Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2022-0243

Vol. **34**, No. 3, pp. 749-786
September 2023

On the Solution Accuracy Downstream of Shocks When Using Godunov-Type Schemes. I. Sources of Errors in One-Dimensional Problems

Alexander V. Rodionov*

Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Moscow, 125047, Russia.

RFNC – All-Russian Scientific Research Institute of Experimental Physics, Sarov, 607188, Russia.

Communicated by Chi-Wang Shu

Received 25 September 2022; Accepted (in revised version) 1 June 2023

Abstract. The article opens a series of publications devoted to a systematic study of numerical errors behind the shock wave when using high-order Godunov-type schemes, including in combination with the artificial viscosity approach. The proposed paper describes the numerical methods used in the study, and identifies the main factors affecting the accuracy of the solution for the case of one-dimensional gas dynamic problems. The physical interpretation of the identified factors is given and their influence on the grid convergence is analyzed.

AMS subject classifications: 65M08, 65M12, 35L65, 35L67

Key words: Shock-capturing methods, Riemann problem, WENO reconstruction, Runge-Kutta method, Hancock-Rodionov method, artificial viscosity approach, Shu-Osher test problem.

1 Introduction

When solving gas dynamic problems in the context of Euler equations, a shock wave is treated as a discontinuity, across which the Rankine-Hugoniot relations are valid. In shock-fitting techniques, shock waves (as well as contact discontinuities) are tracked explicitly, which in principle allows (in one-dimensional problems or in multidimensional problems with a simple wave configuration) obtaining a solution with a high accuracy. However, in multidimensional problems with complex flow structures the use of such methods is problematic.

*Corresponding author. *Email address:* avrodionov@rambler.ru (A. V. Rodionov)

In this connection, shock-capturing methods (schemes) that smear discontinuities over computing cells are in common use in computational fluid dynamics. In such a case, the shock front turns into a shock layer with a width of a few mesh spacing. Although the Rankine-Hugoniot conditions are not explicitly applied in such computations, one can expect that they are approximated integrally (over the shock layer) if the shock-capturing method approximates the conservation laws.

When a shock-capturing method is used, shock waves are smeared due to numerical viscosity (scheme dissipation), which in some sense mimics the effect of physical viscosity. Having functional similarity, numerical and physical viscosities have fundamental distinctions. So, taking into account the physical viscosity we approximate the Navier-Stokes equations within the shock layer. By refining the mesh, we improve the resolution of the shock layer, and the numerical solution tends to the exact solution of the Navier-Stokes equations. In the case of numerical viscosity, the mesh refinement does not lead to any gain in resolution of the shock layer, since its width decreases proportionally. Therefore, fundamental questions arise here: will the numerical solution with the mesh refinement converge to the exact solution of Euler equations supplemented by Rankine-Hugoniot relations across the shock, and, if so, what will be the rate of convergence?

A large number of publications are devoted to the study of such issues. Of the works known to the author, publications [1–16] are of particular interest. Among them, works [1, 4–6, 8, 9, 11, 13, 16] considered the Euler equations, while the remaining papers solved other nonlinear systems of hyperbolic equations, such as the shallow water equations.

Today, it is generally agreed that the convergence rate of high-order shock-capturing schemes reduces to the first order downstream of shocks. This assertion was verified numerically for the third-order Rusanov scheme [1], the second-order MUSCL-type schemes [3,5,9,16], the families of ENO and WENO schemes [2,4,5,9,11,12,14,16], the monotonicity preserving (MP) scheme [11, 16] and the discontinuous Galerkin (DG) method [15]. The relevant analytical studies can be found in [6–8,10,16]. To overcome this problem, the following techniques were suggested: the subcell resolution method [2], the matrix viscosity method [8], the fast sweeping approach [13] and the specific combined scheme [14]. However, these techniques were implemented only in one-dimensional model problems, and their generalization to complex multidimensional cases seems to be just as problematic as in the case of using shock-fitting techniques.

Although the problem of the solution accuracy behind the shock wave has been attracting the attention of researchers for several decades, it is not yet sufficiently studied. This is due to both the wide variety of shock-capturing methods and the problems solved with their help, and the complex nature of the error sources inherent in such methods.

The present article opens a series of publications in which the author intends to report the results of his systematic study of this problem in relation to high-order Godunov-type schemes, including in combination with the artificial viscosity approach [17–20] (this approach not only cures the carbuncle phenomenon, but also reduces substantially the post-shock oscillations). The proposed article (1) describes the numerical methods used in the study and (2), by the example of solving one-dimensional gas dynamic problems, identifies the main sources of errors (factors affecting the solution accuracy downstream of shocks), gives their physical interpretation and analyses their influence on grid convergence. The results of studying the problem in multidimensional simulations will be presented in subsequent publications.

2 Gasdynamic equations and numerical methods

A variety of schemes, which can be classified as high-order Godunov-type schemes, have been developed to date. As distinct from first-order schemes, they differ not only in their Riemann solvers, but also in a variety of techniques for improving their accuracy in space and time. Moreover, the effect of applying a specific high-order scheme to some test problem can also depend on minor (nonprincipal) features of its realization. For this reason, it does not seem possible to conduct a study of the problem under discussion on the whole variety of this class of schemes.

Instead, we restrict the proposed study to considering implementations of some well-known schemes. This section provides their description by the example of solving two-dimensional gas dynamic problems.

2.1 The Euler equations

The equations of compressible gas dynamics in Cartesian coordinates xy have the form

$$\frac{\partial \mathbf{U}}{\partial t}+\frac{\partial \mathbf{F}_x}{\partial x}+\frac{\partial \mathbf{F}_y}{\partial y}=\text{RHS}, \quad \mathbf{U}=\begin{bmatrix}\rho\\ \rho u_x\\ \rho u_y\\ \rho e_0\end{bmatrix}, \quad \mathbf{F}_x=\begin{bmatrix}\rho u_x\\ \rho u_x^2+p\\ \rho u_x u_y\\ \rho u_x h_0\end{bmatrix}, \quad \mathbf{F}_y=\begin{bmatrix}\rho u_y\\ \rho u_x u_y\\ \rho u_y^2+p\\ \rho u_y h_0\end{bmatrix}, \tag{2.1}$$

where t is the time, $\mathbf{u}=(u_x,u_y)$ are the velocity vector and its components, ρ is the density, p is the pressure, $e_0=(u_x^2+u_y^2)/2+e$ is the specific total energy, $e=e(p,\rho)$ is the specific energy, $h_0=(u_x^2+u_y^2)/2+h$ is the specific total enthalpy, $h=h(p,\rho)=e+p/\rho$ is the specific enthalpy, RHS = 0 for the Euler equations (the inviscid gas model).

For a polytropic gas, the specific energy and specific enthalpy are written as

$$e=\frac{p}{(\gamma-1)\rho}, \quad h=\frac{\gamma p}{(\gamma-1)\rho},$$

with γ denoting the ratio of specific heats (constant value).

Remark 2.1. In the framework of the proposed study, in all test cases, we will consider, unless otherwise stated, the flow of a perfect gas with γ = 1.4.

2.2 Space-time discretization

We apply the finite volume approach to solve (2.1) numerically. In so doing we consider a sufficiently smooth structured grid and introduce curvilinear coordinates $\xi\eta$ that transform the grid in physical space xy to a rectangular grid in computational space with the grid spacing $\Delta\xi=\Delta\eta=1$. The following geometric parameters of the grid will take the place of metrics (hereinafter the grid indices i and j correspond to the coordinates ξ and η): $V_{i,j}$, the cell volume; $(\mathbf{S}_\xi)_{i+1/2,j}=(S_{\xi x},S_{\xi y})_{i+1/2,j}$ and $(\mathbf{S}_\eta)_{i,j+1/2}=(S_{\eta x},S_{\eta y})_{i,j+1/2}$, the area vectors of the cell faces in two grid directions (each area vector points in the direction of increase in the associated coordinate).

Assume that the solution at a given time level n is defined, meaning that for the time $t=t^n$ we know the average values $\mathbf{U}_{i,j}^n$ within each cell. In order to update the solution to the next time level $n+1$ we integrate (2.1) over the cell and the time between t^n and $t^{n+1}=t^n+\Delta t$. The result is

$$\mathbf{U}_{i,j}^{n+1}=\mathbf{U}_{i,j}^n-\frac{\Delta t}{V_{i,j}}\left[(\mathbf{F}_\xi)_{i+1/2,j}-(\mathbf{F}_\xi)_{i-1/2,j}+(\mathbf{F}_\eta)_{i,j+1/2}-(\mathbf{F}_\eta)_{i,j-1/2}\right], \tag{2.2}$$

where $\mathbf{F}_\xi=S_{\xi x}\mathbf{F}_x+S_{\xi y}\mathbf{F}_y$ and $\mathbf{F}_\eta=S_{\eta x}\mathbf{F}_x+S_{\eta y}\mathbf{F}_y$ are the flux vectors across the cell faces in two grid directions.

The way of computing the fluxes depends on the specific scheme. In the first-order Godunov-type schemes we assume that the flow variables within each computing cell are constant, and the fluxes are computed from one of the Riemann solvers (the solution of the Riemann problem): the exact solver, as in the Godunov scheme, or the approximate solver, as in the Roe, HLLC and other schemes. Any of these solutions can be written as (hereinafter RS = Riemann solver)

$$(\mathbf{F}_\xi)_{i+1/2,j}=\mathbf{F}^{RS}\left(\mathbf{Q}_{i,j}^n,\mathbf{Q}_{i+1,j}^n,(\mathbf{S}_\xi)_{i+1/2,j}\right),\quad (\mathbf{F}_\eta)_{i,j+1/2}=\mathbf{F}^{RS}\left(\mathbf{Q}_{i,j}^n,\mathbf{Q}_{i,j+1}^n,(\mathbf{S}_\eta)_{i,j+1/2}\right),$$

where $\mathbf{Q}=\left[u_x,u_y,\rho,p\right]^T$ and $\mathbf{Q}_{i,j}^n=\mathbf{Q}\left(\mathbf{U}_{i,j}^n\right)$.

Remark 2.2. Riemann solvers are, although not the main, but obligatory elements of high-order Godunov-type schemes. In the framework of the proposed study, we will use, unless otherwise stated, the HLLC solver [21]. Among approximate solvers, it is one of the most consistent with the exact Riemann solver.

2.3 Reconstructions

Increasing the accuracy of the Godunov-type schemes is usually obtained by data reconstruction.

Piecewise linear reconstructions. In the second-order MUSCL-type schemes, the piecewise linear distribution of data is used instead of the piecewise constant distribution. The distributions of flow variables are reconstructed based on the known values of $\mathbf{Q}_{i,j}^n$ positioned at the cell centers. On a sufficiently smooth structured grid, the reconstruction process reduces to calculating the so-called slopes, $\Delta_\xi \mathbf{Q}_{i,j}^n$ and $\Delta_\eta \mathbf{Q}_{i,j}^n$; these values represent increments of flow variables inside the cell along the grid lines *i* and *j* (i.e. along the directions ξ and η). The boundary extrapolated values are then found as

$$\begin{aligned}
&\mathbf{Q}_{i+1/2-,j}^n = \mathbf{Q}_{i,j}^n + \frac{1}{2}\Delta_\xi \mathbf{Q}_{i,j}^n, \quad \mathbf{Q}_{i-1/2+,j}^n = \mathbf{Q}_{i,j}^n - \frac{1}{2}\Delta_\xi \mathbf{Q}_{i,j}^n,\\
&\mathbf{Q}_{i,j+1/2-}^n = \mathbf{Q}_{i,j}^n + \frac{1}{2}\Delta_\eta \mathbf{Q}_{i,j}^n, \quad \mathbf{Q}_{i,j-1/2+}^n = \mathbf{Q}_{i,j}^n - \frac{1}{2}\Delta_\eta \mathbf{Q}_{i,j}^n,
\end{aligned}$$

and for the fluxes we can write

$$\begin{aligned}
&(\mathbf{F}_\xi)_{i+1/2,j} = \mathbf{F}^{RS}\left(\mathbf{Q}_{i+1/2-,j}^n, \mathbf{Q}_{i+1/2+,j}^n, (\mathbf{S}_\xi)_{i+1/2,j}\right),\\
&(\mathbf{F}_\eta)_{i,j+1/2} = \mathbf{F}^{RS}\left(\mathbf{Q}_{i,j+1/2-}^n, \mathbf{Q}_{i,j+1/2+}^n, (\mathbf{S}_\eta)_{i,j+1/2}\right).
\end{aligned}$$

The slopes are computed by the same algorithm for each grid direction independently (one-dimensional reconstruction). Note that instead of the term *reconstruction*, the term *limiter* is also used, since such algorithms (explicitly or implicitly) impose certain restrictions on the value of the calculated slope.

Several algorithms of piecewise linear reconstruction will be used in our study. We start their description with two basic limiters – minmod [22, 23] and MC [24] (MC = monotonized central-difference limiter, the name borrowed from [25, Chapter 6]). According to these limiters, the slopes of a certain variable f along the i-th grid direction can be written as (hereinafter, the unchanged indices j are omitted for convenience)

$$(\Delta f_i)^{\text{minmod}} = \text{minmod}\left[\Delta f_{i-1/2}, \Delta f_{i+1/2}\right], \tag{2.3}$$

$$(\Delta f_i)^{\text{MC}} = \text{minmod}\left[\Delta f_i, 2\text{minmod}(\Delta f_{i-1/2}, \Delta f_{i+1/2})\right], \tag{2.4}$$

where $\Delta f_{i-1/2} = f_i - f_{i-1}$, $\Delta f_i = (f_{i+1} - f_{i-1})/2$, and the minmod function is

$$\text{minmod}(a,b) = \frac{1}{2}\left[\text{sgn}(a) + \text{sgn}(b)\right]\min(|a|,|b|).$$

The reconstructions based on the limiters (2.3) and (2.4) belong to the class of three-point reconstructions satisfying the TVD condition (TVD = total variation diminishing). The MC limiter is less dissipative than the minmod limiter, however, like all TVD limiters, it leads to a clipping of the solution near local extrema (a TVD scheme degenerates to the first order of accuracy at some points near extrema).

The noted disadvantage is overcome by replacing the TVD limiters with limiters that satisfy the less burdensome TVB property (TVB = total variation bound). Although such limiters allow a limited increase in the total variation of the solution (in the presence of local extrema), they nevertheless preserve the scheme's monotonicity (the aforesaid relates to the case of the linear advection equation).

