Available online at www.sciencedirect.com
ScienceDirect
Journal of Computational Physics 220 (2006) 511–531
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Explicit solutions to a convection-reaction equation and defects of numerical schemes ☆

Youngsoo Ha, Yong Jung Kim *

Division of Applied Mathemtics, KAIST (Korea Advanced Institute of Science and Technology), 373-1 Gusong-dong, Yusong-gu, Taejon 305-701, South Korea

Received 23 September 2005; received in revised form 13 July 2006; accepted 17 July 2006
Available online 6 September 2006

Abstract

We develop a theoretical tool to examine the properties of numerical schemes for advection equations. To magnify the defects of a scheme we consider a convection-reaction equation

$$u_t + (|u|^q/q)_x = u, \quad u, x \in \mathbf{R}, \quad t \in \mathbf{R}^+, \quad q > 1.$$

It is shown that, if a numerical scheme for the advection part is performed with a splitting method, the intrinsic properties of the scheme are magnified and observed easily. From this test we observe that numerical solutions based on the Lax–Friedrichs, the MacCormack and the Lax–Wendroff break down easily. These quite unexpected results indicate that certain undesirable defects of a scheme may grow and destroy the numerical solution completely and hence one need to pay extra caution to deal with reaction dominant systems. On the other hand, some other schemes including WENO, NT and Godunov are more stable and one can obtain more detailed features of them using the test. This phenomenon is also similarly observed under other methods for the reaction part.

Keywords: Numerical schemes; Roll waves; WENO; Godunov; Central schemes; Advection equations; Convection-reaction; Hyperbolic conservation laws

1. Introduction

The scientific computation for the solutions to systems of hyperbolic conservation laws has been successful thanks to the development of accurate numerical schemes (e.g., see [2,11,19,20]). The unfortunate situation we still have here is that there is no meaningful progress in the analysis of such schemes due to their complexity and it is hard to find any useful error estimate. Furthermore, each scheme has features quite its own and

☆ This work was supported in part by the second stage of BK21 project.
* Corresponding author. Tel.: +82 42 869 2739; fax: +82 42 869 5710.
E-mail addresses: young@amath.kaist.ac.kr (Y. Ha), ykim@amath.kaist.ac.kr (Y.J. Kim).

0021-9991/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcp.2006.07.018

produces numerical results that require proper interpretations. For example, Figs. 4.7 and 4.8 in [14] show quite different characters of each scheme considered and it is hard to decide which one is more physically meaningful. Therefore, it is desirable to know the qualitative properties of schemes since one is always forced to choose a proper scheme to a specific problem under consideration.

The main goal in this paper is to survey the properties of numerical schemes which are widely used for hyperbolic conservation laws. Our method starts with a Cauchy problem for a convection-reaction equation

$$u_t + uu_x = u, \quad u(x,0) = u_0(x) \in L^1(\mathbf{R}). \tag{1.1}$$

This equation models the roll wave [6] and several analytic properties have been studied (e.g., see [7,16]). We consider a general case (2.1) under a convex power law (2.2). In Section 2, we derive two kinds of explicit solutions of the problem, (2.8) and (2.9), and compare computed solutions to these exact ones. Here, the convection is nonlinear and the linear reaction term produces an exponential growth. Therefore, the source becomes stiffer exponentially fast.

There are several ways to handle convection-reaction equation numerically. A typical way is a fractional step splitting method (Section 3). For the convection part we employ various schemes for hyperbolic conservation laws widely used nowadays and our main interest is to survey their properties. For the reaction part one can easily obtain an exact solver due to the simplicity of the source term. The main feature of this splitting method with the exact solver is that the intrinsic properties of the scheme for the convection part are magnified exponentially. Proposition 1 indicates that all the properties one may observe from numerical computations for (2.1) are simply the magnification of corresponding properties of schemes for convection part. Therefore, this approach may serve as a microscope to examine properties of numerical schemes.

In Section 4, we observe several interesting blowups of numerical solutions based on Lax–Friedrichs, MacCormack and Lax–Wendroff. As discussed the origin of these undesirable behaviors are from these three well-known schemes. One may ask if a non-splitting method may neutralize the undesirable behaviors of these schemes instead of simply magnifying them. To answer the question we test several more methods such as a non-splitting direct method, a semi-implicit method [13], the Runge–Kutta second-order and fourth-order methods to handle the source term in Section 7. However, tests show that there is no meaningful difference in most of the cases. The authors are not sure if there is a method to handle the source term in a way that these undesirable properties are neutralized. It seems more realistic to modify each scheme to cure the symptoms observed.

The rest of this paper is organized as follows. First we construct two kinds of explicit solutions in Section 2, which are used to test numerical solutions. It is shown in Section 3 that the behavior of each numerical scheme is simply magnified by solving (2.1) with the exact solver for the reaction part, Proposition 1. Properties of three well-known schemes, Lax–Friedrichs, MacCormack and Lax–Wendroff, are tested in Section 4, which show interesting blowups (see Figs. 1–5). In Section 5, three fully discrete schemes are tested, the first and the second-order Godunov schemes and the NT scheme, while Section 6 is devoted for two semi discrete schemes, the KNP and the WENO schemes. These schemes do not show such bizarre behaviors observed previously and hence we could survey their properties in detail. The first test is for the accuracy of the shock location and the second one is to compute the roll wave given explicitly by (2.9), which shows clearly how does a scheme approximate the rarefaction wave.

In Appendix, Section 7, the properties of schemes considered in the paper are compared. It is hard to say one scheme is better than the other. Some scheme may find the shock location very well and, however, it may have certain undesirable behavior. The second-order Godunov scheme provides the most accurate shock locations, Figs. 12 and 13. However, it shows a very unique behavior for the roll wave test. For $q = 1.5$, it is the only one that has error decreasing in x and, for the case $q = 2$, it is the only one that has negative error on the interval $(-1, 0)$, Fig. 14. These indicate that the convection is slightly over estimated in the case and this may cause problems in certain cases.

2. Exact solutions

We consider the entropy solution to a scalar conservation law with a linear source term

$$u_t + f(u)_x = u, \quad u(x,0) = u_0(x) \in L^1(\mathbf{R}), \quad x, u \in \mathbf{R}, \quad t > 0, \tag{2.1}$$

where the flux is given by the convex power law

$$f(u) = \frac{1}{q}|u|^q, \quad q > 1. \tag{2.2}$$

Since the flux considered is convex, $f''(u) \geqslant 0$, the entropy solution is simply the one that satisfies

$$u(x-) \geqslant u(x+), \quad \text{for all } x \in \mathbf{R}. \tag{2.3}$$

Under the change of variables

$$w = \mathrm{e}^{-t/q}u, \quad \xi = \mathrm{e}^{(1-q)t/q}x, \tag{2.4}$$

one can easily check that (2.1) is transformed to

$$w_t + \frac{1}{q}(|w|^q - (q-1)\xi w)_\xi = 0, \quad w(\xi, 0) = u_0(\xi). \tag{2.5}$$

Consider a time independent function

$$\mathscr{W}_{a,b}(\xi) = \begin{cases} g(\xi), & -a < \xi < b, \\ 0, & \text{otherwise}, \end{cases} \tag{2.6}$$

where a, $b \geqslant 0$ and the function $g(\xi)$ is the rarefaction wave profile given by

$$g(\xi) = \operatorname{sign}(\xi)\sqrt[q-1]{(q-1)|\xi|}. \tag{2.7}$$

Then $\mathscr{W}_{a,b}$ is clearly a piecewise smooth function that satisfies the entropy condition (2.3) everywhere and Eq. (2.5) piecewise. Furthermore, one can easily check that the discontinuities at $x = -a$ and $x = b$ are stationary from the Rankine–Hugoniot jump condition. Therefore, $w(\xi, t) = \mathscr{W}_{a,b}(\xi)$ is a steady solution of (2.5). If we return to the original variables, then we obtain our first explicit solution, which is a time dependent N-wave:

$$\mathscr{N}_{a,b}(x,t) = \begin{cases} g(x), & -a\mathrm{e}^{(q-1)t/q} < x < b\mathrm{e}^{(q-1)t/q}, \\ 0, & \text{otherwise}. \end{cases} \tag{2.8}$$

The next explicit solution is a roll wave which is given by

$$\mathscr{R}_1(x) = \begin{cases} g(x+1), & -1 < x < 0, \\ g(x-1), & 0 < x < 1, \\ 0, & \text{otherwise}. \end{cases} \tag{2.9}$$

Then, one can easily check that it is a piecewise smooth solution with zero shock speed satisfying the entropy condition (2.3).

Notice that there is a sensitivity issue related to steady states. To see such a structure consider the total mass $M(t) = \int u(x,t)\,\mathrm{d}x$. Then, its derivative is

$$\frac{\mathrm{d}}{\mathrm{d}t}M(t) = \frac{\mathrm{d}}{\mathrm{d}t}\int u\,\mathrm{d}x = \int u_t\,\mathrm{d}x = -\int f(u)_x\,\mathrm{d}x + \int u\,\mathrm{d}x = M(t).$$

Therefore, the total mass M grows exponentially,

$$M(t) = M(0)\mathrm{e}^t \tag{2.10}$$

and hence all the integrable steady states should have zero total mass. The roll wave in (2.9) is the case. However, even a small rounding off error will grow exponentially and the solution will diverge eventually. Therefore, a numerical computation for the steady state is meaningful only for a certain period of time depending on the precision of the computation. In the rest of this paper these explicit solutions are used to test properties of several numerical schemes.

Consider another set of variables:

$$v = \mathrm{e}^{-t}u, \quad s = \frac{1}{q-1}\left(\mathrm{e}^{(q-1)t} - 1\right). \tag{2.11}$$

Then, (2.1) is transformed to the usual scalar conservation law:

$$v_s + f(v)_x = 0, \quad v(x,0) = u_0(x). \tag{2.12}$$

In the following section, we will observe that the splitting method with the exact solver is a computational version of the change of variables (2.11).

