

# INVARIANT DOMAIN PRESERVING HIGH-ORDER SPECTRAL DISCONTINUOUS APPROXIMATIONS OF HYPERBOLIC SYSTEMS

VALENTIN CARLIER\* AND FLORENT RENAC†

**Abstract.** We propose a limiting procedure to preserve invariant domains with time explicit discrete high-order spectral discontinuous approximate solutions to hyperbolic systems of conservation laws. Provided the scheme is discretely conservative and satisfy geometric conservation laws at the discrete level, we derive a condition on the time step to guaranty that the cell-averaged approximate solution is a convex combination of states in the invariant domain. These states are then used to define local bounds which are then imposed to the full high-order approximate solution within the cell via an a posteriori scaling limiter. Numerical experiments are then presented with modal and nodal discontinuous Galerkin schemes confirm the robustness and stability enhancement of the present approach.

**Key words.** Hyperbolic systems, Convex invariant set, Limiting, Spectral discontinuous method, Discontinuous Galerkin method

**AMS subject classifications.** 65M12, 65M70, 35L65

**1. Introduction.** Let  $D \subset \mathbb{R}^d$  be an open domain with  $d$  the space dimension. We are interested here in high-order numerical solutions to hyperbolic systems of conservations law

$$(1.1) \quad \begin{cases} \partial_t \mathbf{u} + \nabla \cdot \mathbf{f}(\mathbf{u}) = 0, & \text{in } D \times (0, \infty), \\ \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), & \text{in } D, \end{cases}$$

where  $\mathbf{u}(\mathbf{x}, t)$  represents the vector of conserved variables with values in the set of states  $\Omega^a \subset \mathbb{R}^m$  which is assumed to be convex. The flux tensor  $\mathbf{f} = (\mathbf{f}_1, \dots, \mathbf{f}_d) : \Omega^a \ni \mathbf{u} \mapsto \mathbf{f}(\mathbf{u}) \in \mathbb{R}^{m \times d}$  is assumed to be smooth.

Solutions to (1.1) may develop discontinuities in finite time even if the initial data is smooth, therefore the equations are to be understood in the sense of distributions. Nevertheless, in this setting we lose uniqueness of the solution and (1.1) must be supplemented with further admissibility criteria. We here focus on entropy inequalities on some twice differentiable strictly convex function  $\eta : \Omega^a \rightarrow \mathbb{R}$  associated with a smooth entropy flux  $\mathbf{q} : \Omega^a \rightarrow \mathbb{R}^d$  satisfying

$$(1.2) \quad \boldsymbol{\eta}'(\mathbf{u})^\top \mathbf{f}'_i(\mathbf{u}) = \mathbf{q}'_i(\mathbf{u})^\top \quad \forall \mathbf{u} \in \Omega^a, \quad 1 \leq i \leq d.$$

A weak solution to (1.1) is called an entropy weak solution if for every entropy pair of (1.1) we have

$$(1.3) \quad \frac{\partial \eta(\mathbf{u})}{\partial t} + \nabla \cdot \mathbf{q}(\mathbf{u}) \leq 0,$$

in the sense of distributions. Classical (smooth) solutions respect this condition with an equality, since one can apply the chain rule with (1.2). The inequality for discontinuous solutions comes from a vanishing viscosity argument by adding a parabolic perturbation to (1.1), the regularizing effect allows to have a smooth unique solution in this case. Then, under some structural assumptions on (1.1), vanishing viscosity

\*École normale supérieure, F-75230 Paris, France (vcarlier@clipper.ens.fr)

†DAAA, ONERA, Université Paris Saclay F-92322 Châtillon, France (florent.renac@onera.fr)

approximations converge to an entropy measure valued solution to (1.1) and (1.3) [9]. This result is in particular based on the existence of convex invariant domains  $\mathcal{B} \subset \Omega^a$  for (1.1): if  $\mathbf{u}$  is in  $\mathcal{B}$ , then it remains in  $\mathcal{B}$  almost everywhere in  $D \times (0, \infty)$  [28, 13, 24, 40]. This property generalizes the notion of maximum principle for scalar equations. Numerical methods keeping this property at the discrete level are called invariant domain preserving (IDP).

We are here interested in the approximation of (1.1) using high-order discontinuous spectral methods (see, e.g., [7, 14, 11, 6, 8] and references therein) where the solution to (1.1) is sought under the form of discontinuous piecewise truncated series of analytic functions over a partition of the domain  $D$ . Such methods have been applied to a wide range of applications [41, 45], and have the potential to achieve high-order accuracy efficiently on modern parallel architectures [25, 12]. Unfortunately these approximations suffer from spurious oscillations around discontinuities of the exact solution due to Gibbs phenomenon [10, 17] that may cause the approximate solution to become locally nonphysical, leading to robustness issues. A large body of research has been proposed to address such issues with, e.g., solution and flux limiters [31, 49, 21], entropy conservative subcell flux differencing [11, 6, 8] artificial viscosity [20, 2], shock-capturing terms [26, 23].

We here focus on a posteriori limiters from [49, 50] scaling the cellwise approximate solution around its cell average, thus allowing to preserve positivity of the solution (i.e., with  $\mathcal{B} = \Omega^a$ ) and maximum principles in scalar problems, while preserving conservativity of the method. Under some strong assumptions on the mesh, it is indeed possible to derive a condition on the time step of the scheme so that the cell-averaged solution remains in the invariant domains on Cartesian and simplicial grids [49, 50, 27], or on unstructured quadrangular straight-sided grids [38]. We here extend this limiting technique to an IDP limiter for a broad class of spectral discontinuous methods with explicit time stepping on general unstructured meshes with possibly curved elements. Provided, the discretization method is conservative and satisfies geometric conservation laws [43, 29] at the discrete level, we propose a condition on the time step to guaranty that the cell-averaged approximate solution is a convex combination of states lying in the required invariant domains. These states are then used to define local bounds which are then imposed to the full high-order approximate solution within the cell via the scaling limiter. This strategy is closely related to convex limiting [21] based on first-order IDP approximations defining local bounds and then forcing the high-order approximation to satisfy these bounds through flux limiting [4, 48]. This approach has been applied to finite element approximations in [19] and discontinuous Galerkin spectral element method in [34] among others.

The objective of this paper is hence to derive a CFL condition on the time step and to propose an iterative algorithm for its evaluation. These are based first on the existence of a state in the invariant region which satisfies a trivial flux balance over each mesh cell boundaries. We call this state the pseudo-equilibrium state and then use tricks from [36] to expand the the cell-averaged approximate solution is a convex combination of states lying in the invariant domains. The CFL condition is also based on the existence of a quadrature rule to evaluate the cell-averaged solution that includes the traces of the numerical solution used to evaluate numerical fluxes in the scheme. This latter result generalized the work in [51] on triangular grids to general curved polyhedral elements.

The paper is organized as follows. In section 2 we introduce the notion of invariant domain and invariant domain preserving Riemann solver. In section 3 we will state and prove our theorem on the existence of a CFL for high order schemes, present and

discuss the limiting strategy. [section 4](#) will present various schemes that satisfy the hypothesis of our work and state the CFL precisely for those. Numerical experiments will be presented in [section 5](#), and the conclusions follow in [section 6](#).

**2. Approximate Riemann solvers.** In this section we present some basic notions [[5](#), [22](#)] on Riemann problem, approximate Riemann solver (ARS), and convex invariant domain that will be used in the remainder of this work. Throughout this section,  $\mathbf{n}$  in  $\mathbb{R}^d$  is a given unit vector.

**2.1. Riemann problem and invariant domains.** Let two states  $\mathbf{u}_L$  and  $\mathbf{u}_R$  in  $\Omega^a$ , it is convenient for the present analysis to consider the Riemann problem in the direction  $\mathbf{n}$ :

$$(2.1) \quad \partial_t \mathbf{u} + \partial_x \mathbf{f}(\mathbf{u}) \cdot \mathbf{n} = 0, \text{ in } \mathbb{R} \times (0, \infty), \quad \mathbf{u}(x, 0) = \begin{cases} \mathbf{u}_L, & \text{if } x < 0, \\ \mathbf{u}_R, & \text{if } x > 0. \end{cases}$$

We will integrate [\(2.1\)](#) over the space time slab  $[-\frac{h}{2}, \frac{h}{2}] \times [0, \Delta t]$  with  $h > 0$  and  $\Delta t > 0$  the space and time steps. We suppose here that all the Riemann problems we consider have a self-similar entropy weak solution  $\mathcal{W}(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n})$ . Let introduce the self-similar variable  $\xi = \frac{x}{\Delta t}$  and assume that there exist  $\sigma_L, \sigma_R$  such that:  $\mathcal{W}(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) = \mathbf{u}_L$  for  $\xi < \sigma_L$  and  $\mathcal{W}(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) = \mathbf{u}_R$  for  $\xi > \sigma_R$ . We then define the maximum wave speed in [\(2.1\)](#) by

$$(2.2) \quad |\lambda|(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) = \max(|\sigma_L|, |\sigma_R|),$$

and for  $\frac{\Delta t}{h} |\lambda|(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) \leq \frac{1}{2}$ , we define the average over the Riemann fan [[21](#), [22](#)]

$$(2.3) \quad \bar{\mathbf{u}}(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}, \Delta t) := \frac{1}{h} \int_{-\frac{h}{2}}^{\frac{h}{2}} \mathcal{W}\left(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}\right) dx = \frac{\mathbf{u}_L + \mathbf{u}_R}{2} - \Delta t (\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)) \cdot \mathbf{n}.$$

Finally, we will use the definition of invariant domain from [[21](#)]: a convex set  $\mathcal{B} \subset \Omega^a$  is an invariant domain [\(1.1\)](#) if for all  $\mathbf{u}_L$  and  $\mathbf{u}_R$  in  $\mathcal{B}$ , we have

$$\bar{\mathbf{u}}(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}, \frac{\Delta t}{h}) \in \mathcal{B} \quad \forall \frac{\Delta t}{h} |\lambda|(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) \leq \frac{1}{2}.$$

**2.2. Two-point numerical fluxes and approximate Riemann solvers.** The discretization of [\(1.1\)](#) will rely on two-point numerical fluxes [[32](#), [22](#)] for the approximation of  $\mathbf{f} \cdot \mathbf{n}$  and we assume them to be consistent and conservative:

$$(2.4) \quad \mathbf{h}(\mathbf{u}, \mathbf{u}, \mathbf{n}) = \mathbf{f}(\mathbf{u}) \cdot \mathbf{n}, \quad \mathbf{h}(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) = -\mathbf{h}(\mathbf{u}_R, \mathbf{u}_L, -\mathbf{n}) \quad \forall \mathbf{u}, \mathbf{u}_L, \mathbf{u}_R \in \Omega^a,$$

and Lipschitz continuous. We also define the notion of IDP two-point flux in the following definition.

**DEFINITION 2.1.** *A two-point flux is said to be invariant domain preserving (IDP) for  $\mathcal{B}$  an invariant domain if we have*

$$\mathbf{u} - \frac{\Delta t}{h} (\mathbf{h}(\mathbf{u}, \mathbf{u}_R, \mathbf{n}) - \mathbf{h}(\mathbf{u}_L, \mathbf{u}, \mathbf{n})) \in \mathcal{B} \quad \forall \mathbf{u}_L, \mathbf{u}, \mathbf{u}_R \in \mathcal{B},$$

*under the half CFL condition*

$$\frac{\Delta t}{h} \max(|\lambda|(\mathbf{u}_L, \mathbf{u}, \mathbf{n}), |\lambda|(\mathbf{u}, \mathbf{u}_R, \mathbf{n})) \leq \frac{1}{2}.$$

We now introduce the notion of ARS and IDP ARS, which will be used to derive IDP two-point fluxes.

DEFINITION 2.2 (Approximate Riemann solver). *An ARS is a self-similar function  $\mathcal{W}^a(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n})$ , used to approximate the solution  $\mathcal{W}(\frac{x}{t}; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n})$  of the Riemann problem (2.1), that is consistent with the integral form of (1.1) [22, 16]: for any  $\frac{\Delta t}{h}|\lambda|(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) \leq \frac{1}{2}$ , we have*

$$(2.5) \quad \frac{1}{h} \int_{-\frac{h}{2}}^{\frac{h}{2}} \mathcal{W}^a\left(\frac{x}{\Delta t}; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}\right) dx = \frac{\mathbf{u}_L + \mathbf{u}_R}{2} - \Delta t (\mathbf{f}(\mathbf{u}_R) - \mathbf{f}(\mathbf{u}_L)) \cdot \mathbf{n}.$$

Using consistency in (2.4), and setting  $\xi = \frac{x}{\Delta t}$ , one defines a two-point flux from an ARS as

$$(2.6a) \quad \mathbf{h}_{\mathcal{W}^a}(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) = \mathbf{f}(\mathbf{u}_L) \cdot \mathbf{n} - \int_{-\lambda}^0 (\mathcal{W}^a(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) - \mathbf{u}_L) d\xi$$

$$(2.6b) \quad = \mathbf{f}(\mathbf{u}_R) \cdot \mathbf{n} + \int_0^{\lambda} (\mathcal{W}^a(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) - \mathbf{u}_R) d\xi,$$

where  $\lambda = |\lambda|(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n})$ . Both definitions are equivalent due to (2.5). We can now define the notion of IDP ARS.

DEFINITION 2.3. *An ARS  $\mathcal{W}^a(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n})$  is IDP for  $\mathcal{B}$  an invariant domain if we have*

$$\frac{1}{\lambda} \int_{-\lambda}^0 \mathcal{W}^a(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) d\xi \in \mathcal{B}, \quad \frac{1}{\lambda} \int_0^{\lambda} \mathcal{W}^a(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) d\xi \in \mathcal{B} \quad \forall \mathbf{u}_L, \mathbf{u}_R \in \mathcal{B}.$$

As a consequence  $\frac{1}{2\lambda} \int_{-\lambda}^{\lambda} \mathcal{W}^a(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) d\xi$  is also in  $\mathcal{B}$ . We have the following results linking IDP ARS and two-point numerical flux.

