N86-14083

FINITE ELEMENT OR GALERKIN TYPE SEMIDISCRETE SCHEMES †

Kanat Durgun *

ABSTRACT

A finite element or Galerkin type semidiscrete method is proposed for numerical solution of a linear hyperbolic partial differential equation. The question of stability is reduced to the stability of a system of ordinary differential equations for which Dahlquist theory applies.

We also present some results of separating the part of numerical solution which causes the spurious oscillation near shock-like response of semidiscrete scheme to a step function initial condition. In general all methods produce such oscillatory overshoots on either side of shocks. This overshoot pathology, which displays a behaviour similar to Gibb's phenomena of Fourier series, is explained on the basis of dispersion of separated Fourier components which relies on linearized theory to be satisfactory. We present expository results, polished formal proofs will appear elsewhere.

INTRODUCTION

Our model of one and two dimensional linear hyperbolic equations are

$$(1)\quad \frac{\partial U}{\partial t} + c\frac{\partial U}{\partial x} = 0$$

$$(2)\quad \frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} + b\frac{\partial U}{\partial y} = 0$$

---

† NASA Summer Faculty Fellow

* Associate Professor of Mathematics, University of Arkansas at Little Rock.

Introducing $c=\sqrt{a^2+b^2}$, $c_x=c\cos\alpha$ and $c_y=c\sin\alpha$ equation (2) can be written as

$$(2') \qquad \frac{\partial U}{\partial t}+c\cos\alpha\frac{\partial U}{\partial x}+c\sin\alpha\frac{\partial U}{\partial y}=0$$

Galerkin or finite element semidiscretization [1], [2] , [3] , [4] ,seeks an approximate solution for equation (2') in the form

$$(3) \qquad u(x,y,t)=\sum_{m,n}\varphi_{mn}(x,y)\,u_{mn}(t)$$

where

$$(4) \qquad \varphi_{mn}(x,y)=\begin{cases}1 & x=x_m\ ,\ y=y_n\\ 0 & \text{Otherwise.}\end{cases}$$

We obtain a system of ordinary differential equations by requiring that the residual $R=\frac{\partial u}{\partial t}+c\cos\alpha\frac{\partial u}{\partial x}+c\sin\alpha\frac{\partial u}{\partial y}$ be orthogonal to the basis functions φ_{mn} i.e $\langle\varphi_{mn},R\rangle=0$. Candidates for φ_{mn} are too many producing algorithms with increasing complexity proportional with their smoothness. We only present bilinear finite elements on squares. The orthogonality requirement yields, say in one dimensional case

$$(5) \qquad \frac{d}{dt}K_h u_n(t)=L_h u_n(t)$$

where K_h and L_h are discrete Toeplitz operators with eigenvectors $\{e^{i\omega x_n}\}$. If K_h is an identity operator then scheme is explicit, otherwise implicit. If the real part of the corresponding eigenvalue $\lambda(\omega)$ is zero then the scheme is conservative [6] , [7] . The quantity

$$(6) \qquad \tilde{c}(\omega)=-\frac{\operatorname{Im}\lambda(\omega)}{\omega}$$

is the velocity of propagation of numerical solutions in comparison with exact propagation velocity C in (1). The quotient $\tilde{c}(\omega)/c$ or difference $\tilde{c}(\omega)-C$ in an appropriate norm is the measure of spurious oscillations and dispersions in numerical solutions. Purely mathematical treatment without the effects of discretization i.e nonnumerical can be found in [8].

In the next sections, to study the response of semidiscrete scheme to sharp gradient changes we simulate a shock by a step function initial condition in (1), here we present an heuristic argument for the cause of parasitic oscillations around a point of discontinuity.

Consider the weighted Galerkin semidiscretization

$$\frac{\alpha}{2}\frac{du_{n-1}}{dt}+(1-\alpha)\frac{du_n}{dt}+\frac{\alpha}{2}\frac{du_{n+1}}{dt}=-\frac{c}{2h}\left(u_{n+1}-u_{n-1}\right) \tag{7}$$

of our model equation (1), where $\alpha\in[0,1]$ is a parameter. Note that $\alpha=0$ corresponds to the equation

$$\frac{du_n}{dt}=\text{centered difference approximation to }\left(-c\frac{\partial u_n}{\partial x}\right) \tag{8}$$

