Journal of Scientific Computing, Vol. 10, No. 3, 1995

A Note on the Accuracy of Spectral Method Applied to Nonlinear Conservation Laws[1]

Chi-Wang Shu[2] and Peter S. Wong[2]

Received July 26, 1994

Fourier spectral method can achieve exponential accuracy both on the approximation level and for solving partial differential equations if the solutions are analytic. For a linear PDE with discontinuous solutions, Fourier spectral method will produce poor point-wise accuracy without post-processing, but still maintains exponential accuracy for all moments against analytic functions. In this note we assess the accuracy of Fourier spectral method applied to nonlinear conservation laws through a numerical case study. We have found out that the moments against analytic functions are no longer very accurate. However the numerical solution does contain accurate information which can be extracted by a Gegenbauer polynomial based post-processing.

KEY WORDS: Spectral method; accuracy; Gibbs phenomenon; nonlinear conservation laws.

1. INTRODUCTION

In this note we are concerned with the accuracy of Fourier spectral method when applied to a nonlinear conservation law

$$\begin{aligned} &\partial_t u + \partial_x f(u) = 0, \qquad -1 \leqslant x < 1 \\ &u(x, 0) = u^0(x) \end{aligned} \tag{1.1}$$

where the initial condition $u^0(x)$ is 2-periodic. As is well known, solutions to Eq. (1.1) typically contain discontinuities even if the initial condition

[1] Research supported by ARO Grant DAAL03-91-G-0123 and DAAH04-94-G-0205, NSF Grant DMS-9211820, NASA Grant NAG1-1145 and contract NAS1-19480 while the first author was in residence at ICASE, NASA Langley Research Center, Hampton, Virginia 23681-0001, and AFOSR Grant 93-0090.

[2] Division of Applied Mathematics, Brown University, Providence, Rhodes Island 02912.

854/10/3-5

0885-7474/95/0900-0357$07.50/0 © 1995 Plenum Publishing Corporation

$u^0(x)$ is analytic (in this paper, for simplicity of presentations, we will use analytic functions to represent general smooth functions. Similar results can also be obtained for C^k or C^∞ functions). The purpose of this note is to assess accuracy under such situation through a numerical case study.

We start by recalling the Fourier approximation operator S_N to an L^2 function $u(x)$:

$$S_N u(x) = \sum_{k=-N}^{N} \hat{u}_k e^{ik\pi x} \tag{1.2}$$

where the Fourier coefficients $\hat{u}_k$ are defined by

$$\hat{u}_k = \tfrac{1}{2} \int_{-1}^{1} u(x)\, e^{-ik\pi x}\, dx \tag{1.3}$$

for Fourier Galerkin, and by

$$\hat{u}_k = \frac{1}{2N+1} \sum_{j=-N}^{N} u(x_j)\, {}^{-ik\pi x_j}, \qquad x_j = \frac{2j}{2N+1} \tag{1.4}$$

for Fourier collocation. We will also use the notation $u_N(x) = S_N u(x)$. To solve the partial differential Eq. (1.1), the standard Fourier spectral algorithm is

$$\begin{gathered} S_N(\partial_t v_N + \partial_x f(v_N)) = 0, \qquad -1 \leqslant x < 1 \\ v_N(x, 0) = S_N u^0(x) \end{gathered} \tag{1.5}$$

where $v_N(x, t) = \sum_{k=-N}^{N} \hat{v}_k(t)\, e^{ik\pi x}$ is supposed to approximate the exact solution $u(x, t)$ of Eq. (1.1), and S_N is either the Galerkin or the collocation Fourier approximation operator defined by Eqs. (1.2) and (1.3) or by Eqs. (1.2)–(1.4).

