Essentially Nonoscillatory Spectral Fourier Methods for Shock Wave Calculations*

By Wei Cai, David Gottlieb, and Chi-Wang Shu

Dedicated to Professor Eugene Isaacson on the occasion of his 70th birthday

Abstract. In this paper, we present an essentially nonoscillatory spectral Fourier method for the solution of hyperbolic partial differential equations. The method is based on adding a nonsmooth function to the trigonometric polynomials which are the usual basis functions for the Fourier method. The high accuracy away from the shock is enhanced by using filters. Numerical results confirm that essentially no oscillations develop in the solution. Also, the accuracy of the spectral solution of the inviscid Burgers equation is shown to be higher than a fixed order.

1. Introduction. In this paper we discuss shock-capturing techniques using spectral methods. In particular, we would like to present an essentially nonoscillatory version of the spectral Fourier method when applied to a nonlinear hyperbolic equation. The main difficulty in applying spectral methods to discontinuous problems is of course the Gibbs phenomenon. In fact, this problem exists even on the approximation level. It is well known that if a discontinuous function f(x) is approximated by its finite Fourier series $P_N f$,

(1.1a)
$$P_N f = \sum_{k=-N}^N \hat{f}_k e^{ikx},$$

(1.1b)
$$\hat{f}_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} \, dx,$$

then the order of convergence of $P_N f$ to f is only O(1/N) for each fixed point. Moreover, $P_N f$ has oscillations of order 1 in a neighborhood of O(1/N) of the discontinuity.

In the applications, we usually have piecewise C^{∞} functions, and in this paper we will consider only those functions. It is known that it is possible to improve the accuracy of the approximation away from the shocks. There are currently two methods (see [7], [9]) that are being used. The first [9] amounts to modifying the Fourier coefficients by multiplying them by a decreasing function $\tau(k)$. Some of the

Received May 31, 1988.

¹⁹⁸⁰ Mathematics Subject Classification (1985 Revision). Primary 65M99.

^{*}This research was supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-18107 while the second author was in residence at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA 23665. Additional support was provided by the Air Force Office of Scientific Research under Grant No. 85-0303 and DARPA URI contract N00014-86-K0754.

commonly used filters are

(1.2)
$$\begin{aligned} \tau_{k} &= e^{-\alpha ((k-k_{0})/N)^{2m}}, \qquad |k| \geq k_{0}, \\ \tau_{k} &= 1, \qquad |k| < k_{0}. \end{aligned}$$

The second method [1] is based on convolving the approximation with an appropriate C^{∞} function $\psi(x, y)$ such that

(1.3)
$$P_N f * \psi(x,y) \sim f(y).$$

While both (1.2) and (1.3) are effective away from the discontinuity, they do not eliminate the Gibbs phenomenon in the neighborhood of the shock. This is very important for the stability of the spectral method when applied to partial differential equations. In fact, in Section 2 we show that the total variation of $P_N f$ grows like log N. It is easily shown that this is the case also for the filters in (1.2).

In Section 2, we show that by adding a sawtooth function to the basis functions e^{ikx} one can control the Gibbs phenomenon. This, in conjunction with the filters (1.2)-(1.3), yields a higher-order essentially nonoscillatory approximation to a piecewise C^{∞} function. In Theorem 2, we prove that the total variation of the new approximation converges to that of the approximated function. We also prove that the convergence for the new approximation in the L^1 norm is one order higher than that of the usual spectral approximation.

Many modern nonlinear schemes for the solution of the conservation equation

(1.4)
$$u_t + f(u)_x = 0$$

are based on two distinct steps, namely reconstruction and time marching. We use the cell averaging formulation to rewrite (1.4) as

(1.5)
$$\frac{\partial \overline{u}_j}{\partial t} + \frac{1}{\Delta x_j} (f(u_{j+1/2}) - f(u_{j-1/2})) = 0,$$

where

$$\Delta x_j = x_{j+1/2} - x_{j-1/2}, \qquad \overline{u}_j = \frac{1}{\Delta x_j} \int_{x_{j-1/2}}^{x_{j+1/2}} u \, dx_j$$
$$f(u_{j+1/2}) = f(u(x_{j+1/2})).$$

The first step, then, is to reconstruct the function u(x) from $\overline{u}(x)$. It is here that we use the essentially nonoscillatory technique developed in Section 2. For the second step, the time marching, we use the third-order Runge-Kutta scheme developed in [13]. We try to avoid any modification technique, such as the application of limiters, in order to avoid deterioration of the overall accuracy.

We demonstrate in the last section that the procedure applied to several model problems yields indeed essentially nonoscillatory results with an order of accuracy which is higher than algebraic away from the discontinuity.

2. Essentially Nonoscillatory Approximation. In this section, we suggest a method to reconstruct an essentially nonoscillatory approximation to a piecewise C^{∞} periodic function from its first N Fourier coefficients. The approximation is essentially nonoscillatory in the sense that the total variation of the approximation converges to the total variation of the approximated function. Moreover, the approximation converges in the maximum norm outside a small interval around the point of discontinuity. Applying the filters (1.2)-(1.3) will increase the order of convergence away from the discontinuity, thus providing an essentially nonoscillatory spectral approximation.

For simplicity, assume that u(x), $0 \le x \le 2\pi$, is a periodic piecewise C^{∞} function with only one point of discontinuity at x_s , and denote by [u] the value of the jump of u(x) at x_s , namely

(2.1)
$$[u] = \frac{u(x_s^+) - u(x_s^-)}{2\pi}.$$

We assume also that the first 2N + 1 Fourier coefficients \hat{u}_l of u(x) are known:

(2.2)
$$\hat{u}_l = \frac{1}{2\pi} \int_0^{2\pi} u(x) e^{-ilx} dx, \quad -N \le l \le N$$

The objective is to construct an essentially nonoscillatory spectrally accurate approximation to u(x) from the \hat{u}_l 's. We start by noting that the Fourier coefficients \hat{u}_l 's contain information about the shock position x_s and the magnitude [u] of the shock. In fact we can state

LEMMA 1. Let u(x) be a periodic piecewise C^{∞} function with one point of discontinuity x_s ; then for $|l| \ge 1$ and for any n > 0,

(2.3)
$$\hat{u}_l = e^{-ilx_s} \sum_{k=0}^{n-1} \frac{[u^{(k)}]}{(il)^{k+1}} + \frac{1}{2\pi} \int_0^{2\pi} \frac{[u^{(n)}]}{(il)^n} e^{-ilx} dx$$

Proof. Since

$$\hat{u}_l = \frac{1}{2\pi} \int_0^{2\pi} u(x) e^{-ilx} \, dx = \frac{1}{2\pi} \int_0^{x_s} u(x) e^{-ilx} \, dx + \frac{1}{2\pi} \int_{x_s}^{2\pi} u(x) e^{-ilx} \, dx,$$

we can integrate by parts to get

(2.4)
$$\hat{u}_l = e^{-ilx_s} \frac{u(x_s^+) - u(x_s^-)}{2\pi i l} + \frac{1}{2\pi} \int_0^{2\pi} \frac{u'(x)e^{-ilx}}{il} dx;$$

the rest is obtained by induction. This completes the proof. \Box

As an example, consider the sawtooth function $F(x, x_s, A)$ defined by

(2.5)
$$F(x, x_s, A) = A \begin{cases} -x, & x \leq x_s, \\ 2\pi - x, & x > x_s. \end{cases}$$

