Journal of Scientific Computing (2024) 100:83
https://doi.org/10.1007/s10915-024-02628-9

Asymptotic and Invariant-Domain Preserving Schemes for Scalar Conservation Equations with Stiff Source Terms and Multiple Equilibrium Points

Alexandre Ern[1,2] · **Jean-Luc Guermond**[3] · **Zuodong Wang**[1,2]

Received: 21 December 2023 / Revised: 4 July 2024 / Accepted: 17 July 2024 /
Published online: 2 August 2024

Abstract
We propose an operator-splitting scheme to approximate scalar conservation equations with stiff source terms having multiple (at least two) stable equilibrium points. The scheme combines a (reaction-free) transport substep followed by a (transport-free) reaction substep. The transport substep is approximated using the forward Euler method with continuous finite elements and graph viscosity. The reaction substep is approximated using an exponential integrator. The crucial idea of the paper is to use a mesh-dependent cutoff of the reaction time-scale in the reaction substep. We establish a bound on the entropy residual motivating the design of the scheme. We show that the proposed scheme is invariant-domain preserving under the same CFL restriction on the time step as in the nonreactive case. Numerical experiments in one and two space dimensions using linear, convex, and nonconvex fluxes with smooth and nonsmooth initial data in various regimes show that the proposed scheme is asymptotic preserving.

Keywords Stiff sources · Time-integration methods · Conservation equations · Asymptotic preserving · Invariant domain

Mathematics Subject Classification 35L65 · 65M12 · 65M60 · 76V05

✉ Alexandre Ern
alexandre.ern@enpc.fr

1 CERMICS, Ecole des Ponts, 77455 Marne-la-Vallée Cedex 2, France

2 Inria Paris, 48, rue Barrault, CS 61534, 75647 Paris Cedex, France

3 Department of Mathematics, Texas A&M University 3368 TAMU, College Station, TX 77843, USA

1 Introduction

The goal of the paper is to devise approximation schemes for scalar conservation equations with stiff reaction terms having multiple stable equilibrium points. More precisely, we consider the following scalar-valued PDE:

$$\partial_t u^\epsilon + \nabla\cdot \boldsymbol{f}(u^\epsilon) = \frac{1}{\epsilon} R(u^\epsilon) \qquad \text{in } Q, \tag{1}$$

posed in the space-time cylinder $Q := D \times (0, T)$, where D is an open bounded polyhedral subset of $\mathbb{R}^d$, $d \geq 1$, and $T > 0$ is the observation time. The problem is equipped with suitable initial data, u_0, and boundary conditions. Here, the $\mathbb{R}^d$-valued function $\boldsymbol{f}$ is the flux, and the real-valued function R is the reaction term. In many situations, a fundamental property of (1) is that the entropy solution takes values in a bounded interval $\mathcal{B} \subset \mathbb{R}$, which we henceforth call invariant domain. The set $\mathcal{B}$ typically depends on the initial and boundary conditions and on the equilibrium points of R. The property of $\mathcal{B}$ being invariant, also called maximum principle, means that

$$u^\epsilon \in \mathcal{B} \qquad \text{in } Q. \tag{2}$$

Without loss of generality, we assume that $\mathcal{B} := [0, 1]$. More precise statements on the model problem (1) are given in §2.

The stiffness of the system is quantified by the time scale $\epsilon > 0$. We are interested in the stiff regime $\epsilon \ll \min(T, \beta^{-1}\ell_D)$, where $\ell_D := \operatorname{diam}(D)$ is a reference length and $\beta := \operatorname{Lip}_{\mathcal{B}}(\boldsymbol{f})$ a reference speed. Since u^ϵ takes values in the bounded set $\mathcal{B}$, it is possible to identify a limit solution as $\epsilon \to 0$, say $u^0 := \lim_{\epsilon\to 0} u^\epsilon$, at least in the weak$\star$ L^∞-topology. (The limit solution $u^0 : D\times(0, T) \to \mathcal{B}$ should not be confused with the initial data $u_0 : D \to \mathcal{B}$ for the problem (1).) Typically, one expects that the limit solution u^0 takes values in the subset

$$\mathcal{E} := R^{-1}(\{0\}) \subset \mathcal{B}, \tag{3}$$

which contains the equilibrium (or stationary) points of the problem. We henceforth assume that $\mathcal{E}$ is composed of a least three states, two or more being stable equilibria. In this situation, one expects that the limit solution u^0 consists of several constant states in $\mathcal{E}$ separated by shocks moving at different (a priori unknown) speeds. The shock speeds generally differ from those known in the nonreactive case. One crucial issue in the numerical approximation is to predict the correct shock speeds. Albeit simplified, the model problem (1) is representative of some of the difficulties encountered in the numerical simulation of nonequilibrium gas dynamics in hypersonic flows and other combustion problems. We also emphasize that the situation considered in the paper with multiple equilibrium points is more challenging than the setting with a single equilibrium point (as, e.g., in dissipative systems and relaxation problems).

As discussed in Colella et al. [6], LeVeque and Yee [21], devising numerical approximation schemes for the model problem (1) that work well in the limit $\epsilon \to 0$ is quite challenging. Many standard methods yield wrong shock speeds in the reaction dominant regime when the mesh is not fine enough. Our goal is to design approximation schemes endowed with the following two key properties. The first one is to be invariant-domain preserving (IDP), i.e., to deliver a discrete solution u_h^ϵ (the subscript h refers to the mesh size used for the discretization, see §3) such that

$$u_h^\epsilon \in \mathcal{B} \qquad \text{in } Q. \tag{4}$$

The second one is to ensure the so-called asymptotic preserving (AP) property:

$$\lim_{\epsilon\to 0}\lim_{h\to 0} u_h^\epsilon = u^0 = \lim_{h\to 0}\lim_{\epsilon\to 0} u_h^\epsilon. \tag{5}$$

The left equality simply means that the scheme is convergent, whereas the right equality means that the limit solution u^0 can be captured in the under-resolved regime as well, i.e., the scheme is consistent with the limit equation as $\epsilon \to 0$. The reader is referred to Jin [16] for a review on AP schemes.

We propose in the paper a time-stepping scheme that is observed to be AP and is proved to be IDP with the same CFL restriction as in the nonreactive case. The AP property is based on numerical observations, whereas the IDP property is based on theoretical arguments. We follow the well-established paradigm of operator-splitting schemes, i.e., we perform, at each time step, a forward Euler, nonreactive transport substep followed by a transport-free reaction substep. Many schemes are available in the literature to perform the transport substep: finite differences, finite volumes, discontinuous or continuous finite elements. We focus here on continuous finite elements with graph viscosity, as in [11, 12]. A natural idea for the reaction substep consists of employing an implicit scheme, or, often better, an exponential-like integrator (see Hochbruck and Ostermann [14] for a review). This approach has been successful for dissipative systems, kinetic equations, and systems with relaxation; see, among others Chalabi [5], Chainais-Hillairet and Champier [4], Pareschi and Russo [22], Filbet and Rambaud [10], Bulteau et al. [3], Hu and Shu [15]. However, in the present situation, using an implicit scheme or an exponential integrator is not AP, as these schemes usually predict shocks moving with the wrong speed as $\epsilon \to 0$. This phenomenon has been discussed in LeVeque and Yee [21]; see also Colella et al. [6, 7, 17]. The key reason for this odd behavior is that stiffness makes the discrete solution too sensitive to the smeared representation of discontinuities separating equilibrium states.

In order to temper stiffness and achieve the AP property without sacrificing too much accuracy, the main idea of the paper is to introduce a mesh-dependent cutoff on the reaction time-scale when performing the reaction substep. The resulting operator-splitting scheme satisfies the following properties: (i) It is IDP by design; (ii) It satisfies discrete entropy inequalities; (iii) It yields optimal accuracy in the resolved regime; (iv) It is observed to be AP thorough numerical experiments in one and two space dimensions using linear, convex, and nonconvex fluxes.

The literature on IDP-AP schemes for the present problem is relatively scarce. To our knowledge, the few (IDP-)AP schemes available in the literature somehow exploit the knowledge of the limit equation or work only in special situations. Two salient examples are the random projection scheme devised in Bao and Jin [1, 2] and the IMEX scheme proposed in Svärd and Mishra [23]. The projection scheme works for discontinuous (shock-type) initial data and convex flux, and the IMEX scheme is tailored to situations for which the location of the discontinuities can be predicted by the solution to the homogeneous problem (see §6 for other details). In contrast, the scheme proposed in the paper does not require any knowledge on the limit equation and can handle a wide range of situations, including nonconvex fluxes, general initial data, and discontinuities propagating at a priori unknown speeds.

The rest of the paper is organized as follows. The model problem is presented in §2. The discrete setting together with the proposed scheme are discussed in §3. The main results of this section are Proposition 3.1 and Theorem 3.3, which establish, respectively, that the scheme is IDP and that it satisfies entropy inequalities with a residual decaying to zero under some assumptions. Numerical results are presented in §4, §5, §6. All the numerical experiments are conducted with the help of the `Gridap.jl` library developed by Verdugo and Badia [24] in

the `julia` programming language. In §4, we study the cutoff parameters and identify an all-purpose cutoff strategy ensuring that the AP property holds for all our numerical experiments. In §5, we assess the cutoff strategy on challenging test cases. In particular, we highlight that the mesh-dependent cutoff strategy introduced herein allows one to capture the correct shock location even in the under-resolved regime (recall that the presence of multiple equilibrium points causes the shocks to travel at speeds that differ from the nonreactive case). In §6, we finally compare the proposed method to existing schemes from the literature. The main conclusion is that the proposed mesh-dependent cutoff strategy leads to operator-splitting schemes that perform better than existing schemes from the literature when simulating scalar conservation equations with stiff source terms and multiple equilibrium points. Finally, §7 contains the proof of Theorem 3.3.

2 Model Problem

We consider the PDE (1) posed in the space-time cylinder $D \times (0, T)$ together with the initial condition $u_0 \in L^\infty(D; \mathcal{B})$ with $\mathcal{B} := [0, 1]$. The flux and the source term are assumed to be smooth with

$$\boldsymbol{f} \in C^1(\mathcal{B}; \mathbb{R}^d), \qquad R \in C^1(\mathcal{B}; [-1, 1]). \tag{6}$$

Since the time scale $\epsilon > 0$ is used to quantify the strength of the source term, we assume without loss of generality that R takes values in $[-1, 1]$. We assume that there are $0 < \vartheta_0 < \vartheta_1 < 1$ such that

$$\partial\mathcal{B} = \{0, 1\} \subset \mathcal{E}, \qquad R(v) < 0 \ \forall v \in (0, \vartheta_0), \quad R(v) > 0 \ \forall v \in (\vartheta_1, 1), \tag{7}$$

meaning that both 0 and 1 are stable equilibrium points. The values 0 and 1 are chosen for normalization purposes without loss of generality. In general situations, these two values must be replaced by the smallest and largest values of the stable equilibrium points associated with R. The intermediate value theorem then implies that there is at least another equilibrium point $\alpha \in (0, 1)$. The simplest setting is when $\mathcal{E} = \{0, \alpha, 1\}$ and α is an unstable equilibrium point. The following prototypical example considered by LeVeque and Yee [21] meets the above assumptions:

$$R(v) = R_\alpha(v) := v(1 - v)(v - \alpha), \qquad \forall v \in [0, 1]. \tag{8}$$

To avoid distracting technicalities with the boundary conditions, we assume that: (i) either the initial data u_0 is compactly supported in D and the observation time T is short enough so that the solution u^ϵ remains compactly supported at all times $t \in [0, T]$; (ii) or a zero boundary condition is enforced at all times at any inflow boundary and the solution u^ϵ vanishes in a neighborhood of the inflow boundary at all times. Both assumptions are reasonable since 0 is a stable equilibrium point.

For any fixed $\epsilon > 0$, the Cauchy problem admits a unique entropy solution (see e.g., Kružkov [19, Thm. 2]). Specifically, for any convex entropy $\eta \in W^{1,\infty}(\mathcal{B}; \mathbb{R})$ with entropy flux $\boldsymbol{q}(u) := \int_0^u \eta'(v)\boldsymbol{f}'(v)\mathrm{d}v$, and for any test function $\psi \in W_0^{1,\infty}(D \times [0, T]; \mathbb{R}_+)$, the unique entropy solution in $L^\infty(Q; \mathbb{R})$ is such that the following holds:

$$\begin{aligned} &\int_D \eta\big(u^\epsilon(\boldsymbol{x}, T)\big)\psi(\boldsymbol{x}, T)\mathrm{d}\boldsymbol{x} - \int_D \eta\big(u_0(\boldsymbol{x})\big)\psi(\boldsymbol{x}, 0)\mathrm{d}\boldsymbol{x} \\ &\quad - \int_Q \big\{\eta(u^\epsilon)\partial_t\psi + \boldsymbol{q}(u^\epsilon)\cdot\nabla\psi\big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t \le \int_Q \frac{1}{\epsilon}\eta'(u^\epsilon)R(u^\epsilon)\psi\mathrm{d}\boldsymbol{x}\mathrm{d}t. \end{aligned} \tag{9}$$

A characterization of the limit solution u^0 as $\epsilon \to 0$ is available in one space dimension ($d = 1$) for convex fluxes. In this case, and assuming that the reaction term is given by (8) (the result can be extended to more general reaction terms), it is shown in Fan et al. [9, Thm. 1.1] that u^0 takes values in $\{0, 1\}$ (as expected), with shocks moving at speed $\frac{f(1)-f(0)}{1-0}$ when the left and right states are $(1, 0)$ and at speed $f'(\alpha)$ when the left and right states are $(0, 1)$. Notice that in the first case, the shock speed coincides with that given by the Rankin–Hugoniot relation just like for the nonreactive problem. In the second case, the solution to the nonreactive problem is a rarefaction wave, whereas the limit solution u^0 for the reactive problem features a shock whose speed does not satisfy the Rankin–Hugoniot relation (since, in general, $f'(\alpha) \neq \frac{f(1)-f(0)}{1-0}$). To the best of our knowledge, the characterization of the limit solution remains an open problem in more general situations.

3 Discrete Scheme

In this section, we present our scheme and establish that, under some assumptions, the scheme is IDP and that it satisfies an entropy inequality with a residual decaying to zero as the mesh is refined, uniformly in ϵ.

3.1 Discrete Setting

The time discretization is defined by using the collection of discrete time nodes t^n for all $n \in \mathcal{N} := \{0{:}N\}$, with $t^0 = 0$ and $t^N = T$. The time step τ^n is defined as $\tau^n := t^{n+1} - t^n$, and we set $I_n := [t^n, t^{n+1})$ for all $n \in \mathcal{N} := \{0{:}N-1\}$. To simplify the notation, we omit the superscript n and denote the time step by τ.