The approximation error

$$u(x) - S_N u(x) \tag{1.6}$$

is well known to be exponentially small (i.e., it is of the size $O(r^N)$ for some $0 < r < 1$) if $u(x)$ is analytic. However, if $u(x)$ is only piecewise analytic but dicontinuous, the approximation error (1.6) is $O(1)$ near the discontinuity and only first order (i.e., it is of the size $O(1/N)$) elsewhere. This is known as the Gibbs phenomenon. See, e.g., Gottlieb and Orszag (1977) and Canuto *et al.* (1988). Fortunately, even if the accuracy is poor in the pointwise sense, it is still excellent for the moments against any analytic func-

tions. For any L^2 function $u(x)$ and any analytic function $w(x)$, we have Gottlieb and Tadmor (1985):

$$\left| \int_{-1}^{1} (u(x) - u_N(x))\, w(x)\, dx \right| \leqslant Cr^N \tag{1.7}$$

for some constant C and $0 < r < 1$. This property is the basis of all the "reconstruction" or "post-processing" techniques. These techniques try to recover exponential or at least high order accuracy for point values based on the Fourier approximation $S_N u(x)$ of a piecewise analytic function. In other words, one tries to obtain a small post-processed approximation error

$$u(x) - P_N S_N u(x) \tag{1.8}$$

where P_N is some post-processing operator. Examples of P_N include various high frequency filters Madja *et al.* (1978); Kreiss and Oliger (1979); Vandeven (1991) and Cai *et al.* (1992); which are of the form

$$P_N S_N u(x) = \sum_{k=-N}^{N} \sigma\left(\frac{k}{N}\right) \hat{u}_k e^{ik\pi x} \tag{1.9}$$

with $S_N u(x)$ given by Eq. (1.2). The function $\sigma(\xi)$ in Eq. (1.9) is even (or satisfies $\sigma(-\xi) = \overline{\sigma(\xi)}$ if it is complex valued as in Cai *et al.* (1992)), smooth (the accuracy of the filter depends upon its smoothness), supported in $(-1, 1)$ and satisfies $\sigma(0) = 1$ and $\sigma^{(k)}(0) = 0$ for $1 \leqslant k \leqslant K$ (with accuracy of the filter again depends upon K). These filters can recover high order or even exponential accuracy in the smooth regions away from the discontinuities (the filter by Cai *et al.* (1992) can also recover high order accuracy up to the discontinuity from one side). A more recent example of P_N is the Gegenbauer polynomial based procedure discussed by Gottlieb *et al.* (1992) and Gottlieb and Shu (1993, 1994, 1994a, 1995), which can give uniform exponential accuracy for all x right up to the discontinuity for piecewise analytic functions. In this sense spectral Fourier approximation is also exponentially accurate for piecewise analytic functions—one only has to extract the hidden information from the poor approximation $S_N(x)$ using the post-processor P_N.

When spectral method is used to solve the PDE in Eq. (1.1), we can consider the following different types of errors. The strongest is the pointwise error from the exact solution $u(x, t)$:

$$u(x, t) - v_N(x, t) \tag{1.10}$$

which cannot be small even for $t=0$ due to the Gibbs phenomenon. A more reasonable error is the point-wise error of the numerical solution $v_N(x, t)$ from the Fourier approximation of the exact solution $u_N(x, t)$:

$$u_N(x, t) - v_N(x, t) \tag{1.11}$$

If this error is exponentially small, we can claim the spectral method for Eq. (1.1) is exponentially accurate because of the post-processor Eq. (1.8) for the exact solution $u(x, t)$. An even weaker error is defined by the error in the first few Fourier coefficients, i.e.

$$\hat{u}_k(t) - \hat{v}_k(t) \tag{1.12}$$

for the first few k, where $\hat{u}_k(t)$ are the Fourier coefficients of the exact solution $u(x, t)$ of Eq. (1.1). This is actually an example of the more general definition of error in moments against any analytic function $w(x)$:

$$\int_{-1}^{1} (u_N(x) - v_N(x))\, w(x)\, dx \tag{1.13}$$

In fact, as long as this error in moments is exponentially small, we can claim that the spectral method is exponentially accurate in solving Eq. (1.1) by using property of Eq. (1.7) for the exact solution $u(x, t)$ and the post-processing of Eq. (1.8).