LEMMA 2.4 (Interface invariant domain preservation [5]). *The ARS  $\mathcal{W}^a$  is IDP for  $\mathcal{B}$  iff. for all  $\mathbf{u}_L, \mathbf{u}_R$  in  $\mathcal{B}$ , and  $\frac{\Delta t}{h}|\lambda|(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) \leq 1$ , we have*

$$(2.7a) \quad \mathbf{u}_L - \frac{\Delta t}{h} (\mathbf{h}_{\mathcal{W}^a}(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) - \mathbf{f}(\mathbf{u}_L) \cdot \mathbf{n}) \in \mathcal{B},$$

$$(2.7b) \quad \mathbf{u}_R - \frac{\Delta t}{h} (\mathbf{f}(\mathbf{u}_R) \cdot \mathbf{n} - \mathbf{h}_{\mathcal{W}^a}(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n})) \in \mathcal{B}.$$

*Proof.* Let consider (2.7a), a similar argument holds for (2.7b). From (2.6a), we have

$$\mathbf{u}_L - \frac{\Delta t}{h} (\mathbf{h}_{\mathcal{W}^a}(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) - \mathbf{f}(\mathbf{u}_L) \cdot \mathbf{n}) = \left(1 - \frac{\Delta t}{h} \lambda\right) \mathbf{u}_L + \frac{\Delta t}{h} \lambda \frac{1}{\lambda} \int_{-\lambda}^0 (\mathcal{W}^a(\xi; \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) d\xi,$$

and since  $0 < \frac{\Delta t}{h} \lambda \leq 1$  this is a convex combination of states in  $\mathcal{B}$  and therefore is in  $\mathcal{B}$ . Conversely, taking  $\frac{\Delta t}{h} \lambda = 1$  the above integrals are  $\mathcal{B}$  iff. (2.7a) and (2.7b) hold and we conclude from Definition 2.3.  $\square$

This allows us to state the following result.

LEMMA 2.5. *Let  $\mathbf{h}_{\mathcal{W}^a}$  and  $\mathbf{h}_{\mathcal{W}^b}$  be two-point fluxes from two different ARS that are IDP for  $\mathcal{B}$ . Then, we have*

$$(2.8) \quad \mathbf{u} - \frac{\Delta t}{h} (\mathbf{h}_{\mathcal{W}^a}(\mathbf{u}, \mathbf{u}_R, \mathbf{n}) - \mathbf{h}_{\mathcal{W}^b}(\mathbf{u}_L, \mathbf{u}, \mathbf{n})) \in \mathcal{B} \quad \forall \mathbf{u}_L, \mathbf{u}, \mathbf{u}_R \in \mathcal{B},$$

under the half CFL condition

$$\frac{\Delta t}{h} \max(|\lambda|(\mathbf{u}_L, \mathbf{u}, \mathbf{n}), |\lambda|(\mathbf{u}, \mathbf{u}_R, \mathbf{n})) \leq \frac{1}{2}.$$

*Proof.* We rewrite (2.8) as

$$\begin{aligned} \mathbf{u} - \frac{\Delta t}{h} (\mathbf{h}_{\mathcal{W}^a}(\mathbf{u}, \mathbf{u}_R, \mathbf{n}) - \mathbf{h}_{\mathcal{W}^b}(\mathbf{u}_L, \mathbf{u}, \mathbf{n})) &= \frac{1}{2} \left( \mathbf{u} - 2 \frac{\Delta t}{h} (\mathbf{h}_{\mathcal{W}^a}(\mathbf{u}, \mathbf{u}_R, \mathbf{n}) - \mathbf{f}(\mathbf{u}) \cdot \mathbf{n}) \right) \\ &\quad + \frac{1}{2} \left( \mathbf{u} - 2 \frac{\Delta t}{h} (\mathbf{f}(\mathbf{u}) \cdot \mathbf{n} - \mathbf{h}_{\mathcal{W}^b}(\mathbf{u}, \mathbf{u}_R, \mathbf{n})) \right), \end{aligned}$$

and apply Lemma 2.4 to both terms of the right-hand side with  $2 \frac{\Delta t}{h} \lambda \leq 1$ .  $\square$

The previous lemma with  $\mathcal{W}^b = \mathcal{W}^a$  also proves that the three-point scheme built from an IDP ARS is also IDP [5].

**3. Invariant domain preserving limiter.** We here state and prove our main results on the existence of an explicit condition on the time step to ensure that the cell-averaged solution from a high-order spectral discontinuous scheme is IDP. In subsection 3.1 we clarify the schemes we are considering in this work. Our results are based on the existence of a pseudo-equilibrium state allowing a balance of the numerical fluxes at faces of each element which is introduced in subsection 3.2 where we prove its existence. The main result giving the condition on the time step is given in subsection 3.3. A limiting strategy based on convex bounds is described in subsection 3.4, while subsection 3.5 introduces a fast algorithm to evaluate the time step.

**3.1. Cell-averaged fully discrete scheme.** We now describe the main properties of the numerical methods we are considering in this work. We consider here discretely conservative high-order approximations of (1.1). Without loss of generality, we use an explicit forward Euler discretization in time. High-order time integration is then performed using strong-stability preserving Runge-Kutta methods [18] that are convex combinations of explicit first-order schemes in time and thus keep their stability properties. For the spatial discretization, the approximate solution  $\mathbf{u}_h(\mathbf{x}, t)$  is defined locally over each element  $\kappa$  of the partition  $\mathcal{T}_h$  of the domain  $D$  in a local function cell space  $\mathcal{V}_h^p(\kappa)$ . By  $\mathbf{u}_h^{(n+1)}(\cdot) = \mathbf{u}_h(\cdot, t^{(n+1)})$  we denote the solution at time  $t^{(n+1)} = t^{(n)} + \Delta t^{(n)}$  with  $t^{(0)} = 0$  and  $\Delta t^{(n)} > 0$  the time step. The approximate solution is assumed to satisfy the following relation for the cell-averaged solution  $\langle \mathbf{u}_h \rangle_\kappa$ :

$$(3.1) \quad \langle \mathbf{u}_h^{(n+1)} \rangle_\kappa = \langle \mathbf{u}_h^{(n)} \rangle_\kappa - \Delta t^{(n)} \sum_{k=1}^{N_f} s_k^\kappa \mathbf{h}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa) \quad \forall \kappa \in \mathcal{T}_h,$$

where the  $\mathbf{x}_k^\kappa$  are some points on the faces  $f$  in  $\partial\kappa$  and  $\mathbf{u}_h^\pm(\mathbf{x}_k^\kappa, t^{(n)}) = \lim_{\epsilon \rightarrow 0^+} \mathbf{u}_h(\mathbf{x}_k^\kappa \pm \epsilon \mathbf{n}_k^\kappa, t^{(n)})$  denote evaluations of the traces of the solutions at  $\mathbf{x}_k^\kappa$  (see Fig. 1). The  $s_k^\kappa > 0$  are local contributions to  $\frac{|f|}{|\kappa|}$  with  $|f|$  and  $|\kappa|$  approximations of the face surface and element volume, and we introduce

$$(3.2) \quad \mathcal{S}^\kappa := \sum_{k=1}^{N_f} s_k^\kappa.$$

The geometrical quantities depend on the numerical method under consideration and examples will be given in section 4. By  $\mathbf{h}$  we denote a consistent, conservative

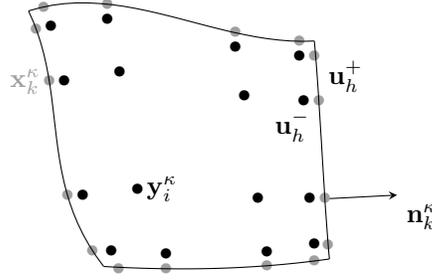


Fig. 1: Notations for  $d = 2$  on a quadrangle: definitions of the unit outward normal vector  $\mathbf{n}_\kappa^\kappa$ , element quadrature nodes  $\mathbf{y}_i^\kappa$  (black bullets), surface quadrature node  $\mathbf{x}_k^\kappa$  (gray bullets), and inner and outer traces  $\mathbf{u}_h^\pm$  at  $\mathbf{x}_k^\kappa$ .

(2.4), and IDP (see Definition 2.1) two-point flux. The cell-averaged solution  $\langle \mathbf{u}_h^{(n)} \rangle_\kappa$  is supposed to be evaluated through a suitable quadrature rule, that includes  $N_v$  volume quadrature points  $\mathbf{y}_i^\kappa$  in  $\kappa$  together with the  $N_f$  surface points  $\mathbf{x}_i^\kappa$  on  $\partial\kappa$  (see Fig. 1) introduced in (3.1):

$$(3.3) \quad \langle \mathbf{u}_h^{(n)} \rangle_\kappa = \sum_{i=1}^{N_v} \nu_i^\kappa \mathbf{u}_h(\mathbf{y}_i^\kappa, t^{(n)}) + \sum_{i=1}^{N_f} \beta_i^\kappa \mathbf{u}_h^-(\mathbf{x}_i^\kappa, t^{(n)}),$$

where the weights  $\nu_{1 \leq i \leq N_v}^\kappa \geq 0$  and  $\beta_{1 \leq i \leq N_f}^\kappa > 0$  are assumed to satisfy

$$(3.4) \quad \sum_{i=1}^{N_v} \nu_i^\kappa + \sum_{i=1}^{N_f} \beta_i^\kappa = 1.$$

Lemma 3.1 and Corollary 3.3 give a theoretical basis on existence of such a quadrature on polygonal or polyhedral mesh elements. An explicit example is also given in [51] in the case of triangles and in section 4.

LEMMA 3.1. *Let  $\kappa$  be a compact subset of  $\mathbb{R}^d$  and  $\mathcal{P}_\kappa$  a finite dimensional subspace of  $\mathcal{C}^0(\kappa, \mathbb{R})$ , that contains the constant function  $f \equiv 1_\kappa$ . Suppose that there exists a quadrature  $(\varpi_i^\kappa, \mathbf{y}_i^\kappa)_{1 \leq i \leq N_v}$  with positive weights  $\varpi_i^\kappa > 0$  and points  $\mathbf{y}_i^\kappa$  in  $\kappa$  that integrates exactly products of functions in  $\mathcal{P}_\kappa$ :*

$$(3.5) \quad \int_\kappa f(\mathbf{y})g(\mathbf{y})dV = \sum_{i=1}^{N_v} \varpi_i^\kappa f(\mathbf{y}_i^\kappa)g(\mathbf{y}_i^\kappa) \quad \forall f, g \in \mathcal{P}_\kappa.$$

*Then, for given points  $(\mathbf{x}_i^\kappa)_{1 \leq i \leq N_f}$  in  $\kappa$  there exist nonnegative  $(\nu_i^\kappa)_{1 \leq i \leq N_v}$  and positive  $(\beta_i^\kappa)_{1 \leq i \leq N_f}$  coefficients such that*

$$(3.6) \quad \langle f \rangle_\kappa = \sum_{i=1}^{N_v} \nu_i^\kappa f(\mathbf{y}_i^\kappa) + \sum_{i=1}^{N_f} \beta_i^\kappa f(\mathbf{x}_i^\kappa) \quad \forall f \in \mathcal{P}_\kappa.$$

*Proof.* It is known that  $(f, g) \mapsto \int_\kappa f(\mathbf{y})g(\mathbf{y})dV$  defines a scalar product on  $\mathcal{C}^0(\kappa, \mathbb{R})$  and therefore on  $\mathcal{P}_\kappa$ . Then using the Riesz representation theorem, every linear form  $\varphi : \mathcal{P}_\kappa \rightarrow \mathbb{R}$  can be represented using this scalar product: there exists

$f_\varphi \in \mathcal{P}_\kappa$  such that for every  $g \in \mathcal{P}_\kappa$ ,  $\varphi(g) = \int_\kappa f_\varphi(\mathbf{y})g(\mathbf{y})dV$ , then let  $\alpha_i^\varphi = \varpi_i^\kappa f_\varphi(\mathbf{y}_i^\kappa)$ , we obtain for every  $g \in \mathcal{P}_\kappa$ ,  $\varphi(g) = \sum_{i=1}^{N_v} \alpha_i^\varphi g(\mathbf{y}_i^\kappa)$ . Now, since  $f \mapsto \sum_{i=1}^{N_f} s_i^\kappa f(\mathbf{x}_i^\kappa)$  defines a linear form on  $\mathcal{P}_\kappa$ , for any  $s_i^\kappa > 0$ , it can be represented this way: there exist  $(\alpha_i^\kappa)_{1 \leq i \leq N_v}$  such that:

$$(3.7) \quad \sum_{i=1}^{N_f} s_i^\kappa f(\mathbf{x}_i^\kappa) = \sum_{i=1}^{N_v} \alpha_i^\kappa f(\mathbf{y}_i^\kappa) \quad \forall f \in \mathcal{P}_\kappa.$$

As the constant function is in  $\mathcal{P}_\kappa$ , we have for  $f$  in  $\mathcal{P}_\kappa$ ,  $\langle f \rangle_\kappa = \sum_{i=1}^{N_v} \varpi_i^\kappa f(\mathbf{y}_i^\kappa)$ , so for  $\varepsilon_\kappa > 0$

$$\begin{aligned} \langle f \rangle_\kappa &= \sum_{i=1}^{N_v} \varpi_i^\kappa f(\mathbf{y}_i^\kappa) = \sum_{i=1}^{N_v} \varpi_i^\kappa f(\mathbf{y}_i^\kappa) - \varepsilon_\kappa \sum_{i=1}^{N_f} s_i^\kappa f(\mathbf{x}_i^\kappa) + \varepsilon_\kappa \sum_{i=1}^{N_f} s_i^\kappa f(\mathbf{x}_i^\kappa) \\ &= \sum_{i=1}^{N_v} \varpi_i^\kappa f(\mathbf{y}_i^\kappa) - \varepsilon_\kappa \sum_{i=1}^{N_v} \alpha_i^\kappa f(\mathbf{y}_i^\kappa) + \varepsilon_\kappa \sum_{i=1}^{N_f} s_i^\kappa f(\mathbf{x}_i^\kappa) \\ &= \sum_{i=1}^{N_v} (\varpi_i^\kappa - \varepsilon_\kappa \alpha_i^\kappa) f(\mathbf{y}_i^\kappa) + \varepsilon_\kappa \sum_{i=1}^{N_f} s_i^\kappa f(\mathbf{x}_i^\kappa). \end{aligned}$$

Since the  $\varpi_i^\kappa$  are positive, for  $\varepsilon_\kappa = \min_{\{i: \alpha_i^\kappa > 0\}} \left( \frac{\varpi_i^\kappa}{\alpha_i^\kappa} \right) > 0$ , the  $(\varpi_i^\kappa - \varepsilon_\kappa \alpha_i^\kappa)$  are nonnegative, then (3.6) holds with

$$(3.8) \quad \nu_i^\kappa = \varpi_i^\kappa - \varepsilon_\kappa \alpha_i^\kappa \geq 0 \quad \forall 1 \leq i \leq N_v, \quad \beta_i^\kappa = \varepsilon_\kappa s_i^\kappa > 0 \quad \forall 1 \leq i \leq N_f. \quad \square$$

*Remark 3.2.* Note that  $\kappa$  is usually a polyhedron and  $\mathcal{P}_\kappa$  a polynomial space, we can subdivide  $\kappa$  into simplices and since there are quadrature rules integrating exactly arbitrary order polynomials on simplices, the previous lemma can be applied. Likewise, the existence of the quadrature in Lemma 3.1 is required only for modal methods or when the DOFs are not defined at the  $\mathbf{x}_k^\kappa$  in (3.1), so the present framework also holds for non polynomial nodal approximations with DOFs at the faces as in [8].

