JOURNAL OF COMPUTATIONAL PHYSICS **125,** 1–15 (1996)
ARTICLE NO. 0075

On Nonsmooth Solutions of Linear Hyperbolic Systems

KNUT S. ECKHOFF AND JENS H. ROLFSNES

Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, N-5008 Bergen, Norway

Received July 15, 1994; revised May 22, 1995

Initial value problems for linear hyperbolic systems with smooth 2π-periodic coefficients are solved numerically by a modified Fourier–Galerkin method when the initial values are nonsmooth. The described approach is seen to give substantially improved accuracy compared to more traditional methods. The discontinuities are accurately resolved already on coarse grids, and the fine-structure of structured solutions is resolved on relatively coarse grids as well. The accuracy is seen to be of high order and, even for very long term integrations, the global error can be kept very small if the grid is sufficiently refined.

1. INTRODUCTION

It is well known that traditional spectral methods have to be modified if satisfactory results shall be obtainable for linear hyperbolic partial differential equations with discontinuous initial data [17]. The major cause of the problems is the oscillatory behaviour of a truncated expansion near a discontinuity, known as the Gibbs phenomenon. Various filtering techniques for curing this deficiency of spectral methods have been suggested in the literature [4]. The results obtained by methods utilizing step-functions in the reconstruction of discontinuous functions have been particularly promising in this connection. The approach utilizing step-functions was initiated by Gottlieb *et al.* [14] and has been further developed in [1–3, 8–10, 13]. In the present paper we shall consider further improvements obtainable by such methods.

More specifically, we consider applications of the modified Fourier method presented in [10] to the problem of obtaining accurate numerical solutions of well-posed initial value problems with nonsmooth initial data for linear hyperbolic systems of the form

$$L[\mathbf{u}] = \mathbf{u}_t + A(x,t)\mathbf{u}_x + B(x,t)\mathbf{u} = \mathbf{0}, \tag{1}$$

where $\mathbf{u} = \{u_1, ..., u_m\}^{\mathrm{T}}$ are the dependent variables which we shall allow to be complex-valued, while A, B are given $m \times m$ matrices with smooth coefficients. We shall restrict ourselves to consider the case where the solutions, as well as all the coefficients in (1), are 2π-periodic with respect to the spatial variable x. Since we are concerned with discontinuous initial values associated with (1), we have to consider generalized solutions in the form of generalized functions (or distributions) [5; 6; 4, Appendix A]. In this paper, however, the generalized solutions considered for (1) will for each t be assumed to be piecewise smooth with respect to x.

Utilizing Fourier methods, numerical solutions of (1) subject to discontinuous initial data have been studied, for instance by Majda *et al.* [17]. In [17] it was proved that by application of appropriate smoothing (or filtering) techniques, the Fourier-collocation method could be modified such that spectral accuracy is obtainable in regions where the solution is smooth, while regions of low accuracy due to the Gibbs phenomenon are localized to small neighbourhoods of the discontinuities. The idea of utilizing filters has been further refined in a number of later papers, e.g. [2, 19]. At least away from the discontinuities, the main conclusion is that for a stable algorithm, the solution for the Fourier–Galerkin approximation contains sufficient information that the pointwise values of the exact solution can be recovered within spectral accuracy by application of a proper postprocessing filter. As partly discussed in [4, Section 8.3], however, these methods in practice give a relatively broad region of large error near the discontinuities. Furthermore, for the variable-coefficient case additional filtering is normally necessary on every time step in order to stabilize the computations, and this will normally result in inaccuracies which are increasing with time. Thus for long term integrations in the variable-coefficient case, the existing methods are often not satisfactory. It seems difficult in particular to retain the small structure in structured solutions [4, Sections 8.3, 8.5.3].

In order to amend some of the drawbacks of existing methods, we shall in this paper consider a different modification of the Fourier–Galerkin method for (1). The method is based on an accurate capturing of the propagating discontinuities of the nonsmooth solutions and of their derivatives. Whenever required, the solutions are accurately reconstructed from their spectral approximations by adding step-functions to the Fourier basis [9]. Special measures are taken towards minimizing numerical dispersion and numerical diffusion associated with the spatial discretization of the linear operator L by invoking a process

0021-9991/96 $18.00

called de-truncation [7, 10]. The approach is aimed at giving high order global accuracy and accurate resolution of the discontinuities even for long term integration. Since discontinuities are accurately represented by the utilized step-functions, we may also expect to achieve good accuracy on relatively coarse grids. Moreover, since smoothing filters are not applied, the small structure is retained in the solutions and no ''buffer-zone'' in the high-frequency part of the spectrum is required. An additional advantage is that with the same accuracy requirements, longer time steps may be used in our approach for an explicit time integrator than in the traditional approaches.

The paper is organized in the following way: After a presentation of some basic mathematical tools in Section 2, we derive in Section 3 the equations governing the propagation of singularities of nonsmooth solutions of (1) by the method of characteristics, and then we propose a first-version modified Fourier method for studying such solutions. In Section 4 we present two more modified Fourier–Galerkin methods which are of an analogous nature, but which are less intimately connected with the method of characteristics and therefore anticipated to be more flexible in applications. In Sections 5 and 6 we study the performance of the two latter methods when applied to selected nonsmooth problems for (1) with constant and variable coefficients, respectively.

2. FUNDAMENTALS

To a Riemann integrable complex-valued 2π-periodic function $u(x)$ we may for any given even integer $N > 0$ associate the Nth order *truncated* Fourier series

$$P_N u(x) = \sum_{k=-N/2+1}^{N/2-1} \hat{u}_k e^{ikx}, \tag{2}$$

where

$$\hat{u}_k = \frac{1}{2\pi} \int_0^{2\pi} u(x) e^{-ikx}\, dx; \quad k = 0, \pm 1, \pm 2, ... \tag{3}$$

As is well known [4], the error involved when we approximate $u(x)$ by the truncated Fourier series expansion (2) is strongly dependent on the smoothness of the function $u(x)$. We shall in this paper limit our discussion to functions $u(x)$ which are piecewise smooth on $[0, 2\pi]$. If we in addition assume that $u(x)$ is everywhere continuous and has continuous derivatives of order $p = 1, 2, ..., m - 1$, it can be shown [4] that

$$\hat{u}_k = O(|k|^{-(m+1)}) \quad \text{as } k \to \pm\infty, \tag{4}$$

and the best available global estimate is

$$\max_{0 \le x \le 2\pi} |u(x) - P_N u(x)| = O(N^{-m}) \quad \text{as } N \to \infty. \tag{5}$$

At a fixed point x, where $u^{(m)}(x)$ is continuous, however, it can be shown that

$$u(x) - P_N u(x) = O(N^{-(m+1)}) \quad \text{as } N \to \infty. \tag{6}$$

Moreover, it follows from (5) that as $N \to \infty$

$$\begin{aligned} \|u - P_N u\|_{L^2(0,2\pi)} &= \left\{ \int_0^{2\pi} |u(x) - P_N u(x)|^2\, dx \right\}^{1/2} \\ &= O(N^{-m}). \end{aligned} \tag{7}$$

The above estimates give an important part of the explanation for why spectral methods work so well when all functions involved are smooth. When $u(x)$ is discontinuous, however, the estimates (4), (5), (6), and (7) hold with $m = 0$, and consequently the convergence of the associated Fourier series is very slow. The resulting oscillatory behaviour of the truncated Fourier series expansion (2) near a discontinuity is known as the Gibbs phenomenon.

In the reconstruction of discontinuous 2π-periodic functions described in [9], a family of 2π-periodic functions $U_n(\xi)$, $n = 0, 1, 2, ...$, is utilized, which on the interval $0 \le \xi < 2\pi$ are given by

$$U_n(\xi) = -\frac{(2\pi)^n}{(n+1)!} B_{n+1}\left(\frac{\xi}{2\pi}\right), \tag{8}$$

where $B_j(x)$, $j = 1, 2, ...$, are the Bernoulli polynomials [11]. From (8) it follows that for each $n = 1, 2, ...$, $U_n(\xi)$ is a 2π-periodic function of finite regularity with derivatives $U_n^{(p)}(\xi) = U_{n-p}(\xi)$ continuous everywhere for $p = 0, ..., n - 1$, but with $U_n^{(n)}(\xi) = U_0(\xi)$ only piecewise continuous with jump-discontinuities of magnitude $+1$ at $\xi = 2m\pi$, $m = 0, \pm 1, \pm 2, ...$ In fact, from (8) it follows that $U_0(\xi)$ is a 2π-periodic step function (or rather a saw-tooth function), which on the interval $(-2\pi, 2\pi)$ is given by

$$U_0(\xi) = \begin{cases} \dfrac{1}{2\pi}(-\pi - \xi) & \text{if } -2\pi < \xi < 0 \\ \dfrac{1}{2\pi}(\pi - \xi) & \text{if } 0 < \xi < 2\pi. \end{cases} \tag{9}$$

For the higher order derivatives $U_n^{(p)}(\xi)$, $p \ge n + 1$, there are no jumps at the singularity locations. In fact, we see that

$$U_{-1}(\xi) \stackrel{\text{def}}{=} U_0'(\xi) = U_n^{(n+1)}(\xi) = -\frac{1}{2\pi} + \sum_{m=-\infty}^{+\infty} \delta(\xi - 2m\pi), \tag{10}$$

$$U_{-l}(\xi) \stackrel{\text{def}}{=} U_0^{(l)}(\xi) = U_n^{(n+l)}(\xi) = \sum_{m=-\infty}^{+\infty} \delta^{(l-1)}(\xi - 2m\pi), \quad l = 2, 3, ..., \tag{11}$$

where δ denotes Dirac's δ-function.

