THE SULICIU APPROXIMATE RIEMANN SOLVER IS NOT INVARIANT DOMAIN PRESERVING*

JEAN-LUC GUERMOND†, CHRISTIAN KLINGENBERG‡, BOJAN POPOV†, AND IGNACIO TOMAS

Abstract. We show in this note that the first-order finite volume technique based on the Suliciu approximate Riemann solver, while being positive, violates the invariant domain properties of the p-system.

Key words. Conservation equations, Second-order, maximum principle, entropy-viscosity, finite element method.

AMS subject classifications. 65M60, 65M10, 65M15, 35L65

1. Introduction.

The objective of this paper is to investigate the approximation of the p-system using a finite volume technique based on the so-called Suliciu relaxation method and explicit time stepping. This technique, initially introduced in Suliciu [10] to study phase transition in fluid flows, has been adopted in the numerical community to design approximate Riemann solvers; we refer the reader to Bouchut [1, §4.7] and Coquel et al. [4] and the references therein for more details on the method. We restrict ourselves in the present paper to the p-system and show that the first-order finite volume technique based on Suliciu's approximate Riemann solver, while being positive under a standard CFL assumption, violates the invariant domain properties of the PDE.

One motivation for the present work is the construction of robust schemes. We say that a scheme is robust if, under reasonable CFL condition and if the data are admissible, it never fails to produce a a solution that satisfies some reasonable (physical) bounds. Of course, one would want such a scheme to be at least second-order accurate in space (accuracy in time is easily achieved by using strong stability preserving Runge Kutta techniques). One possible route to construct such a scheme consists of computing at each time step a high-order solution and then limiting the high-order solution is some way if it violates some local physical bounds. The natural question that follows is what to limit and how to limit it? The strategy proposed in Guermond et al. [7] consists of using the notion of local convex invariant domain to do the limiting. We recall that convex invariant domains are convex sets in the phase space that are invariant by the PDE. This notion is the natural generalization of the maximum principle for scalar equations to hyperbolic systems. For instance, positivity of the density, positivity of the internal energy, and the local minimum principle on the specific entropy are convex invariant properties for the compressible Euler system. The Riemann invariants define convex invariant domains for the p-system. The technique proposed in Guermond et al. [7] consists at each time step to compute a low-order solution that is guaranteed to be invariant domain preserving and to limit the high-order solution by forcing it to be inside some local invariant domain generated by the

*This material is based upon work supported in part by the National Science Foundation grants DMS-1217262, by the Air Force Office of Scientific Research, USAF, under grant/contract number FA99550-12-0358, and by the Army Research Office under grant/contract number W911NF-15-1-0517. Draft version, January 31, 2018

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

‡Dept. of Mathematics, Würzburg University, Emil Fischer Str. 40, Würzburg, Germany

low-order solution. This method guarantees that the high-order method is as robust as the low-order one. Of course this strategy works well only if the low-order method is robust. The purpose of the present note is to show that the first-order finite volume technique based on the Suliciu approximate Riemann solver is not robust in the sense defined above. More specifically, while the method is definitely positive, we show that it violates the invariant domain properties of the p-system.

The paper is organized as follows. We introduce the problem and notation, and recall key results that are used in the rest of the paper in §2. Suliciu's approximate Riemann solver is recalled in §3. Positivity of this method is established in this section. The main result of this paper is reported in §4. It is proved therein that the first-order finite volume technique based on Suliciu's approximate Riemann solver violates the invariant domain property of the p-system. This statement is proved by producing a counterexample. Originality is claimed only for the material presented in §4.

2. Preliminaries. The objective of this section is to introduce notation and preliminary results that will be useful in the rest of the paper. We use the notation and the terminology of Hoff [8, 9] and Chueh et al. [3, §6].

2.1. p-system. The so-called p-system describes the one-dimensional motion of an isentropic gas in Lagrangian coordinates

$$\begin{cases} \partial_t \tau - \partial_x u = 0, \\ \partial_t u + \partial_x p(\tau) = 0, \quad \text{for } (x,t) \in \mathbb{R}\times\mathbb{R}_+. \end{cases} \tag{2.1}$$

The dependent variables are the velocity u and the specific volume τ, i.e., the reciprocal of density. The mapping $\tau \mapsto p(\tau)$ is the pressure and is assumed to be of class $C^2(\mathbb{R}_+;\mathbb{R})$ and to satisfy the following properties:

$$p' < 0, \qquad 0 < p'', \qquad \int_1^\infty p(s)\,\mathrm{d}s < \infty. \tag{2.2}$$

A typical example is the so-called gamma-law, $p(\tau) = r\tau^{-\gamma}$, where $r > 0$ and $\gamma > 1$. The PDE system (2.1) is supplemented with the initial data

$$\tau(x,0) = \tau_0(x) > 0, \quad u(x,0) = u_0(x), \quad \text{for } x \in \mathbb{R}. \tag{2.3}$$

We further assume that the fluid at infinity approaches constants states. We shall be using these boundary conditions in the rest of the paper without explicitly mentioning it.

2.2. Invariant domain. Defining $\boldsymbol{U} := (\tau,u)^\mathsf{T}$, $\boldsymbol{F}(U) := (-u, p(\tau))^\mathsf{T}$, we can re-write the p-system in vector form: $\partial_t \boldsymbol{U} + \partial_x \boldsymbol{F}(\boldsymbol{U}) = 0$. The Jacobian matrix

$$D\boldsymbol{F} = \begin{pmatrix} 0 & -1 \\ p'(\tau) & 0 \end{pmatrix} \tag{2.4}$$

is diagonalizable with eigenpairs

$$\lambda_1(\boldsymbol{U}) = -\sqrt{-p'(\tau)}, \qquad \boldsymbol{r}_1(\boldsymbol{U}) = (1, -\lambda_1(\boldsymbol{U}))^\mathsf{T}, \tag{2.5}$$

$$\lambda_2(\boldsymbol{U}) = \sqrt{-p'(\tau)}, \qquad \boldsymbol{r}_2(\boldsymbol{U}) = (-1, \lambda_2(\boldsymbol{U}))^\mathsf{T}. \tag{2.6}$$

The two eigenvalues are distinct and real, thereby showing that this nonlinear system is strictly hyperbolic for all $\tau > 0$. Moreover the identities $D\lambda_1(\boldsymbol{U})\cdot\boldsymbol{r}_1 = D\lambda_2(\boldsymbol{U})\cdot\boldsymbol{r}_2 = \frac{p''(\tau)}{2\sqrt{-p'(\tau)}}$ show that the system is genuinely nonlinear under the condition $p''(\tau) > 0$.

