1990

A Study of Numerical Methods for Hyperbolic Conservation Laws with Stiff Source Terms

R. J. LeVeque
University of Washington

Helen C. Yee
NASA Ames Research Center, yee@nas.nasa.gov

JOURNAL OF COMPUTATIONAL PHYSICS 86, 187–210 (1990)

A Study of Numerical Methods for Hyperbolic Conservation Laws with Stiff Source Terms

R. J. LEVEQUE*

Department of Mathematics, University of Washington, Seattle, Washington 98195

AND

H. C. YEE

Research Scientist, Computational Fluid Dynamics Branch, NASA Ames Research Center, Moffett Field, California 94035.

Received June 6, 1988; revised December 20, 1988

The proper modeling of nonequilibrium gas dynamics is required in certain regimes of hypersonic flow. For inviscid flow this gives a system of conservation laws coupled with source terms representing the chemistry. Often a wide range of time scales is present in the problem, leading to numerical difficulties as in stiff systems of ordinary differential equations. Stability can be achieved by using implicit methods, but other numerical difficulties are observed. The behavior of typical numerical methods on a model advection equation with a parameter-dependent source term is studied. Two approaches to incorporate the source terms are utilized: MacCormack type predictor-corrector methods with flux limiters and splitting methods in which the fluid dynamics and chemistry are handled in separate steps. Comparisons over a wide range of parameter values are made. On the whole, the splitting methods perform somewhat better. In the stiff case, a numerical phenomenon of incorrect propagation speeds of discontinuities is observed and explained. Similar behavior was reported by Colella, Majda, and Roytburd (*SIAM J. Sci. Stat. Comput.* **7**, 1059 (1986)) on a model combustion problem. Using the model scalar equation, we show that this is due to the introduction of nonequilibrium values through numerical dissipation in the advection step.

1. INTRODUCTION

In nonequilibrium gas dynamics, chemical reactions between the constituent gases must be modeled along with the fluid dynamics. This added complexity is required in certain regimes of hypersonic aerodynamic modeling, for example, in the bow shock of hypersonic vehicles.

* Supported in part by NASA-Ames University Consortium NCA2-185 and NSF Grant DMS-8657319.

0021-9991/90 $3.00

Coupled systems of this form also arise in combustion problems. In particular the modeling of scramjet engines that might be used in hypersonic vehicles requires the numerical simulation of supersonic combustion.

Restricting our attention to inviscid flow, we have essentially the Euler equations of gas dynamics, coupled with source terms representing the chemistry. In two space dimensions these equations take the form

$$u_t + f(u)_x + g(u)_y = \psi(u) \tag{1}$$

where u is the vector of dependent variables including momentum, energy, and densities or concentrations for each species in the reacting mixture. The flux functions f and g describe the fluid dynamics as in the Euler equations while the source term $\psi(u)$ arises from the chemistry of the reacting species.

A variety of such systems are possible, depending on the level of detail of chemical modeling included. Examples and more discussions of these equations may be found in various references, e.g., [2, 13, 18].

When we attempt to solve the reacting flow equations numerically, new difficulties arise that are absent in non-reacting flows. Aside from the increase in the number of equations, the main difficulties stem from the possible "stiffness" of the reaction terms. Although many excellent numerical methods are now available for the nonreacting case ($\psi = 0$) which give high resolution and sharp shocks, it is not clear to what extent these methods can be used in the reacting case.

The kinetics equations often include reactions with widely varying time scales. Moreover, many of the chemical time scales may be orders of magnitude faster than the fluid dynamical time scales. This can lead to problems of stiffness akin to the classical stiffness problems of ordinary differential equations (ODEs). Stiff ODEs arise, for example, in modeling chemical kinetics in a uniform stirred reactor where the fluid dynamics terms drop out. The numerical difficulty with such problems is that some time scales will typically be much faster than the scale on which the solution is evolving and on which one would like to compute. This occurs when the fast reactions are in near-equilibrium during most of the computation. With many numerical methods, including all explicit methods, taking a time step appropriate for the slower scale of interest can result in violent numerical instability caused by the faster scales.

Of course, if the fast reactions are always in equilibrium it may be possible to eliminate these reactions from the system. In the extreme case one obtains equilibrium gas dynamics in which the kinetics are not explicitly modeled but the equation of state varies with the mixture. In many problems, however, nonequilibrium effects play an important role and must be included.

In practice, it is common to take time steps which resolve the fastest scales. Boris and Oran [2] suggest as a rule of thumb that time steps must be restricted so that the energy release from chemical reactions does not change the total energy in any cell by more than 10–20%. It is more desirable, however, to develop robust methods that can allow larger time steps. This will naturally incur some loss of

resolution—reaction fronts will lose their structure and approach discontinuities, for example. What we should demand is that the *correct* discontinuities are obtained. They may be smeared out due to numerical diffusion, but should represent the correct jumps in the correct locations. This goal can be achieved for the nonreacting case by using "conservative" numerical methods [12] (See Section 4).

In this paper we investigate the extent to which this goal can be achieved for Eq. (1) using various popular finite difference techniques. In particular, we introduce and study a simple one-dimensional scalar model equation which illuminates some of the difficulties sure to be encountered also in solving more realistic equations. We investigate the following questions: (i) Can we develop stable methods? (ii) Can we obtain "high resolution" results, with sharp discontinuities and second order accuracy in smooth regions, and (iii) Do we obtain the correct jumps in the correct locations?

Numerical stability is typically not a problem. A variety of excellent implicit methods have been developed for solving stiff systems of ODEs, and many of the same techniques can be applied to the stiff source terms in (1) to obtain stable methods for solving this system.

The second question is investigated in Sections 2 and 3, where we will see that with some care, second-order accuracy and reasonably sharp discontinuities can be obtained.

The third question is the most interesting. For stiff reactions it is possible to obtain stable solutions that look reasonable and yet are completely wrong, because the discontinuities are in the wrong locations. Stiff reaction waves move at nonphysical wave speeds, often at the rate of one grid cell per time step regardless of their proper speed.

This phenomenon has also been observed by Colella, Majda, and Roytburd [5] who made a similar study of the limiting behavior with increasing stiffness for various model systems. In particular, they look at the Euler equations coupled with a single chemistry variable representing the mass fraction of unburnt gas in a detonation wave. These waves have the structure of a fluid dynamic shock that raises the pressure to some peak value, followed immediately by a reaction zone that brings the pressure back down to a new equilibrium value. On coarse grids it is not possible to resolve this combustion spike and the best one can hope for a single discontinuity linking the two equilibrium values and moving at the correct speed.

Colella, Majda, and Roytburd apply Godunov's method and a high resolution extension of Godunov's method [6] to this problem. The source terms are handled by splitting and solving the resulting ODEs exactly, so that stability is not a problem. However, they observe that on coarse grids the numerical solution is qualitatively incorrect. The computed solution consists of a weak detonation wave, in which all the chemical energy is released, followed by a fluid dynamic shock traveling more slowly. The reaction wave always travels at the speed of one mesh cell per time step, which is totally nonphysical.

A simpler model system is also studied in [5] and is shown to exhibit similar

behavior numerically. This system is essentially Burgers' equation coupled with a single reaction equation. By studying this system and its numerical solution theoretically, progress has been made in understanding the structure of numerical solutions of the reacting Euler equations.

However, the essential numerical difficulty can be identified and studied most easily by looking at even simpler equations. This same numerical behavior of discontinuities traveling at incorrect speeds can be observed in scalar problems. We have found it illuminating to study the model problem

$$u_t + u_x = \psi(u) \tag{2}$$

with

$$\psi(u) = -\mu u(u-1)(u-\tfrac{1}{2}). \tag{3}$$

This is the linear advection equation with a source term that is stiff for large μ. Along the characteristic $x = x_0 + t$, the solution to (2) evolves according to the ODE

$$\frac{d}{dt} u(x_0 + t, t) = \psi(u(x_0 + t, t)) \tag{4}$$

with initial data $u(x_0, 0)$. This equation has stable equilibria at $u = 0$ and $u = 1$ and an unstable equilibrium at $u = \frac{1}{2}$. For large μ and arbitrary initial data the ODE solution consists of a rapid transient with u approaching 0 (if $u(x_0, 0) < \frac{1}{2}$) or 1 (if $u(x_0, 0) > \frac{1}{2}$).