We note that $U_n(\xi)$ is an odd function when n is an even number and $U_n(\xi)$ is even when n is odd. The Fourier coefficients (3) for the functions $U_n(x)$, $n = 0, 1, 2, ...$, are given by

$$(\widehat{U_n})_0 = 0, \quad (\widehat{U_n})_k = \frac{1}{2\pi (ik)^{n+1}}; \quad k = \pm 1, \pm 2, ... \tag{12}$$

If we now consider a 2π-periodic function $u(x)$ which is known to be piecewise smooth on $[0, 2\pi]$, the interval $[0, 2\pi]$ can be divided into a finite number of subintervals on which $u(x)$ is smooth. At the endpoints of those subintervals, however, the function $u(x)$ and/or some (or all) of its derivatives may have jump-discontinuities. Following [9], the assumption that $u(x)$ is piecewise smooth on $[0, 2\pi]$, is clearly equivalent with the assumption that $u(x)$ for any given integer $Q \geq 0$ can be written, for some finite integer M,

$$u(x) = u^Q(x) + \sum_{n=0}^{Q} \sum_{j=1}^{M} A_j^n U_n(x - \gamma_j), \tag{13}$$

where $u^Q(x)$ is some Q times continuously differentiable 2π-periodic function which is piecewise smooth on $[0, 2\pi]$, the functions $U_n(\xi)$ are given by (8), and A_j^n, γ_j are some constants for $j = 1, ..., M$, $n = 0, 1, ..., Q$. From the properties of the functions $U_n(\xi)$ discussed above, it is clear that the points $x = \gamma_j$, $j = 1, ..., M$, are the locations for the singularities on the interval $[0, 2\pi]$ for the function $u(x)$ given by (13), while A_j^0 are the associated jumps for the function $u(x)$ itself and A_j^n are the associated jumps for the derivatives $u^{(n)}(x)$, $n = 1, ..., Q$, at those singularity locations.

Assuming that $P_N u(x)$ is known for some N and that we can determine all the quantities γ_j and A_j^n occurring in (13), we may in view of (12) compute $P_N u^Q(x)$ from

$$(\widehat{u^Q})_0 = \hat{u}_0, \quad (\widehat{u^Q})_k = \hat{u}_k - \sum_{n=0}^{Q} \sum_{j=1}^{M} \frac{A_j^n e^{-ik\gamma_j}}{2\pi (ik)^{n+1}}; \quad k \neq 0. \tag{14}$$

As described in [9], this may be used to accurately reconstruct $u(x)$ from $P_N u(x)$. A brief review of the reconstruction algorithm presented in [9] is given in an appendix. It follows from (5), (7) that the reconstructed function is globally $O(N^{-(Q+1)})$ accurate, provided that γ_j and the A_j^n can be determined with sufficient accuracy. In the following we shall describe various ways of utilizing these observations in order to obtain accurate numerical solutions of nonsmooth initial value problems for (1).

3. PROPAGATION OF SINGULARITIES

We shall in this section give a brief discussion of how singularities of nonsmooth solutions of (1) propagate when we restrict ourselves to *strictly hyperbolic* systems [6]. Thus, the characteristic equation associated with (1)

$$\det(-\lambda I + A) = 0 \tag{15}$$

is at every point x, t assumed to have m distinct real-valued roots $\lambda_k(x, t)$, $k = 1, ..., m$. To each eigenvalue λ_k we choose a left eigenvector $\mathbf{l}_k(x, t)$ and a right eigenvector $\mathbf{r}_k(x, t)$ by

$$\mathbf{l}_k(-\lambda_k I + A) = (-\lambda_k I + A)\mathbf{r}_k = \mathbf{0}. \tag{16}$$

As is well known, the set of left eigenvectors $\{\mathbf{l}_1, ..., \mathbf{l}_m\}$ are then linearly independent, and so is the set of right eigenvectors $\{\mathbf{r}_1, ..., \mathbf{r}_m\}$. Furthermore [6], there is no loss of generality by assuming that those eigenvectors at every point x, t satisfy the orthogonality relations

$$\mathbf{l}_k \cdot \mathbf{r}_i = \delta_{ki} \quad \text{for } k, i = 1, ..., m. \tag{17}$$

In view of (13), a solution $\mathbf{u}(x, t)$ of (1) which for each t is piecewise smooth on $[0, 2\pi]$ and 2π-periodic with respect to x, may for any given integer $Q \geq 0$ and for some integer $R \geq 0$ be written

$$\mathbf{u}(x, t) = \mathbf{u}^Q(x, t) + \sum_{n=0}^{Q} \sum_{j=1}^{R} \mathbf{a}_j^n(t) U_n(x - x_j(t)). \tag{18}$$

Here $x = x_j(t)$, $j = 1, ..., R$, denote curves Γ_j in the x, t-plane across which the solution itself and/or some (or all) of its spatial derivatives suffer jump-discontinuities. The function $\mathbf{u}^Q(x, t)$ is for each t at least Q times continuously differentiable and 2π-periodic with respect to x, and $\mathbf{a}_j^n(t)$ denotes the jump in the nth spatial derivative of $\mathbf{u}$ across Γ_j at the time t. Following essentially [6], we get by substitution of (18) into (1)

$$L[\mathbf{u}] = L[\mathbf{u}^Q] + \mathbf{F} = \mathbf{0}, \tag{19}$$

where

$$\begin{aligned}\mathbf{F}(x,t) = &\sum_{j=1}^{R} U_{-1}(x-x_j(t))\left[A(x,t) - \frac{dx_j}{dt}I\right]\mathbf{a}_j^0(t)\\ &+ \sum_{n=0}^{Q-1}\sum_{j=1}^{R} U_n(x-x_j(t))\left\{\left[A(x,t)\right.\right.\\ &\left.\left. - \frac{dx_j}{dt}I\right]\mathbf{a}_j^{n+1}(t) + L[\mathbf{a}_j^n]\right\}\\ &+ \sum_{j=1}^{R} U_Q(x-x_j(t))L[\mathbf{a}_j^Q].\end{aligned} \tag{20}$$

From the assumed regularity of the functions involved, it follows from (19), (20) that on each of the curves Γ_j, $j = 1, ..., R$, the following equations must be satisfied:

$$\left[A - \frac{dx_j}{dt}I\right]\mathbf{a}_j^0 = \mathbf{0}, \tag{21}$$

$$\left[A - \frac{dx_j}{dt}I\right]\mathbf{a}_j^{n+1} + L[\mathbf{a}_j^n] = \mathbf{0} \quad \text{for } n = 0, 1, ..., Q-1. \tag{22}$$

For each t we have by assumption that $\mathbf{a}_j^k \neq \mathbf{0}$ for one or more values of $k = 0, 1, ..., Q$. By considering the lowest value of k for which $\mathbf{a}_j^k \neq \mathbf{0}$, it follows from (21), (22) that the matrix $[A - (dx_j/dt)I]$ must be singular. From (15) we therefore see that the singularity curve Γ_j must be a characteristic curve for (1) determined by

$$\frac{dx_j}{dt} = \lambda_i(x,t), \tag{23}$$

where $i = i(j)$ is some integer such that $i \in \{1, ..., m\}$. When i is known, the singularity curve Γ_j is uniquely determined by (23) subject to some given initial value $x_j(0)$, say. From (21) it follows that $\mathbf{a}_j^0(t) = \alpha_{ji}^0(t)\mathbf{r}_i(x_j(t), t)$, and multiplication from the left by $\mathbf{l}_i(x_j(t), t)$ in (22) when $n = 0$, shows that on Γ_j we have the ordinary differential equation

$$\mathbf{l}_i L[\mathbf{a}_j^0] = \frac{d\alpha_{ji}^0}{dt} + \mathbf{l}_i \mathscr{L}^{(i)}[\mathbf{r}_i]\alpha_{ji}^0 = 0, \tag{24}$$

where $\mathscr{L}^{(i)} = I(\partial/\partial t + \lambda_i(\partial/\partial x)) + B$. Clearly, the discontinuity jump $\mathbf{a}_j^0(t)$ is uniquely determined along Γ_j by (24), subject to the actual initial data for $\mathbf{a}_j^0$ at $t = 0$, say.

We may now successively determine $\mathbf{a}_j^k$, $k = 1, ..., Q$, from (22). In fact, let us assume that $\mathbf{a}_j^0, ..., \mathbf{a}_j^n$ have been determined, and let us write

$$\mathbf{a}_j^{n+1}(t) = \sum_{l=1}^{m} \alpha_{jl}^{n+1}(t)\mathbf{r}_l(x_j(t), t). \tag{25}$$

In view of (16), (17), substitution of (25) into (22), and left multiplication by $\mathbf{l}_l$, then gives on Γ_j when $l \neq i$

$$\alpha_{jl}^{n+1} = -\frac{\mathbf{l}_l L[\mathbf{a}_j^n]}{\lambda_l - \lambda_i} \quad \text{for } n = 0, 1, ..., Q-1. \tag{26}$$

On the other hand, if we replace n by $n+1$, left multiplication of (22) by $\mathbf{l}_i$ is seen to give on Γ_j

$$\frac{d\alpha_{ji}^{n+1}}{dt} + \mathbf{l}_i \mathscr{L}^{(i)}[\mathbf{r}_i]\alpha_{ji}^{n+1} = -\mathbf{l}_i \sum_{\substack{l=1\\ l\neq i}}^{m} \mathscr{L}^{(i)}[\mathbf{r}_l]\alpha_{jl}^{n+1}, \tag{27}$$

for $n = 0, 1, ..., Q-2$ after substitution of (25). By considering the above construction with Q replaced by $Q+1$, it is clear that we must insist that (27) holds for $n = Q-1$ also, if the required assumption that $\mathbf{a}_j^Q(t)$ shall equal the jump at Γ_j of the Qth order spatial derivative of $\mathbf{u}$ shall be fulfilled. It is also clear that the jump $\mathbf{a}_j^{n+1}(t)$ is uniquely determined along Γ_j by (25), (26), and (27), subject to the actual initial data for α_{ji}^{n+1} at $t = 0$, say.

