École des Ponts
ParisTech

École des Ponts ParisTech

2022
Internship report for:
Modélisation, Analyse, Simulation (MAS)

Zuodong Wang

Élève ingénieur

Analysis of an invariant-domain-preserving scheme for conservation laws and hyperbolic systems

(Analyse d’un schéma préservant le domaine invariant pour des lois de conservation et des systèmes hyperboliques)

Centre Inria de Paris

2 Rue Simone Iff, 75012 Paris

From 01/05/2022 to 31/09/2022

Under kind guidance of
Zhaonan Dong, Alexandre Ern & Jean-Luc Guermond

Fiche de synthèse

- Type de stage : stage de M2

- Année : 2022

- Auteur (Nom, prénom) : Wang, Zuodong

- Formation 3ème année: IMI

- Titre du rapport : Analysis of an invariant-domain-preserving scheme for conservation laws and hyperbolic systems

- Titre en français: Analyse d'un schéma préservant le domaine invariant pour des lois de conservation et des systèmes hyperboliques

- Organisme d'accueil : Centre Inria de Paris

- Pays d'accueil : France

- Responsables de stage : Zhaonan Dong, Alexandre Ern & Jean-Luc Guermond

- Mots-clés: lois de conservation, systèmes hyperboliques, domaine invariant, convergence du schéma

Synthèse du rapport en français

Ce stage porte sur la résolution numérique des lois de conservation et des systèmes hyperboliques. Plus précisément, l'objectif est de démontrer la stabilité et la convergence d'un schéma qui préserve la propriété du domaine invariant, proposé par Jean-Luc Guermond et ses collègues entre 2014 et 2016. La discrétisation temporelle du schéma est basée sur la méthode d'Euler explicite et la discrétisation spatiale utilise des éléments finis H^1-conformes. Il peut être démontré de manière informelle que le schéma est précis au premier ordre en temps et en espace, et qu'il préserve chaque ensemble invariant associé à une loi de conservation ou à un système hyperbolique. L'ensemble du stage est grosso modo divisé en deux parties. Durant la première moitié du stage (environ deux mois), mon travail principal consistait à implémenter numériquement ce schéma. La seconde moitié du stage fut principalement consacrée à explorer les propriétés du schéma au point théorique.

Les propriétés mathématiques des lois de conservation et des systèmes hyperboliques sont étudiées depuis longtemps. L'existence de chocs et de discontinuités pour les lois de conservation et les systèmes hyperboliques est un énorme défi pour les études théoriques et numériques. Les chercheurs utilisent la notion de solution faible pour traiter le problème de la discontinuité et utilisent l'inégalité d'entropie pour résoudre le problème de l'unicité à un certain niveau. Pour le cas scalaire, toute fonction convexe peut jouer le rôle de l'entropie, et la condition d'entropie résout le problème d'unicité. Mais pour les systèmes, la condition d'entropie ne garantit pas toujours l'unicité de la solution.

Dans la partie théorique, nous avons analysé la convergence du schéma. Pour le cas scalaire, nous avons d'abord établi une borne uniforme supérieure sur le terme de dissipation lié à la viscosité du graphe sous certaines hypothèses raisonnables. Ensuite, nous avons utilisé ce résultat pour prouver que la limite de notre solution numérique (si elle existe) est une solution faible et satisfait les inégalités d'entropie, à nouveau, sous certaines hypothèses raisonnables. Après, les systèmes hyperboliques sont également analysés avec un raisonnement similaire. Une différence importante est que, dans le cas scalaire, la propriété de préservation du domaine invariant est une propriété locale, et cette localité joue un rôle essentiel pour établir la stabilité du schéma. Mais dans le cas des systèmes, bien que nous pouvons observer cette localité dans des expériences numériques, le résultat théorique n'est pas encore clair. A cause de cela, dans notre analyse numérique, nous avons ajouté l'hypothèse de localité. D'autre part, l'hypothèse de localité peut être enlevée si nous introduisons une borne inférieure uniforme pour la borne supérieure de la vitesse maximale des ondes, et cela est discuté dans le rapport. Une difficulté pour l'analyse de convergence est la non-linéarité du flux. Puisque la solution numérique est uniformément bornée, il est possible de trouver une sous-suite convergente au sens faible*. Mais, à cause de la non-linéarité du flux, on ne peut rien dire sur la convergence du flux, et c'est la raison pour laquelle il faut ajouter une hypothèse sur la convergence de la solution numérique. Une possibilité d'enlever cette hypothèse est d'introduire les solutions à valeur de mesure, mais cela reste à explorer.

Dans la partie numérique, j'ai réalisé un code sur Python en suivant la structure du code en Matlab de Zhaonan Dong. Ce code contient cinq parties: lecture et traitement du maillage, intégration numérique, calcul matriciel, implémentation du schéma et estimation des erreurs. La première partie consiste à lire le maillage à partir du fichier Matlab écrit par Zhaonan Dong et à calculer les informations géométriques liées au schéma. La partie d'intégration numérique est consacrée au calcul des opérateurs différentiels et des intégrations numériques pour des bases Lagrange et Bernstein. La partie de calcul matriciel se concentre sur le calcul des matrices utilisées dans le schéma en utilisant la parallélisation de multi-thread. Pour l'implémentation du schéma, j'ai fourni deux façons de calculer l'évolution temporelle : (i) en utilisant la multiplication matrices-vecteurs et matrices-matrices, (ii) en utilisant le stencil de chaque degré de liberté et l'expression algébrique du schéma. La condition aux limites peut être imposée au sens fort, au sens faible ou en résolvant un problème de Riemann. Enfin, la partie d'estimation des erreurs est réalisée en utilisant une intégration numérique d'ordre élevé.

Contents

Abstract

This report is devoted to the study of an invariant-domain-preserving scheme to approximate the solution to conservation laws and hyperbolic systems under the simplifying assumption that the solution is compactly supported. This scheme consists of an explicit Euler time-stepping and continuous finite elements in space under a CFL stability condition. Our main result is to prove under some suitable assumptions that, if a subsequence of the numerical solutions converges in L^1 to a certain function in L^∞, then this function is a weak solution and an entropy solution. Finally, we present numerical solutions illustrating the scheme on conservation laws and hyperbolic systems.

1 Introduction

1.1 Conservation laws and hyperbolic systems

The mathematical properties of conservation laws and hyperbolic systems are studied in Chapter 1 in [2], [3], pages 1-104 in [11], Chapter 5 in [17], and Chapter 6 in [19]. The existence of shocks and discontinuities for conservation laws and hyperbolic systems is a huge challenge for theoretical and numerical studies. Indeed, even if the initial condition is smooth, it is possible to observe a shock during the time evolution. Therefore, one should interpret the solution in a weak sense instead of a strong sense. These weak solutions are merely bounded locally in space and in time. But the notion of weak solution is not sufficient to ensure the uniqueness of the solution. In many cases, one can construct more than one weak solution with the same initial condition. A reasonable weak solution should be a solution which correctly represents the physical properties. This is usually done by requiring the satisfaction of an entropy condition, whenever there are entropies associated with the conservation law or the hyperbolic system. For the scalar case, any convex function can play the role of the entropy, and the entropy condition resolves the problem of uniqueness. But for systems, the entropy condition cannot guarantee the uniqueness of the solution (see [4]). This well-posedness problem can be resolved to some level by introducing the so-called entropy measure-valued (EMV) solution as a probability measure, see [8].

1.2 Existing numerical methods in literature

From the numerical point of view, one wants to fit the properties observed at the PDE level, for instance, the invariant-domain-preserving properties (e.g. the positivity of mass density for Euler equations), and the entropy inequality. Moreover, if possible, one hopes to establish convergence for the scheme, for instance to ensure that the limit of a (sub)sequence of numerical solutions coincides with a suitable weak solution.

A huge body of literature is devoted to the numerical approximations of conservation laws and hyperbolic systems. For instance, [5, 6] studies the convergence to the entropy weak solution for scalar problems with finite volume method, and continuous/discontinuous Galerkin method are investigated in [14, 18].

1.3 An invariant-domain-preserving scheme

In this report, we focus on a scheme designed by Guermond and Nazarov [12] and Guermond and Popov [15]. It can be informally shown to be first-order accurate in time and in space and to preserve every invariant set of a conservation law or a hyperbolic system. The time discretization is based on the forward Euler method and the space discretization employs H^1-conforming finite elements. A overview can be found in Chapters 81, 82 and 83 in [10].

1.4 Outline

Our main result is to prove under some suitable assumptions that, if a subsequence of the numerical solutions converges in L^1 to a certain function in L^∞, then this function is a weak solution. This report is organized as follows: In Section 2, we introduce the model problem, and the basic properties of conservation laws and hyperbolic systems. In Section 3, we introduce the discrete setting and investigate the convergence of the scheme for conservation laws under some reasonable assumptions. In Section 4, we prove the convergence for hyperbolic systems, but with more assumptions. Section 5 is devoted to two further analyses: (i) proving that a subsequence of the numerical solutions converges to an entropy solution; (ii) removing some assumptions, by slightly modifying the scheme (increasing the graph viscosity). Finally, numerical experiments are presented in Section 6.

2 Model problem

In this section, we introduce the models investigated. A more detailed discussion of these models can be found in Chapter 79 and 80 of [10].

2.1 General setting

Let Ω be an open bounded polyhedral subset of $\mathbb{R}^d, d \geq 1$. Let us denote by $\partial\Omega$ its boundary, n_Ω the unit normal vector to $\partial\Omega$ outward to Ω, T the given final time. We denote the time variable by t and the spatial variable by x. The subscript of differential operators with respect to x is omitted, e.g., we note $\mathrm{div} := \mathrm{div}_x$. We also denote the Euclidean norm for all $\xi \in \mathbb{R}^n$ by $\|\xi\|_2 := \sqrt{\sum_{k=1}^n \xi_k^2}$.

2.2 Conservation laws

We introduce the model problem firstly.

Definition 2.1. (Model problem) We consider the following scalar conservation law:

$$\partial_t u + \mathrm{div}\boldsymbol{f}(u) = 0,\ \forall (x,t) \in \Omega \times (0,T), \tag{1}$$

with the initial condition

$$u(x,0) = u_0(x), \forall x \in \Omega.$$

Assumption 2.1. (Model assumptions) We assume the problem data are such that it is meaningful to consider the following boundary condition:

$$u(x,t) = 0,\ \forall (x,t) \in \partial\Omega \times (0,T).$$

Moreover, we assume that there is an invariant domain $\mathcal{B} \subset \mathbb{R}$, which is an interval such that the invariant-domain-preserving property is satisfied:

$$u(x,t) \in \mathcal{B},$$

for all $(x,t) \in \Omega \times (0,T)$. For conservation laws, $\mathcal{B}$ is a bounded interval and this property is called maximum principle. Finally, we make the following hypotheses on the initial data and on the flux:

$$\begin{aligned} &u_0 \in L^2(\Omega),\\ &\boldsymbol{f} \in W^{1,\infty}(\mathcal{B};\mathbb{R}^d). \end{aligned}$$

Then we introduce the notions of weak and entropy solutions.

Definition 2.2. (Weak solution) We say that $u \in L^\infty(\Omega \times (0,T))$ is a weak solution to (1) if for all $\varphi \in C_c^\infty(\Omega \times [0,T))$, we have

$$\int_0^T \int_\Omega u\partial_t\varphi + \int_0^T \int_\Omega \boldsymbol{f}(u) \cdot \nabla\varphi + \int_\Omega u_0\varphi(x,0) = 0. \tag{2}$$

Definition 2.3. (Entropy pair) For any convex function $\eta \in C^1(\mathcal{B})$ with associated flux $\boldsymbol{q} \in C^1(\mathcal{B};\mathbb{R}^d)$ such that $\partial_l \boldsymbol{q}(v) = \eta'(v)\partial_l \boldsymbol{f}(v)$ for all $1 \leq l \leq d$ and all $v \in \mathcal{B}$, we say that $(\eta, \boldsymbol{q})$ is an entropy pair for the conservation law (1).

Definition 2.4. (Entropy solution) We say that $u \in L^\infty(\Omega \times (0,T))$ is an entropy solution to (1) if for any entropy pair $(\eta, \boldsymbol{q})$, and for all $\varphi \in C_c^\infty(\Omega \times [0,T); \mathbb{R}_+)$, we have

$$-\int_0^T \int_\Omega \eta(u)\partial_t\varphi - \int_0^T \int_\Omega \boldsymbol{q}(u) \cdot \nabla\varphi - \int_\Omega \eta(u_0)\varphi(x,0) \leq 0. \tag{3}$$

The invariant-domain-preserving property is related to the Riemann problem and the Riemann average. Note that the Riemann average is an important ingredient for the design of our scheme, so we introduce it here.

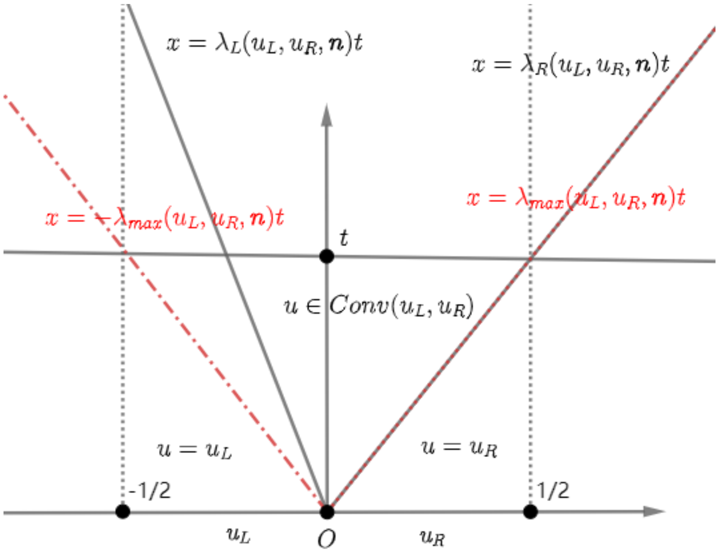

$x = \lambda_L(u_L, u_R, \boldsymbol{n})t$
$x = \lambda_R(u_L, u_R, \boldsymbol{n})t$
$x = -\lambda_{max}(u_L, u_R, \boldsymbol{n})t$
$x = \lambda_{max}(u_L, u_R, \boldsymbol{n})t$
t
$u \in Conv(u_L, u_R)$
$u = u_L$
$u = u_R$
-1/2
1/2
u_L
O
u_R

Figure 1: Solution of Riemann problem for scalar case

Definition 2.5. (Riemann problem) The Riemann problem is defined as follows: Find the entropy solution u such that

$$\partial_t u + \partial_x(\boldsymbol{f}(u) \cdot \boldsymbol{n}) = 0, \quad u(x, 0) := \begin{cases} u_L \text{ if } x < 0, \\ u_R \text{ if } x > 0, \end{cases} \tag{4}$$

where $u_L, u_R \in \mathbb{R}$ and $\boldsymbol{n}$ is an arbitrary unit vector in $\mathbb{R}^d$.

Definition 2.6. (Maximum wave speed) We refer to $\lambda_L(u_L, u_R, \boldsymbol{n})$ and $\lambda_R(u_L, u_R, \boldsymbol{n})$ as the left and right extreme wave speeds for the Riemann problem, respectively. The precise definition of the wave speed can be found in Section 79.2.4 of [10]. Any real number $\lambda_{max}(u_L, u_R, \boldsymbol{n})$ satisfying the inequality

$$\lambda_{max}(u_L, u_R, \boldsymbol{n}) \geq \max(|\lambda_L(u_L, u_R, \boldsymbol{n})|, |\lambda_R(u_L, u_R, \boldsymbol{n})|)$$

is called upper bound on the maximum wave speed.

An important property of the maximum wave speed is the following:

Lemma 2.1. *(Riemann average) Let $(\eta, \boldsymbol{q})$ be an entropy pair, u be the entropy solution to* (4)*, and define the Riemann average as $\bar{u}(t, u_L, u_R) := \int_{-\frac{1}{2}}^{\frac{1}{2}} u(x, t)dx$. Let $\lambda_{max}(u_L, u_R, \boldsymbol{n})$ be any upper bound on the maximum wave speed. Then, for all $t \in [0, \frac{1}{2\lambda_{max}(u_L, u_R, \boldsymbol{n})}]$,*

$$\bar{u}(t, u_L, u_R) = \frac{1}{2}(u_L + u_R) - t(\boldsymbol{f}(u_R) - \boldsymbol{f}(u_L)) \cdot \boldsymbol{n} \in Conv(u_L, u_R),$$
$$\eta(\bar{u}(t, u_L, u_R)) \leq \frac{1}{2}(\eta(u_L) + \eta(u_R)) - t(\boldsymbol{q}(u_R) - \boldsymbol{q}(u_L)) \cdot \boldsymbol{n},$$

where $Conv(u_L, u_R)$ denotes the set of all convex combinations of u_L and u_R.