To stay general, we do not specify the space discretization scheme yet; more details are given in §3.3 in the context of continuous finite elements. Possible space discretization methods are, e.g., finite volumes, finite differences, discontinuous or continuous finite elements. At this stage, we just assume that the space discretization is based on a mesh $\mathcal{T}_h$ that belongs to a quasi-uniform mesh sequence. Here, h denotes the mesh size, i.e., the largest diameter of the mesh cells. The space discretization is characterized by a collection of degrees of freedom (dofs) which we enumerate with the index set $\mathcal{V}$. The set $\mathcal{V}$ is partitioned as $\mathcal{V} = \mathcal{V}^\circ \cup \mathcal{V}^\partial$, where $\mathcal{V}^\circ$ collects the interior dofs and $\mathcal{V}^\partial$ the (inflow) boundary dofs. We denote $I := \mathrm{card}(\mathcal{V})$.

The operator-splitting scheme is composed of a (reaction-free) transport substep followed by a (transport-free) reaction substep at each time step $n \in \mathcal{N}$:

$$(\mathsf{U}_i^n)_{i\in\mathcal{V}} \xrightarrow{\text{transport}} (\mathsf{W}_i^{n+1})_{i\in\mathcal{V}} \xrightarrow{\text{reaction}} (\mathsf{U}_i^{n+1})_{i\in\mathcal{V}}, \tag{10}$$

where $(\mathsf{U}_i^n)_{i\in\mathcal{V}}$ is obtained from the previous time step if $n \geq 1$ or by a suitable approximation of the initial condition if $n = 0$. For both transport and reaction substeps, boundary conditions can be enforced by requiring that $\mathsf{U}_i^{n+1} = \mathsf{W}_i^{n+1} = \mathsf{U}_i^n = 0$ for all $i \in \mathcal{V}^\partial$.

We henceforth assume that the transport substep is IDP under a CFL restriction on the time step: There exists a real number τ^* depending on β and h so that for all

$$\tau \leq \tau^*, \tag{11}$$

the following holds true:

$$\left(\mathsf{U}^n := (\mathsf{U}_i^n)_{i\in\mathcal{V}} \in \mathcal{B}^I\right) \implies \left(\mathsf{W}^{n+1} := (\mathsf{W}_i^{n+1})_{i\in\mathcal{V}} \in \mathcal{B}^I\right). \tag{12}$$

We briefly show in §3.3 how (12) is achieved using continuous finite elements.

3.2 Details on the Reaction Substep and IDP Property

We describe in this section a method to perform the reaction substep in (10). The central idea of the paper is to regularize the stiffness parameter ϵ in the reaction substep by using a mesh-dependent cutoff. Recall that there are two reference times, T and $\beta^{-1}\ell_D$. For the sake of simplicity, we assume that both times are of the same order of magnitude, and we use T as the reference time. Up to straightforward modifications, everything that is said hereafter remains valid if one replaces T by $\min(T, \beta^{-1}\ell_D)$ in (13).

We define a regularized stiffness time using two user-dependent parameters (θ, γ), both in $(0, 1]$, as follows:

$$\Phi_{\epsilon,\gamma,\theta} := \max\left(\epsilon, \gamma T\left(\frac{h}{\beta T}\right)^{\theta}\right). \tag{13}$$

The time scale T and the velocity scale β are introduced for dimensional consistency; in the nondimensional setting, one simply obtains $\Phi_{\epsilon,\gamma,\theta} := \max(\epsilon, \gamma h^{\theta})$, which better highlights that the two parameters available to tune the cutoff are γ and θ. How to choose the parameters (θ, γ) is thoroughly discussed in §4. Setting

$$h_{\epsilon,\gamma,\theta} := \beta T\left(\frac{\epsilon}{\gamma T}\right)^{\frac{1}{\theta}}, \tag{14}$$

we have $\Phi_{\epsilon,\gamma,\theta} := \epsilon \max(1, (\frac{h}{h_{\epsilon,\gamma,\theta}})^{\theta})$. We say that we are in the resolved regime when $h \lesssim h_{\epsilon,\gamma,\theta}$ and in the under-resolved regime when $h \gtrsim h_{\epsilon,\gamma,\theta}$. Hence, $\Phi_{\epsilon,\gamma,\theta} = \epsilon$ in the resolved regime and $\Phi_{\epsilon,\gamma,\theta} = \gamma T\left(\frac{h}{\beta T}\right)^{\theta}$ in the under-resolved regime. Selecting the parameters (θ, γ) deserves some attention. Ideally, one would like to pick θ close (or equal) to one to make the resolved regime as large as possible (i.e., when $h \lesssim \frac{1}{\gamma}\beta\epsilon$). However, numerical experiments reported in §4 show that it may happen that $\lim_{h\to 0}\lim_{\epsilon\to 0} u_h^{\epsilon} \neq u^0$ (see (5)) when $\theta > \frac{1}{2}$, which means that the scheme is not AP when $\theta > \frac{1}{2}$. On the other hand, numerical experiments suggest that the scheme is indeed AP for $\theta \in (0, \frac{1}{2}]$. Moreover, we establish in §3.4 a bound on the entropy residual for which $\theta < \frac{1}{2}$ is a sufficient condition for the residual to decay to zero.

The reaction substep is based on the general idea of exponential integrators with two salient differences with respect to what is usually done in the literature. First, the part of the source term that is integrated exactly is quadratic and is based on the two equilibrium states $\{0, 1\}$ composing the boundary of the invariant domain $\mathcal{B}$. Second, and more importantly, the time integration is not performed over the time interval $[0, \frac{\tau}{\epsilon}]$ but over the (generally) shorter time interval $[0, \tau_{\epsilon,\gamma,\theta}]$ with

$$\tau_{\epsilon,\gamma,\theta} := \frac{\tau}{\Phi_{\epsilon,\gamma,\theta}}. \tag{15}$$

We start by defining the function $\tilde{R}(v) := \frac{R(v)}{v(1-v)}$ for all $v \in \mathcal{B}$ (this function is well-defined on $(0, 1)$ and is continuously extended to $\mathcal{B} = [0, 1]$ using l'Hôpital's rule: $\tilde{R}(0) = R'(0)$, $\tilde{R}(1) = -R'(1)$). For instance, when R is defined by (8), we have $\tilde{R}(v) = v - \alpha$ for all $v \in \mathcal{B}$. The next step is to consider the ODE

$$\begin{cases} \dfrac{d}{ds}\vartheta(v; s) = \vartheta(v; s)(1 - \vartheta(v; s))\tilde{R}(v), \quad s \geq 0, \\ \vartheta(v; 0) = v \in \mathcal{B}. \end{cases} \tag{16}$$

Finally, the reaction substep in (10) is defined by setting

$$\mathsf{U}_i^{n+1} = \vartheta(\mathsf{W}_i^{n+1}; \tau_{\epsilon,\gamma,\theta}), \qquad \forall i \in \mathcal{V}. \tag{17}$$

As the solution to (16) is $\vartheta(v; s) = v\exp(s\tilde{R}(v))/(1 + v(\exp(s\tilde{R}(v)) - 1)$, we obtain

$$\mathsf{U}_i^{n+1} = \frac{\mathsf{W}_i^{n+1}\exp\left(\tau_{\epsilon,\gamma,\theta}\tilde{R}(\mathsf{W}_i^{n+1})\right)}{1 + \mathsf{W}_i^{n+1}\left(\exp\left(\tau_{\epsilon,\gamma,\theta}\tilde{R}(\mathsf{W}_i^{n+1})\right) - 1\right)}, \qquad \forall i \in \mathcal{V}. \tag{18}$$

Notice that for the boundary dofs, the above expression gives $\mathsf{U}_i^{n+1} = \mathsf{W}_i^{n+1} = 0$ since $\vartheta(0; s) = 0$ for all $s \geq 0$.

Proposition 3.1 (IDP) *Assume that the CFL restriction (11) holds. Let the reaction step be defined in (17). Then, the operator-splitting scheme (10) is IDP.*

Proof Assume that $\mathsf{U}^n \in \mathcal{B}^I$. The property (12) holds owing to the CFL restriction (11), and we infer that $\mathsf{W}^{n+1} \in \mathcal{B}^I$. Furthermore, the solution $\vartheta(v; s)$ of the ODE (16) stays in $\mathcal{B}$ for all $s \geq 0$ whenever $v \in \mathcal{B}$, whence $\mathsf{U}_i^{n+1} \in \mathcal{B}$ for all $i \in \mathcal{V}$. Thus, $\mathsf{U}^{n+1} \in \mathcal{B}^I$, i.e., the operator-splitting scheme (10) is IDP. □

Remark 3.2 *(Alternative)* An alternative to the reaction substep defined in (18) is to use a forward Euler substep with an additional clipping on the time step to ensure that the update is IDP. Specifically, we observe that there is $\chi > 0$ such that

$$v \in \mathcal{B} \implies v + \rho R(v) \in \mathcal{B}, \ \forall \rho \in [-\chi, \chi].$$

For instance, one can take $\chi = \min(\frac{1}{\alpha}, \frac{1}{1-\alpha}) \leq 2$ when $R(v) := v(1-v)(v-\alpha)$. Then, the reaction substep is defined by setting

$$\mathsf{U}_i^{n+1} = \mathsf{W}_i^{n+1} + \min(\chi, \tau_{\epsilon,\gamma,\theta})R(\mathsf{W}_i^{n+1}), \qquad \forall i \in \mathcal{V}.$$

By construction, $\mathsf{U}_i^{n+1} \in \mathcal{B}$ whenever $\mathsf{W}_i^{n+1} \in \mathcal{B}$, and therefore the operator-splitting scheme (10) is IDP under the CFL restriction (11). Note, however, that the clipping of the time step can become a hindrance if χ is very small.

3.3 Finite-Element Transport Substep

The discretization of the transport step using continuous finite elements can be done in many ways. We follow here the technique described in [11, 12]; see also [8, Chaps. 79-83] for an easy introduction to the method. Recall that the mesh $\mathcal{T}_h$ belongs to a quasi-uniform mesh sequence. We assume that the mesh is composed of (affine) simplices. We focus, for simplicity, on continuous, piecewise affine finite elements. Thus, the dofs are the values at the mesh vertices, and the boundary dofs are the values at the mesh vertices located at the inflow boundary. The global shape functions are denoted by $\{\varphi_i\}_{i\in\mathcal{V}}$. The stencil associated with the dof $i \in \mathcal{V}$ is defined as

$$\mathcal{I}(i) := \{j \in \mathcal{V} \mid \varphi_i\varphi_j \neq 0\}, \tag{19}$$

The notion of stencil is symmetric, i.e., $j \in \mathcal{I}(i)$ iff $i \in \mathcal{I}(j)$. The global shape functions satisfy the following partition of unity property: $\sum_{i\in\mathcal{V}}\varphi_i(\boldsymbol{x}) = 1$, for all $\boldsymbol{x} \in D$. The matrix with entries $m_{ij} := \int_D \varphi_i(\boldsymbol{x})\varphi_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, for all $i, j \in \mathcal{V}$ is called the consistent mass matrix. The lumped mass matrix has entries equal to $m_i := \sum_{j\in\mathcal{I}(i)} m_{ij} = \int_D \varphi_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x} > 0$, for all

$i \in \mathcal{V}$. For all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)\backslash\{i\}$, we define the vectors $\boldsymbol{c}_{ij} := \int_D \varphi_i(\boldsymbol{x})\nabla\varphi_j(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{n}_{ij} := \frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|_2} \in \mathbb{R}^d$. Finally, we also define the first-order graph-viscosity coefficients

$$d_{ij}^n := \max(\lambda_{\max}(\mathsf{U}_i^n, \mathsf{U}_j^n, \boldsymbol{n}_{ij})\|\boldsymbol{c}_{ij}\|_2, \lambda_{\max}(\mathsf{U}_j^n, \mathsf{U}_i^n, \boldsymbol{n}_{ji})\|\boldsymbol{c}_{ji}\|_2), \tag{20}$$

where $\lambda_{\max}(\mathsf{U}_i^n, \mathsf{U}_j^n, \boldsymbol{n}_{ij})$ is any upper bound on the maximum wave speed in the Riemann problem with data $(\mathsf{U}_i^n, \mathsf{U}_j^n)$ and flux $\boldsymbol{f}\cdot\boldsymbol{n}_{ij}$.

With the above definitions, the finite element realization of the transport substep reads as follows: For all $n \in \mathcal{N}$,

$$\mathsf{W}_i^{n+1} = \mathsf{U}_i^n - \frac{\tau}{m_i}\sum_{j\in\mathcal{I}(i)\backslash\{i\}}\big\{\boldsymbol{f}(\mathsf{U}_j^n)\cdot\boldsymbol{c}_{ij} - d_{ij}^n(\mathsf{U}_j^n - \mathsf{U}_i^n)\big\}, \qquad \forall i \in \mathcal{V}, \tag{21}$$

where $\mathsf{U}^n = (\mathsf{U}_i^n)_{i\in\mathcal{V}}$ is either known from the previous time step if $n \geq 1$ or prescribed by the initial condition (e.g., $\mathsf{U}_i^0 := \frac{1}{m_i}\int_D u^0(\boldsymbol{x})\varphi_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, for all $i \in \mathcal{V}$). Recall that owing to the assumptions made on the initial data and the boundary conditions, the update (21) also holds true for the boundary dofs and gives $\mathsf{W}_i^{n+1} = \mathsf{U}_i^n = 0$.

A crucial property of the transport substep (21) is that it is IDP (i.e., (12) holds true) under the CFL restriction

$$\tau \leq \tau^* := \min_{i\in\mathcal{V}^\circ}\frac{m_i}{2\sum_{j\in\mathcal{I}(i)\backslash\{i\}} d_{ij}^n}, \qquad \forall n \in \mathcal{N}. \tag{22}$$

3.4 Bound on Entropy Residual

The main result of this section is that, under reasonable assumptions, the above scheme satisfies entropy inequalities with a residual that decays to zero with the mesh size. The proof is postponed to §7. For all $n \in \{0{:}N\}$, we reconstruct from the dofs $(\mathsf{U}_i^n)_{i\in\mathcal{V}}$ a continuous function that is piecewise affine in space by setting

$$u_h^n(\boldsymbol{x}) := \sum_{i\in\mathcal{V}}\mathsf{U}_i^n\varphi_i(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \overline{D}. \tag{23}$$

Then, we reconstruct a piecewise constant function in time by setting

$$u_h^\epsilon(\boldsymbol{x}, t)|_{[t^n, t^{n+1})} := u_h^n(\boldsymbol{x}), \ \forall n \in \mathcal{N}, \qquad u_h^\epsilon(\boldsymbol{x}, T) := u_h^N(\boldsymbol{x}). \tag{24}$$

The Lagrange interpolant is defined as $\mathcal{I}_h(v)(\boldsymbol{x}) := \sum_{i\in\mathcal{V}} v(\boldsymbol{a}_i)\varphi_i(\boldsymbol{x})$, for any function $v \in C^0(\overline{D})$ and all $\boldsymbol{x} \in \overline{D}$, where $\boldsymbol{a}_i$ denotes the mesh vertex associated with the global shape function φ_i. The same definition is used componentwise for vector-valued fields.