An examination of Eqs. (24), (26), and (27), known as the transport equations for the system (1), easily reveals that a solution of (1) with nonsmooth initial data will contain singularities of the same order as the initial data for every t. In particular, this means that if a solution $\mathbf{u}(x, t)$ of (1) initially is, say, $k-1$ times continuously differentiable, but suffer one or more jumps in the kth derivative, then the solution $\mathbf{u}(x, t)$ will at later times still be $k-1$ times continuously differentiable, but will suffer one or more jumps in the kth derivative. On the other hand, it is not difficult to see that jumps in higher order derivatives of the solution $\mathbf{u}(x, t)$ may develop along some of the characteristic curves in cases with $m > 1$, even if there are no jumps in the corresponding higher order derivatives of the initial data. This point will be considered further in later examples.

Let us now be more specific by considering the above construction for the case where the nonsmooth initial data associated with (1) are given on a form consistent with (18),

$$\mathbf{u}(x,0) = \mathbf{f}(x) = \mathbf{f}^Q(x) + \sum_{n=0}^{Q}\sum_{l=1}^{M} \mathbf{A}_l^n U_n(x-\gamma_l), \tag{28}$$

where, as usual, $\mathbf{f}^Q(x)$ is at least Q times continuously differentiable and the constants $\mathbf{A}_l^n$ denote the jumps in the function $\mathbf{f}$ and its derivatives at the singularity locations $x = \gamma_l$. According to the results obtained above, the only places where the solution $\mathbf{u}(x, t)$ of (1), (28) can be singular, i.e., not Q times continuously differentiable, is then for each $l = 1, ..., M$ clearly along the m characteristic curves $\Gamma_{j,l}$, $j = 1, ..., m$, which are passing through the initial

singularity locations $x = \gamma_l$ at $t = 0$. Consequently, we will in (18) have $R \le mM$.

Since we in the following construction may allow any jump $\mathbf{a}_j^n(t)$ in (18) to vanish, there will clearly be no loss of generality by taking $R = mM$. With this convention it is convenient here to replace (18) by

$$\mathbf{u}(x,t) = \mathbf{u}^Q(x,t) + \sum_{n=0}^{Q}\sum_{l=1}^{M}\sum_{j=1}^{m} \mathbf{a}_{j,l}^n(t) U_n(x - x_{j,l}(t)), \quad (29)$$

where the characteristic curves $\Gamma_{j,l}$ are given by $x = x_{j,l}(t)$. Since by construction $x_{j,l}(0) = \gamma_l$ for every $l = 1, ..., M$ and $j = 1, ..., m$, comparison of (28) and (29) gives

$$\sum_{j=1}^{m} \mathbf{a}_{j,l}^n(0) = \mathbf{A}_l^n \quad \text{for } n = 0, 1, ..., Q; l = 1, ..., M. \quad (30)$$

In view of (26), it is not difficult to see that the initial conditions associated with (24), (27) along each of the characteristic curves $\Gamma_{j,l}$ are uniquely determined by (30). As soon as the triple sum on the right-hand side in (29) (i.e., the singular part of a nonsmooth solution of (1), (28)) has been constructed satisfying (30), we may solve the residual problem for the function $\mathbf{u}^Q(x, t)$

$$L[\mathbf{u}^Q] = -\mathbf{F}, \quad \mathbf{u}^Q(x,0) = \mathbf{f}^Q(x), \quad (31)$$

where $\mathbf{F}(x, t)$ is given by (20) with the singular part substituted. For sufficiently smooth coefficients $A(x, t)$ and $B(x, t)$, it follows from the construction described above that $\mathbf{F}(x, t)$ is 2π-periodic with respect to x and at least $Q - 1$ times continuously differentiable everywhere. For Q sufficiently large, $\mathbf{u}^Q(x, t)$ may therefore be calculated numerically from (31) by utilizing any traditional method suitable for solving smooth hyperbolic problems.

We note here that in solving (31) numerically, it has been established [12, 15] that spectral methods involve less numerical dispersion and numerical diffusion than finite difference methods. For the purpose of maintaining high order accuracy for long term integration, a Fourier-collocation or Fourier–Galerkin method [4] therefore seems preferable in our case. We shall refer to this combination as Method 0, and in view of (6) we expect it to be a method of order $Q + 1$. Although it seems very probable that Method 0 can provide accurate numerical solutions in applications, we shall not utilize it in the form described above. Instead, we shall in the following section describe two alternative methods, Method 1 and Method 2, which are more suited for generalizations to systems (1) which are not strictly hyperbolic, as well as to nonlinear systems, employing some of the same underlying ideas as in Method 0. As it turns out, it is in fact possible to show that Method 0 is essentially equivalent with Method 1 for strictly hyperbolic systems (1) with coefficients that do not depend on x, while Method 2 seems to be the most flexible of the described methods.

4. A MODIFIED FOURIER-GALERKIN METHOD

From the elaboration in [9], it is clear that the Fourier coefficients in a truncated Fourier series representation (2) for a piecewise smooth function $u(x)$ carries information about the singularity locations and the corresponding jumps of $u(x)$ and the derivatives of $u(x)$. In fact, for sufficiently large N, $u(x)$ may be accurately reconstructed on the form (13) from $P_N u(x)$ by application of the reconstruction algorithm. Hence, even though truncated Fourier series do not by themselves provide accurate approximations, it may be constructive also for nonsmooth solutions of (1) to associate the truncated spatial Fourier series at each instant t

$$\mathbf{u}(x,t) \sim P_N\mathbf{u}(x,t) = \sum_{k=-N/2+1}^{N/2-1} \hat{\mathbf{u}}_k(t) e^{ikx}. \quad (32)$$

In this section we shall briefly describe modified Fourier–Galerkin methods for initial value problems for (1) with nonsmooth initial data which exploit these facts. The methods imply in practice that as far as possible the emphasis will be put on maintaining the accuracy of the Fourier coefficients in (32) as we march forward in time. Accordingly [7, 10], we represent the terms occurring in (1) by their Nth-order truncated Fourier series

$$A(x,t)\mathbf{u}_x(x,t) = \mathbf{f}(x,t) \sim P_N\mathbf{f}(x,t) = \sum_{k=-N/2+1}^{N/2-1} \hat{\mathbf{f}}_k(t) e^{ikx}, \quad (33)$$

$$B(x,t)\mathbf{u}(x,t) = \mathbf{g}(x,t) \sim P_N\mathbf{g}(x,t) = \sum_{k=-N/2+1}^{N/2-1} \hat{\mathbf{g}}_k(t) e^{ikx}. \quad (34)$$

Substitution of (32), (33), and (34) into (1) leads to

$$\frac{d\hat{\mathbf{u}}_k}{dt}(t) + \hat{\mathbf{f}}_k(t) + \hat{\mathbf{g}}_k(t) = \mathbf{0} \quad \text{for } k = 0, \pm 1, ..., \pm(N/2 - 1). \quad (35)$$

This set of ordinary differential equations for the Fourier coefficients in (32) should be solved subject to the initial data

$$\hat{\mathbf{u}}_k(0), \quad k = 0, \pm 1, ..., \pm(N/2 - 1), \quad (36)$$

which can be determined by (3) from the initial conditions associated with (1).

If the Fourier coefficients $\hat{\mathbf{f}}_k(t)$ and $\hat{\mathbf{g}}_k(t)$ in (33), (34) can be determined with sufficient accuracy, the Fourier method now reduces to solving (35), (36) by some accurate numerical integration scheme. In fact, at whatever time we wish to accurately determine the solution $\mathbf{u}(x, t)$, we then merely have to apply the reconstruction algorithm. In the cases where the coefficients A, B in (1) are independent of x, it is easily seen that $P_N\mathbf{f}(x, t) = A(t)P_N\mathbf{u}_x(x, t)$, $P_N\mathbf{g}(x, t) = B(t)P_N\mathbf{u}(x, t)$. Hence, the accuracy of the computed Fourier coefficients by (35), (36) will in these cases depend solely on the accuracy of the applied time stepping scheme; no numerical dispersion or numerical diffusion is introduced through the spatial discretization in these cases. As we shall see in the following section, these expectations are confirmed by numerical experiments.

If, on the other hand, the coefficients A, B in (1) are dependent of x, it is not difficult to see [7, 10] that the traditional methods for calculating $\hat{\mathbf{f}}_k(t)$ and $\hat{\mathbf{g}}_k(t)$ in (33), (34) may lead to intolerable inaccuracies if $\mathbf{u}(x, t)$ is not smooth. For a detailed discussion concerning the resulting numerical dispersion and numerical diffusion, and how it can be avoided, we refer to [7, 10]. Here we shall limit ourselves to give a brief review of the conclusions when the coefficients A, B in (1) are smooth and therefore may be approximated with spectral accuracy by their truncated expansions

$$P_N A(x,t) = \sum_{k=-N/2+1}^{N/2-1} \hat{A}_k(t)e^{ikx},$$
$$P_N B(x,t) = \sum_{k=-N/2+1}^{N/2-1} \hat{B}_k(t)e^{ikx}. \tag{37}$$

The modification advocated in [7, 10], and called the method of de-truncation, can formally be expressed as

$$P_N\mathbf{f} \approx P_N[(P_N A)(P_{2N-2}\mathbf{u}_x)],$$
$$P_N\mathbf{g} \approx P_N[(P_N B)(P_{2N-2}\mathbf{u})]. \tag{38}$$

The key to avoid numerical dispersion and numerical diffusion is here to calculate an accurate approximation for the de-truncated expansion $P_{2N-2}\mathbf{u}$ from the known truncated expansion $P_N\mathbf{u}$ given by (32).