If the PDE (1.1) is linear (i.e., $f(u) = a(x, t)u$), it is proven by Gottlieb and Tadmor (1985), Abarbanel *et al.* (1986) that spectral Fourier method is exponentially accurate in the sense that Eq. (1.13) is exponentially small. A post-processing Eq. (1.8) applied to $v_N(x, t)$ would then yield an exponentially accurate pointwise approximation to the exact solution $u(x, t)$. However, if Eq. (1.1) is nonlinear, it is still a theoretically open problem whether spectral Fourier method, equipped with either high frequency filtering or vanishing viscosity Tadmor (1989); Maday and Tadmor (1989), is exponentially (or high order) accurate in the sense of Eq. (1.13). Computational evidence in Maday *et al.* (1993) seems to suggest that, even in this nonlinear case, highly accurate information is still implicitly contained in the numerical solution and can be extracted (at least away from the discontinuity) by a post-processing using high frequency filtering. In the next section we will perform a detailed numerical case study about this accuracy issue for Burgers' equation ($f(u) = u^2/2$). We use a high frequency solution filter to stablize the algorithm, and post process the numerical result using the Gegenbauer polynomial based procedure Gottlieb *et al.* (1992); Gottlieb and Shu (1994a). We observe that the spectral Fourier method is not very accurate in the sense of moments against analytic

functions Eq. (1.13). However, numerical evidence does indicate the possibility of very high accuracy under some weaker definition of accuracy, perhaps some average of Fourier coefficients, since the post-processed result $P_N v_N(x, t)$ is much more accurate than the Fourier coefficients themselves, and accurate Fourier coefficients can be "reconstructed" for this post-processed solution $P_N v_N(x, t)$.

2. A NUMERICAL CASE STUDY ABOUT ACCURACY

In our numerical solution reported in this section, time discretization is by a third order Runge-Kutta method, with a time step Δt sufficiently small such that spatial error is dominant in all cases. We compute the exact solutions of the PDE by Newton iterations on the implicit characteristic equations, and compute the Fourier coefficients of a function (if not analytically given) by using a sufficiently accurate numerical quadrature.

We first solve a linear equation

$$\begin{gathered} \partial_t u + \frac{3}{5 - 4\cos(\pi x)} \partial_x u = 0, \qquad -1 \leqslant x < 1 \\ u(x, 0) = x \end{gathered} \tag{2.1}$$

with periodic boundary conditions, up to $t = 1$, using the Fourier Galerkin method:

$$\begin{gathered} S_N\left(\partial_t v_N + \frac{3}{5 - 4\cos(\pi x)} \partial_x v_N\right) = 0 \\ v_N(x, 0) = S_N x = \sum_{\substack{k=-N \\ k \neq 0}}^{N} \frac{(-1)^k i}{k\pi} e^{ik\pi x} \end{gathered} \tag{2.2}$$

Standard Galerkin method is stable for this linear problem but produces poor point value accuracy (Fig. 1, top). However, the accuracy in the first few Fourier coefficients, as representatives of moments against analytic functions, are computed more accurately (Fig. 1, bottom).

In order to compare with the nonlinear case reported later, we solve the same linear equation in Eq. (2.1) using the filtered Fourier method, i.e., after each Runge-Kutta time step, the numerical solution is filtered by Eq. (1.9) with the exponential filter:

