Multi-dimensional asymptotically stable 4^{th}-order accurate schemes for the diffusion equation.

Saul Abarbanel*
Adi Ditkowski*
School of Mathematical Sciences
Department of Applied Mathematics
Tel-Aviv University
Tel-Aviv, ISRAEL

Abstract

An algorithm is presented which solves the multi-dimensional diffusion equation on complex shapes to 4^{th}-order accuracy and is asymptotically stable in time. This bounded-error result is achieved by constructing, on a rectangular grid, a differentiation matrix whose symmetric part is negative definite. The differentiation matrix accounts for the Dirichlet boundary condition by imposing penalty like terms.

Numerical examples in 2-D show that the method is effective even where standard schemes, stable by traditional definitions, fail.

*This research was supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-19480 while the authors were in the residence of the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA 23681-0001. S. Abarbanel was also supported in part by the Air Force Office of Scientific research Grant No. AFOSR-F49620-95-1-0074, and by the Department of Energy under grant DOE-DE-FG02-95ER25239.

1 Introduction

Recently there has been renewed interest in finite-difference algorithms of high order of accuracy (4^{th} and above), both for hyperbolic and parabolic p.d.e's (see for example, [1], [2], [3]). The advantages of high-order accuracy schemes, especially for truly time dependent problems, are often offset by the difficulty of imposing stable boundary conditions. Even when the scheme is shown to be G.K.S.-stable the error may increase exponentially in time.

This paper is concerned with 4^{th}-order approximations to the long time solutions of the diffusion equation in one and two dimensions, on irregular domains. By an irregular domain, we mean a body whose boundary points do not coincide with nodes of a rectangular mesh.

In section 2 we develop the theory for the one-dimensional semi-discrete system resulting from the spatial differentiation used in the finite difference algorithm. Energy methods are used in conjunction with "SAT" type terms (see [1]), in order to find boundary conditions that preserve the accuracy of the scheme while constraining an energy norm of the error to be temporally bounded for all $t > 0$ by a constant proportional to the truncation error.

In section 3 it is shown how the methodology developed in section 2 is used as a building block for the multi-dimensional algorithm, even for irregular shapes containing "holes."

Section 4 presents numerical results in two space dimensions illustrating the long-time temporal stability of the method, in contradistinction to "standard" methods for cartesian grid on irregular shapes.

2 The One Dimensional Case

We consider the following problem

$$\frac{\partial u}{\partial t} = k\frac{\partial^2 u}{\partial x^2} + f(x,t); \qquad \Gamma_L \leq x \leq \Gamma_R, \ \ t \geq 0, \ \ k > 0 \tag{2.1a}$$

$$u(x,0) = u_0(x) \tag{2.1b}$$

$$u(\Gamma_L, t) = g_L(t) \tag{2.1c}$$

$$u(\Gamma_R, t) = g_R(t) \tag{2.1d}$$

and $f(x,t)\epsilon\, C^4$.

Let us spatially discritize (2.1a) on the following uniform grid:

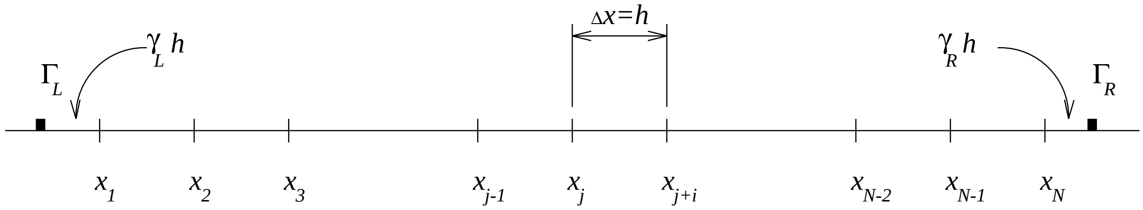

Figure 1: One dimensional grid.

Note that the boundary points do not necessarily coincide with x_1 and x_N. Set $x_{j+1} - x_j = h$, $1 \leq j \leq N-1$; $x_1 - \Gamma_L = \gamma_L h$, $0 \leq \gamma_L < 1$; $\Gamma_R - x_N = \gamma_R h$, $0 \leq \gamma_R < 1$.

The projection unto the above grid of the exact solution $u(x,t)$ to (2.1), is $u_j(t) = u(x_j, t) \triangleq \mathbf{u}(t)$. Let $\tilde{D}$ be a matrix representing the second partial derivative with respect to x, at internal points without specifying yet how it is being built. Then we may write

$$\frac{d}{dt}\mathbf{u}(t) = k[\tilde{D}\mathbf{u}(t) + \mathbf{B} + \mathbf{T}] + \mathbf{f}(t) \tag{2.2}$$

where $\mathbf{T}$ is the truncation error due to the numerical differentiation and $\mathbf{f}(t) = f(x_j, t)$, $1 \leq j \leq N$. The boundary vector $\mathbf{B}$ has entries whose values depend on g_L, g_R, γ_L, γ_R in such a way that $\tilde{D}\mathbf{u} + \mathbf{B}$ represents the 2^{nd} derivative everywhere to the desired accuracy. The standard way of finding a numerical approximate solution to (2.1) is to omit $\mathbf{T}$ from (2.2) and solve

$$\frac{d}{dt}\mathbf{v}(t) = k(\tilde{D}\mathbf{v}(t) + \mathbf{B}) + \mathbf{f}(t) \tag{2.3}$$

where $\mathbf{v}(t)$ is the numerical approximation to the projection $\mathbf{u}(t)$. An equation for the solution error vector, $\vec{\epsilon}(t) = \mathbf{u}(t) - \mathbf{v}(t)$, can be found by subtracting (2.3) from (2.2):

$$\frac{d}{dt}\vec{\epsilon} = k\tilde{D}\vec{\epsilon}(t) + k\mathbf{T}(t) \tag{2.4}$$

Our requirement for *temporal stability is that* $\| \vec{\epsilon} \|$*, the* L_2 *norm of* $\vec{\epsilon}$*, be bounded by a "constant" proportional to* h^m (m being the spatial order of accuracy) *for all* $t < \infty$. Note that this definition is more severe than either the G.K.S. stability criterion [4] or the definition in [1].

It can be shown that if $\tilde{D}$ is constructed in a standard manner, i.e., the numerical second derivative is symmetric away from the boundaries, and near the boundaries one uses non symmetric differentiation, then there are ranges of values of γ_R and γ_L for which $\tilde{D}$ is not negative definite. Since in the multi-dimensional case one may encounter all values of $0 \leq \gamma_L, \gamma_R < 1$, this is unacceptable.

The rest of this section is devoted to the construction of a scheme of 4^{th} order spatial accuracy, which is temporally stable for all γ_L, γ_R.

The basic idea is to use a penalty-like term as in the SAT procedure of ref [1]; here, however, it will be modified and applied in a different manner.

Note first that the solution projection $u_j(t)$ satisfies, besides (2.2), the following differential equation:

$$\frac{d\mathbf{u}}{dt} = kD\mathbf{u} + k\mathbf{T}_e + \mathbf{f}(t) \tag{2.5}$$

where now D is indeed a differentiation matrix, that does not use the boundary values, and therefore $\mathbf{T}_e \neq \mathbf{T}$ but it too is a truncation error due to differentiation.

Next let the semi-discrete problem for $\mathbf{v}(t)$ be, instead of (2.3),

$$\frac{d\mathbf{v}}{dt} = k[D\mathbf{v} - \tau_L(A_L\mathbf{v} - \mathbf{g}_L) - \tau_R(A_R\mathbf{v} - \mathbf{g}_R)] + \mathbf{f}(t) \tag{2.6}$$

where $\mathbf{g}_L = (1, \ldots, 1)^T g_L(t)$; $\mathbf{g}_R = (1, \ldots 1)^T g_R(t)$, are vectors created from the left and right boundary values as shown. The matrices A_L and A_R are defined by the relations:

$$A_L\mathbf{u} = \mathbf{g}_L - \mathbf{T}_L; \qquad A_R\mathbf{u} = \mathbf{g}_R - \mathbf{T}_R, \tag{2.7}$$

i.e., each row in $A_L(A_R)$ is composed of the coefficients extrapolating $\mathbf{u}$ to its boundary value $\mathbf{g}_L(\mathbf{g}_R)$, at $\Gamma_L(\Gamma_R)$ to within the desired order of accuracy. (The error is then $\mathbf{T}_L(\mathbf{T}_R)$.) The diagonal matrices τ_L and τ_R are given by

$$\tau_L = \text{ diag } (\tau_{L_1}, \tau_{L_2} \ldots, \tau_{L_N}); \qquad \tau_R = \text{ diag } (\tau_{R_1}, \ldots \tau_{R_N}) \tag{2.8}$$

