

# UNIVERSIDAD DE CONCEPCIÓN



## CENTRO DE INVESTIGACIÓN EN INGENIERÍA MATEMÁTICA (CI<sup>2</sup>MA)



A random sampling approach for a family of Temple-class  
systems of conservation laws

FERNANDO BETANCOURT, RAIMUND BÜRGER,  
CHRISTOPHE CHALONS, STEFAN DIEHL,  
SEBASTIAN FARÅS

PREPRINT 2015-16

SERIE DE PRE-PUBLICACIONES



# A RANDOM SAMPLING APPROACH FOR A FAMILY OF TEMPLE-CLASS SYSTEMS OF CONSERVATION LAWS

FERNANDO BETANCOURT<sup>A</sup>, RAIMUND BÜRGER<sup>B</sup>, CHRISTOPHE CHALONS<sup>C</sup>,  
STEFAN DIEHL<sup>D</sup>, AND SEBASTIAN FARÅS<sup>D,\*</sup>

**ABSTRACT.** Several applications, including the Aw-Rascle-Zhang traffic model and a model of sedimentation of small particles in a viscous fluid, give rise to nonlinear  $2 \times 2$  systems of conservation laws that are governed by a single scalar system velocity, which is associated to a scalar flux function. Such systems are of the Temple class since rarefaction wave curves and Hugoniot curves coincide. Moreover, one characteristic field is genuinely nonlinear (with the exception of a manifold of inflection points of the scalar flux function) and the other is linearly degenerate. For systems of this family, there are two well-known problems. The vacuum state, which may form naturally even from positive initial data, gives rise to potential problems of non-uniqueness and instability. This is resolved by the introduction of two alternative solution concepts of the Riemann problem. The other problem are spurious oscillations produced by Godunov's method near contact discontinuities. This behaviour actually arises with many standard conservative schemes since the numerical solution invariably leaves the invariant region of the exact solution. It is demonstrated that a strategy consisting of alternating between averaging (Av) and remap steps similar to the approach by C. Chalons and P. Goatin [Commun. Math. Sci. 5:533–551, 2007] generates numerical solutions that satisfy an invariant region property (in contrast to, for instance, Godunov's method). For the case that the remap step is done by random sampling (RS), techniques due to J. Glimm [Comm. Pure Appl. Math. 18(4):697–715, 1965], R. J. LeVeque and B. Temple [Trans. Amer. Math. Soc. 288(1):115–123, 1985] are combined to prove that the resulting statistically conservative Av-RS scheme converges to a weak solution. Numerical examples illustrate the performance of the Av-RS scheme, and its superiority over Godunov's method in terms of accuracy and resolution.

---

*Date:* June 11, 2015.

*AMS subject classifications.* 35L65, 35L45, 65M12.

*Key words and phrases.* Averaging-remap schemes, anti-diffusive, statistically conservative, convergence, Aw-Rascle-Zhang traffic model, sedimentation.

\*Corresponding author.

<sup>A</sup>CI<sup>2</sup>MA and Departamento de Ingeniería Metalúrgica, Facultad de Ingeniería, Universidad de Concepción, Casilla 160-C, Concepción, Chile. E-Mail: [fbetancourt@udec.cl](mailto:fbetancourt@udec.cl).

<sup>B</sup>CI<sup>2</sup>MA and Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Casilla 160-C, Concepción, Chile. E-Mail: [rburger@ing-mat.udec.cl](mailto:rburger@ing-mat.udec.cl).

<sup>C</sup>Laboratoire de Mathématiques de Versailles, UMR 8100, Université de Versailles Saint-Quentin-en-Yvelines, UFR des Sciences, bâtiment Fermat, 45 avenue des Etats-Unis, 78035 Versailles cedex, France. E-mail: [christophe.chalons@uvsq.fr](mailto:christophe.chalons@uvsq.fr).

<sup>D</sup>Centre for Mathematical Sciences, Lund University, P.O. Box 118, S-221 00 Lund, Sweden. E-Mail: [diehl@maths.lth.se](mailto:diehl@maths.lth.se) (S. Diehl), [faras@maths.lth.se](mailto:faras@maths.lth.se) (S. Farås).