$$\sigma(\xi) = e^{-\alpha |\xi|^r} \tag{2.3}$$

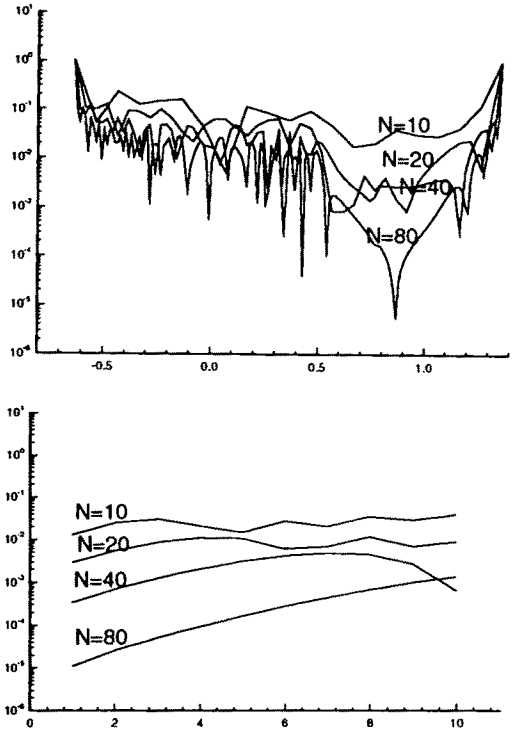

Fig. 1. Errors in log scale, linear PDE (2.1). Fourier Galerkin using $2N+1$ modes, for $N=10$; $N=20$; $N=40$ and $N=80$. Top: Point-wise errors; Bottom: errors in the first 10 Fourier coefficients.

where r is increasing with N and is related to the order of the filter, and α is chosen such that $e^{-\alpha}$ equals machine zero (10^{-16} for double precision). The exponential filter in Eq. (2.3) has the advantage of simplicity, and usually it works equally well as more complicated filters Vandeven (1991). For this linear problem, as well as for the nonlinear Burgers' equation later, we will use the Fourier method with the following choice of filter orders: $r=4$ for $N=10$; $r=6$ for $N=20$; $r=8$ for $N=40$ and $r=12$ for $N=80$. The result is shown in Fig. 2. Comparing with Fig. 1, we can see better point value accuracy in the smooth region because of the filters, and similar (good) accuracy for the first few Fourier coefficients.

The computational result for the linear equation is not surprising since it just shows the proven fact Gottlieb and Tadmor (1985), Abarbanel *et al.* (1986) that Fourier coefficients, as representatives of moments against analytic functions, are computed with exponential accuracy by the spectral Fourier method, and filtering will recover exponential point value accuracy

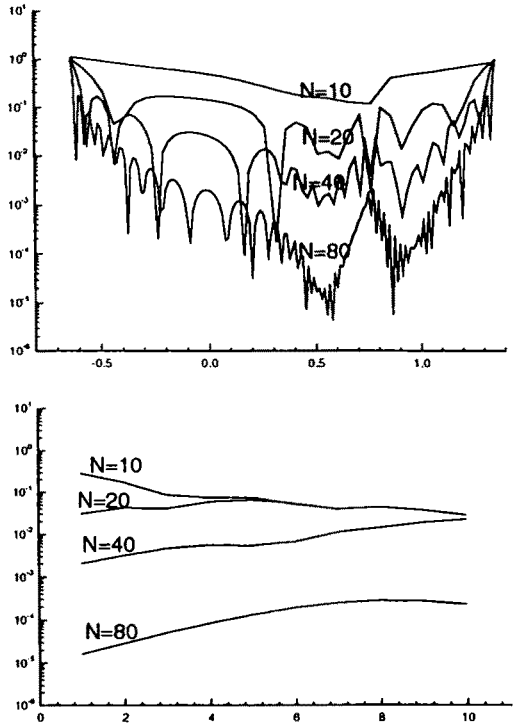

Fig. 2. Errors in log scale, linear PDE (2.1). Fourier Galerkin using $2N+1$ modes with exponential solution filters of order r. $r=4$ for $N=10$; $r=6$ for $N=20$; $r=8$ for $N=40$ and $r=12$ for $N=80$. Top: point-wise errors; Bottom: errors in the first 10 Fourier coefficients.

in smooth regions away from the discontinuity. It should be noticed that, for the same N, the accuracy for the first few Fourier coefficients is at the same level at or better than the best point value accuracy in the smooth region after filtering. This is again not surprising since point value accuracy is obtained from the Fourier coefficients through filtering.