3. A splitting method and a microscopic view of schemes

In numerical computations, a convection-reaction equation such as (2.1) is usually treated by a fractional step splitting method, in which one alternates between solving the convection equation

$$u_t + f(u)_x = 0 \tag{3.1}$$

and the ordinary differential equation

$$u_t = u \tag{3.2}$$

in each time step. First, we introduce notations. Consider a fixed width $\Delta x > 0$ and uniform mesh points $x_{j+1/2} = (j+1/2)\Delta x$, $j \in \mathbf{Z}$. Since the actual wave speed of an N-wave type solution increases exponentially in time, the time step $\Delta t^n := t^{n+1} - t^n$ and time mesh $t^n = \sum_{k=0}^{n-1}\Delta t^k$ are decided by setting the CFL number to a constant $0 < \nu \leqslant 1$, i.e.,

$$\frac{\Delta t^n}{\Delta x} f'(\tilde{u}_n) = \nu, \tag{3.3}$$

where $\Delta x/\Delta t^n$ is the numerical wave speed and $\tilde{u}_n = \max_x |u(x,t^n)|$. In the case that the exact solution is unknown this maximum is usually replaced by the maximum of its approximation $\max_j |U_j^n|$. However, in our case the exact solution is given explicitly by (2.8) and (2.9) and hence the maximum $\tilde{u}_n = \max_x |u(x,t^n)|$ is explicit. We view the approximation U_j^n as the cell average of the true solution, i.e.,

$$U_j^n \simeq \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u(x,t^n)\,\mathrm{d}x. \tag{3.4}$$

We also view $u_{j+1/2}^n$ as the approximation of $u(x,t)$ at the interface $x_{j+1/2}$ of each cell. In a conservative numerical scheme the interface $u_{j+1/2}^n$ is approximated from its neighboring cell averages and we may set

$$u_{j+1/2}^n \equiv I(U_{j-p}^n, \cdots, U_{j+q}^n) \sim u(x_{j+1/2}, t), \quad t^n \leqslant t \leqslant t^{n+1}. \tag{3.5}$$

Then, after integrating (3.1) over the mesh $[x_{j-1/2}, x_{j+1/2}]$, one obtains

$$\frac{\partial}{\partial t}(U_j^n) = \frac{f(u_{j-1/2}^n) - f(u_{j+1/2}^n)}{\Delta x} =: \mathscr{L}(U^n; j). \tag{3.6}$$

Note that in many numerical schemes approximations of the flux at the interface, $f(u_{j+1/2}^n)$, are given instead of the interface approximation, $u_{j+1/2}^n$, itself. In either cases the numerical scheme is based on (3.6) and the operator $\mathscr{L}$ is well defined.

For the time discretization we first employ the forward time difference,

$$\overline{U}_j^{n+1} = U_j^n + \frac{\Delta t^n}{\Delta x}(f(u_{j-1/2}^n) - f(u_{j+1/2}^n)) (= U_j^n + \Delta t^n \mathscr{L}(U^n; j)), \tag{3.7}$$

which computes the first part (3.1). On the other hand, the ordinary differential equation (3.2) can be exactly solved and we obtain an exact solver for the reaction part

$$U_j^{n+1} = \mathrm{e}^{\Delta t^n}\overline{U}_j^{n+1}. \tag{3.8}$$

Combining these two steps we obtain a fully discrete numerical scheme

$$U_j^{n+1} = \mathrm{e}^{\Delta t^n}\left(U_j^n - \frac{\Delta t^n}{\Delta x}\left(f(u_{j+1/2}^n) - f(u_{j-1/2}^n)\right)\right). \tag{3.9}$$

Notice that the second equation (3.2) is treated exactly and hence the numerical error will be caused from the first step (3.7) only. Since the focus of this paper is to study the behavior of the numerical schemes for advection equations, this approach serves our goal well.

Finally, we provide a proposition which indicates that any unexpected behavior of the numerical approximation based on (3.9) is just a mirror image of such a behavior of the numerical scheme (3.7) for the convection equation (3.1).

Proposition 1. *Let $V_j^n \sim v(x, s^n)$ be the approximation of the solution to the homogeneous problem (2.12) satisfying*

$$V_j^{n+1} = V_j^n - \frac{\Delta s^n}{\Delta x}\left(f(v_{j+1/2}^n) - f(v_{j-1/2}^n)\right), \quad \frac{\Delta s^n}{\Delta x} f'(\max_x |v(x, s^n)|) = \nu,$$

where $\Delta s^n = s^{n+1} - s^n$, $0 < \nu < 1$ is a fixed CFL number, and v is the exact solution of (2.12). If the interface approximation (3.5) satisfies

$$I(CU_{j-p}, \cdots, CU_{j+q}) = CI(U_{j-p}, \cdots, U_{j+q}), \quad C > 0, \tag{3.10}$$

then the approximation $U_j^n \sim u(x, t^n)$ given by (3.9) with (3.3) satisfy

$$U_j^n = \mathrm{e}^{t^n} V_j^n \quad \text{for any } n \geqslant 0. \tag{3.11}$$

Proof. We use inductive arguments. The relation (3.11) holds for $n = 0$ since U_j^0 and V_j^0 are the initial discretization of the same initial value $u_0(x)$. Now we show (3.11) for $n+1$ assuming that it holds for n. Let $U_j^n = CV_j^n$ by setting $C = \mathrm{e}^{t^n}$. Then the relation between Δt^n and Δs^n is given by

$$\Delta t^n = \frac{\nu \Delta x}{f'(\max_x |u(x, t^n)|)} = \frac{\nu \Delta x}{f'(\max_x |Cv(x, t^n)|)} = \frac{\nu \Delta x}{C^{q-1} f'(\max_x |v(x, t^n)|)} = \frac{\Delta s^n}{C^{q-1}},$$

where ν is the fixed CFL number. Then under the assumption on the interface approximation (3.10) the numerical scheme (3.9) becomes

$$\begin{aligned} U_j^{n+1} &= \mathrm{e}^{\Delta t^n}\left(CV_j^n - \frac{\Delta s^n}{C^{q-1}\Delta x}\left(f(Cv_{j+1/2}^n) - f(Cv_{j-1/2}^n)\right)\right) = \mathrm{e}^{\Delta t^n} C\left(V_j^n - \frac{\Delta t^n}{\Delta x}\left(f(v_{j+1/2}^n) - f(v_{j-1/2}^n)\right)\right) \\ &= \mathrm{e}^{t^{n+1}} V_j^{n+1}, \end{aligned} \tag{3.12}$$

which implies (3.11) for $n+1$. □

Notice that the relation between V_j^n and U_j^n in (3.11) exactly reflects the change of variable $u = \mathrm{e}^t v$ in (2.11). The second part of the change of variable for the time variable is not immediate. However, assuming Δt^ns are constant and small, $\Delta t^n = k \ll 1$, we can easily check that

$$s^n = k_1' + \cdots + k_{n-1}' = \sum_{i=0}^{n-1} \mathrm{e}^{\mathrm{i}(q-1)k} k \sim \int_0^{t^n} \mathrm{e}^{(q-1)t}\,\mathrm{d}t = \frac{1}{q-1}\left(\mathrm{e}^{(q-1)t^n} - 1\right),$$

which approximates the other half of the change of variables in (2.11).

The assumption (3.10) on the interface approximation is natural since it only implies that the interface approximation of $Cu(x, t)$ should be simply C times the interface approximation of $u(x, t)$. One can easily check that the interface of the Godunov method (5.3) clearly satisfies this interface assumption.

However, unfortunately, many other schemes do not satisfy this simple and natural assumption. It seems that, if a numerical scheme is consistent, this assumption should be satisfied up to the leading order (see an

example in Section 4.2). Furthermore, for the Lax–Friedrichs scheme, the assertion (3.11) of the proposition holds (see Section 4.1).

Finally, we discuss about initial values. In this paper, we consider three kinds of explicit solutions discussed in Section 2. The first example comes with a positive initial value $u(x,0) = \mathscr{N}_{0,1}(x,0)$. Then the exact solution is the positive N-wave

$$u(x,t) = \mathscr{N}_{0,1}(x,t), \tag{3.13}$$

which is given explicitly by (2.8). Since the numerical approximation is for the cell average (3.4), the discretization of the initial data is taken as

$$U_j^0 := \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} \mathscr{N}_{0,1}(x,0)\,\mathrm{d}x. \tag{3.14}$$

In the second example, the initial value is taken as $u(x,0) = \mathscr{N}_{1,1}(x,0)$ which has both of negative and positive parts. Then the exact solution is the sign-changing N-wave

$$u(x,t) = \mathscr{N}_{1,1}(x,t). \tag{3.15}$$

We may take the initial discretization in the same way as in (3.14). Let $P(t) = \int_0^\infty u(x,t)\,\mathrm{d}x$ which give the total mass of the positive part. Then as we obtained (2.10) we can easily show

$$P(t) = P(0)\mathrm{e}^t.$$

This shows the sensitivity of the wave propagation on the initial discretization. We want to assign U_j^0 the same sizes of positive and negative parts as those of the initial value and hence take

$$U_j^0 := \frac{1}{\Delta x}\int_{x_j}^{x_{j+1}} \mathscr{N}_{1,1}(x,0)\,\mathrm{d}x. \tag{3.16}$$

In this way we may reduce other source of computation errors and observe the properties of numerical schemes. We may observe that some numerical schemes provide the shock place equally well for both of the cases. However, some schemes show a poor approximation for a sign-changing case.

The initial value of the last example is the roll wave $u(x,0) = \mathscr{R}_1(x)$. We consider the problem under a periodic boundary condition, $u(x,t) = u(x+2,t)$, $t > 0$. Then the exact solution is the time invariant roll wave,

$$u(x,t) = \mathscr{R}_1(x), \quad -1 < x < 1, \quad t > 0. \tag{3.17}$$

The initial value is discretized in the same manner as the one in (3.16) that is

$$U_j^0 := \frac{1}{\Delta x}\int_{x_j}^{x_{j+1}} \mathscr{R}_1(x)\,\mathrm{d}x. \tag{3.18}$$

We may also consider initial discretization of type (3.16) for this example. In the case similar phenomena are observed. The difference is specific properties of numerical schemes are reduced due to the extra zero points which suppress the solution at the boundary. Since our goal is to observe the property of numerical schemes we employ (3.18). This example is particularly good to observe how does the scheme approximates rarefaction waves.

To test properties of numerical schemes we also consider three different powers in the flux function, which are $q = 1.5$, 2 and 3. In the following sections we have selected a few examples from these nine possible cases which show interesting properties of each scheme.