An example of the Riemann problem is given in Figure 1, as well as an illustration of the maximum wave speeds and their upper bound. As we can see from the figure, when t is sufficiently small, the solution u is a convex combination between u_L and u_R.

2.3 Hyperbolic systems

Definition 2.7. (Models) We denote m the dimension of the system, and we consider the following hyperbolic system:

$$\partial_t \boldsymbol{u} + \text{div}\mathbb{f}(\boldsymbol{u}) = 0, \ \forall (x, t) \in \Omega \times (0, T), \tag{5}$$

with the initial condition

$$\boldsymbol{u}(x,0) = \boldsymbol{u}_0(x),\ \forall x \in \Omega.$$

Assumption 2.2. (Model assumptions) We assume the solution is compactly supported. Moreover, the invariant-domain-preserving property is satisfied, i.e., $\boldsymbol{u} \in \mathcal{B}$ for all $(x,t) \in \Omega \times (0,T)$. We emphasis that $\mathcal{B}$ is convex and can be unbounded. Finally, we make the following hypotheses on the data and on the flux:

$$\begin{aligned} &\boldsymbol{u}_0 \in L^2(\Omega; \mathbb{R}^m), \\ &\mathbb{f} \in W^{1,\infty}(\mathcal{B}; \mathbb{R}^{m\times d}). \end{aligned}$$

Then we introduce the notions of weak and entropy solutions.

Definition 2.8. (Weak solution) We say that $\boldsymbol{u} \in L^\infty(\Omega \times (0,T); \mathbb{R}^m)$ is a weak solution to (5) if for all $\boldsymbol{\varphi} \in C_c^\infty(\Omega \times [0,T); \mathbb{R}^m)$, we have

$$\int_0^T \int_\Omega \boldsymbol{u}\partial_t \boldsymbol{\varphi} + \int_0^T \int_\Omega \mathbb{f}(\boldsymbol{u}) : \nabla \boldsymbol{\varphi} + \int_\Omega \boldsymbol{u}_0 \boldsymbol{\varphi}(x,0) = 0. \tag{6}$$

Definition 2.9. (Entropy pair) We say that $(\eta, \boldsymbol{q})$ is an entropy pair for (5) if the function $\eta \in C^1(\mathcal{B}; \mathbb{R})$ is convex and if the function $\boldsymbol{q} \in C^1(\mathcal{B}; \mathbb{R}^d)$ is such that $\partial_l \boldsymbol{q}_k(\boldsymbol{v}) = \sum_{1\le i\le m} \partial_i \eta(\boldsymbol{v}) \partial_l \mathbb{f}_{ik}(\boldsymbol{v})$, for all $1 \le l \le m$, all $1 \le k \le d$, and all $\boldsymbol{v} \in \mathcal{B}$. In other words, $D_u \boldsymbol{q}(\boldsymbol{v}) = D_u \eta(\boldsymbol{v}) D_u \mathbb{f}(\boldsymbol{v})$, where D_u is the differential operator with respect to u.

Definition 2.10. (Entropy solution) We say that $\boldsymbol{u} \in L^\infty(\Omega\times(0,T); \mathbb{R}^m)$ is an entropy solution to (5) if for any entropy pair $(\eta, \boldsymbol{q})$, and for all $\varphi \in C_c^\infty(\Omega \times [0,T); \mathbb{R}_+)$, we have

$$-\int_0^T \int_\Omega \eta(\boldsymbol{u})\partial_t \varphi - \int_0^T \int_\Omega \boldsymbol{q}(\boldsymbol{u}) \cdot \nabla\varphi - \int_\Omega \eta(\boldsymbol{u}_0)\varphi(x,0) \le 0. \tag{7}$$

Similarly to the scalar case, the Riemann problem plays an important role, and we give the definition and properties here.

Definition 2.11. (Riemann problem) The Riemann problem is defined as follows: Find the entropy solution $\boldsymbol{u}$ such that

$$\partial_t \boldsymbol{u} + \partial_x(\mathbb{f}(\boldsymbol{u}) \cdot \boldsymbol{n}) = 0, \quad \boldsymbol{u}(x,0) := \begin{cases} \boldsymbol{u}_L \text{ if } x < 0, \\ \boldsymbol{u}_R \text{ if } x > 0, \end{cases} \tag{8}$$

where $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathcal{B}$ and $\boldsymbol{n}$ is an arbitrary unit vector in $\mathbb{R}^d$.

Definition 2.12. (Maximum wave speed) We refer to $\lambda_1^- \le \lambda_1^+ \le \lambda_2^- \le .. \le \lambda_m^- \le \lambda_m^+$ as the wave speeds for the Riemann problem. The precise definition can be found in Section 80.2.1 of [10]. Any real number $\lambda_{max}(\boldsymbol{u}_L, \boldsymbol{u}_R, \boldsymbol{n})$ satisfying the inequality

$$\lambda_{max}(\boldsymbol{u}_L, \boldsymbol{u}_R, \boldsymbol{n}) \ge \max(|\lambda_1^-|, |\lambda_m^+|)$$

is called upper bound on the maximum wave speed.

One important property of the maximum wave speed is the following:

Lemma 2.2. *(Riemann average) Let* $(\eta, \boldsymbol{q})$ *be an entropy pair,* $\boldsymbol{u}$ *be the entropy solution to* (8)*, and define the Riemann average as* $\bar{\boldsymbol{u}}(t, \boldsymbol{u}_L, \boldsymbol{u}_R) := \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{u}(x,t)dx$*. Let* $\lambda_{max}(\boldsymbol{u}_L, \boldsymbol{u}_R, \boldsymbol{n})$ *be any upper bound on the maximum wave speed. Then, for all* $t \in [0, \frac{1}{2\lambda_{max}(\boldsymbol{u}_L, \boldsymbol{u}_R, \boldsymbol{n})}]$,

$$\begin{aligned} \bar{\boldsymbol{u}}(t, \boldsymbol{u}_L, \boldsymbol{u}_R) &= \frac{1}{2}(\boldsymbol{u}_L + \boldsymbol{u}_R) - t(\mathbb{f}(\boldsymbol{u}_R) - \mathbb{f}(\boldsymbol{u}_L)) \cdot \boldsymbol{n} \in \mathcal{B}, \\ \eta(\bar{\boldsymbol{u}}(t, \boldsymbol{u}_L, \boldsymbol{u}_R)) &\le \frac{1}{2}(\eta(\boldsymbol{u}_L) + \eta(\boldsymbol{u}_R)) - t(\boldsymbol{q}(\boldsymbol{u}_R) - \boldsymbol{q}(\boldsymbol{u}_L)) \cdot \boldsymbol{n}. \end{aligned}$$

3 Numerical analysis for conservation laws

In the first part of this section, we introduce the finite element space. The second part is devoted to describing the scheme. In the third part, we prove that some properties satisfied for solution at the PDE level (e.g., invariant-domain-preserving, entropy inequality) can also be satisfied for discrete solution in some sense. Moreover, we also prove that the numerical solution converges to a weak solution at the PDE level.

3.1 Discrete setting

Definition 3.1. (Time discretization) We introduce the discrete time nodes t_n for all $n \in \mathcal{N}_T$, where $\mathcal{N}_T = \{0, ..., N-1\}$, $t_0 = 0$ and $t_N = T$. The time step τ_n satisfies $t_{n+1} = t_n + \tau_n$ and we set $I_n = [t_n, t_{n+1})$.

Definition 3.2. (Mesh and FEM space) We consider a shape-regular sequence of matching meshes $\{\mathcal{T}_h\}_h$, $h = \max_{K \in \mathcal{T}_h} h_K$ where h_K denotes the diameter of cell K. For simplicity, we assume that the sequence $\{\mathcal{T}_h\}_h$ is quasi-uniform. The reference element is denoted by $\hat{K}$. We denote by $\Phi_K : \hat{K} \to K$ the diffeomorphism mapping $\hat{K}$ to an arbitrary element $K \in \mathcal{T}_h$. We introduce the reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ and we define the scalar-valued finite elements space

$$V_h := \{v \in C^0(\Omega; \mathbb{R}) : v|_K \circ \Phi_K \in \hat{P}, \ \forall K \in \mathcal{T}_h\},$$

where $\hat{P}$ is the reference space. We also define the finite elements space with zero boundary condition:

$$V_h^0 := V_h \cap H_0^1(\Omega).$$

Letting $n_{sh} = \dim \hat{P}$, the shape functions on the reference element are denoted by $\{\hat{\theta}_i\}_{1 \le i \le n_{sh}}$. We assume that the basis $\{\hat{\theta}_i\}_{1 \le i \le n_{sh}}$ has the partition of unity property:

$$\sum_{1 \le i \le n_{sh}} \hat{\theta}_i(\hat{x}) = 1, \ \forall \hat{x} \in \hat{K}. \tag{9}$$

The global shape functions are denoted by $\{\phi_i\}_{i \in \mathcal{A}_h}$, where $\mathcal{A}_h$ is the set of degrees of freedom. We also denote the interior degrees of freedom as $\mathcal{A}_h^0 := \{i \in \mathcal{A}_h : \phi_i|_{\partial\Omega} = 0\}$. These functions form a basis of V_h, and the partition of unity property implies that $\sum_{i \in \mathcal{A}_h} \phi_i(x) = 1$ for all $x \in \Omega$. The support of ϕ_i is denoted by ω_i, for all $i \in \mathcal{A}_h$. The set of indices of shape functions whose support on E is of nonzero measure is denoted by $\mathcal{I}(E) := \{j \in \mathcal{A}_h : |\omega_j \cap E| \neq 0\}$, where $|\cdot|$ denotes the measure of a set. The set of indices of shape functions whose support on ω_i is of nonzero measure is denoted by

$$\mathcal{I}(i) := \mathcal{I}(\omega_i) = \{j \in \mathcal{A}_h : |\omega_j \cap \omega_i| \neq 0\}.$$

This set defines the stencil for the finite element scheme.

The matrix with entries $m_{ij} := \int_\Omega \phi_i(x)\phi_j(x)dx, \ i, j \in \mathcal{A}_h$ is called the consistent mass matrix and is denoted by $\mathcal{M}$. The diagonal matrix with entries equal to $m_i := \int_\Omega \phi_i(x)dx$ is called the lumped mass matrix and is denoted by $\mathcal{M}^L$. The partition of unity property implies that $\sum_{j \in \mathcal{I}(i)} m_{ij} = m_i$. One key assumption used in the rest of the report is that

$$m_i > 0, \ \forall i \in \mathcal{A}_h. \tag{10}$$

The assumptions (9) and (10) hold for many Lagrange elements and for Bernstein-Bezier finite elements of any polynomial degree, as mentioned in [14]. For simplicity, we only consider P_1 Lagrange elements in this report.

In various bounds, we denote by C any generic constant (its value can change at each occurrence) that is independent of h and N, but may depend on d, m, T, the Lipschitz constants of the flux and the entropy, and some constants related to the approximation properties of Lagrange finite element.

Owing to the mesh assumptions, we have, for all $v_h = \sum_{i \in \mathcal{A}_h} V_i \phi_i \in V_h$, the following norm equivalence:

$$C\|v_h\|_{L^2(\Omega)}^2 \le \|v_h\|_{l_h^2}^2 \le C\|v_h\|_{L^2(\Omega)}^2,$$

where $\|v_h\|_{l_h^2}^2 := \sum_{i \in \mathcal{A}_h} m_i (V_i)^2$.

3.2 Scheme

Definition 3.3. (Scheme) We denote the spatial approximation of u in the interval I_n as

$$u_h^n(x) := \sum_{i\in\mathcal{A}_h} U_i^n \phi_i(x),$$

for all $n \in \mathcal{N}_T$. The global approximation is defined as $u_{h,\tau}(x,t)|_{I_n} := u_h^n(x)$. The scheme is defined as follows:

$$m_i \frac{U_i^{n+1} - U_i^n}{\tau_n} + \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \Big(\boldsymbol{f}(U_j^n)\boldsymbol{c}_{ij} + d_{ij}^n (U_i^n - U_j^n) \Big) = 0, \tag{11}$$

for all $i \in \mathcal{A}_h^0$ and all $n \in \mathcal{N}_T$, where $\boldsymbol{c}_{ij} := \int_\Omega \phi_i \nabla\phi_j \in \mathbb{R}^d$ and

$$d_{ij}^n := \max(\lambda_{max}(U_i^n, U_j^n, \boldsymbol{n}_{ij})\|\boldsymbol{c}_{ij}\|_2, \lambda_{max}(U_j^n, U_i^n, \boldsymbol{n}_{ji})\|\boldsymbol{c}_{ji}\|_2) \in \mathbb{R}_+,$$

with $\lambda_{max}(U_i^n, U_j^n, \boldsymbol{n}_{ij})$ any upper bound on the maximum wave speed in the Riemann problem with data (U_i^n, U_j^n) and the normal vector $\boldsymbol{n}_{ij} := \frac{\boldsymbol{c}_{ij}}{\|\boldsymbol{c}_{ij}\|_2}$. Moreover, the following CFL condition should be satisfied with a constant $\rho \in (0,1]$:

$$\tau_n \le \rho \min_{i\in\mathcal{A}_h^0} \frac{m_i}{2d_{ii}^n}, \tag{12}$$

where $d_{ii}^n := \sum_{j\in\mathcal{I}(i)\backslash\{i\}} d_{ij}^n$. The initial data is approximated by the L^2-projection onto V_h^0:

$$\int_\Omega (u_h^0 - u_0)\phi_i = 0,$$

for all $i \in \mathcal{A}_h^0$. And to be consistent with our assumption on the compactness of the support of u at the boundary, the boundary coefficients are set to zero, i.e., $U_i^n = 0$ for all $n \in \mathcal{N}_T$ and $i \in \mathcal{A}_h \setminus \mathcal{A}_h^0$.

3.3 Basic properties

We introduce two basic properties here: local maximum principle and discrete entropy inequality. The first one is important for establishing the stability of the scheme and for proving that the limit of numerical solution (up to a subsequence) is a weak solution; the second one is important for proving that the limit of numerical solution (up to a subsequence) is the entropy solution. This will be further discussed in Section 5.

Firstly, we can rewrite the scheme for all $i \in \mathcal{A}_h^0$ as follows:

$$U_i^{n+1} = \sum_{j\in\mathcal{I}(i)} \theta_{ij}^n \bar{U}_{ij}^n, \tag{13}$$

where $\theta_{ij}^n := \frac{2\tau_n d_{ij}^n}{m_i}$ for all $j \in \mathcal{I}(i)\backslash\{i\}$, $\theta_{ii}^n := 1 - \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \theta_{ij}^n$, and $\bar{U}_{ii}^n := U_i^n$,

$$\bar{U}_{ij}^n := \frac{1}{2}(U_i^n + U_j^n) - (\boldsymbol{f}(U_j^n) - \boldsymbol{f}(U_i^n))\frac{\boldsymbol{c}_{ij}}{2d_{ij}^n}.$$

An important property is that $\bar{U}_{ij}^n = \int_{-\frac{1}{2}}^{\frac{1}{2}} u_{ij}^n(x, t_{ij}^n)dx$, where u_{ij}^n is the exact solution for Riemann problem with the initial data (U_i^n, U_j^n), the normal vector $\boldsymbol{n}_{ij}$, and the artificial time $t_{ij}^n := \frac{\|\boldsymbol{c}_{ij}\|_2}{2d_{ij}^n}$. Owing to the definition of d_{ij}^n, we infer that t_{ij}^n is sufficiently small so that $\bar{U}_{ij}^n \in Conv(U_i^n, U_j^n)$. Moreover, the CFL condition (12) gives $\theta_{ij}^n \in [0,1]$ for all $j \in \mathcal{I}(i)$.

The following lemmas are proved in [12, 15]. See also Theorem 81.8, Corollary 81.9 and Theorem 81.12 in [10].

Lemma 3.1. *(Maximum principle) Under CFL condition* (12)*, we have the local maximum principle and global maximum principle, i.e.,*

$$U_i^{n+1} \in Conv_{j\in\mathcal{I}(i)}(U_j^n), \quad u_h^n(x) \in Conv_{j\in\mathcal{I}(i)}(U_j^0),$$

for all $i \in \mathcal{A}_h$ and all $n \in \mathcal{N}_T$, where $Conv_{j\in\mathcal{I}(i)}(U_j^n)$ denotes the set of all convex combinations for $\{U_j^n\}_{j\in\mathcal{I}(i)}$.

