 for a full discretization of (1.1), leading to a nonlinear Sylvester equation. As was done in the spatial domain, a sequence of simpler problems leading up to the discretization of (1.1) is addressed. This begins with the heat equation subject to Dirichlet boundary conditions which was addressed in [12] via a fully Sinc–Galerkin scheme. The choice of weight function in these schemes does not allow one to address nonzero steady states which is one of the goals of this thesis. The method developed in this thesis can compute both zero and nonzero steady states. The method discussed in [2] adds an advective term to the heat equation and discusses the efficiency of solving Sylvester equations. The Sylvester equation, its solvability, and a method of solution are discussed for the discretization of the the linear problem. A method for tracking steady state solutions gives rise to bordered matrices in the Sylvester equation for the same problem.

When considering the boundary layer effects in the partial differential equation (ϵ small), the same problems as those occurring in the boundary value problem of Chapter 3 arise. A nonlinear Sylvester equation appears and simplicity of computer implementation dictates an iterative solution procedure. For moderately large ϵ this procedure works fine but comes at the expense of the inability of the procedure to

compute solutions for small values of ϵ. Use of the concatenation operator allows one to view the nonlinear Sylvester system in a block structure. Each of the blocks in this system is similar to that arising from the scalar problem discussed in Example 3.5. The Newton method given in Example 3.5 is used to outline a block iterative procedure which could be used to solve the concatenated system. A similar point of view was taken in [15] when dealing with linear elliptic equations. As advocated in that work and supported here, these block calculations should be done on a parallel computing machine. This author does not underestimate this programming task, and has therefore included an outline for the algorithm.

CHAPTER 2

Temporal Discretization

In this chapter, a Sinc–collocation method for the initial value problem

$$\begin{aligned} \frac{du(t)}{dt} &= f(t,u(t)), \quad t > a \\ u(a) &= 0 \end{aligned} \tag{2.1}$$

is developed. A global approximation of the solution of (2.1), which is valid for $t \in [a,b)$, is obtained using 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

$$S_j(x) \equiv \begin{cases} \dfrac{\sin\left[(\frac{\pi}{h})(x-jh)\right]}{\left[(\frac{\pi}{h})(x-jh)\right]}, & x \neq jh \\ 1, & x = jh \end{cases} \tag{2.2}$$

The sinc functions form an interpolatory set of functions. In other words,

$$S_j(kh) = \delta_{jk}^{(0)} = \begin{cases} 1\,, & \text{if } j = k \\ 0\,, & \text{if } j \neq k \end{cases} \tag{2.3}$$

Since these 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

$$\begin{aligned} \frac{du(x)}{dx} &= f(x,u(x)), \quad -\infty < x < \infty \\ \lim_{x\to-\infty} u(x) &= 0. \end{aligned} \tag{2.4}$$

The basis functions in (2.2) automatically satisfy the limiting condition in (2.4) so that the assumed approximate solution

$$u_m(x) = \sum_{j=-M}^{M-1} c_j S_j(x) \ , \quad m = 2M \tag{2.5}$$

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

$$\lim_{x \to \infty} u(x) = 0 \quad . \tag{2.6}$$

The assumed approximate solution (2.5) automatically satisfies (2.6) as well. Until otherwise stated, it is assumed that the solution of (2.4) satisfies (2.6).

A collocation scheme is defined by substituting (2.5) into (2.4) and evaluating the result at $x_k = kh$, $k = -M, \ldots, M-1$. This gives the equation

$$\frac{1}{h} I_m^{(1)} \vec{c} = -f(\vec{x}, \vec{c}) \quad , \tag{2.7}$$

where the $m \times 1$ vectors $\vec{x} = [x_{-M}, \ldots, x_{M-1}]^t$ and $\vec{c} = [c_{-M}, \ldots, c_{M-1}]^t$ denote the vectors of nodes and coefficients in (2.5), respectively. The coefficient matrix in (2.7) is obtained from the explicit values for the derivative of the sinc basis functions at the nodes:

$$\left. \frac{dS_j(x)}{dx} \right|_{x=x_k=kh} = \frac{1}{h} \delta_{jk}^{(1)} = \frac{1}{h} \begin{cases} 0 \ , & \text{if } \ j = k \\ \dfrac{(-1)^{k-j}}{k-j} \ , & \text{if } \ j \neq k \end{cases} \ . \tag{2.8}$$

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 (2.7)

$$I_m^{(1)} = \begin{bmatrix} 0 & -1 & \frac{1}{2} & -\frac{1}{3} & \cdots & -\frac{1}{2M-1} \\ 1 & 0 & -1 & \frac{1}{2} & \cdots & \frac{1}{2M-2} \\ -\frac{1}{2} & 1 & 0 & -1 & \cdots & -\frac{1}{2M-3} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ -\frac{1}{2M-2} & \frac{1}{2M-3} & \cdots & 1 & 0 & -1 \\ \frac{1}{2M-1} & -\frac{1}{2M-2} & \frac{1}{2M-3} & \cdots & 1 & 0 \end{bmatrix}_{m \times m} \ . \tag{2.9}$$

The procedure then is to solve the system (2.7) for the $m \times 1$ vector of coefficients $\vec{c}$ in (2.5). The discrete system in (2.7) can also be obtained via a Sinc–Galerkin procedure as outlined in [13]. 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], [15], and for the two–point boundary value problem in [1], [17]. These procedures, as well as an extensive summary of the properties of sinc approximation, can be found in [21].

In this chapter it is shown that if the function $f(x, u(x))$ is continuously differentiable and $u(x)$ is in the appropriate class of functions for which sinc interpolation is exponentially accurate, then there exists a unique solution $\vec{c}$ to (2.21) so that

$$\|\vec{u} - \vec{c}\| \leq KM^2 \exp(-\kappa\sqrt{M}) \quad , \tag{2.10}$$

where $\vec{u} = [u_{-M}, \ldots, u_{M-1}]^t$. Furthermore, the error between the approximation defined by (2.5) and the solution $u(x)$ to (2.4) satisfies

$$\|u - u_m\| \leq \widehat{K}M^2 \exp(-\kappa\sqrt{M}) \quad , \tag{2.11}$$

where $K, \widehat{K}$ and κ are positive constants. The notation $\| \ \|$ used throughout this thesis denotes the discrete or continuous two norm. In the discrete case,

$$\|\vec{u}\| = \left(\sum_{k=1}^{n} u_k^2 \right)^{\frac{1}{2}} \quad ,$$

where $\vec{u}$ is a vector of length n. In the continuous case,

$$\|u\| = \left(\int_a^b (u(x))^2 \, dx \right)^{\frac{1}{2}} \quad ,$$

where $u(x)$ is a function defined on the interval (a, b). The proof of the estimate (2.11) 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.

The convergence proof which gives the order statement in (2.10) also applies to problems on an interval (a, b) via the method of conformal mapping. The case of the mapping $x = \Upsilon(t) = \ln(t)$, $t \in (0, \infty)$ is addressed in the final section of this chapter. 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 a rational rate.

If the time domain is the half–line, the sinc basis functions in (2.2) are replaced by

$$S_j \circ \Upsilon(t) \equiv \begin{cases} \dfrac{\sin[(\pi/h)(\Upsilon(t) - jh)]}{[(\pi/h)(\Upsilon(t) - jh)]}, & \Upsilon(t) \neq jh \\ 1, & \Upsilon(t) = jh \end{cases} . \tag{2.12}$$

With this alteration the approximation procedure is the same. Assume an approximate solution of (2.1) of the form

$$u_m(t) = \sum_{j=-M}^{M-1} c_j S_j \circ \Upsilon(t) \ , \quad m = 2M \ . \tag{2.13}$$

Substitute (2.13) into (2.1) and evaluate the result at the nodes $t_k = \Upsilon^{-1}(x_k)$ for $k = -M, \ldots, M-1$. This leads to the equation

$$\frac{1}{h} I_m^{(1)} \vec{c} = -\mathcal{D}\left(\frac{1}{\Upsilon'}\right) f(\vec{t}, \vec{c}) \tag{2.14}$$

where, given a function $g(t)$ defined on the nodes $t_k, k = -M, \ldots, M-1$, the notation $\mathcal{D}(g)$ denotes a $2M \times 2M$ (or $m \times m$) diagonal matrix with the k^{th} diagonal entry given by $g(t_k)$. One of the implementation conveniences of this sinc procedure is that the only alteration in (2.14) to the numerical procedure given in (2.7) is the introduction of a diagonal matrix on the right–hand side. 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. The method is implemented in the last section of this chapter.