4. Blowup of numerical solutions

In this section, we test three well-known schemes, the Lax–Friedrichs (LxF for short), the MacCormack and the Lax–Wendroff. The test results show interesting blowups of numerical solutions which reflects serious defects of each scheme.

4.1. Lax–Friedrichs (LxF) scheme

In this section, we consider the LxF scheme and its second-order modification. The numerical flux of the LxF scheme at the interface $x_{j+1/2}$ is given by

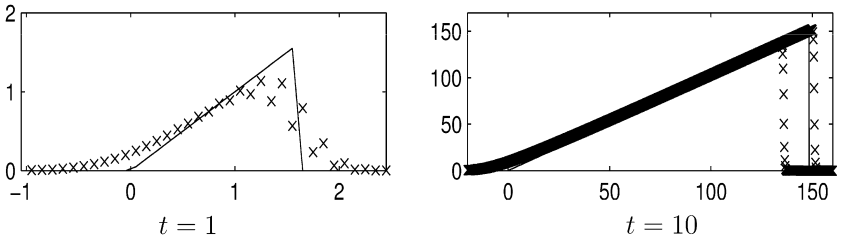

Fig. 1. LxF Scheme (4.2) for the positive solution (3.13) with $q = 2$. The computed solution evolves into two separated N-waves. ((×) Numerical, (−) exact).

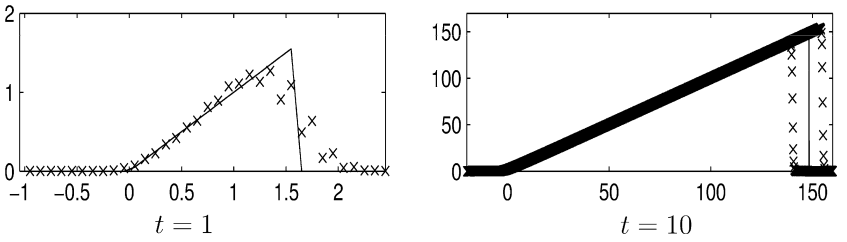

Fig. 2. Second-order LxF scheme (4.3) for the positive solution (3.13) with $q = 2$. The same phenomenon of separation is observed. The numerical viscosity is smaller.

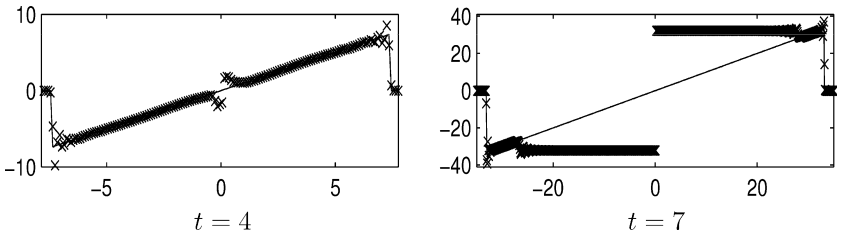

Fig. 3. MacCormack's scheme, (3.9) with (4.7), for the sign-changing solution (3.15) with $q = 2$. A non-physical shock emerges from the sign-changing point and finally destroys the computed solution completely.

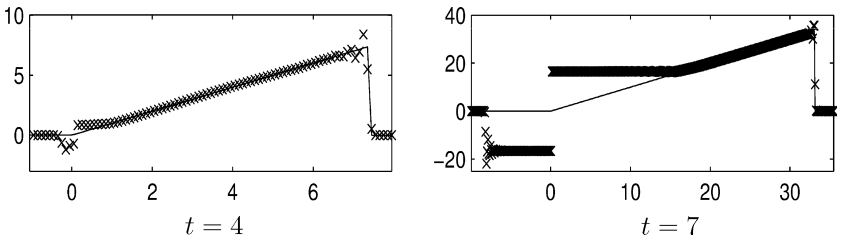

Fig. 4. MacCormack's scheme for the positive solution (3.13) with $q = 2$. Even for this positive case the similar non-physical shock appears due to the oscillation.

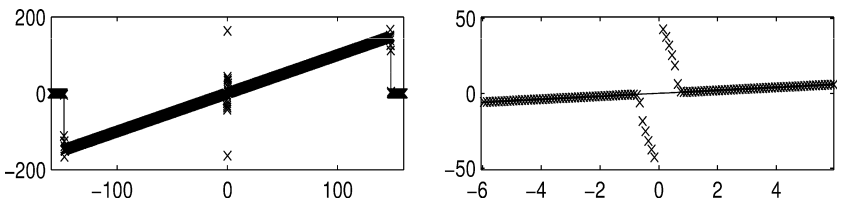

Fig. 5. Richtmyer two-step Lax–Wendroff scheme, (3.9) with (4.9), at time $t = 10$. Non-physical blowup appears at the sonic point and the computed solution collapse.

$$f(u^n_{j+1/2}) = \frac{\Delta x}{2\Delta t^n}(U^n_j - U^n_{j+1}) + \frac{1}{2}(f(U^n_j) + f(U^n_{j+1})). \tag{4.1}$$

Then the numerical scheme (3.9) is written as

$$U^{n+1}_j = \mathrm{e}^{\Delta t^n}\left(\frac{1}{2}(U^n_{j+1} + U^n_{j-1}) - \frac{\Delta t^n}{2\Delta x}(f(U^n_{j+1}) - f(U^n_{j-1}))\right). \tag{4.2}$$

Unfortunately, the interface approximation given by (4.1) does not satisfy the assumption (3.10). However, one can easily check that the inductive arguments in the proof of Proposition 1 still hold. For example, the key step (3.12) can be replaced by

$$\begin{aligned} U^{n+1}_j &= \mathrm{e}^{\Delta t^n}\left(\frac{1}{2}(CV^n_{j+1} + CV^n_{j-1}) - \frac{\Delta s^n}{2C^{q-1}\Delta x}(f(CV^n_{j+1}) - f(CV^n_{j-1}))\right) \\ &= \mathrm{e}^{\Delta t^n} C\left(\frac{1}{2}(V^n_{j+1} + V^n_{j-1}) - \frac{\Delta s^n}{2\Delta x}(f(V^n_{j+1}) - f(V^n_{j-1}))\right) = \mathrm{e}^{t^{n+1}} V^{n+1}_j. \end{aligned}$$

Therefore, the assertion of Proposition 1 is valid for the LxF scheme.

In Fig. 1, exact and computed solutions to the reaction-convection equation (2.1) are given using the LxF method. In the following figures, exact and computed solutions are displayed using lines and dots, respectively. In the figure at time $t = 1$ one can observe a kind of oscillation. However, it is not exactly an oscillation. It is a separation. In Scheme (4.2) only odd numbered cells are used to compute the even numbered ones. Hence, even numbered grids and odd numbered ones generate two different solutions. In the figure at time $t = 10$, we may clearly observe a separation of two N-waves.

High resolution central schemes were proposed by Nessyahu and Tadmor in 1990 in [17], which is based on the Lax–Friedrichs method. Modified from the LxF scheme using Van Leer's MUSCL-type interpolant [21] a direct second-order modification of LxF scheme for (2.1) can be written as:

$$\begin{aligned} U^{n+1/2}_j &= U^n_j - \frac{\Delta t^n}{2\Delta x} f'_j, \\ U^{n+1}_j &= \mathrm{e}^{\Delta t^n}\left(\frac{1}{2}\left[U^n_{j-1} + U^n_{j+1}\right] - \frac{1}{4}\left[U'_{j-1} - U'_{j+1}\right] - \frac{\Delta t^n}{2\Delta x}\left[f\left(U^{n+1/2}_{j+1}\right) - f\left(U^{n+1/2}_{j-1}\right)\right]\right), \end{aligned} \tag{4.3}$$

where the numerical derivatives of the flux $\frac{1}{\Delta x} f'_j = f(u)_x|_{u=U_j} + \mathrm{O}(\Delta x)$ and of the solution $\frac{1}{\Delta x} U'_j = u_x|_{u=U_j} + \mathrm{O}(\Delta x)$ are given by

$$U'_j = \mathrm{minmod}\left(\alpha \Delta U_{j-1/2}, \quad \frac{1}{2}(U_{j+1} - U_j), \quad \alpha \Delta U_{j+1/2}\right), \tag{4.4}$$

$$f'_j = \mathrm{minmod}\left(\alpha \Delta f_{j-1/2}, \quad \frac{1}{2}(f_{j+1} - f_j), \quad \alpha \Delta f_{j+1/2}\right), \tag{4.5}$$

where $\alpha \in [1,2]$, $\Delta U_{j+1/2} = U_{j+1} - U_j$ and

$$\mathrm{minmod}(x_1, x_2, \ldots) = \begin{cases} \min\limits_j \{x_j\} & \text{if } x_j > 0 \text{ for all } j, \\ \max\limits_j \{x_j\} & \text{if } x_j < 0 \text{ for all } j, \\ 0 & \text{otherwise.} \end{cases} \tag{4.6}$$

In Fig. 2, exact and computed solutions are similarly presented using this second-order LxF scheme. In this example, we observe the same phenomenon of separation. This separation is also due to separation of the odd numbered and odd numbered cells. Therefore, a staggering scheme in Section 5.1 is a natural recipe to fix this kind of phenomenon.

4.2. Maccormack

In this section, we consider two oscillatory second-order schemes. In the MacCormack's method the numerical flux at the interface is given by

$$f(u^n_{j+1/2}) = \frac{1}{2}\left(f(U^n_{j+1}) - f\left(U^n_j - \frac{\Delta t^n}{\Delta x}[f(U^n_{j+1}) - f(U^n_j)]\right)\right). \tag{4.7}$$

First, we check that this interface approximation does not satisfy the assumption (3.10). Let $U^n_j = CV^n_j$. Then, for $C = e^{t^n} > 1$ and $q > 1$,

$$\begin{aligned} f(u^n_{j+1/2}) &= \frac{1}{2}\left(f(CV^n_{j+1}) - f\left(CV^n_j - \frac{\Delta t^n}{\Delta x}[f(U^n_{j+1}) - f(U^n_j)]\right)\right) \\ &= \frac{C^q}{2}\left(f(V^n_{j+1}) - f\left(V^n_j - \frac{\Delta t^n}{\Delta x}C^{-1}[f(U^n_{j+1}) - f(U^n_j)]\right)\right) \\ &= \frac{C^q}{2}\left(f(V^n_{j+1}) - f\left(V^n_j - \frac{\Delta t^n}{\Delta x}C^{q-1}[f(V^n_{j+1}) - f(V^n_j)]\right)\right). \end{aligned}$$

However, since

$$f\left(Cv^n_{j+1/2}\right) = C^q f(v^n_{j+1/2}) = \frac{C^q}{2}\left(f(V^n_{j+1}) - f\left(V^n_j - \frac{\Delta t^n}{\Delta x}[f(V^n_{j+1}) - f(V^n_j)]\right)\right),$$

$u^n_{j+1/2} - Cv^n_{j+1/2} \neq 0$ in general. Now we show that the order of this difference is less than the order of $Cv^n_{j+1/2}$, which implies that at least the leading order part of the interface approximation of the MacCormack's method satisfies the assumption on the interface approximation (3.10).