Consequently, the solution $u(x, t)$ to (2) with initial data $u(x, 0)$ rapidly approaches a piecewise constant traveling wave solution $w(x - t)$, where

$$w(x) = \begin{cases} 0 & \text{if } u(x,0) < \frac{1}{2} \\ \frac{1}{2} & \text{if } u(x,0) = \frac{1}{2} \\ 1 & \text{if } u(x,0) > \frac{1}{2}. \end{cases}$$

In particular, the solution with piecewise constant initial data

$$u(x, 0) = \begin{cases} 1 & \text{if } x < x_0 \\ 0 & \text{if } x > x_0 \end{cases} \tag{5}$$

is simply $u(x, t) = u(x - t, 0)$. In this case the ODE solution is in equilibrium on each side of the discontinuity, which theoretically behaves as it would if the source term were not present and we simply solved the linear advection equation $u_t + u_x = 0$.

This linear discontinuity could easily be converted to a shock by replacing u_x in (2) by $f(u)_x$ for some nonlinear flux function f. However, the numerical behavior is qualitatively the same in either case and nonlinearity of the flux is not the source of the difficulties of primary interest here.

Other functions $\psi(u)$ could also be considered. A model corresponding more closely to the "ignition temperature kinetics" of [5] is obtained by using

$$\psi(u) = \begin{cases} -\mu(u-1) & \text{if } u > \frac{1}{2} \\ 0 & \text{if } u \leqslant \frac{1}{2}. \end{cases}$$

This gives numerical behavior similar to what is reported here for (3) and will not be considered further.

All of the methods studied in this paper give propagation of the step function (5) at incorrect speeds when the source term is sufficiently stiff, i.e., when μ is sufficiently large. We identify the quantity $k\mu$, where k is the time step, as the critical parameter affecting the propagation speed. Unless $k\mu$ is much smaller than 1, numerical difficulties are observed.

Note that $\tau \equiv 1/\mu$ is the relaxation time scale for the source term. Typically $k = O(h)$, where h is the spatial mesh width, and therefore k is the appropriate time scale for advection on the grid. Consequently, we can view $k\mu = k/\tau$, the ratio of the advection time scale to the relaxation scale, as a sort of "cell Damköhler number."

The numerical phenomenon of incorrect propagation speeds is studied in Section 4. A simple explanation is found for the scalar model that also carries over to systems of equations such as the model system studied in [5].

The basic explanation is that numerical advection of the discontinuity gives a smeared representation, which includes intermediate states $0 < u < 1$ that are not in equilibrium. When $k\mu$ is large, the source term restores near equilibrium in each time step, shifting the value in each cell towards 0 or 1 and consequently shifting the discontinuity to a cell boundary. It is thus not surprising that nonphysical propagation speeds of one cell per time step can be observed for large $k\mu$.

Clearly this scalar model is inadequate as a full test of any numerical method. However, it does model one essential difficulty encountered in reacting flow problems and is sufficient to point out difficulties that may arise also on more complicated systems of equations. Moreover, due to the simplicity of this equation, numerical problems that do arise can be easily understood and their source identified, yielding insight that may be valuable in developing better methods.

We will discuss two different approaches to constructing numerical methods for (1) and compare their numerical behavior on the model problem (2) for various values of μ. For simplicity we only discuss the one-dimensional version of (1), in which $g \equiv 0$, but in each case two-dimensional analogues are easily defined. Forward and backward differences of g-fluxes can be included in the predictor-corrector methods along with differences of the f-fluxes.

The first method we consider is based on MacCormack's predictor-corrector method for conservation laws [14]. This second-order accurate method can be modified to include the source terms, which appear in each step of the method. Stiff source terms are usually handled in a semi-implicit manner to obtain stability with reasonable time steps. We have found, however, that one very natural and commonly used modification does not preserve the second-order accuracy of the semi-

implicit method on time-dependent problems, although steady states are accurately computed. Based on a truncation error analysis, we show how this can be easily rectified in Section 2.

In order to avoid oscillations near discontinuities, MacCormack's method can be modified by adding a flux-correction step motivated by the theory of TVD methods [20, 21]. We will compare two different forms of this correction.

The second approach we study is the splitting method, in which one alternates between solving the conservation laws (with no source terms) in one step and the stiff systems of ODEs modeling the chemistry (with no fluid motion) in the second step. This approach has certain advantages, in that high quality numerical methods exist for each of the subproblems. Combining these via splitting can yield stable, second-order accurate methods for the full problem. This is demonstrated in Section 3.

Numerical tests on the model problem (2) reveal that methods can be devised by either of these approaches that will be stable and second-order accurate as the mesh is refined. However, for realistic choices of grid and time step, stiff reaction waves will have the nonphysical behavior described above. This is investigated in Section 4.

2. Extensions of MacCormack's Method.

MacCormack's method for a system of conservation laws is a two step predictor-corrector method in which backward differences are used in the first step and forward differences in the second step (or vice versa). The method is easily modified to include source terms in an explicit manner and maintain second-order accuracy [19]. For the one-dimensional system

$$u_t + f(u)_x = \psi(u) \tag{6}$$

this explicit method takes the form

$$\begin{aligned} \Delta U_j^{(1)} &= -\frac{k}{h}(f(U_j^n) - fU_{j-1}^n)) + k\psi(U_j^n) \\ U_j^{(1)} &= U_j^n + \Delta U_j^{(1)} \\ \Delta U_j^{(2)} &= -\frac{k}{h}(f(U_{j+1}^{(1)}) - f(U_j^{(1)})) + k\psi(U_j^{(1)}) \\ U_j^{n+1} &= U_j^n + \tfrac{1}{2}(\Delta U_j^{(1)} + \Delta U_j^{(2)}). \end{aligned} \tag{7}$$

Here h is the grid spacing in x and k is the time step. Computing the truncation error for this method shows that it is second-order accurate in both space and time, as the grid is refined with k/h fixed.

Note that if we set $f(u) \equiv 0$, so that (6) reduces to a system of ODEs, then (7) reduces to the standard two-stage Runge–Kutta method. Clearly this explicit

method will be inadequate if the system is stiff, in that the time step k required for stability will be much smaller than desirable for accuracy.

It is natural to try to improve the stability of the method by making it semi-implicit, so that the source terms are handled implicitly while the flux terms are still explicit. In order to avoid solving nonlinear systems of equations in each step, a linearly implicit method is frequently used. Methods of this form have been used by many workers (e.g., Bussing and Murman [3], Drummond, Rogers, and Hussaini [7], and Yee and Shinn [21]). This method takes the form

$$\begin{aligned}
\left[I-\frac{1}{2}k\psi'(U_j^n)\right]\Delta U_j^{(1)} &= -\frac{k}{h}(f(U_j^n)-f(U_{j-1}^n))+k\psi(U_j^n)\\
U_j^{(1)} &= U_j^n+\Delta U_j^{(1)}\\
\left[I-\frac{1}{2}k\psi'(\hat{U}_j)\right]\Delta U_j^{(2)} &= -\frac{k}{h}(f(U_{j+1}^{(1)})-f(U_j^{(1)}))+k\psi(\bar{U}_j)\\
U_j^{n+1} &= U_j^n+\tfrac{1}{2}(\Delta U_j^{(1)}+\Delta U_j^{(2)}).
\end{aligned} \tag{8}$$

The values of $\hat{U}_j$ and $\bar{U}_j$ are still unspecified. In most of the papers referenced above, $\hat{U}_j=\bar{U}_j\equiv U_j^{(1)}$ is used, as motivated by (7). However, another possibility is to use $\hat{U}=\bar{U}_j=U_j^n$. A truncation error analysis for the method shows that this latter choice is in fact preferable, since it gives a method that is second-order accurate in both space and time. The traditional choice is second order in space but only first order in time. Thus it gives second-order steady state solutions but would only be first order in time for unsteady problems.

