ELSEVIER

Journal of Computational and Applied Mathematics 66 (1996) 359–375

JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

Extrapolation methods for hyperbolic systems with relaxation

M. Morandi Cecchi [a], M. Redivo-Zaglia [b,*], G. Russo [c]

[a] *Dipartimento di Matematica Pura ed Applicata, University of Padova, Via Belzoni 7, 35131 Padova, Italy*

[b] *Dipartimento di Elettronica e Informatica, University of Padova, Via Gradenigo 6/A, 35131 Padova, Italy*

[c] *Dipartimento di Matematica, University of L'Aquila, Via Vetoio, Loc. Coppito, 67010 L'Aquila, Italy*

Received 5 October 1994; revised 18 May 1995

Abstract

A splitting scheme is used for numerical solution of hyperbolic systems with a stiff relaxation. High-order flux in the convection step guarantees a good spatial accuracy. The relaxation term is treated implicitly. Extrapolation is used to combine the steps, obtaining high accuracy in time. A third-order scheme both in space and in time is presented and applied to the Broadwell model of gas dynamics.

Keywords: Broadwell model; Conservation laws; Extrapolation methods; Finite differences; Relaxation; Richardson extrapolation

AMS classification: 34A65; 35L65; 43M25; 65B05; 82B40

1. Introduction

Hyperbolic systems with relaxation are used to describe many physical problems that involve both convection and nonlinear interaction.

Relaxation occurs in water waves when the gravitational force balances the frictional forces of the river bed. In the Boltzmann equation from the kinetic theory of rarefied gas dynamics the collision (relaxation) terms describe the interaction of particles. Relaxation also occurs in other problems ranging from magnetohydrodynamics to traffic flow.

In kinetic theory the relaxation term is large when the mean free path between collisions is small. By analogy with the kinetic theory, we shall refer to the limit of large relaxation rate (or small relaxation time) for a general hyperbolic system with relaxation as the fluid dynamic limit.

Such a limit is characterized by the fact that in this regime, the relaxation terms become stiff. In particular, a standard numerical scheme might fail to give physically correct solution once the relaxation distance is smaller than the spatial discretization.

* Corresponding author. E-mail: elen@elett1.dei.unipd.it.

SSDI 0377-0427(95)00194-8

Schemes for hyperbolic systems with stiff relaxation have been discussed by Pember [11, 12].

A class of numerical methods using implicit finite difference equations have been presented in [4] working with uniform accuracy from the rarefied regime to the fluid dynamic limit for the Broadwell model of kinetic theory. A splitting scheme is used for the convection and collision steps. High-order flux in the convection step allows high spatial accuracy. A second-order time discretization can be obtained either by using Richardson extrapolation or by a suitable combination of relaxation and convection steps.

In this paper one pushes the Richardson extrapolation process to its second column. Such an approach seems preferable to the one used in [4], because of the complicated algebraic conditions imposed on the parameters for obtaining a high-order scheme.

The motivation differs from some earlier approaches to the solution of systems of equations with relaxation in that we seek to develop robust numerical schemes that work with a wide range of relaxation rate. We improve the convergence of the schemes when using a splitting by improving the order of convergence of the time discretization.

The Broadwell model describes a 2-D (3-D) gas as composed of particles of only four (six) velocities with a binary collision law [3]. When looking for one-dimensional solutions of the 2-D gas, the evolution equations for the density function in phase space are given by

$$\begin{aligned} &\partial_t f + \partial_x f = \frac{1}{\varepsilon}(h^2 - fg), \\ &\partial_t h = -\frac{1}{\varepsilon}(h^2 - fg), \\ &\partial_t g - \partial_x g = \frac{1}{\varepsilon}(h^2 - fg), \end{aligned} \tag{1}$$

where ε is the mean free path, f, h and g denote the mass densities of gas particles with x-velocity 1, 0 and -1, respectively, in space x and time t.

The fluid dynamic moment variables are density ϱ, momentum m and velocity u defined by

$$\begin{aligned} &\varrho = f + 2h + g, \\ &m = f - g, \\ &u = \frac{m}{\varrho}. \end{aligned}$$

In addition, we define

$$z = f + g.$$

Then the Broadwell equations can be rewritten as

$$\begin{aligned} &\partial_t \varrho + \partial_x m = 0, \\ &\partial_t m + \partial_x z = 0, \\ &\partial_t z + \partial_x m = \frac{1}{2\varepsilon}(\varrho^2 + m^2 - 2\varrho z). \end{aligned}$$

The Broadwell system is an example of hyperbolic system of conservation law with a nonlinear source term. Such systems can be written in the general form

$$\partial_t U + \partial_x F(U) = -\frac{1}{\varepsilon} R(U), \quad U \in \mathbb{R}^N \tag{2}$$

They are called *relaxation systems* in the sense of Whitham [14] and Liu [8] if there exists a constant matrix Q with rank $n < N$, such that

$$QR(U) = 0, \quad \forall U \in \mathbb{R}^N.$$

This implies that n-independent quantities $v = QU$ are conserved. Also we assume that v uniquely determines a local equilibrium value, $U = \mathscr{E}(v)$, satisfying $R(\mathscr{E}(v)) = 0$ and such that $Q\mathscr{E}(v) = v, \forall v \in \mathbb{R}^n$.

As $\varepsilon \to 0$, system (2) reduces to a system of n conservation laws

$$\partial_t v + \partial_x e(v) = 0,$$

where $e(v) \equiv QF(\mathscr{E}(v))$.

2. Splitting schemes: convection and relaxation steps

A splitting scheme for system (2) consists in the solution of a convection step and a relaxation step, and in the combination of them.

2.1. Convection step: upwind schemes

Let us consider the convection part of system (2):

$$\partial_t U + \partial_x F(U) = 0, \quad U \in \mathbb{R}^N. \tag{3}$$

A conservative spatial discretization of (3) is

$$\partial_t U_j + \frac{F_{j+1/2} - F_{j-1/2}}{\Delta x} = 0,$$

where U_j may be an approximation of the cell average

$$U_j(t) \approx \frac{1}{\Delta x} \int_{x_{j-1/2}}^{x_{j+1/2}} U(x,t)\,\mathrm{d}x \tag{4}$$

or an approximation of a pointwise value of the exact solution and F_j is a suitably defined "flux function".