Lemma 3.2. *(Discrete entropy inequality) Assume the CFL condition* (12) *is satisfied. Let $(\eta, \boldsymbol{q})$ be an entropy pair for* (1)*. Then, the following discrete entropy inequality holds true for all $i \in \mathcal{A}_h^0$ and all $n \in \mathcal{N}_T$:*

$$\frac{m_i}{\tau_n}(\eta(U_i^{n+1}) - \eta(U_i^n) + \int_\Omega \mathrm{div}(\mathcal{I}_h(\boldsymbol{q}(u_h^n)))\phi_i + \sum_{j\in\mathcal{I}(i)} d_{ij}^n(\eta(U_i^n) - \eta(U_j^n)) \leq 0, \tag{14}$$

where $\mathcal{I}_h(\boldsymbol{q}(u_h^n)) := \sum\limits_{i\in\mathcal{A}_h} \boldsymbol{q}(U_i^n)\phi_i$ is the Lagrange interpolation of $\boldsymbol{q}(u_h^n)$.

3.4 Convergence analysis

In this section, we present the convergence result derived during this internship.

Lemma 3.3. *(control on flux) For any function $v_h = \sum_{i\in\mathcal{A}_h} V_i\phi_i \in V_h^0$ and any function $g \in W^{1,\infty}(\mathcal{B}; \mathbb{R}^d)$, we have the following bound:*

$$\|g(v_h) - \mathcal{I}_h(g(v_h))\|_{L^2(\Omega)} + h\|\mathrm{div}(g(v_h) - \mathcal{I}_h(g(v_h)))\|_{L^2(\Omega)} \leq \|g\|_{W^{1,\infty}(\mathcal{B})} C_I C_d C_{inv} h \|\nabla v_h\|_{L^2(\Omega)}, \tag{15}$$

where C_I is the constant from the approximation theorems for Lagrange interpolation, C_{inv} is the constant from the discrete inverse inequality and C_d is the constant for inequality $\|vw\|_{L^2(\Omega)} \leq C_d\|v\|_{L^\infty(\Omega)}\|w\|_{L^2(\Omega)}$ for all $v \in L^\infty(\Omega; \mathbb{R}^{m\times d})$ and $w \in L^2(\Omega; \mathbb{R}^d)$.

Proof. We first prove $g(v_h) \in W^{1,\infty}(\Omega)$ by considering the lemma 1.23 in [7]: if $g(v_h) \in W^{1,\infty}(\mathcal{T}_h)$ and the jump of $g(v_h)$ is zero, then $g(v_h) \in W^{1,\infty}(\Omega)$.

Preciously, on each cell $K \in \mathcal{T}_h$, we have $v_h|_K \in C^\infty(K)$ and $\boldsymbol{f} \in W^{1,\infty}(\mathcal{B})$, so $g(v_h)|_K \in W^{1,\infty}(K)$. This implies $g(v_h) \in W^{1,\infty}(\mathcal{T}_h)$. Since $W^{1,\infty}(\mathcal{B}) \hookrightarrow C^0(\mathcal{B})$ and $v_h \in C^0$, we infer that $g(v_h) \in C^0(\Omega)$. So, on any face of K, the jump is zero. combining these two results proves $g(v_h) \in W^{1,\infty}(\Omega)$.

We then apply the approximation result from [1] to prove the bound for $\|g(v_h) - \mathcal{I}_h(g(v_h))\|_{L^2(\Omega)}$: For any $p > d$, we have $W^{1,p} \hookrightarrow C^0$ and we have $W^{1,p} \hookrightarrow L^q$ for all $q < \infty$. So, on each cell K,we apply the theorem 2.9 in this article with $p > d, q = 2$ to get

$$\begin{aligned}
\|g(v_h) - \mathcal{I}_h(g(v_h))\|_{L^2(K)} &\leq C_I h^{d(\frac{1}{2}-\frac{1}{p})} h \|\nabla g(v_h)\|_{L^p(K)} \\
&\leq C_I C_d h^{d(\frac{1}{2}-\frac{1}{p})} h \|g\|_{W^{1,\infty}(\mathcal{B})} \|\nabla v_h\|_{L^p(K)} \\
&\leq C_I C_d C_{inv} h \|g\|_{W^{1,\infty}(\mathcal{B})} \|\nabla v_h\|_{L^2(K)},
\end{aligned}$$

where we used the discrete inverse inequality(e.g. lemma 1.50 in[7]) for the last inequality. Then, we get the expected bound by summing over all cells $K \in \mathcal{T}_h$.

For $h\|\mathrm{div}(g(v_h) - \mathcal{I}_h(g(v_h)))\|_{L^2(\Omega)}$, we use the stability property of Lagrange interpolation mentioned in theorem 11.13 of [9] to derive the bound: On each cell K, we use the Hölder inequality to get

$$\begin{aligned}
\|\mathrm{div}(g(v_h) - \mathcal{I}_h(g(v_h))\|_{L^2(K))} &\leq h^{d(\frac{1}{2}-\frac{1}{p})} \|\mathrm{div}(g(v_h) - \mathcal{I}_h(g(v_h))\|_{L^p(K))} \\
&\leq C_I h^{d(\frac{1}{2}-\frac{1}{p})} \|\nabla g(v_h)\|_{L^p(K)} \\
&\leq C_I C_d h^{d(\frac{1}{2}-\frac{1}{p})} \|g\|_{W^{1,\infty}(\mathcal{B})} \|\nabla v_h\|_{L^p(K)} \\
&\leq C_I C_d C_{inv} \|g\|_{W^{1,\infty}(\mathcal{B})} \|\nabla v_h\|_{L^2(K)}.
\end{aligned}$$

Then, summing over $K \in \mathcal{T}_h$ concludes the proof. □

Lemma 3.4. *(Estimate on temporal accumulation) Under the CFL condition* (12)*, we have the following estimate for all $n \in \mathcal{N}_T$:*

$$\|u_h^{n+1} - u_h^n\|_{l_h^2}^2 \leq 2\rho\tau_n b_h^n(u_h^n, u_h^n), \tag{16}$$

where $b_h^n(v_h, w_h) := \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n(V_i - V_j)(W_i - W_j)$, for all $v_h, w_h \in V_h^0$.

Proof. With the help of (13) and the CFL condition (12), we obtain

$$m_i(U_i^{n+1} - U_i^n)^2 = m_i\Big(\sum_{j\in\mathcal{I}(i)} \theta_{ij}^n(\bar{U}_{ij}^n - U_i^n)\Big)^2 = \frac{4\tau_n^2}{m_i}\Big(\sum_{j\in\mathcal{I}(i)} d_{ij}^n(\bar{U}_{ij}^n - U_i^n)\Big)^2 \leq \rho\tau_n\frac{2}{d_{ii}^n}\Big(\sum_{j\in\mathcal{I}(i)} d_{ij}^n(\bar{U}_{ij}^n - U_i^n)\Big)^2.$$

Then, the Cauchy-Schwarz inequality implies that

$$m_i(U_i^{n+1} - U_i^n)^2 \leq 2\rho\tau_n \sum_{j\in\mathcal{I}(i)} d_{ij}^n(\bar{U}_{ij}^n - U_i^n)^2.$$

Now, we take the sum over all $i \in \mathcal{A}_h$, notice that $\bar{U}_{ij}^n = \bar{U}_{ji}^n$ and $d_{ij}^n = d_{ji}^n$, so that

$$\begin{aligned}\sum_{i\in\mathcal{A}_h} m_i(U_i^{n+1} - U_i^n)^2 &\leq 2\rho\tau_n \sum_{i\in\mathcal{A}_h, j\in\mathcal{I}(i)} d_{ij}^n(\bar{U}_{ij}^n - U_i^n)^2 \\ &= 2\rho\tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n\Big((\bar{U}_{ij}^n - U_i^n)^2 + (\bar{U}_{ij}^n - U_j^n)^2\Big).\end{aligned}$$

Then, we introduce the quadratic function $\Phi(x, a, b) := (x-a)^2 + (x-b)^2$ which takes values in $\big[\frac{1}{2}(b-a)^2, (b-a)^2\big]$ for all $x \in Conv(a, b)$. Thus, by considering this function and noticing that $\bar{U}_{ij}^n \in Conv(U_i^n, U_j^n)$, we have

$$\begin{aligned}\sum_{i\in\mathcal{A}_h} m_i(U_i^{n+1} - U_i^n)^2 &\leq 2\rho\tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n\Big((\bar{U}_{ij}^n - U_i^n)^2 + (\bar{U}_{ij}^n - U_j^n)^2\Big) \\ &= 2\rho\tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n\Phi(\bar{U}_{ij}^n, U_i^n, U_j^n) \\ &\leq 2\rho\tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n(U_i^n - U_j^n)^2.\end{aligned}$$

This proves (16). □

Assumption 3.1. (BV-like estimate) We assume that $\sum_{n\in\mathcal{N}_T} \tau_n h\|\nabla u_h^n\|_{L^2(\Omega)}^2 \leq C$.

Lemma 3.5. *(Bound on dissipation) Assume the CFL condition* (12) *holds true with $\rho < 1$ and Assumption 3.1 holds. Then we have the following stability property:*

$$\|u_h^N\|_{l_h^2}^2 + 2(1-\rho)\sum_{n\in\mathcal{N}_T} \tau_n b_h^n(u_h^n, u_h^n) \leq \|u_h^0\|_{l_h^2}^2 + C. \tag{17}$$

Proof. We multiply the scheme by $2\tau_n U_i^n$ and sum over all $i \in \mathcal{A}_h$ and $n \in \mathcal{N}_T$, and use the CFL condition (12) to obtain the expected bound. More precisely, we start by multiplying (11) by $2\tau_n U_i^n$, and notice that $2ab - 2b^2 = a^2 - b^2 - (a-b)^2$. This gives

$$m_i(U_i^{n+1})^2 + 2\tau_n\int_\Omega \mathrm{div}\mathcal{I}_h(\boldsymbol{f}(u_h^n))\phi_i U_i^n + 2\tau_n\sum_{j\in\mathcal{I}(i)} d_{ij}(U_i^n - U_j^n)U_i^n = m_i(U_i^n)^2 + m_i(U_i^{n+1} - U_i^n)^2.$$

Then, we sum over all $i \in \mathcal{A}_h$ and $n \in \mathcal{N}_T$ to get

$$\|u_h^N\|_{l_h^2}^2 + 2\sum_{n\in\mathcal{N}_T}\tau_n\int_\Omega \mathrm{div}\mathcal{I}_h(\boldsymbol{f}(u_h^n))u_h^n + 2\sum_{n\in\mathcal{N}_T}\tau_n\sum_{i\in\mathcal{A}_h, j<i} d_{ij}(U_i^n - U_j^n)^2 = \|u_h^0\|_{l_h^2}^2 + \sum_{n\in\mathcal{N}_T}\|u_h^{n+1} - u_h^n\|_{l_h^2}^2. \tag{18}$$

We first estimate $\int_\Omega \mathrm{div}\mathcal{I}_h(\boldsymbol{f}(u_h^n))u_h^n$. We remove and add some terms to get

$$\mathrm{div}\mathcal{I}_h(\boldsymbol{f}(u_h^n))u_h^n = \mathrm{div}\Big(\mathcal{I}_h(\boldsymbol{f}(u_h^n)) - \boldsymbol{f}(u_h^n)\Big)u_h^n + \mathrm{div}\boldsymbol{f}(u_h^n)u_h^n.$$

Noticing that $u_h^n|_{\partial\Omega} = 0$, introducing $\boldsymbol{q}(u) = \int_0^u \boldsymbol{f}(s)ds$ and using the Stokes formula, we get

$$\int_\Omega \mathrm{div}\boldsymbol{f}(u_h^n)u_h^n = -\int_\Omega \boldsymbol{f}(u_h^n)\nabla u_h^n = -\int_\Omega \mathrm{div}\boldsymbol{q}(u_h^n) = -\int_{\partial\Omega} \boldsymbol{q}(u_h^n)n_\Omega = \int_{\partial\Omega} \boldsymbol{q}(0)n_\Omega = 0.$$

Moreover, integrating by parts and using the approximation properties of Lagrange interpolation, we infer that

$$\begin{aligned}
\Big|\int_\Omega \mathrm{div}\Big(\mathcal{I}_h(\boldsymbol{f}(u_h^n)) - \boldsymbol{f}(u_h^n)\Big)u_h^n\Big| &= \Big|\int_\Omega \Big(\mathcal{I}_h(\boldsymbol{f}(u_h^n)) - \boldsymbol{f}(u_h^n)\Big)\nabla u_h^n\Big| \\
&\le \|\mathcal{I}_h(\boldsymbol{f}(u_h^n)) - \boldsymbol{f}(u_h^n)\|_{L^2(\Omega)}\|\nabla u_h^n\|_{L^2(\Omega)} \\
&\le Ch\|\nabla \boldsymbol{f}(u_h^n)\|_{L^2(\Omega)}\|\nabla u_h^n\|_{L^2(\Omega)} \\
&\le Ch\|\nabla u_h^n\|_{L^2(\Omega)}^2,
\end{aligned}$$

where C depends on the Lipschitz constant of $\boldsymbol{f}$. So we get

$$\sum_{n\in\mathcal{N}_T} \tau_n\Big|\int_\Omega \mathrm{div}\mathcal{I}_h(\boldsymbol{f}(u_h^n))u_h^n\Big| \le C\sum_{n\in\mathcal{N}_T} \tau_n h\|\nabla u_h^n\|_{L^2(\Omega)}^2. \tag{19}$$

For the last term on the right-hand side of (18), we use (16) to infer that

$$\sum_{n\in\mathcal{N}_T} \|u_h^{n+1} - u_h^n\|_{l_h^2}^2 \le 2\rho\sum_{n\in\mathcal{N}_T} \tau_n b_h^n(u_h^n, u_h^n). \tag{20}$$

After obtaining the estimates (19) and (20), we put them into (18) and obtain

$$\|u_h^N\|_{l_h^2}^2 + 2(1-\rho)\sum_{n\in\mathcal{N}_T}\tau_n\sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n(U_i^n - U_j^n)^2 \le \|u_h^0\|_{l_h^2}^2 + C\sum_{n\in\mathcal{N}_T}\tau_n h\|\nabla u_h^n\|_{L^2(\Omega)}^2.$$

We conclude by using Assumption 3.1.

□

Assumption 3.2. (Convergence in L^1) We assume that the sequence of numerical solutions $\{u_{h,\tau}\}_h$ has a subsequence (still denoted by $\{u_{h,\tau}\}_h$) which converges strongly to some function $u \in L^\infty(\Omega\times(0,T))$ in L^1, i.e., that $\lim_{h\to 0}\|u_{h,\tau} - u\|_{L^1(\Omega\times(0,T))} = 0$.

Theorem 3.1. *(Convergence of the scheme) Assume that the CFL condition* (12) *holds with $\rho < 1$, and that Assumptions 3.1 and 3.2 hold. Then, the limit u of the sequence of numerical solutions (up to a subsequence) is a weak solution of* (1)*, i.e., for all $\varphi \in C_c^\infty(\Omega\times[0,T))$, we have*

$$\int_0^T\int_\Omega u\partial_t\varphi + \int_0^T\int_\Omega \boldsymbol{f}(u)\cdot\nabla\varphi + \int_\Omega u_0\varphi(x,0) = 0.$$

Proof. We introduce the the Lagrange interpolant $\varphi_h^n := \mathcal{I}_h(\varphi(\cdot,t_n)) = \sum_{i\in\mathcal{A}_h}\varphi_i^n\phi_i$, with $\varphi_i^n := \varphi(x_i,t_n)$ where x_i denotes the ith interpolation node, and we introduce the notation $\varphi^n := \varphi(\cdot,t_n)$. We multiply the scheme (11) by φ_i^n and sum over all $i\in\mathcal{A}_h$ and all $n\in\mathcal{N}_T$ to get

$$T_{1,h} + T_{2,h} + T_{3,h} = 0, \tag{21}$$

where

$$\begin{aligned}
T_{1,h} &:= \sum_{n\in\mathcal{N}_T}\sum_{i\in\mathcal{A}_h} m_i(U_i^{n+1} - U_i^n)\varphi_i^n, \\
T_{2,h} &:= \sum_{n\in\mathcal{N}_T}\tau_n\int_\Omega \mathrm{div}\mathcal{I}_h(\boldsymbol{f}(u_h^n))\varphi_h^n, \\
T_{3,h} &:= \sum_{n\in\mathcal{N}_T}\tau_n\sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n(U_i^n - U_j^n)(\varphi_i^n - \varphi_j^n).
\end{aligned}$$

Similarly, we introduce two terms:

$$\begin{aligned}\tilde{T}_{1,h} &:= -\int_0^T \int_\Omega u_{h,\tau} \partial_t \varphi - \int_\Omega u_0 \varphi(x,0),\\ \tilde{T}_{2,h} &:= -\int_0^T \int_\Omega \boldsymbol{f}(u_{h,\tau}) \cdot \nabla \varphi.\end{aligned}$$

We will prove that $T_{1,h} - \tilde{T}_{1,h} \xrightarrow{h\to 0} 0$, $T_{2,h} - \tilde{T}_{2,h} \xrightarrow{h\to 0} 0$ and $T_{3,h} \xrightarrow{h\to 0} 0$. Then, passing to the limit in (21) gives $\tilde{T}_{1,h} + \tilde{T}_{2,h} \xrightarrow{h\to 0} 0$. After that, we decompose the left-hand side of (2) into two parts:

$$\begin{aligned}T_{1,0} &:= -\int_0^T \int_\Omega u \partial_t \varphi - \int_\Omega u_0 \varphi(x,0),\\ T_{2,0} &:= -\int_0^T \int_\Omega \boldsymbol{f}(u) \cdot \nabla \varphi.\end{aligned}$$

Then, using Assumption 3.2, we conclude by observing that $\tilde{T}_{1,h} \xrightarrow{h\to 0} T_{1,0}$ and $\tilde{T}_{2,h} \xrightarrow{h\to 0} T_{2,0}$.