Here, we consider the Burgers case $q = 2$ for simplicity. From the mean value theorem, there exists ξ between $V^n_j - \frac{\Delta t^n}{\Delta x}[f(V^n_{j+1}) - f(V^n_j)]$ and $V^n_j - \frac{\Delta t^n}{\Delta x}C^1[f(V^n_{j+1}) - f(V^n_j)]$ such that

$$|f(u^n_{j+1/2}) - C^2 f(v^n_{j+1/2})| = \frac{C^2}{2}(C-1)\frac{\Delta t^n}{\Delta x}|[f(V^n_{j+1}) - f(V^n_j)]f'(\xi)| \leqslant \frac{\Delta t^n}{2\Delta x}|[f(U^n_{j+1}) - f(U^n_j)]f'(C\xi)|.$$

Employing the exact solutions in (2.8) and the mean value theorem again, we obtain $x_0 \in (x_j, x_j + \Delta x)$ such that

$$|f(U^n_{j+1}) - f(U^n_j)| \cong \frac{1}{2}|(x_j + \Delta x)^2 - x_j^2| \cong \Delta x|x_0| \leqslant \Delta x \max_j |U^n_j|.$$

From the exact solution in (2.8) with $q = 2$ we can easily see that $\max_j|U^n_j| = O(e^{t/2}) = O(C^{1/2})$. Since $C\xi$ is between $U^n_j - \frac{\Delta t^n}{\Delta x}C^{-1}[f(U^n_{j+1}) - f(U^n_j)]$ and $U^n_j - \frac{\Delta t^n}{\Delta x}[f(U^n_{j+1}) - f(U^n_j)]$, $f'(C\xi)$ is of order $O(C^{1/2})$. Therefore, since $f(u^n_{j+1/2})$ is of order C, we obtain

$$\frac{|f(u^n_{j+1/2}) - C^2 f(v^n_{j+1/2})|}{|f(u^n_{j+1/2})|} = O(\Delta x). \tag{4.8}$$

This estimate indicates that, even if the assumption (3.10) does not hold for the MacCormack's scheme, its leading order approximation satisfies the assumption.

Now we examine the properties of MacCormack's method from numerical experiments. In Fig. 3, exact and computed solutions to the reaction-convection equation (2.1) are given using the fully discrete method (3.9) with the MacCormack's numerical flux (4.7). In the figure at $t = 4$, one can observe a small discontinuity that violates the entropy condition (2.3). This non-physical shock grows fast and eventually all the meaningful information of the solution disappears.

Proposition 1 claims that this strange behavior of the numerical experiment is due to the property of the scheme for the convection equation (3.1). In fact, this behavior is related to the well-known fact of the MacCormack's scheme that for any $c > 0$ the discrete function

$$U^n_j = \begin{cases} -c, & j \leqslant 0 \\ c, & j > 0 \end{cases}$$

is a steady state of the MacCormack's scheme for the convection equation (2.12). This example indicates that this kind of steady state is not a rare case and one should deal with such a blowup if reaction terms play a crucial role. In fact, since the approximation is a piecewise constant function, such a discontinuity may appear

easily across a sign-changing point. Furthermore, even for a positive solution case this kind of inadmissible discontinuity may appear due to the oscillating property of the scheme. In Fig. 4 we may observe a non-physical shock developed from a positive solution.

4.3. Lax–Wendroff scheme

Next we consider the Richtmyer two-step Lax–Wendroff method. In this scheme the interface approximation is given by

$$u^n_{j+1/2} = \frac{1}{2}(U^n_j + U^n_{j+1}) - \frac{\Delta t^n}{2\Delta x}[f(U^n_{j+1}) - f(U^n_j)]. \tag{4.9}$$

As we did for the MacCormack's scheme we can similarly show that at least the leading order term satisfies the assumption (3.10). In Fig. 5, the numerical solution looks fine until $t = 9$. However, when it reaches $t = 10$, we can observe a non-physical pick around the sign changing point and the computed solution blows up quickly.

Similar behaviors of computed solutions are observed with other powers $q > 1$. One may observe such blowups more easily for the periodic case (3.17).

5. Fully-discrete schemes

In this section, we test three more fully discrete numerical schemes, NT scheme and the first and the second-order Godunov schemes. These schemes do not show the bizarre behavior observed previously. Therefore, the tests of the accuracy of the shock location and the rarefaction profile are now meaningful. The properties of schemes observed in this section are tabled in Section 7.1 for easier comparison.

5.1. NT scheme

The separation into two waves of computed solutions in Figs. 1 and 2 is due to the separation of even numbered and the odd numbered cells in the schemes. Therefore, the best way to cure such a symptom is to consider a scheme in a staggered form. The second order modification (4.3) can be written in a staggered form (see [4]):

$$U^{n+1}_{j+1/2} = \frac{1}{2}[U^n_j + U^n_{j+1}] + \frac{1}{8}[U'_j - U'_{j+1}] - \frac{\Delta t^n}{\Delta x}[f(U^{n+1/2}_{j+1}) - f(U^{n+1/2}_j)], \tag{5.1}$$

where U'_j and $U^{n+1/2}_j$ are given by (4.3)–(4.6). This scheme is usually called the NT scheme and is employed in this section for numerical computations.

First, we compare the shock location. In Figs. 6 and 7, computed solutions of the NT scheme are given for the positive N-wave (3.13) and the sign-changing N-wave (3.15), respectively. In the figures, solutions are plotted near the shock to check the performance. Most of the cases the NT scheme provides reasonably correct shock locations. The only exception is the case for the sign-changing solution with $q = 3$. In this case, the

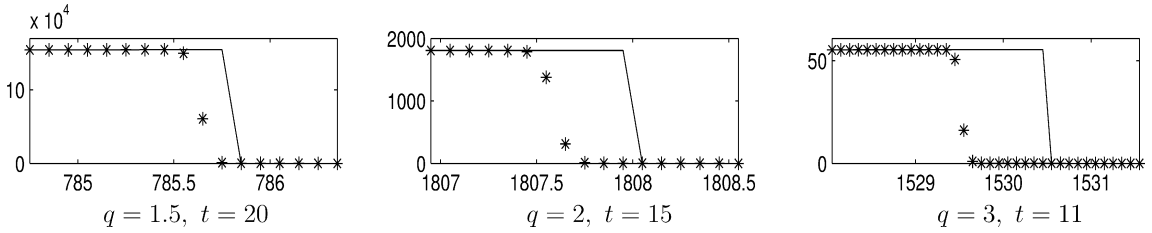

Fig. 6. NT scheme (5.1) for the positive solution (3.13). The magnifications of computed solution near the shock show reasonably correct shock locations.

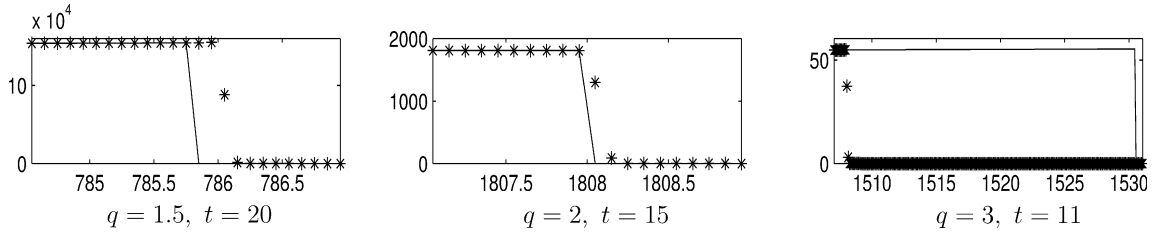

Fig. 7. NT scheme for the sign changing solution (3.15). Shock locations are correct except the case $q = 3$ when the exact solution has a singularity at the origin.

numerical solution gives slow shock propagation. This behavior seems related to the numerical viscosity and the singularity of the solution near the sign change.

Computed solutions for the roll wave (3.17) with the periodic boundary condition show how the scheme approximates the rarefaction wave. In the first row of Fig. 8, computed solutions are given with exact ones using the NT scheme (5.1). In the second row, the errors of computed solutions are given. Since the computed solution is an approximation of cell averages under the initial discretization (3.18), the error is taken as

$$\text{Error} = U_j^n - \int_j^{j+1} u(x, t^n)\,\mathrm{d}x, \tag{5.2}$$

where u is the exact solution. In the figures this error is plotted at the grid points $x_{j+1/2}$.

Now proper interpretations for the computation result is required. First, the error is an increasing function away from discontinuities for all three cases. This implies that the graph of the approximation is steeper and hence the numerical flux is slightly weaker than the exact one. We may also say that the sonic glitch phenomenon is not observed from all of three cases. The numerical error for the case $q = 3$ shows a discontinuity at the sonic points. Notice that the continuity is simply due to the singularity of the exact solution at the sonic point.

Remark 2. Note that the linear reaction term makes the solution grow and the nonlinear convection term makes it flat. Therefore, if the error function in (5.2) increases away from the discontinuity (i.e., the computed solution is flatter than the exact one), then we may say that the flux is overestimated numerically. Another factor that decides the signs of the error function is so called the sonic glitch phenomenon. Note that $x = \pm 1$ are sonic points for the periodic cases. Certain numerical schemes generate entropy violating discontinuities at such points, which is called a sonic glitch.