For a scalar equation ($N = 1$), an upwind scheme is constructed by taking

$$F_{j+1/2} = \begin{cases} F_j & \text{if } F'(U_j) > 0, \\ F_{j+1} & \text{if } F'(U_j) < 0. \end{cases} \tag{5}$$

In the case of a system where F is a linear function of U, i.e. if $A = \nabla_U F$ is a constant matrix, then the system is diagonalized when expressed in terms of the Riemann variables:

$$\partial_t V + \Lambda \partial_x V = 0,$$

where $U = PV$, V is the vector of Riemann variables, and P is the matrix of the eigenvectors of A. According to (5), one defines

$$V_{j+1/2}^{(k)} = \begin{cases} V_j^{(k)} & \text{if } \lambda^{(k)} > 0 \\ V_{j+1}^{(k)} & \text{if } \lambda^{(k)} < 0 \end{cases} \quad k = 1,\ldots,N, \tag{6}$$

where $\lambda^{(k)}$ is the eigenvalues corresponding to the Riemann variable $V^{(k)}$. The vector flux function $F_{j+1/2}$ is obtained by expressing (6) back in the original field variable U.

Let us consider first schemes which are based on cell average approximation (4).

In a first-order upwind method, the Riemann variables are approximated by a piecewise constant function over each cell. A second-order scheme can be constructed by using a piecewise linear approximation of the Riemann variables over the cell. However, near discontinuities, a scheme must necessarily be of first order. Therefore the high-order flux functions are equipped by a suitable flux limiter, which makes the scheme first-order near discontinuities, and reduces oscillations. An extensive treatment of such methods can be found in [7].

Spatial accuracy may be improved, and third-order schemes can be constructed by a piecewise quadratic approximation of the Riemann variables (piecewise parabolic method, or PPM).

By using forward Euler time discretization one obtains a step which is accurate in space and first-order in time. We shall construct schemes for system (2) which are based on van Leer second-order scheme and on PPM third-order scheme.

High-order shock capturing schemes can be constructed also by starting from pointwise approximation of the solution as in the case of ENO schemes [13]. As we shall see, this approach is in fact more convenient for obtaining third-order schemes for hyperbolic systems with relaxation.

As a simple illustration of this approach, we consider second and third order upwind schemes applied to the wave equation

$$u_t - a u_x = 0,$$

where a is a positive constant. They are obtained, respectively, by second- and third-order upwind approximation of space derivative. The semidiscrete schemes are given by the following.
Second order:

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} - \frac{a}{2\,\Delta x}(-3u_j + 4u_{j+1} - u_{j+2}) = 0. \tag{7}$$

Third order:

$$\frac{\mathrm{d}u_j}{\mathrm{d}t} - \frac{a}{12\,\Delta x}(-3u_{j-1} - 10u_j + 18u_{j+1} - 6u_{j+2} + u_{j+3}) = 0. \tag{8}$$

We shall denote by P2 and P3, respectively, the space discretization (7) and (8).

2.2. Relaxation and combination of the steps

The relaxation step is

$$\begin{cases} \partial_t U = -\dfrac{1}{\varepsilon} R(U), \\ U(0) = U_0. \end{cases} \tag{9}$$

Because of the stiffness of the problem, it is necessary to solve the step by an implicit scheme. An essential feature of the relaxation step is that

$$U(\Delta t) \to \mathscr{E}(Q(U_0)) \quad \text{as } \varepsilon \to 0, \tag{10}$$

that is, the field must relax to equilibrium as the relaxation time vanishes. This is because the time asymptotic solution of Eq. (6) satisfies $R(U) = 0$.

Sometimes (9) can be explicitly solved analytically. The analytical solution can be used since it satisfies property (10). Implicit Euler is usually also a good choice:

$$U_j^{n+1} - U_j^n = -\frac{1}{\varepsilon} R(U_j^{n+1}).$$

Relaxation and convection steps can be combined together as

$$\begin{aligned} \tilde{U} &= U^n - \frac{\Delta t}{\varepsilon} R(\tilde{U}), \\ U^{n+1} &= \tilde{U} - \Delta t \mathscr{D} F \tilde{U}, \end{aligned} \tag{11}$$

where $\mathscr{D}$ is some high-order discrete convection operator. The resulting scheme is first order in time. Note that the use of the exact solution of the relaxation step or a higher-order time discretization of the convection step would not improve the accuracy of scheme (11), since the latter is limited by the structure of the splitting.

Higher order in time could be obtained by combining the two steps in a suitable form. In paper [4] a second-order scheme has been derived for an accurate solution of system (2), which works over a large range of the parameter ε. The scheme is a combination of convection and relaxation steps

$$\begin{aligned} U_1 &= U^n - \alpha \frac{\Delta t}{\varepsilon} R(U_1), \\ U_2 &= U_1 - \tilde{\alpha}\, \Delta t\, \mathscr{D} F(U_1), \\ U_3 &= U_2 - \beta \frac{\Delta t}{\varepsilon} R(U_3) - \gamma \frac{\Delta t}{\varepsilon} R(U_1), \\ U_4 &= U_3 - \tilde{\beta}\, \Delta t\, \mathscr{D} F(U_3), \\ U_5 &= \xi U^n + \eta U_4, \\ U^{n+1} &= U_5 - \mu \frac{\Delta t}{\varepsilon} R(U^{n+1}), \end{aligned} \tag{12}$$

where $\mathscr{D}$ is a second-order accurate discrete convection operator, and the parameters α, $\tilde{\alpha}$, β, $\tilde{\beta}$, γ, ξ, η, μ have been determined by imposing that the scheme is second order both in space and time.

In the actual implementation, a second-order van Leer scheme has been used for the convection step [7].

We shall call this scheme VLSP (van Leer splitting).

Better spatial accuracy could be obtained by using a third-order discrete operator in the convection step (for example the piecewise parabolic method, PPM [7], or schemes based on discretization (8)). However, in order to obtain a third-order scheme in space and time, an accurate time discretization is required.

It seems difficult to device a third-order scheme by using an approach based on scheme (12), because it requires the derivation and solution of a complicated system of algebraic equations in order to match the first terms in the Taylor expansion of the exact solution of system (2).

Because of these difficulties we propose an approach which is based on the use of Richardson extrapolation to improve the accuracy in time.

3. Extrapolation methods

In numerical analysis there are many methods producing sequences. Such is the case, for instance, of discretization methods as the one we discussed above. In these methods, extrapolation procedures can be used for improving the accuracy of the solution.

Let (S_n) be the sequence to be accelerated. It is assumed to converge to a limit S. A sequence transformation T consists in transforming this sequence into a new one called (T_n), that is $T:(S_n) \longmapsto (T_n)$. The transformation T is said to accelerate the convergence of the sequence (S_n) if and only if

$$\lim_{n\to\infty} \frac{T_n - S}{S_n - S} = 0.$$

In that case, one can also say that (T_n) converges to S faster than (S_n).

