

A NEW SINC-GALERKIN METHOD FOR CONVECTION-DIFFUSION EQUATIONS WITH MIXED BOUNDARY CONDITIONS

JENNIFER L. MUELLER AND THOMAS S. SHORES

ABSTRACT. A new Sinc-Galerkin method is developed for approximating the solution of convection-diffusion equations with mixed boundary conditions on half-infinite intervals. The method avoids differentiation of the coefficients of the PDE, rendering it appropriate as a forward solver in an inverse coefficient problem. The method has the advantage that no functions are appended to the sinc basis in the discretization. An error analysis is included, and it is shown that the error in the approximate solution is bounded in the infinity norm by the norm of the inverse of the coefficient matrix multiplied by a factor that decays exponentially with the size of the system. We demonstrate the exponential convergence of the method on several test problems.

1. INTRODUCTION

Sinc methods for the numerical solution of ordinary and partial differential equations have been extensively studied and found to be a very effective technique, particularly for problems with singular solutions and those on unbounded domains. The first Sinc-Galerkin method was presented in [11] to solve two-point boundary-value problems for second-order differential equations with Dirichlet boundary conditions. The books [5] and [12] provide excellent overviews of existing methods based on sinc functions for solving ODE's, PDE's, and integral equations. Sinc methods have also been employed as forward solvers in the solution of inverse problems (see, for example, [6],[10],[4],[8]).

This work was motivated by the fact that applying the existing sinc methods to convection-diffusion equations with mixed boundary conditions results in the differentiation of the coefficient in the leading order term. This can be undesirable, for instance when the method is being used as a forward solver in the solution of

1991 *Mathematics Subject Classification.* 65L60, 65M70.

Key words and phrases. Sinc-Galerkin method, convection-diffusion equation, contaminant transport.

an inverse coefficient problem. Furthermore, in the existing methods the ease of implementation is hampered by the need to append non-sinc basis functions to the expansion of the solution. In this paper, a new Sinc-Galerkin method for the solution of convection-diffusion equations with mixed boundary conditions is presented which has the advantage of being an appropriate forward solver in the solution of an inverse coefficient problem, as it does not require the differentiation of the coefficients. It also does not append any functions to the sinc basis. Furthermore, the method is suitable for unbounded domains and singularities in the coefficients.

The convection-diffusion equation (CDE) in one space dimension for transport of a nonreactive chemical tracer through a saturated heterogeneous porous medium is [2, p.254ff]

$$(1.1) \quad (\omega(x) + (1 - \omega(x))\rho_s K_{di}(x))c_t = (D(x)c_x)_x - (v(x)c)_x - \lambda\omega(x)c + f(x),$$

where $c(x, t)$ is the concentration of the tracer, x is distance, t is time, $v(x)$ is the advective velocity, $\omega(x)$ is the porosity, λ is the decay rate, $f(x)$ is a source term and $D(x)$ is the dispersivity. The term $(1 - \omega(x))\rho_s K_{di}(x)c_t$ arises from modeling linear reversible adsorption. In the case where porosity is constant and the transported substances are in weak concentration, the coefficient of c_t in (1) may be taken to be a constant.

We abbreviate the coefficient of c_t to $\rho(x)$ and absorb the $\omega(x)$ term into $\lambda(x)$. Thus we obtain the equation

$$(1.2) \quad \rho c_t - (D(x)c_x)_x + (v(x)c)_x + \lambda c = f(x), \quad t > 0 \quad x > 0.$$

We will assume that $f \in L^2(0, \infty)$, $v \in H^1(0, \infty)$, $v'(x) \geq 0$, $0 < \lambda_0 \leq \lambda(x) \in L^\infty(0, \infty)$, and that $0 < d_0 \leq D(x)$, $\rho(x) \in L^\infty(0, \infty)$ are bounded and

$$(1.3) \quad \lim_{x \rightarrow \infty} D(x) = D_+ = \text{CONSTANT}.$$

To complete the mathematical model, the CDE is subject to the initial condition

$$(1.4) \quad c(x, 0) = g(x), \quad x > 0,$$

the decay condition

$$(1.5) \quad c(\infty, t) = 0, \quad t > 0,$$

and Fourier boundary condition

$$(1.6) \quad -D(0)c_x(0, t) + v(0)c(0, t) = v(0)G(t), \quad t > 0$$

where $g(x) \in H^1(0, \infty)$ and $G(t) \in L^2(0, T)$ is the concentration in the entrance reservoir, which is assumed to be perfectly mixed. The suitability of various choices of boundary conditions was studied in [13] where it was shown that the Fourier type condition preserves mass balance. In [7] it is shown that in the case of constant ρ , the problem (1.2), (1.3), (1.4), (1.5), (1.6) has a unique solution in $L^2([0, T], H^1(0, \infty))$ which has norm

$$(1.7) \quad \|c\| = \left(\int_0^T \left(\int_0^\infty |c(x, t)|^2 + |c_x(x, t)|^2 dx \right) dt \right)^{1/2}.$$

Another existence-uniqueness proof in fractional Hölder spaces can be found in [9]. The Sinc-Galerkin method presented here begins with the variational form of the CDE. In obtaining the variational form, only one integration by parts is performed to avoid differentiation of the coefficients. The key idea of our approach is to transform the problem as follows: we introduce a rapidly decaying transformation function which enables us to solve for a function \tilde{c} that can be approximated by a sinc expansion interpolating \tilde{c} and its derivatives with exponential accuracy. This circumvents the need to append any functions to the sinc basis at the expense of explicitly introducing the value of $c(0)$ into the vector of unknowns. Hence the first step in the algorithm is to solve for $c(0)$. This method was applied in [7] as a forward-solver for the inverse problem of determining $D(x)$ from measurements of $c(x)$ in the steady-state case.

We compare our method to the Sinc-Galerkin method presented in Section 4.4 of [5] to solve equations of the form

$$(1.8) \quad L_1 u(x) \equiv -u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \quad a < x < b$$

with mixed boundary conditions at a and b . The idea in [5] is to express u in terms of an appended sinc basis to obtain $u_A \approx u$ and apply a Petrov-Galerkin approach to $L_1 u_A$. While this method is highly accurate, its application to the CDE would involve differentiating the dispersion coefficient $D(x)$. In Section 5 of this paper, we include some numerical comparisons with this method. Section 5.5 of [5] contains another approach for the discretization of self-adjoint forms such as $-(D(x)c_x(x))_x$, but relies on the applicability of a sinc expansion of $D(x)$. In [1] a Sinc-Collocation method is presented for equations of the form (1.8) with mixed boundary conditions at a and b . This method, while accurate and efficient, also involves appended basis functions and differentiation of $D(x)$ when applied to the CDE.

The paper is organized as follows. In Section 2 we present some notation and background from sinc theory. The definitions and theorems of Section 2 are all taken from [5] and are included for the reader's convenience since they are used in the derivation and convergence analysis of the new Sinc-Galerkin method. Section 3 contains the construction of the new Sinc-Galerkin method for the steady-state problem. The convergence analysis is found in Section 4, and numerical experiments are found in Section 5. Section 6 addresses the time-dependent problem, which was solved using a weighted implicit/explicit method in the time variable. The corresponding numerical experiments are found at the end of Section 6.

2. NOTATION AND BACKGROUND

The methods of sinc approximation for differential and integral equations rest on substantial foundations which have been laid by F. Stenger and his students, a complete development of which can be found in the texts [12] and [5]. In this section some definitions and pertinent theorems from [5] are presented for the reader's convenience.

Definition 2.1. The *sinc function* is defined for all $z \in \mathbb{C}$ by

$$\text{sinc}(z) = \begin{cases} \frac{\sin(\pi z)}{\pi z} & z \neq 0 \\ 1 & z = 0. \end{cases}$$

Let h be a positive constant. We will denote the *sinc basis functions* by

$$S(k, h)(x) = \text{sinc}\left(\frac{x - kh}{h}\right) \quad k \in \mathbb{Z}, \quad -\infty < x < \infty.$$

Theorem 2.2. [Theorem 4.1 [5]] Let $x_k = kh$, $k = -M, \dots, N$.

$$(2.1) \quad \delta_{jk}^{(0)} \equiv S(j, h)(x)|_{x=x_k} = \begin{cases} 1, & \text{if } j = k \end{cases}$$

$$(2.2) \quad \delta_{jk}^{(1)} \equiv hS'(j, h)(x)|_{x=x_k} = \begin{cases} 0 & \text{if } j = k \end{cases}$$

For the assembly of discrete systems, it is convenient to define the following matrices:

$$(2.3) \quad I^{(l)} \equiv [\delta_{jk}^{(l)}], \quad l = 0, 1$$

The matrix $I^{(0)}$ is the $m \times m$ identity matrix. The matrix $I^{(1)}$ is the skew symmetric Toeplitz matrix

$$I^{(1)} = \begin{pmatrix} 0 & -1 & \frac{1}{2} & \cdots & \frac{(-1)^{m-1}}{m-1} \\ 1 & & & & \vdots \\ -\frac{1}{2} & & \ddots & & \frac{1}{2} \\ \vdots & & & & -1 \\ \frac{(-1)^m}{m-1} & \cdots & -\frac{1}{2} & 1 & 0 \end{pmatrix}$$

Definition 2.3. [Definition 3.1 [5]] Let \mathcal{D} be a domain in the $w = u + iv$ plane with points $a \neq b$ on the boundary of D . Let $z = \phi(w)$ be a one-to-one conformal map of \mathcal{D} onto the infinite strip $\mathcal{D}_d \equiv \{z \in \mathbb{C} : z = x + iy, |y| < d\}$ where $\phi(a) = -\infty$ and $\phi(b) = \infty$. Denote by $w = \psi(z)$ the inverse of the mapping ϕ and let

$$(2.4) \quad \Gamma \equiv \{w \in \mathbb{C} : w = \psi(x), x \in \mathbb{R}\} = \psi(\mathbb{R}).$$

Let $B(\mathcal{D})$ denote the class of functions F analytic in \mathcal{D} which satisfy for some constant $a \in [0, 1)$,

$$(2.5) \quad \int_{\Psi(x+L)} |F(w)dw| = O(|x|^a), \quad x \rightarrow \pm\infty$$

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

$$(2.6) \quad N(F, \mathcal{D}) \equiv \lim_{\gamma \rightarrow \partial\mathcal{D}} \int_{\gamma} |F(w)dw| < \infty$$

where this limit means that the contour γ is $\partial\mathcal{D}$ in the limit. Further, for $h > 0$, define the nodes

$$(2.7) \quad w_k = \psi(kh), \quad k = 0, \pm 1, \pm 2, \dots$$

In our example calculations we shall take $z = \phi(w) = \ln w$ and $\mathcal{D} = \{w = re^{i\theta} \in \mathbb{C} \mid |\theta| < d\}$ so that $\psi(z) = e^z$. Formulas for interpolation involving the infinite cardinal series for a function f are discussed in [5] and [12]; here, we consider only the truncated series.

