JOURNAL OF COMPUTATIONAL PHYSICS 114, 265–279 (1994)

Finite Difference Schemes for Long-Time Integration

ZIGO HARAS AND SHLOMO TA'ASAN*

Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science and The Institute for Computer Applications in Science and Engineering, 76100 Rehovot, Israel

Received June 1, 1993; revised March 7, 1994

A general method for constructing finite difference schemes for long-time integration problems is presented. It is demonstrated for discretizations of first and second space derivatives; however, the approach is not limited to these cases. The schemes are constructed so as to minimize the global truncation error, taking into account the initial data. The resulting second-order compact schemes can be used for integration times fourfold or more longer than previously studied schemes with similar computational complexity. A similar approach was used to obtain improved integration schemes. © 1994 Academic Press, Inc.

1. INTRODUCTION

The simulation of hyperbolic partial differential equations often requires long-time integration. The physical phenomena described by these equations typically possess a range of space and time scales; turbulent fluid flow is a common example. Accurate numerical simulation of this type of process requires proper representation of all the relevant physical scales in the numerical model. These requirements led recently to new interest in Padé approximations also known as compact finite difference schemes [7].

Compact finite difference schemes had long been known and used in numerical analysis [1–3]. They offer a means of obtaining high order approximations to differential operators using narrow stencils. This is achieved by treating the sought derivatives as unknowns and solving a system of equations for them. Typically, the resulting matrices are tridiagonal or pentadiagonal and can be efficiently solved. A detailed exposition of compact schemes and derivation techniques can be found in [12] and will not be pursued here.

In [7] a class of highly accurate compact schemes for first, second, and higher derivatives were presented and analyzed. A notion of *resolving efficiency* was introduced to measure the accuracy of the finite difference approximation of the exact solution over the full range of length scales realizable on a given grid. This criterion was then used to compare various schemes and motivated the design of a new class of schemes, the so-called *schemes with spectral-like resolution.* These are fourth-order pentadiagonal systems with seven-point stencils. Their improved resolution characteristics were obtained by giving up on high formal accuracy, instead requiring that the symbol of the discrete difference operator should agree with the differential operator at three prescribed high frequencies. These fourth-order schemes had better resolution than tenth-order schemes (the highest order obtainable) with the same computational complexity.

A more adequate measure for evaluating finite difference schemes is the L_2 norm of the local truncation error. This measure which takes into account the Fourier components present in the solution and their amplitude was used in [9] to design explicit time marching schemes (i.e., discretizing time and space simultaneously) by analytically solving constrained minimization problems with quadratic cost. This approach has several limitations which severely restrict its usability. The simultaneous treatment of time and space discretizations yields very complex optimization problems. A generalization of this approach to equations in higher dimensions yields large nonlinear constrained minimization problems which are difficult to solve. It, also, seems inadequate to design compact schemes. These difficulties and the use of analytic, rather than numerical, methods makes the suggested approach impractical.

The use of the L_2 norm to evaluate the global truncation error and its application to find an optimal discretization from a one-parameter family of compact schemes can be found in [12].

In [4] a heuristic derivation was done by minimizing the weighted error (in the Fourier space) of the discrete and continuous operators.

The present paper focuses on the construction of finite difference schemes that minimize the discretization error as

* This research was made possible in part by funds granted to the second author through a fellowship program sponsored by the Charles H. Revson Foundation and in part by the National Aeronautics and Space Administration under NASA Contract No. NAS1-19480 and NAS1-18605 while the authors were in residence at ICASE, NASA Langley Research Center, Hampton, Va 23681.

0021-9991/94 $6.00

581/114/2-9

in [9], but with several important differences. First, improved bounds on the truncation error were derived. These enabled us to treat time and space discretizations separately. Specifically, first the space operator is discretized; then a stable time marching scheme is optimized for this discretization. This reduces the minimization problem to two lower dimensional problems that are significantly easier to solve. Further simplification is obtained by optimizing each partial derivative separately, rather than approximating the whole differential operator as was done in [9]. These reductions of problem complexity resulted in a simple and general approach to synthesis of discretization schemes. It enabled us to design highly accurate compact difference schemes and integration formulas for various operators and initial data. The resulting second-order approximations proved to be robust to perturbations in the spectrum of the initial data, exhibiting resolution superior to other known schemes of formally higher order. In particular, we show that the notion of resolving efficiency [7] is too crude a measure as it assumes that all frequencies occur with similar amplitude in the initial data.

It should be emphasized that when considering a refinement process where the mesh size $\Delta x \to 0$, higher order schemes are superior asymptotically. However, for any finite mesh size and given the initial data, there exists a scheme (of the type constructed in this paper) of lower formal accuracy with superior accuracy on that grid for an appropriate initial data set.

The organization of the paper is as follows. In Section 2, Fourier analysis is used to obtain bounds on the truncation error. In Section 3, approximations to derivatives are presented, for the first and second derivatives and first derivative at mid-cell points. Tables I–III list coefficients for these derivatives for various stencils and initial conditions. Improved time integration schemes are developed in Section 4, and their coefficients are listed in Table IV. Section 5 discusses generalization of the present approach to more complex problems. Numerical results are presented in Section 6. Concluding remarks are made in Section 7.

2. BOUNDS ON THE TRUNCATION ERROR

The application of Fourier analysis for the design and evaluation of finite difference schemes can be found in many sources, e.g., [9, 10]. A comprehensive discussion on the use of Fourier analysis in the numerical approximation of hyperbolic problems can be found in [12].

In the following section bounds on the L_2 norm of the error in the discrete solution are derived, accounting for the effect of discretization both in space and time. These estimates are used in subsequent sections to design schemes with improved accuracy on a given grid.

Consider a linear constant coefficient scalar partial differential equation with periodic boundary conditions of the form:

$$\frac{\partial u}{\partial t} = Lu \tag{2.1}$$

$$u(x, 0) = u_0(x). \tag{2.2}$$

Further assume that there exist constants $C > 0$, α such that

$$\|u(t)\| \leqslant Ce^{\alpha t} \|u(0)\|, \qquad t \geqslant 0. \tag{2.3}$$

The discrete analog of this equation can be written as

$$u_h^{n+1} = P(h, \Delta t)\, u_h^n. \tag{2.4}$$

$$u_h^0 = u_0, \tag{2.5}$$

where h is the mesh size in space, and $P(h, \Delta t)$ is a stable finite difference approximation.

We would like to bound the L_2 norm of the error in the discrete solution, for the initial value u_0, given by

$$\begin{aligned} e^2(n\,\Delta t; u_0) &= \|u(n\,\Delta t) - u_h^n\|^2 \\ &= \|e^{Ln\,\Delta t} u_0 - P^n(h, \Delta t)\, u_0\|^2. \end{aligned} \tag{2.6}$$

We will use in the rest of section t, instead of $n\,\Delta t$, to simplify notation. The Fourier transform of Eq. (2.6) yields

$$\int_{-\pi/h}^{\pi/h} |e^{\hat{L}(\omega h)t} - \hat{P}^n(\omega h)|^2\, |\hat{u}_0(\omega h)|^2\, d\omega \tag{2.7}$$

$$\leqslant \int_{-\pi/h}^{\pi/h} (|e^{\hat{L}(\omega h)t} - e^{\hat{L}^h(\omega h)t}| + |e^{\hat{L}^h(\omega h)t} - \hat{P}^n(\omega h)|)^2\, |\hat{u}_0(\omega h)|^2\, d\omega, \tag{2.8}$$

where L^h is the discrete operator approximating L, and $\hat{L}$, $\hat{L}^h$ are their corresponding symbols. Thus, the space and time discretization errors can be bounded separately.

Denote $\hat{L}(\omega h) = \hat{L}_R(\omega h) + \hat{L}_I(\omega h)$ for the real and imaginary parts of $\hat{L}(\omega h)$, respectively; and use a similar decomposition for $\hat{L}^h(\omega h)$. Then

$$\begin{aligned} &e^{\hat{L}(\omega h)t} - e^{\hat{L}^h(\omega h)t} \\ &\quad = e^{\hat{L}(\omega h)t}(1 - e^{(\hat{L}_I^h(\omega h) - \hat{L}_I(\omega h))t} e^{(\hat{L}_R^h(\omega h) - \hat{L}_R(\omega h))t}). \end{aligned} \tag{2.9}$$

The assumption on the growth rate of the solution in time implies that

$$|e^{\hat{L}(\omega h)t}| \leqslant Ce^{\alpha t}, \qquad t \geqslant 0. \tag{2.10}$$

Assume that the discretization is stable, i.e., satisfies a similar bound,

$$|e^{\hat{L}^h(\omega h)t}| \leqslant Ce^{\alpha t}, \qquad t \geqslant 0, \tag{2.11}$$

for the same C and α. The assumption that the differential and discrete solutions are bounded by the same function is not restrictive as one can take the larger bound.

For real numbers θ, γ with $\gamma < 0$, simple geometric considerations yield the bound

$$|1 - e^{i\theta}e^{\gamma}| \leqslant |\theta| + |1 - e^{\gamma}|. \tag{2.12}$$

Combining bounds (2.10) and (2.12) and assuming that $\hat{L}_R^h(\omega h) - \hat{L}_R(\omega h) < 0$ results in

$$\begin{aligned} |e^{\hat{L}(\omega h)t} - e^{\hat{L}^h(\omega h)t}| &\leqslant Ce^{\alpha t}[\,|\hat{L}_I^h(\omega h) - \hat{L}_I(\omega h)|\, t \\ &\quad + |1 - e^{(\hat{L}_R^h(\omega h) - \hat{L}_R(\omega h))t}|\,] \qquad (2.13) \\ &\leqslant Ce^{\alpha t}[\,|\hat{L}_I^h(\omega h) - \hat{L}_I(\omega h)|\, t \\ &\quad + |\hat{L}_R^h(\omega h) - \hat{L}_R(\omega h)|\, t]. \qquad (2.14) \end{aligned}$$

If $\hat{L}_R^h(\omega h) - \hat{L}_R(\omega h) > 0$, this bound can be obtained using the same argument when factoring $e^{\hat{L}^h(\omega h)t}$ in (2.9).