In the study of a sequence transformation the first question to be asked and solved (before those of convergence and acceleration) is an algebraic one: it concerns the so-called kernel $\mathscr{K}_T$ of the transformation that is the set of sequences for which $\exists S$ such that $\forall n \geqslant N$, $T_n = S$. We are able to construct a sequence transformation starting from a given kernel $\mathscr{K}_T$. It is more difficult to find the kernel of a given sequence transformation.

A sequence transformation $T:(S_n) \longmapsto (T_n)$ is also called an extrapolation method.

Among the extrapolation methods (for a review, see [2]), the most well known are certainly Richardson's extrapolation algorithm and Aitken's Δ^2 process. Sometimes the Richardson extrapolation process is called h-extrapolation, where we take $x_n = h_n$ that is the step length of the subdivision intervals used in the computation of the terms of the sequence.

The kernel of the Richardson extrapolation process is the set of sequences of the form

$$S_n = S + a_1 x_n + a_2 x_n^2 + \cdots + a_k x_n^k,$$

where (x_n) is an auxiliary sequence such that $\forall i$ and $\forall j \neq i$, $x_i \neq x_j$. Thus Richardson process corresponds to polynomial extrapolation at the point 0.

The rule of this extrapolation algorithm is

$$T_0^{(n)} = S_n, \quad n = 0, 1, \ldots,$$

$$T_k^{(n)} = \frac{x_{n+k} T_{k-1}^{(n)} - x_n T_{k-1}^{(n+1)}}{x_{n+k} - x_n}, \quad k = 1, 2, \ldots; \ n = 0, 1, \ldots. \tag{13}$$

$T_k^{(n)}$ is the value at the point 0 of the interpolation polynomial $P_k^{(n)}$, of degree at most k, which satisfies

$$P_k^{(n)}(x_{n+i}) = S_{n+i}, \quad i = 0, \ldots, k.$$

It is well known that these interpolation polynomials $P_k^{(n)}$ can be recursively computed by the Neville–Aitken scheme. Setting $x = 0$ in this scheme leads to Richardson process.

Instead of computing recursively the $T_k^{(n)}$'s we can also directly solve the system of linear equations

$$S_{n+i} = T_k^{(n)} + a_1 x_{n+i} + \cdots + a_k x_{n+i}^k, \quad i = 0, \ldots, k.$$

Multiplying equation i by b_i and adding them leads to

$$T_k^{(n)} = b_0 S_n + \cdots + b_k S_{n+k}, \tag{14}$$

where the b_i (which depend on n) are the solution of the system

$$\begin{cases} b_0 + \cdots + b_k = 1 \\ b_0 x_n + \cdots + b_k x_{n+k} = 0 \\ \vdots \qquad\qquad\qquad\quad \vdots \\ b_0 x_n^k + \cdots + b_k x_{n+k}^k = 0. \end{cases}$$

Marchuk and Shaidurov [9] showed that, for some particular choices of x_i and for $n = 0$, the b_i's can be obtained in closed form. In particular, when $x_i = (i+1)^{-1}$, we have

$$b_i = \frac{(-1)^{k-i} \cdot (i+1)^{k+1}}{(i+1)! \cdot (k-i)!}, \quad i = 0, \ldots, k. \tag{15}$$

Marchuk and Shaidurov [10] devoted a whole monograph to the extrapolation of finite difference methods by Richardson process. This book studies the application of Richardson method to first-order ordinary differential equations, to the one-dimensional stationary diffusion equation, to elliptic equations, to nonstationary problems, to integral equations, to quasilinear equations, to eigenvalue problems and to boundary layer problems.

If we know that the sequence has an asymptotic expansion of the form

$$S_n - S = a_1 x_n + a_2 x_n^2 + \cdots + a_k x_n^k + \cdots, \tag{16}$$

with $a_i \neq 0$ for $i = 1, 2, \ldots$, we can however apply Richardson extrapolation process and obtain some interesting properties.

If, for instance, we consider the first sequence $T_1^{(n)}$, we have

$$\begin{aligned} T_1^{(n)} - S &= \frac{x_{n+1} T_0^{(n)} - x_n T_0^{(n+1)}}{x_{n+1} - x_n} - S \\ &= \frac{x_{n+1}\left(a_1 x_n + a_2 x_n^2 + a_3 x_n^3 + \cdots\right) - x_n\left(a_1 x_{n+1} + a_2 x_{n+1}^2 + a_3 x_{n+1}^3 + \cdots\right)}{x_{n+1} - x_n} \\ &= \frac{x_{n+1}\left(a_2 x_n^2 + a_3 x_n^3 + \cdots\right) - x_n\left(a_2 x_{n+1}^2 + a_3 x_{n+1}^3 + \cdots\right)}{x_{n+1} - x_n} \\ &= -a_2 x_n x_{n+1} - a_3 x_n x_{n+1}\left(x_n + x_{n+1}\right) - \cdots, \end{aligned}$$

and for the second column

$$\begin{aligned} T_2^{(n)} - S &= \frac{x_{n+2} T_1^{(n)} - x_n T_1^{(n+1)}}{x_{n+2} - x_n} - S \\ &= \left(x_{n+2}\left(-a_2 x_n x_{n+1} - a_3 x_n x_{n+1}\left(x_n + x_{n+1}\right) - \cdots\right)\right. \\ &\quad \left. - x_n\left(-a_2 x_{n+1} x_{n+2} - a_3 x_{n+1} x_{n+2}\left(x_{n+1} + x_{n+2}\right) - \cdots\right)\right) / \left(x_{n+2} - x_n\right) \\ &= a_3 x_n x_{n+1} x_{n+2} + \cdots \end{aligned}$$

and so on.

4. Application of Richardson extrapolation process

In this section we apply truncation analysis to a linear version of system (2)

$$\frac{\partial U}{\partial t} = (A + B)\, U, \tag{17}$$

where A and B are two linear operators. We may assume that $U \in \mathbb{R}^N$, and A and B are two constant matrices.

Let us write the exact solution of (17) as

$$U(t) = \mathscr{S}(t) U(0), \tag{18}$$

where $\mathscr{S}(t)$ is the evolution operator and $U(0)$ is the initial value. At time $t_1 = \Delta t$ the exact solution is given by

$$\begin{aligned} U(t_1) &= \mathscr{S}(\Delta t)\, U(0) \\ &= e^{E_1 \Delta t} U(0) \\ &= \left(I + E_1 \Delta t + \tfrac{1}{2} E_1^2 \Delta t^2 + \tfrac{1}{6} E_1^3 \Delta t^3 + \mathrm{O}\left(\Delta t^4\right)\right) U(0), \end{aligned}$$

where $E_1 = A + B$.