Theorem 3.3 (Bound on entropy residual) *Let the transport substep be defined in (21). Let the reaction substep be defined in (17). Assume that the CFL restriction (11) holds true. Then, there exists a constant C independent of h, τ, and ϵ, but that can depend on the mesh shape-regularity, the functions R, η, and ψ, and the cutoff parameters (θ, γ), such that, for any convex entropy $\eta \in C^2(\mathcal{B}; \mathbb{R})$ with associated flux $\boldsymbol{q}$, and for any test function $\psi \in W_0^{1,\infty}(D \times [0, T]; \mathbb{R}_+)$, we have*

$$\begin{aligned}&\int_D \mathcal{I}_h\big(\eta(u_h^N)\big)(\boldsymbol{x})\psi(\boldsymbol{x}, T)d\boldsymbol{x} - \int_D \mathcal{I}_h\big(\eta(u_h^0)\big)(\boldsymbol{x})\psi(\boldsymbol{x}, 0)d\boldsymbol{x} \\ &- \int_Q\Big\{\eta(u_h^\epsilon)\partial_t\psi + \boldsymbol{q}(u_h^\epsilon)\cdot\nabla\psi + \frac{1}{\Phi_{\epsilon,\gamma,\theta}}\eta'(u_h^\epsilon)R(u_h^\epsilon)\psi\Big\}d\boldsymbol{x}dt \leq C\Lambda(h),\end{aligned} \tag{25}$$

where

$$\Lambda(h) := \frac{h}{\beta\Phi^2_{\epsilon,\gamma,\theta}}\|u_h^\epsilon\|_{L^1(Q)} + \left(\frac{h}{\Phi_{\epsilon,\gamma,\theta}} + \frac{h^2}{\beta\Phi^2_{\epsilon,\gamma,\theta}}\right)\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)}. \tag{26}$$

Remark 3.4 ($\Lambda(h)$) Notice that $\|u_h^\epsilon\|_{L^1(Q)}$ is bounded since u_h^ϵ takes values in the bounded set $\mathcal{B}$. If a uniform bound is available on $\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)}$, (26) shows that, in the under-resolved regime where $\Phi_{\epsilon,\gamma,\theta} \sim h^\theta$, we have $\Lambda(h) \sim h^{1-2\theta}$, and this quantity decays to zero if $\theta \in (0, \frac{1}{2})$ (the first term in (26) is the dominant one). A more realistic assumption is $\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)} \leq Ch^{-\frac{1}{2}}$ (this bound is a consequence of the L^2-estimate, but a sharper BV-estimate is possible in 1D). In this case, one has $\Lambda(h) \sim h^{\frac{1}{2}-\theta}$, which again decays to zero if $\theta \in (0, \frac{1}{2})$ (the second term in (26) now becomes dominant). Finally, in the resolved regime, one obtains $\Lambda(h) \sim h^{\frac{1}{2}}/\epsilon$. The half-order decay in h with fixed ϵ is typical of the nonreactive case.

4 Numerical Study on the Cutoff Parameters

The goal of this section is to numerically study the impact of the cutoff parameters (θ, γ) on the computational performance of the scheme, and therefore propose a rationale for choosing these parameters. We proceed in three steps. First, we show that it is indeed beneficial to use a cutoff on the source term. Second, we find optimal values for the cutoff parameters (θ, γ) on a series of test cases. However, we shall see that these values depend on the flux type (linear, convex, nonconvex) and the form of the reaction term (quantified by the parameter α, see (8)), whereas the dependence on the smoothness of the initial condition appears to be marginal. The third step consists of selecting all-purpose values of the cutoff parameters. Although optimal values of the cutoff parameters are problem-dependent, our numerical experiments indicate that it is still possible to identify all-purpose values for these parameters that produce results that are reasonably close to those produced by the optimal ones.

4.1 Overview of the Test Cases

We consider 1D test cases, all posed on the interval $D := (-1, 1)$, and we are going to explore linear, convex (Burgers), and nonconvex (sine) fluxes, defined respectively as follows:

$$f(v) := v, \qquad f(v) := \frac{1}{2}v^2, \qquad f(v) := \frac{1}{2\pi}\sin(2\pi v). \tag{27}$$

We select the source term to be that defined in (8), and we are going to explore $\alpha \in \{0.5, 0.7, 0.9\}$. We are also going to explore three types of initial data:

- IC1 (smooth (C^0) IC)

$$u_0(x) = \begin{cases} x+1, & \text{if } x \in (-1, 0), \\ 1, & \text{otherwise.} \end{cases} \tag{28}$$

- IC2 (nonsmooth IC with one shock)

$$u_0(x) = \begin{cases} 2(x+1), & \text{if } x \in (-1, -\frac{1}{2}), \\ 1, & \text{if } x \in (-\frac{1}{2}, \frac{2}{5}), \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

- IC3 (nonsmooth IC with two shocks)

$$u_0(x) = \begin{cases} 2(x+1), & \text{if } x \in (-1, -\frac{1}{2}), \\ 0.95, & \text{if } x \in (-\frac{1}{2}, \frac{2}{5}), \\ 0.3, & \text{otherwise.} \end{cases} \tag{30}$$

In all cases, the reference velocity is $\beta := 1$, and we set the final time to $T := 0.5$. The time step is defined by

$$\tau := C_{\text{CFL}} \min_{i \in \mathcal{V}^\circ} \frac{m_i}{2 \sum_{j \in \mathcal{I}(i) \setminus \{i\}} d_{ij}^n}, \tag{31}$$

and, unless stated otherwise, we use $C_{\text{CFL}} = \frac{1}{2}$ in the simulations.

Errors are measured in the relative L^1-norm at the final time (i.e., normalized by the L^1-norm of the exact solution). For the linear flux, the exact solution is computed by the method of characteristics and an implicit Runge–Kutta integrator along the characteristics. For the nonlinear fluxes, the exact solution is obtained on a fine grid with mesh size $h_{\text{ref}} := 0.1 \times 2^{-13} \approx 1.2 \times 10^{-5}$. We consider two values for the stiffness parameter $\epsilon \in \{10^{-2}, 10^{-3}\}$. For these values, and since $T = 0.5$, the difference between u^ϵ and u^0 measured in the relative L^1-norm is of the order of the machine precision for the linear flux, whereas it scales like $\mathcal{O}(\epsilon)$ for the nonlinear fluxes; for instance, this difference is in the range $[1, 5] \times 10^{-3}$ for $\epsilon = 10^{-3}$.

The mesh sizes sampled are $h_j := 0.1 \times 2^{-j}$ for $j \in \{0{:}10\}$, thus we mainly focus on the under-resolved regime. The mesh size h_{ref} we use to approximate the exact solution is eight times smaller than the smallest mesh size h_{10} explored. As the cutoff (13) makes the source term in the numerical scheme independent of ϵ in the under-resolved regime (i.e., $h_{\epsilon,\gamma,\theta} \le h$), and $\|u_h^\epsilon - u^\epsilon\|_{L^1(D)} \approx \|u_h^\epsilon - u^0\|_{L^1(D)}$ for very small values of ϵ, we expect that the error $\|u_h^\epsilon - u^\epsilon\|_{L^1(D)}$ varies very little with respect to ϵ in the under-resolved regime when ϵ is smaller than 10^{-3}. We have numerically tested this statement, and observed that it is indeed the case. We do not report these tests for brevity. In conclusions, we do not report tests done with values of ϵ smaller than 10^{-3}.

4.2 On the Benefits of Using a Cutoff in the Source Term

Recall that we are in the resolved regime when $h \lesssim h_{\epsilon,\gamma,\theta}$ and in the under-resolved regime when $h \gtrsim h_{\epsilon,\gamma,\theta}$, with $h_{\epsilon,\gamma,\theta}$ defined in (14). For the linear flux, we expect the asymptotic convergence rate to be of order 1 for the initial condition IC1 and of order $\frac{1}{2}$ for the initial conditions IC2 and IC3. For the nonlinear fluxes, the asymptotic convergence rate is expected to be between $\frac{1}{2}$ and 1 for the three initial conditions. Recall, however, that we are mainly considering mesh sizes in the under-resolved regime.

Figure 1 shows tests with the stiffness parameter $\epsilon = 10^{-3}$ for the problems defined in §4.1. The relative L^1-errors at the final time are represented as a function of the mesh size for three values of the cutoff parameter $\theta \in \{0.2, 0.4, 0.8\}$, the choice $\gamma = 0.1$, and the three initial conditions. Each panel corresponds to one value of the reaction parameter $\alpha \in \{0.5, 0.7, 0.9\}$ (from left to right) and to one flux (from top to bottom: linear, convex, nonconvex). In each panel, the red dashed curve corresponds to the numerical results without cutoff, i.e., setting $\Phi_{\epsilon,\gamma,\theta} := \epsilon$ (plain exponential integrator, labeled $\Phi = \epsilon$ in the legend). Vertical lines in each panel indicate the value of the mesh size corresponding to the transition from the under-resolved to the resolved regime; the color of the vertical line corresponds to

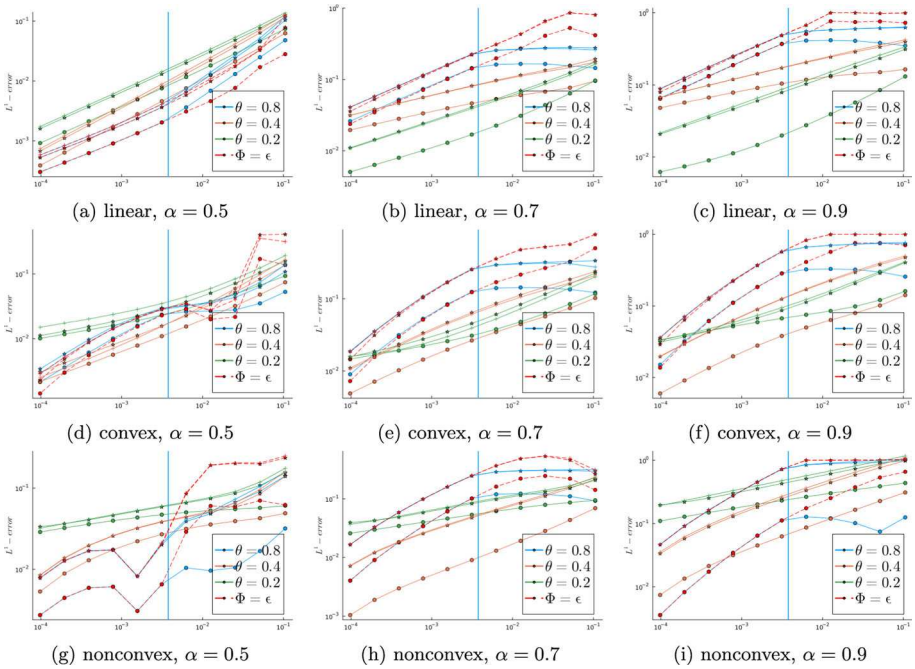


Fig. 1 L^1-errors as a function of the mesh size for $\epsilon = 10^{-3}$, three values of the cutoff parameter ($\theta = 0.8$ in blue, $\theta = 0.4$ in brown, $\theta = 0.2$ in green), $\gamma = 0.1$, and three ICs (IC1: $\circ$, IC2: $\star$, IC3: $\times$). From left to right: $\alpha = 0.5$, $\alpha = 0.7$, and $\alpha = 0.9$. From top to bottom: linear, convex, and nonconvex fluxes. The red dashed line labeled '$\Phi = \epsilon$' is obtained without any cutoff (plain exponential integrator) (Color figure online)

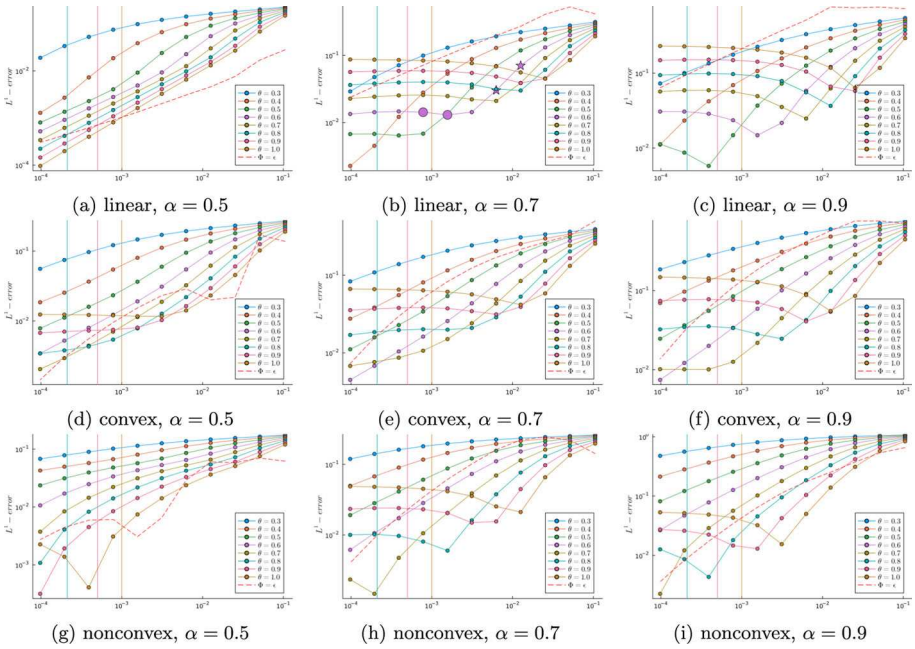
the value of θ. We observe in Fig. 1 that the resolved regime can be reached only for $\theta = 0.8$ with the mesh sizes considered here. The red and blue curves overlap for mesh sizes smaller than the value indicated by the vertical blue line.

Several observations can be made from the results displayed in Fig. 1. Let us focus first on the reaction parameters $\alpha \in \{0.7, 0.9\}$ (central and right columns).

- The errors with no cutoff are generally larger than those obtained with cutoff. The errors level off on the coarser meshes if no cutoff is used.
- Choosing $\theta = 0.8$ is always less effective than choosing $\theta < \frac{1}{2}$. A plateau is observed on the coarser meshes for $\theta = 0.8$. This observation is consistent with the main conclusion of Theorem 3.3 which recommends to select $\theta < \frac{1}{2}$.
- The most effective choice of θ in $\{0.2, 0.4\}$ depends on the flux type. The value $\theta = 0.4$ generally performs better for the nonlinear fluxes, whereas the value $\theta = 0.2$ generally performs better for the linear flux.