The following theorem gives the error resulting from differentiating the truncated cardinal series. A weight function g is needed to ensure the existence of the derivative. As usual, $\lceil x \rceil$ denotes the *ceiling* of x .

Theorem 2.4. [Theorem 3.17 [5]] Let $\phi'F/g \in B(\mathcal{D})$ and $h > 0$. Let ϕ be a one-to-one conformal map of the domain \mathcal{D} onto \mathcal{D}_d . Let $\psi = \phi^{-1}$, $w_k = \psi(kh)$, and

$\Gamma = \psi(R)$. Assume that there exist constants K and L so that for all $\xi \in \Gamma$

$$(2.8) \quad \left| \frac{d^m}{d\xi^m} \left[\frac{g(\xi) \sin(\pi\phi(\xi)/h)}{2\pi i(\phi(x) - \phi(\xi))} \right] \right|_{x \in \partial\mathcal{D}} \leq Kh^{-m}$$

and

$$(2.9) \quad \left| \frac{d^m}{d\xi^m} \left[g(\xi) \operatorname{sinc} \left(\frac{\phi(\xi) - kh}{h} \right) \right] \right| \leq Lh^{-m}$$

for all $m = 0, 1, \dots, n$. Assume that there are positive constants α, β and C so that

$$\left| \frac{F(\xi)}{g(\xi)} \right| \leq K \quad \begin{cases} \exp(-\alpha|\phi(\xi)|), & \xi \in \Gamma_a \\ \exp(-\beta|\phi(\xi)|), & \xi \in \Gamma_b \end{cases}$$

where $\Gamma_a \equiv \{\xi \in \Gamma : \phi(\xi) = x \in (-\infty, 0)\}$ and $\Gamma_b \equiv \{\xi \in \Gamma : \phi(\xi) = x \in [0, \infty)\}$.

Make the selections

$$(2.10) \quad N = \lceil \frac{\alpha}{\beta} M \rceil \quad \text{and} \quad \left(h = \frac{\pi d}{\alpha M} \right)^{\frac{1}{2}} \leq \frac{2\pi d}{\ln(2)}.$$

Then for all $\xi \in \Gamma$ and $m = 0, 1, \dots, n$,

$$(2.11) \quad \left\| \frac{d^m}{d\xi^m} F(\xi) - \sum_{k=-M}^N \frac{F(w_k)}{g(w_k)} \frac{d^m}{d\xi^m} [g(\xi) S(k, h) \circ \phi(\xi)] \right\|_{\infty} \leq KM^{(m+1)/2} e^{-\sqrt{\pi d \alpha M}}.$$

The following theorem is useful in the development of the Sinc-Galerkin scheme.

Theorem 2.5. [5, Theorem 4.4] Assume that there are positive constants α, β , and K so that

$$|F(\xi)| \leq K \quad \begin{cases} \exp(-\alpha|\phi(\xi)|), & \xi \in \Gamma_a \\ \exp(-\beta|\phi(\xi)|), & \xi \in \Gamma_b \end{cases}$$

where $\Gamma_a \equiv \{\xi \in \Gamma : \phi(\xi) = x \in (-\infty, 0)\}$ and $\Gamma_b \equiv \{\xi \in \Gamma : \phi(\xi) = x \in [0, \infty)\}$

and $F = upw$ or $u\phi'w$. Let $B_T = (u'[S(j, h) \circ \phi]w)(x) - (u([S(j, h) \circ \phi]w'))(x)|_a^b$.

Select N and h as in (2.10).

(a) Let $vw \in B(\mathcal{D})$ for $v = f(x)$ or qu . Then

$$(2.12) \quad \left| \int_a^b (vw[S(j, h) \circ \phi](x) dx - h \frac{vw}{\phi'}(x_j)) \right| \leq L_0 M^{-1/2} \exp(-(\pi d \alpha M)^{1/2})$$

where L_0 is a constant depending on v and d .

(b) Let $u(p[S(j, h) \circ \phi]w)' \in B(\mathcal{D})$ and $B_T = 0$. Then

$$\left| \int_a^b (pu'[S(j, h) \circ \phi]w)(x) dx + h \sum_{k=-M}^N (upw)(x_k) \frac{\delta_{jk}^{(1)}}{h} + h \left(\frac{u(pw)'}{\phi'} \right)(x_j) \right| \leq L_1 M^{1/2} \exp(-(\pi d \alpha M)^{1/2})$$

where L_1 is a constant depending on u, p, w , and d .

Finally, we shall need the following result which is suitable for more general sinc quadrature.

Theorem 2.6. [5, Theorem 3.8] *Assume the notation of Theorem 2.5. Suppose there are positive constants α, β , and K so that*

$$\left| \frac{F(\xi)}{\phi'(\xi)} \right| \leq K \begin{cases} \exp(-\alpha|\phi(\xi)|), & \xi \in \Gamma_a \\ \exp(-\beta|\phi(\xi)|), & \xi \in \Gamma_b \end{cases}$$

where $\Gamma_a \equiv \{\xi \in \Gamma : \phi(\xi) = x \in (-\infty, 0)\}$ and $\Gamma_b \equiv \{\xi \in \Gamma : \phi(\xi) = x \in [0, \infty)\}$. Make the selections

$$N = \lceil \frac{\alpha}{\beta} M \rceil \quad \text{and} \quad h = \left(\frac{2\pi d}{\alpha M} \right)^{\frac{1}{2}} \leq \frac{2\pi d}{\ln(2)}.$$

Then there is a constant L depending only on F , d , ϕ and \mathcal{D} such that

$$\left| \int_a^b F(x) dx - h \sum_{k=-M}^N \frac{F(x_k)}{\phi'(x_k)} \right| \leq L \exp(-(2\pi d \alpha M)^{1/2}).$$

We observe that Theorem 2.6 remains valid if (2.10) is used to select N and h . One can see from the proof of this theorem that the only change is a loss of the factor of 2 in the exponential bound, which results in the same sort of exponential bound as in Theorem 2.5. We will primarily be interested in the case that $\phi(x) = \ln x$, whereupon the inequalities on $F(\xi)$ reduce to

$$|F(\xi)| \leq K \begin{cases} \xi^{\alpha-1}, & \xi \in \Gamma_a \\ \xi^{-\beta-1}, & \xi \in \Gamma_b \end{cases}.$$

3. NUMERICAL FORMULATION OF STEADY STATE PROBLEM

We first consider the steady-state problem

$$(3.1) \quad Lc \equiv -(D(x)c_x)_x + (v(x)c)_x + \lambda(x)c = f(x), \quad x > 0$$

$$(3.2) \quad Ac(0) + Bc_x(0) = G, \quad B \neq 0$$

$$(3.3) \quad c(\infty) = 0$$

where $c = c(x)$. Recall that only the natural boundary conditions $A = v(0)$ and $B = -D(0)$, guarantee a well posed problem, so we must assume that the constants of (3.2) have been properly chosen. We define the integrals

$$(3.4) \quad T(c, u) = \int_0^\infty (Dc_x u_x - vcu_x + \lambda cu) dx$$

and

$$(3.5) \quad R(c, u) = \int_0^\infty f(x)u dx + \left\{ c(0) \left(v(0) + \frac{A}{B}D(0) \right) - D(0)\frac{G}{B} \right\} u(0).$$

Multiply (3.1) by u and integrate by parts to obtain that (3.1)–(3.3) can be written in the variational form

$$(3.6) \quad \int_0^\infty (Dc_x u_x - vcu_x + \lambda cu) dx = \int_0^\infty f(x)u dx + (-D(0)c_x(0) + v(0)c(0)) u(0).$$

Use the boundary condition (3.2) to eliminate $c_x(0)$ and we can express this variational form as $T(c, u) = R(c, u)$.

Next the problem is transformed in two steps to obtain what we will refer to as the *numerical variational form*. We select a function $p(x)$ such that $p(0) = 1$ and $\lim_{x \rightarrow \infty} p(x) = 0$ at approximately the same rate as $c(x)$. Asymptotic methods such as WKB can be used to estimate this rate, but we note that in practice the method was somewhat insensitive to the particular choice of p .