If we denote by $\mathscr{T}(t)$ a first-order difference operator, then the numerical solution is given by

$$u(t) = \mathscr{T}(t) U(0).$$

The operator $\mathscr{T}(\Delta t)$ such that $U^1 = \mathscr{T}(\Delta t)U^0$ can be constructed, for example, by a splitting scheme as in (11):

$$\begin{aligned} \tilde{U} &= U^0 + \Delta t\, B\, \tilde{U}, \\ U^1 &= \tilde{U} + \Delta t\, A\, \tilde{U}, \end{aligned} \tag{19}$$

where the first step is implicit Euler.

Let us consider now the numerical solution of the splitting scheme obtained at the first iteration ($t_1 = \Delta t$). Using a step of Δt we have

$$\begin{aligned} u_1(t_1) &= \mathscr{T}(\Delta t)\, U(0) \\ &= (I + E_1 \Delta t + E_2 \Delta t^2 + E_3 \Delta t^3 + \mathrm{O}(\Delta t^4))\, U(0), \end{aligned} \tag{20}$$

since the splitting scheme is globally of the first order (locally second order).

Using two- and three-step iterations we have

$$\begin{aligned} u_2(t_1) &= \mathscr{T}(\tfrac{1}{2}\Delta t)^2 U(0) \\ &= (I + E_1 \Delta t + (2E_2 + E_1^2)\, \tfrac{1}{4}\Delta t^2 + (2E_3 + E_1E_2 + E_2E_1)\, \tfrac{1}{8}\Delta t^3 + \mathrm{O}(\Delta t^4))\, U(0) \end{aligned} \tag{21}$$

and

$$\begin{aligned} u_3(t_1) &= \mathscr{T}(\tfrac{1}{3}\Delta t)^3\, U(0) \\ &= (I + E_1 \Delta t + (E_2 + E_1^2)\, \tfrac{1}{3}\Delta t^2 + (3E_3 + 3E_1E_2 + 3E_2E_1 + E_1^3)\, \tfrac{1}{27}\Delta t^3 + \mathrm{O}(\Delta t^4))\, U(0). \end{aligned} \tag{22}$$

Their related residuals with respect to the exact solution $U(t_1)$ are the following:

$$\begin{aligned} u_1(t_1) - U(t_1) &= ((2E_2 - E_1^2)\, \tfrac{1}{2}\Delta t^2 + (6E_3 - E_1^3)\, \tfrac{1}{6}\Delta t^3 + \mathrm{O}(\Delta t^4))\, U(0), \\ u_2(t_1) - U(t_1) &= ((2E_2 - E_1^2)\, \tfrac{1}{4}\Delta t^2 + (6E_3 + 3E_1E_2 + 3E_2E_1 - 4E_1^3)\, \tfrac{1}{24}\Delta t^3 + \mathrm{O}(\Delta t^4))\, U(0), \\ u_3(t_1) - U(t_1) &= ((2E_2 - E_1^2)\, \tfrac{1}{6}\Delta t^2 + (6E_3 + 6E_1E_2 + 6E_2E_1 - 7E_1^3)\, \tfrac{1}{54}\Delta t^3 + \mathrm{O}(\Delta t^4))\, U(0). \end{aligned}$$

We set

$$T_0^{(n)} = u_{n+1}(t_1) \quad \text{and} \quad x_n = h_n = \frac{\Delta t}{n+1} \qquad \text{for } n = 0, 1, 2, \tag{23}$$

that is, we choose x_n as the step length of the different subdivision intervals. For $n \to \infty$, $x_n = h_n \to 0$ and S is the exact solution $U(t_1)$.

In [4], although the given sequence (20)–(22) apparently does not have an asymptotic expansion of the error like (16), the authors anyway use the rule (13) to compute the first value given in the

first column of Richardson extrapolation process. That is, they compute

$$\begin{aligned}\bar{u}(t_1) = T_1^{(0)} &= \frac{x_1 T_0^{(0)} - x_0 T_0^{(1)}}{x_1 - x_0} = \frac{\frac{1}{2}\Delta t u_1(t_1) - \Delta t\, u_2(t_1)}{\frac{1}{2}\Delta t - \Delta t} \\ &= 2u_2(t_1) - u_1(t_1) \end{aligned} \tag{24}$$

and they proved that they obtain a locally third- (globally second-) order approximation.

Now we shall prove that, using again Richardson extrapolation process, but by considering as the first time step the value $T_2^{(0)}$ given in the second column, we obtain a locally fourth- (globally third-) order approximation.

We have

$$T_1^{(1)} = \frac{x_2 T_0^{(1)} - x_1 T_0^{(2)}}{x_2 - x_1} = \frac{\frac{1}{3}\Delta t u_2(t_1) - \frac{1}{2}\Delta t u_3(t_1)}{\frac{1}{3}\Delta t - \frac{1}{2}\Delta t} = 3u_3(t_1) - 2u_2(t_1)$$

and thus

$$\begin{aligned}\tilde{u}(t_1) = T_2^{(0)} &= \frac{x_2 T_1^{(0)} - x_0 T_1^{(1)}}{x_2 - x_0} = \frac{\frac{1}{3}\Delta t\,(2u_2(t_1) - u_1(t_1)) - \Delta t\,(3u_3(t_1) - 2u_2(t_1))}{\frac{1}{3}\Delta t - \Delta t} \\ &= \tfrac{1}{2}u_1(t_1) - 4u_2(t_1) + \tfrac{9}{2}u_3(t_1). \end{aligned} \tag{25}$$