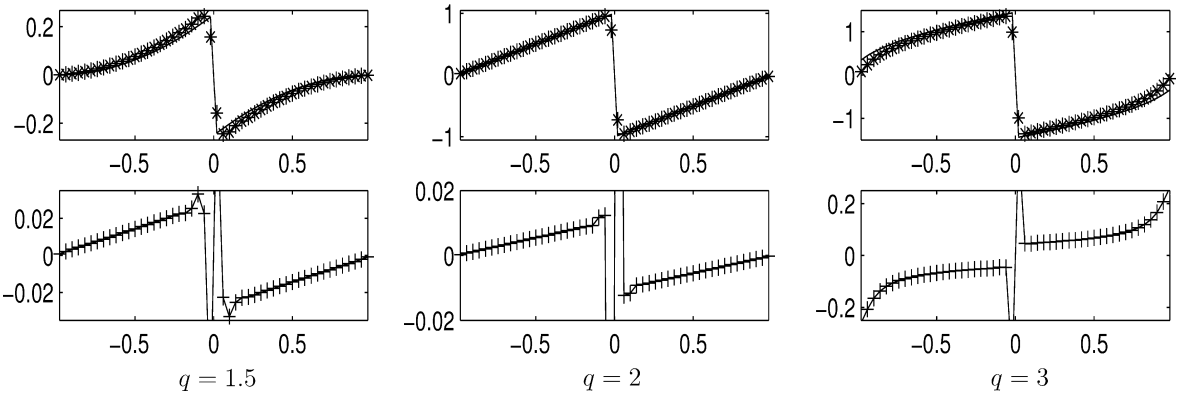

Fig. 8. NT scheme for the periodic solution (3.17). The error in the second row is an increasing function away from the shock. There is no sonic glitch. The discontinuity at the sonic point $x = \pm 1$ for the case $q = 3$ is due to the singularity of the exact solution.

5.2. First and second-order Godunov methods

Godunov schemes [1,10] are based on either the exact or an approximate solution of the Riemann problem using characteristic information within the framework of a conservative method. Since the flux is convex $f''(u) > 0$, the interface approximation of the first-order Godunov method is simply given by

$$u^n_{j+1/2} = \begin{cases} U_j & \text{if } U_j > 0 \text{ and } s > 0, \\ U_{j+1} & \text{if } U_{j+1} < 0 \text{ and } s < 0, \\ 0 & \text{if } U_j < 0 < U_{j+1}. \end{cases} \tag{5.3}$$

Here, $s = [f(U_{i+1}) - f(U_i)]/(U_{i+1} - U_i)$ is the shock speed at the interface $x_{j+1/2}$. Clearly, this interface approximation satisfies the condition (3.10) and the assertion in Proposition 1 is valid.

In Figs. 9 and 10, computed solutions of the first-order Godunov scheme are given for the positive N-wave (3.13) and the sign-changing N-wave (3.15), respectively. One can observe that the Godunov method gives equally correct shock locations for both of positive and sign-changing cases. In particular, the shock location is pretty correct even for the sign-changing case with $q = 3$ which is the case that all the other schemes considered give pretty poor performance. In Fig. 11, one may observe sonic glitches for the cases $q = 2$ and $q = 3$. For the case $q = 1.5$ the rarefaction wave has slope zero at the sonic point and the glitch does not develop. The error function is decreasing for the case $q = 3$, which implies that the flux is slightly over estimated numerically. Notice that Godunov schemes are the only ones that produce increasing error functions for certain cases.

The Godunov method can be modified to a second-order scheme by employing a proper limiter. For this example we simply use the CLAWPACK [12] with monotonized centered limiter. From Figs. 12 and 13, one may observe that this second-order Godunov scheme provides very accurate shock location for all of six cases. They are almost exact.

On the other hand, from Fig. 14, one may observe that the structure of the error function is exactly the opposite of those of the first-order one. First sonic glitches are not observed. The overall error is smaller than most of other schemes or is competitive with others at least. However, the error function is decreasing for the case $q = 1.5$ and is increasing for the case $q = 3$. The case $q = 2$ is somewhat between them. So $q = 1.5$ is the case that the flux is over estimated numerically even if the size is small. Note that the property of the scheme is

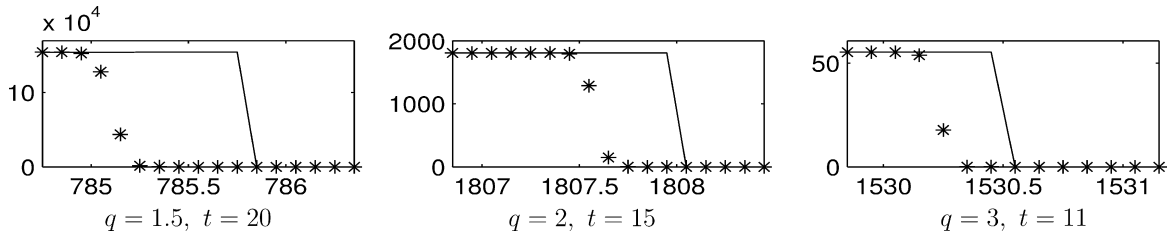

Fig. 9. The first-order Godunov scheme, (3.9) with (5.3), for the positive solution (3.13). The shock locations are reasonably correct even with this first-order scheme.

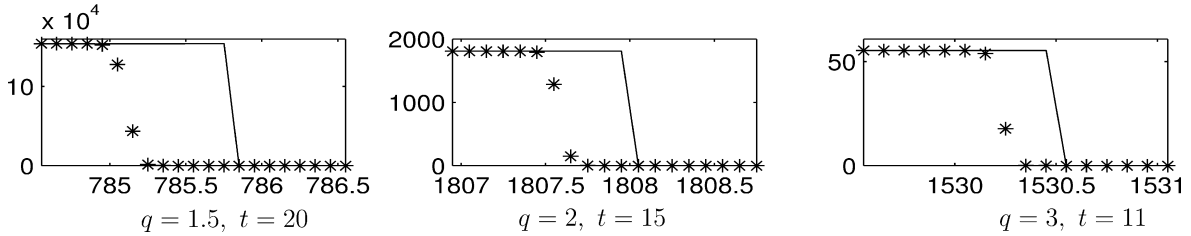

Fig. 10. The first order Godunov scheme for the sign changing solution (3.15). The shock locations are correct even for the singular case $q = 3$.

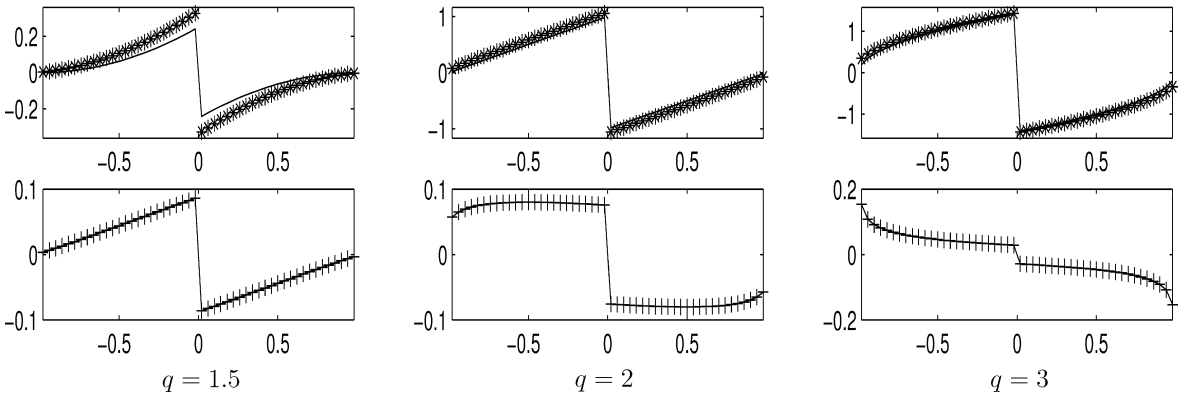

Fig. 11. The first-order Godunov scheme for the periodic solution (3.17). Sonic glitches are observed for $q = 2$, 3. The error increases if $q = 1.5$ or 2, but decreases if $q = 3$. Godunov is the only scheme with a decreasing error function in our tests.

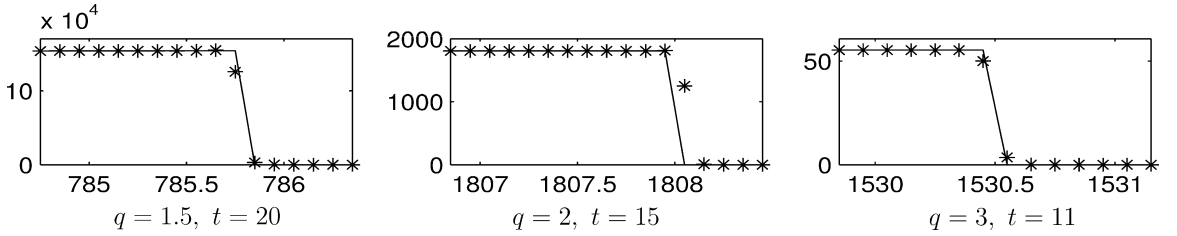

Fig. 12. The second-order Godunov (CLAWPACK) for the positive solution (3.13). The shock locations are almost exact for all of three cases.

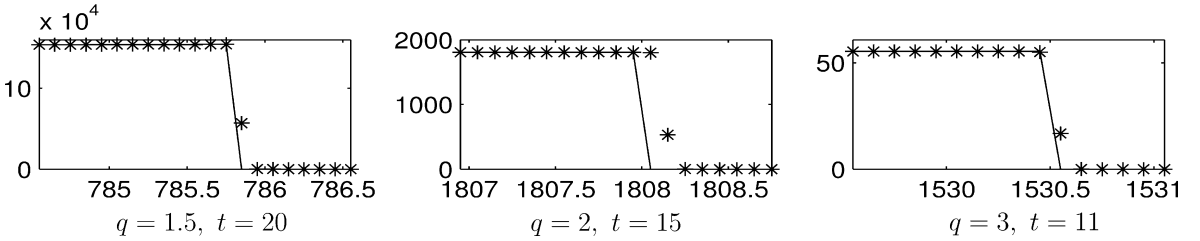

Fig. 13. The second-order Godunov for the sign changing solution (3.15). The shock locations are almost exact even for the singular case $q = 3$.

also strongly depending on the property of the limiter. Using different limiter one may obtain different phenomenon.