Within the framework of our study, two five-point reconstructions belonging to the class of TVB limiters will be used; they were previously described by the author in [26]. The first reconstruction is a modification of the MC limiter, differing from it only at points near local extrema (where the TVB condition replaces the TVD condition); therefore, we will call it the MC+ limiter. The MC+ limiter is written as follows

$$(\Delta f_i)^{\text{MC+}} = \text{minmod}\left[\Delta f_i, 2\text{minmod}\left(\Delta f^-_{i-1/2}, \Delta f^+_{i+1/2}\right)\right], \tag{2.5}$$

where

$$\Delta f^-_{i-1/2} = \begin{cases} \Delta f_{i-1/2} & \text{if } \Delta f_i \Delta f_{i-1} \geq 0, \\ \Delta f_{i-1/2} - 0.5\Delta f_{i-1} & \text{if } \Delta f_i \Delta f_{i-1} < 0, \end{cases}$$

$$\Delta f^+_{i+1/2} = \begin{cases} \Delta f_{i+1/2} & \text{if } \Delta f_i \Delta f_{i+1} \geq 0, \\ \Delta f_{i+1/2} - 0.5\Delta f_{i+1} & \text{if } \Delta f_i \Delta f_{i+1} < 0. \end{cases}$$

The second reconstruction (limiter), called NOLD (NOLD = non-oscillatory low-dissipative), consists of several steps. First, the slope limits inside the cell are calculated

$$\Delta f_i^{min} = \min(0, \Delta f_{i-1/2}, \Delta f_{i+1/2}), \quad \Delta f_i^{max} = \max(0, \Delta f_{i-1/2}, \Delta f_{i+1/2}),$$

which are then modified into

$$\Delta f_i^{min2} = \max\left(2\Delta f_{i+1/2} - \Delta f_{i+1}^{max}, 2\Delta f_{i-1/2} - \Delta f_{i-1}^{max}\right),$$
$$\Delta f_i^{max2} = \min\left(2\Delta f_{i+1/2} - \Delta f_{i+1}^{min}, 2\Delta f_{i-1/2} - \Delta f_{i-1}^{min}\right).$$

Now the condition $\Delta f_i^{min2} < \Delta f_i^{max2}$ is checked and then

$$(\Delta f_i)^{\text{NOLD}} = \begin{cases} \text{median}\left[\Delta f_i, \Delta f_i^{min2}, \Delta f_i^{max2}\right] & \text{if } \Delta f_i^{min2} < \Delta f_i^{max2}, \\ \text{minmod}\left[\Delta f_i^{min2}, \Delta f_i^{max2}\right] & \text{otherwise.} \end{cases} \tag{2.6}$$

Here, of the three arguments, the median function selects the one that lies between the two remaining ones (if at least two arguments are equal, then their value will be chosen as the "median"). This function can be written as follows (the order of the function arguments does not matter)

$$\text{median}(a,b,c) = a + \text{minmod}(b-a, c-a).$$

Remark 2.3. The NOLD limiter has a number of positive points that were demonstrated in [26]. In particular, it possesses enhanced dissipative properties on discontinuous solutions (as compared to other limits), which makes the contact discontinuities less smeared.

Reconstructions based on characteristic variables. The effect of the reconstruction depends not only on the chosen slope limiter, but also on the variables, based on which the reconstruction is performed. The most straightforward way is to perform a reconstruction based on *primitive variables*, i.e. components of the vector $\mathbf{Q}$. This way, however, often leads to poor results. It is more consistent to perform a reconstruction based on *characteristic variables*. In our study, this approach is implemented as follows.

Suppose we need to perform a reconstruction in cell (i,j) in grid direction ξ. First, the required sets of primitive slopes are calculated: $\Delta_\xi \mathbf{Q}^n_{i-1/2,j} = \mathbf{Q}^n_{i,j} - \mathbf{Q}^n_{i-1,j}$, $\Delta_\xi \mathbf{Q}^n_{i,j} = (\mathbf{Q}^n_{i+1,j} - \mathbf{Q}^n_{i-1,j})/2$ and so on. Then, using the transformation matrix $\mathbf{A} = (\partial \mathbf{Z}/\partial \mathbf{Q})^n_{i,j}$ they are converted into characteristic slopes: $\Delta_\xi \mathbf{Z}^n_{i-1/2,j}$, $\Delta_\xi \mathbf{Z}^n_{i,j}$ and so on. Next, one of the limiters (2.3) – (2.6) is applied to the characteristic slopes and the resulting vector $(\Delta_\xi \mathbf{Z}^n_{i,j})^{lim}$ (here *lim* is the limiter name) is converted into $(\Delta_\xi \mathbf{Q}^n_{i,j})^{lim}$ by the inverse transformation matrix $\mathbf{A}^{-1}$; this completes the reconstruction process. The two transformation matrices are given by

$$\mathbf{A} = \begin{bmatrix} n_x & n_y & 0 & 1/\rho a \\ n_x & n_y & 0 & -1/\rho a \\ 0 & 0 & 1 & -1/a^2 \\ n_y & -n_x & 0 & 0 \end{bmatrix}, \quad \mathbf{A}^{-1} = \begin{bmatrix} n_x/2 & n_x/2 & 0 & n_y \\ n_y/2 & n_y/2 & 0 & -n_x \\ \rho/2a & -\rho/2a & 1 & 0 \\ \rho a/2 & -\rho a/2 & 0 & 0 \end{bmatrix},$$

where $a = \sqrt{\gamma p/\rho}$ is the sound speed, $\mathbf{n} \equiv (n_x, n_y) = \mathbf{S}_\xi / |\mathbf{S}_\xi|$ is normal to the cell interface.

Fifth-order reconstructions. In our study we also employ two fifth-order reconstructions, WENO5 and MP5, detailed descriptions of which can be found in [27] and [28] (their FORTRAN codes are given in the latter work). Both reconstructions are applied to characteristic variables based on conservative variables, but in each case it is done differently.

In the WENO5 reconstruction (WENO = weighted essentially non-oscillatory), the values of $\mathbf{U}_{i+1/2-,j}$ and $\mathbf{U}_{i+1/2+,j}$ (vectors of conservative variables on both sides of the cell interface) are calculated using forward and inverse transformation matrices, $\mathbf{B} = (\partial \mathbf{Z}/\partial \mathbf{U})^n_{i+1/2,j}$ and $\mathbf{B}^{-1}$, which are the matrices of left and right eigenvectors of the Jacobian $\partial \mathbf{F}_\xi / \partial \mathbf{U}$ defined at the cell interface $(i+1/2,j)$ (frozen values). The transformation matrices have the form

$$\mathbf{B} = \begin{bmatrix} (b_2 + u_1/a)/2 & -(b_1 u_x + n_x/a)/2 & -(b_1 u_y + n_y/a)/2 & b_1/2 \\ 1 - b_2 & b_1 u_x & b_1 u_y & -b_1 \\ (b_2 - u_1/a)/2 & -(b_1 u_x - n_x/a)/2 & -(b_1 u_y - n_y/a)/2 & b_1/2 \\ -u_2 & -n_y & n_x & 0 \end{bmatrix},$$

$$\mathbf{B}^{-1} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ u_x - a n_x & u_x & u_x + a n_x & -n_y \\ u_y - a n_y & u_y & u_y + a n_y & n_x \\ h_o - a u_1 & u^2/2 & h_0 + a u_1 & u_2 \end{bmatrix},$$

where

$$b_1=(\gamma-1)/a^2, \quad b_2=b_1u^2/2, \quad h_0=u^2/2+1/b_1,$$
$$u^2=u_x^2+u_y^2, \quad u_1=u_xn_x+u_yn_y, \quad u_2=u_yn_x-u_xn_y.$$

In the MP5 reconstruction (MP = monotonicity preserving), the same transformation matrices are calculated in the cell center (i,j); they are used for computing the values of $\mathbf{U}_{i-1/2+,j}$ and $\mathbf{U}_{i+1/2-,j}$, the vectors of conservative variables at the inner sides of the opposite interfaces of the cell.

Remark 2.4. Note that changing the location of applying (freezing) the transformation matrices (from the cell interface to the cell center or vice versa) does not have a material effect on the results obtained by a specific reconstruction (WENO5 or MP5). Nevertheless, we will used the original versions proposed in [27] (for WENO5) and in [28] (for MP5).

2.4 Time integration

In our study we will used several techniques to increase the accuracy in time.

Runge-Kutta methods. Let us rewrite the equation (2.2) in a compact form:

$$\mathbf{U}_{i,j}^{n+1}=\mathbf{U}_{i,j}^{n}+\Delta t\mathbf{L}_{i,j}(\mathbf{Q}^n), \tag{2.7}$$

where the spatial operator $\mathbf{L}_{i,j}$ expresses the total flux across all the cell interfaces. Its calculation implies a data reconstruction in each grid direction and solution of the Riemann problem on each cell interface.

The use of Runge-Kutta methods for solving Euler equations (or other systems of partial differential equations) involves dividing the discretization procedure into two stages: first discretizing only in space (it is described by the operator $\mathbf{L}_{i,j}$), and then discretizing in time using one of the methods for systems of ordinary differential equations (see [29] and [25, Section 10.4] for details). Within such a concept, Eq. (2.7) presents the forward Euler time discretization method which is first-order accurate.

In the Runge-Kutta methods, the accuracy in time is increased using the stepping procedure. In our study we will use second- and third-order TVD Runge-Kutta methods [29] (RK2 and RK3, respectively).

The RK2 method has the following form

$$\mathbf{U}_{i,j}^{(1)}=\mathbf{U}_{i,j}^{n}+\Delta t\mathbf{L}_{i,j}(\mathbf{Q}^n),$$
$$\mathbf{U}_{i,j}^{n+1}=\frac{1}{2}\mathbf{U}_{i,j}^{n}+\frac{1}{2}\mathbf{U}_{i,j}^{(1)}+\frac{1}{2}\Delta t\mathbf{L}_{i,j}\left(\mathbf{Q}^{(1)}\right).$$

The RK3 method has the form

$$\begin{aligned}
\mathbf{U}_{i,j}^{(1)} &= \mathbf{U}_{i,j}^{n} + \Delta t \mathbf{L}_{i,j}(\mathbf{Q}^{n}), \\
\mathbf{U}_{i,j}^{(2)} &= \frac{3}{4}\mathbf{U}_{i,j}^{n} + \frac{1}{4}\mathbf{U}_{i,j}^{(1)} + \frac{1}{4}\Delta t \mathbf{L}_{i,j}\left(\mathbf{Q}^{(1)}\right), \\
\mathbf{U}_{i,j}^{n+1} &= \frac{1}{3}\mathbf{U}_{i,j}^{n} + \frac{2}{3}\mathbf{U}_{i,j}^{(2)} + \frac{2}{3}\Delta t \mathbf{L}_{i,j}\left(\mathbf{Q}^{(2)}\right).
\end{aligned}$$

Note that compared to the Euler method, the RK2 and RK3 methods are respectively two and three times more time-consuming.

Hancock-Rodionov method. To achieve the second order of accuracy in time, other methods can be used, including the predictor-corrector technique described in [30, 31]. This specific technique is a kind of the Hancock method [32, 33], so in our study we call it the Hancock-Rodionov method (or HR method for short).

The predictor-corrector procedure of the HR scheme can be written in its compact form as

$$\begin{aligned}
&\hat{\mathbf{U}}_{i,j}^{n+1} = \mathbf{U}_{i,j}^{n} + \Delta t \mathbf{L}_{i,j}^{pred}(\mathbf{Q}^{n}), \\
&\mathbf{U}_{i,j}^{n+1} = \mathbf{U}_{i,j}^{n} + \Delta t \mathbf{L}_{i,j}^{corr}\left(\mathbf{Q}^{n+1/2}\right), \quad \mathbf{Q}_{i,j}^{n+1/2} = \frac{1}{2}\left(\mathbf{Q}_{i,j}^{n} + \hat{\mathbf{Q}}_{i,j}^{n+1}\right).
\end{aligned}$$

The spatial operators $\mathbf{L}_{i,j}^{pred}$ and $\mathbf{L}_{i,j}^{corr}$ are different here. As distinct from $\mathbf{L}_{i,j}$, the operator $\mathbf{L}_{i,j}^{pred}$ does not involve solving the Riemann problem. Instead, the fluxes are calculated based on the boundary extrapolated values, which are interior with respect to the cell being integrated. Thus, at the predictor step, the fluxes for the cell (i,j) are calculated by simple relations:

$$\begin{aligned}
&(\mathbf{F}_{\xi})_{i-1/2,j} = \mathbf{F}\left(\mathbf{Q}_{i-1/2+,j}^{n}, (\mathbf{S}_{\xi})_{i-1/2,j}\right), \quad (\mathbf{F}_{\xi})_{i+1/2,j} = \mathbf{F}\left(\mathbf{Q}_{i+1/2-,j}^{n}, (\mathbf{S}_{\xi})_{i+1/2,j}\right), \\
&(\mathbf{F}_{\eta})_{i,j-1/2} = \mathbf{F}\left(\mathbf{Q}_{i,j-1/2+}^{n}, (\mathbf{S}_{\eta})_{i,j-1/2}\right), \quad (\mathbf{F}_{\eta})_{i,j+1/2} = \mathbf{F}\left(\mathbf{Q}_{i,j+1/2-}^{n}, (\mathbf{S}_{\eta})_{i,j+1/2}\right).
\end{aligned}$$

At the corrector step, we use the operator $\mathbf{L}_{i,j}^{corr}$, which involves solving the Riemann problem (like in the basic operator $\mathbf{L}_{i,j}$), but do not recompute the slopes: at an intermediate time level $n+1/2$, the same slopes are used as those obtained at the predictor step ($\Delta_{\xi}\mathbf{Q}_{i,j}^{n}$ and $\Delta_{\eta}\mathbf{Q}_{i,j}^{n}$).

Remark 2.5. Although the HR method requires less computing time than the RK2 method, it usually provides better accuracy. Thus, in the case of solving the linear advection equation, the HR method, unlike the RK2 method, is exact when the Courant number is one. See also the results of comparative studies in Section 2.8.

2.5 Constraint on the time step

In this study we consider explicit finite-difference schemes, for which there exists a constraint on the time step Δt (the Courant-Friedrichs-Lewy condition). For the Euler equations we write this condition in the form

$$\Delta t = C_{cfl}\cdot\min_{i,j}\left\{(\Delta t^{conv})_{i,j}\right\},\quad \Delta t^{conv}=\frac{V}{\left|(\mathbf{u}\cdot\mathbf{S}_\xi)\right|+a|\mathbf{S}_\xi|+\left|(\mathbf{u}\cdot\mathbf{S}_\eta)\right|+a|\mathbf{S}_\eta|},\tag{2.8}$$

where C_{cfl} is the Courant number ($C_{cfl}\le 1$).

Remark 2.6. In the framework of the proposed study, all simulations by the HR and RK2 methods will be done, unless otherwise stated, with the Courant number of 0.9; for the RK3 method we take $C_{cfl}=0.6$.

2.6 Navier-Stokes equations and their approximation

In the case of using the Navier-Stokes equations, the right-hand side in (2.1) reads

$$\mathrm{RHS}=\frac{\partial\mathbf{F}_x^{\mathrm{V}}}{\partial x}+\frac{\partial\mathbf{F}_y^{\mathrm{V}}}{\partial y},\quad \mathbf{F}_x^{\mathrm{V}}=\begin{bmatrix}0\\ \tau_{xx}\\ \tau_{yx}\\ u_x\tau_{xx}+u_y\tau_{yx}-q_x\end{bmatrix},\quad \mathbf{F}_y^{\mathrm{V}}=\begin{bmatrix}0\\ \tau_{xy}\\ \tau_{yy}\\ u_x\tau_{xy}+u_y\tau_{yy}-q_y\end{bmatrix}.\tag{2.9}$$

The components of the stress tensor ($\boldsymbol{\tau}$) and the heat flux ($\mathbf{q}$) can be written as

$$\tau_{xx}=\mu\left(\frac{4}{3}\frac{\partial u_x}{\partial x}-\frac{2}{3}\frac{\partial u_y}{\partial y}\right),\quad \tau_{xy}=\tau_{yx}=\mu\left(\frac{\partial u_x}{\partial y}+\frac{\partial u_y}{\partial x}\right),\quad \tau_{yy}=\mu\left(\frac{4}{3}\frac{\partial u_y}{\partial y}-\frac{2}{3}\frac{\partial u_x}{\partial x}\right),$$
$$q_x=-\lambda\frac{\partial T}{\partial x}=-\frac{\mu}{\mathrm{Pr}}\frac{\partial h}{\partial x},\quad q_y=-\lambda\frac{\partial T}{\partial y}=-\frac{\mu}{\mathrm{Pr}}\frac{\partial h}{\partial y},$$

where T is the gas temperature, μ and λ are the viscosity and thermal conductivity coefficients related by the equation $\lambda=C_p\mu/\mathrm{Pr}$, $C_p=\partial h/\partial T$ is the specific heat capacity at constant pressure, Pr is the Prandtl number.