Denote by $\tilde{e}(t; u_0)$ the error due to spatial discretization only, when the initial data is u_0. For a final time T, using (2.14) yields

$$\begin{aligned} \tilde{e}^2(T; u_0) \leqslant (CTe^{\alpha T})^2 \int_{-\pi/h}^{\pi/h} &(|\hat{L}_I^h(\omega h) - \hat{L}_I(\omega h)| \\ &+ |\hat{L}_R^h(\omega h) - \hat{L}_R(\omega h)|)^2\, |\hat{u}_0(\omega h)|^2\, d\omega. \end{aligned} \tag{2.15}$$

Therefore, a difference scheme minimizing the integral (2.15) for a given initial value u_0 will better resolve, in the L_2 norm, the frequencies occurring in the solution.

A similar argument is made with respect to the time integration operator. Consider the difference,

$$\begin{aligned} e^{\hat{L}^h(\omega h)n\,\Delta t} - \hat{P}^n(\omega h) &= (e^{\hat{L}^h(\omega h)\,\Delta t} - \hat{P}(\omega h)) \\ &\quad \times \sum_{j=1}^{n-1} e^{\hat{L}^h(\omega h)j\,\Delta t}\hat{P}^{n-1-j}(\omega h). \end{aligned} \tag{2.16}$$

Under the previous assumptions

$$|e^{\hat{L}^h(\omega h)\,\Delta t}| \leqslant Ce^{\alpha\,\Delta t} \tag{2.17}$$

$$|\hat{P}(\omega h)| \leqslant Ce^{\alpha\,\Delta t}. \tag{2.18}$$

Therefore,

$$\left|\sum_{j=1}^{n-1} e^{\hat{L}^h(\omega h)j\,\Delta t}\hat{P}^{n-1-j}(\omega h)\right| \leqslant nCe^{\alpha T}. \tag{2.19}$$

Hence,

$$|e^{\hat{L}^h(\omega h)t} - \hat{P}^n(\omega h)| \leqslant \frac{CTe^{\alpha T}}{\Delta t}\,|e^{\hat{L}^h(\omega h)\,\Delta t} - \hat{P}(\omega h)|. \tag{2.20}$$

Denote by $\bar{e}(t; u_0)$ the error due to time discretization only, when the initial value is u_0. For a final time T, the bound (2.20) implies that

$$\begin{aligned} \bar{e}^2(T, u_0) \leqslant \left(\frac{CTe^{\alpha T}}{\Delta t}\right)^2 &\int_{-\pi/h}^{\pi/h} |e^{\hat{L}^h(\omega h)\,\Delta t} - \hat{P}(\omega h)|^2 \\ &\times |\hat{u}_0(\omega h)|^2\, d\omega. \end{aligned} \tag{2.21}$$

Combining these estimates yields a bound on the L_2 norm of the global truncation error:

$$\begin{aligned} e^2(T; u_0) \leqslant (CTe^{\alpha T})^2 \int_{-\pi/h}^{\pi/h} \Bigg(&[\,|\hat{L}_I^h(\omega h) - \hat{L}_I(\omega h)| \\ &+ |\hat{L}_R^h(\omega h) - \hat{L}_R(\omega h)|\,]^2 \\ &+ \left[\frac{1}{\Delta t}\,|e^{\hat{L}^h(\omega h)\,\Delta t} - \hat{P}(\omega h)|\right]^2 \Bigg) |\hat{u}_0(\omega h)|^2\, d\omega. \end{aligned} \tag{2.22}$$

In order to minimize the global error in the numerical solution one should use schemes with the least L_2 error bounds. A simple technique to design such schemes consists of the following two steps. First solve the minimization problem,

$$\begin{aligned} \min_{L^h \in \mathscr{A}} \int_{-\pi/h}^{\pi/h} &(|\hat{L}_I^h(\omega h) - \hat{L}_I(\omega h)| \\ &+ |\hat{L}_R^h(\omega h) - \hat{L}_R(\omega h)|)^2\, |\hat{u}_0(\omega h)|^2\, d\omega, \end{aligned} \tag{2.23}$$

where $\mathscr{A}$ is a class of finite difference operators considered, e.g., compact schemes with at least second-order formal accuracy. Let L_*^h denote the optimal discretization found in this stage.

Next, solve the minimization problem

$$\min_{P^h \in \mathscr{B}} \int_{-\pi/h}^{\pi/h} |e^{L_*^h(\omega h)\,\Delta t} - \hat{P}(\omega h)|^2\, |\hat{u}_0(\omega h)|^2\, d\omega, \tag{2.24}$$

where $\mathscr{B}$ is a class of stable time marching schemes and P depends on L_*^h.

The spatial operator is optimized over a parameterized family of finite difference schemes which approximates a derivative appearing in the differential operator. Those parameters are used to obtain the desired formal accuracy, by imposing algebraic relations on them and thus reducing

the problem dimensionality, while the remaining parameters are used to minimize the error (2.15), by solving a quadratic minimization problem.

The same approach may be utilized to design schemes with other particular properties by either modifying the cost function or imposing constraints on the resulting scheme; however, these possibilities will not be pursued here.

The basic observation that the global error is determined by the L_2 norm of the truncation error rather than by the formal accuracy of the scheme implies that one should impose the least order of accuracy required for the scheme consistency. Thus, more parameters may be used to achieve improved performances.

In the present work compact schemes were investigated since they provide a large number of parameters while maintaining narrow stencils.

The design of the time marching operator is similar to the derivation of the space operator. The basic approach remains to minimize the L_2 norm (2.21), where $\hat{L}^h$ is the already designed spatial discretization, over a parameterized family of schemes. Some of the parameters will be used to enforce the order of accuracy, while minimizing over the free ones will yield the desired performance. The stability of the fully discrete operator is accomplished by adding to this minimization problem a constraint that for a prescribed CFL number, the scheme should be stable; then looking for the maximal CFL possible. According to the previous observations it seems reasonable to require only low-order accuracy and use the remaining parameters for other purposes.

At this stage one might discover that the spatial operator discretization allows only very small CFL numbers, yielding the computation inefficient. Therefore, it is necessary to redesign the space operator by adding a constraint which will prohibit such behavior. Thus, although the time and space discretizations are performed separately a close feedback should be maintained between the two design processes.

3. APPROXIMATING SPATIAL DERIVATIVES

3.1. Approximation of the First Derivative

Consider a uniformly spaced mesh whose nodes are indexed by i and its mesh size is given by $h = 1/N$, where $N+1$ is the number of grid points. The variable at node i is $x_i = ih$ and the function value at the nodes, $f_i = f(x_i)$, are given for $0 \leqslant i \leqslant N$. An approximation f'_i to the first derivative $(df/dx)(x_i)$ should be computed as a linear combination of the function values at neighboring grid points. Compact finite difference schemes regard the approximation f'_i as unknown and a system of equations is solved to approximate the first derivative at all nodes, simultaneously. Thus, unlike in finite difference approximations, the derivative at node i depends on function values at all other nodes.

Following [7] we use approximations of the form:

$$\beta f'_{i-2} + \alpha f'_{i-1} + f'_i + \alpha f'_{i+1} + \beta f'_{i+2} = c\frac{f_{i+3} - f_{i-3}}{6h} + b\frac{f_{i+2} - f_{i-2}}{4h} + a\frac{f_{i+1} - f_{i-1}}{2h}. \quad (3.1)$$

A second-order approximation can be obtained by adding a constraint that the Taylor expansion on both sides should agree up to the second-order term, i.e.,

$$a + b + c = 1 + 2\alpha + 2\beta. \quad (3.2)$$

Higher order schemes may be obtained by further matching the next terms in the expansion [7]. However, in this paper merely second-order accuracy is enforced, and the remaining free parameters are chosen so as to improve the accuracy on a given grid.

The symbol of the differentiation operator is given by

$$\hat{L}(\omega h) = i\omega h, \quad (3.3)$$

whereas the symbol of the discrete approximation (3.1) is

$$\hat{L}^h(\omega h) = i\frac{a\sin(\omega h) + (b/2)\sin(2\omega h) + (c/3)\sin(3\omega h)}{1 + 2\alpha\cos(\omega h) + 2\beta\cos(2\omega h)}. \quad (3.4)$$

In view of the bound (2.15), define the following constrained minimization problem whose solution should yield a compact scheme with improved resolution properties,

$$\min_{a,b,c,\alpha,\beta} \int_{-\pi/h}^{\pi/h} |\hat{L}^h(\omega h) - \hat{L}(\omega h)|^2 \, |\hat{u}_0(\omega h)|^2 \, d\omega, \quad (3.5)$$

under the constraint

$$a + b + c = 1 + 2\alpha + 2\beta, \quad (3.6)$$

where $\hat{L}(\omega h)$ and $\hat{L}^h(\omega h)$ are given by (3.3) and (3.4), respectively.