Since for any n, in equation (7), indices take three successive integer values we may relabel them for n even as u_n and for n odd as v_n we then obtain respectively the following systems

$$\begin{aligned}\frac{du_n}{dt}&=-\frac{c}{2h}\left(v_{n+1}-v_{n-1}\right)\\ \frac{dv_n}{dt}&=-\frac{c}{2h}\left(u_{n+1}-u_{n-1}\right)\end{aligned} \tag{9}$$

for $\alpha=0$, and

$$\begin{aligned}\frac{\alpha}{2}\left(\frac{dv_n}{dt}+\frac{dv_{n-1}}{dt}\right)+(1-\alpha)\frac{du_n}{dt}&=-\frac{c}{2h}\left(v_{n+1}-v_{n-1}\right)\\ (1-\alpha)\frac{dv_n}{dt}+\frac{\alpha}{2}\left(\frac{du_{n+1}}{dt}+\frac{du_{n-1}}{dt}\right)&=-\frac{c}{2h}\left(u_{n+1}-u_{n-1}\right)\end{aligned} \tag{10}$$

These equations are consistent approximations for the following systems

$$\begin{aligned}\frac{\partial u}{\partial t}&=-c\frac{\partial v}{\partial x}\\ \frac{\partial v}{\partial t}&=-c\frac{\partial u}{\partial x}\end{aligned} \tag{11}$$

$$\begin{aligned}\alpha\frac{\partial v}{\partial t}+(1-\alpha)\frac{\partial u}{\partial t}&=-c\frac{\partial v}{\partial x}\\ (1-\alpha)\frac{\partial v}{\partial t}+\alpha\frac{\partial u}{\partial t}&=-c\frac{\partial u}{\partial x}\end{aligned} \tag{12}$$

Eliminating u or v in (11) we obtain respectively

$$\begin{aligned}\frac{\partial^2 u}{\partial t^2}&=c^2\frac{\partial^2 u}{\partial x^2}\\ \frac{\partial^2 v}{\partial t^2}&=c^2\frac{\partial^2 v}{\partial x^2}\end{aligned} \tag{13}$$

Showing that in a doubly spaced grid wave equation is satisfied. This indicates that finite differencing is consistent with (13) rather than (1). Also adding the equations in (11)

(14) $\frac{\partial}{\partial t}\left(\frac{u+v}{2}\right) = -c\frac{\partial}{\partial x}\left(\frac{u+v}{2}\right)$

we see that discretization is consistent for the avarage of the solutions at two successive grid points. However subtracting equations in (11) we obtain

(15) $\frac{\partial}{\partial t}(u-v) = c\frac{\partial}{\partial x}(u-v)$

This shows that due to discretization difference, however small, of two successive solutions propagates as an error wave in the discrete medium in the opposite direction.

For (12), adding we obtain

(16) $\frac{\partial}{\partial t}\left(\frac{u+v}{2}\right) = -c\frac{\partial}{\partial x}\left(\frac{u+v}{2}\right)$

and subtracting we find

(17) $\frac{\partial}{\partial t}(v-u) = \frac{-c}{1-2\alpha}\frac{\partial}{\partial x}(v-u) \qquad \alpha \neq \frac{1}{2},$

which is the cause of oscillations in general.

GALERKIN SEMIDISCRETIZATION FOR EQUATION (2')

On the square with vertices (x_{m-1}, y_{n+1}), (x_{m+1}, y_{n+1}), (x_{m+1}, y_{n-1}) and (x_{m-1}, y_{n-1}) we take basis functions to be

$$(18)\quad \varphi_{mn}(x,y) = \begin{cases} 1 + \frac{x-x_m}{h} & \text{for} \quad \begin{array}{l} x_{m-1} \le x \le x_m \\ y_n-(x-x_m) \le y \le y_n+(x-x_m) \end{array} \\ 1 - \frac{y-y_n}{h} & \text{for} \quad \begin{array}{l} y_n \le y \le y_{n+1} \\ x_m-(y-y_n) \le x \le x_m+(y-y_n) \end{array} \\ 1 - \frac{x-x_m}{h} & \text{for} \quad \begin{array}{l} x_m \le x \le x_{m+1} \\ y_n-(x-x_m) \le y \le y_n+(x-x_m) \end{array} \\ 1 + \frac{y-y_n}{h} & \text{for} \quad \begin{array}{l} y_{n-1} \le y \le y_n \\ x_m+(y-y_n) \le x \le x_m-(y-y_n) \end{array} \\ 0 & \text{otherwise} \end{cases}$$

These are pyramids whose base is a square with vertices are given above, centered at (x_m, y_n) with unit height. Forming the inner products with the residual we obtain

$$(19)\quad \left\langle \varphi_{m,n}, \sum_{k,\ell} \left[\varphi_{k\ell} \frac{du_{k\ell}}{dt} + c_x u_{k\ell} \frac{\partial \varphi_{k\ell}}{\partial x} + c_y u_{k\ell} \frac{\partial \varphi_{k\ell}}{\partial y} \right\rangle = 0 \quad \forall\ m,n.$$

Only nonvanishing terms come for the values of indices $k = m-1, m, m+1$ and $\ell = n-1, n, n+1$.