Actually, in order to achieve second-order accuracy overall, it is only necessary to use $\bar{U}_j=U_j^n$. The choice of $\hat{U}_j$ is immaterial so long as $\psi'(\hat{U}_j)=\psi'(U_j^n)+O(k)$. In particular, $\hat{U}_j=U_j^n$ or $\hat{U}_j=U_j^{(1)}$ are both allowed. In view of this it seem most efficient to take $\hat{U}_j=U_j^n$ in practice, since then the matrix $[I-\frac{1}{2}k\psi'(U_j^n)]$ need only be computed and factored once and the resulting factorization used in each fractional step.

These statements are justified in the Appendix, where we study the truncation error for this class of methods. These results have also been verified numerically for smooth initial data.

Flux Limiters

The method (8) is spatially centered and hence will typically give oscillatory behavior on problems involving steep gradients. To minimize this problem, one can introduce flux limiter terms into the method, as motivated by the theory of TVD methods. This is described in more detail in [20, 21] and so we will be brief in our description here.

Let $A_{j+1/2}$ represent an average of $f'(u)$ between U_j^n and U_{j+1}^n, e.g., the Roe approximation [15]. Let $R_{j+1/2}$ be the matrix of right eigenvectors of $A_{j+1/2}$ and $\lambda_{j+1/2}$ the vector of corresponding eigenvalues. Also, let

$$\alpha_{j+1/2}=R_{j+1/2}^{-1}(U_{j+1}^n-U_j^n).$$

The components of this vector give the coefficients of the decomposition of the jump $U^n_{j+1} - U^n_j$ into eigenvectors of $A_{j+1/2}$. Corresponding to the lth component of this vector, $\alpha^l_{j+1/2}$, define a limited version by

$$\hat{Q}^l_{j+1/2} = \text{minmod}\,(\alpha^l_{j-1/2}, \alpha^l_{j+1/2}, \alpha^l_{j+3/2}).$$

where the minmod function is defined by

$$\text{minmod}(a, b, c) = \begin{cases} s \min(|a|, |b|, |c|) & \text{if} \quad s \equiv \text{sgn}(a) = \text{sgn}(b) = \text{sgn}(c) \\ 0 & \text{otherwise.} \end{cases}$$

Other versions of this limiter can also be used (see [20, 21]) but we will restrict our attention to this one. Finally, we define

$$v^l_{j+1/2} = \frac{k}{h} \lambda^l_{j+1/2}$$

and the smoothed absolute value

$$q(z) = \begin{cases} |z| & \text{if} \quad |z| \geqslant \varepsilon \\ (z^2 + \varepsilon^2)/2\varepsilon & \text{if} \quad |z| < \varepsilon \end{cases}$$

for some positive parameter ε. In our present example this plays no role, and in fact $q(z)$ simply reduces to the absolute value in the context where we use it.

We now replace the last line of (8) by

$$U^{(2)}_j = U^n_j + \tfrac{1}{2}(\Delta U^{(1)}_j + \Delta U^{(2)}_j)$$

and then set

$$U^{n+1}_j = U^{(2)}_j + [R_{j+1/2}\phi_{j+1/2} - R_{j-1/2}\phi_{j-1/2}], \tag{9}$$

where $\phi_{j+1/2}$ has components

$$\phi^l_{j+1/2} = \tfrac{1}{2}[q(v^l_{j+1/2}) - (v^l_{j+1/2})^2](\alpha^l_{j+1/2} - \hat{Q}^l_{j+1/2}).$$

Note that for smooth solutions, $\alpha^l_{j+1/2} = O(h)$ and $\alpha^l_{j+1/2} - \hat{Q}^l_{j+1/2} = O(h^2)$ and is also smooth. The perturbation to $U^{(2)}_j$ in (9) will then be $O(h^3)$, leaving the method second-order accurate. Near discontinuities, however, this modification serves to introduce an upwind bias, dropping the method to first-order accuracy but reducing oscillations. For a scalar problem with no source terms, the resulting method is TVD. When source terms are present, the true solution may no longer be TVD, and it is not clear what the correct theoretical criterion should be.

The above flux correction procedure can be modified by basing the correction terms on $U^{(2)}$ rather then U^n. For example, in place of $\alpha_{j+1/2}$ we would use

$$\alpha^{(2)}_{j+1/2} = (R^{(2)}_{j+1/2})^{-1}\,(U^{(2)}_{j+1} - U^{(2)}_j).$$

This approach is advocated in [21] and has the advantage that U^n need not be saved for this correction. However, this is no real advantage if we intend to save U^n for the second stage of MacCormack's method, as we have argued that one should do for time dependent problems. When based on $U^{(2)}$ rather than U^n, the method is no longer strictly TVD on scalar problems without source terms. This might argue for the use of U^n. When source terms are present it is not clear which approach is superior. Good results for a steady state reacting flow problem were achieved in [21] with corrections based on $U^{(2)}$. This approach has also been successfully used for unsteady problems in the nonreacting case [1, 16, 23].

The experiments below indicate that for the model problem (2), limiting based on U^n is preferable for small values of $k\mu$ but that limiting based on $U^{(2)}$ may be more robust for larger values of $k\mu$.

Numerical Results on Discontinuous Data.

The method described above is second order accurate on smooth solutions, but eventually fronts sharpen and become nearly discontinuous. To investigate the ability of this method to deal with propagating discontinuities, we consider the following initial data for Eq. (2):

$$u(x, 0) = \begin{cases} 1 & \text{if } x \leqslant 0.3 \\ 0 & \text{if } x > 0.3. \end{cases} \tag{10}$$

We now use $\hat{U}_j = \bar{U}_j = U_j^n$ exclusively and compare the effects of the different limiters (no limiter, limiting based on U^n, limiting based on $U^{(2)}$). We take $h = 0.02$, $k/h = 0.75$, and various values of μ. Note that due to scaling properties of the equation and method, results at time T with a particular value of μ can equally well be regarded as results with μ replaced by μ/β, for arbitrary β, at time βT with time step βk and grid spacing βh (with x rescaled so that $[0, 1]$ becomes $[0, \beta]$). Indeed, the critical dimensionless parameters that determine the performance of the method are the mesh ratio $\lambda = k/h$ and the product $k\mu$ of the time step and reaction rate. The value $k\mu$ determines the stiffness of the system. When $k\mu$ is large, relaxation to equilibrium occurs on a time scale that cannot be temporally resolved on the grid.

Figure 1 shows computed results at $t = 0.3$ for $\mu = 1$, 10, 100, and 1000 ($k\mu = 0.015$, 0.15, 1.5, and 15). Each row of figures illustrates a different value of $k\mu$. The three figures in each row correspond to different choices of limiter. We see several interesting things from these graphs:

- For small $k\mu$ (0.015) oscillations are visible if no limiter is used and to a lesser extent if the limiter is based on $U^{(2)}$, while limiting on U^n gives monotone results. This agrees with what is expected for the pure convection case ($k\mu = 0$).

- For larger $k\mu$ (0.15–1.5), there is a slight overshoot in all cases, of similar magnitude regardless of the limiter. Note that for the case of no limiter there is less oscillation here than with smaller $k\mu$, due to the stabilizing effect of the source terms that tend to restore μ towards 1.