Using the notation $\mathrm{d}\mu := \sqrt{-p'(s)}\,ds$, and recalling that we assumed $\int_1^\infty \mathrm{d}\mu < \infty$, the system also has two families of global Riemann invariants:

$$W_1(\boldsymbol{U}) := u + \int_\tau^\infty \mathrm{d}\mu, \quad \text{and} \quad W_2(\boldsymbol{U}) := u - \int_\tau^\infty \mathrm{d}\mu. \tag{2.7}$$

We call $\mathcal{A} := \mathbb{R}_+ \times \mathbb{R}$ the admissible set for (2.1). The reasons for this terminology are as follows. The Riemann problem with any data in $\mathcal{A}$ is uniquely solvable, see Young [11, 12]. For any smooth initial data with value in a bounded subset of $\mathcal{A}$ there is short time existence of a smooth solution to (2.1). Finally, for any smooth initial data with value in a bounded subset of $\mathcal{A}$, the parabolic regularization of the (2.1) stays in $\mathcal{A}$, see Chueh et al. [3, p. 385].

For any set $A \subset \mathcal{A}$ such that $\sup_{\boldsymbol{U}\in A} W_1(\boldsymbol{U}) < \infty$ and $-\infty < \inf_{\boldsymbol{U}\in A} W_2(\boldsymbol{U})$ we define the mappings $W_1^{\max}, W_2^{\min} : \mathcal{A} \to \mathbb{R}$ by setting

$$W_1^{\max}(A) := \sup_{\boldsymbol{U}\in A} W_1(\boldsymbol{U}), \qquad W_2^{\min}(A) := \inf_{\boldsymbol{U}\in A} W_2(\boldsymbol{U}). \tag{2.8}$$

This then leads us to introduce the following set:

$$C(A) := \{\boldsymbol{U} \in \mathcal{A} \mid W_2^{\min}(A) \le W_2(\boldsymbol{U}),\ W_1(\boldsymbol{U}) \le W_1^{\max}(A)\}. \tag{2.9}$$

It is known that W_1 is convex and W_2 is concave. These two conditions imply that $C(A)$ is convex for any admissible set A and $A \subset C(A) \subset \mathcal{A}$.

In the rest of the paper we abuse the notation and view the initial data $\boldsymbol{U}_0$ of (2.1) as a set in the phase space $\mathbb{R}_+ \times \mathbb{R}$, i.e., $\{\boldsymbol{U}_0(x) \mid x \in \mathbb{R}\}$, and using this abuse of notation we consider the set $C(\boldsymbol{U}_0)$. A remarkable fact is that $C(\boldsymbol{U}_0)$ is invariant for smooth solutions of (2.1), meaning that $\boldsymbol{U}(x,t) \in C(\boldsymbol{U}_0)$ for all $x \in \mathbb{R}$ and all t until smoothness is lost. Also, the invariance property holds for the parabolic regularization of (2.1) as shown in Chueh et al. [3, p. 385]. A natural expectation is that any physically relevant solution of (2.1) should satisfy this invariance property, which we henceforth refer to as invariant domain property. One now faces the question of constructing numerical approximations that also satisfy the invariant domain property. For instance, it is known that $C(\boldsymbol{U}_0)$ is invariant for a variety of first-order explicit numerical methods based on finite volumes on uniform grids, see e.g., Hoff [9, Thm. 4.1,4.2] and Hoff [8, Thm 2.1]; this property holds true also for the continuous finite element technique introduced in Guermond and Popov [6]. The purpose of this paper is to show that the first-order finite volume technique based on the Suliciu's approximate Riemann solver, while being positive, violates the invariant domain property of the p-system.

2.3. Riemann problem. Let us consider (2.1) equipped with Riemann data, $\boldsymbol{U}_0(x) = (\tau_R, u_R)^{\mathsf{T}} =: \boldsymbol{U}_L \in \mathcal{A}$ if $x < 0$, $\boldsymbol{U}_0(x) = (\tau_R, u_R)^{\mathsf{T}} =: \boldsymbol{U}_R \in \mathcal{A}$ if $0 < x$:

$$\partial_t \boldsymbol{u} + \partial_x \boldsymbol{F}(\boldsymbol{u}) = 0, \qquad \boldsymbol{u}(\cdot, 0) = \boldsymbol{U}_0. \tag{2.10}$$

It is well-known that this problem has a unique entropy satisfying solution; we refer the reader to Young [11, 12] for the details.

Let us denote by $A_{LR} := \{\boldsymbol{U}_L, \boldsymbol{U}_R\} \subset \mathcal{A}$. It is known that the entropy solution to the Riemann problem stays in the set $C(A_{LR})$. A schematic representation of the set $C(A_{LR})$ is shown in the right panel of Figure 1. Let us denote by $\lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R)$ the maximum wave speed in the problem; that is, let $\lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R) := \max(|\lambda_1^-|, |\lambda_2^+|)$ where λ_1^- is the maximum wave speed of the 1-wave and λ_2^+ is the maximum wave speed of the 2-wave. In general one needs to solve exactly the Riemann problem to estimate $\lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R)$, but in practice it is often enough to have an upper bound on $\lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R)$ to devise numerical schemes that guarantee that the approximate solution to (2.10) stays in $C(A_{LR})$. This can be done without solving the Riemann problem; for instance, the following result established in Guermond and Popov [6, Lem. 2.5] gives such an upper bound.