Subtracting (2.6) from (2.5) we get

$$\frac{d\vec{\epsilon}}{dt} = k[D\vec{\epsilon} - \tau_L A_L\vec{\epsilon} - \tau_R A_R\vec{\epsilon} + \mathbf{T}_1] \tag{2.9}$$

where

$$\mathbf{T}_1 = \mathbf{T}_e + \tau_L \mathbf{T}_L + \tau_R \mathbf{T}_R$$

Taking the scalar product of $\vec{\epsilon}$ with (2.9) one gets:

$$\begin{aligned} \frac{1}{2}\frac{d}{dt} \parallel \vec{\epsilon} \parallel^2 &= k(\vec{\epsilon}, (D - \tau_L A_L - \tau_R A_R)\vec{\epsilon}) + k(\vec{\epsilon}, \mathbf{T}_1) \\ &= k(\vec{\epsilon}, M\vec{\epsilon}) + k(\vec{\epsilon}, \mathbf{T}_1) \end{aligned} \qquad (2.10)$$

We notice that $(\vec{\epsilon}, M\vec{\epsilon})$ is $(\vec{\epsilon}, (M + M^T)\vec{\epsilon}/2$, where

$$M = D - \tau_L A_L - \tau_R A_R. \qquad (2.11)$$

If $M + M^T$ can be made negative definite then

$$(\vec{\epsilon}, (M + M^T)\vec{\epsilon}/2 \leq -c_0 \parallel \vec{\epsilon} \parallel^2, \qquad (c_0 > 0). \qquad (2.12)$$

Equation (2.10) then becomes

$$\frac{1}{2}\frac{d}{dt} \parallel \vec{\epsilon} \parallel^2 \leq -kc_0 \parallel \epsilon \parallel^2 + k(\vec{\epsilon}, \mathbf{T}_1)$$

and using Schwartz's inequality we get after dividing by $\parallel \vec{\epsilon} \parallel$

$$\frac{d}{dt} \parallel \vec{\epsilon} \parallel \leq -kc_0 \parallel \vec{\epsilon} \parallel + k \parallel \mathbf{T}_1 \parallel$$

and therefore (using the fact that $\mathbf{v}(0) = \mathbf{u}(0)$)

$$\parallel \vec{\epsilon} \parallel \leq \frac{\parallel \mathbf{T_1} \parallel_M}{c_0}(1 - e^{-kc_0 t}) \qquad (2.13)$$

where the "*constant*$''$ $\| \mathbf{T_1} \|_M = \max_{0 \le \tau \le t} \| \mathbf{T_1}(\tau) \|$.

If we indeed succeed in constructing M such that $M + M^T$ is negative definite, with $c_0 > 0$ independent of the size of the matrix M as it increases, then it follows from (2.13) that the norm of the error will be bounded for all t by a constant which is $O(h^m)$ where m is the spatial accuracy of the finite difference scheme (2.6). The numerical solution is then temporally stable.

The rest of this section is devoted to this task for the case of $m = 4$, i.e, a fourth order accurate finite difference algorithm.

Let the $n \times n$ differentiation matrix, D, be given by

$$
\frac{1}{12h^2}
\begin{bmatrix}
45 & -154 & 214 & -156 & 61 & -10 & & & & & & \\
10 & -15 & -4 & 14 & -6 & 1 & & & & & & \\
-1 & 16 & -30 & 16 & -1 & & & & & & & \\
 & -1 & 16 & -30 & 16 & -1 & & & & & & \\
 & & -1 & 16 & -30 & 16 & -1 & & & & & \\
 & & & & & \ddots & & & & & & \\
 & & & & & & \ddots & & & & & \\
 & & & & & & & \ddots & & & & \\
 & & & & & & -1 & 16 & -30 & 16 & -1 & \\
 & & & & & & & -1 & 16 & -30 & 16 & -1 \\
 & & & & & & 1 & -6 & 14 & -4 & -15 & 10 \\
 & & & & & & -10 & 61 & -156 & 214 & -154 & 45
\end{bmatrix}
\tag{2.14}
$$

The upper two rows and the lower two rows represent non-symmetric fourth order accurate approximation to the second derivative without using boundary values. The internal rows are symmetric and represent central differencing approximation to u_{xx} to the same order. Note that D is not negative definite, and neither is the symmetric part of $\frac{1}{2}(D+D^T)$ which is given by:

$$\frac{1}{24h^2}\left[\begin{array}{rrrrrrrrrrrrr}
90 & -144 & 213 & -156 & 61 & -10 & & & & & & & \\
-144 & -30 & 12 & 13 & -6 & 1 & & & & & & & \\
213 & 12 & -60 & 32 & -2 & 0 & & & & & & & \\
-156 & 13 & 32 & -60 & 32 & -2 & & & & & & & \\
61 & -6 & -2 & 32 & -60 & 32 & -2 & & & & & & \\
-10 & 1 & 0 & -2 & 32 & -60 & 32 & -2 & & & & & \\
 & & & & & & & & & & & & \\
 & & & & & & & & & & & & \\
 & & & & & -2 & 32 & -60 & 32 & -2 & 0 & 1 & -10 \\
 & & & & & & -2 & 32 & -60 & 32 & -2 & -6 & 61 \\
 & & & & & & & -2 & 32 & -60 & 32 & 13 & -156 \\
 & & & & & & & 0 & -2 & 32 & -60 & 12 & 213 \\
 & & & & & & & 1 & -6 & 13 & 12 & -30 & -144 \\
 & & & & & & & -10 & 61 & -156 & 213 & -144 & 90
\end{array}\right] \tag{2.15}$$

In order to construct M we need to specify A_L, A_R, τ_L and τ_R. We construct A_L as follows:

$$A_L = A_\alpha^{(L)} + c_L A_e^{(L)} \tag{2.16}$$

where

$$A_\alpha^{(L)} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & 0 & \dots & 0 \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & 0 & \dots & 0 \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & 0 & \dots & 0 \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & 0 & \dots & 0 \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & 0 & \dots & 0 \\ \vdots & & & & & & & \\ \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 & 0 & \dots & 0 \end{bmatrix}, \tag{2.17}$$

$$c_L = \text{ diag } [-20\alpha_1/71, \quad 0, \dots, 0] \tag{2.18}$$

$$A_e^{(L)} = \begin{bmatrix} -1 & 5 & -10 & 10 & -5 & 1 & 0 & \dots & 0 \\ -1 & 5 & -10 & 10 & -5 & 1 & 0 & \dots & 0 \\ \vdots & & & & & & & & \\ -1 & 5 & -10 & 10 & -5 & 1 & 0 & \dots & 0 \end{bmatrix}. \tag{2.19}$$

The α's are given by

$$\begin{aligned} \alpha_1 &= 1 + \frac{25}{12}\gamma_L + \frac{35}{24}\gamma_L^2 + \frac{5}{12}\gamma_L^3 + \frac{1}{24}\gamma_L^4 \\ \alpha_2 &= -\left(4\gamma_L + \frac{13}{3}\gamma_L^2 + \frac{3}{2}\gamma_L^3 + \frac{1}{6}\gamma_L^4\right) \\ \alpha_3 &= 3\gamma_L + \frac{19}{4}\gamma_L^2 + 2\gamma_L^3 + \frac{1}{4}\gamma_L^4 \qquad (2.20) \\ \alpha_4 &= -\left(\frac{4}{3}\gamma_L + \frac{7}{3}\gamma_L^2 + \frac{7}{6}\gamma_L^3 + \frac{1}{6}\gamma_L^4\right) \\ \alpha_5 &= \frac{1}{4}\gamma_L + \frac{11}{24}\gamma_L^2 + \frac{1}{4}\gamma_L^3 + \frac{1}{24}\gamma_L^4 \end{aligned}$$

Note that $A_\alpha^{(L)}\mathbf{v}$ gives a vector whose components are the extrapolated value of $\mathbf{v}$ at $x = \Gamma_L$ (i.e., $v_{\Gamma_L}(t)$), to fifth order accuracy; while $A_e^{(L)}\mathbf{v}$ gives a vector whose components represents $(\partial^5 v_1/\partial x^5)h^5$. Since C_L (see 2.18) is of order unity, then $A_L\mathbf{v} = (A_\alpha^{(L)} + c_L A_e^{(L)})\mathbf{v}$ represents an extrapolation of $\mathbf{v}$ to v_{Γ_L} to fifth order.