Collocation on $\mathbb{R}$

In this section the convergence rate given in (2.10) is obtained for the problem

$$\begin{aligned} u'(x) &= f(x,u(x)), \quad -\infty < x < \infty \\ \lim_{x\to-\infty} u(x) &= 0 \quad . \end{aligned} \tag{2.15}$$

The space of functions where the sinc approximate given by (2.5) yields an exponential discretization error is given in the following definition.

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

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

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 [21]. For the present work, the following interpolation and quadrature theorems play a key role.

Theorem 2.2 Interpolation: Assume that the function $u \in \mathcal{H}^2(\mathcal{D}_d)$. Then for all $z \in \mathcal{D}_d$,

$$\begin{aligned} E(u,h)(z) &\equiv u(z) - \sum_{k=-\infty}^{\infty} u(kh) S_k(z) \\ &= \frac{\sin(\pi z)}{2\pi i} \int_{-\infty}^{\infty} \left\{ \frac{u(s-id^-)}{(s-z-id^-)\sin(\pi(s-id^-)/h)} \right. \\ &\quad \left. - \frac{u(s+id^-)}{(s-z+id^-)\sin(\pi(s+id^-)/h)} \right\} ds \end{aligned} \tag{2.16}$$

and

$$\|E(u,h)\| \leq \frac{\mathcal{N}^2(u,\mathcal{D}_d)}{\sinh(\pi d/h)} = \mathcal{O}\left(e^{-\pi d/h}\right) . \tag{2.17}$$

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

$$|u(x)| \leq K_1 \exp(-\alpha|x|) \quad x \in \mathbb{R} . \tag{2.18}$$

If the mesh selection

$$h = \sqrt{\frac{\pi d}{\alpha M}} , \tag{2.19}$$

is made in the finite sinc expansion

$$C_m(u)(x) = \sum_{j=-M}^{M-1} u(x_j) S_j(x) \tag{2.20}$$

that interpolates $u(x)$, then the error is bounded by

$$\|u - C_m(u)\| \leq K_2 M \exp(-\sqrt{\pi d \alpha M}) . \tag{2.21}$$

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,

$$|\eta| \leq \frac{\mathcal{N}^2(u,\mathcal{D}_d)e^{-\pi d/h}}{2\sinh(\pi d/h)} = \mathcal{O}\left(e^{-2\pi d/h}\right) . \tag{2.22}$$

One obtains, upon differentiating (2.16), the identity

$$\begin{aligned} u'(x) &- \sum_{j=-M}^{M-1} u(jh) S_j'(x) = \sum_{\substack{|j|>M \\ j=M}} u(jh) S_j'(x) \\ &+ \frac{d}{dx}\left[\frac{\sin(\pi x)}{2\pi i} \int_{-\infty}^{\infty} \frac{u(s-id^-)}{(s-x-id^-)\sin(\pi(s-id^-)/h)} \right. \\ &- \left. \frac{u(s+id^-)}{(s-x+id^-)\sin(\pi(s+id^-)/h)} ds \right] , \end{aligned} \tag{2.23}$$

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 [16] that

$$\begin{aligned} & \left| \frac{d}{dx} \left[\frac{\sin(\pi x)}{2\pi i} \int_{-\infty}^{\infty} \frac{u(s-id^-)}{(s-x-id^-)\sin(\pi(s-id^-)/h)} \right.\right. \\ - & \left.\left. \frac{u(s+id^-)}{(s-x+id^-)\sin(\pi(s+id^-)/h)} ds \right] \right| \\ \leq & \frac{K_3}{h} \exp(-\pi d/h) \, . \end{aligned} \tag{2.24}$$

A short calculation gives the bound

$$|S_j'(x)| = \left| \frac{dS_j(x)}{dx} \right| \leq \frac{\pi}{2h}, \quad x \in \mathbb{R}. \tag{2.25}$$

There will be a need for a similar bound on the second derivative of the sinc function later in this work and so it is displayed here:

$$\begin{aligned} & \left| \frac{d^2}{dx^2} \left[\frac{\sin(\pi x)}{2\pi i} \int_{-\infty}^{\infty} \frac{u(s-id^-)}{(s-x-id^-)\sin(\pi(s-id^-)/h)} \right.\right. \\ - & \left.\left. \frac{u(s+id^-)}{(s-x+id^-)\sin(\pi(s+id^-)/h)} ds \right] \right| \\ \leq & \frac{K_3}{h^2} \exp(-\pi d/h) \, , \end{aligned} \tag{2.26}$$

and

$$|S_j''(x)| = \left| \frac{d^2 S_j(x)}{dx^2} \right| \leq \frac{\pi^2}{3h^2}, \quad x \in \mathbb{R}. \tag{2.27}$$

Combining (2.25) with (2.18) gives the following bound on the truncation error:

$$\begin{aligned} \left| \sum_{|j|>M} u(jh) S_j'(x) \right| & \leq \sum_{|j|>M} \left| u(jh) S_j'(x) \right| \leq \frac{\pi}{2h} \sum_{j=M+1}^{\infty} |u(jh)| \\ & \leq \frac{K_1}{h} \sum_{j=M+1}^{\infty} |\exp(-\alpha j h)| \\ & = \frac{K_1}{h} \left(\frac{\exp(-\alpha h)}{1-\exp(-\alpha h)} \right) \exp(-\alpha M h) \\ & \leq \frac{K_1}{\alpha h^2} \exp(-\alpha M h) \leq \frac{K_4}{h^2} \exp(-\alpha M h) \, , \end{aligned} \tag{2.28}$$

where the fact that

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

yields the first inequality in the last line of (2.28).

Collocation, when applied to the initial value problem (2.15), requires that $u'(x_k) = f(x_k, u(x_k))$. Evaluating (2.23) at the nodes, and using the approximation implied there, one gets the system

$$N_m(\vec{u}) = \frac{1}{h} I_m^{(1)} \vec{u} + f(\vec{x}, \vec{u}) \quad . \tag{2.29}$$

The inequalities in (2.24) and (2.28) show that the k^{th} component of (2.29) is bounded by

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

where the mesh selection h in (2.19) was used to obtain the second inequality. Therefore,

$$\begin{aligned} \|N_m(\vec{u})\| &= \left(\sum_{k=-M}^{M-1} |N_m(u_k)|^2 \right)^{1/2} \\ &\leq \sqrt{2M} \max_{-M \leq k \leq M-1} |N_m(u_k)| \\ &\leq K_5 M^{3/2} \exp(-\sqrt{\pi d \alpha M}) \; . \end{aligned} \tag{2.30}$$