The spatial discretization of (2.9) takes the form

$$\mathrm{RHS}=\frac{1}{V_{i,j}}\left[\left(\mathbf{F}_\xi^{\mathrm{V}}\right)^n_{i+1/2,j}-\left(\mathbf{F}_\xi^{\mathrm{V}}\right)^n_{i-1/2,j}+\left(\mathbf{F}_\eta^{\mathrm{V}}\right)^n_{i,j+1/2}-\left(\mathbf{F}_\eta^{\mathrm{V}}\right)^n_{i,j-1/2}\right],$$

where $\mathbf{F}_\xi^{\mathrm{V}}=S_{\xi x}\mathbf{F}_x^{\mathrm{V}}+S_{\xi y}\mathbf{F}_y^{\mathrm{V}}$ and $\mathbf{F}_\eta^{\mathrm{V}}=S_{\eta x}\mathbf{F}_x^{\mathrm{V}}+S_{\eta y}\mathbf{F}_y^{\mathrm{V}}$.

When computing the components of viscous fluxes $\mathbf{F}_\xi^{\mathrm{V}}$ and $\mathbf{F}_\eta^{\mathrm{V}}$, explicit central difference approximations are used (see [20] for details). The resulting value of the right-hand side of Navier-Stokes equations is introduced into the operator $\mathbf{L}_{i,j}$, or any of its variants, as explicit source terms, which remain frozen (not recomputed) at all stages of marching the solution to the next time level $n+1$.

In case of solving the Navier-Stokes equations we should account for an additional constraint on the time step, and relation (2.8) then changes to

$$\Delta t = C_{cfl} \cdot \min_{i,j} \left\{ \left[\frac{1}{(\Delta t^{conv})_{i,j}} + \frac{1}{(\Delta t^{diff})_{i,j}} \right]^{-1} \right\}, \quad \Delta t^{diff} = \frac{3\rho}{8\mu} \cdot \frac{V^2}{|\mathbf{S}_\xi|^2 + |\mathbf{S}_\eta|^2}. \tag{2.10}$$

Remark 2.7. Although in the framework of the proposed research we will study the properties of shock-capturing methods when applied to the Euler equations, the Navier-Stokes equations (their right-hand side) will also be used in computations in some cases. The first of such cases concerns the use of a basic version of the artificial viscosity approach, the second is the use of a reference model for shock-capturing methods. The details of these techniques will be disclosed, respectively, in Sections 2.7 and 3.1.

2.7 Artificial viscosity approach

When applying Godunov-type methods for multidimensional simulations, a specific problem, called the *carbuncle phenomenon* or the *shock-wave instability*, may occur. This kind of instability arises if the following conditions are met: (1) the physical viscosity is zero or its influence on smearing the shock is negligibly small (compare to the numerical viscosity); (2) the shock Mach number exceeds some threshold value that depends on the numerical scheme and the problem specification; (3) the grid (one family of grid lines) aligns with the flow lines at the shock front; (4) the numerical scheme employs a *complete* Riemann solver (a Riemann solver that can capture a contact discontinuity).

In our study we will use the artificial viscosity approach (or AV approach for short) to cure the carbuncle phenomenon. This approach was developed by the author in [17–20], and it has two versions: *basic* and *simplified*.

Basic version. The idea of the original AV approach is to introduce some dissipation in the form of right-hand side of the Navier-Stokes equations into the basic method of solving Euler equations. In so doing, the molecular viscosity coefficient is replaced by the artificial viscosity coefficient, and the Prandtl number is set equal to 3/4. As suggested in [17], the expression for the artificial viscosity coefficient reads

$$\mu_{AV} = \begin{cases} C_{AV}\rho\Delta l^2 \sqrt{(\nabla\mathbf{u})^2 - (C_{th}a/\Delta l)^2}, & \text{if } -(\nabla\mathbf{u}) > C_{th}a/\Delta l, \\ 0, & \text{otherwise,} \end{cases} \tag{2.11}$$

where Δl is the characteristic mesh size, C_{AV} is the dimensionless parameter, $C_{th} = 0.05$ is the coefficient in the compression intensity threshold that restricts the effect of artificial viscosity to the shock layer only. In the two-dimensional case with non-square cells the characteristic mesh size is computed as $\Delta l = \max(d_1, d_2)\sqrt{2}$, with d_1 and d_2 denoting the lengths of the cell diagonals. For rectangular cells in Cartesian coordinates it reduces to $\Delta l = \sqrt{(\Delta x^2 + \Delta y^2)/2}$.

The artificial viscosity coefficient is first defined at the cell centers, and its values at the cell interfaces are then found by linear interpolation: $f_{i+1/2,j} = 0.5(f_{i,j} + f_{i+1,j})$. The velocity vector divergence in (2.11) is calculated by Gauss's theorem; in so doing the values of the velocity vector components at the cell interfaces are also found by linear interpolation between the cell center values.

For the principal coefficient of the AV approach the value of $C_{AV} = 0.5$ chosen in [19] ensures the suppression of shock instability if two recommendations for the high-order schemes are followed. The first recommendation relates to the procedure of data reconstruction; it is essential to use the characteristic variables here (as described in Section 2.3). The second recommendation refers to the data reconstruction in the cells that fall inside the shock layer. One should adhere to the concept of piecewise linear distribution of the data within such cells and use the minmod limiter (see [19] for details). To denote the latter technique, we add the character (m) to the basic reconstruction: MC+(m), WENO5(m) and so on.

Simplified version. Considering that many of the existing CFD codes support simulations in the framework of both Euler and Navier-Stokes equations, the basic AV approach could be readily implemented within such codes. However, introducing this approach into the codes developed solely for solving the Euler equations may seem to be overcomplicated (see the cumbersome expressions for the viscous fluxes in [20]). In this case, on can use a simplified version of the artificial viscosity approach (sAV approach for short), which is comparable to the original one in its efficiency.

In the sAV method, the viscous flux vector in a particular grid direction (ξ-direction, for instance) has a simple form

$$\mathbf{F}_{\xi}^{\mathrm{V}} = \left[\ 0,\quad \hat{\tau}_{x\xi},\quad \hat{\tau}_{y\xi},\quad \hat{q}_{\xi}\ \right]^{T},\quad \hat{\tau}_{x\xi} = \frac{\mu S_{\xi}^{2}}{V}\frac{\partial u_x}{\partial \xi},\quad \hat{\tau}_{y\xi} = \frac{\mu S_{\xi}^{2}}{V}\frac{\partial u_y}{\partial \xi},\quad \hat{q}_{\xi} = \frac{\mu S_{\xi}^{2}}{V}\frac{1}{\mathrm{Pr}}\frac{\partial h}{\partial \xi},$$

and the space derivatives at the cell interface $(i+1/2,j)$ are approximated as $\partial f/\partial \xi = f_{i+1,j} - f_{i,j}$, where $f = u_x, u_y, h$.

All other elements of the basic AV approach are retained in the sAV approach without modification.

Remark 2.8. The artificial viscosity approach not only cures the carbuncle phenomenon, but also improves the solution behind shocks in both one-dimensional and multidimensional simulations. Besides, it can be used in CFD codes without switching or tuning to any specific problem: the artificial viscosity terms will be introduced only in those computational cells that are located within the shock layer. Because of this, in a shock-free problem the artificial viscosity approach will have no impact on the solution.

2.8 Testing the schemes on a shock-free problem

Before studying the numerical errors behind the shock wave, let us demonstrate the accuracy of the above-described schemes in the case of simulating a shock-free flow. For

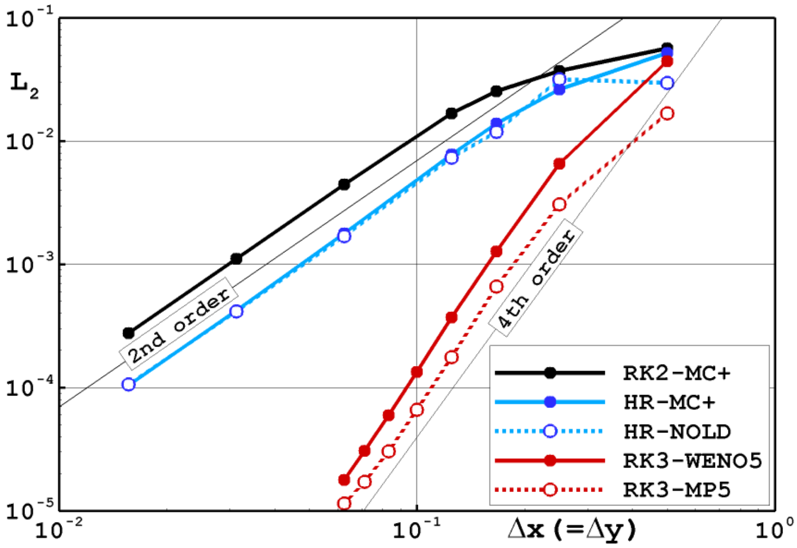

Figure 1: Vortex evolution problem at $t = 50$: convergence of density errors with the mesh size.

this purpose, the two-dimensional vortex evolution problem in the following formulation was chosen.

The computational domain of $[0,10]\times[0,10]$ in xy plane is covered with a square mesh of size $\Delta x = \Delta y$; the boundary conditions are periodic in the Cartesian directions. The initial flow field in the domain is a superposition of two flow patterns: a mean diagonal flow with $u_{x1} = u_{y1} = \rho_1 = p_1 = 1$ and an isentropic vortex (perturbation of the mean flow) defined as

$$V(r) = \frac{r}{r_0} f, \quad f = V_m \exp\left\{\frac{1-(r/r_0)^2}{2}\right\},$$

where $V(r)$ is the vortex velocity profile, $V_m = 5/2\pi$ is the maximum value in the vortex velocity profile (at $r = r_0$), $r = \sqrt{(x-x_c)^2+(y-y_c)^2}$ is the distance to the vortex center, $(x_c,y_c) = (5,5)$ are the coordinates of the vortex center, $r_0 = 1$ is the vortex effective radius.

In line with this definition, and assuming anticlockwise rotation of the vortex, the initial flow values are distributed over the computational domain as follows

$$u_x = u_{x1} - \frac{y-y_c}{r_0} f, \quad u_y = u_{y1} + \frac{x-x_c}{r_0} f, \quad \rho = \rho_1 \left(1 - \frac{(\gamma-1)\rho_1}{2\gamma p_1} f^2\right)^{1/(\gamma-1)}, \quad p = p_1 \left(\frac{\rho}{\rho_1}\right)^{\gamma}.$$

The vortex evolution problem was computed by the schemes under consideration until the time $t = 50$ (the vortex crossed the domain boundaries and returned to the initial position five times; the exact solution corresponds to the initial data). Fig. 1 shows the grid convergence study for density errors in the L_2-norm. As one can see, the second-order schemes (RK2 and HR methods with the MC+ and NOLD limiters) achieve their formal order of accuracy; in so doing, the RK2 method is inferior in accuracy to the HR method by about 2.5 times. The RK3 method with the WENO5 and MP5 reconstructions is much more accurate, revealing the convergence rate between four and five. It should be noted here that in terms of the CPU time the RK3-WENO5 and RK3-MP5 schemes are much more expensive as compared with the second-order schemes: by nearly a factor of 5. Nevertheless, for the problem under consideration this rise in the cost seems justified.

3 Numerical results and their analysis

Now let us move on to solving test problems with discontinuities using the schemes described in the previous section. These are the HR method with the MC+ and NOLD limiters and the RK3 method with the WENO5 and MP5 reconstructions. They will be applied either in basic versions or in combination with the AV approach as described in Section 2.7.

In this section, we will limit ourselves to considering one-dimensional problems. We will move from one test problem to another (in the order of increasing complexity), revealing and analyzing various factors affecting the solution accuracy downstream of shocks. When analyzing the disclosed factors, we will use a specially designed *reference model*, with the description of which we begin this section.

3.1 Reference model for shock-capturing methods

Although the shock-capturing methods related to the class of Godunov-type schemes are extremely diverse, they usually behave in a similar way when simulating shocks. By this we mean the following: there is a limited set of basic factors that affect the solution accuracy behind the shock, and the level of each factor is governed by a specific scheme. In order to reveal and analyze these basic factors in our study, it is useful to build a reference model that would include the main features of shock-capturing methods, but at the same time would be free of the properties of a specific scheme. As such a reference model, it seems reasonable to propose a model that has a physical basis. Let the shock front be smeared not due to numerical viscosity, but due to physical viscosity (by default, in association with thermal conductivity), and the effect of physical viscosity will be limited only to the shock layer and its immediate vicinity. In relation to one-dimensional problems, such a model can be formulated as follows.

When simulating a test problem by the reference model, we will divide the computational domain into two subdomains. In the first one, containing the shock-free part of the flow, we will use the Euler equations, and in the second one, containing the smeared shock layer, we will use the Navier-Stokes equations. When solving the Navier-Stokes equations, for simplicity, we will not take into account the dependence of the molecular viscosity coefficient, μ, on temperature, and we will assume the Prandtl number to be equal to 3/4. We will select the coefficient μ for each test problem in such a way that the width of the shock smearing due to physical viscosity corresponds to the width of its smearing due to numerical viscosity when computed by a shock-capturing method with the grid spacing Δx.

Simulations by the reference model will be done using the HR-MC+ scheme and fine grid, so that the shock layer is well resolved, and the solution can be considered as grid independent. As practice has shown, it is enough to refine the base grid by a factor of 5 to 10, that is, to use the grid spacing $\Delta x^{refined} \sim 0.1\Delta x$. Note that in this case, due to constraint (2.10), we have $\Delta t^{diff} \sim \Delta x^2$, which means that an explicit approximation of the right-hand side of the Navier-Stokes equations, as described in Section 2.6, does not reduce the overall second order of accuracy of the scheme.

When using the reference model, the computational cells in which the Euler equations are replaced by the Navier-Stokes equations (i.e., where the right-hand side (2.9) is added to the computation) will be ascertained from the condition $|x-x_s|<\Delta_s/2$, where x_s is the coordinate of the shock layer center (where the velocity divergence is minimal), and Δ_s is an upper estimate of the width of smearing the shock front due to physical viscosity (selected for a specific problem in association with the coefficient μ).