Thus equation (19) reduces to

$$(20)\quad \sum_{k=-1}^{1} \sum_{\ell=-1}^{1} \left\langle \varphi_{mn}, \varphi_{m-k,n-\ell} \frac{du_{m-k,n-\ell}}{dt} + c_x u_{m-k,n-\ell} \frac{\partial \varphi_{m-k,n-\ell}}{\partial x} + c_y u_{m-k,n-\ell} \frac{\partial \varphi_{m-k,n-\ell}}{\partial y} \right\rangle = 0$$

Computation of inner products as double integrals are straightforward but tedious. Replacing the values of various integrals in equation (20), we obtain

$$(21)\quad \frac{1}{36}\frac{d}{dt}\left[u_{m-1,n-1} + 4u_{m,n-1} + u_{m+1,n-1} + 4u_{m-1,n} + 16u_{mn} + 4u_{m+1,n} + u_{m-1,n+1} + 4u_{m,n+1} + u_{m+1,n+1}\right]$$

$$= -\frac{1}{2h}\left[-\theta u_{m-1,n-1} - \alpha_y u_{m,n-1} + \mu u_{m+1,n-1} - \alpha_x u_{m-1,n} + \alpha_x u_{m+1,n} - \mu u_{m-1,n+1} + \alpha_y u_{m,n+1} + \theta u_{m+1,n+1}\right]$$

where $\theta = \frac{c_x + c_y}{6}$, $\mu = \frac{c_x - c_y}{6}$, $\alpha_x = \frac{2c_x}{3}$ and $\alpha_y = \frac{2c_y}{3}$.

System of equations (21) can be written in matrix notation on a rectangle $[0,(M+1)h] \times [0,(N+1)h]$ in various ways. Let $U_k = [u_{k1}, u_{k2} \cdots u_{kN}]^T$, $k = 1,2,\ldots N$, I_N be the $N \times N$ identity matrix, and $L_N = [\ell_{ij}]_{N \times N}$ where

$$\ell_{ij} = \begin{cases} 1 & j = i-1 \\ 0 & \text{otherwise,} \end{cases}$$

superscript T indicates transposition. Then

$$\frac{1}{36}\frac{d}{dt}\left[4(L_N + 4I_N + L_N^T)U_1 + (L_N + 4I_N + L_N^T)U_2\right] = -\frac{1}{2h}\left[\alpha_y(-L_N + L_N^T)U_1 + (\mu L_N + \alpha_x I_N + \theta L_N^T)U_2\right] - \frac{1}{2h}\left[v_0 + w_1 + x_2 - (\theta L_N + \alpha_x I_N + \mu L_N^T)U_0\right] - \frac{1}{36}\frac{d}{dt}\left[(L_N + 4I_N + L_N^T)U_0 + y_0 + 4y_1 + y_2\right]$$

$$(22)\quad \frac{1}{36}\frac{d}{dt}\left[(L_N + 4I_N + L_N^T)U_{m-1} + 4(L_N + 4I_N + L_N^T)U_m + (L_N + 4I_N + L_N^T)U_{m+1}\right] =$$

$$-\frac{1}{2h}\left[-(\theta L_N + \alpha_x I_N + \mu L_N^T)U_{m-1} + \alpha_y(-L_N + L_N^T)U_m + (\mu L_N + \alpha_x I_N + \theta L_N^T)U_{m+1}\right]$$

$$-\frac{1}{2h}[v_{m-1}+w_m+x_{m+1}]-\frac{1}{36}\frac{d}{dt}[y_{m-1}+4y_m+y_{m+1}] \text{ for } 1<m<M.$$

$$\frac{1}{36}\frac{d}{dt}\left[(L_N+4I_N+L_N^T)U_{M-1}+4(L_N+4I_N+L_N^T)U_M\right]=$$

$$-\frac{1}{2h}\left[-(\theta L_N+\alpha_x I_N+\mu L_N^T)U_{M-1}+\alpha_y(-L_N+L_N^T)U_M\right]-\frac{1}{2h}\left[v_{M-1}+w_M+x_{M+1}\right.$$

$$\left.+(\mu L_N+\alpha_x I_N+\theta L_N^T)U_{M+1}\right]-\frac{1}{36}\frac{d}{dt}\left[(L_N+4I_N+L_N^T)U_{M+1}+y_{M-1}+4y_M+y_{M+1}\right].$$