Define

$$(3.7) \quad q_0(x) = \frac{(G/B)x}{x+1} p(x)$$

so that $q_0(0) = 0$ and $q'_0(0) = G/B$, and q_0 satisfies the Fourier boundary condition (3.2). Define

$$(3.8) \quad q_1(x) = \frac{-(A/B + p'(0) - 1)x + 1}{x+1} p(x)$$

so that $q_1(0) = 1$ and $q'_1(0) = -A/B$. Hence q_1 satisfies the homogeneous boundary condition

$$(3.9) \quad Aq_1(0) + Bq'_1(0) = 0.$$

Now define

$$(3.10) \quad \tilde{c}(x) \equiv c(x) - q_0(x) - c(0)q_1(x).$$

Then \tilde{c} satisfies the homogeneous condition

$$(3.11) \quad A\tilde{c}(0) + B\tilde{c}'(0) = 0,$$

and since $\tilde{c}(0) = 0$ we have that $\tilde{c}'(0) = 0$.

Thus, from (3.1) and (3.10) the numerical variational form is

$$(3.12) \quad T(\tilde{c}, u) + T(q_0, u) + c(0)T(q_1, u) = R(c, u).$$

For a Sinc-Galerkin approximation we choose test functions $u_j(x) = w(x)S(j, h) \circ \phi(x)$ where $w(x)$ is a weight function and $\phi(x)$ is a conformal map from $(0, \infty)$ to

$(-\infty, \infty)$, such as $\phi(x) = \ln(x)$. This is not the only possible choice for $\phi(x)$ and for further discussion of the alternatives, we refer the reader to [5, p. 79ff]. The unknown $c(0)$ will be obtained via an auxillary equation derived at the end of this section, and will be found before solving the linear system obtained by orthogonalizing the residual in (3.12). Thus, for the present derivation, we will regard $c(0)$ as known. Assuming \tilde{c} satisfies the hypotheses of Theorem 2.4, we can approximate \tilde{c} by

$$(3.13) \quad \tilde{c}_m(x) = \sum_{j=-M_x}^{N_x} \frac{d_j}{\gamma(x_j)} \gamma(x) S(j, h) \circ \phi(x)$$

where γ is an appropriately chosen weight function so that $\tilde{c}'(x)$ is accurately approximated by the derivative of the cardinal sum, and h , M_x and N_x are chosen according to Theorem 2.4. We define the sinc nodes $x_k = \phi^{-1}(kh)$, $k = -M_x, \dots, N_x$ and let $m = M_x + N_x + 1$. Note that

$$\tilde{c}_m(x_k) = d_k.$$

For a true Galerkin method, we choose the weight function $w(x) = \gamma(x)$, but the following derivation is carried out for a general w .

Note that differentiating the approximation (3.13) to \tilde{c} and evaluating at x_k yields

$$\tilde{c}'_m(x_k) = \sum_{j=-M_x}^{N_x} \frac{d_j}{\gamma(x_j)} (\gamma(x_k) S(j, h) \circ \phi(x_k))'$$

which is exponentially accurate by Theorem 2.4. Now by the definition of $\delta_{jk}^{(i)}$ in Theorem 2.2,

$$\tilde{c}'_m(x_k) = \frac{d_k}{\gamma_k} \gamma'_k + \frac{1}{h} \sum_{j=-M_x}^{N_x} \frac{d_j}{\gamma_j} \gamma_k \delta_{jk}^{(1)} \phi'_k.$$

where a subscript of k denotes evaluation at the node x_k , *e.g.*, $\gamma_k \equiv \gamma(x_k)$.

For the remainder of this section we adopt the Einstein summation notation where we sum over any repeated index in a product or quotient unless otherwise indicated. Thus, applying the sinc quadrature formula of Theorem 2.6 to $T(\tilde{c}, u_j)$ yields

$$(3.14) \quad T_{quad}(\tilde{c}, u_j) = \frac{h}{\phi'_k} (D_k \tilde{c}'_k(u_j)'_k - v_k \tilde{c}_k(u_j)'_k + \lambda_k \tilde{c}_k(u_j)_k).$$

Thus,

$$T_{quad}(\tilde{c}_m, u_j) = \frac{h}{\phi'_k} \left(D_k \left\{ \frac{\gamma'_k d_k}{\gamma_k} + \frac{d_r}{\gamma_r} \gamma_k \delta_{rk}^{(1)} \frac{\phi'_k}{h} \right\} (u'_j)_k - v_k d_k (u'_j)_k + \lambda_k d_k (u_j)_k \right).$$

Note that for our choice $u_j = wS(j, h) \circ \phi$,

$$(u_j)_k = w_k \delta_{jk}^{(0)} \quad \text{and} \quad (u'_j)_k = w'_k \delta_{jk}^{(0)} + w_k \delta_{jk}^{(1)} \frac{\phi'_k}{h}, \quad (\text{no sum on } k)$$

So

$$\begin{aligned} (3.15) T_{quad}(\tilde{c}_m, u_j) &= \frac{h}{\phi'_k} \left[D_k \left\{ \frac{\gamma'_k d_k}{\gamma_k} + \frac{d_r}{\gamma_r} \gamma_k \delta_{rk}^{(1)} \frac{\phi'_k}{h} \right\} \left(w'_k \delta_{jk}^{(0)} + w_k \delta_{jk}^{(1)} \frac{\phi'_k}{h} \right) \right. \\ &= \left. -v_k d_k \left(w'_k \delta_{jk}^{(0)} + w_k \delta_{jk}^{(1)} \frac{\phi'_k}{h} \right) + \lambda_k d_k w_k \delta_{jk}^{(0)} \right] \end{aligned}$$

For ease in building the discrete system, we use the property that $\delta_{ij}^{(1)} = -\delta_{ji}^{(1)}$ and regroup the terms in (3.15) to obtain the following expression (no sum on j).

(3.16)

$$\begin{aligned} T_{quad}(\tilde{c}_m, u_j) &= h \frac{D_j w'_j \gamma'_j}{\phi'_j \gamma_j} d_j - \gamma_j D_j w'_j \delta_{jr}^{(1)} \frac{1}{\gamma_r} d_r + \delta_{jk}^{(1)} \frac{w_k D_k \gamma'_k}{\gamma_k} d_k \\ &- \frac{1}{h} \delta_{jk}^{(1)} w_k D_k \gamma_k \phi'_k \delta_{kr}^{(1)} \frac{1}{\gamma_r} d_r - h \frac{v_j w'_j}{\phi'_j} d_j - \delta_{jk}^{(1)} v_k w_k d_k + \frac{h}{\phi'_j} \lambda_j w_j d_j. \end{aligned}$$

Let $\text{Diag}(\cdot)$ denote a diagonal matrix of node evaluations. Also, in general we let $\text{Vec}(f)$ denote a vector of node evaluations of the function $f(x)$ at nodes x_j , $j = -M_x, \dots, N_x$. To avoid subscript ambiguity, let $d(x) = \tilde{c}_m(x)$. We see from (3.13) that $\text{Vec}(d) = \{d_j\}_{j=-M_x}^{N_x}$. Letting $j = -M_x, \dots, N_x$ in (3.15) yields the system

$$(3.17) \quad \{T_{quad}(\tilde{c}_m, u_j)\}_{j=-M_x}^{N_x} = M \cdot \text{Vec}(d)$$

where the matrix M is given by

$$\begin{aligned} M &\equiv h \text{Diag} \left(\frac{D\gamma'w'}{\gamma\phi'} \right) + I^{(1)} \text{Diag} \left(\frac{D\gamma'w}{\gamma} \right) - \text{Diag}(D\gamma w') I^{(1)} \text{Diag} \left(\frac{1}{\gamma} \right) \\ &- I^{(1)} \text{Diag}(D\gamma\phi'w) \frac{1}{h} I^{(1)} \text{Diag} \left(\frac{1}{\gamma} \right) - h \text{Diag} \left(\frac{vw'}{\phi'} \right) - I^{(1)} \text{Diag}(vw) + h \text{Diag} \left(\frac{\lambda w}{\phi'} \right) \end{aligned}$$

and where $I_{ij}^{(1)} = \delta_{ij}^{(1)}$ as defined in (2.3). The integrals $T(q_i, wS(j, h) \circ \phi)$, $i = 0, 1$ are approximated by $T_{quad}(q_i, u_j)$ as in (3.14). Since the functions q_i are known, denoting the discrete approximations to $\{T(q_i, u_j)\}_{j=-M_x}^{N_x}$ by T_i , $i = 0, 1$, (3.14)

yields the vectors

$$\begin{aligned} T_i &= I^{(1)} \text{Diag}(Dw) \text{Vec}(q'_i) - I^{(1)} \text{Diag}(vw) \text{Vec}(q_i) + h \text{Diag}\left(\frac{\lambda w}{\phi'}\right) \text{Vec}(q_i) \\ &+ h \text{Diag}\left(\frac{Dw'}{\phi'}\right) \text{Vec}(q'_i) - h \text{Diag}\left(\frac{vw'}{\phi'}\right) \text{Vec}(q_i). \end{aligned}$$

Since $u(0) = 0$, the discretized $R(c, u_j)$, $j = -M_x, \dots, N_x$ does not involve $c(0)$ and is simply

$$(3.18) \quad R_{dis} \equiv h \text{Diag}\left(\frac{w}{\phi'}\right) \text{Vec}(f).$$

Thus, the discrete system corresponding to the variational form (3.6) with $u_j = wS(j, h) \circ \phi$, $j = -M_x, \dots, N_x$ is

$$(3.19) \quad M \text{Vec}(d) + T_0 + c_0 T_1 = R_{dis}$$

where $c_0 = c(0)$. Let $w_0 \equiv M^{-1}(R_{dis} - T_0)$ and $w_1 \equiv M^{-1}T_1$ and we may write this equation in the form

$$(3.20) \quad \text{Vec}(d) = w_0 - c_0 w_1.$$

We apply the following scheme for computing c_0 : Integrating (3.1) from 0 to ∞ and applying the boundary condition yields

$$\begin{aligned} (3.21) \quad \int_0^\infty (f(x) - \lambda(x)c)dx &= D(0)c_x(0) - v(0)c(0) \\ &= D(0)\frac{G}{B} - c_0 \left(v(0) + \frac{A}{B}D(0) \right) \end{aligned}$$