3.2 Test 1. Uniform shock wave

Now we turn to the consideration of the first test problem – the propagation of a plane shock through a uniform medium in the one-dimensional formulation. Initially (at time $t=0$), the front of the Mach 3 shock is located at $x=0$. Ahead of the shock ($x>0$), the flow values are as follows: $\rho_1=p_1=1$, behind the shock ($x<0$): $\rho_2=27/7$, $p_2=31/3$; the shock front speed is $u_s=3\sqrt{1.4}$. The problem is computed in two formulations (using different coordinate systems) till time $t=0.36$.

Advancing shock. In the first formulation, the coordinate system (and the associated computational grid) is chosen so that $u_1=0$ (motionless gas ahead of the shock). In this case, $u_2=20\sqrt{1.4}/9$, and the speed of shock propagation through the grid is $u_0=u_s$. The computational domain $[-0.2,1.4]$ is covered by a uniform grid with spacing Δx. The left boundary condition is the inflow with (u_2,ρ_2,p_2); the right boundary is a solid wall.

Slowly moving shock. In the second formulation, the shock moves through the grid slowly, with the speed u_0=0.1. In this case, $u_1=-3\sqrt{1.4}+0.1$ and $u_2=-7\sqrt{1.4}/9+0.1$. The computational domain $[-1.5,0.1]$ is covered by a uniform grid with spacing Δx. The left boundary condition is the outflow at a given pressure p_2; the right boundary condition is the inflow with (u_1, ρ_1, p_1). When saving the computational results obtained with this formulation, we adjust, for convenience, the velocity and the grid coordinates in accordance with the first formulation, i.e. we transform: $u_x-u_1\to u_x$, $x-u_1t\to x$.

3.3 Factor 1. Start-up errors

Fig. 2 shows the density distributions in Test 1 (advancing shock) obtained by the HR-MC+ and RK3-WENO5 schemes with the AV approach and the grid spacing Δx = 1/300.

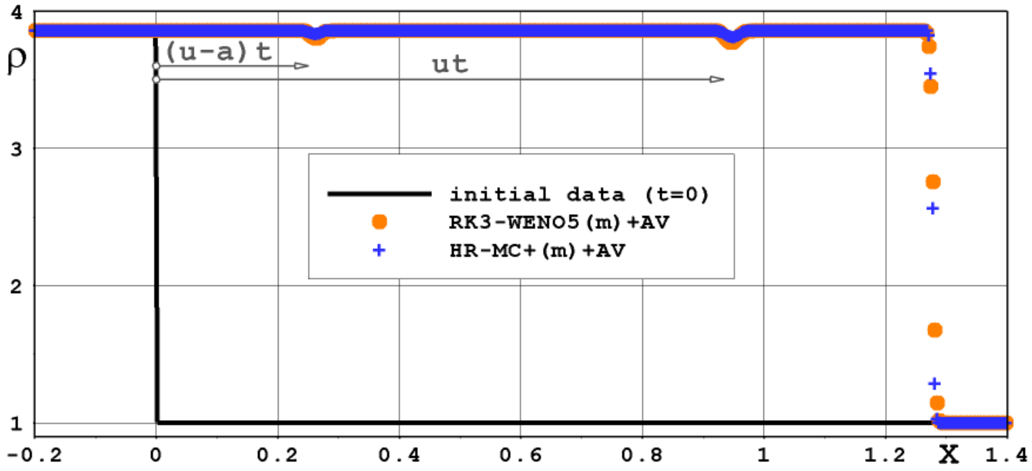

Figure 2: Density profiles in Test 1 (advancing shock) obtained by two schemes with $\Delta x = 1/300$.

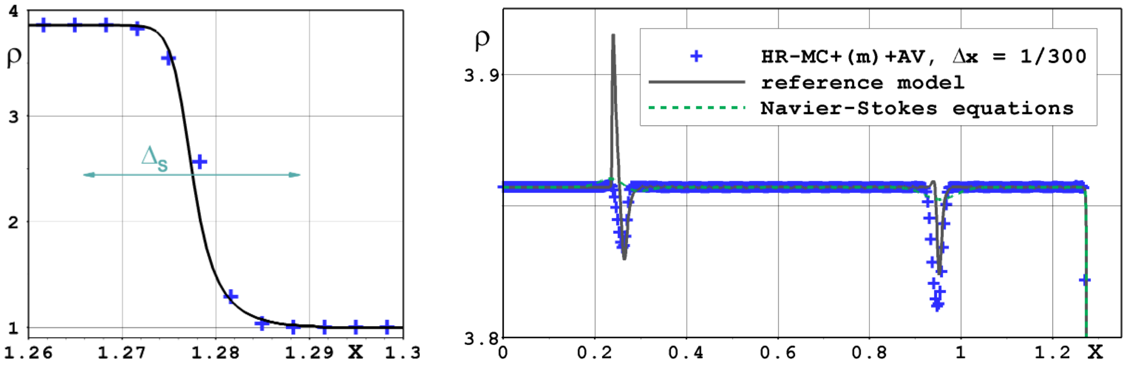

Figure 3: Density profiles in Test 1 (advancing shock) obtained within different approximations.

Here, first of all, two dips in the density profile attract attention. These are start-up errors that arise due to inappropriate (inconsistent with numerical viscosity) smearing of the shock in the initial data (see [25, Section 15.8.4]). The fact is that in the course of forming the shock layer at the initial stage of computation, perturbations of the solution are generated that correspond to the characteristics traveling at speeds $(u-a)$, u and $(u+a)$ (hereinafter referred to as the $(u-a)$-, u- and $(u+a)$-characteristics). We see the first two perturbations in the density profile, and the last perturbation is absorbed by the shock wave. Such errors will be considered the first factor affecting the solution accuracy behind the shock. They are local in nature and are inherent in all shock-capturing methods.

Factor 1 and the reference model. It may seem that the start-up errors are a numerical artifact in its pure form, but this is not the case. Fig. 3 shows a comparison of two computations: (1) the solution of the Euler equations obtained by the HR-MC+(m)+AV scheme with $\Delta x = 1/300$ (symbols) and (2) the reference model solution with $\mu = 5\times 10^{-3}$ and $\Delta_s = 0.024$ (solid line). In the second solution the mesh was refined ($\Delta x^{refined} = 1/1500$), and the physical viscosity coefficient was selected so that the shock layer width corresponded to the first solution (see the left plot of Fig. 3). In the right plot one can see disturbances arising during the initial formation of the shock layer, both due to the numerical viscosity and due to the physical viscosity. Since the numerical and physical viscosities have fundamental differences, the forms of perturbations are different. Note that if, in the second computation, the Navier-Stokes equations are used in the entire computational domain (the reference model with $\Delta_s = \infty$), the perturbations are strongly smoothed (the dash line in Fig. 3). From this we can conclude that if any shock-capturing method (scheme) shows smaller errors due to the first factor relative to other schemes, then it is most likely more dissipative in the smooth part of the solution, which cannot be considered as an advantage of the scheme.

3.4 Factor 2. Errors in reproducing the flat post-shock profile

Now let us consider the second factor affecting the accuracy of the solution behind the shock – the errors in reproducing the flat post-shock profile. Fig. 4 shows the density profiles on an enlarged scale in the range $1 < x < 1.3$, obtained by the HR-MC+ and RK3-WENO5 schemes with $\Delta x = 1/300$. The data are presented separately for the two formulations of Test 1 (advancing or slowly moving shock) and two variants of using schemes (with or without the AV approach). The shaded area shows an error range of 1%. From the analysis of these data, one can conclude the following.

1. When computing the problem in the first formulation without using the AV approach (left top plot), the errors in reproducing the post-shock profile by the RK3-WENO5 scheme are very small (hundredths of a percent). This suggests that in this case, the Rankine-Hugoniot relations are approximated (integrally) with very high accuracy. The HR-MC+ scheme is less accurate, especially in the cells located just behind the shock front.

2. When computing the problem in the second formulation without using the AV approach (left bottom plot), the errors deteriorate markedly (up to $0.8-1.5\%$) and the post-shock oscillations fade very weakly as they move away from the shock front. Here, the RK3-WENO5 scheme already shows the worst accuracy.

3. The use of the AV approach in both problem formulations significantly improves the accuracy of reproducing the post-shock profile (right top and bottom plots).

Factor 2 and the reference model. Although the reference model reproduces the solution behind the uniform shock wave with perfect accuracy (excluding, of course, local errors associated with Factor 1), its modification can help us explain the nature of errors associated with Factor 2. To do this, we will make changes to the reference model that will

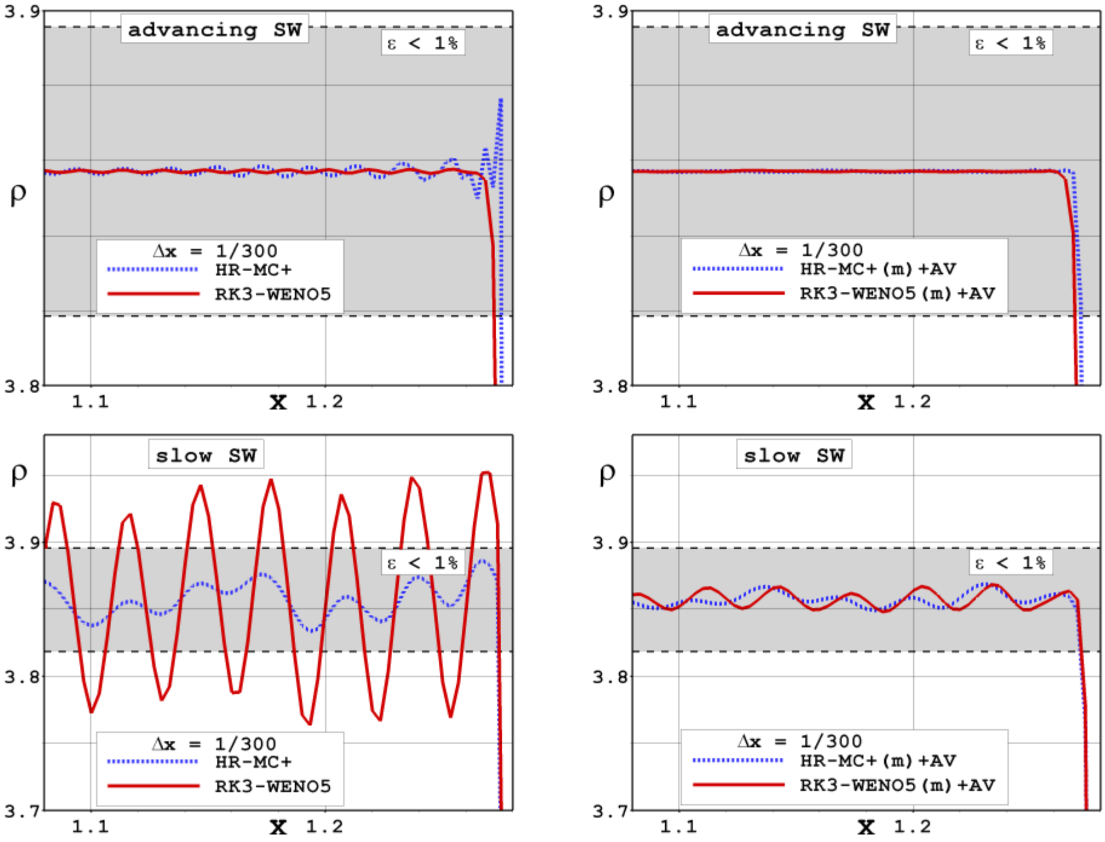

Figure 4: Density profiles in Test 1 obtained by two schemes with $\Delta x = 1/300$. Top: advancing shock; bottom: slowly moving shock; left: without the AV approach; right: with the AV approach.

simulate the unsteady behavior of the numerical viscosity. Before doing this, let us take a close look at the evolution of the smeared shock front over time.

Fig. 5 shows separately the data obtained by the HR-MC+ scheme for the two above formulations of Test 1 (advancing or slowly moving shock) and two variants of using the scheme (with or without the AV approach). The dots show the density values at cell centers at different points in time $t = t^n$, depending on the dimensionless distance from the nominal location of the shock front: $x^* = (x - u_0 t)/\Delta x$. In each plot, the data for the two selected time points are highlighted – see gray circles connected by solid lines and white circles connected by dashed lines.

Note that the travel time of the distance Δx by the shock is $\tau_0 \approx 9.4 \times 10^{-4}$ and $\tau_0 \approx 3.3 \times 10^{-2}$ for the first and second formulations of the problem, respectively. Thus, we can say that the time interval τ_0 determines the cyclicity of the shock passing through the grid, and the time points highlighted in the plots correspond to the opposite moments of such a cycle. From the data presented in Fig. 5, we can conclude the following.

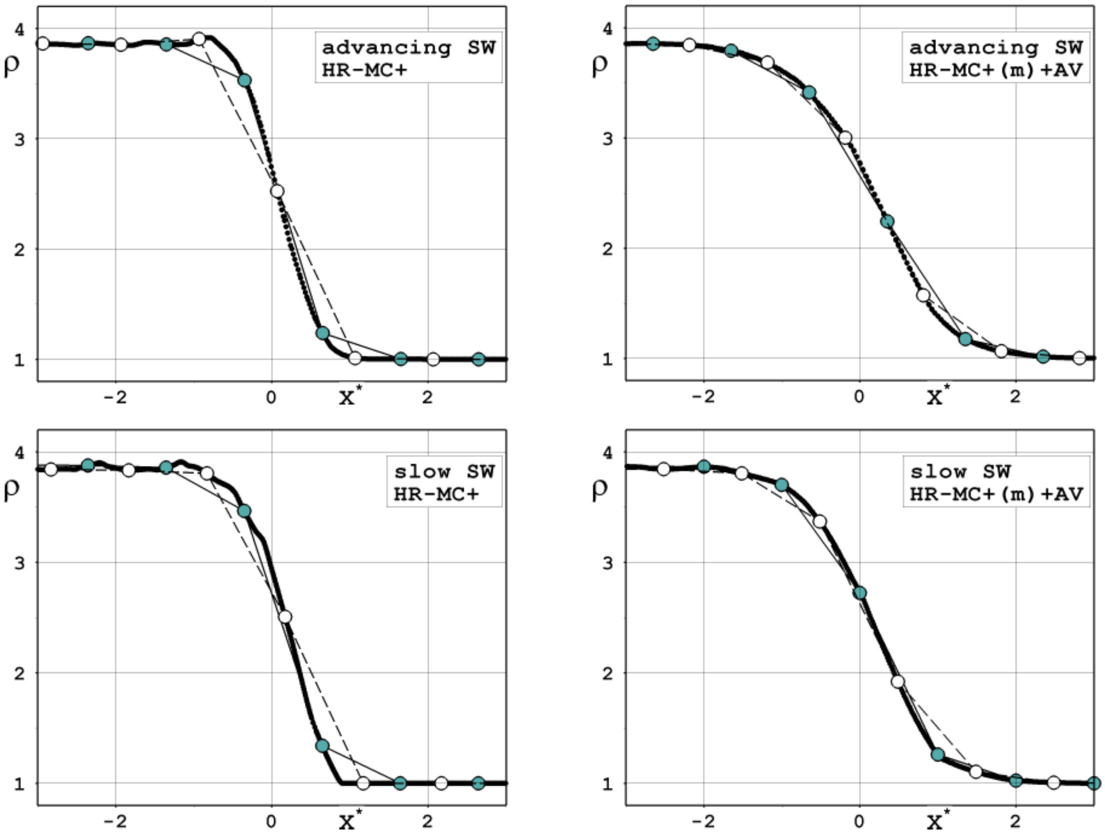

Figure 5: Scatter plots of density in the computations of Test 1 by the HR-MC+ scheme. Top: advancing shock; bottom: slowly moving shock; left: without the AV approach; right: with the AV approach.