Theorem 2.5 Assume that the function $u \in \mathcal{H}^2(\mathcal{D}_d)$, u solves (2.15), and u satisfies (2.18). 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. Then in a sufficiently small ball about $u(x)$, the function

$$u_m(x) = \sum_{j=-M}^{M-1} c_j S_j(x) \; , \tag{2.31}$$

where the coefficients are determined by solving the equation

$$N_m(\vec{c}) \equiv \frac{1}{h} I_m^{(1)} \vec{c} + f(\vec{x}, \vec{c}) = \vec{0} \,, \tag{2.32}$$

satisfies

$$\|u_m - u\| \leq K_6 M^2 \exp(-\sqrt{\pi d \alpha M}) \,. \tag{2.33}$$

The proof of Theorem 2.5 depends on the orthogonality of the sinc basis. To see this, let $\vec{u} = [u(x_{-M}), \ldots, u(x_{M-1})]^t$ be the vector of coefficients in the sinc expansion (2.20). The equality of function and vector norms

$$\|u_m - C_m(u)\| = \|\vec{c} - \vec{u}\|$$

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

$$\begin{aligned} \|u_m - u\| &\leq \|u_m - C_m(u)\| + \|C_m(u) - u\| \\ &= \|\vec{c} - \vec{u}\| + \|C_m(u) - u\| \\ &\leq \|\vec{c} - \vec{u}\| + K_2 M \exp(-\sqrt{\pi d \alpha M}) \quad , \end{aligned} \tag{2.34}$$

where the last inequality follows from (2.21). It remains to bound the error in the coefficients $\|\vec{c} - \vec{u}\|$ which is addressed in the following two lemmas. These two lemmas will then complete the proof of Theorem 2.5.

Lemma 2.6 Assume that the function $u \in \mathcal{H}^2(\mathcal{D}_d)$ and satisfies (2.18). 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. Then in a sufficiently small ball about $\vec{u}$ there is a unique solution $\vec{c}$ to (2.32) which satisfies the inequality

$$\|\vec{c} - \vec{u}\| \leq K_5 M^2 \exp(-\sqrt{\pi d \alpha M}) \,. \tag{2.35}$$

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

$$L_m[\vec{u}] \equiv \frac{1}{h} I_m^{(1)} + \mathcal{D}(f_u(\vec{x}, \vec{u})) \tag{2.36}$$

which, in turn, depends on the norm of the inverse of the matrix $I_m^{(1)}$. This estimate is obtained with the help of the following lemma.

Lemma 2.7 Let ie_1 be the pure 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

$$\|(I_m^{(1)} + \mathcal{D})^{-1}\| \leq \frac{1}{e_1} = \|(I_m^{(1)})^{-1}\| \leq \frac{1}{\cos\left(\frac{M\pi}{2M+1}\right)} \leq 2M \quad . \tag{2.37}$$

Since $I_m^{(1)}$ has real entries and is skew-symmetric, its eigenvalues are pure 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 $\vec{z} \in \mathbb{C}^{2M}$

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

since e_1 and $\mathcal{D}$ are real. This implies that

$$\|(I_m^{(1)} + \mathcal{D})^{-1}\| \leq \frac{1}{|e_1|} = \|(I_m^{(1)})^{-1}\|$$

and yields the first inequality in (2.37). The proof of the second inequality in (2.37) is not so straightforward and follows as a consequence of the Toeplitz structure of the matrix $I_m^{(1)}$. A proof of the last inequality in (2.37) follows the proof of Lemma 2.6.

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

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

Lemma 2.7 shows that the function $L_m^{-1}[\vec{u}]$ in (2.36) exists and its norm is bounded by

$$\|L_m^{-1}[\vec{u}]\| = \left\| \left[\frac{1}{h} I_m^1 + \mathcal{D}(f_u(\vec{x},\vec{u})) \right]^{-1} \right\| \leq h(2M) = K_6\sqrt{M} \tag{2.38}$$

where the mesh size in (2.20) yields the last inequality. It follows that a fixed point of F_m gives a solution of (2.32). Let $\vec{v} \in B_r(\vec{u})$, then the calculation

$$\begin{aligned} & \|F_m(\vec{v}) - \vec{u}\| = \|\vec{v} - \vec{u} - L_m^{-1}[\vec{u}]N_m(\vec{v})\| \\ = & \left\| \vec{v} - \vec{u} - L_m^{-1}[\vec{u}] \left[N_m(\vec{u}) + \left(\int_0^1 \frac{\partial}{\partial u} N_m(t\vec{v} + (1-t)\vec{u}) dt \right) (\vec{v} - \vec{u}) \right] \right\| \\ \leq & \ \|L_m^{-1}[\vec{u}]N_m(\vec{u})\| \\ + & \left\| L_m^{-1}[\vec{u}] \left(\int_0^1 L_m[\vec{u}] - \frac{\partial}{\partial u} N_m(t\vec{v} + (1-t)\vec{u}) dt \right) (\vec{v} - \vec{u}) \right\| \end{aligned} \tag{2.39}$$

follows from the Taylor polynomial for the function N_m and the triangle inequality. The first term following the last inequality in (2.39) can be bounded by the product of the right–hand sides of (2.30) and (2.38).

Now consider bounding the second term following the last inequality on the right–hand side of (2.39). Using the assumed Lipschitz continuity of f_u leads to

$$\begin{aligned} & \left\| L_m^{-1}[\vec{u}] \left(\int_0^1 L_m[\vec{u}] - \frac{\partial}{\partial u} N_m(t\vec{v} + (1-t)\vec{u}) dt \right) (\vec{v} - \vec{u}) \right\| \\ = & \left\| L_m^{-1}[\vec{u}] \left(\int_0^1 \mathcal{D}\left(f_u(\vec{x},\vec{u})\right) - f_u(\vec{x}, t\vec{v} + (1-t)\vec{u}) \, dt \right) (\vec{v} - \vec{u}) \right\| \\ \leq & \ \|L_m^{-1}[\vec{u}]\| K_L r^2 \ . \end{aligned} \tag{2.40}$$

Substituting (2.40) in the right hand side of (2.39) leads to the inequality

$$\|F_m(\vec{v}) - \vec{u}\| \ \leq \ \|L_m^{-1}[\vec{u}]\| \left(\|N_m(\vec{u})\| + K_L r^2 \right)$$

$$\leq K_6\sqrt{M}\left[\left(K_4M^{3/2}\right)\exp(-\sqrt{\pi d\alpha M})+hK_Lr^2\right]$$
$$\leq K_7M^2\exp(-\sqrt{\pi d\alpha M})+\sqrt{M}\hat{K}_Lr^2 \qquad (2.41)$$

where (2.30) and (2.38) yield the second inequality. The quadratic inequality

$$K_7M^2\exp(-\sqrt{\pi d\alpha M})+\sqrt{M}\hat{K}_Lr^2<r$$

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

$$r_0=\mathcal{O}(M^2\exp(-\sqrt{\pi d\alpha M}))<r_1=\mathcal{O}\left(\frac{1}{\sqrt{M}}\right)\ . \qquad (2.42)$$

since $M^2\exp(-\sqrt{\pi d\alpha M})\to 0$ 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 $\vec{c},\vec{v}\in B_r(\vec{u})$, then

$$\|F_m(\vec{v})-F_m(\vec{c})\|=\|\vec{v}-\vec{c}-L_m^{-1}[\vec{u}]\,(N_m(\vec{v})-N_m(\vec{c}))\,\|$$
$$=\|L_m^{-1}[\vec{u}]\,(L_m[\vec{u}](\vec{v}-\vec{u})-(N_m(\vec{v})-N_m(\vec{c}))\,\|$$
$$\leq\|L_m^{-1}[\vec{u}]\|\,\|\mathcal{D}\left\{f_u(\vec{x},\vec{u})(\vec{v}-\vec{c})-[f(\vec{x},\vec{v})-f(\vec{x},\vec{c})]\right\}\|$$
$$=\|L_m^{-1}[\vec{u}]\|\,\left\|\int_0^1\mathcal{D}\left\{f_u(\vec{x},t\vec{u}+(1-t)\vec{u})-f_u(\vec{x},t\vec{v}+(1-t)\vec{c})\right\}dt\,(\vec{v}-\vec{c})\right\|$$
$$\leq 2rK_L\|L_m^{-1}[\vec{u}]\|\,\|\vec{v}-\vec{c}\|\quad,$$

where K_L is a Lipschitz constant for f_u. From (2.38), (2.42) and

$$2hrK_L\|L_m^{-1}[\vec{u}]\|=\mathcal{O}(M^2\exp(-\sqrt{\alpha\pi dM}))\quad,$$

it follows for sufficiently small $r=\mathcal{O}(M^2\exp(-\sqrt{\alpha\pi dM}))$ that F_m is a contraction on $B_r(\vec{u})$ and F_m has a unique fixed point. This completes the proof of Lemma 2.6.