The following corollary allows to explicitly define the quadrature (3.3) for modal polynomial based methods.

**COROLLARY 3.3.** *Given a basis  $(\phi_j)_{1 \leq j \leq N_p}$  of  $\mathcal{P}_\kappa$  which is orthonormal with respect to the inner product (i.e.,  $\int_\kappa \phi_i(\mathbf{x})\phi_j(\mathbf{x})dV = \delta_{ij}$  where  $\delta_{ij}$  is the Kronecker symbol), then the  $\alpha_i^\kappa$  in (3.8) read*

$$(3.9) \quad \alpha_i^\kappa = \varpi_i^\kappa \sum_{j=1}^{N_p} \sum_{k=1}^{N_f} s_k^\kappa \phi_j(\mathbf{x}_k^\kappa) \phi_j(\mathbf{y}_i^\kappa).$$

*Proof.* We know that there exists  $g \in \mathcal{P}_\kappa$  such that (see proof of Lemma 3.1)

$$(3.10) \quad \sum_{i=1}^{N_f} s_i^\kappa f(\mathbf{x}_i^\kappa) = \sum_{i=1}^{N_v} \varpi_i^\kappa f(\mathbf{y}_i^\kappa) g(\mathbf{y}_i^\kappa) \quad \forall f \in \mathcal{P}_\kappa.$$

Expanding  $g \equiv \sum_{k=1}^{N_p} g_k \phi_k$  in the orthonormal basis and using (3.10) with  $f \equiv \phi_j$

we get

$$\sum_{i=1}^{N_f} s_i^\kappa \phi_j(\mathbf{x}_i^\kappa) = \sum_{i=1}^{N_v} \varpi_i^\kappa \phi_j(\mathbf{y}_i^\kappa) \sum_{k=1}^{N_p} g_k \phi_k(\mathbf{y}_i^\kappa) = g_j,$$

for all  $1 \leq j \leq N_p$ , by orthonormality of the basis. Substituting  $g$  in (3.10) by its expansion in the basis gives

$$\sum_{i=1}^{N_f} s_i^\kappa f(\mathbf{x}_i^\kappa) = \sum_{i=1}^{N_v} \varpi_i^\kappa f(\mathbf{y}_i^\kappa) \sum_{j=1}^{N_p} \sum_{k=1}^{N_f} s_k^\kappa \phi_j(\mathbf{x}_k^\kappa) \phi_j(\mathbf{y}_i^\kappa) \quad \forall f \in \mathcal{P}_\kappa,$$

and we conclude by comparing this result with (3.7).  $\square$

Finally, the scheme is assumed to preserve uniform states in the following sense:

$$(3.11) \quad \sum_{k=1}^{N_f} s_k^\kappa \mathbf{n}_k^\kappa = 0,$$

which is a discrete version of  $\frac{1}{|\kappa|} \oint_{\partial\kappa} \mathbf{n} dS = 0$  for a closed contour. Relation (3.11) is closely related to the discrete geometric conservation laws [43] and is required for the numerical scheme to preserve free-stream states [29]. In section 5 we will present schemes that satisfies assumptions (3.1), (3.3) and (3.11).

**3.2. The pseudo-equilibrium state.** We first introduce the Rusanov flux [39]

$$(3.12) \quad \mathbf{h}_\lambda(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) = \frac{\mathbf{f}(\mathbf{u}_L) \cdot \mathbf{n} + \mathbf{f}(\mathbf{u}_R) \cdot \mathbf{n}}{2} - \frac{\lambda}{2}(\mathbf{u}_R - \mathbf{u}_L),$$

for  $\lambda \geq |\lambda|(\mathbf{u}_L, \mathbf{u}_R, \mathbf{n})$  defined in (2.2). Note that the Rusanov flux is derived from the following ARS:

$$\mathcal{W}_\lambda(\xi, \mathbf{u}_L, \mathbf{u}_R, \mathbf{n}) = \begin{cases} \mathbf{u}_L, & \xi < -\lambda, \\ \frac{\mathbf{u}_L + \mathbf{u}_R}{2} - \frac{1}{2\lambda}(\mathbf{f}(\mathbf{u}_R) \cdot \mathbf{n} - \mathbf{f}(\mathbf{u}_L) \cdot \mathbf{n}), & -\lambda < \xi < \lambda, \\ \mathbf{u}_R, & \lambda < \xi, \end{cases}$$

so from (2.3) we know that it is IDP, see also [13]. We now state a result that will allow us to rewrite (3.1) with updates of three-point schemes.

**LEMMA 3.4** (pseudo-equilibrium state). *Suppose that the numerical scheme satisfies (3.1), (3.3) and (3.11). Let  $\mathcal{B}$  be a invariant domain and suppose that the internal traces  $\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)})_{1 \leq k \leq N_f}$  are in  $\mathcal{B}$ , then there exists  $\mathbf{u}_\kappa^* = \mathbf{u}_\kappa^*(t^{(n)})$  in  $\mathcal{B}$  and  $\lambda_\kappa^* = \lambda_\kappa^*(t^{(n)}) > 0$  finite such that*

$$(3.13) \quad \sum_{k=1}^{N_f} s_k^\kappa \mathbf{h}_{\lambda_\kappa^*}(\mathbf{u}_\kappa^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa) = 0,$$

with  $\mathbf{h}_{\lambda_\kappa^*}$  defined in (3.12) with  $\lambda = \lambda_\kappa^*$  where

$$(3.14) \quad \lambda_\kappa^* \geq \max_{1 \leq k \leq N_f} (|\lambda|(\mathbf{u}_\kappa^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa)),$$

and the pseudo-equilibrium state is defined by

$$(3.15) \quad \mathbf{u}_\kappa^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \left( \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}) - \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)})) \cdot \mathbf{n}_k^\kappa}{\lambda_\kappa^*} \right), \quad \tilde{\gamma}_k^\kappa := \frac{s_k^\kappa}{S^\kappa}.$$

*Proof.* For the sake of clarity, we remove the time dependance of  $\mathbf{u}_h$  since all evaluations are done at  $t^{(n)}$  and write  $\mathbf{u}_h^-(\mathbf{x}_k^\kappa)$  for  $\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)})$ . We first remark that from (3.2) and (3.15), we have

$$(3.16) \quad \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa = 1.$$

We introduce the following two sequences:

$$(3.17) \quad \begin{cases} \mathbf{u}_0^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa), & \lambda_0 = \max_{1 \leq k \leq N_f} (|\lambda|(\mathbf{u}_0^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa)), \\ \lambda_{p+1} = \max \left( \lambda_p, \frac{1}{d(\mathbf{u}_p^*, \partial \mathcal{B})} + \max_{1 \leq k \leq N_f} (|\lambda|(\mathbf{u}_p^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa)) \right), & p \geq 0, \\ \mathbf{u}_{p+1}^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \left( \frac{\mathbf{u}_h^-(\mathbf{x}_k^\kappa) + \mathbf{u}_p^*}{2} - \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa)) \cdot \mathbf{n}_k^\kappa - \mathbf{f}(\mathbf{u}_p^*) \cdot \mathbf{n}_k^\kappa}{2\lambda_{p+1}} \right), & p \geq 0, \end{cases}$$

where  $d(\mathbf{u}, \partial \mathcal{B}) = \inf_{\mathbf{v} \in \partial \mathcal{B}} \|\mathbf{u} - \mathbf{v}\|$  is the distance from  $\mathbf{u}$  to the boundary of  $\mathcal{B}$ . We will show that both sequences converge and the limits satisfy (3.13) and (3.14). We first remark that  $(\lambda_p)$  is non decreasing and therefore converges to some  $\lambda_\kappa^*$  in  $\mathbb{R} \cup \{+\infty\}$  and that for all  $p \geq 0$ ,  $\mathbf{u}_p^*$  is in  $\mathcal{B}$  since  $\lambda_{p+1} \geq |\lambda|(\mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{u}_p^*, \mathbf{n}_k^\kappa)$  so we can use (2.3) with  $\Delta t = \frac{1}{2\lambda_{p+1}}$ . Using successively (3.11) and the definition of  $\tilde{\gamma}_k^\kappa$  in (3.15), then (3.16), we have

$$\begin{aligned} \mathbf{u}_{p+1}^* &= \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \left( \frac{\mathbf{u}_h^-(\mathbf{x}_k^\kappa) + \mathbf{u}_p^*}{2} - \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa)) \cdot \mathbf{n}_k^\kappa}{2\lambda_{p+1}} \right) \\ &= \frac{\mathbf{u}_p^*}{2} + \frac{1}{2} \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa) - \frac{1}{2\lambda_{p+1}} \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa)) \cdot \mathbf{n}_k^\kappa. \end{aligned}$$

Then the first step of the sequence (3.17) for  $\mathbf{u}_\kappa^*$  reads

$$\mathbf{u}_1^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa) - \frac{1}{2\lambda_1} \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa)) \cdot \mathbf{n}_k^\kappa,$$

so applying the recurrence relation  $p$  more times, we obtain

$$(3.18) \quad \mathbf{u}_{p+1}^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa) - \sum_{i=1}^{p+1} \frac{1}{2^i \lambda_{(p+2-i)}} \times \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa)) \cdot \mathbf{n}_k^\kappa.$$

Let show that  $\sum_{i=1}^p \frac{1}{2^i \lambda_{(p+1-i)}}$  above converges to  $\frac{1}{\lambda_\kappa^*}$ . Assume  $\lambda_\kappa^*$  is finite and let  $\epsilon > 0$ , since  $\lambda_p$  converges to  $\lambda_\kappa^*$ , there exists  $p_0$  such that  $|\frac{1}{\lambda_p} - \frac{1}{\lambda_\kappa^*}| < \epsilon$  for all  $p > p_0$ .

Then for  $p > p_0$ , we set

$$\begin{aligned} \sum_{i=1}^p \frac{1}{2^i \lambda_{(p+1-i)}} - \frac{1}{\lambda_\kappa^*} &= \sum_{k=1}^p \frac{1}{2^{(p+1-k)} \lambda_k} - \frac{1}{\lambda_\kappa^*} \\ &= \sum_{k=1}^{p_0} \frac{1}{2^{(p+1-k)} \lambda_k} + \sum_{k=p_0+1}^p \frac{1}{2^{(p+1-k)} \lambda_k} - \frac{1}{\lambda_\kappa^*} \\ &= \frac{1}{2^{p+1}} \sum_{k=1}^{p_0} \frac{2^k}{\lambda_k} + \sum_{k=p_0+1}^p \frac{2^k}{2^{p+1}} \left( \frac{1}{\lambda_k} - \frac{1}{\lambda_\kappa^*} \right) - \frac{1}{\lambda_\kappa^*} \left( 1 - \sum_{k=p_0+1}^p \frac{2^k}{2^{p+1}} \right) \end{aligned}$$

and using a triangle inequality we obtain

$$\begin{aligned} \left| \sum_{i=1}^p \frac{1}{2^i \lambda_{(p+1-i)}} - \frac{1}{\lambda_\kappa^*} \right| &\leq \frac{1}{2^{(p+1)}} \sum_{k=1}^{p_0} \frac{2^k}{\lambda_k} + \sum_{k=p_0+1}^p \frac{2^k}{2^{p+1}} \left| \frac{1}{\lambda_k} - \frac{1}{\lambda_\kappa^*} \right| \\ &\quad + \frac{1}{\lambda_\kappa^*} \left( 1 - \sum_{k=p_0+1}^p \frac{2^k}{2^{p+1}} \right) \\ &\leq \frac{1}{2^{(p+1)}} \sum_{k=1}^{p_0} \frac{2^k}{\lambda_k} + \sum_{i=1}^{p-p_0} \frac{1}{2^i} \epsilon + \frac{1}{\lambda_\kappa^*} \left( 1 - \sum_{i=1}^{p-p_0} \frac{1}{2^i} \right) \\ &= \frac{1}{2^{p+1}} \sum_{k=1}^{p_0} \frac{2^k}{\lambda_k} + \left( 1 - \frac{1}{2^{p-p_0}} \right) \epsilon + \frac{1}{2^{p-p_0}} \frac{1}{\lambda_\kappa^*} \end{aligned}$$

which clearly converges to  $\epsilon$  when  $p \rightarrow \infty$ , proving our statement and (3.15).

Let now prove that  $\lambda_\kappa^*$  is finite by contradiction. Suppose that  $\lambda_p \rightarrow +\infty$ , then  $\mathbf{u}_p^* \rightarrow \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa)$ , but the application

$$\mathbf{u} \mapsto \max_{1 \leq k \leq N_f} (|\lambda|(\mathbf{u}, \mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa)) + \frac{1}{d(\mathbf{u}, \partial B)}$$

is continuous in the interior of  $\mathcal{B}$  and is hence locally bounded around  $\sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa)$ , implying that  $\lambda_p$  is bounded and  $\lambda_\kappa^*$  is finite which is a contradiction. By (3.17), we obviously have  $\lambda_{p+1} \geq \max_{1 \leq k \leq N_f} (|\lambda|(\mathbf{u}_p^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa))$  and passing the inequality to the limit we obtain (3.14).

