LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-PROC-209167

Global Error Bounds for the Petrov-Galerkin Discretization of the Neutron Transport Equation

B. Chang, P.N. Brown, A. Greenbaum, E. Machorro

January 24, 2005

Nuclear Explosive Code Design Conference
Livermore , CA, United States
October 4, 2004 through October 7, 2004

This work was performed under the auspices of the U. S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Global Error Bounds for the Petrov-Galerkin Discretization of the Neutron Transport Equation

Britton Chang [*] Peter Brown [*] Anne Greenbaum [†]
Eric Machorro [‡]

January 20, 2005

Abstract. In this paper, we prove that the numerical solution of the mono-directional neutron transport equation by the Petrov-Galerkin method converges to the true solution in the L^2 norm at the rate of h^2. Since consistency has been shown elsewhere, the focus here is on stability. We prove that the system of Petrov-Galerkin equations is stable by showing that the 2-norm of the inverse of the matrix for the system of equations is bounded by a number that is independent of the order of the matrix. This bound is equal to the length of the longest path that it takes a neutron to cross the domain in a straight line. A consequence of this bound is that the global error of the Petrov-Galerkin approximation is of the same order of h as the local truncation error. We use this result to explain the widely held observation that the solution of the Petrov-Galerkin method is second accurate for one class of problems, but is only first order accurate for another class of problems.

Key words: linear hyperbolic equation, neutron transport

AMS subject classifications. 82D75, 65M99, 35Q99

Introduction

It is easy to prove that the local truncation errors of the Petrov-Galerkin Method [4] and the closely related Diamond Difference Method [2] in multi-dimensions are second order with respect to the grid spacing h [4], [6], [7] and [8]. However much less is known about their global errors except in 1-D slab

[*]Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, P.O. Box 808 L-561, Livermore, CA 94551. *email*: bchang@llnl.gov. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract no. W-7405-Eng-48.

[†]Dept. of Mathematics, Box 354350 University of Washington, Seattle, WA 98195-4350.

[‡]Dept. of Applied Mathematics, Box 352420 University of Washington, Seattle, WA 98195-2420.

geometry. For the Diamond Difference Method, the global error in 1-D slab geometry has been shown to be second order [7] and [8]. Since the Diamond Difference Method and the Petrov-Galerkin Method in 1-D slab geometry are equivalent within the discretization error of h^2, then the global error of the Petrov-Galerkin solution in 1-D slab geometry is also second order. In transport problems involving spherical coordinates, the authors of [4] proved that the Petrov-Galerkin Method is locally second order accurate in the 1-D spherical coordinate system, but did not provide an estimate for the global error.

The global error is defined to be the difference between the computed solution and the true solution, while the local truncation error is defined to be the error by which the true solution fails to satisfy the discrete system of equation of the approximation method. We will derive a system of equations that relates the global error to the local truncation error. As we shall see, this system of equations is the Petrov-Galerkin system of equations in which the unknown is the the global error and the source is the local truncation error. This equation shows that the global error is the accumulation of local truncation errors. This accumulation occurs for the same reason that the angular flux is the accumulation of source particles. As we shall also see that the Petrov-Galerkin system of equations is equivalent to a system of finite difference equations. It is important to recognize that while the local truncation error of a finite difference method may be proportional to a power of the grid spacing, the finite difference solution may diverge from the true solution as the grid spacing is made vanishingly small [5]. The reason is because the finite difference solution may accumulate so much local truncation error that it diverges from the true solution. A stable finite difference method is one in which the accumulation of the local truncation errors is kept to a minimum.

An a-priori estimate of the global error can be used to predict the accuracy of the numerical results for problems that are without analytical solutions. For example, we can use the results of this paper to to explain why the Petrov-Galerkin Method gives second order accurate solutions for one class of problems, but gives only first order accurate solutions for another class of problems. Similar dichotomous results have been observed in the solutions of the Diamond Difference Method [9]. The reason for these differences in the accuracy of the Petrov-Galerkin solution is because the global error depends on the smoothness of the true solution through the local truncation error. As we shall see, if the true solution does not have enough continuous derivatives to yield a second order accurate local truncation error, then the Petrov-Galerkin solution converges at a first order rate. We will show that the smoothness of the true solution depends on the smoothness of the source term on the right hand side (rhs) of the transport equation. A consequence of this work is that the Petrov-Galerkin solution converges at a first order rate for problems with discontinuous sources.

We use finite difference theory to prove our results even though the Petrov-Galerkin approximation is a finite element method. As it was recognized in [4], the Petrov-Galerkin system of equations is, within the local truncation error,

equivalent to a system of finite difference equations. In 1-D , the finite difference equations are the Diamond Difference equations. The Petrov-Galerkin system of equations differs from a finite difference system of equation by the rhs's of equations. The rhs of a Petrov-Galerkin equation is the average integral of the source over a finite element, while the rhs of a finite difference equation is the value of the source which is evaluated at a point inside of the finite element. If we approximate the integral of the source by the mid-point rule [4], we find that the difference between the two rhs's to be of the order h^2.

We prove the Petrov-Galerkin solution converges to the true solution by showing that the equivalent system of finite difference equations is consistent and stable. The Petrov-Galerkin system of equation is consistent because the local truncation error vanishes as h^2, if the true solution has continuous second derivatives. Since consistency has been shown in [4], we will not dwell on it in this paper. However, for completeness, we provide a derivation of the local truncation error in §. We prove the Petrov-Galerkin system of equations is stable by showing that the linear algebraic norm of the matrix that represents the Petrov-Galerkin system of equations is independent of the grid spacing (or independent of the order of the matrix since the order of the matrix depends on the grid spacing). This result shows that the norm of the global error of the Petrov-Galerkin solution is proportional to the norm of the local truncation error, which shows that the Petrov-Galerkin solution converges to the true solution at the rate that is determined by the local truncation error.

In §, we derive the Petrov-Galerkin system of equations for neutron transport. In §, we define the truncation error, the global error and derive the equation that relates these two types of errors. In §, we prove stability for the Petrov-Galerkin system of equations in 1-D and 2-D. In §, we provide numerical results to support the main conclusion of this paper. The convergence rate of the Petrov-Galerkin method is second order for the solutions of problems with continuous sources, but falls to first order for the solutions of problems with discontinuous sources. We summarize the results of this paper in §.

The Single Group and Mono-Directional Neutron Transport Equation

Let $\mathcal{D}$ be the 2-dimensional Cartesian domain, $\mathcal{D} \equiv [0, a_1] \times [0, a_2]$, $\mathbf{x}$ be the 2-dimensional vector in $\mathcal{D}$, Ω be a unit vector in R^2, $\sigma \geq 0$ be the absorption cross section, and $q(\mathbf{x}) \geq 0$ be an external source of neutrons. In order to derive the matrix inverse of the Petrov-Galerkin system of equations, we make the simplifying assumption that the absorption cross section is constant. The transport of a single group of mono-directional neutrons in a non-scattering medium is described by

$$\Omega \cdot \nabla \psi(\mathbf{x}) + \sigma \psi(\mathbf{x}) = q(\mathbf{x}), \tag{1}$$

with the boundary condition

$$\psi(\mathbf{x}) = \psi_b(\mathbf{x}), \quad \mathbf{n}\cdot\Omega < 0, \quad \mathbf{x}\in\partial\mathcal{D}. \tag{2}$$

Let μ and η, the components of $\Omega \equiv (\mu,\eta)^T$, be positive for this presentation. In this paper we assume that the second derivatives of $q(\mathbf{x})$ are bounded so that the local truncation can be second order.