In order to establish the invertability of the matrix $I_m^{(1)}, m=2M$ in (2.9), it is convenient to use the theorem of Otto Toeplitz [9].

Theorem 2.8 Toeplitz Denote the Fourier coefficients of the real-valued function $f \in L^1(-\pi, \pi)$ by

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

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

$$\mathcal{C}_m(f) \equiv \begin{bmatrix} f_0 & f_1 & f_2 & \cdots & f_{m-1} \\ f_{-1} & f_0 & f_1 & \cdots & f_{m-2} \\ f_{-2} & f_{-1} & f_0 & \ddots & f_{m-3} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ f_{-m+1} & \cdots & f_{-2} & f_{-1} & f_0 \end{bmatrix}_{m \times m} . \tag{2.43}$$

Denote the real eigenvalues of the Hermitian matrix $\mathcal{C}_m(f)$ 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,

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

Further, if $\mathcal{C}_m(g)$ is the Toeplitz matrix of the real–valued function $g \in L^1(-\pi, \pi) \cap C[\pi, \pi]$ and $g(x) \leq f(x)$, then for all j,

$$c_j^m \leq e_j^m \quad , \tag{2.44}$$

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

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

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

so that upon comparing these coefficients with the entries of the matrix $I_m^{(1)}, m = 2M$ in (2.9), one sees that for $f(x) = x$ the Toeplitz matrix $C_m(f) = iI_m^{(1)}$. The eigenvalues of the real skew–symmetric matrix $I_m^{(1)}$ occur in conjugate pairs $\{\pm ie_p^m\}_{p=1}^M$ and the nonnegative real numbers, e_p^m, satisfy the inequality

$$-\pi \leq -e_M^m \leq \ldots \leq -e_1^m \leq e_1^m \leq \ldots \leq e_M^m \leq \pi \ . \tag{2.45}$$

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

$$C_m(g) \equiv \frac{1}{2i}\begin{bmatrix} 0 & -1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & -1 \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}_{m\times m} . \tag{2.46}$$

The eigenvalues of the real skew–symmetric matrix $iC_m(g)$ also occur in conjugate pairs $\{\pm ic_p^m\}_{p=1}^M, m = 2M$. 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

$$0 < c_1^m \leq c_2^m \leq \ldots \leq c_M^m < 1 \ . \tag{2.47}$$

The inequality $g(x) = \sin(x) \leq x = f(x)$ is satisfied on the interval $[0, \pi]$, so that using (2.44) and (2.45) with (2.47) gives

$$\begin{aligned} \min_{j=1,\ldots,M} e_j^m = e_1^m \geq c_1^m &= \cos\left(\frac{M\pi}{2M+1}\right) \\ &= \sin\left(\frac{\pi}{2(2M+1)}\right) \\ &\geq \frac{1}{2M} \ . \end{aligned} \tag{2.48}$$

Hence, it follows that

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

This completes the proof of Lemma 2.7.

Due to the upper bound in (2.45) for the eigenvalues of the matrix $I_m^{(1)}$, the spectral condition number of this matrix is

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

The following example clearly exposes the various parameter selections yielding the mesh selection h in (2.19) and also illustrates the close connection of this method with the method found in [21].

Example 2.9 The function

$$u(z) = \frac{1}{\cosh(\pi z)} \tag{2.49}$$

is analytic in a strip of width one (the poles of $u(z)$ closest to the real line occur 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 (2.18) with $K_1 = 2$ and $\alpha = \pi$ and is the unique solution to the problem

$$\begin{aligned} u'(x) &= \frac{-\pi \sinh(\pi x)}{\cosh^2(\pi x)} \quad -\infty < x < \infty \\ \lim_{x \to -\infty} u(x) &= 0 \quad . \end{aligned} \tag{2.50}$$

The function in (2.49) satisfies the auxiliary assumption $\lim_{x \to \infty} u(x) = 0$ so that Theorem 2.5 applies. Hence, setting $d = 1/2$ and $\alpha = \pi$ leads to the mesh size $h = \sqrt{1/(2M)}$. The coefficients $\{c_j\}_{j=-M}^{M-1}$ in (2.31) are obtained by solving the system

$$\frac{1}{h} I_m^{(1)} \vec{c} = -g(\vec{x}) = \frac{\pi \sinh(\pi \vec{x})}{\cosh^2(\pi \vec{x})} \quad . \tag{2.51}$$

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

$$ERR(M) = \|\vec{u} - \vec{c}\| \quad , \tag{2.52}$$

which, due to the factor M^2 in (2.35) and the inequality in (2.34), represents the dominant error contribution to $\|u - u_m\|$.

M	$ERR(M)$
4	7.9514e-02
8	1.6165e-02
16	1.6267e-03
32	5.6978e-05
64	4.3819e-07
128	3.9179e-10

Table 1: Results for (2.50)

The development to this point has assumed that the solution of the initial value problem (2.15) vanishes at infinity. This limiting assumption is removed by appending an auxiliary basis function to the sinc expansion in (2.31). Define the basis function

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

and form the augmented approximate sinc solution

$$u_m(x) = \sum_{j=-M}^{M-2} c_j S_j(x) + c_{M-1}\omega_\infty(x) \ . \tag{2.53}$$

The additional basis function $\omega_\infty(x)$ satisfies

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

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

$$v(x) = u(x) - u_\infty\omega_\infty(x) \tag{2.54}$$

transforms the problem

$$\begin{aligned} u'(x) &= f(x, u(x)), \quad -\infty < x < \infty \qquad (2.55) \\ \lim_{x \to -\infty} u(x) &= 0 \end{aligned}$$

to the problem

$$v'(x) = f(x, v(x) + u_\infty \omega_\infty(x)) - u_\infty \omega'_\infty(x), \quad -\infty < x < \infty \qquad (2.56)$$

$$\lim_{x \to \pm\infty} v(x) = 0\,. \qquad (2.57)$$

If u_∞ is known then the method defined by (2.32) determines the $\{c_j\}_{j=-M}^{M-1}$ in the expansion

$$u_m(x) = \sum_{j=-M}^{M-1} c_j S_j(x)$$

and the result of Theorem 2.5 applies to the approximation of $v(x)$ in (2.56) by $u_m(x)$. If u_∞ is unknown, one approach which preserves the error of Theorem 2.6 is to replace this unknown by c_{M-1} in (2.54) and use the Quadrature Theorem 2.4 to write

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

Add this equation to the solution procedure to obtain the approximate value for c_{M-1}. Since the error in the quadrature theorem is the square of the error of interpolation, this procedure introduces no more error then the error in the method defined by (2.32).