where we introduced vectors in $\mathbb{R}^N$

$$v_k=[-\theta u_{k0},0,\ldots,0,-\mu u_{k,N+1}]^T, \quad k=0,1,\ldots,M-1$$

$$w_k=[-\alpha_y u_{k0},0,\ldots,0,\alpha_y u_{k,N+1}]^T, \quad k=1,2,\ldots,M$$

$$x_k=[\mu u_{k0},0,\ldots,0.\ \theta u_{k,N+1}]^T, \quad k=2,3,\ldots,M+1$$

$$y_k=[u_{k0},0,\ldots,0,u_{,N+1}]^T, \quad k=0,1,\ldots,M+1.$$

We let

$$T_N=L_N+4I_N+L_N^T=\begin{bmatrix}4&1&&&\\1&4&1&&\bigcirc\\&\ddots&\ddots&\ddots&\\&\bigcirc&&\ddots&1\\&&&1&4\end{bmatrix}, \quad \alpha_N=\alpha_y(-L_N+L_N^T)=\begin{bmatrix}0&\alpha_y&&&\\-\alpha_y&0&&&\bigcirc\\&\ddots&\ddots&\ddots&\\&\bigcirc&&\ddots&\alpha_y\\&&&-\alpha_y&0\end{bmatrix}$$

$$\beta_N=\mu L_N+\alpha_x I_N+\theta L_N^T=\begin{bmatrix}\alpha_x&\theta&&&\\\mu&\alpha_x&&&\bigcirc\\&\ddots&\ddots&\ddots&\\&\bigcirc&&\ddots&\theta\\&&&\mu&\alpha_x\end{bmatrix}$$

Then the system (22) in vector and block tridiagonal matrix notation becomes

$$\frac{1}{36}\frac{d}{dt}\begin{bmatrix} 4T_N & T_N & & & \\ T_N & 4T_N & \cdot & & \bigcirc \\ & \cdot & \cdot & \cdot & \\ \bigcirc & & \cdot & \cdot & T_N \\ & & & T_N & 4T_N \end{bmatrix}\begin{bmatrix} U_1 \\ U_2 \\ \cdot \\ \cdot \\ \cdot \\ U_M \end{bmatrix} = -\frac{1}{2h}\begin{bmatrix} \alpha_N & \beta_N & & & \\ -\beta_N^T & \alpha_N & \cdot & & \bigcirc \\ & \cdot & \cdot & \cdot & \\ \bigcirc & & \cdot & \cdot & \beta_N \\ & & & -\beta_N & \alpha_N \end{bmatrix}\begin{bmatrix} U_1 \\ U_2 \\ \cdot \\ \cdot \\ U_M \end{bmatrix}$$

(23)

$$-\frac{1}{2h}\left\{\begin{bmatrix} -\beta_N^T & & & \\ & \cdot & & \bigcirc \\ & & \cdot & \\ \bigcirc & & & \cdot \\ & & & & \beta_N \end{bmatrix}\begin{bmatrix} U_0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ U_{M+1} \end{bmatrix} + \begin{bmatrix} v_0+w_1+x_2 \\ v_1+w_2+x_3 \\ \cdot \\ \cdot \\ \cdot \\ v_{M-1}+w_M+x_{M+1} \end{bmatrix}\right\} - \frac{1}{36}\frac{d}{dt}\left\{\begin{bmatrix} T_N & & & \\ & 0 & & \bigcirc \\ & & \cdot & \\ \bigcirc & & & 0 \\ & & & & T_N \end{bmatrix}\begin{bmatrix} U_0 \\ 0 \\ \cdot \\ \cdot \\ 0 \\ U_{M+1} \end{bmatrix} + \mathcal{Y}\right\}$$

where

$$\mathcal{Y} = \left[y_0+y_1+y_2,\; y_1+y_2+y_3,\; \cdots,\; y_{M-1}+y_M+y_{M+1}\right]^T$$

Here entries of matrices are $N\times N$ matrices and entries of vectors are N vectors. Note that for time independent boundary conditions the last term in this equation vanishes. Further simplification is obtained by introducing NM dimensional vectors or M dimensional compound vectors i.e vectors whose components are N dimensional vectors,

$$\mathcal{U} = \begin{bmatrix} U_1 \\ U_2 \\ \cdot \\ \cdot \\ U_M \end{bmatrix},\; \mathcal{V} = \begin{bmatrix} v_0 \\ v_1 \\ \cdot \\ \cdot \\ v_{M-1} \end{bmatrix},\; \mathcal{W} = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ w_M \end{bmatrix},\; \mathcal{X} = \begin{bmatrix} x_2 \\ x_3 \\ \cdot \\ \cdot \\ \cdot \\ x_{M+1} \end{bmatrix},\; \mathcal{Z} = \begin{bmatrix} -\beta_N^T U_0 \\ 0 \\ \cdot \\ \cdot \\ \beta_N U_{M+1} \end{bmatrix},\; \mathcal{C} = -\frac{1}{2h}\left[\mathcal{Z}+\mathcal{V}+\mathcal{W}+\mathcal{X}\right]$$

and the square matrices of order NM, $\mathcal{A}$ for the matrix on the left and $\mathcal{B}$ for the matrix on the right hand side of equation (23).