By the sinc quadrature rule (and the Einstein summation notation) we have

$$\int_0^\infty f(x)dx \approx h \frac{f_k}{\phi'_k} \quad \text{and} \quad \int_0^\infty \lambda(x)c(x)dx \approx h \frac{\lambda_k c_k}{\phi'_k}$$

so that

$$(3.22) \quad h \frac{f_k}{\phi'_k} - h \frac{\lambda_k c_k}{\phi'_k} \approx D(0)\frac{G}{B} - c_0 \left(v(0) + \frac{A}{B}D(0) \right).$$

We have

$$(3.23) \quad c_k = \tilde{c}_k + (q_0)_k + c(0)(q_1)_k, \quad k = -M_x, \dots, N_x$$

so that

$$(3.24) \quad c_k \approx (w_0 - c_0 w_1)_k + (q_0)_k + c_0 (q_1)_k$$

Substituting for c_k in (3.22) yields the following equation :

$$(3.25) \quad h \frac{f_k}{\phi'_k} - h \frac{\lambda_k}{\phi'_k} \{(w_0 - c_0 w_1)_k + (q_0)_k + c_0 (q_1)_k\} \approx D(0) \frac{G}{B} - c_0 \left(v(0) + \frac{A}{B} D(0) \right).$$

We solve for c_0 to obtain

$$(3.26) \quad c_0 = \frac{D(0) \frac{G}{B} + \frac{h}{\phi'_k} \{\lambda_k ((w_0)_k + (q_0)_k) - f_k\}}{v(0) + \frac{A}{B} D(0) + h \frac{\lambda_k}{\phi'_k} \{(w_1)_k - (q_1)_k\}}$$

We summarize the algorithm for the solution to the CDE:

Algorithm

- (1) Form M , T_0 , and T_1 .
- (2) Compute $w_0 \equiv M^{-1}(R - T_0)$ and $w_1 \equiv M^{-1}T_1$.
- (3) Calculate c_0 by (3.26).
- (4) Calculate $\text{Vec}(d)$ from (3.20).
- (5) Compute $\text{Vec}(c)$ from (3.24).

4. CONVERGENCE ANALYSIS

In order for the discretizations we have employed to be valid, the hypotheses of Theorems 2.4-2.6 have to be satisfied for many functions in our discussion. Throughout this section we assume that the coefficients D, v, λ and f in (3.1) – (3.3) and the unique solution c are analytic in the simply connected domain D containing 0 and ∞ on its boundary and that ϕ is a conformal map of \mathcal{D} onto the strip $\mathcal{D}_d \equiv \{z \in \mathbb{C} : z = x + iy, |y| < d\}$ such that $\phi(0) = -\infty$ and $\phi(\infty) = \infty$. Assume also that $q_0, q_1, D\tilde{c}'(wS(k, h) \circ \phi)', v\tilde{c}(wS(k, h) \circ \phi)', \lambda\tilde{c}, \lambda\tilde{c}w, \lambda\tilde{c}w \in B(\mathcal{D})$. We assume that constants α, β can be found such that the exponential inequality

$$|F(\xi)| \leq K \begin{cases} \exp(-\alpha|\phi(\xi)|), & \xi \in \Gamma_a \\ \exp(-\beta|\phi(\xi)|), & \xi \in \Gamma_b \end{cases}$$

where $\Gamma_a \equiv \{\xi \in \Gamma : \phi(\xi) = x \in (-\infty, 0)\}$ and $\Gamma_b \equiv \{\xi \in \Gamma : \phi(\xi) = x \in [0, \infty)\}$ is satisfied for all functions F needed to validate the sinc quadratures and interpolations used in our algorithm. Given M_x , choose

$$(4.1) \quad N_x = \lceil \frac{\alpha}{\beta} M_x + 1 \rceil \quad \text{and} \quad h = \left(\frac{\pi d}{\alpha M_x} \right)^{1/2}.$$

Denote the Sinc-Galerkin solution obtained by the method of Section 3 by $c_m(x)$, $m = M_x + N_x + 1$. Then we wish to find a bound on $\|c - c_m\|_\infty$. Recall, $u_j(x) = w(x)S(j, h) \circ \phi(x)$. Here we will choose $w(x) \equiv \gamma(x) = x/(10 + x)^2$, as used in the

computations. (Other choices are possible. Some justification for this particular $w(x)$ is given at the beginning of Section 5.) From (3.12) we have

$$(4.2) \quad T(\tilde{c}, u) = R(u_j) - GT(q_0, u_j) - c(0)T(q_1, u_j).$$

Let the vector g be defined to have j th component

$$g_j \equiv R(u_j) - GT(q_0, u_j) - c(0)T(q_1, u_j).$$

We define the *discretization error vector*

$$e_1 \equiv T_{quad}(\tilde{c}, u_j) - T(\tilde{c}, u_j)$$

where \tilde{c} is the exact solution to (3.12), i.e., $T(\tilde{c}, u_j) = g_j$. We define the *right-hand side error vector*

$$e_2 \equiv g_j - \tilde{g}_j$$

where $\tilde{g}_j \equiv R_{dis}(u_j) - GT_{quad}(q_0, u_j) - c(0)T_{quad}(q_1, u_j)$.

Now let \tilde{c}_m solve $T_{quad}(\tilde{c}_m, u_j) = \tilde{g}_j$. Then

$$(4.3) \quad \begin{aligned} T_{quad}(\tilde{c} - \tilde{c}_m, u_j) &= T(\tilde{c}, u_j) + e_1 - \tilde{g}_j \\ &= g_j + e_1 - (g_j - e_2) \\ &= e_1 + e_2. \end{aligned}$$

Throughout this section we shall abbreviate the infinity norm to $\|\cdot\|_\infty = \|\cdot\|$.

Lemma 4.1. *Assume that the coefficients D, v, λ and f in (3.1)–(3.3) and the unique solution c are analytic in the simply connected domain \mathcal{D}_w . Let ϕ be the conformal map of \mathcal{D}_w onto \mathcal{D}_S . Given M_x , choose*

$$N_x = \lceil \frac{\alpha}{\beta} M_x + 1 \rceil \quad \text{and} \quad h = \left(\frac{\pi d}{\alpha M_x} \right)^{1/2}.$$

Assume also that $D\tilde{c}'(wS(k, h) \circ \phi)'$, $v\tilde{c}(wS(k, h) \circ \phi)'$, $\lambda\tilde{c}w \in B(\mathcal{D})$. Let $N = \min(M_x, N_x)$. Then

$$(4.4) \quad \|e_1\|_\infty \leq K(M_x^{1/2} + M_x^2 + N_x^{3/4}) \exp(-\sqrt{\pi d \alpha N}).$$

Proof. We let K be a constant, independent of M_x which will be permitted to change without relabeling. Define

$$(4.5) \quad \eta_m(x) = \sum_{k=-M_x}^{N_x} \frac{\tilde{c}(x_k)}{\gamma(x_k)} \gamma(x) S(k, h) \circ \phi(x).$$

Then

$$\begin{aligned} |T_{quad}(\tilde{c}, u_j) - T(\tilde{c}, u_j)| &\leq |T_{quad}(\tilde{c}, u_j) - T_{quad}(\eta_m, u_j)| + |T_{quad}(\eta_m, u_j) - T(\eta_m, u_j)| \\ &\quad + |T(\eta_m, u_j) - T(\tilde{c}, u_j)|. \end{aligned}$$

By Theorem 2.5

$$|T_{quad}(\eta_m, u_j) - T(\eta_m, u_j)| \leq L_2 M_x^{1/2} \exp(-\sqrt{\pi d \alpha M_x}).$$

By (3.14) and Theorem 2.4

$$\begin{aligned} |T_{quad}(\tilde{c}, u_j) - T_{quad}(\eta_m, u_j)| &\leq h \left(\left| \frac{D_k(u'_j)_k}{\phi'_k} \right| |(\tilde{c}' - \eta'_m)_k| + \left| \frac{v_k(u'_j)_k}{\phi'_k} \right| |(\tilde{c} - \eta_m)_k| + \left| \frac{\lambda_k(u_j)_k}{\phi'_k} \right| |(\tilde{c} - \eta_m)_k| \right) \\ &\leq h K M_x e^{-\sqrt{\pi d \alpha M_x}} \left(\left| \frac{D_k(u'_j)_k}{\phi'_k} \right| + \left| \frac{v_k(u'_j)_k}{\phi'_k} \right| + \left| \frac{\lambda_k(u_j)_k}{\phi'_k} \right| \right) \end{aligned}$$

Since $h \leq K M_x^{-1/2}$ and $|1/\phi'_k| \leq K M_x^{1/2}$, we have

$$|T_{quad}(\eta_m, u_j) - T_{quad}(\tilde{c}, u_j)| \leq K M_x e^{-\sqrt{\pi d \alpha M_x}} (|D_k(u'_j)_k| + |v_k(u'_j)_k| + |\lambda_k(u_j)_k|)$$

Since w and w' are bounded and $|S(k, h) \circ \phi(x)| \leq 1$ and $|(S(k, h) \circ \phi)'(x)| \leq K M_x$,

$$(4.6) \quad |T_{quad}(\eta_m, u_j) - T_{quad}(\tilde{c}, u_j)| \leq K M_x e^{-\sqrt{\pi d \alpha M_x}} (K_1 M_x + K_2 M_x + K_3).$$