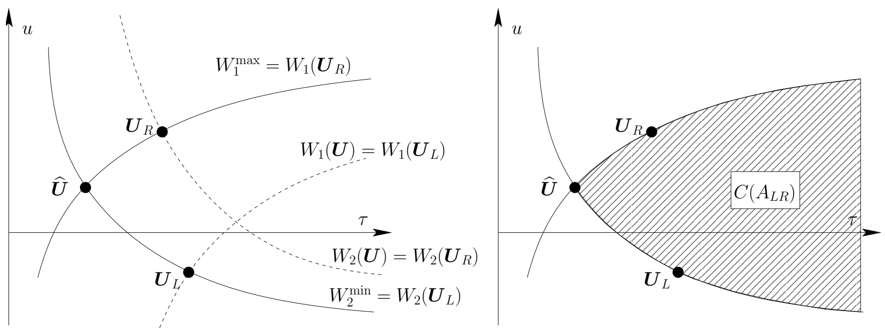


FIG. 1. *Left: Riemann invariants of two states* $(\boldsymbol{U}_L, \boldsymbol{U}_R)$ *for the p-system; the state* $\widehat{\boldsymbol{U}}$ *is obtained by solving* $W_1(\widehat{\boldsymbol{U}}) = W_1^{\max}(A_{LR})$ *and* $W_2(\widehat{\boldsymbol{U}}) = W_2^{\min}(A_{LR})$*. Right: the shaded region is the invariant domain* $C(A_{LR})$ *for the states* $\boldsymbol{U}_L, \boldsymbol{U}_R$*.*

LEMMA 2.1. *Assume that* $p(\tau) = r\tau^{-\gamma}$ *with* $\gamma > 1$ *and* $r > 0$*. Let*

$$\widehat{\tau} := (\gamma r)^{\frac{1}{\gamma-1}} \left(\frac{4}{(\gamma-1)(W_1^{\max}(A_{LR}) - W_2^{\min}(A_{LR}))} \right)^{\frac{2}{(\gamma-1)}} .$$

then $\lambda_{\max}(\boldsymbol{U}_L,\boldsymbol{U}_R) \le \sqrt{-p'(\widehat{\tau})}$*.*

In the rest of the paper we denote by $\widehat{\lambda}_{\max}(\boldsymbol{U}_L,\boldsymbol{U}_R)$ any upper bound on the maximum wave speed $\lambda_{\max}(\boldsymbol{U}_L,\boldsymbol{U}_R)$; for instance, for the γ-law, $p(\tau) = r\tau^{-\gamma}$, $\widehat{\lambda}_{\max}(\boldsymbol{U}_L,\boldsymbol{U}_R) := \sqrt{-p'(\widehat{\tau})}$ is such an upper bound as stated in Lemma 2.1. The computation of $\widehat{\tau}$ is illustrated in the left panel of Figure 1; the state $\widehat{\boldsymbol{U}}$ is obtained by solving $W_1(\widehat{\boldsymbol{U}}) = W_1^{\max}(A_{LR})$ and $W_2(\widehat{\boldsymbol{U}}) = W_2^{\min}(A_{LR})$.

3. Suliciu's approximate Riemann solver.

We recall in this section important properties of the approximate Riemann solver that we are going to use. No originality is claimed on the material presented in this section.

3.1. The approximate Riemann solution.

In this section we produce a consistent approximate Riemann solution to (2.1). To this end we consider the so-called relaxation/projection approximation to the p-system (2.1) described in Bouchut [1], Coquel et al. [4]. The relaxation system in question is written as follows:

$$\begin{cases} \partial_t \tau^\epsilon - \partial_x u^\epsilon = 0, \\ \partial_t u^\epsilon + \partial_x \pi = 0, \\ \partial_t \pi + a^2 \partial_x u^\epsilon = \frac{1}{\epsilon}(p(\tau^\epsilon) - \pi), \end{cases} \tag{3.1}$$

where we choose a large enough, and $\epsilon > 0$ is a small parameter (relaxation time). We are going to be more precise on how large a should be in the next section. In Carbou et al. [2] it is proven under the assumption that if $\inf_{s\in\mathbb{R}_+} p'(s) > 0$, $\sup_{s\in\mathbb{R}_+} p'(s) <$

∞, and $a^2 > \sup_{s\in\mathbb{R}_+} p'(s)$, then for any smooth initial data there exists a time interval (depending on the data) such that the solution to the system (3.1) converges to that of (2.1) as $\epsilon \to 0$.

In order to construct an approximate solution to the Riemann problem (2.10) with the initial data $\boldsymbol{U}_L = (\tau_L, u_L), \boldsymbol{U}_R = (\tau_R, u_R)$, we consider (3.1) with zero right-hand side and with the extended initial data $\widetilde{\boldsymbol{U}}_L := (\tau_L, u_L, p(\tau_L)), \widetilde{\boldsymbol{U}}_R := (\tau_R, u_R, p(\tau_R))$:

$$\begin{cases} \partial_t \widetilde{\tau} - \partial_x \widetilde{u} = 0, \\ \partial_t \widetilde{u} + \partial_x \widetilde{\pi} = 0, \\ \partial_t \widetilde{\pi} + a^2 \partial_x \widetilde{u} = 0. \end{cases} \tag{3.2}$$

The solution to this linear first order PDE consists of four constant states separated by three contact lines: $\frac{x}{t} = -a < \frac{x}{t} = 0 < \frac{x}{t} = a$. Denoting by $\xi = \frac{x}{t}$ the self-similarity variable, the solution to the above problem is described as follows:

(3.3)