If the nonsmooth solution $\mathbf{u}(x, t)$ of (1) at some instant t is given by (18), the Fourier coefficients associated with $P_N\mathbf{u}$ are in view of (12) given by

$$\hat{\mathbf{u}}_k(t) = (\widehat{\mathbf{u}^Q})_k(t) + \sum_{n=0}^{Q}\sum_{j=1}^{R} \frac{\mathbf{a}_j^n(t)e^{-ikx_j(t)}}{2\pi(ik)^{n+1}}$$
$$\text{for } k = \pm 1, ..., \pm(N/2 - 1), \tag{39}$$

and $\hat{\mathbf{u}}_0(t) = (\widehat{\mathbf{u}^Q})_0(t)$. Assuming that accurate approximations for the Fourier coefficients $\hat{\mathbf{u}}_k(t)$ for $k = 0, \pm 1, ..., \pm(N/2 - 1)$ can be calculated, there are two apparent alternatives for calculating the terms on the right-hand side in (39). The first, which we shall call Method 1, is to calculate the locations of the singularities $x_j(t)$ and the associated jumps $\mathbf{a}_j^n(t)$ by the method described in the preceding section, i.e., by integrating the characteristic equations (23) and the transport equations (24), (27). The second method, which we shall call Method 2, is to calculate approximate singularity locations and jumps by the reconstruction algorithm. For both methods, we deduce from (4) that for t arbitrarily given and for any appropriate norm $\|\cdot\|$, we have asymptotically

$$\|(\widehat{\mathbf{u}^Q})_k(t)\| = O(|k|^{-(Q+2)}) \quad \text{as } |k| \to \infty. \tag{40}$$

Thus the additional Fourier coefficients needed in the method of de-truncation can be approximated to the order $O(N^{-(Q+1)})$ by

$$\hat{\mathbf{u}}_k(t) \approx \sum_{n=0}^{Q}\sum_{j=1}^{R} \frac{\mathbf{a}_j^n(t)e^{-ikx_j(t)}}{2\pi(ik)^{n+1}}$$
$$\text{for } k = \pm N/2, \pm(N/2 + 1), ..., \pm(N - 2). \tag{41}$$

Method 1, using characteristic data, is optimal in the sense that for any given Q the singular part of $\mathbf{u}(x, t)$, i.e., the last double sum in (18), (39), (41), can be determined with any desired accuracy. Consequently, the function $\mathbf{u}^Q(x, t)$ in (18) is also determined with the highest obtainable accuracy by this method. Although Method 1 essentially is equivalent with the method of characteristics (Method 0) described in the preceding section, it seems to have a somewhat wider scope. In fact, for nonlinear problems one may think of applying established shock relations (the Rankine–Hugoniot relations) instead of the characteristic equation and the transport equations. We do regard Method 2, however, as even more flexible than Method 1. Method 2 is clearly not restricted to strictly hyperbolic systems, and it can also be applied to nonlinear systems without major alterations [7, 10].

If the dimension m of the system (1) is even, and it is easily seen that we may arrange real-valued unknowns and real-valued coefficients such that (38) may be efficiently implemented by employing complex $2N$-point FFT-transforms [7, 10]. From a computational point of view, the method of de-truncation requires at each time step in the "worst" variable coefficient case $2(m/2 + m^2/2)$ inverse $2N$-point complex FFT-transforms, $4Nm^2$ multiplications, $2Nm(2m - 1)$ additions, and $2(m/2)$ $2N$-point complex FFT-transforms. An additional $O(2Nm)$ operations are needed in order to de-truncate $P_N\mathbf{u}$ to $P_{2N-2}\mathbf{u}$ and to differentiate $P_{2N}\mathbf{u}$ in Fourier space. The operational count for

the computation of the singularity locations and the jumps from $P_N\mathbf{u}$ when Method 2 is applied, depends on the parameters M and Q. Since M and Q here are considered to be small numbers, it is not difficult to see that the extra work normally will be insignificant in this connection.

In the examples which we are going to present in the following sections, the solutions of (1) are computed both by Method 1 and by Method 2. As previously noted, Method 2 is considered to be the most flexible of the two with respect to generalizations; the performance of Method 2 is therefore of particular interest. Method 1, on the other hand, can for the problems considered in this paper be expected to produce more accurate results and may, therefore, be considered primarily as a reference method. Since the emphasis in this paper is put on the spatial approximation of the solutions, we shall seek to minimize the temporal discretization error and to choose the range of N in the computations such that the spatial discretization error can be considered dominant. The scheme used for the integration of (35) is an explicit Runge–Kutta method of order (4)5 with step size control due to Dormand and Prince [16]. The scheme can be used with *local* tolerances as low as about 10^{-9}. The accuracy of the computed solutions is calculated by employing the RMS-error at the grid points $x_i = 2\pi i/N$, $i = 0, 1, ..., N-1$. We shall present plots generated by gnuplot routines where interpolation between the grid point values are employed. The plots will therefore not show the subgrid accuracy obtained by the described methods.

5. CONSTANT COEFFICIENTS

As a consequence of the discussion in the preceding section, it follows that the necessary modifications needed for studying nonsmooth solutions by the Fourier method in the constant coefficient case consist essentially only of a postprocessing filter. In fact, no explicit information about the singularity locations and the discontinuity jumps of a nonsmooth solution $\mathbf{u}$ is needed during the integration of (35) in order to maintain an accurate approximation for $P_N\mathbf{u}$. Information concerning the singularity locations and the discontinuity jumps is for the constant coefficient case implicitly carried by the Fourier coefficients and needs only be extracted at the actual time when we wish to reconstruct $\mathbf{u}$ from $P_N\mathbf{u}$ by the reconstruction algorithm. The accuracy of the reconstructed solution at a certain instant depends only on the time discretization error and on the accuracy achieved by the reconstruction algorithm. In the example we are going to present in this section, time discretization errors can be considered insignificant with our choice of integration scheme.

As a test case we consider a Cauchy problem similar to one discussed in [17],

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_t + \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_x + \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} f(x) \\ 0 \end{pmatrix} \quad \text{at } t = 0, \tag{42}$$

where $f(x)$ is a 2π-periodic function which on $[0, 2\pi)$ is a rectangular pulse defined by

$$f(x) = \begin{cases} 1 & \text{if } 0 \le x < \pi, \\ 0 & \text{if } \pi < x < 2\pi. \end{cases} \tag{43}$$

Clearly, the two different families of characteristics for the system (42) are given by $\phi_1(x, t) = x - t = c_1$ and $\phi_2(x, t) = x + t = c_2$. By introduction of the characteristic variables $\xi = x + t$, $\eta = x - t$, the system (42) can be written

$$\frac{\partial u_1}{\partial \eta} + \frac{1}{2}u_2 = 0, \qquad \frac{\partial u_2}{\partial \xi} + \frac{1}{2}u_1 = 0, \qquad u_1 = f, \quad u_2 = 0 \quad \text{for } \xi = \eta. \tag{44}$$

From (44) it follows that both components u_1, u_2 satisfy the Klein–Gordon equation

$$\frac{\partial^2 u}{\partial \xi \partial \eta} - \frac{1}{4}u = 0. \tag{45}$$

Solving (45) for u_1 by the Riemann method [6] and substituting into (44) produces the solution

$$u_1(\xi, \eta) = f(\xi) - \frac{1}{2}\int_\eta^\xi \sqrt{(\tau - \eta)/(\xi - \tau)} \times J_1(\sqrt{[\tau - \eta][\xi - \tau]})f(\tau)\, d\tau, \tag{46}$$

$$u_2(\xi, \eta) = -\frac{1}{2}\int_\eta^\xi J_0(\sqrt{[\tau - \eta][\xi - \tau]})f(\tau)\, d\tau, \tag{47}$$

where J_0, J_1 are Bessel functions of the first kind. An accurate approximate solution can be easily obtained from (46), (47) by numerical integration, utilizing, for instance, routines from the NAG-library.

According to the theory described in Section 3, the singularities of the solution of (42) will propagate along the characteristic curves $\phi_j(x, t) = k\pi$, $k = 0, \pm 1, \pm 2, ...$, for $j = 1, 2$. The solutions of the corresponding transport equations (24)–(27) along those curves are easily found to be for the jumps $[\cdot]_j$ of $\mathbf{u}$ and its first two derivatives,

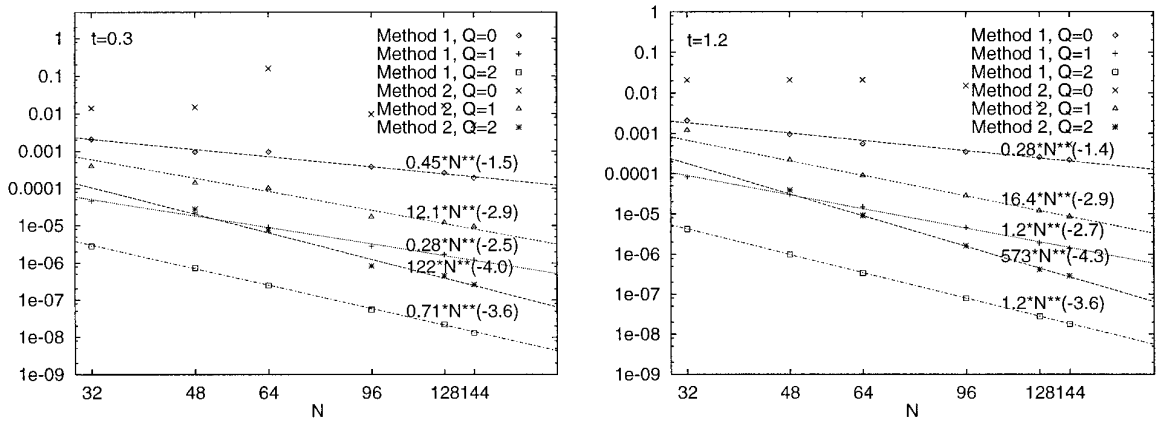

FIG. 1. Maximum RMS-error in the two computed solution components of (42), (43).