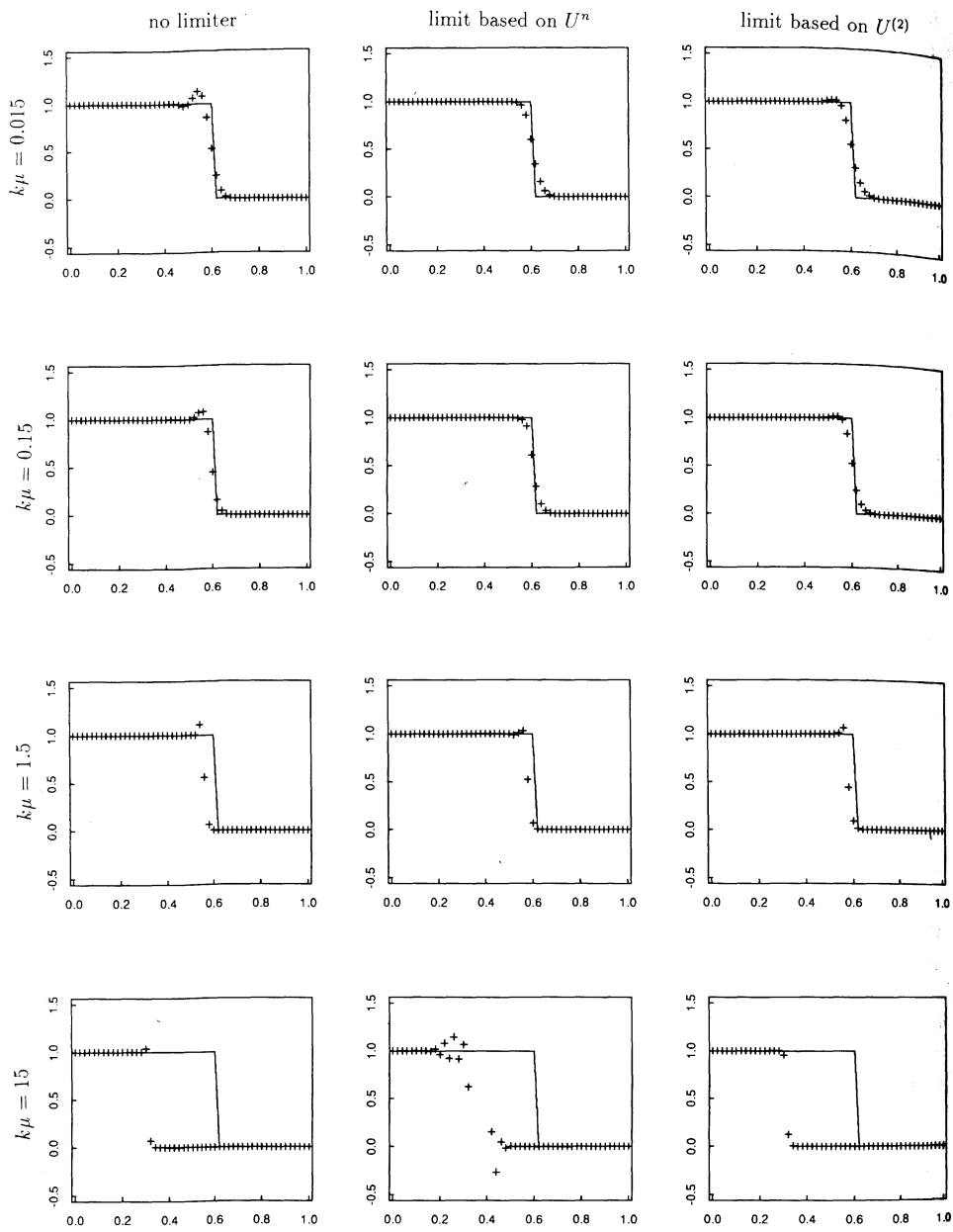

FIG. 1. Numerical results using extended MacCormack method with discontinuous initial data:—, true solution; +, computed solution.

• For large $k\mu$ (15), limiting on U^n appears to be unstable (there are large scale oscillations near $x = 0.3$ not visible in the figure) whereas limiting on $U^{(2)}$ or using no limiter gives stable results. In each case, however, the solution is completely wrong! The discontinuity has remained at its initial location $x = 0.3$ rather than propagating.

Note that for $k\mu = 1.5$ there is also some discrepancy in the location of the discontinuity. The speed of propagation is slightly too small. For intermediate values of $k\mu$ it is possible to obtain results with the discontinuity anywhere between 0.3 and 0.6. This phenomenon of wrong propagation speeds for large $k\mu$ will be discussed in more detail in Section 4.

3. Splitting Methods.

The semi-implicit predictor-corrector method (8) attempts to handle the fluid dynamics and chemistry simultaneously. An alternative approach is to employ a time-splitting in which one alternates between solving a system of conservation laws, with no source terms, and a system of ordinary differential equations modeling the chemistry. In the simplest case this splitting takes the form

$$U^{n+1} = S_\psi(k)\, S_f(k) U^n. \tag{11}$$

Here $S_f(k)$ represents the numerical solution operator for the system of conservation laws

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

over a time step of length k, and $S_\psi(k)$ is the numerical solution operator for the ODE system

$$u_t = \psi(u).$$

To maintain second-order accuracy, the Strang splitting [17] can be used, in which the solution U^{n+1} is computed from U^n by

$$U^{n+1} = S_\psi(k/2)\, S_f(k)\, S_\psi(k/2)\, U^n. \tag{12}$$

Naturally, when several time steps are taken the adjacent operators $S_\psi(k/2)$ can be combined to give

$$U^n = S_\psi(k/2)\, S_f(k) [S_\psi(k)\, S_f(k)]^{n-1}\, S_\psi(k/2)\, U^0.$$

In this form the method is nearly as efficient as (11).

The splitting approach has also frequently been used to solve reacting flow problems [2, 4, 5]. At first glance it may appear to be less satisfactory than an unsplit method such as (8), since in reality the fluid dynamics and chemistry are

strongly coupled and cannot be separated. However, the fact that the splitting (12) is second-order accurate suggests that the interaction of different effects is adequately modeled by a split method, at least for smooth solutions. Moreover, there are distinct advantages to the splitting from the standpoint of algorithm design. High quality numerical methods have been developed both for systems of conservation laws and for stiff systems of ordinary differential equations. By decomposing the problem into subproblems of these types, it is possible to take advantage of these methods directly. To some extent the mathematical theory that supports them can also be carried over. By alternating between using a high resolution method for the conservation law and a stable stiff solver for the system of ODEs, one can easily derive a method with excellent prospects of stability on the full problem. By constrast, attempting to devise a good hybrid method handling both effects simultaneously with good accuracy and stability properties can be difficult, as has been seen in the previous section. (But in the stiff case, we will still see the problem of incorrect wave speeds with the splitting method.)

A split version of the method studied in Section 2 might take the form

$$
\begin{aligned}
S_\psi(k/2): \quad & [I - \tfrac{1}{4}k\psi'(U_j^n)]\,\Delta U_j^* = \tfrac{1}{2}k\psi(U_j^n) \\
& U_j^* = U_j^n + \Delta U_j^* \\
S_f(k): \quad & \Delta U_j^{(1)} = -\frac{k}{h}(f(U_j^*) - f(U_{j-1}^*)) \\
& U_j^{(1)} = U_j^* + \Delta U_j^{(1)} \\
& \Delta U_j^{(2)} = -\frac{k}{h}(f(U_{j+1}^{(1)}) - f(U_j^{(1)})) \\
& U_j^{(2)} = U_j^* + \tfrac{1}{2}(\Delta U_j^{(1)} + \Delta U_j^{(2)}) \\
& U_j^{**} = U_j^{(2)} + (R^*_{j+1/2}\phi^*_{j+1/2} - R^*_{j-1/2}\phi^*_{j-1/2}) \\
S_\psi(k/2): \quad & [I - \tfrac{1}{4}k\psi'(U_j^{**})]\,\Delta U_j^{**} = \tfrac{1}{2}k\psi(U_j^{**}) \\
& U_j^{n+1} = U_j^{**} + \Delta U_j^{**}.
\end{aligned}
\tag{13}
$$

Here ϕ^* involves limited fluxes as before, based on U^*. Alternatively, we can compute the limited value U_j^{**} based on $U^{(2)}$ and replace R^* and ϕ^* by $R^{(2)}$ and $\phi^{(2)}$, respectively.

Each of these methods could be replaced by other well-known methods for the respective problems. For example, any implicit stiff solver, such as the trapezoidal method, could be used for S_ψ and any of a wide variety of high resolution methods used for S_f. We consider the present form first as the logical choice for comparison with the previous results. The ODE method used in (13) for S_ψ will be referred to as the "linearized implicit method."