	$\xi \le -a$	$-a < \xi \le 0$	$0 < \xi < a$	$a < \xi$
$\widetilde{\tau}$	τ_L	τ_L^*	τ_R^*	τ_R
$\widetilde{u}$	u_L	u^*	u^*	u_R
$\widetilde{\pi}$	$p(\tau_L)$	π^*	π^*	π_R

with the notation

$$\begin{cases} u^* := u^*(\boldsymbol{U}_L, \boldsymbol{U}_R) := \frac{u_L + u_R}{2} - \frac{p(\tau_R) - p(\tau_L)}{2a} \\ \pi^* := \pi^*(\boldsymbol{U}_L, \boldsymbol{U}_R) := \frac{p(\tau_L) + p(\tau_R)}{2} - \frac{a}{2}(u_R - u_L) \\ \tau_L^* := \tau_L^*(\boldsymbol{U}_L, \boldsymbol{U}_R) := \tau_L + \frac{u_R - u_L}{2a} + \frac{p(\tau_L) - p(\tau_R)}{2a^2} \\ \tau_R^* := \tau_R^*(\boldsymbol{U}_L, \boldsymbol{U}_R) = \tau_R + \frac{u_R - u_L}{2a} + \frac{p(\tau_R) - p(\tau_L)}{2a^2}. \end{cases}$$

We then consider the following expression as an approximation of the flux $\boldsymbol{F}(\boldsymbol{u}(0,t))$, where $\boldsymbol{u}$ is the exact solution of the Riemann problem (2.10) with the Riemann data $\boldsymbol{U}_L = (\tau_L, u_L), \boldsymbol{U}_R = (\tau_R, u_R)$:

$$\boldsymbol{F}^*(\boldsymbol{U}_L, \boldsymbol{U}_R) := (-u^*(\boldsymbol{U}_L, \boldsymbol{U}_R), \pi^*(\boldsymbol{U}_L, \boldsymbol{U}_R))^{\mathsf{T}}. \tag{3.4}$$

Notice that denoting by $\widetilde{\boldsymbol{F}}(\widetilde{\boldsymbol{u}}(x,t))$ the flux of the extended system (3.2), $\boldsymbol{F}^*(\boldsymbol{U}_L, \boldsymbol{U}_R)$ is the vector composed of the first two components of $\widetilde{\boldsymbol{F}}(\widetilde{\boldsymbol{u}}(0,t))$.

3.2. Positivity. We now want to establish that the solution defined by (3.3) is positive in the sense that $\widetilde{\tau}(x,t) \ge 0$ for all $x \in \mathbb{R}$ and all $t > 0$. To do so we have to establish that $\tau_L^* \ge 0$ and $\tau_R^* \ge 0$. Let us introduce the state $\overline{\boldsymbol{U}}$ defined by

$$\overline{\boldsymbol{U}} = \frac{\boldsymbol{U}_L + \boldsymbol{U}_R}{2} - \frac{\boldsymbol{F}(\boldsymbol{U}_R) - \boldsymbol{F}(\boldsymbol{U}_L)}{2a}. \tag{3.5}$$

It is well-known that if $a \ge \lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R)$, then $\overline{\boldsymbol{U}}$ belongs to the invariant set $C(A_{LR})$, see e.g., [6, Lem. 2.1]. In particular, setting $\overline{\boldsymbol{U}} =: (\overline{\tau}, \overline{u})^{\mathsf{T}}$, we have

$$\inf_{(\tau,u)\in C(\boldsymbol{U}_L,\boldsymbol{U}_R)} \tau \le \overline{\tau}, \tag{3.6}$$

$$W_2^{\min}(A_{LR}) = \inf_{(\tau,u)\in C(\boldsymbol{U}_L,\boldsymbol{U}_R)} u \le \overline{u} \le \sup_{(\tau,u)\in C(\boldsymbol{U}_L,\boldsymbol{U}_R)} u = W_1^{\max}(A_{LR}). \tag{3.7}$$

LEMMA 3.1. *$\boldsymbol{U}_L, \boldsymbol{U}_R$ be two states in the admissible set of the p-system. Let $\Delta W := W_1^{\max}(A_{LR}) - W_2^{\min}(A_{LR})$. Let a be such that*

$$a \geq \max(\lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R), \frac{\Delta W}{\min(\tau_L, \tau_R)}), \tag{3.8}$$

then $\tau_L^*(\boldsymbol{U}_L, \boldsymbol{U}_R) \geq 0$ and $\tau_R^*(\boldsymbol{U}_L, \boldsymbol{U}_R) \geq 0$.

Proof. We first notice that

$$\tau_L^* = \tau_L + \frac{1}{a}(\overline{u} - u_L), \qquad \tau_R^* = \tau_R + \frac{1}{a}(u_R - \overline{u}).$$

As a result, positivity holds if $a \geq \max(\frac{(u_L-\overline{u})_+}{\tau_L}, \frac{(\overline{u}-u_R)_+}{\tau_R})$. Notice that if $a \geq \lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R)$ then $\max(|\overline{u} - u_L|, |u_R - \overline{u}|) \leq \Delta W$ owing to (3.7). Therefore the desired result holds true if $a \geq \Delta W / \min(\tau_L, \tau_R)$. ∎

Remark 3.2 (Expansion wave). In order to have some intuition on the relative magnitude of the quantities appearing on the right-hand side of (3.8), let us assume that $\boldsymbol{U}_L$ and $\boldsymbol{U}_R$ are located on a 1-wave and $\tau_L < \tau_R$; i.e., the Riemann solution is an expansion wave. This case will be used to construct the counterexample in §4.2. Let us further assume that the equation of state is a γ-law $p(\tau) = r\tau^{-\gamma}$. Then $\lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R) = \sqrt{-p'(\tau_L)} = (\gamma r)^{\frac{1}{2}} \tau_L^{-\frac{\gamma+1}{2}}$. Moreover, $\Delta W = W_1(\boldsymbol{U}_L) - W_2(\boldsymbol{U}_L) = 2\int_{\tau_L}^{\infty} \sqrt{-p'(s)}\,\mathrm{d}s$; that is, $\min(\tau_L, \tau_R)^{-1} \Delta W = \frac{4}{\gamma-1}(\gamma r)^{\frac{1}{2}} \tau_L^{-\frac{\gamma+1}{2}}$. In this case we have $\min(\tau_L, \tau_R)^{-1}\Delta W = \frac{4}{\gamma-1}\lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R)$; in particular, for $\gamma \in (1, 5)$, we have $\min(\tau_L, \tau_R)^{-1}\Delta W > \lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R)$. No claim is made on the optimality of the bound (3.8). The results reported at the end of §4.2 have been obtained with $a = \max(\widehat{\lambda}_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R) \geq \max(\lambda_{\max}(\boldsymbol{U}_L, \boldsymbol{U}_R)$. □

4. The main result. We describe in this section the Godunov-type finite volume scheme using the approximate Riemann solver defined in §3 to solve (2.1), and we show that the scheme is positive but violates the invariant domain property.