## 1. INTRODUCTION

**1.1. Scope.** The purpose of this paper is to present and analyze numerical schemes for the approximation of solutions to the initial value problem

$$\begin{pmatrix} \phi \\ k\phi \end{pmatrix}_t + \left( V(\phi, k) \begin{pmatrix} \phi \\ k\phi \end{pmatrix} \right)_x = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad (x, t) \in \mathbb{R} \times (0, T), \quad (1)$$

$$\phi(x, 0) = \phi_0(x), \quad k(x, 0) = k_0(x), \quad x \in \mathbb{R}, \quad (2)$$

where  $T > 0$  is fixed and the unknowns are  $\phi = \phi(x, t)$  and  $k = k(x, t)$ . The system of conservation laws (1) is said to be of the Temple class since the rarefaction wave curves of each characteristic field coincide with the corresponding Hugoniot curves [40]. It appears – with different assumptions on the scalar system velocity  $V$  – in the modelling of one-dimensional two-phase flow [23, 39], elasticity theory [22, 33], traffic-flow theory [1, 42] and sedimentation of particles in a liquid [4].

We consider a family of systems (1), for which the velocity function  $V \in C^2$  is such that  $V(\cdot, k)$  has one zero in  $[0, 1]$  for every  $k \in [0, 1]$  and

$$V_\phi < 0 \text{ and } V_k > 0 \text{ on } (0, 1) \times (0, 1]. \quad (3)$$

Moreover, we restrict the discussion to initial data with values in

$$\Omega := \{(\phi, k) \in [0, 1]^2 : V(\phi, k) \geq 0, V(\cdot, k) \text{ invertible}\}.$$

The numerical schemes and the theoretical results are derived on a general level for the entire family, with no further conditions imposed on  $V$ , and the analysis is illustrated by two particular function expressions yielding the Aw-Rascle-Zhang (ARZ) traffic model [1] and the sedimentation model in [4]. It is common in the literature to assume concavity of the scalar flux function  $\phi \mapsto f(\phi, k) := V(\phi, k)\phi$ , but the results given herein are not restricted to this special case.

Introducing the conserved quantity  $w := k\phi$ , the vector  $\mathbf{y} := (\phi, w)^T$  and  $\tilde{V}(\mathbf{y}) := V(\phi, w/\phi)$ , we can write (1) as

$$\mathbf{y}_t + (\tilde{V}(\mathbf{y})\mathbf{y})_x = \mathbf{0}. \quad (4)$$

This form becomes useful when we need standard results on first-order systems. However, vacuum ( $\phi = 0$ ) complicates the definition of  $\tilde{V}$ , the fundamental difficulty being how to understand  $k = w/\phi$  in this case. Following Temple [39], we therefore always assume that data are given in terms of  $\mathbf{u} := (\phi, k)^T$  and move to the conserved variables in  $\mathbf{y}$  via the mapping  $\mathbf{Y}(\mathbf{u}) := (\phi, k\phi)^T$  whenever needed.

For a general  $2 \times 2$  system of conservation laws, Temple [40] proved that a Hugoniot curve coincides with an integral curve of the same characteristic field if and only if the curve is either a straight line or a level curve of the corresponding wave speed (eigenvalue of the Jacobian). The system reduces to a scalar conservation law on each such curve. The two types of characteristic fields are called line and contact field, respectively. The line field may contain shock waves and rarefaction waves, while the second field with constant wave speed is called a contact field since the only waves are contact discontinuities.

As LeVeque and Temple [28] recognized, the solution to the Riemann problem for (4) is contained in an invariant region in the phase plane. This region – which we will refer to as

$\mathcal{M}$  – bounds the Riemann invariants ( $k$  and  $V$ ) by their values at the initial datum. For the numerical schemes, we define a counterpart  $\mathcal{M}^0$  of  $\mathcal{M}$  for general initial data with values in  $\Omega$ . If both characteristic fields are line fields, then the set  $\mathcal{M}$  is convex, which means that the Riemann invariants of the averages produced by Godunov's method are bounded by their values for the local Riemann data at each time step and  $\mathcal{M}^0$  is globally invariant. These were the basic properties that allowed the authors of [28] to prove convergence of Godunov's method for initial data with bounded total variation. Here, the assumptions (3) on  $V$  do not necessarily imply two line fields but the system may have a contact field in which the Hugoniot curves have non-zero curvature. Under such circumstances, the set  $\mathcal{M}$  need not be convex and the invariance property for Godunov's method is lost (both locally and globally). One consequence is that spurious oscillations may develop near contact discontinuities in the numerical solution, but even more importantly, the proof in [28] breaks down.