Regardless of the error levels, the above conclusions are fairly independent of the initial conditions. The errors obtained with the smooth initial condition IC1 are smaller than those obtained with the nonsmooth initial conditions IC2 and IC3. The differences on the results obtained with the initial data IC2 and IC3 are marginal.

Perhaps a bit surprisingly, the conclusions are less clear cut for $\alpha = 0.5$ (left column in Fig. 1). The most salient observation is that for the linear flux, the scheme without cutoff (i.e., setting $\Phi_{\epsilon,\gamma,\theta} := \epsilon$) generally leads to lower errors. It is, however, still beneficial to use a



(a) linear, $\alpha = 0.5$ (b) linear, $\alpha = 0.7$ (c) linear, $\alpha = 0.9$

(d) convex, $\alpha = 0.5$ (e) convex, $\alpha = 0.7$ (f) convex, $\alpha = 0.9$

(g) nonconvex, $\alpha = 0.5$ (h) nonconvex, $\alpha = 0.7$ (i) nonconvex, $\alpha = 0.9$

Fig. 2 L^1-errors as a function of the mesh size for $\epsilon = 10^{-3}$ and IC1, $\alpha \in \{0.5, 0.7, 0.9\}$ (from left to right), and linear, convex, and nonconvex fluxes (from top to bottom). The cutoff parameters are $\theta \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and $\gamma = 1$. The red dashed line labeled '$\Phi = \epsilon$' is obtained without any cutoff (Color figure online)

cutoff for the nonlinear fluxes in the under-resolved regime. Some clarifications about these observations are given in §4.3.

4.3 Optimizing the Cutoff Parameters

The above results lead us to wonder whether it is possible to devise an optimal strategy to define the cutoff parameters (θ, γ). For this purpose, we fix $\gamma = 1$ and we perform a sampling of θ in the interval $[0.3, 1.0]$ with step $\delta\theta = 0.1$. We explore again the three reaction coefficients $\alpha \in \{0.5, 0.7, 0.9\}$ and the three fluxes (linear, convex, and nonconvex) defined in §4.1. For brevity, we focus on the initial condition IC1, the results for the initial conditions IC2 and IC3 being essentially similar.

We first discuss the results for $\epsilon = 10^{-3}$ and with $\Phi_{\epsilon,\gamma,\theta} := \gamma T \left(h/\beta T\right)^{\theta}$ instead of (13). The results are reported in Fig. 2. For comparison, all the panels in Fig. 2 also include the errors corresponding to the plain exponential integrator ($\Phi_{\epsilon,\gamma,\theta} = \epsilon$, red dashed curve labeled $\Phi = \epsilon$ in the legend). As before, vertical lines indicate the start of the resolved regime for each value of θ. The most striking observation is that, in most situations, there is an interval of mesh size in the under-resolved regime, say $[h_\flat, h_\sharp]$, where the error curves corresponding to various values of θ reach smaller errors than the curve corresponding to $\Phi_{\epsilon,\gamma,\theta} = \epsilon$. In particular, for those values of θ, the error has two different behaviors as h spans $[h_\flat, h_\sharp]$: There is first a super-convergence phase, then the error stagnates until the resolved regime is reached (as indicated by the vertical lines). As expected, the error levels off in the resolved

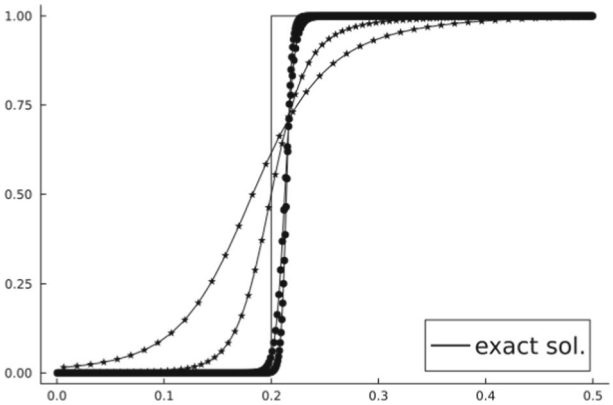


Fig. 3 1D linear transport with $\epsilon = 10^{-3}$ and IC1: solution profiles corresponding to the symbols (star, circle) shown in Fig. 2b

regime since using $\Phi_{\epsilon,\gamma,\theta} := \gamma T\big(h/\beta T\big)^{\theta}$, instead of $\Phi_{\epsilon,\gamma,\theta} := \epsilon$, is not consistent. We loosely refer to the above behavior as a resonance phenomenon. The resonance phenomenon is clearly visible for the three fluxes and $\alpha \in \{0.7, 0.9\}$. It is also visible for the nonlinear fluxes when $\alpha = 0.5$ (up to some oscillations of the reference solution corresponding to $\Phi_{\epsilon,\gamma,\theta} = \epsilon$).

To gain some insight into the resonance phenomenon, we report in Fig. 3 some solution profiles for the linear flux with $\alpha = 0.7$ and $\theta = 0.6$. We consider four mesh sizes identified by circle and star symbols in Fig. 2b. We observe that the super-convergent phase of the under-resolved regime (star symbols) corresponds to a swift reduction of the smearing of the discrete solution near the shock, whereas the stagnation phase (circle symbols) corresponds to the stabilization of the shock position at an incorrect location. As expected, the shock eventually moves to its correct location in the resolved regime (i.e., when $\Phi_{\epsilon,\gamma,\theta} = \epsilon$).

For each triple consisting of a flux, a value of α and an initial condition, we construct a list $\{(\theta_i, h_i)\}_{i\in\mathcal{L}}$ where for each index i in this list, θ_i is such that a resonance occurs in the under-resolved regime and the value h_i is the mesh size giving the smallest error. Plotting these points in a graph (not shown for brevity), we find that a good fit is obtained in the form $\theta_i \approx a + b(\log(h_i/\beta T))^{-1}$ (recall that, in all the cases, we have $\beta T = \frac{1}{2}$). In other words, multiplying by $\log(h_i/\beta T)$ and taking the exponential, the above fit implies that

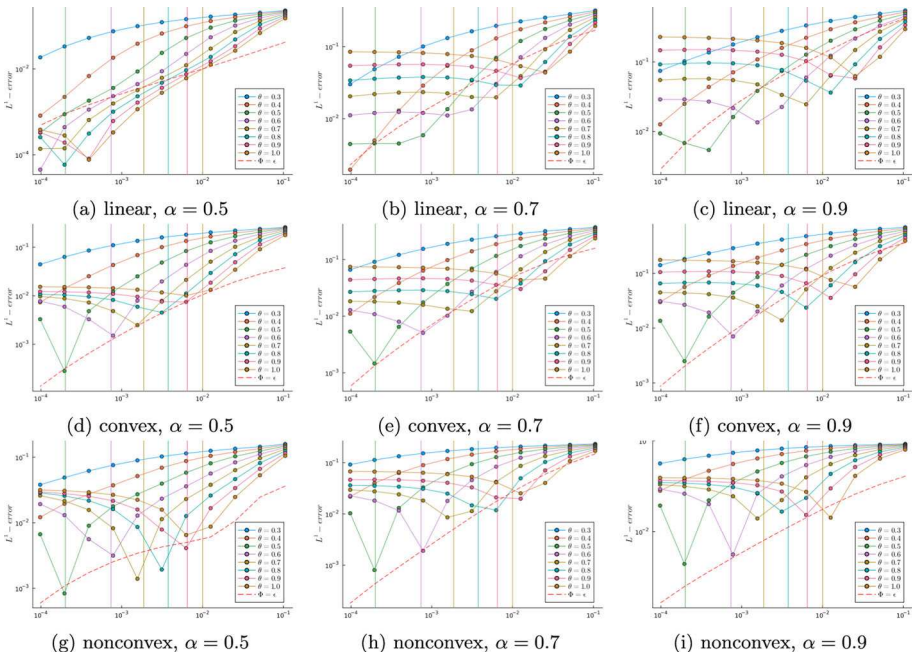
$$(h_i/\beta T)^{\theta_i} \approx e^b (h_i/\beta T)^a. \tag{32}$$

This, in turn, implies that the optimal expression of the cutoff function $\Phi_{\epsilon,\gamma,\theta}$ is indeed of the form $\max\left(\epsilon, \gamma_{\text{opt}} T\left(\frac{h}{\beta T}\right)^{\theta_{\text{opt}}}\right)$ as proposed in (13) with $\gamma_{\text{opt}} := e^b$ and $\theta_{\text{opt}} := a$. These optimal values are reported in Table 1 for the linear, convex, and nonconvex flux, respectively. Entries with a dash in the tables mean that optimal values were not found, i.e., resonance did not occur in the under-resolved regime. For $\epsilon = 10^{-3}$, this is only the case for the linear flux and $\alpha = 0.5$. In this case, we observe in Fig. 2a that the resonance phenomenon only occurs in the resolved regime.

The numerical experiments discussed above for $\epsilon = 10^{-3}$ are repeated for $\epsilon = 10^{-2}$. The results are reported in Fig. 4. The main observations regarding the presence of a resonance phenomenon and the possibility to devise optimal values for the cutoff parameters remain unchanged. The only relevant difference is that the value of the reaction parameter $\alpha = 0.5$

Table 1 Optimal cutoff parameters θ_{opt} and γ_{opt}

ϵ	α	0.5			0.7			0.9		
	IC	IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
Linear										
10^{-2}	θ	–	–	–	0.05	0.1	0.1	0.1	0.1	0.05
	γ	–	–	–	0.05	0.1	0.1	0.05	0.1	0.05
10^{-3}	θ	–	–	–	0.05	0.1	0.1	0.1	0.1	0.05
	γ	–	–	–	0.05	0.1	0.1	0.05	0.1	0.05
Burgers										
10^{-2}	θ	–	–	–	0.3	0.3	0.3	0.3	0.3	0.3
	γ	–	–	–	0.15	0.15	0.15	0.15	0.15	0.15
10^{-3}	θ	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
	γ	0.1	0.1	0.1	0.2	0.2	0.2	0.1	0.2	0.2
Sine										
10^{-2}	θ	–	–	–	0.15	0.15	0.15	0.2	0.15	0.15
	γ	–	–	–	0.05	0.05	0.05	0.05	0.05	0.05
10^{-3}	θ	0.4	0.4	0.4	0.25	0.25	0.25	0.3	0.4	0.4
	γ	0.03	0.03	0.03	0.05	0.05	0.05	0.03	0.1	0.1



(a) linear, $\alpha = 0.5$ (b) linear, $\alpha = 0.7$ (c) linear, $\alpha = 0.9$

(d) convex, $\alpha = 0.5$ (e) convex, $\alpha = 0.7$ (f) convex, $\alpha = 0.9$

(g) nonconvex, $\alpha = 0.5$ (h) nonconvex, $\alpha = 0.7$ (i) nonconvex, $\alpha = 0.9$

Fig. 4 L^1-errors as a function of the mesh size for $\epsilon = 10^{-2}$ and IC1, $\alpha \in \{0.5, 0.7, 0.9\}$ (from left to right), and linear, convex, and nonconvex fluxes (from top to bottom). The cutoff parameters are $\theta \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and $\gamma = 1$. The red dashed line labeled '$\Phi = \epsilon$' is obtained without any cutoff

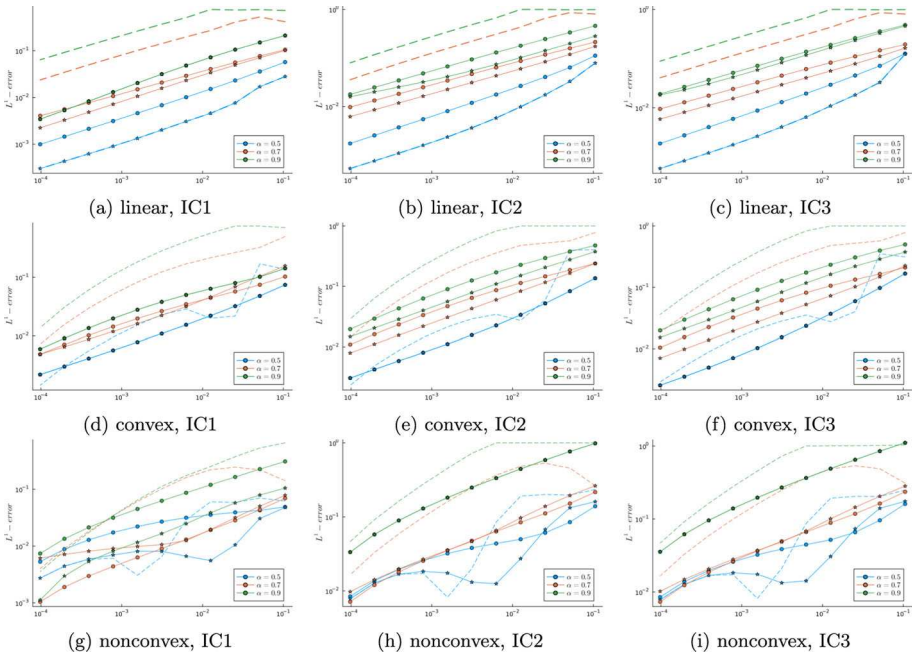


Fig. 5 L^1-errors as a function of the mesh size for $\epsilon = 10^{-3}$ and $\alpha \in \{0.5, 0.7, 0.9\}$. From top to bottom: linear, convex, and nonconvex fluxes. From left to right: IC1, IC2, and IC3. Comparison between optimal values (star symbols, $\star$) and all-purpose values (circle symbols, $\bullet$) of the cutoff parameters. The errors obtained with $\Phi_{\epsilon,\gamma,\theta} = \epsilon$ are reported using dashed lines with color matching the value of α (Color figure online)

now eludes the possibility of devising optimal cutoff parameters for the three fluxes (see the panels (a,d,g) in Fig. 4). The reason is that the resonance phenomenon is observed in the resolved regime for the linear flux and the nonconvex flux, and the resonance phenomenon fails to deliver lower errors than those obtained with $\Phi_{\epsilon,\gamma,\theta} = \epsilon$ for the convex flux. The optimal cutoff parameters for $\epsilon = 10^{-2}$ are again collected in Table 1 for the linear, convex, and nonconvex flux, respectively.

4.4 Selection of All-Purpose Cutoff Parameters

The next step in our investigation is to propose pairs of all-purpose cutoff parameters (θ, γ) that perform reasonably well uniformly over all the test cases. The inspection of Table 1 suggests to use

$$(\theta, \gamma) := \begin{cases} (0.1, 0.05) & \text{linear flux,} \\ (0.4, 0.1) & \text{nonlinear (convex and nonconvex) fluxes.} \end{cases} \tag{33}$$

Slightly different values can be chosen without significantly impacting the computational performance of the cutoff procedure. Figure 5 compares the errors obtained using the optimal cutoff parameters (star symbols, $\star$) with the errors obtained using the all-purpose cutoff parameters defined above (circle symbols, $\bullet$). For these tests, we have set $\epsilon = 10^{-3}$ and used $\alpha \in \{0.5, 0.7, 0.9\}$, the three fluxes, and the three initial conditions. The main observation is that using the all-purpose cutoff parameters instead of the optimal ones only leads to a

marginal deterioration of the errors. This fortunately indicates that despite the diversity of the behaviors observed when varying the flux, the initial conditions, and the reaction parameter, reasonable all-purpose values of the cutoff parameters can be found. The results for $\epsilon = 10^{-2}$ are similar to those displayed in Fig. 5 and are omitted for brevity.

