in ``Numerical Methods for Fluid Dynamics II",
Proceedings of the 1985 Conference on Numerical Methods for Fluid Dynamics (K. W. Morton and M. J. Baines, eds.),
Clarendon Press, Oxford (1986), 129-153.

NASA Contractor Report 177974

ICASE REPORT NO. 85-38

ICASE

SPECTRAL METHODS FOR DISCONTINUOUS PROBLEMS

Saul Abarbanel

David Gottlieb

Eitan Tadmor

NASA Contract No. NAS1-17070
August 1985

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

SPECTRAL METHODS FOR DISCONTINUOUS PROBLEMS

Saul Abarbanel

David Gottlieb

Eitan Tadmor*

Tel-Aviv University, Tel-Aviv, Israel
and
Institute for Computer Applications in Science and Engineering

Abstract

We show that spectral methods yield high-order accuracy even when applied to problems with discontinuities, though not in the sense of pointwise accuracy. Two different procedures are presented which recover pointwise accurate approximations from the spectral calculations.

Research was supported in part by the National Aeronautics and Space Administration under NASA Contract No. NAS1-17070 while the authors were in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665. Additional support was also provided in part by the Air Force Office of Scientific Research under Contract No. AFOSR 83-0089 for the first and second authors, and by the National Science Foundation under Grant No. DMS85-03294 and Army Research Office under Grant No. DAAG29-85-K-0190 for the third author.

*Bat-Sheva Foundation Fellow

1. INTRODUCTION

Consider the evolution partial differentiation equation $u_t = Lu$, on a finite interval, where L is a hyperbolic operator. The solution u has a projection $P_N u$ on a finite subspace (which may for example consist of the first N modes in a Galerkin method, or N collocating points in the interval), and a numerical approximation u_N generated by some spectral method. For linear operators it is known from the Lax equivalence theorem that if the scheme is consistent and stable, then u_N approximates $P_N u$ in some appropriate norm. If u is smooth, then the theorem implies that u_N approximates the solution u in the same sense.

In practice, one looks at the point values of u_N at the grid points and takes it as an approximation to the values of the true solution u at these points. We shall call this approach the realization of the computed solution via its grid-points value. The aims of the paper are: 1) demonstrate that when u is a complicated function, this realization will not produce acceptable results; 2) to suggest different ways for the realization of the solution in such cases.

The following examples give a very clear illustration of the misleading results that may be obtained by pointwise realization.

<u>Example 1</u>

Consider the equation

$$\begin{aligned} u_t &= u_x \qquad\qquad 0 < x < 2\pi \\ u(x,0) &= u_0(x) \end{aligned} \tag{1}$$

where $u(x)$ and $u_0(x)$ are periodic functions and <u>$u_0(x)$ is a discontinuous function</u>. If we expand $u_0(x)$ in Fourier series we get

$$u_0(x) = \sum_{k=-\infty}^{\infty} a_k \, e^{ikx} \tag{2a}$$

where

$$a_k = \frac{1}{2\pi} \int_0^{2\pi} u_0(x) e^{-ikx}. \tag{2b}$$

The solution $u(x)$ is thus given by

$$u(x) = \sum_{k=-\infty}^{\infty} a_k \, e^{ikt} \, e^{ikx}.$$

Suppose that (1) is solved numerically by the Fourier-Galerkin method, namely we seek a trigonometric polynomial of the form

$$u_N(x,t) = \sum_{k=-N}^{N} b_k(t) e^{ikx}$$

that satisfies

$$\begin{aligned} &\left(\frac{\partial u_N}{\partial t} - \frac{\partial u_N}{\partial x} , \, e^{ikx} \right) = 0 \, , \qquad -N \leq k \leq N \\ &u_N(x,0) = \sum_{k=-N}^{N} a_k \, e^{ikx}. \end{aligned} \tag{3}$$

From (3) it is clear that

$$\frac{db_k(t)}{dt} = ik \; b_k(t), \qquad -N \leq k \leq N \tag{4}$$

and

$$b_k(0) = a_k$$

yielding the solution

$$b_k(t) = a_k\, e^{ikt}.$$

Therefore

$$u_N(x,t) = \sum_{k=-N}^{N} a_k\, e^{ikt}\, e^{ikx}. \tag{5}$$

Equation (5) implies that $u_N(x,t)$, obtained from the numerical solution (3), coincides with $P_N\, u(x,t)$, the Galerkin projection of u, thus yielding the best possible convergence of u_N to $P_N\, u$. However, since the Fourier series of $u(x,t)$ converges very slowly, the point values $u_N(x_j,t)$ will not approximate well $u(x_j,t)$. In general, one would witness the Gibbs phenomenon of overshoot in the neighborhood of the discontinuity and global oscillations all over the domain. In fact, even the initial approximation, $u_N(x,0)$, displays the same behavior in relation to $u_0(x)$.

In the second example we show that the same phenomenon occurs even if the numerical initial point values do approximate the true initial point values to a high degree of accuracy.

<u>Example 2</u>

Consider the equation (1) where $u_0(x)$ is the saw-tooth function

$$u_0(x,\bar{x}) = \begin{cases} Ax & x < \bar{x} \\ A(2x - \pi) & x > \bar{x} \end{cases} \tag{6}$$

for some k, $0 < k < 2N-1$, $\bar{x} = \frac{\pi}{N}\,(k+1/2)$. .