1. When calculating the problem without using the AV approach (left plots), from one to two cell centers fall into the shock layer, depending on the position of the shock relative to the grid. Naturally, in this case, the effect of numerical viscosity on the solution at different time points will vary substantially.

2. When calculating the problem with using the AV approach (right plots), the shock layer becomes smoother, and approximately three cell centers fall into it at any position of the shock relative to the grid. In this case, the effect of numerical viscosity on the solution will be more uniform in time.

To simulate the unsteady nature of the numerical viscosity we modify the reference model. Now the coefficient of physical viscosity will change over time as $\mu=\mu_0+\Delta\mu\cdot\cos(2\pi t/\tau_0)$.

The left plot of Fig. 6 shows the density profiles in Test 1 obtained by the modified reference model with fixed parameters $\mu_0=3\times10^{-3}$ and $\Delta_s=0.024$, which correspond

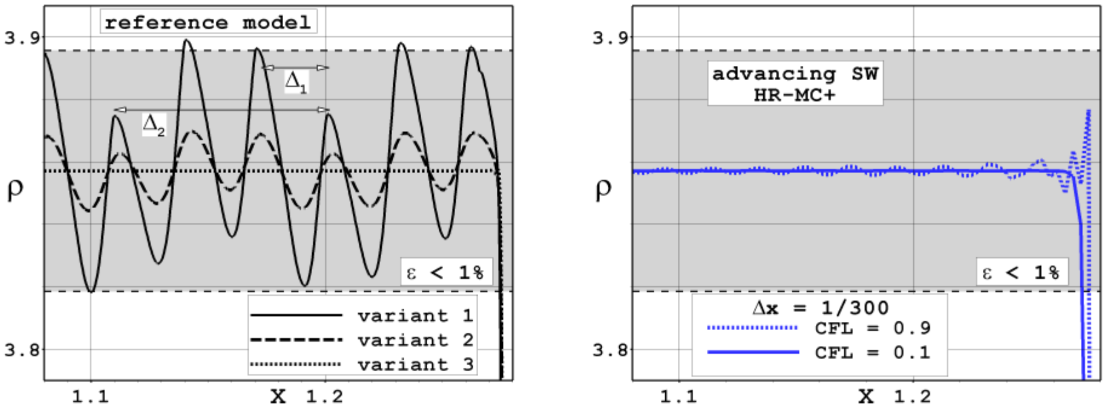

Figure 6: Density profiles in Test 1 obtained by the modified reference model (left) and by the HR-MC+ scheme at different Courant numbers (right).

to the shock smearing by the HR-MC+ scheme with $\Delta x = 1/300$ (in the reference model, $\Delta x^{refined} = 1/3000$ was taken). Three computational variants were considered:

○ *variant 1* with $\Delta\mu = 1.5\times 10^{-3}$ and $\tau_0 = 3.3\times 10^{-2}$ (slowly moving shock);

○ *variant 2* with $\Delta\mu = 0.5\times 10^{-3}$ and $\tau_0 = 3.3\times 10^{-2}$ (slowly moving shock);

○ *variant 3* with $\Delta\mu = 1.5\times 10^{-3}$ and $\tau_0 = 9.4\times 10^{-4}$ (advancing shock).

Variants 1 and 2. As one can see, the first two variants of the modified reference model (solid and dashed lines in the left plot) demonstrate noticeable oscillations behind the shock. The oscillation amplitude is governed by the magnitude of $\Delta\mu_0$ (which was expected), and two types of oscillations with the wavelengths $\Delta_1 = u_2^*\tau_0 \approx 0.030$ and $\Delta_2 = (u_2^* + a_2)\tau_0 \approx 0.094$ are clearly visible in the plot; here a_2 and $u_2^* = u_0 - u_2$ are the sound speed and the relative gas velocity behind the shock, respectively. This kind of oscillations is similar to what we observed in the bottom plots of Fig. 4. These data throw light on the nature of the errors associated with Factor 2 when computing Test 1 in the second formulation. It turns out that the magnitude of such errors depends mainly on how sensitive the numerical viscosity of a particular scheme is to the shock position relative to the grid (nonuniformity in time). It is also clear that the use of the AV approach, although it increases the shock layer width, makes the numerical viscosity more uniform, thereby significantly reducing the magnitude of errors.

Variant 3. It can be seen that this variant of the modified reference model (the dotted line in the left plot) demonstrates the absence of oscillations behind the shock wave. To explain this phenomenon, it is necessary to introduce parameter τ_1, denoting the residence time of the fluid particle in the shock layer. In our case $\tau_1 \approx 3\times 10^{-3}$, which means that for the first formulation of the problem (advancing shock wave) we have $\tau_1 \approx 3\tau_0$, whereas for the second formulation it is $\tau_1 \ll \tau_0$. Now, during the time the fluid particle is in the shock layer, several cycles of fluctuations of the numerical viscosity take place, and their effect on the particle is averaged. Thus, different fluid particles have about the same history of passing through the shock layer (when solving the problem in the second formulation, such averaging does not occur).

The analysis left one question unclear, namely: why does the density profile obtained by the HR-MC+ scheme (the dotted line in the left top plot of Fig. 4) demonstrate noticeable oscillations in the cells located just behind the shock? To answer this question, it is necessary to pay attention to the value of the time step. In the case under consideration, $\Delta t \approx 6.6\times 10^{-4}$, which means that the shock passes the distance Δx in 1.4 time steps (on average), and there is no complete averaging. If a significantly smaller time step is used in the computation (when $C_{cfl} \ll 1$), then the averaging will become complete and the oscillations behind the shock will weaken significantly. This is confirmed by the data shown in the right plot of Fig. 6.

Let us make one more remark about the errors under discussion (Factor 2). As the computation continues and the gas moves away from the shock front (or as the mesh is refined), such errors will slowly decrease due to numerical dissipation; in so doing the schemes of the higher-order accuracy will suppress the errors to a lesser extent. Moreover, sometimes there may be no convergence to the exact solution when refining the mesh at all. Let us consider such a case on the example of solving a test problem in which the wave is reflected from the wall.

3.5 Test 2. Reflected shock wave

The formulation of this test problem differs from the first formulation of Test 1 (advancing shock) only in that the computational domain $[-0.2,1]$ is used here (instead of $[-0.2,1.4]$). In this case, during the computation of the problem, the shock manages to reach the right boundary of the computational domain (where a solid wall is specified) and reflect from it. By the final time ($t=0.36$), the reflected shock arrives in the section $x=1-|u_0^*|(t-1/u_0)\approx 0.8868$; here $u_0=3\sqrt{1.4}$ and $u_0^*=-11\sqrt{1.4}/9$ are the velocities of the shock wave before and after reflection from the wall, respectively. Behind the reflected shock, the flow values are as follows: $u_3=0$, $\rho_3=31\rho_2/11$, $p_3=5p_2$.

Before proceeding to the demonstration and analysis of the computational data, it is worth giving a clarification on the realization of the boundary condition at $x=1$. All computations of this test problem were obtained using the symmetry condition, namely: in the area $x>1$, the required number of computing cells was added, in which the gas parameters were maintained at each stage of the calculation based on the condition of mirror symmetry.

Fig. 7 shows the computational data of Test 2 obtained by the HR-MC+(m)+AV scheme on a very detailed grid. The density and pressure profiles behind the reflected shock are shown on an enlarged scale (the displayed ranges of density and pressure are $\sim 0.5\%$).

Arrows 1 and 2 in the figure show the flow disturbances, which we will discuss in the next section. Now let us pay attention to the interval $0.97<x<0.98$ (the neighborhood

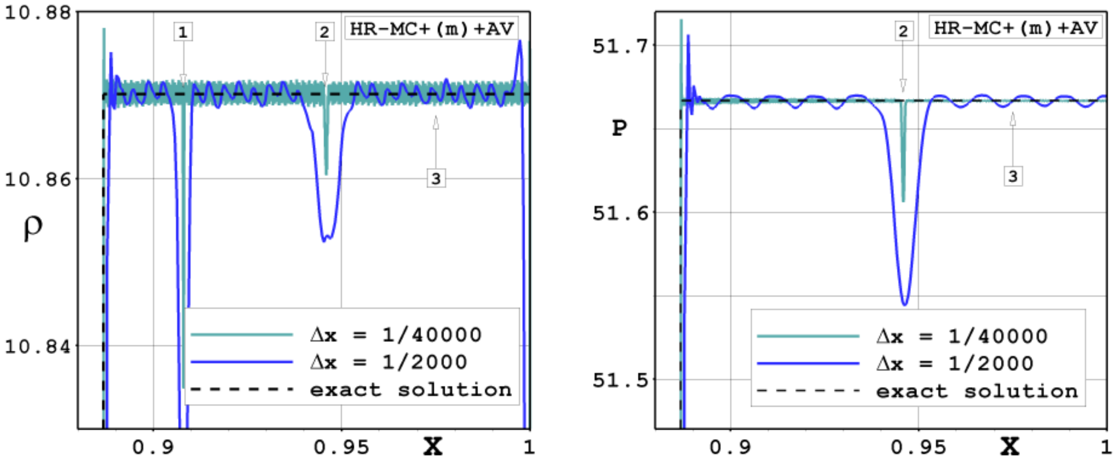

Figure 7: Density (left) and pressure (right) profiles behind the reflected shock in Test 2, obtained using the HR-MC+(m)+AV scheme at high grid resolution.

of arrow 3). Here the oscillations of the parameters are visible, and the amplitudes of the density and pressure fluctuations behave differently. The pressure amplitude decreases with the mesh refinement, and here we can talk about convergence to the exact solution. The density amplitude practically does not change, remaining at the level of 0.015%. (Note that for the RK3-WENO5(m)+AV scheme, this value is 0.01%; in computations without using the AV approach, the density amplitude for both schemes increases about 4 times.) Strictly speaking, there is no convergence to the exact solution here (we are not talking about convergence in integral quantities). This is understandable, because after the shock is reflected from the wall, the gas becomes motionless relative to the grid and the oscillations in the density cease to fade immediately after equalization in pressure and (at zero level) speed.

Thus, ending the discussion of errors related to Factor 2, we formulate the necessary condition for using one or another shock capturing method in practical computations. Given that the level of such errors is weakly dependent on the grid resolution, the method used should ensure such an accuracy of reproducing the flat post-shock profile that would be guaranteed to meet the requirements of the conducted research.

3.6 Factor 3. Entropy trace

Now let us discuss the disturbances shown in Fig. 7 by arrows 1 and 2. The first of them is a disturbance caused by the start-up errors (Factor 1) and corresponding to the u-characteristic. In its vicinity, the velocity and pressure are almost constant (up to the errors caused by Factor 2), but there is a dip in the density profile, which corresponds to an increase in the temperature and *entropy* profiles. When this perturbation passes through the reflected shock, a new perturbation corresponding to the $(u+a)$-characteristic is born;

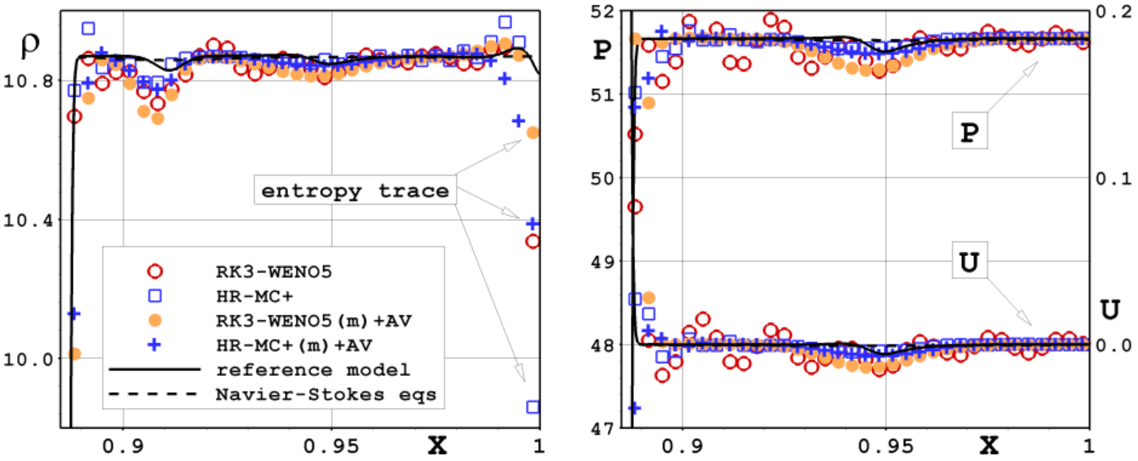

Figure 8: Density (left), pressure and velocity (right) profiles behind the reflected shock in Test 2, obtained by: the HR-MC+ and RK3-WENO5 schemes (with and without the AV method) with $\Delta x = 1/300$ (symbols); the reference model (solid lines); the Navier-Stokes equations (dash lines).

it is shown in Fig. 7 by arrow 2. Here we can find an analogy with the disturbances caused by the initial formation of the shock layer (Factor 1).

The term *entropy trace* [34] or *wall heating error* [35] is used to denote local disturbances (errors) corresponding to the u-characteristic and arising after the interaction of the shock with the wall or symmetry plane (this implies solving the Euler equations by one of the shock-capturing methods).

Now let us turn to Fig. 8, which shows the computational data of Test 2 obtained by the HR-MC+ and RK3-WENO5 schemes (with and without the AV approach) on a grid with $\Delta x = 1/300$. As can be seen, there are significant dips in density near the boundary $x = 1$ (left plot), while the pressure and velocity (right plot) are reproduced with high accuracy. It is precisely such errors that are called the *entropy trace*. In our case, the error in density reaches $5-9\%$ for basic variants of the schemes, and $2-4\%$ for the schemes with adding the AV approach.

Factor 3 and the reference model. Fig. 8 also presents data obtained from the reference model and within the framework of the Navier-Stokes equations with the same parameters that were used in the analysis of Factor 1. These data indicate the following.

1. When the shock is reflected from the wall, the parameter profiles in the shock layer undergo significant restructuring, which occurs due to the action of the numerical (Euler equation) or physical (reference model) viscosity. In each case, this restructuring occurs in different ways, but when using the reference model (solid line), the intensity of the entropy trace turns out to be significantly lower (the error in density is $\sim 0.5\%$) than in the case of solving the Euler equations.

2. In the case of using the Navier-Stokes equations, at time $t=0.29$ (immediately after the reflection of the shock from the wall), an entropy trace of low intensity is also observed (these data are not given here). However, it dissipates afterwards due to the action of thermal conductivity and by the final time ($t=0.36$) the density profile turns out to be almost flat (dashed line in the left plot of Fig. 8). In the reference model, dissipation works only inside the shock layer and, therefore, the entropy trace remains.