Note that the jump of the function, [F], is A and all the derivatives are continuous: $[d^k F/dx^k] = 0$ for $k \ge 1$. That means that the expansion (2.3) can be terminated after the first term, yielding the following results for \hat{f}_k , the Fourier coefficients of $F(x, x_s, A)$:

(2.6)
$$\hat{f}_k(x_s, A) = A \frac{e^{-ikx_s}}{ik}, \qquad |k| \ge 1,$$
$$\hat{f}_0(x_s, A) = A(\pi - x_s).$$

This example suggests that we can rewrite (2.3) as

$$(2.7) \quad \hat{u}_l = \hat{f}_l(x_s, [u]) + e^{-ilx_s} \sum_{k=1}^{n-1} \frac{[u^{(k)}]}{(il)^{k+1}} + \frac{1}{2\pi} \int_0^{2\pi} \frac{u^{(n)}(x)e^{-ilx}}{(il)^n} \, dx, \qquad |l| \ge 1.$$

The order 1 oscillations in approximating u(x) by its finite Fourier sum

$$(2.8) P_N u = \sum_{l=-N}^N \hat{u}_l e^{ilx}$$

are caused by the slow convergence of

(2.9)
$$F_N(x, x_s, [u]) = \sum_{l=-N}^N \hat{f}_l(x_s, [u]) e^{ilx}$$

to the sawtooth function $F(x, x_s, [u])$. Therefore, those oscillations can be eliminated by adding a sawtooth function to the basis of the space to which u(x) is projected. To be specific, we seek an expansion of the form

(2.10)
$$v_N(x) = \sum_{|l| \le N} a_l e^{ilx} + \sum_{|l| > N} \frac{A}{il} e^{-ily} e^{ilx}$$

to approximate u(x). The 2N + 3 unknowns a_l $(|l| \le N)$, A and y are determined by the orthogonality condition

(2.11)
$$\int_0^{2\pi} (u - v_N) e^{-ikx} \, dx = 0, \qquad |k| \le N + 2$$

The system of equations (2.11) leads to the following conditions:

$$(2.12) a_l = \hat{u}_l, |l| \le N$$

(where \hat{u}_l are the usual Fourier coefficients of u(x), see (2.2)) and

(2.13a)
$$\frac{A}{i(N+1)}e^{-i(N+1)y} = \hat{u}_{N+1}$$

(2.13b)
$$\frac{A}{i(N+2)}e^{-i(N+2)y} = \hat{u}_{N+2}.$$

Solving (2.13) for A and y one gets

(2.14a)
$$e^{iy} = \frac{(N+1)\hat{u}_{N+1}}{(N+2)\hat{u}_{N+2}}$$

(2.14b)
$$|A| = (N+1)|\hat{u}_{N+1}|.$$

The sign of A is determined by (2.13).

Note that in the expansion presented in (2.10) the second sum starts at |l| = N + 1. This is due to the fact that we make the additional basis function F(x, y, A) orthogonal to e^{ikx} , thus we use $F(x, y, A) - F_N(x, y, A)$ in the expansion (2.10). The procedure described in (2.14) is second-order accurate in the location and jump of the shock. In fact, we can state

THEOREM 1. Let u(x) be a piecewise C^{∞} function with one discontinuity at x_s . Let y and A be defined in (2.14); then

(2.15)
$$|y - x_s| = O(1/N^2), \qquad |A - [u]| = O(1/N^2).$$

Proof. From (2.3) we get

$$e^{iy} = \frac{(N+1)\hat{u}_{N+1}}{(N+2)\hat{u}_{N+2}} = \frac{e^{-i(N+1)x_s} \left[[u] + \frac{[u']}{i(N+1)} + O\left(\frac{1}{(N+1)^2}\right) \right]}{e^{-i(N+2)x_s} \left[[u] + \frac{[u']}{i(N+2)} + O\left(\frac{1}{(N+2)^2}\right) \right]}$$
$$= e^{ix_s} \left[1 + O\left(\frac{1}{N^2}\right) \right].$$

By the same token,

$$|A| = (N+1)|\hat{u}_{N+1}| = \left[\left\{[u] - \frac{[u'']}{(N+1)^2}\right\}^2 + \frac{[u']^2}{(N+1)^2}\right]^{1/2}$$
$$= |[u]| \left[1 + O\left(\frac{1}{N^2}\right)\right]. \quad \Box$$

It should be noted that a better approximation to the shock location x_s and its magnitude [u] can be obtained if we add to the basis functions a function of the form

(2.16)
$$\sum_{|l|>N} \left[\frac{A}{il} + \frac{B}{(il)^2}\right] e^{-ily} e^{ilx}$$

and extend (2.11) to $|k| \leq N + 3$. In practice, however, (2.10) is enough to get an essentially nonoscillatory scheme.

In order to demonstrate that the procedure described in (2.10), (2.12), and (2.14) is indeed essentially nonoscillatory, we recall the definition of the total variation of a function.

Definition. The total variation of u over $[0, 2\pi]$ —TV[u]—is defined as

(2.17)
$$TV[u] = \sup \sum_{i=1}^{n} |u(x_i) - u(x_{i-1})|$$

where $0 \le x_0 < x_1 < \cdots < x_n \le 2\pi$ is a partition of $[0, 2\pi]$. The supremum is taken over all partitions.

It is clear that if $u'(x) \in L^1$ then

(2.18)
$$\mathrm{TV}[u] = \int_0^{2\pi} |u'(\xi)| \, d\xi.$$

If we approximate the function u(x) by its finite Fourier series $P_N u$ defined in (2.8), then it is well known that the total variation of $P_N u$ need not approximate that of u. In fact we can state

LEMMA 2. Let the sawtooth function F(x,0,1) and its Nth Fourier approximation $F_N(x,0,1)$ be defined by (2.5) and (2.9); then

$$(2.19) TV[F] = 4\pi,$$

(2.20)
$$\operatorname{TV}[F_N] = O(\log N).$$

Proof. Equation (2.19) follows directly from the definition of total variation. As for (2.20), we note that

(2.21)
$$TV[F_N(x,0,1)] = \int_0^{2\pi} |F'_N(x,0,1)| \, dx = \int_0^{2\pi} \left| \sum_{\substack{l=-N\\l\neq 0}}^N e^{ilx} \right| \, dx$$
$$= \int_0^{2\pi} \left| \frac{\sin(N+\frac{1}{2})x}{\sin\frac{1}{2}x} - 1 \right| \, dx$$
$$= \int_0^{2\pi} \left| \frac{\sin(N+\frac{1}{2})x}{\sin\frac{1}{2}x} \right| \, dx + O(1).$$

The first term on the right-hand side of (2.21) is the Lebesgue constant. It is known [14, p. 67] that it grows like log N. Hence (2.20). \Box

We can therefore conclude that $TV(P_N u)$ does not converge to TV[u]. This reflects the existence of large oscillations in the neighborhood of the discontinuities.