The linear system (22) or equivalently (23) can be written as

(24) $\frac{1}{36}\frac{d}{dt}\mathcal{A}\mathcal{U} = -\frac{1}{2h}\mathcal{B}\mathcal{U}+\mathcal{C}$

with the initial condition $\mathcal{U}=\mathcal{U}_0$ when t=0.

This system has a unique solution for $\frac{d\mathcal{U}}{dt}$. Letting $\mathcal{A}=A_1+A_2$ with

$$A_1 = 4\begin{bmatrix} T_N & & & O \\ & T_N & & \\ & & \ddots & \\ O & & & T_N \end{bmatrix}, \quad A_2 = \begin{bmatrix} O & T_N & & & O \\ T_N & O & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & T_N \\ O & & & T_N & O \end{bmatrix}$$

direct multiplication shows that A_1 and A_2 commute, this is a direct consequence of both being Toeplitz matrices, therefore they have the same eigenvectors. Eigenvalues of A_1, as easily verified, are

$$\lambda_k = 8\left(2+\cos\frac{k\pi}{NM+1}\right), \quad k=1,2,\ldots,NM.$$

with corresponding eigenvectors

$$X_k = \left[\sin\frac{k\pi}{NM+1}, \ldots, \sin\frac{NMk\pi}{NM+1}\right]^T.$$

Let μ_k be an eigenvalue of A_2 associated with the eigenvector X_k, since $A_2=A_2^T$ we have

$$\mu_k\langle X_k,X_k\rangle=\langle \mu_k X_k,X_k\rangle=\langle A_2X_k,X_k\rangle=\langle X_k,A_2^TX_k\rangle=\langle X_k,A_2X_k\rangle=\bar{\mu}_k\langle X_k,X_k\rangle$$

and $X_k\neq 0$ implies $\mu_k=\bar{\mu}_k$ so $\mu_k\in\mathbb{R}$. Gerschgorin theorem applied to A_2 yields $\mu_k+\lambda_k>0$, hence $\mathcal{A}$ is nonsingular [9]. It is known that $\mathcal{A}$ is unitarily similar to a diagonal matrix $\mathcal{D}$ with eigenvalues of $\mathcal{A}$, which are the sum of the eigenvalues of A_1 and A_2, are the diagonal entries. This similarity transformation is performed by taking $S=[X_1,\ldots,X_{NM}]$ i.e columns of S are eigenvectors of $\mathcal{A}$. Letting $S^{-1}\mathcal{U}=\mathbb{U}$ and multiplying (24) by S^{-1} the initial value problem reduces to

$$(25)\quad \begin{aligned} &\frac{1}{36}\frac{d}{dt}[S^{-1}\mathcal{A}S\mathbb{U}]=\frac{1}{36}\frac{d}{dt}\mathcal{D}\mathbb{U}=-\frac{1}{2h}S^{-1}\mathcal{B}S\mathbb{U}+S^{-1}C \\ &S^{-1}\mathcal{U}_0=\mathbb{U} \end{aligned}$$

Note that one does not need to compute S^{-1}, since $S^{-1}=S^T$.

For the solution of (25) one step methods such as Runge-Kutta method can be used. Also a large number of multistep methods, implicit or explicit in time (predictorcorrector), o ce a starting procedure is realized by a one step method, are available, and their stability theory is well understood and detailed treatment can be found in [10] , [11] .

To show that the finite differencing scheme is conservative, we must show that the eigenvalues $\lambda(\omega,\alpha)$ of Galerkin difference operators in (21) belonging to eigenvectors $\exp i[\omega_x x_m+\omega_y y_n]$, are purely imaginary where $\omega_x=\omega\cos\alpha$ and $\omega_y=\omega\sin\alpha$. Substituting $u_{mn}(t)=a_\omega(t)\exp i[\omega_x x_m+\omega_y y_n]$ in (21) after some manipulation yields for the left hand side

$$\begin{aligned} L.H.S=\frac{1}{36}a'_\omega(t)e^{i[\omega_x x_m+\omega_y y_n]}[&e^{-ih(\omega_x+\omega_y)}+4e^{-ih\omega_y}+e^{ih(\omega_x-\omega_y)}+4e^{-ih\omega_x}+16+4e^{ih\omega_x}\\ &+e^{-ih(\omega_x-\omega_y)}+4e^{ih\omega_y}+e^{ih(\omega_x+\omega_y)}]=\frac{1}{9}a'_\omega(t)\,e^{i[\omega_x x_m+\omega_y y_n]}[2+\cos\omega_x h][2+\cos\omega_y h], \end{aligned}$$

