ICASE-84-8

NASA Contractor Report 172302

ICASE

SPECTRAL METHODS IN TIME FOR HYPERBOLIC EQUATIONS

Hillel Tal-Ezer

Contract No. NAS1-17070
February 1984

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING
NASA Langley Research Center, Hampton, Virginia 23665

Operated by the Universities Space Research Association

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23665

9 1 1 RN/NASA-CR-172302
DISPLAY 09/2/1
84N20274*# ISSUE 10 PAGE 1568 CATEGORY 64 RPT#: NASA-CR-172302
ICASE-84-8 NAS 1.26:172302 CNT#: NAS1-17070 DAJA38-80-C-0032 84/02/00
35 PAGES UNCLASSIFIED DOCUMENT
UTTL: Spectral methods in time for hyperbolic equations TLSP: Final Report
AUTH: A/TAL-EZER, H. PAA: A/(Tel-Aviv Univ.)
CORP: National Aeronautics and Space Administration. Langley Research Center, Hampton, Va. AVAIL.NTIS SAP: HC A03/MF A01
MAJS: /*ALGORITHMS/*HYPERBOLIC FUNCTIONS/*SPECTRUM ANALYSIS
MINS: / NUMERICAL ANALYSIS/ PERIODIC FUNCTIONS
ABA: Author
ABS: A pseudospectral numerical scheme for solving linear, periodic, hyperbolic problems is described. It has infinite accuracy both in time and in space. The high accuracy in time is achieved without increasing the computational work and memory space which is needed for a regular, one step explicit scheme. The algorithm is shown to be optimal in the sense that among all the explicit algorithms of a certain class it requires the least amount of work to achieve a certain given resolution. The class of algorithms referred to consists of all explicit schemes which may be represented as a polynomial in the spatial operator.

ENTER:

SPECTRAL METHODS IN TIME FOR HYPERBOLIC EQUATIONS

Hillel Tal-Ezer
School of Mathematical Sciences
Tel-Aviv University, Tel-Aviv, Israel

Abstract

A pseudospectral numerical scheme for solving linear, periodic, hyperbolic problems is described. It has infinite accuracy both in time and in space. The high accuracy in time is achieved without increasing the computational work and memory space which is needed for a regular, one step explicit scheme. The algorithm is shown to be optimal in the sense that among all the explicit algorithms of a certain class it requires the least amount of work to achieve a certain given resolution. The class of algorithms referred to consists of all explicit schemes which may be represented as a polynomial in the spatial operator.

Research supported in part by the U.S. Army Research and Standardization Group (Europe) under Contract DAJA 38-80-C-0032 and in part by the National Aeronautics and Space Administration under NASA Contract No. NAS1-17070 while the author was in residence at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA 23665.

1. Introduction

In recent years, it has been shown that spectral methods can provide a very useful tool for the solution of time dependent partial differential equations. A standard scheme uses spectral methods to approximate the space derivatives and a finite difference approach to march the solution in time. This tactic results in an unbalanced scheme; it has infinite accuracy in space and finite accuracy in time. It is obvious that the overall accuracy is influenced strongly by the relatively poor approximation of the time derivative.

In this paper we present an alternative approach that also yields spectral accuracy in time and is optimal in terms of efficiency.

The finite difference approach for the time discretization of the P.D.E. is discussed in Section 2. In Section 3 we present the new approach for marching the solution in time in order to get an overall infinite accuracy.

The method presented in Section 3 is based on expanding the evolution operator by orthogonal polynomials. In Sections 4 and 5 we discuss the resolution and stability properties of the method. The scheme is compared to the leap-frog type schemes in order to clarify its properties. In Section 6 we give a proof of the infinite accuracy of our approach. Section 7 presents the algorithm in detail and in Section 8 we demonstrate its validity for the variable coefficients case. Section 9 concludes the discussion by presenting numerical results which confirm the theoretical results developed in the previous sections.

2. Finite Difference Approach

Consider the differential equation

$$U_t - GU = 0 \qquad 0 \leq x \leq 2\pi$$

$$U(x,0) = U^0(x) \tag{2.1}$$

where G is a linear spatial differential operator. We assume that the coefficients of the derivatives appearing in G are time independent and 2π - periodic. Suppose further that (2.1) is discretized in space by using the pseudospectral Fourier method, [2], [4]. This involves seeking a trigonometric polynomial $U_N(x)$ of degree N that satisfies

$$\frac{\partial U_N}{\partial t} - R_N G P_N U_N = 0$$

$$U_N(x,0) = P_N U^0(x) \tag{2.2}$$

where for any function $f(x)$, $P_N f(x)$ is its trigonometric interpolant at the collocation points

$$X_j = j\frac{\pi}{N} \qquad j = 0,1,\cdots,2N-1;$$

more precisely,

$$P_N f(x) = \sum_{k=-N}^{N} a_K e^{ikx} \tag{2.3}$$

where

$$a_k = \frac{1}{2N} \sum_{j=0}^{2N-1} f(x_j) e^{-ikx_j} \quad .$$

The solution of (2.2) is given by

$$U_N(x,t) = \exp\left(tP_N G P_N\right) U^o(x). \tag{2.4}$$

Except for very simple operators G, it is impractical to construct the exponential matrix $\exp\left(tP_N G P_N\right)$ explicity. Usually an approximation to the exponent is used. Most frequently an explicit or implicit finite difference scheme is used to march the solution over a time step Δt. All these algorithms are based on a Taylor expansion of the evolution operator $\exp\left(tP_N G P_N\right)$. Essentially, the scalar function e^z is expanded either as a Taylor series of the for $\sum_{\ell=0}^{m} z^\ell/\ell!$ or by a Pade approximation

$$e^z = \frac{\sum_{\ell=0}^{P} b_\ell z^\ell}{\sum_{\ell=0}^{q} c_\ell z^\ell} \tag{2.5}$$

where b_ℓ, c_ℓ, are so chosen that the expansion of the right side of (2.5) agrees with the Taylor expansion and then z is replaced by the matrix $\Delta t\left(P_N G P_N\right)$.