So

$$(4.7) \quad |T_{quad}(\eta_m, u_j) - T_{quad}(\tilde{c}, u_j)| \leq K M_x^2 \exp(-\sqrt{\pi d \alpha M_x}).$$

Finally, by Holder's inequality

$$\begin{aligned} |T(\eta_m, u_j) - T(\tilde{c}, u_j)| &\leq \left| \int_0^\infty D(\eta'_m - \tilde{c}') u'_j dx \right| + \left| \int_0^\infty v(\eta_m - \tilde{c}) u'_j dx \right| + \left| \int_0^\infty \lambda(\eta_m - \tilde{c}) u_j dx \right| \\ &\leq \|D u'_j\|_2 \|\eta'_m - \tilde{c}'\|_2 + \|v u'_j\|_2 \|\eta_m - \tilde{c}\|_2 + \|\lambda u_j\|_2 \|\eta_m - \tilde{c}\|_2 \end{aligned}$$

In Chapter 4, [12] one finds the bound

$$(4.8) \quad \left| \frac{d}{dx} S(k, h) \circ \log(x) \right| \leq C \pi x^{-1}/h,$$

which implies, for our choice $\gamma = x/(10 + x)^2$,

$$(4.9) \quad \|D u'_j\| \leq \|D\|_\infty \left\| \frac{d\gamma}{dx} S(j, h) \circ \phi + \gamma \frac{d}{dx} S(j, h) \circ \phi \right\|_2 < \infty.$$

Similarly, $\|vu'_j\| < \infty$. By equation (4.4.13), [12],

$$(4.10) \quad \|\eta'_m - \tilde{c}'\|_2 \leq C_1 N_x^{3/4} e^{(-\sqrt{\pi d \alpha N_x})}$$

$$(4.11) \quad \|\eta_m - \tilde{c}\|_2 \leq C_0 N_x^{1/4} e^{(-\sqrt{\pi d \alpha N_x})}$$

Finally, since $\|u_j\|_2 < \infty$, we have

$$|T(\eta_m, u_j) - T(\tilde{c}, u_j)| \leq K N_x^{3/4} e^{(-\sqrt{\pi d \alpha N_x})}.$$

This proves the lemma. \square

Lemma 4.2. $\|e_2\|_\infty \leq K M_x^{1/2} \exp(-\sqrt{\pi d \alpha M_x})$

Proof. By Theorem 2.5

$$\begin{aligned} |g_j - \tilde{g}_j| &\leq |R(u_j) - R_{dis}(u_j)| + |G| |T(q_0, u_j) - T_{quad}(q_0, u_j)| \\ &\quad + |c(0)| |T_{quad}(q_1, u_j) - T(q_1, u_j)| \\ &\leq L_0 M_x^{-1/2} \exp(-\sqrt{\pi d \alpha M_x}) + |G| L_2 M_x^{1/2} \exp(-\sqrt{\pi d \alpha M_x}) \\ &\quad + |c(0)| L_2 M_x^{1/2} \exp(-\sqrt{\pi d \alpha M_x}) \\ &\leq K_1 M_x^{1/2} \exp(-\sqrt{\pi d \alpha M_x}). \end{aligned}$$

This proves the lemma. \square

Theorem 4.3. *With the assumptions of Lemma 4.1 there exists a constant K independent of M_x such that*

$$\|\tilde{c} - \tilde{c}_m\| \leq K \|M^{-1}\| M_x^2 \exp(-\sqrt{\pi d \alpha N}).$$

If, in addition, $\left\|\frac{1}{\phi_k}\right\|_\infty$ is bounded by a multiple of m and q_1 is chosen such that $\int_0^\infty \lambda(x) q_1(x) dx \neq v(0) + \frac{A}{B} D(0)$, then there exists a constant C such that

$$\|c(0) - c_0\| \leq C \|M^{-1}\| M_x^{5/2} \exp(-\sqrt{\pi d \alpha N}).$$

Proof. Note that from (4.3) we have that

$$\|\tilde{c} - \tilde{c}_m\| \leq \|M^{-1}\| (\|e_1\| + \|e_2\|).$$

Since N_x is less than a constant multiple of M_x , the first assertion of the theorem follows from Lemmas 4.1 and 4.2.

To estimate the error in approximating $c(0)$, we let $H = v(0) + \frac{A}{B} D(0)$, so that (3.21) becomes

$$(4.12) \quad \int_0^\infty (f - \lambda(\tilde{c} + Gq_0 + c(0)q_1)) dx = D(0) \frac{G}{B} - c(0)H$$

Thus we have

$$(4.13) \quad -h \frac{f_k}{\phi'_k} - h \frac{\lambda_k}{\phi'_k} (\tilde{c} + Gq_0 + c(0)q_1)_k - e_3 = D(0) \frac{G}{B} - c(0)H$$

where e_3 is the error produced by performing sinc-quadrature on (4.12) as specified by Theorem 2.6. If in (4.13) we replace $c(0)$ by c_0 , \tilde{c} by \tilde{c}_m , and delete the error term e_3 , we obtain (3.25), which is the equation used to solve for c_0 . Subtract (3.25) from (4.13) and there results

$$(4.14) \quad -h \frac{\lambda_k}{\phi'_k} ((\tilde{c} - \tilde{c}_m) + (c(0) - c_0)q_1)_k - e_3 = -(c(0) - c_0)H.$$

So the error $e = c(0) - c_0$ satisfies

$$(4.15) \quad \left(H - h \frac{(\lambda q_1)_k}{\phi'_k} \right) e = -h \frac{\lambda_k}{\phi'_k} (\tilde{c} - \tilde{c}_m)_k + e_3.$$

Note that as $m \rightarrow \infty$, $h \frac{\lambda_k}{\phi'_k} \rightarrow \int_0^\infty \lambda(x) dx$ the coefficient of e tends to $H - \int_0^\infty \lambda(x) q_1(x) dx$, which is a nonzero constant by our hypotheses. Furthermore, e_3 is bounded by a constant multiple of $\exp(-\sqrt{\pi d \alpha M_x})$ and $h \frac{\lambda_k}{\phi'_k}$ is bounded by a multiple of $M^{1/2}$ by our hypothesis on ϕ' , the definition of h and the fact that $\lambda(x)$ is bounded. Thus, if we take absolute values of both sides, divide by the coefficient of e and use the first assertion of this theorem, we obtain that for some constant C independent of M_x ,

$$(4.16) \quad |c(0) - c_0| \leq C M_x^{5/2} \exp(-\sqrt{\pi d \alpha M_x}).$$

□

We remark that the technical hypotheses of the second part of the preceding theorem are easily satisfied by most problems. If λ is nonzero anywhere, the first condition can be satisfied. And the condition on ϕ' is easily seen to be true for the two most commonly used choices of ϕ , namely $\phi(x) = \ln x$ and $\phi(x) = \ln(\sinh(u))$.

We conclude this section with a discussion of the matrix M . We use the infinity norm in this discussion. The ideal situation would be for $\text{cond}(M)$ to be polynomial. However, numerical evidence from each matrix used in this paper suggests that $\|M\|$ grows exponentially. Unfortunately, this is demonstrably true in some cases.

Example 4.4. Let $D(x) = 1, v(x) = 0, \lambda(x) = 1, \gamma(x) = w(x) = \phi'(x) = 1/x$. The matrix M is given by

$$(4.17) \quad M = h \operatorname{Diag} \left(\frac{1}{x_k^2} \right) + I^{(1)} \operatorname{Diag} \left(\frac{-1}{x_k^2} \right) - \operatorname{Diag} \left(\frac{-1}{x_k^3} \right) I^{(1)} \operatorname{Diag}(x_k) \\ - I^{(1)} \operatorname{Diag} \left(\frac{1}{x_k^3} \right) \frac{1}{h} I^{(1)} \operatorname{Diag}(x_k) + h I^{(0)}$$

Let us estimate the $(1, 1)$ -th entry of the matrix M . Since $I^{(1)}$ is skew-symmetric, diagonal entries are zero and so the second and third terms of the right-hand side make no contribution to this entry. Also, the contribution of the last term tends to zero with h , and one can show (we omit the details) that the first term makes a contribution that is negligible relative to the contribution of the fourth term, which we now examine. Let $D = \operatorname{Diag}(1/x_k^{3/2})$ so that this term becomes $(1/h)B$, where

$$(4.18) \quad B = (DI^{(1)})^T DI^{(1)} \operatorname{Diag}(x_k).$$

Certainly $1/h > 1$, so it suffices to examine the matrix B . We see that the $(1, 1)$ -th entry of B is product of a sum of squares of the entries of the first column of $DI^{(1)}$ and the first entry of the diagonal matrix $\operatorname{Diag}(x_k)$. Now recall that $x_k = \exp(-kh)$, $k = -M_x \dots N_x$. Thus, the second entry in the first column of $DI^{(1)}$ is $\exp((M_x - 1)h(3/2))$. Hence the $(1, 1)$ -th entry of B is a sum of squares, one of which is $\exp(3(M_x - 1)h)$ multiplied by $\exp(-M_x h)$. This entry can be shown to be of order $\exp(AM_x^{1/2})$ for some positive constant A (we omit the details here). Thus $\|M\|_\infty$ grows exponentially with M_x .