and for the right hand side

$$\begin{aligned} R.H.S=-\frac{1}{2h}a_\omega(t)e^{i[\omega_x x_m+\omega_y y_n]}[&\theta(e^{ih(\omega_x+\omega_y)}-e^{-ih(\omega_x+\omega_y)})+\alpha_y(e^{ih\omega_y}-e^{-ih\omega_y})\\ &+\alpha_x(e^{ih\omega_x}-e^{-ih\omega_x})+\mu(e^{ih(\omega_x-\omega_y)}-e^{-ih(\omega_x-\omega_y)})]=-\frac{i}{3h}a_\omega(t)\,e^{i[\omega_x x_m+\omega_y y_n]}\\ &[C_x\sin\omega_x h\,(2+\cos\omega_y h)+C_y\sin\omega_y h\,(2+\cos\omega_x h)] \end{aligned}$$

Hence

$$a'_\omega(t)=a_\omega(t)\,\lambda(\omega,\alpha)$$

where

$$(26)\quad \lambda(\omega,\alpha)=-i\,\omega c\left[\cos^2\alpha\,\frac{\sin\omega_x h}{\omega_x h}\,\frac{1}{\frac{2}{3}+\frac{1}{3}\cos\omega_x h}+\sin^2\alpha\,\frac{\sin\omega_y h}{\omega_y h}\,\frac{1}{\frac{2}{3}+\frac{1}{3}\cos\omega_y h}\right]$$

which is imaginary. Setting $\omega\tilde{C}(\omega,\alpha)=-\operatorname{Im}\lambda(\omega,\alpha)$ we find the numerical solution

$$u_{mn}(t)=a_\omega(0)\exp i[\omega_x x_m+\omega_y y_n-\omega\tilde{C}(\omega,\alpha)t].$$

The discrepancy between $\tilde{C}(\omega,\alpha)$ and C or more precisely the order of zero of $\tilde{C}(\omega,\alpha)-C$ about $\omega h=0$ is the the order of accuracy of the semidiscrete method. To show that this method is of order four, we expand $\tilde{C}(\omega,\alpha)$ in a Taylor series

and a straightforward computation shows that

$$(27)\quad \tilde{C}(\omega,\alpha)-C=-\frac{C(\cos^6\alpha+\sin^6\alpha)}{180}(\omega h)^4+\theta\left((\omega_x h)^6+(\omega_y h)^6\right).$$

To justify the heuristic argument presented earlier we may assume that this spurious oscillations are rapresented by small perturbations in ω , and replace ω by $\omega+\varepsilon$ in trial solutions

$$(28)\quad u(x,t)=a_\varepsilon e^{i(\omega+\varepsilon)[x-\tilde{c}(\omega+\varepsilon)t]}$$

Expanding $\tilde{c}(\omega+\varepsilon)$ in a Taylor series about $\varepsilon=0$ and retaining only the linear terms we obtain

$$\tilde{c}(\omega+\varepsilon)\simeq\tilde{c}(\omega)+\varepsilon\tilde{c}'(\omega)$$

Since $\omega\gg\varepsilon$, terms of order ε^2 can be neglected and introducing group velocity

$$(29)\quad g(\omega)=\frac{d}{d\omega}(\omega\tilde{c}(\omega))$$

(28) can be written as

$$(30)\quad u(x,t)=a_\varepsilon e^{i\omega(x-\tilde{c}(\omega)t)}\; e^{i\varepsilon(x-g(\omega)t)}$$

Straightforward computation shows that eigenvalues of (7) corresponding to eigenvectors $\{e^{i\omega x_n}\}$, are

$$\lambda(\omega)=\frac{-ic}{1-\alpha+\alpha\cos\omega h}\;\frac{\sin\omega h}{h}$$

and therefore

$$\tilde{c}(\omega)=\frac{c}{1-\alpha+\alpha\cos\omega h}\;\frac{\sin\omega h}{\omega h}$$

Using (29), $g(\omega)$ is easily computed as

$$(31)\quad g(\omega)=c\,\frac{\alpha+(1-\alpha)\cos\omega h}{(1-\alpha+\alpha\cos\omega h)^2}$$

Due to the discretization of the domain of the equation, the group velocity corresponding to 2h wavelenght, from equation (31) is

$$(32)\quad g\left(\frac{\pi}{h}\right)=-\frac{c}{1-2\alpha}$$

which is the same as depicted in (17) and for $\alpha=0$ in (15).