Before using A_L in (2.11) or (2.6) we must define τ_L:

$$\tau_L = \frac{1}{12h^2}\text{diag}[\tau_1, \tau_2, \tau_3, \tau_4, \tau_5, 0, \ldots, 0] \tag{2.21}$$

where

$$\begin{aligned} \tau_1 &= 71/2\alpha_1 \\ \tau_2 &= (-94 - \alpha_2\tau_1)/\alpha_1 \\ \tau_3 &= (113 - \alpha_3\tau_1)/\alpha_1 \\ \tau_4 &= (-56 - \alpha_4\tau_1)/\alpha_1 \\ \tau_5 &= (11 - \alpha_5\tau_1)/\alpha_1 \end{aligned} \tag{2.22}$$

The right boundary treatment is constructed in a similar fashion, and the formulae corresponding to (2.16) - (2.22) become:

$$A_R = A_\alpha^{(R)} + c_R A_e^{(R)}, \tag{2.23}$$

$$A_\alpha^{(R)} = \begin{bmatrix} 0\ldots & \ldots & \ldots & 0 & 0 & \alpha_{N-4} & \alpha_{N-3} & \alpha_{N-2} & \alpha_{N-1} & \alpha_N \\ 0\ldots & \ldots & \ldots & 0 & 0 & \alpha_{N-4} & \alpha_{N-3} & \alpha_{N-2} & \alpha_{N-1} & \alpha_N \\ 0\ldots & \ldots & \ldots & 0 & 0 & \alpha_{N-4} & \alpha_{N-3} & \alpha_{N-2} & \alpha_{N-1} & \alpha_N \\ 0\ldots & \ldots & \ldots & 0 & 0 & \alpha_{N-4} & \alpha_{N-3} & \alpha_{N-2} & \alpha_{N-1} & \alpha_N \\ 0\ldots & \ldots & \ldots & 0 & 0 & \alpha_{N-4} & \alpha_{N-3} & \alpha_{N-2} & \alpha_{N-1} & \alpha_N \\ 0\ldots & \ldots & \ldots & 0 & 0 & \alpha_{N-4} & \alpha_{N-3} & \alpha_{N-2} & \alpha_{N-1} & \alpha_N \\ 0\ldots & \ldots & \ldots & 0 & 0 & \alpha_{N-4} & \alpha_{N-3} & \alpha_{N-2} & \alpha_{N-1} & \alpha_N \\ 0\ldots & \ldots & \ldots & 0 & 0 & \alpha_{N-4} & \alpha_{N-3} & \alpha_{N-2} & \alpha_{N-1} & \alpha_N \end{bmatrix}, \tag{2.24}$$

$$C_R = \mathrm{diag}[0, 0, \ldots, 0, \quad -20\alpha_N/71] \tag{2.25}$$

$$A_e^{(R)} = \begin{bmatrix} 0 & 0 & \ldots & 0 & 1 & -5 & 10 & -10 & 5 & -1 \\ 0 & 0 & \ldots & 0 & 1 & -5 & 10 & -10 & 5 & -1 \\ \vdots & & & & & & & & & \\ 0 & 0 & \ldots & 0 & 1 & -5 & 10 & -10 & 5 & -1 \end{bmatrix} \tag{2.26}$$

The α's are here:

$$\begin{aligned} \alpha_N &= 1 + \frac{25}{12}\gamma_R + \frac{35}{24}\gamma_R^2 + \frac{5}{12}\gamma_R^3 + \frac{1}{24}\gamma_R^4 \\ \alpha_{N-1} &= -\left(4\gamma_R + \frac{13}{3}\gamma_R^3 + \frac{3}{2}\gamma_R^3 + \frac{1}{6}\gamma_R^4\right) \\ \alpha_{N-2} &= 3\gamma_R + \frac{19}{4}\gamma_R^2 + 2\gamma_R^3 + \frac{1}{4}\gamma_R^4 \qquad (2.27) \\ \alpha_{N-3} &= -\left(\frac{4}{3}\gamma_R + \frac{7}{3}\gamma_R^2 + \frac{7}{6}\gamma_R^3 + \frac{1}{6}\gamma_R^4\right) \\ \alpha_{N-4} &= \frac{1}{4}\gamma_R + \frac{11}{24}\gamma_R^2 + \frac{1}{4}\gamma_R^3 + \frac{1}{24}\gamma_R^4, \end{aligned}$$

$$\tau_R = \frac{1}{12h^2}\mathrm{diag}[0, \ldots, \tau_{N-4}, \tau_{N-3}, \tau_{N-2}, \tau_{N-1}, \tau_N], \tag{2.28}$$

$$\tau_N = 71/2\alpha_N$$

$$
\begin{aligned}
\tau_{N-1} &= (-94 - \alpha_{N-1}\tau_N)/\alpha_N \\
\tau_{N-2} &= (113 - \alpha_{N-2}\tau_N)/\alpha_N \qquad (2.29)\\
\tau_{N-3} &= (-36 - \alpha_{N-3}\tau_N)/\alpha_N \\
\tau_{N-4} &= (11 - \alpha_{N-4}\tau_N)/\alpha_N
\end{aligned}
$$

We are now ready to construct

$$
\begin{aligned}
\frac{1}{2}(M+M^T) = \frac{1}{2}\Big\{D + D^T \quad &- \quad [\tau_L(A_\alpha^{(L)} + c_L A_e^{(L)}) + \tau_R(A_\alpha^{(R)} + c_R A_e^{(R)})] \\
&- \quad [\tau_L(A_\alpha^{(L)} + c_L A_e^{(L)}) + \tau_R(A_\alpha^{(R)} + c_R A_e^{(R)}]^T\Big\} \qquad (2.30)
\end{aligned}
$$

Upon using equations (2.14)-(2.29) in (2.30) one gets:

$$
\frac{M+M^T}{2} =
$$

$$
\frac{1}{24h^2}\left[\begin{array}{ccccccccccccccccc}
 & & & & & & 0 & 6 & & & & & & & & & \\
 & & W^{(L)} & & & & 0 & & & & & & & & & & \\
 & & & & & & -2 & 0 & & & & & & & & & \\
 & & & & & & 32 & -2 & & & & & 0 & & & & \\
0 & \dots & & 0 & -2 & 32 & -60 & 32 & -2 & & & & & & & & \\
 & & & & & -2 & 32 & -60 & 32 & -2 & & & & & & & \\
 & & & & & & -2 & 32 & -60 & 32 & & & & & & & \\
 & & & & & & & \ddots & \ddots & \ddots & \ddots & \ddots & & & & & \\
 & & & & & & & & -2 & 32 & -60 & -32 & 2 & & 0 & \dots & 0 \\
 & & & & 0 & & & & & & & & & & & & \\
 & & & & & & & & & -2 & 32 & & & & & & \\
 & & & & & & & & & & -2 & & & W^{(R)} & & & \\
 & & & & & & & & & & 0 & & & & & &
\end{array}\right]
$$

(2.31)

where $W^{(L)}$ and $W^{(R)}$ are 6×6 blocks given by:

$$W^{(L)} = W_1^{(L)} + W_2^{(L)} \tag{2.32}$$

$$W^{(R)} = W_1^{(R)} + W_2^{(R)} \tag{2.33}$$

$$W_{1_{ij}}^{(L)} = \left\{ \begin{array}{ll} 0 & i = 1 \text{ or } j = 1 \\ \\ \\ -(\alpha_i \tau_j + \alpha_j \tau_i) & i, j \neq 1 \end{array} \right\} \quad 1 < i, j < 5 \tag{2.34}$$

$$W_{1_{ij}}^{(L)} = \left\{ \begin{array}{ll} 0 & i = N \text{ or } j = N \\ \\ \\ -(\alpha_{N-i} \tau_{N-j} + \alpha_{N-j} \tau_{N-i}) & \end{array} \right\} \quad 0 \leq N - i, N - j \leq 4 \tag{2.35}$$

$$W_2^{(L)} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -30 & 12 & 13 & -6 & 1 \\ 0 & 12 & -60 & 32 & -2 & 0 \\ 0 & 13 & 32 & -60 & 32 & -2 \\ 0 & -6 & -2 & 32 & -60 & 32 \\ 0 & 1 & 0 & -2 & 32 & -60 \end{bmatrix} \tag{2.36}$$

$$W_2^{(R)} = \begin{bmatrix} -60 & 32 & -2 & 0 & 1 & 0 \\ 32 & -60 & 32 & -2 & -6 & 0 \\ -2 & 32 & -60 & 32 & 13 & 0 \\ 0 & -2 & 32 & -60 & 12 & 0 \\ 1 & -6 & 13 & 12 & -30 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix} \tag{2.37}$$

The next task is to show that $\tilde{M} = \frac{1}{2}(M + M^T)$ is negative definite. We write the symmetric matrix $\tilde{M}$ as a sum of five symmetric matrices,

$$\tilde{M} = \frac{1}{24h^2}\left[\beta_0\tilde{M}_1 + 2\tilde{M}_2 + (24 - \beta_0)\tilde{M}_3 + \tilde{M}_4 + \tilde{M}_5\right]. \tag{2.38}$$

We shall show that $\tilde{M}_1$ is negative definite, and that $\tilde{M}_j (j = 2, \ldots 5)$ are non-positive definite. The $\tilde{M}$'s are given by

$$\tilde{M}_1 = \left[\begin{array}{ccccccccc} -\frac{1}{2\beta_0} & 0 & 0 & & & & & & \\ 0 & -2 & 1 & 0 & 0 & & & & \\ 0 & 1 & -2 & 1 & 0 & & & & \\ 0 & 0 & 1 & -2 & 1 & & & & \\ 0 & 0 & 0 & 1 & -2 & 1 & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & & 1 & -2 & 1 & 0 \\ & & & & & & 1 & -2 & 0 \\ & & & & & & 0 & 0 & -\frac{1}{2\beta_0} \end{array}\right] = M_1^L + \hat{M}_1 + M_1^R \tag{2.39}$$

where $M_1^L = \left[\begin{array}{cc} -1/2\beta_0 & 0 \\ 0 & 0 \end{array}\right]$, $M_2^R = \left[\begin{array}{cc} 0 & 0 \\ 0 & -1/2\beta_0 \end{array}\right]$ and $\hat{M}_1$ is the remaining $(N-2)\times(N-2)$ middle block.