So, the experience of using the reference model and the artificial viscosity approach in Test 2 suggests the following. Although the intensity of the entropy trace strongly depends on the specific numerical technique and, in principle, can be significantly reduced, it does not seem possible to completely resolve this problem within the concept of shock-capturing.

3.7 Test 3. Shu-Osher test problem

The next test problem is computing the propagation of a shock through a nonuniform medium, as formulated by Shu and Osher in [36]. Here, just as in the previous test problem, the front of the Mach 3 shock is initially located at $x=0$. However, the shock wave then propagates through a gas with a sinusoidal density distribution. Namely, ahead of the shock ($x>0$), the flow values are as follows: $\rho_1=1+0.2\sin(5\pi x)$, $p_1=1$; behind the shock ($x<0$): $\rho_2=27/7$, $p_2=31/3$. The problem is also computed in two coordinate systems till time $t=0.36$.

Advancing shock. In the first formulation, the coordinate system is chosen so that $u_1=0$ (motionless gas ahead of the shock). In this case, $u_2=20\sqrt{1.4}/9$, and the initial speed of shock propagation through the grid is $u_0=3\sqrt{1.4}$. The computational domain [-0.2, 1.4] is covered by a uniform grid with spacing Δx. The left boundary condition is the inflow with (u_2,ρ_2,p_2); the right boundary is a solid wall.

Slowly moving shock. In the second formulation, the shock is initially motionless ($u_0=0$), but then it starts slow periodic motion relative to the grid (a consequence of the sinusoidal density distribution in the flow ahead of the wave). In this case, $u_1=-3\sqrt{1.4}$ and $u_2=-7\sqrt{1.4}/9$. The computational domain $[-1.2,1.6]$ is covered by a uniform grid with spacing Δx. The left boundary condition is the outflow at a given pressure p_2; the right boundary condition is the inflow with (u_1,ρ_1,p_1), and here we can assume $\rho_1=1$ (without taking into account the sinusoidal distribution), since the gas flowing through this boundary has no time to reach the shock front during the computation. As in Test 1, when saving the computational results obtained with this formulation, we adjust the velocity and the grid coordinates in accordance with the first formulation, i.e. we transform: $u_x-u_1\to u_x$, $x-u_1t\to x$.

Reference solution. Fig. 9 shows the reference solution of the problem at time $t=0.36$ obtained using the HR-MC+(m)+AV scheme with the grid spacing $\Delta x=1/30\,000$. Two

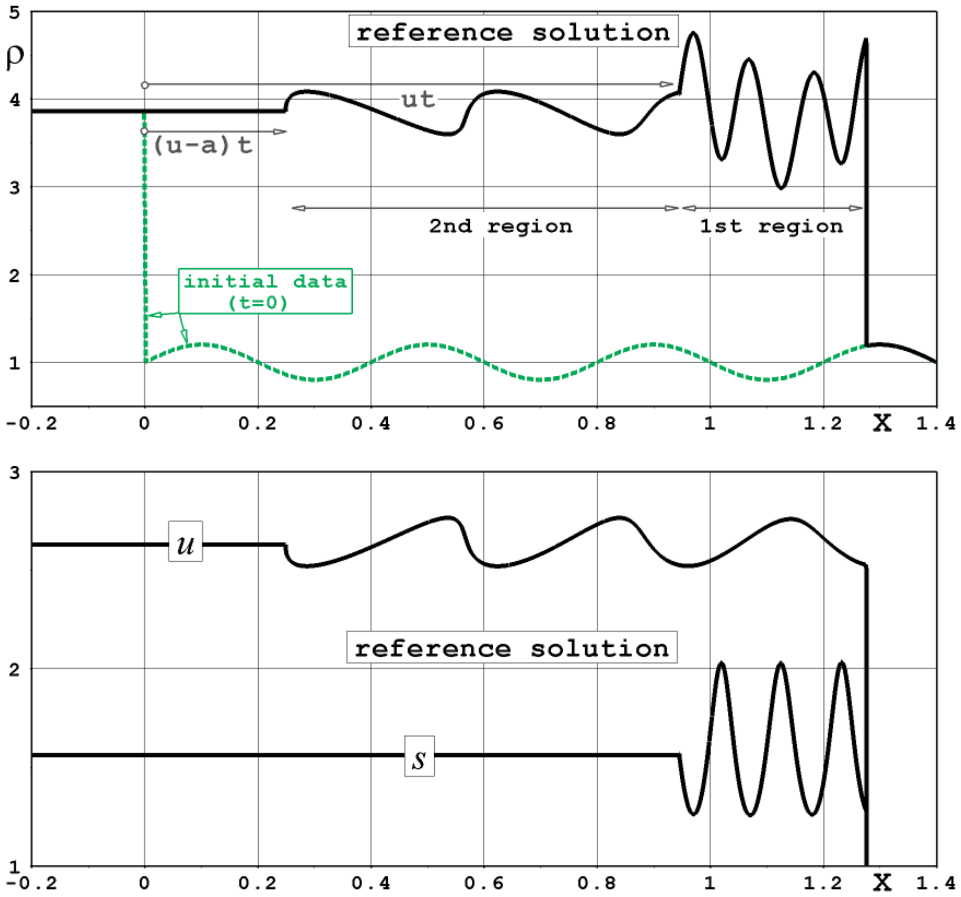

Figure 9: Density (top),velocity and entropy (bottom) profiles in the Shu-Osher test problem (reference solution).

regions are clearly visible in the density profile (top plot). The first region ($0.945 < x < 1.276$) contains gas that passed through the shock front during the computation. This clearly confirms the entropy profile ($s = p/\rho^\gamma$) in the bottom plot: the initial nonuniform density distribution before the shock leads to a nonuniform entropy profile. The region with uniform entropy to the left of $x = 0.945$ refers to the gas initially located behind the shock. If we look at the velocity profile (bottom plot), we will see acoustic waves arising when a shock wave moves through a nonuniform medium. These waves also penetrate into the region where the entropy is constant; this is how the second region ($0.25 < x < 0.945$) is formed in the density profile. In what follows, the first region will be of interest – we will analyze the solution accuracy in the part of the gas that passed through the shock during the computation.

Fig. 10 shows the density profile in the first region; it indicates the interval $L = [1, 1.2]$, in which we will analyze the error of the numerical solution of the problem, changing the

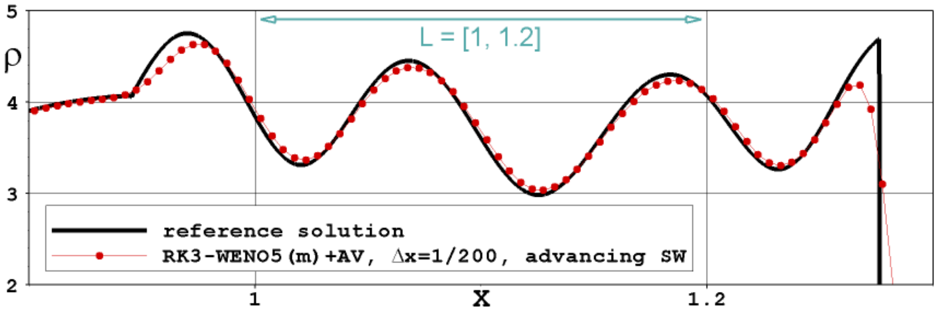

Figure 10: Fragment of the density profile in the Shu-Osher test problem.

problem formulation, the numerical technique and the grid spacing. The numerical error will be calculated using the formula

$$\epsilon=\sqrt{\frac{\sum_{i=i_1}^{i_2}\left(\rho_i-\rho_i^{ref}\right)^2}{i_2-i_1+1}}\times\frac{100\%}{\rho_2},$$

where ρ_i is the gas density in the i-th cell in a numerical solution with the grid spacing Δx, ρ_i^{ref} is the projection of the reference solution onto the grid (density averaged over the cell), i_1 and i_2 are the initial and final numbers of cells whose centers fall in the interval L.

Fig. 11 shows the numerical error in the interval $L=[1,1.2]$ for the selected schemes depending on the grid spacing Δx. The data are displayed separately for the two above formulations of Test 3 (advancing or slowly moving shock) and two variants of using the scheme (with or without the AV approach). Before analyzing these data, we note that the considered interval L does not include local disturbances (errors) associated with Factor 1, and Factor 3 does not take place in this problem. However, disturbances associated with Factor 2 are present here, and their impact on the numerical error may be significant.

Left bottom plot of Fig. 11. First, we consider the case when Test 3 is computed in the second formulation (slowly moving shock) by schemes without adding artificial viscosity; here Factor 2 manifests itself to the maximum extent. Based on the data presented in the left bottom plot, the following conclusions can be drawn.

1. In the range under consideration, the error level ϵ is $0.2-2\%$, which means that the errors associated with Factor 2 (see the left bottom plot of Fig. 4) play a crucial role here.

2. The grid convergence is very uneven, and the order of convergence on average is noticeably less than one.

3. The higher-order schemes are inferior in accuracy to the second-order schemes, which explains the left bottom plot of Fig. 4: the RK3-WENO5 scheme generates oscillations of noticeably greater intensity than the HR-MC+ scheme.

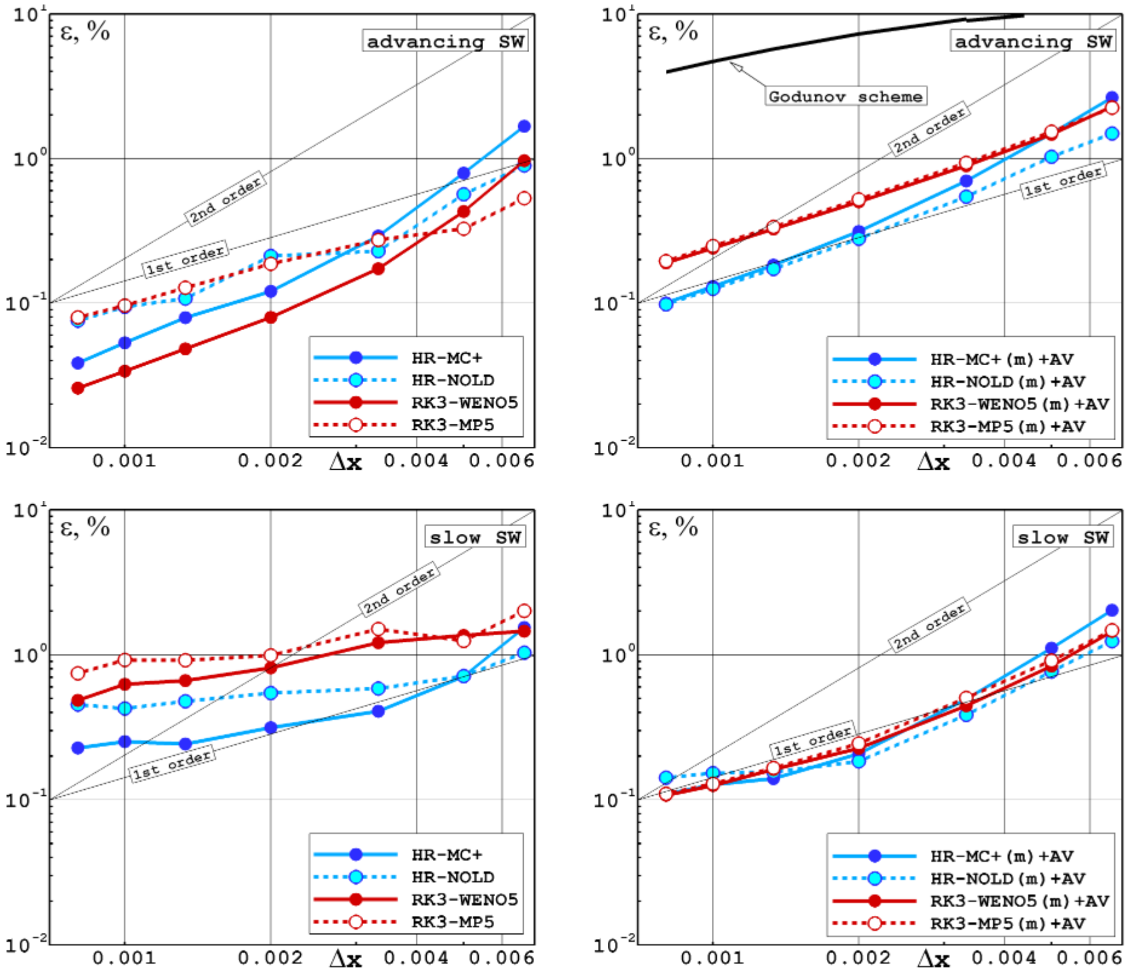

Figure 11: Shu-Osher test problem. Solution errors versus Δx for the interval $L=[1,1.2]$. Top: advancing shock; bottom: slowly moving shock; left: without the AV approach; right: with the AV approach.

Right bottom plot of Fig. 11. Now we consider the data for computing the problem in the same formulation, but with the use of the AV approach. Here we note the following.

1. The error level has decreased and the grid convergence has become more uniform.
2. All schemes show close accuracy. As this takes place, the HR-MC+(m)+AV scheme shows slightly less accurate results on the coarsest grid, while on the detailed grid the HR-NOLD(m)+AV scheme is slightly inferior to the others.
3. In the range $\Delta x > 0.003$ the convergence rate of all schemes is around two; this is most clearly demonstrated by the scheme HR-MC+(m)+AV. With the mesh refinement the convergence rate decreases so that on the most detailed grid (where

$\epsilon \sim 0.1\%$) it becomes less than one. This is due to the influence of the errors associated with Factor 2; their level corresponds exactly to the errors that can be observed in the right bottom plot of Fig. 4.

Left top plot of Fig. 11. Now let us turn our attention to the results of computing Test 3 in the first formulation (advancing shock). According to the data obtained by the schemes without the AV approach, the following conclusions can be drawn.

1. In comparison with the solution of the problem in the second formulation (left bottom plot), the level of error ϵ has decreased markedly, which is explained by the lower level of the oscillations caused by Factor 2.

2. The HR-MC+ and RK3-WENO5 schemes demonstrate better accuracy and their grid convergence curves look smoother. At the same time, the HR-MC+ scheme is inferior in accuracy to the RK3-WENO5 scheme, which is quite understandable: its approximation order is lower, and the level of oscillations (see the left top plot in Fig. 4) is higher.

3. On a coarse grid, the HR-NOLD scheme is superior in accuracy to the HR-MC+ scheme, and the RK3-MP5 scheme is superior to the RK3-WENO5 scheme. This can be explained by their reduced dissipation, which is confirmed by the data shown earlier in Fig. 1 (see the results of computations on the coarsest grid). However, with the mesh refinement, this property of these schemes begins to play a negative role – the oscillations associated with Factor 2 are suppressed to a lesser extent (these data are not given in Fig. 4), which is why the grid convergence noticeably worsens.

Right top plot of Fig. 11. Finally, we consider the data for computing the problem in the first formulation, but with the use of the AV approach. As demonstrated above (see the right top plot of Fig. 4), the errors associated with Factor 2 are extremely small ($<0.01\%$) in this computational case, which means that their influence on the error ϵ in the range under consideration can be neglected. Therefore, this computational case is the most suitable for identifying and studying other factors (besides those already considered) that affect the solution accuracy downstream of the shock. Before proceeding to this part of our study, we will formulate three questions that arise first of all when considering the data in the right top plot.