We start with $T_{1,h} - \tilde{T}_{1,h}$ which can be considered as the error in time. The main idea is to bound the temporal accumulation term $m_i|U_i^{n+1} - U_i^n|$ using (16) and to use the regularity of the test function. In details, integrating by parts in $\tilde{T}_{1,h}$ in time gives

$$\tilde{T}_{1,h} = \sum_{n\in\mathcal{N}_T} \sum_{i\in\mathcal{A}_h} (U_i^{n+1} - U_i^n) \int_{\omega_i} \phi_i \varphi^n + \int_\Omega (u_h^0 - u_0)\varphi^0.$$

So,

$$T_{1,h} - \tilde{T}_{1,h} = \sum_{n\in\mathcal{N}_T} \sum_{i\in\mathcal{A}_h} (U_i^{n+1} - U_i^n) \int_{\omega_i} \phi_i (\varphi_i^n - \varphi^n) - \int_\Omega (u_h^0 - u_0)\varphi^0.$$

The L^2-orthogonality of the initial condition ensures the second term tends to zero:

$$\Big| \int_\Omega (u_h^0 - u_0)\varphi^0 \Big| = \Big| \int_\Omega (u_h^0 - u_0)(\varphi^0 - \mathcal{I}_h(\varphi^0)) \Big| \leq C\|u_0\|_{L^2(\Omega)} \|\varphi^0 - \mathcal{I}_h(\varphi^0)\|_{L^2(\Omega)} \xrightarrow{h\to 0} 0.$$

For the first term, we need the following relation which is valid owing to the mesh assumptions:

$$C|\omega_i| \leq m_i \leq |\omega_i|, \quad \forall i \in \mathcal{A}_h.$$

We invoke this relation to obtain

$$\Big| \sum_{n\in\mathcal{N}_T} \sum_{i\in\mathcal{A}_h} (U_i^{n+1} - U_i^n) \int_{\omega_i} \phi_i (\varphi_i^n - \varphi^n) \Big| \leq C \sum_{n\in\mathcal{N}_T} \sum_{i\in\mathcal{A}_h} |U_i^{n+1} - U_i^n| \int_{\omega_i} |\phi_i| h \leq C \sum_{n\in\mathcal{N}_T} h \sum_{i\in\mathcal{A}_h} m_i |U_i^{n+1} - U_i^n|.$$

Using the Cauchy-Schwarz inequality, Young's inequality, $m_i > 0$ for all $i \in \mathcal{A}_h$ and $\sum_{i\in\mathcal{A}_h} m_i = |\Omega|$, we get

$$\sum_{i\in\mathcal{A}_h} m_i |U_i^{n+1} - U_i^n| \leq |\Omega|^{1/2} \Big(\sum_{i\in\mathcal{A}_h} m_i (U_i^{n+1} - U_i^n)^2 \Big)^{1/2} \leq \frac{1}{2} h^{1/2} |\Omega| + \frac{1}{2} h^{-1/2} \sum_{i\in\mathcal{A}_h} m_i (U_i^{n+1} - U_i^n)^2.$$

Inserting this estimate into the above inequality, using (16) and $\sum_{n\in\mathcal{N}_T} \tau_n b_h^n(u_h^n, u_h^n) \leq C$, we infer that

$$\begin{aligned}\Big| \sum_{n\in\mathcal{N}_T} \sum_{i\in\mathcal{A}_h} (U_i^{n+1} - U_i^n) \int_{\omega_i} \phi_i (\varphi_i^n - \varphi^n) \Big| &\leq C \sum_{n\in\mathcal{N}_T} h \sum_{i\in\mathcal{A}_h} m_i |U_i^{n+1} - U_i^n| \\ &\leq C h^{1/2} |\Omega| \sum_{n\in\mathcal{N}_T} h + C h^{1/2} \sum_{n\in\mathcal{N}_T} \|u_h^{n+1} - u_h^n\|_{l_h^2}^2 \leq C h^{1/2}.\end{aligned}$$

To estimate $T_{2,h} - \tilde{T}_{2,h}$ which can be considered as the error in space, the idea is to use Assumption 3.1. More precisely,

$$\begin{aligned} T_{2,h} - \tilde{T}_{2,h} &= \sum_{n\in\mathcal{N}_T} \int_{I_n} \int_\Omega \boldsymbol{f}(u_h^n) \cdot \nabla\varphi^n - \mathcal{I}_h(\boldsymbol{f}(u_h^n)) \cdot \nabla\varphi_h^n \\ &= \sum_{n\in\mathcal{N}_T} \int_{I_n} \int_\Omega \Big(\boldsymbol{f}(u_h^n) - \mathcal{I}_h(\boldsymbol{f}(u_h^n))\Big) \cdot \nabla\varphi^n + \mathcal{I}_h(\boldsymbol{f}(u_h^n)) \cdot \nabla(\varphi^n - \varphi_h^n). \end{aligned}$$

The second term converges to zero owing to the stability of Lagrange interpolation and $\boldsymbol{f}(u_{h,\tau}) \in W^{1,\infty} \hookrightarrow C^0$. For the first term, we use the approximation properties of Lagrange interpolation to obtain

$$\Big|\int_\Omega \Big(\boldsymbol{f}(u_h^n) - \mathcal{I}_h(\boldsymbol{f}(u_h^n))\Big) \cdot \nabla\varphi^n\Big| \le \|\boldsymbol{f}(u_h^n) - \mathcal{I}_h(\boldsymbol{f}(u_h^n))\|_{L^2(\Omega)} \|\nabla\varphi^n\|_{L^2(\Omega)} \le C(h\|\nabla u_h^n\|_{L^2(\Omega)}^2)^{1/2} h^{1/2}.$$

Summing over $n \in \mathcal{N}_T$ gives

$$\begin{aligned} \sum_{n\in\mathcal{N}_T} \int_{I_n} \Big|\int_\Omega \Big(\boldsymbol{f}(u_h^n) - \mathcal{I}_h(\boldsymbol{f}(u_h^n))\Big) \cdot \nabla\varphi^n\Big| &\le C \sum_{n\in\mathcal{N}_T} \tau_n (h\|\nabla u_h^n\|_{L^2(\Omega)}^2)^{1/2} h^{1/2} \\ &\le h^{1/2} \Big(\sum_{n\in\mathcal{N}_T} \tau_n\Big)^{1/2} \Big(\sum_{n\in\mathcal{N}_T} \tau_n h \|\nabla u_h^n\|_{L^2(\Omega)}^2\Big)^{1/2} \le Ch^{1/2}. \end{aligned}$$

For the viscosities dissipation term $T_{3,h}$, the Cauchy-Schwarz inequality implies that:

$$\begin{aligned} |T_{3,h}| &\le \Big|\sum_{n\in\mathcal{N}_T} \tau_n b_h^n(u_h^n, \varphi_h^n)\Big| \\ &\le \sum_{n\in\mathcal{N}_T} \tau_n b_h^n(u_h^n, u_h^n)^{1/2} b_h^n(\varphi_h^n, \varphi_h^n)^{1/2} \\ &\le \Big(\sum_{n\in\mathcal{N}_T} \tau_n b_h^n(u_h^n, u_h^n) b_h^n(\varphi_h^n, \varphi_h^n)\Big)^{1/2} \Big(\sum_{n\in\mathcal{N}_T} \tau_n\Big)^{1/2}. \end{aligned}$$

To estimate $b_h^n(\varphi_h^n, \varphi_h^n)$, we use the mesh assumptions that imply $|\mathcal{A}_h| = O(h^{-d})$ and $d_{ij}^n \le Ch^{d-1}$. The regularity of φ implies $(\varphi_i^n - \varphi_j^n)^2 \le Ch^2$. So, we have $b_h^n(\varphi_h^n, \varphi_h^n) = \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n (\varphi_i^n - \varphi_j^n)^2 \le Ch^{-d} \cdot (h^{d-1}h^2) \le Ch$, which gives $|T_{3,h}| \le Ch^{1/2}$.

Collecting all the estimates gives $\tilde{T}_{1,h} + \tilde{T}_{2,h} \xrightarrow{h\to 0} 0$.

Since $\{u_{h,\tau}\}_h$ converges to u in L^1 owing to Assumption 3.2 and $\boldsymbol{f}$ is Lipschitz continuous, we have

$$\begin{aligned} &\Big|\int_0^T \int_\Omega (u_{h,\tau} - u)\partial_t\varphi\Big| \le C\|u_{h,\tau} - u\|_{L^1(\Omega\times(0,T))} \xrightarrow{h\to 0} 0, \\ &\Big|\int_0^T \int_\Omega \big(\boldsymbol{f}(u_{h,\tau}) - \boldsymbol{f}(u)\big) \cdot \nabla\varphi\Big| \le C\|\boldsymbol{f}(u_{h,\tau}) - \boldsymbol{f}(u)\|_{L^1(\Omega\times(0,T))} \le C\|u_{h,\tau} - u\|_{L^1(\Omega\times(0,T))} \xrightarrow{h\to 0} 0. \end{aligned}$$

Then, invoking $\tilde{T}_{1,h} + \tilde{T}_{2,h} \xrightarrow{h\to 0} 0$, we complete the proof.

□

4 Numerical analysis for hyperbolic systems

This section is the generalization of Section 3, where the scheme for hyperbolic systems is presented and analyzed. We follow the notation introduced in Section 3 when there is no ambiguity. Since the proof is similar to the one from Section 3, we do not give detailed proofs for most results.

4.1 Scheme

Definition 4.1. (Scheme) We denote the spatial approximation of $\boldsymbol{u}$ in the interval I_n as

$$\boldsymbol{u}_h^n(x) := \sum_{i\in\mathcal{A}_h} \boldsymbol{U}_i^n \phi_i(x),$$

for all $n \in \mathcal{N}_T$. The global approximation is defined as $\boldsymbol{u}_{h,\tau}(x,t)|_{I_n} := \boldsymbol{u}_h^n(x)$. The scheme is defined as follows:

$$m_i \frac{\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n}{\tau_n} + \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \Big(\mathbb{f}(\boldsymbol{U}_j^n)\boldsymbol{c}_{ij} + d_{ij}^n(\boldsymbol{U}_i^n - \boldsymbol{U}_j^n) \Big) = 0, \tag{22}$$

for all $i \in \mathcal{A}_h^0$ and all $n \in \mathcal{N}_T$. Moreover, the following CFL condition should be satisfied with a constant $\rho \in (0,1]$:

$$\tau_n \le \rho \min_{i\in\mathcal{A}_h^0} \frac{m_i}{2d_{ii}^n}. \tag{23}$$

As before, the initial data is approximated by the L^2-projection onto V_h^0. And boundary coefficients are set to zero.

4.2 Basic properties

Firstly, we can rewrite the scheme for all $i \in \mathcal{A}_h^0$ as follows:

$$\boldsymbol{U}_i^{n+1} = \sum_{j\in\mathcal{I}(i)} \Theta_{ij}^n \bar{\boldsymbol{U}}_{ij}^n, \tag{24}$$

where $\Theta_{ij}^n := \frac{2\tau_n d_{ij}^n}{m_i}$ for all $j \in \mathcal{I}(i)\backslash\{i\}$, $\Theta_{ii}^n := 1 - \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \Theta_{ij}^n$, and $\bar{\boldsymbol{U}}_{ii}^n := \boldsymbol{U}_i^n$,

$$\bar{\boldsymbol{U}}_{ij}^n := \frac{1}{2}(\boldsymbol{U}_i^n + \boldsymbol{U}_j^n) - (\mathbb{f}(\boldsymbol{U}_j^n) - \mathbb{f}(\boldsymbol{U}_i^n))\frac{\boldsymbol{c}_{ij}}{2d_{ij}^n}.$$

An important property is that $\bar{\boldsymbol{U}}_{ij}^n = \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{u}_{ij}^n(x, t_{ij}^n)dx$, where $\boldsymbol{u}_{ij}^n$ is the exact solution for Riemann problem with the initial data $(\boldsymbol{U}_i^n, \boldsymbol{U}_j^n)$, the normal vector $\boldsymbol{n}_{ij}$, and the artificial time $t_{ij}^n := \frac{\|c_{ij}\|_2}{2d_{ij}^n}$. Owing to the choice of d_{ij}^n, t_{ij}^n is sufficiently small so that $\bar{U}_{ij}^n \in \mathcal{B}$. Moreover, the CFL condition (23) gives $\Theta_{ij}^n \in [0,1]$ for all $j \in \mathcal{I}(i)$.

Lemma 4.1. *(Invariant-domain-preserving) Under CFL condition* (23)*, we have* $\boldsymbol{u}_{h,\tau} \in \mathcal{B}$*.*

Lemma 4.2. *(Discrete entropy inequality) Assume the CFL condition* (23) *is satisfied. Let* $(\eta, \boldsymbol{q})$ *be an entropy pair for* (5)*. Then, the following discrete entropy inequality holds true for all* $i \in \mathcal{A}_h^0$ *and all* $n \in \mathcal{N}_T$*:*

$$\frac{m_i}{\tau_n}(\eta(\boldsymbol{U}_i^{n+1}) - \eta(\boldsymbol{U}_i^n) + \int_\Omega \operatorname{div}(\mathcal{I}_h(\boldsymbol{q}(\boldsymbol{u}_h^n)))\phi_i + \sum_{j\in\mathcal{I}(i)} d_{ij}^n(\eta(\boldsymbol{U}_i^n) - \eta(\boldsymbol{U}_j^n)) \le 0. \tag{25}$$

4.3 Convergence analysis

Hyperbolic systems are slightly different from conservation laws, and these differences cause the problem for our analysis. Thus we need to add a few assumptions. Since the square entropy $\eta(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^2$ is not always relevant for hyperbolic systems, we need to add some hypotheses on the entropy for establishing the bound on dissipation:

Assumption 4.1. (Convexity of the entropy) The entropy satisfies the following properties:

$$\begin{aligned} &\eta \in C^2(\mathcal{B}; \mathbb{R}), \\ &\alpha\xi^2 \le \xi^T D_u^2\eta\xi \le \xi^2, \ \forall \xi \in \mathbb{R}^m, \end{aligned} \tag{26}$$

where $\alpha > 0$, and $D_u^2\eta$ is the Hessian matrix of η with respect to u. Note that the second bound in (26) is just a normalization of the entropy

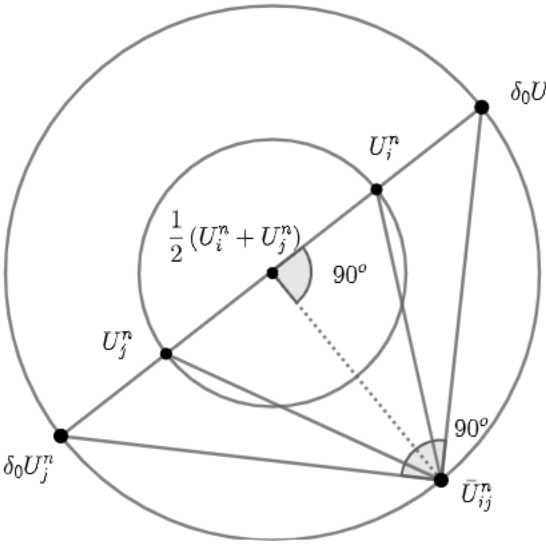


Figure 2: Illustration of Assumption 4.2

Moreover, since the locality of the invariant-domain-preserving property (local maximum principal in scalar case) cannot be guaranteed, we need one more assumption which guarantees quasi-locality:

Assumption 4.2. (Local invairiant-domain-preserving property) We assume there exists a constant $\delta_0 > 0$ such that $\bar{\boldsymbol{U}}_{ij}^n \in B(\frac{\boldsymbol{U}_i^n+\boldsymbol{U}_j^n}{2}, \frac{1}{2}\delta_0\|\boldsymbol{U}_i^n - \boldsymbol{U}_j^n\|_2)$ for all $i, j \in \mathcal{A}_h$ and all $n \in \mathcal{N}_T$, where $B(x_0, r)$ denotes the ball in $\mathbb{R}^m$ with center x_0 and radius r. An example is given in Figure 2.