$$\tilde{M}_2 = \left[\begin{array}{cccccccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & \text{\Large 0} & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\
0 & 0 & 0 & 0 & -1 & 2 & -1 & & & & & & & & & \\
0 & 0 & 0 & 0 & 2 & -5 & 4 & -1 & & & & & & & & \\
 & & & & -1 & 4 & -6 & 4 & -1 & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & -1 & 4 & -6 & 4 & -1 & & & & \\
 & & & & & & & & -1 & 4 & -5 & 2 & 0 & 0 & 0 & 0 \\
 & & & \text{\Large 0} & & & & & & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right] \tag{2.40}$$

$$\tilde{M}_3 = \left[\begin{array}{ccccccccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & \text{\Large 0} & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & -1 & 1 & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 & & & & & & & & & \\
 & & & & & & 1 & -2 & 1 & & & & & & & & \\
 & & & & & & & & & & & & & & & & \\
 & & & & & & & & & & & & & & & & \\
 & & & & & & & & & 1 & -2 & 1 & & & & & \\
 & & & & & & & & & & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & \text{\Large 0} & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right] \tag{2.41}$$

$$\tilde{M}_4 = \left[\begin{array}{c|c|c|c|c|c} -1/2 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \begin{array}{c} -30+2\beta \\ 2\alpha_2\tau_2 \end{array} & \begin{array}{c} 12-\beta \\ -(\alpha_2\tau_3+\alpha_3\tau_2) \end{array} & \begin{array}{c} 13 \\ -(\alpha_2\tau_4+\alpha_4\tau_2) \end{array} & \begin{array}{c} -6 \\ -(\alpha_2\tau_5+\alpha_5\tau_2) \end{array} & 1 \\ \hline 0 & \begin{array}{c} 12-\beta \\ -(\alpha_2\tau_3+\alpha_3\tau_2) \end{array} & \begin{array}{c} -60+2\beta \\ -2\alpha_3\tau_3 \end{array} & \begin{array}{c} 32-\beta \\ -(\alpha_3\tau_4+\alpha_4\tau_3) \end{array} & \begin{array}{c} -2 \\ -(\alpha_3\tau_5+\alpha_5\tau_3) \end{array} & 0 \\ \hline 0 & \begin{array}{c} 13 \\ -(\alpha_2\tau_4+\alpha_4\tau_2) \end{array} & \begin{array}{c} 32-\beta \\ -(\alpha_3\tau_4+\alpha_4\tau_3) \end{array} & \begin{array}{c} -60+2\beta \\ -2\alpha_4\tau_4 \end{array} & \begin{array}{c} 32-\beta \\ -(\alpha_4\tau_5+\alpha_5\tau_4) \end{array} & -2 \\ \hline 0 & \begin{array}{c} -6 \\ -(\alpha_2\tau_5+\alpha_5\tau_2) \end{array} & \begin{array}{c} -2 \\ -(\alpha_3\tau_5+\alpha_5\tau_3) \end{array} & \begin{array}{c} 32-\beta \\ -(\alpha_4\tau_5+\alpha_5\tau_4) \end{array} & \begin{array}{c} 58+\beta \\ -2\alpha_5\tau_5 \end{array} & 28-\beta \\ \hline 0 & 1 & 0 & -2 & 28-\beta & 26+\beta \end{array}\right] \tag{2.42}$$

$$\tilde{M}_5 = \left[\begin{array}{c|c|c|c|c|c}
-26+\beta & 28-\beta & -2 & 0 & 1 & 0 \\ \hline
28-\beta & \begin{array}{c} -58+2\beta \\ -2\alpha_{N-4}\tau_{N-4} \end{array} & \begin{array}{c} 32-\beta \\ -(\alpha_{N-3}\tau_{N-4} \\ +\alpha_{N-4}\tau_{N-3}) \end{array} & \begin{array}{c} -2 \\ -(\alpha_{N-2}\tau_{N-4} \\ +\alpha_{N-4}\tau_{N-2}) \end{array} & \begin{array}{c} -6 \\ -(\alpha_{N-1}\tau_{N-4} \\ +\alpha_{N-4}\tau_{N-1}) \end{array} & 0 \\ \hline
-2 & \begin{array}{c} 32-\beta \\ -(\alpha_{N-3}\tau_{N-4} \\ +\alpha_{N-4}\tau_{N-3} \end{array} & \begin{array}{c} -60+2\beta \\ -2\alpha_{N-3}\tau_{N-3} \end{array} & \begin{array}{c} 32-\beta \\ -2(\alpha_{N_2}\tau_{N-3} \\ -\alpha_{N-3}\tau_{N-2}) \end{array} & \begin{array}{c} 13 \\ -(\alpha_{N-1}\tau_{N-3} \\ +\alpha_{N-3}\tau_{N-1}) \end{array} & 0 \\ \hline
0 & \begin{array}{c} -2 \\ -(\alpha_{N-2}\tau_{N-4} \\ +\alpha_{N-4}\tau_{N-2}) \end{array} & \begin{array}{c} 32-\beta \\ -2(\alpha_{N-2}\tau_{N-3} \\ +\alpha_{N-3}\tau_{N-2}) \end{array} & \begin{array}{c} -60+2\beta \\ -2\alpha_{N-2}\tau_{N-2} \end{array} & \begin{array}{c} 12-\beta \\ -(\alpha_{N-1}\tau_{N-2} \\ +\alpha_{N-2}\tau_{N-1}) \end{array} & 0 \\ \hline
1 & \begin{array}{c} -6 \\ -(\alpha_{N-1}\tau_{N-4} \\ +\alpha_{N-4}\tau_{N-1}) \end{array} & \begin{array}{c} 13 \\ -(\alpha_{N-1}\tau_{N-3} \\ +\alpha_{N-3}\tau_{N-1} \end{array} & \begin{array}{c} 12-\beta \\ -(\alpha_{N-1}\tau_{N-2} \\ (\alpha_{N-2}\tau_{N-1}) \end{array} & \begin{array}{c} -30+2\beta \\ -2\alpha_{N-1}\tau_{N-1} \end{array} & 0 \\ \hline
0 & 0 & 0 & 0 & 0 & -1/2
\end{array}\right]. \tag{2.43}$$

Let us consider $\hat{M}_1$ - see (2.39); it may be decomposed as follows:

$$\hat{M}_1 = -\begin{bmatrix} 1 & -1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & -1 \\ & & & & 1 \end{bmatrix} \begin{bmatrix} 1 & & & & \\ -1 & \ddots & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} + \begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & -1 \end{bmatrix} \tag{2.44}$$

The last matrix in non-positive definite. The first term is a product of a regular matrix with its transpose, hence its negative is a negative definite matrix. Thus we established that $\hat{M}_1$

is negative definite for any finite dimension N. All its eigenvalues are negative. It remains to show that the eigenvalues of $\tilde{M}_1/h^2$ (see (2.38) are bounded away from zero by a constant as $h \rightarrow 0$ $(N \rightarrow \infty)$.

Consider a symmetric tridiagonal matrix S with, like $\hat{M}_1$, constant diagonals:

$$S = \begin{bmatrix} b & a & 0 & & & \\ a & b & a & & & \\ 0 & a & b & a & & \\ & \ddots & \ddots & \ddots & & \\ & & & a & b & a \\ & & & & a & b \end{bmatrix}. \tag{2.45}$$

Designate by D_j the determinant of the upper-left $j \times j$ sub-matrix. Thus $D_1 = b$, $D_2 = \det \begin{bmatrix} b & a \\ a & b \end{bmatrix}$, etc.

We have then $D_1 = b$, $D_2 = b^2 - a^2$ and in general

$$D_j = bD_{j-1} - a^2 D_{j-2} \tag{2.46}$$

It can be shown (see Appendix I) that the solution to the recursion relation (2.46) is

$$D_j = -\frac{1}{a^2}\left[\frac{A}{\mu_1^j} + \frac{B}{\mu_2^j}\right] \tag{2.47}$$

where

$$\mu_1 = \frac{1}{2a^2}\left[b + \sqrt{b^2 - 4a^2}\right] \tag{2.48}$$

$$\mu_2 = \frac{1}{2a^2}\left[b - \sqrt{b^2 - 4a^2}\right] \tag{2.49}$$

$$A = \frac{1}{\mu_1 - \mu_2}\left[(D_2 - bD_1)\,\mu_1 + D_1\right] \tag{2.50}$$

$$B = \frac{1}{\mu_1 - \mu_2}\left[(D_2 - bD_1)\,\mu_2 + D_1\right] \tag{2.51}$$

We have already shown that $\tilde{M}_1$ is negative definite. The eigenvalue of $\hat{M}_1$ are found from

$$\det(\tilde{M}_1 - I\lambda) = \left(-\frac{1}{2\beta_0} - \lambda\right) \cdot \det(\hat{M}_1 - \lambda I) \cdot \left(-\frac{1}{2\beta_0} - \lambda\right) = 0 \tag{2.52}$$

thus either $\lambda = -1/2\beta_0 < 0$ (because β_0 will be taken positive) or $\lambda =$ eigenvalue of $\hat{M}_1 < 0$. We would like to investigate the behavior of the eigenvalues of $\frac{\beta_0}{24h^2}\tilde{M}_1$. In particular we would like to show that these eigenvalues (which are negative) are bounded away from zero. To show this we analyze the behavior of $\hat{M}_1 - \lambda I$ as N increases. We now take $S = \hat{M}_1 - \lambda I$. Its determinant is given by D_{N-2}. Substituting (2.48)-(2.51) into (2.47) with $j = N-2$ we get after some elementary manipulations

$$D_{N-2} = \frac{2^{N-2}}{\rho r^{N-3}} \sin(N-1)\theta \tag{2.53}$$

where

$$\begin{aligned} \rho &= \sqrt{4 - b^2}; \quad b = -2 - \lambda; \quad a = 1 \\ r &= \sqrt{b^2 + \rho^2} = 2 \\ \theta &= \tan^{-1}(\rho/b) \end{aligned} \tag{2.54}$$