Figure 2 shows the same set of experiments as in Fig. 1, now with the splitting method. We observe that

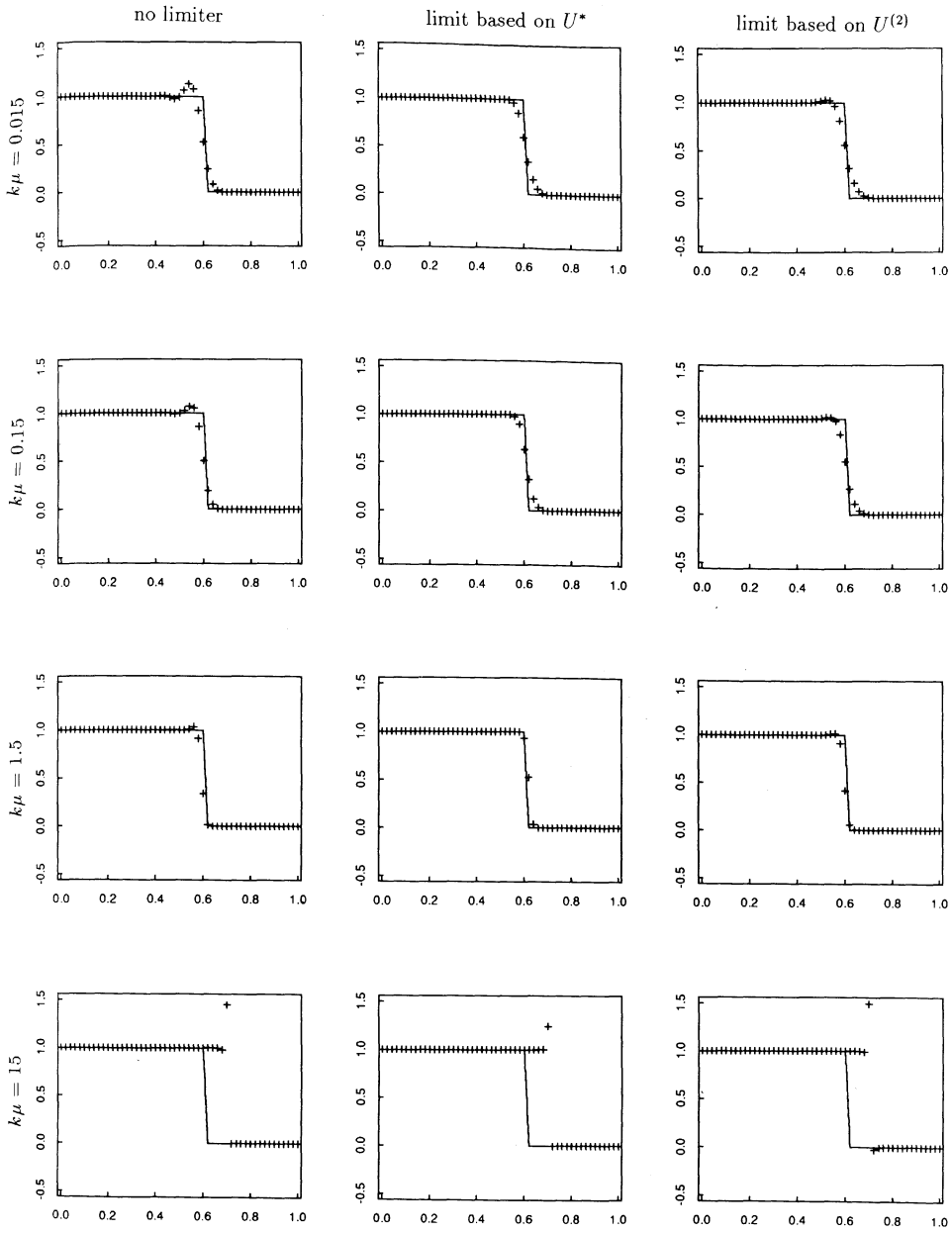

FIG. 2. Numerical results using splitting method with discontinuous initial data: —, true solution; +, computed solution.

- For small $k\mu$ either choice of limiter (based on $U^{(2)}$ or U^*) works well, and good results are obtained.
- For $k\mu = 15$ the discontinuity again moves at the wrong speed, now too fast. In fact, it has moved to $x = 0.7$ and so is moving at speed 4/3 rather than 1. Since $k/h = 3/4$, this indicates that the wave is moving at the speed of one mesh cell per time step.
- A large overshoot occurs in one mesh cell behind the discontinuity for $k\mu = 15$, regardless of the limiter used.

With regard to this last observation, it appears that the overshoot must originate within the ODE-solving step. The flux-limiter method is applied only to the homogeneous conservation law and should give no overshoots, at least in the case where we limit based on U^*. In other words, $S_f(k)$ keeps monotone data monotone and therefore the lack of monotoniciy must be generated by $S_\psi(k)$. Note that this solution operator works pointwise (for example, U_j^* is a function only of U_j^n, independent of U_i^n for $i \neq j$), and so is oblivious to the gradient in u. What it does see, however, is a nonequilibrium value of u near the discontinuity. The linearized ODE method used in (13) is stable but converges in an oscillatory manner to the steady state of a stiff equation and we are seeing this here. In ODE terminology, the S_ψ is not an L-stable method (see, e.g., [11]).

These overshoots can be avoided by switching to a different ODE method. For example, if we leave $S_f(k)$ unchanged but change $S_\psi(k/2)$ to the trapezoidal method, then these overshoots disappear for this value of $k\mu$ (Fig. 3), but note that the propagation speed is still wrong. With the trapezoidal method we compute, for example, U_j^* from U_j^n by solving the nonlinear equation

$$U_j^* = U_j^n + \tfrac{1}{4}k(\psi(U_j^n) + \psi(U_j^*)).$$

Although we obtain monotone profiles in Fig. 3, the trapezoidal method also experiences overshoots if we go to still larger values of $k\mu$. The use of an L-stable method such as the backward Euler method might eliminate this problem more generally, but backward Euler is only first-order accurate. One might consider the use of higher order BDF methods (the "Backward Differentiation Methods," also called "stiffly stable methods" in [8]), but the second-order BDF method is already

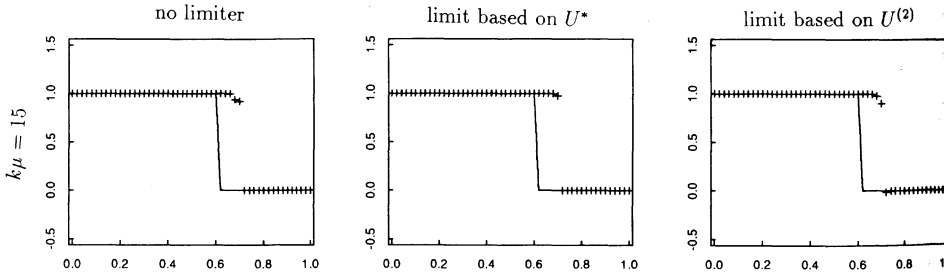

FIG. 3. Results for $k\mu = 15$ when the trapezoidal method is used: —, true solution; +, computed solution.

a two step method and in the present context we appear to require a one step method, because of the nature of the splitting method. Implicit Runge–Kutta methods are a possibility. For reaction equations a special asymptotic method has been developed by Young and Boris [22] (see also [2]) which may avoid this problem. Another possibility is to use *several* steps of an ODE solver for $S_\psi(k/2)$, i.e., subdivide the time interval to a point where we can more adequately resolve the transient approach to equilibrium. It appears that this fails to achieve our goal of using time step large relative to the fast time scales, but note that we would need to do this refinement in time only in regions where nonequilibrium conditions hold. At grid points where u starts out close to equilibrium (e.g., those for which $k\psi(u)$ is small, presumably most grid points), a single step of the linearized implicit method used in (13) is adequate to maintain stability.

This is another advantage of the splitting method—since the ODE solver is decoupled from the fluid solver and is applied at each grid point independently, it is easy to change the ODE solver or even to use different solvers at different points depending on the character of the flow. This approach is also advocated by Young and Boris [22], who suggest using their asymptotic integration method at stiff points and explicit Euler elsewhere.

We stress again, however, that improvements to the ODE solver cannot cure the problem of incorrect propagation speeds. In the next section we will investigate the source of this difficulty.