The situation is different for v_N defined in (2.10). In fact we can state

THEOREM 2. Let u(x) be a piecewise C^{∞} periodic function with one point of discontinuity x_s , and a jump of [u]. Let A and y be such that

(2.22)
$$|y - x_s| = \Delta_1, \qquad |A - [u]| = \Delta_2.$$

Let v_N be defined in (2.10); then

(2.23)
$$\operatorname{TV}[v_N] \le \operatorname{TV}[u] + L_0 \frac{\log N}{N} + L_1 \Delta_1 N \log N + L_2 \Delta_2 \log N,$$

(2.24)
$$\|v_N - u\|_{L^1} \le C_0 \frac{\log N}{N^2} + C_1 \Delta_1 \log N + C_2 \Delta_2.$$

We present the proof in a series of lemmas in order to clarify the role of each one of the terms on the right-hand sides of (2.23) and (2.24).

LEMMA 3. Let $F_N(x, \alpha, 1)$ and $F_N(x, \beta, 1)$ be defined by (2.6)-(2.9) and $\Delta = \alpha - \beta > 0$. Then, if $\Delta \leq 1/N$,

(2.25)
$$\operatorname{TV}[F_N(x,\alpha,1) - F_N(x,\beta,1)] = O(\Delta N \log N),$$

(2.26)
$$||F_N(x,\alpha,1) - F_N(x,\beta,1)||_{L^1} = O(\Delta \log N)$$

Proof. Since $F_N(x, \alpha, 1), F_N(x, \beta, 1)$ are trigonometric polynomials, they are C^{∞} functions. Therefore

$$TV[F_N(x,\alpha,1) - F_N(x,\beta,1)] = \int_0^{2\pi} |F'_N(x,\alpha,1) - F'_N(x,\beta,1)| dx$$

$$(2.27) \qquad \qquad = \int_0^{2\pi} \left| \sum_{|l| \le N} [e^{il(x-\alpha)} - e^{il(x-\beta)}] \right| dx$$

$$= 4 \int_0^{2\pi} \left| \sum_{l=1}^N \sin l \left(x - \frac{\alpha+\beta}{2} \right) \sin l \frac{\alpha-\beta}{2} \right| dx.$$

Upon defining $\sigma_l = \sin l \frac{\alpha - \beta}{2}$, we can rewrite (2.27) as

(2.28)
$$TV[F_N(x,\alpha,1) - F_N(x,\beta,1)] = 4 \int_0^{2\pi} \left| \sum_{l=1}^N \sigma_l \sin l\xi \right| d\xi;$$

we note that the σ_l are positive and monotone in l, $\sigma_{l-1} - \sigma_l < 0$. Define now

$$(2.29) B_l(\xi) = \sum_{k=0}^l \sin k\xi$$

to get

(2.30)
$$\sum_{l=1}^{N} \sigma_l \sin l\xi = \sum_{l=1}^{N} \sigma_l (B_l(\xi) - B_{l-1}(\xi))$$
$$= \sum_{l=1}^{N} (\sigma_{l-1} - \sigma_l) B_{l-1} + \sigma_N B_N.$$

Therefore,

(2.31)
$$\int_{0}^{2\pi} \left| \sum_{l=1}^{N} \sigma_{l} \sin l\xi \right| d\xi \leq \sigma_{N} \int_{0}^{2\pi} |B_{N}(\xi)| d\xi + \sum_{l=1}^{N} (\sigma_{l} - \sigma_{l-1}) \int_{0}^{2\pi} |B_{l-1}(\xi)| d\xi.$$

Denote now

(2.32)
$$\mu = \max_{1 \le l \le N} \int_0^{2\pi} |B_l(\xi)| \, d\xi;$$

from (2.28) and (2.31) we obtain

(2.33)
$$\operatorname{TV}[F_N(x,\alpha,1) - F_N(x,\beta,1)] \leq 8\mu\sigma_N.$$

In order to estimate μ , we first note that

(2.34)
$$B_l(\xi) = \sin \frac{(l+1)\xi}{2} \cdot \frac{\sin l\xi/2}{\sin \xi/2}$$

Therefore,

(2.35)
$$\int_0^{2\pi} |B_l(\xi)| \, d\xi \le \int_0^{2\pi} \frac{|\sin l\xi/2|}{\sin \xi/2} \, d\xi = \mu_l.$$

But μ_l is exactly the Lebesgue constant; therefore,

(2.36)
$$\mu_l = O(\log l).$$

Since $\sigma_N \leq N\Delta$, we get

$$\mathrm{TV}[F_N(x,\alpha,1) - F_N(x,\beta,1)] = O(\Delta N \log N).$$

To obtain (2.26), we follow the same arguments as above. Similar to (2.27) and (2.28) we have

(2.37)
$$\int_{0}^{2\pi} |F_N(x,\alpha,1) - F_N(x,\beta,1)| \, dx \le 2\pi\Delta + 4 \int_{0}^{2\pi} \left| \sum_{l=1}^{N} \hat{\sigma}_l \cos l\xi \right| \, d\xi,$$

where $\hat{\sigma}_l = \sigma_l/l$, $1 \le l \le N$. The $\hat{\sigma}_l$'s are positive and monotone in l, $\sigma_l - \sigma_{l-1} < 0$. If we define

(2.38)
$$\hat{B}_{l}(\xi) = \sum_{k=1}^{l} \cos k\xi = \cos \frac{(l+1)\xi}{2} \cdot \frac{\sin l\xi/2}{\sin \xi/2},$$
$$\hat{B}_{0}(\xi) = 0,$$

then similar to (2.33) we have

(2.39)
$$\int_0^{2\pi} |F_N(x,\alpha,1) - F_N(x,\beta,1)| \, dx \le 2\pi\Delta + 8\mu\hat{\sigma}_1,$$

where μ is defined in (2.32) with B_l replaced by \hat{B}_l of (2.38). Notice that (2.35) also holds for $\hat{B}_l(x)$ and $|\hat{\sigma}_1| \leq \Delta$. (2.26) now follows from (2.35), (2.37) and (2.39). \Box

LEMMA 4. Let $F_N(x, \alpha, A)$ and $F_N(x, \beta, B)$ be defined in (2.6)-(2.9). Denote

(2.40)
$$|\alpha - \beta| = \Delta_1, \qquad |A - B| = \Delta_2.$$

Then

(2.41)
$$\operatorname{TV}[F_N(x,\alpha,A) - F_N(x,\beta,B)] \le K_1 \Delta_1 N \log N + K_2 \Delta_2 \log N,$$

(2.42)
$$\|F_N(x,\alpha,A) - F_N(x,\beta,B)\|_{L^1} \le C_1 \Delta_1 \log N + C_2 \Delta_2$$

for K_1, K_2, C_1, C_2 independent of N.