For example, in the case of the modified Euler scheme one advances from the time leve n to $n + 1$ by using

$$I + \Delta t P_N G P_N + \frac{\Delta t^2}{2}\left(P_N G P_N\right)^2$$

to approximate $\exp\left[\Delta t\left(P_N G P_N\right)\right]$. Let V^n be the approximation to $U_N(n\cdot\Delta t)$. Then

$$V^{n+1} = \left[I+\Delta t\left(P_N G P_N\right)+ \frac{\Delta t^2}{2}\left(P_N G P_N\right)^2\right]V^n \tag{2.6}$$

or

$$V^n = \left(I+\Delta t G_N+ \frac{\Delta t^2}{2} G_N^2\right)^n U_N(0) \tag{2.7}$$

where

$$G_N = P_N G P_N. \tag{2.8}$$

Equation (2.7) can be rewritten as

$$V^n = \left(\sum_{k=0}^{2n} \beta_k \left(G_N \Delta t\right)^k\right) U_N(0) \tag{2.9}$$

or

$$V^n = \left(\sum_{k=0}^{2n} \overline{\beta}_k \left(G_N t\right)^k\right) U_N(0)$$

where

$$\overline{\beta}_k = \frac{\beta_k}{n^k}.$$

These types of approximations result in a limitation on the allowable time step Δt since Taylor expansion possesses high accuracy for small Δt and this accuracy deteriorates rapidly as Δt increases.

For problems in which the solution changes over a time scale which is comparable to the spatial scale, as in hyperbolic problems, the limitation on Δt can make the scheme impractical. We therefore explore other possible expansions that do not suffer from this drawback. A natural candidate is an expansion based on orthogonal polynomials. We thus restrict our discussion to polynomial schemes, i.e., to schemes that employ an algorithm which approximates the numerical solution at time level t in the following way

$$V(x,t) = \sum_{k=0}^{m} \alpha_k (G_N t)^k \; V^o(x) \quad .$$

The modified Euler algorithm (2.7) is an example of such a scheme, as well as are most explicit methods.

3. Orthogonal Polynomial Approach

We start by explaining how the new method is constructed in the case of the simple hyperbolic equation

$$\begin{aligned} & U_t - aU_x = 0 \\ & \qquad\qquad\qquad\qquad 0 \leq x \leq \partial\pi \\ & U(x,0) = U_0(x) \\ & U(2\pi,t) = U(0,t) \qquad t \geq 0 \end{aligned} \tag{3.1}$$

where a is constant.

The semi-discrete pseudospectral Fourier method can be written in the form

$$\begin{aligned} & \frac{\partial \overline{V}}{\partial t} = G_N \overline{V} \\ & \overline{V}(t=0) = \overline{V}^o \end{aligned} \tag{3.2}$$

where $\overline{V}(t)$ is the column vector

$$(V(x_0),\ V(x_1),\cdots,V(x_{2N-1}))^T$$

and

$$\overline{V}^{o} = \left(U_0(x_0),\cdots,U_0(x_{2N-1})\right)^T \quad .$$

The $2N \times 2N$ matrix G_N is given by

$$(3.3) \qquad \left(G_N\right)_{jk} = \begin{cases} \frac{a}{2}(-1)^{j+k}\operatorname{ctg}\frac{x_j - x_k}{2} & j \neq k \\ 0 & j = k \ . \end{cases}$$

(In practice $G_N\overline{V}$ is calculated using two fast Fourier transforms.) G_N is a skew-symmetric matrix and therefore has a complete set of $2N$ eigenvectors which will be deonted by $\overline{\omega}_k$ $k = 1,2,\cdots 2N$. Let

$$\overline{V}^{o} = \sum_{k=1}^{2N} b_k \overline{\omega}_k$$

then

$$(3.4) \qquad e^{G_N t}\overline{V}^{o} = \sum_{k=1}^{2N} b_k e^{G_N t}\overline{\omega}_k = \sum_{k=1}^{2N} b_k e^{\lambda_k t}\overline{\omega}_k$$

where λ_k are the eigenvalues of G_N corresponding to $\overline{\omega}_k$. In our case λ_k are purely imaginery. Let $H_m(G_N t)$ be a polynomial approximation of $\exp(G_N t)$ of the form

$$H_m(G_N t) = \sum_{\ell=0}^{m} \alpha_\ell (G_N t)^\ell \quad ;$$

then

$$H_m(G_N t)\overline{V}^{o} = \sum_{k=1}^{2N} b_k H_m(\lambda_k t)\overline{\omega}_k$$

and therefore

$$||[e^{G_N t}-H_m(G_N t)]\overline{v}^o||^2 = \sum_{k=1}^{2N} |b_k|^2 |e^{\lambda_k t}-H_m(\lambda_k t)|^2 \quad ,$$

if b_k are arbitrary (see however the remark at the end of this section),

$$||e^{G_N t}-H_m(G_N t)||^2 = \max_k |e^{\lambda_k t}-H_m(\lambda_k t)|^2 \leq \max_z |e^z-H_m(z)|^2 \tag{3.5}$$

where

$$z \in [iat(N-1),\ iat(N-1)]$$

We therefore seek a polynomial approximation with real coefficients to the function e^z that will minimize the expression on the R.H.S. of (3.5).