We have

$$\begin{aligned}\widetilde{\mathcal{F}}(\Delta t) = {} & \tfrac{1}{2}\left(I + E_1\Delta t + E_2\Delta t^2 + E_3\Delta t^3 + \mathrm{O}(\Delta t^4)\right) \\ & - 4\left(I + E_1\tfrac{1}{2}\Delta t + E_2\tfrac{1}{4}\Delta t^2 + E_3\tfrac{1}{8}\Delta t^3 + \mathrm{O}(\Delta t^4)\right) \\ & \times \left(I + E_1\tfrac{1}{2}\Delta t + E_2\tfrac{1}{4}\Delta t^2 + E_3\tfrac{1}{8}\Delta t^3 + \mathrm{O}(\Delta t^4)\right) \\ & + \tfrac{9}{2}\left(I + E_1\tfrac{1}{3}\Delta t + E_2\tfrac{1}{9}\Delta t^2 + E_3\tfrac{1}{27}\Delta t^3 + \mathrm{O}(\Delta t^4)\right) \\ & \times \left(I + E_1\tfrac{1}{3}\Delta t + E_2\tfrac{1}{9}\Delta t^2 + E_3\tfrac{1}{27}\Delta t^3 + \mathrm{O}(\Delta t^4)\right) \\ & \times \left(I + E_1\tfrac{1}{3}\Delta t + E_2\tfrac{1}{9}\Delta t^2 + E_3\tfrac{1}{27}\Delta t^3 + \mathrm{O}(\Delta t^4)\right) \\ = {} & \tfrac{1}{2}I + \tfrac{1}{2}E_1\Delta t + \tfrac{1}{2}E_2\Delta t^2 + \tfrac{1}{2}E_3\Delta t^3 - 4I - 4E_1\Delta t - 2E_2\Delta t^2 - E_1^2\Delta t^2 - E_3\Delta t^3 \\ & - \tfrac{1}{2}E_1E_2\Delta t^3 - \tfrac{1}{2}E_2E_1\Delta t^3 + \tfrac{9}{2}I + \tfrac{9}{2}E_1\Delta t + \tfrac{3}{2}E_2\Delta t^2 + \tfrac{3}{2}E_1^2\Delta t^2 + \tfrac{1}{2}E_3\Delta t^3 \\ & + \tfrac{1}{2}E_1E_2\Delta t^3 + \tfrac{1}{2}E_2E_1\Delta t^3 + \tfrac{1}{6}E_1^3\Delta t^3 + \mathrm{O}(\Delta t^4) \\ = {} & I + E_1\Delta t + \tfrac{1}{2}E_1^2\Delta t^2 + \tfrac{1}{6}E_1^3\Delta t^3 + \mathrm{O}(\Delta t^4). \end{aligned}$$

and thus, we obtain a locally fourth- (globally third-) order approximation.

The same is true if we consider the asymptotic expansions of the error and we have

$$\tilde{u}(t_1) - U(t_1) = \tfrac{1}{2}(u_1(t_1) - U(t_1)) - 4(u_2(t_1) - U(t_1)) + \tfrac{9}{2}(u_3(t_1) - U(t_1)) = \mathrm{O}(\Delta t^4).$$

The problem is now to understand why our sequence which does not have the form (16), when we set (23), gives

$$T_k^{(0)} - S = \mathrm{O}(\Delta t^{k+2}) \quad \text{for } k = 0, 1, 2.$$

To explain this fact it is necessary to express our sequence in a different form. In fact, we may transform it to obtain the form (16) and we have, for $n = 0, 1, 2$

$$u_{n+1}(t_1) - U(t_1) = \frac{1}{n+1}\left[(2E_2 - E_1^2)\tfrac{1}{2}\Delta t^2 + (E_1E_2 + E_2E_1 - E_1^3)\tfrac{1}{2}\Delta t^3 + \mathrm{O}(\Delta t^4)\right] + \frac{1}{(n+1)^2}\left[(6E_3 - E_1^3)\tfrac{1}{6}\Delta t^3 - (E_1E_2 + E_2E_1 - E_1^3)\tfrac{1}{2}\Delta t^3 + \mathrm{O}(\Delta t^4)\right] + \cdots,$$

where in fact the role of x_n is assumed by $1/(n+1)$ and the constants $a_1, a_2, \ldots$ are formed by a sum of terms of an order, respectively, greater or equal to $\Delta t^2, \Delta t^3, \ldots$ and thus we are exactly in the case (15).

In order to construct a third-order scheme over a finite interval of time there are two possible strategies. The first one is to solve the problem three times with a first-order scheme over the interval $[0, T]$, using the different step size Δt, $\frac{1}{2}\Delta t$ and $\frac{1}{3}\Delta t$ and then apply extrapolation at the common points $t_n = n\Delta t$ (for $n = 1, \ldots, T/\Delta t$). This can be, in practice, quite inaccurate. In fact, after a long time the global error can be quite large, and one is not close to the asymptotic regime, therefore the expected third-order accuracy occurs only after short times, or for extremely short time step.

A second possibility (which is the one we use in this paper) is to use Richardson extrapolation to derive a third-order step and to use it to update the numerical solution when going from t_n to t_{n+1}. The use of the extrapolated result at the beginning of the next step is the standard way it is used in time-dependent ODEs, and was first studied in detail by Gragg [6].

In our case, this is performed by applying (25) but using the extrapolated solution $\tilde{u}(t_n)$ obtained at the time t_n. That is, we compute

$$\begin{aligned} u_1(t_{n+1}) &= \mathscr{T}(\Delta t)\tilde{u}(t_n), \\ u_2(t_{n+1}) &= \mathscr{T}\left(\tfrac{1}{2}\Delta t\right)^2 \tilde{u}(t_n), \\ u_3(t_{n+1}) &= \mathscr{T}\left(\tfrac{1}{3}\Delta t\right)^3 \tilde{u}(t_n) \end{aligned} \tag{26}$$

and therefore

$$\tilde{u}(t_{n+1}) = \tfrac{1}{2}u_1(t_{n+1}) - 4u_2(t_{n+1}) + \tfrac{9}{2}u_3(t_{n+1}). \tag{27}$$

In this way we use a locally fourth-order approximation of the solution instead of a solution of a local second order.

A third-order scheme for system (17) is obtained using third-order Richardson extrapolation (26, 27), and the first-order operator (in time) $\mathscr{T}(\Delta t)$, defined by Eq. (19).

The operator $\mathscr{T}(\Delta t)$ such that $U^1 = \mathscr{T}(\Delta t)U^0$ relative to system (2) is defined as in (11):

$$\tilde{U} = U^0 - \frac{\Delta t}{\varepsilon} R(\tilde{U}),$$
$$U^1 = \tilde{U} - \Delta t \mathscr{D} F\tilde{U},$$

where $\mathscr{D}$ is a third-order accurate discrete convection operator, such as the one used in the piecewise parabolic method (PPM) or a scheme based on third-order upwind approximation of derivative, Eq. (8).

The schemes obtained in this way will be referred, respectively, as PPM-R3 and P3-R3.

5. Stability considerations

In this section we make some remarks about the stability of the various schemes proposed.

Let us consider first the schemes based on extrapolation and applied to the linear system (17). We assume that U is a scalar and A and B are two complex numbers. Here A represents the convection operator, and B represents the relaxation rate.

If $A = 0$ then the schemes are implicit schemes for the equation

$$\frac{\partial U}{\partial t} = BU.$$

Such schemes for ordinary differential equations, based on implicit Euler and extrapolation methods, have been studied in the literature. First- and second-order schemes are A-stable, and third-order schemes are A-α stable [5].