Let now prove that  $\mathbf{u}_\kappa^*$  is in  $\mathcal{B}$ , since for all  $p \geq 0$ ,  $\mathbf{u}_p^*$  is in  $\mathcal{B}$ , we already know that  $\mathbf{u}_\kappa^*$  is in the closure  $\bar{\mathcal{B}}$ . Now by contradiction, assuming that  $\mathbf{u}_\kappa^*$  is not in  $\mathcal{B}$ , we necessarily have  $\mathbf{u}_\kappa^*$  in  $\partial \mathcal{B}$ , so  $d(\mathbf{u}_\kappa^*, \partial \mathcal{B}) \rightarrow 0$  inducing  $\lambda_p \rightarrow +\infty$  by (3.17) which a contradiction. It remains to prove (3.13). Using (3.11), we add  $\frac{1}{\lambda_\kappa^*} \mathbf{f}(\mathbf{u}_\kappa^*) \cdot \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{n}_k^\kappa = 0$  to (3.15) and get

$$\mathbf{u}_\kappa^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \left( \mathbf{u}_h^-(\mathbf{x}_k^\kappa) - \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa)) \cdot \mathbf{n}_k^\kappa + \mathbf{f}(\mathbf{u}_\kappa^*) \cdot \mathbf{n}_k^\kappa}{\lambda_\kappa^*} \right).$$

Moving  $\mathbf{u}_\kappa^*$  to the right-hand side and using (3.16) we obtain

$$\sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \left( \mathbf{u}_h^-(\mathbf{x}_k^\kappa) - \mathbf{u}_\kappa^* - \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa)) \cdot \mathbf{n}_k^\kappa + \mathbf{f}(\mathbf{u}_\kappa^*) \cdot \mathbf{n}_k^\kappa}{\lambda_\kappa^*} \right) = 0,$$

and multiplying the above quantity by  $-\frac{\lambda_\kappa^*}{2} \sum_{i=1}^{N_f} s_i^\kappa = -\frac{\lambda_\kappa^*}{2} \mathcal{S}^\kappa$  we finally get

$$\sum_{k=1}^{N_f} s_k^\kappa \left( \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa)) \cdot \mathbf{n}_k^\kappa + \mathbf{f}(\mathbf{u}_\kappa^*) \cdot \mathbf{n}_k^\kappa}{2} - \lambda_\kappa^* \frac{(\mathbf{u}_h^-(\mathbf{x}_k^\kappa) - \mathbf{u}_\kappa^*)}{2} \right) = 0,$$

which is exactly (3.13) from the definition of the Rusanov flux (3.12).  $\square$

**3.3. Invariant domain preserving schemes.** Using Lemma 3.4 we now state and prove the main result of this work in the theorem below.

**THEOREM 3.5** (Time step condition). *Assume that the numerical scheme satisfies (3.1), (3.3) and (3.11) and assume that  $u_h^\pm(\mathbf{x}_k^\kappa, t^{(n)})_{1 \leq k \leq N_f}$  and  $u_h(\mathbf{y}_i^\kappa, t^{(n)})_{1 \leq i \leq N_v}$  are in  $\mathcal{B}$ , then under the following condition on the time step*

$$(3.19) \quad \Delta t^{(n)} \max_{\kappa \in \mathcal{T}_h} \max_{1 \leq k \leq N_f} \frac{s_k^\kappa}{\beta_k^\kappa} \max \left( \lambda_\kappa^*, |\lambda| (\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa) \right) \leq \frac{1}{2},$$

where  $\lambda_\kappa^*$  is defined by (3.14),  $\langle u_h^{(n+1)} \rangle_\kappa$  is also in  $\mathcal{B}$ .

*Proof.* Once again we remove the time dependance of  $\mathbf{u}_h$  for the sake of clarity, except when explicitly needed in the evaluation of  $\langle \mathbf{u}_h^{(n+1)} \rangle_\kappa$ . Using Lemma 3.4 we add the trivial quantity  $\Delta t^{(n)} \times$  (3.13) to (3.1) and use (3.3) to get

$$\begin{aligned} \langle \mathbf{u}_h^{(n+1)} \rangle_\kappa &= \langle \mathbf{u}_h^{(n)} \rangle_\kappa - \Delta t^{(n)} \sum_{k=1}^{N_f} s_k^\kappa \left( \mathbf{h}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{u}_h^+(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa) - \mathbf{h}_{\lambda_\kappa^*}(\mathbf{u}_\kappa^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa) \right) \\ &= \sum_{i=1}^{N_v} \nu_i \mathbf{u}_h(\mathbf{y}_i^\kappa) + \sum_{k=1}^{N_f} \left( \beta_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa) \right. \\ &\quad \left. - \Delta t^{(n)} s_k^\kappa \left( \mathbf{h}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{u}_h^+(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa) - \mathbf{h}_{\lambda_\kappa^*}(\mathbf{u}_\kappa^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa) \right) \right) \\ (3.20) \quad &= \sum_{i=1}^{N_v} \nu_i \mathbf{u}_h(\mathbf{y}_i^\kappa) + \sum_{k=1}^{N_f} \beta_k^\kappa \mathcal{U}_k^{\kappa, n}, \end{aligned}$$

where, from Lemma 2.5 and the condition (3.19), the updates

$$(3.21) \quad \mathcal{U}_k^{\kappa, n} := \mathbf{u}_h^-(\mathbf{x}_k^\kappa) - \frac{\Delta t^{(n)} s_k^\kappa}{\beta_k^\kappa} \left( \mathbf{h}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{u}_h^+(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa) - \mathbf{h}_{\lambda_\kappa^*}(\mathbf{u}_\kappa^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa), \mathbf{n}_k^\kappa) \right),$$

are in  $\mathcal{B}$ . Then by (3.4)  $\langle \mathbf{u}_h^{(n+1)} \rangle_\kappa$  is a convex combination of quantities in  $\mathcal{B}$ , which concludes the proof.  $\square$

*Remark 3.6.* Since the set of states  $\Omega^a$  is in general a convex invariant domain, Theorem 3.5 can then be applied to  $\mathcal{B} = \Omega^a$  to ensure robustness of the scheme.

*Remark 3.7.* In the case where we are using the quadrature on the volume defined by Lemma 3.1, from (3.8) and (3.9) and the definition  $\varepsilon_\kappa = \min_{\{i: \alpha_i^\kappa > 0\}} \left( \frac{\varpi_i^\kappa}{\alpha_i^\kappa} \right)$ , we have

$$(3.22) \quad \frac{s_k^\kappa}{\beta_k^\kappa} = \frac{1}{\varepsilon_\kappa} = \max_{\{i: \alpha_i^\kappa > 0\}} \left( \frac{\alpha_i^\kappa}{\varpi_i^\kappa} \right) = \max_{1 \leq i \leq N_v} \sum_{j=1}^{N_p} \sum_{l=1}^{N_f} s_l^\kappa \phi_j(\mathbf{x}_l^\kappa) \phi_j(\mathbf{y}_i^\kappa) \quad \forall 1 \leq k \leq N_f,$$

and the CFL condition (3.19) now reads

$$(3.23) \quad \Delta t^{(n)} \max_{\kappa \in \mathcal{T}_h} \frac{1}{\varepsilon_\kappa} \max \left( \lambda_\kappa^*, |\lambda| (\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa) \right) \leq \frac{1}{2}.$$

**3.4. Limiting strategy.** Convex limiting enforces the numerical solution to preserve invariant domains [21] through quasiconcave constraints [1]. Let recall that a function  $\psi : \mathcal{B} \rightarrow \mathbb{R}$  is quasiconcave iff. for every family of convex coefficients  $(\lambda_i) \geq 0$ , with  $\sum \lambda_i = 1$ , we have  $\psi(\sum_i \lambda_i \mathbf{u}_i) \geq \min_i \psi(\mathbf{u}_i)$  for all  $\mathbf{u}_i$  in  $\mathcal{B}$ . From [Theorem 3.5](#) we see that for any quasiconcave function  $\psi$  we have

$$(3.24) \quad \psi(\langle \mathbf{u}_h^{(n+1)} \rangle_\kappa) \geq m_\kappa^\psi := \min \left( \psi(\mathbf{u}_h(\mathbf{y}_i^\kappa, t^{(n)}))_{1 \leq i \leq N_v}, \psi(\mathbf{U}_k^{\kappa, n})_{1 \leq k \leq N_f} \right),$$

where the updates  $\mathbf{U}_k^{\kappa, n}$  are defined in (3.21). We now limit the solution around its cell-average  $\langle \mathbf{u}_h^{(n+1)} \rangle_\kappa$  so that it satisfies the same bounds and we rely on scaling limiters introduced in [50, 49] to enforce the bounds from quasiconcave functions to points where  $\mathbf{u}_h$  needs to be evaluated. The limited solution is thus defined as

$$(3.25) \quad \tilde{\mathbf{u}}_h^{(n+1)} \equiv (1 - \theta_\kappa) \mathbf{u}_h^{(n+1)} + \theta_\kappa \langle \mathbf{u}_h^{(n+1)} \rangle_\kappa,$$

where

$$\theta_\kappa = \min_{\mathbf{z} \in (\mathbf{y}_{1 \leq i \leq N_v}^\kappa) \cup (\mathbf{x}_{1 \leq i \leq N_f}^\kappa)} \max \{ 0 \leq t \leq 1 : \psi((1-t)\mathbf{u}_h(\mathbf{z}, t^{(n+1)}) + t\langle \mathbf{u}_h^{(n+1)} \rangle_\kappa) \geq m_\kappa^\psi \}.$$

This strategy may be applied to a finite family  $(\psi_i)_{1 \leq i \leq n_c}$  of  $n_c$  quasiconcave functions by using the minimum value  $\theta_\kappa = \min\{\theta_\kappa(\psi_i) : 1 \leq i \leq n_c\}$ . The limiter (3.25) is then applied locally to each cell  $\kappa$  and preserves high-order accuracy of the scheme in smooth domains [50]. The cell-average is not modified,  $\langle \tilde{\mathbf{u}}_h^{(n+1)} \rangle_\kappa = \langle \mathbf{u}_h^{(n+1)} \rangle_\kappa$ , and cellwise discrete conservation (3.1) still holds.

**3.5. Practical evaluation of  $\mathbf{u}_\kappa^*$ .** The evaluation of  $m_\kappa^\psi$  in (3.24) requires to first evaluate both  $\lambda_\kappa^*$  and  $\mathbf{u}_\kappa^*$  as limits of the sequences in (3.17) which might be cumbersome. Here we propose another strategy where we only need to check the convergence of  $(\lambda_p)$  to ensure that  $(\mathbf{u}_p^*)$  has also converged to  $\mathbf{u}_\kappa^*$  which is in  $\mathcal{B}$ . Let introduce the sequences

$$(3.26) \quad \begin{cases} \lambda_0 = \frac{1}{\vartheta} \max_{1 \leq k \leq N_f} |\lambda|(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa), \mathbf{u}_0^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \\ \lambda_{p+1} = \max \left( \lambda_p, \frac{1}{\vartheta} \max_{1 \leq k \leq N_f} (|\lambda|(\mathbf{u}_p^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa)) \right), \quad p \geq 0, \\ \mathbf{u}_{p+1}^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \left( \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}) - \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)})) \cdot \mathbf{n}_k^\kappa}{\lambda_{p+1}} \right), \quad p \geq 0, \end{cases}$$

where

$$\vartheta = \max \left\{ 0 \leq t \leq 1 : \mathbf{u}_0^* - \frac{t}{\tilde{\lambda}_1} \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)})) \cdot \mathbf{n}_k^\kappa \in \mathcal{B} \right\},$$

$$\tilde{\lambda}_1 = \max_{1 \leq k \leq N_f} \left( |\lambda|(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa), |\lambda|(\mathbf{u}_0^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa) \right).$$

Since  $\mathbf{u}_0^* \in \mathcal{B}$ , we have  $0 < \vartheta \leq 1$  and then  $\mathbf{u}_1^* \in \mathcal{B}$  by definition of  $\vartheta$ . Now  $\lambda_p$  is increasing and  $\mathbf{u}_{p+1}^* \in [\mathbf{u}_0^*, \mathbf{u}_p^*] \subset [\mathbf{u}_0^*, \mathbf{u}_1^*]$  for all  $p \geq 1$  which is enough to prove convergence of both sequences and ensure that  $\mathbf{u}_p^* \in \mathcal{B}$  and since  $[\mathbf{u}_0^*, \mathbf{u}_1^*]$  is closed the limit is also in  $\mathcal{B}$  (no need to add the distance term as in (3.17)). Obviously, the limits satisfy (3.13), (3.14) and (3.15). But this time, if  $\lambda_{p+1} = \lambda_p$  for  $p \geq 1$ , then

$\mathbf{u}_{p+1}^* = \mathbf{u}_p^*$  and from this point both sequences are stationary. Now, the evaluation of  $\mathbf{u}_p^*$  is really fast and we only need to evaluate  $\max_{1 \leq k \leq N_f} (|\lambda|(\mathbf{u}_p^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa))$ .