1. Why does the use of the AV approach in this problem formulation lead to the deterioration in solution accuracy?

2. Why are the second-order schemes (HR-MC+(m)+AV and HR-NOLD(m)+AV) more accurate than the higher-order schemes (RK3-WENO5(m)+AV and RK3-MP5(m) +AV) in this computational case?

3. Since, with sufficient mesh refinement ($\Delta x < 0.002$) all the schemes demonstrate approximately first-order convergence, isn't it justified to use first-order schemes, say, the Godunov scheme?

We will give answers to the first two questions later, and to answer the last question, we should once again refer to the right top plot of Fig. 11, which also shows the data obtained by the Godunov scheme (black solid line). It can be seen that the errors by this scheme are one and a half orders of magnitude higher, which means that its use in problems of this type is ineffective.

Important notice. Starting from this point, we will continue our study, limiting ourselves to the case of solving the Shu-Osher test problem in the first formulation (advancing shock) using the AV approach. In so doing, for short, in the text we will omit the abbreviation (m)+AV, which is common to all the schemes.

Continuing the analysis of the data in the right top plot of Fig. 11, one can say that the HR-MC+ scheme is inferior in accuracy to the other methods on a coarse grid, but at the same time its convergence rate is close to the second order. Therefore, its accuracy is growing rapidly with the mesh refining. For all schemes, it can be seen that, on a coarse grid, the convergence rate is higher than the first order. Thus, the dependence of the numerical error behind shocks on the grid spacing can be approximately described by the formula

$$\epsilon \approx c_1 \Delta x + c_2 \Delta x^2. \tag{3.1}$$

With a fine mesh spacing (for small values of Δx), the first term on the right-hand side of (3.1) dominates, and we see the first order of convergence. On a coarse grid (for large Δx), the second term may prevail, increasing the rate of convergence to the second order.

Now let us take a closer look at Fig. 10, where the numerical solution obtained by the RK3-WENO5 scheme on the grid with $\Delta x = 1/200$ is compared with the reference solution. It can be noticed that its difference from the reference profile consists in (1) shifting and (2) smoothing of the density profile. These are two other factors affecting the solution accuracy downstream the shock. The analysis showed that they are responsible, respectively, for the first and second terms in (3.1). Let us focus on this in more detail.

3.8 Factor 4. Errors caused by the profile shift

So, the fourth factor affecting the accuracy of the solution behind the shock is related to the profile shift. Previously, when Factor 1 (the start-up errors) was discussed, it was noted that the perturbation corresponding to the $(u+a)$-characteristic is absorbed by the shock itself. Such absorption, however, leaves its imprint, a slight shift of the shock.

To calculate the magnitude of this shift, let us go back to Test 1 (propagation of a shock through a uniform medium) and take advantage of the fact that the flat profile behind the shock is reproduced very accurately. We integrate the density over the interval $L=[x_a,x_b]$, which includes the shock layer (the smeared shock front), and equate the result to an exact solution, in which the density jumps at the point $x=x_s$. As a result, we get

$$\int_{x_a}^{x_b} \rho dx = \Delta x \sum_{i=i_1}^{i_2} \rho_i = \rho_2 (x_s - x_a) + \rho_1 (x_b - x_s).$$

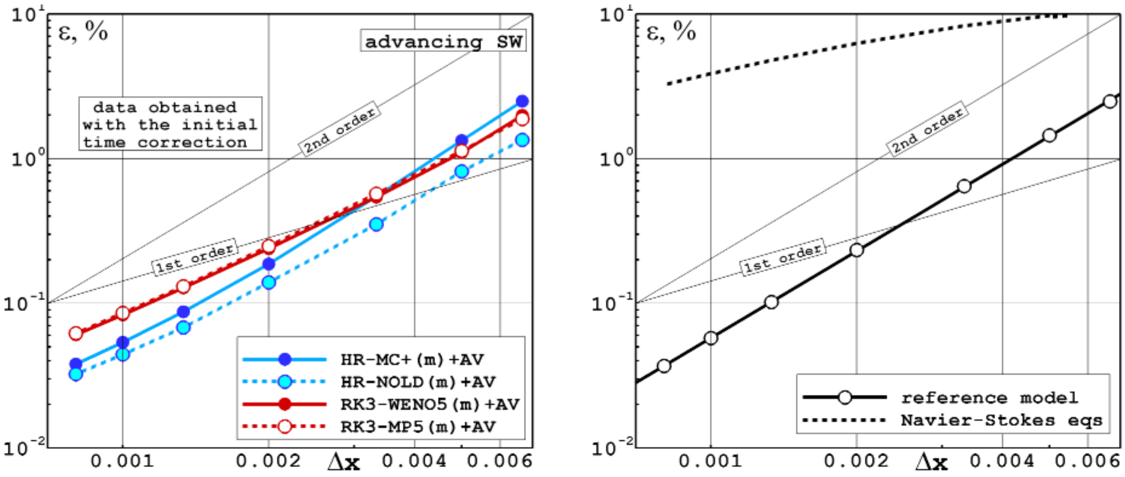

Figure 12: Shu-Osher test problem. Solution errors versus Δx for the interval $L=[1,1.2]$. Left: computations with adjusting the initial time; right: data obtained from the reference model and the Navier-Stokes equations.

In this expression, all values are known, except for the x_s coordinate, by determining which one can calculate the shift of the shock position relative to the exact value $x_s^{\text{exact}}=u_s t$. Computations have shown that: (1) the value of displacement $\delta=x_s-x_s^{\text{exact}}$ is proportional to Δx; (2) for the second-order schemes under discussion, it is approximately equal to $0.17\Delta x$; (3) for the higher-order schemes, it is about twice as large: $\delta\approx 0.36\Delta x$. It is worth recalling here that the AV approach was used in all computations, which assumes switching to the minmod limiter within the shock layer. Therefore, the computational procedure in such cells differs exclusively in the way of integration over time: using the HR method or the RK3 method. This explains the discovered difference in the values of δ in the schemes of the second and higher orders.

If you set out to get rid of the errors caused by the profile shift, you can suggest a simple way – to adjust the initial time: instead of $t=0$, assume $t=\delta/u_s$. Let us see how this affects the dependence of the numerical error on Δx in the Shu-Osher test problem.

The left plot of Fig. 12 shows the updated data that were obtained taking into account the proposed correction of the initial time. As one can see, with this correction, the accuracy of all computations on the coarsest grid ($\Delta x\approx 0.006$) changes slightly. However, the second order of convergence is now clearly visible here, and with the grid refining, the accuracy of the schemes increases significantly. On the finest grid ($\Delta x\approx 0.0008$), it is hundredths of a percent versus tenths of a present in the previous (basic) case. Nevertheless, even in the adjusted computations, the order of convergence approaches the first one with mesh refining. Thus, the estimated formula (3.1) for the numerical error has not changed, but the coefficient c_1 in it has become noticeably smaller.

Let us try to explain the fact that the first term on the right-hand side of (3.1), the first-order term, has not disappeared at all. To do this, recall that, when calculating the value of the shock wave shift, we used the law of conservation of mass. However, other laws of conservation, of momentum or total energy, can be used for the same purpose. It turns out that the value of displacement depends on the conservation law used. For the second-order schemes, we obtain: $\delta \approx 0.17\Delta x$, $0.10\Delta x$ and $0.06\Delta x$ (respectively, for the laws of conservation of mass, momentum and energy). For the higher-order schemes: $\delta \approx 0.36\Delta x$, $0.20\Delta x$ and $0.14\Delta x$. Thus, it becomes clear that it will not be possible to completely get rid of the errors associated with Factor 4 (it is impossible to simultaneously level the shifts in three conservative variables). The obtained data also provide an answer to the second of the questions posed earlier as to why higher-order approximation schemes with grid refinement show worse accuracy compared to second-order schemes. They simply have a larger value of displacement δ.

3.9 Factor 5. Errors caused by the profile smoothing

Now we will discuss the fifth factor affecting the solution accuracy downstream of shocks – the smoothing of the profile due to numerical viscosity when it passes through the shock layer.

Let us draw an analogy between the shock front smearing due to the physical viscosity and due to the numerical viscosity. First, we estimate the difference between two exact solutions: the first is obtained within the framework of the Euler equations and the Rankine-Hugoniot relations on the shock front, and the second is obtained within the framework of the Euler equations in the smooth part of the solution and the Navier-Stokes equations inside the smeared shock front (the reference model). The Euler equations differ from the Navier-Stokes equations only in their right-hand side, the intensity of which, $|\mathrm{RHS}|$, is proportional to the physical viscosity coefficient μ. At the same time, the shock layer thickness, Δ_s, is also proportional to μ, and the residence time of the gas particle in this layer is $\tau_1 \sim \Delta_s$. From this it can be concluded that the influence of the right-hand side on the gas particle as it passes through the shock layer will be proportional to $|\mathrm{RHS}| \cdot \tau_1 \sim \Delta_s^2$. That is, the two compared solutions behind the shock will differ by the magnitude of $\sim \Delta_s^2$.

Let us now consider the solution obtained numerically by the shock-capturing scheme. In regions where the solution is smooth, the scheme dissipation may be minor (for high-order schemes), but within the shock layer it will in any case be significant, and, by analogy with the effect of the physical viscosity, the effect of scheme dissipation in the shock layer will be proportional to Δ_s^2. Since $\Delta_s \sim \Delta x$, the influence of Factor 5 (profile smoothing) on the accuracy of the solution will be $\sim \Delta x^2$.

Thus, the accuracy of solving the Shu-Osher problem in the first formulation using the AV approach is governed by two factors: shifting and smoothing of the profile when passing through the shock layer. The estimate (3.1) is valid for the numerical error here, and the higher-order schemes have no advantages over the second-order schemes.

Factors 4 and 5 and the reference model. Now let us consider the influence Factors 4 and 5 on the accuracy of solving the Shu-Osher problem in the case of using the reference model. In this part of our study, we will use the following dependencies of the reference model parameters on the grid resolution: $\mu = 5\times 10^{-3}\cdot(\Delta x/\Delta x^*)$, $\Delta_s = 0.024\cdot(\Delta x/\Delta x^*)$, $\Delta x^{refined} = \Delta x/10$, where $\Delta x^* = 1/300$.

First, we turn to Test 1 and find the values of displacement δ using three conservation laws. Computations revealed that these values are $\delta \approx 0.03\Delta x$, $0.03\Delta x$ and $0.005\Delta x$, which is significantly less than the values obtained earlier, both for the second-order schemes and for the higher-order schemes. Hence, we can assume that in the considered range of Δx, Factor 4 will not have a noticeable effect on the value of ϵ.

Now we return to Test 3 and look at the right plot of Fig. 12, which shows the solution error for the reference model depending on the grid spacing Δx (solid line). As we can see, the convergence rate is uniformly close to the second order, which means that: (1) the influence of Factor 5 on the solution accuracy corresponds to the estimate made above; (2) the influence of Factor 4 is negligible, just as we assumed. Comparing these data with the data shown in the left plot of Fig. 12, note their close agreement.

The right plot of Fig. 12 also shows the data for computing the problem in the framework of the Navier-Stokes equations (dashed line) with the same coefficient μ that was used in the reference model. It can be seen that in this case the computational accuracy is substantially lower and it corresponds to the accuracy of the first-order scheme (compare with the data of the Godunov scheme in the right top plot of Fig. 11). This result is understandable, because in this case, the right-hand side (2.9) acts in the entire computational domain, and the following estimate is valid for it: $|\mathrm{RHS}| \sim \mu \sim \Delta x$.

Concluding the discussion of Factors 4 and 5, we will answer the first of the questions posed earlier: why, when solving Test 3 in the first formulation, adding the artificial viscosity approach leads to the deterioration in accuracy? The fact is that in the case of using the AV approach, the width of the shock layer grows, which means that the negative effect of Factor 5 on the solution accuracy increases (in proportion to Δ_s^2). In addition, the evaluation of the influence of Factor 4 in the computations without the AV approach showed that in this case the values of displacement δ become smaller (by about 1.5 times).

3.10 Factor 6. Errors associated with the order of accuracy of the scheme

In Section 2.8, it was shown that in the absence of discontinuities in the solution, the accuracy of the higher-order schemes significantly exceeds the accuracy of the second-order schemes. In the Shu-Osher test problem we did not observe such superiority because we studied the solution accuracy behind the shock wave directly. However, the situation may change if we extend the computation time. The solution accuracy at a long distance from the shock in this case will depend on the accuracy of the scheme in the lengthy region, where the solution is presumably smooth. In this case, to evaluate the accuracy we can write

$$\epsilon \approx c_1\Delta x + c_2\Delta x^2 + t\cdot c_k\Delta x^k,$$

where k is the order of the scheme.

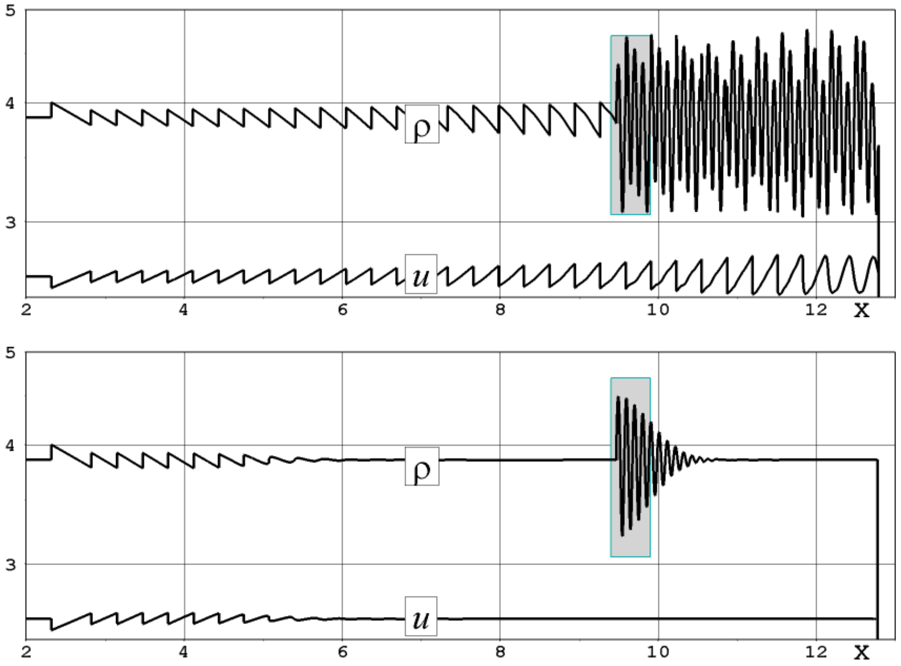

Figure 13: Density and velocity profiles in the Shu-Osher problem at $t=3.6$. Top: the basic formulation of the problem; bottom: the modified formulation of the problem.

According to the assessment made, the advantage of the higher-order schemes can become significant with large t and Δx, that is, with a long computation on a sufficiently coarse grid. Let us check this assumption. To do this, we will increase the computation time in the Shu-Osher test problem by an order of magnitude, while enlarging the computational domain.