To obtain estimates on local and global error of numerical solution we recall the definitions. [5] p.43, [13]. We say an infinite series $\sum u_k$ is (C,1) summable if $\lim_{n\to\infty}\sigma_n=\lim_{n\to\infty}\frac{S_0+S_1+\cdots+S_n}{n+1}=S$ exists where $S_k=\sum_{\ell=0}^{k}u_\ell$, in this case we write $\sum u_k=S$ (C,1) sense.

An infinite series $\sum u_k$ is said tobe summable by Abel's method (some say Poisson's) or simply A-summable to s, if $\sum u_k r^k$ is convergent for $|r|<1$ and $\lim_{r\to 1}\sum u_k r^k=\lim_{r\to 1}(1-r)\sum S_k r^k=S$ where S_k is defined above.

We need two results, the first is that the series

(33) $\frac{1}{2}+\sum_{n=1}^{\infty}\cos nx$

is (C,1) and also A-summable to zero.

It is known that, [5] p.20

$$S_n=\frac{1}{2}+\sum_{k=1}^{n}\cos kx=\frac{1}{2}\,\frac{\sin(n+\frac{1}{2})x}{\sin\frac{x}{2}}\qquad x\neq 2\ell\pi$$

and from the trigonometric identity

$$2\sin\frac{x}{2}\sin(k+\tfrac{1}{2})x=\cos kx-\cos(k+1)x$$

it follows that

$$\sum_{k=0}^{n-1}\sin(k+\tfrac{1}{2})x=\frac{\sin^2\frac{nx}{2}}{\sin\frac{x}{2}}$$

Thus

$$\sigma_n=\frac{1}{n+1}\sum_{k=0}^{n}S_k=\frac{1}{n+1}\,\frac{1}{2\sin\frac{x}{2}}\sum_{k=0}^{n}\sin(k+\tfrac{1}{2})x=\frac{1}{n+1}\,\frac{\sin^2\frac{n+1}{2}x}{\sin^2\frac{x}{2}}\qquad x\neq 2\ell\pi$$

and $\lim_{n\to\infty}\sigma_n=0$. To show A-summability, recall the Poisson's formula [5] , p.61;

$$\frac{1}{2}+\sum_{n=1}^{\infty}r^n\cos nx=\frac{1-r^2}{2(1-2r\cos x+r^2)}\qquad |r|<1\quad ,$$

Letting $r\to 1$ we see that the assertion is true.

The second result is that

(34) $\sum_{n=1}^{\infty}\sin nx$

is (C,1) and also A-summable to $\frac{1}{2}\cot\frac{x}{2}$

It is known that [5], p.21;

$$S_n=\sum_{k=1}^{n}\sin kx=\frac{1}{2}\cot\frac{x}{2}-\frac{1}{2}\,\frac{\cos(n+\frac{1}{2})x}{\sin\frac{x}{2}}$$

Using the trigonometric identity

$$2\sin\frac{x}{2}\left[\cos\frac{3x}{2}+\cos\frac{5x}{2}+\cdots+\cos(n+\tfrac{1}{2})x\right]=\sin(n+1)x-\sin x$$

we find

$$\sigma_n=\frac{S_1+S_2+\cdots S_n}{n+1}=\frac{n}{n+1}\,\frac{1}{2}\cot\frac{x}{2}-\frac{1}{4\sin^2\frac{x}{2}}\,\frac{\sin(n+1)x-\sin x}{n+1}$$

and the result follows by letting $n\to\infty$

To show A-summability we use Poisson's formula

$$\sum_{n=1}^{\infty} r^n \sin nx = \frac{r \sin x}{1-2r\cos x + r^2}$$

and let $r \to 1$.

We now estimate the $\mathcal{L}_2$ norm of the global error. As a direct consequence of Parseval's identity, it is known that Fourier transform is an isometric isomorphism between the Hilbert spaces involved [16] p.51-52, [15] p.25. Therefore it suffices to compute the $\mathcal{L}_2$ norm of the Fourier transform of the error. To simulate the shock, we let the initial condition to be the step function

$$(35)\quad U(x,o) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Without loss of generality we may assume that the discrete Fourier transform of the net initial condition $u_n(o)$ is equal to the Fourier transform of (35), and we obtain

$$(36)\quad \hat{U}(\omega,o) = \hat{u}_n(\omega,o) = h\sum_{-\infty}^{\infty} u_n(o)\, e^{-i\omega x_n} = h\sum_{n=0}^{\infty} e^{-i\omega n h} = h\left[1 + \sum_{1}^{\infty} \cos \omega n h - i\sum_{1}^{\infty} \sin \omega n h\right]$$