Petrov-Galerkin System of Equations

We introduce the uniform mesh spacing $\Delta x = a_1/n_x,\ \Delta y = a_2/n_y$. Let $\{(x_i,y_j) = (i\Delta x, j\Delta y)\ :\ i = 0,1,\ldots,n_x;\ j = 0,1,\ldots,n_y\}$ be the nodes of our grid. The Petrov-Galerkin Method consists of approximating the true solution by piece-wise bi-linear trial functions

$$\psi^h(x,y) \equiv \sum_{j=0}^{n_2}\sum_{i=0}^{n_1} u_{i,j}\ \phi_i(x)\,\phi_j(y)\ , \tag{3}$$

where $\phi_i(x)$ is the hat function

$$\phi_i(x) = \begin{cases} \frac{x-x_{i-1}}{\Delta x}\ , & x_{i-1}\le x\le x_i\ , \\ \frac{x_{i+1}-x}{\Delta x}\ , & x_i\le x\le x_{i+1}\ . \end{cases}$$

To obtain the Petrov-Galerkin equations, substitute (3) into (1) and then average over zone $x_{i-1}\le x\le x_i,\ \ y_{j-1}\le y\le y_j$ to give

$$\frac{\mu}{\Delta x}\left(\frac{u_{i,j}+u_{i,j-1}}{2} - \frac{u_{i-1,j}+u_{i-1,j-1}}{2}\right) + \frac{\eta}{\Delta y}\left(\frac{u_{i,j}+u_{i-1,j}}{2} - \frac{u_{i,j-1}+u_{i-1,j-1}}{2}\right) + \sigma\left(\frac{u_{i,j}+u_{i,j-1}}{4} + \frac{u_{i-1,j}+u_{i-1,j-1}}{4}\right) = q_{i,j} \tag{4}$$

for $i = 1,\ldots,n_1$, and $j = 1,\ldots,n_2$, where

$$q_{i,j} = \frac{1}{\Delta x\Delta y}\int_{y_{j-1}}^{y_j} dy \int_{x_{i-1}}^{x_i} dx\ q(x,y)\ . \tag{5}$$

For the boundary conditions (2), we have

$$\begin{aligned} u_{i,\,0} &= \psi_b(x_i,y_0)\ , && i = 0,1,\ldots,n_1\ , \\ u_{0,j} &= \psi_b(x_0,y_j)\ , && j = 1,2,\ldots,n_2\ . \end{aligned} \tag{6}$$

The Global Error and the Local Truncation Error

We now turn to the consideration of the global error $\{\epsilon_{i,j}\}$ defined by

$$\epsilon_{i,j} \equiv u_{i,j} - \psi(x_i,y_j)\ . \tag{7}$$

We also need a measure of the error by which the true solution fails to satisfy the difference equations (4). This error is called the local truncation error $\{\tau_{i,j}\}$ and is defined by the substitution of the true solution into (4)

$$\begin{aligned}
&\frac{\mu}{\Delta x}\left(\frac{\psi(x_i,y_j)+\psi(x_i,y_{j-1})}{2}-\frac{\psi(x_{i-1},y_j)+\psi(x_{i-1},y_{j-1})}{2}\right)\\
+&\frac{\eta}{\Delta y}\left(\frac{\psi(x_i,y_j)+\psi(x_{i-1},y_j)}{2}-\frac{\psi(x_i,y_{j-1})+\psi(x_{i-1},y_{j-1})}{2}\right)\\
&+\sigma\left(\frac{\psi(x_i,y_j)+\psi(x_i,y_{j-1})}{4}+\frac{\psi(x_{i-1},y_j)+\psi(x_{i-1},y_{j-1})}{4}\right) \;\equiv\; q_{i,j}-\tau_{i,j} \quad (8)
\end{aligned}$$

It was shown in [4] that the local truncation error $\{\tau_{i,j}\}$ is of the order h^2. For completeness, a derivation is given in the § of this paper.

We also need the local truncation errors $\{\tau_i^{(s)}\}$ and $\{\tau_j^{(w)}\}$ by which the true solution fails to satisfy the discrete boundary conditions (6). The substitution of the true solution into (6) yields

$$\begin{aligned}
\psi(x_i,y_0) &= \psi_b(x_i,y_0)+\tau_i^{(s)}\;, \qquad & i=0,1,\ldots,n_1\;,\\
\psi(x_0,y_j) &= \psi_b(x_0,y_j)+\tau_j^{(w)}\;, \qquad & j=1,2,\ldots,n_2\;.
\end{aligned} \quad (9)$$

Comparing (9) and (2), we find the truncation errors that are on the boundary to be zero, i.e. $\{\tau_i^{(s)}=0, i=0,\ldots,n_1\}$ and $\{\tau_j^{(w)}=0, j=1,\ldots,n_2\}$.

The subtraction of (8) from (4) yields

$$\begin{aligned}
&\frac{\mu}{\Delta x}\left(\frac{\epsilon_{i,j}+\epsilon_{i,j-1}}{2}-\frac{\epsilon_{i-1,j}+\epsilon_{i-1,j-1}}{2}\right)+\frac{\eta}{\Delta y}\left(\frac{\epsilon_{i,j}+\epsilon_{i-1,j}}{2}-\frac{\epsilon_{i,j-1}+\epsilon_{i-1,j-1}}{2}\right)\\
&\qquad+\sigma\left(\frac{\epsilon_{i,j}+\epsilon_{i,j-1}}{4}+\frac{\epsilon_{i-1,j}+\epsilon_{i-1,j-1}}{4}\right)=\tau_{i,j}\;, \quad (10)
\end{aligned}$$

and the subtraction of (9) from (6) gives

$$\begin{aligned}
\epsilon_{i,0} &= \tau_i^{(s)} = 0\;, \qquad & i=0,1,\ldots,n_1\;,\\
\epsilon_{0,j} &= \tau_j^{(w)} = 0\;, \qquad & j=1,2,\ldots,n_2\;.
\end{aligned} \quad (11)$$

We see that the global error $\{\epsilon_{i,j}\}$ is determined by a Petrov-Galerkin system of equations (10) with homogeneous boundary conditions (11) and a source term which is the local truncation error $\{\tau_{i,j}\}$. If we eliminate the boundary terms $\{\epsilon_{i,0} : i=0,\ldots,n_1\}$ and $\{\epsilon_{0,j} : j=1,\ldots,n_2\}$ in (10), we are left with $n_1 n_2$ unknowns $\{\epsilon_{i,j} \;:\; i=1,\ldots,n_1,\; j=1,\ldots,n_2\}$. We will show below that this reduced system can be written in matrix-vector notation as $H\epsilon=\tau$. An estimate of the global error is $\|\,\epsilon\,\|_2 \le \|H^{-1}\|_2\,\|\,\tau\,\|_2$, where $\|\cdot\|_2$ is the linear algebraic l_2-norm, i.e.

$$\|\,\epsilon\,\|_2^2=\sum_{j=1}^{n_2}\sum_{i=1}^{n_1}\epsilon_{i,j}^2 \quad .$$

In order to compare our theoretical estimates with numerical results, we define the *average* l_2-norm, $\|\cdot\|_{L^2(\mathcal{D})}$,

$$\|\cdot\|^2_{L^2(\mathcal{D})} \equiv \frac{1}{n_1 n_2}\|\cdot\|_2^2 \tag{12}$$

Dividing the inequality $\|\epsilon\|_2 \leq \|H^{-1}\|_2 \|\tau\|_2$, by $(n_1 n_2)^{1/2}$ and using (12), we have

$$\|\epsilon\|_{L^2(\mathcal{D})} \leq \|H^{-1}\|_2 \|\tau\|_{L^2(\mathcal{D})}.$$

There are three reasons for normalizing the l_2-norm in this way. The first is because $\|\tau\|_2^2$ is approximately $n_1 n_2 \cdot O(h^{2r})$ if each element of τ is of $O(h^r)$. The second is because

$$\|\epsilon\|_{L^2(\mathcal{D})} = \left(\frac{1}{n_1 n_2}\sum_{j=1}^{n_2}\sum_{i=1}^{n_1}(u_{i,j} - \psi(x_i, y_j))^2\right)^{\frac{1}{2}}$$

can be thought of as the root mean square error. The third is because $\|H^{-1}\|_2$ is the $h-$independent constant, C, of the inequality $\|\epsilon\|_{L^2(\mathcal{D})} \leq C\|\tau\|_{L^2(\mathcal{D})}$.