In conclusion, the all-purpose values for the cutoff parameters indicated in (33) will be used in the rest of the paper.

5 Tests with All-Purpose Cutoff Parameters

In this section, we focus on the nonlinear fluxes (convex and nonconvex) and we test the proposed scheme with the all-purpose cutoff parameters identified in (33), i.e., we set $(\theta, \gamma) = (0.4, 0.1)$. We consider two 1D test cases and two 2D test cases.

5.1 High-Order Viscosity

To improve the performance of the scheme, we now make use of a high-order graph viscosity instead of the low-order one defined in (20). Following [13], [8, §82.2], we replace d_{ij}^n in the transport substep by the quantity $d_{ij}^{*,n} := d_{ij}^n \max(\psi(\alpha_i^n), \psi(\alpha_j^n))$, for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$, where α_i^n is a local, linearity preserving, smoothness indicator based on the discrete solution $(\mathsf{U}_i^n)_{i\in\mathcal{V}}$ (see, e.g., [8, Eq. (82.24)]) and $\psi : [0, 1] \to [0, 1]$ is any smooth increasing function satisfying $\psi(0) = 0$ and $\psi(1) = 1$ (we set $\psi(t) = t^2$ in our numerical experiments). When the flux is nonconvex, we actually set $d_{ij}^{*,n} := d_{ij}^n$ for any pair (i, j) such that an inflexion point of the flux lies between U_i^n and U_j^n. Moreover, we now set $C_{\mathrm{CFL}} = 0.1$ in (31).

5.2 1D Tests with Nonlinear Fluxes

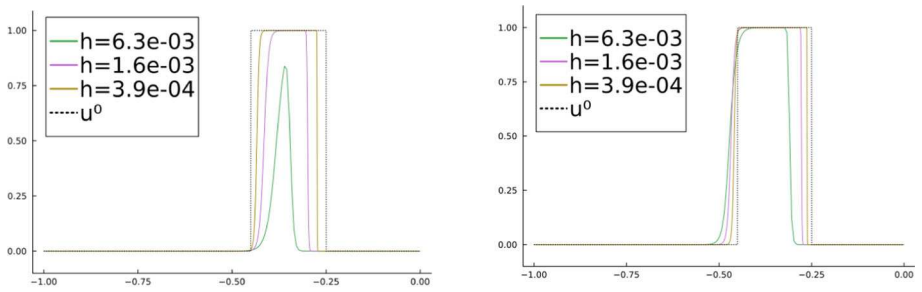
This section is devoted to 1D tests.

5.2.1 Convex (Burgers) Flux

We consider nonlinear transport with Burgers flux, $\epsilon = 10^{-3}$, and $\alpha = 0.9$. The initial condition is

$$u_0(x) = \begin{cases} 0, & \text{if } x \in (-1, -0.9), \\ 1, & \text{if } x \in (-0.9, -0.5), \\ 0.3, & \text{otherwise.} \end{cases} \tag{34}$$

The limit solution u^0 in the time interval $(0, 1)$ is composed of the two equilibrium states $\{0, 1\}$ separated by two moving shocks. The shock originating at $x = -0.9$ travels at speed $f'(\alpha) = 0.9$, and the shock originating at $x = -0.5$ travels at speed $\frac{f(1)-f(0)}{1-0} = \frac{1}{2}$. The two shocks meet at time $T = 1$, and the limit solution is identically zero thereafter. To illustrate the importance of using a high-order graph viscosity, we compare in Fig. 6 the solutions obtained using low-order and high-order viscosity at $T = 0.5$. Three solutions using the mesh sizes $h \in \{0.39, 1.6, 6.3\} \times 10^{-3}$ are shown (we are still in the under-resolved regime since $h_{\epsilon,\gamma,\theta} \approx 2.9\times10^{-5}$, see (14)). We observe that the higher-order solutions are significantly



(a) low-order graph viscosity (b) high-order graph viscosity

Fig. 6 1D Burgers with $\epsilon = 10^{-3}$, $\alpha = 0.9$ and IC (34). Solution profiles at $T = 0.5$ with mesh sizes $h \in \{0.39, 1.6, 6.3\} \times 10^{-3}$. Left: first-order graph viscosity; right: high-order graph viscosity

more accurate that the low-order ones. Using the high-order graph viscosity allows us to capture well the two shocks (recall that the speed of the first shock is reaction-dependent).

5.2.2 Nonconvex (sine) Flux

We now consider the nonconvex flux (sine) with $\epsilon = 10^{-3}$ and $\alpha = 0.7$. The initial condition is

$$u_0(x) = \begin{cases} 1, & \text{if } x \in (-\frac{1}{2}, -\frac{1}{5}) \cup (\frac{1}{5}, 1), \\ 0, & \text{otherwise.} \end{cases} \tag{35}$$

Figure 7 shows the profiles at times $T \in \{0.1, 0.2, 0.4, 1.0\}$ of the exact solutions of the nonreactive ($\epsilon = \infty$) and limit ($\epsilon = 0$) equations. We observe that u^∞ develops composite waves combining shocks and rarefaction waves; this is the expected behavior. The limit u^0 takes values in $\{0, 1\}$ with three shocks moving at the same speed, $s \approx -0.216$. This speed is the value of the derivative of both the upper concave and the lower convex envelopes of the flux f evaluated at α. Thus, the fact that a shock separating two states corresponding to a rarefaction wave moves at the speed $f'(\alpha)$ when the flux f is convex, as shown in Fan et al. [9], carries over to the present nonconvex case. This can be seen by adapting the arguments in [9], whereby f is replaced by its suitable envelope. Another interesting observation drawn from Fig. 7 is that the shocks in the nonreactive and reactive solutions move at the same speed at short times ($T \in \{0.1, 0.2\}$), but the speeds differ as soon as shocks originating from different locations start to interact in the nonreactive case.

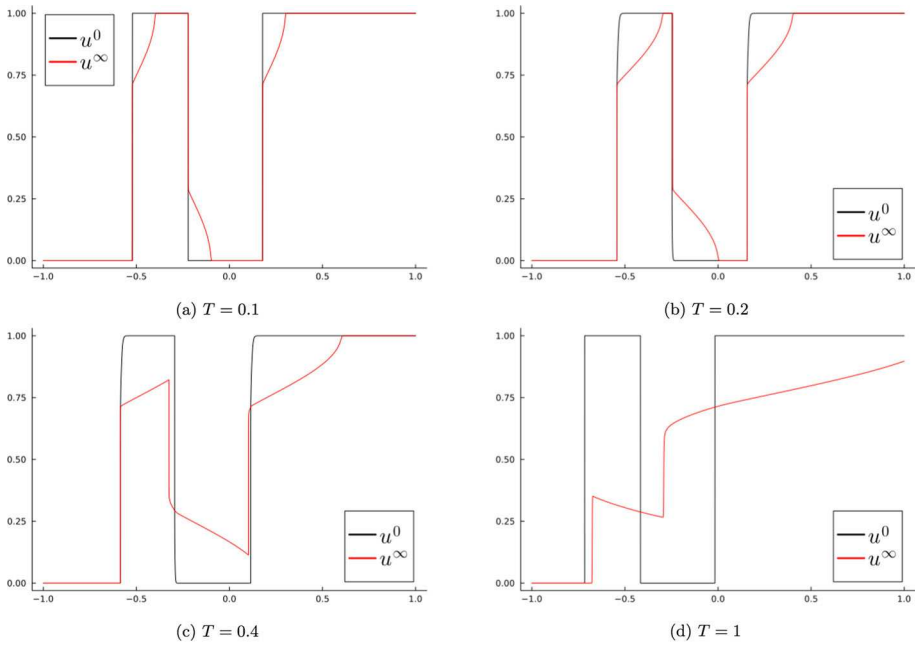
Figure 8 shows the profiles at $T \in \{0.1, 0.2, 0.4, 1.0\}$ of approximate solutions obtained with the mesh sizes $h \in \{0.39, 1.6, 6.3\} \times 10^{-3}$ and the high-order graph viscosity. The three shock positions are well captured on all meshes at the short times $T \in \{0.1, 0.2\}$, and that this is still the case for the longer times $T \in \{0.4, 1.0\}$ on the finer meshes.

We close this section with a more challenging situation where the source term leads to five equilibrium states:

$$R_{\frac{1}{4},\frac{3}{4}}(v) := 4^4 v(v - \tfrac{1}{4})(v - \tfrac{1}{2})(v - \tfrac{3}{4})(1 - v). \tag{36}$$

The states $\{0, \frac{1}{2}, 1\}$ are stable, whereas the states $\{\frac{1}{4}, \frac{3}{4}\}$ are unstable. We still consider the nonconvex flux $f(v) = \frac{1}{2\pi}\sin(2\pi v)$, but we now use the smooth initial condition

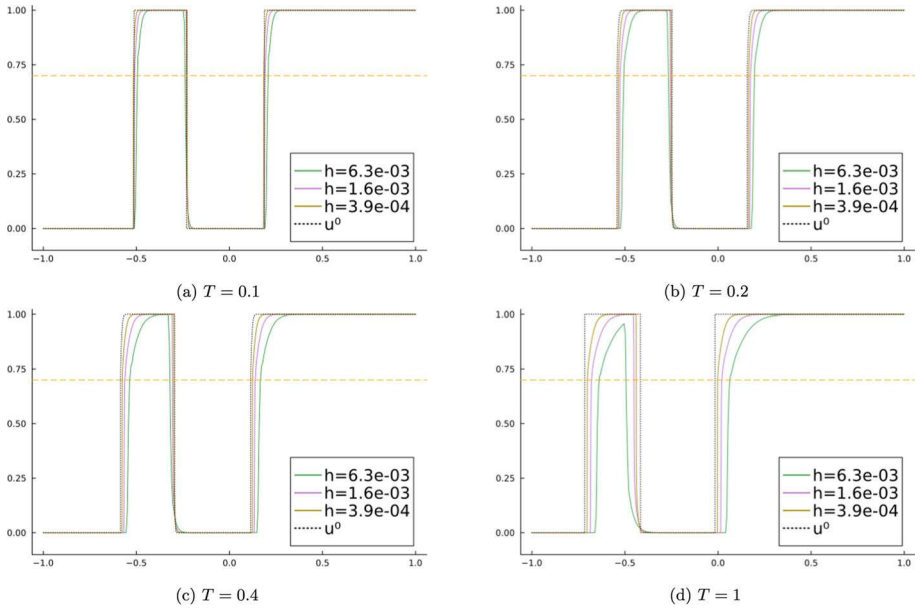
$$u_0(x) = 0.5\big(1 + \sin(\pi(x + 0.5))\big). \tag{37}$$



(a) $T = 0.1$ (b) $T = 0.2$

(c) $T = 0.4$ (d) $T = 1$

Fig. 7 1D test with nonconvex flux (sine), IC (35), and $\alpha = 0.7$. Comparison of nonreactive ($\epsilon = \infty$) and limit ($\epsilon = 0$) solutions at times $T \in \{0.1, 0.2, 0.4, 1.0\}$



(a) $T = 0.1$ (b) $T = 0.2$

(c) $T = 0.4$ (d) $T = 1$

Fig. 8 1D test with nonconvex flux (sine), IC (35), and $\alpha = 0.7$. Discrete solutions at times $T \in \{0.1, 0.2, 0.4, 1.0\}$ with mesh sizes $h \in \{0.39, 1.6, 6.3\} \times 10^{-3}$

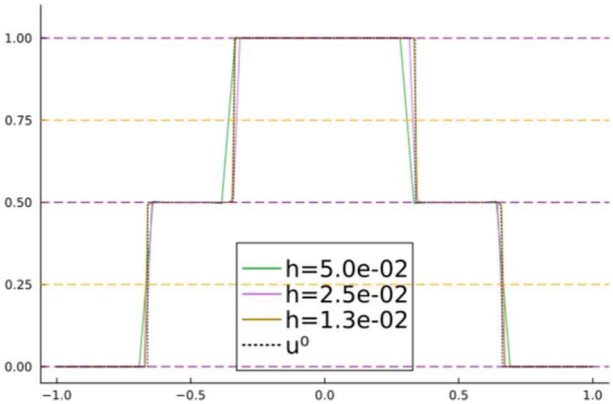


Fig. 9 1D nonconvex flux (sine), IC (37), and source term (36): Discrete solution profiles at $T = 0.1$ using the mesh sizes $h \in \{1.25, 2.5, 5.0\} \times 10^{-2}$; the horizontal dashed orange (resp., violet) line materializes the unstable (resp., stable) values of α (Color figure online)

We consider the short observation time $T := 0.1$ so as to allow the solution to take intermediate values in $[0, 1]$. The profiles of the limit solution u^0 and the approximate solutions using the mesh sizes $h \in \{1.25, 2.5, 5.0\} \times 10^{-2}$ are reported in Fig. 9. The limit solution takes values in the set $\{0, \frac{1}{2}, 1\}$ as expected. The four shocks are well captured by the discrete solutions, even using relatively coarse meshes, thereby giving again credence to the cutoff strategy proposed in the paper.

5.3 2D Numerical Tests

This section is devoted to 2D numerical tests. As, to our knowledge, there is no mathematical theory identifying the limit solution in this setting, these 2D results should be considered as illustrative of the capacity of the present scheme to capture reasonable solutions.

5.3.1 Convex (Burgers-Like) Flux

In this section, we consider the reactive Burgers equation in the square $D := (-0.25, 1.75)^2$ with the flux function $\boldsymbol{f}(v) := (\frac{1}{2}v^2, \frac{1}{2}v^2)$, and the following initial condition:

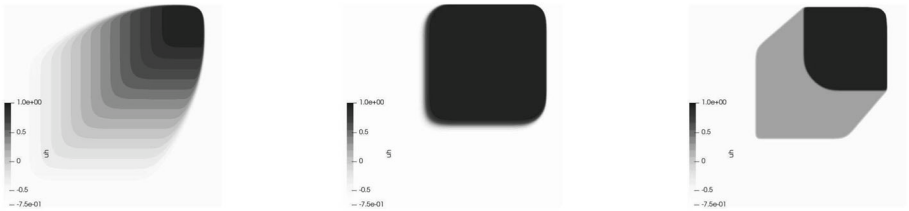
$$u_0(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \|\boldsymbol{x} - (1,1)^\mathsf{T}\|_\infty \le \frac{1}{2}, \\ -\frac{3}{4}, & \text{otherwise.} \end{cases} \tag{38}$$

This test, considered in [13, §6.1] in the nonreactive regime, is challenging since it exhibits sonic points. The nonreactive solution is given by Eqs. (52)–(53) therein.