It is possible to use a local  $\lambda_\kappa^{*k}$  at each vertex  $\mathbf{x}_\kappa^k$  in  $\partial\kappa$  in the above algorithm to lower the artificial dissipation of the Rusanov flux (3.12) and thus avoid the induced restriction on the time step as well as to reduce over-diffusion of the updates  $\mathcal{U}_k^{\kappa, n}$  in (3.21). We thus look for  $\lambda_\kappa^{*k}$ ,  $1 \leq k \leq N_f$ , and  $\mathbf{u}_\kappa^*$  satisfying

$$(3.27) \quad \sum_{k=1}^{N_f} s_k^\kappa \mathbf{h}_{\lambda_\kappa^{*k}}(\mathbf{u}_\kappa^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa) = 0,$$

and

$$(3.28) \quad \lambda_\kappa^{*k} \geq |\lambda|(\mathbf{u}_\kappa^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa).$$

We therefore introduce new sequences as

$$(3.29) \quad \begin{cases} \lambda_0^k = \frac{1}{\vartheta} |\lambda|(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa), & \mathbf{u}_0^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \\ \lambda_{p+1}^k = \max\left(\lambda_p^k, \frac{1}{\vartheta} |\lambda|(\mathbf{u}_p^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa)\right), & p \geq 0, \\ \mathbf{u}_{p+1}^* = \sum_{k=1}^{N_f} \frac{\tilde{\gamma}_k^\kappa \lambda_{p+1}^k}{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_{p+1}^i} \left( \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}) - \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)})) \cdot \mathbf{n}_k^\kappa}{\tilde{\lambda}_1^k} \right), & p \geq 0, \end{cases}$$

where the index  $k$  ranges from 1 to  $N_f$  and

$$\vartheta = \max \left\{ 0 \leq t \leq 1 : \sum_{k=1}^{N_f} \frac{\tilde{\gamma}_k^\kappa \lambda_1^k}{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_1^i} \left( \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}) - t \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)})) \cdot \mathbf{n}_k^\kappa}{\tilde{\lambda}_1^k} \right) \in \mathcal{B} \right\},$$

$$\tilde{\lambda}_1^k = \max \left( |\lambda|(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa), |\lambda|(\mathbf{u}_0^*, \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{n}_k^\kappa) \right).$$

Now all the sequences  $(\lambda_p^k)_p$  are non-decreasing and will converge. Also for  $p \geq 1$ , we compute

$$\left( \sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_{p+1}^i \right) \mathbf{u}_{p+1}^* - \left( \sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_p^i \right) \mathbf{u}_p^* = \sum_{k=1}^{N_f} \tilde{\gamma}_k^\kappa (\lambda_{p+1}^k - \lambda_p^k) \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}),$$

so  $\mathbf{u}_{p+1}^*$  can be recast as a convex combination

$$\mathbf{u}_{p+1}^* = \frac{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_p^i}{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_{p+1}^i} \mathbf{u}_p^* + \sum_{k=1}^{N_f} \frac{\tilde{\gamma}_k^\kappa (\lambda_{p+1}^k - \lambda_p^k)}{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_{p+1}^i} \mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}).$$

with  $N_f + 1$  positive weights such that

$$\frac{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_p^i}{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_{p+1}^i} + \sum_{k=1}^{N_f} \frac{\tilde{\gamma}_k^\kappa (\lambda_{p+1}^k - \lambda_p^k)}{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_{p+1}^i} = 1,$$

and by recurrence  $\mathbf{u}_{p+1}^*$  is also a convex combination of  $\mathbf{u}_1^*$  and the  $\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)})$ . Therefore, if  $\mathbf{u}_1^*$  is in  $B$ , then  $\mathbf{u}_p^*$  is also in  $\mathcal{B}$  for all  $p \geq 1$  and stays in a compact

subset of  $\mathcal{B}$  which ensures that the  $\lambda_p^k$  are bounded. So they converge to some finite  $\lambda_{\kappa}^{*k}$  and  $\mathbf{u}_p^*$  converges to

$$(3.30) \quad \mathbf{u}_{\kappa}^* = \sum_{k=1}^{N_f} \frac{\tilde{\gamma}_k^{\kappa} \lambda_{\kappa}^{*k}}{\sum_{i=1}^{N_f} \tilde{\gamma}_i \lambda_{\kappa}^{*i}} \left( \mathbf{u}_h^-(\mathbf{x}_k^{\kappa}, t^{(n)}) - \frac{\mathbf{f}(\mathbf{u}_h^-(\mathbf{x}_k^{\kappa}, t^{(n)})) \cdot \mathbf{n}_k}{\lambda_{\kappa}^{*k}} \right) \in \mathcal{B}.$$

By definition of the  $\lambda_p^k$ , (3.28) holds and with the definition of  $\mathbf{u}_{\kappa}^*$ , (3.27) is also satisfied, while Theorem 3.5 still holds with  $\lambda_{\kappa}^{*k}$  instead of  $\lambda_{\kappa}^*$  in (3.19). We summarize these results in the following lemma.

**LEMMA 3.8.** *Suppose that the numerical scheme satisfies (3.1), (3.3) and (3.11), let  $\mathcal{B}$  be a invariant domain and suppose that for all  $1 \leq i \leq N_f$ ,  $\mathbf{u}_h^{\pm}(x_i^{\kappa}, t^{(n)})$  are in  $\mathcal{B}$ , then  $\mathbf{u}_{\kappa}^*$  defined by (3.30) is in  $\mathcal{B}$ , satisfies (3.27), and there exists a family of finite and positive estimates  $(\lambda_{\kappa}^{*k})_{1 \leq k \leq N_f}$  satisfying (3.28).*

**4. Examples of high-order spectral discontinuous methods.** We here review some high-order spectral discontinuous approximations of (1.1) which satisfy the assumptions of discrete conservation (3.1), existence of quadrature rule (3.3) and preservation of uniform states (3.11). As a consequence there exists pseudo-equilibrium states  $\mathbf{u}_{\kappa}^*$  such that Lemma 3.4, Lemma 3.8, and Theorem 3.5 hold, and the limiter (3.25) can be applied. In the following, we consider a partition  $\mathcal{T}_h$  of  $D \subset \mathbb{R}^d$ , composed of non-overlapping and non-empty elements  $\kappa$ , and by  $\mathcal{F}_h$  we denote the set of faces in the partition. The approximate solution is sought under the form

$$(4.1) \quad \mathbf{u}_h(\mathbf{x}, t) = \sum_{k=1}^{N_p} \phi_k^{\kappa}(\mathbf{x}) \mathbf{U}_k^{\kappa}(t) \quad \forall \mathbf{x} \in \kappa, \kappa \in \mathcal{T}_h, \forall t \geq 0,$$

where the basis functions  $\phi_k^{\kappa}$  span the function space  $\mathcal{V}_h^p(\kappa)$  restricted onto  $\kappa$  and the discrete scheme may be written as

$$(4.2) \quad M_k^{\kappa} \frac{\mathbf{U}_{k,n+1}^{\kappa} - \mathbf{U}_{k,n}^{\kappa}}{\Delta t^{(n)}} + \mathbf{R}_k^{\kappa}(\mathbf{u}_h^{(n)}) = 0 \quad \forall \kappa \in \mathcal{T}_h, 1 \leq k \leq N_p, n \geq 0,$$

where  $\mathbf{U}_{k,n}^{\kappa} = \mathbf{U}_k^{\kappa}(t^{(n)})$  and the  $M_k^{\kappa}$  are the entries of the mass matrix.

**4.1. Discontinuous Galerkin Method.** The first numerical scheme we describe here is the discontinuous Galerkin method with modal basis [37, 3, 7]. We look for approximate solutions in the function space of discontinuous polynomials

$$\mathcal{V}_h^p = \bigoplus_{\kappa \in \mathcal{T}_h} \mathcal{V}_h^p(\kappa) = \{ \phi \in L^2(D) : \phi|_{\kappa} \circ \mathbf{x}_{\kappa} \in \mathcal{P}^p(\hat{K}) \forall \kappa \in \mathcal{T}_h \},$$

where  $\mathcal{P}^p(\hat{K})$  is a polynomial space over a reference element  $\hat{K}$ . Each physical element  $\kappa$  is the image of  $\hat{K}$  through the mapping  $\mathbf{x} = \mathbf{x}_{\kappa}(\boldsymbol{\xi})$  with  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_d)$ . Likewise, each face  $f$  in  $\mathcal{F}_h$  is the image of a reference face  $\hat{F}$  through the mapping  $\mathbf{x} = \mathbf{x}_f(\xi_1, \dots, \xi_{d-1})$ . We suppose that we have a quadrature  $(\boldsymbol{\xi}_i^V, \omega_i^V)_{1 \leq i \leq N_v}$  on  $\hat{K}$  and denote  $\mathbf{y}_i^{\kappa} = \mathbf{x}_{\kappa}(\boldsymbol{\xi}_i)$  (see Fig. 1). Similarly, we consider a quadrature  $(\boldsymbol{\xi}_k^f, \omega_k^f)_{1 \leq k \leq n_f}$  on  $\hat{F}$  and denote the faces of  $\kappa$   $(f_{\kappa}^j)_{1 \leq j \leq N_{\kappa}}$  where  $N_{\kappa}$  is the number of faces of  $\kappa$  and the  $f_{\kappa}^j \in \mathcal{F}$  are distinct. Then define the  $\mathbf{x}_i^{\kappa}$  in (3.1) by  $\mathbf{x}_{(j-1)n_f+k}^{\kappa} = \mathbf{x}_{f_{\kappa}^j}(\boldsymbol{\xi}_k^f)$  for

$1 \leq j \leq N_\kappa$  and  $1 \leq k \leq n_f$  and  $N_f = N_\kappa \times n_f$ . We further define Jacobians of the transformations by  $J_\kappa(\mathbf{x}) = |\mathbf{x}'_\kappa(\boldsymbol{\xi})|$  and  $J_f(\mathbf{x}) = |\mathbf{x}'_f(\xi_1, \dots, \xi_{d-1})|$ .

The DG method consists in defining a discrete weak formulation of (1.1) by multiplying it with test functions  $\phi_k^\kappa$  spanning  $\mathcal{V}_h^p$  and integrating over  $\kappa$ , using integration by parts and approximating  $\mathbf{f}(\mathbf{u}_h) \cdot \mathbf{n}$  by two-point numerical fluxes and the integrals by the quadrature rules. The space discretization in (4.2) reads

$$(4.3) \quad \mathbf{R}_k^\kappa(\mathbf{u}_h) = - \sum_{i=1}^{N_v} \omega_i^V J_\kappa(\mathbf{y}_i^\kappa) \mathbf{f}(\mathbf{u}_h(\mathbf{y}_i^\kappa, t^{(n)})) \cdot \nabla \phi_k^\kappa(\mathbf{y}_i^\kappa) \\ + \sum_{i=1}^{N_f} \omega_i^f J_f(\mathbf{x}_i^\kappa) \mathbf{h}(\mathbf{u}_h^-(\mathbf{x}_i^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_i^\kappa, t^{(n)}), \mathbf{n}_i^\kappa) \phi_k^\kappa(\mathbf{x}_i^\kappa).$$

We use an orthonormal basis such that  $M_\kappa^k = |\kappa| := \sum_{i=1}^{N_v} \omega_i^V J_\kappa(\mathbf{y}_i^\kappa)$  and further setting  $\phi_k^\kappa = 1_\kappa$  the indicator function of  $\kappa$ , the first sum vanishes and we obtain (3.1) with  $s_k^\kappa = \frac{\omega_k^f J_f(\mathbf{x}_k^\kappa)}{|\kappa|}$  and  $\beta_k^\kappa = \varepsilon_\kappa s_k^\kappa$  in (3.3) where  $\varepsilon_\kappa$  is evaluated from (3.22). Then by (3.23), the scheme (4.2) and (4.3) is IDP under the condition

$$(4.4) \quad \Delta t^{(n)} \max_{\kappa \in \mathcal{T}_h} \frac{1}{\varepsilon_\kappa} \max_{1 \leq k \leq N_f} \max \left( \lambda_\kappa^*, |\lambda|(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)})) \right) \leq \frac{1}{2}.$$

**4.2. Discontinuous Galerkin Spectral Element Method.** In the DGSEM, the reference element is an hypercube :  $\hat{K} = I^d := \{\boldsymbol{\xi} = (\xi_1, \dots, \xi_d) : -1 \leq \xi_j \leq 1\}$  and the polynomial space  $\mathcal{P}^p(I^d)$  is formed by tensor products of polynomials of degree at most  $p$  in each direction. The approximate solution is sought under the form (4.1) where  $(\mathbf{U}_k^\kappa)_{1 \leq k \leq N_p}$  are the  $N_p = (p+1)^d$  DOFs in the element  $\kappa$  with indexing

$$k = \bar{k}(i_1, \dots, i_d) := 1 + \sum_{j=1}^d i_j (p+1)^{j-1} \quad 0 \leq i_1, \dots, i_d \leq p.$$

We define a basis  $(\phi_k^\kappa)_{1 \leq k \leq N_p}$  of  $\mathcal{V}_h^p(\kappa)$  by using tensor products:  $\phi_k^\kappa(\mathbf{x}) = \phi_k^\kappa(\mathbf{x}_\kappa(\boldsymbol{\xi})) = \prod_{j=1}^d \ell_{i_j}(\xi_j)$ , where  $\ell_{0 \leq i \leq p}$  denote the  $i$ th Lagrange interpolation polynomial associated to  $\zeta_i$  the  $i$ th Gauss-Lobatto quadrature node with  $\zeta_0 = -1 < \zeta_1 < \dots < \zeta_p = 1$  (see Fig. 2) and by  $\omega_i$  we denote the associated weight. We therefore have the following cardinality relation at quadrature points  $\boldsymbol{\xi}_{k'} = (\xi_{i'_1}, \dots, \xi_{i'_d})$  in  $\hat{K}$ :  $\phi_k^\kappa(\mathbf{x}_{k'}) = \phi_k^\kappa(\mathbf{x}_\kappa(\boldsymbol{\xi}_{k'})) = \delta_{i_1, i'_1} \dots \delta_{i_d, i'_d}$  with  $\delta_{i, i'}$  the Kronecker symbol, so the DOFs correspond to the point values of the solution:  $\mathbf{U}_k^\kappa(t) = \mathbf{u}_h(\mathbf{y}_k^\kappa, t)$  and interpolation and quadrature points are collocated, hence  $N_v = N_p$ .