We now come to the nonlinear problem we are really interested in: we solve the nonlinear Burgers' equation

$$\begin{aligned} &\partial_t u + \partial_x\left(\frac{u^2}{2}\right) = 0, \qquad -1 \leqslant x < 1 \\ &u(x, 0) = 0.3 + 0.7\sin(\pi x) \end{aligned} \tag{2.4}$$

The solution develops a shock at $t = 1/0.7\pi$ and we compute the solution up to $t = 1$. The initial condition is chosen such that shock is moving with time. For this nonlinear PDE, the standard Galerkin method cannot

converge to the entropy solution Tadmor (1989). One would need to add dissipations either by the high frequency solution filtering Eq. (1.9) or by the spectral vanishing viscosity Tadmor (1989); Maday and Tadmor (1989); and Maday *et al.* (1993). Numerical results for the Burgers' equation with the vanishing viscosity method can be found in, e.g., Maday *et al.* (1993). Here we will only report the results obtained by solution filtering, using the same r as in the previous linear case Eq. (2.1). We have also computed with the vanishing viscosity methods and have obtained similar results.

In Fig. 3, we plot the point-wise error $u(x, t) - v_N(x, t)$ (top), and the error for the first 10 Fourier coefficients (bottom). While the pattern of the point-wise errors are similar to the linear case in Fig. 2, the errors for the Fourier coefficients are clearly much worse in comparison. As a matter of fact, for the same N, the errors for the first few Fourier coefficients are a few magnitudes larger than the smallest point value error in the smooth

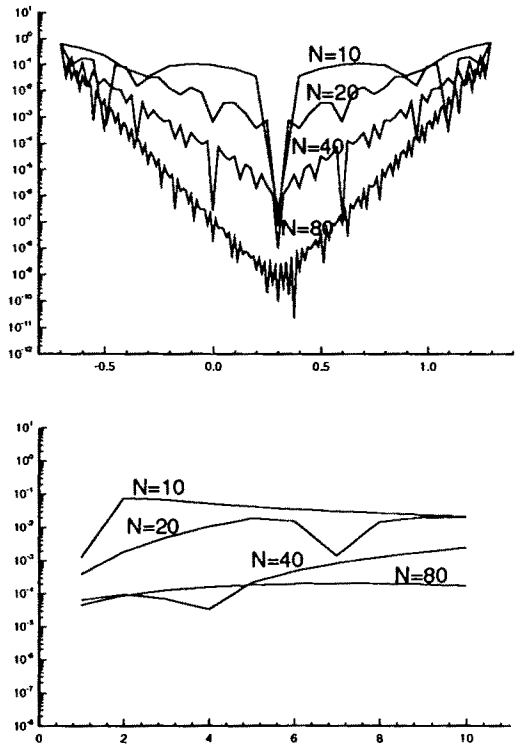

Fig. 3. Errors in log scale, Burgers Eq. (2.4). Fourier Galerkin using $2N+1$ modes with exponential solution filters of order r. $r=4$ for $N=10$; $r=6$ for $N=20$; $r=8$ for $N=40$ and $r=12$ for $N=80$. Tip: point-wise errors; Bottom: errors in the first 10 Fourier coefficients.

region. This is clearly different from what we observe in the linear case in Fig. 2, and suggests that the first few Fourier coefficients, again as representatives of moments against analytical functions, are no longer computed with exponential or high order accuracy. It is sort of puzzling that each difference in the Fourier coefficients $\hat{u}_k(t) - \hat{v}_k(t)$ is relatively large (Fig. 3 bottom), but the point-wise error $u(x, t) - v_N(x, t)$, which is just an average of $\hat{u}_k(t) - \hat{v}_k(t)$ (against an $O(1)$ function $e^{ik\pi x}$), is much smaller in the smooth region (Fig. 3 top). Clearly some cancellation is present.