From (2.52) we require

$$D_{N-2} = 0 \tag{2.55}$$

This is equivalent, see (2.53), to requiring

$$\theta = \frac{k\pi}{N-1}, \qquad k = 1, \ldots, N-2. \tag{2.56}$$

From the definition of θ and (2.54) we obtain

$$\tan\left(\frac{k\pi}{N-1}\right) = -\frac{\sqrt{-\lambda(\lambda+4)}}{2+\lambda}, \qquad (\lambda < 0). \tag{2.57}$$

Squaring (2.57) we get a quadratic equation for λ, the solution of which is

$$\begin{aligned} \lambda &= -2\left[1 \pm \left(1 + \tan^2\left(\frac{k\pi}{N-1}\right)\right)^{-1/2}\right] \\ &= -2\left[1 \pm \cos\left(\frac{k\pi}{N-1}\right)\right]. \end{aligned} \tag{2.58}$$

For any fixed N, the smallest values of $|\lambda|$ is given by (2.58) for $k = 1$,

$$\lambda_{\max} = \min_k |\lambda| = -2\left[1 - \cos\left(\frac{\pi}{N-1}\right)\right]. \tag{2.59}$$

As N increases, we have

$$\begin{aligned} \lambda_{\max} &\to -2\left[1 - \left(1 - \frac{\pi^2}{2(N-1)^2} + O\left(\frac{1}{N^4}\right)\right)\right] \\ &= -\frac{\pi^2}{(N-1)^2} \approx -\pi^2 h^2. \end{aligned} \tag{2.60}$$

Thus the eigenvalues of $\hat{M}_1/24h^2$ (and hence of $\tilde{M}_1/24h^2$) are bounded away from zero by the value $-\left(\frac{\pi^2}{24}\right)$.

We now consider $\tilde{M}_2$. One can verify that

$$\tilde{M}_2 = -\hat{M}_2 \hat{M}_2^T \tag{2.61}$$

where

$$
\hat{M}_2 = \left[\begin{array}{cccccccccccccccc}
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & & & & & \Large 0 & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & -2 & 1 & 0 & & & & & & & \\
 & & & & & 0 & 1 & -2 & 1 & & & & & & & \\
 & & & & & & & 1 & -2 & 1 & & & & & & \\
 & & & & & & & & & & & & & & & \\
 & & & & & & & & & & 1 & 0 & & & & \\
 & & & & & & & & 0 & 1 & -2 & 1 & 0 & & & \\
\Large 0 & & & & & & & & & 0 & 1 & -2 & 0 & & & \\
 & & & & & & & & & & 0 & 1 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 0 & 0 & 0 & 0 & 0 \\
 & & & & & & & & & & & 0 & 0 & 0 & 0 & 0
\end{array}\right] \tag{2.62}
$$

Therefore $\tilde{M}_2$ is non-positive definite. In a similar fashion $\tilde{M}_3$ is non-positive definite because

$$
M_3 = -\hat{M}_3 \hat{M}_3^T \tag{2.63}
$$

with

$$\hat{M}_3 = \left[\begin{array}{cccccccccccccccc} 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & & & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & & & & & & & & & \\ & & & & & & 1 & -1 & & & & & & & & \\ & & & & & & & & & & & & & & & \\ & & & & & & & & & 1 & -1 & & & & & \\ & & & & & & & & & & 1 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right]. \qquad (2.64)$$

The matrices $\tilde{M}_4$ and $\tilde{M}_5$ are $N \times N$ matrices with zero entries except for 6×6 upper-left (lower-right) blocks. It is sufficient to show that these blocks are negative definite. This was done symbolically using the Mathematica software and plotted for $0 \leq \gamma_L, \gamma_R < 1$ and $\beta_0 = 1$. $\tilde{M}_4$ and $\tilde{M}_5$ are indeed negative definite for, $0 \leq \gamma_R, \gamma_L < 1$. Thus we have shown that $\tilde{M} = \frac{1}{2}(M + M^T)$ is indeed negative definite, and its eigenvalues are bounded away from zero by $(-\pi^2/24)$, even as $N \to \infty$, and the error estimate (2.13) is valid.

3 The Two Dimensional Case

We consider the inhomogeneous diffusion equation, with constant coefficients, in a domain Ω. To begin with we shall assume that Ω is convex and has a boundary curve $\partial\, \Omega \,\epsilon\, \mathcal{C}^2$. The convexity restriction is for the sake of simplicity in presenting the basic idea; it will be removed later. We thus have

$$\frac{\partial u}{\partial t} = k\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + f(x,y;t); \qquad x, y \epsilon\, \Omega; \;\; t \geq 0; \;\; k > 0 \tag{3.1a}$$

$$u(x,y,0) = u_0(x,y) \tag{3.2b}$$

$$u(x,y,t)|_{\partial\Omega} = u_B(t) \tag{3.1c}$$

We shall refer to the following grid representation:

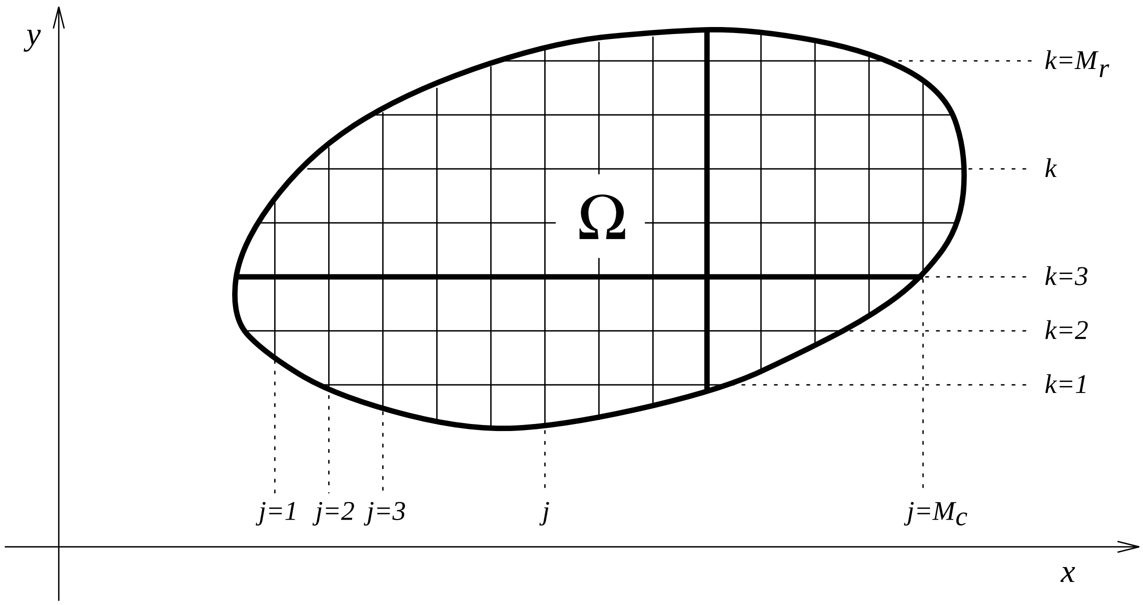

Figure 2: Two dimensional grid.

We have M_R rows and M_c columns inside Ω. Each row and each column has a discreitized structure as in the one 1-D case, see figure 1. Let the number of grid points in the k^{th} row be denoted by R_k and similarly let the number of grid points in the j^{th} column be C_j. Let the solution projection be designated by $U_{j,k}(t)$. By $\mathbf{U}(t)$ we mean, by analogy to the 1-D case,

$$\begin{aligned}\mathbf{U}(t) &= (u_{1,1}, u_{2,1}, \ldots, u_{R_1,1}; u_{1,2}, u_{2,2}, \ldots, u_{R_2,2}; \ldots; u_{1,M_R}, u_{2,M_R}, \ldots u_{R_{M_R},M_R}) \\ &\equiv (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{M_R}) \qquad (3.2)\end{aligned}$$

Thus, we have arranged the solution projection array in vectors according to rows, starting from the bottom.