It was shown in [4] that $\|\tau\|_{L^2(\mathcal{D})}$ is of order h^2. However τ depends also on the second and higher derivatives of the true solution ψ. As shown in § of this paper and [4], τ is derived by expanding ψ of (8) in a Taylor series and is the leading term of the expansion. If there is a break in the Taylor series expansion, e.g. if a second derivative of ψ is unbounded, then leading term of (8) will be lowered than second order. This means that the true solution is not smooth enough to be interpolated by the linear basis functions of the Petrov-Galerkin method.

We calculate an upper bound for $\|H^{-1}\|_2$ by deriving an explicit formula for H^{-1} from which we can calculate $\|H^{-1}\|_1$ and $\|H^{-1}\|_\infty$. We then use the norm inequality, $\|A\|_2^2 \leq \|A\|_1 \cdot \|A\|_\infty$ for any matrix A, provided by linear algebra [3], to bound $\|H^{-1}\|_2$. The formulas for $\|H^{-1}\|_1$ and $\|H^{-1}\|_\infty$ are derived by exploiting the rich set of symmetries possessed by the Petrov-Galerkin system of equations which is not obvious in the form in which it is written. In order to gain insight into the symmetries, we investigate (10) and (11) in increasing levels of difficulty. We start with the simplest case, which is the 1-D transport in a vacuum ($\sigma = 0$). The second, but more difficult case, is the 1-D transport in an absorber ($\sigma \neq 0$). The most difficult case addressed in this paper, and is the general case of (1), is the 2-D transport in an absorber.

The Global Error for 1-D Transport in a Vacuum

The 1-D transport in a vacuum gives us insight into how the norm of the matrix inverse depends on the order of the matrix and on the mesh spacing. The Petrov-Galerkin equations (10) and (11) in 1-D and in a vacuum $\sigma = 0$ simplify to

$$\epsilon_0 \;=\; 0 \;, \tag{13}$$

$$\frac{\mu}{\Delta x}\left(\epsilon_i - \epsilon_{i-1}\right) \;=\; \tau_i \;, \qquad i = 1, \ldots, n_1 \;. \tag{14}$$

We eliminate ϵ_0 from (14) by substituting (13) into the $i = 1$ equation of (14) to give the reduced system

$$\begin{aligned} \frac{\mu}{\Delta x}\epsilon_1 &= \tau_1 \\ \frac{\mu}{\Delta x}\left(\epsilon_i - \epsilon_{i-1}\right) &= \tau_i \;, \qquad i = 2, \ldots, n_1 \;. \end{aligned} \tag{15}$$

If we introduce the $n_1 th$ order identity matrix I_{n_1}

$$(I_{n_1})_{i,j} = \delta_{i,j} \;, \qquad i, j = 1, 2, \ldots, n_1 \;,$$

where $\delta_{i,j}$ is the Kronecker delta, and L_{n_1}, the $n_1 th$ order square nilpotent matrix of degree n_1, i.e. $L_{n_1}^{n_1} = 0$,

$$(L_{n_1})_{i,j} \;=\; \delta_{i,j+1} \;, \quad i, j = 1, 2, \ldots, n_1 \;,$$

$$L_{n_1} \;=\; \begin{pmatrix} 0 & & & . & . & . \\ 1 & 0 & & . & . & . \\ 0 & 1 & 0 & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & 0 & 1 & 0 & . \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

we can write (15) in the matrix-vector notation

$$H_{n_1}\epsilon = \tau \;,$$

where

$$H_{n_1} \equiv \frac{\mu}{\Delta x}\left(I_{n_1} - L_{n_1}\right) \;,$$

$\epsilon = (\epsilon_1, \epsilon_2, \ldots., \epsilon_{n_1})^T$, and

$$\tau = \begin{pmatrix} \tau_1 \\ \tau_2 \\ . \\ . \\ . \\ \tau_{n_1} \end{pmatrix} \;.$$

The matrix inverse of H_{n_1}, can be written as

$$H_{n_1}^{-1} = \frac{\Delta x}{\mu}\left(I_{n_1} - L_{n_1}\right)^{-1} \;,$$

but is devoid of any intuition. However by recognizing that the $n_1 th$ power of L_{n_1} vanishes, i.e. $L_{n_1}^{n_1} = 0$, then the Neumann series of $(I_{n_1} - L_{n_1})^{-1}$,

$$H_{n_1}^{-1} = \frac{\Delta x}{\mu} \left(I_{n_1} + L_{n_1} + L_{n_1}^2 + \ldots + L_{n_1}^{(n_1-1)} \right) , \tag{16}$$

truncates after the $n_1 th$ term, which can be verified by the simple calculation

$$\begin{aligned} \left(I_{n_1} + L_{n_1} + L_{n_1}^2 + \ldots + L_{n_1}^{(n_1-1)} \right) (I_{n_1} - L_{n_1}) \quad &= \\ \left(I_{n_1} + L_{n_1} + L_{n_1}^2 + \ldots + L_{n_1}^{(n_1-1)}\right) - \quad &\left(L_{n_1} + L_{n_1}^2 + \ldots + L_{n_1}^{(n_1-1)}\right) = I_{n_1}. \end{aligned}$$

With the aid of (16), we can derive the 1-norm and the $\infty-$norm of $H_{n_1}^{-1}$.

The powers of L_{n_1} are easy to calculate since the multiplication of L_{n_1} with a power of L_{n_1} shifts the non-zero diagonal of the power of L_{n_1} down by one diagonal position. The powers of L_4 are, for example,

$$L_4 = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & 0 \end{pmatrix}, \quad L_4^2 = \begin{pmatrix} 0 & & & \\ 0 & 0 & & \\ 1 & 0 & 0 & \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad L_4^3 = \begin{pmatrix} 0 & & & \\ 0 & 0 & & \\ 0 & 0 & 0 & \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Therefore H_4^{-1} is explicitly

$$H_4^{-1} = \frac{\Delta x}{\mu} \begin{pmatrix} 1 & & & \\ 1 & 1 & & \\ 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 \end{pmatrix} .$$

The above equation shows that H_4^{-1} is a lower triangular non-negative matrix, and that the row sum of the ith column is smaller than the row sum of the $(i-1)th$ column. Therefore the 1-norm of $H_{n_1}^{-1}$ is the row sum of first column of $H_{n_1}^{-1}$, i.e.

$$\left\| H_{n_1}^{-1} \right\|_1 = \frac{\Delta x}{\mu} \, n_1 = \frac{a_1}{\mu} \ . \tag{17}$$

Thus $\left\| H_{n_1}^{-1} \right\|_1$ is independent of Δx.

From the symmetry of $H_{n_1}^{-1}$, we observe that the column sum of last row of $H_{n_1}^{-1}$ is equal to the row sum of the first column of $H_{n_1}^{-1}$. Therefore we have the extraordinary coincidence that

$$\left\| H_{n_1}^{-1} \right\|_\infty = \left\| H_{n_1}^{-1} \right\|_1 .$$

Since any matrix A satisfies the inequality, $\| A \|_2 \ \leq \sqrt{\|A\|_1 \, \|A\|_\infty}$, [3], then

$$\left\| H_{n_1}^{-1} \right\|_2 \leq \frac{a_1}{\mu} \ , \tag{18}$$

which shows that 2-norm of the matrix $H_{n_1}^{-1}$ of any order n_1 is bounded by a_1 the diameter of the domain. For this simple example, we can even derive an explicit expression for $\|H_{n_1}^{-1}\|_2$. Let us take a moment to derive this result, so that we can use it to show that (18) is a tight upper bound for $\|H_{n_1}^{-1}\|_2$.