Although the problem was formulated as a constrained minimization problem, it can be transformed by substitution to an unconstrained minimization problem over a reduced set of parameters. Moreover, setting some of the parameters to zero further reduces the problem dimensionality. Since tridiagonal systems of equations are more amenable to numerical solution than pentadiagonal ones, setting $\beta = 0$ seems a plausible choice. Similar considerations might suggest using a narrower stencil obtained by

setting $c = 0$, as well. All those possibilities are presented in Section 6, and several sets of coefficients for different initial data are listed in Table I.

3.2. Approximation of the Second Derivative

The derivation of compact schemes for the second derivative proceeds in an analogous way to the first derivative. The starting point is an approximation of the form

$$\beta f''_{i-2} + \alpha f''_{i-1} + f''_i + \alpha f''_{i+1} + \beta f''_{i+2} = c\frac{f_{i+3} - 2f_i + f_{i-3}}{9h^2} + b\frac{f_{i+2} - 2f_i + f_{i-2}}{4h^2} + a\frac{f_{i+1} - 2f_i + f_{i-1}}{h^2}, \quad (3.7)$$

where f''_i is the approximation to the second derivative at node i. Matching the Taylor series coefficients on both sides of (3.7) yields condition (3.2) for the second-order accuracy.

The symbol of the second derivative is given by

$$\hat{L}(\omega h) = -\omega^2 h^2. \quad (3.8)$$

The symbol of the discrete approximation (3.7) is

$$\hat{L}^h(\omega h) = -[2a(\cos(\omega h) - 1) + (b/2)(\cos(2\omega h) - 1) + (9c/2)(\cos(3\omega h) - 1)]/[1 + 2\alpha\cos(\omega h) + 2\beta\cos(2\omega h)]. \quad (3.9)$$

The constrained minimization problem whose solution is the sought scheme can be formulated as

$$\min_{a,b,c,\alpha,\beta} \int_{-\pi/h}^{\pi/h} |\hat{L}^h(\omega h) - \hat{L}(\omega h)|^2 \, |\hat{u}_0(\omega h)|^2 \, d\omega \quad (3.10)$$

under the constraint

$$a + b + c = 1 + 2\alpha + 2\beta. \quad (3.11)$$

Now, however, $\hat{L}(\omega h)$ and $\hat{L}^h(\omega h)$ are given by (3.8) and (3.9), respectively.

3.3. Approximating the First Derivative on a Cell-Centered Mesh

The approximation of the first derivative at the cell-centered mesh is

$$\beta f'_{i-2} + \alpha f'_{i-1} + f'_i + \alpha f'_{i+1} + \beta f'_{i+2} = c\frac{f_{i+5/2} - f_{i-5/2}}{5h} + b\frac{f_{i+3/2} - f_{i-3/2}}{3h} + a\frac{f_{i+1/2} - f_{i-1/2}}{h}. \quad (3.12)$$

The second order of the approximation is guaranteed by condition (3.2).

The symbol of the differentiation operator is

$$\hat{L}(\omega h) = i\omega h \quad (3.13)$$

while the symbol of the discrete approximation (3.12) is

$$\hat{L}^h(\omega h) = i\,[2\alpha\sin(\omega h/2) + (2b/3)\sin(3\omega h/2) + (2c/5)\sin(5\omega h/2)]/[1 + 2\alpha\cos(\omega h) + 2\beta\cos(2\omega h)]. \quad (3.14)$$

A constrained minimization problem of the same type as in the previous sections was formulated and solved for these symbols.

4. APPROXIMATION OF THE INTEGRATION OPERATOR

The design of integration schemes is substantially limited by the stability requirement which renders high order schemes computationally costly. Therefore, efforts have been made to obtain schemes of lower order with improved characteristics.. Within this approach, the free variables in the Runge–Kutta schemes were set to yield better truncation error [5] or extended stability region [6]. The idea of giving up on formal accuracy in order to obtain better approximation of the wavenumbers relevant to the problem solved may be viewed as a generalization of these ideas.

The discrete time integration of linear constant coefficient partial differential equation

$$\frac{\partial u}{\partial t} = Lu \quad (4.1)$$

amounts to approximation of the exact discrete solution $e^{L^h t}u_0$. Therefore, the integration scheme may be written as

$$P_n(L^h \Delta t) = \sum_{i=0}^{n} a_i (L^h \Delta t)^i, \quad (4.2)$$

where a_i may depend on L^h. The order of the integration scheme is determined by the number of first terms a_i which agrees with the Taylor expansion of e^x.

The derivation of the integration schemes is similar to that of spatial derivative discretization; i.e., a constrained quadratic optimization problem is formulated based on the error estimate (2.21). The solution of this minimization problem yields an improved integration scheme. However, the derivation of integration schemes is more involved than the generation of compact spatial discretization schemes since the stability condition leads to a nonlinearly constrained minimization problem.

Following (2.21), the next optimization problem is defined as

$$\min_{a_i} \int_{-\pi/h}^{\pi/h} |e^{\hat{L}^h(\omega h)\,\Delta t} - P_n(\hat{L}^h(\omega h)\,\Delta t|^2\, |\hat{u}_0(\omega h)|^2\, d\omega, \quad (4.3)$$

subject to the constraints

$$a_i = 1/i!, \qquad 0 \leqslant i \leqslant p, \quad (4.4)$$

$$|\hat{P}_n(\hat{L}^h(\omega h)\,\Delta t)|^2 \leqslant 1, \qquad \omega \in \left[-\frac{\pi}{h}, \frac{\pi}{h}\right], \quad (4.5)$$

where L^h is a discrete approximation of L and p is the order of the n stage formula. Condition (4.4) can be treated by substitution, but the stability condition requires an explicit treatment.

In accordance with our general approach, we argue that second-order formal accuracy suffices. It remains to determine the number of stages in the integration formula. This should be chosen to assure that the error in space and time discretizations (2.15) and (2.21), respectively, will be of similar magnitude. In the present work, five stage schemes of second order were investigated, i.e., $n=5$ and $p=2$. Integration formulas were obtained for optimized seven-point tridiagonal compact schemes approximating the first derivative and were tested for the advection equation in one and two space dimensions.

An important feature of the present appraoch is that once a feasible minimum has been found for a prescribed initial value and a given CFL number, the resulting scheme will be stable for this data. This might enable the use of somewhat larger time steps.

5. APPROXIMATION OF DIFFERENTIAL OPERATORS

The method introduced in the previous sections for generating optimal finite difference approximations for derivatives and time integration schemes for one-dimensional scalar linear constant coefficient equations can be extended to more general cases. In this section, a few straightforward and robust generalizations will be presented. The guiding principle was to maintain simplicity of application, even at the cost of losing some of the attainable accuracy. Clearly, there are other generalizations, and the trade-off between accuracy versus simplicity and robustness should be carefully investigated. Nevertheless, our numerical experiments (see Section 6) demonstrate that the ideas presented here yield significant improvements to previously studied schemes.

The error bounds derived in Section 2 can be generalized for d-dimensional problems; noting that the same proof holds for the d-dimensional case when changing the integration over $[-\pi/h, \pi/h]$ to multi-integration over the box $[-\pi/h, \pi/h]^d$. This suggests that approximation of the differential equation should be obtained by solving constrained optimization problems in d-dimensional Fourier space for a large set of parameters. For some equations, solving this large minimization problem might be essential to achieve accurate schemes. Quite often, though, a set of simpler minimization problems can be obtained by optimizing each partial derivative separately, resulting in highly accurate approximations.

An approach which was successfully tested in the present paper, divides the optimization process into two stages. First, a set of schemes are designed for a large enough variety of typical initial data (e.g., Gaussians with different parameters, in our examples). Once this precomputation is performed its results are saved to be used in subsequent simulations. In the actual simulation, the Fourier transform $\hat{u}_0$ of the initial data u_0 is computed. This $\hat{u}_0$ should be used to design the optimal space and time discrete operators, by solving the corresponding optimization problems. Alternatively, we suggest choosing, from the previously designed schemes, one corresponding to the initial data which best approximates $\hat{u}_0$. A further simplification can be obtained if $\hat{u}_0$ is approximated by a product of two Gaussians. This construction might introduce larger errors, but it is very simple and seems to yield substantial improvement over the standard schemes. Thus, the discretization of the partial derivatives is determined by approximating $\hat{u}_0$ as a product of one-dimensional functions (one-dimensional Gaussians, in our examples) for which optimized schemes were designed. Each partial derivative is discretized using the corresponding one-dimensional optimized scheme. The time marching scheme is selected from the set of schemes corresponding to the approximating one-dimensional functions. In the present work, the selection was done by computing the L_2 error norm of each candidate integration scheme when applied to the approximate initial data with the already determined discretizations, then selecting the scheme which yields the minimum error norm. This computation, too, can be done prior to the actual simulation for a large set of typical initial data. Thus, the marching scheme selection can be done by looking up in a precomputed table. The robustness of the proposed schemes to perturbations in the initial data yields this optimization very efficient, as can be seen in the numerical results presented in Section 6. It should be noted that the time required to obtain an appropriate scheme using this approach is negligible relative to simulation time.

When the frequencies present in the solution change with time, e.g., due to a time dependent source term, the computation of the optimized schemes should be repeated, once a large cumulative change has occurred. Still, the relative cost of this computation is minimal.

The Fourier transform gives the energy content of the

whole initial data. It may occur that the initial data is smooth at some regions of the computational domain and oscillatory in others, in which case the designed approximation will give good performance over the whole domain. One can do better by computing a different scheme for each region and using a smooth weighted sum of the resulting schemes near region boundaries. This requires computing the Fourier transform locally in each region. The localization to a particular region can be achieved by multiplying $\hat{u}_0$ by a C^∞ function with a compact support which encloses the region.