6. Semi-discrete schemes

We may employ the third-order TVD Runge–Kutta discretization for time stepping (see [18]). In the case, the scheme is written as

$$\begin{aligned} U_j^{(1)} &= U_j^n + \Delta t^n \mathscr{L}(U^n; j) \\ U_j^{(2)} &= \frac{3}{4} U_j^n + \frac{1}{4} U_j^{(1)} + \frac{1}{4} \Delta t^n \mathscr{L}(U^{(1)}; j) \\ U_j^{n+1} &= \mathrm{e}^{\Delta t^n} \left(\frac{1}{3} U_j^n + \frac{2}{3} U_j^{(2)} + \frac{2}{3} \Delta t^n \mathscr{L}(U^{(2)}; j) \right), \end{aligned} \tag{6.1}$$

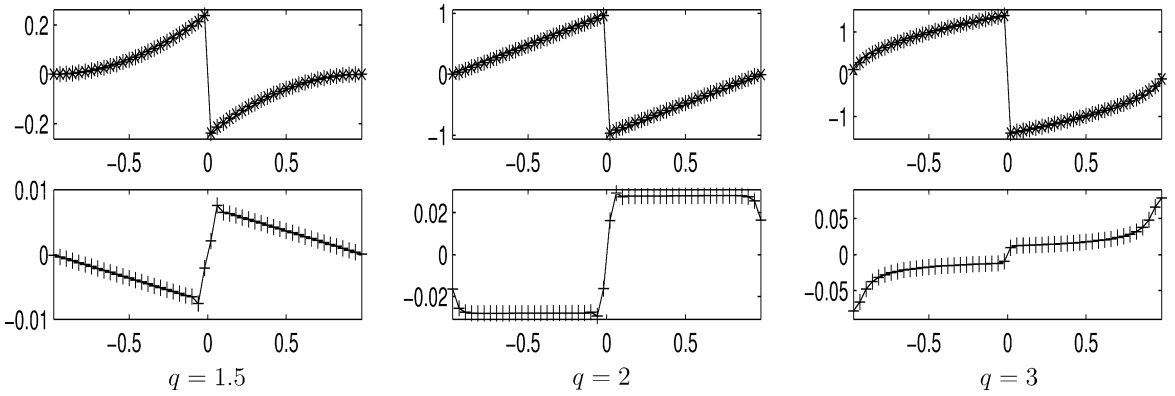

Fig. 14. The second-order Godunov for the periodic solution (3.17). The sonic glitch is observed for the case $q = 2$. The error function decreases for the case $q = 1.5$.

where the finite difference operator $\mathscr{L}$ is given by (3.6). This kind of semi-discrete schemes usually provide better performance than the fully discrete methods (3.9). However, the unexpected behaviors of schemes presented in this paper are observed from both of fully and semi discrete methods for the scheme tested in Section 4. In this section, we test the central upwind method and the WENO scheme using this semi-discrete time marching.

6.1. Central upwind scheme

Next we consider one more central scheme which shares some properties with the Godunov scheme. In [8,9], a semi-discrete central upwind scheme (KNP scheme for short) was introduced. This scheme is constructed based on a piecewise linear approximation like the NT scheme. We compute the local speeds of propagation at the interface $x = x_{j+1/2}$ which may have discontinuities. Since the speed of propagation is related to the CFL condition, we can estimate the local speeds of the right and left side at the cell boundary. The local speeds of wave propagation are bounded by $a^n_{j+1/2,r}$ and $a^n_{j+1/2,l}$ which are given by

$$a^n_{j+1/2,r} = \max_{\mathscr{C}}(f'(u), 0), \quad a^n_{j+1/2,l} = \min_{\mathscr{C}}(f'(u), 0), \tag{6.2}$$

where $\mathscr{C}$ is a relevant range for u. Employing this local speed of propagation the flux at the interface is approximated by

$$f(u^n_{j+1/2}) = \frac{a_{j+1/2,r} f(u^-_{j+1/2}) - a_{j+1/2,l} f(u^+_{j+1/2})}{a_{j+1/2,r} - a_{j+1/2,l}} + \frac{a_{j+1/2,r} a_{j+1/2,l}}{a_{j+1/2,r} - a_{j+1/2,l}} [u^+_{j+1/2} - u^-_{j+1/2}], \tag{6.3}$$

where $u^+_{j+1/2}$ and $u^-_{j+1/2}$ are computed as

$$u^+_{j+1/2} \equiv U^n_{j+1} - \frac{\Delta x}{2}(u_x)_{j+1}(t^n),$$

$$u^-_{j+1/2} \equiv U^n_j + \frac{\Delta x}{2}(u_x)_j(t^n),$$

$$(u_x)_j = \operatorname{minmod}\left(\alpha \frac{U^n_{j+1} - U^n_j}{\Delta x}, \frac{U^n_{j+1} - U^n_{j-1}}{\Delta x}, \alpha \frac{U^n_j - U^n_{j-1}}{\Delta x}\right), \quad 1 \leqslant \alpha \leqslant 2.$$

From Figs. 15 and 16, one may observe that the KNP scheme provides good shock location except the sign-changing case with $q = 3$. In this case the shock location is about the middle of the ones of the NT scheme and of the exact solution. Considering the fact that the Godunov scheme provides almost exact shock location for this case (see Fig. 10) we may feel that the KNP scheme is placed between the NT scheme and the Godunov scheme.

In Fig. 17, we may observe sonic glitches for the case $q = 2$ and $q = 3$ which is a property of the first-order Godunov scheme. The error functions are increasing away from the discontinuity which is a property of the NT scheme.

6.2. WENO scheme

The last numerical scheme considered is a high-order weighted essentially non-oscillatory (WENO for short) method (see [3,2,5,15,18]). For this example, we employ the semi-discrete Runge–Kutta type method (6.1). To avoid entropy violating solutions and obtain the numerical stability we split the flux $f(u)$ into two components f^+ and f^- such that

$$f(u) = f^+(u) + f^-(u), \tag{6.4}$$

where $\frac{\partial f^+}{\partial u} \geqslant 0$ and $\frac{\partial f^-}{\partial u} \leqslant 0$. One of the simplest flux splitting is the Lax–Friedrichs splitting which is given by

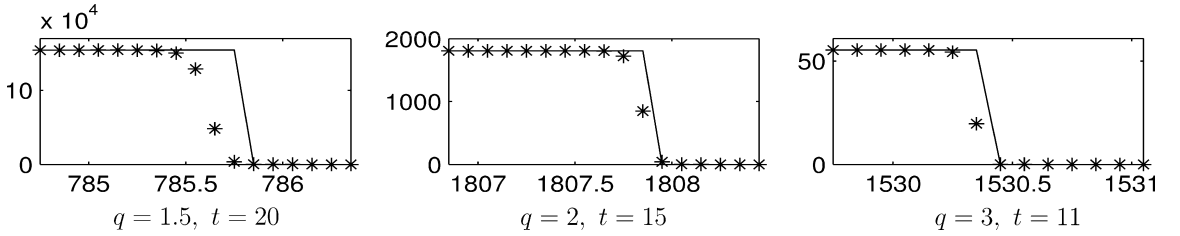

Fig. 15. KNP scheme, (6.1) with (6.3), for the positive solution (3.13). The method provides correct shock locations.

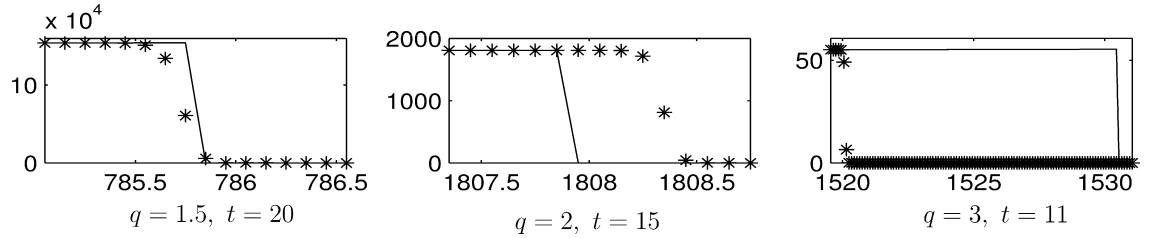

Fig. 16. KNP scheme for the sign changing solution (3.15). Shock locations are correct except the case $q = 3$ due to the singularity at the sonic point.

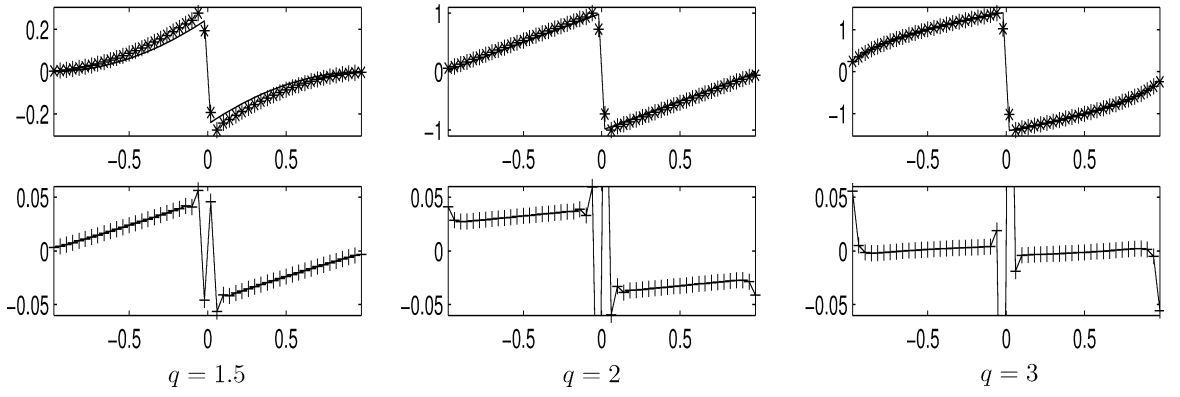

Fig. 17. KNP scheme for the periodic solution (3.17). The error functions are increasing in all three cases away from the shock. Sonic glitches are observed if $q = 2$ or 3.

$$f^{\pm}(u) = \frac{1}{2}(f(u) \pm \alpha u), \tag{6.5}$$

where $\alpha = \max_u |f'(u)|$ over the pertinent range of u which can be decided a priori using the explicit formula for the exact solution.