The invariant set is now $\mathcal{B} = [a, b] := [-\frac{3}{4}, 1]$. Defining the linear map $\varphi : \mathcal{B} \to [0, 1]$ with $\varphi(v) := \frac{v-a}{b-a}$ (and $\varphi^{-1}(z) = (b-a)z + a$), we consider the reaction terms:

$$R_\alpha^\varphi(v) := R_\alpha(\varphi(v)), \qquad R_{\frac{1}{4},\frac{3}{4}}^\varphi(v) := R_{\frac{1}{4},\frac{3}{4}}(\varphi(v)), \qquad \forall v \in \mathcal{B}, \tag{39}$$

with R_α defined in (8) and $R_{\frac{1}{4},\frac{3}{4}}$ defined in (36). Thus, R_α^φ gives two stable equilibrium states $\{a, b\} = \{-\frac{3}{4}, 1\}$ and one unstable equilibrium state $\{\varphi^{-1}(\alpha)\}$. In what follows, we choose $\alpha := -\frac{a}{b-a} = \frac{3}{7}$, so that $\varphi^{-1}(\alpha) = 0$, which corresponds to the sonic point for Burgers



(a) nonreactive solution (b) two stable equilibria (c) three stable equilibria

Fig. 10 2D Burgers equation with IC (38). From left to right: **a** nonreactive solution; **b** reactive solution with two stable equilibria; **c** reactive solution with three stable equilibria

flux. The source $R^{\varphi}_{\frac{1}{4},\frac{3}{4}}$ gives three stable equilibrium states $\{-\frac{3}{4}, \frac{1}{8}, 1\}$ and two unstable equilibrium states $\{-\frac{5}{16}, \frac{7}{16}\}$. Finally, recalling that the solution to the ODE (16), $\vartheta(v; s)$, maps $[0, 1]$ to $[0, 1]$, we formulate the reaction substep by using the change of variable $[a, b] \ni v \mapsto z := \varphi(v) \in [0, 1]$. We obtain

$$\mathsf{U}_i^{n+1} = \varphi^{-1}\big(\vartheta(\varphi(\mathsf{W}_i^{n+1}); (b-a)\tau_{\epsilon,\gamma,\theta})\big), \qquad \forall i \in \mathcal{V}, \tag{40}$$

which gives, with $\mathsf{Z}_i^{n+1} := \varphi(\mathsf{W}_i^{n+1}) = \frac{\mathsf{W}_i^{n+1}-a}{b-a}$ and R^{φ} denoting either R^{φ}_{α} or $R^{\varphi}_{\frac{1}{4},\frac{3}{4}}$,

$$\mathsf{U}_i^{n+1} = a + (b-a)\frac{\mathsf{Z}_i^{n+1}\exp\big((b-a)\tau_{\epsilon,\gamma,\theta}R^{\varphi}(\mathsf{Z}_i^{n+1})\big)}{1+\mathsf{Z}_i^{n+1}\big(\exp\big((b-a)\tau_{\epsilon,\gamma,\theta}R^{\varphi}(\mathsf{Z}_i^{n+1})\big)-1\big)}, \qquad \forall i \in \mathcal{V}. \tag{41}$$

Figure 10 shows isocontours of the nonreactive solution (left panel), the reactive solution with R^{φ}_{α} (center panel), and the reactive solution with $R^{\varphi}_{\frac{1}{4},\frac{3}{4}}$ (right panel). The computations are done with $\epsilon = 10^{-3}$ on a fine mesh composed 400^2 grid points. The nonreactive solution matches well with the analytical solution given in [13]. In the reactive case with two stable equilibrium states, the shocks separating the two states are very well resolved. The shocks propagate differently than the nonreactive shocks. In the reactive case with three stable equilibrium states, the numerical solution takes the three values in the set $\{-\frac{3}{4}, \frac{1}{8}, 1\}$. Notice that the level set $\{u = 1\}$ is different when considering two or three stable equilibrium states.

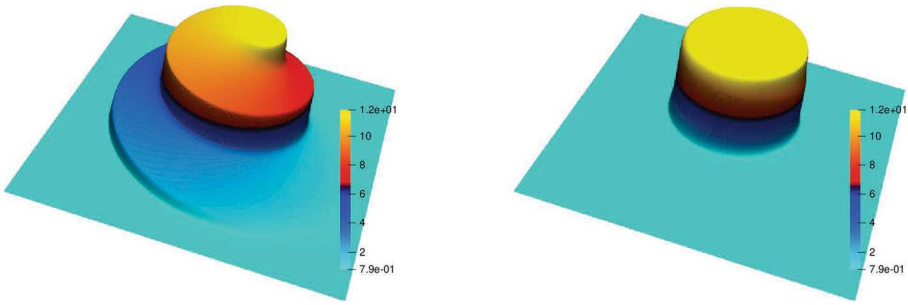
5.3.2 Nonconvex Flux

For the second 2D test case, we set $D := (-2, 2) \times (-2.5, 1.5)$, with the nonconvex flux $\boldsymbol{f}(v) := (\sin(v), \cos(v))$, and the initial condition

$$u_0(\boldsymbol{x}) = \begin{cases} \frac{15\pi}{4}, & \text{if } \|\boldsymbol{x} - (0.5, 0.5)^{\mathsf{T}}\|_2 \le 1, \\ \frac{\pi}{4}, & \text{otherwise.} \end{cases} \tag{42}$$

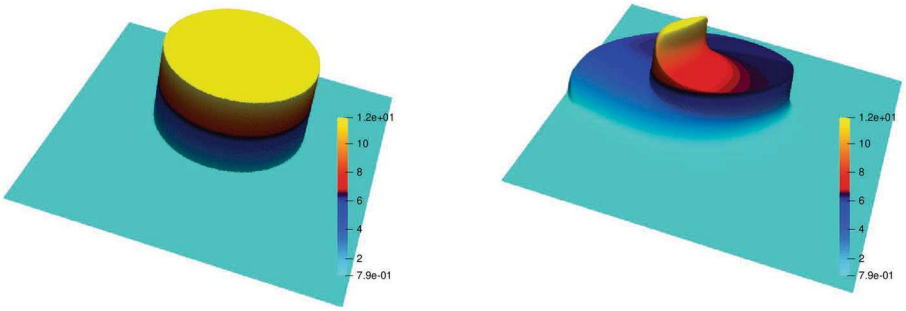
The nonreactive solution to this problem, proposed by Kurganov et al. [20], is a composite wave composed of a shock followed by a rarefaction wave. This solution is shown in Fig. 11a at $T = 1$.

The invariant domain associated with the initial condition is $\mathcal{B} = [a, b] := [\frac{\pi}{4}, \frac{15\pi}{4}]$. We proceed as in (39) to define the reaction terms R^{φ}_{α} and $R^{\varphi}_{\frac{1}{4},\frac{3}{4}}$ on $\mathcal{B}$. Here, we set $\alpha = \frac{10\pi}{4}$. Figure 11b shows isocontours of the reactive solution with two stable equilibrium states. The computation is done with $\epsilon = 10^{-3}$ on a fine mesh composed of 400^2 grid points. As above,



(a) nonreactive solution (b) two stable equilibria

Fig. 11 2D KPP test case with IC (42). Solution isocontours at $T = 1$. Left: nonreactive solution. Right: reactive solution with two stable equilibria, $\epsilon = 10^{-3}$



(a) $\epsilon = 10^{-3}$, $T = 1$ (b) $\epsilon = 10$, $T = 4$

Fig. 12 2D KPP test case with IC (42) and three stable equilibria. Left: solution isocontours for $\epsilon = 10^{-3}$ and $T = 1$. Right: solution isocontours for $\epsilon = 10$ and $T = 4$

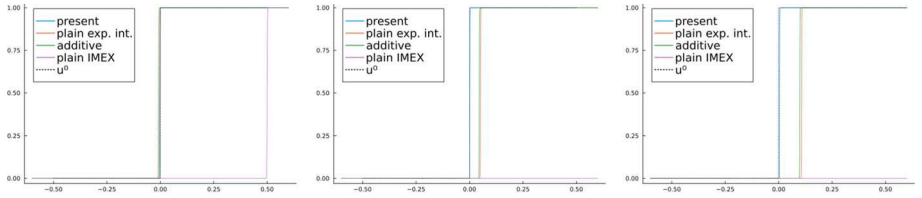
we observe a sharp resolution of the shocks separating the stable equilibrium states. Here again, the shocks in the reactive case propagate differently from the nonreactive case. Finally, Fig. 12 shows isocontours of the solution with the source term giving three stable equilibrium states. The left panel corresponds to $\epsilon = 10^{-3}$ and $T = 1$. The right panel corresponds to $\epsilon = 10$ and $T = 4$. In the first case, we observe that the intermediate equilibrium state is absent, whereas it can be observed in the second case where the stiffness parameter ϵ is much milder. This illustrates the complex interaction between shock dynamics and reaction terms.

6 Comparison to Other Schemes

We now compare the proposed algorithm with methods published in the literature. The tests are done in one space dimension with the three-state reaction term R_α defined in (8).

6.1 Other IMEX-Based Schemes

We first compare the present scheme to three IMEX-based methods from the literature. The reaction substep for the first method consists of using an exponential integrator, i.e.,



(a) $\alpha = 0.5$ (b) $\alpha = 0.7$ (c) $\alpha = 0.9$

Fig. 13 1D linear transport with IC (43) and $\epsilon = 10^{-3}$: solution profiles with mesh size $h = 3.9 \times 10^{-4}$. Left: $\alpha = 0.5$; center: $\alpha = 0.7$; right: $\alpha = 0.9$

setting $\Phi_{\epsilon,\gamma,\theta} = \epsilon$ in (17). The reaction substep for the second method consists of using the implicit Euler scheme with the nonlinear equation for each dof solved using Newton's method. Finally, the third scheme is inspired from the additive schemes proposed in Kennedy and Carpenter [18] for convection-diffusion equations. In the present setting, it amounts to writing the reaction substep as

$$\mathsf{U}_i^{n+1} = \mathsf{W}_i^{n+1} + \vartheta(\mathsf{U}_i^n; \epsilon^{-1}\tau) - \mathsf{U}_i^n, \qquad \forall i \in \mathcal{V}. \tag{43}$$

Notice that the exponential integrator is used in (43), but with the initial data U_i^n instead of W_i^{n+1} as in (17). Another significant difference is that there is no cutoff on the time step, i.e., $\frac{\tau}{\epsilon}$ is used instead of $\tau_{\epsilon,\gamma,\theta}$ as in (17). Notice in passing that neither the plain IMEX scheme nor the additive scheme are IDP.

We solve the 1D linear transport with the discontinuous initial data

$$u_0(x) = \begin{cases} 0, & \text{if } x < -\frac{1}{2}, \\ 1, & \text{otherwise.} \end{cases} \tag{44}$$

Figure 13 compares solution profiles at $T = 0.5$ obtained with the above schemes and the present one with $\epsilon = 10^{-3}$, $\alpha \in \{0.5, 0.7, 0.9\}$. The mesh size is $h = 3.9 \times 10^{-4}$. We observe that, for $\alpha \in \{0.7, 0.9\}$ (central and right panels), the three IMEX-based schemes fail to locate the correct shock position (the predicted shock location actually falls outside the figure for the plain IMEX scheme), whereas the present scheme correctly locates the shock. As already mentioned above, correctly capturing the shock location is less challenging when $\alpha = 0.5$ (left panel); in particular, the plain exponential integrator and the additive scheme now work well, whereas the plain IMEX scheme still fails.

6.2 Another AP Scheme

A scheme to approximate the model problem (1) has been proposed by Svärd and Mishra [23]. The method is tested therein for scalar conservation laws with convex fluxes and for the compressible the Euler equations coupled with a scalar conservation equation for a reactive species with a dissipative source term, i.e., 0 is the only equilibrium state.

To discuss the scheme (henceforth referred to with the letters SM), we focus on the reaction term defined in (8) and introduce the projector $\Pi_\alpha : \mathcal{B} \to \mathcal{B}$ such that $\Pi_\alpha(v) = 0$ if $v \in [0, \alpha)$, $\Pi_\alpha(\alpha) = \alpha$, and $\Pi_\alpha(v) = 1$ if $v \in (\alpha, 1]$. The key idea in [23] is to simultaneously approximate (1) and the nonreactive problem ($\epsilon = \infty$). An IMEX scheme is used to advance in time the reactive equation, and an explicit Euler scheme is used for the nonreactive equation. The IMEX scheme for the reactive equation gives a nonlinear problem

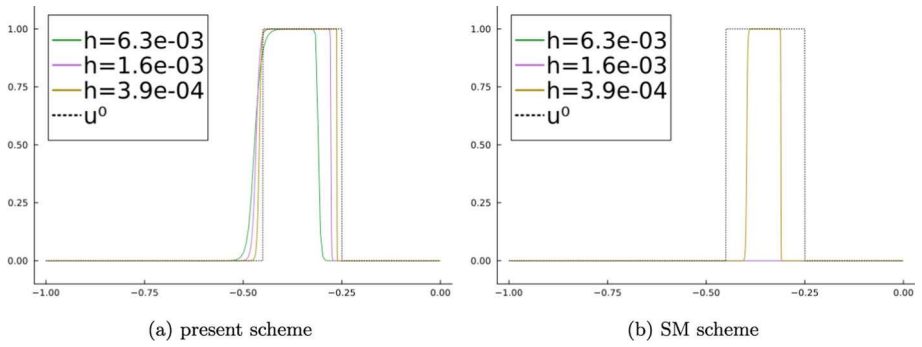


Fig. 14 1D Burgers with $\epsilon = 10^{-3}$, $\alpha = 0.9$ and IC (34). Solution at $T = 0.5$ with mesh sizes $h \in \{0.39, 1.6, 6.3\} \times 10^{-3}$. Left: present scheme; right: SM scheme

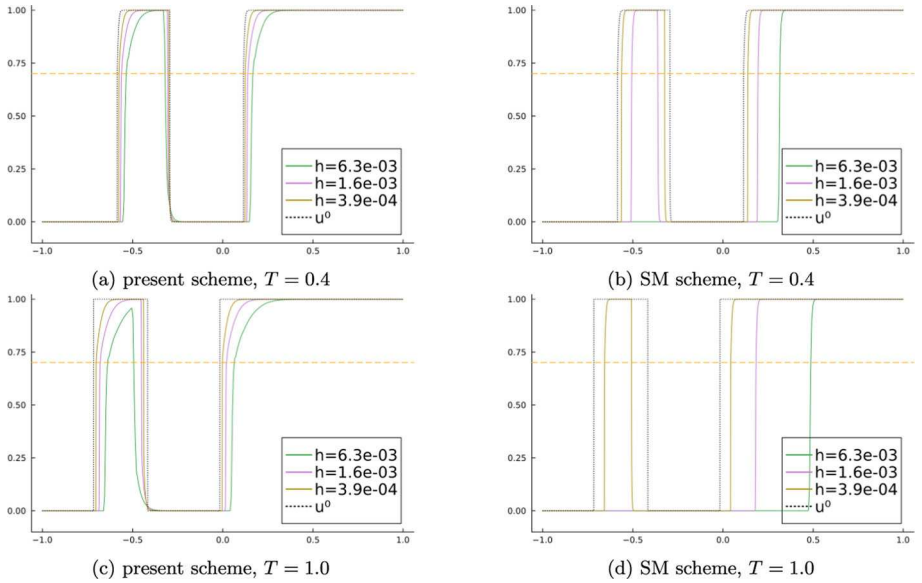


Fig. 15 1D sine flux with $\epsilon = 10^{-3}$, $\alpha = 0.7$ and IC (35). Solution profiles at $T = 0.4$ (top row) and $T = 1.0$ (bottom row) with mesh sizes $h \in \{0.39, 1.6, 6.3\} \times 10^{-3}$. Left column: present scheme; right column: SM scheme

at every dof, which is solved using a Newton method initialized using the image by Π_α of the dof corresponding to the nonreactive solution.