Nonphysical Wave Speeds

The numerical results presented above indicate a disturbing feature of this problem—it is possible to obtain perfectly reasonable results that are stable and free of oscillations and yet are completely incorrect. Needless to say, this can be misleading. In order to understand how the phenomenon occurs, it is sufficient to consider a simpler version of the splitting method, in which we use the splitting (11) with the first order upwind difference method for S_f and the exact solution operator $\bar{S}_\psi$ of the ODE for S_ψ. The method is then

$$\begin{aligned} U_j^* &= U_j^n - \frac{k}{h}(U_j^n - U_{j-1}^n) \\ U_j^{n+1} &= \bar{S}_\psi(k)\, U_j^*. \end{aligned} \tag{14}$$

We use the exact solution operator for $\bar{S}_\psi$ to avoid the suspicion that difficulties are caused by the ODE solver.

We also want to stress that for this scalar problem the splitting itself should not be suspect. In fact, it can be argued that the splitting method is the correct approach in the following sense. If $\bar{S}_f$ and $\bar{S}_\psi$ represent the *exact* solution operators, then they commute and the true solution $u(x, t)$ in fact satisfies

$$u(x, t) = \bar{S}_\psi(t)\, \bar{S}_f(t)\, u(x, 0).$$

This is just a restatement of the fact that the solution is obtained by integrating the ODE along characteristics, since $\bar{S}_f(t)\, u\,(x, 0) = u(x-t, 0)$. But this says that the true solution at time $t+k$ can be obtained from the solution at time t via the split method

$$u^*(x) = \bar{S}_f(k)\, u\,(x, t) \qquad [= u(x-k, t)]$$
$$u(x, t+k) = \bar{S}_\psi(k)\, u^*(x).$$

The method (14) is a direct discretization of this in which we replace $\bar{S}_f$ by the upwind method. This amounts to replacing $u(x-k, t)$ by the interpolated value

$$\left(1-\frac{k}{h}\right) U_j^n + \frac{k}{h} U_{j-1}^n.$$

To see why this apparently reasonable method gives incorrect results when $k\mu$ is large, suppose we take initial data

$$U_j^n = \begin{cases} 1 & \text{if } j < J \\ 0 & \text{if } j \geqslant J \end{cases} \tag{15}$$

for some J. Applying the first step of (14) gives

$$U_j^* = \begin{cases} 1 & \text{if } j < J \\ \lambda & \text{if } j = J, \\ 0 & \text{if } j > J. \end{cases} \tag{16}$$

where $\lambda = k/h$. In the second step we solve the ODE, which gives $U_j^{n+1} = U_j^*$ for $j \neq J$. For $j = J$ the value we obtain depends on λ and the size of $k\mu$. The interesting case is when $k\mu \gg 1$, so that U_J is restored to near equilibrium at the end of the time step. The equilibrium value reached depends on λ, which by (16) is the initial condition for the ODE at grid point J. If $\lambda < 1/2$ then the solution rapidly decays to zero and so $U_J^{n+1} \approx 0$. If $\lambda > 1/2$ then the solution rapidly approaches 1, so $U_J^{n+1} \approx 1$. We thus obtain the results, depending on the mesh ratio λ: If $\lambda < 1/2$,

$$U_j^{n+1} \approx \begin{cases} 1 & \text{if } j < J \\ 0 & \text{if } j \geqslant J; \end{cases}$$

if $\lambda > 1/2$,

$$U_j^{n+1} \approx \begin{cases} 1 & \text{if } j < J+1 \\ 0 & \text{if } j \geqslant J+1. \end{cases}$$

The same behavior occurs in each time step and so we obtain a wave moving with speed 0 if $\lambda < 1/2$ or with speed $1/\lambda$ (i.e., one mesh cell per time step) if $\lambda > 1/2$. (If λ is very close to $1/2$ this argument is not valid. In fact, it is easy to verify that if

$\lambda = 1/2$ the method gives the correct speed 1. However, this is an unlikely special case.)

In general we obtain propagation at a nonphysical speed that is purely an artifact of the numerical method. The problem lies with the smearing of the discontinuity caused by the advection, which introduces a nonequilibrium state ($U_j^* = \lambda$) into the calculation. Unfortunately, any conservative shock-capturing method for the conservation laws will necessarily introduce some smearing since the true shock location almost never coincides with a cell boundary. At least one point in a shock is necessary in order to represent a discontinuity within a cell. As soon as a nonequilibrium value is introduced in this manner, the source terms turn on and immediately restore equilibrium, thus shifting the discontinuity to a cell boundary.

It is difficult to see how this problem can be avoided using standard finite difference methods of the type used here, short of increasing the resolution considerably so that $k\mu$ is small. To see how small $k\mu$ must be to obtain reasonable results, it is interesting to plot the observed wave speed as a function of $k\mu$ for fixed k/h. Because of the form of the true solution it is natural to define the wave speed in time step n by

$$\text{wave speed} = \frac{h}{k}\left(\sum_j U_j^n - \sum_j U_j^{n-1}\right).$$

For large $k\mu$ this is essentially equal to 0 (if $\lambda < 1/2$) or $1/\lambda = h/k$ (if $\lambda > 1/2$) in each step. For smaller $k\mu$ this varies with n in a regular but generally oscillatory manner. To compare wave speeds for various $k\mu$, we define an average wave speed by averaging this function over a fixed time interval t_0 to t_n, or equivalently as

$$\text{average speed} = \frac{h}{(t_n - t_0)}\left(\sum_j U_j^n - \sum_j U_j^0\right). \tag{17}$$

Figure 4 shows this average speed as a function of $k\mu$ for several values of λ.

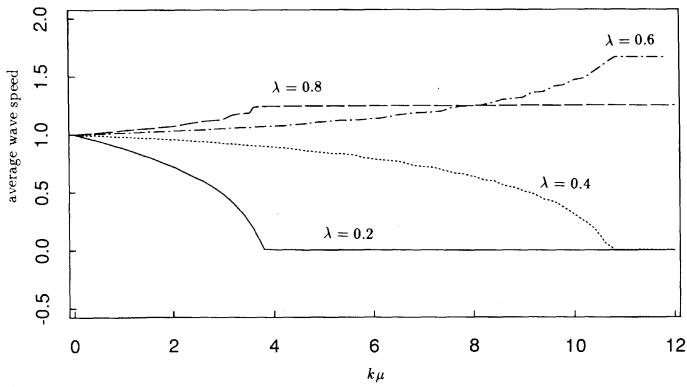

FIG. 4. Average wave speed as function of $k\mu$ for $\lambda = k/h$ fixed.

This can be equivalently viewed as giving results on a fixed grid as μ varies or as giving results with fixed μ as k and h are varied (with λ fixed), i.e., as the grid is refined. The latter viewpoint is more relevant to the discussion here and shows that a substantial refinement (e.g., $k\mu < 1$) is necessary to obtain reasonable results.

In these calculations k/h was held fixed. One might also consider fixing h but taking k much smaller in an attempt to resolve the nonequilibrium effects. Figure 5 shows that this is unsuccessful. Here h is held fixed at 0.01 and for various values of μ the speed is plotted as a function of $\lambda = k/h$, as k is varied. As expected, the correct speed is obtained only at $\lambda = 1/2$ and $\lambda = 1$. Note in particular that letting $k \to 0$ with h fixed is detrimental and that if $h\mu > 1$ there is little hope of obtaining the correct speed. These results indicate that spatial resolution is as important as temporal resolution. This is not surprising, since it is the smearing of the discontinuity that is the source of the difficulty, and the extent of the smearing depends on the spatial resolution.

If we wish to solve such problems without refining the grid to the extent indicated above, we must consider alternatives to the uniform finite difference methods considered so far. We must find methods that are capable of essentially increasing the spatial resolution without excessive refinement of the overall grid.

One possibility is to use local refinement only near the reaction fronts. This is certainly more efficient than global refinement and may be practical in situations where the value of $k\mu$ on the coarse grid is moderate, so that a reasonable degree of refinement will give greatly improved results. Note that refinement in both space and time by a factor of 10 $k\mu$, for example, would reduce the fine grid value of $k\mu$ to 0.1. According to our numerical results, accurate propagation can be achieved at this point.

In situations where $k\mu$ is several orders of magnitude larger than 1, this degree of refinement may not be practical, and is certainly not desirable if we can find another approach that achieves the correct propagation speed without resolving the fastest time scales.