Proof. We have

(2.43)
$$TV[F_N(x,\alpha,A) - F_N(x,\beta,B)] \leq TV[F_N(x,\alpha,A) - F_N(x,\beta,A)]$$
$$+ TV[F_N(x,\beta,A) - F_N(x,\beta,B)].$$

The first term in the right-hand side of (2.43) is bounded by (2.25); for the second term,

(2.44)

$$TV[F_N(x,\beta,A) - F_N(x,\beta,B)] = TV \left[(A-B) \sum_{\substack{l=-N\\l\neq 0}}^{N} \frac{e^{il(x-\beta)}}{il} \right]$$

$$\leq |A-B| \int_0^{2\pi} \left| \sum_{\substack{l=-N\\l\neq 0}}^{N} e^{il(x-\beta)} \right| dx$$

$$\leq K_2 \Delta_2 \log N.$$

Similarly, we have

(2.45)
$$\begin{split} \|F_N(x,\alpha,A) - F_N(x,\beta,B)\|_{L^1} \\ &\leq \|F_N(x,\alpha,A) - F_N(x,\beta,A)\|_{L^1} + \|F_N(x,\beta,A) - F_N(x,\beta,B)\|_{L^1} \\ &= \|F_N(x,\alpha,A) - F_N(x,\beta,A)\|_{L^1} + |A - B| \|F_N(x,\beta,1)\|_{L^1}. \end{split}$$

The first term on the right side of (2.45) is bounded by (2.26). For the second term,

(2.46)
$$|A - B| ||F_N(x,\beta,1)||_{L^1} = \Delta_2 ||F_N(x,\beta,1)||_{L^1} \le C_2 \Delta_2.$$

The assertion (2.42) now follows from (2.45), (2.46).

LEMMA 5. Let $S(x, \alpha, A)$ and $S_N(x, \alpha, A)$ be defined by

(2.47)
$$S(x,\alpha,A) = A \sum_{\substack{l=-\infty\\l\neq 0}}^{\infty} \frac{e^{il(x-\alpha)}}{(il)^2}, \qquad S_N(x,\alpha,A) = A \sum_{\substack{l=-N\\l\neq 0}}^{N} \frac{e^{il(x-\alpha)}}{(il)^2}.$$

396

Then

(2.48)
$$\operatorname{TV}[S(x,\alpha,A) - S_N(x,\alpha,A)] \le K_3 \frac{\log N}{N}$$

(2.49)
$$||S(x,\alpha,A) - S_N(x,\alpha,A)||_{L^1} \le K_4 \frac{\log N}{N^2}$$

for K_3, K_4 independent of N.

Proof. It is clear that

$$\mathrm{TV}(S(x,\alpha,A)-S_N(x,\alpha,A))=\int_0^{2\pi}\left|F(x,\alpha,A)-F_N(x,\alpha,A)\right|dx.$$

The estimates (2.48)–(2.49) then follow from [14, p. 185]. \Box

We are now ready to prove Theorem 2.

Proof of Theorem 2. First we prove (2.23). In view of (2.3) we can write

(2.50)
$$u = F(x, x_s, [u]) + S(x, x_s, [u']) + g(x)$$

and therefore

(2.51)
$$P_N u = F_N(x, x_s, [u]) + S_N(x, x_s, [u']) + P_N g(x).$$

We can also rewrite (2.10) as

(2.52)
$$v_N(x) = P_N u + F(x, y, A) - F_N(x, y, A)$$

hence

$$v_N(x) = F_N(x, x_s, [u]) + S_N(x, x_s, [u']) + P_N g(x) + F(x, y, A) - F_N(x, y, A)$$

or

(2.53)
$$v_N(x) = [F_N(x, x_s, [u]) - F_N(x, y, A)] + [F(x, y, [u]) + S(x, y, [u']) + g(x - y + x_s)] + [S_N(x, x_s, [u']) - S(x, y, [u'])] + [P_Ng(x) - g(x - y + x_s)] + [F(x, y, A) - F(x, y, [u])].$$

The second term on the right-hand side is just the original function u shifted,

(2.54)
$$F(x,y,[u]) + S(x,y,[u']) + g(x-y+x_s) = u(x-y+x_s);$$

also from (2.48),

(2.55)

$$TV[S_N(x, x_s, [u']) - S(x, y, [u'])] \leq TV[S_N(x, x_s, [u']) - S_N(x, y, [u'])] \\
 + TV[S_N(x, y, [u']) - S(x, y, [u'])] \\
 \leq K \frac{\log N}{N}$$

and finally, since g(x) is smooth enough,

(2.56)
$$\operatorname{TV}[P_N g - g(x - y + x_s)] \leq \frac{K}{N}.$$

Therefore from (2.53) and Lemmas 4 and 5,

(2.57)
$$\operatorname{TV}[v_N] \le \operatorname{TV}[u] + L_0 \frac{\log N}{N} + L_1 \Delta_1 N \log N + L_2 \Delta_2 \log N.$$

Next we prove (2.24) following the same argument above. From (2.50), (2.51) and (2.52),

(2.58)
$$v_N(x) - u(x) = [F_N(x, x_s, [u]) - F_N(x, y, A)] + [S_N(x, x_s, [u']) - S(x, x_s, [u'])] + [F(x, y, A) - F(x, x_s, [u])] + [P_N g(x) - g(x)].$$

The first term will be bounded by (2.42), the second term by (2.49); for the third term,

(2.59)
$$\int_0^{2\pi} |F(x,y,A) - F(x,x_s,[u])| \, dx \le C_1 \Delta_1 + C_2 \Delta_2.$$

Now since g(x) is smooth enough, we have

(2.60)
$$||P_Ng - g||_{L^1} = O\left(\frac{1}{N^2}\right)$$

Therefore from Lemma 4, Lemma 5 and (2.59)-(2.60),

$$\|v_N - u\|_{L^1} \le C_0 \frac{\log N}{N^2} + C_1 \Delta_1 \log N + C_2 \Delta_2,$$

and the proof is completed. \Box

COROLLARY. The method suggested in (2.15) yields

$$|A-[u]| = O\left(\frac{1}{N^2}\right)$$
 and $|y-x_s| = O\left(\frac{1}{N^2}\right)$,

and therefore

(2.61)
$$\mathrm{TV}[v_N] \leq \mathrm{TV}[u] + K \frac{\log N}{N},$$

(2.62)
$$\|v_N - u\|_{L^1} \le C \frac{\log N}{N^2}$$

Thus, the total variation of v_N converges to that of u. Convergence of v_N to u in the L^1 norm is one order higher than in the case of $P_N u$, for which the rate of convergence in L^1 is $O(\log N/N)$. The method therefore yields a reconstruction technique which is total variation bounded.

We conclude this section by pointing out that a similar result for collocation method and/or for Chebyshev expansions can be developed along the same lines. Computationally, we observe similar results for Galerkin and collocation methods (see Section 4). In practice, collocation is used more often than Galerkin, especially when solving a nonlinear PDE (Section 3).