In some cases, systems of equations may be treated in a similar way. Let us first look at a one-dimensional first-order system,

$$\begin{aligned} u_t &= Au_x \\ u(x, 0) &= u_0(x), \end{aligned} \tag{5.6}$$

where A is a $p \times p$ symmetric matrix. Let $\Lambda = P^{-1}AP$ be a diagonal matrix and denote $v = Pu$. The discretization of the system

$$\begin{aligned} v_t &= \Lambda v_x \\ v(x, 0) &= Pu_0(x) \end{aligned} \tag{5.7}$$

can be done in an analogous way to the scalar case, except for the time marching scheme which should be chosen from a set of candidate schemes (as for the multidimensional scalar equations). Thus, highly accurate discretization of the system (5.6) can be achieved by first discretizing (5.7) and using the identity $u_x = P^{-1}v_x$. For systems in higher dimensions,

$$\begin{aligned} u_t &= \sum_{i=1}^{n} A_i \frac{\partial u}{\partial x_i} \\ u(x, 0) &= u_0(x), \end{aligned} \tag{5.8}$$

we require that all A_i be symmetric and simultaneously diagonalizable. For this case the proof in Section 2 applies and one obtains a similar error estimate.

The proposed schemes might be useful for nonlinear equations, as well. There, one should design the schemes for the linearized equation and will be obliged to modify them, once a large change in the amplitude of the wavenumbers appearing in the solution occurs.

6. NUMERICAL RESULTS

The numerical examples in the following section clearly demonstrate the accuracy and robustness of the proposed schemes. The spatial discretizations are compared to those presented in [7], where the same families of parameterized schemes were investigated. In particular, our second-order pentadiagonal seven-point stencil scheme is compared to the fourth-order *spectral-like* scheme with the same computational complexity, which was shown [7] to have resolution superior to the tenth-order scheme having the same structure.

The accuracy of our approximations is demonstrated by applying the optimized scheme to the initial data that it was designed to best approximate, comparing the results to the exact solution and the solution obtained by using a *spcetral-like* scheme. First, all solutions are integrated until the error in the solution approximated with the *spectral-like* scheme is visible. Typically at this stage the solution discretized with the optimized scheme is almost indistinguishable from the exact solution. This gives a rough estimate on the time that the spectral-like scheme could be efficiently used. Next, the solutions are further integrated, until an error of similar magnitude prevails in the solution approximated with the optimized scheme. Typically by now, the solution corresponding to the other scheme greatly differs from the exact solution. The time that this occurs is an order of magnitude larger than the spectral-like scheme effectiveness time. Thus, we may conclude that the optimized schemes may be used for integration times an order of magnitude longer than *spectral-like* schemes with similar computational complexity.

The robustness of an optimized scheme is shown by using it with initial data different from those it was designed to approximate. In these examples the solutions are integrated until a visible error appears in the solution corresponding to the optimized scheme. At this time the solution corresponding to the spectral-like scheme bears only a little similarity to the exact solution.

Although these comparisons are not quantitatively precise, the resulting qualitative conclusions are surely valid. The accuracy and robustness of the time marching schemes is demonstrated by examples that are similarly constructed, where the optimized fully discrete scheme is compared with a tridiagonal example with the same stencil size using fourth-order Runge–Kutta. The last two-dimensional example with variable coefficients demonstrates the robustness of the schemes and the applicability of the generalization to higher dimensions as suggested in Section 5.

6.1. Approximation of Derivatives

The constrained minimization problem (3.5) for the space discretization can be easily solved by substitution using (3.2). Differentiation of the resulting quadratic form provides a set of necessary conditions holding at the minimum. This nonlinear system can be solved using the Newton method, yielding a local minimum. Since the schemes obtained using this process significantly improve

TABLE I

Coefficients of First Derivative Approximations for Various Initial Conditions

$\hat{u}_0$	Schemes with $\beta = c = 0$		
$e^{-\omega^2}$	$\alpha = 0.3793894912$,	$a = 1.575573790$,	$b = 0.1832051925$
$e^{-2\omega^2}$	$\alpha = 0.3534620435$,	$a = 1.566965775$,	$b = 0.1399583152$
$e^{-3\omega^2}$	$\alpha = 0.3461890571$,	$a = 1.5633098070$,	$b = 0.1290683071$
$e^{-4\omega^2}$	$\alpha = 0.3427812069$,	$a = 1.5614141543$,	$b = 0.124148259$
$e^{-5\omega^2}$	$\alpha = 0.3408027739$,	$a = 1.5602604992$,	$b = 0.121345048$
$e^{-6\omega^2}$	$\alpha = 0.3395099051$,	$a = 1.5594855939$,	$b = 0.119534216$
$e^{-7\omega^2}$	$\alpha = 0.3385987444$,	$a = 1.5589295176$,	$b = 0.1182679712$

$\hat{u}_0$	Schemes with $\beta = 0$			
$e^{-\omega^2}$	$\alpha = 0.4303030674$,	$a = 1.5567577428$,	$b = 0.3451622238$,	$c = -0.0413138317$
$e^{-2\omega^2}$	$\alpha = 0.3991476265$,	$a = 1.5636386371$,	$b = 0.2563784492$,	$c = -0.0217218334$
$e^{-3\omega^2}$	$\alpha = 0.3904091387$,	$a = 1.5638887738$,	$b = 0.2348222711$,	$c = -0.0178927675$
$e^{-4\omega^2}$	$\alpha = 0.3863287472$,	$a = 1.5637497712$,	$b = 0.2252138483$,	$c = -0.0163061252$
$e^{-5\omega^2}$	$\alpha = 0.3839604005$,	$a = 1.5635937780$,	$b = 0.21976694619$,	$c = -0.0154399233$
$e^{-6\omega^2}$	$\alpha = 0.3824122042$,	$a = 1.5634617985$,	$b = 0.21625718276$,	$c = -0.0148945794$

[redacted]

$e^{-7\omega^2}$	$\alpha = 0.3813206436$,	$a = 1.5633544597$,	$b = 0.21380659696$,	$c = -0.0145197694$

$\hat{u}_0$	General schemes				
$e^{-\omega^2}$	$\alpha = 0.5779403671$,	$\beta = 0.0890143475$,	$a = 1.3030269541$,	$b = 0.994883769$,	$c = 0.0359987066$
$e^{-2\omega^2}$	$\alpha = 0.5801818925$,	$\beta = 0.0877284887$,	$a = 1.3058941939$,	$b = 0.9975884963$,	$c = 0.0323380724$
$e^{-3\omega^2}$	$\alpha = 0.5821143744$,	$\beta = 0.0867224075$,	$a = 1.3086733956$,	$b = 0.9990906893$,	$c = 0.0299094788$
$e^{-4\omega^2}$	$\alpha = 0.5831688320$,	$\beta = 0.0862000893$,	$a = 1.3102698137$,	$b = 0.9997174262$,	$c = 0.0287506026$
$e^{-5\omega^2}$	$\alpha = 0.5838221871$,	$\beta = 0.0858844217$,	$a = 1.3112828763$,	$b = 1.0000513827$,	$c = 0.0280789585$
$e^{-6\omega^2}$	$\alpha = 0.58426518608$,	$\beta = 0.0856735831$,	$a = 1.31197935750$,	$b = 1.00025665126$,	$c = 0.02764152958$
$e^{-7\omega^2}$	$\alpha = 0.58458494112$,	$\beta = 0.08552292859$,	$a = 1.31248665912$,	$b = 1.00039487751$,	$c = 0.02733420278$

previously known schemes [7], no attempts were made to find the other zeroes of the nonlinear system, searching for better minima.

Three types of schemes were studies: (a) tridiagonal with five-point stencil, i.e., $\beta = c = 0$; (b) tridiagonal with seven-point stencil, i.e., $\beta = 0$; (c) pentadiagonal with seven-point stencil. The initial approximation to the Newton iteration was, typically, a compact scheme with the same structure, taken from [7].

It can be observed, in Figs. 1 and 5, that the modulas of the symbol of the optimized pentadiagonal scheme for the first and second derivatives is larger than the modulus of the differential symbol. This error is larger for schemes generated to approximate narrower spectra. The overshooting occurs in the highest end of the spectrum for wavenumbers not appearing in the solution. However, since the stability of a scheme is determined by the values assumed by $\hat{L}^h(\omega h)$ [11], this type of scheme is applicable only with small CFL. Moreover, the desired robustness is limited by this phenomenon. Therefore, this behavior of the approximation cannot be ignored. A possible remedy can be found by searching for other minimizers of the quadratic this limiting property, but with reduced resolution, similar to the tridiagonal schemes. Other possible directions, e.g., further looking for other minima or penalizing in the cost function for this behavior were not explored. This is because we believe that for practical applications pentadiagonal systems are too costly to solve, whereas the tridiagonal

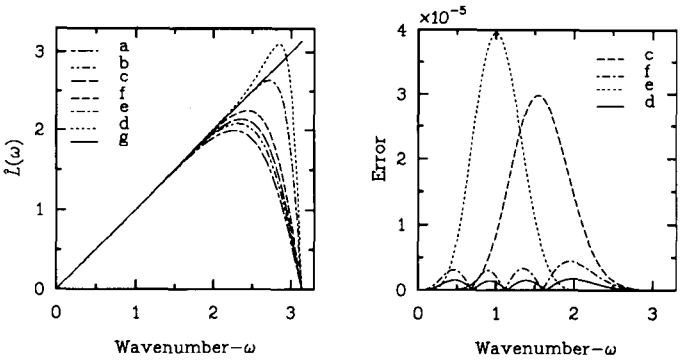

FIG. 1. Symbols (left) and absolute value of error (right) for d/dx. $\hat{u}_0 = e^{-2\omega^2}$: (a) sixth-order tridiagonal scheme ($\beta = c = 0$); (b) second-order optimized tridiagonal scheme ($\beta = c = 0$); (c) eight-order tridiagonal scheme ($\beta = 0$); (d) second-order optimized tridiagonal scheme ($\beta = 0$);