Since the 2-norm of $H_{n_1}^{-1}$ is the square root of the largest eigenvalue of $H_{n_1}^{-1,T} H_{n_1}^{-1}$, then the 2-norm of $H_{n_1}^{-1}$ is the inverse of the square root of the smallest eigenvalue of $H_{n_1}^T H_{n_1}$. Noting that

$$L_{n_1}^T L_{n_1} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 0 \end{pmatrix} ,$$

we have

$$\begin{aligned} H_{n_1}^T H_{n_1} &= \left(\frac{\mu}{\Delta x}\right)^2 \left(I_{n_1} - L_{n_1} - L_{n_1}^T + L_{n_1}^T L_{n_1}\right) \\ &= \left(\frac{\mu}{\Delta x}\right)^2 \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} . \end{aligned} \tag{19}$$

Recall that (19) is the discrete matrix for Laplace's equation,

$$-\mu^2 \frac{d^2 v}{dx^2} = 0 ,$$

with the Dirichlet and the Neumann boundary conditions,

$$v(0) = 0 \quad \text{and} \quad \frac{\mathrm{dv(a_1)}}{\mathrm{dx}} = 0 ,$$

on the left and the right end points respectively. The eigenvalues of this discrete Laplacian are

$$\lambda_k\left(H_{n_1}^T H_{n_1}\right) = \left(\frac{\mu}{\Delta x}\right)^2 \sin^2\left(\frac{2k-1}{4n_1+2}\,\pi\right), \qquad k = 1, \ldots, n_1 .$$

Substituting $\Delta x = a_1/n_1$ into the above equation, we find

$$\|H_{n_1}^{-1}\|_2 = \lambda_1^{-\frac{1}{2}} = \frac{a_1}{\mu} \frac{1}{n_1 \sin\left(\frac{\pi}{4n_1+2}\right)} .$$

As the order of H_{n_1} increases, we find (with (17))

$$\lim_{n_1 \to \infty} \|H_{n_1}^{-1}\|_2 = \frac{2}{\pi} \frac{a_1}{\mu} = \frac{2}{\pi} \|H_{n_1}^{-1}\|_1 ,$$

which is $2/\pi$ times smaller than the upper bound for $\|H_{n_1}^{-1}\|_2$ estimated in (18). Thus (18) over-estimates $\|H_{n_1}^{-1}\|_2$ by approximately 36%.

The Global Error for 1-D Transport in an Absorbing Medium

The goal of this section is to show that $\|H_{n_1}^{-1}\|_1$ is bounded by a function that decreases with increasing absorption cross section. When $\sigma \neq 0$, (10) in 1-D is

$$\frac{\mu}{\Delta x}\left(\epsilon_i - \epsilon_{i-1}\right) + \frac{\sigma}{2}\left(\epsilon_i + \epsilon_{i-1}\right) \;=\; \tau_i \;, \qquad i = 1, \ldots, n_1 \;. \tag{20}$$

The elimination of the boundary condition (13) from (20) gives

$$\begin{aligned} \left(\frac{\mu}{\Delta x} + \frac{\sigma}{2}\right)\epsilon_1 &= 0 \\ \left(\frac{\mu}{\Delta x} + \frac{\sigma}{2}\right)\epsilon_i - \left(\frac{\mu}{\Delta x} - \frac{\sigma}{2}\right)\epsilon_{i-1} &= \tau_i \;, \qquad i = 2, \ldots, n_1 \;, \end{aligned} \tag{21}$$

which can be written in the matrix-vector notation, $H_{n_1}\epsilon = \tau$, where

$$H_{n_1} = \left(\frac{\mu}{\Delta x} + \frac{\sigma}{2}\right)\left(I_{n_1} - \alpha L_{n_1}\right) \;,$$

and

$$\alpha = \frac{\frac{\mu}{\Delta x} - \frac{\sigma}{2}}{\frac{\mu}{\Delta x} + \frac{\sigma}{2}} \;.$$

The matrix inverse $H_{n_1}^{-1}$ is

$$\begin{aligned} H_{n_1}^{-1} &= \frac{1}{\frac{\mu}{\Delta x} + \frac{\sigma}{2}}\left(I_{n_1} + (\alpha L_{n_1}) + (\alpha L_{n_1})^2 + \ldots + (\alpha L_{n_1})^{(n_1-1)} \right) \;, \\ &= \frac{1}{\frac{\mu}{\Delta x} + \frac{\sigma}{2}} \begin{pmatrix} 1 & & & & & \\ \alpha & 1 & & & & \\ \alpha^2 & \alpha & 1 & & & \\ \vdots & \vdots & \vdots & \ddots & & \\ \alpha^{n_1-2} & \alpha^{n_1-3} & . & \ldots & 1 & \\ \alpha^{n_1-1} & \alpha^{n_1-2} & \alpha^{n_1-3} & \ldots & \alpha & 1 \end{pmatrix} . \end{aligned}$$

Note that $|\alpha| < 1$ for $\sigma > 0$. However, α can be negative (when $\frac{\sigma}{2} > \frac{\mu}{\Delta x}$). For these cases $H_{n_1}^{-1}$ is not a positive matrix. It is oscillatory, i.e. the adjacent diagonals of $H_{n_1}^{-1}$ are of opposite signs. An easy calculation gives

$$\|H_{n_1}^{-1}\|_1 = \frac{1}{\frac{\mu}{\Delta x} + \frac{\sigma}{2}}\left(1 + |\alpha| + |\alpha|^2 + \cdots + |\alpha|^{(n-1)}\right) \;.$$

Since $|\alpha|$ is bounded by 1, then $1 + |\alpha| + |\alpha|^2 + \cdots + |\alpha|^{(n_1-1)} \leq n_1$. Therefore an upper bound for $\|H_{n_1}^{-1}\|_1$ is

$$\|H_{n_1}^{-1}\|_1 \leq \frac{1}{\frac{\mu}{\Delta x} + \frac{\sigma}{2}}\, n_1 \;.$$

which shows that $\|H_{n_1}^{-1}\|_1$ is bounded by a function that decreases with increasing absorption. Furthermore, since

$$\frac{1}{\frac{\mu}{\Delta x}+\frac{\sigma}{2}}\, n_1 \leq \frac{1}{\frac{\mu}{\Delta x}} n_1 = \frac{a_1}{\mu}\ ,$$

then $\|H_{n_1}^{-1}\|_1$ is bounded by a number that is independent of Δx.

From the symmetry of $H_{n_1}^{-1}$, we find that the equality $\|H_{n_1}^{-1}\|_\infty = \|H_{n_1}^{-1}\|_1$, and the inequality $\|H_{n_1}^{-1}\|_2 \leq \|H_{n_1}^{-1}\|_1$ hold for $\sigma \geq 0$. Thus the $1, 2$ and ∞ norms of $H_{n_1}^{-1}$ decrease with increasing σ, and their bounds are independent of Δx.