Define

$$R = |at(N-1)| \tag{3.6}$$

$$\theta = -iz/R \quad (|\theta| \leq 1) \quad . \tag{3.7}$$

Then

$$|e^z-H_m(z)|^2 = |e^{i\theta R}-H_m(i\theta R)|^2 =$$

$$|\cos(\theta R)-H_m^R(\theta R)|^2 + |\sin(\theta R)-H_m^I(\theta R)|^2$$

where

$$H_m(i\theta R) = H_m^R(\theta R) + iH_m^I(\theta R) \tag{3.8}$$

(H_m^R, H_m^I are polynomials with real coefficients.)

The polynomials that minimize (3.8) are the "best approximation" to $\cos(\theta R)$ and $\sin(\theta R)$. It is known that Chebyshev polynomials provide an approximation which is "almost" as good as the "best approximation". In fact, we can quote the following result [7].

Theorem: Suppose that $f \in c[-1,1]$ and $S_n(f) = \|f-q_n^*\|_\infty$ where q_n^* is the least-square approximation with respect to the weight function $(1-z^2)^{-1/2}$ then

$$S_n(f) < \left(4+\frac{4}{\pi^2}\log n\right)E_n(f)$$

($E_n(f) = \|f-q_n\|_\infty$, q_n is the best polynomial approximation.)

It follows that the improved accuracy of the best polynomial approximation does not make up for the added computational complexity. Taking $H_m^R(\theta R)$, $H_M^I(\theta R)$ as the Chebyshev polynomial approximations to the trigonometric functions we have [8]

$$H_m^R(\theta R) = J_o(R) + 2\sum_{k=1}^{\infty}(-1)^k J_{2k}(R)T_{2k}(\theta)$$

(3.9)

$$H_m^I(\theta R) = 2\sum_{k=1}^{\infty}(-1)^k J_{2k+1}(R)T_{2k+1}(\theta)$$

where $J_k(R)$ is Bessel function of order K. Hence

$$H_m(i\theta R) = \sum_{k=0}^{m}(i)^k c_k J_k(R)T_k(\theta) \tag{3.10}$$

$c_0 = 1$, $c_k = 2 \quad k \geq 1$.

Since (3.7) we have

$$T_k(\theta) = T_k(-iw) \quad w = z/R \quad w \in [-i,i] \tag{3.11}$$

Define

$$Q_k(w) = (i)^k T_k(-iw) \quad ; \tag{3.12}$$

using the recurrence relation satisfied by Chebyshev polynomials

(3.13) $$T_{k+1}(x) = 2xT_{k-1}(x) - T_{k-1}(x) \qquad T_0(x) = 1, \quad T_1(x) = x$$

it is easily verified that $Q_k(w)$ satisfies the following recurrence relation

(3.14) $$Q_{k+1}(w) = 2wQ_k(w) + Q_{k-1}(w)$$
$$Q_0(w) = 1, \quad Q_1(w) = w \quad .$$

Thus, Q_k's are polynomials in z/R with real coefficients so that

(3.15) $$H_m(z) = \sum_{k=0}^{m} c_k J_k(R) Q_k(z/R)$$

which is the desired approximation.

Remark 1: The polynomials Q_k are the imaginary analog of Chebyshev polynomials. They are orthogonal on the interval $[-i,i]$ with respect to the following inner product

(3.16) $$\langle f,g \rangle = -i\int_{-i}^{i} f(w)g(w)\left(1-|w|^2\right)^{-1/2}$$

Remark 2: It is apparent from (3.5) that in using the maximum norm we did not take into account the fact that the b_k's are decreasing. To do so requires us to consider the larger set of Gegenbauer polynomials. When the degree m is large it can be shown that the improvement thus achieved is negligible so that only for small values of m is the larger set relevant. A detailed analysis of the use of Gegenbauer polynomials will be carried out in a future paper which will deal with nonlinear problems.

4. Resolution in Time

Let us define first the notion of resolution. The accuracy of a polynomial approximation is defined by its asymptotic rate of convergence as m (the degree of the polynomial) tends to infinity. Denote by $[m_o,\infty)$ the interval of the asymptotic behavior, m has then to be greater or equal to m_o in order to have resolution. This is a necessary condition but not sufficient. For example, if the relative error is of order 1, the results are meaningless and we have no resolution. Therefore, we define the condition of having a meaningful resolution as one in which $m \geq m_o$ and the relative error norm is less than 10%. To be precise, assume that for $m \in [m_o,\infty)$, the minimal m which achieves this accuracy is $\bar{m}_o$, we then say that a necessary and sufficient condition for resolution is

$$m \geq \bar{m}_o \tag{4.1}$$

Applying the above definition to our case means that one has to apply the spatial operator tG_N, $\bar{m}_o$ times in order to resolve N modes of the exact solution of (3.1) at time level t.

Let us see what is $\bar{m}_o$ in the case of Chebyshev polynomials approximation.