Remark 4.1. (Bound of δ_0) In the following analysis, we assume without loss of generality $\delta_0 \geq 1$, since in scalar case, we have $\delta_0 = 1$.

Lemma 4.3. *(Estimate on temporal accumulation) Under the CFL condition* (23) *and Assumption 4.2, we have the following estimate for all $n \in \mathcal{N}_T$:*

$$\|\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n\|_{l_h^2}^2 \leq (1+\delta_0^2)\rho\tau_n b_h^n(\boldsymbol{u}_h^n, \boldsymbol{u}_h^n), \tag{27}$$

where $b_h^n(\boldsymbol{v}_h, \boldsymbol{w}_h) := \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n(\boldsymbol{V}_i - \boldsymbol{V}_j)^T(\boldsymbol{W}_i - \boldsymbol{W}_j)$, for all $\boldsymbol{v}_h, \boldsymbol{w}_h \in \left(V_h^0\right)^m$.

Proof. We use (24) and the CFL condition (23) to infer that

$$m_i(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n)^2 = m_i\Big(\sum_{j\in\mathcal{I}(i)} \Theta_{ij}^n(\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_i^n)\Big)^2 = \frac{4\tau_n^2}{m_i}\Big(\sum_{j\in\mathcal{I}(i)} d_{ij}^n(\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_i^n)\Big)^2 \leq \rho\tau_n\frac{2}{d_{ii}^n}\Big(\sum_{j\in\mathcal{I}(i)} d_{ij}^n(\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_i^n)\Big)^2.$$

The Cauchy-Schwarz inequality implies that

$$m_i(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n)^2 \leq 2\rho\tau_n \sum_{j\in\mathcal{I}(i)} d_{ij}^n(\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_i^n)^2.$$

Now, we take the sum over all $i \in \mathcal{A}_h$, notice that $\bar{\boldsymbol{U}}_{ij}^n = \bar{\boldsymbol{U}}_{ji}^n$ and $d_{ij}^n = d_{ji}^n$. This gives

$$\begin{aligned}\sum_{i\in\mathcal{A}_h} m_i(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n)^2 &\leq 2\rho\tau_n \sum_{i\in\mathcal{A}_h, j\in\mathcal{I}(i)} d_{ij}^n(\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_i^n)^2 \\ &= 2\rho\tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n\Big((\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_i^n)^2 + (\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_j^n)^2\Big).\end{aligned}$$

Then, we denote $\Phi(\boldsymbol{v}, \boldsymbol{a}, \boldsymbol{b}) := (\boldsymbol{v}-\boldsymbol{a})^2+(\boldsymbol{v}-\boldsymbol{b})^2$. Owing to Assumption 4.2, we have $\bar{\boldsymbol{U}}_{ij}^n \in B(\frac{\boldsymbol{U}_i^n+\boldsymbol{U}_j^n}{2}, \frac{1}{2}\delta_0\|\boldsymbol{U}_i^n-\boldsymbol{U}_j^n\|_2)$, so that $(\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_i^n)^2 + (\bar{\boldsymbol{U}}_{ij}^n - \boldsymbol{U}_j^n)^2 \leq \frac{1+\delta_0^2}{2}(\boldsymbol{U}_i^n - \boldsymbol{U}_j^n)^2$ as illustrated in Figure 2. This implies that $\Phi(\bar{\boldsymbol{U}}_{ij}^n, \boldsymbol{U}_i^n, \boldsymbol{U}_j^n) \leq \frac{1+\delta_0^2}{2}(\boldsymbol{U}_i^n - \boldsymbol{U}_j^n)^2$. Hence, we get

$$\sum_{i\in\mathcal{A}_h} m_i(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n)^2 \leq (1+\delta_0^2)\rho\tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n(\boldsymbol{U}_i^n - \boldsymbol{U}_j^n)^2.$$

This completes the proof. □

Assumption 4.3. (BV-like estimate) We assume that $\sum_{n\in\mathcal{N}_T} \tau_n h\|\nabla \boldsymbol{u}_h^n\|_{L^2(\Omega)}^2 \leq C$.

Lemma 4.4. *(Bound on dissipation) Assume that the CFL condition* (23) *holds true with* $\rho < \frac{2\alpha}{1+\delta_0^2}$ *and that Assumptions 4.1, 4.2 and 4.3 hold true. Thus, we have the following stability property:*

$$\sum_{i\in\mathcal{A}_h} m_i\eta_i^N + \big(2\alpha - \rho(1+\delta_0^2)\big) \sum_{n\in\mathcal{N}_T} \tau_n b_h^n(\boldsymbol{u}_h^n, \boldsymbol{u}_h^n) \leq \sum_{i\in\mathcal{A}_h} m_i\eta_i^0 + C, \tag{28}$$

where $\eta_i^n := \eta(\boldsymbol{U}_i^n)$.

Proof. We denote $\eta_h^n := \sum_{i\in\mathcal{A}_h} \eta_i^n\phi_i$, $D_u\eta_h^n := \sum_{i\in\mathcal{A}_h} D_u\eta_i^n\phi_i$, with $D_u\eta_i^n := D_u\eta(\boldsymbol{U}_i^n)$, and $\mathbb{f}_h^n := \sum_{i\in\mathcal{A}_h} \mathbb{f}(\boldsymbol{U}_i^n)\phi_i$. We multiply the scheme (22) by $2\tau_n D_u\eta_i^n$, sum over all $i \in \mathcal{A}_h$ and all $n \in \mathcal{N}_T$, and use the CFL condition (23) to obtain the expected bound. More precisely, we start by multiplying (22) by $2\tau_n D_u\eta_i^n$, notice that $(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n)D_u\eta_i^n = \eta_i^{n+1} - \eta_i^n - \frac{1}{2}(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n)^T\mathcal{H}_i^n(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n)$, where $\mathcal{H}_i^n := \int_0^1(1-t)D_u^2\eta(\boldsymbol{U}_i^n + t(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n))$. This gives

$$m_i\eta_i^{n+1} + 2\tau_n\int_\Omega \mathrm{div}\mathbb{f}_h^n \cdot \phi_i D_u\eta_i^n + 2\tau_n \sum_{j\in\mathcal{I}(i)} d_{ij}(U_i^n - U_j^n)D_u\eta_i^n \leq m_i\eta_i^n + m_i(U_i^{n+1} - U_i^n)^2,$$

Since $\|\mathcal{H}_i^n\|_2 \leq 1$ by (26) for all $i \in \mathcal{A}_h$ and all $n \in \mathcal{N}_T$. Then, we sum over all $i \in \mathcal{A}_h$ and all $n \in \mathcal{N}_T$ to get

$$\begin{aligned}
&\sum_{i\in\mathcal{A}_h} m_i\eta_i^N + 2\sum_{n\in\mathcal{N}_T}\tau_n\int_\Omega \mathrm{div}\mathbb{f}_h^n \cdot D_u\eta_h^n + 2\alpha\sum_{n\in\mathcal{N}_T}\tau_n\sum_{i\in\mathcal{A}_h, j<i} d_{ij}(\boldsymbol{U}_i^n - \boldsymbol{U}_j^n)^2 \\
&\leq \sum_{i\in\mathcal{A}_h} m_i\eta_i^N + 2\sum_{n\in\mathcal{N}_T}\tau_n\int_\Omega \mathrm{div}\mathbb{f}_h^n \cdot D_u\eta_h^n + 2\sum_{n\in\mathcal{N}_T}\tau_n\sum_{i\in\mathcal{A}_h, j<i} d_{ij}(\boldsymbol{U}_i^n - \boldsymbol{U}_j^n)(D_u\eta_i^n - D_u\eta_j^n) \\
&\leq \sum_{i\in\mathcal{A}_h} m_i\eta_i^0 + \sum_{n\in\mathcal{N}_T}\|\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n\|_{l_h^2}^2.
\end{aligned}\tag{29}$$

We first estimate $\int_\Omega \mathrm{div}\mathbb{f}_h^n D_u\eta_h^n$. The entropy relation $D_u\eta D_u\mathbb{f} = D_u\boldsymbol{q}$ plays an important role. We add and remove some terms to get

$$\begin{aligned}
\mathrm{div}\mathbb{f}_h^n \cdot D_u\eta_h^n = {}&\mathrm{div}(\mathbb{f}(\boldsymbol{u}_h^n) - \mathbb{f}_h^n) \cdot (D_u\eta_h^n - D_u\eta(\boldsymbol{u}_h^n)) \\
&+ \mathrm{div}(\mathbb{f}(\boldsymbol{u}_h^n) - \mathbb{f}_h^n) \cdot D_u\eta(\boldsymbol{u}_h^n) + \mathrm{div}\mathbb{f}(\boldsymbol{u}_h^n) \cdot (D_u\eta_h^n - D_u\eta(\boldsymbol{u}_h^n)) \\
&+ \mathrm{div}\mathbb{f}(\boldsymbol{u}_h^n) \cdot D_u\eta(\boldsymbol{u}_h^n)
\end{aligned}$$

Noticing that $\boldsymbol{u}_h^n|_{\partial\Omega} = 0$ and using the Stokes formula, we get

$$\int_\Omega \mathrm{div}\mathbb{f}(\boldsymbol{u}_h^n) \cdot D_u\eta(\boldsymbol{u}_h^n) = \int_\Omega \mathrm{div}\boldsymbol{q}(\boldsymbol{u}_h^n) = \int_{\partial\Omega} \boldsymbol{q}(\boldsymbol{u}_h^n)n_\Omega = \int_{\partial\Omega} \boldsymbol{q}(0)n_\Omega = \int_\Omega \mathrm{div}\boldsymbol{q}(0) = 0.$$

Moreover, integrating by parts and using the approximation properties of Lagrange interpolation, we infer that

$$\begin{aligned}
&\Big|\int_\Omega \mathrm{div}(\mathbb{f}(\boldsymbol{u}_h^n) - \mathbb{f}_h^n) \cdot (D_u\eta_h^n - D_u\eta(\boldsymbol{u}_h^n))\Big| + \Big|\int_\Omega \mathrm{div}(\mathbb{f}(\boldsymbol{u}_h^n) - \mathbb{f}_h^n) \cdot D_u\eta(\boldsymbol{u}_h^n)\Big| + \Big|\int_\Omega \mathrm{div}\mathbb{f}(\boldsymbol{u}_h^n) \cdot (D_u\eta_h^n - D_u\eta(\boldsymbol{u}_h^n))\Big| \\
&\leq Ch\|\nabla\boldsymbol{u}_h^n\|_{L^2(\Omega)}^2.
\end{aligned}$$

Thus, we have

$$2\sum_{n\in\mathcal{N}_T}\tau_n\Big|\int_\Omega \mathrm{div}\mathbb{f}_h^n\cdot D_u\eta_h^n\Big|\le C\sum_{n\in\mathcal{N}_T}\tau_n h\|\nabla \boldsymbol{u}_h^n\|^2_{L^2(\Omega)}. \tag{30}$$

For the last term on the right-hand side of (29), we use (27) to infer that

$$\sum_{n\in\mathcal{N}_T}\|\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n\|^2_{l_h^2}\le(1+\delta_0^2)\rho\sum_{n\in\mathcal{N}_T}\tau_n b_h^n(\boldsymbol{u}_h^n,\boldsymbol{u}_h^n). \tag{31}$$

After obtaining the estimates (30) and (31), we put them into (29) and infer that

$$\big(2\alpha-\rho(1+\delta_0^2)\big)\sum_{n\in\mathcal{N}_T}\tau_n\sum_{i\in\mathcal{A}_h,j<i}d_{ij}^n(\boldsymbol{U}_i^n-\boldsymbol{U}_j^n)^2\le\sum_{i\in\mathcal{A}_h}m_i\eta_i^0-\sum_{i\in\mathcal{A}_h}m_i\eta_i^N+C\sum_{n\in\mathcal{N}_T}\tau_n h\|\nabla\boldsymbol{u}_h^n\|^2_{L^2(\Omega)}.$$

We conclude using Assumption 4.3.

□

Assumption 4.4. (Upper bound of maximum wave speed) We assume that $\lambda_{max}(\boldsymbol{U}_i^n,\boldsymbol{U}_j^n,\boldsymbol{n}_{ij})$ is uniformly bounded from above by a constant for all $i,j\in\mathcal{A}_h$ and for all $n\in\mathcal{N}_T$.

Assumption 4.5. (Convergence in L^1) We assume that the sequence of numerical solutions $\{\boldsymbol{u}_{h,\tau}\}_h$ has a subsequence (still denoted by $\{\boldsymbol{u}_{h,\tau}\}_h$) which converges strongly to a function $\boldsymbol{u}\in L^\infty(\Omega\times(0,T);\mathbb{R}^m)$ in L^1, i.e., that $\|\boldsymbol{u}_{h,\tau}-\boldsymbol{u}\|_{L^1(\Omega\times(0,T);\mathbb{R}^m)}\xrightarrow{h\to 0}0$.

Theorem 4.1. *(Convergence of the scheme) Assume that the CFL condition* (23) *holds with $\rho<\frac{2\alpha}{1+\delta_0}$, and that Assumptions 4.1, 4.2, 4.3, 4.4 and 4.5 hold true. Then, the limit of the sequence of numerical solutions (up to a subsequence) is a weak solution of* (5)*, i.e., for all $\boldsymbol{\varphi}\in C_c^\infty(\Omega\times[0,T);\mathbb{R}^m)$, we have*

$$\int_0^T\int_\Omega \boldsymbol{u}\partial_t\boldsymbol{\varphi}+\int_0^T\int_\Omega\mathbb{f}(\boldsymbol{u}):\nabla\boldsymbol{\varphi}+\int_\Omega\boldsymbol{u}_0\boldsymbol{\varphi}(x,0)=0.$$

Proof. The proof is the same as in Section 3, the only difference is that, in Section 3, we have $d_{ij}^n\le Ch^{d-1}$, but for hyperbolic systems, we do not know an a priori upper bound of maximum wave speed, so we need Assumptions 4.4 to ensure that $d_{ij}^n\le C\|\boldsymbol{c}_{ij}\|_2\le Ch^{d-1}$. □

5 Further analysis

Since the analysis for conservation laws is similar to the analysis for hyperbolic systems, we focus in this section on hyperbolic systems.

5.1 Entropy inequality

In this subsection, we prove that the sequence of numerical solutions (up to a subsequence) converges to an entropy solution, i.e., the entropy inequality (7) is satisfied by the limit. The argument for proving this result is similar to the proof of Theorem 3.1.

We use the scheme (22) for this subsection.