Fortunately, our error estimates involve $\|M^{-1}\|$ and not $\operatorname{cond}(M)$. Now it is customary to consider the condition number of the (invertible) coefficient matrix of a linear system $Mx = b$ as the deciding factor for accuracy, as it appears as a coefficient that amplifies error. However, we can get a different perspective if we have to go back to the classical forward error inequality and express it in this form: If $Mx = b$ and $(M + \Delta M)(x + \Delta x) = b + \Delta b$ and $r = \|\Delta M M^{-1}\| < 1$ then we have

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|M^{-1}\|}{1 - r} \left\{ \|\Delta M\| + \|M\| \frac{\|\Delta b\|}{\|b\|} \right\}.$$

Now when we solve a system $Mx = b$, by backwards error analysis, we effectively solve a system with perturbed coefficient matrix as above. Only input or independent calculation of b will introduce the error term Δb . Thus, given that $\|M^{-1}\|$ grows more slowly with its size than $\|M\|$ and $\|M\|$ grows exponentially with size,

we see the importance of starting with an accurate right hand side b order to have the term $\|M^{-1}\|$ be the major source of amplification of error rather than $\text{cond}(M)$.

Numerical experiments indicate that the behavior of $\|M^{-1}\|$ is quite complex. In Figure 4.1 we plot norms versus matrix dimension for different coefficient matrices M that come from the test problems in the following section, and dimension $m = 1, \dots, 320$. (In sinc applications, most examples are satisfactorily solved by matrices in this range.) In each case, the exponential growth of $\|M\|$ is apparent. On the other hand, $\|M^{-1}\|$ appears to grow linearly over a large range of dimensions, then abruptly changes to quadratic growth followed by sub-quadratic growth. For matrices in the linear range we expect to see obvious exponential convergence rates of solutions, and our test problems will bear this out.

The matrix $I^{(1)}$ is central in the construction of M . There is a well known conjecture in sinc theory that $\|(I^{(1)})^{-1}\|$ grows linearly with the size of the matrix, when it is invertible, and thus $\|(I^{(1)})^{-2}\|$ would grow quadratically. Even invertibility was unknown until recently when it was proved [3] that $I^{(1)}$ is invertible if the matrix is of even order. Thus, the exponential convergence of the method remains a conjecture. However, exponential convergence is demonstrated on the test problems in the following section.

5. NUMERICAL RESULTS

In Examples 5.1–5.4 below, we consider the steady-state problem

$$(5.1) \quad -((1 - 0.5e^{-2x})c_x)_x + c_x + c = f(x), \quad x > 0$$

$$(5.2) \quad c(0) - 0.5c_x(0) = G$$

$$(5.3) \quad c(\infty) = 0.$$

In each example the weight function γ of (3.13) was chosen to be

$$(5.4) \quad \gamma(x) = \frac{x}{(10 + x)^2}.$$

This choice ensures that $(\gamma(x)S(k, h) \circ \phi(x))'$ is continuous and integrable on $[0, \infty)$. Other choices of γ are possible, and the 10 in the denominator is arbitrary, but serves to move the singularity of $\gamma(x)$ away from the domain $(0, \infty)$ which results in greater accuracy in derivative approximation. However, this increase in accuracy tapers off as the singularity approaches $-\infty$.

In regards to the implementation of the code, there are several details we mention. First, if a solution is assumed to have exponential decay and the conformal

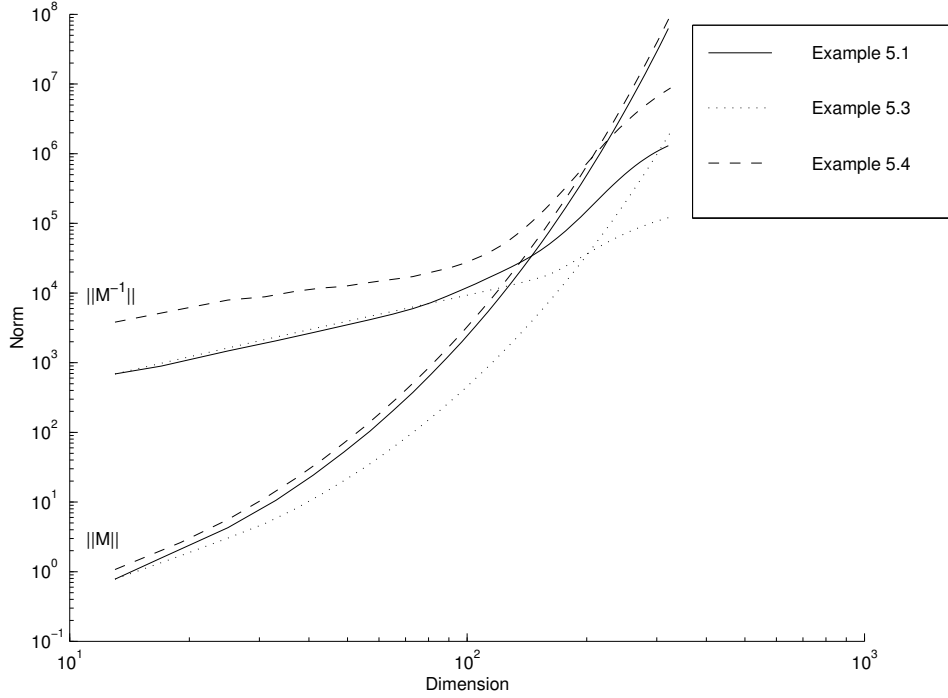


FIGURE 4.1. Log-log plot of norms of $\|M\|$ and $\|M^{-1}\|$ for the coefficient matrices of Examples 5.1, 5.3 and 5.4 against matrix dimension.

mapping $\phi(x)$ is used, then the prescription for computing N in (2.10), namely $N = \lceil \frac{\alpha}{\beta} M \rceil$, can be replaced by the formula

$$N = \lceil \frac{1}{h} \ln \left(\frac{\alpha}{\beta} M h \right) \rceil$$

while preserving the exponential accuracy of sinc approximations (see [5, p. 77] for a discussion of this point.) There results a considerable reduction in the size of the systems needed for the examples. This economization is applied to each of the following examples except Example 5.3, where the solution has polynomial decay. Moreover, there are cases in which $I^{(1)}$ could actually be a factor of the coefficient matrix M . To ensure nonsingularity, we adjust the size of the coefficient matrix by one if necessary, so that size is even. These results were generated using Octave. Versions of the programs used can be obtained from the authors (e.g., tshores@math.unl.edu).

Finally, there is the issue of determining the optimum convergence rate parameters d , α and β . In the absence of additional information it is usually safe to take the minimum values $\alpha = \beta = 1$ and $d \leq \pi/4$. These are conservative choices which work reasonably well in most cases. In fact, it is no accident that $\alpha = 1$ is the optimum choice in most cases. Consider the systems we solve. In the terminology of Section 3, the coefficient matrix is M and the right hand sides are $R_{dis} - T_0$ and T_1 . As we saw at the end of the preceding section, it is imperative that these terms be calculated exponentially accurately. Therefore, choices of α and β must be well suited to these integrals. There is no difficulty with the term $R_{dis} - T_0$. However, one of the terms in the integrand that is discretized in T_1 is the function

$$(d(x)q_1'(x) - v(x)q_1(x))w(x)\phi'(x)S(k, h)' \circ \phi(x).$$

It is easy to see that $F(x) = w(x)\phi'(x)S(k, h)' \circ \phi(x)$ tends to 0 as $x \rightarrow 0^+$ as $1/|\ln x|$. Recall that sinc quadrature mandates that α be chosen so that $F(x) = \mathcal{O}(x^{\alpha-1})$, $x \rightarrow 0^+$. Therefore, we should not choose α larger than 1 in general, though there may be cases, such as when $c(0)$ and $c'(0)$ are both 0, where $\alpha > 1$ works well.

In the following examples we denote the error at 0, $|c_0 - \tilde{c}(0)|$, by $|e(0)|$ and the sup-norm of the error at the sinc nodes by $\|e(x_k)\|_\infty$. In the first three examples, $c(0) = 1$, and in the last example $c(1.3) \approx 0.4$, so that in all cases absolute error is a good measure of relative error.

Example 5.1. (erfc-like decay) Here we chose $c(x) = e^{-x^2/4}$, and computed $f(x)$ accordingly. We also calculate $G = 1$. The function $p(x)$ was chosen as $p(x) = e^{-x^2}$. The sinc parameters of Theorem 2.4 were chosen to be $\alpha = 1, \beta = 1, d = \pi/3$. The algorithm was tested for several values of M_x . Norms of the coefficient matrix M and M^{-1} , along with the resulting errors are tabulated in Table 2.

Example 5.2. (exponential decay) Here we chose $c(x) = (1 + x^2)e^{-x}$ computed $f(x)$ accordingly, and chose $p(x) = e^{-x}$. We also calculate $G = 3/2$. Note that the coefficient matrix M is the same as in the previous example since the same choices were made for the sinc parameters α , β , and d . The algorithm was tested for several values of M_x , and the condition number of the coefficient matrix M and resulting error is tabulated in Table 2.

M_x	$\alpha = 1, \beta = 1, d = \pi/3$					
	m	h	$\ M\ _\infty$	$\ M^{-1}\ _\infty$	$ e(0) $	$\ e(x_k)\ _\infty$
8	12	0.64	1.49e+00	8.94e+02	5.67e-03	1.11e-02
16	22	0.45	1.22e+01	2.06e+03	1.75e-03	3.39e-03
32	40	0.32	2.53e+02	5.06e+03	4.89e-04	5.04e-04
64	78	0.22	1.46e+04	3.07e+04	2.27e-05	2.78e-05
128	148	0.16	5.78e+06	1.03e+05	1.69e-08	1.28e-06

TABLE 1. Example 5.1 with erfc-like decay.

M_x	$\alpha = 1, \beta = 1, d = \pi/3$					
	m	h	$\ M\ _\infty$	$\ M^{-1}\ _\infty$	$ e(0) $	$\ e(x_k)\ _\infty$
8	12	0.64	1.49e+00	8.94e+02	1.01e-03	5.76e-03
16	22	0.45	1.22e+01	2.06e+03	6.52e-04	1.12e-03
32	40	0.32	2.53e+02	5.06e+03	1.33e-03	1.33e-03
64	78	0.22	1.46e+04	2.88e+04	8.18e-06	3.01e-05
128	148	0.16	5.78e+06	1.03e+05	5.05e-08	2.14e-07

TABLE 2. Example 5.2 with exponential decay.