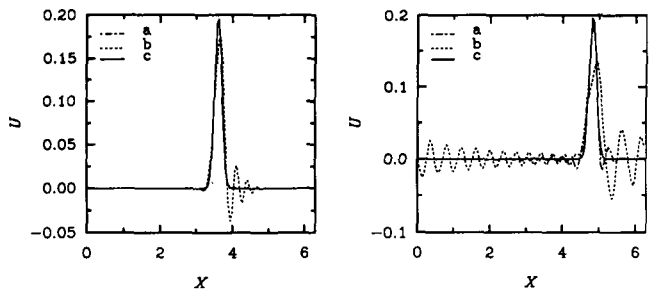

FIG. 2. Long-time integration of the equation $u_t = u_x$, $\hat{u}_0 = e^{-2\omega^2}$, $\sigma = 0.8$: (a) pentadiagonal scheme optimized for $\hat{u}_0 = e^{-2\omega^2}$; (b) spectral-like pentadiagonal scheme; (c) exact solution.

schemes offer similar resolution characteristics, are easier to solve, and do not suffer from this deficiency. The pentadiagonal schemes are given mainly for theoretical reasons as a counterpart to the spectral-like approximations.

A proper appreciation of the superiority of the proposed schemes can be gained by using them to integrate hyperbolic equations for long times, provided the integration introduces only negligible numerical errors. This requirement necessitates either using high order integration schemes or employing exact integration, as was done in the present work. The experiments described in the next subsections clearly demonstrate the superior behavior of the proposed optimized schemes.

6.1.1. *Approximation of the First Derivative*

Compact finite difference schemes were designed and tested for initial data having a Fourier transform of the form $e^{-\alpha\omega^2}$ for several values of α. In Fig. 1, the symbols of schemes corresponding to $\alpha = 2$ are plotted, as well as the weighted error

$$|\hat{L}(\omega h) - \hat{L}^h(\omega h)|\,|\hat{u}_0(\omega h)| \tag{6.9}$$

for the more accurate schemes. The coefficients of the optimized schemes can be found in Table I. The coefficients

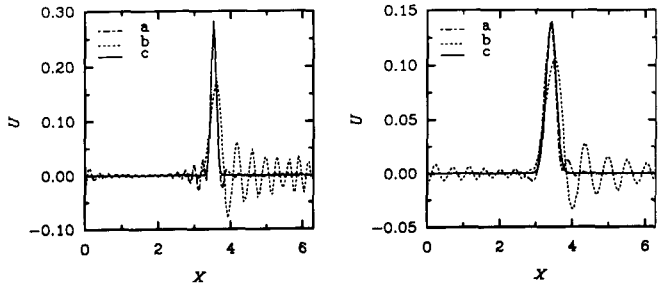

FIG. 3. Long-time integration of the equation $u_t = u_x$, $\sigma = 0.8$. Initial solution on the left figure was $\hat{u}_0 = e^{-\omega^2}$; on the right figure it was $\hat{u}_0 = e^{-4\omega^2}$: (a) pentadiagonal scheme optimized for $\hat{u}_0 = e^{-2\omega^2}$; (b) spectral-like pentadiagonal scheme; (c) exact solution.

of the other schemes were taken from [7]. For scheme (a) the coefficients were

$$\alpha = \tfrac{1}{3}, \qquad \beta = 0, \qquad a = \tfrac{14}{9}, \qquad c = 0. \tag{6.10}$$

The coefficients of scheme (c) were

$$\alpha = \tfrac{3}{8}, \qquad \beta = 0, \qquad a = \tfrac{25}{16}, \qquad b = \tfrac{1}{5}, \qquad c = -\tfrac{1}{80}. \tag{6.11}$$

The coefficients of the spectral-like scheme (e) were

$$\alpha = 0.5771439, \qquad \beta = 0.0896406, \qquad a = 1.3025166, \qquad b = 0.99355, \qquad c = 0.03750245. \tag{6.12}$$

It can be seen that each optimized scheme better approximates the differential operator than its non-optimized counterpart. In Fig. 1, one can observe that although the symbol of the spectral-like pentadiagonal scheme follows the differential symbol for more wavenumbers than the tridiagonal scheme, the L_2 norm of truncation error of tridiagonal scheme is somewhat smaller for this data. This can be explained by noting that the error of the tridiagonal scheme is mainly in the high frequencies while the spectral-like scheme has a large error at the smoother Fourier components, where the present initial data has more energy. The spectral-like scheme attains better resolution at the expense of larger error in lower frequencies. The error in the optimized schemes is significantly smaller than in their counterparts. More precisely, computing the error norms reveals that the error in the tridiagonal scheme is about six times larger than in the optimized tridiagonal scheme while the error norm of the spectral-like scheme is about 17 times larger than in the optimized pentadiagonal. The plot of the absolute value of the error reveals that the L_2 norm was used as a minimization criteria. This can be

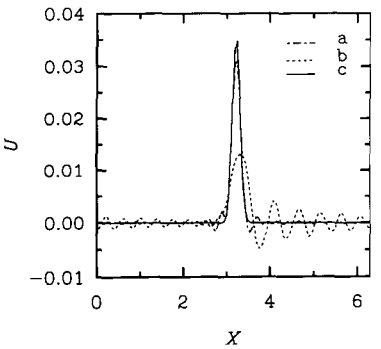

FIG. 4. Integration of the equation $u_t = u_{x_2} + u_{y_3}$ $\sigma = 0.8$ using pentadiagonal schemes. Initial solution was $\hat{u}_0 = e^{-(\omega_1^2 + 5\omega_2^2)}$ rotated at an angle of $\pi/4$. This data was approximated by unrotated gaussian $e^{-(\omega_1^2 + 2\omega_2^2)}$: (a) optimized pentadiagonal scheme; (b) spectral-like pentadiagonal scheme; (c) exact solution.

seen from the several sign changes of the error of the optimized schemes, being in accordance with the averaging property of the chosen norm.

Figure 2 demonstrates the better resolution of the optimized scheme by exact integration in Fourier space on a 128-point grid with the pentadiagonal spectral-like scheme and the pentadiagonal optimized scheme, the equation

$$\frac{\partial U}{\partial t}=\frac{\partial U}{\partial x}, \tag{6.13}$$

$\sigma=0.8$ (σ being the CFL number) was used. It is shown that at time $T=5000$, the error in the solution using the optimized scheme is smaller than the error at time $T=500$ when using the spectral-like scheme. This suggests that the optimized scheme can be used for the integration time at least 10 times longer than the spectral-like scheme, in close accordance with the ratio of the error norms.

Figure 3 displays the scheme's robustness to perturbation in initial condition. The solution integrated with the optimized scheme far better approximates the exact solution than the one employing the spectral-like approximation, even for initial data different from the ones it was designed to resolve. This holds for both smoother and more oscillatory initial data. Although those examples do not give a quantitative view on the relative efficiency of the schemes for those initial data, one can see in both figures that, by the time the solution with the optimized scheme developed significant error, the error in the one corresponding to the spectral-like scheme is so large that it no longer approximates the exact solution.

Figure 4 shows a two-dimensional equation which demonstrates the robustness of the proposed schemes. In this example the initial data was taken to be the Gaussian $e^{-(\omega_1^2+5\omega_2^2)}$ rotated at an angle of $\pi/4$. Then the program searched for initial data of the form $e^{-(n_1\omega_1^2+n_2\omega_2^2)}$, for the integers $1\leqslant n_1, n_2\leqslant 7$, which yielded the best approximation to the initial data. The pentadiagonal schemes optimized for initial data $e^{-n_1\omega^2}$ and $e^{-n_2\omega^2}$ were then used to compute u_x and u_y, respectively. In this example $n_1=3$ and $n_2=2$. The resulting semi-discrete system was solved by exact integration in Fourier space on a 128×128 grid. The plot shows a cut through the solution in the x direction through the maximum point of the solution. While the solution corresponding to the optimized discretization closely approximates the exact solution, the solution discretized

TABLE II

Coefficients of Second Derivative Approximations for Various Initial Conditions

$\hat{u}_0$	Schemes with $\beta=c=0$		
$e^{-\omega^2}$	$\alpha=0.2285657609$,	$a=1.0139538409$,	$b=0.4431776810$
$e^{-2\omega^2}$	$\alpha=0.2028150072$,	$a=1.0598135170$,	$b=0.3458164974$
$e^{-3\omega^2}$	$\alpha=0.1952770765$,	$a=1.0716695072$,	$b=0.3188846458$
$e^{-4\omega^2}$	$\alpha=0.1917151916$,	$a=1.0770076313$,	$b=0.3064227519$
$e^{-5\omega^2}$	$\alpha=0.1896428309$,	$a=1.0800332355$,	$b=0.29925242633$
$e^{-6\omega^2}$	$\alpha=0.18828772017$,	$a=1.0819792783$,	$b=0.29459616204$
$e^{-7\omega^2}$	$\alpha=0.18733255632$,	$a=1.0833354275$,	$b=0.29132968512$