If $A \neq 0$ then these results are generalized as follows. Application of first-, second- and third-order schemes give

$$U^{n+1} = P_k(z_1, z_2)U^n,$$

where $U^n \approx U(n\Delta t), z_1 = A\Delta t$, $z_2 = B\Delta t$, and

$$P_1(z_1, z_2) = \frac{1 + z_1}{1 - z_2},$$

$$P_2(z_1, z_2) = 2\left(\frac{1 + \frac{1}{2}z_1}{1 - \frac{1}{2}z_2}\right)^2 - \frac{1 + z_1}{1 - z_2},$$

$$P_3(z_1, z_2) = \frac{9}{2}\left(\frac{1 + \frac{1}{3}z_1}{1 - \frac{1}{3}z_2}\right)^3 - 4\left(\frac{1 + \frac{1}{2}z_1}{1 - \frac{1}{2}z_2}\right)^2 + \frac{1}{2}\frac{1 + z_1}{1 - z_2}.$$

The following propositions generalize the stability results known for ordinary differential equations.

Proposition 5.1. *If* $|1 + z_1| < 1$ *and* $\Re(z_2) < 0$ *then* $|P_1(z_1, z_2)| < 1$.

The proof is straightforward.

Proposition 5.2. *If* $|1+z_1| < 1$ *and* $\Re(z_2) < 0$ *then* $|P_2(z_1,z_2)| < 1$.

Proof. Let

$$S_2 = \{(z_1,z_2) \in \mathbb{C}^2\colon\ |z_1+1| \leqslant 1, \Re(z_2) \leqslant 0\}.$$

For the maximum principle of analytic functions, the maximum of $|P_2(z_1,z_2)|$ in S_2 is reached on the boundary, and therefore it is a point of the form $z_1 = e^{i\theta} - 1$, $z_2 = iy$. It is

$$|P_2(e^{i\theta}-1, iy)|^2 = \frac{\mathcal{N}}{\mathcal{D}},$$

and

$$F = \mathcal{N} - \mathcal{D} = 16(1+y^2)(\cos^2\theta - 1) + 8y^2(\cos\theta - 1) - 8y^4 - y^6 \leqslant 0,$$

therefore

$$|P_2(z_1,z_2)| \leqslant 1 \quad \forall (z_1,z_2) \in S_2. \qquad \square$$

Using a similar technique the following can be proved.

Proposition 5.3. *If* $|z_1+1| \leqslant 1$, *then the method is A-α stable in* z_2, *i.e. there exists* $m > 0$ *such that* $|P_3(e^{i\theta}-1, -m|y| + iy)| \leqslant 1$, $\forall y \in \mathbb{R}$, $\theta \in [0, 2\pi]$.

The proof is omitted.

A particular care must be used when time discretization is applied to Eqs. (7) and (8). By performing von Neumann stability analysis, one can show that the amplification factor for forward Euler scheme applied to (7) and (8) is greater than one for any value of the Courant number $c = a\,\Delta t/\Delta x$. However, if one uses second-order Richardson extrapolation on Eq. (7), then the amplification factor is smaller than one if $c < \frac{1}{2}$, and using third-order extrapolation on Eq. (8), stability is guaranteed if $c < 0.9$. We omit the proof here.

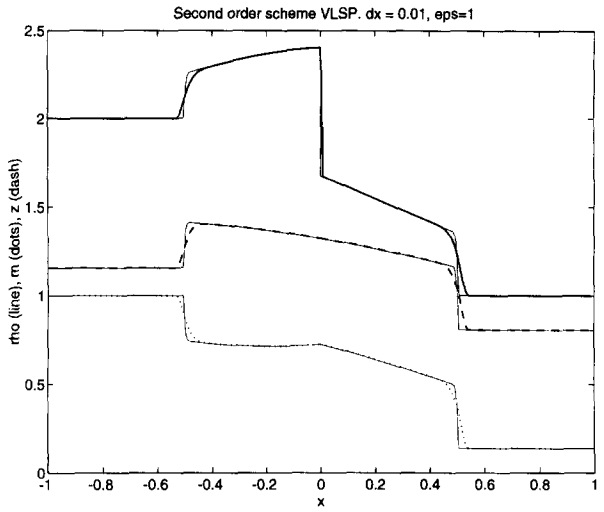
6. Numerical results

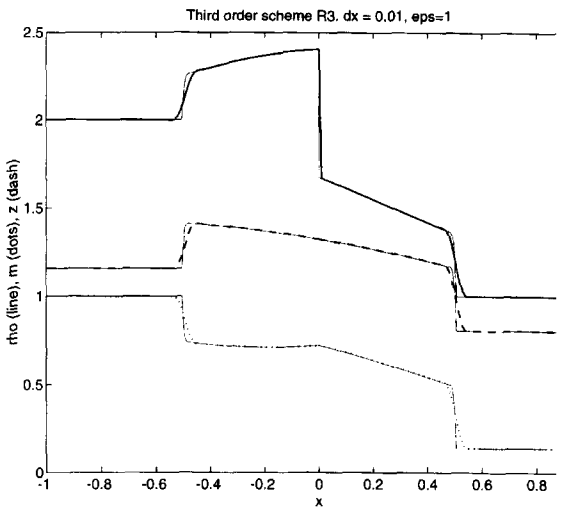
In this section we present some numerical results in which we compare the third-order schemes based on Richardson extrapolation (PPM-R3 and P3-R3) with the previous second-order scheme VLSP (equation (12)).

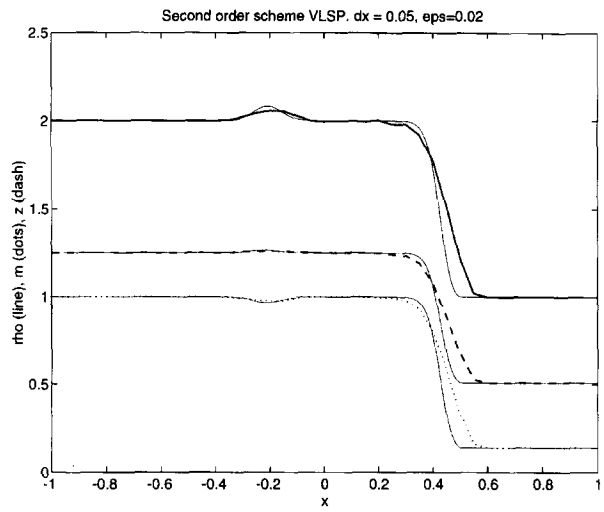
We consider two different test problems. The first is a Riemann problem with the following initial data:

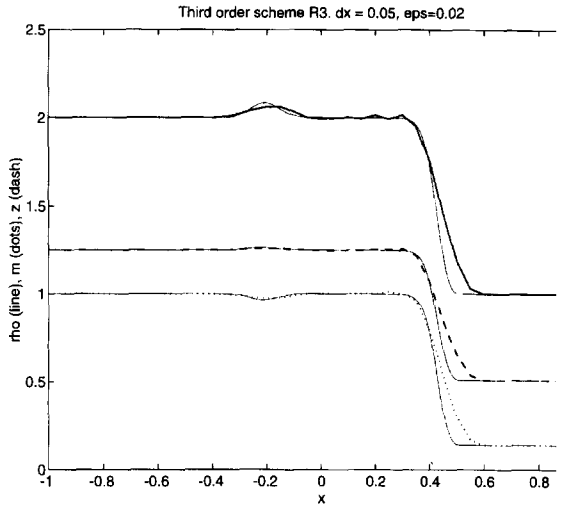
$$\begin{aligned} &\varrho = 2, \quad m = 1, \quad z = 1 \quad \text{for } x < 0, \\ &\varrho = 1, \quad m = 0.13962, \quad z = 1 \quad \text{for } x > 0. \end{aligned} \tag{28}$$

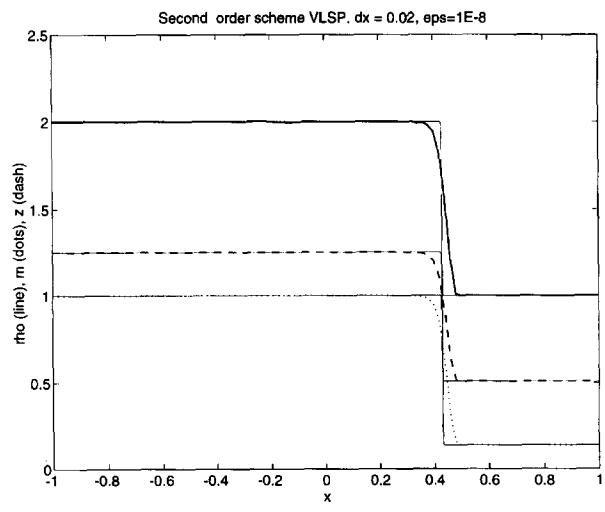
We integrate system (1) over the domain $[-1,1]$ with reflecting boundary conditions. The "exact solution" is obtained using the second-order scheme and a fine grid with $\Delta x = 0.001$.

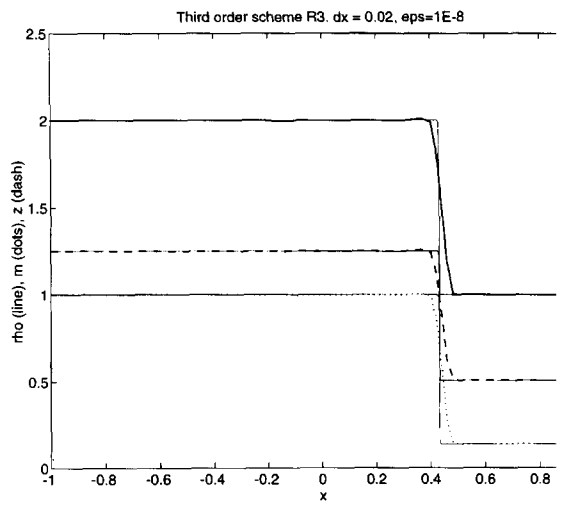

Second order scheme VLSP. dx = 0.01, eps=1
rho (line), m (dots), z (dash)
x


Third order scheme R3. dx = 0.01, eps=1
rho (line), m (dots), z (dash)
x


Second order scheme VLSP. dx = 0.05, eps=0.02
rho (line), m (dots), z (dash)
x


Third order scheme R3. dx = 0.05, eps=0.02
rho (line), m (dots), z (dash)
x


Second order scheme VLSP. dx = 0.02, eps=1E-8
rho (line), m (dots), z (dash)
x


Third order scheme R3. dx = 0.02, eps=1E-8
rho (line), m (dots), z (dash)
x

The numerical results for ϱ (solid line), m (dotted line) and z (dashed line) are reported in Fig. 1, together with the exact solution (thin line). The pictures on the left represent the result obtained by scheme VLSP, while the figures on the right are obtained by using scheme PPM-R3, based on Richardson extrapolation. The same initial data have been used with three different values of ε (top to bottom: $\varepsilon = 1$, 0.02, 10^{-8}, respectively).

As it is evident from the figures, both schemes give an accurate solution of the shock profile for every regime of ε. The solution obtained by scheme PPM-R3 is slightly sharper than the one obtained by VLSP.

Next we perform the numerical convergence study. We consider an initial value problem with periodic boundary conditions, such that the solution is smooth in a time interval $[0, T]$ for any value of the parameter ε. We compute the error at time T by differencing, i.e. by comparing the result obtained with a given grid $(\Delta x, \Delta t)$ with the one obtained with the grid $(\frac{1}{2}\Delta x, \frac{1}{2}\Delta t)$.

The goal of the test is to perform a numerical study of the convergence rate for a wide range of ε, and check whether the convergence is uniform in ε. The test problem is given by equations (1) with periodic boundary conditions: $s(x + L, t) = s(x, t)$ with $s = f, g, h$. The initial data is given by

$$\varrho(x,0) = 1 + a_\varrho \sin\frac{2\pi x}{L}, \qquad u(x,0) = \frac{1}{2} + a_u \sin\frac{2\pi x}{L},$$

$$m(x,0) = \varrho(x,0)u(x,0), \qquad z(x,0) = z_E(\varrho(x,0), m(x,0))\theta_M,$$

where θ_M is a real parameter and $z_E(\rho, m) = (\rho^2 + m^2)/(2\rho)$ is the equilibrium distribution. If $\theta_M = 1$ then the initial condition is a local Maxwellian, otherwise it is not. If $\theta_M \neq 1$, $\varepsilon \ll 1$, there is an initial layer. The system is integrated for $t \in [0, T]$. The values of the parameters used in the computations are:

$$L = 20, \quad T = 32, \quad a_\varrho = 0.3, \quad a_u = 0.1, \quad \theta_M = 0.2.$$

The values of Δx used in the computations are

$$\Delta x = 0.8, 0.4, 0.2, 0.1, 0.05.$$

The time step is chosen in such a way that CFL condition is satisfied: $\Delta t = \frac{1}{2}\Delta x$. The convergence rate is computed from the error according to the formula:

$$\text{convergence rate}_i = \frac{\log(\text{error}_i/\text{error}_{i+1})}{\log(\Delta x_i/\Delta x_{i+1})},$$

where error_i is obtained by comparing the solution obtained with Δx_i to that obtained with Δx_{i+1}. The errors and convergence rate are computed and plotted as function of ε. For each value of ε, five runs have been done for five different values of Δx, resulting in four error curves and three curves of convergence rate.