The Global Error for 2-D Transport

The Petrov-Galerkin equations (10) for the global error in 2-D can be written more simply by collecting terms

$$a\,\epsilon_{i,j} - b\,\epsilon_{i-1,j} - c\,\epsilon_{i,j-1} - d\,\epsilon_{i-1,j-1} = \tau_{i,j}\ , \qquad i = 1,\ldots,n_1,\ j = 1,\ldots,n_2, \tag{22}$$

where

$$\begin{aligned} a &= \frac{\mu}{2\Delta x} + \frac{\eta}{2\Delta y} + \frac{\sigma}{4} \\ b &= \frac{\mu}{2\Delta x} - \frac{\eta}{2\Delta y} - \frac{\sigma}{4} \\ c &= -\frac{\mu}{2\Delta x} + \frac{\eta}{2\Delta y} - \frac{\sigma}{4} \\ d &= \frac{\mu}{2\Delta x} + \frac{\eta}{2\Delta y} - \frac{\sigma}{4}\ . \end{aligned} \tag{23}$$

The elimination of the boundary conditions by the substitution of (6) into (22) yields the reduced set of equations

$$a\,\epsilon_{i,j} - b\,\epsilon_{i-1,j} - c\,\epsilon_{i,j-1} - d\,\epsilon_{i-1,j-1} = \tau_{i,j}\ , \qquad i = 2,\ldots,n_1,\ j = 2,\ldots,n_2, \tag{24}$$

and

$$a\,\epsilon_{1,1} = \tau_{1,1}\ , \tag{25}$$

$$a\,\epsilon_{i,1} - b\,\epsilon_{i-1,1} = \tau_{i,1}\ , \qquad i = 2,\ldots,n_1,$$

$$a\,\epsilon_{1,j} - c\,\epsilon_{1,j-1} = \tau_{1,j}\ , \qquad j = 2,\ldots,n_2.$$

Up to this point, there is no preferred ordering of the i, j indices in the system of equations (24) and (25). We choose the ordering which simplifies our presentation, and that choice depends on the relative sizes of $\mu/\Delta x$ and $\eta/\Delta y$. Without lost of

generality, let us assume that $\mu/dx \geq \eta/dy$. Under this condition, j is chosen to be the "outer" indices of the tensor product matrices and the tensor product vectors when casting (24) and (25) in matrix-vector notation. Let the $n_1 n_2$ dimensional vector ϵ be a n_2-dimensional compound vector whose components are n_1 dimensional vectors, i.e.

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_j \\ \vdots \\ \epsilon_{n_2} \end{pmatrix} \in R^{n_1 n_2}, \quad \text{with} \quad \epsilon_{\mathrm{j}} = \begin{pmatrix} \epsilon_{1,j} \\ \epsilon_{2,j} \\ \vdots \\ \epsilon_{i,j} \\ \vdots \\ \epsilon_{n_1,j} \end{pmatrix} \in \mathrm{R}^{\mathrm{n_1}}.$$

We also introduce the $n_1 n_2$-dimensional compound vector τ of truncation errors

$$\tau = \begin{pmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_j \\ \vdots \\ \tau_{n_2} \end{pmatrix} \in R^{n_1 n_2}, \quad \text{with} \quad \tau_{\mathrm{j}} = \begin{pmatrix} \tau_{1,j} \\ \tau_{2,j} \\ . \\ . \\ . \\ \tau_{n_1,j} \end{pmatrix} \in \mathrm{R}^{\mathrm{n_1}} \ .$$

By introducing the $n_1 th$ order identity and nilpotent matrices I_{n_1} and L_{n_1}, and the $n_2 th$ order identity and nilpotent matrices I_{n_2} and L_{n_2}, we can write (24) and (25) as the matrix-vector equation

$$H\epsilon = \tau \ ,$$

where H is the $(n_1 n_2) th$ order tensor product matrix,

$$H = H_d \otimes I_{n_2} - H_l \otimes L_{n_2} \ , \tag{26}$$

in which

$$H_d = a\, I_{n_1} - b\, L_{n_1} \ , \quad \text{and} \quad \mathrm{H_l} = \mathrm{c\ I_{n_1}} + \mathrm{d\ L_{n_1}} \ . \tag{27}$$

The matrix inverse of (26) can be written as

$$\begin{aligned} H^{-1} &= \left(I_{n_1} \otimes I_{n_2} - H_d^{-1} H_l \otimes L_{n_2}\right)^{-1} \left(H_d^{-1} \otimes I_{n_2}\right) \\ &= \left(I_{n_1} \otimes I_{n_2} + H_d^{-1} H_l \otimes L_{n_2} + \ldots + (H_d^{-1} H_l)^{(n_2-1)} \otimes L_{n_2}^{(n_2-1)}\right) \left(H_d^{-1} \otimes I_{n_2}\right). \end{aligned}$$

A bound for the 2-norm of H^{-1},

$$\|H^{-1}\|_2 \leq \left(1 + \|H_d^{-1} H_l\|_2 + \|H_d^{-1} H_l\|_2^2 + \ldots + \|H_d^{-1} H_l\|_2^{(n_2-1)} \right) \ \|H_d^{-1}\|_2 \ , \tag{28}$$

can be determined from the bounds on $\|H_d^{-1}H_l\|_2$ and $\|H^{-1}\|_2$.

We found from our work in 1-D that $\|H^{-1}\|_2$ is majorized by the vacuum, i.e. $\sigma = 0$. This is also true in 2-D, because both $\|H_d^{-1}H_l\|_2$ and $\|H_d^{-1}\|_2$ decrease with increasing σ. So from hereon, we will focus on the case that $\sigma = 0$. Let's consider the estimation of $\|H_d^{-1}\|_2$ first. Since the matrix H_d^{-1} has same form as the matrices which were investigated in §, we use those techniques developed there to bound $\|H_d^{-1}\|_2$. It follows from § that

$$\|H_d^{-1}\|_2 \leq \|H_d^{-1}\|_1 = \frac{1 - \left|\frac{b}{a}\right|^{n_1}}{|a| - |b|} < \frac{1}{|a| - |b|} = \frac{\Delta y}{\eta} \,, \tag{29}$$

where the assumption $\mu/\Delta x \geq \eta/\Delta y$ is used to derive the equality $|a| - |b| = \eta/\Delta y$.

Turning to the calculation of $\|H_d^{-1}H_l\|_2$. When $\sigma = 0$, the coefficients a, b, c and d simplify. We find from (23) that $c = -b$ and $d = a$ for $\sigma = 0$. If we substitute these simplifying relations into (27), and if we let $\beta \equiv b/a$ (note that $0 \leq \beta < 1$ because $\mu/\Delta x \geq \eta/\Delta y$), then we can write H_d and H_l respectively as

$$H_d = a\,(I_{n_1} - \beta\, L_{n_1}) \,, \qquad \text{and} \qquad \mathrm{H_l = a\,(-\beta\, I_{n_1} + L_{n_1})} \,.$$

Multiplying the series representations of H_d^{-1} and of H_l, and using the fact that L_{n_1} is nilpotent ($L_{n_1}^{n_1} = 0$), we find that

$$\begin{aligned} H_d^{-1}H_l &= \frac{1}{a}\left(I_{n_1} + \beta L_{n_1} + \beta^2 L_{n_1}^2 + \ldots + \beta^{(n_1-1)} L_{n_1}^{(n_1-1)}\right)\, a\,(-\beta I_{n_1} + L_{n_1}) \\ &= -\beta I_{n_1} + \left(1 - \beta^2\right) L_{n_1}\left(I_{n_1} + \beta L_{n_1} + \beta^2 L_{n_1}^2 + \ldots + \beta^{(n_1-2)} L_{n_1}^{(n_1-2)}\right) \\ &= -\beta \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{pmatrix} + (1-\beta^2) \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ \beta & 1 & 0 & & \\ \vdots & \vdots & \ddots & \ddots & \\ \beta^{(n_1-2)} & \beta^{(n_1-3)} & \ldots & 1 & 0 \end{pmatrix} \end{aligned} \tag{30}$$

is the upper left hand section of a Toeplitz matrix. We use Toeplitz's theorem to prove that $\|\cdot\|_2$ of this $n_1 \times n_1$ matrix,

$$\|H_d^{-1}H_l\|_2 \leq 1 \,, \tag{31}$$

for any order n_1. However it is instructive to calculate $\|H_d^{-1}H_l\|_2$ for the cases $n_1 = 1$ and 2 by alternative methods. For the case $n_1 = 1$, we can deduce by inspection that $\|H_d^{-1}H_l\|_2 = |\beta| < 1$.