$\hat{u}_0$	Schemes with $\beta=0$			
$e^{-\omega^2}$	$\alpha=0.3125176074$,	$a=0.7701351999$,	$b=0.9469577413$,	$c=-0.0920577265$
$e^{-2\omega^2}$	$\alpha=0.2702488609$,	$a=0.8863525584$,	$b=0.7065172637$,	$c=-0.0523721002$
$e^{-3\omega^2}$	$\alpha=0.2580699154$,	$a=0.9170322739$,	$b=0.6425330979$,	$c=-0.0434255409$
$e^{-4\omega^2}$	$\alpha=0.2523894606$,	$a=0.9308701065$,	$b=0.6135153110$,	$c=-0.0396064963$
$e^{-5\omega^2}$	$\alpha=0.2491062584$,	$a=0.9387256232$,	$b=0.59698635859$,	$c=-0.0374994649$
$e^{-6\omega^2}$	$\alpha=0.2469677390$,	$a=0.9437849227$,	$b=0.5863166347$,	$c=-0.0361660397$
$e^{-7\omega^2}$	$\alpha=0.2454642305$,	$a=0.9473144209$,	$b=0.5788609571$,	$c=-0.0352469171$

$\hat{u}_0$	General schemes				
$e^{-\omega^2}$	$\alpha=0.5024750577$,	$\beta=0.0554440666$,	$a=0.2150536435$,	$b=1.7246523136$,	$c=0.1761322914$
$e^{-2\omega^2}$	$\alpha=0.5041582074$,	$\beta=0.0527585356$,	$a=0.2120465713$,	$b=1.7488409942$,	$c=0.1529459205$
$e^{-3\omega^2}$	$\alpha=0.5053986368$,	$\beta=0.05124444502$,	$a=0.2112256102$,	$b=1.7609579037$,	$c=0.1411026601$
$e^{-4\omega^2}$	$\alpha=0.5061009898$,	$\beta=0.0504756862$,	$a=0.2110263782$,	$b=1.7667867767$,	$c=0.1353401973$
$e^{-5\omega^2}$	$\alpha=0.5065435817$,	$\beta=0.0500170894$,	$a=0.21097836343$,	$b=1.7701652358$,	$c=0.1319777431$
$e^{-6\omega^2}$	$\alpha=0.5068465815$,	$\beta=0.0497133535$,	$a=0.2109761550$,	$b=1.7723629924$,	$c=0.1297807226$
$e^{-7\omega^2}$	$\alpha=0.5070666579$,	$\beta=0.0494975852$,	$a=0.2109890794$,	$b=1.7739051293$,	$c=0.1282342776$

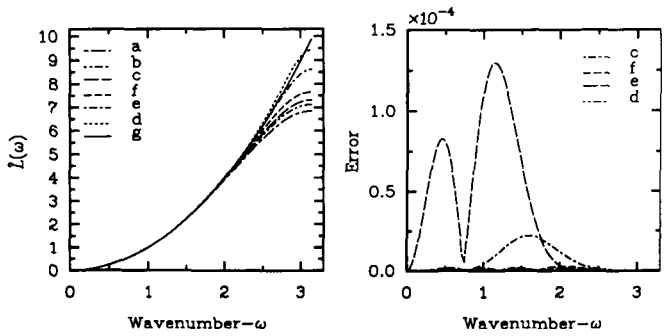

FIG. 5. Symbol (left) and absolute value of error (right) for d^2/dx^2. $\hat{u}_0 = e^{-2\omega^2}$: (a) sixth-order tridiagonal scheme ($\beta = c = 0$); (b) second-order optimized tridiagonal scheme ($\beta = c = 0$); (c) eight-order tridiagonal scheme ($\beta = 0$); (d) second-order optimized tridiagonal scheme ($\beta = 0$); (e) spectral-like pentadiagonal; (f) optimized pentadiagonal; (g) exact symbol. Schemes were optimized for $\hat{u}_0 = e^{-2\omega^2}$.

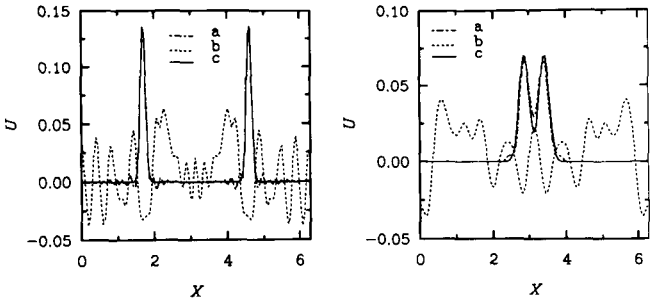

FIG. 7. Long-time integration for $u_{tt} = u_{xx}$, $\sigma = 0.8$. Initial solution on the left figure was $\hat{u}_0 = e^{-\omega^2}$; on the right figure it was $\hat{u}_0 = e^{-4\omega^2}$: (a) Pentadiagonal scheme optimized for $\hat{u}_0 = e^{-2\omega^2}$; (b) Spectral-like pentadiagonal scheme; (c) exact solution.

with the spectral-like scheme bears very little resemblance to the exact solution.

6.1.2. *Approximation of the Second Derivative*

The coefficients of compact schemes for various initial conditions having a Fourier transform of the form $e^{-\alpha\omega^2}$ can be found in Table II.

Figure 5 plots absolute value of the symbol of the second derivative and the weighted error, for $\alpha = 2$. The parameters of the optimized schemes can be found in Table II. The coefficients of the other schemes were taken from [7]. Scheme (a) is given by

$$\alpha = \tfrac{2}{11}, \qquad \beta = 0, \qquad a = \tfrac{12}{11}, \qquad b = \tfrac{3}{11}, \qquad c = 0. \tag{6.14}$$

The coefficients of scheme (c) are

$$\alpha = \frac{9}{38}, \qquad \beta = 0, \qquad a = \frac{696 - 1191\alpha}{428}, \qquad b = \frac{2454\alpha - 294}{535}, \qquad c = -\frac{1179\alpha - 344}{2140}. \tag{6.15}$$

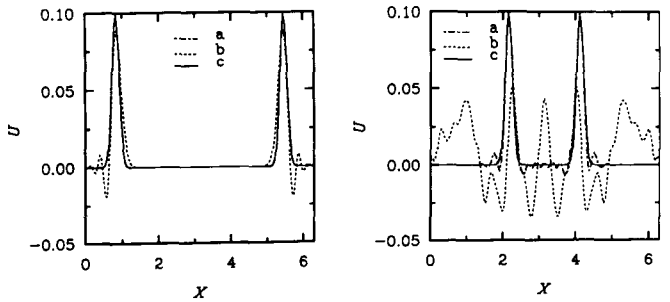

FIG. 6. Long-time integration for $u_{tt} = u_{xx}$, $\hat{u}_0 = e^{-2\omega^2}$, $\sigma = 0.8$: (a) Pentadiagonal scheme optimized for $\hat{u}_0 = e^{-2\omega^2}$; (b) spectral-like pentadiagonal scheme; (c) exact solution.

The coefficients of the spectral-like pentadiagonal scheme (e) are

$$\alpha = 0.50209266, \qquad \beta = 0.05669169, \qquad a = 0.21564935,$$

$$b = 1.723322, \qquad c = 0.1765973. \tag{6.16}$$

It can be seen that the error in the spectral-like schemes is significantly larger than in the optimized ones. It is interesting to note that, again, for this specific data the L_2 error norm of the spectral-like scheme is about an order of magnitude larger than the non-optimized tridiagonal scheme. This phenomenon suggests that the resolution efficiency is a poor estimate for discretizations evaluation. Computing the error norms reveals that the error in the optimized tridiagonal scheme is about seven times smaller than in the non-optimized scheme, whereas the error in the optimized pentadiagonal scheme is 70 times smaller than the spectral-like scheme, for this given data.

The efficiency of the pentadiagonal schemes was compared by integrating the wave equation:

$$\frac{\partial^2 U}{\partial t^2} = \frac{\partial^2 U}{\partial x^2}. \tag{6.17}$$

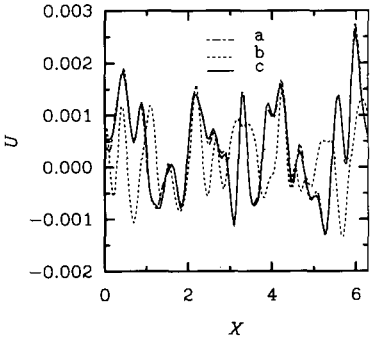

FIG. 8. Long-time integration for $u_{tt} = u_{xx} + u_{yy}$, $\sigma = 0.8$ using pentadiagonal schemes. Initial solution on the left figure was $\hat{u}_0 = e^{-(\omega_1^2 + 5\omega_2^2)}$ rotated at an angle of $\pi/4$. This data was approximated by unrotated gaussian $e^{-(3\omega_1^2 + 2\omega_2^2)}$: (a) optimized pentadiagonal scheme; (b) Spectral-like pentadiagonal; (c) exact solution.

This equation was put in a system form:

$$\begin{pmatrix} u \\ v \end{pmatrix}_t = \begin{pmatrix} 0 & 1 \\ (\partial^2/\partial x^2) & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}. \tag{6.18}$$

This form of discretization of the wave equation was successfully applied in [4] for problems in elasticity. This system was solved using exact integration on a 128-point grid and the results are given in Figs. 6–7 demonstrating the improved accuracy of the optimized scheme and its robustness, respectively. Figure 6 demonstrates that the optimized scheme can be used for integration time 50 times, or more, longer than the spectral-like scheme. In Fig. 7 the scheme robustness is clearly shown for initial data smoother or more oscillatory than the data for which the scheme was designed. In both cases, by the time a significant error occurs in the solution discretized with the optimized scheme, the solution corresponding to the spectral-like scheme totally differs from the exact solution.