Theorem 5.1. *(Entropy inequality) Assume the CFL condition* (23) *holds with $\rho<\frac{2\alpha}{1+\delta_0^2}$, and Assumptions 4.1, 4.2, 4.3, 4.4 and 4.5 hold. Then, the limit of our numerical solution (up to a subsequence) is the entropy solution of* (5)*, i.e., for all $\varphi\in C_c^\infty(\Omega\times[0,T);\mathbb{R}_+)$, we have*

$$-\int_0^T\int_\Omega\eta(\boldsymbol{u})\partial_t\boldsymbol{\varphi}-\int_0^T\int_\Omega\boldsymbol{q}(\boldsymbol{u})\cdot\nabla\boldsymbol{\varphi}-\int_\Omega\eta(\boldsymbol{u}_0)\boldsymbol{\varphi}(x,0)\le 0.$$

Proof. We follow the same arguments as in the proof of Theorem 3.1, with slight modifications. We multiply the discrete entropy inequality (25) by $\varphi_i^n \geq 0$ and sum over all $i \in \mathcal{A}_h$ and all $n \in \mathcal{N}_T$ to get

$$E_{1,h} + E_{2,h} + E_{3,h} \leq 0, \tag{32}$$

where

$$\begin{aligned}
E_{1,h} &:= \sum_{n\in\mathcal{N}_T} \sum_{i\in\mathcal{A}_h} m_i(\eta_i^{n+1} - \eta_i^n)\varphi_i^n, \\
E_{2,h} &:= \sum_{n\in\mathcal{N}_T} \tau_n \int_\Omega \mathrm{div}\mathcal{I}_h(\boldsymbol{q}(\boldsymbol{u}_h^n))\varphi_h^n, \\
E_{3,h} &:= \sum_{n\in\mathcal{N}_T} \tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}^n(\eta_i^n - \eta_j^n)(\varphi_i^n - \varphi_j^n).
\end{aligned}$$

Similarly, we introduce two terms:

$$\begin{aligned}
\tilde{E}_{1,h} &:= -\int_0^T \int_\Omega \eta(\boldsymbol{u}_h^n)\partial_t\varphi - \int_\Omega \eta(\boldsymbol{u}_0)\varphi(x,0), \\
\tilde{E}_{2,h} &:= -\int_0^T \int_\Omega \boldsymbol{q}(\boldsymbol{u}_h^n)\cdot\nabla\varphi.
\end{aligned}$$

We will prove that $E_{1,h} - \tilde{E}_{1,h} \xrightarrow{h\to 0} 0$, $E_{2,h} - \tilde{E}_{2,h} \xrightarrow{h\to 0} 0$ and $E_{3,h} \xrightarrow{h\to 0} 0$. Then, passing to the limit in (32) gives $\lim_{h\to 0} \tilde{E}_{1,h} + \tilde{E}_{2,h} \leq 0$. After that, we decompose the left-hand side of (7) into two parts:

$$\begin{aligned}
E_{1,0} &:= -\int_0^T \int_\Omega \eta(\boldsymbol{u})\partial_t\varphi - \int_\Omega \eta(\boldsymbol{u}_0)\varphi(x,0), \\
E_{2,0} &:= -\int_0^T \int_\Omega \boldsymbol{q}(\boldsymbol{u})\cdot\nabla\varphi.
\end{aligned}$$

Then, using Assumption 4.5, we conclude by proving $\tilde{E}_{1,h} \xrightarrow{h\to 0} E_{1,0}$ and $\tilde{E}_{2,h} \xrightarrow{h\to 0} E_{2,0}$.

We start with $E_{1,h} - \tilde{E}_{1,h}$. The strategy is slightly different from Theorem 3.1. We add and remove the Lagrange interpolant η_h^n of $\eta(\boldsymbol{u}_h^n)$ in $\tilde{E}_{1,h}$ and estimate the errors. In details, adding and removing η_h^n in $\tilde{E}_{1,h}$ gives

$$\tilde{E}_{1,h} = -\int_0^T \int_\Omega \big(\eta(\boldsymbol{u}_h^n) - \eta_h^n\big)\partial_t\varphi - \int_0^T \int_\Omega \eta_h^n\partial_t\varphi - \int_\Omega \eta(\boldsymbol{u}_0)\varphi(x,0) := \tilde{E}_{1,h}^{error} + \tilde{E}_{1,h}^{Lagrange},$$

with obvious notation. The first term $\tilde{E}_{1,h}^{error}$ converges to zero since

$$\Big|\int_0^T \int_\Omega \big(\eta(\boldsymbol{u}_h^n) - \eta_h^n\big)\partial_t\varphi\Big| \leq C \sum_{n\in\mathcal{N}_T} \tau_n h\|\nabla\boldsymbol{u}_h^n\|_{L^2(\Omega)} \leq Ch^{1/2}\Big(\sum_{n\in\mathcal{N}_T} \tau_n h\|\nabla\boldsymbol{u}_h^n\|_{L^2(\Omega)}^2\Big)^{1/2}\Big(\sum_{n\in\mathcal{N}_T}\tau_n\Big)^{1/2} \leq Ch^{1/2}.$$

We integrate by parts in time for $\tilde{E}_{1,h}^{Lagrange}$ and infer that

$$\tilde{E}_{1,h}^{Lagrange} = \sum_{n\in\mathcal{N}_T} \sum_{i\in\mathcal{A}_h} (\eta_i^{n+1} - \eta_i^n)\int_{\omega_i} \phi_i\varphi^n + \int_\Omega \big(\eta(\boldsymbol{u}_h^0) - \eta(\boldsymbol{u}_0)\big)\varphi^0.$$

So,

$$E_{1,h} - \tilde{E}_{1,h}^{Lagrange} = \sum_{n\in\mathcal{N}_T} \sum_{i\in\mathcal{A}_h} (\eta_i^{n+1} - \eta_i^n)\int_{\omega_i} \phi_i(\varphi_i^n - \varphi^n) - \int_\Omega \big(\eta(\boldsymbol{u}_h^0) - \eta(\boldsymbol{u}_0)\big)\varphi^0.$$

The approximation properties of L^2-projection ensure the second term tends to zero:

$$\Big|\int_\Omega \big(\eta(\boldsymbol{u}_h^0) - \eta(\boldsymbol{u}_0)\big)\varphi^0\Big| \leq C\|\eta(\boldsymbol{u}_h^0) - \eta(\boldsymbol{u}_0)\|_{L^2(\Omega)} \leq C\|\boldsymbol{u}_h^0 - \boldsymbol{u}_0\|_{L^2(\Omega)} \xrightarrow{h\to 0} 0,$$

where the limit result is from a classical density argument. For the first term, using the Lipschitz continuity of η and following the same reasoning as in the proof of Theorem 3.1 gives

$$\Big|\sum_{n\in\mathcal{N}_T}\sum_{i\in\mathcal{A}_h}(\eta_i^{n+1}-\eta_i^n)\int_{\omega_i}\phi_i(\varphi_i^n-\varphi^n)\Big|\le C\sum_{n\in\mathcal{N}_T}\sum_{i\in\mathcal{A}_h}|U_i^{n+1}-U_i^n|\int_{\omega_i}|\phi_i(\varphi_i^n-\varphi^n)|\le Ch^{1/2}.$$

To estimate $E_{2,h}-\tilde{E}_{2,h}$, the idea is to use Assumption 4.3. More precisely, we have

$$\begin{aligned}E_{2,h}-\tilde{E}_{2,h}&=\sum_{n\in\mathcal{N}_T}\int_{I_n}\int_{\Omega}\boldsymbol{q}(\boldsymbol{u}_h^n)\cdot\nabla\varphi^n-\mathcal{I}_h(\boldsymbol{q}(\boldsymbol{u}_h^n))\cdot\nabla\varphi_h^n\\&=\sum_{n\in\mathcal{N}_T}\int_{I_n}\int_{\Omega}\Big(\boldsymbol{q}(\boldsymbol{u}_h^n)-\mathcal{I}_h(\boldsymbol{q}(\boldsymbol{u}_h^n))\Big)\cdot\nabla\varphi^n+\mathcal{I}_h(\boldsymbol{q}(\boldsymbol{u}_h^n))\cdot\nabla(\varphi^n-\varphi_h^n).\end{aligned}$$

Noticing that $\|\boldsymbol{q}(\boldsymbol{u}_h^n)-\mathcal{I}_h(\boldsymbol{q}(\boldsymbol{u}_h^n))\|_{L^2(\Omega)}\le Ch\|\nabla\boldsymbol{u}_h^n\|_{L^2(\Omega)}$, the same reasoning as in the proof of Theorem 3.1 gives $|E_{2,h}-\tilde{E}_{2,h}|\le Ch^{1/2}$.

For the viscosities dissipation term $E_{3,h}$, noticing that $(\eta_i^n-\eta_j^n)^2\le C(\boldsymbol{U}_i^n-\boldsymbol{U}_j^n)^2$, the Cauchy-Schwarz inequality gives:

$$\begin{aligned}|E_{3,h}|&\le C\Big|\sum_{n\in\mathcal{N}_T}\tau_n b_h^n(\eta_h^n,\varphi_h^n)\Big|\\&\le C\sum_{n\in\mathcal{N}_T}\tau_n b_h^n(u_h^n,u_h^n)^{1/2}b_h^n(\varphi_h^n,\varphi_h^n)^{1/2}\\&\le C\Big(\sum_{n\in\mathcal{N}_T}\tau_n b_h^n(u_h^n,u_h^n)b_h^n(\varphi_h^n,\varphi_h^n)\Big)^{1/2}\Big(\sum_{n\in\mathcal{N}_T}\tau_n\Big)^{1/2}\\&\le Ch^{1/2},\end{aligned}$$

since $b_h^n(\varphi_h^n,\varphi_h^n)\le Ch$ as discussed in the proof of Theorem 3.1. Collecting all the estimates gives $\lim_{h\to0}\tilde{E}_{1,h}+\tilde{E}_{2,h}\le 0$.

Since $\{\boldsymbol{u}_h^n\}_h$ converges to $\boldsymbol{u}$ in L^1 owing to Assumption 4.5, we have,

$$\begin{aligned}&\Big|\int_0^T\int_\Omega\big(\eta(\boldsymbol{u}_h^n)-\eta(\boldsymbol{u})\big)\partial_t\varphi\Big|\le C\|\boldsymbol{u}_h^n-\boldsymbol{u}\|_{L^1(\Omega\times(0,T))}\xrightarrow{h\to0}0,\\&\Big|\int_0^T\int_\Omega(\boldsymbol{q}(\boldsymbol{u}_h^n)-\boldsymbol{q}(\boldsymbol{u}))\cdot\nabla\varphi\Big|\le C\|\boldsymbol{q}(\boldsymbol{u}_h^n)-\boldsymbol{q}(\boldsymbol{u})\|_{L^1(\Omega\times(0,T))}\le C\|\boldsymbol{u}_h^n-\boldsymbol{u}\|_{L^1(\Omega\times(0,T))}\xrightarrow{h\to0}0.\end{aligned}$$

Then, invoking $\lim_{h\to0}\tilde{E}_{1,h}+\tilde{E}_{2,h}\le0$, we complete the proof.

□

5.2 Analysis without sonic points

In this subsection, we want to remove the assumptions on $\sum_{n\in\mathcal{N}_T}\tau_n h\|\nabla\boldsymbol{u}_h^n\|^2_{L^2(\Omega)}$, and the assumption on the Riemann average, i.e., Assumptions 4.3 and 4.2. This can be realized by assuming that there are no sonic points, i.e., assuming that the upper bound on the maximum wave speed is uniformly bounded from below.

Definition 5.1. (scheme) We use the notation from the previous section. The scheme should be modified slightly as follows:

$$m_i\frac{\boldsymbol{U}_i^{n+1}-\boldsymbol{U}_i^n}{\tau_n}+\sum_{j\in\mathcal{I}(i)\setminus\{i\}}\Big(\mathbb{f}(U_j^n)\boldsymbol{c}_{ij}+\kappa d_{ij}^n(\boldsymbol{U}_i^n-\boldsymbol{U}_j^n)\Big)=0,\tag{33}$$

with κ a sufficiently large constant precised in Lemma 5.6. Moreover, the following CFL condition should hold:

$$\tau_n\le\rho\min_{i\in\mathcal{A}_h^0}\frac{m_i}{2\kappa d_{ii}^n},\tag{34}$$

with a given constant $\rho\in(0,1]$.

Lemma 5.1. *(Convex combination) We can rewrite the scheme as follows:*

$$\boldsymbol{U}_i^{n+1} = \sum_{j\in\mathcal{I}(i)} \Theta_{ij}^n \bar{\boldsymbol{U}}_{ij}^n, \tag{35}$$

where $\Theta_{ij}^n := \frac{2\tau_n \kappa d_{ij}^n}{m_i}$ *for all* $j \in \mathcal{I}(i)\backslash\{i\}$ *and* $\Theta_{ii}^n := 1 - \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \Theta_{ij}^n$. *By the same argument as in the previous section, all the coefficients are in* $[0,1]$, *and the Riemann average* $\bar{\boldsymbol{U}}_{ij}^n := \frac{1}{2}(\boldsymbol{U}_i^n + \boldsymbol{U}_j^n) - \left(\mathbb{f}(\boldsymbol{U}_j^n) - \mathbb{f}(\boldsymbol{U}_i^n)\right)\frac{\boldsymbol{c}_{ij}}{2\kappa d_{ij}^n}$ *takes values in the invariant set* $\mathcal{B}$.

Assumption 5.1. (Boundedness of maximum wave speed) We assume that $\lambda_{max}(\boldsymbol{U}_i^n, \boldsymbol{U}_j^n, \boldsymbol{n}_{ij})$ is uniformly bounded from below and above for all $i, j \in \mathcal{A}_h$ and all $n \in \mathcal{N}_T$, i.e., there exists a constant C_λ such that

$$C_\lambda^{-1} L(\mathbb{f}) \leq \lambda_{max}(\boldsymbol{U}_i^n, \boldsymbol{U}_j^n, \boldsymbol{n}_{ij}) \leq C_\lambda L(\mathbb{f}), \tag{36}$$

where $L(\mathbb{f})$ is the Lipschitz constant of $\mathbb{f}$.

Lemma 5.2. *(Boundedness of Riemann average) Under Assumption 5.1, the Riemann average stays in a ball with center* $\frac{\boldsymbol{U}_i^n+\boldsymbol{U}_j^n}{2}$ *and radius* $\frac{C_\lambda}{2\kappa}\|\boldsymbol{U}_i^n - \boldsymbol{U}_j^n\|_2$ *for all* $i, j \in \mathcal{A}_h$ *and all* $n \in \mathcal{N}_T$.

Proof. Owing to Assumption 5.1, we have $d_{ij}^n \geq C_\lambda^{-1} L(\mathbb{f})\|\boldsymbol{c}_{ij}\|_2$. Hence,

$$\left\|\left(\mathbb{f}(\boldsymbol{U}_j^n) - \mathbb{f}(\boldsymbol{U}_i^n)\right)\frac{\boldsymbol{c}_{ij}}{2\kappa d_{ij}^n}\right\|_2 \leq \|\mathbb{f}(\boldsymbol{U}_j^n) - \mathbb{f}(\boldsymbol{U}_i^n)\|_2 \frac{\|\boldsymbol{c}_{ij}\|_2}{2\kappa C_\lambda^{-1} L(\mathbb{f})\|\boldsymbol{c}_{ij}\|_2} \leq \frac{C_\lambda}{2\kappa}\|\boldsymbol{U}_i^n - \boldsymbol{U}_j^n\|_2.$$

□

Lemma 5.3. *(estimate on temporal accumulation) Under the CFL condition* (34) *and Assumption 5.1, we have the following estimate for all* $n \in \mathcal{N}_T$*:*

$$\|\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n\|_{l_h^2}^2 \leq (\kappa + \frac{C_\lambda^2}{\kappa})\rho\tau_n b_h^n(\boldsymbol{u}_h^n, \boldsymbol{u}_h^n). \tag{37}$$

Proof. The proof is the same as the proof in Section 4, the only differences are that we replace δ_0 in (27) by $\frac{C_\lambda}{\kappa}$ and that the definition of the CFL condition is changed. □

Lemma 5.4. *(Norm equivalence) Under Assumption 5.1, there exists a constant* $C_1 > 0$ *such that*

$$(C_1)^{-1} L(\mathbb{f}) b_h^n(\boldsymbol{v}_h, \boldsymbol{v}_h) \leq h\|\nabla \boldsymbol{v}_h\|_{L^2(\Omega)}^2 \leq C_1 L(\mathbb{f}) b_h^n(\boldsymbol{v}_h, \boldsymbol{v}_h), \tag{38}$$

for all $\boldsymbol{v}_h \in \left(V_h^0\right)^m$.