3. Essentially Nonoscillatory Spectral Schemes. In this section we apply the techniques discussed in Section 2 to solve the PDE (1.4):

(3.1a)
$$u_t + f(u)_x = 0,$$

(3.1b)
$$u(x,0) = u^0(x).$$

If the cell average of u is defined by

(3.2)
$$\overline{u}(x,t) = \frac{1}{\Delta x} \int_{x-\Delta x/2}^{x+\Delta x/2} u(\xi,t) \, d\xi,$$

then (3.1) can be rewritten as

(3.3a)
$$\frac{\partial}{\partial t}\overline{u}(x,t) + \frac{1}{\Delta x} \left[f\left(u\left(x + \frac{\Delta x}{2}, t\right) \right) - f\left(u\left(x - \frac{\Delta x}{2}, t\right) \right) \right] = 0,$$

(3.3b) $\overline{u}(x,0) = \overline{u}^0(x).$

Hence a semidiscrete conservative scheme

(3.4)
$$\frac{d}{dt}\overline{u}_j = L(\overline{u})_j = -\frac{1}{\Delta x}(\hat{f}_{j+1/2} - \hat{f}_{j-1/2})$$

will be of high order if the numerical flux $\hat{f}_{j+1/2}$ approximates $f(u(x_j + \Delta x/2, t))$ to high order. This is the approach used in the MUSCL type semidiscrete finite difference TVD and ENO schemes [11], [4]. Notice that (3.4) is a scheme for the cell averages \overline{u}_j . However, in evaluating $\hat{f}_{j+1/2}$, which should approximate $f(u(x_j + \Delta x/2, t))$, we also need accurate point values $u_{j+1/2} = u(x_j + \Delta x/2, t)$. For finite difference schemes the reconstruction from cell averages to point values is a major issue and causes difficulties, especially in several space dimensions [4], [5]. For spectral methods, this is very simple because \overline{u} is just the convolution of u with the characteristic function of $(x_{j-1/2}, x_{j+1/2})$. To be precise, if

(3.5)
$$u(x) = \sum_{l=-N}^{N} a_l e^{ilx}$$

(we have suppressed the time variable t), then

(3.6)
$$\overline{u}(x) = \sum_{l=-N}^{N} \overline{a}_l e^{ilx}$$

with

(3.7)
$$\overline{a}_l = \sigma_l a_l, \qquad \sigma_l = \frac{\sin(l\Delta x/2)}{l\Delta x/2} \quad \text{for } 0 < |l| \le N, \ \sigma_0 = 1.$$

Notice that for collocation or Galerkin with $\Delta x = 2\pi/2N$, we have $|l\Delta x/2| \leq \pi/2$ for $|l| \leq N$, hence $2/\pi \leq \sigma_l \leq 1$. The division or multiplication by σ_l thus causes no stability difficulty. We point out that σ_l resembles the Lanczos filter [8, p. 65], which in our notations is $\sin(l\Delta x)/l\Delta x$, and approaches zero when $|l| \rightarrow N$.

The easy transform between u and \overline{u} is also valid in several space dimensions and for other spectral expansions (e.g., Chebyshev expansions). We omit the details.

We now state our scheme as (3.4) with

(3.8)
$$\hat{f}_{j+1/2} = f(v_N(x_{j+1/2}, t)),$$

where v_N is defined by (2.10). We obtain the Fourier coefficients \overline{a}_l of \overline{u} from $\{\overline{u}_j\}$ by collocation, and obtain a_l of u needed in (2.10) by (3.7). The main difference between the conventional spectral method and the current approach is that we use the essentially nonoscillatory reconstruction v_N instead of the oscillatory $P_N u$ in (3.8).

The scheme, as it stands, can only treat a solution of not more than one discontinuity. However, it can be easily generalized.

We remark that if u is smooth, (2.10) keeps spectral accuracy because A determined by (2.14) will be spectrally small.

To discretize (3.4) in time, we use the high-order TVD Runge-Kutta methods in [13]:

(3.9)
$$\overline{u}^{(i)} = \sum_{k=0}^{i-1} [\alpha_{ik}\overline{u}^{(k)} + \beta_{ik}\Delta t L(\overline{u}^{(k)})], \qquad i = 1, \dots, r,$$
$$\overline{u}^{(0)} = \overline{u}^n, \qquad \overline{u}^{(r)} = \overline{u}^{n+1}.$$

In Section 4, we use a third-order scheme r = 3 with $\alpha_{10} = \beta_{10} = 1$, $\alpha_{20} = \frac{3}{4}$, $\beta_{20} = 0$, $\alpha_{21} = \beta_{21} = \frac{1}{4}$, $\alpha_{30} = \frac{1}{3}$, $\beta_{30} = \alpha_{31} = \beta_{31} = 0$, $\alpha_{32} = \beta_{32} = \frac{2}{3}$. We use a small Δt so that the temporal error can be neglected. These methods are TVD (or TVB) if the Euler forward version of (3.4) is TVD (or TVB). In light of (2.61) we expect the total variation of (3.4)-(3.8)-(3.9) to grow at most at the rate of $O(\ln N)$. In practice, we observe stable results (Section 4).

In summary, a suggested algorithm can be:

(1) Starting with $\{\overline{u}_j\}$, compute its collocation Fourier coefficients $\{\overline{a}_l\}$ and the Fourier coefficients $\{a_l\}$ of u by (3.7).

- (2) Compute the shock location y and the shock strength A by (2.14).
- (3) Compute $v_N(x)$ by

$$v_N(x) = a_0 + \sum_{\substack{l=-N\\l\neq 0}}^{N} \tau_l \left(a_l - \frac{A}{il} e^{-ily} \right) e^{ilx} + F(x, y, A)$$

and a filtered $\{\overline{u}_i^*\}$ by

$$\overline{u}_j^* = \overline{a}_0 + \sum_{\substack{l=-N\\l\neq 0}}^N \tau_l \left(\overline{a}_l - \sigma_l \frac{A}{il} e^{-ily} \right) e^{ilx_j} + \overline{F}(x_j, y, A),$$

where τ_l is some filter, e.g. (1.2).

(4) Use $\hat{f}_{j+1/2} = f(v_N(x_{j+1/2}, t))$ in (3.4), and use

$$\overline{u}^{(i)} = \sum_{k=0}^{i-1} [\alpha_{ik} \overline{u}^{*(k)} + \beta_{ik} \Delta t L(\overline{u}^{(k)})], \qquad i = 1, \dots, r,$$

in (3.9).