Incorporating the above side condition in the approximate method to determine the coefficients in (2.56) is less convenient to implement than the following approach. Directly substitute the augmented approximate sinc solution (2.53) into

the differential equation (2.55) and evaluate this expansion at the $m = 2M$ nodes $x_k, k = -M, \ldots, 0, \ldots M-1$. This leads to the bordered matrix system

$$A\vec{c} = \left[\frac{1}{h} I^{(1)}_{m\times(m-1)} \mid -\vec{\omega^I_\infty}\right] \vec{c} = -f(\vec{x}, T_{\omega_\infty}\vec{c}) \ . \tag{2.58}$$

The notation $I^{(1)}_{m\times(m-1)}$ denotes a copy of $I^{(1)}_m$ without the last column. In (2.58) the vector $\vec{c} = [c_{-M}, \ldots, c_0, \ldots c_{M-2}, c_{M-1}]^t$ are the coefficients in (2.53). The approximate solution $\vec{u}_m$ is obtained from the transformation

$$\vec{u}_m = T_{\omega_\infty}\vec{c} \quad , \tag{2.59}$$

where the matrix T_{ω_∞} is defined by

$$T_{\omega_\infty} = \begin{bmatrix} 1 & 0 & \cdots & 0 & (\omega_\infty)_{-M} \\ 0 & 1 & \cdots & 0 & (\omega_\infty)_{-M+1} \\ 0 & 0 & 1 & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & (\omega_\infty)_{M-2} \\ 0 & 0 & \cdots & 0 & (\omega_\infty)_{M-1} \end{bmatrix} . \tag{2.60}$$

Since the matrix T_{ω_∞} has the explicit inverse

$$T^{-1}_{\omega_\infty} = \begin{bmatrix} 1 & 0 & \cdots & 0 & -\frac{(\omega_\infty)_{-M}}{(\omega_\infty)_{M-1}} \\ 0 & 1 & \cdots & 0 & -\frac{(\omega_\infty)_{-M+1}}{(\omega_\infty)_{M-1}} \\ 0 & 0 & 1 & \cdots & \vdots \\ 0 & 0 & \cdots & 1 & -\frac{(\omega_\infty)_{M-2}}{(\omega_\infty)_{M-1}} \\ 0 & 0 & \cdots & 0 & \frac{1}{(\omega_\infty)_{M-1}} \end{bmatrix} , \tag{2.61}$$

one may regard either the vector $\vec{c}$ or $\vec{u}_m$ as the unknown in (2.59).

The system in (2.58) is solved for the coefficients by applying Newton's method to the function

$$N_m(\vec{c}) = A\vec{c} + f(\vec{x}, T_{\omega_\infty}\vec{c}) \ . \tag{2.62}$$

If the matrix A satisfies the conclusion of Lemma 2.7, then Theorem 2.5 applies to the function $N_m(\vec{c})$ so that the rate of convergence of the present method is also given by (2.33). Although an argument verifying 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 $I_m^{(1)}$, the numerical results displayed in the next example provide compelling evidence for a version of Lemma 2.7 with $I_m^{(1)}$ replaced by the matrix A in (2.62).

Example 2.10 In this example, the function

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

is a solution to

$$\begin{aligned} u'(x) &= -u(x)^2 + g(x), \quad -\infty < x < \infty \qquad (2.63) \\ \lim_{x\to-\infty} u(x) &= 0 \end{aligned}$$

provided $g(x) = u(x)$. The coefficients $\vec{c}$ in the approximation $u_m(x)$ are found by solving (2.62), which takes the form

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

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

$$ERR(M) = \|\vec{u} - \vec{u}_m\| \qquad (2.64)$$

is displayed in the second column of Table 2.

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 (2.62) is

M	n	$ERR(M)$	$R((I_m^{(1)})^{-1})$	$R(A^{-1})$
4	6	1.2284e-01	5.19e-01	6.71e-01
8	6	2.5326e-02	5.13e-01	6.02e-01
16	7	2.6765e-03	5.09e-01	5.51e-01
32	8	9.7673e-05	5.06e-01	5.23e-01
64	9	7.7053e-07	5.03e-01	5.11e-01
128	10	6.9836e-10	5.02e-01	5.05e-01

Table 2: Results for (2.63)

independent of the problem (it only depends on the choice of $\omega_\infty(x)$), this comparison remains the same for other initial value problems.

Collocation on $\mathbb{R}^+$

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

$$\begin{aligned} u'(t) &= f(t,u(t)), \quad t>0 \\ u(0) &= 0 \end{aligned} \tag{2.65}$$

via the method of conformal mapping. Specifically, the map

$$z = \Upsilon(w) = \ell n(w), \quad w = e^z$$

is a conformal equivalence of the strip $\mathcal{D}_d$ in Definition 2.1 and the wedge

$$\mathcal{D}_W = \{w \in \mathbb{C}\colon w = re^{i\theta}, \ \ |\theta| < d \leq \pi/2\}\ . \tag{2.66}$$

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

Definition 2.11 The function $u(z)$ is in the space $\mathcal{H}^2(\mathcal{D}_W)$ if u is analytic in $\mathcal{D}_W$ and satisfies

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

and

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

A sinc approximate solution of (2.65) takes the form

$$u_m(t) = \sum_{j=-M}^{M-1} c_j S_j \circ \Upsilon(t), \quad m = 2M \quad , \tag{2.67}$$

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

$$S_j \circ \Upsilon(t) \equiv \frac{\sin[(\pi/h)\Upsilon(t) - jh]}{[(\pi/h)\Upsilon(t) - jh]} \ . \tag{2.68}$$

With this alteration, the derivation of the approximation procedure is the same as in the previous section. Substitute (2.67) into (2.65) and evaluate at the $m = 2M$ sinc nodes $\Upsilon^{-1}(x_k) \equiv t_k = \exp(kh), k = -M, \ldots, M-1$ to arrive at the discrete system

$$\frac{1}{h} I_m^{(1)} \vec{c} = -\mathcal{D}\left(\frac{1}{\Upsilon'}\right) f(\vec{t}, \vec{c}) \ . \tag{2.69}$$

The only difference between this matrix equation and the one presented in (2.32) is the diagonal matrix $\mathcal{D}(\frac{1}{\Upsilon'})$.

The importance of the class of analytic functions in Definition 2.11 lies in the fact that if $\Upsilon'(w)u(w) \in \mathcal{H}^2(\mathcal{D}_W)$ and there are positive constants α and K_1 so that

$$|u(t)| \leq K_1 \frac{t^\alpha}{(1+t)^{2\alpha}}, \quad t > 0 \quad , \tag{2.70}$$

then the sinc interpolant to $u(t)$ also satisfies (2.21) and (2.23). Since $u'(t_k) = f(t_k, u(t_k))$, it again follows that the error in the k^{th} component of the function

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

is bounded by

$$|N_m(u_k)| \leq \frac{K_3}{h}\exp(-\pi d/h) + \frac{K_4}{h^2}\exp\left(-\alpha M h\right) \quad . \tag{2.71}$$

Finally, the mesh selection

$$h = \sqrt{\frac{\pi d}{\alpha M}} \quad ,$$

when substituted into the right–hand side of (2.71), leads to the bound in (2.30) for $\|N_m(\vec{u})\|$ in (2.71).

Theorem 2.12 Assume that the function $\Upsilon'(w)u(w) \in \mathcal{H}^2(\mathcal{D}_W)$ and that the solution u of (2.65) satisfies (2.70). 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 $\vec{c}$ which provides the coefficients for $u_m(t)$ in (2.67) and $u_m(t)$ satisfies the inequality

$$\|u_m - u\| \leq KM^2\exp(-\sqrt{\pi d\alpha M}) \; . \tag{2.72}$$

The proof follows from Lemma 2.6 and Lemma 2.7 which remain valid with the stated assumptions and due to the fact that the coefficient matrix in (2.69) remains the same as in the previous section.

The assumed approximate solution $u_m(t)$ in (2.67) has the property that $\lim_{t\to\infty} u_m(t) = 0$ so that the method can only be expected to approximate initial value problems with the same property. This limiting assumption is removed by appending an auxiliary basis function to the sinc expansion in (2.67) and is discussed in the next example.