Let introduce the discrete derivative matrix with entries  $D_{ij} = \ell'_j(\zeta_i)$  with  $0 \leq i, j \leq p$ . The DGSEM discretization takes the form (4.2) with  $M_k^\kappa = \omega_k^V J_\kappa(\mathbf{k}_k^\kappa)$ ,  $\omega_k^V = \prod_{j=1}^d \omega_{i_j}$  for  $k = \bar{k}(i_1, \dots, i_d)$  and

$$(4.5) \quad \mathbf{R}_k^\kappa(\mathbf{u}_h) = 2\omega_k^V \sum_{j=1}^d \sum_{l=0}^p D_{ijl} \mathbf{h}_{sym}(\mathbf{U}_k^\kappa, \mathbf{U}_{k'_j}^\kappa, \{J_\kappa \nabla \xi_j\}_{(k, k'_j)}) \\ + \sum_{i=1}^{N_f} \phi_k^\kappa(\mathbf{x}_i^\kappa) \omega_i^f J_f(\mathbf{x}_i^\kappa) \left( \mathbf{h}(\mathbf{U}_{k, n}^\kappa, \mathbf{u}_h^+(\mathbf{x}_i^\kappa, t), \mathbf{n}_i^\kappa) - \mathbf{f}(\mathbf{U}_{k, n}^\kappa) \cdot \mathbf{n}_i^\kappa \right)$$

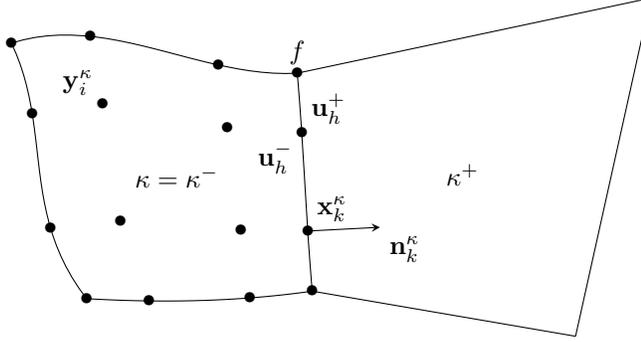


Fig. 2: Notations for the DGSEM and  $d = 2$ : inner and outer elements,  $\kappa^-$  and  $\kappa^+$ , for  $d = 2$ ; definitions of traces  $\mathbf{u}_h^\pm$  on the interface  $f$  and of the unit outward normal vector  $\mathbf{n}_k^\kappa$ . Element quadrature nodes  $\mathbf{y}_i^\kappa$  and surface quadrature node  $\mathbf{x}_k^\kappa$  that are also included in the  $\mathbf{y}_i^\kappa$ .

where for  $d = 3$   $k'_1 = \bar{k}(l, i_2, i_3)$ ,  $k'_2 = \bar{k}(i_1, l, i_3)$  and  $k'_3 = \bar{k}(i_1, i_2, l)$ , hence  $k'_j = k + (i'_j - i_j)(p + 1)^{j-1}$ ,  $N_f = 2d(p + 1)^{d-1}$ ,  $\omega_i^f = \prod_{j=1}^{d-1} \omega_{i_j}$ , and

$$\{J_\kappa \nabla \xi\}_{(k, k'_j)} = \frac{1}{2} (J_\kappa(\mathbf{y}_k^\kappa) \nabla \xi_j(\boldsymbol{\xi}_k) + J_\kappa(\mathbf{y}_{k'_j}^\kappa) \nabla \xi_j(\boldsymbol{\xi}_{k'_j})),$$

have been introduced to keep conservation of the scheme [46]. By  $\mathbf{h}_{sym}$  we denote a two-point flux supposed to be symmetric in the sense that  $\mathbf{h}_{sym}(\mathbf{u}, \mathbf{v}, \mathbf{n}) = \mathbf{h}_{sym}(\mathbf{v}, \mathbf{u}, \mathbf{n})$ . Note that  $\phi_k^\kappa(\mathbf{x}_i^\kappa) = 1$  if  $\mathbf{x}_i^\kappa = \mathbf{y}_k^\kappa$  and  $\phi_k^\kappa(\mathbf{x}_i^\kappa) = 0$  else.

The choice of the  $\mathbf{y}_k^\kappa$  in the quadrature (3.3) is not unique and we here use the following decomposition

$$\begin{aligned} \langle \mathbf{u}_h^{(n)} \rangle_\kappa &= \frac{1}{|\kappa|} \sum_{\mathbf{y}_i^\kappa \in \kappa} \omega_i^V J_\kappa(\mathbf{y}_i^\kappa) \mathbf{U}_{i,n}^\kappa \\ &= \frac{1}{|\kappa|} \sum_{\mathbf{y}_i^\kappa \in \text{int}(\kappa)} \omega_i^V J_\kappa(\mathbf{y}_i^\kappa) \mathbf{U}_{i,n}^\kappa + \frac{1}{|\kappa|} \sum_{i=1}^{N_f} \tilde{\omega}_i^f J_\kappa(\mathbf{x}_i^\kappa) \mathbf{u}_h(\mathbf{x}_i^\kappa, t^{(n)}), \end{aligned}$$

where  $\text{int}(\kappa)$  denotes the interior of  $\kappa$ , while  $\tilde{\omega}_i^f J_\kappa(\mathbf{x}_i^\kappa) = \frac{1}{d} \omega_j^V J_\kappa(\mathbf{y}_j^\kappa)$  if  $\mathbf{y}_j^\kappa = \mathbf{x}_i^\kappa$  is a vertex of the  $d$ -dimensional hexahedron,  $\tilde{\omega}_i^f J_\kappa(\mathbf{x}_i^\kappa) = \frac{1}{d-1} \omega_j^V J_\kappa(\mathbf{y}_j^\kappa)$  if  $\mathbf{y}_j^\kappa = \mathbf{x}_i^\kappa$  is on some edge, and  $\tilde{\omega}_i^f J_\kappa(\mathbf{x}_i^\kappa) = \omega_j^V J_\kappa(\mathbf{y}_j^\kappa)$  else.

Summing (4.2) over  $1 \leq k \leq N_v$  gives for the cell-averaged solution

$$\begin{aligned} \langle \mathbf{u}_h^{(n+1)} \rangle_\kappa &= \langle \mathbf{u}_h^{(n)} \rangle_\kappa - \frac{\Delta t^{(n)}}{|\kappa|} \sum_{k=1}^{N_v} \mathbf{R}_k^\kappa(\mathbf{u}_h^{(n)}) \\ &= \langle \mathbf{u}_h^{(n)} \rangle_\kappa - \frac{\Delta t^{(n)}}{|\kappa|} \sum_{i=1}^{N_f} \omega_i^f J_f(\mathbf{x}_i^\kappa) \mathbf{h}(\mathbf{u}_h^-(\mathbf{x}_i^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_i^\kappa, t^{(n)}), \mathbf{n}_i^\kappa), \end{aligned}$$

by conservation of the DGSEM [15, 46] and providing that the so-called metric identities are satisfied at the discrete level [29]. This relation can be identified with (3.1)

with  $s_k^\kappa = \frac{\omega_k^f J_f(\mathbf{x}_k^\kappa)}{|\kappa|}$ . Then we can apply [Theorem 3.5](#) with  $\beta_k^\kappa = \frac{\tilde{\omega}_k^f J_\kappa(\mathbf{x}_k^\kappa)}{|\kappa|}$  and the DGSEM scheme [\(4.2\)](#) is IDP under the condition

$$(4.6) \quad \Delta t^{(n)} \max_{\kappa \in \Omega_h} \max_{1 \leq k \leq N_f} \frac{\omega_k^f J_f(\mathbf{x}_k^\kappa)}{\tilde{\omega}_k^f J_\kappa(\mathbf{x}_k^\kappa)} \max \left( \lambda_\kappa^*, |\lambda|(\mathbf{u}_h^-(\mathbf{x}_k^\kappa, t^{(n)}), \mathbf{u}_h^+(\mathbf{x}_k^\kappa, t^{(n)})) \right) \leq \frac{1}{2}.$$

**4.3. Other methods.** Properties [\(3.1\)](#), [\(3.3\)](#), and [\(3.11\)](#) also hold for other discretely conservative spectral difference methods on general curved elements provided the discretization operators satisfy the metric identities at the discrete level, which imposes some limits on the order of approximation of the mesh elements compared to the approximation order of the solution [\[29\]](#). The limiter [\(3.25\)](#) may hence be applied to make these methods invariant domain preserving. We list some examples below.

The skew-symmetric entropy stable modal discontinuous Galerkin methods [\[6\]](#) uses skew-hybridized summation-by-parts (SBP) operators allowing conservation and free-stream preservation under the standard accuracy requirements of volume and surface quadratures. In [\[8\]](#) multidimensional discretization schemes based on SBP operators on general curved elements generalize the staggered finite differences from [\[11\]](#). The discretizations with curved elements remain accurate, conservative, and entropy stable. Spectral differences [\[33\]](#) and staggered Chebyshev [\[30\]](#) methods on curved elements belong to a same family of conservative approximations satisfying the discrete metric identities. The methods use two sets of interpolation points for the solution and fluxes and impose the discrete residuals to be satisfied at solution points. Flux points contain points at faces of the elements where two-point numerical fluxes are used. Then, the space derivatives at solution points are evaluated by differencing the polynomials interpolating the fluxes.

**5. Numerical experiments.** Let consider the compressible Euler equations of gas dynamics. The conservative variables and fluxes in [\(1.1\)](#) are

$$(5.1) \quad \mathbf{u} = \begin{pmatrix} \rho \\ \rho \mathbf{v} \\ \rho E \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} \rho \mathbf{v}^\top \\ \rho \mathbf{v} \mathbf{v}^\top + p \mathbf{I}_d \\ (\rho E + p) \mathbf{v}^\top \end{pmatrix},$$

where  $\rho$ ,  $\mathbf{v}$ , and  $E$  denote the density, velocity vector, and specific total energy, respectively. The system is closed by defining the equation of state  $\mathbf{p} = \mathbf{p}(\frac{1}{\rho}, e)$  with  $e = E - \frac{1}{2} \mathbf{v} \cdot \mathbf{v}$  the specific internal energy and the system is hyperbolic over the set of states  $\Omega^a = \{\mathbf{u} \in \mathbb{R}^{d+2} : \rho > 0, \mathbf{v} \in \mathbb{R}^d, e > 0\}$ . We focus here on the polytropic ideal gas law  $\mathbf{p} = (\gamma - 1)\rho e$  where  $\gamma = \frac{C_p}{C_v} = \frac{7}{5}$  is the ratio of specific heats. The compressible Euler equations [\(1.1\)](#) and [\(5.1\)](#) possess the natural entropy – entropy flux pair

$$\eta = -\rho s, \quad \mathbf{q} = -\rho s \mathbf{v}, \quad s = C_v \ln \left( \frac{p}{\rho^\gamma} \right),$$

and  $\mathcal{B} = \{\mathbf{u} \in \Omega^a : s(\mathbf{u}) \geq s_0\}$ , with  $s_0$  in  $\mathbb{R}$ , is an invariant domain for [\(1.1\)](#) and [\(5.1\)](#) [\[13\]](#). We use our convex limiting strategy with the quasiconcave functions  $\psi_1 \equiv \rho$  and  $\psi_2 \equiv \rho e$ .

We now test the robustness and efficiency of the CFL condition on simulations of [\(1.1\)](#) and [\(5.1\)](#) with discontinuous solutions on one-dimensional and unstructured two-dimensional grids. We use the modal DG method in [subsection 4.1](#) and the DGSEM in [subsection 4.2](#). The CFL conditions [\(4.4\)](#) and [\(4.6\)](#) guaranty the cell-averaged solution

Table 1: Initial conditions of Riemann problems (2.1) where  $x_0$  indicates the abscissa separating the states.

problem	left state $(\rho_L, u_L, p_L)^\top$	right state $(\rho_R, u_R, p_R)^\top$	$x_0$
Sod [42]	$(1, 0, 1)^\top$	$(0.125, 0, 0.1)^\top$	0
Lax	$(0.445, 0.698, 3.528)^\top$	$(0.5, 0, 0.571)^\top$	0
Toro 4 [44]	$(5.99924, 19.5975, 460.894)^\top$	$(5.99242, -6.19633, 46.0950)^\top$	-0.1

to be IDP and we then apply the limiter (3.25) to further impose the high-order solution to be IDP. We will compare the following limiting strategies: a positivity limiter (POS) which imposes the solution to remain in  $\Omega^a$  thus extending [50] to unstructured grids; an IDP limiter which imposes the solution to remain in the convex hull of the states in (3.20) and computing the time step in (4.4) and (4.6) with either the global wave estimate  $\lambda_\kappa^*$  (IDP) from algorithm (3.26), or the local wave estimate  $\lambda_\kappa^*$  from (3.29) (IDPloc). Imposing the IDP property may result in over-limiting of the solution and some strategies are usually applied such as bound relaxation [21], or subcell smoothness indicator [34]. We here follow the second strategy which relies on the smoothness indicator from [35] (see [34, Sec. 4.4] for details). Finally, we use the Suliciu pressure relaxation based numerical flux from [5, Sec. 2.4.6] at interfaces, while for  $\mathbf{h}_{sym}$  in the DGSEM scheme (4.5) we use the Kennedy and Grubber splitting from [15].

**5.1. Riemann problems.** We here consider Riemann problems (2.1) with initial data given in Tab. 1. We first consider computations with the DGSEM (see section 4) and the three limiting strategies. Results are shown in Figs. 3 to 5. The POS limiter only ensures that the solution remains in the set of states  $\Omega^a$  and does not modify non-physical oscillations, with the IDP and IDPloc strategies succeed in damping spurious oscillations and result in very close oscillations. All numerical experiments always show that there is no sensible improvement to evaluate the pseudo-equilibrium state  $\mathbf{u}_\kappa^*$  with local wave estimates  $\lambda_\kappa^{*k}$  in (3.26) instead of a global estimate  $\lambda_\kappa^*$  in (3.29), while it leads to a more expensive algorithm. Besides, our observation show that using a local estimates require more iterations for algorithm (3.29) to converge with a global average between 2.8 and 3.2 iterations evaluated over the whole computations compared to between 1.11 and 1.17 when using (3.26). In the latter case, for most of the computations the initial guess  $\mathbf{u}_0^*$  given in (3.26) satisfies the requirement (3.27) and (3.28) so no more step is needed.

We now reproduce the same numerical experiments but using the classical modal DG (see subsection 4.1) with the IDP strategy. Figure 6 presents the results for the three Riemann problems and we observe a good resolution of the waves with only lower amplitude oscillations compared to the DGSEM results. Imposing the IDP property at all quadrature points may result in stronger limiting of the solution with the modal DG scheme as these are more quadrature points (compare Figs. 1 and 2). Again, algorithm (3.26) converges very fast with a global average between of 1.11 and 1.18 iterations evaluated over the whole computations. Finally note that all the computations (with either DGSEM, or modal DG scheme) require to apply one of the limiting strategy to avoid non-physical solution.