As in the finite difference case [11], [12], we may also apply limiters to obtain provable TVB schemes while still keeping spectral accuracy. Let

where $u_{j+1/2} = v_N(x_{j+1/2}, t)$ in (3.8). We limit the increments by

(3.11)
$$\tilde{u}_{j}^{(\mathrm{mod})} = m(\tilde{u}_{j}, \Delta_{+}\overline{u}_{j}, \Delta_{-}\overline{u}_{j}), \qquad \tilde{\tilde{u}}_{j+1}^{(\mathrm{mod})} = m(\tilde{\tilde{u}}_{j+1}, \Delta_{+}\overline{u}_{j}, \Delta_{-}\overline{u}_{j}),$$

where m is the minmod function with TVB correction:

(3.12)
$$m(a_1,\ldots,a_k) = \begin{cases} a_1, & \text{if } |a_1| \le M\Delta x^2, \\ s \cdot \min_{1 \le i \le k} |a_i|, & \text{if } |a_1| > M\Delta x^2 \text{ and } \operatorname{sign}(a_i) = s \ \forall i, \\ 0, & \text{otherwise}, \end{cases}$$

with $M = \frac{2}{3}M_2$ or $M = M_j = \frac{2}{9}(3 + 10M_2)M_2 \cdot \Delta x^2/(\Delta x^2 + |\Delta_+\overline{u}_j| + |\Delta_-\overline{u}_j|)$. Here, M_2 is the maximum of $|u_{xx}^0|$ in some region around the smooth critical points of $u^0(x)$. See [12], [2].

The flux (3.8) is modified to

(3.13)
$$\hat{f}_{j+1/2} = h(\overline{u}_j + \tilde{u}_j^{(\text{mod})}, \overline{u}_{j+1} - \tilde{\tilde{u}}_{j+1}^{(\text{mod})}),$$

where h is any monotone flux [3]. We then have the following lemma.

LEMMA 6. Scheme (3.9)-(3.13) is TVB and formally spectrally accurate in space (i.e., the spatial local truncation error in smooth regions is spectrally small), if the filtering (1.3) is used.

Proof. The proof for TVB can be found in [11], [12]. By [1], the local truncation error is spectrally small in smooth regions if the limiter (3.11) returns the first argument. The proof that (3.11) always returns the first argument in smooth regions, including at critical points, can be found in [2]. \Box

We remark that the scheme (3.4)-(3.8)-(3.9), with or without the TVB limiting (3.11), yields almost identical results in our numerical examples (Section 4). This indicates the good stability property of the scheme (3.4)-(3.8)-(3.9). We also remark that (3.13) yields a TVB scheme regardless of the underlying method (3.4). However, accuracy in smooth regions may be lost if the underlying method (3.4) is globally oscillatory, because the limiters (3.11) may be enacted in *smooth* regions to counterbalance these spurious oscillations. Numerical examples in Section 4 verify these remarks. In [10], McDonald also used some limiters to obtain a TVD spectral scheme. However, the accuracy in smooth regions is questionable in view of the above remarks.

4. Numerical Results. We use several numerical examples to illustrate the methods introduced in the previous sections.

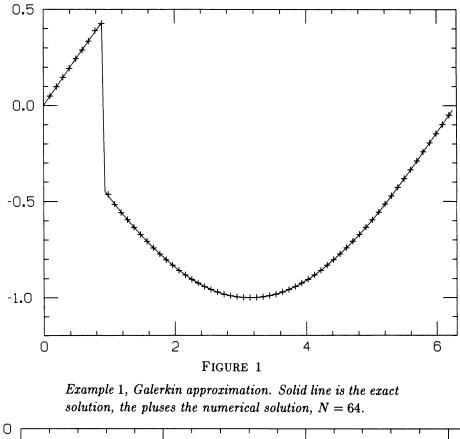
Example 1. We use the approximation (2.10)-(2.12)-(2.14) on the following function

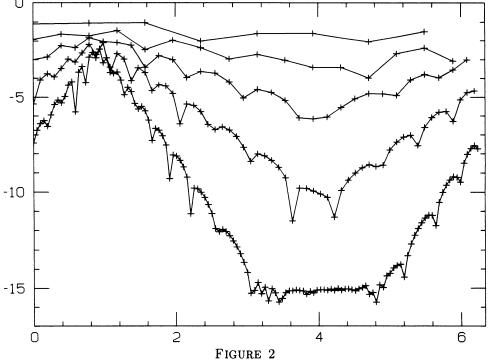
(4.1)
$$u(x) = \begin{cases} \sin \frac{x}{2}, & 0 \le x \le 0.9, \\ -\sin \frac{x}{2}, & 0.9 < x < 2\pi. \end{cases}$$

Notice that $[u^{(k)}] \neq 0$ for all $k \geq 0$. Both Galerkin and collocation methods are tested. Exponential filters (1.2) with m = 4, $k_0 = 0$ are used.

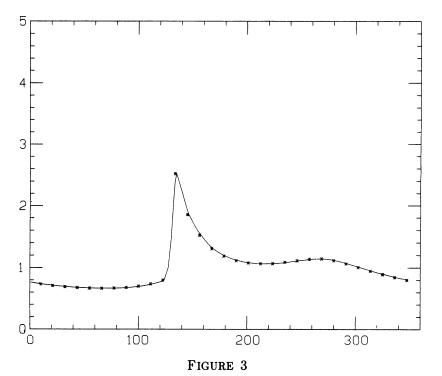
In Table 1, we list the errors of the shock location and shock strength determined by (2.14). Notice that the second-order accuracy (2.15) is verified.

Figure 1 displays the numerical solution of the Galerkin approximation (2.10)-(2.12)-(2.14) with N = 64. Figure 2 is the error of the approximation on a logarithm scale. We have found the same kind of results for collocation approximation. In Table 2 we list the L^1 error and numerical order in a smooth region (in this case, we define the smooth region to be 0.8 away from the discontinuity). We can see Galerkin and collocation have the same order of accuracy. There is no O(1) error near the discontinuity, overall we achieve $O(\log N/N^2)$ for L^1 convergence, verifying (2.62). For comparison we refer the reader to [7].





Example 1, error of the Galerkin approximation on logarithm scale for N = 8, 16, 32, 64, 128.



Galerkin approximation of (2.11) with N = 32 for the steady state solution of the astrophysics problem [6].

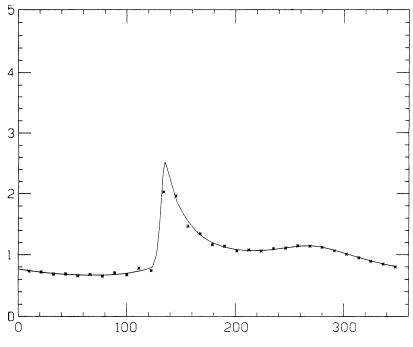


FIGURE 4

Usual Galerkin approximation for the steady state solution of the astrophysics problem, N = 32.

Example 2. We apply (2.10)-(2.12)-(2.14) on a discontinuous function which is the steady state solution of an astrophysics problem [6]. Figure 3 is v_N in (2.10) with N = 32. For comparison, Figure 4 is the usual Galerkin approximation $P_N u$ with N = 32. The improvement is apparent.