For the $n_1 = 2$ case, we can derive explicitly $\tilde{\sigma}_+$ and $\tilde{\sigma}_-$, the singular values of $H_d^{-1}H_l$, by setting the determinant of the 2×2 matrix,

$$\left(H_d^{-1}H_l\right)^T \left(H_d^{-1}H_l\right) - \tilde{\sigma} I_2 = \begin{pmatrix} -\beta & 1-\beta^2 \\ 0 & -\beta \end{pmatrix} \begin{pmatrix} -\beta & 0 \\ 1-\beta^2 & -\beta \end{pmatrix} - \tilde{\sigma} \begin{pmatrix} 1 & \\ & 1 \end{pmatrix} \,,$$

equal to zero, which gives

$$\tilde{\sigma}^2 - (1+\beta^4)\tilde{\sigma} + \beta^4 = 0 \; .$$

Solving this quadratic equation, we have

$$\tilde{\sigma}_{\pm} = \left\{ \begin{array}{l} 1 \\ \beta^4 \; . \end{array} \right.$$

Since $\tilde{\sigma}_+$ is the larger of $\tilde{\sigma}_+$ and $\tilde{\sigma}_-$, then $\|H_d^{-1}H_l\|_2 = \tilde{\sigma}_+^{1/2} = 1$, which completes our proof for the $n_1 = 2$ case.

We apply the theory of Toeplitz matrices [1] to prove that $\|H_d^{-1}H_l\|_2 = 1$ for the $n_1 = \infty$ case. A Toeplitz matrix, $T(g)$, is an infinite dimensional matrix, which is constant along the diagonals

$$T(g) = \left(\begin{array}{cccc} g_0 & g_{-1} & g_{-2} & \cdots \\ g_1 & g_0 & g_{-1} & \cdots \\ g_2 & g_1 & g_0 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{array} \right) \; ,$$

with matrix elements $\{g_n\}_{n=-\infty}^{\infty}$ that are the Fourier coefficients,

$$g_n = \frac{1}{2\pi} \int_0^{2\pi} e^{-i\,n\theta}\, g(e^{i\theta})\, d\theta \; ,$$

of the (complex) symbol $g(e^{i\theta})$. The theorem of Toeplitz [1] states that the 2-norm of the infinite matrix $T(g)$ is equal to the maximum modulus of $g(e^{i\theta})$ in the interval $0 \le \theta < 2\pi$, i.e.

$$\|T(g)\|_2 \; = \; \max_{0 \le \theta < 2\pi} |g(e^{i\theta})| \; . \tag{32}$$

In order to apply (32) to the calculation of $\|H_d^{-1}H_l\|_2$ for the infinite dimensional case, we need the symbol of $H_d^{-1}H_l$. Given the coefficients $\{g_n\}_{n=-\infty}^{\infty}$ of a Toeplitz matrix, the symbol g that generates them is, by Fourier's composition theorem,

$$g(z) = \sum_{n=-\infty}^{\infty} g_n z^n, \quad z = e^{i\,\theta} \; . \tag{33}$$

When $g(e^{\,i\theta})$ is expressed as a Fourier series, we can use elementary calculus to determine the maximum of $|g(e^{\,i\theta})|$ in the interval $0 \le \theta < 2\pi$.

Referring to (30), the coefficients of $H_d^{-1}H_l$ are;

$$g_n = 0, \;\; \forall\, n < 0 \; , \qquad g_0 = -\beta \; , \qquad g_n = (1-\beta^2)\beta^{(n-1)}, \;\; \forall\, n > 0 \; .$$

Substituting these coefficients into (33), we have

$$\begin{aligned} g(z) \;\; &= \;\; -\beta + (1-\beta^2)\, z \left(1 + \beta z + \beta^2 z^2 + \beta^3 z^3 + \ldots\ldots \right) \\ &= \;\; -\beta + (1-\beta^2)\, z \, \frac{1}{1-\beta z} \; = \frac{z-\beta}{1-\beta z} \; . \end{aligned}$$

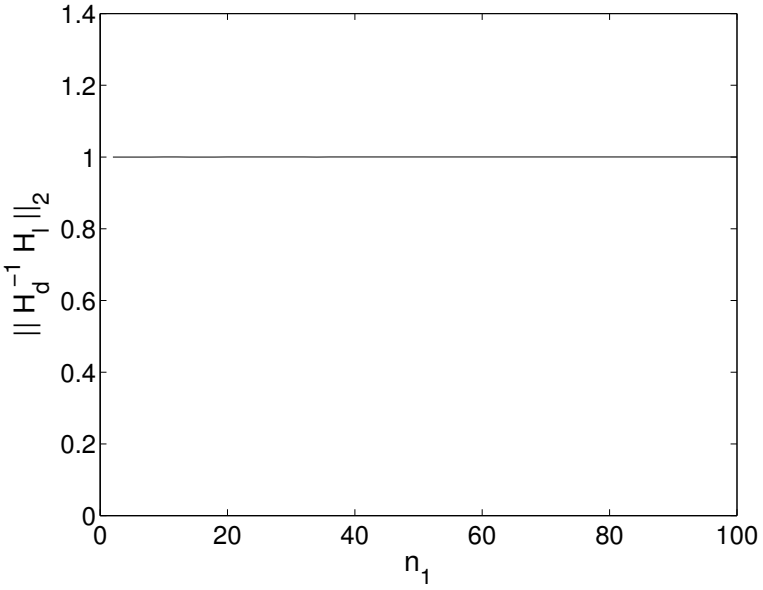

Figure 1: A plot of $\|H_d^{-1}H_l\|_2$ as a function of the matrix order n_1.

Using the fact that $z^*z = e^{-i\,\theta}e^{\,i\theta} = 1$, the squared modulus of this symbol is

$$g^*(z)g(z) = \frac{z^* - \beta}{1 - \beta z^*}\,\frac{z - \beta}{1 - \beta z} = \frac{z^*z - \beta z^* - \beta z + \beta^2}{1 - \beta z^* - \beta z + \beta^2 z^* z} = \frac{1 - \beta z^* - \beta z + \beta^2}{1 - \beta z^* - \beta z + \beta^2} = 1\ .$$

Then by Toeplitz theorem, (32), we have

$$\|H_d^{-1}H_l\|_2 = \max_{0\le\theta<2\pi}\left|\frac{e^{\,i\theta} - \beta}{1 - \beta e^{\,i\theta}}\right| = 1\ ,$$

which completes our derivation for the infinite dimensional case.

Since it was shown in [1] that the 2-norm of T_n, the $n \times n$ upper left hand section of the infinite dimensional matrix T, is majorized by $\|T\|_2$, i.e. $\|T_n\|_2 \le = \max_{0\le\theta<2\pi}|g(e^{i\theta})|$, then (31) holds by the above equality. We found from numerical experiments that $\|H_d^{-1}H_l\|_2$ is precisely 1 for orders of $H_d^{-1}H_l$ between 2 and 100. In Fig. 1, we plot the results of these calculations of $\|H_d^{-1}H_l\|_2$ for the parameters $\mu = .9$, $a_2 = a_1 = 1$ and $n_2 = n_1$. Since our numerical experiments show that $\|H_d^{-1}H_l\|_2 = 1$ for any order between 2 and 100, then we conjecture that this equality can be derived theoretically.

From (28), (29) and (31), we have the upper bound

$$\|H^{-1}\|_2 \le\ n_2\,\frac{\Delta y}{\eta} = \frac{a_2}{\eta}\ . \tag{34}$$

This upper bound for $\|H^{-1}\|_2$ is asymmetric with respect to the mesh parameters a_1 and a_2, yet H is symmetric with respect to Δx and Δy. This asymmetry was introduced into (28) by the triangle inequality which was used to derive it. Let's

examine how this symmetry breaking affects our estimate of $\|H^{-1}\|_2$. Suppose $\mu/\Delta x$ is not greater than $\eta/\Delta y$, as assumed, and j is chosen unwittingly as the "outer" index. Under these conditions, we have $\|H_d^{-1}\|_2 \leq \Delta x/\mu$, rather than the inequality in (29), which leads to the clumsy result

$$\|H^{-1}\|_2 \leq \; n_2 \, \frac{\Delta x}{\mu} = \frac{n_2}{n_1} \frac{a_1}{\mu} \, . \tag{35}$$

How does this estimate compared to the one in (34)? For $\mu/\Delta x \leq \eta/\Delta y$, we have $n_1\mu/a_1 \leq n_2\eta/a_2$. The substitution of this inequality into the rhs of (35) leads to

$$\frac{n_2}{n_1} \frac{a_1}{\mu} \; = \frac{n_2}{n_1} \frac{a_1}{\mu} \frac{\eta a_2}{\eta a_2} \; = \frac{n_2 \eta}{a_2} \frac{a_1}{n_1 \mu} \frac{a_2}{\eta} \; \geq \frac{a_2}{\eta} \, ,$$

which shows that the bound in (34) is smaller than the bound in (35).