Using the results from the previous section we have

$$e^z = \sum_{k=0}^{\infty} c_k J_k(R) Q_k(z/R) \tag{4.2}$$

It is known [10] that $J_k(R)$ converges to zero exponentially fast when k increases beyond R. It implies that the interval of asymptotic behavior is

$[m_o,\infty)$ while $m_o = [R]$. Because of the exponentially rate of convergence $\overline{m}_o$ is close to m_o, and when R is large we can consider $\overline{m}_o$ as m_o for any practical use. Thus, we obtain that in order to resolve N modes one has to use the spatial operator at least $|at(N-1)|$ times.

For comparison let us analyze the resolution qualities of the leap-frog scheme. This scheme is a typical explicit scheme which evaluates the numerical solution at the $n + 1$ time set, using data from the two previous time levels

$$\overline{V}^{n+1} = \overline{V}^{n-1} + 2\Delta t G\overline{V}^{n} \tag{4.3}$$

A straightforward eigenvalues analysis implies that there are two solutions for the amplification factor of each mode w_k.

$$\mu_1 = \Delta t\lambda_k + \sqrt{(\Delta t\lambda_k)^2 + 1}$$

$$\mu_2 = \Delta t\lambda_k - \sqrt{(\Delta t\lambda_k)^2 + 1} \tag{4.4}$$

with

$$\lambda_k = iak \tag{4.5}$$

The scheme is stable when $|ak\Delta t| \leq 1$ and we get

$$|\mu_{1,2}| = 1 \tag{4.6}$$

which means that the error of the scheme is only a phase error.

Let us assume that we choose the initial data at the first two levels in such a way that only μ_1 is relevant. Therefore

(4.7) $$\overline{V}^n = \mu_1^n \overline{V}^o$$

or

(4.8) $$\overline{V}^n = e^{in\psi}\overline{V}^o = e^{i\frac{t}{\Delta t}\psi}\overline{V}^o$$

where

(4.9) $$\psi = tg^{-1}\left[\frac{\theta}{(1-\theta^2)^{1/2}}\right] = \theta + \frac{\theta^3}{6} + O(\theta^5)$$

is the phase shift of the numerical scheme after one time step. The quantity $\theta = ak\Delta t$ is the phase shift of the exact solution after $t = \Delta t$. Hence, the phase error at time level t is

(4.10) $$\Delta\theta = \frac{t}{\Delta t}\frac{\theta^3}{6} + O(\theta^5) = \frac{tk^3a^3}{6}\Delta t^2 + O(\Delta t^4) \quad .$$

The largest mode is $w_N\left(N = \frac{\pi}{\Delta x}\right)$ so that the maximum phase error is

(4.11) $$\Delta\theta_{max} = \frac{t}{6}(a\pi)^3\frac{\Delta t^2}{\Delta x^3} + O(\Delta t^4) \quad .$$

This scheme is obviously second order in time and error E is

(4.12) $$E = |e^{i\theta} - e^{i(\theta+\Delta\theta)}| = |1-e^{i\Delta\theta}||e^{i\theta}| \quad .$$

When $\Delta\theta > \pi$, decreasing Δt would not necessarily decrease the error. Thus resolution is achieved when $\Delta\theta_{max}$ is at least less than π. Therefore since (4.11) we obtain

(4.13) $$a^3\frac{\Delta t^2}{\delta x^3} \leq \frac{6}{t}\frac{1}{\pi^2} + O(\Delta t^4)$$

or, using the notion of m_0 defined at the beginning of the section,

$$m_0 = \frac{\pi}{6^{1/2}}\left(\frac{a\cdot t}{\Delta x}\right)^{3/2} = \frac{1}{(6\pi)^{1/2}}(taN)^{3/2} + O(\Delta t^4) \tag{4.14}$$

and the sufficient condition for 10% error results in

$$|\Delta\theta| \leq 10^{-1} \quad . \tag{4.15}$$

Hence, using (4.10) we have

$$\bar{m}_0 = \left(\frac{5}{3}\right)^{1/2}(taN)^{3/2} + O(\Delta t^4) \quad .$$

Thus we have obtained the following result: in order to get a resolution of N modes by the leap-frog type scheme one has to operate with tG_N at least $\left(\frac{5}{3}\right)^{1/2}|(taN)|^{3/2}$ times, a requirement much more stringent than the previous result of $|ta(N-1)|$ operations for the Chevyshev approximation. For example, when $t = 2\pi$, $a = 1$ and $N = 32$ one has to apply the spatial operator, in the leap-frog case, approximately 1300 times compared to 100 times in the Chebyshev approximation case.

As stated previously, for any practical use one can identify Chebyshev polynomials approximation as the "best polynomial". Thus we conclude that from resolution point of view the scheme based on these polynomials yields the best results.

5. Stability

In the last section we discussed in detail the notion of resolution. It is clear from the above discussion that resolution implies stability. In fact, since

$$|H_m(G_N t) - e^{G_N t}| \leq \text{const} \tag{5.1}$$

$H_m(G_N t)$ must be bounded independently of m and N.
The converse is not true in general. Consider for example the leap-frog time difference scheme. It has been shown in Section 4 that in order to get resolution we need

$$m \geq \left(\frac{5}{3}\right)^{1/2} (taN)^{3/2}$$

or, equivalently,

$$\frac{a\Delta t}{\Delta t} \leq \left(\frac{3}{5m}\right)^{1/3} \frac{1}{\pi} ; \tag{5.2}$$

this is in contrast to the stability condition

$$\frac{a \cdot \Delta t}{\Delta t} \leq \frac{N}{N-1} \frac{1}{\pi} \sim \frac{1}{\pi} \tag{5.3}$$

for the leap-frog scheme [5] which allows a much larger Δt.