$$[\mathbf{u}]_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad [\mathbf{u}_x]_1 = (-1)^k \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix},$$

$$[\mathbf{u}_{xx}]_1 = (-1)^k \begin{pmatrix} -\frac{1}{4} \\ \frac{t}{4} \end{pmatrix}, \tag{48}$$

$$[\mathbf{u}]_2 = \begin{pmatrix} (-1)^k \\ 0 \end{pmatrix}, \quad [\mathbf{u}_x]_2 = (-1)^k \begin{pmatrix} -\frac{t}{2} \\ -\frac{1}{2} \end{pmatrix},$$

$$[\mathbf{u}_{xx}]_2 = (-1)^k \begin{pmatrix} \frac{t^2}{8} + \frac{1}{4} \\ \frac{t}{4} \end{pmatrix}. \tag{49}$$

Similarly, jump discontinuities in the derivatives of higher orders, $\mathbf{u}_{xxx}$, $\mathbf{u}_{xxxx}$, etc., develop along the same characteristic curves. When reconstructing the solution at some chosen instant, we have to choose the smoothness parameter Q according to the desired accuracy. We note that the initial jump discontinuities in $\mathbf{u}$ are transported with constant magnitude only along the curves $\phi_2(x, t) = k\pi$, whereas along the curves $\phi_1(x, t) = k\pi$ the solution itself is continuous while jumps in its spatial derivatives are generated. Figure 2 shows plots of the computed solution, and the calculated global error is presented in Fig. 1. For $N = 32$, Method 2 can only be applied for $Q \leq 1$, due to the limited number of available Fourier coefficients (see the Appendix). Straight lines are fitted to the error data using the method of least squares, and for Method 2 we have only used the points at $N = 64, 96, 128, 144$. The calculated errors indicate that the accuracy for both methods is at least $O(N^{-(Q+1)})$ in the convergent cases. We note that we have no convergence of Method 2 when $Q = 0$, $R = 2$. Indeed, since the solution has jump discontinuities in its derivatives at locations distinct from the points where the solution itself is discontinuous, the reconstruction algorithm is not accurate in this case [9]. However, by putting $R = 4$ and thus also counting higher order singularities, Method 2 for the case $Q = 0$ appeared to be at least $O(N^{-1})$ accurate in our computations.

The solution of (42) was computed in [17] utilizing a Fourier method incorporating a smoothing filter applied to the initial data. The obtained solution shows relatively high accuracy in parts of the domain where the exact solution is continuous, while regions of low accuracy are localized to small neighborhoods of the discontinuities. The results shown in Fig. 1 for Method 2 with $N = 128$ may be compared to the results presented for the coarsest of the two grids considered in [17]. At $t = 0.3$, we find that our *global* error results for $Q = 1$ and $Q = 2$ are approximately two and three orders of magnitude better, respectively, than the error observed in [17] at specific points lying in the part of the influence domain of the initial discontinuities where the solution is continuous. Since we in contrast to [17] also obtain accurate results near the discontinuities, the superiority of our approach should therefore be well established.

6. VARIABLE COEFFICIENTS

As a consequence of the discussion in Section 4, it follows that in the case of variable coefficients, it is advisable to apply the method of de-truncation at every time step during the integration of (35). Thus the singularity locations and

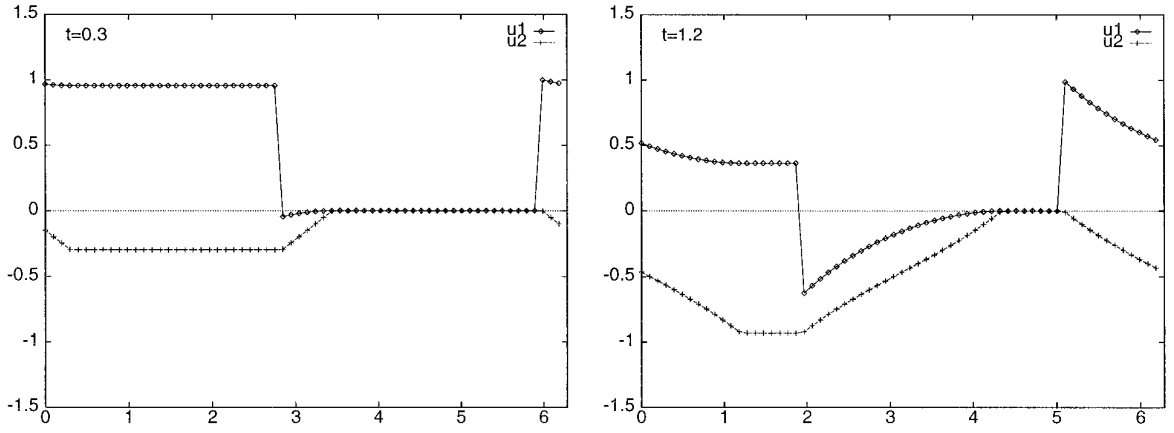

FIG. 2. Solution of (42), (43) computed by Method 2 with $N = 64$, $Q = 2$.

the associated jumps of the solution must be calculated at every time step, utilizing either Method 1 or Method 2.

In this section, we shall first look at the symmetric Cauchy problem

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_t + \begin{pmatrix} \cos(x-t) & \sin(x-t) \\ \sin(x-t) & -\cos(x-t) \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}_x = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} f(x) \\ 0 \end{pmatrix} \quad \text{at } t = 0, \tag{50}$$

where the characteristic curves again are given by $\phi_1(x, t) = x - t = c_1$ and $\phi_2(x, t) = x + t = c_2$. By a change of dependent variables

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \cos\left(\frac{x-t}{2}\right) & -\sin\left(\frac{x-t}{2}\right) \\ \sin\left(\frac{x-t}{2}\right) & \cos\left(\frac{x-t}{2}\right) \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}, \tag{51}$$

(50) can be transformed into the canonical form

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix}_t + \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}_x = \begin{pmatrix} 0 \\ w_1 \end{pmatrix}, \qquad \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \cos\left(\frac{x}{2}\right) f(x) \\ -\sin\left(\frac{x}{2}\right) f(x) \end{pmatrix} \quad \text{at } t = 0. \tag{52}$$

Hence, it is easy to show that the exact solution of (50) is given by

$$u_1(x,t) = \frac{1}{2}\left\{ [1 + \cos(x-t)] f(x-t) - \sin\left(\frac{x-t}{2}\right) \int_{x-t}^{x+t} \cos\left(\frac{\tau}{2}\right) f(\tau)\, d\tau + [\cos(t) - \cos(x)] f(x+t) \right\}, \tag{53}$$

$$u_2(x,t) = \frac{1}{2}\left\{ \sin(x-t) f(x-t) + \cos\left(\frac{x-t}{2}\right) \int_{x-t}^{x+t} \cos\left(\frac{\tau}{2}\right) f(\tau)\, d\tau - [\sin(t) + \sin(x)] f(x+t) \right\}. \tag{54}$$

By examining (53), (54), it is not difficult to see that no singularity in the solution **u** can propagate along the characteristic curves $\phi_2(x, t) = x + t = 2k\pi$, $k = 0, \pm 1, \pm 2, \ldots$. When $f(x)$ is given by (43), the singularities will therefore propagate along the characteristic curves $\phi_1(x, t) = k\pi$ and $\phi_2(x, t) = (2k - 1)\pi$, $k = 0, \pm 1, \pm 2, \ldots$. Along $\phi_1(x, t) = 2k\pi$, the initial discontinuity in **u** at $x = 2k\pi$ propagates undisturbed for $k = 0, \pm 1, \pm 2, \ldots$, and no jump discontinuities develop in the derivatives of **u** along those curves. Along $\phi_2(x, t) = (2k - 1)\pi$, however, both **u** and its derivatives are discontinuous, and along $\phi_1(x, t) = (2k - 1)\pi$, **u** is continuous, but its derivatives are discontinuous. In Fig. 3 and Fig. 4 we present some computational results, and plots of the numerical solution are shown in Fig. 5. We have not presented results for Method 2 with $Q = 0$, $R = 2$ here, since we expect no convergence in this case in view of the results obtained in the preceding section. If we let $R = 3$ in the computations with Method 2 for

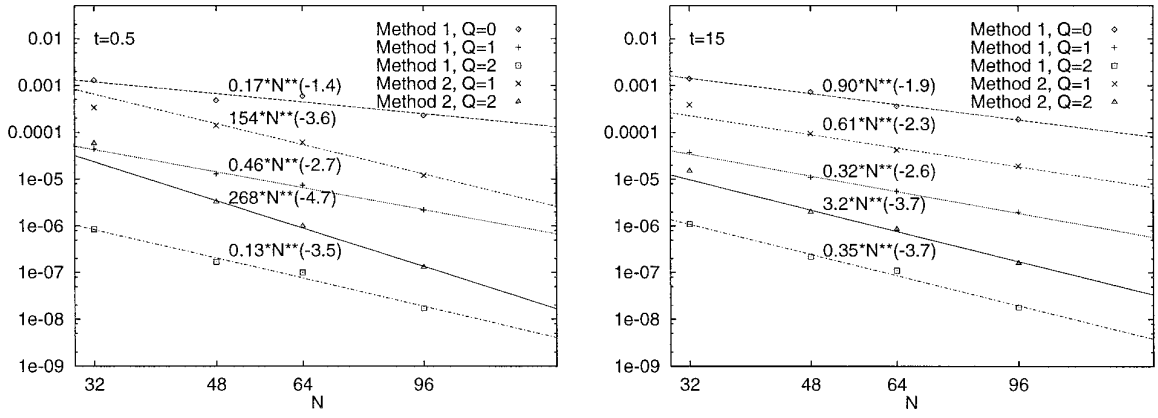

FIG. 3. Maximum RMS-error in the two components of the computed solution of (50), (43). The lines corresponding to Method 2 are obtained by linear regression with respect to the data at $N = 48, 64, 96$.