Note that, when $\mu/\Delta x \leq \eta/\Delta y$, we can cleverly derive the upper bound

$$\| \, H^{-1} \|_2 \; \leq \frac{a_1}{\mu} \, , \tag{36}$$

a result which is tidier than (35), by switching the indices of the tensor product representation of H (so that i is the "outer" index). However, if i is chosen as the "outer" index of H, and if the condition $\mu/\Delta x \leq \eta/\Delta y$ is not met, then we get the awkward result

$$\|H^{-1}\|_2 \leq \; n_1 \, \frac{\Delta y}{\eta} = \frac{n_1}{n_2} \frac{a_2}{\eta} \, .$$

An argument, which is similar to the one used to show that the bound in (34) is smaller than the bound in (35), can be applied to prove the bound in (36) is smaller than the bound in the above inequality.

Since both (34) and (36) are upper bounds for $\|H^{-1}\|_2$, then we take the smaller of the two

$$\| \, H^{-1} \|_2 \; \leq \min \left(\frac{a_1}{\mu} \, , \frac{a_2}{\eta} \right) . \tag{37}$$

Note that $\min \left(\frac{a_1}{\mu} \, , \frac{a_2}{\eta} \right)$ is the length of the longest line which crosses the domain $[0, a_1] \times [0, a_2]$ in the direction (μ, η). Furthermore, $\min \left(\frac{a_1}{\mu} \, , \frac{a_2}{\eta} \right)$ is the 2-D generalization of a_1/μ, the length of the path of a neutron through a slab in the direction μ in 1-D.

Numerical Tests

We have done a series of numerical experiments to verify these conclusions. We present the results of two test problems to illustrate the main conclusion of this paper. The convergence rate of the *Petrov-Galerkin method is independent of the*

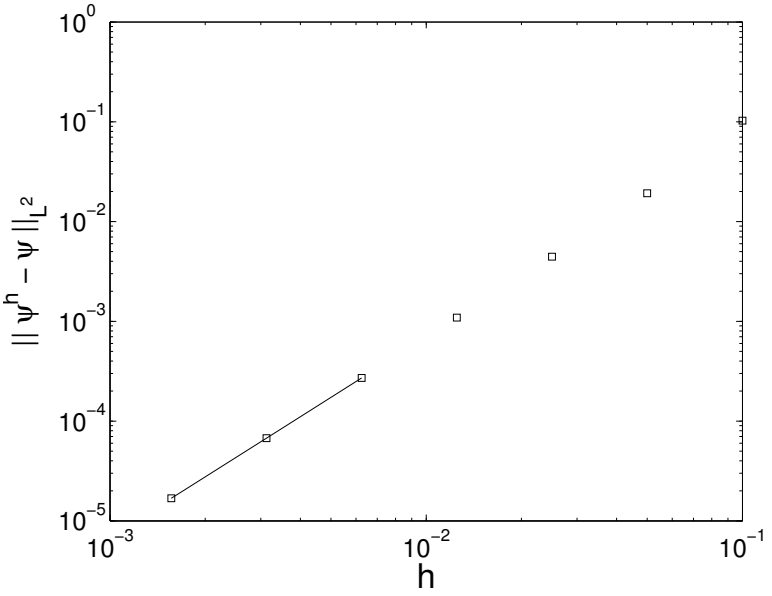

Figure 2: This plot shows that the global error $\|\psi^h - \psi\|_2$ is of order $h^{2.002}$. The slope of the line that passes through the three leftmost points is 2.002.

smoothness of the cross section but depends on the smoothness of the external source. The convergence rate is second order for the solutions of problems with continuous sources, but falls to first order for the solutions of problems with discontinuous sources.

We demonstrate this point with two test problems. The first problem consist of a continuous source and a discontinuous cross section. The second consists of a discontinuous source and a continuous cross section. The domain of both test problems is the unit square, $\{\mathcal{D} : 0 \leq x \leq 1\ ,\ 0 \leq y \leq 1\}$, and is divided into two regions by a concentric inner square of length $\frac{1}{5}$. We model the discontinuous cross section of the first problem by assigning the cross section unequal constants in the two regions. The source of the second problem is modeled by a similar discontinuous function. In both problems we use $(\mu, \eta) = (\frac{\sqrt{3}}{2}, \frac{1}{2})$.

The cross section σ of the first problem is the discontinuous function

$$\sigma(x,y) = \begin{cases} 0, & \frac{2}{5} \leq x \leq \frac{3}{5}\ ,\ \frac{2}{5} \leq y \leq \frac{3}{5} \\ 100, & \text{otherwise}\ . \end{cases} \tag{38}$$

In the first problem, we assume the exact solution to be the smooth function

$$\psi(x,y) = \sin^2(2\pi x)\sin^2(2\pi y)\ , \tag{39}$$

and we determine the source by substituting (39) into the transport equation (1) to get

$$q(x,y) \quad = \quad 4\ \mu\ \pi\ \sin(2\pi x)\cos(2\pi x)\sin^2(2\pi y)$$

$$
\begin{aligned}
&+ \quad 4\,\eta\,\pi\,\sin^2(2\pi x)\sin(2\pi y)\cos(2\pi y) \\
&+ \quad \sigma(x,y)\sin^2(2\pi x)\sin^2(2\pi y)\ . \qquad (40)
\end{aligned}
$$

We substitute the cross section of (38) and the source of (40) into the Petrov-Galerkin system of equations (4), and then solve it. In Fig. 2, we plot the log of the global error as a function of the log of the grid spacing. In this plot, an error that is proportional to h^2 should have a slope of 2. We have also drawn the straight line through the three leftmost points in this figure which was determined least squares fitting these three points to a straight line. From the least squares fit, we find a slope of 2.002 for the straight line. The reader may question why we consider the source of (40) to be continuous when its third term is comprised of the discontinuous function $\sigma(x,y)$. The reason is because this term appears also as the absorption term $\sigma(x,y)\psi(x,y)$ on the lhs of the transport equation (1). As a result, the two discontinuous terms in the transport equation cancel.

In the second problem, the source is the discontinuous function,

$$
q(x,y) = \begin{cases} 1, & \frac{2}{5} \le x \le \frac{3}{5}\ ,\ \frac{2}{5} \le y \le \frac{3}{5} \\ 0, & \text{otherwise}\ , \end{cases} \qquad (41)
$$

and the cross section is the constant $\sigma(x,y) = 1$. In this problem, the exact solution is determined by the method of characteristics. We also solve this transport problem by the Petrov-Galerkin method. In Fig 3, we plot the root mean square difference between these solutions as a function of the grid spacing. Here we find a slope of 1.0403. We expect the slope to approach 1 as we refine the grid further. This test problem shows that the Petrov-Galerkin solution converges to the true solution at a first order rate when the source is discontinuous.

Conclusions

We proved in § and § that $\|H^{-1}\|_2$ is bounded by a number that is independent of the order of the matrix. This number is the length of the longest line that crosses the domain, $[0,a_1]\times[0,a_2]$, in the (μ,η) direction. The lengths of these lines in 1-D and 2-D are respectively a_1/μ and $\min(a_1/\mu\,,\,a_2/\eta)$. These bounds are universal, and are applicable to situations in which absorption is non-zero. The 2-norm of the H^{-1} in an absorbing medium is, in fact, even smaller than these bounds.