Three schemes have been compared:

Fig. 1. The numerical solutions of ϱ (solid line), m (dotted line) and z (dashed line) at $t = 0.5$ in $x \in [-1, 1]$ for initial data (28) by VLSP (left column) and PPM-R3 (right column). Exact solution in thin line. The values of ε and Δx are (top to bottom) $\varepsilon = 1$, $\Delta x = 0.01$, $\varepsilon = 0.02$, $\Delta x = 0.05$, $\varepsilon = 10^{-8}$, $\Delta x = 0.02$.

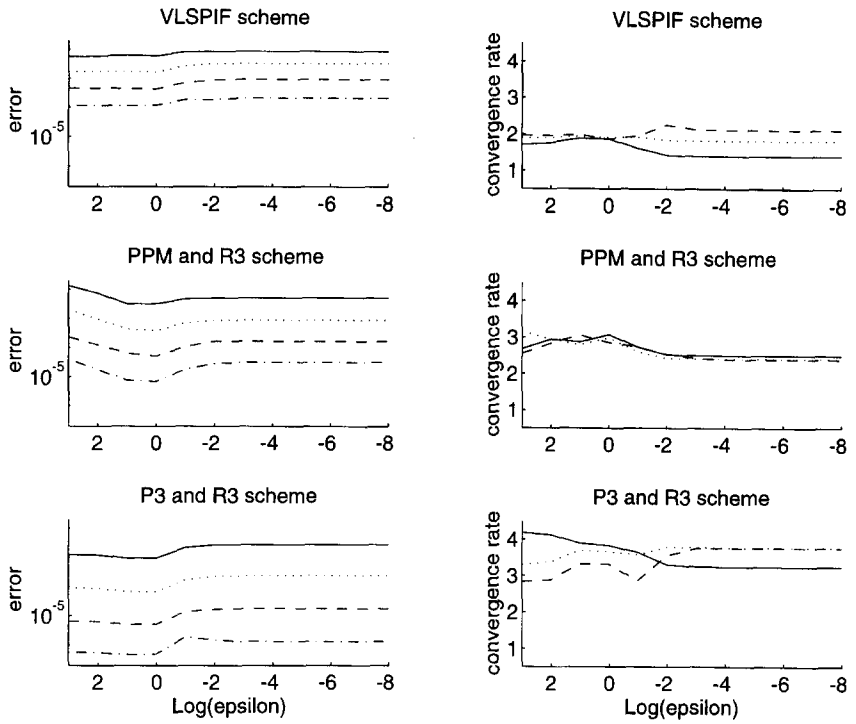


Fig. 2. Numerical convergence study. L_1 relative error in ϱ (left column) and convergence rate (right column) versus ε for several values of the step size. Schemes: VLSPIF (top), PPM-R3 (middle), P3-R3 (bottom).

VLSPIF is the splitting scheme (12), where the convection step is solved by the second-order van Leer flux scheme [4, 7], and a Richardson extrapolation ($T_1^{(0)}$) is used for the first step in order to avoid the problem of the initial layer (VLSPIF means "VLSP scheme with Initial layer Fix").

PPM-R3 is similar to the previous one, but a third-order Richardson extrapolation is used for the time discretization.

P3-R3 is obtained by third-order upwind approximation of space derivative (8), and third-order Richardson extrapolation.

The results of the convergence study are summarized in Fig. 2. It is evident that the use of a third-order space discretization improves the accuracy, but an improvement in the rate of convergence is observed only if a third-order time discretization is used.

The scheme PPM-R3 is more accurate than the previous scheme VLSP, but it does not reach uniform third-order accuracy. This behavior is explained as follows. The convection step is based on a cell-average approximation of the field, while the collision step is based on a pointwise, cell-centered approximation. Now the cell average of the field is a second-order approximation of the point value of the field in the middle of the cell, and this gives a limit to the space accuracy available with such mixed scheme.

In order to improve the accuracy, the scheme should be based on cell-averaged or cell-centered values on both steps. Of the two approaches, we chose the second one, which is easier to implement and provides more efficient schemes.

Scheme P3-R3 is the most accurate in the whole range of ε, and is at least third order. We remark that the scheme presented here is constructed without flux limiter, since our main concern is to study the accuracy of the scheme on smooth solutions. It is possible to include flux limiter and make the scheme suitable for shock capturing.

Acknowledgements

We would like to thank Shi Jin for helpful discussions. We also thank the referee for the comments which helped us to improve the quality of the paper.

References

[1] L.V. Ahlfors *Complex Analysis* (McGraw-Hill, New York, 3rd ed., 1979) 134.
[2] C. Brezinski and M. Redivo-Zaglia, *Extrapolation Methods. Theory and Practice* (North-Holland, Amsterdam, 1991).
[3] J.E. Broadwell, Shock structure in a simple discrete velocity gas, *Phys. Fluids* **7** (1964) 1013–1037.
[4] R.E. Caflisch, Shi Jin and G. Russo, Uniformly accurate schemes for hyperbolic systems with relaxation, *SIAM J. Numer. Anal.*, to appear.
[5] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems* (Springer, New York, 1991) 46.
[6] W.B. Gragg, On extrapolation algorithms for ordinary initial value problems, *SIAM J. Numer. Anal.* **2** (1965) 384–403.
[7] R.J. LeVeque, *Numerical Methods for Conservation Laws* (Birkhäuser, Basel, 1994).
[8] T.P. Liu, Hyperbolic conservation laws with relaxation, *Commun. Math. Phys.* **108** (1987) 153–175.
[9] G. Marchouk and V. Shaydurov, *Raffinement des Solutions des Schémas aux Différences* (Mir, Moscow, 1983).
[10] G.I. Marchuk and V.V. Shaidurov, *Difference Methods and their Extrapolation* (Springer, Berlin, 1983).
[11] R.B. Pember, Numerical methods for hyperbolic conservation laws with stiff relaxation I. Spurious solutions, *SIAM J. Appl. Math.* **53** (1993) 1293–1330.
[12] R.B. Pember, Numerical methods for hyperbolic conservation laws with stiff relaxation II. Higher order Godunov methods, *SIAM J. Sci. Comput.* **14** (1993) 824–859.
[13] C.W. Shu and S. Osher, Efficient implementation of essentially non-oscillatory shock capturing schemes, II, *J. Comput. Phys.* **83** (1989) 32–78.
[14] G.B. Whitham, *Linear and Nonlinear Waves* (Wiley, New York, 1974).