Next, we apply the Gegenbauer post-processor Gottlieb *et al.* (1992) to $v_N(x, t)$. This procedure can be roughly described as follows: given the Fourier partial sum $u_N(x)$ of an analytic but not periodic function $u(x)$, one first finds the approximations to the first m Gegenbauer expansion coefficients of the function $u(x)$. Here Gegenbauer polynomials are orthogonal polynomials in $[-1, 1]$ under the weight function $(1-x^2)^{\lambda-1/2}$. One then uses this Gegenbauer series with the first m terms to approximate $u(x)$ everywhere in $[-1, 1]$. To use this procedure, one must know the location of the discontinuity (however, the procedures by Gottlieb and Shu (1993) allows one to handle the case where the location is not known exactly), and to choose the parameters λ and m. It is proven by Gottlieb *et al.* (1992) that when m and λ are both chosen proportional to (but less than) N, the reconstructed point values are exponentially accurate everywhere inside $[-1, 1]$. Thus Gibbs phenomenon is completely removed. The details can be found in Gottlieb *et al.* (1992); Gottlieb and Shu (1993, 1994, 1994a, 1995), and unpublished works.

We would like to point out that there is no theoretical justification in doing this post-processing for the current nonlinear case, since the post-processing procedure assumes that the Fourier coefficients are accurate, which is not true any more. However, the post-processed result is surprisingly good (Fig. 4). Just like in the approximation test cases Gottlieb *et al.* (1992). We can observe good accuracy everywhere including at the discontinuity $x = \pm 1 + 0.3$. The reconstructed Fourier coefficients, namely the Fourier coefficients of $P_N v_N(x, t)$, are much more accurate than before the post-processing (compare Fig. 4 bottom with Fig. 3 bottom).

This suggests that, even if $v_N(x, t)$ or its Fourier coefficients $\hat{v}_k(t)$ are not very accurate, it contains accurate information which is extracted in this case by the Gegenbauer polynomial based post-processor P_N. This numerical evidence suggests that in the nonlinear PDE case, Fourier coefficients $\hat{v}_k(t)$, just like point-wise values in the linear (or nonlinear) PDE case, are no longer good indicators of accuracy. They themselves are not very accurate, but they implicitly contain accurate information which can be extracted by adequate post-processors P_N. This accurate information might be contained in some averages of the Fourier coefficients (since the

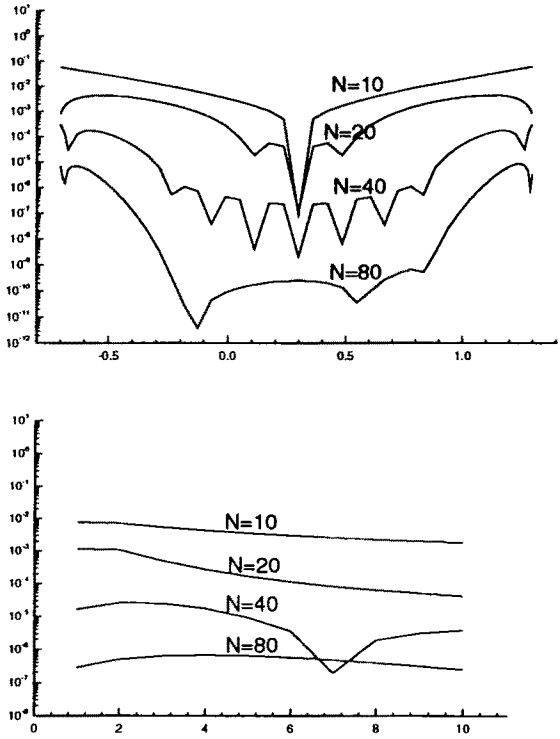

Fig. 4. Errors in log scale, Burgers Eq. (2.4). Fourier Galerkin using $2N+1$ modes with exponential solution filters of order r. $r=4$ for $N=10$; $r=6$ for $N=20$; $r=8$ for $N=40$ and $r=12$ for $N=80$. Gegenbauer post-processed, with parameters $\lambda=2$, $m=1$ for $N=10$; $\lambda=3$, $m=3$ for $N=20$; $\lambda=12$, $m=7$ for $N=40$ and $\lambda=62$, $m=15$, for $N=80$. Top: point-wise errors; Bottom: errors in the first 10 Fourier coefficients.

post-processing procedure based on Gegenbauer polynomials Gottlieb *et al.* (1992) uses certain averages of Fourier coefficients rather than the coefficients themselves).