In the pseudospectral Fourier method we seek a trigonometric polynomial $v_N(x,t)$

$$v_N(x,t) = \sum_{\ell=-N}^{N} b_\ell(t)e^{i\ell x} \tag{7}$$

such that

$$\frac{\partial v_N}{\partial t} = \frac{\partial v_N}{\partial x} \qquad \text{at the points} \quad x_j = \frac{\pi j}{N}, \quad j = 0,\ldots,2N-1 \tag{8a}$$

$$v_N(x,0) = u_0(x,\bar{x}). \tag{8b}$$

Since v_N is a polynomial of degree N, (8a) implies that

$$\frac{\partial v_N}{\partial t} = \frac{\partial v_N}{\partial x} \tag{9}$$

for all x $(0,2\pi)$. Moreover, from (8b) it is clear that $v_N(x,0)$ is the (unique) trigonometric polynomial of order N that interpolates $u_0(x)$ at the points x_j, $j = 0,\ldots,2N-1$, thus

$$v_N(x,0) = \sum_{\ell=-N}^{N} a_\ell(\bar{x})e^{i\ell x} = AF_N(x,\bar{x}) \tag{10}$$

where

$$a_\ell(\bar{x}) = \frac{1}{2Nc_\ell}\sum_{j=0}^{2N-1} u_0(x_j,\bar{x})e^{-i\ell x_j}. \tag{11}$$

Performing (11) we get

$$a_0(\bar{x}) = A\,\frac{\pi}{N}\left[k - N + .5\right] \tag{12}$$

$$a_\ell(\bar{x}) = A\frac{\pi}{2Nc_\ell}\ 2\,\frac{1 - e^{\frac{-i\pi\ell}{N}(k+1)}}{1 - e^{\frac{-i\pi\ell}{N}}} + i\operatorname{ctn}\frac{\pi\ell}{N} - 1 \quad , \quad \ell \neq 0 \tag{13}$$

where

$$c_{-N} = c_N = 2, \qquad c_\ell = 1, \qquad |\ell| \neq N.$$

The numerical solution $v_N(x,t)$ of (9), (10) is

$$v_N(x,t) = v_N(x + t,0) = AF_N(x + t,\bar{x}) \tag{14}$$

and upon manipulating (12), (13) one gets

$$v_N(x,t) = AF_N(x,\bar{x} - t) + At. \tag{15}$$

The trigonometric interpolant $F_N(x,\bar{x})$ collocates $u_0(x,\bar{x})$ at the grid points x_j. However, in between the grid points it oscillates. If we read the values of $v_N(x,t)$ at the grid points, then by (14)

$$v_N(x_j,t) = AF_N\left(x_j + t,\bar{x}\right)$$

and unless $t = \frac{\pi m}{N}$ for some integer m, we will get solution that looks oscillatory. Thus, even though the initial approximation looks smooth at the grid points, when it evolves in time the oscillations will present themselves at the points x_j.

The conclusion one might draw from the above examples is that spectral methods (or any higher-order methods) are useless when applied to discontinuous function. A different approach is to look at a different realization of the numerical solution rather than the pointwise one. We will argue that high-order accurate information is contained in the numerical solution and demonstrate how that information can be extracted in such a way that accurate pointwise approximation to the true solution can be obtained.

2. INFORMATION AND HOW TO EXTRACT IT

Consider the linear equation

$$\begin{aligned} u_t &= Lu \\ u(0) &= u_0 \end{aligned} \tag{16}$$

where L is a linear hyperbolic operator with variable coefficients and u_0 is a discontinuous function. For simplicity, we will restrict ourselves to a periodic, one (space) dimensional problem though the results are more general, (see Gottlieb and Tadmor [2]). Let v be the solution of the auxiliary problem

$$\begin{aligned} v_t &= - L^* v \\ v(0) &= v_0, \end{aligned} \tag{17}$$

where v_0 is a C^∞ function. Because of the hyperbolicity of L, (17) is a well-posed problem. In Lemma 1 we quote the well-known Green's identity.

Lemma 1: Let $u(t)$ and $v(t)$ be the solutions of (16) and (17) at some level t, then

$$\big(u(t),\ v(t)\big) = \big(u_0, v_0\big). \tag{18}$$

Assume now that (16) and (17) are discretized by the Fourier-Galerkin method. That is, we seek u_N and v_N that are trigonometric polynomials of degree N such that for every k, $|k| \leq N$

$$\left(\frac{\partial u_N}{\partial t} - L\ u_N,\ e^{ikx}\right) = 0 \tag{19a}$$

$$\left(u_N(0) - u_0,\ e^{ikx}\right) = 0, \tag{19b}$$

$$\left(e^{ikx},\ \left(\frac{\partial v_N}{\partial t} + L^*\ v_N\right)\right) = 0 \tag{19c}$$

$$e^{ikx},\ \left(v_N(0) - v_0\right) = 0. \tag{19d}$$

We have also a Green identity for u_N and v_N.

Lemma 2:

$$\big(u_N(t), v_N(t)\big) = \big(u_N(0), v_N(0)\big). \tag{20}$$

Proof: Since $v_N(t)$ and $u_N(t)$ are N^{th}-order trigonometric polynomials we use (19a) and (19c) to get

$$\left(\frac{\partial u_N}{\partial t} - Lu_N,\ v_N\right) = 0$$

$$\left(u_N,\ \frac{\partial v_N}{\partial t} + L^*\ v_N\right) = 0,$$

and therefore

$$\frac{\partial}{\partial t}\left(u_N, v_N\right) = \left(Lu_N, v_N\right) - \left(u_N, L^*\ v_N\right) = 0$$

which implies (20).

We will proceed by showing the relation of the RHS of (20) to that of (18).

Lemma 3:

$$\left(u_N(0), v_N(0)\right) = \left(u_0, v_0\right) + \varepsilon_1 \tag{21}$$

where

$$|\varepsilon_1| < K\,\frac{\|v_0\|_s}{N^s} \tag{22}$$

for every s.