Inspection of Table 2 reveals a curious phenonemon: the error is not monotone decreasing in step size h . The spacing of nodes in sinc methods is not uniform, so that halving step size does not uniformly halve node separation. Thus, one might expect that decrease in error need not be strictly monotone in every case. However, doubling the error in going from a system of dimension 22 to dimension 40 seems disconcerting. Typically, such irregular convergence rates (or no convergence at all) signals a bad choice of parameters in sinc methods, which can be sensitive to the choice of parameters. Here are some guidelines that we have found useful, especially in situations in which the solution to the differential equation to be solved is really unknown (as opposed to our examples of systems generated by prescribed solutions.) We assume, of course, that *some* choice of parameters satisfies the conditions of Theorem 4.3:

- (1) Use available information to estimate convergence rate parameters α, β of the solution at 0 and infinity. As we noted earlier, in our setup, $\alpha = 1$ is

a good choice for most problems and β has to be estimated by behavior of the solution at ∞ .

- (2) Start with a conservative (small) choice of d and increase it to a maximum of $d = \pi/2$. (The motivation here is that a valid choice of d which is larger will improve the exponentially decreasing term of Theorems 2.4 and 2.5.)
- (3) Use inspection of an error table to search for improved estimates for α, β and d found in (1) and (2).

It is possible to observe rough convergence rates with an unknown solution. One way to do so is to find a large dimension M_x for which the system matrix is numerically nonsingular and use the resulting computed solution as the “true” solution by which error is measured. Equations (3.10) and (3.13) give us a formula for the “true” solution at any point in the interval of interest. Then calculate the node “error” for smaller choices of M_x and make out tables as we have done in Examples 5.1-5.2. Good *a posteriori* evidence that reasonable choices of the dimensions M_x have been made will be indicated by an error table that suggests exponential convergence rates.

To illustrate this strategy, we revisit Example 5.2 and imagine that we do not know the solution in advance. A crude asymptotic model for Equation (5.1) can be obtained by letting the terms of the differential equation pass to the limit as $x \rightarrow \infty$, resulting in

$$-c_{xx} + c_x + c = 0.$$

The only decaying solutions to this equation are scalar multiples of $e^{\lambda x}$ with $\lambda = (1 - \sqrt{5})/2 \approx -0.62$. This suggests that $\beta = 0.6$ might be a safer choice than $\beta = 1$. We used $\alpha = 1$, $\beta = 0.6$ and $d = \pi/4$ to start our investigation and found that we could increase d to the value $d = \pi/2.5$ with nonsingular coefficient matrix for $M_x = 150$ to generate our “true” solution. The results are recorded in Table 3 with $\|\hat{e}(x_k)\|_\infty$ denoting errors found by using the estimated “true” solution. These numbers compare quite favorable with the actual errors which are listed to the left of the estimated errors. Note also that the error for $M_x = 128$ in Table 3 is about half the corresponding error in Table 2. The overly optimistic estimate of Table 3 should be expected, since the approximate solution with $M_x = 150$ is used in place of the exact solution.

Example 5.3. (polynomial decay) Here we chose $c(x) = \frac{1+x^2}{1-x+x^4}$, computed $f(x)$, and chose $p(x) = e^{-x}$. We also calculate $G = 1/2$. The sinc parameters were taken

M_x	$\alpha = 1, \beta = 0.6, d = \pi/2.5$							
	m	h	$\ M\ _\infty$	$\ M^{-1}\ _\infty$	$ e(0) $	$ \widehat{e}(0) $	$\ e(x_k)\ _\infty$	$\ \widehat{e}(x_k)\ _\infty$
8	12	0.64	1.49e+00	8.94e+02	1.90e-02	1.90e-02	3.22e-02	3.22e-02
16	22	0.45	1.22e+01	2.06e+03	2.72e-03	2.72e-03	5.12e-03	5.12e-03
32	40	0.32	2.53e+02	5.06e+03	5.50e-05	5.50e-05	1.98e-04	1.98e-04
64	78	0.22	1.46e+04	2.88e+04	3.78e-06	3.78e-06	1.36e-05	1.36e-05
128	148	0.16	5.78e+06	1.03e+05	2.56e-08	1.33e-08	1.03e-07	1.02e-07

TABLE 3. Example 5.2 with exponential decay.

M_x	$\alpha = 1, \beta = 2, d = \pi/6$						
	m	h	$\ M\ _\infty$	$\ M^{-1}\ _\infty$	$ e(0) $	$\ e(x_k)\ _\infty$	
8	14	0.45	8.26e-01	6.91e+02	1.27e-01	6.15e-01	
16	26	0.32	2.80e+00	1.63e+03	1.65e-02	6.98e-02	
32	50	0.27	2.02e+01	3.82e+03	3.59e-03	1.11e-02	
64	98	0.16	3.90e+02	8.99e+03	2.76e-04	8.11e-04	
128	194	0.13	2.69e+04	3.23e+04	7.43e-06	5.63e-05	
256	386	0.08	1.09e+07	1.52e+05	7.74e-07	7.74e-07	

TABLE 4. Example 5.3 with polynomial decay.

to be $\alpha = 1, \beta = 2$, and $d = \pi/6$. The reason for a choice of d different from the preceding examples is that in the absence of solution knowledge, one would still be able to observe terms with denominator $q(x) = 1 - x + x^4$ in the right hand side $f(x)$. Now we want $f(x)$ to be analytic in $B(D)$, which is a wedge centered along the positive x -axis, with vertex at the origin and angle d from the positive x -axis. Since one of the roots of $q(x)$ is approximately $0.727 + 0.43i$, we choose a smaller angle to exclude this point from the wedge. Also, one could observe quadratic decay in $f(x)$, motivating the choice $\beta = 2$. The results are detailed in Table 4. One observes that due to the slower decay, in this example more nodes are needed to achieve accuracy comparable to previous examples.

M_x	m	Method 1			Method 2		
		$\ M\ _\infty$	$\ M^{-1}\ _\infty$	$\ e(x_k)\ _\infty$	$\ M\ _\infty$	$\ M^{-1}\ _\infty$	$\ e(x_k)\ _\infty$
8	14	9.89e-01	3.72e+03	3.52e-02	1.97e-02	6.31e+01	1.27e-03
16	26	5.55e+00	6.91e+03	9.60e-03	1.82e+03	4.17e+02	9.68e-05
32	50	7.93e+01	1.22e+04	1.13e-04	4.02e+04	1.16e+04	1.76e-06
64	98	2.70e+03	2.63e+04	2.08e-07	2.73e+06	1.18e+06	1.08e-08
128	194	4.84e+05	5.82e+05	2.41e-10	8.19e+08	singular	5.06e-11

TABLE 5. The new Sinc-Galerkin method (Method 1) and the traditional Sinc-Galerkin method of [5] (Method 2) applied to Example 5.4 with sinc parameters $\alpha = 1, \beta = 2, d = \pi/4$. The error $|e(0)|$ is not applicable for the traditional method, so not listed.

Example 5.4. (singular coefficient) In this example we consider

$$\begin{aligned}
-c_{xx} + c_x + \frac{3}{4x^2}c &= e^{-x}\left(\frac{9}{2}x^{1/2} - 2x^{3/2}\right), \quad x > 0 \\
c(0) - c_x(0) &= 0 \\
c(\infty) &= 0.
\end{aligned}$$

This problem has solution $c(x) = x^{3/2}e^{-x}$. The results of our method with $p(x) = e^{-x}$ and $\alpha = 1, \beta = 2, d = \pi/4$ are found in Table 5, and compared to the results of the Sinc-Galerkin method in Section 4.6 of [5] with $\alpha = 1, \beta = 2, d = \pi/4$. Also included is the condition number of the coefficient matrix M of each method. From the table, it is evident that the convergent rates are comparable, with the traditional Sinc-Galerkin method exhibiting slightly more rapid convergence and a more ill-conditioned coefficient matrix.

6. THE TIME DEPENDENT PROBLEM

Consider the time dependent problem

$$(6.1) \quad \rho(x)c_t + Lc = f(x, t), \quad x > 0, t > 0$$

$$(6.2) \quad Ac(0, t) + Bc_x(0, t) = G(t), \quad B \neq 0$$

$$(6.3) \quad c(\infty, t) = 0,$$

$$(6.4) \quad c(x, 0) = g(x)$$

where $Lc \equiv -(D(x)c_x)_x + (v(x)c)_x + \lambda(x)c$ and $c = c(x, t)$. Again we assume the parameters are chosen so that the problem is well posed.

Given a steady state solver such as we have developed, there are many time marching schemes one could develop for the general problem. Alternately, one could develop a fully Sinc-Galerkin method in space and time. In this section we examine a very simple example of time marching, namely, we apply a convex combination of the standard first order explicit and implicit time marching schemes to the problem above. With a time stepping increment of Δt and superscripts denoting values at time $t = k\Delta t$, such methods are described by the equation

$$(6.5) \quad \rho \frac{c^{k+1} - c^k}{\Delta t} + Lc^{k+1} = f,$$

where $f = f^k$ yields the explicit method and $f = f^{k+1}$ the implicit method. Multiply the former equation by $1 - \mu$ and the latter by μ to obtain the family of implicit methods

$$(6.6) \quad \rho \frac{c^{k+1} - c^k}{\Delta t} + \mu Lc^{k+1} + (1 - \mu)Lc^k = \mu f^{k+1} + (1 - \mu)f^k, \quad 0 < \mu \leq 1.$$

Regroup terms to obtain the form

$$(6.7) \quad \left(L + \frac{\rho}{\mu\Delta t}\right)c^{k+1} = f^{k+1} + \frac{1 - \mu}{\mu} \left(f^k - \left(L - \frac{\rho}{(1 - \mu)\Delta t}\right)c^k\right),$$

where it is understood that the second term on the right hand side is $\frac{\rho}{\Delta t}c^k$ for the fully implicit method $\mu = 1$.