The interface approximation of the fifth-order WENO with Lax–Friedrichs splitting (WENO-LF5 for short) is given by

$$\begin{aligned} f(u^n_{j+1/2}) = {} & \frac{1}{12}(-f_{j-1} + 7f_j + 7f_{j+1} - f_{j+2}) - \Phi_N(\Delta f^+_{j-\frac{3}{2}}, \Delta f^+_{j-\frac{1}{2}}, \Delta f^+_{j+\frac{1}{2}}, \Delta f^+_{j+\frac{3}{2}}) \\ & + \Phi_N\left(\Delta f^-_{j+\frac{5}{2}}, \Delta f^-_{j+\frac{3}{2}}, \Delta f^-_{j+\frac{1}{2}}, \Delta f^-_{j-\frac{1}{2}}\right), \end{aligned} \tag{6.6}$$

where $f_j = f(u^n_j)$, $f^{\pm}_j = f^{\pm}(u^n_j)$, $\Delta f^{\pm}_{i+\frac{1}{2}} = f^{\pm}_{i+1} - f^{\pm}_i$ and

$$\Phi_N(a,b,c,d) = \frac{1}{3}\omega_0(a - 2b + c) + \frac{1}{6}\left(\omega_2 - \frac{1}{2}\right)(b - 2c + d). \tag{6.7}$$

The nonlinear weights ω_0 and ω_2 are defined by

$$\omega_j = \frac{\alpha_j}{\sum_{l=0}^{k-1}\alpha_l}, \quad \alpha_l = \frac{d_l}{(\varepsilon + \beta_l)^2}, \quad d_0 = \frac{1}{10}, \quad d_1 = \frac{3}{5}, \quad d_2 = \frac{3}{10},$$

where $0 < \varepsilon \ll 1$ is taken to prevent singularity and the smoothness indicators β_j's are given by

$$\begin{aligned} \beta_0 &= \frac{13}{12}(f_{i-2} - 2f_{i-1} + f_i)^2 + \frac{1}{4}(f_{i-2} - 4f_{i-1} + 3f_i)^2 \\ \beta_1 &= \frac{13}{12}(f_{i-1} - 2f_i + f_{i+1})^2 + \frac{1}{4}(f_{i-1} - f_{i+1})^2 \\ \beta_2 &= \frac{13}{12}(f_i - 2f_{i+1} + f_{i+2})^2 + \frac{1}{4}(3f_i - 4f_{i+1} + f_{i+2})^2. \end{aligned} \tag{6.8}$$

From Figs. 18 and 19, one may observe that the scheme gives reasonable shock location except the case $q = 3$ for the sign-changing solution. In this case, the shock location is pretty same as the NT scheme. In the roll wave computation, Fig. 20, the sonic glitch is not observed. In all of the three cases the error functions are increasing away from the shock. Therefore, we may say that the flux is slightly under estimated.

7. Appendix

The properties of each scheme tested in the previous sections are tabled in Section 7.1. This comparison is not to say that one scheme is better than the other, which is not possible at all. Instead it is simply to compare their properties more clearly. It is natural to ask if there is a method that may neutralize the undesirable behaviors of numerical schemes instead of magnifying them. In Section 7.2 we apply several other methods and compare to previous results which employed the splitting method with the exact solver. In most of the cases, the properties of each scheme are magnified similarly and there is no meaningful difference.

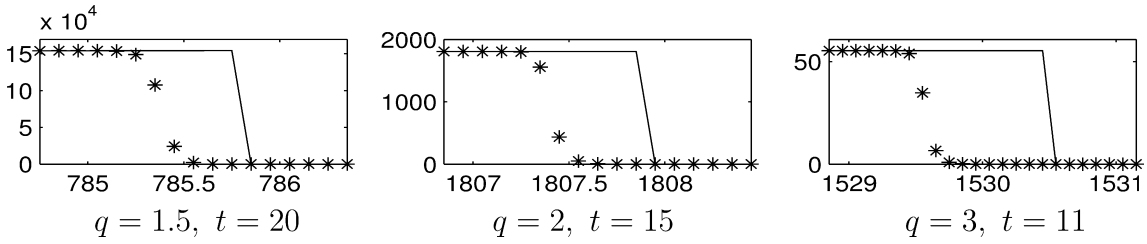

Fig. 18. WENO scheme, (6.1) with (6.6), for the positive solution (3.13). This method provides reasonably correct shock locations.

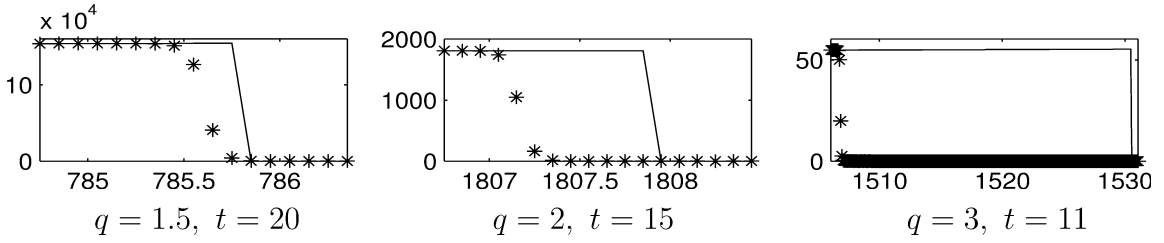

Fig. 19. WENO scheme for the sign changing solution (3.15). Shock locations are correct except the case $q = 3$ due to the singularity at the sonic point.

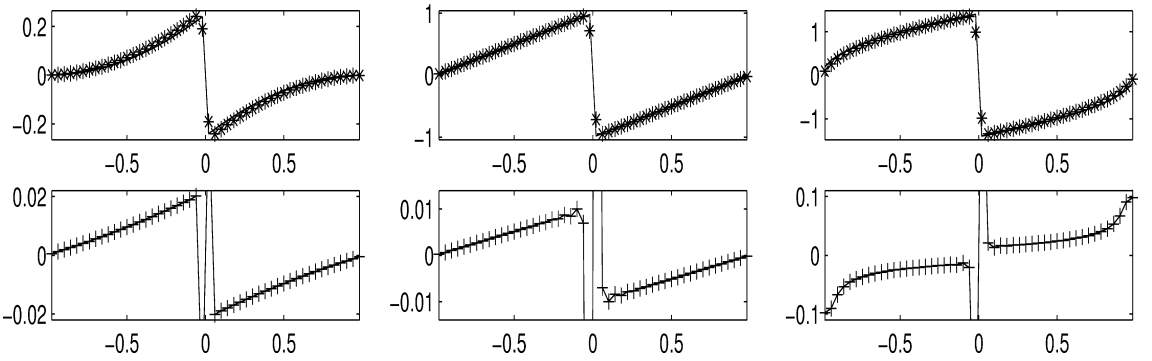

Fig. 20. WENO scheme for the periodic solution (3.17). The error functions are increasing. There is no sonic glitch. The discontinuity at the sonic point for the case $q = 3$ is due to the singularity of the exact solution at the sonic point.

7.1. Summery of test results

The tables in this section are made from the computation results in the previous sections. In Tables 1 and 2, the accuracy of the shock location of each scheme are compared for the positive and the sign-changing cases, respectively. In the tables the numbers of grid points between the exact and the numerical shock are given. The negative or positive sign in the table imply the numerical shock is before or after the exact one, respectively. One may observe that the Godunov second-order method gives the best shock location. However, considering the huge number of grid points used in the computation, one can say other schemes are also equally good.

By computing the roll wave (3.17) one may test if a numerical scheme approximates rarefaction waves well. In Tables 3 and 4, a few properties of schemes are tabled for the power $q = 1.5$ and $q = 2$, respectively. The tables are about the structure of the error functions for $-1 < x < -0.1$. One may easily observe that the second-order Godunov scheme has a unique property. It is the only scheme that the error function is negative and decreasing. This indicates that the scheme provide numerical flux which is slightly bigger than the actual advection. This property may cause vacuum or negative density in the computation of gas dynamics. Any sonic glitch is not observed for the case $q = 1.5$. For the Burgers case sonic glitches are observed from the Godunov schemes and the KNP scheme.

Table 1
Shock location test #1: the numbers of grid points between the exact and the numerical shock for positive solutions (3.13) are tabled

Scheme	$q = 1.5,\ t = 20$	$q = 2,\ t = 15$	$q = 3,\ t = 11$
First-order Godunov	−7	−4	−3
Second-order Godunov	0	+1	0
NT scheme	−2	−4	−10
KNP scheme	−2	−1	−1
WENO-LF5	−4	−5	−10

The negative or positive sign in the table indicate that the numerical shock is slower or faster than the exact one, respectively.

Table 2
Shock location test #2: the numbers of grid points between the exact and the numerical shock for sign-changing solution (3.15) are listed

Scheme	$q = 1.5$, $t = 20$	$q = 2$, $t = 15$	$q = 3$, $t = 11$
First-order Godunov	−7	−4	−3
Second-order Godunov	0	+2	+1
NT scheme	+3	+1	−220
KNP scheme	−1	+5	−210
WENO-LF5	−2	−8	−230

The negative or positive sign in the table indicate that the numerical shock is slower or faster than the exact one, respectively.

7.2. Other algorithms for the source term

In the previous sections, we employed the splitting method with the exact solver to handle the source term. In this section, we consider four other algorithms to discretize

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

and compare their behaviors to the previous results. First, instead of the exact solver, splitting methods with RK 2 and RK 4 are tested.

A natural non-splitting method is the direct substitution. For the fully discrete method (LxF, MacCormack, Godunov and NT), it is written as

$$U_j^{n+1} = U_j^n - \frac{\Delta t^n}{\Delta x}\left(f(u_{j+1/2}^n) - f(u_{j-1/2}^n)\right) + \Delta t^n U_j^n, \tag{7.1}$$

where $u_{j+1/2}^n$ is the interface approximation of each scheme. For the semi-discrete methods (KNP and WENO-LF5), the operator $\mathscr{L}$ in (3.6) can be modified as

$$\mathscr{L}(U^n; j) := \frac{f(u_{j-1/2}^n) - f(u_{j+1/2}^n)}{\Delta x} + U_j^n$$

and the third line in (6.1) is replaced with

Table 3
Roll wave (or Rarefaction wave) test with $q = 1.5$: the structures of the difference, the exact solution minus the numerical approximation for (3.17), are listed for each scheme

Scheme	Maximum error	Monotonicity	Sonic glitch
Godunov First-order	+0.1	Increasing	None
Godunov Second-order	−0.01	Decreasing	None
NT scheme	+0.03	Increasing	None
KNP scheme	+0.05	Increasing	None
WENO-LF5	+0.02	Increasing	None

The structure on (−1, −0.1) is stated for comparison.