<u>Proof</u>: From (19b) it is clear that

$$\left(u_N(0) - u_0,\ v_N(0)\right) = 0. \tag{23}$$

Also,

$$\left|\left(u_0, v_N(0) - v_0\right)\right| < K\|u_0\|\ \|v_N(0) - v_0\|$$

and since v_0 is a C^∞ function,

$$\|v_N(0) - v_0\| \leq K \frac{\|v_0\|_s}{N^s}, \qquad \text{for every } s. \tag{24}$$

Now

$$\left(u_N(0),\ v_N(0)\right) = (u_0,v_0) + \left(u_N(0) - u_0,\ v_N(0)\right) + \left(u_0, v_N(0) - v_0\right)$$

and in view of (23) and (24),

$$\left(u_N(0),\ v_N(0)\right) = (u_0,v_0) + \varepsilon_1$$

where

$$|\varepsilon_1| \leq K \frac{\|v_0\|_s}{N^s}$$

and this proves the Lemma.

From Lemmas 1 - 3 we can conclude:

Theorem 1: Let $u(t)$ and $v(t)$ be the solutions of (16) and (17), respectively. Let $u_N(t)$ and $v_N(t)$ be the solutions of the Fourier-Galerkin approximations of (16) and (17). Then

$$\left|\left(u_N(t),v_N(t)\right) - \left(u(t),v(t)\right)\right| \leq K \frac{\|v_0\|_s}{N^s}, \qquad \text{for every } s. \tag{25}$$

The proof is an immediate consequence of (18), (20), and (21).

Assume now that the Fourier-Galerkin method described in (19c) and (19d) is stable, then $v_N(t)$ approximates $v(t)$ within spectral accuracy, that is

$$\|v_N(t) - v(t)\| = \varepsilon_2 \leq K \frac{\|v\|_s}{N^{s-1}} .$$

We can, therefore, replace $v_N(t)$ in (25) and get

$$\left(u_N(t), v(t)\right) = \left(u(t), v(t)\right) + \varepsilon$$

where ε is spectrally small. We use now the fact that every C^∞ function $v(t)$ can be obtained from some v_0 in (17). This is, in fact, one of the definitions of hyperbolicity. We can, therefore, state:

Theorem 2: Let $u(t)$ be the (nonsmooth) solution of (16) and let $u_N(t)$ be the solution of the spectral Galerkin approximation to (16). Then for any C^∞ function $v(t)$

$$\left(u_N(t), v(t)\right) = \left(u(t), v(t)\right) + \varepsilon \tag{26}$$

where ε is spectrally small.

Thus, $u_N(t)$ approximates <u>weakly</u> $u(t)$ within spectral accuracy. It is in this sense that $u_N(t)$ contains a highly accurate information about $u(t)$. We will show later how to use this information in order to obtain spectral accurate approximation to the grid-point values of $u(t)$.

We turn now to the pseudospectral Fourier case. Here we need some preprocessing of the initial data in order to prove the same result as in Theorem 2.

Theorem 3: Let $u_N(x,t)$ be a trigonometric polynomial of order N that satisfies

$$\begin{aligned} &\frac{\partial u_N}{\partial t} = Lu_N \quad \text{at } x = x_j, \quad x_j = \frac{\pi j}{N}, \quad j = 0,\cdots,2N-1 \\ &\left(u_N(0) - u_0, e^{ikx}\right) = 0, \qquad |k| \leq N, \end{aligned} \tag{27}$$

(i.e., $u_N(x,t)$ is the solution of the pseudospectral Fourier scheme, but initially $u_N(x,0)$ is obtained by the Galerkin projection).

Then for every smooth function $u(x,t)$

$$\frac{\pi}{N}\sum_{j=0}^{2N-1} u_N(x_j,t)\, v(x_j,t) = \int_0^{2\pi} u(x,t)\, v(x,t)dx + \varepsilon \tag{28}$$

where ε is spectrally small, provided that the pseudospectral approximation is stable.

<u>Proof:</u> Let v_N be the solution of the pseudospectral Fourier approximation of (17a) and let $v_N(0)$ be the Galerkin projection of v_0, that is

$$\left(v_N(0) - v_0, e^{ikx}\right) = 0, \qquad |k| \leq N. \tag{29}$$

From (27) and the analog equation for v_N, one gets

$$\frac{\pi}{N} \sum_{j=0}^{2N-1} u_N(x_j,t)\ v_N(x_j,t) = \frac{\pi}{N} \sum_{j=0}^{N} u_N(x_j,0) v_N(x_j,0). \tag{30}$$

From the exactness of the trapezoidal rule for polynomials of degree 2N, we conclude

$$\frac{\pi}{N} \sum_{j=0}^{2N-1} u_N(x_j,t)\ u_N(x_j,t) = \int_0^{2\pi} u_N(x,0)\ v_N(x,0)dx = \left(u_N(0),v_N(0)\right). \tag{31}$$

Note that the initial functions $v_N(x,0)$ and $u_N(x,0)$ are not the interpolants of u_0 and v_0 as in the usual pseudospectral methods but rather the Galerkin approximation to u_0 and v_0. We recall now Lemma 3 and equality (18) to establish (28). The proof is thus completed.

It is interesting to note the way in which the information is contained. The interpolant of u_0 looks smooth at the grid points, whereas the Galerkin approximation of u_0 looks oscillatory on the grid points. It means that in order to preserve the information one has to require initially oscillatory-looking solution. The information is preserved in the structure of the oscillations.

We will show now a way of using (26) and (28) in order to construct a better approximation to $u(x_j,t)$ then the one given by $u_N(x_j,t)$ (here $u_N(x,t)$ is given by either the Galerkin method or the pseudospectral method).