4.1. Finite volume discretization. Let $\mathcal{T}_h := \{x_{i+\frac{1}{2}}\}_{i\in\mathbb{Z}}$ be a sequence of distinct points in $\mathbb{R}$. We denote $I_i := [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, $h_i := x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$. We are going to solve (2.1) with a Godunov-type finite volume technique using the approximation space $\boldsymbol{P}_0(\mathcal{T}_h) := \{\boldsymbol{v}_h \in L^\infty(\mathbb{R}; \mathbb{R}^2) \mid \boldsymbol{v}_{h|I_i} \in \mathbb{P}_0 \times \mathbb{P}_0,\ \forall i \in \mathbb{Z}\}$, where $\mathbb{P}_0$ denotes the real vector space composed of the constant univariate polynomials. The interface flux will be computed by using the approximate flux (3.4).

Given cell average $\{\boldsymbol{U}_i^n\}_{i\in\mathbb{Z}}$ at time t^n, $n \in \mathbb{N}$, we define the update $\{\boldsymbol{U}_i^{n+1}\}_{i\in\mathbb{Z}}$ by setting

$$h_i(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n) + \Delta t(\boldsymbol{F}^*(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n) - \boldsymbol{F}^*(\boldsymbol{U}_{i-1}^n, \boldsymbol{U}_i^n)) = 0, \tag{4.1}$$

where we recall that the interface flux is given by (3.4):

$$\boldsymbol{F}^*(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n) := (-u^*(u_i^n, u_{i+1}^n), \pi^*(u_i^n, u_{i+1}^n))^\mathsf{T}, \tag{4.2}$$

where the speed a in (3.2) is denoted $a_{i+\frac{1}{2}}^n$, $i \in \mathbb{Z}$. This quantity is chosen by the user and should be large enough; for instance, based on Lemma 3.1 one could take

$$a_{i+\frac{1}{2}}^n = \max(\lambda_{\max}(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n), \frac{\Delta W_{i+\frac{1}{2}}^n}{\min(\tau_i^n, \tau_{i+1}^n)}), \tag{4.3}$$

with $\Delta W_{i+\frac{1}{2}}^n := \max(W_1(\boldsymbol{U}_i^n), W_1(\boldsymbol{U}_{i+1}^n)) - \min(W_2(\boldsymbol{U}_i^n), W_2(\boldsymbol{U}_{i+1}^n))$.

LEMMA 4.1 (Positivity). *Given admissible states* $\boldsymbol{U}^n_{i-1}, \boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1}$, *assume that the condition* (4.3) *on* $a^n_{i-\frac12}$ *and* $a^n_{i+\frac12}$ *holds for the pairs* $(\boldsymbol{U}^n_{i-1}, \boldsymbol{U}^n_i)$ *and* $(\boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1})$. *Assume also that* $(a^n_{i-\frac12} + a^n_{i+\frac12})\Delta t < h_i$, *then the scheme is positive, i.e.,* $\tau^{n+1}_i > 0$.

Proof. Since $(a^n_{i-\frac12} + a^n_{i+\frac12})\Delta t < h_i$, the definition of the flux (4.2) implies that

$$\boldsymbol{U}^{n+1}_i = \frac{a^n_{i-\frac12}\Delta t}{h_i}\boldsymbol{U}^{*,R}_{i-\frac12} + \frac{a^n_{i+\frac12}\Delta t}{h_i}\boldsymbol{U}^{*,L}_{i+\frac12} + \Big(1 - \frac{a^n_{i-\frac12}\Delta t}{h_i} - \frac{a^n_{i+\frac12}\Delta t}{h_i}\Big)\boldsymbol{U}^n_i,$$

where

$$\boldsymbol{U}^{*,R}_{i-\frac12} := (\tau^*_R(\boldsymbol{U}^n_{i-1}, \boldsymbol{U}^n_i), u^*(\boldsymbol{U}^n_{i-1}, \boldsymbol{U}^n_i))^\mathsf{T},$$
$$\boldsymbol{U}^{*,L}_{i+\frac12} := (\tau^*_L(\boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1}), u^*(\boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1}))^\mathsf{T},$$

and the functions τ^*_L, τ^*_R, and u^* are defined in (3.3). We have established in Lemma 3.1 that $\tau^*_R(\boldsymbol{U}^n_{i-1}, \boldsymbol{U}^n_i) \geq 0$ and $\tau^*_L(\boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1}) \geq 0$ under the condition (4.3) for the pairs $(\boldsymbol{U}^n_{i-1}, \boldsymbol{U}^n_i)$ and $(\boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1})$. Then τ^{n+1}_i is a convex combination of the three states $\tau^*_R(\boldsymbol{U}^n_{i-1}, \boldsymbol{U}^n_i) \geq 0$, $\tau^n_i > 0$, and $\tau^*_L(\boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1}) \geq 0$ under the CFL condition $(a^n_{i-\frac12} + a^n_{i+\frac12})\Delta t < h_i$, which proves the result. ∎

4.2. Violation of the invariant domain property. We show in this section that it is possible to find initial data such that the scheme defined in (4.1)-(4.2) violates the invariant domain property of the p-system. The counterexample in question is built by considering an expansion wave.