We finally make two remarks:

Remark 2.1. In the previous Gegenbauer reconstruction procedure, we have used the exact shock location. The procedure by Gottlieb and Shu (1993) allows us to use an approximate shock location, determined from the Fourier coefficients themselves [e.g., Cai *et al.* (1992)]. Similarly good results can be obtained when the reconstruction is performed in a slightly smaller sub-interval inside which the solution is guaranteed to be analytic. For example, we use the shock detector by Cai *et al.* (1992), which in this

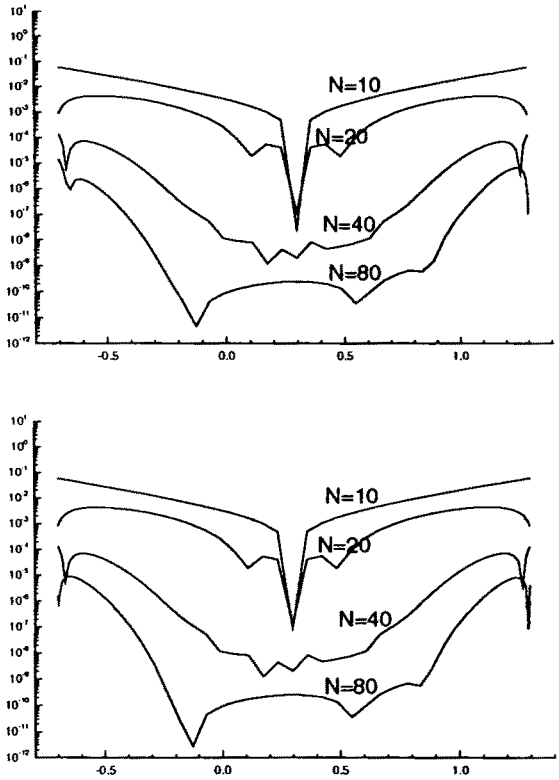

Fig. 5. Point-wise Errors in log scale, Burgers Eq. (2.4). Fourier method using $2N+1$ modes with exponential solution filters of order r. $r=4$ for $N=10$; $r=6$ for $N=20$; $r=8$ for $N=40$ and $r=12$ for $N=80$. Top: Galerkin, Gegenbauer post-processed with a numerically determined shock location using the techniques in [2], which for this problem produce shock locations to within 0.0000025 for all the N used. The reconstruction subinterval is $[-0.999997, 0.999997]$ when the numerical shock is shifted to $x=-1$. Parameters: $\lambda=2$, $m=1$ for $N=10$; $\lambda=3$, $m=3$ for $N=20$; $\lambda=26$, $m=9$ for $N=40$ and $\lambda=52$, $m=17$, for $N=80$. Bottom: collocation. Gegenbauer post-processed, with parameters $\lambda=2$, $m=1$ for $N=10$; $\lambda=3$, $m=3$ for $N=20$; $\lambda=26$, $m=9$ for $N=40$ and $\lambda=60$, $m=15$, for $N=80$.

case detects the shock location to within 0.0000025 for all the N values used, and a reconstruction inside the sub-interval $[-0.999997, 0.999997]$, which is just slightly smaller than $[-1, 1]$ (when numerically detected shock is shifted to $x=-1$) and guarantees that the true shock is outside this region. The result is shown in Fig. 5 (top). It is clearly as good as the case where the exact shock location is used (compare with Fig. 4 top).

Remark 2.2. If we use collocation in Eq. (1.4) instead of Galerkin [for the reconstruction procedure, see Gottlieb and Shu (1994a)], the result is almost identically good: Compare Fig. 5 (bottom) with Fig. 4 (top).