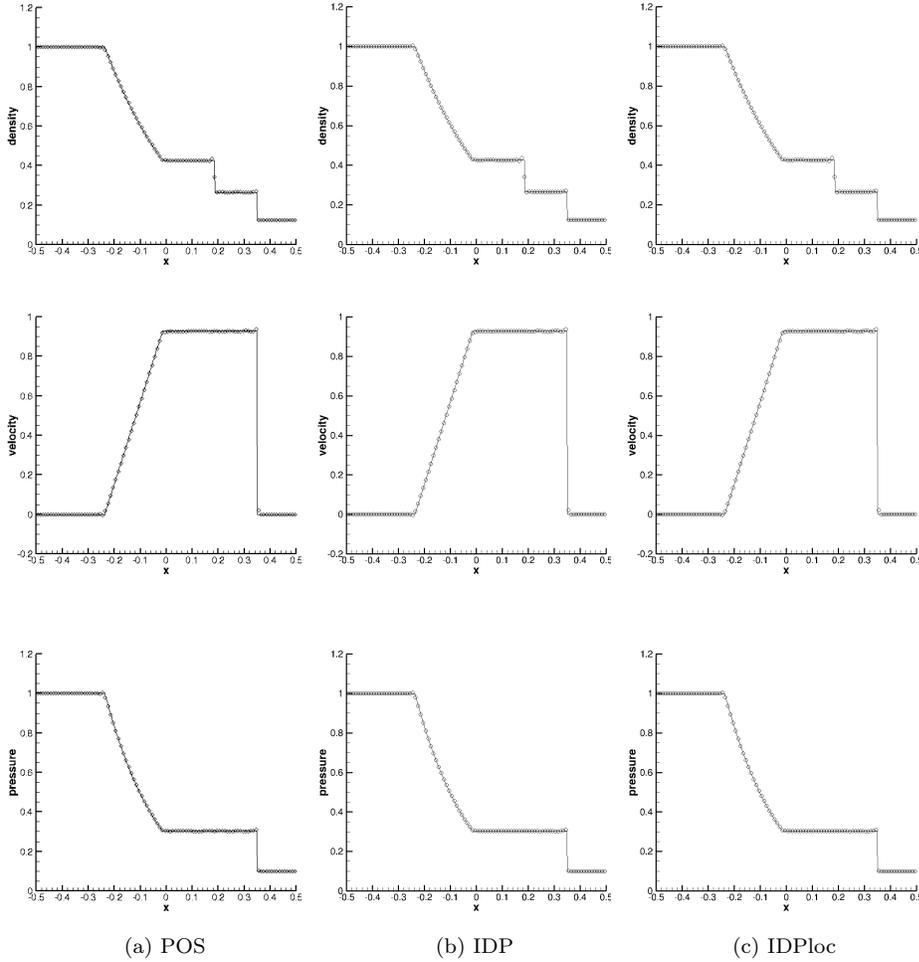


Fig. 3: DGSEM computations of the Sod problem at  $t = 0.2$  with  $p = 3$  and  $N = 100$  elements for the density (top), velocity (middle), and pressure (bottom).

**5.2. Double Mach Reflection problem.** We now consider the two-dimensional problem of a Mach 10 shock reflection over a  $30^\circ$  wedge [47]. Ahead of the shock, the gas is at rest and has a density of 1.4 and pressure of 1. Inflow and outflow conditions are applied at the left and bottom boundaries, while a symmetry condition is applied at the top boundary. Initially, the shock is located at  $x = 0$  corresponding to the beginning of the wedge. We use an unstructured mesh with 132800 quadrangles to solve the horizontally moving shock interacting with the inclined wall where slip conditions are applied. In Fig. 7 we can observe qualitatively similar results for the three tests. The limiting strategy induce some spurious oscillations, but the computations proved to be robust. Here again, algorithm (3.26) proves to be cheap in term of iterations to converge with global averages of 1.18 for both the DGSEM (IDP) and modal DG (IDP) computations.

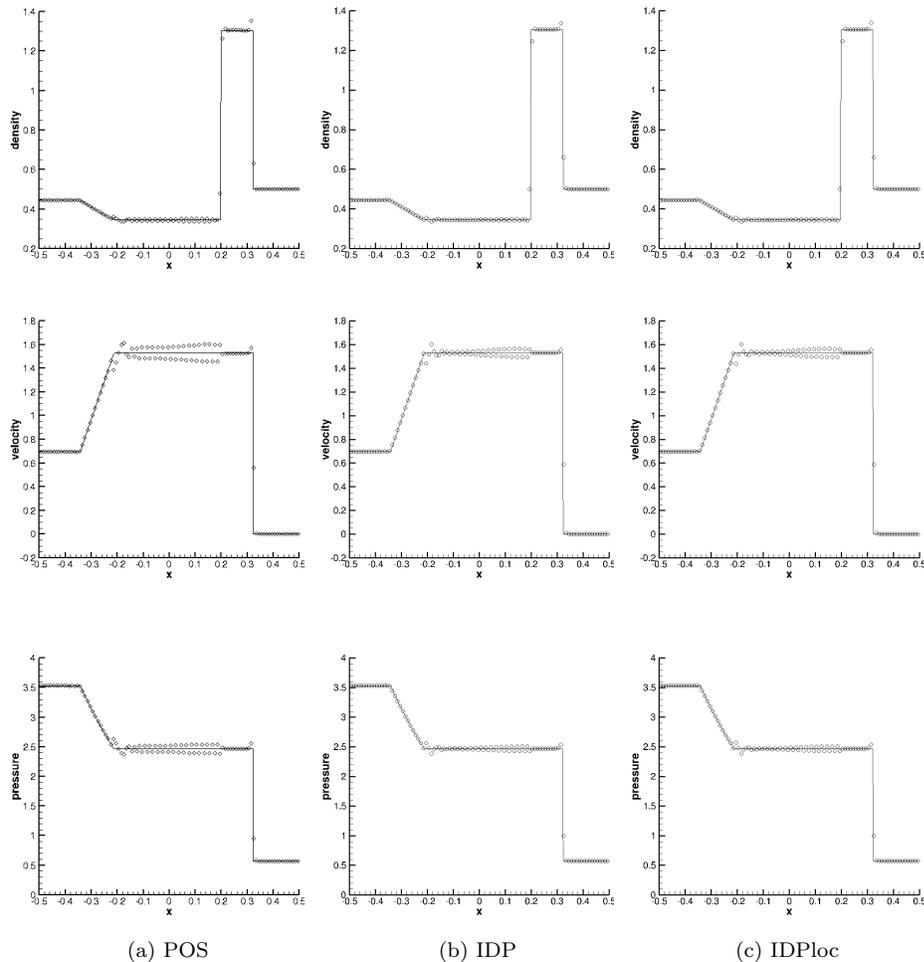


Fig. 4: DGSEM computations of the Lax problem at  $t = 0.2$  with  $p = 3$  and  $N = 100$  elements for the density (top), velocity (middle), and pressure (bottom).

**6. Conclusions.** We here investigate robustness and stability properties of discretely conservative high-order spectral discontinuous methods with explicit time stepping for the approximation of hyperbolic systems of conservation laws. We derive a condition on the time step to guaranty that the cell-averaged approximate solution is a convex combination of DOFs at preceding time step and updates of invariant domain preserving and entropy stable three-point schemes. As a consequence, the cell-averaged solution lies in some convex invariant domain and we apply a posteriori scaling limiting techniques [50] that impose to all the DOFs to satisfy the same invariant domain properties.

The condition on the time step is evaluated from the traces of the solution at faces of the mesh and can be easily evaluated on the fly. Provided the scheme satisfies the discrete metric identities, the condition is fairly general and holds for general unstructured grid with possibly curved elements. It relies on the existence of a so-called

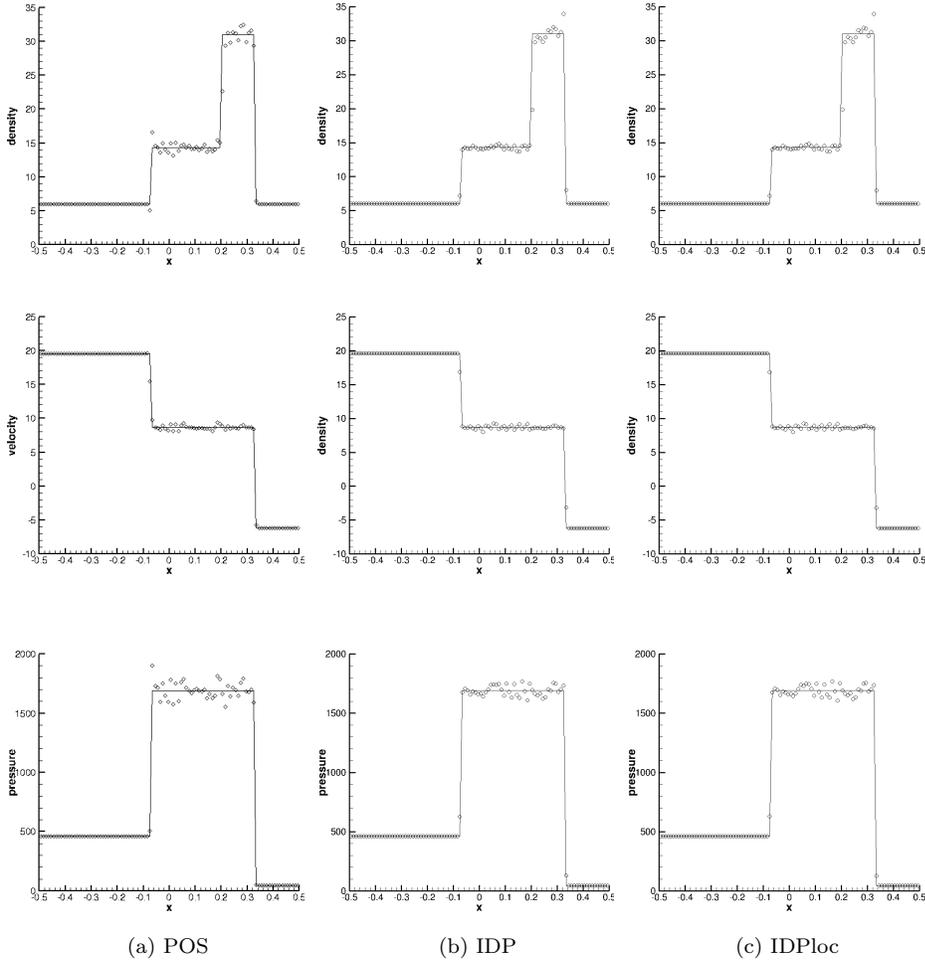


Fig. 5: DGSEM computations of the Toro 4 problem at  $t = 0.2$  with  $p = 3$  and  $N = 100$  elements for the density (top), velocity (middle), and pressure (bottom).

pseudo-equilibrium state which satisfies a flux balance over each mesh element, and the existence of a quadrature rule including the traces to evaluate the cell-averaged solution. We here prove their existence in the general case and provide an iterative algorithm to evaluate the pseudo-equilibrium state. We illustrate these results with the classical modal discontinuous Galerkin and DGSEM schemes. Numerical experiments in one and two space dimensions are provided to illustrate the robustness and stability of the present approach. The extension of this framework to parabolic systems of conservation laws, e.g., the compressible Navier-Stokes, is a possible direction of future research.

REFERENCES

[1] M. AVRIEL, *r-convex functions*, Math. Program., 2 (1972), pp. 309–323.

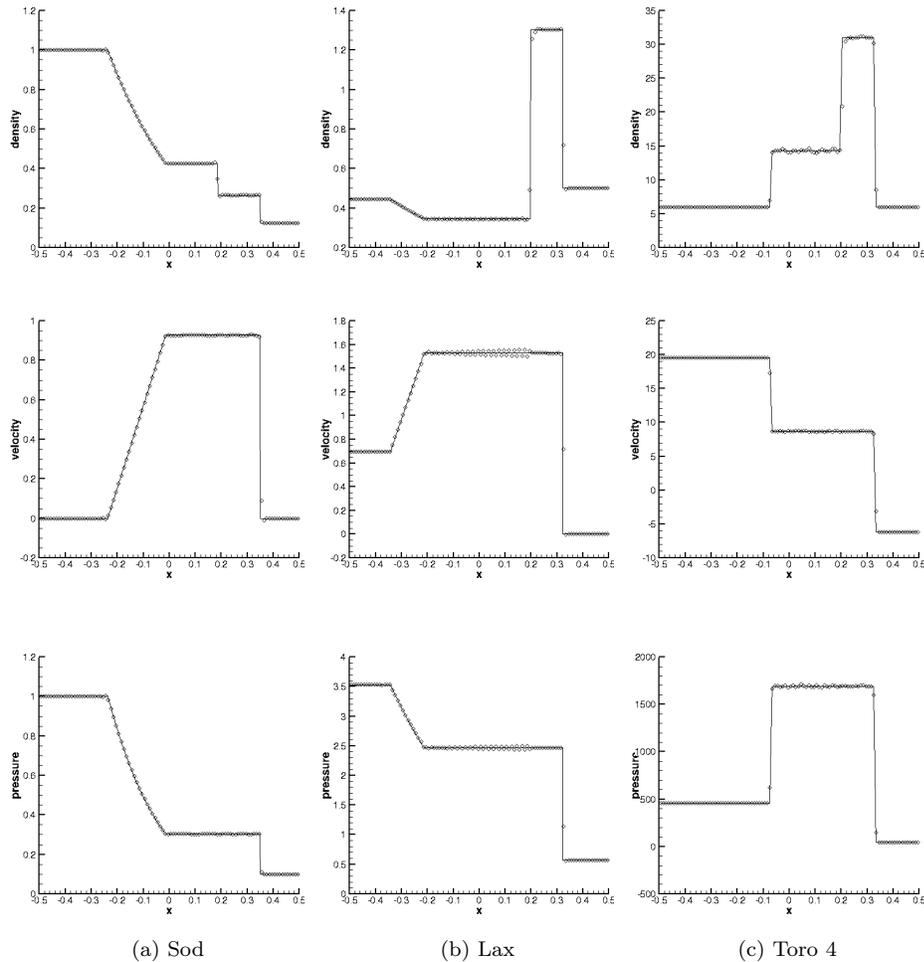


Fig. 6: Modal DG (IDP) computations of Riemann problems with  $p = 3$  and  $N = 100$  elements for the density (top), velocity (middle), and pressure (bottom).