Let $u_L, u_R \in \mathbb{R}$, and $\tau_L, \tau_R \in \mathbb{R}_+$. We set the initial data to (2.1) to be

$$\boldsymbol{u}_{0h|I_i} =: \boldsymbol{U}^0_i := \begin{cases} (\tau_L, u_L)^\mathsf{T} & \text{if } i < 1, \\ (\tau_R, u_R)^\mathsf{T} & \text{if } 1 \leq i. \end{cases} \tag{4.4}$$

Then, the following result demonstrates that the (4.1)-(4.2) is not invariant domain preserving.

THEOREM 4.2. *Assume that* $\tau_L < \tau_R$ *and* $W_1(\boldsymbol{U}_L) = W_1(\boldsymbol{U}_R)$. *Assume that* $a^0_{\frac12}$ *satisfies* (4.3) *and* $\frac{a^0_{\frac12}\Delta t}{h_0} \leq 1$. *Then we have* $W^{\max}_1(A_{LR}) < W_1(\boldsymbol{U}^1_0)$, *i.e., the scheme* (4.1)-(4.2) *violates the invariant domain property of the p-system at the first time step.*

Proof. After observing that $\boldsymbol{U}^{*,R}_{-\frac12} = \boldsymbol{u}_L$, we infer that

$$\boldsymbol{U}^1_0 = \frac{a^0_{\frac12}\Delta t}{h_0}\boldsymbol{U}^{*,L}_{\frac12} + \big(1 - \frac{a^0_{\frac12}\Delta t}{h_0}\big)\boldsymbol{U}^0_0.$$

Denoting $\alpha := \frac{a^0_{\frac12}\Delta t}{h_0}$ and $a := a^0_{\frac12}$ we can write the components of $\boldsymbol{U}^1_0$ as follows:

$$\tau^1_0 = \tau_L + \frac{\alpha(u_R - u_L)}{2a} + \frac{\alpha(p(\tau_L) - p(\tau_R))}{2a^2},$$
$$u^1_0 = u_L + \frac{\alpha(u_R - u_L)}{2} + \frac{\alpha(p(\tau_L) - p(\tau_R))}{2a}.$$

We take $u_R - u_L = \int_{\tau_L}^{\tau_R} \mathrm{d}\mu$ which corresponds to the states $\boldsymbol{U}_0^0 := \boldsymbol{U}_L$ and $\boldsymbol{U}_1^0 := \boldsymbol{U}_R$ being on a left expansion wave. Then $W_1^{\max}(A_{LR}) = W_1(\boldsymbol{U}_0^0) = W_1(\boldsymbol{U}_1^0)$. Let us denote $\Delta W := W_1(\boldsymbol{U}_0^1) - W_1^{\max}(A_{LR})$. We have that

$$\Delta W = \frac{\alpha(u_R - u_L)}{2} + \frac{\alpha(p(\tau_L) - p(\tau_R))}{2a} - \int_{\tau_L}^{\tau_0^1} \mathrm{d}\mu.$$

Observing that $\tau_L < \tau_R$ implies that $u_R > u_L$, $p(\tau_L) > p(\tau_R)$, and $\tau_0^1 > \tau_L$. Using that $\mathrm{d}\mu := \sqrt{-p'(s)}\,ds$ and $\sqrt{-p'(s)}$ is a strictly decreasing function we have

$$\Delta W > \frac{\alpha(u_R - u_L)}{2} + \frac{\alpha(p(\tau_L) - p(\tau_R))}{2a} - \sqrt{-p'(\tau_L)}(\tau_0^1 - \tau_L).$$

Recalling that $\tau_0^1 - \tau_L = \frac{\alpha(u_R - u_L)}{2a} + \frac{\alpha(p(\tau_L) - p(\tau_R))}{2a^2}$, we conclude that

$$\Delta W > \big(\frac{\alpha(u_R - u_L)}{2} + \frac{\alpha(p(\tau_L) - p(\tau_R))}{2a}\big)\big(1 - \frac{\sqrt{-p'(\tau_L)}}{a}\big).$$

Notice that $\frac{\alpha(u_R - u_L)}{2} + \frac{\alpha(p(\tau_L) - p(\tau_R))}{2a}$ is positive. Recalling that a is an upper bound on the maximum speed of propagation in the Riemann problem, we have $\sqrt{-p'(\tau_L)} \le a$. Hence, $\Delta W > 0$ for any $\sqrt{-p'(\tau_L)} \le a$ and therefore $\boldsymbol{U}_0^1$ is not in the local invariant domain of the states $\boldsymbol{U}_0^0$ and $\boldsymbol{U}_1^0$. Notice in passing that we actually have established an upper bound and a lower bound on ΔW

$$1 > \frac{\Delta W}{\frac{\alpha(u_R - u_L)}{2} + \frac{\alpha(p(\tau_L) - p(\tau_R))}{2a}} > \big(1 - \frac{\sqrt{-p'(\tau_L)}}{a}\big), \tag{4.5}$$

and these two bounds are independent on the mesh size. This completes the proof. ☐