Example 3. We solve the Burgers' equation

(4.2)
$$u_t + \left(\frac{u^2}{2}\right)_x = 0,$$
$$u(x,0) = 0.3 + 0.7 \sin x,$$

using scheme (3.4)-(3.8)-(3.9) and (3.4)-(3.9)-(3.13). We find the shock location and strength with (2.14). In our computation, we find that the coefficients of modes in the range of $\sqrt{N} \sim N^{3/4}$ give us the best results to detect shock location and strength. It can be proven that in this range of modes (2.14) will not fail in the presence of possible transition points in the numerical solutions. The errors of (3.4)-(3.8)-(3.9) in smooth regions (1.6 away from shock when it appears), at t = 0.8 (before shock), t = 1.42 (when the shock just develops), and t = 2.00 (after shock) are listed in Table 3. The numerical solutions are displayed in Figures 5–6. The error at t = 2.00, in logarithm scale, is displayed in Figure 7.

We seem to observe higher than algebraic order in smooth regions both before and after the shock develops. This might be the first time superalgebraic accuracy is observed in a shock-capturing spectral scheme solving a nonlinear PDE with shocks. The usual O(1) Gibbs oscillation near the shock is also absent in all of our calculations. We also notice that the TVB limiter (3.11) does not change the numerical results significantly in the smooth region (see Table 4). Actually, we observe the same order of accuracy in the smooth region, comparing Table 4 with Table 3. This indicates that the scheme (3.4)-(3.8)-(3.9) is by itself very stable.

Finally, we run the usual spectral scheme (i.e., with v_N in (3.8) replaced by $P_N u$) with the TVB limiter (3.11). The errors in smooth regions (1.6 away from shock) are listed in Table 5 (compare with Table 4). Clearly we only get first-order accuracy in smooth regions after the shock develops. This indicates that TVB limiting can make a scheme stable but may not preserve the accuracy.

Example 4. 2-D Steady State. We solve a 2-dimensional scalar conservation law

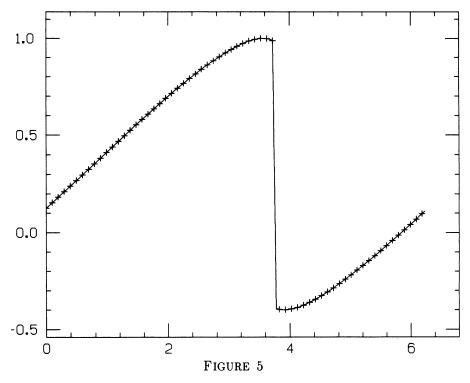
(4.3)
$$\begin{cases} u_t + (\frac{u^2}{2})_x + u_y = 0, & (x, y) \in [0, 2\pi] \times [-1, 1], \\ u(x, 0, t) = \sin x, \\ u(0, y, t) = u(2\pi, y, t), & y \in [-1, 1], t \ge 0. \end{cases}$$

We know that (4.3) has a steady state solution $u_{\infty}(x, y)$. $u_{\infty}(x, y)$ actually will be the solution to (4.2) if we replace t by y and set $u(x, 0) = \sin x$ in (4.2).

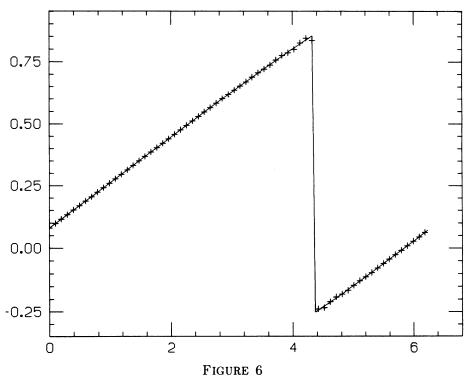
As mentioned in Section 3, (3.4) (3.8) (3.8) (3.9) can be extended to 2-dimensional cases and we can use either the Fourier or Chebyshev method in each of the spatial directions. To solve for the steady state of (4.3), we use the Fourier method in the *x*-direction and the Chebyshev method in the *y*-direction. The criterion we set for the steady state is that the relative L^1 residue between two consecutive time stages be less than 10^{-6} , i.e.,

(4.4)
$$\frac{\|u^{n+1}-u^n\|_{L^1}}{\|u^n\|_{L^1}} \le 10^{-6}.$$

4



Example 3, inviscid Burgers' equation with initial data $u(x,0) = 0.3+0.7*\sin(x)$, time t = 2.0, N = 64. Solid line is the exact solution, the pluses the numerical solution.



Example 3, same as Figure 5, except time t = 4.0.

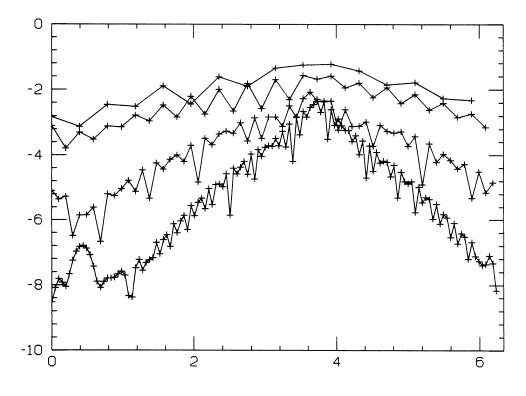


FIGURE 7

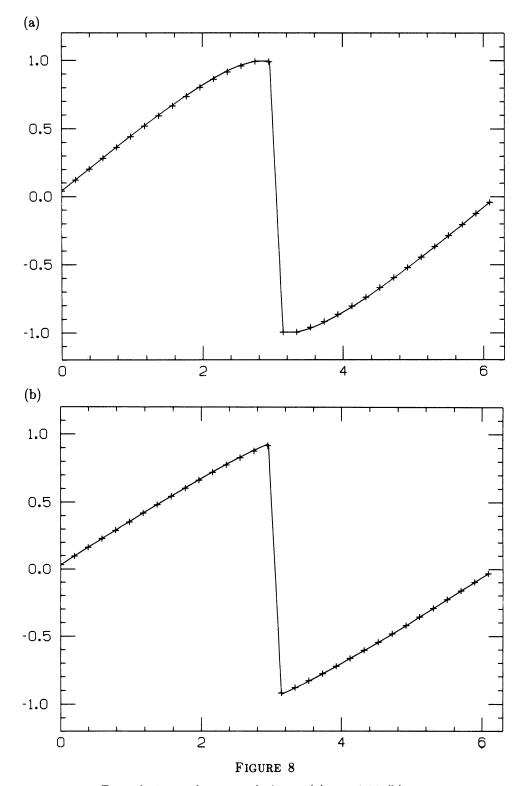
Example 3, inviscid Burgers' equation, $u(x, 0) = 0.3 + 0.7 * \sin(x)$. Errors of numerical solutions at time t = 2.00 in the logarithm scale for N = 16, 32, 64, 128.

Figure 8 displays the profile of the steady state at y = 0.38 and y = 1.00. We used 32 points in the x-direction and 8 points in the y-direction. Figure 9 is the contour plot for the numerical steady state solution.

TABLE 1

Errors of shock location and strength, Example 1.