We can now apply the methods of Section 3} to this problem after making a few slight adjustments in notation. Define operators and functions

$$(6.8) \quad L^+ = L + \frac{\rho}{\mu\Delta t}$$

$$(6.9) \quad L^- = L - \frac{\rho}{(1 - \mu)\Delta t}$$

$$(6.10) \quad \tilde{f} = f^{k+1} + \frac{1 - \mu}{\mu} \left(f^k - \left(L - \frac{\rho}{(1 - \mu)\Delta t}\right)c^k\right)$$

and corresponding bilinear forms T^+, T^- as in Section 3 to obtain the steady state differential equation at the k th time step in the unknown c^{k+1}

$$(6.11) \quad L^+ c^{k+1} = \tilde{f}$$

and variational form (with R, u as in Section 3)

$$(6.12) \quad T^+(c^{k+1}, u) = R(\tilde{f})$$

$$(6.13) \quad = R(f^{k+1}) + \frac{1 - \mu}{\mu} (R(f^k) - T^-(c^k, u))$$

The suitable choice of weighting factor depends on the problem. If the problem exhibits stiff behavior in time, μ close to 1 might be a good choice. On the other hand, the choice $\mu = 1/2$ gives a method which is second order in time (Crank-Nicolson is a special case.) If one uses finite difference methods, this second order accuracy will not materialize unless the operator L is carefully discretized and possibly other restrictions on step sizes such as the Courant number $\mu = \Delta t / \Delta x^2 \leq 1/2$ are applied. If the discretization error is even in powers of Δt and Δx , one can employ Richardson extrapolation to increase the order of accuracy in even powers of the step sizes. An advantage of using sinc methods on the spatial operator L is that spatial step size discretization error is exponentially small for sufficiently large node number M_x . Thus, one can expect that one step of Richardson extrapolation should lead to a fourth order method in time.

It should be emphasized in general one can only expect extrapolation to reduce the time stepping error to the level of spatial discretization error. Further time extrapolation should have little effect on the spatial discretization error, so that for a sufficiently small time step size, we should not see any improvement by extrapolating the solution.

We illustrate these points in the following example, where fairly modest values of M_x are employed. We have constructed this example so that at $T = 2$ the spatial problem is exactly equivalent to the problem of Example 4.4. Therefore, Table 1 should offer us an approximate floor on the error level that we can attain by reducing step size and/or extrapolating.

Example 6.1. Consider the following time dependent test problem.

$$(6.14) \quad \left(\frac{2+x}{1+x} \right) c_t - ((1 - 0.5e^{-2x})c_x)_x + c_x + c = f(x, t), \quad t > 0, \quad x > 0,$$

$$(6.15) \quad c(0, t) - 0.5c_x(0, t) = 1 - \cos(\pi t/4), \quad t > 0$$

$$(6.16) \quad c(\infty, t) = 0, \quad t > 0$$

$$(6.17) \quad c(x, 0) = 0, \quad x > 0.$$

The function $f(x, t)$ was computed for $c(x, t) = e^{-tx^2/8}(1 - \cos(\pi t/4))$, and the solution at time $T = 2$ was computed for $p(x) = e^{-x^2}$, $\alpha = 1, \beta = 1, d = \pi/3$, $M_x = 32$ and $M_x = 64$, time-stepping parameters $\Delta t = 1, 0.5, 0.4, 0.2, 0.1, 0.05$ and $\mu = 0.5$. We expect halving step size to reduce error by about four, which is evident from Table 6. In the absence of effect from spatial discretization, we also expect Richardson extrapolation to reduce the error by a factor of about 16. The

Δt	M_x	$\ e(x_k)\ _\infty$	EERF	Δt	M_x	$\ e(x_k)\ _\infty$	EERF
0.5	32	7.45e-03		0.5	64	7.20e-03	
0.25	32	1.99e-03		0.25	64	1.79e-03	
R.E.		5.39e-04	13.8	R.E.		2.40e-04	30
0.1	32	6.31e-04		0.1	64	2.96e-04	
0.05	32	4.72e-04		0.05	64	8.15e-05	
R.E.		4.19e-04	1.5	R.E.		1.97e-05	15.4

TABLE 6. Example 6.1: error table for the solution at time $T = 2$ with sinc parameters $\alpha = 1$, $\beta = 1$, and $d = \pi/3$ including the error after Richard extrapolation (R.E.) and the extrapolation error reduction factor (EERF).

table reflects these expectations: in the last column we exhibit the extrapolation error reduction factors (EERF) obtained by dividing the full step error $\|e(x_k)\|_\infty$ by the corresponding error after Richardson extrapolation. Further extrapolations with smaller step sizes only worsened the error in the case of $M_x = 32$. Notice that the lowest error of $4.72e - 04$ is to be compared with the counterpart error of $5.04e - 04$ listed in Table 1. It is somewhat surprising to see a slightly smaller error when time stepping is used as opposed to the steady state problem of Example 4.4. However, this result is an artifact of time stepping for this particular problem and should not be expected in general. The improvement can be explained by examining the actual error at $x = 0$. In the case of Example 4.4 we found that the approximate solutions (any M_x) undershot the correct value. On the other hand, the solutions in our time stepping example overshoot the exact value at $x = 0$. Richardson extrapolation then gave an approximation that undershot the exact value by a serendipitously smaller amount than the steady state problem with $M_x = 32$.

Notice in Table 1 that the EERF of 1.5 for Richardson extrapolation when $M_x = 32$ and $\Delta t = 0.1, 0.05$ is disappointingly small. This suggests that the time discretization error is dominated by spatial error, which time extrapolation cannot expect to reduce. However, in the case of $M_x = 64$ with $\Delta t = 0.1, 0.05$ the table suggests that spatial error is dominated by time discretization error since Richardson extrapolation gives a better EERF of 15.4. Table 1 lists an error of $2.78e - 05$ for the corresponding $M_x = 64$ steady state problem. This suggests that further improvements will not be realized by reducing time step sizes below the

values $\Delta t = 0.1, 0.05$ displayed in Table 6, and further calculations (which we did not display) confirmed this speculation.

7. CONCLUSIONS

We have presented a new Sinc-Galerkin method for solving the convection-diffusion equation on a half-infinite interval with mixed boundary conditions. The method is developed by recasting the original problem into a variational form. The method is suitable as a forward solver for the inverse coefficient problem of determining the dispersivity coefficient as it does not involve differentiation of the coefficients. An exponential rate of convergence was demonstrated on several test problems, and this rapid rate of convergence is even maintained in the presence of end-point singularities in the coefficients of the convection-diffusion equation.

8. ACKNOWLEDGMENTS

The author J.M. was supported in part by EPA Graduate Fellowship U-914813-01-0 and the University of Nebraska. The authors thank the anonymous referees for their careful reading of the paper and constructive comments.

REFERENCES

- [1] B. Bialecki. Sinc-collocation methods for two-point boundary value problems. *IMA J. Numer. Anal.*, 11:357–375, 1991.
- [2] G. de Marsily. *Quantitative Hydrogeology*. Academic Press, San Diego, 1986.
- [3] P. Gierke. *Discrete Approximations of Differential Operators by Sinc Methods*. PhD thesis, University of Nebraska, Lincoln, NE, USA, 1999.
- [4] J. Lund. Sinc approximation method for coefficient identification in parabolic systems. In J. van Schuppen M. Kaashoek and A. Ran, editors, *Robust Control of Linear Systems and Nonlinear Control*, volume 4 of *Progress in Systems and Control Theory*, pages 507–514, Boston, 1990. Birkhauser.
- [5] J. Lund and K. Bowers. *Sinc Methods for Quadrature and Differential Equations*. SIAM, Philadelphia, 1992.
- [6] J. Lund and C. Vogel. A fully-galerkin method for the numerical solution of an inverse problem in a parabolic partial differential equation. *Inverse Problems*, 6:205–217, 1990.
- [7] J. Mueller. *Inverse Problems for Singular Differential Equations*. PhD thesis, University of Nebraska, Lincoln, NE, USA, 1997.
- [8] J. Mueller and T. Shores. Uniqueness and numerical recovery of a potential on the real line. *Inverse Problems*, 13:289–303, 1997.
- [9] N.N. Ural'ceva O.A. Ladyzenskaja, V.A. Solonnikov. *Linear and Quasilinear Equations of Parabolic Type*. American Mathematical Society, Providence, 1968.

- [10] R. Smith and K. Bowers. Sinc-galerkin estimation of diffusivity in parabolic problems. *Inverse Problems*, 9:113–135, 1993.
- [11] F. Stenger. A sinc-galerkin method of solution of boundary value problems. *Math. Comp.*, 33:85–109, 1979.
- [12] F. Stenger. *Numerical Methods Based on Sinc and Analytic Functions*. Springer-Verlag, New York, 1993.
- [13] M. Th. van Genuchten and J.C. Parker. Boundary conditions for displacement experiments through short laboratory soil columns. *Soil Sci. Soc. Am. J.*, 48:703–708, 1984.

Jennifer Mueller: jmueller@math.colostate.edu

Department of Mathematics

Colorado State University

Fort Collins, CO 80523

Thomas Shores: tshores@math.unl.edu

Department of Mathematics

University of Nebraska

Lincoln, NE 68588