- [2] G. E. BARTER AND D. L. DARMOFAL, *Shock capturing with pde-based artificial viscosity for dgfm: Part i. formulation*, J. Comput. Phys., 229 (2010), pp. 1810–1827, <https://doi.org/https://doi.org/10.1016/j.jcp.2009.11.010>.
- [3] F. BASSI AND S. REBAY, *High-order accurate discontinuous finite element solution of the 2d Euler equations*, J. Comput. Phys., 138 (1997), pp. 251–285.
- [4] J. P. BORIS AND D. L. BOOK, *Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works*, J. Comput. Phys., 11 (1973), pp. 38–69, [https://doi.org/https://doi.org/10.1016/0021-9991\(73\)90147-2](https://doi.org/https://doi.org/10.1016/0021-9991(73)90147-2).
- [5] F. BOUCHUT, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources*, Frontiers in Mathematics, Birkhäuser Basel, 2004.
- [6] J. CHAN, *Skew-symmetric entropy stable modal discontinuous Galerkin formulations*, J. Sci. Comput., 81 (2019), pp. 459–485, <https://doi.org/https://doi.org/10.1007/s10915-019-01026-w>.
- [7] B. COCKBURN AND C. SHU, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws ii: general framework*, Math. Comput., 52 (1989), pp. 411–435, <https://doi.org/https://doi.org/10.1090/S0025-5718-1989-0983311-4>.

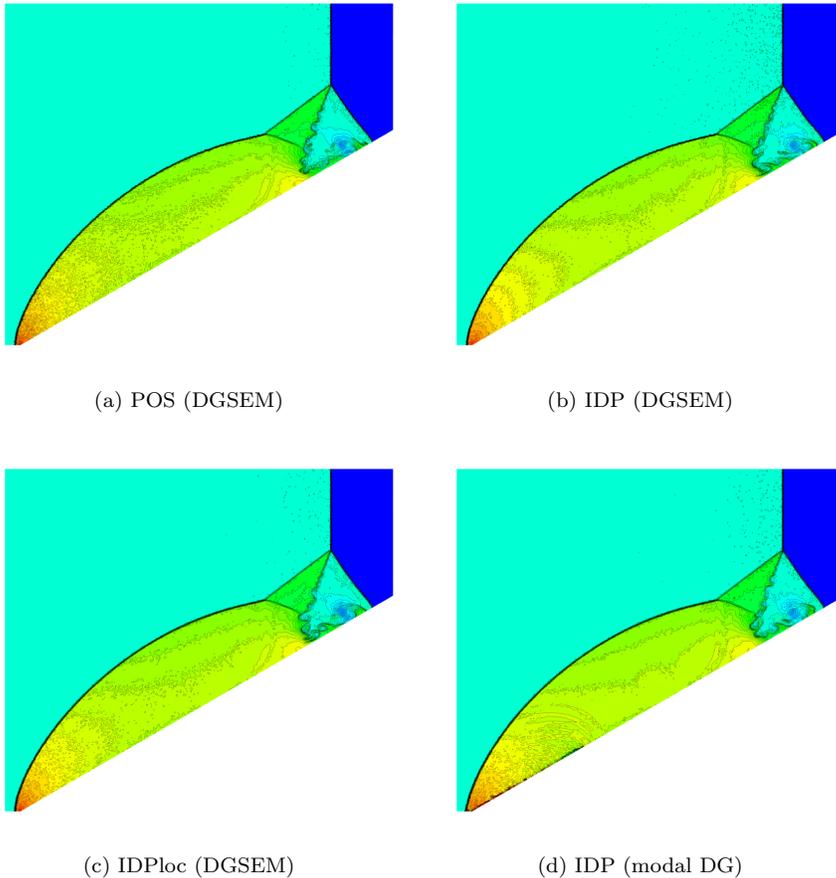


Fig. 7: Comparison of the results for the double Mach reflection problem: 39 equispaced density contours between 2.1 and 22.1.

- [8] J. CREAN, J. E. HICKEN, D. C. DEL REY FERNÁNDEZ, D. W. ZINGG, AND M. H. CARPENTER, *Entropy-stable summation-by-parts discretization of the Euler equations on general curved elements*, J. Comput. Phys., 356 (2018), pp. 410–438, <https://doi.org/https://doi.org/10.1016/j.jcp.2017.12.015>.
- [9] R. J. DIPERNA, *Convergence of approximate solutions to conservation laws*, Arch. Rat. Mech. Anal., 82 (1983), pp. 27–70.
- [10] W. S. DON, *Numerical study of pseudospectral methods in shock wave applications*, J. Comput. Phys., 110 (1994), pp. 103–111.
- [11] T. C. FISHER AND M. H. CARPENTER, *High-order entropy stable finite difference schemes for nonlinear conservation laws: Finite domains*, J. Comput. Phys., 252 (2013), pp. 518–557.
- [12] M. FRANCO, J.-S. CAMIER, J. ANDREJ, AND W. PAZNER, *High-order matrix-free incompressible flow solvers with GPU acceleration and low-order refined preconditioners*, Comput. Fluids, 203 (2020), p. 104541.
- [13] H. FRID, *Maps of convex sets and invariant regions for finite-difference systems of conservation laws*, Arch. Rational Mech. Anal., 160 (2001), pp. 245–269, <https://doi.org/https://doi.org/10.1007/s002050100166>.
- [14] G. J. GASSNER, *A skew-symmetric discontinuous Galerkin spectral element discretization and its relation to SBP-SAT finite difference methods*, SIAM J. Sci. Comput., 35 (2013), pp. A1233–A1253, <https://doi.org/10.1137/120890144>.

- [15] G. J. GASSNER, A. R. WINTERS, AND D. A. KOPRIVA, *Split form nodal discontinuous Galerkin schemes with summation-by-parts property for the compressible Euler equations*, J. Comput. Phys., 327 (2016), pp. 39–66.
- [16] S. GODUNOV, *A difference scheme for numerical computation of discontinuous solutions of equations of fluid dynamics*, Math. USSR Sbornik, 47 (1959), pp. 271–306.
- [17] D. GOTTLIEB AND C.-W. SHU, *On the Gibbs phenomenon and its resolution*, SIAM Rev., 39 (1997), pp. 644–668.
- [18] S. GOTTLIEB, C.-W. SHU, AND E. TADMOR, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.
- [19] 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, <https://doi.org/10.1137/17M1149961>.
- [20] J.-L. GUERMOND, R. PASQUETTI, AND B. POPOV, *Entropy viscosity method for nonlinear conservation laws*, J. Comput. Phys., 230 (2011), pp. 4248–4267, <https://doi.org/https://doi.org/10.1016/j.jcp.2010.11.043>. Special issue High Order Methods for CFD Problems.
- [21] J.-L. GUERMOND, B. POPOV, AND I. TOMAS, *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems*, Comput. Methods Appl. Mech. Engrg., 347 (2019), pp. 143–175, <https://doi.org/https://doi.org/10.1016/j.cma.2018.11.036>.
- [22] A. HARTEN, P. D. LAX, AND B. VAN LEER, *On upstream differencing and Godunov-type schemes for hyperbolic conservation laws*, SIAM Rev., 25 (1983), pp. 35–61.
- [23] A. HILTEBRAND AND S. MISHRA, *Entropy stable shock capturing space–time discontinuous Galerkin schemes for systems of conservation laws*, Numer. Math., 126 (2014), pp. 103–151.
- [24] D. HOFF, *Invariant regions for systems of conservation laws*, Trans. Amer. Math. Soc., 289 (1985), pp. 591–610, <https://doi.org/https://doi.org/10.2307/2000254>.
- [25] M. HUTCHINSON, A. HEINECKE, H. PABST, G. HENRY, M. PARSANI, AND D. KEYES, *Efficiency of high order spectral element methods on petascale architectures*, in International Conference on High Performance Computing, Springer, 2016, pp. 449–466.
- [26] J. JAFFRE, C. JOHNSON, AND A. SZEPESSY, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci., 05 (1995), pp. 367–386, <https://doi.org/10.1142/S021820259500022X>.
- [27] Y. JIANG AND H. LIU, *Invariant-region-preserving dg methods for multi-dimensional hyperbolic conservation law systems, with an application to compressible Euler equations*, J. Comput. Phys., 373 (2018), pp. 385–409, <https://doi.org/https://doi.org/10.1016/j.jcp.2018.03.004>.
- [28] J. S. K. CHUEH, C. CONLEY, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.
- [29] D. A. KOPRIVA, *Metric identities and the discontinuous spectral element method on curvilinear meshes.*, J. Sci. Comput., 26 (2006), pp. 302–327, <https://doi.org/https://doi.org/10.1007/s10915-005-9070-8>.
- [30] D. A. KOPRIVA AND J. H. KOLIAS, *A conservative staggered-grid chebyshev multidomain method for compressible flows*, J Comput. Phys., 125 (1996), pp. 244–261, <https://doi.org/https://doi.org/10.1006/jcph.1996.0091>.
- [31] L. KRIVODONOVA, J. XIN, J.-F. REMACLE, N. CHEVAUGEON, AND J. FLAHERTY, *Shock detection and limiting with discontinuous galerkin methods for hyperbolic conservation laws*, Applied Numerical Mathematics, 48 (2004), pp. 323–338, <https://doi.org/https://doi.org/10.1016/j.apnum.2003.11.002>.
- [32] P. D. LAX, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, Society for Industrial and Applied Mathematics, 1973, <https://doi.org/10.1137/1.9781611970562>, <https://arxiv.org/abs/https://epubs.siam.org/doi/pdf/10.1137/1.9781611970562>.
- [33] Y. LIU, M. VINOKUR, AND Z. WANG, *Spectral difference method for unstructured grids i: Basic formulation*, J Comput. Phys., 216 (2006), pp. 780–801, <https://doi.org/https://doi.org/10.1016/j.jcp.2006.01.024>.
- [34] W. PAZNER, *Sparse invariant domain preserving discontinuous galerkin methods with subcell convex limiting*, Comput. Methods Appl. Mech. Engrg., 382 (2021), p. 113876, <https://doi.org/https://doi.org/10.1016/j.cma.2021.113876>.
- [35] P.-O. PERSSON AND J. PERAIRE, *Sub-Cell Shock Capturing for Discontinuous Galerkin Methods*, 2006, <https://doi.org/10.2514/6.2006-112>.
- [36] B. PERTHAME AND C.-W. SHU, *On positivity preserving finite volume schemes for Euler equations*, Numer. Math., 73 (1996), pp. 119–130, <https://doi.org/10.1007/s002110050187>.
- [37] W. H. REED AND T. R. HILL, *Triangular mesh methods for the neutron transport equation*,

- Los Alamos Report LA-UR-73-479, (1973).
- [38] F. RENAC, *Entropy stable, robust and high-order DGSEM for the compressible multicomponent Euler equations*, J. Comput. Phys., 445 (2021), p. 110584, <https://doi.org/10.1016/j.jcp.2021.110584>.
  - [39] V. RUSANOV, *Calculation of interaction of non-steady shock waves with obstacles*, J. Comp. Math. Phys. USSR, 1 (1961), pp. 267–279.
  - [40] D. SERRE, *Domaines invariants pour les systèmes hyperboliques de lois de conservation*, J. Differ. Equ., 69 (1987), pp. 46–62, [https://doi.org/10.1016/0022-0396\(87\)90102-1](https://doi.org/10.1016/0022-0396(87)90102-1).
  - [41] C.-W. SHU, *High order WENO and DG methods for time-dependent convection-dominated PDEs: A brief survey of several recent developments*, J. Comput. Phys., 316 (2016), pp. 598–613.
  - [42] G. A. SOD, *A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws*, Journal of computational physics, 27 (1978), pp. 1–31.
  - [43] P. D. THOMAS AND C. K. LOMBARD, *Geometric conservation law and its application to flow computations on moving grids*, AIAA J., 17 (1979), pp. 1030–1037, <https://doi.org/10.2514/3.61273>.
  - [44] E. F. TORO, *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction. Third Edition*, Springer-Verlag Berlin Heidelberg, 2009.
  - [45] Z. J. WANG, K. FIDKOWSKI, R. ABGRALL, F. BASSI, D. CARAENI, A. CARY, H. DECONINCK, R. HARTMANN, K. HILLEWAERT, H. T. HUYNH, ET AL., *High-order CFD methods: current status and perspective*, International Journal for Numerical Methods in Fluids, 72 (2013), pp. 811–845.
  - [46] N. WINTERMEYER, A. R. WINTERS, G. J. GASSNER, AND D. A. KOPRIVA, *An entropy stable nodal discontinuous Galerkin method for the two dimensional shallow water equations on unstructured curvilinear meshes with discontinuous bathymetry*, J. Comput. Phys., 340 (2017), pp. 200–242.
  - [47] P. WOODWARD AND P. COLELLA, *The numerical simulation of two-dimensional fluid flow with strong shocks*, J. Comput. Phys., 54 (1984), pp. 115–173, [https://doi.org/https://doi.org/10.1016/0021-9991\(84\)90142-6](https://doi.org/https://doi.org/10.1016/0021-9991(84)90142-6).
  - [48] S. T. ZALESAK, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362.
  - [49] X. ZHANG AND C.-W. SHU, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, J. Comput. Phys., 229 (2010), pp. 3091–3120.
  - [50] X. ZHANG AND C.-W. SHU, *On positivity-preserving high order discontinuous Galerkin schemes for compressible Euler equations on rectangular meshes*, J. Comput. Phys., 229 (2010), pp. 8918–8934.
  - [51] X. ZHANG, Y. XIA, AND C.-W. SHU, *Maximum-principle-satisfying and positivity-preserving high order discontinuous galerkin schemes for conservation laws on triangular meshes*, Journal of Scientific Computing, 50 (2012), pp. 29–62.