The initial solution and its approximation, for the two-dimensional problem in Fig. 8 were obtained similarly to those of the example in Fig. 4. While the solution integrated with the optimized scheme closely approximates the exact solution, it is hard to see that the solution corresponding to the spectral-like scheme indeed approximates the same problem.

TABLE III

Coefficients of Mid-Cell Approximation of the First Derivative for Various Initial Conditions

$\hat{u}_0$	Schemes with $\beta = c = 0$		
$e^{-\omega^2}$	$\alpha = 0.1824466564$,	$a = 0.9847348088$,	$b = 0.3801585039$
$e^{-2\omega^2}$	$\alpha = 0.1621215357$,	$a = 1.0026558711$,	$b = 0.3215872003$
$e^{-3\omega^2}$	$\alpha = 0.1560892225$,	$a = 1.0076143702$,	$b = 0.3045640747$
$e^{-4\omega^2}$	$\alpha = 0.1532174394$,	$a = 1.0099120548$,	$b = 0.2965228240$
$e^{-5\omega^2}$	$\alpha = 0.1515399131$,	$a = 1.0112348225$,	$b = 0.2918450036$
$e^{-6\omega^2}$	$\alpha = 0.1504402935$,	$a = 1.0120939889$,	$b = 0.2887865980$
$e^{-7\omega^2}$	$\alpha = 0.1496639344$,	$a = 1.0126967653$,	$b = 0.2866311035$

$\hat{u}_0$	Schemes with $\beta = 0$			
$e^{-\omega^2}$	$\alpha = 0.2803531992$,	$a = 0.8656018611$,	$b = 0.7202754832$,	$c = -0.0251709460$
$e^{-2\omega^2}$	$\alpha = 0.2421691108$,	$a = 0.9108711860$,	$b = 0.5897758895$,	$c = -0.0163088538$
$e^{-3\omega^2}$	$\alpha = 0.2311768224$,	$a = 0.9233491904$,	$b = 0.5531540626$,	$c = -0.0141496081$
$e^{-4\omega^2}$	$\alpha = 0.2260281312$,	$a = 0.9290969691$,	$b = 0.5361564844$,	$c = -0.0131971911$
$e^{-5\omega^2}$	$\alpha = 0.2230456380$,	$a = 0.9323967450$,	$b = 0.5263571378$,	$c = -0.0126626068$
$e^{-6\omega^2}$	$\alpha = 0.2211004185$,	$a = 0.9345368034$,	$b = 0.5199847302$,	$c = -0.0123206966$
$e^{-7\omega^2}$	$\alpha = 0.2197316282$,	$a = 0.9360368674$,	$b = 0.5155096846$,	$c = -0.0120832956$

$\hat{u}_0$	General schemes				
$e^{-\omega^2}$	$\alpha = 0.3392424034$,	$\beta = 0.0126851467$,	$a = 0.7880308119$,	$b = 0.8956208871$,	$c = 0.0202034010$
$e^{-2\omega^2}$	$\alpha = 0.3364203680$,	$\beta = 0.0159838314$,	$a = 0.7894607720$,	$b = 0.8790559502$,	$c = 0.0362916767$
$e^{-3\omega^2}$	$\alpha = 0.3359766282$,	$\beta = 0.0164557610$,	$a = 0.7895453413$,	$b = 0.8768367139$,	$c = 0.0384827231$
$e^{-4\omega^2}$	$\alpha = 0.3358345755$,	$\beta = 0.0166014190$,	$a = 0.78955615727$,	$b = 0.87616875736$,	$c = 0.03914707436$
$e^{-5\omega^2}$	$\alpha = 0.33577201042$,	$\beta = 0.01666433833$,	$a = 0.78955722003$,	$b = 0.87588406207$,	$c = 0.03943141540$
$e^{-6\omega^2}$	$\alpha = 0.33573907328$,	$\beta = 0.01669706335$,	$a = 0.78955658369$,	$b = 0.87573722427$,	$c = 0.03957846531$
$e^{-7\omega^2}$	$\alpha = 0.33571963682$,	$\beta = 1.67162152050$,	$a = 0.78955572985$,	$b = 0.87565178238$,	$c = 0.03966419181$

$\hat{u}_0$	Schemes with $\beta = 0$, designed to approximate d^2/dx^2			
$e^{-\omega^2}$	$\alpha = 0.2949304593$,	$a = 0.8473898079$,	$b = 0.7718938474$,	$c = -0.0294227367$
$e^{-2\omega^2}$	$\alpha = 0.2482825125$,	$a = 0.9037600128$,	$b = 0.6104318128$,	$c = -0.0176268008$
$e^{-3\omega^2}$	$\alpha = 0.2349387889$,	$a = 0.9190859222$,	$b = 0.5656763757$,	$c = -0.0148847202$
$e^{-4\omega^2}$	$\alpha = 0.2287385754$,	$a = 0.9260661646$,	$b = 0.5451127700$,	$c = -0.0137017838$
$e^{-5\omega^2}$	$\alpha = 0.2251628777$,	$a = 0.9300484196$,	$b = 0.5333228985$,	$c = -0.0130455627$
$e^{-6\omega^2}$	$\alpha = 0.2228371850$,	$a = 0.9326209622$,	$b = 0.5256822919$,	$c = -0.0126288841$
$e^{-7\omega^2}$	$\alpha = 0.2212037269$,	$a = 0.9344193325$,	$b = 0.52032910793$,	$c = -0.0123409866$

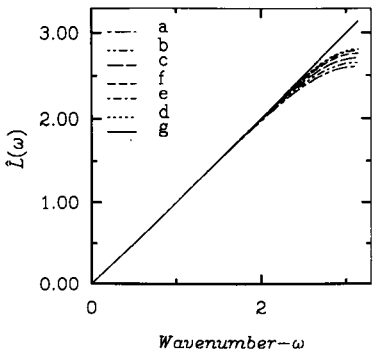

FIG. 9. Symbol for mid-cell discretizations of d/dx, $\hat{u}_0 = e^{-2\omega^2}$: (a) sixth-order tridiagonal scheme ($\beta = c = 0$); (b) second-order optimized tridiagonal scheme ($\beta = c = 0$); (c) eight-order tridiagonal scheme ($\beta = 0$); (d) second-order optimized tridiagonal scheme ($\beta = 0$); (e) tenth-order pentadiagonal; (f) optimized pentadiagonal; (g) exact symbol. Schemes were optimized for $\hat{u}_0 = e^{-2\omega^2}$.

6.1.3. *Mid-Cell Approximation of the First Derivative*

Table III lists the coefficients of schemes designed for various initial data. The coefficients of the schemes taken from [7] are listed below. Scheme (a) is given by

$$\alpha = \tfrac{9}{62}, \qquad \beta = 0, \qquad a = \tfrac{3}{8}(3 - 2\alpha),$$
$$b = \tfrac{1}{8}(22\alpha - 1), \qquad c = 0. \tag{6.19}$$

The coefficients of scheme (c) are

$$\alpha = \tfrac{75}{354}, \qquad \beta = 0, \qquad a = \frac{37950 - 39725\alpha}{31368},$$
$$b = \frac{65115\alpha - 3350}{20912}, \qquad c = \frac{25669\alpha - 6114}{62736}. \tag{6.20}$$

The coefficients of the tenth-order pentadiagonal scheme (f) are

$$\alpha = \tfrac{96850}{288529}, \qquad \beta = \tfrac{9675}{577058}, \qquad a = \tfrac{683425}{865587},$$
$$b = \tfrac{505175}{577058}, \qquad c = \tfrac{69049}{11731174}. \tag{6.21}$$

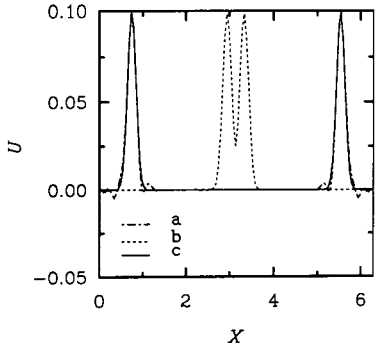

FIG. 10. Long-time integration for $u_{tt} = u_{xx}$, $\hat{u}_0 = e^{-2\omega^2}$, $\sigma = 0.8$: (a) tridiagonal mid-cell discretization scheme of d/dx, optimized to approximate d^2/dx^2, when $\hat{u}_0 = e^{-2\omega^2}$; (b) non-optimized tridiagonal scheme; (c) exact solution.

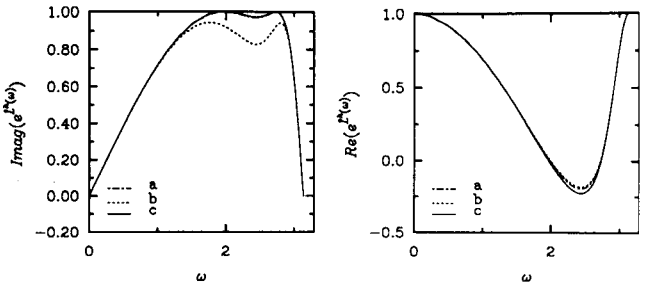

FIG. 11. Real and imaginary parts of approximations to $e^{L^h(\omega)}$, where $L^h(\omega)$ is the symbol of the tridiagonal scheme for d/dx, optimized for $\hat{u}_0 = e^{-2\omega^2}$ and $\sigma = 0.8$: (a) five-stage scheme optimized for the same σ; (b) fourth-order Runge–Kutta; (c) exact time integration.

The standard compact schemes give very good resolution in this form (see Fig. 9); thus, the improvement introduced by the optimized schemes is smaller. Optimizing the tridiagonal scheme yields a 6.5 times smaller error norm while optimizing the pentadiagonal scheme yields a 2.5 times smaller norm. In this case, the error norm of the optimized tridiagonal scheme is very close to that of the non-optimized pentadiagonal scheme.