Example 2.13 Let γ be a real parameter in the family of initial value problems

$$\begin{aligned} u'(t) &= (1-\gamma t)\exp(-t), \quad t > 0 \\ u(0) &= 0 \quad . \end{aligned} \tag{2.73}$$

The solution is given by

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

and satisfies

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

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

Add the basis function

$$\omega_\infty(t) = \frac{t}{t+1} \tag{2.74}$$

to the sinc approximate (2.67) to obtain the approximate

$$u_m(t) = \sum_{j=-M}^{M-2} c_j S_j \circ \Upsilon(t) + c_{M-1}\omega_\infty(t) \; . \tag{2.75}$$

Substitute (2.75) into (2.65) and evaluate this result at the sinc nodes $t_k = \exp(kh)$, $k = -M, \ldots, M-1$. This yields the matrix system

$$A\vec{c} = -h\mathcal{D}\left(\frac{1}{\Upsilon'}\right) f(\vec{t}) \tag{2.76}$$

where

$$A = \left[\frac{1}{h} I^{(1)}_{m\times(m-1)} \Big| -\frac{\vec{\omega}'_\infty}{\vec{\Upsilon}'}\right] \; . \tag{2.77}$$

The approximate solution $\vec{u}_m$ is obtained from the transformation $\vec{u}_m = T_{\omega_\infty}\vec{c}$. The coefficients $c_k, k = -M, \ldots, M-1$, are assembled in the $m\times 1$ vector $\vec{c}$ and the matrix

$$T_{\omega_\infty} = \left[I_{m\times(m-1)} \Big| \vec{\omega}_\infty\right] \tag{2.78}$$

is the same as in (2.60) with ω_∞ replaced by (2.74). It is important that the system (2.76) calculates the limiting value when $\gamma = 1$, namely zero. For purposes of illustration, the system without the augmented basis function, (2.69), has also been used. The results of solving that system for the coefficients in (2.68) are given in Table 3 as well. If the bound on the inverse of A in (2.77) satisfies the conclusion of Lemma 2.7 then the results displayed in the above table are not specific to this example.

	Bordered	Unbordered
M	$ERR(M)$	$ERR(M)$
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

Table 3: Results using augmented and non–augmented approximation for the solution of (2.73) with $\gamma = 1$

In the general case, the discretization of the problem (2.65) takes the form

$$A\vec{c} = -\mathcal{D}\left(\frac{1}{\Upsilon'}\right) f(\vec{t}, T_{\omega_\infty}\vec{c}) \quad , \tag{2.79}$$

from which the coefficients in (2.75) are calculated and the approximation to the solution at the nodes is given by $u_m(t_k) = c_k + c_{M-1}\omega_\infty(t_k)$. In each of the following examples Newton's method is applied to the function

$$N_m(\vec{c}) = A\vec{c} + \mathcal{D}\left(\frac{1}{\Upsilon'}\right) f(\vec{t}, T_{\omega_\infty}\vec{c}) \; . \tag{2.80}$$

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

$$\vec{c}^{\,n+1} = \vec{c}^{\,n} + \vec{\delta}^{\,n} \quad , \tag{2.81}$$

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

$$-\mathcal{J}(N_m)(\vec{c}^{\,n})\vec{\delta}^{\,n} = N_m(\vec{c}^{\,n}) \quad . \tag{2.82}$$

The Jacobian of (2.80) is

$$\mathcal{J}(N_m)(\vec{c}) = A + \mathcal{D}\left(\frac{1}{\Upsilon'}\right)\mathcal{D}\left(\frac{\partial f}{\partial u}(\vec{t}, T_{\omega_\infty}\vec{c})\right) T_{\omega_\infty} \; . \tag{2.83}$$

Note that, besides the exponential rate of convergence given by (2.72), the computation involved for the Jacobian of the nonlinear system involves little work. In fact, from (2.83), the update of the Jacobian is simply a diagonal evaluation.

Example 2.14 The initial value problem

$$\begin{aligned} u'(t) &= -\left(\frac{u^2+4u+1}{2u+4}\right), \quad t>0 \\ u(0) &= 0 \end{aligned} \tag{2.84}$$

has the solution

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

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

$$u(t) = (2-\sqrt{3}) - \mathcal{O}(\exp(-t)) \quad \text{as} \quad t \to \infty. \tag{2.85}$$

The results in Table 4 display the number of Newton steps n in (2.81) and the two-norm error

$$ERR(M) = \|\vec{u}_m - \vec{u}\| .$$

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

Table 4: Results for (2.84)

A particularly useful application of the present procedure is in those initial value problems where the convergence to the asymptotic state is only of a rational rate. For example, an autonomous differential equation that has a non-hyperbolic rest point. The sinc approximation to such solutions also assumes rational decay at infinity so that the convergence estimate in (2.72) is maintained. This is illustrated in the following example.

Example 2.15 For small positive parameters β, the problem

$$\begin{aligned} u'(t) &= \beta(1-u)^2, \quad t>0 \\ u(0) &= 0 \end{aligned} \tag{2.86}$$

has the solution

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

The asymptotic behavior

$$u(t) - 1 \sim \frac{1}{\beta t} \quad \text{as} \quad t \to \infty \tag{2.87}$$

shows the rational rate of approach to the asymptotic state. In particular, for small β, this rate is quite slow compared to the rate of approach in the previous example given by (2.85).

	$ERR(M)$		$ERR(M)$		$ERR(M)$	
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

Table 5: Results for (2.86)

In Table 5 the error in the calculated solution of (2.86) is displayed for several values of β. As one reads the table from left to right (decreasing β), there are fewer Newton steps computed to achieve the error due 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 (2.71). For t large, the inequality in (2.70) implies

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

As seen from (2.87), $K_1 \sim 1/\beta$. Hence, as β is decreasing, the constant K_1 is increasing. In these cases (a rational rate of approach to the asymptotic state) a simple change in the definition of the mesh selection (2.73) produces an error bounded by $\exp(-(\delta\sqrt{M}))$, where $\delta \leq \alpha$. This alternative mesh selection, which defines a mesh reallocation, is also used in boundary layer problems and will be discussed and developed in Chapter 3.

CHAPTER 3

Spatial Discretization

Having discussed a method for the temporal domain in Chapter 2, attention is now turned toward a discretization of the spatial operator. Both the Galerkin and collocation methods are reviewed and discussed. Attention is given to the implementation of the two approaches when dealing with radiation boundary conditions, resolution of steep fronts, and nonlinearities.

A Sinc–Galerkin procedure first developed by Stenger [19] is reviewed with the focus of attention on problems of the form

$$\begin{aligned} -u''(x)+p(x)u'(x) &= f(x) \quad , \quad 0<x<1 \\ u(0)=u(1) &= 0 \quad . \end{aligned} \tag{3.1}$$

Numerous different approaches to this problem have been proposed in [3], [8], [10], and [18].

In order to have the sinc translates given by (2.2) defined on the interval $(0,1)$, consider the conformal map

$$\phi(z)=\ell n\left(\frac{z}{1-z}\right) \quad . \tag{3.2}$$

This map carries the eye-shaped region

$$\mathcal{D}_E=\left\{z=u+iv : \left|\arg\left(\frac{z}{1-z}\right)\right|<d\leq\frac{\pi}{2}\right\}$$

onto the infinite strip

$$\mathcal{D}_d=\left\{w=x+iy : |y|<d\leq\frac{\pi}{2}\right\} \quad .$$

A Sinc–Galerkin or Sinc–collocation approximate solution of (3.1) takes the form

$$u_m(x) = \sum_{k=-M}^{N} c_k S_k \circ \phi(x) \quad , \quad m = M + N + 1 \quad . \tag{3.3}$$

For a Galerkin scheme, the coefficients $\{c_k\}$ in (3.3) are determined by orthogonalizing the residual with respect to the basis functions

$$(-u_m'' + pu_m' - f, S_j \circ \phi) = 0 \quad , \quad -M \leq j \leq N \tag{3.4}$$

with the inner product given by

$$(u, v) = \int_0^1 u(x)v(x)w(x)dx \tag{3.5}$$

where $w(x)$ is, for the moment, an unspecified weight function.