Proof. The proof follows the idea of [14]. We take two functions $\boldsymbol{v}_h = \sum_{i\in\mathcal{A}_h} \boldsymbol{V}_i\phi_i, \boldsymbol{w}_h = \sum_{i\in\mathcal{A}_h} \boldsymbol{W}_i\phi_i \in \left(V_h^0\right)^m$, and define the local bilinear form on each cell K:

$$b_K(\boldsymbol{v}_h, \boldsymbol{w}_h) = \frac{1}{2}\sum_{i,j\in\mathcal{I}(K)} d_{ij}^n (\boldsymbol{V}_i - \boldsymbol{V}_j)^T(\boldsymbol{W}_i - \boldsymbol{W}_j).$$

Up to the change of variable $\hat{\boldsymbol{v}}_h|_{\hat{K}} = \boldsymbol{v}_h|_K \circ \Phi_K$, the definition of b_K, Assumption 5.1 and $\|\boldsymbol{c}_{ij}\|_2 = O(h^{d-1})$ imply that $\left(\frac{h}{m(K)} b_K(\cdot,\cdot)\right)^{\frac{1}{2}}$ is a norm on $\hat{P}/\mathbb{R}$. Since all norms are equivalent on $\hat{P}/\mathbb{R}$, we infer that

$$CL(\mathbb{f})\|\nabla\hat{\boldsymbol{v}}_h\|_{L^2(\hat{K})}^2 \leq \frac{h}{m(K)} b_K(\boldsymbol{v}_h, \boldsymbol{v}_h) \leq CL(\mathbb{f})\|\nabla\hat{\boldsymbol{v}}_h\|_{L^2(\hat{K})}^2.$$

After using the change of variable $\boldsymbol{v}_h = \hat{\boldsymbol{v}}_h \circ \Phi_K^{-1}$ and $Ch^d \leq m_i \leq Ch^d$ for all $i \in \mathcal{A}_h$, we infer that

$$CL(\mathbb{f})h\|\nabla \boldsymbol{v}_h\|_{L^2(K)}^2 \leq b_K(\boldsymbol{v}_h, \boldsymbol{v}_h) \leq CL(\mathbb{f})h\|\nabla \boldsymbol{v}_h\|_{L^2(K)}^2.$$

Since the pairs $(i,j) \in \mathcal{A}_h \times \mathcal{A}_h$ are counted twice, when we sum over all cells, we get

$$b_h^n(\boldsymbol{v}_h, \boldsymbol{v}_h) = \frac{1}{2} \sum_K b_K(\boldsymbol{v}_h, \boldsymbol{v}_h).$$

This concludes the proof. □

Lemma 5.5. *(BV-like estimate) Under Assumption 5.1, we have* $h\|\nabla \boldsymbol{u}_h^n\|^2_{L^2(\Omega)} \leq C_1 L(\mathbb{f}) b_h^n(\boldsymbol{u}_h^n, \boldsymbol{u}_h^n)$.

Proof. This is the direct corollary of the previous lemma. □

Lemma 5.6. *(Bound on dissipation) Under the CFL condition* (34) *with* $\rho < 2\alpha$ *and Assumptions 4.1 and 5.1, we have the following stability property by choosing* κ *sufficiently large so that* $(2\alpha - \rho)\kappa - \frac{\rho C_\lambda^2}{\kappa} - C_0 > 0$ *and* $\kappa \geq 1$.

$$\sum_{i\in\mathcal{A}_h} m_i \eta_i^N + \left((2\alpha - \rho)\kappa - \frac{\rho C_\lambda^2}{\kappa} - C_0\right) \sum_{n\in\mathcal{N}_T} \tau_n b_h^n(\boldsymbol{u}_h^n, \boldsymbol{u}_h^n) \leq \sum_{i\in\mathcal{A}_h} m_i \eta_i^0, \tag{39}$$

where C_0 *is a constant depending on the approximation properties of* P_1 *Lagrange finite element,* $L(\mathbb{f})$*,* C_1 *and the final time* T.

Proof. The argument is the same as in Section 22 with the following modifications:

(i) Replacing $2\alpha \sum_{n\in\mathcal{N}_T} \tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}(\boldsymbol{U}_i^n - \boldsymbol{U}_j^n)^2$ by $2\kappa\alpha \sum_{n\in\mathcal{N}_T} \tau_n \sum_{i\in\mathcal{A}_h, j<i} d_{ij}(\boldsymbol{U}_i^n - \boldsymbol{U}_j^n)^2$ in (29).

(ii) Replacing $\sum_{n\in\mathcal{N}_T} \tau_n h \|\nabla \boldsymbol{u}_h^n\|^2_{L^2(\Omega)}$ by $C_0 \sum_{n\in\mathcal{N}_T} \tau_n b_h^n(\boldsymbol{u}_h^n, \boldsymbol{u}_h^n)$ in (30).

(iii) Replacing $1 + \delta_0^2$ by $\kappa + \frac{C_\lambda^2}{\kappa}$ in (31).

Then, a similar reasoning proves the present lemma. □

After obtaining the above bound on dissipation, we can get the convergence result as in the previous section.

6 Numerical experiments

In this section, we numerically illustrate the first-order invariant-domain-preserving method on conservation laws and hyperbolic systems. Since the scheme defined in the previous sections only considers zero boundary conditions, we introduce here a scheme that accounts for nonzero the boundary conditions. We denote $\Gamma_{in} \subset \partial\Omega$ a part of the boundary where the boundary condition is enforced at the PDE level. For conservation laws, it is typically the problem-dependent inflow part, i.e., $\Gamma_{in} := \{x \in \partial\Omega : \boldsymbol{f}(u(x,t)) \cdot \boldsymbol{n} < 0\}$. The set $\mathcal{A}_h^{\Gamma_{in}}$ is defined as the set of degrees of freedom lying on Γ_{in}, and $\mathcal{A}_h^o := \mathcal{A}_h \backslash \mathcal{A}_h^{\Gamma_{in}}$. Then we compute U_i^{n+1} by

$$m_i \frac{U_i^{n+1} - U_i^n}{\tau_n} + \sum_{j\in\mathcal{I}(i)\backslash\{i\}} \left(\boldsymbol{f}(U_j^n)\boldsymbol{c}_{ij} + d_{ij}^n(U_i^n - U_j^n)\right) = 0,$$

for all $i \in \mathcal{A}_h^o$ and all $n \in \mathcal{N}_T$. The boundary condition is strongly enforced on U_i^{n+1}. For hyperbolic systems, the boundary condition may be enforced only on some components of $\boldsymbol{U}_i^{n+1}$. The L^1- and L^2-relative errors are estimated at the final time T for all the examples. In all the tests, the upper bound on the maximum wave speed is set to a constant value for all pairs $(U_i^n, U_j^n) \in \mathcal{B} \times \mathcal{B}$ for simplicity. Provided this upper bound is large enough, the invariant-domain-preserving property is satisfied.

6.1 Details about code

I implemented a Python program to realize the scheme by following Zhaonan Dong's idea. The whole code is roughly divided into five parts: mesh reading and treatment, numerical integration, matrix computation, schema implementation, and error estimate.

The data format follows the idea of Zhaonan Dong. The program can read a file in .mat format, and it can convert the mesh information to Python's data of type 'Ndarray', and store it in a class named 'Geometry'. In addition to reading data from .mat files, this class also provides functions for calculating useful geometric information in various schemes.

For numerical integration, I wrote the bases of finite elements of the Lagrange type of order 1, 2 and 3 on a reference element. The realization of numerical integration uses the tensor product of the point of Gauss and the point of Legendre on the rectangular unit, then apply a linear transformation between the rectangle and the triangle for computing the integration on the triangle element. For various differential operators, I have calculated the exact first and second order differentials of each base. In addition to the Lagrange bases, the first to third order Bernstein bases were also implemented by a linear transformation from Lagrange basis to Bernstein basis.

In the matrix calculation part, the matrices are calculated in parallel. This includes the stiffness matrix of the Poisson problem, the mass matrix and the matrices $\boldsymbol{c}_{ij}$ and d_{ij}^n used in the implementation of the scheme. In the calculation of d_{ij}^n, in order to simplify the problem, I only consider the case where the upper bound of the maximum wave speed is a constant. In all calculations, the numerical integral on each mesh is parallel, but the final assembling process is not in parallel.

The implementation part of the scheme mainly contains the following parts: implementation of the initial condition, calculation of the CFL condition, time evolution and boundary conditions. The calculation of the initial condition uses the L^2-projection, i.e., an approximation of the initial condition is obtained by inverting a mass matrix. The CFL condition calculation is computed by using the information from inflow part. There are two ways to update coefficients: one is by matrix-vector multiplication (scalar case) and matrix-matrix multiplication (system case), the advantage of this method is that the 'scipy' module optimizes these algebraic operators. Another method is to use the algebraic expressions given by the scheme and the locality of stencil. As the update of each degree of freedom is only related to its stencil, we only use the local information for the coefficient updates. The advantage of this method is that its parallel implementation is much simpler than matrix multiplication. The imposition of boundary conditions depends on the problem. In the scalar case, I implemented three different methods: imposing in the strong sense, in the weak sense, and by solving the Riemann problem. The first method consists in imposing the boundary conditions directly on the degree of freedom on the inflow part by using the nodal value (if the boundary data is sufficiently smooth). The second method is achieved by adding a surface integral (on inflow part) on the left and right side of the scheme. The last method is to determine the coefficients by solving a Riemann problem in which the left state is the updated value calculated by the scheme, and the right state is the nodal value of the boundary data. In the numerical experiments, their results are very similar, so in the report I only present the experiments by strongly imposing boundary condition. In the system case, I only implemented the first method, because the boundary conditions of hyperbolic systems are more complicated and depend on the problem.

The error estimate is realized by the numerical integration of an order higher than the polynomial order of the finite elements, I implemented the estimate for L^1, L^2 and H^1.

6.2 Convergence tests with smooth solutions

We illustrate the method by solving model problems with smooth solutions in 1D and 2D.

6.2.1 Linear transport in 1D

We consider the model problem

$$u_t + u_x = 0,$$

in the domain $\Omega := (-1, 1)$, with the initial condition $u_0(x) = u_{ex}(x, 0)$, the inflow boundary condition $u(-1, t) = t + 1$, and the exact solution $u_{ex}(x, t) := t - x$. The upper bound on the maximum wave speed is set to $\lambda_{max} := 1$. The final time is $T := 1$.

We show in Table 1 the relative errors in the L^1- and the L^2-norms. We observe from the table a super-convergence phenomenon, although the time discretization is only first-order accurate.

Table 1: Linear transport, 1D

	L^1		L^2	
DoFs	error	rate	error	rate
21	2.50E-03	-	9.12E-03	-
41	6.25E-04	2.00	3.23E-03	1.50
81	1.56E-04	2.00	1.14E-03	1.50
161	3.91E-05	2.00	4.03E-04	1.50
321	9.77E-06	2.00	1.43E-04	1.50

Table 2: Linear transport, 2D

	L^1		L^2	
DoFs	error	rate	error	rate
21	3.73E-01	-	9.12E-01	-
41	2.10E-01	0.98	3.23E-01	0.90
81	1.14E-01	0.96	1.14E-01	0.90
161	6.00E-02	0.97	4.03E-02	0.93
321	3.08E-02	0.98	1.43E-02	0.96

6.2.2 Linear transport in 2D

We consider the model problem

$$u_t + \operatorname{div}(\boldsymbol{\beta} u) = 0,$$

in the domain $\Omega := (-1,1)^2$, with $\boldsymbol{\beta} := (2,-1)^T$ the fixed transport velocity. The initial condition is $u_0(x) = u_{ex}(x,0)$, the inflow boundary condition $u(x,t) = u_{ex}(x,t)$ are imposed at the inflow part $\Gamma_{in} = \{(x_1,x_2) \in \Omega : \ x_1 = -1 \text{ or } x_2 = 1\}$, the exact solution is $u_{ex}(x,t) := \exp(x_1 + x_2 - t)$. In the numerical experiment, the upper bound on the maximum wave speed is set to $\lambda_{max} := 1.5$. The final time is $T := 0.75$.

We show in Table 2 the relative errors in the L^1-norm and L^2-norm.

6.2.3 Wave equation in 1D

We consider the model problem

$$\begin{cases} u_t + v_x = 0, \\ v_t + u_x = 0, \end{cases}$$

in the domain $\Omega := (-1,1)$, with the initial condition $\big(u_0(x), v_0(x)\big) = \big(u_{ex}(x,0), v_{ex}(x,0)\big)$, and the boundary condition $u(x,t) = u_{ex}(x,t)$ is enforced at the whole boundary. The exact solution is defined as $u_{ex}(x,t) := \sin(x)\sin(t)$, $v_{ex}(x,t) := \cos(x)\cos(t)$. The upper bound on the maximum wave speed is set to $\lambda_{max} := 1$. The final time is $T := 1$.

We show in Table 3 the relative errors in the L^1-norm and L^2-norm for (u,v), i.e., we estimate errors defined by $\|\big(u_h^N, v_h^N\big) - \big(u_{ex}(\cdot,T), v_{ex}(\cdot,T)\big)\|_{L^p(\Omega)}$ for $p = 1,2$.

6.2.4 Wave equation in 2D with small and large graph viscosity

We consider the model problem

$$\begin{cases} u_t - \operatorname{div}\boldsymbol{v} = 0, \\ \boldsymbol{v}_t - c^2 \nabla u = 0, \end{cases}$$

Table 3: Wave equation, 1D

	L^1		L^2	
DoFs	error	rate	error	rate
21×2	3.14E-01	-	1.39E-01	-
41×2	1.36E-01	1.21	6.12E-02	1.18
81×2	6.60E-02	1.04	3.00E-02	1.03
161×2	3.16E-02	1.06	1.44E-02	1.06
321×2	1.57E-02	1.01	7.15E-03	1.01

Table 4: Wave equation, 2D

small viscosity	L^1		L^2	
DoFs	error	rate	error	rate
25×3	1.74E-00	-	1.70E-00	-
81×3	1.24E-00	0.58	1.21E-00	0.57
289×3	8.09E-01	0.65	7.98E-01	0.67
1089×3	4.83E-01	0.74	4.88E-01	0.78
4225×3	2.69E-01	0.80	2.85E-01	0.86
16641×3	1.44E-01	0.82	1.61E-02	0.91

large viscosity	L^1		L^2	
DoFs	error	rate	error	rate
25×3	2.87E-00	-	2.89E-00	-
81×3	2.55E-00	0.20	2.56E-00	0.21
289×3	2.08E-00	0.32	2.06E-00	0.34
1089×3	1.53E-00	0.47	1.50E-00	0.48
4225×3	9.95E-01	0.63	9.75E-01	0.63
16641×3	5.91E-01	0.76	5.87E-02	0.74

in the domain $\Omega := (-1,1)^2$, $c := \frac{1}{\sqrt{2\pi}}$ the wave speed, The initial condition is $\big(u_0(x), \boldsymbol{v}_0(x)\big) = \big(u_{ex}(x,0), \boldsymbol{v}_{ex}(x,0)\big)$, and the boundary condition $u(x,t) = u_{ex}(x,t)$ is enforced at the whole boundary, and the exact solution is

$$u_{ex}(x,t) = \sin(\pi x_1)\sin(\pi x_2)\sin(t), \boldsymbol{v}_{ex}(x,t) = \Big(\tfrac{-1}{2\pi}\cos(\pi x_1)\sin(\pi x_2)\cos(t), \tfrac{-1}{2\pi}\sin(\pi x_1)\cos(\pi x_2)\cos(t)\Big).$$

The upper bound on the maximum wave speed is set to $\lambda_{max} := c$ (small viscosity) in the first experiment and $\lambda_{max} := 1$ (large viscosity) in the second experiment. The final time is $T := 1$.

We show in Table 4 the relative errors in the L^1-norm and L^2-norm for $(u, \boldsymbol{v})$. As observed from the Table, if we add too much viscosity, the scheme still converges but the rate may be influenced.