Front tracking is one possibility, in which the reaction fronts are replaced by

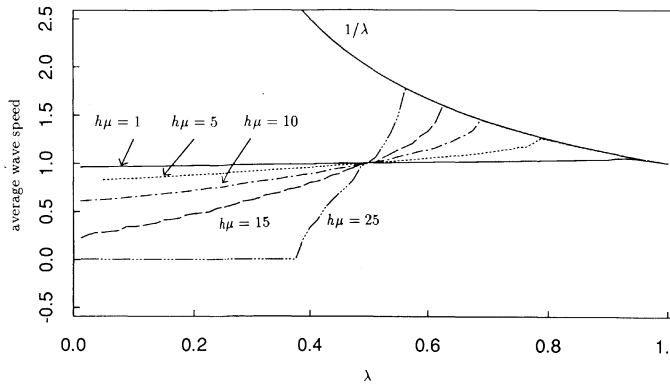

FIG. 5. Average wave speed as function of λ for h fixed.

sharp discontinuities that are explicitly tracked as the solution evolves. This would probably give the best results, but is quite complicated in multi-dimensional problems. It would be nice to develop methods that can deal with stiff reaction fronts more robustly without requiring explicit tracking.

It is illuminating to compare the present situation with that of a homogeneous system of conservation laws with no source terms. In the latter case, the use of a "conservative" numerical method (as defined below) guarantees that an isolated numerical shock of the type considered here must propagate at the correct speed. It may be smeared out over several mesh points, but the speed, defined in manner analogous to (17), must be correct by conservation. To see this, note that by defining the cell average

$$u_j^n = \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} u(x, t_n)\, dx$$

and integrating the conservation law $u_t + f(u)_x = 0$ over $[x_{j-1/2}, x_{j+1/2}] \times [t_n, t_{n+1}]$, we obtain

$$u_j^{n+1} = u_j^n - \frac{1}{h}\left[\int_{t_n}^{t_{n+1}} f(u(x_{j+1/2}, t))\, dt - \int_{t_n}^{t_{n+1}} f(u(x_{j-1/2}, t))\, dt\right]. \tag{18}$$

Summing this expression over j gives cancellation of the flux terms so that we are left only with fluxes at the boundaries of our region.

Our numerical values U_j^n are approximations to u_j^n. A finite difference method is said to be conservative if it can be written in the conservation form

$$U_j^{n+1} = U_j^n - \frac{k}{h}[F_{j+1/2}^n - F_{j-1/2}^n], \tag{19}$$

where $F_{j\pm 1/2}^n$ are the numerical fluxes based on U at neighboring points, and $kF_{j\pm 1/2}^n$ approximates the corresponding integral in (18). Summing (19) over j gives the same cancellation of fluxes as in the true solution. Provided the fluxes at the boundaries of the region are correct, we maintain the correct total sum in each time step and hence the correct speed in the simple case of piecewise constant data with a single discontinuity. Lax and Wendroff [12] have shown more generally that a convergent conservative method must converge to a weak solution of the conservation laws and thus must give discontinuities in the correct locations.

If we now include a source term and integrate

$$u_t + f(u)_x = \psi(u)$$

as before, we obtain

$$u_j^{n+1} = u_j^n - \frac{1}{h}\left[\int_{t_n}^{t_{n+1}} f(u(x_{j+1/2}, t))\, dt - \int_{t_n}^{t_{n+1}} f(u(x_{j-1/2}, t))\, dt\right]$$
$$+ \frac{1}{h}\int_{t_n}^{t_{n+1}}\int_{x_{j-1/2}}^{x_{j+1/2}} \psi(u(x, t))\, dx\, dt. \tag{20}$$

The new term appearing here does not undergo cancellation when we sum over j, and consequently it is important that this term is modeled accurately if we are to obtain the correct behavior.

Note that for the model problem we are considering, where the true u is everywhere in equilibrium except at the discontinuity, we have $\psi(u) \equiv 0$ almost everywhere and so the integral of ψ in (20) is zero. In the numerical methods we have been considering, this integral is approximated by something analogous to $k\psi(u_j^n)$. This is a reasonable approximation if u is smooth, but very poor in the present context. We are replacing the average value of $\psi(u)$ (which should be zero) by ψ evaluated at the average value of u (which may be far from zero).

One possible approach toward deriving better numerical methods is to attempt to model the integral of ψ in (20) more accurately than simply using $k\psi(U_j^n)$. One possibility is to compute some approximation $v(x)$ to $u(x, t_n)$ based on the grid values U_j^n, and then integrate $\psi(v(x))$. One way to obtain a local reconstruction is to use the "subcell resolution" approach of Harten [10]. This method was originally proposed as a way to obtain sharp contact discontinuities in nonreacting flows, but appears promising in our context as well. The idea is to construct a piecewise polynomial function based on the data U_j^n that may have discontinuities within the cells. Smoothness criteria and conservation are used to locate the discontinuities. Harten [9] has tested a version of his method on the model problem considered here and reports excellent results, as would be expected on this scalar problem with piecewise constant solution. It is not yet clear to what extent this approach can be extended to systems of equations and ultimately to multidimensional problems.

5. Conclusions

We have proposed a very simple scalar equation as a model problem for understanding the behavior of numerical methods on reacting flow problems. We have considered two classes of numerical methods for this problem: MacCormack style predictor-corrector methods and splitting methods. In either case it is possible to derive second-order accurate methods that are stable even for very stiff problems. However, all of these methods are subject to another numerical difficulty in the stiff case—incorrect propagation speeds of discontinuities. We have shown that this results from a lack of spatial resolution in evaluating the source terms. A nonequilibrium value in the numerical representation of the discontinuity, when viewed as the average value of u over a large mesh cell, will cause the source terms to be activated over this entire cell in a nonphysical manner. In order to avoid this difficulty, it will be necessary to increase the resolution of the discontinuity, at least for the purpose of evaluating $\psi(u)$. One possibility is to use some form of mesh refinement or shock tracking. A more appealing alternative is to attempt to model the integral of ψ in (20) more accurately using, for example, subcell resolution. It

is not yet clear to what extent these approaches are practical for multidimensional systems of equations. The development of new methods along these lines is the subject of ongoing research.

6. Appendix: Accuracy of the Method (8)

The truncation error $T(x, t)$ of (8) is defined by

$$T(x, t) = \frac{1}{k}(u(x, t+k) - u(x, t)) - \frac{1}{2k}(\Delta u^{(1)}(x, t) + \Delta u^{(2)}(x, t)), \tag{21}$$

where

$$\Delta u^{(1)}(x, t) = \left[I - \frac{1}{2}k\psi'(u(x, t))\right]^{-1} \times \left\{-\frac{k}{h}[f(u(x, t)) - f(u(x-h, t))] + k\psi(u(x, t))\right\}$$

$$u^{(1)}(x, t) = u(x, t) + \Delta u^{(1)}(x, t)$$

and

$$\Delta u^{(2)}(x, t) = \left[I - \frac{1}{2}k\psi'(\hat{u}(x, t))\right]^{-1} \times \left\{-\frac{k}{h}[f(u^{(1)}(x+h, t)) - f(u^{(1)}(x, t))] + k\psi(\bar{u}(x, t))\right\}.$$

Here the choice of $\hat{u}$ and $\bar{u}$ correspond to the choice of $\hat{U}$ and $\bar{U}$ in (8), i.e., either $u(x, t)$ or $u^{(1)}(x, t)$. To compute the order of accuracy we must expand in Taylor series and simplify the expression (21) for $T(x, t)$. This is easiest to do if we first consider the choice $\hat{U} = \bar{U} = U^n$. Then (21) becomes

$$T(x, t) = \frac{1}{k}(u(x, t+k) - u(x, t)) + \frac{1}{2h}\left[I - \frac{1}{2}k\psi'(u(x, t))\right]^{-1} \times \{f(u(x,, t)) - f(u(x-h, t)) + f(u^{(1)}(x+h, t)) - f(u^{(1)}(x, t)) - 2h\psi(u(x, t))\}. \tag{22}$$