Since $\|H^{-1}\|_2$ is independent of h, and the global error is related to the truncation error by $\|\epsilon\|_2 \le \|H^{-1}\|_2\,\|\tau\|_2$, then $\|\epsilon\|_2$ is of the same order as $\|\tau\|_2$. We can draw two conclusions from this result. First, the Petrov-Galerkin solution converges to the true solution in the sense that the root mean square difference between the two goes to zero as h^2, if the true solution has continuous second derivatives. Second, the truncation error depends on the smoothness of the true solution which in turn depends on the smoothness of the source term on the rhs of the transport equation. First order truncation errors arise when a second derivative

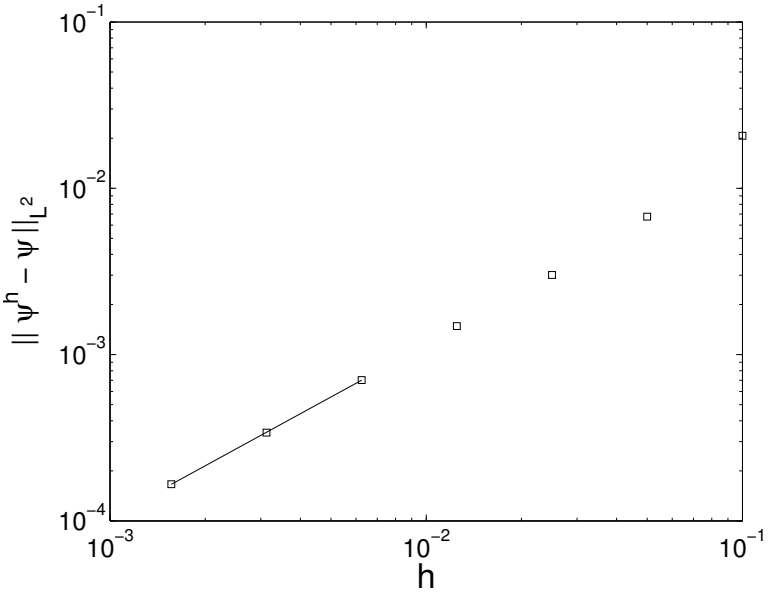

Figure 3: This plot shows that the global error $\|\psi^h - \psi\|_2$ is of order $h^{1.0403}$. The slope of the line that passes through the three leftmost points is 1.0403.

of the true solution is discontinuous. This situations occurs in problems with discontinuous sources. These conclusions were borne out by our test problems in §.

Appendix: Local Truncation Error

This appendix is divided into two parts. In the first part, we show the system of Petrov-Galerkin equations is, within the local truncation error, equivalent to a system of finite difference equations. In the second part, we derive the local truncation error by expanding the true solution in a Taylor series.

For the convenience of the reader, we rewrite the definition of the local truncation error, (8), below

$$\begin{aligned} &\frac{\mu}{\Delta x}\left(\frac{\psi(x_i,y_j)+\psi(x_i,y_{j-1})}{2}-\frac{\psi(x_{i-1},y_j)+\psi(x_{i-1},y_{j-1})}{2}\right)\\ +&\frac{\eta}{\Delta y}\left(\frac{\psi(x_i,y_j)+\psi(x_{i-1},y_j)}{2}-\frac{\psi(x_i,y_{j-1})+\psi(x_{i-1},y_{j-1})}{2}\right)\\ &+\sigma\left(\frac{\psi(x_i,y_j)+\psi(x_i,y_{j-1})}{4}+\frac{\psi(x_{i-1},y_j)+\psi(x_{i-1},y_{j-1})}{4}\right) \equiv q_{i,j}-\tau_{i,j}\ , \end{aligned} \tag{42}$$

where $q_{i,j}$, defined in (5), is also rewritten here

$$q_{i,j}=\frac{1}{\Delta x\Delta y}\int_{y_{j-1}}^{y_j}dy\int_{x_{i-1}}^{x_i}dx\ q(x,y)\ . \tag{43}$$

Now if we approximate the above integral by the mid-point rule, i.e. the integrand is expanded in a Taylor series at the zone mid-point

$x_{i-\frac{1}{2}} = \frac{1}{2}(x_{i-1} + x_i)$, $y_{j-\frac{1}{2}} = \frac{1}{2}(y_{j-1} + y_j)$ and the result integrated,

$$q_{i,j} \approx q(x_{i-\frac{1}{2}}, y_{j-\frac{1}{2}}) + \frac{(\Delta x)^2}{24} \left.\frac{\partial^2 q}{\partial x^2}\right|_{x_{i-\frac{1}{2}}, y_{j-\frac{1}{2}}} + \frac{(\Delta y)^2}{24} \left.\frac{\partial^2 q}{\partial y^2}\right|_{x_{i-\frac{1}{2}}, y_{j-\frac{1}{2}}} , \tag{44}$$

we find the quadrature error is of order h^2. As we shall see, this quadrature error can be subsumed in the definition of the local truncation error $\tau_{i,j}$.

When the Taylor series expansion of ψ at each node in (42) is substituted into (42) and (44) is also substituted into (42), we find

$$\begin{aligned} \tau_{i,j} &= \mu \left(\frac{(\Delta x)^2}{24} \psi_{x,x,x} + \frac{(\Delta y)^2}{8} \psi_{x,y,y} \right) + \eta \left(\frac{(\Delta y)^2}{24} \psi_{y,y,y} + \frac{(\Delta x)^2}{8} \psi_{x,x,y} \right) \\ &+ \sigma \left(\frac{(\Delta x)^2}{8} \psi_{x,x} + \frac{(\Delta y)^2}{8} \psi_{y,y} \right) - \frac{(\Delta x)^2}{24} q_{x,x} - \frac{(\Delta y)^2}{24} q_{y,y} , \end{aligned} \tag{45}$$

where all partial derivatives are evaluated at the zone mid-point.

Acknowledgments

Britton Chang would like to thank Prof. Tom Manteufel for his many insights that led to the results in this paper.

References

[1] A. BOTTCHER and B. SILBERMANN, *Introduction to Large Truncated Toeplitz Matrices*, Springer-Verlag, New York, 1998.

[2] B. G. CARLSON and K. D. LATHROP, *Transport Theory The Method of Discrete Ordinates*, in Computing Methods in Reactor Physics, ed. H. Greenspan, C. N. Kelber, and D Okrent, Gordon and Breach, New York, 1968, pp. 167-266.

[3] G. H. GOLUB and C. F. Van LOAN, *Matrix Computations* John Hopkins, Baltimore, 1991.

[4] A. GREENBAUM AND J. M. FERGUSON, *A Petrov-Galerkin Finite Element for Solving the Neutron Transport Equation*, J. Comput. Phys., **64** (1986), pp. 97-111.

[5] E. ISAACSON AND H. B. KELLER, *Analysis of Numerical Methods*, John Wiley, New York, 1966.

[6] E. W. LARSEN AND W. F. MILLER, *Convergence Rates of Spatial Difference Equationf for the Discrete-Ordinates Neutron Transport Equations in Slab Geometry*, Nucl. Sci. Eng., 73,(1980), pp. 76-83.

[7] E. W. LARSEN AND P. NELSON, *Finite DIfference Approximations and Superconvergence for the Discrete-Ordinate Equations in Slab Geometry*, Siam J. Numer. Anal., 19, 2, (1982), 334-348.

[8] S. M. LEE AND R. VAIDYANATHAN, *Comparision of the Order of Approximation in Several Difference Schemes for the Discrete-Ordinates Transport Equation in One-Dimensional Plane Geometry*, Nucl. Sci. Eng., 76, (1980), pp. 1-9.

[9] E. E. LEWIS AND W. F. MILLER, *Computational Methods of Neutron Transport*, John Wiley, New York, 1984.