Table 4
Roll wave (or rarefaction wave) test with $q = 2$: the structures of the difference, the exact solution minus the numerical approximation for (3.17), are listed for each scheme

Scheme	Maximum error	Monotonicity	Sonic glitch
Godunov First-order	+0.1	Increasing	Yes
Godunov Second-order	−0.03	Decreasing	None
NT scheme	+0.02	Increasing	None
KNP scheme	+0.05	Increasing	Yes
WENO-LF5	+0.01	Increasing	None

The structure on (−1, −0.1) is stated for comparison.

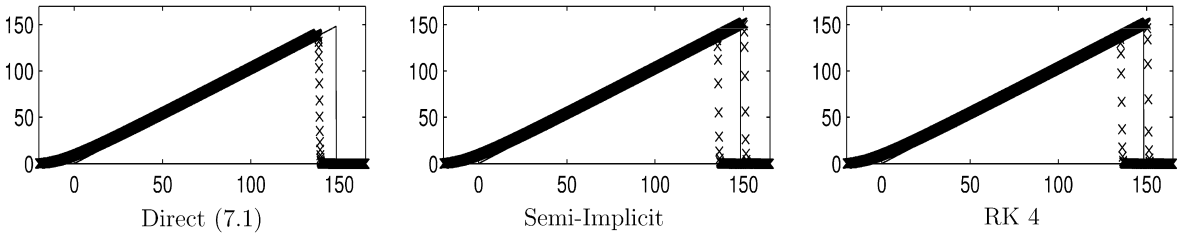

Fig. 21. LxF Scheme (4.2) for the positive solution (3.13) with $q = 2$. Separation phenomenon does not appear for the direct method.

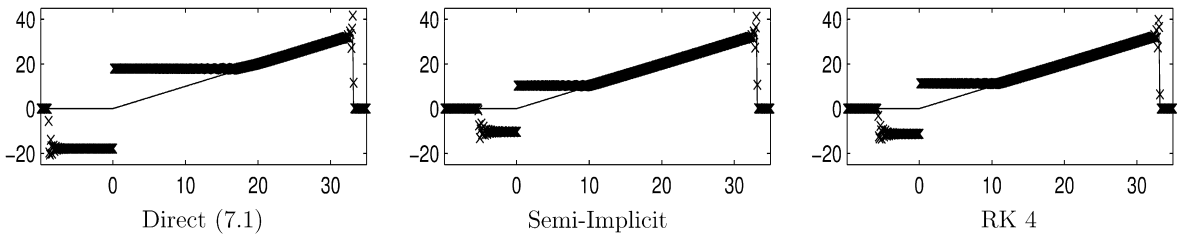

Fig. 22. MacCormack's scheme for the positive solution (3.13) with $q = 2$ and $t = 7$. A non-physical shock appears similarly.

$$U_j^{n+1} = \frac{1}{3}U_j^n + \frac{2}{3}U_j^{(2)} + \frac{2}{3}\Delta t^n \mathscr{L}(U^{(2)}; j). \tag{7.1′}$$

Unlike the fully discrete case the information of the source term is also updated with RK 3 accuracy and, hence, one may expect better resolution. We also consider the splitting version of the semi-implicit method discussed in [13]. As mentioned in the paper the non-splitting version is somewhat worse and results of splitting version are included.

First, we test if the separation phenomenon of the LxF scheme similarly appears under these methods. In Fig. 21, one can observe similar separations under the semi-implicit method and RK 4. The direct method is the only successful case. We know that the odd numbered grid points and even numbered ones evolve independently in the LxF scheme. In the direct method (7.1) they are not separated anymore and the separation phenomenon is resolved. However, it is just a lucky case. This kind of luck does not happen in the following example.

Next we check if the non-physical shock of the MacCormack's scheme appears similarly. In Fig. 22 one can observe that such a phenomenon appears in all three cases. The shape of the numerical solutions are similar. Therefore, one can say that there is no meaningful difference.

The test for computing the roll wave (3.17) does not give any noticeable difference. Therefore, we include the shock location comparison only in Table 5. One can observe that RK 4 shows almost the same shock location as the one with exact solver and one of the RK 2 is several grid points behind. These indicate that if one increase the order of splitting method, the approximation converges to the one with the exact solver. In the meantime, the shock propagation under the semi-implicit method is several grid points faster. The direct

Table 5
Shock location test with $q = 2$: the numbers of grid points between the exact and the numerical shock for sign-changing solution (3.15) are listed

Method	Godunov 1	Godunov 2	NT	KNP	WENO-LF5
Exact solver	−4	+1	+1	+4	−8
Non-splitting (7.1)	−265	−260	−260	+4	−8
Semi-implicit	−2	+3	+2	+9	−5
RK 2	−8	−3	−1	−4	−12
RK 4	−4	+1	+1	+4	−8

The negative or positive sign in the table imply the numerical shock is before or after the exact one, respectively.

substitution (7.1) shows poor performance for the fully-discrete case. However, the accuracy of the semi-discrete case is as good as RK 4. The pattern one can observe is almost same for all the schemes tested. Considering the huge number of grid points used in the computation these differences are relatively small.

8. Conclusion

In this paper, we have introduced a theoretical microscope to examine the properties of numerical schemes for advection equations and applied it to LxF, MacCormack, Lax–Wendroff, Godunov, NT, KNP and WENO-LF5 schemes. Unusual blowups of numerical solutions are observed from the test, see Figs. 1–5. It is shown in Proposition 1 that a defect of a scheme is exponentially magnified together with the linear source term in (2.1) if a splitting method is applied with the exact solver for the source term. In many cases, these defects are overlooked since they are small. However, if the evolution is dominated by the reaction, those small bugs may grow into a monster as observed in the previous examples. Therefore, one should pay more attention to the properties of each scheme when the reaction plays a dominant role.

Even though we used the model equation (2.1) to test the properties of schemes for the advection, it can be employed to test schemes for reaction. For example, one may ask if there is a better way to handle the source term so that defects of each scheme are neutralized instead of being magnified. In Section 7.2, we tested semi-implicit, non-splitting and higher order splitting methods and obtained similar behaviors. It seems a challenge to find a method to handle a source term in a way that those defects are neutralized. If that is not possible, it can be a more realistic approach to find a cure to fix the defects of each scheme.

References

[1] S.K. Godunov, A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics, Mat. Sb. (NS) 47 (1959) 271–306.
[2] Y. Ha, C.L. Gardner, A. Gelb, C.-W. Shu, Numerical simulation of high mach number astrophysical jets with radiative cooling, J. Sci. Comput. 24 (2005) 29–44.
[3] A. Harten, B. Engquist, S. Osher, S. Chakravarthy, Uniformly high order essentially non-oscillatory schemes, III, J. Comput. Phys. 231 (1987) 231–303.
[4] G.-S. Jiang, D. Levy, C.-T. Lin, S. Osher, E. Tadmor, High-resolution nonoscillatory central schemes with nonstaggered grids for hyperbolic conservation laws, SIAM J. Numer. Anal. 35 (1998) 2147–2168.
[5] G.-S. Jiang, C.-W. Shu, Efficient implementation of weighted ENO schemes, J. Comput. Phys. 126 (1996) 202–228.
[6] S. Jin, M. Katsoulakis, Hyperbolic systems with supercharacteristic relaxations and roll waves, SIAM J. Appl. Math. 61 (2000) 271–292.
[7] S. Jin, Y.-J. Kim, On the computation of roll waves, M2AN Math. Model. Numer. Anal. 35 (3) (2001) 463–480.
[8] A. Kurganov, S. Noelle, G. Petrova, Semi-discrete central-upwind schemes for hyperbolic conservation laws and Hamilton–Jacobi equations, SIAM J. Sci. Comput. 23 (2001) 703–740.
[9] A. Kurganov, E. Tadmor, New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations, J. Comput. Phys. 160 (2000) 241–282.
[10] R.J. LeVeque, Numerical methods for conservation lawsLectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 1990.
[11] R.J. LeVeque, High-resolution conservative algorithms for advection in incompressible flow, SIAM J. Numer. Anal. 33 (1996) 627–665.
[12] R.J. LeVeque, Clawpack Version 4.0 Users Guide, Technical report, University of Washington, Seattle, 1999. Available online at: http://www.amath.washington.edu/-claw/.
[13] R.J. Leveque, H.C. Yee, A study of numerical methods for hyperbolic conservation laws with stiff source terms, J. Comput. Phys. 86 (1) (1990) 187–210.
[14] R. Liska, B. Wendroff, Comparison of several difference schemes on 1D and 2D test problems for the Euler equations, SIAM J. Sci. Comput. 25 (3) (2003) 995–1017.
[15] X.-D. Liu, S. Osher, T. Chan, Weighted essentially non-oscillatory schemes, J. Comput. Phys. 115 (1994) 200–212.
[16] A.N. Lyberopoulos, Asymptotic oscillations of solutions of scalar conservation laws with convexity under the action of a linear excitation, Quar. Appl. Math. XLVIII (1990) 755–765.
[17] H. Nessyahu, E. Tadmor, Nonoscillatory central differencing for hyperbolic conservation laws, J. Comput. Phys. 87 (1990) 408–463.
[18] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock capturing schemes, J. Comput. Phys. 77 (1988) 439–471.
[19] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock capturing schemes, II, J. Comput. Phys. 83 (1989) 32–78.

[20] C.-W. Shu, S. Osher, Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws, in: B. Cockburn, C. Johnson, C.-W. Shu, E. Tadmor (Ed.: A. Quarteroni), Advanced numerical approximation of nonlinear hyperbolic equations, Lecture Notes in Mathematics, vol. 1697, Springer, 1998, p. 325.
[21] B. van Leer, MUSCL, a new approach to numerical gas dynamics, in: Computing in Plasma Physics and Astrophysics, Max-Planck-Institut fűr Plasma Physik, Garchung, Germany, 1976.