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

$$\int_{\phi^{-1}(x+L)} |F(w)dw| = \mathcal{O}(|x|^a), \quad x \to \pm\infty, \quad 0 \leq a < 1 \tag{3.6}$$

where $L = \{iy : |y| < d\}$ and for γ a simple closed contour in $\mathcal{D}_E$

$$N^2(F, \mathcal{D}_E) \equiv \lim_{\gamma \to \partial\mathcal{D}_E} \int_\gamma |F(w)dw| < \infty. \tag{3.7}$$

Substituting (3.3) into (3.4) leads, after integrating by parts the terms involving derivatives of the dependent variable and choosing the weight function $w(x) = \frac{1}{\phi'(x)}$ to ensure that the boundary terms vanish, to the discrete linear system

$$A_G \vec{c} = \mathcal{D}\left(\frac{1}{(\phi')^2}\right)\vec{f} \tag{3.8}$$

where

$$A_G = \frac{-1}{h^2} I_m^{(2)} - \frac{1}{h} I^{(1)} \mathcal{D}\left(\frac{\phi''}{(\phi')^2} + \frac{p}{\phi'}\right) - \mathcal{D}\left(\frac{1}{\phi'}\left(\frac{1}{\phi'}\right)'' + \frac{1}{\phi'}\left(\frac{p}{\phi'}\right)'\right) \tag{3.9}$$

The matrix A_G is the matrix B in [21] on page 470 and is also found in [13] on page 166 using $r = 1$. A discussion of other choices for weight functions is found in [13].

The one matrix that hasn't been introduced yet in the above is $I_m^{(2)}$ which has the entries

$$\delta_{jk}^{(2)} \equiv h^2 \frac{d^2}{d\phi^2} [S_j \circ \phi(x)]\Big|_{x=x_k} = \begin{cases} \frac{-\pi^2}{3}, & j = k \\ \frac{-2(-1)^{k-j}}{(k-j)^2}, & j \neq k, \end{cases} \tag{3.10}$$

with $x_k = \phi^{-1}(kh)$. Define $I_m^{(2)} = [\delta_{jk}^{(2)}]$ where $m = M + N + 1$ giving,

$$I_m^{(2)} = \begin{bmatrix} -\frac{\pi^2}{3} & 2 & \frac{-2}{2^2} & \cdots & \frac{-2(-1)^{m-1}}{(m-1)^2} \\ 2 & & & & \vdots \\ \frac{-2}{2^2} & \ddots & \ddots & \ddots & \frac{-2}{2^2} \\ \vdots & & & & 2 \\ \frac{-2(-1)^{m-1}}{(m-1)^2} & \cdots & \frac{-2}{2^2} & 2 & -\frac{\pi^2}{3} \end{bmatrix}_{m \times m} . \tag{3.11}$$

Solutions obtained from (3.9) then have the exponential convergence rate guaranteed by the following theorem given in Chapter 7, Section 2.4 of [21].

Theorem 3.2 Assume that the functions p and f in

$$\begin{aligned} -u''(x) + p(x)u'(x) &= f(x), \quad 0 < x < 1 \\ u(0) = u(1) &= 0 \end{aligned}$$

and the unique solution u are analytic in the simply connected domain $\mathcal{D}_E$. Let ϕ be the conformal one-to-one map of $\mathcal{D}_E$ onto $\mathcal{D}_d$ given in (3.2). Assume also that $f/\phi' \in \mathcal{H}^2(\mathcal{D}_E)$ and $uF \in \mathcal{H}^2(\mathcal{D}_E)$ for each of

$$F = \phi', \quad (p/\phi')', \quad p. \tag{3.12}$$

Suppose there are positive constants $\hat{K}_\alpha$, $\hat{K}_\beta$, α, and β so that

$$|u(x)| \leq \begin{cases} \hat{K}_\alpha x^\alpha, & x \in (0; 1/2) \\ \hat{K}_\beta (1-x)^\beta, & x \in [1/2, 1). \end{cases} \tag{3.13}$$

If the $\{c_k\}_{k=-M}^{N}$ in

$$u_m(x) = \sum_{k=-M}^{N} c_k S_k \circ \phi(x)$$

are determined by solving (3.9) then

$$\begin{aligned} \|u - u_m\| &\leq K_\alpha M \exp(-\alpha M h) \\ &+ K_\beta N \exp(-\beta N h) \\ &+ K_I M^{5/2} \exp(-\pi d/h) \end{aligned} \tag{3.14}$$

where K_α, K_β, are constant multiples of $\hat{K}_\alpha$ and $\hat{K}_\beta$. These constants and K_I are independent of M, N, and h. Balancing the exponential contributions of the three terms on the right hand side of (3.14) yields the proper choices of h and N as

$$h = \left(\frac{\pi d}{\alpha M}\right)^{1/2}, \quad N = \left[\left|\frac{\alpha}{\beta} M\right|\right] \quad . \tag{3.15}$$

These choices then yield the error statement

$$\|u - u_m\| \leq K M^{5/2} \exp(-(\pi d \alpha M)^{1/2}) \tag{3.16}$$

where K is independent of M and h.

An outline of the proof proceeds as follows. If u is in $\mathcal{H}^2(\mathcal{D}_E)$ then the sinc interpolant to $u(x)$ satisfies (2.21). Moreover, its first derivative satisfies the bound in (2.24) with a similar bound for the second derivative. Let $\vec{c}$ be the unique solution of (3.8), then the two–norm of the vector $A_G\vec{u} - A_G\vec{c}$, which corresponds to the discretization error, is of the order $M^{1/2}\exp(-(\pi d\alpha M)^{1/2})$. From (3.8) it follows that

$$A_G\vec{u} - A_G\vec{c} \;=\; A_G\vec{u} - \mathcal{D}\left(\frac{1}{(\phi')^2}\right)\vec{f} \quad ,$$

and hence

$$\|\vec{u} - \vec{c}\| \leq \|A_G^{-1}\| \|A_G\vec{u} - \mathcal{D}\left(\frac{1}{(\phi')^2}\right)\vec{f}\| \quad . \tag{3.17}$$

Stenger in [21] shows that $\|h^2 A_G^{-1}\|$ is $\mathcal{O}(M^2)$. Curiously, the proof of this order statement depends upon considering a collocation scheme for (3.1).

This collocation scheme can be developed by substituting (3.3) into (3.1). Evaluating at the nodes x_k, $k = -M, \ldots, N$, yields the system

$$A\vec{c} = \vec{f} \tag{3.18}$$

where

$$A = \mathcal{D}\left((\phi')^2\right) A_C \tag{3.19}$$

and

$$A_C = \frac{-1}{h^2} I_m^{(2)} + \frac{1}{h}\mathcal{D}\left(\frac{\phi''}{(\phi')^2} - \frac{p}{\phi'}\right) I_m^{(1)} \tag{3.20}$$

The matrix A_C is the matrix A in [21] on page 468 and is the matrix $C(0)$ in [13] on page 171.

The matrices A_G and A_C are quite similar, and in fact, this similarity is used in [21] to show that the solution of the linear system (3.18) yields an approximate solution which satisfies (3.16). Thus, whether using a collocation or Galerkin procedure for (3.1), one obtains an approximation whose error satisfies (3.16). That is, $h^2\|A_C^{-1}\|$ is also $\mathcal{O}(M^2)$. The following example, which illustrates the similarity of the two methods, records

$$ERR(M) = \|\vec{u} - \vec{c}\| \quad .$$

The error given by (3.16) is the difference in the functions while the error displayed in the following tables is the difference in the coefficients. As in the discussion leading to (2.34), the error in the coefficients provides the dominant contribution to the error.