From (28) and (26) it is clear that in order to get a good approximation to $u(y,t)$ at some point y $(0,2\pi)$, we need to find a function $v_y(x,t)$ such that

$$\int_0^{2\pi} u(x,t)\ v_y(x,t)dx = u(y,t) + \varepsilon_1,$$

where ε_1 is spectrally small. By (26) we will have

$$\int_0^{2\pi} u_N(x,t)\, v_y(x,t)dx = u(y,t) + \varepsilon + \varepsilon_1 \tag{32}$$

for the Galerkin method and

$$\frac{\pi}{N} \sum_{j=0}^{2N-1} u_N(x_j,t)\, v_y(x_j,t) = u(y,t) + \varepsilon + \varepsilon_1 \tag{33}$$

for the pseudospectral method.

For conveniency we will shift the interval $[0,2\pi]$ to $[-\pi,\pi]$. Let $\rho(x)$ be a C^∞-function vanishing outside the interval $[-\pi,\pi]$ satisfying

$$\rho(0) = 1. \tag{34}$$

Let $D_p(x)$ be the Dirichlet kernel, namely

$$D_p(y) = \frac{1}{2\pi} \sum_{|k|<p} e^{ikx} = \frac{1}{2\pi} \frac{\sin(p+ \frac{1}{2})\, y}{\sin\,(y/2)}\ . \tag{35}$$

We set now

$$v_y(x) = \psi^{\theta,p}(x) = \theta^{-1}\ \rho(\theta^{-1}y)D_p(\theta^{-1}y). \tag{36}$$

One can prove (see [2]) that

$$\int_{-\pi}^{\pi} u(x)\psi^{\theta,p}(y-x)dx = u(y) + \varepsilon_2 \tag{37}$$

where ε_2 is spectrally small.

Thus, it is possible to extract accurate pointwise values from $u_N(x)$.

3. NUMERICAL RESULTS

In this section we demonstrate the efficacy of the smoothing procedure outlined above. As a test function we have chosen the piecewise C^∞-function

$$f(x) = \begin{cases} \sin \frac{x}{2} & 0 < x < \pi \\ -\sin \frac{x}{2} & \pi < x < 2\pi. \end{cases} \tag{38}$$

Denote its spectral approximation by $\hat{f}_N(x)$, and let $\tilde{f}_N(x)$ be the pseudospectral approximation to $f(x)$. It is evident from the first column of Tables I and III that $\hat{f}_N(y_\nu)$ - the spectral approximation sampled at $y_\nu = \nu\pi/N$ - do not approximate $f(y_\nu)$ within spectral accuracy. In fact, the error committed by $\hat{f}_{128}(y_\nu)$ is only half of that committed by $\hat{f}_{64}(y_\nu)$. Regarding the pseudospectral approximation, $\tilde{f}_N(x)$, it, of course, <u>collocates</u> the exact values at the sampling grid points, $\tilde{f}_N(y_\nu) = f(y_\nu)$; yet, <u>in between</u> these gridpoints, $\tilde{f}_N\left(y_{\nu+1/2} = (\nu + 1/2)\pi/N\right)$ approximate $f(y_{\nu+1/2})$ within first-order accuracy only, as shown in the first column of Tables II and IV.

In order to construct our regularization kernel in (36), we define the cut-off function $\rho(\xi) = \rho_\alpha(\xi)$ to be

$$\rho_\alpha(\xi) = \begin{cases} \exp \dfrac{\alpha\xi^2}{\xi^2 - 1} & |\xi| < 1 \\ 0 & \text{otherwise} \end{cases} , \tag{39}$$

namely, $\rho_\alpha(\xi)$ is a C^∞- function whose support is the interval $|\xi| < 1$. ψ to be used in (36) is of the form

$$\psi^{\theta,p}(y) = \frac{1}{2\pi\theta} \rho_\alpha(\theta^{-1} y) \frac{\sin (p+1/2)y/\theta}{\sin y/2\theta} . \tag{40}$$

The post-processing procedure of the spectral approximation $\hat{f}_N$ involves convoluting $\hat{f}_N$ against $\psi^{\theta,p}$, namely

$$f(x) \sim \frac{1}{2\pi\theta} \int_0^{2\pi} \hat{f}_N(y)\rho\left(\frac{x-y}{\theta}\right) \frac{\sin\ (p+1/2)(x-y)/\theta}{\sin\ (x-y)/2\theta}\, dy \tag{41}$$

where x is a fixed point of interest. (In practice, we use the trapezoidal rule to evaluate the right-hand-side of (41) taking a large number of quadrature points.)

The parameter θ was chosen as

$$\theta = \pi \cdot |x - \pi|; \tag{42}$$

this guarantees that ψ is so localized that it does not interact with regions of discontinuity.

It should be noted, in this stage, that if θ was so chosen to be the same for each x, (and not as in (42)), the formula (41) admits a simpler form; that is, if

$$\psi^{\theta,p}(y) = \sum_{k=-\infty}^{\infty} \sigma_k\, e^{iky} \tag{43}$$

then

$$f(x) \sim \sum_{k=-N}^{N} \hat{f}(k)\sigma_k\, e^{ikx}. \tag{44}$$

This procedure can be carried out efficiently in the Fourier space.

Next, we turn to the post-processing for the pseudospectral approximation $\hat{f}_N(x)$ which is simpler than (41). In fact, in this case

$$f(x) \sim \frac{2\pi}{2N} \sum_{\nu=0}^{2N-1} \tilde{f}(y_\nu)\psi^{\theta,p}(x-y_\nu). \tag{45}$$

Note that carrying out the smoothing procedure defined in (45) does not involve any extra evaluation of $\tilde{f}(y)$ in points other than y_ν, in contrast to spectral smoothing procedure in (41). As before, the parameter θ was chosen according to (42). We have yet to determine the parameters p and α. The parameter p must be equal to N^β for $0 < \beta < 1$, in order to assure infinite accuracy. (In our computations, $\beta \approx .8$.) Finally, we feel that α is problem dependent and we chose $\alpha = 10$. We have not tuned the parameters to get optimal results; further tuning may improve the quality of our filtering procedure.