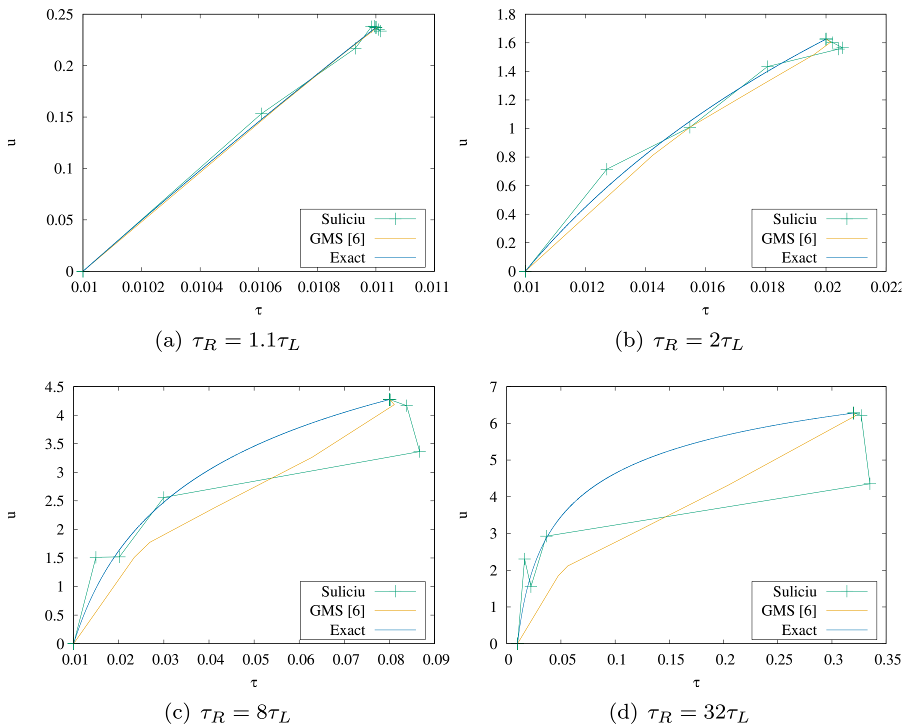
To illustrate Theorem 4.2, we compare the scheme (4.1)-(4.2) with the so-called GMS-GV1 scheme descried in Guermond and Popov [6]. (GMS stands for Guaranteed Maximum Speed and GV1 stands for first-order graph viscosity.) In the present context, the GMS-GV1 scheme can be rewritten as follows:

$$h_i(\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n) + \Delta t(\boldsymbol{F}^{\mathrm{GMS}}(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n) - \boldsymbol{F}^{\mathrm{GMS}}(\boldsymbol{U}_{i-1}^n, \boldsymbol{U}_i^n)) = 0, \tag{4.6}$$

where

$$\boldsymbol{F}^{\mathrm{GMS}}(\boldsymbol{U}, \boldsymbol{V}) := \frac{1}{2}(\boldsymbol{F}(\boldsymbol{U}) + \boldsymbol{F}(\boldsymbol{V})) + \frac{1}{2}\widehat{\lambda}_{\max}(\boldsymbol{U}, \boldsymbol{V})(\boldsymbol{U} - \boldsymbol{V}).$$

The initial data that we use is similar to that invoked in the proof of Theorem 4.2: the states $\boldsymbol{U}_L$, $\boldsymbol{U}_R$ are parts of an expansion (1-wave). We take $\tau_L := 0.01$ and $u_L := 0$. The following ratios $\tau_R/\tau_L \in \{1.1, 2, 8, 32\}$ are tested, and the quantity u_R is given by $u_R := u_L + \int_{\tau_L}^{\tau_R} \mathrm{d}\mu$. We use the equation of state $p(\tau) := 1/(\gamma\tau^\gamma)$ with $\gamma := 1.4$. The speed $a_{i+\frac{1}{2}}^n$ is computed by setting $a_{i+\frac{1}{2}}^n := \widehat{\lambda}(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n)$ using the estimate of $\widehat{\lambda}(\boldsymbol{U}, \boldsymbol{V})$ given in Lemma 2.1. The time step is defined by $\Delta t := \mathrm{CFL} h_i/\sqrt{-p'(\tau_L)}$ where we set CFL $:= 0.9$. The results shown in Figure 2 compare in the phase space ($u(x,t)$ vs. $\tau(x,t)$) the GMS-GV1 solution and the solution given by the scheme (4.1)-(4.2). The comparison is done after 3 time steps. Notice that the GMS-GV1 solution is invariant domain preserving, as proved in Guermond and Popov [6, Thm. 4.1]. The scheme (4.1)-(4.2) clearly steps out of the invariant domain; that is, there are states $\boldsymbol{U}_j$ such that $W_1(\boldsymbol{U}_j) > W_1^{\max}(A_{LR})$, on the plots these states sit above the blue curve, which is the graph of the exact solution in the phase space and is also the upper boundary of the invariant domain. Let us emphasize that the results shown in Figure 2 are independent of the number of grid points; More precisely, the amount of violation only depends on the CFL number and the number of time step, as established in (4.5).



(a) $\tau_R = 1.1\tau_L$ (b) $\tau_R = 2\tau_L$

(c) $\tau_R = 8\tau_L$ (d) $\tau_R = 32\tau_L$

FIG. 2. *Illustration of Theorem 4.2. Comparison in the phase space* (τ, u) *of the GMS-GV1 solution and the solution given by the scheme* (4.1)-(4.2) *after 3 time steps:* $\tau_L = 0.01$; $u_L = 0$; $p(\tau) = 1/(\gamma\tau^\gamma)$; $\gamma = 1.4$; $a^n_{i+\frac{1}{2}}$ *computed by setting* $a^n_{i+\frac{1}{2}} = \widehat{\lambda}(\boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1})$; $\Delta t = 0.9 h_i/\sqrt{-p'(\tau_L)}$.

4.3. Artificial viscosity interpretation. In this section we reinterpret the scheme (4.1)-(4.2) in term of artificial viscosity and put the scheme in perspective with the parabolic regularization theory of Chueh et al. [3].

We start by mentioning a result that will help us understand why the scheme (4.1)-(4.2) is not invariant domain preserving.

LEMMA 4.3 (Parabolic regularization). *The following parabolic regularization of the system* (2.1) $\partial_t \boldsymbol{u}^{\epsilon,\mu} + \partial_x \boldsymbol{F}(\boldsymbol{u}^{\epsilon,\mu}) = (\epsilon\partial_{xx}\tau^{\epsilon,\mu}, \mu\partial_{xx}u^{\epsilon,\mu})^\mathsf{T}$ *with* $\epsilon, \mu > 0$ *preserves the invariant domains of* (2.1) *if and only if* $\epsilon = \mu$.