If we arrange this array by columns (instead of rows) we will have the following structure

$$\begin{aligned}\mathbf{U}^{(c)}(t) &= (u_{1,1}, u_{1,2}, \ldots, u_{1,c_1}; u_{2,1}, u_{2,2}, \ldots, u_{2,c_2}; \ldots; u_{M_c,1}, u_{M_c,2}, \ldots, u_{M_c,c_{M_c}}) \\ &\equiv (\mathbf{u}_1^{(c)}, \mathbf{u}_2^{(c)}, \ldots, \mathbf{u}_{M_c}^{(c)}) \qquad (3.3)\end{aligned}$$

Since $\mathbf{U}^{(c)}(t)$ is just a permutation of $\mathbf{U}(t)$, there must exist an orthogonal matrix $\mathbf{P}$ such that

$$\mathbf{U}^{(c)}(t) = \mathbf{PU} \qquad (3.4)$$

If the length of $\mathbf{U}(t)$ is ℓ, then P is an $\ell \times \ell$ matrix whose each row contains $\ell - 1$ zeros and a single 1 somewhere.

The second derivative operator $\partial^2/\partial x^2$ in (3.1a) is represented on the k^{th} row by the differentiation matrix $D_k^{(x)}$, whose structure is given by (2.14). Similarly let $\partial^2/\partial y^2$ be given on the j^{th} column by $D_j^{(y)}$, whose structure is also given by (2.14). With this notation the Laplacian of the solution projection is:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) u_{ij}(t) = \mathcal{D}^{(x)}\mathbf{U} + \mathcal{D}^{(y)}\mathbf{U}^{(c)} + \mathbf{T}_e^{(x)} + \mathbf{T}_e^{(y)} \qquad (3.5)$$

where

$$\mathcal{D}^{(x)} = \begin{bmatrix} D_1^{(x)} & & \\ & D_2^{(x)} & \\ & & D_{M_R}^{(x)} \end{bmatrix} ; \mathcal{D}^{(y)} = \begin{bmatrix} D_1^{(y)} & & \\ & D_2^{(y)} & \\ & & D_{M_c}^{(y)} \end{bmatrix} \tag{3.6}$$

where $\mathcal{D}^{(x)}$ and $\mathcal{D}^{(y)}$ are $(\ell \times \ell)$ matrices and have the block structures shown. $\mathbf{T}_e^{(x)}$ and $\mathbf{T}_e^{(y)}$ are the truncation errors associated with $\mathcal{D}^{(x)}$ and $\mathcal{D}^{(y)}$, respectively. We now call attention to the fact that $\mathcal{D}^{(x)}$ and $\mathcal{D}^{(y)}$ do not operate on the same vector. This is fixed using (3.4):

$$\nabla^2 u_{ij}(t) = \nabla^2 \mathbf{U} = (\mathcal{D}^{(x)} + P^T \mathcal{D}^{(y)} P)\mathbf{U} + \mathbf{T}_e^{(x)} + P^T \mathbf{T}_e^{(y)} \tag{3.7}$$

Thus (3.1a) becomes, by analogy to (2.5),

$$\frac{d\mathbf{U}}{dt} = k(\mathcal{D}^{(x)} + P^T \mathcal{D}^{(y)} P)\mathbf{U} + k(\mathbf{T}_e^{(x)} + P^T T_e^{(y)}) + \mathbf{f}(t) \tag{3.8}$$

where $\mathbf{f}(t)$ is $f(x, y; t)$ arranged by *rows* as a vector.

Before proceeding to the semi-discrete problem let us define:

$$M_k^{(x)} = D_k^{(x)} - \tau_{L_k} A_{L_k} - \tau_{R_k} A_{R_k} \tag{3.9}$$

where τ_{L_k}, A_{L_k} are the τ_L and A_L defined in section 2, appropriate to the k^{th} row; similarly for τ_{R_k} and A_{R_k}. In the same way, define

$$M_j^{(y)} = D_j^{(y)} - \tau_{L_j} A_{L_j} - \tau_{R_j} A_{R_j} \tag{3.10}$$

where the notation should be self explanatory by now.

We can now write the semi-discrete problem by analogy to (2.6)

$$\frac{d\mathbf{V}}{dt} = k(\mathcal{M}^{(x)} + P^T \mathcal{M}^{(y)} P)\mathbf{V} + k\mathbf{G}^{(x)} + kP^T \mathbf{G}^{(y)} + \mathbf{f}(t) \tag{3.11}$$

where $\mathbf{V}$ is the numerical approximation to $\mathbf{U}$;

$$\mathcal{M}^{(x)} = \begin{bmatrix} M_1^{(x)} & & \\ & M_2^{(x)} & \\ & & M_{M_R}^{(x)} \end{bmatrix} ; \mathcal{M}^{(y)} = \begin{bmatrix} M_1^{(y)} & & \\ & M_2^{(y)} & \\ & & M_{M_c}^{(y)} \end{bmatrix} ; \tag{3.12}$$

and

$$\begin{aligned} \mathbf{G}^{(x)} &= \Big[(\tau_{L_1}\mathbf{g}_{L_1} + \tau_{R_1}\mathbf{g}_{R_1}), \ldots, (\tau_{L_k}\mathbf{g}_{L_k} + \tau_{R_k}\mathbf{g}_{R_k}), \ldots, (\tau_{L_{M_r}}\mathbf{g}_{L_{M_R}} + \tau_{R_{M_R}}\mathbf{g}_{R_{M_R}})\Big] , \\ \mathbf{G}^{(y)} &= \Big[(\tau_{L_1}\mathbf{g}_{L_1} + \tau_{R_1}g_{R_1}), \ldots, (\tau_{L_j}\mathbf{g}_{L_j} + \tau_{R_j}\mathbf{g}_{R_j}), \ldots, (\tau_{L_{M_c}}\mathbf{g}_{L_{M_c}} + \tau_{R_{M_c}}\mathbf{g}_{R_{M_c}})\Big] . \end{aligned} \tag{3.13}$$

Note that in $\mathbf{G}^{(y)}$ the indices "L" and "R" designate, of course, bottom and top, rather than left and right. Subtracting (3.11) from (3.8) we get in a fashion similar to the derivation of (2.9):

$$\frac{d\mathbf{E}}{dt} = k[\mathcal{M}^{(x)} + P^T\mathcal{M}^{(y)}P]\mathbf{E} + k\mathbf{T}_2 \tag{3.14}$$

where $\mathbf{E} = \mathbf{U} - \mathbf{V}$ is the two dimensional array of the errors, ϵ_{ij}, arranged by rows as a vector. $\mathbf{T}_2$ is proportional to the truncation error.

The time change of $\| \mathbf{E} \|^2$ is given by

$$\frac{1}{2}\frac{d}{dt} \| \mathbf{E} \|^2 = k(\mathbf{E}, (\mathcal{M}^{(x)} + P^T\mathcal{M}^{(y)}P)\mathbf{E}) + k(\mathbf{E}, \mathbf{T}_2) \tag{3.15}$$

The symmetric part of $\mathcal{M}^{(x)} + P^T\mathcal{M}^{(y)}P$ is given by

$$\frac{1}{2}[(\mathcal{M}^{(x)} + \mathcal{M}^{(x)^T}) + P^T(\mathcal{M}^{(y)} + \mathcal{M}^{(y)^T})P] \tag{3.16}$$

Clearly $\mathcal{M}^{(x)} + \mathcal{M}^{(x)^T}$ and $\mathcal{M}^{(y)} + \mathcal{M}^{(y)^T}$ are block-diagonal matrices with typical blocks given by $M_k^{(x)} + M_k^{(x)^T}$ and $M_j^{(y)} + M_j^{(y)^T}$. We have already shown in the one dimensional case that each one of those blocks is negative definite and bounded away from zero by $\pi^2/24$. Therefore the operator (3.16) is also negative definite and bounded away from zero. The rest of the proof follows the one dimensional case and thus the norm of the error $\| E \|$ is bounded by a constant.

If the domain Ω is not convex or simply connected then either rows or columns, or both, may be "interrupted" by $\partial\Omega$. In that case the values of the solution on each "internal" interval (see figure [3] below) are taken as *separate* vectors.

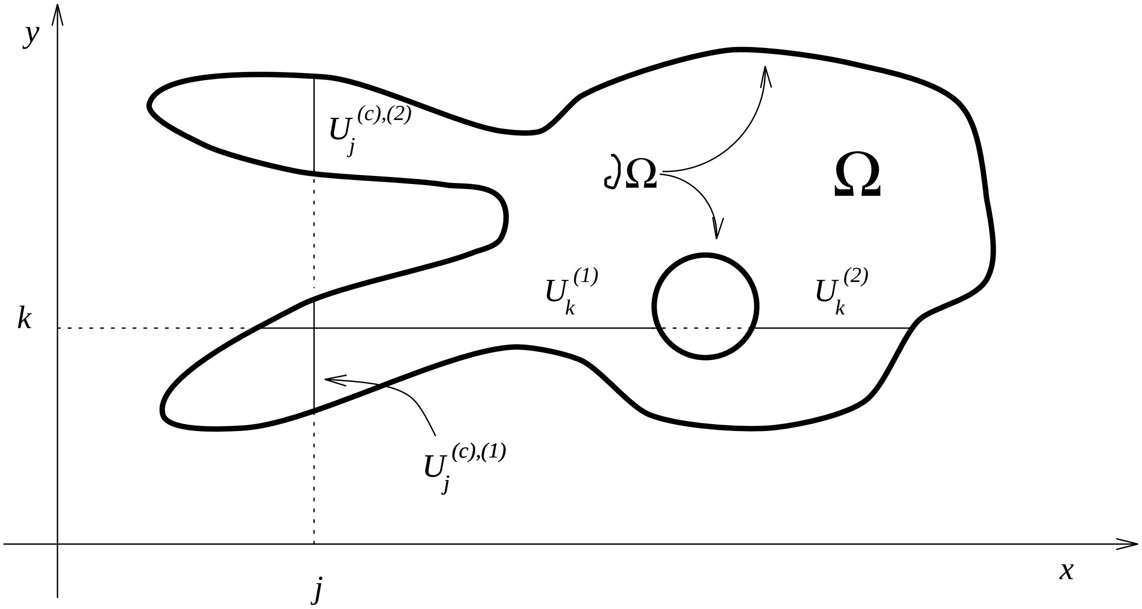

Figure 3: Two dimensional grid, non convex domain.