When a time step Δt is chosen based on the condition (5.3) rather than (5.2), one may get meaningless results in spite of the stability of the scheme. To illustrate this we solved the equation

$$\begin{aligned} u_t - u_x &= 0 \qquad && 0 \leq x \leq 2\pi \\ u_0(x) &= \cos kx \qquad && 1 \leq k \leq 7 \end{aligned} \tag{5.4}$$

numerically.

The exact solution is

(5.5) $$u(x,t) = \cos(k(x+t)).$$

Using a grid of 16 points in space assures us that the error at time level t comes solely from the time discretization.

The results are:

K	1	2	3	4	5	6	7
L2ERROR	$.9236 \times 10^{-2}$	$.767 \times 10^{-1}$	.2722	.6070	1.291	2.096	.9453

$t = 3.625$; $\Delta x = .3927$; $\Delta t = \Delta x/\pi = .1250$

These computations illustrate the above claim: meaningful results are achieved only for $1 \leq k \leq 3$, while for $4 \leq k \leq 7$ the results are meaningless despite the fact that we have used a spatial approximation that resolves all the modes exactly.

The conclusion is obvious: for nonstiff problems, the important property is resolution rather than stability. It is inefficient to use a scheme which is stable but does not resolve all the modes. A scheme with less modes and the same degree polynomial in time will produce the same results. In the leap-frog case, for example, any results achieved by using the maximum time step allowed by the stability condition could be achieved with less amount of work by using coarser grid and the same time step.

Condition (5.2) can also be written as

$$a^3 \frac{\Delta t^2}{\Delta x^3} \leq \frac{3}{5} \frac{1}{\pi^3 t} \quad ;$$

thus, it is obvious that in order to get resolution in the leap-frog case t has to be proportional to $(\Delta x)^{3/2}$ and not to Δx as required by stability. A same proportion between Δt and $\Delta x^{3/2}$ is needed to get resolution for any scheme second order in time, and it does not matter if the scheme is stable or not for Δt proportional to Δx.

For any scheme of order P the truncation error can be written as

$$E = c \cdot \Delta t^{P+1} \cdot N^{P+1} + O(\Delta t^{P+2} N^{P+2});$$

thus, the overall error in time t is

$$E = \frac{t}{\Delta t} \cdot c \cdot \Delta t^{P+1} \cdot N^{P+1} + O(\Delta t^{P+2} N^{P+2})$$

or

$$E = t \cdot c \cdot \pi^{P+1} \frac{\Delta t^P}{\Delta x^{P+1}} + O\left(\frac{\Delta t^{P+1}}{\Delta x^{P+2}}\right) \quad .$$

It follows that for a scheme of order P, Δt has to be proportional to $\Delta x^{\frac{P+1}{P}}$ in order to get resolution.

Considering the result of Lemma 2 and using the relations $\Delta t = t/m$, $\Delta x = \pi/n$ the requirement for resolution in the Chebyshev polynomial case is equivalent to

$$\frac{a \cdot \Delta t}{\Delta x} \leq \frac{N}{N-1} \frac{1}{\pi} \sim \frac{1}{\pi} \tag{5.6}$$

which means that Δt has to be proportional to Δx. Thus, this new algorithm can be regarded as a limit case of finite order schemes.

We would like to mention here another important result which follows this stability discussion. The stability condition (5.3) for the leap-frog type schemes is much more stringent than the C.F.L. condition of a similar algorithm based on finite difference approximations despite the fact that spectral approximation uses all the previous mesh points. Thus, it was regarded as an artificial condition which may be overcome by properly designed time algorithm. Observing the fact that this stability restriction is exactly (5.6) which is the resolution condition for the orthogonal polynomials algorithm we conclude that this severe stability conditon is an essential one which can not be violated. The reason is due to the fact that the spectral radius of the operator G_N is increasing. For example, in the leap-frog type algorithm the eigenvalues of G_N, using finite difference approximation in space, are

$$\lambda_k = i\frac{\sin(K\Delta x)}{\Delta x} \qquad -N \leq k \leq N \tag{5.7}$$

and the spectral radius is

$$r_{FD} = \max_k \left| i\frac{\sin(K\Delta x)}{\Delta x} \right| < \frac{1}{\Delta x} = \frac{N}{\pi} \quad .$$

On the other hand, for spectral (in space) approximation, the eigenvalues are

$$\lambda_k = ik \qquad -N \leq k \leq N \tag{5.8}$$

hence, the spectral radius is

(5.9) $$r_{SP} = \max_k |ik| = N - 1$$

so that

$$\frac{r_{SP}}{r_{FD}} = \pi\frac{N-1}{N} \sim \pi$$

6. Accuracy

According to (4.4) we have the following expression for the coefficients

$$b_k = (i)^k c_k J_k(R) \qquad c_0 = 2,\ c_k = 1 \quad k \geq 1$$

Bessel functions satisfy the following inequality [10]

(6.1) $$|J_m(m\phi)| \leq \left| \frac{\phi \cdot \exp\left(\sqrt{1-\phi^2}\right)}{1+\sqrt{1-\phi^2}} \right|^m \qquad |\phi| < 1 \quad .$$