In Tables I, II, III, and IV we give the results of the smoothing procedure at several points in the domain. The pointwise values are now recovered with high accuracy. The first column in each table indicates the points in which the procedure was performed. We limited ourselves to four points in the interval $(0,\pi)$ because of the symmetry of the function $f(x)$.

The second column gives either the spectral approximation $\hat{f}_N(x)$ or the pseudospectral approximation $\tilde{f}_N(x)$, $N = 128$ in Table I and II and $N = 64$ in Tables III and IV. The third column gives the smoothed results, when filtered by (41) on (45), at the same points as in column I.

The results indicate the dramatic improvement obtained by the smoothing procedure. Moreover, note that the error committed by $\tilde{f}_{128}$ (or $\hat{f}_{128}$) is better than the one committed by $\tilde{f}_{64}$ (or $\hat{f}_{64}$) only by a factor of 2 whereas after the post-processing the error improves by a factor of 10^4.

Table I. Results of smoothing of the spectral approximation of f(x), N = 128

$x_\nu = \frac{\pi\nu}{8}$ ν equals	$\lvert f(x_\nu) - \hat{f}_N(x_\nu)\rvert$	$\lvert f - \hat{f}_N * \psi\rvert$ at $x - x_\nu$
2	3.2 (-3)	5.8 (-10)
3	5.2 (-3)	7.9 (-10)
4	7.8 (-3)	6.3 (-10)
5	1.1 (-2)	1.1 (-10)

Table II. Same as Table I for the pseudo-spectral approximation $\tilde{f}_N(x)$.

$x_{\nu+1/2} = \frac{\pi}{8}(\nu + 1/2)$ ν equals	$\lvert f(x_{\nu+1/2}) - \tilde{f}_N(x_{\nu+1/2})\rvert$	$\lvert f - \tilde{f}_N * \psi\rvert$ at $x = x_{\nu+1/2}$
2	5 (-3)	7 (-10)
3	8.1 (-3)	7.9 (-10)
4	1.2 (-2)	6.4 (-10)
5	1.8 (-2)	1.2 (-10)

Table III. Results of smoothing of the spectral approximation of f(x), N = 64

$x_\nu = \frac{\pi\nu}{8}$ ν equals	$\lvert f(x_\nu) - \hat{f}_N(x_\nu)\rvert$	$\lvert f - \hat{f}_N * \psi\rvert$ at $x = x_\nu$
2	6.4 (-3)	4.8 (-6)
3	1 (-2)	5.9 (-6)
4	1.5 (-2)	7.7 (-6)
5	2.3 (-2)	8.9 (-6)

Table IV. Same as Table III for the pseudo-spectral approximation, $\tilde{f}_N(x)$.

$x_{\nu+1/2} = \frac{\pi}{8}(\nu+1/2)$ ν equals	$\lvert f(x_{\nu+1/2}) - \tilde{f}_N(x_{\nu+1/2})\rvert$	$\lvert f - \tilde{f}_N * \psi\rvert$ at $x = x_{\nu+1/2}$
2	1 (-2)	4.1 (-6)
3	1.6 (-2)	6 (-6)
4	2.4 (-2)	7.8 (-6)
5	3.6 (-2)	8.9 (-6)

4. A DIFFERENT METHOD FOR EXTRACTING INFORMATION

In this section we would like to present a different approach for extracting the information from an oscillatory solution. The idea is to subtract from the solution those oscillations that correspond to the saw-tooth function discussed in Example 2. This leads to the following procedure:

Let $u_N(x,t) = \sum_{\ell=-N}^{N} \hat{u}_\ell e^{i\ell x}$, be the solution of the pseudospectral approximation to a hyperbolic problem. We try to find an unknown smooth function and a (oscillatory) saw-tooth function $F_N(x-t,x_s)$ with an unknown jump $2\pi A$ at an unknown location x_s such that

$$H = \left[\sum_{j=0}^{2N-1} u_N(x_j,t) - AF_N(x_j,x_s) - c - \sum_{\substack{\ell=-p \\ \ell \neq 0}}^{p} b_\ell e^{i\ell k_j} \right]^2 \tag{46}$$

is minimized. Note that we have $2p + 3$ unknowns in (46): A, x_s, c and $2P$ values of $b_\ell (\ell \neq 0)$.

The conditions for local minima of H are found from the following $2p + 3$ equations:

$$\frac{\partial H}{\partial A} = 0 \Longrightarrow \sum_{j=0}^{2N-1} u_j F_j - AF_j^2 - cF_j - F_j \sum_{\substack{\ell=-p \\ \ell \neq 0}}^{p} b_\ell e^{i\ell x_j} = 0 \tag{47}$$

where $F_j = F_N(x_j,x_s)$, $u_j = u_N(x_j,t)$. Also,

$$\frac{\partial H}{\partial c} = 0 \Longrightarrow \sum_{j=0}^{2N-1} u_j - AF_j - c - \sum_{\substack{\ell=-p \\ \ell\neq 0}}^{p} b_\ell \, e^{i\ell x_j} = 0 \tag{48}$$

$$\frac{\partial H}{\partial s} = 0 \Longrightarrow \sum_{j=0}^{2N-1} F_j' u_j - AF_j' F_j - cF_j' - F_j' \sum_{\substack{\ell=-p \\ \ell\neq 0}}^{p} b_\ell \, e^{ix_j\ell} = 0 \tag{49}$$