Example 3.3 For a simple comparison of the Sinc–Galerkin and Sinc–collocation methods, consider

$$\begin{aligned} -u'' + \pi u' &= \sin(\pi x) + \cos(\pi x) \\ u(0) &= u(1) = 0 \quad , \end{aligned} \tag{3.21}$$

which has as a true solution $u(x) = \frac{1}{\pi^2}\sin(\pi x)$. Sinc–collocation and Sinc–Galerkin solutions are obtained by solving (3.18) and (3.8), respectively. The choices of $d = \frac{\pi}{2}$ and $\alpha = \beta = 1$ leads to the mesh selection $h = \frac{\pi}{\sqrt{2M}}$ as given by (3.15) and $N = M$. The results in Table 6 indicate, as shown in [21], that these procedures are virtually

	$ERR(M)$	
M	Collocation	Galerkin
4	1.5526e-03	2.4501e-03
8	3.0255e-04	3.5012e-04
16	2.7151e-05	2.7835e-05
32	7.9038e-07	7.9165e-07
64	4.7922e-09	4.7925e-09

Table 6: Error in the approximation (3.4) where the coefficients are obtained from (3.18) and (3.8) respectively

identical.

Boundary Layers

The study of Burgers' equation leads one to consider parabolic partial differential equations with large Reynolds numbers which corresponds to $\epsilon \ll 1$ in (1.1). That is to say the ratio of the convective term to the diffusive term is large. In terms of the scalar equation under consideration given by (3.1), this implies $|p(x)u'(x)| \gg |u''(x)|$. The manifestation of this inequality is geometrically seen in a boundary layer being introduced into the function u. Analytically this is characterized by an abrupt change in the derivative of the solution.

A standard method in numerical schemes to handle this abrupt change is to allocate more computational nodes near the boundary layer. This idea, resulting in a redistribution of the nodes, was developed in [4] by incorporating the boundary layer effect into the parameter selections of the method. This redistribution is incorporated

into the collocation procedure and the increased accuracy via the new mesh selection is displayed in the following example

Example 3.4 For positive κ, consider the model problem

$$\begin{aligned} -u''(x) + \kappa u'(x) &= \kappa, \quad 0 < x < 1 \\ u(0) = u(1) &= 0. \end{aligned} \tag{3.22}$$

This problem exhibits a boundary layer near $x = 1$ if $\kappa \gg 1$. The true solution to this problem is given by

$$u(x) = x - \frac{\exp(\kappa x) - 1}{\exp(\kappa) - 1}. \tag{3.23}$$

A finite element approach for this problem is discussed in [7]. Figure 1 displays (3.23) for increasing values of κ. For $\kappa = 1000$ the solution graphically appears to be discontinuous and not much different from $\kappa = 100$ and is therefore not plotted. An

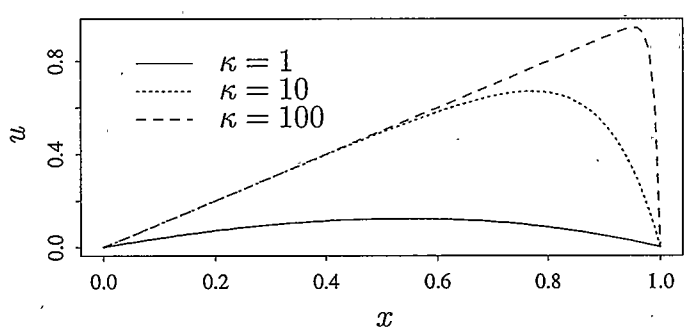

Figure 1: True solution of (3.22) for $\kappa = 1, 10, 100$

inspection of (3.23), or a Taylor series analysis of (3.22), shows that for x near 1

$$u(x) \approx \kappa(1 - x) \tag{3.24}$$

and for x near zero

$$u(x) \approx x. \tag{3.25}$$

This shows that $\alpha = \beta = 1$ are appropriate choices for exponents in (3.13). Designating h by h_s and balancing the exponential contributions to the error yields the "standard" choice for mesh size

$$h_s = \left(\frac{\pi d}{\beta N}\right)^{1/2} = \frac{\pi}{\sqrt{2N}}, \tag{3.26}$$

when using $d = \pi/2$. In the balancing of the error terms in (3.14), the integers M and N play interchangeable roles. Here, the selection of the mesh size h_s is based on N due to the boundary layer occurring at the right-hand end–point. Choosing $N = M$ yields an exponential convergence rate of $\exp(-Nh_s) = \exp(-\pi\sqrt{N/2})$. These choices of M and N are independent of κ, so that increasing values of κ are not reflected in the error statement. However, geometric considerations dictate that κ should play a role in the error analysis. From (3.25) $\hat{K}_\alpha \approx 1$ but from (3.24) $\hat{K}_\beta \approx \kappa$ so that a more accurate error representation ensues if κ is factored from the K_β. To do this, rewrite κ as $\kappa = \exp(\delta\ell n(10))$. Now consider the exponential error contributions in (3.14) as

$$\exp(-\alpha Mh), \quad \exp(-\beta Nh + \delta\ell n(10)), \quad \text{and} \quad \exp(-\pi d/h) \quad .$$

Again, the goal is to balance the error contribution from these terms. Equating the exponents in the last two terms, one finds a different h, dependent upon δ and denoted by h_δ, where

$$h_\delta = \frac{\delta\ell n(10) + \sqrt{(\delta\ell n(10))^2 + 4\pi d\beta N}}{2\beta N} \geq h_s, \quad \delta \geq 0. \tag{3.27}$$

Substituting h_δ into $\exp(-\beta Nh + \delta\ell n(10))$ and equating this term with $\exp(-\alpha Mh)$ leads to a balancing of the error terms if one defines

$$M_\delta = \left[\left|\frac{\beta}{\alpha}N - \frac{\delta\ell n(10)}{\alpha h_\delta}\right|\right], \quad \delta \geq 0. \tag{3.28}$$

Note that the selection h_δ has placed more sinc nodes $\frac{\exp(jh_\delta)}{\exp(jh_\delta)+1}$ near the boundary layer at $x = 1$ because $h_\delta \geq h_s$. From Figure 1 this is geometrically the correct thing to do. Also, since $h_\delta \geq h_s$, a comparison of the error terms shows $\exp(-\beta Nh) \geq \exp(-\beta Nh_\delta)$ so that a more accurate solution is expected. This increased accuracy is displayed in Table 7. Implementing the new mesh selection h_δ in the collocation

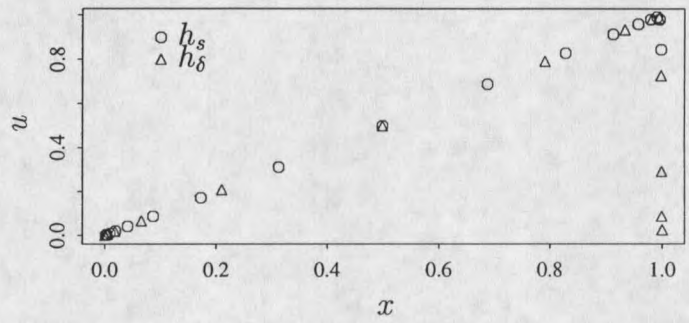

Figure 2: Effect of new node placement for $N = 8$ and $\kappa = 1000$.

procedure requires only a change in the points at which the diagonal matrices of (3.18) are evaluated.

Nonlinear terms

The study of Burgers' equation naturally leads one to consider a method that is able to accurately deal with the nonlinearity present in the spatial operator. The purpose of this section is to consider the nonlinear term in

$$\begin{aligned} -\epsilon u''(x) + u(x)u'(x) &= f(x), \quad 0 < x < 1 \\ u(0) = u(1) &= 0 \quad . \end{aligned} \tag{3.29}$$