The top plot of Fig. 13 shows the density and velocity profiles in the Shu-Osher test at time $t=3.6$. Inspecting the velocity profile we can see how the acoustic waves, being initially smooth, transform into weak shocks as they move to the left. That is, with a long computation time of the problem under consideration, the solution behind the shock loses its smoothness. Moreover, during the computation (up to $t=3.6$), multiple weak shocks run across the part of the gas that passed through the shock at the initial stage of the computation (shaded region in Fig. 13), which could not but affect the accuracy of the solution (it is known that the convergence rate of any scheme in the vicinity of discontinuities is always ≤ 1). To avoid this, we will make a correction to the problem formulation.

3.11 Test 4. Modified Shu-Osher test problem

In the new, modified formulation, the initial density distribution ahead of the shock ($x>0$) is given by: $\rho_1=1+0.2\sin(5\pi x)\exp(-0.2x^2)$. The reference solution corresponding to the

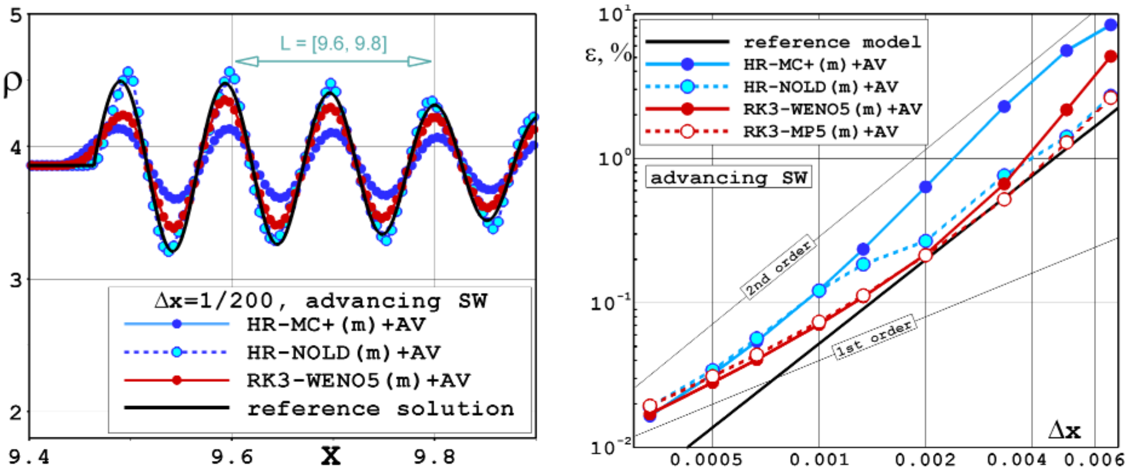

Figure 14: Modified Shu-Osher problem. Left: fragment of the density profile; right: solution errors versus Δx for the interval $L=[9.6,9.8]$ (computations with the adjusted initial time).

modified formulation of the problem is shown in the bottom plot of Fig. 13. The part of the density profile we are interested in is shown on an enlarged scale in the left plot of Fig. 14, where (in addition to the reference solution) the data obtained on a coarse grid (with $\Delta x=1/200$) using the HR-MC+, HR-NOLD and RK3-WENO5 schemes are given. The numerical errors in the interval $L=[9.6,9.8]$ versus Δx for the four selected schemes are demonstrated in the right plot of Fig. 14. From the analysis of these data we can conclude the following.

1. In the range $\Delta x > 0.001$, the HR scheme with the MC+ limiter turns out to be the least accurate, but it consistently demonstrates the rate of convergence close to the second order. On a very fine grid, this scheme demonstrates the convergence rate close to the first order, as well as all the other schemes.

2. The higher-order schemes show better accuracy, and the RK3-WENO5 scheme on a coarse grid clearly demonstrates the convergence rate higher than the second order. However, for these schemes the convergence rate drops rapidly with the mesh refinement.

3. The HR-NOLD scheme on a coarse grid is not inferior in accuracy to the higher-order schemes, This point is explained by the fact that the NOLD limiter, like the MP5 reconstruction, has low dissipation not only near discontinuities, but also on smooth solutions in the case of coarse grids (see Fig. 1 and the left plot of Fig. 14).

Factor 6 and the reference model. The right plot of Fig. 14 also shows the computational data of Test 4 obtained from the reference model (black solid line) with the same parameters that were used to solve Test 3. It can be seen that in Test 4, the reference model also shows the second order of grid convergence. However, here it clearly manifests itself as a benchmark, relative to which the effectiveness of the schemes under discussion can be determined. In this regard, the previous analysis can be supplemented with the following considerations.

1. As the grid is refined, the RK3-WENO5 scheme reaches the ”reference” accuracy somewhere in the vicinity of $\Delta x = 0.002$, but then its accuracy begins to decrease (relative to the reference model) due to errors associated with Factor 4 (first-order errors).

2. The less dissipative RK3-MP5 scheme shows good agreement with the reference model in the range $0.002 < \Delta x < 0.0067$, but in the range $\Delta x < 0.02$ it follows the RK3-WENO5 scheme.

3. The second-order HR-NOLD scheme in the range $\Delta x > 0.002$ competes in accuracy with the RK3-MP5 scheme, however, as the grid is further refined, it departs from the higher-order schemes and, starting at $\Delta x = 0.001$, follows the HR-MC+ scheme.

Concluding the discussion of Factor 6, we emphasize the following. When computing flows without gasdynamic discontinuities, the higher-order schemes most clearly demonstrate their advantage over the second-order schemes in the case of using a very detailed grid. However, when computing flows with discontinuities on a detailed grid, the errors associated with Factors 4 and 5 dominate behind the shock, which does not allow the higher-order schemes to demonstrate high accuracy. Therefore, these schemes turn out to be noticeably more accurate than the second-order schemes only in a small range of Δx. It should also be born in mind that the RK3-WENO5 and RK3-MP5 schemes require significantly more CPU time, which downplays their advantage over the second-order schemes.

4 Conclusion

In this paper we presented a systematic study of the solution accuracy behind the shock wave when using Godunov-type schemes. The study was conducted on two schemes of the second-order accuracy and two schemes of the higher-order accuracy (third-order accurate in time and fifth-order accurate in space). The schemes were applied both in their basic versions and in combination with the artificial viscosity approach. By the example of solving one-dimensional gas dynamic problems, the main factors affecting the solution accuracy were identified and analyzed. Let us list these factors and give a brief description of each of them.

Factor 1: start-up errors arising due to inappropriate (inconsistent with numerical viscosity) smearing of the shock in the initial data. These errors are local in nature; they are located near the $(u-a)$- and u-characteristics that come from the initial discontinuity. Whenever they appear, the error in density may be $\sim 1-3\%$.

Factor 2: errors in reproducing the flat post-shock profile. These errors characterize the accuracy of the integral approximation of the Rankine-Hugoniot relations. When computing unsteady flows, the magnitude of such errors is governed mainly by the extent to which the numerical viscosity of a particular scheme is sensitive to the position of the shock relative to the grid (nonuniformity in time). In our study, their level reached $\sim 0.1\%$ in the case of computing an advancing shock wave, and $\sim 2\%$ in the case of computing a slowly moving shock wave. The errors associated with Factor 2 can lead to uneven grid convergence with the convergence rate (on average) noticeably lower than the first one.

Factor 3: entropy trace (or wall heating errors) arising after the interaction of the shock with the wall or symmetry plane. Similar to the errors associated with Factor 1, they are local in nature. In the regions where they occur (near the interaction surface), the error in density may be $\sim 10\%$.

Factor 4: profile shift upon passing through the shock layer. These errors are unavoidable, since the displacement values calculated according to different conservation laws (mass, momentum or total energy) are different. The level of such errors depends linearly on the grid spacing Δx.

Factor 5: profile smoothing upon passing through the shock layer. These errors are also unavoidable due to the effect of scheme dissipation within the shock layer (the smeared shock front). The magnitude of such errors depends on the grid spacing as Δx^2.

Factor 6: the order of accuracy of the scheme. This is the only factor that gives an advantage to higher-order approximation schemes over second-order schemes. It manifests itself at large distances from the shock when using a sufficiently coarse grid. The presence of discontinuities in the flow behind the shock can noticeably weaken the effect of this factor.

The artificial viscosity approach provides for a noticeable decrease in the level of oscillations of the flow parameters behind the shock front. For example, the level of errors associated with Factor 2 may decrease several times, which will lead to a noticeable improvement of the solution in those problems where the influence of this factor is significant. At the same time, adding the artificial viscosity leads to an increase in the shock layer width, thereby increasing the negative impact of Factors 4 and 5 on the accuracy of the solution.

In conclusion, we may say that the extent of influence of each of the above factors depends on the specific problem, the coordinate system in which it is solved, the numerical technique (including all its aspects – the spatial and temporal discretizations, etc.), the degree of grid refinement. In some cases, the second-order schemes may exceed the higher-order schemes in accuracy, in other cases vice versa. However, in our study, we did not find much superiority of higher-order schemes, which can be observed when solving problems with smooth solutions (without gasdynamic discontinuities).

The results of studying this problem in multidimensional simulations will be presented in subsequent publications.

Acknowledgments

This work has been supported by the Russian Science Foundation (project 22-11-00199). The author would like to thank Tatiana Zezyulina for the qualified assistance in English and the reviewers for their helpful comments.

References

[1] M.Ya. Ivanov, A.N. Kraiko, The approximation of discontinuous solutions by using through calculation difference schemes, USSR Comput. Math. Math. Phys. 18 (3) (1978) 259-262.

[2] R. Donat, S. Osher, Propagation of error into regions of smoothness for non-linear approximations to hyperbolic equations, Comput. Meth. Appl. Mech. Eng. 80 (1990) 59-64.

[3] V.V. Ostapenko, Convergence of finite-difference schemes behind a shock front, Comput. Math. Math. Phys. 37 (10) (1997) 1161-1172.

[4] J. Casper, M.H. Carpenter, Computational considerations for the simulation of shock-induced sound, SIAM J. Sci. Comput. 19 (3) (1998) 813-828.

[5] B. Engquist, B. Sjögreen, The convergence rate of finite difference schemes in the presence of shock, SIAM J. Numer. Anal. 35 (6) (1998) 2464-2485.

[6] G. Efraimsson, G. Kreiss, A remark on numerical errors downstream of slightly viscous shocks, SIAM J. Numer. Anal. 1999. Vol. 36 (3). P. 853-863.

[7] G. Kreiss, G. Efraimsson, J. Nordström, Elimination of first order errors in shock calculation, SIAM J. Numer. Anal. 38 (6) (2001) 1986-1998.

[8] M. Siklosi, G. Kreiss, Elimination of first order errors in time-dependent shock calculation, SIAM J. Numer. Anal. 41 (6) (2003) 2131-2148.

[9] J.A. Greenough, W.J. Rider, A quantitative comparison of numerical methods for the compressible Euler equations: Fifth-order WENO and piecewise-linear Godunov, J. Comput. Phys. 196 (2004) 259-281.

[10] M. Siklosi, G. Efraimsson, Analysis of first order errors in shock calculations in two space dimension, SIAM J. Numer. Anal. 43 (2) (2005) 672-685.

[11] A. Suresh, Interaction of a shock with a density disturbance via shock fitting, J. Comput. Phys. 206 (2005) 6-15.

[12] N.A. Mikhailov, The convergence order of WENO schemes behind a shock front, Math. Models Comput. Simul. 7 (5) (2015) 467-474.

[13] B, Engquist, B.D. Froese, Y.-H. R. Tsai, Fast sweeping methods for hyperbolic systems of conservation laws at steady state II, J. Comput. Phys. 286 (2015) 70-86.

[14] O.A. Kovyrkina, V.V. Ostapenko, On the construction of combined finite-difference schemes of high accuracy, Dokl. Math. 97 (1) (2018) 77-81.

[15] M.E. Ladonkina, O.A. Neklyudova, V.V. Ostapenko, V.F. Tishkin, On the accuracy of the discontinuous Galerkin method in calculation of shock waves, Comput. Math. Math. Phys. 58 (8) (2018) 1344-1353.

[16] G. Zhao, M. Sun, A. Memmolo, S. Pirozzoli, A general framework for the evaluation of shock-capturing schemes, J. Comput. Phys. 376 (2019) 924-936.

[17] A.V. Rodionov, Artificial viscosity in Godunov-type schemes to cure the carbuncle phenomenon, J. Comput. Phys. 345 (2017) 308-329.

[18] A.V. Rodionov, Artificial viscosity to cure the carbuncle phenomenon: The three-dimensional case, J. Comput. Phys. 361 (2018) 50-55.

[19] A.V. Rodionov, Artificial viscosity to cure the shock instability in high-order Godunov-type schemes, Comput. Fluids. 190 (2019) P. 77-97.

[20] A.V. Rodionov, Simplified artificial viscosity approach for curing the shock instability, Comput. Fluids. 219 (2021) 104873.

[21] E.F. Toro, M. Spruce, W. Speares, Restoration of the contact surface in the HLL-Riemann solver, Shock Waves 4 (1994) 25-34.

[22] V.P. Kolgan, Application of the principle of minimizing the derivative to the construction of finite-difference schemes for computing discontinuous solutions of gas dynamics, Uch. Zap. TsAGI (Sci. Notes TsAGI) 3(6) (1972) 68-77 [in Russian]; translation: J. Comput. Phys. 230 (2011) 2384-2390.

[23] A. Harten, High resolution schemes for hyperbolic conservation laws, J. Comput. Phys. 49 (1983) 357-393.

[24] B. van Leer, Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection, J. Comput. Phys. 23 (1977) 276-299.

[25] R.J. LeVeque, Finite volume methods for hyperbolic problems, Cambridge University Press, 2002.

[26] A.V. Rodionov, A comparison of the CABARET and MUSCL-type schemes, Math. Models Comput. Simul. 6 (2) (2014) 203-225.

[27] G.-S. Jiang, C.-W. Shu, Efficient implementation of weighted ENO schemes, J. Comput. Phys. 126 (1996) 202-228.

[28] A. Suresh, H.T. Huynh, Accurate monotonicity-preserving schemes with Runge–Kutta time stepping, J. Comput. Phys. 136 (1997) P. 83-99.

[29] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, J. Comput. Phys. 77 (1988) 439-471.

[30] A.V. Rodionov, Monotonic scheme of the second order of approximation for the continuous calculation of non-equilibrium flows, USSR Comput. Math. Math. Phys. 27 (2) (1987) 175-180.

[31] A.V. Rodionov, Methods of increasing the accuracy in Godunov's scheme, USSR Comput. Math. Math. Phys. 27 (6) (1987) 164-169.

[32] G.D. van Albada, B. van Leer, W.W. Roberts, A comparative study of computational methods in cosmic gas dynamics, Astron. Astrophys. 108 (1982) 76-84.

[33] B. van Leer, On the relation between the upwind-differencing schemes of Godunov, Engquist–Osher and Roe, SIAM J. Sci. Stat. Comput. 5 (1) (1984) 1-20.

[34] B.L. Rozhdestvenskii, N.N. Yanenko, Systems of quasilinear equations and their applications to gas dynamics. Amer. Math. Soc., 1983 (translated from the Russian).

[35] W.F. Noh, Errors for calculations of strong shocks using an artificial viscosity and an artificial heat flux, J. Comput. Phys. 72 (1987) 78-120.

[36] C.-W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, II, J. Comput. Phys. 83 (1989) 32-78.