where $F_j' = \partial F_N(x_j,x_s)/\partial s = \sum_{\ell=-N}^{N} \frac{\partial a_\ell(s)}{\partial s} \cdot e^{i\ell x_j}$; and

$$\frac{\partial H}{\partial b_m} = 0 \Longrightarrow b_m = \hat{u}_m - Aa_m, \quad |m| = 1,2,\ldots,p \tag{50}$$

where

$$\hat{u}_m = \frac{1}{2Nc_m} \sum_{j=0}^{2N-1} u_N(x_j) e^{-i\ell x_j}.$$

Substituting (50) into (47), (48), and (49) we get, respectively:

$$\hat{u}_0 - Aa_0 - c = 0 \tag{51}$$

$$\sum_{|\ell|>p} \left(c_\ell \; a_{-\ell} \; \hat{u}_\ell - A\right) \sum_{|\ell|>p} \left(c_\ell \; a_{-\ell} \; a_\ell\right) = 0 \tag{52}$$

$$\sum_{|\ell|>p} \left(c_\ell \; a_{-\ell}' \; \hat{u}_\ell - A\right) \sum_{|\ell|>p} \left(c_\ell \; a_{-\ell}' \; a_\ell = 0\right) \tag{53}$$

where $a'(s) = \partial a_\ell(x)/\partial s$. Next, we combine (52) and (53) to get a single nonlinear equation for s:

$$\sum c_\ell \, a'_{-\ell} \, \hat{u}_\ell \sum c_\ell \, a_{-\ell} \, a_\ell - \sum c_\ell \, a_{-\ell} \, \hat{u}_\ell \sum c_\ell \, a'_{-\ell} \, a_\ell = 0 \tag{54}$$

where all sums run over $p < |\ell| \leq N$.

Equation (54) is solved iteratively for s. Having found s, one immediately obtains from Example 2 all the $a_\ell(s)'$s. Then from (50) we have the b_m's, and A from (52). Finally, having A we find c from (51).

The minimum thus obtained may be a local one while we are seeking a global minimum. This means that in practice one searches for the global minimum.

We now give an example that illustrates the efficacy of the procedure. We solve the following problem:

$$\frac{\partial u_N}{\partial t} + \frac{\partial u_N}{\partial x} = 0, \qquad 0 < x < 2\pi, \quad t > 0 \tag{55}$$

$$u_N(x,0) = \begin{cases} \sin \frac{x}{2} & 0 \leq x \leq \pi \\ -\sin \frac{x}{2} & \pi \leq x \leq 2\pi \end{cases} \tag{56}$$

$$u_N(0,t) = u_N(2\pi,t). \tag{57}$$

We ran the problem on several grids and exhibit here the numerical results for the case $N = 8$ (i.e., 16 subintervals in the domain $(0,2\pi)$). The unadulterated results at $t = \pi/2N$ on the grid points are shown in Figure 1.

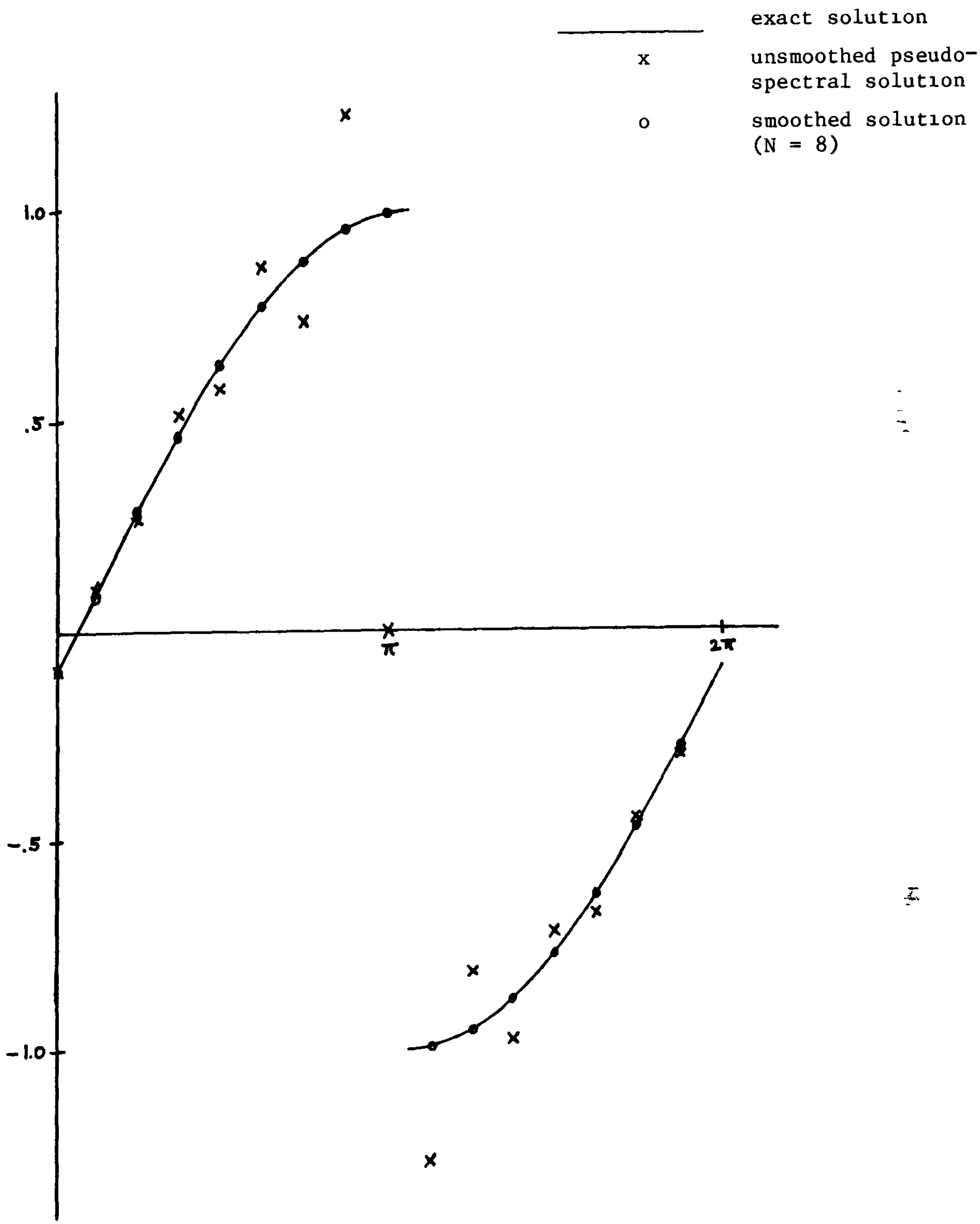
exact solution
x
unsmoothed pseudo-spectral solution
o
smoothed solution (N = 8)
1.0
.5
-.5
-1.0
π
2π