This results is proved in Chueh et al. [3, p. 385]. A somewhat similar result has been proved in Guermond and Popov [5, Thm. 4.1] for the Euler equations.

Let us now rewrite the flux $\boldsymbol{F}^*(\boldsymbol{U}^n_i, \boldsymbol{U}^n_{i+1})$ introduced in (4.2) as the sum of the centered flux plus a "viscous" perturbation:

$$\boldsymbol{F}^*(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n) = \begin{pmatrix} -\frac{u_i^n + u_{i+1}^n}{2} + \frac{p(\tau_{i+1}^n) - p(\tau_i^n)}{2a_{i+\frac12}^n} \\ \frac{p(\tau_i^n) + p(\tau_{i+1}^n)}{2} - \frac{a_{i+\frac12}^n}{2}(u_{i+1}^n - u_i^n) \end{pmatrix}$$

$$= \frac{1}{2}(\boldsymbol{F}(\boldsymbol{U}_i^n) + \boldsymbol{F}(\boldsymbol{U}_{i+1}^n)) + \frac{1}{2} a_{i+\frac12}^n \begin{pmatrix} \frac{p(\tau_{i+1}^n) - p(\tau_i^n)}{(a_{i+\frac12}^n)^2} \\ u_i^n - u_{i+1}^n \end{pmatrix}.$$

This expression shows that using the approximate flux $\boldsymbol{F}^*(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n)$ is strictly equivalent to using the centered flux augmented with the heterogenous viscous flux

$$\frac{1}{2} a_{i+\frac12}^n \begin{pmatrix} -\frac{p(\tau_{i+1}^n) - p(\tau_i^n)}{(a_{i+\frac12}^n)^2 (\tau_{i+1}^n - \tau_i^n)} (\tau_i^n - \tau_{i+1}^n) \\ u_i^n - u_{i+1}^n \end{pmatrix}.$$

This argument shows in turn that the scheme (4.1)-(4.2) is a discrete realization of the following perturbed PDE:

$$\partial_t \boldsymbol{u}^\epsilon + \partial_x \boldsymbol{F}(\boldsymbol{u}^\epsilon) = \begin{pmatrix} \partial_x \left(\frac12 a\epsilon \frac{|p'(\tau^\epsilon)|}{a^2} \partial_x \tau^\epsilon \right) \\ \partial_x \left(\frac12 a\epsilon \partial_x u^\epsilon \right) \end{pmatrix},$$

where the quantity ϵ plays the role of the meshsize. In the light of Lemma 4.3, we now understand that to make the scheme (4.1)-(4.2) invariant domain preserving one should set $(a_{i+\frac12}^n)^2 = -\frac{p(\tau_{i+1}^n) - p(\tau_i^n)}{\tau_{i+1}^n - \tau_i^n}$. But this choice is not good enough, since one should also have $a_{i+\frac12}^n \geq \lambda_{\max}(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n)$ (see [5, Thm. 4.1]), which implies $(a_{i+\frac12}^n)^2 > -\frac{p(\tau_{i+1}^n) - p(\tau_i^n)}{\tau_{i+1}^n - \tau_i^n}$ because $\lambda_{\max}^2(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n) = -p'(\tau_L)$ and p is a strictly decreasing function. Hence the requirements $(a_{i+\frac12}^n)^2 = -\frac{p(\tau_{i+1}^n) - p(\tau_i^n)}{\tau_{i+1}^n - \tau_i^n}$ and $(a_{i+\frac12}^n)^2 \geq \lambda_{\max}(\boldsymbol{U}_i^n, \boldsymbol{U}_{i+1}^n)$ cannot be achieved at the same time. In conclusion, we conjecture that the scheme (4.1)-(4.2) cannot be made invariant domain preserving for any choice of $a_{i+\frac12}^n$.

References.

[1] 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.

[2] G. Carbou, B. Hanouzet, et al. Relaxation approximation of some initial-boundary value problem for p-systems. *Communications in Mathematical Sciences*, 5(1):187–203, 2007.

[3] K. N. Chueh, C. C. Conley, and J. A. Smoller. Positively invariant regions for systems of nonlinear diffusion equations. *Indiana Univ. Math. J.*, 26(2):373–392, 1977.

[4] F. Coquel, E. Godlewski, and N. Seguin. Relaxation of fluid systems. *Math. Models Methods Appl. Sci.*, 22(8):1250014, 52, 2012.

[5] J.-L. Guermond and B. Popov. Viscous regularization of the Euler equations and entropy principles. *SIAM J. Appl. Math.*, 74(2):284–305, 2014.

[6] J.-L. Guermond and B. Popov. Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM J. Numer. Anal.*, 54 (4):2466–2489, 2016.

[7] J.-L. Guermond, M. Nazarov, B. Popov, and I. Tomas. Second-order invariant domain preserving approximation of the euler equations using convex limiting. *J. Sci. Comput.*, 2018.

[8] D. Hoff. A finite difference scheme for a system of two conservation laws with artificial viscosity. *Math. Comp.*, 33(148):1171–1193, 1979.

[9] D. Hoff. Invariant regions for systems of conservation laws. *Trans. Amer. Math. Soc.*, 289(2):591–610, 1985.

[10] I. Suliciu. On modelling phase transitions by means of rate-type constitutive equations. Shock wave structure. *International Journal of Engineering Science*, 28(8):829–841, 1990.

[11] R. Young. The p-system. I. The Riemann problem. In *The legacy of the inverse scattering transform in applied mathematics (South Hadley, MA, 2001)*, volume 301 of *Contemp. Math.*, pages 219–234. Amer. Math. Soc., Providence, RI, 2002.

[12] R. Young. The p-system. II. The vacuum. In *Evolution equations (Warsaw, 2001)*, volume 60 of *Banach Center Publ.*, pages 237–252. Polish Acad. Sci., Warsaw, 2003.