Define

(6.2) $$\alpha \equiv \left| \frac{\phi \cdot \exp\left(\sqrt{1-\phi^2}\right)}{1+\sqrt{1-\phi^2}} \right| \qquad |\phi| < 1$$

so that

$$\frac{d\alpha}{d\phi} = \begin{cases} \dfrac{\beta e^{\beta}}{1+\beta} & \phi \geq 0 \\ -\dfrac{\beta e^{\beta}}{1+\beta} & \phi < 0 \end{cases}$$

where

$$\beta = \sqrt{1-\phi^2} \quad .$$

Thus α is monotone decreasing for $-1 < \phi \leq 0$ and increases monotonically for $0 \leq \phi < 1$; and also

(6.5) $$\alpha(0) = 0, \qquad \alpha(1) = \alpha(-1) = 1 \quad .$$

Hence it follows that

(6.6) $$0 \leq \alpha < 1.$$

In our case $m = R$ which implies

(6.7) $$\phi = R/m = ta\frac{N-1}{m} \sim ta\frac{N}{m}$$

Thus, refinement of the approximation in space and time keeping the same proportion of N/m reproduces a scheme whose error in time converges to zero as

(6.8) $$\alpha^m \xrightarrow[m \to \infty]{} 0 \qquad 0 \leq \alpha < 1$$

Hence, we have produced with a scheme which has spectral accuracy both in time and space.

We would like to point out an interesting result which can be concluded from this analysis. When T is large m is large as well since it has to be greater than R in order to have resolution. According to (6.8), once we obey this requirement of resolution ($\phi<1$) the time error is negligible, and the error of the numerical solution of the (2.1) comes only from the spatial approximation.

7. The New Algorithm

In this section we describe the actual construction of the algorithm.

In order to obtain our scheme we use the expansion (4.2)

$$e^z \approx \sum_{k=0}^{m} c_k J_k(R) Q_k(z/R)$$

$$c_0 = 1, \quad c_k = 2 \qquad k \geq 1.$$

Substituting $G_N t$ for z results in

$$\overline{V}(t) = e^{G_N t}\overline{V}^o = \sum_{k=0}^{m} c_k J_k(R)\left(Q_k(\overline{G}_N)\overline{V}^o\right) \tag{7.1}$$

where $\overline{G}_N = \frac{t}{R}G_N$. Using the recurrence relation (3.14) and (3.11) we get

$$Q_k(\overline{G}_N)\overline{V}^o = 2\overline{G}_N Q_{k-1}(\overline{G}_N)\overline{V}^o + Q_{k-2}(\overline{G}_N)\overline{V}^o$$

$$Q_o(\overline{G}_N)\overline{V}^o = V_o \qquad Q_1(\overline{G}_N)\overline{V} = \overline{G}_N\overline{V}^o \quad . \tag{7.2}$$

Using (7.2) in (7.1) enables us to compute $\overline{V}(t)$ by operating with $\overline{G}_N$ m times. It is obvious that because of the use of the recurrence relation this scheme may be regarded as a two level scheme. Therefore, it has the disadvantage of requiring extra memory. This disadvantage can't be overcome by converting (7.1) to a power series in $\overline{G}_N$ and using Horner scheme to compute $\overline{V}(t)$ because huge roundoff errors result.

A useful way to compute $\overline{V}(t)$ by a one level scheme is to calculate the roots of the polynomial

(7.3) $$p(z) = \sum_{k=0}^{m} c_k J_k(R) Q_k(z).$$

Let us assume that the roots are

(7.4) $$\lambda_1, \lambda_2, \cdots, \lambda_m \quad .$$

Since $c_k J_k(R)$ are real, every complex root appears with its conjugate. Rearranging (7.4) in such a way that the first $2p$ roots are p couples of a complex number and its conjugate we get

(7.5) $$\mu_1, \bar{\mu}_1, \cdots, \mu_p, \bar{\mu}_p, \mu_{2p+1}, \cdots, \mu_m \quad .$$

Thus, (7.3) can be written as

(7.6) $$p(z) = \alpha_o \prod_{i=1}^{p} \left(1 - \alpha_i z + \beta_i z^2\right) \prod_{i=2p+1}^{m} \left(1 - \gamma_i z\right)$$

while

$$\alpha_0 = \sum_{k=0}^{m/2} c_{2k} J_{2k}(R) \quad , \quad \beta_i = \frac{1}{|\mu_i|^2}$$

(7.7)

$$\alpha_i = \frac{2R_e \mu_i}{|\mu_i|^2} \quad , \quad \gamma_i = \frac{1}{\mu_i} \quad .$$

Hence we get

(7.8) $$p(z) = \alpha_o \prod_{i=1}^{p} \left[I - \alpha_i(\bar{G}_N) + \beta_i(\bar{G}_N)^2\right] \prod_{i=2p+1}^{m} \left[I - \gamma_i(\bar{G}_N)\right] V^o \quad .$$

This is obviously a scheme that uses the minimal memory required for an explicit scheme.

Our algorithm can be used as one step method by getting the solution at the final time t directly from the initial data. It can also be used as a marching scheme if one is interested in intermediate results. The size of the time step Δt depends only on the information one wants to get out of the numerical procedure. Δt enters instead of t in the expressions above and the parameters R,m are determined accordingly. In any case, the refinement of the algorithm is done by increasing the degree of the polynomial and not by decreasing the size of the time step.