Figure 1

Table V

j	exact solution	error 1 = $\lvert$exact-unsmoothed$\rvert$	error 2 = $\lvert$exact-smoothed$\rvert$	$\frac{\text{error 1}}{\text{error 2}}$
0	9.80×10^{-2}	5.86×10^{-5}	5.86×10^{-5}	1.00
1	9.80×10^{-2}	1.24×10^{-2}	5.86×10^{-5}	211
2	2.90×10^{-1}	2.57×10^{-2}	6.30×10^{-5}	408
3	4.71×10^{-1}	4.13×10^{-2}	7.33×10^{-5}	563
4	6.34×10^{-1}	6.15×10^{-2}	9.30×10^{-5}	661
5	7.73×10^{-1}	9.11×10^{-2}	1.31×10^{-4}	695
6	8.82×10^{-1}	1.43×10^{-1}	2.16×10^{-4}	662
7	9.57×10^{-1}	2.70×10^{-1}	4.42×10^{-4}	611
8	-9.95×10^{-1}	1.00×10^{0}	1.10×10^{-2}	91
9	-9.95×10^{-1}	2.68×10^{-1}	1.34×10^{-3}	200
10	-9.57×10^{-1}	1.42×10^{-1}	4.42×10^{-4}	321
11	-8.82×10^{-1}	9.07×10^{-2}	2.16×10^{-4}	420
12	-7.73×10^{-1}	6.12×10^{-2}	1.32×10^{-4}	464
13	-6.34×10^{-1}	4.11×10^{-2}	9.30×10^{-5}	442
14	-4.71×10^{-1}	2.55×10^{-2}	7.32×10^{-5}	348
15	-2.90×10^{-1}	1.22×10^{-2}	6.30×10^{-5}	194

We then post-processed these $u_N(x_j, \pi/2N)$ values according to the procedure described above. The filtered values are shown on the same graph, and the errors listed in Table V are computed before and after processing. The dramatic improvement is evident.

Next we demonstrate the procedure in the case of the Euler equation for gas dynamics. Because the physical problem involves inflow, outflow, and no-flow boundary conditions, periodicity could not be imposed and we use the Tchebyshev, rather than Fourier, pseudospectral method.

The physical problem is that of a wedge, inserted as a zero angle of attack, into a uniform supersonic flow of an ideal gas with $\gamma = 1.4$. An oblique shock develops in time and the flow reaches, after a while, a steady state. The time-dependent Euler equations in two-space dimensions were discretized by the pseudospectral Tchebyshev method in space with an 8×8 grid and a modified Euler scheme was used for the time discretization. Since we are interested in the steady state only, the accuracy for the time integration is of little importance. In order to be sure that a steady state is reached, the code was run until all physical quantities did not change to 11 significant figures over a span of 100 time steps. The values of the density in the steady state at the grid points together with the grid points themselves are given in Table VI.

Table VI.

	ρ									Y
	1.862	1.851	1.869	1.871	1.837	1.865	1.892	1.885	1.878	1.
	1.862	1.870	1.867	1.820	1.870	1.954	1.899	1.803	1.759	.961
	1.862	1.854	1.852	1.904	1.877	1.770	1.782	1.864	1.900	.853
	1.862	1.871	1.876	1.812	1.838	1.969	1.975	1.884	1.841	.691
	1.862	1.848	1.842	1.935	1.899	1.703	1.710	1.890	1.984	.5
	1.862	1.883	1.894	1.729	1.832	2.429	2.994	3.255	3.316	.308
	1.862	1.808	1.810	2.387	3.133	3.375	3.224	3.054	3.002	.146
	1.862	2.115	2.868	3.288	3.176	2.965	3.006	3.136	3.187	.038
	1.862	3.083	3.046	2.975	3.087	3.108	3.024	3.013	3.016	0
X	0	.038	.146	.308	.5	.691	.853	.961	1.	

Note that the raw data in Table VI seems to indicate roughly the same y-shock location at $x_0 = 1$, $x_1 = .961$, and $x_2 = .853$, namely between the grid points $y_4 = .3086$ and $y_5 = .500$. This means that because of the coarse Tchebyshev grid the shock location cannot be resolved to better than 20% of the domain. In fact, the correct shock locations at those x-stations are $y = .434$ for x_0, $y = .417$ for x_1 and $y = .370$ for x_2.