It follows from the proofs of statements concerning equations (33) and (34) that series on the right of equation (36) is (C,1) and hence A-summable to

$$(37)\quad \hat{U}(\omega,o) = \hat{u}_n(\omega,o) = \frac{h e^{\frac{i\omega h}{2}}}{2i \sin \frac{\omega h}{2}}$$

From equation (1), Fourier transform of the exact solution is easily computed

$$\hat{U}(\omega,t) = \hat{U}(\omega,o)\, e^{-ic\omega t}$$

$u_n(t)$ is the solution of semidiscrete equation, for simplicity we assume K_h to be the identity operator, taking the discrete Fourier transform of the semidiscrete equation and solving the resulting differential equation one obtains

$$\hat{u}(\omega,t) = \hat{u}(\omega,o)\, e^{\lambda(\omega)t}$$

For conservative schemes $\lambda(\omega) = -i\omega\tilde{c}(\omega)$, therefore the $\mathcal{L}_2$ norm of the global error is

$$\|E\|_2^2 = \frac{1}{2\pi}\int_{-\frac{\pi}{h}}^{\frac{\pi}{h}} |\hat{u}(\omega,o)|^2 \left| e^{-i\omega\tilde{c}(\omega)t} - e^{-i\omega c t}\right|^2 d\omega$$

Introducing dimensionless variables $\tau = \frac{tc}{h}$ and $y = \omega h$, a straightforward computation shows that

$$\|E\|_2^2 = \frac{h}{\pi}\int_0^{\pi} \frac{\sin^2 \frac{y\tau}{c}\left(\tilde{c}\left(\frac{y}{h}\right) - c\right)}{\sin^2 \frac{y}{2}}\, dy .$$

CONCLUSION

The semidiscrete method proposed here has a reasonable Courant number and a fourth order accuracy. Results are theoretically conclusive. Computational evidence for detailed comparison of this method with conventional methods will await our numerical experiments.

The measure of oscillations in the numerical solution, in a neighborhood of sharp changes is the pointwise error. We were able to show with a lenghty argument, although there are some gaps in details of proofs, that maxima of the difference between the exact and the numerical solutions continually diminish and minima continually increase in an interval of lenght 4h on each side of the sharp gradient change. Numerical solution is approximately 0.28h larger in the upstream direction.

ACKNOWLEDGEMENT

I wish to express my sincere appreciation to Dr. W. Goodrich for providing me the encouragement and this gratifying experience. I would also like to acknowledge the pleasant working environment provided for NASA at Johnson Space Center.

REFERENCES

1. Babuska, I. and Aziz, A. K.: Survey Lectures on the Mathematical Foundations of the Finite Elements Method with Applications to Partial Differential Equations Academic Press, NY 1972.

2. Mitchell, A. R. and Wait, R.: The Finite Element Method in Partial Differential Equations, Wiley, 1977.

3. Thomee, V.: Convergence Estimates for Semidiscrete Galerkin Methods for Variable Coefficient Initial Value Problems. Lecture notes in Math., 333, Springer, Berlin, 1973, pp. 243-262.

4. Layton, W. J.: Stable Galerkin Methods for Hyperbolic Systems. SIAM J. Numerical Analysis, Vol. 20, April 1983, pp. 221-233.

5. Zygmund, A.: Trigonometrical Series. Dover, 1955.

6. Hyman, J. M.: A Method of Lines Approach to the Numerical Solution of Conservation Laws. Advances in Computer Methods for Partial Differential Equations. IMACS, New Brunswick, NJ, 1979.

7. Lax, P. D.: Hyperbolic Systems of Conservation Laws II. Comm. on Pure and Applied Math., Vol. 7, 1959, pp. 159-193.

8. Morawetz, C. S.: Notes on Time Decay and Scattering for Some Hypersolic Problems. SIAM Regional Conference Series in Applied Mathematics.

9. Varga, R. S.: Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1962.

10. Dahlquist, G.: Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations. Inaugural Dissertation. Appsala, Sweden, 1958.

11. Henrici, P.: Error Propagation for Difference Methods. SIAM Series in Applied Mathematics. John Wiley, 1963.

12. Vichnevetsky, R. and Bowles, J. B.: Fouries Analysis of Numerical Approximations of Hyperbolic Equations. SIAM Studies in Applied Mathematics. 1982.

13. Hardy, G. H. and Rogosinski, W. W.: Fourier Series. Cambridge. 1968.

14. Whitham, G. D.: Linear and Nonlinear Waves, Wiley. 1974.

15. Sneddon, I. N.: Fourier Transforms. McGraw Hill. 1951.

16. Richtmyer, R. D.: Difference Methods for Initial Value Problems. Interscience. 1962.