	Galerkin				Collocation			
	Location		Strength		Location		Strength	
N	Error	Order	Error	Order	Error	Order	Error	Order
8	0.15(0)		0.12(-1)		0.36(0)		0.20(-1)	
16	0.24(-1)	2.6	0.22(-2)	2.4	-0.21(0)	0.8	0.12(-1)	0.7
32	0.49(-2)	2.3	0.48(-3)	2.2	-0.14(-1)	3.8	0.38(-2)	1.7
64	0.11(-2)	2.1	0.11(-3)	2.1	-0.32(-2)	2.2	0.11(-2)	1.8
128	0.26(-3)	2.1	0.27(-4)	2.1	-0.77(-3)	2.0	0.28(-3)	1.9



Example 4, steady state solution at (a) y = 0.38 (b) y = 1.0. Solid lines are the exact solution, the pluses the numerical solution.

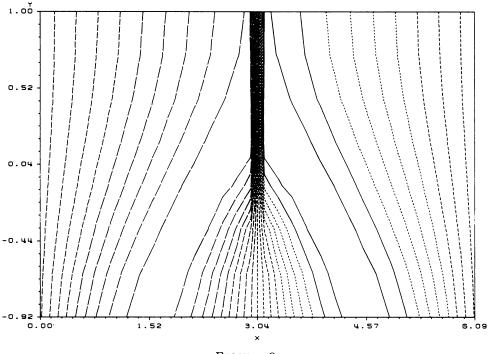


FIGURE 9

Example 4, contour plot of the steady state solution.

TABLE 2

 L^1 Error in Region I = { $x \in [0, 2\pi], |x - x_s| > 0.8$ }

and Region II = $[0, 2\pi]$, Example 1.

	Galerkin				Collocation			
	Region I		Region II		Region I		Region II	
Ν	Error	Order	Error	Order	Error	Order	Error	Order
8	0.32(-1)		0.14(0)		0.23(-1)		0.31(-1)	
16	0.32(-2)	3.30	0.75(-2)	4.27	-0.21(-2)	3.46	0.61(-2)	2.34
32	0.24(-3)	3.75	0.17(-2)	2.11	0.23(-3)	3.20	0.17(-2)	1.79
64	0.51(-5)	5.55	0.39(-3)	2.14	0.54(-5)	5.40	0.49(-3)	1.86
128	0.12(-7)	8.67	0.96(-4)	2.04	0.12(-7)	8.82	0.13(-3)	1.92

TABLE 3

Errors in smooth region for (4.2). At t = 0.8, the smooth region is $[0, 2\pi]$.

At t = 1.42, 2.0, the smooth region is 1.6 away from the shock.

	t = 0.8		t =	1.42	t = 2.0	
Ν	L^1 Error	Order	L^1 Error	Order	L^1 Error	Order
16			_			
	0.94(-2)		0.39(-2)		0.44(-2)	
32		3.57		1.66		1.40
	0.79(-3)		0.13(-2)		0.16(-2)	
64		5.00		5.17		5.28
	0.25(-4)		0.35(-4)		0.42(-4)	
128		7.58		7.71		6.58
	0.13(-6)		0.16(-6)		0.44(-6)	

TABLE 4

Errors in smooth regions for (4.2) of new spectral scheme with TVB limiting (3.11). For both t = 1.42 and 2.0, the L^1 errors are taken in the region 1.6 away from the shock.

	t =	1.42	t = 2.0		
N	L^1 Error	Order	L^1 Error	Order	
16					
	0.64(-2)		0.63(-2)		
32		1.90		1.67	
	0.17(-2)		0.20(-2)		
64		5.55		5.29	
	0.36(-4)		0.50(-4)		
128		7.75		7.11	
	0.17(-6)		0.36(-6)		

TABLE 5

Errors in smooth regions for (4.2) of the usual spectral scheme with TVB limiting (3.11). For both t = 1.42 and 2.0, the L^1 errors are taken in the region 1.6 away from the shock.

	t =	1.42	t = 2.0		
Ν	L^1 Error	Order	L^1 Error	Order	
16					
	0.25(-1)		0.16(-1)		
32		0.98		*	
	0.98(-2)		0.17(-1)		
64		1.50		1.17	
	0.34(-2)		0.79(-2)		
128		0.84		0.76	
	0.19(-2)		0.47(-2)		

Division of Applied Mathematics Brown University Providence, Rhode Island 02912 *E-mail* (Gottlieb): am420000@brownvm.bitnet

1. S. ABARBANEL, D. GOTTLIEB & E. TADMOR, "Spectral methods for discontinuous problems," in *Numerical Methods for Fluid Dynamics*, II (K. Morton and M. Baines, eds.), Oxford University Press, New York, 1986, pp. 129–153.

2. B. COCKBURN & C.-W. SHU, TVB Runge-Kutta Local Projection Discontinuous Galerkin Finite Element Methods for Conservation Laws II: General Framework, IMA Preprint Series #392, University of Minnesota, 1988; Math. Comp., this issue, pp. 411-435.

3. M. CRANDALL & A. MAJDA, "Monotone difference approximations for scalar conservation laws," Math. Comp., v. 34, 1980, pp. 1-21.

4. A. HARTEN, B. ENGQUIST, S. OSHER & S. CHAKRAVARTHY, "Uniformly high order accurate non-oscillatory schemes, III," J. Comput. Phys., v. 71, 1987, pp. 231-303.

5. A. HARTEN, Preliminary Results on the Extension of ENO Schemes to Two-Dimensional Problems, Proc. Internat. Conf. on Hyperbolic Problems, Saint-Etienne, January 1986.

6. M. HUSSAINI, D. KOPRIVA, M. SALAS & T. ZANG, "Spectral method for Euler equation, Part 1: Fourier method and shock capturing," AIAA J., v. 23, 1985, pp. 234-240.

7. D. KOPRIVA, "A practical assessment of spectral accuracy for hyperbolic problems with discontinuity," J. Sci. Comput., v. 2, 1987, pp. 249-262.

8. C. LANCZOS, Discourse on Fourier Series, Oliver & Boyd, Edinburgh, 1966.

9. A. MAJDA, J. MCDONOUGH & S. OSHER, "The Fourier method for nonsmooth initial data," *Math. Comp.*, v. 32, 1978, pp. 1041-1081.

10. B. MCDONALD, "Flux corrected pseudospectral methods for scalar hyperbolic conservation laws," J. Comput. Phys. (To appear.)

11. S. OSHER, "Convergence of generalized MUSCL schemes," SIAM J. Numer. Anal., v. 22, 1985, pp. 947-961.

12. C.-W. SHU, "TVB uniformly high-order schemes for conservation laws," Math. Comp., v. 49, 1987, pp. 105-121.

13. C.-W. SHU & S. OSHER, "Efficient implementation of essentially non-oscillatory shock-capturing schemes," J. Comput. Phys., v. 77, 1988, pp. 439-471.

14. A. ZYGMUND, Trigonometric Series, v. 1, Cambridge Univ. Press, New York, 1959.