6.2.5 Euler equations in 2D

We consider the following equations:

$$\begin{cases} \rho_t + \mathrm{div}\boldsymbol{m} = 0, \\ \boldsymbol{m}_t + \mathrm{div}(\boldsymbol{v} \otimes \boldsymbol{m}) + \nabla p = 0, \\ E_t + \mathrm{div}(\boldsymbol{v}(E + p)) = 0, \end{cases}$$

in the domain $\Omega := (-1,1)^2$, with the initial data $\big(\rho_0(x), \boldsymbol{m}_0(x), E_0(x)\big) = \big(\rho_{ex}(x,0), \boldsymbol{m}_{ex}(x,0), E_{ex}(x,0)\big)$, and the boundary data $\big(\rho(x,t), \boldsymbol{m}(x,t), E(x,t)\big) = \big(\rho_{ex}(x,t), \boldsymbol{m}_{ex}(x,t), E_{ex}(x,t)\big)$ enforced at the whole boundary, where the exact solution $(\rho_{ex}, \boldsymbol{m}_{ex}, E_{ex})$ is a two-dimensional isentropic vortex from Section 5.6 in [13] with $x_1^0 = x_2^0 = 0$. More precisely, the exact solution is constructed as follows: Let $\rho_\infty := T_\infty := 1$, $\boldsymbol{u}_\infty := (u_\infty, v_\infty)$, $u_\infty := v_\infty := 1$, be the free-stream values; then the exact solution is a passive convection of a vortex with mean velocity $\boldsymbol{u}_\infty$:

$$\rho_{ex}(x,t) = (T_\infty + \delta T)^{1/(\gamma-1)}, \quad \boldsymbol{u}_{ex}(x,t) = \boldsymbol{u}_\infty + \delta\boldsymbol{u}, \quad p(x,t) = \rho_{ex}^\gamma,$$

$$\delta\boldsymbol{u}(x,t) = \frac{\beta}{2\pi}\exp(\frac{1-r^2}{2})(-\bar{x}_2, \bar{x}_1)^T, \quad \delta T(x,t) = -\frac{(\gamma-1)\beta^2}{8\gamma\pi^2}\exp(1-r^2),$$

Table 5: Euler equations, 2D

ρ	L^1		L^2	
DoFs	error	rate	error	rate
25×4	1.46E-02	-	2.23E-01	-
81×4	1.81E-02	-0.36	3.02E-02	-0.51
289×4	1.76E-02	0.05	2.99E-02	0.01
1089×4	1.52E-02	0.22	2.51E-02	0.27
4225×4	1.26E-02	0.27	1.96E-02	0.35

E	L^1		L^2	
DoFs	error	rate	error	rate
25×4	4.57E-01	-	3.34E-01	-
81×4	3.46E-01	0.47	2.27E-01	0.66
289×4	2.66E-01	0.41	1.67E-01	0.48
1089×4	1.89E-01	0.51	1.14E-01	0.57
4225×4	1.29E-01	0.56	7.55E-02	0.61

Table 6: Shock for Burgers' equation

	L^1		L^2	
DoFs	error	rate	error	rate
21	9.37E-02	-	1.91E-01	-
41	4.70E-02	0.99	1.36E-01	0.50
81	2.30E-02	1.03	9.34E-02	0.54
161	1.15E-02	1.00	6.60E-02	0.50
321	5.88E-03	0.97	4.79E-02	0.46

with $\bar{x} := (x_1 - u_\infty t, x_2 - v_\infty t)$, $r^2 := \|\bar{x}\|_2^2$, $\gamma := 7/5$, and $\beta := 5$. Moreover,

$$\boldsymbol{m}_{ex} = \rho_{ex}\boldsymbol{u}_{ex}, \quad E_{ex} = \frac{1}{2}\Big(\frac{\boldsymbol{m}_{ex}^2}{\rho_{ex}} + \frac{p}{\gamma - 1}\Big).$$

The upper bound on the maximum wave speed is set to $\lambda_{max} := 5$ and the final time is $T := 0.5$.

We show in Table 5 the relative errors in the L^1-norm and L^2-norm for ρ and E.

6.3 Convergence tests with non-smooth solutions

In this section, we present simulations for non-smooth solutions, including a shock for Burgers' equation in 1D, a curved shock for Burgers' equation in 2D, and a curved shock for a nonconvex flux in 2D.

6.3.1 Shock for Burgers' equation in 1D

In this example, we consider Burgers' equation

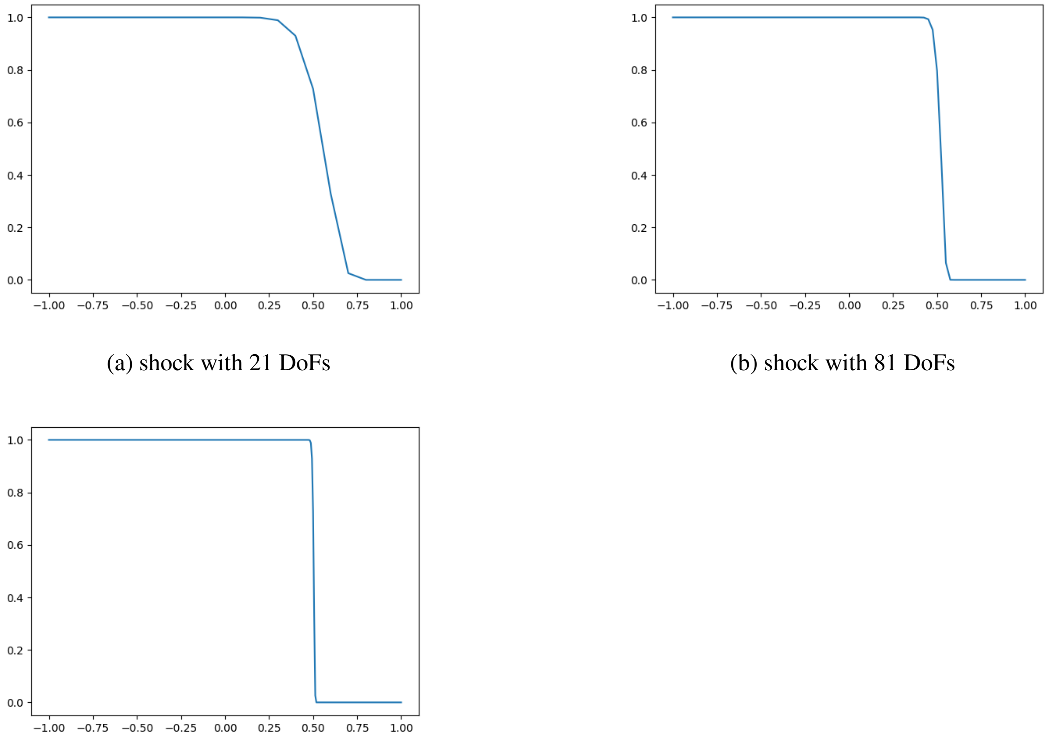
$$u_t + \Big(\frac{1}{2}u^2\Big)_x = 0,$$

with the initial data $u := 1$ if $x < 0$ and $u := 0$ if $x > 0$. In other words, we want to solve a Riemann problem for Burgers' equation. The exact solution is $u_{ex} = 1$ if $x < 0.5t$ and $u = 0$ otherwise. We estimate the errors at time $T := 1$. In Table 6, we observe that the rate of convergence is $h^{1/p}$, as predicted in [14]. The upper bound on the maximum wave speed is set to 1. We present solutions on a sequence of uniform meshes at the final time $T = 1$ in Figure 3 to illustrate how the mesh refinement improves the shock resolution.

6.3.2 Curved shock for Burgers' equation in 2D

We still consider Burgers' equation, but in 2D. This example is from [16]. We consider the model problem

$$u_t + \operatorname{div}\boldsymbol{f}(u) = 0,$$



(a) shock with 21 DoFs

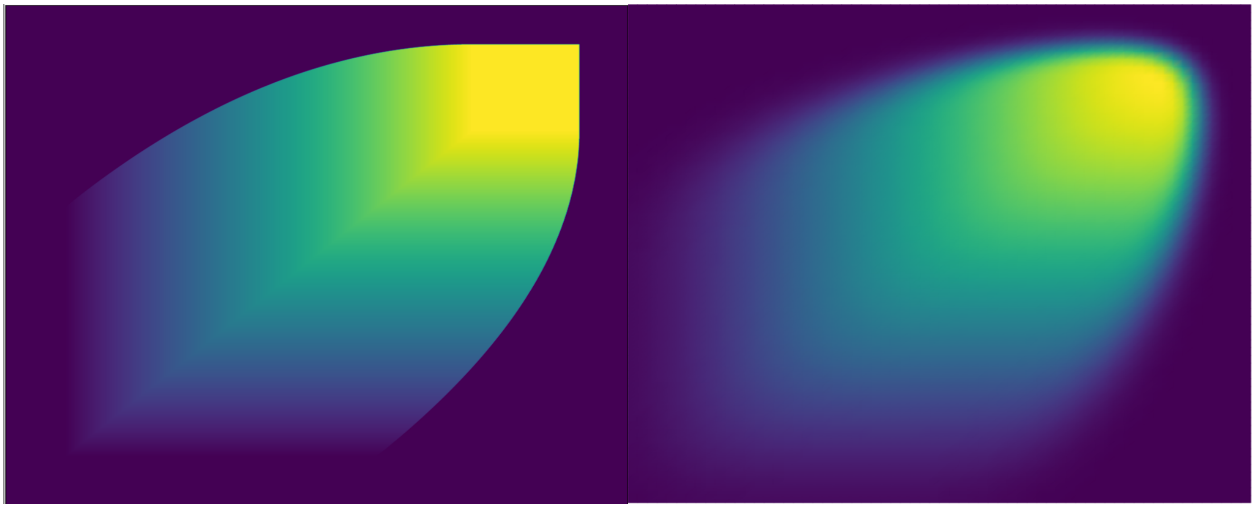
(b) shock with 81 DoFs

(c) shock with 321 DoFs

Figure 3: Shocks of Burgers' equation

Table 7: Burgers' equation, 2D

	L^1		L^2	
DoFs	error	rate	error	rate
25	6.40E-01	-	8.96E-01	-
81	4.86E-01	0.47	6.90E-01	0.44
289	3.44E-01	0.55	5.03E-01	0.49
1089	2.28E-01	0.62	3.53E-01	0.53
4225	1.43E-01	0.68	2.57E-01	0.47
16641	8.65E-02	0.74	1.88E-01	0.45



(a) exact solution

(b) numerical solution

Figure 4: Solutions to Burgers' equation

in the domain $\Omega := (-0.25, 1.75)^2$, with $\boldsymbol{f}(u) := \frac{1}{2}(u^2, u^2)^T$, and the initial condition $u_0(x) = u_{ex}(x, 0)$. The boundary condition is imposed on the inflow part. The final time is $T := 0.75$. The exact solution is constructed as follows. We take $a = 0.75$. We assume firstly $x_2 \leq x_1$. Then, we set $z_1 = x_1 - \frac{1}{2}$, $z_2 = x_2 - \frac{1}{2}$ and $\alpha = z_1 - z_2$. There are three cases depending on the value of α:

$$\text{if } \alpha \leq 1 - \frac{t(1+a)}{2}, \quad u_{ex}(x,t) := \begin{cases} \frac{z_2}{t} & \text{if } -at \leq z_2 < t, \\ 1, & \text{if } t \leq z_2 < 1 - a + (1-a)\frac{t}{2}, \\ -a, & \text{otherwise;} \end{cases}$$

$$\text{if } 1 - \frac{t(1+a)}{2} < \alpha \leq 1, \quad u_{ex}(x,t) := \begin{cases} \frac{z_2}{t}, & \text{if } -at \leq z_2 < \sqrt{2(1+a)t(1-a)} - at, \\ -a, & \text{otherwise;} \end{cases}$$

if $1 < \alpha$, then $u_{ex}(x,t) := -a$. Finally, we set $u_{ex}((x_1, x_2), t) := u_{ex}((x_2, x_1), t)$ if $x_1 \leq x_2$. The final time is $T = 0.75$ and the upper bound on the maximum wave speed is set to be 1. The rate of convergence is reported in Table 7.

Moreover, in Figure 4, we present the P_1-interpolant of the exact solution and our numerical simulation with 8192 cells. We observe that the shape of the exact solution is well reproduced by the numerical solution.

6.3.3 KPP flux in 2D

This example is also from [16]. We consider here a nonconvex flux $\boldsymbol{f}(u) := \Big(\sin(u), \cos(u)\Big)^T$. This is a challenging test, because of the loss of the convexity of the flux. The final time is $T := 1$, the domain is $\Omega := (-2, 2) \times (-2.5, 1.5)$,

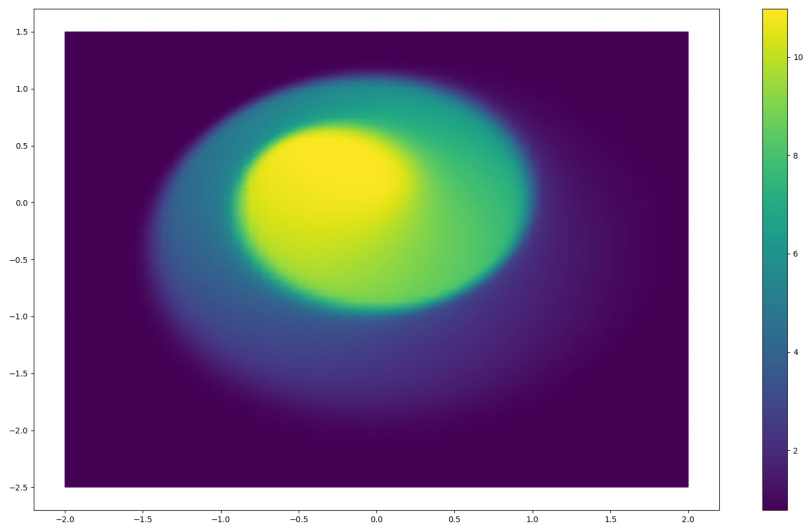


Figure 5: Numerical simulation with KPP flux

and the initial data is

$$u_0(x) := \begin{cases} \frac{14\pi}{4}, & \text{if } \|x\|_2^2 \leq 1, \\ \frac{\pi}{4}, & \text{otherwise.} \end{cases}$$

The boundary condition is imposed on the inflow part where $\boldsymbol{f}(u) \cdot \boldsymbol{n} < 0$. We observe from Figure 5 and comparing to the results reported in [16] that the correct shape of the solution is captured.

7 Further work

7.1 In practice

- We will rewrite the mesh reading part by using a well-developed mesh package in Python instead of reading the Matlab file.
- High order basis will be implemented.
- A general method for transforming Lagrange basis to Bernstein basis will be considered.
- We will find a way for imposing the boundary condition for Bernstein basis.
- Higher order scheme will be implemented.
- The scheme will be rewritten in parallel.

7.2 In theory

- We will try to remove the assumption of BV-like bound 3.1 and 4.3 without modifying the scheme.
- The general boundary condition will be considered.
- A more general assumption on entropy pairs will be investigated.

References

[1] T. APEL AND J. M. MELENK, *Interpolation and quasi-interpolation in h-and hp-version finite element spaces (extended version)*, in ASC Report 39/2015, Vienna University of Technology, 2015, pp. 1–60.

[2] F. BOUCHUT, *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources*, Frontiers in Mathematics, Birkhäuser Verlag, Basel, 2004.

[3] A. BRESSAN, *Hyperbolic conservation laws: an illustrated tutorial*, in Modelling and optimisation of flows on networks, vol. 2062 of Lecture Notes in Math., Springer, Heidelberg, 2013, pp. 157–245.

[4] E. CHIODAROLI AND O. KREML, *An overview of some recent results on the Euler system of isentropic gas dynamics*, Bull. Braz. Math. Soc. (N.S.), 47 (2016), pp. 241–253.

[5] B. COCKBURN, F. COQUEL, AND P. G. LEFLOCH, *Convergence of the finite volume method for multidimensional conservation laws*, SIAM J. Numer. Anal., 32 (1995), pp. 687–705.

[6] M. G. CRANDALL AND A. MAJDA, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[7] D. A. DI PIETRO AND A. ERN, *Mathematical aspects of discontinuous Galerkin methods*, vol. 69 of Mathématiques & Applications (Berlin) [Mathematics & Applications], Springer, Heidelberg, 2012.

[8] R. J. DIPERNA, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223–270.

[9] A. ERN AND J.-L. GUERMOND, *Finite elements I—Approximation and interpolation*, vol. 72 of Texts in Applied Mathematics, Springer, Cham, [2021] ©2021.

[10] ——, *Finite elements III—first-order and time-dependent PDEs*, vol. 74 of Texts in Applied Mathematics, Springer, Cham, [2021] ©2021.

[11] E. GODLEWSKI AND P.-A. RAVIART, *Numerical approximation of hyperbolic systems of conservation laws*, vol. 118 of Applied Mathematical Sciences, Springer-Verlag, New York, 1996.

[12] J.-L. GUERMOND AND M. NAZAROV, *A maximum-principle preserving C^0 finite element method for scalar conservation equations*, Comput. Methods Appl. Mech. Engrg., 272 (2014), pp. 198–213.

[13] J.-L. GUERMOND, M. NAZAROV, B. POPOV, AND I. TOMAS, *Second-order invariant domain preserving approximation of the Euler equations using convex limiting*, SIAM J. Sci. Comput., 40 (2018), pp. A3211–A3239.

[14] J.-L. GUERMOND AND B. POPOV, *Error estimates of a first-order Lagrange finite element technique for nonlinear scalar conservation equations*, SIAM J. Numer. Anal., 54 (2016), pp. 57–85.

[15] ——, *Invariant domains and first-order continuous finite element approximation for hyperbolic systems*, SIAM J. Numer. Anal., 54 (2016), pp. 2466–2489.

[16] ——, *Invariant domains and second-order continuous finite element approximation for scalar conservation equations*, SIAM J. Numer. Anal., 55 (2017), pp. 3120–3146.

[17] H. HOLDEN AND N. H. RISEBRO, *Front tracking for hyperbolic conservation laws*, vol. 152 of Applied Mathematical Sciences, Springer, Heidelberg, second ed., 2015.

[18] J. JAFFRÉ, C. JOHNSON, AND A. SZEPESSY, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci., 5 (1995), pp. 367–386.

[19] P. G. LEFLOCH, *Hyperbolic systems of conservation laws*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2002. The theory of classical and nonclassical shock waves.