Decomposing "interrupted" vectors in this fashion leaves the previous analysis unchanged. The length of $\mathbf{U}$ (or $\mathbf{U}^{(c)}$) is again ℓ, where ℓ is the number of grid nodes inside Ω. The differentiation and permutation matrices remain $\ell \times \ell$. Note that adding more "holes" inside $\partial\Omega$ does not change the general approach.

4 Numerical Example

In this section we describe numerical results for the following problem:

$$\frac{\partial u}{\partial t} = k(u_{xx} + u_{yy}) + f(x, y, t), \qquad (x, y)\epsilon\Omega, \;\; t > 0, \tag{4.1}$$

where Ω is the region contained between a circle of radius $r_0 = 1/2$ and inner circle of radius $r_i \leq 0.1$. The inner circle is *not concentric* with the outer one. Specifically Ω is described by

$$\Big\{(x - .5)^2 + (y - .5)^2 \leq 1/4\Big\} \cap \Big\{(x - .6)^2 + (y - .5)^2 \geq (.1 - \delta)^2; 0 < \delta < .1\Big\} \tag{4.2}$$

The cartesian grid in which Ω is embedded spans $0 \leq x, y \leq 1$. We took $\Delta x = \Delta y$, and ran several cases with $\Delta x = 1/50$, $1/75$, $1/100$. The geometry thus looks as follows:

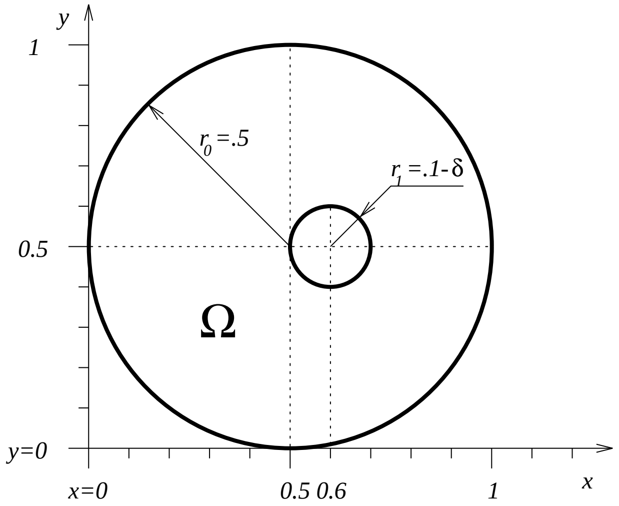

Figure 4:

The source function $f(x, y, t)$ was chosen different from zero so that we could assign an exact analytic solution to (4.1). This enables one to compute the error $E_{ij} = U_{ij} - V_{ij}$ "exactly" (to machine accuracy). We chose $k = 1$ and

$$u(x, y, t) = 1 + \cos(10t - 10x^2 - 10y^2) \tag{4.3}$$

This leads to

$$\begin{aligned} f(x, y, t) &= 400(x^2 + y^2)\cos(10t - 10x^2 - 10y^2) \\ &- 50\sin(10t - 10x^2 - 10y^2) \end{aligned} \tag{4.4}$$

From the expression for $u(x, y, t)$ one obtains the boundary and initial conditions.

The problem (4.1), (4.2), (4.4) was solved using both a "standard" fourth order algorithm (a 2-D version of (2.3)) and the new "SAT," or "bounded error," approach described in Section 3. The temporal advance was via a fourth order Runge-Kutta.

The standard algorithm was run for $\Delta x = 1/50$ and a range of $0 \leq \delta < .01$ $(.09 < r_i \leq .1)$. We found that for $\delta \geq .0017323$, the runs were stable and the error bounded for "long" times (10^5 time steps, or equivalently $t = 2$). For $0 \leq \delta < .0017233$ the results began to diverge exponentially from the analytic solution. The "point of departure" depended on δ. A discussion of these results is deferred to the next section. Figures 5,6,7 show the L_2-norm of the error vs. time for different radii of the inner "hole."

The same configurations were also run using the "bounded error" algorithm described in Section 3 (see eq. (3.5)), and the results are shown in figures 8,9,10,11. It is seen that for

δ's for which the standard methods fails, the new algorithm still has a bounded error, as predicted by the theory.

To check on the order of accuracy, the "SAT" runs (with $\delta = 0$) were repeated for $\Delta x = \Delta y = 1/75$ and $1/100$. Figure 12,13, and 14 show the logarithmic slope of the L_2, L_1 and L_∞ errors to be less than -4; i.e., we indeed have a 4^{th} order method. That the slopes are larger in magnitude than 4.5 is attributed to the fact that as $\Delta x = \Delta y$ decreases the percentage of "internal" points increases (the boundary points have formally only 3^{rd} oder accuracy). It is therefore possible that if the number of grid points was increased much further, the slope would tend to -4. Lack of computer resources prevented checking this point further. (For $\Delta x = 0.01$, running 20,000 time steps, $t = .1$, cpu time on a CRAY YMP is about 5 hours). It should also be noted that the "bounded-error" algorithm was run with a time step, Δt, twice as large as the one used in the standard scheme. At this larger Δt the standard scheme "explodes" immediately.