The SM scheme captures well the reactive solution in the under-resolved regime in various situations. This happens when the nonreactive solution has shocks or rarefactions, possibly many of them but not interacting, and in some cases, when the exact solution features composite waves. More precisely, the nonreactive solution, u^∞, is informative about the shocks appearing in the reactive solution, u^ϵ, as $\epsilon \to 0$, if the following holds true:

$$\Pi_\alpha(u^\infty) = u^0. \tag{45}$$

(This condition appears not to be explicitly identified in [23].) Although a bit unexpected at first sight, the condition (45) turns out to be satisfied in several situations. For instance, it is

the case for Burgers equation, as long as the shocks and rarefaction waves present in u^∞ do not interact. The SM scheme performs well in these situations, and actually better than the present scheme on coarse meshes since it does not have any cutoff. On the other hand, the SM scheme behaves poorly when u^∞ is composed of two shocks moving at different speeds, which eventually interact.

We now illustrate the above argumentation with the 1D Burgers equation with $\epsilon = 10^{-3}$, R_α with $\alpha = 0.9$, and the initial condition (34), as in §5.2.1. Figure 14 presents solution profiles at $T = 0.5$ on the three mesh sizes $h \in \{0.39, 1.6, 6.3\} \times 10^{-3}$. The solution obtained with the present scheme is shown in the left panel (this is the same as in Fig. 6b). The solution obtained with the SM scheme is shown in the right panel. The present scheme predicts well the shock locations on the considered mesh sizes (recall that we focus on the under-resolved regime). We also observe that the SM scheme captures (at least part of) the subset $\{u = 1\}$ only on the finest mesh.

We further illustrate the method by considering the 1D nonconvex flux (sine), with $\epsilon = 10^{-3}$, $\alpha = 0.7$, and the initial condition (35), as in §5.2.2. Figure 15 shows solution profiles at $T = 0.4$ (top row) and $T = 1.0$ (bottom row) using the same mesh sizes as above. The solution obtained with the present scheme are shown in the left column. The solutions obtained with the SM scheme are shown in the right column. We draw the same conclusions as above. The present scheme achieves a much sharper prediction of the shock locations than the SM scheme in the under-resolved regime, and the SM scheme meets with some difficulties in capturing the first connected component of the subset $\{u = 1\}$ on the coarser meshes.

7 Proof of Theorem 3.3

Recall that we want to prove that, for any test function $\psi \in W_0^{1,\infty}(D \times [0, T]; \mathbb{R}_+)$, we have

$$\int_D \mathcal{I}_h\big(\eta(u_h^N)\big)(\boldsymbol{x})\psi(\boldsymbol{x}, T)\mathrm{d}\boldsymbol{x} - \int_D \mathcal{I}_h\big(\eta(u_h^0)\big)(\boldsymbol{x})\psi(\boldsymbol{x}, 0)\mathrm{d}\boldsymbol{x} \tag{46}$$
$$- \int_Q \Big\{\eta(u_h^\epsilon)\partial_t\psi + \boldsymbol{q}(u_h^\epsilon)\cdot\nabla\psi + \frac{1}{\Phi_{\epsilon,\gamma,\theta}}\eta'(u_h^\epsilon)R(u_h^\epsilon)\psi\Big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t \le C\Lambda(h),$$

where

$$\Lambda(h) := \frac{h}{\beta\Phi_{\epsilon,\gamma,\theta}^2}\|u_h^\epsilon\|_{L^1(Q)} + \left(\frac{h}{\Phi_{\epsilon,\gamma,\theta}} + \frac{h^2}{\beta\Phi_{\epsilon,\gamma,\theta}^2}\right)\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)}.$$

The symbol C denotes a generic positive real number whose value can change at each occurrence as long as it is independent of h, τ, and ϵ. The value of C can, in particular, depend on the norm $\|\psi\|_{L^\infty(Q)} + T\|\partial_t\psi\|_{L^\infty(Q)} + \beta T\|\nabla\psi\|_{\boldsymbol{L}^\infty(Q)}$, as well as bounds over $\mathcal{B}$ on the functions R, η, $\beta^{-1}\boldsymbol{q}$, and their derivatives. Notice also that the CFL condition (11) on the time step with τ^* defined in (22) amounts to

$$\tau \lesssim \beta^{-1}h. \tag{47}$$

Proof (1) It is well-known (see, e.g., [8, Theorem 81.12]) that the update W_i^{n+1} from the transport substep satisfies the following discrete entropy inequality: For all $n \in \mathcal{N}$ and all $i \in \mathcal{V}$,

$$m_i\frac{\eta(\mathsf{W}_i^{n+1}) - \eta(\mathsf{U}_i^n)}{\tau} - \int_D \mathcal{I}_h\big(\boldsymbol{q}(u_h^n)\big)\cdot\nabla\varphi_i\mathrm{d}\boldsymbol{x} - \sum_{j\in\mathcal{I}(i)} d_{ij}^n\big(\eta(\mathsf{U}_j^n) - \eta(\mathsf{U}_i^n)\big) \le 0.$$

Moreover, we can rewrite the reaction substep as

$$\mathsf{U}_i^{n+1} = \mathsf{W}_i^{n+1} + \tau_{\epsilon,\gamma,\theta} R_{\epsilon,\gamma,\theta}(\mathsf{W}_i^{n+1}), \qquad R_{\epsilon,\gamma,\theta}(v) := \frac{1}{\tau_{\epsilon,\gamma,\theta}}\big(\vartheta(v;\tau_{\epsilon,\gamma,\theta}) - v\big)\ \forall v \in \mathcal{B}. \tag{48}$$

Recalling that $\tau_{\epsilon,\gamma,\theta} := \frac{\tau}{\Phi_{\epsilon,\gamma,\theta}}$, the convexity of η then implies that

$$\begin{aligned}\eta(\mathsf{W}_i^{n+1}) &\geq \eta(\mathsf{U}_i^{n+1}) + \eta'(\mathsf{U}_i^{n+1})(\mathsf{W}_i^{n+1} - \mathsf{U}_i^{n+1}) \\ &= \eta(\mathsf{U}_i^{n+1}) - \frac{\tau}{\Phi_{\epsilon,\gamma,\theta}}\eta'(\mathsf{U}_i^{n+1}) R_{\epsilon,\gamma,\theta}(\mathsf{W}_i^{n+1}).\end{aligned}$$

We infer that

$$\begin{aligned}&m_i \frac{\eta(\mathsf{U}_i^{n+1}) - \eta(\mathsf{U}_i^n)}{\tau} - \int_D \mathcal{I}_h\big(\boldsymbol{q}(u_h^n)\big)\cdot\nabla\varphi_i \mathrm{d}\boldsymbol{x} \\ &\quad - \sum_{j\in\mathcal{I}(i)} d_{ij}^n\big(\eta(\mathsf{U}_j^n) - \eta(\mathsf{U}_i^n)\big) - \frac{m_i}{\Phi_{\epsilon,\gamma,\theta}}\eta'(\mathsf{U}_i^{n+1}) R_{\epsilon,\gamma,\theta}(\mathsf{W}_i^{n+1}) \leq 0.\end{aligned} \tag{49}$$

(2) Let us set $\psi_h^n(\boldsymbol{x}) := \sum_{i\in\mathcal{V}} \Psi_i^n \varphi_i(\boldsymbol{x})$ with $\Psi_i^n := \frac{1}{m_i}\int_D \psi^n(\boldsymbol{x})\varphi_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ and $\psi^n(\boldsymbol{x}) := \psi(\boldsymbol{x}, t^n)$ for all $\boldsymbol{x} \in \overline{D}$. Multiplying the inequality (49) by $\tau\Psi_i^n \geq 0$ and summing over $n \in \mathcal{N}$ and $i \in \mathcal{V}$, we infer that

$$E_{1,h} + E_{2,h} + E_{3,h} + E_{4,h} \leq 0, \tag{50}$$

where

$$\begin{aligned}E_{1,h} &:= \sum_{n\in\mathcal{N}} \int_D \Big\{\big(\mathcal{I}_h\big(\eta(u_h^{n+1})\big) - \mathcal{I}_h\big(\eta(u_h^n)\big)\big)\psi^n\Big\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \\ E_{2,h} &:= -\sum_{n\in\mathcal{N}} \tau \int_D \Big\{\mathcal{I}_h\big(\boldsymbol{q}(u_h^n)\big)\cdot\nabla\psi_h^n\Big\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \\ E_{3,h} &:= \sum_{n\in\mathcal{N}} \tau \sum_{i\in\mathcal{V}} \sum_{j\in\mathcal{I}(i)} \frac{1}{2} d_{ij}^n(\eta(\mathsf{U}_i^n) - \eta(\mathsf{U}_j^n))(\Psi_i^n - \Psi_j^n), \\ E_{4,h} &:= -\sum_{n\in\mathcal{N}} \tau \int_D \frac{1}{\Phi_{\epsilon,\gamma,\theta}} \Big\{\mathcal{I}_h\big(\eta'(u_h^{n+1}) R_{\epsilon,\gamma,\theta}(w_h^{n+1})\big)\psi^n\Big\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\end{aligned}$$

where we used the symmetry of d_{ij}^n to re-arrange the expression of $E_{3,h}$ and where we have set $w_h^{n+1}(\boldsymbol{x}) := \sum_{i\in\mathcal{V}} \mathsf{W}_i^{n+1}\varphi(\boldsymbol{x})$. Moreover, denoting LHS the left-hand side of (46), we have

$$\mathrm{LHS} = \overline{E}_{1,h} + \overline{E}_{2,h} + \overline{E}_{4,h},$$

with

$$\begin{aligned}
\overline{E}_{1,h} &:= \int_D \{\mathcal{I}_h\big(\eta(u_h^N)\big)\psi^N\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_D \{\mathcal{I}_h\big(\eta(u_h^0)\big)\psi^0\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad - \int_Q \{\eta(u_h^\epsilon)\partial_t\psi\}\mathrm{d}\boldsymbol{x}\mathrm{d}t, \\
\overline{E}_{2,h} &:= -\int_Q \{\boldsymbol{q}(u_h^\epsilon)\cdot\nabla\psi\}\mathrm{d}\boldsymbol{x}\mathrm{d}t, \\
\overline{E}_{4,h} &:= -\int_Q \frac{1}{\Phi_{\epsilon,\gamma,\theta}}\{\eta'(u_h^\epsilon)R(u_h^\epsilon)\psi\}\mathrm{d}\boldsymbol{x}\mathrm{d}t.
\end{aligned}$$

Since $E_{1,h} + E_{2,h} + E_{3,h} + E_{4,h} \le 0$, we have

$$\begin{aligned}
\mathrm{LHS} &= \overline{E}_{1,h} + \overline{E}_{2,h} + \overline{E}_{4,h} \\
&= E_{1,h} + E_{2,h} + E_{3,h} + E_{4,h} \\
&\quad + (\overline{E}_{1,h} - E_{1,h}) + (\overline{E}_{2,h} - E_{2,h}) - E_{3,h} + (\overline{E}_{4,h} - E_{4,h}) \\
&\le |\overline{E}_{1,h} - E_{1,h}| + |\overline{E}_{2,h} - E_{2,h}| + |E_{3,h}| + |\overline{E}_{4,h} - E_{4,h}|.
\end{aligned}$$

Hence, to prove that LHS $\le C\Lambda(h)$, it suffices to establish that $|E_{3,h}| \le C\Lambda(h)$ and $|E_{k,h} - \overline{E}_{k,h}| \le C\Lambda(h)$ for all $k \in \{1, 2, 4\}$.