Using the approximation

$$[I - \tfrac{1}{2}k\psi'(u(x, t))]^{-1} = I + \tfrac{1}{2}k\psi'(u(x, t)) + O(k^2),$$

we obtain

$$u^{(1)}(x,t) = u(x,t) + \left[I + \frac{1}{2} k\psi'(u) + \dots \right] \left\{ -\frac{k}{h} \left[hf(u)_x - \frac{1}{2} h^2 f(u)_{xx} + \dots \right] + k\psi(u) \right\}$$
$$= u + k[\psi(u) - f(u)_x] + O(k^2)$$
$$= u + ku_t + O(k^2),$$

where $u \equiv u(x,t)$. Consequently,

$$f(u^{(1)}(x,t)) = f(u) + kf'(u)u_t + O(k^2).$$

Moreover, the $O(k^2)$ terms here are smooth functions of x and so will cancel to $O(k^3)$ when we compute $f(u^{(1)}(x+h,t)) - f(u^{(1)}(x,t))$, giving

$$\begin{aligned} f(u^{(1)}(x+h,t)) - f(u^{(1)}(x,t)) &= \{[f(u) + hf(u)_x + \tfrac{1}{2}h^2 f(u)_{xx} + \dots] \\ &\quad + k[f'(u) + hf'(u)_x + \dots][u_t + hu_x + \dots] + O(k^2)\} \\ &\quad - \{f(u) + kf'(u)u_t + O(k^2)\} \\ &= hf(u)_x + \tfrac{1}{2}h^2 f(u)_{xx} + hk(f'(u)_x u_t + f'(u)u_{tx}) + O(k^3) \\ &= hf(u)_x + \tfrac{1}{2}h^2 f(u)_{xx} + hkf(u)_{xt} + O(k^3). \end{aligned}$$

We also have that

$$f(u(x,t)) - f(u(x-h,t)) = hf(u)_x - \tfrac{1}{2}h^2 f(u)_{xx} + O(k^3)$$

and so (22) becomes

$$\begin{aligned} T(x,t) &= \left\{ u_t + \frac{1}{2} ku_{tt} + O(k^2) \right\} + \frac{1}{2h} \left[I + \frac{1}{2} k\psi'(u) + O(k^2) \right] \\ &\quad \times \{2hf(u)_x + khf(u)_{tx} - 2h\psi(u) + O(k^2)\} \\ &= u_t + \tfrac{1}{2}ku_{tt} + (f(u)_x - \psi(u)) + \tfrac{1}{2}k[\psi'(u)f(u)_x + f(u)_{tx} - \psi'(u)\psi(u)] + O(k^2) \\ &= O(k^2), \end{aligned}$$

since $u_t = \psi(u) - f(u)_x$ and $u_{tt} = \psi'(u)u_t - f(u)_{tx} = \psi'(u)\psi(u) - \psi'(u)f(u)_x - f(u)_{tx}$. This shows that the method (8) is second-order accurate provided we use $\bar{U}_j = \hat{U}_j = U_j^n$.

Now consider what happens if we use $\bar{U}_j = U_j^{(1)}$ instead of $\bar{U}_j = U_j^n$. Then the term $2h\psi(u(x,t))$ in (22) will become

$$\begin{aligned} h[\psi(u(x,t)) + \psi(u^{(1)}(x,t))] &= 2h\psi(u(x,t)) + h\psi'(u(x,t))\,\Delta u^{(1)}(x,t) + \cdots \\ &= 2h\psi(u(x,t)) + hk\psi'(u(x,t))\,u_t + O(k^3). \end{aligned}$$

Since this factor is multiplied by $1/h$ in computing $T(x, t)$, this will cause an $O(k)$ change in the truncation error and hence a reduction to first-order accuracy in general. But note that in computing a steady state, where $u_t = 0$, this perturbation drops out and so $\bar{U}_j = U_j^{(1)}$ can be used in that case.

To justify our other claim, that alternative values of $\hat{U}_j$ are allowed provided that $\psi'(\hat{U}_j) = \psi'(U_j^n) + O(k)$, consider the effect that using a different $\hat{U}_j$ would have on $T(x, t)$. For analytical purposes, we can rewrite (8) in this case as

$$\left[I - \frac{1}{2}k\psi'(U_j^n)\right] \Delta U_j^{(1)} = -\frac{k}{h}[f(U_j^n) - f(U_{j-1}^n)] + k\psi(U_j^n)$$

$$U_j^{(1)} = U_j^n + \Delta U_j^{(1)}$$

$$\left[I - \frac{1}{2}k\psi'(U_j^n)\right] \Delta U_j^{(2)} = -\frac{k}{h}[f(U_{j+1}^{(1)}) - f(U_j^{(1)})] + k\psi(U_j^n)$$

$$\Delta \hat{U}_j = [I - \tfrac{1}{2}k\psi'(\hat{U}_j)]^{-1}\,[I - \tfrac{1}{2}k\psi'(U_j^n)]\,\Delta U_j^{(2)}$$

$$U_j^{n+1} = U_j^n + \tfrac{1}{2}(\Delta U_j^{(1)} + \Delta \hat{U}_j).$$

The first three lines are identical to the method already analyzed, i.e., (8) with $\hat{U}_j = \bar{U}_j = U_j^n$. But now we compute a modified increment $\Delta \hat{U}_j$ and use this to update U_j^n rather than $\Delta U_j^{(2)}$. Clearly the method remains second order accurate provided $\Delta \hat{U}_j = \Delta U_j^{(2)} + O(k^3)$. But this follows easily by Taylor series expansion of the definition of $\Delta \hat{U}_j$, since $\psi'(\hat{U}_j) = \psi'(U_j^n) + O(k)$ and $\Delta U_j^{(2)} = O(k)$.

Acknowledgments

Valuable discussions with Ami Harten, Elaine Oran, Chul Park, Philip Roe, and Robert Warming during the course of this work are gratefully acknowledged.

References

1. T. Aki, National Aerospace Laboratory Technical Report, Tokyo, 1987 (unpublished).
2. J. P. Boris and E. S. Oran, *Numerical Simulation of Reactive Flow* (Elsevier, Amsterdam/New York, 1987).
3. T. R. A. Bussing and E. M. Murman, AIAA Paper 85-0331, 1985 (unpublished).
4. G. C. Carofano, Technical Report ARLCB-TR-84029, 1984 (unpublished).
5. P. Colella, A. Majda, and V. Roytburd, *SIAM J. Sci. Stat. Comput.* **7**, 1059 (1986).
6. P. Colella and P. Woodward, *J. Comput. Phys.* **54**, 174 (1984).
7. J. P. Drummond, R. C. Rogers, and M. Y. Hussaini, AIAA Paper 86-1327, 1986 (unpublished).
8. C. W. Gear, *Numerical Initial Value Problems in Ordinary Differential Equations* (Prentice–Hall, Englewood Cliffs, NJ, 1971).
9. A. Harten, pivate communication (1987).
10. A. Harten, ICASE Report No. 87-56, NASA Langley Research Center, 1987 (unpublished).
11. J. D. Lambert, *Computational Methods in Ordinary Differential Equations* (Wiley, New York, 1973).

12. P. D. LAX AND B. WENDROFF, *Commun Pure Appl. Math.* **13**, 217 (1960).
13. J. LEE, AIAA Paper 84-1729 (1984).
14. R. W. MACCORMACK, AIAA Paper 69-354 (1969).
15. P. L. ROE, *J. Comput. Phys.* **43**, 357 (1981).
16. N. D. SANDHAM AND H. C. YEE, NASA Technical Memorandum 102194, 1989 (unpublished).
17. G. STRANG, *SIAM J. Num. Anal.* **5**, 506 (1968).
18. W. G. VINCENTI AND C. H. KRUGER, JR., *Introduction to Physical Gas Dynamics* (Wiley, New York, 1967).
19. R. F. WARMING, P. KUTLER, AND H. LOMAX, *AIAA J.* **189** (1973).
20. H. C. YEE, NASA Ames Technical Memoranda 89464, 1987 (unpublished), and 101088, 1989 (unpublished).
21. H. C. YEE AND J. L. SHINN, AIAA Paper 87-1116, 1987 (unpublished).
22. T. R. YOUNG AND J. P. BORIS, *J. Phys. Chem.* **81**, 2424 (1977).
23. V. Y. C. YOUNG AND H. C. YEE, AIAA Paper 87-0112, 1987 (unpublished).