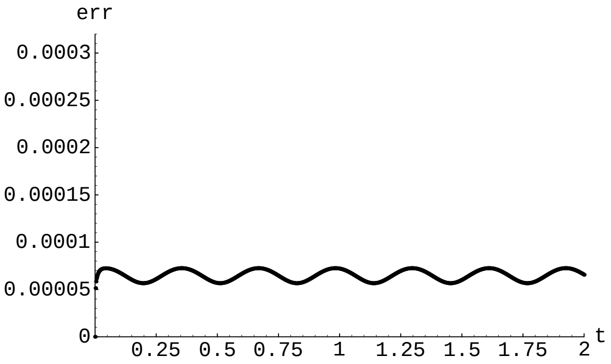

Figure 5: $\delta = 0.0017325$, Standard scheme

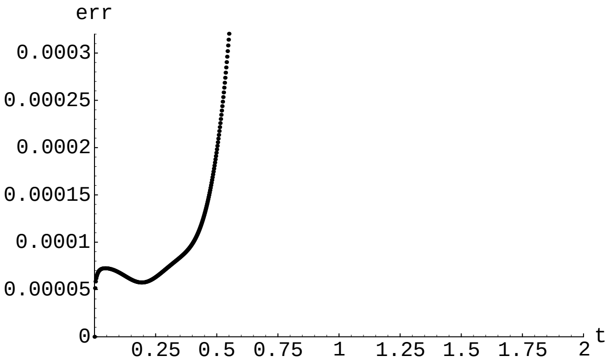

Figure 6: $\delta = 0.0017323$, Standard scheme

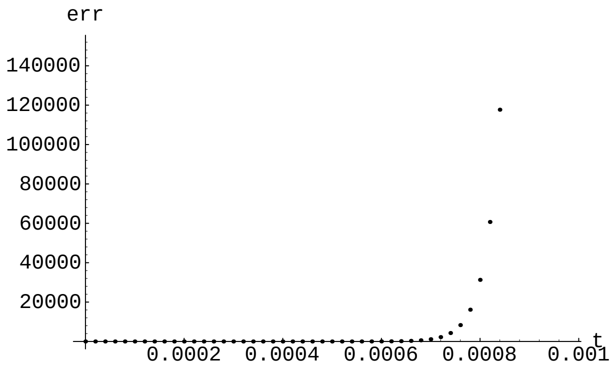

Figure 7: $\delta = 0.0015$, Standard scheme

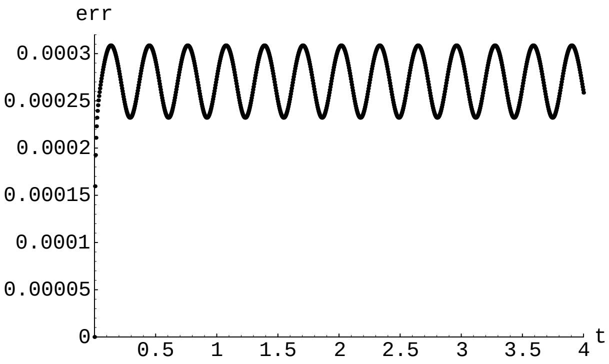

Figure 8: $\delta = 0$, SAT scheme

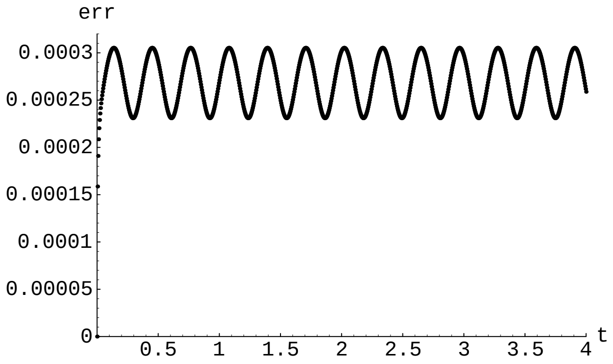

Figure 9: $\delta = 0.0015$, SAT scheme

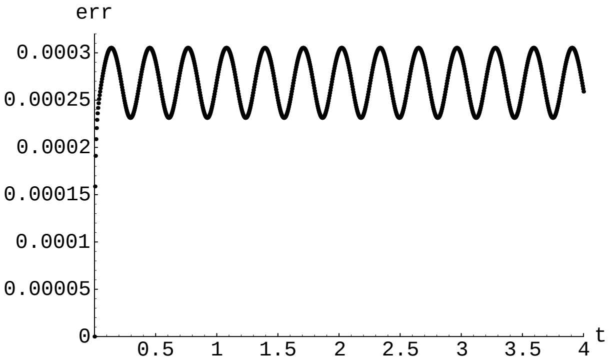

Figure 10: $\delta = 0.0017323$, SAT scheme

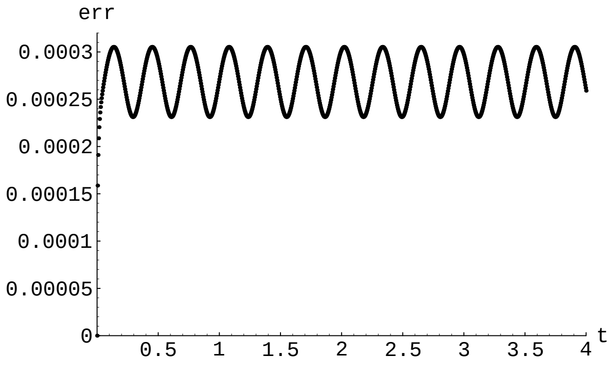

Figure 11: $\delta = 0.0017325$, SAT scheme

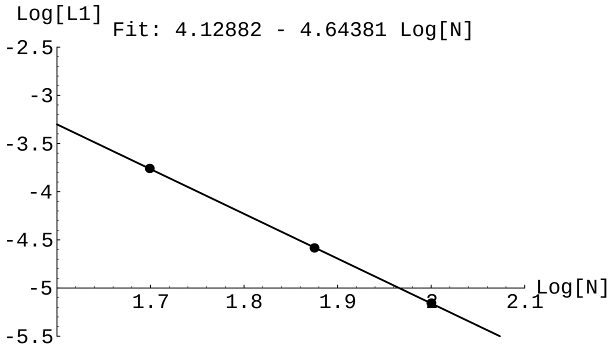

Figure 12: Order of accuracy L_1

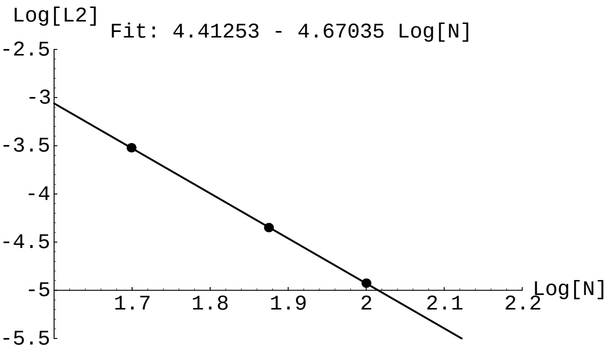

Figure 13: Order of accuracy L_2

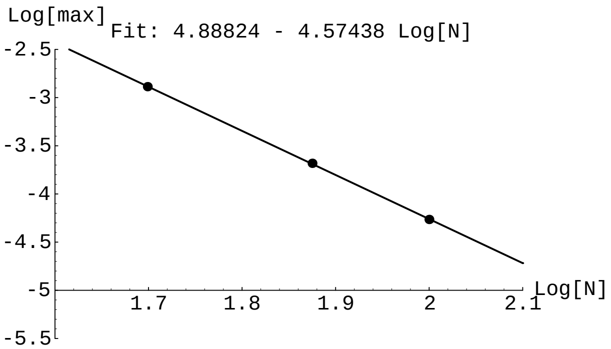

Figure 14: Order of accuracy L_∞

A study of the effect of size of Δt shows that the instabilities exhibited above are due to the time-step being near the C.F.L.-limit. It is interesting that this C.F.L.-limit depends so strongly on the geometry.

5 Conclusions

(i) The theoretical results show that one has to be very careful when using an algorithm whose differentiation matrix, or rather its symmetric part, is not negative definite. For some problems, such "standard" schemes will give good answers (i.e., bounded errors) and for others instability will set in. Thus, for example, the "standard" scheme for the 1-D case has a matrix which, for all $0 < \gamma_L, \gamma_R < 1$, though not negative definite has eigenvalues with negative real parts. This assures, in the 1-D case, the temporally asymptotic stability. In the 2-D case, even though each of the block sub-matrices of the $\ell \times \ell$ x-and-y differentiation matrices has only negative (real-part) eigenvalues, it is not assured that the sum of the two $\ell \times \ell$ matrices will have this property. This depends, among other things, on the shape of the domain and the mesh size (because the mesh size determines, for a given geometry, the γ_L and γ_R's along the boundaries). Thus that we might have the "paradoxical" situation, that for a given domain shape, successive mesh refinement could lead to instability due to the occurrence of destabilizing γ's. This cannot happen if one constructs, as was done here, a scheme whose differentiation matrices have symmetric parts that are negative definite.

It is also interesting to note that if one uses explicit standard method then the allowable C.F.L. may decrease extremely rapidly with change in the geometry that causes decrease in the γ's. This point is brought out in figures 5 to 7.

(ii) Note that the construction of the 2-D algorithm, and its analysis, which were based on the 1-D case, can be extended in a similar (albeit more complex) fashion to higher dimensions.

(iii) Also note that if the diffusion coefficient k, in the equation

$$u_t = k\Delta^2 u$$

is a function of the spatial coordinates, $k = k(x, y, z)$, the previous analysis goes through but the energy estimate for the error is now for a different, but equivalent norm.

Appendix I

We start with

$$D_j = bD_{j-1} - a^2 D_{j-2} \tag{A.1}$$

with

$$D_1 = b, D_2 = b^2 - a^2 \tag{A.2}$$

We associate with (A.1) a generating function $f(x)$,

$$f(x) = \sum_{j=0}^{\infty} D_{j+1} x^j \tag{A.3}$$

Multiplying (A.1) by x^{j-2} for each $j \geq 3$, and summing both sides we obtain:

$$\frac{f - D_1 - D_x x}{x^2} = b\frac{f - D_1}{x} - a^2 f \tag{A.4}$$

leading to

$$\begin{aligned} f &= \frac{1}{a^2}\left[\frac{D_1 + (D_2 - bD_1)x}{x^2 - (b/a^2)x + (1/a^2)}\right] \\ &= \frac{1}{a^2}\frac{D_1 + (D_2 - bD_1)x}{(x - u_1)(x - u_2)} \end{aligned} \tag{A.5}$$

where u_1, u_2 are given by (2.48), (2.49).

We may also present f by

$$f = \frac{1}{a^2}\left[\frac{A}{(x - u_1)} + \frac{B}{(x - u_2)}\right] \tag{A.6}$$

Comparing (A.6) and (A.5) we get expression for A and B as given in (2.50), (2.51). Expanding the denominator in (A.6) we get the following series for f

$$f(x) = -\frac{1}{a^2} \sum_{j=0}^{\infty} \left(\frac{A}{u_1^{j+1}} + \frac{B}{u_2^{j+1}} \right) x^j, \tag{A.7}$$

from which it immediately follows (see (A.3)) that

$$D_j = -\frac{1}{a^2} \left(\frac{A}{u_1^j} + \frac{B}{u_2^j} \right) \tag{A.8}$$

References

[1] M.H. Carpenter, D.Gottlieb and S. Abarbanel, Time Stable Boundary Conditions for Finite Difference Schemes Solving Hyperbolic Systems: Methodology and Application to High Order Compact Schemes. NASA Contractor Report 191436, ICASE Report 93-9, To appear *JCP*.

[2] P. Olson, Summation by Parts, Projections and Stability. RIACS Technical Report 93.04. (June 1993).

[3] B. Strad, Summation by Parts for Finite Difference Approximations for d/dx, Dept of Scientific Computing, Upsala University, Upsala, Sweden, August,1991

[4] B. Gustafsson, H.O. Kreiss, and A. Sundström, Stability Theory of Difference Approximations for Mixed Initial Boundary Value Problems. II, *Math. Comp.* **26**, 1972. 649-686.