(3) Bound on $|\overline{E}_{1,h} - E_{1,h}|$. Let us set

$$\bar{u}_h^\epsilon(\boldsymbol{x}, t)|_{(t^n, t^{n+1}]} := u_h^{n+1}(\boldsymbol{x}),\ \forall n \in \mathcal{N}, \qquad \bar{u}_h^\epsilon(\boldsymbol{x}, 0) := u_h^0(\boldsymbol{x}).$$

Since $\bar{u}_h^\epsilon$ is piecewise constant in time, we infer that

$$\begin{aligned}
\int_Q \{\mathcal{I}_h\big(\eta(\bar{u}_h^\epsilon)\big)\partial_t\psi\}\mathrm{d}\boldsymbol{x}\mathrm{d}t &= \sum_{n\in\mathcal{N}}\int_D \{\mathcal{I}_h\big(\eta(u_h^{n+1})\big)(\psi^{n+1} - \psi^n)\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \int_D \{\mathcal{I}_h\big(\eta(u_h^N)\big)\psi^N\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \int_D \{\mathcal{I}_h\big(\eta(u_h^0)\big)\psi^0\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&\quad - \sum_{n\in\mathcal{N}}\int_D \{\big(\mathcal{I}_h\big(\eta(u_h^{n+1})\big) - \mathcal{I}_h\big(\eta(u_h^n)\big)\big)\psi^n\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.
\end{aligned}$$

This gives

$$\overline{E}_{1,h} - E_{1,h} = \int_Q \Big\{\big(\mathcal{I}_h\big(\eta(\bar{u}_h^\epsilon) - \eta(u_h^\epsilon)\big)\big)\partial_t\psi\Big\} + \Big\{\big(\mathcal{I}_h\big(\eta(u_h^\epsilon)\big) - \eta(u_h^\epsilon)\big)\partial_t\psi\Big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

We bound the right-hand side using the triangle inequality. The second term on the right-hand side is bounded by $C\Lambda(h)$ using the approximation properties of $\mathcal{I}_h$ and since η is of class C^2. Invoking the L^1-stability of $\mathcal{I}_h$ and the smoothness of η, the first term is bounded as

$$\Big|\int_Q \Big\{\big(\mathcal{I}_h\big(\eta(\bar{u}_h^\epsilon) - \eta(u_h^\epsilon)\big)\big)\partial_t\psi\Big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t\Big| \le CT^{-1}\sum_{n\in\mathcal{N}}\tau\sum_{i\in\mathcal{V}} m_i|\mathsf{U}_i^{n+1} - \mathsf{U}_i^n|.$$

Using $m_i \sim h^d$, $\|\boldsymbol{c}_{ij}\|_{\ell^2} \sim h^{d-1}$, $\lambda_{\max}(\mathsf{U}_i^n, \mathsf{U}_j^n, \boldsymbol{n}_{ij}) \le \beta$ together with the CFL restriction (47), we obtain

$$\sum_{i\in\mathcal{V}} m_i|\mathsf{W}_i^{n+1} - \mathsf{U}_i^n| \le Ch\|\nabla u_h^n\|_{\boldsymbol{L}^1(D)}. \tag{51}$$

Recalling (48), observing that $|R_{\epsilon,\gamma,\theta}(v)| \le Cv$ for all $v \in \mathcal{B}$, and using the triangle inequality together with (51), the reaction substep gives

$$\begin{aligned}\sum_{i\in\mathcal{V}} m_i|\mathsf{U}_i^{n+1} - \mathsf{W}_i^{n+1}| &\le C\tau_{\epsilon,\gamma,\theta}\|w_h^{n+1}\|_{L^1(D)}\\ &\le C'(\beta\Phi_{\epsilon,\gamma,\theta})^{-1}h\big(\|u_h^n\|_{L^1(D)} + h\|\nabla u_h^n\|_{\boldsymbol{L}^1(D)}\big),\end{aligned} \tag{52}$$

where we used the definition (15) of $\tau_{\epsilon,\gamma,\theta}$ and the CFL restriction (47). In conclusion, we have

$$\begin{aligned}&T^{-1}\sum_{n\in\mathcal{N}}\tau\sum_{i\in\mathcal{V}} m_i|\mathsf{U}_i^{n+1} - \mathsf{U}_i^n|\\ &\quad\le C\big(T^{-1}h\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)} + (\beta T\Phi_{\epsilon,\gamma,\theta})^{-1}h\big(\|u_h^\epsilon\|_{L^1(Q)} + h\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)}\big)\big) \le C'\Lambda(h),\end{aligned}$$

where the last bound follows from $\Phi_{\epsilon,\gamma,\theta} \le T$. Putting everything together, we obtain

$$|\overline{E}_{1,h} - E_{1,h}| \le C\Lambda(h).$$

(4) Bound on $|\overline{E}_{2,h} - E_{2,h}|$. We have $\overline{E}_{2,h} - E_{2,h} = A_{2,1} + A_{2,2} + A_{3,3}$ with

$$\begin{aligned}A_{2,1} &= \sum_{n\in\mathcal{N}}\tau\int_D\Big\{\big(\mathcal{I}_h\big(\boldsymbol{q}(u_h^n)\big) - \boldsymbol{q}(u_h^n)\big)\cdot\nabla\psi_h^n\Big\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\ A_{2,2} &= -\sum_{n\in\mathcal{N}}\tau\int_D\Big\{\nabla\cdot\boldsymbol{q}(u_h^n)(\psi_h^n - \psi^n)\Big\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\\ A_{2,3} &= -\sum_{n\in\mathcal{N}}\tau\int_D\Big\{\nabla\cdot\boldsymbol{q}(u_h^n)(\psi^n - \psi)\Big\}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Using the approximation properties of the $\mathcal{I}_h$, the smoothness of ψ, and the CFL restriction (47), we infer that

$$|\overline{E}_{2,h} - E_{2,h}| \le CT^{-1}(h+\beta\tau)\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)} \le C'T^{-1}h\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)}.$$

(5) Bound on $|E_{3,h}|$. We have $|\eta(\mathsf{U}_i^n) - \eta(\mathsf{U}_j^n)| \le C|\mathsf{U}_i^n - \mathsf{U}_j^n|$ and $|\Psi_i^n - \Psi_j^n| \le C(\beta T)^{-1}h$ for all $i \in \mathcal{V}$ and all $j \in \mathcal{I}(i)$. Since $d_{ij}^n \le C\beta h^{d-1}$, we infer that

$$\begin{aligned}|E_{3,h}| &\le \sum_{n\in\mathcal{N}}\tau\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)} d_{ij}^n|\mathsf{U}_i^n - \mathsf{U}_j^n||\Psi_i^n - \Psi_j^n|\\ &\le C\sum_{n\in\mathcal{N}}\tau\sum_{i\in\mathcal{V}}\sum_{j\in\mathcal{I}(i)} T^{-1}h^d|\mathsf{U}_i^n - \mathsf{U}_j^n| \le C'T^{-1}h\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)},\end{aligned}$$

where we used the shape-regularity of the mesh sequence in the last bound.

(6) Bound on $|\overline{E}_{4,h} - E_{4,h}|$. We observe that $\overline{E}_{4,h} - E_{4,h} = -\sum_{k\in\{1:5\}} A_{4,k}$ with

$$A_{4,1} := \sum_{n\in\mathcal{N}} \int_{I^n} \frac{1}{\Phi_{\epsilon,\gamma,\theta}} \int_D \big\{\eta'(u_h^n(\boldsymbol{x}))R(u_h^n(\boldsymbol{x}))(\psi(\boldsymbol{x}) - \psi^n(\boldsymbol{x},t))\big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

$$A_{4,2} := \sum_{n\in\mathcal{N}} \tau \frac{1}{\Phi_{\epsilon,\gamma,\theta}} \int_D \big\{\big(\eta'(u_h^n)R(u_h^n) - \mathcal{I}_h\big(\eta'(u_h^n)R(u_h^n)\big)\big)\psi^n\big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

$$A_{4,3} := \sum_{n\in\mathcal{N}} \tau \frac{1}{\Phi_{\epsilon,\gamma,\theta}} \int_D \big\{\mathcal{I}_h\big(\big(\eta'(u_h^n) - \eta'(u_h^{n+1})\big)R(u_h^n)\big)\psi^n\big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

$$A_{4,4} := \sum_{n\in\mathcal{N}} \tau \frac{1}{\Phi_{\epsilon,\gamma,\theta}} \int_D \big\{\mathcal{I}_h\big(\eta'(u_h^{n+1})\big(R(u_h^n) - R(w_h^{n+1})\big)\big)\psi^n\big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t,$$

$$A_{4,5} := \sum_{n\in\mathcal{N}} \tau \frac{1}{\Phi_{\epsilon,\gamma,\theta}} \int_D \big\{\mathcal{I}_h(\eta'(u_h^{n+1})\big(R(w_h^{n+1}) - R_{\epsilon,\gamma,\theta}(w_h^{n+1})\big))\psi^n\big\}\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Using the smoothness of Ψ in time and the CFL restriction (47) gives

$$|A_{4,1}| \le C\Phi_{\epsilon,\gamma,\theta}^{-1}\tau T^{-1}\|u_h^\epsilon\|_{L^1(Q)} \le C'\Phi_{\epsilon,\gamma,\theta}^{-1}h(\beta T)^{-1}\|u_h^\epsilon\|_{L^1(Q)}.$$

Using the approximation properties of $\mathcal{I}_h$ we obtain

$$|A_{4,2}| \le C\Phi_{\epsilon,\gamma,\theta}^{-1}h\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)}.$$

The shape-regularity of the mesh sequence and the triangle inequality yield

$$|A_{4,3}| + |A_{4,4}| \le C\sum_{n\in\mathcal{N}} \tau \frac{1}{\Phi_{\epsilon,\gamma,\theta}} \sum_{i\in\mathcal{V}} m_i\big(|\mathsf{W}_i^{n+1} - \mathsf{U}_i^n| + |\mathsf{U}_i^{n+1} - \mathsf{W}_i^{n+1}|\big).$$

Recalling inequalities (51) and (52), and invoking the CFL restriction (47), this gives

$$|A_{4,3}| + |A_{4,4}| \le C\Phi_{\epsilon,\gamma,\theta}^{-1}\big(\Phi_{\epsilon,\gamma,\theta}^{-1}\beta^{-1}h\|u_h^n\|_{L^1(Q)} + h(1+\Phi_{\epsilon,\gamma,\theta}^{-1}\beta^{-1}h)\|\nabla u_h^n\|_{\boldsymbol{L}^1(Q)}\big).$$

Setting $\zeta(v) := v(1-v)$, we have, for all $v \in \mathcal{B}$,

$$R(v) - R_{\epsilon,\gamma,\theta}(v) = \frac{1}{\tau_{\epsilon,\gamma,\theta}} \int_0^{\tau_{\epsilon,\gamma,\theta}} \big(\zeta(\vartheta(v;s)) - \zeta(v)\big)\tilde{R}(v)\mathrm{d}s.$$

Since ζ is Lipschitz in $\mathcal{B}$, we infer that $\big|R(v) - R_{\epsilon,\gamma,\theta}(v)\big| \le C\tau|v|$. As a result, invoking again the CFL restriction (47), we obtain

$$|A_{4,5}| \le C\Phi_{\epsilon,\gamma,\theta}^{-2}\beta^{-1}h\big(\|u_h^n\|_{L^1(Q)} + h\|\nabla u_h^n\|_{\boldsymbol{L}^1(Q)}\big).$$

Putting everything together and since $\Phi_{\epsilon,\gamma,\theta} \le T$, we infer that

$$|\overline{E}_{4,h} - E_{4,h}| \le C\Big(\tfrac{h}{\beta\Phi_{\epsilon,\gamma,\theta}^2}\|u_h^\epsilon\|_{L^1(Q)} + \big(\tfrac{h}{\Phi_{\epsilon,\gamma,\theta}} + \tfrac{h^2}{\beta\Phi_{\epsilon,\gamma,\theta}^2}\big)\|\nabla u_h^\epsilon\|_{\boldsymbol{L}^1(Q)}\Big).$$

(7) Combining the bounds established in Steps (3)–(6) and using $\Phi_{\epsilon,\gamma,\theta} \le T$ to simplify the upper bound completes the proof. □

Funding This material is based upon work supported in part by the National Science Foundation grant DMS2110868, the Air Force Office of Scientific Research, USAF, under grant/contract number FA9550-18-1-0397, the Army Research Office, under grant number W911NF-19-1-0431, and the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contracts B640889. The support of INRIA through the International Chair program is acknowledged.

Data Availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors do not have any conflicting interest to declare.

References

1. Bao, W., Jin, S.: The random projection method for hyperbolic conservation laws with stiff reaction terms. J. Comput. Phys. **163**(1), 216–248 (2000)
2. Bao, W., Jin, S.: Error estimates on the random projection methods for hyperbolic conservation laws with stiff reaction terms. Appl. Numer. Math. **43**(4), 315–333 (2002)
3. Bulteau, S., Berthon, C., Bessemoulin-Chatard, M.: Convergence rate of an asymptotic preserving scheme for the diffusive limit of the p-system with damping. Commun. Math. Sci. **17**(6), 1459–1486 (2019)
4. Chainais-Hillairet, C., Champier, S.: Finite volume schemes for nonhomogeneous scalar conservation laws: error estimate. Numer. Math. **88**(4), 607–639 (2001)
5. Chalabi, A.: On convergence of numerical schemes for hyperbolic conservation laws with stiff source terms. Math. Comp. **66**(218), 527–545 (1997)
6. Colella, P., Majda, A., Roytburd, V.: Theoretical and numerical structure for reacting shock waves. SIAM J. Sci. Statist. Comput. **7**(4), 1059–1080 (1986)
7. Engquist, B., Sjögreen, B.: The convergence rate of finite difference schemes in the presence of shocks. SIAM J. Numer. Anal. **35**(6), 2464–2485 (1998)
8. Ern, A., Guermond, J.-L.: Finite elements III. First-order and time-dependent PDEs, volume 74 of Texts in Applied Mathematics. Springer, Cham (2021)
9. Fan, H., Jin, S., Teng, Z.-H.: Zero reaction limit for hyperbolic conservation laws with source terms. J. Differ. Equ. **168**(2), 270–294 (2000)
10. Filbet, F., Rambaud, A.: Analysis of an asymptotic preserving scheme for relaxation systems. ESAIM Math. Model. Numer. Anal. **47**(2), 609–633 (2013)
11. Guermond, J.-L., Nazarov, M.: A maximum-principle preserving C^0 finite element method for scalar conservation equations. Comput. Methods Appl. Mech. Engrg. **272**, 198–213 (2014)
12. Guermond, J.-L., Popov, B.: Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J. Numer. Anal. **54**(4), 2466–2489 (2016)
13. Guermond, J.-L., Popov, B.: Invariant domains and second-order continuous finite element approximation for scalar conservation equations. SIAM J. Numer. Anal. **55**(6), 3120–3146 (2017)
14. Hochbruck, M., Ostermann, A.: Exponential integrators. Acta Numer. **19**, 209–286 (2010)
15. Hu, J., Shu, R.: A second-order asymptotic-preserving and positivity-preserving exponential Runge-Kutta method for a class of stiff kinetic equations. Multiscale Model. Simul. **17**(4), 1123–1146 (2019)
16. Jin, S.: Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review. Riv. Math. Univ. Parma (N.S.) **3**(2), 177–216 (2012)
17. Jin, S., Levermore, C.D.: Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. J. Comput. Phys. **126**(2), 449–467 (1996)
18. Kennedy, C.A., Carpenter, M.H.: Additive Runge-Kutta schemes for convection-diffusion-reaction equations. Appl. Numer. Math. **44**(1–2), 139–181 (2003)
19. Kružkov, S.N.: First order quasilinear equations with several independent variables. Mat. Sb. (N.S.) **81**(123), 228–255 (1970)
20. Kurganov, A., Petrova, G., Popov, B.: Adaptive semidiscrete central-upwind schemes for nonconvex hyperbolic conservation laws. SIAM J. Sci. Comput. **29**(6), 2381–2401 (2007)
21. LeVeque, R.J., Yee, H.C.: A study of numerical methods for hyperbolic conservation laws with stiff source terms. J. Comput. Phys. **86**(1), 187–210 (1990)
22. Pareschi, L., Russo, G.: Implicit-Explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation. J. Sci. Comput. **25**(1–2), 129–155 (2005)
23. Svärd, M., Mishra, S.: Implicit-explicit schemes for flow equations with stiff source terms. J. Comput. Appl. Math. **235**(6), 1564–1577 (2011)
24. Verdugo, F., Badia, S.: The software design of Gridap: a finite element package based on the Julia JIT compiler. Comput. Phys. Commun. **276**, 108341 (2022)