In the present case it is not necessry to employ a saw-tooth piecewise smooth function, as was done in the previous section, because there is no need to preserve periodicity. Instead, we subtract from the oscillatory data an expansion of the Heaviside function, $S(y,y_s)$:

$$S(y,y_s) = \begin{cases} d_1 + d_2 & -1 < y < y_s \\ d_1 & y_s < y < 1 \end{cases} \tag{58}$$

where d_1, the state ahead of the shock, and d_2, the magnitude of the discontinuity, are constant. The description here of $S(y,y_s)$, as if independent of x, has to do with the fact that the two-dimensional results of the pseudospectral algorithm were post-processed at fixed x-stations. The expansion of $S(y,y_s)$ is given by

$$S_N(y,y_s) = \sum_{\ell=0}^{N} A_\ell(s)T_\ell(y) \tag{59}$$

where $T_\ell(y)$ is the Tchebyshev polynomial of order ℓ,
$T_\ell(y) = \cos[\ell \cos^{-1}(y)]$, and

$$A_0(s) = (s + \frac{1}{2})/N$$

$$A_\ell(s) = \sin[\frac{\pi\ell}{N}(s + \frac{1}{2})] / N \sin\frac{\pi\ell}{2N}; \quad 1 \leq \ell \leq N-1$$

$$A_N(s) = \sin[(s + \frac{1}{2})]/2N.$$

If s is an integer, then on the grid points, $y_j = \cos(\pi j/N)$.

$$S_N(y_j, y_s) = S(y_j, y_s). \tag{60}$$

The L_2-norm which we wish to minimize is now, at any given x-station:

$$H = \sum_{j=0}^{N} \frac{1}{c_j} [\rho_N(y_j) - d_1 - d_2 S_N(y_j, y_s) - \sum_{\ell=1}^{p<N} b_\ell T_\ell(y_j)]^2 \tag{61}$$

$$c_j = \begin{cases} 1 & 1 \leq j \leq N-1 \\ 2 & j = 0, N \end{cases} . \tag{62}$$

Differentiating (61) with respect to the parameters d_1, d_2, s and $b_\ell (1 \leq \ell \leq p < N)$, using the orthogonality relations for the Tchebyshev polynomials and manipulations similar to those used in the previous section, we get $p + 3$ nonlinear algebraic equations which are completely analogous to (50) - (53). They are:

$$b_\ell = \hat{\rho}_\ell - d_2 A_\ell, \qquad \ell = 1,2,\ldots,p. \tag{63}$$

$$\hat{\rho}_0 - d_2 A_0 - d_1 = 0 \tag{64}$$

$$\sum_{\ell=p+1}^{N} c_\ell A_\ell \hat{\rho}_\ell - d_2 \sum_{\ell=p+1}^{N} c_\ell A_\ell^2 = 0 \tag{65}$$

$$\sum_{\ell=p+1}^{N} c_\ell A_\ell' \hat{\rho}_\ell - d_2 \sum_{\ell=p+1}^{N} c_\ell A_\ell A_\ell' = 0 \tag{66}$$

where

$$\hat{\rho}_\ell = \frac{2}{Nc_\ell} \sum_{j=0}^{N} \frac{1}{c_j} \rho(y_j) T_\ell(y_j) \tag{67}$$

$$A' = \frac{\partial}{\partial s} A_\ell(s). \tag{68}$$

Again, we combine (65) and (66) into a single nonlinear equation for the shock location index, s:

$$\sum c_\ell A_\ell' \rho_\ell \sum c_\ell A_\ell^2 - \sum c_\ell A_\ell \hat{\rho}_\ell \sum c_\ell A_\ell A_\ell' = 0 \tag{69}$$

where all the sums are from $\ell = p+1$ to $\ell = N$.

The procedure for extracting the shock location, jump magnitude and smooth part of the solution from the raw data $\rho(x,y_j)$ (given in Table VI) is exactly the same as described above for the Fourier problem.

For the wedge-flow problem considered here, this procedure applied in the case of a coarse net (N = 8), located the shock with an error only in the fourth significant figure. The smooth part was recovered to within 1% at the worst field point.

Conclusion

We have demonstrated that the realization of a numerical solution via its grid-point value may be misleading when the true solution has a complicated structure which is not resolved by the grid. We have shown, however, that the numerical solution does contain highly accurate information about the solution and we suggested two ways of extracting this information.

The analysis outlined in this chapter is a linear one (though the procedure was applied also to nonlinear problems). However, in [28] Lax has argued that more information about the solution is contained in high resolution schemes even in the nonlinear case. In fact, using notions from information theory, Lax has shown that the ε-capacity of the set of approximate solutions is closer to the ε-capacity of the set of the projections of exact solutions if the numerical scheme is a high-order scheme.

In the area of digital filters one always processes the data in order to overcome the Gibbs phenomenon. If we look at the initial conditions as an input signal and at the final result as the output signal, the idea of filtering is a natural one.

REFERENCES

[1] GOTTLIEB, D., M. Y. HUSSAINI, and S. A. ORSZAG, "Introduction to the Proc. of the Symposium on Spectral Methods," SIAM CBMS Series, 1983.

[2] GOTTLIEB, D. and E. TADMOR, "Recovering pointwise values of discontinuous data within spectral accuracy," ICASE Report No. 85-3, NASA CR-172535, 1985.

[3] LAX, P. D., "Accuracy and resolution in the computation of solutions of linear and nonlinear equations," Recent Advances in Numerical Analysis, Proc. Symp. Mathematical Research Center, University of Wisconsin, Academic Press, 1978, pp. 107-117.

[4] MAJDA, A. and S. OSHER, "The Fourier method for nonsmooth initia data," Math. Comp., Vol. 32, 1978, pp. 1041-1081.

[5] MOCK, M. S. and LAX, P. D., "The computation of discontinuous solutions of linear hyperbolic equations," Comm. Pure Appl. Math., Vol. 31, 1978, pp. 423-430.

End of Document