8. Variable Coefficients

In the variable coefficient case, the operator G is

$$G = a(x)\frac{\partial}{\partial x} \quad . \tag{8.1}$$

It is approximated in the numerical procedure by the matrix G_N which is a multiplication of two matrices

$$G_N = A_N \cdot D_N \tag{8.2}$$

where A_N is a diagonal matrix whose elements are

$$(A_N)_{ij} = a(x_j)\delta_{ij} \tag{8.3}$$

and D_N is a skew-hermition matrix which approximates the derivative spectrally. It is clear from this representation that G_N is no longer normal. Nevertheless the main results of our approach are still valid. In the non-normal case, the set of eigenvectors is not always complete and one cannot use (3.4) - (3.5).

In order to justify our approach in the non-normal case we use the following definition for a function fo a matrix A [3]:

$$f(A) = \sum_{k=1}^{s} \left[f(\lambda_k) z_{k1} + f^{(1)}(\lambda_k) z_{k2} + \cdots + f^{(m_k-1)}(\lambda_k) z_{km_k} \right] \tag{8.4}$$

where λ_k are the eigenvalues of the matrix A and m_k is the multiplicity of λ_k in the minimal polynomial of A. The matrices z_{kj} are completely determined when A is given and do not depend on the choice of the function f. Expression (8.4) has a meaning when f and its required derivatives are defined on the spectrum of A. In our case f is the exponent function which is a well defined analytic function. Hence, using definition (8.4) it is also obvious that in the general case approximating the exponent matrix is equivalent to approximating the scalar exponent and its derivatives on a domain which includes all the eigenvalues of the matrix.

In the constant coefficients case the domain is

$$I = [-iaN, iaN]. \tag{8.5}$$

The following theorem implies that this result is valid also in the variable coefficients case when a(x) doesn't change sign in the interval and

(8.6) $$a = \max_x |a(x)| \qquad x \in [0,2\pi]$$

Theorem: If $a(x) > 0$ and λ_k is an eigenvalue of $A_N D_N$ then $\lambda_k \in I$. (I is defined by (8.5), (8.6).)

Proof: Define the following inner product

(8.7) $$[u,v]_{A^{-1}} = (u,A^{-1}v)$$

Then we have

(8.8) $$[u_n,(u_N)_t]_{A_N^{-1}} = [u_N,A_N D_N u_N]_{A_N^{-1}} = (u_N,D_N u_N)$$

where $(u_N,D_N u_N)$ is real; thus

(8.9) $$(u_N,D_N u_N) = (D_N u_N,u_N) = (u_N,D_N^* u_N) = -(u_N,D_N u_N).$$

Hence

(8.10) $$(u_N,D_N u_N) = 0$$

Using this result in (8.8) we get

(8.11) $$\frac{d}{dt}\left\{\frac{1}{2}[u_N,u_N]_{A_N^{-1}}\right\} = 0$$

thus

(8.12) $$\|u_N\|_{A_N^{-1}} = \text{const},$$

Assuming that w_k is the eigenvector of $A_N D_N$ corresponding to λ_k we can use it as the initial data so that

(8.13) $$u_N = e^{A_N D_N t} w_k = e^{\lambda_k t} w_k \quad ;$$

hence

(8.14) $$[u_N, u_N]_{A_N^{-1}} = [e^{\lambda_k t} w_k, e^{\lambda_k t} w_k]_{A_N^{-1}} = e^{2(R_e \lambda_k) t} [w_k, w_k]_{A_N^{-1}}$$

and according to (8.12) this is constant. This implies that

(8.15) $$R_e \lambda_k = 0$$

hence λ_k is pure imaginery. In addition

(8.16) $$\max_k |\lambda_k| \leq \|A_N D_N\| \leq \|A_N\| \ \|D_N\| = a \cdot N$$

and the proof is concluded.

The proof is essentially the same when $a(x) < 0$ in the interval.

The case when $a(x)$ changes sign is more complicated and not much is yet known. [6] gives a proof of stability for the simple case when $a(x)$ is a trigonometric polynomial of order 1. It implies that in this simple situation

(8.17) $$-\alpha \leq R_e \lambda_k \leq \alpha \qquad a > 0$$

while α doesn't depend on N. Because of (8.16) there exists an ellipse whose larger axis is $[-iaN, iaN]$ and the small one is $[-\alpha, \alpha]$ which contains the eigenvalues of $A_N D_N$. The theory of Chebyshev polynomial approximations guaranties convergence in this domain [9].

Remark Numerical experiments lead us to the assumption that (8.17) is also valid in the general case that $a(x)$ is any periodic function.

9. Numerical Results

In order to illustrate the spectral convergence of our scheme we shall consider the following scalar problem

$$u_t - a(x)u_x = 0 \qquad 0 \leq x \leq 2\pi \quad . \tag{9.1}$$

In Table 1 we take

$$a(x) = \frac{1}{2 + \cos x}, \qquad u_o(x) = \sin(2x+\sin x) \quad . \tag{9.2}$$

The exact solution to this problem is

$$u(x,t) = \sin(2x+\sin x+t) \quad . \tag{9.3}$$

The numerical solution is computed at time leve $T = 6.283$.

N = Number of mesh points in space.

M = The degree of the ploynomial approximation.

The ratio of the L_2 errors illustrates very clearly the spectral convergence of the scheme.

Table 1

N	M	L•2 Errors	Ratio
8	36	1.605×10^{-1}	9.2×10^{3}
16	72	1.740×10^{-5}	4.6×10^{7}
32	144	3.756×10^{-13}	

In Table 2 we take

(9.4) $$a(x) = \sin(x), \quad u_o(x) = \sin(x)$$

(a(x) changes sign in the interval).

The exact solution is

(9.5) $$u(x,t) = \sin\left(2\mathrm{tg}^{-1}\left(e^{t}\mathrm{tg}\frac{x}{2}\right)\right)$$

The sollution is computed at T = 1.571.