3. CONCLUDING REMARKS

Through a careful numerical case study for the Burgers' equation, we have found that the Fourier spectral method, equipped with spectrally small dissipations in the form of high frequency filters or vanishing viscosities, are not accurate in the first few Fourier coefficients, or in moments against smooth functions. However, accurate information is indeed contained in the numerical solution, and can be extracted by using the Gegenbauer polynomial based post-processor by Gottlieb *et al.* (1992) and Gottlieb and Shu (1993, 1994, 1994a, 1995).

REFERENCES

Abarbanel, A., Gottlieb, D., and Tadmor, E. (1986). Spectral methods for discontinuous problems, in Morton, W., and Baines, M. J. (eds.), *Numerical Methods for Fluid Dynamics II*, Oxford University Press, London, pp. 129–153.

Cai, W., Gottlieb, D., and Shu, C.-W. (1992). On One-Sided Filters for Spectral Fourier Approximations of Discontinuous Functions, *SIAM J. Numer. Anal.* **29**, 905–916.

Canuto, C., Hussaini, M. Y., Quarternoni, A., and Zang, T. A. (1988). *Spectral Methods in Fluid Dynamics*, Springer-Verlag.

Gottlieb, D., and Orszag, S. (1977). *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM-CBMS, Philadelphia.

Gottlieb, D., and Tadmor, E. (1985). Recovering Pointwise Values of Discontinuous Data Within Spectral Accuracy, in Murman, E. M., and Abarbanel, S. S. (eds.), *Progress and Supercomputing in Computational Fluid Dynamics*, Birkhäuser, Boston, pp. 357–375.

Gottlieb, D., and Shu, C.-W. (1994). Resolution properties of the Fourier method for discontinuous waves, *Meth. Appl. Mech. Engin.* **116**, 27–37.

Gottlieb, D., and Shu, C.-W. (1993). On the Gibbs Phenomenon III: Recovering exponential accuracy in a sub-interval from the spectral partial sum of a piecewise analytic function, ICASE Report No. 93-82, NASA Langley Research Center, *SIAM J. Numer. Anal.* (to appear).

Gottlieb, D., and Shu, C-W. (1995). On the Gibbs Phenomenon IV: Recovering exponential accuracy in a sub-interval from the Gegenbauer partial sum of a piecewise analytic function, *Math. Comp.* **64**, 1081–1095.

Gottlieb, D., and Shu, C.-W. (1994a). On the Gibbs Phenomenon V: Recovering exponential accuracy from collocation point values of a piecewise analytic function, ICASE Report 94-61, NASA Langley Research Center, *Numer. Math.*, to appear.

Gottlieb, D., Shu, C.-W., Solomonoff, A., and Vandeven, H. (1992). On the Gibbs Phenomenon I: recovering exponential accuracy from the Fourier partial sum of a non-periodic analytic function, *J. Comput. Appl. Math.* **43**, 81–92.

Kreiss, H., and Oliger, J. (1979). Stability of the Fourier Method, *SIAM J. Numer. Anal.* **16**, 421–433.

Maday, Y., and Tadmor, E. (1989). Analysis of the spectral vanishing viscosity method for periodic conservation laws, *SIAM J. Numer. Anal.* **26**, 854–870.

Maday, Y., Ould Kaber, S., and Tadmor, E. (1993). Legendre pseudospectral viscosity method for nonlinear conservation laws, *SIAM J. Numer. Anal.* **30**, 321–342.

Madja, A., McDonough, J., and Osher, S. (1978). The Fourier Method for Nonsmooth Initial Data, *Math. Comput.* **32**, 1041–1081.

Tadmor, E. (1989). Convergence of spectral methods for nonlinear conservation laws, *SIAM J. Numer. Anal.* **26**, 30–44.

Vandeven, H. (1991). Family of Spectral Filters for Discontinuous Problems, *J. Sci. Comput.* **8**, 159–192.

Printed in Belgium
Verantwoordelijke uitgever:
Hubert Van Maele
Altenastraat 20 - B-8310 St.-Kruis