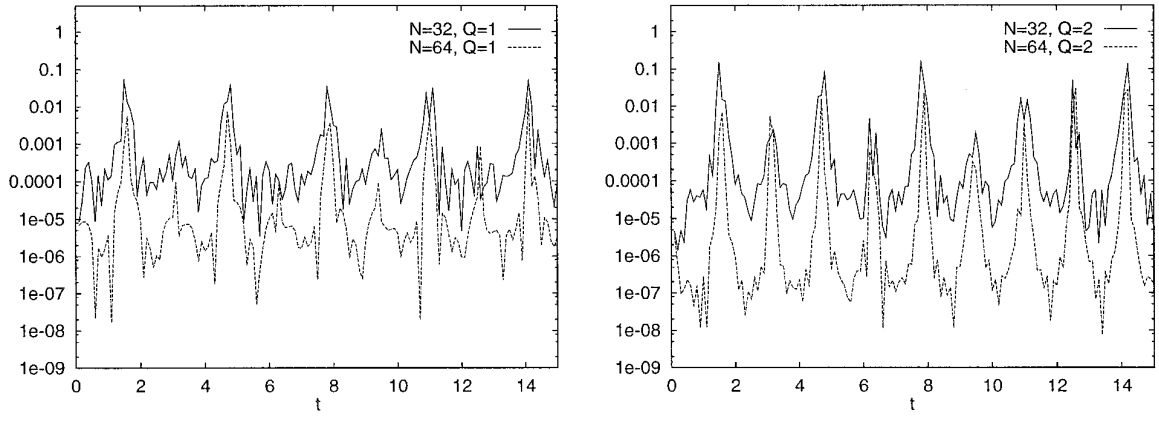

FIG. 4. Absolute error in computed singularity location by Method 2.

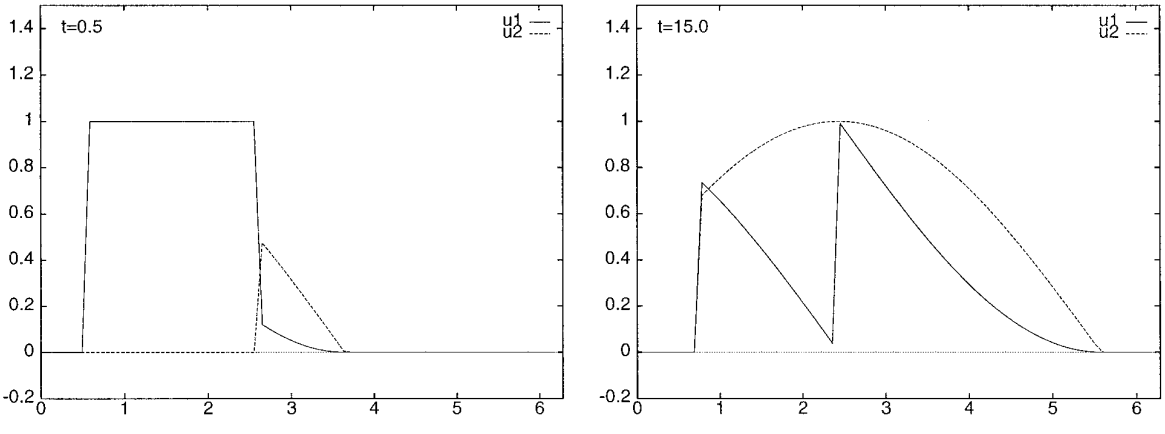

FIG. 5. Solution of (50), (43) computed by Method 2 with $N = 64$, $Q = 2$.

the case $Q = 0$, however, the accuracy appears to be at least $O(N^{-1})$.

From Fig. 3, the accuracy of Method 1 and Method 2 appears to be at least $O(N^{-(Q+1)})$ for the two presented times. In Fig. 4, we have plotted the error in the singularity location computed by Method 2 relative to the corresponding characteristic curve $\phi_2(x, t) = \pi$. In our computations we have applied the reconstruction with a fixed number $R = 3$ of singularities in each period. Thus, we have not taken into account the reduction of the number of singularities to $R = 2$ at the special times $t = k\pi/2, k = 1, 2, ...$, where crossings of different characteristics carrying singularities occur. Furthermore, in a small time interval around each of those critical times, we have clustering of the singularity locations and can in view of the discussion in [8, 9] expect that the system of equations for the determination of the singularity locations is ill-conditioned and, hence, that the accuracy deteriorates there. This explains the peaks of low accuracy seen in Fig. 4. It is interesting to note, however, that high accuracy is recovered in each time interval between two such peaks and no essential loss of overall accuracy is observed in the number of periods considered here. A natural conclusion to be drawn from these observations is that a quite accurate approximation for the Fourier coefficients in $P_N\mathbf{u}(x, t)$ is maintained during the time integration, even when there is clustering of the singularities, and consequently the method of de-truncation with Method 2 is fairly robust. This conclusion does not necessarily mean, however, that the reconstructed solution itself is particularly accurate at times when the singularities are clustered.

For the special test case (50) considered above, it is not difficult to see that the corresponding coefficient matrix $A(x, t)$ in (1) for any given $N \geq 2$ can be written

$$A(x, t) \equiv P_N A(x, t) = \hat{A}_{-1}(t)e^{-ix} + \hat{A}_1(t)e^{ix}. \tag{55}$$

From the discussion in [7, 10] it is therefore clear that we can calculate $P_N A\mathbf{u}_x$ exactly if we, in addition to $P_N\mathbf{u}$, know the two Fourier coefficients $\hat{\mathbf{u}}_{\pm N/2}$. To show the significance of those two "missing" Fourier coefficients, we have computed the solution for $N = 64$ using the traditional Galerkin method with de-aliasing in the approximation of the convolution sums. Figure 6 shows a plot of the computed solution reconstructed with $Q = 2$ at $t = 0.5$ and the corresponding RMS-error is found to be 1.9×10^{-2}. This illustrates the error introduced by ignoring the Fourier coefficients $\hat{v}_k$ for $k = \pm N/2, \pm(N/2 + 1), ...$ in the approximation of the convolutions corresponding to the product of an irregular function $v(x)$ and a smooth function $a(x)$ from the knowledge of $P_N a$ and $P_N v$ [7, 10]. It clearly motivates the use of the de-truncation method for such calculations.

As our final test case we shall consider the problem

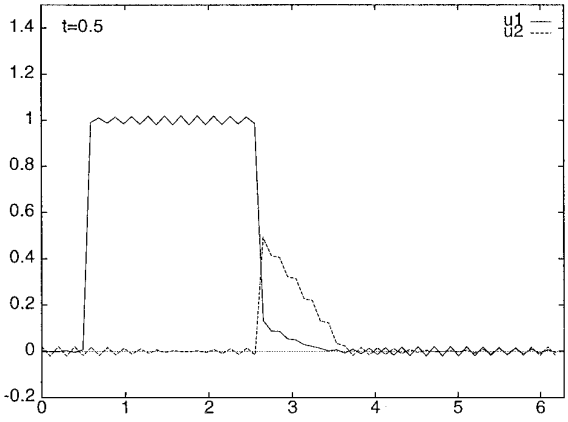

FIG. 6. Computed solution of (50) using the Galerkin approximation, $N = 64$.

$$u_t + a(x)u_x = 0, \quad u(x, 0) = h(x), \tag{56}$$

where the 2π-periodic coefficient $a(x)$ is given by

$$a(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 2, \\ 1 + c[(x-2)(x-4)]^m & \text{if } 2 < x \leq 4, \\ 1 & \text{if } 4 \leq x < 2\pi. \end{cases} \tag{57}$$

Here $-1 < c < 1$ is a constant and m is a positive integer. Clearly, $a(x)$ is $m - 1$ times continuously differentiable everywhere, and in addition to the value $a = 1$, $a(3) = 1 + (-1)^m c$ is the extremal value of a. If the initial function $h(x)$ in (56) is 2π-periodic, the solution of (56) is seen to be periodic in t with period

$$T = b(2\pi), \quad \text{where } b(x) = \int_0^x \frac{d\tau}{a(\tau)}, \tag{58}$$

and one easily finds by the method of characteristics that the solution of (56) is given by

$$u(x, t) = g(b(x) - t), \tag{59}$$

where $g(\xi)$ is the T-periodic function which is such that $g(b(x)) \equiv h(x)$. Since $a(x)$ is positive everywhere, the solution (59) constitutes a wave propagating to the right. If the 2π-periodic initial function in (56) is chosen by

$$h(x) = \begin{cases} 1 + 0.1 \sin^4\left(\dfrac{\pi x}{2}\right) & \text{if } 0 < x < 2, \\ 0 & \text{if } 2 < x < 2\pi, \end{cases} \tag{60}$$

we see that $h(x)$ has jump discontinuities of magnitude ± 1

at $x = 0$ and $x = 2$, respectively, whereas its first three derivatives suffer no jump discontinuities. In this case the solution (59) is seen to retain essentially the form of the initial function $h(x)$, but in the primary period $[0, 2\pi]$ modified by a stretching or shrinking in the subinterval $2 < x < 4$ due to the variable coefficient $a(x)$. The wave regains its exact original form after its trailing front has passed $x = 4$.