Table 2

N	M	L2 Errors	Ratio
16	18	5.968×10^{-2}	2.9×10^{1}
32	36	2.031×10^{-3}	8.7×10^{2}
64	72	2.345×10^{-6}	

Table 3 inllustrates the resolution properties of the scheme. In this case we

$$a(x) = \frac{1}{2 + \cos x}, \qquad T = 50.27, \qquad N = 16$$

Because of the high resolution in space only time errors occur. According to the theory developed previously (Lemma 2) the degree of the polynomial approximation has to be at least R. We have

$$R = at(N-1) = 50.27 \times 15 = 754.05$$

Table 3

M	L2 Error
740	1.120
750	5.981×10^{-1}
760	1.354×10^{-1}
770	1.476×10^{-2}
780	1.048×10^{-3}
840	5.391×10^{-13}

The result for M = 840 shows that, while the minimal M required for resolution is 755, increasing M by only 11% gives machine accuracy.

In Tables 4 and 5 we compare our scheme to the leap-frog scheme. The model problem is (9.1), (9.2) with N = 32. In Table 4 we compute the numerical solution at different time levels. The results clearly shows the high accuracy of the Chebyshev polynomials approach. Another interesting

phenomenon which is illustrated by the table is the following: while the leap-frog case the error increased due to accumulation this is not so in our case. In contrast it decreased geometrically. The explanation is obvious using (6.7), (6.8). Increasing t and m, keeping the same proportion between them, results in ϕ = constant and therefore (6.8) is valid.

In the following table

C.P. = Chebyshev polynomials.

L.F. = leap-frog.

Table 4

Time	M	L2 Error (L.F.)	L2 Error (C.P.)
1.571	35	5.247×10^{-4}	8.726×10^{-5}
3.142	70	1.108×10^{-3}	2.084×10^{-7}
6.283	140	2.184×10^{-2}	1.429×10^{-13}

In Table 5 we compare our scheme to the leap-frog scheme from the point of view of the amount of work needed to achieve a certain degree of accuracy. The L2 Error is computed at time level $T = 6.283$.

Table 5

L2 Error	u(L.F.)	u(C.P.)
1.0×10^{-4}	580	110
1.0×10^{-6}	3480	117
1.0×10^{-8}	56000	122

Acknowledgement

I would like to thank my advisor, Professor David Gottlieb, for his helpful guidance.

References

[1] M. Abramowitz, I. A. Stegun, Handbook of Mathematical Functions, Dover Publications, Inc., New York, 1972.

[2] P. J. Davis, Interpolation and Approximation, Dover Publications, Inc., New York, 1975.

[3] A. Erdelyi, W. Magnus, F. Oberhettinger and F. G. Tricomi, Higher Transcendental Funtions, Vol. 2, McGraw-Hill, New York, 1953.

[4] B. Forenberg, On a Fourier method for the integration of hyperbolic equations, SIAM J. Numer. Anal., Vol. 12, (1975), pp. 509.

[5] F. R. Gantmacher, The Theory of Matrices, Vol. 1k, Chelsea Publishing Company, New York, 1959.

[6] D. Gottlieb and S. Orszag, Numerical Analysis of Spcetral Methods; Theory and Applications, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM Publisher, Philadelphia, PA 1977.

[7] D. Gottlieb and E. Turkel, On time discretization for spectral methods, Stud. Appl. Math., Vol. 63, No. 1, (1980), pp. 67-86.

[8] D. Gottlieb, S. Orszag and E. Turkel, Stability of pseudospectral and finite difference methods for variable coefficient problems, Math. Comp., Vol. 37, No. 156, (1981), pp. 293-305.

[9] T. J. Rivlin, An Introduction to the Approximation of Functions, Blaisdell Publishing Company, Waltham, MA, 1969.

[10] M. A. Snyder, Chebyshev Methods in Numerical Approximation, Prentice Hall, Inc., Englewood Cliffs, N.Y., 1966.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
NASA CR-172302		

4. Title and Subtitle	5. Report Date
Spectral Methods in Time for Hyperbolic Equations	February 1984
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Hillel Tal-Ezer	84-8
	10. Work Unit No.

9. Performing Organization Name and Address	11. Contract or Grant No.
Institute for Computer Applications in Science and Engineering Mail Stop 132C, NASA Langley Research Center Hampton, VA 23665	NAS1-17070
	13. Type of Report and Period Covered
12. Sponsoring Agency Name and Address	contractor report
National Aeronautics and Space Administration Washington, D.C. 20546	14. Sponsoring Agency Code

15. Supplementary Notes

Langley Technical Monitor: Robert H. Tolson
Final Report

16. Abstract

A pseudospectral numerical scheme for solving linear, periodic, hyperbolic problems is described. It has infinite accuracy both in time and in space. The high accuracy in time is achieved without increasing the computational work and memory space which is needed for a regular, one step explicit scheme. The algorithm is shown to be optimal in the sense that among all the explicit algorithms of a certain class it requires the least amount of work to achieve a certain given resolution. The class of algorithms referred to consists of all explicit schemes which may be represented as a ploynomial in the spatial operator.

17. Key Words (Suggested by Author(s))	18. Distribution Statement
spectral methods hyperbolic equations time marching	64 Numerical Analysis Unclassified-Unlimited

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	34	A03

For sale by the National Technical Information Service, Springfield, Virginia 22161 NASA-Langley, 1984