An interesting option suggested by this approach was to optimize the d/dx operator in order to obtain the best approximation for d^2/dx^2, for given initial values. This has been done for the tridiagonal scheme which was used to integrate Eq. (6.17). It was compared, in Fig. 10, to the tridiagonal scheme from [7], where both are used to approximate the second derivative in solving the one-dimensional wawe equation in the system form (6.18) on a 128-point grid. Again, the optimized scheme gives significantly better approximation.

6.2. Approximate Time Integration

The constrained minimization (4.3)–(4.5) was solved by requiring that the solution will touch the stability constraint at one point while maintaining global stability and minimizing the functional. The point which gives the least error norm was found by exhaustive search. This straightforward approach yielded the local minima reported in this paper.

TABLE IV

Coefficients of Time Integration Scheme

$\hat{u}_0$	Third-order schemes designed for $\sigma = 0.9$ having $a_0 = 1$, $a_1 = 1$, $a_2 = \frac{1}{2}$		
$e^{-2\omega^2}$	$a_3 = 0.166281$,	$a_4 = 0.0397196$,	$a_5 = 0.0076705$
$e^{-3\omega^2}$	$a_3 = 0.166407$,	$a_4 = 0.0409525$,	$a_5 = 0.0074510$
$e^{-4\omega^2}$	$a_3 = 0.1664488$,	$a_4 = 0.04111513$,	$a_5 = 0.00739737$
$e^{-5\omega^2}$	$a_3 = 0.1664805$,	$a_4 = 0.04121264$,	$a_5 = 0.00736302$
$e^{-6\omega^2}$	$a_3 = 0.1665028$,	$a_4 = 0.04128218$,	$a_5 = 0.00733301$
$e^{-7\omega^2}$	$a_3 = 0.1665207$,	$a_4 = 0.04133150$,	$a_5 = 0.00731074$

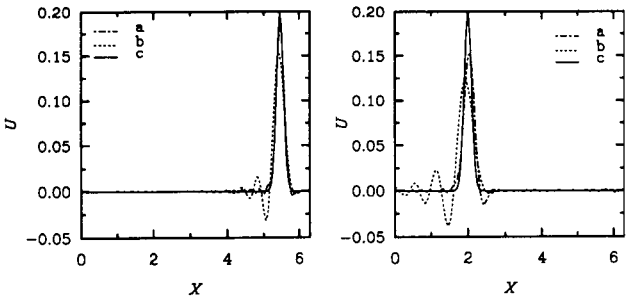

FIG. 12. Integration of $u_t = u_x$, $\hat{u}_0 = e^{-2\omega^2}$, $\sigma = 0.8$. The space derivative is computed using the tridiagonal compact scheme optimized for the same initial data and σ: (a) five-stage scheme optimized for this scheme and CFL; (b) fourth-order Runge–Kutta; (c) exact time integration.

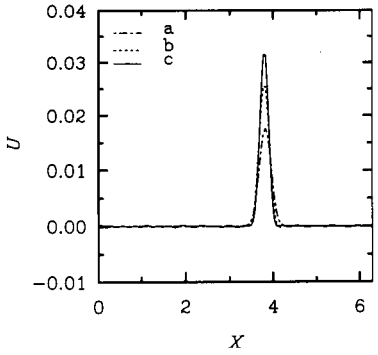

FIG. 14. Integration of $u_t = u_x + 0.5(1 + 0.6\sin(2\pi y))\,u_y$, $\sigma = 0.8$ using tridiagonal schemes. Initial solution was $\hat{u}_0 = e^{-(\omega_1^2 + 5\omega_2^2)}$ rotated at an angle of $\pi/4$. This data was approximated by unrotated gaussian $e^{-(3\omega_1^2 + 2\omega_2^2)}$: (a) optimized tridiagonal scheme and optimized marching scheme; (b) tridiagonal scheme integrated with fourth-order Runge–Kutta; (c) a fine grid solution (practically exact).

Somewhat better integration schemes might be achieved by using more advanced optimization techniques [8].

According to the general approach outlined in Section 5, one should choose an integration scheme which yields a truncation error of similar magnitude in time and space. Since the stability region for several fifth-order six-stage explicit Runge–Kutta schemes intersects the imaginary axis only in a small neighborhood of the origin [5, 6], disabling time marching with a large CFL, the optimized scheme was compared with the four-stage fourth-order Runge–Kutta. We preferred this five-stage scheme, which has an error norm about five times larger than the space discretization, to the seventh-order scheme, which yields an error norm about 11 times smaller than the space discretization, because of its lower computational cost.

The analysis performed in Section 2 suggests that the integration operator should be optimized with respect to the spatial discrete operator employed, i.e., to minimize $\|P(L^h(\omega h)\,\Delta t) - e^{L^h(\omega h)\,\Delta t}\|_{L_2}$. In the following examples L^h is the tridiagonal approximation for d/dx, when the initial data is $e^{-2\omega^2}$ and $\sigma = 0.8$. Table IV contains the coefficients

versus the four-stage fourth-order Runge–Kutta and the optimized scheme. The norm of the imaginary part of the error was reduced by a factor of 31 while its real part was reduced by a factor of merely 2.3.

Figures 12–13 shows the inegration of the advection equation with those schemes on a 128-point grid, demonstrating the superior efficiency and robustness of the proposed schemes. In Figure 12 one can see that the optimized scheme can be used for at least a four times longer integration time than the Runge–Kutta scheme applied to the tridiagonal scheme from [7]. The computed error norms suggest that the time marching error is dominant in all examples.

The two-dimensional example in Fig. 14 summarizes the approach suggested in this work. It compares the optimized tridiagonal scheme combined with the appropriate integration formula, to a fourth-order Runge–Kutta applied to a non-optimized tridiagonal discretization. Although the

of integration scheme for various initial data, when L^h is the tridiagonal scheme optimized for the same initial data and $\sigma = 0.8$.

Figure 11 plots the real and imaginary parts of $e^{L^h(\omega h)\,\Delta t}$

analysis in Section 2 applies only to constant coefficient problems, this example shows that it holds, heuristically, to variable coefficient equations, as well. The initial data for this problem was obtained in a similar manner to that in Fig. 4. However, instead of comparing the solutions computed on the 128×128 grid to the exact solution, they are compared to the solution on a 256×256 grid which was

while formal accuracy of a discrete approximation describes the error in the smoothest components realizable on a grid and predicts the global error reduction as the mesh is refined, it is inadequate to capture the error on a given finite grid with arbitrary initial data. Therefore, although asymptotically, higher order schemes are superior to lower order ones, on finite computational grids they might not be. This observation led to deriving improved error estimates on the L_2 norm of the global error, when the initial data and grid resolution are given. Thus, the relative amplitude of the physical frequencies representable on the grid can be taken into account. This property, as well as the fact that the error estimate separately bounds the spatial and temporal errors, enabled us to devise a general approach for designing finite difference schemes with optimal performance for the given grid and data.

These error bounds were used to design compact finite difference schemes for derivative evaluation. The resulting schemes combined adaptivity to the specific initial data by the nature of their design and robustness to perturbations in the initial data. The improved resolution had been demonstrated for several problems and was compared to previously known similar schemes, schemes with spectral-like resolution. It was shown that the optimized schemes can be efficiently used for an integration time that was an order of magnitude larger than the spectral-like scheme with similar computational complexity.

A similar approach was used to design improved and robust integration schemes, taking into account the spatial discretization, as well as the initial data and grid resolution.

The robustness of the resulting schemes to perturbation in the initial data enabled us to extend them to more general differential operators in a simple straightforward way by giving up on some of the obtainable accuracy. Other extensions which do not compromise on efficiency should be investigated and trade-off between accuracy, robustness, and ease of use, for these generalizations should be better understood. This robustness enables us to obtain improved resolution even when only approximate knowledge of the energy distribution is available, e.g., a probability density function of the wavenumber amplitude (as in turbulence).

The approach suggested in this paper for optimizing discrete operators can be similarly applied to higher derivatives. Its applicability to more general and complex operators should be further investigated. The use of these ideas to design boundary conditions will be presented elsewhere.

REFERENCES

1. Y. Adam, *J. Comput. Phys.* **24**, 10 (1977).
2. R. S. Hirsh, *J. Comput. Phys.* **19**, 90 (1975).
3. Z. Kopal, *Numerical Analysis*, 2nd ed. (Wiley, New York, 1961), p. 552.
4. D. Kishoni and S. Ta'asan, "Improved Finite Difference Method for Long Distance Propagation of Waves," in *Review of Progress in Quantitative NDE*, Vol. 12A, edited by D. O. Thompson and Chimenti (Plenum, New York, 1993), p. 139.
5. L. Lapidus and J. Seinfeld, *Numerical Solution of Ordinary Differential Equations* (Academic Press, New York, 1971), p. 298.
6. J. D. Lawson, *SIAM J. Numer. Anal.* **3**, 593 (1966).
7. S. K. Lele, *J. Comput. Phys.* **103**, 16 (1992).
8. S. McCormick, *Nonlinear Programming: Theory, Algorithms and Applications* (Wiley, New York, 1983), p. 444.
9. W. L. Miranker, *Numer. Math.* **17**, 124 (1971).
10. B. Swartz and B. Wendroff, *SIAM J. Numer. Anal.* **11**, 979 (1974).
11. R. Vichenevetsy, *Math. Comput. Simulation* **25**, 170 (1979).
12. R. Vichenevetsy and J. B. Bowles, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations* (SIAM Philadelphia, 1982), p. 140.