In accordance with the discussion in Section 4, we shall in the numerical computations approximate $a(x)$ in (56) by $P_N a(x)$. Thus, the exact solution of the corresponding modified equation (56) is given by (59) when $b(x)$ is replaced by

$$\tilde{b}(x) = \int_0^x \frac{d\tau}{P_N a(\tau)}. \tag{61}$$

The resulting error can be easily estimated. In fact, by restricting ourselves to functions (57) with $|c| < 0.5$, we have $a(x) > 0.5$ and $P_N a > 0.4$ everywhere for every $N \geq 16$ and every $m \geq 1$. From the Cauchy–Schwarz inequality it follows that

$$\begin{aligned} |b(x) - \tilde{b}(x)|^2 &= \left| \int_0^x \left(\frac{1}{a(\tau)} - \frac{1}{P_N a(\tau)} \right) d\tau \right|^2 \\ &= \left| \int_0^x \frac{P_N a(\tau) - a(\tau)}{P_N a(\tau) a(\tau)}\, d\tau \right|^2 \\ &\leq \left(\int_0^x |P_N a(\tau) - a(\tau)|^2\, d\tau \right) \\ &\quad \times \left(\int_0^x \left| \frac{1}{P_N a(\tau) a(\tau)} \right|^2 d\tau \right) \\ &\leq C_1 \| P_N a(x) - a(x) \|^2_{\tilde{L}^2(0,2\pi)}, \end{aligned} \tag{62}$$

where the constant C_1 is such that $C_1 < 50\pi$. It now follows from the properties of $a(x)$ and (7) that for some constant C we have

$$|b(x) - \tilde{b}(x)| \leq C N^{-m}. \tag{63}$$

In order to thoroughly test the accuracy and robustness of the methods described in this paper, we let in the following $m = 2$. The above estimates then suggest that we should not expect better accuracy than $O(N^{-2})$, since the numerical solution is based on approximating $a(x)$ by $P_N a(x)$. In our implementation of Method 1, we have calculated the singularity locations by the characteristic equation with the exact $a(x)$ instead of $P_N a(x)$, and we have in fact observed the above expected accuracy. We do note here that the effect of the oscillatory behaviour of the approximation $P_N a(x)$ relative to the exact values of $a(x)$ has not been accounted for in (63). We may therefore hope for a faster convergence of Method 2 than that indicated by (63), and this is also confirmed by the numerical experiments. If we for Method 1 calculate the singularity locations by the characteristic equation with $P_N a(x)$, instead of the exact $a(x)$, we may anticipate a similar improvement in accuracy. It is therefore reasonable to believe that Method 1 normally performs better than Method 2 also for this case.

In Fig. 7 we present results for computations with two different choices of the parameter c determining the amount of stretching. The modified solution (59) has no jumps in its first three derivatives, so we let $Q = 0$. To ensure stability within the whole time range considered, we had to introduce a moderate form of filtering in the computations employing Method 2. In fact, in contrast to all previous cases, we have here used the highest order Fourier coefficients of positive wave number in the truncated series $P_{N-2D} u$ rather than in $P_N u$, when we computed the discontinuity locations and the jumps needed for the de-truncation at each time step (see the Appendix). This is done since the highest order modes of $P_N u$ are polluted with errors which presumably stem from the numerical dispersion resulting from the relatively slow convergence of $P_N a$. If we let $D = 0$, the numerical solution obtained by Method 2 has in our computations been found to break down in the time range $1.0 < t < 2\pi$ for $N = 32$ and $N = 64$. The choices of the parameter D used in the computations are listed in Table I.

The rates of convergence deducible from Fig. 7 strongly indicate that for the times presented, the computed solution is at least $O(N^{-2})$ accurate for Method 1 while for Method 2 convergence is considerably faster. Figure 8 shows the error in the discontinuity location computed in Method 2 and corresponding to the characteristic curve $b(x) - t = 2$. It may be noted that Method 2 gives a good resolution of the discontinuity locations even for $N = 16$, $(2\pi/N \approx 0.39)$, while more grid points are needed in order to get resolution of the fine structure of the wave. Figure 9 shows plots of the solution computed by Method 2 with $N = 64$, $c = 0.2$. The time $t = 30.5$ then corresponds to approximately 5 periods. We note here that, in view of the presented error results, the computed solutions shown in the plots are indistinguishable from the exact ones. For comparison, we have also for this test case computed a solution using the standard Galerkin method with dealiasing for the case $N = 64$, $c = 0.2$. The solution reconstructed with $D = 9$ is shown in Fig. 10. The RMS-error at $t = 2.5$ was now found to be 1.8×10^{-2}. Clearly, the fine structure of the solution is not properly resolved in this case unless the de-truncation method is employed.

7. DISCUSSION

We have in this paper presented modified Fourier–Galerkin methods for studying nonsmooth solutions of

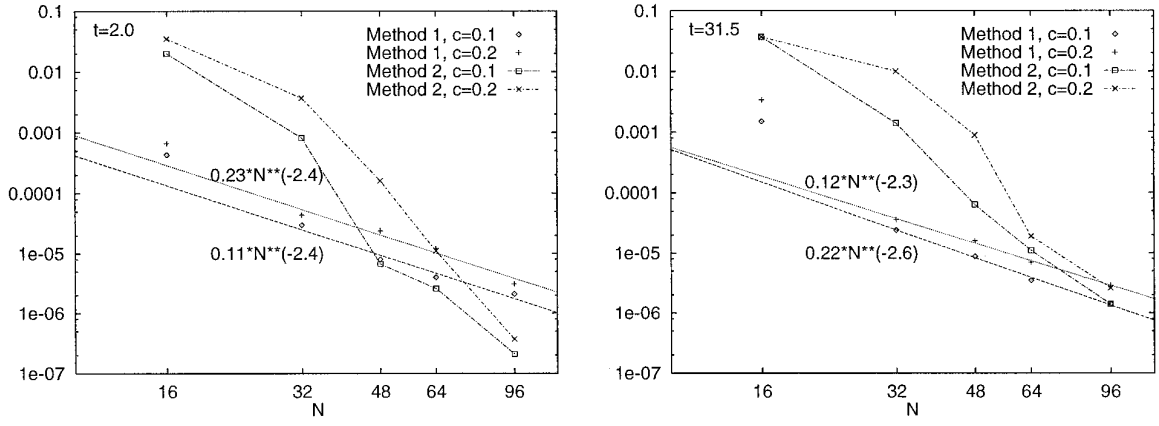

FIG. 7. Calculated error for the computed solution of (56). The straight lines fitted to the data for Method 1 are obtained by linear regression with respect to the data at $N = 32, 48, 64, 96$.

linear hyperbolic systems. The methods are based on a spatial discretization utilizing truncated Fourier series combined with a carefully chosen family of 2π-periodic piecewise polynomials. The general idea is that a piecewise smooth 2π-periodic function may be decomposed into a smooth part having a fast and uniformly convergent Fourier series representation and a singular part which is a linear combination of piecewise polynomials. That singular part of the function is determined by the discontinuity locations and the associated jumps of the piecewise smooth function itself and of its derivatives up to the arbitrarily specified order $Q \geq 0$. We may then according to the theory reconstruct the function with $O(N^{-(Q+1)})$ global accuracy, where N is the order of the known truncated Fourier series representation for the function.

For the two modified Fourier methods, Method 1 and Method 2, which are applied in the test cases presented in this paper, emphasis is put on two issues in particular. On the one hand, the maintenance of accurate Fourier coefficients in the truncated spatial Fourier series representation for the nonsmooth solution during the time integration and, on the other hand, the accurate computation of the singular part of the solution at instants where it is required. The two methods differ in the latter issue in that Method 1 determines the singularity locations from the characteristic equations and the jumps from the transport equations of the linear hyperbolic system, while Method 2 employs the reconstruction algorithm [9] which extracts the same information from the truncated Fourier series representation of the solution. In cases where the coefficients of the linear hyperbolic system are independent of the spatial variable, the methods act only as postprocessing filters for the removal of the Gibbs phenomenon. In cases of space-dependent coefficients, however, the two issues mentioned are closely related since the method of de-truncation is then applied. In fact, the method of de-truncation requires knowledge of an accurate approximation for the singular part of the solution at every time step during the time integration. Even with that interplay between the two issues, the fact that the two issues still are somewhat independent of each other is illustrated by the observation

TABLE I

Choices of Parameter D in the Computations with Method 2 for (56)

	Choice of D	
N	$c = 0.1$	$c = 0.2$
16	0	0
32	3	5
48	5	7
64	7	9
96	9	11

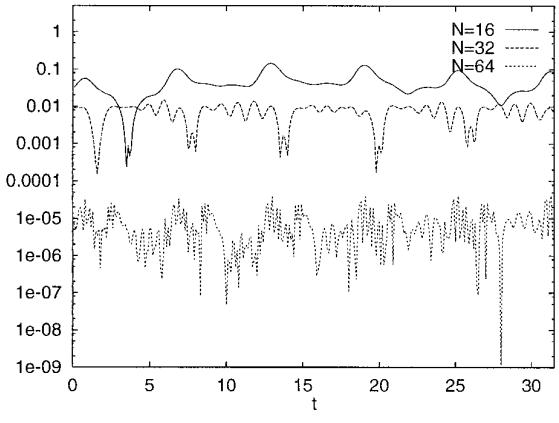

FIG. 8. Absolute error in computed discontinuity location, $c = 0.2$.

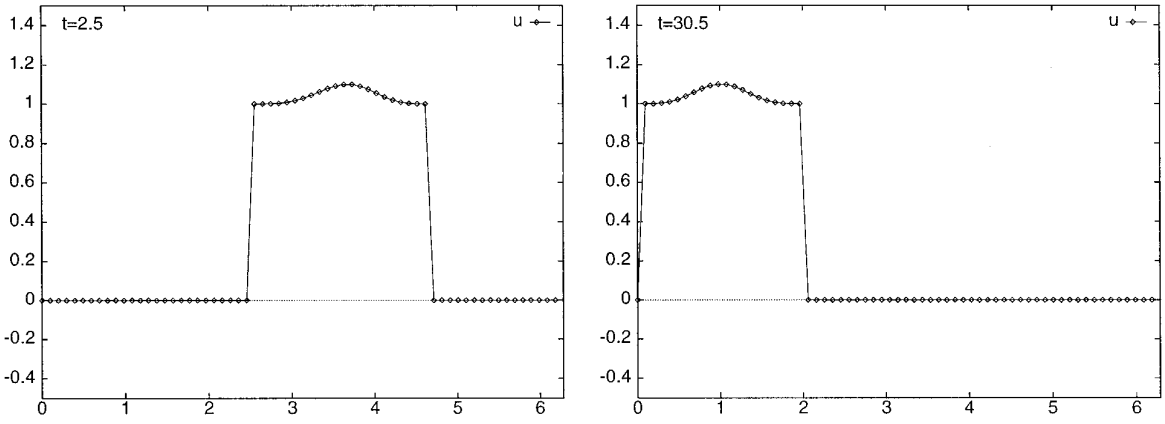

FIG. 9. Solution of (56) computed by Method 2 with $N = 64$, $c = 0.2$.

that for Method 2 the accuracy of the obtained, reconstructed numerical solution is not steadily decreasing when the time increases, as would be the normal behaviour of traditional methods. In fact, the accuracy for Method 2 will normally fluctuate as the time increases, and the main reason for that is the sensitivity of the reconstruction algorithm with respect to the varying distribution of the singularity locations.

The computational results obtained for the test problems considered in Sections 5 and 6, indicate that both for systems with constant and variable coefficients the two methods achieve $O(N^{-(Q+1)})$ accuracy as long as the coefficients are sufficiently smooth. This is in accordance with the theoretical estimates given. The methods also show their applicability to nonsmooth problems for systems where the variable coefficients are not particularly smooth, i.e., problems where the error associated with the truncation of the Fourier series representation for the coefficients cannot be neglected. The results obtained for the selected nonsmooth problems for systems with variable coefficients show that the method of de-truncation, either by Method 1 or by Method 2, represents a considerable improvement relative to the standard Galerkin method for the same problems. In all the test problems presented in this paper Method 1 is seen to perform better than Method 2. However, we consider Method 2 to be more flexible than Method 1 with respect to applications to more general hyperbolic systems and in particular to nonlinear systems, which will be considered elsewhere.

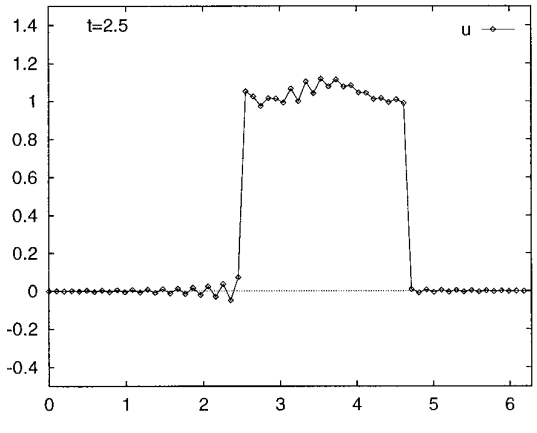

FIG. 10. Computed solution of (56) with the standard Galerkin method, $N = 64$, $c = 0.2$.

APPENDIX: THE RECONSTRUCTION ALGORITHM

We shall here give a brief description of the reconstruction algorithm utilized in what is referred to as Method 2 in this paper. A detailed derivation of the algorithm is given in [9]. We let $u(x)$ be a 2π-periodic discontinuous function which is piecewise smooth on $[0, 2\pi]$ with M singularities in each period. We assume that the truncated Fourier series expansion (2) for $u(x)$ is known. From this we want to reconstruct $u(x)$ by calculating the representation (13), where the quantities Q and γ_j, A_j^n; $j = 1, ..., M$, $n = 0, 1, ..., Q$, and the function $u^Q(x)$ are as defined in Section 2.

We first want to approximately construct the algebraic equation

$$z^M + X_1 z^{M-1} + X_2 z^{M-2} + \cdots + X_{M-1} z + X_M = 0, \quad (64)$$

which has the roots $z_j = e^{-i\gamma_j}$, $j = 1, ..., M$, by calculating the unknown coefficients $X_1, X_2, ..., X_M$. For this purpose we introduce the notation

$$\tilde{C}_k \stackrel{\text{def}}{=} 2\pi i k \hat{u}_k, \quad \tilde{G}_k^Q(0) \stackrel{\text{def}}{=} k^Q \tilde{C}_k, \quad (65)$$

and then successively for $m = 1, 2, ..., Q + 1$,

$$\tilde{G}_k^Q(m) \stackrel{\text{def}}{=} \tilde{G}_k^Q(m-1) + \tilde{G}_{k-1}^Q(m-1)X_1 + \tilde{G}_{k-2}^Q(m-1)X_2 + \cdots + \tilde{G}_{k-M}^Q(m-1)X_M. \tag{66}$$

For each k and m, $\tilde{G}_k^Q(m)$ is seen to be an algebraic expression of degree m with respect to the coefficients X_j, $j = 1, ..., M$, in (64). In [9] it is shown that the following approximate equations hold for $|k|$ large and $k \neq 0, 1, 2, ..., (Q+1)M$,

$$\tilde{G}_k^Q(Q+1) = 0. \tag{67}$$

Since from the relations (66), it follows by induction that for each $j = 1, ..., M$,

$$\frac{\partial}{\partial X_j}\tilde{G}_k^Q(Q+1) = (Q+1)\tilde{G}_{k-j}^Q(Q), \tag{68}$$

(67) can be solved iteratively by the Newton–Raphson method in the case $Q > 0$. More specifically, an integer $D \geq 0$ is chosen as small as possible such that $\hat{u}_k$ is reliable for $|k| < N/2 - D$ and such that $N/2 - D > M(Q+2)$. Then the following approximate system of equations for X_j, $j = 1, ..., M$, is considered:

$$\tilde{G}_k^Q(Q+1) = 0; \quad k = N/2 - M - D, ..., N/2 - 1 - D. \tag{69}$$

In view of (68), it is straightforward to obtain the Jacobian matrix for the system (69). As an initial approximation for the Newton–Raphson iterations, we take the solution of the corresponding linear system resulting from letting $Q = 0$. Having computed the approximate coefficients X_j, $j = 1, ..., M$, (64) is solved by Laguerre's method (see, e.g., [18]), and the approximate singularity locations are obtained from $\gamma_j = -\arg(z_j/|z_j|)$, $j = 1, ..., M$. The jumps are then computed from the approximate linear system

$$\sum_{n=0}^{Q}\sum_{j=1}^{M}\frac{A_j^n e^{-ik\gamma_j}}{(ik)^n} = \tilde{C}_k; \quad k = N/2 - M(Q+1) - D, ..., N/2 - 1 - D. \tag{70}$$

Finally, whenever a complete reconstruction of $u(x)$ is required, the approximate $P_N u^Q(x)$ is obtained from (14).

ACKNOWLEDGMENTS

This paper is partly based on work done while the first author was engaged at the SINTEF Multiphase Flow Laboratory, Trondheim, Norway. The second author has been supported by the Research Council of Norway. The support from SINTEF and the Research Council of Norway is thankfully acknowledged.

REFERENCES

1. S. Abarbanel and D. Gottlieb, in *Progress in Scientific Computing*, edited by E. M. Murman and S. S. Abarbanel, Proceedings, U.S.-Israel Workshop, 1984 (Birkhäuser, Boston, 1985) Vol. 6, p. 345.
2. S. Abarbanel, D. Gottlieb, and E. Tadmor, Technical Report 85-38, NASA-CR-177974, ICASE, 1985; in *Numerical Methods for Fluid dynamics II*, *Proceedings*, *Conf. Reading, 1985*, edited by K. W. Morton and M. J. Baines (Clarendon Press, Oxford, 1986), p. 129.
3. W. Cai, D. Gottlieb, and C.-W. Shu, *Math. Comput.* **52,** 389 (1989).
4. C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics* (Springer-Verlag, New York, 1988).
5. D. C. Champeney, *A Handbook of Fourier Theorems* (Cambridge Univ. Press, Cambridge, 1987).
6. R. Courant and D. Hilbert, *Methods of Mathematical Physics. Vol. II. Partial Differential Equations* (Interscience, New York, 1962).
7. K. S. Eckhoff, Technical Report STF11 F91053, SINTEF Multiphase Flow Laboratory, Trondheim, 1991 (unpublished).
8. K. S. Eckhoff, *Math. Comput.*, **61,** 745 (1993).
9. K. S. Eckhoff, *Math. Comput.*, **64,** 671, (1995).
10. K. S. Eckhoff, *Comput. Methods Appl. Mech. Eng.*, **116,** 103, (1994).
11. A. Erdélyi, W. Magnus, F. Oberhettinger, and F. C. Tricomi, *Higher Transcendental Functions* (McGraw–Hill, New York, 1953).
12. B. Fornberg, *SIAM J. Numer. Anal.* **12,** 509 (1975).
13. D. Gottlieb, Spectral methods for compressible flow problems, in *Proceedings, 9. Int. Conf. Numer. Methods Fluid. Dynamics, Saclay, France, 1984*, edited by Soubbaramayer and J. P. Boujot, *Lecture Notes in Physics*, Vol. 218 (Springer-Verlag, New York/Berlin, 1985), p. 48.
14. D. Gottlieb, L. Lustman, and S. A. Orszag, *SIAM J. Sci. Stat. Comput.* **2,** 296 (1981).
15. D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, Regional Conference Series in Applied Mathematics (SIAM, Philadelphia, 1977).
16. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I, Nonstiff Problems* (Springer-Verlag, Berlin, 1987).
17. A. Majda, J. McDonough, and S. Osher, *Math. Comput.*, **32,** 1041, (1978).
18. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Wettering, *The Art of Scientific Computing. Numerical Recipes*. (Cambridge Univ. Press, Cambridge, 1989).
19. H. Vandeven, *J. Sci. Comput.* **6,** 159 (1991).