The main novelty of the present work is a so-called anti-diffusive numerical scheme, which relies on averaging (Av) and random sampling (RS) and is referred to as the *Av-RS scheme*. This method is closely related to the one proposed by Chalons and Goatin [9], and is constructed in two steps (averaging and remap) to obtain the local and global invariant region properties associated with  $\mathcal{M}$ . System (1) has at least one line field and the first step of the Av-RS scheme consists of a Godunov-like averaging along this field. Such an averaging stays within  $\mathcal{M}$  locally because of the convexity of straight lines, but it comes at the cost of a non-uniform refinement of the spatial mesh. To remap the numerical approximations to the original mesh, random sampling is applied in the second step.

Due to the averaging in the first step, the Av-RS scheme needs no detailed information about rarefaction waves and therefore becomes simpler than the random-choice method by Glimm [16], which has frequently been used for similar systems (see Section 1.4). As a consequence of the sampling in the remap step, the Av-RS scheme does not conserve mass (the integral of  $\phi$  with respect to  $x$ ) exactly. The *expected value* of the mass is however conserved, wherefore we say that this scheme is *statistically conservative*. By combining the ideas in [28] with those of Glimm [16] (also outlined by Smoller [38]), we prove convergence for the Av-RS scheme to weak solutions for initial data with bounded total variation. We reach this result assuming Lipschitz continuity of  $V(\cdot, k)$  and its inverse  $\Phi(\cdot, k)$  on each line  $k = \text{constant}$  in  $\mathcal{M}^0$  and its image on the plane of Riemann invariants, respectively.

The delicate vacuum case when  $\phi = 0$  may be formed naturally in the solutions to (1)–(2) even for initial data with  $\phi_0 > 0$ . In traffic-flow applications, the non-uniqueness and the instability problems that appear near the vacuum states are well known; see e.g. [1]. We distinguish between two types of Riemann solutions: type A and type B, where the latter is unstable with respect to disturbances of initial data with  $\phi_0|_{x>0} = 0$  and  $k_0|_{x>0} < k_0|_{x<0}$ . For all other Riemann initial data, the two types of solutions coincide.

Unstable solutions to mathematical models in general are often disregarded as unphysical (e.g. via some entropy condition), but the instability in the type B solution is a consequence of problems associated with the interpretation of  $k$  – and thus  $V$  – in vacuum. For example, if  $\phi$  denotes the density of vehicles on a road, what is the system velocity on empty sections of the road? The answer is in some (non-unique) sense given by the prescribed values of  $\phi_0$  and  $k_0$ . In the type B solution, such problematic is avoided by not taking values of  $V$  at

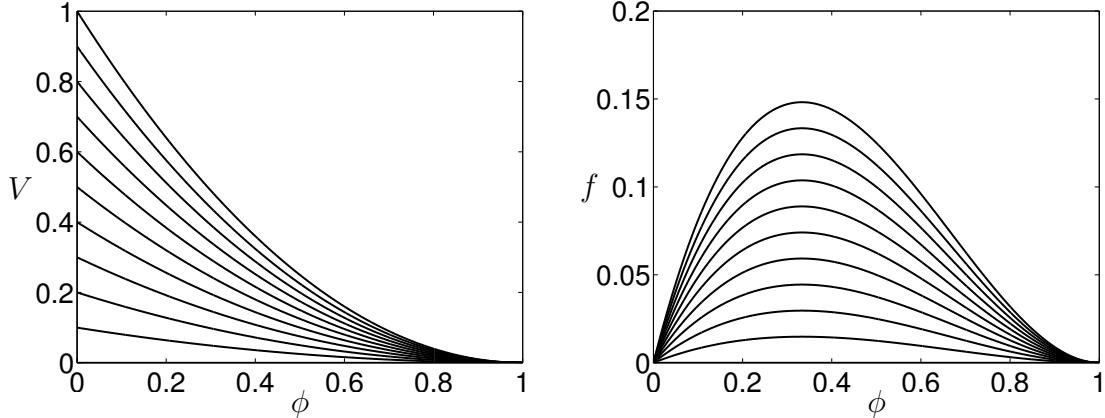


FIGURE 1. The system velocity  $V(\cdot, k)$  of the sedimentation model (6) (left) and the corresponding scalar flux  $f(\cdot, k)$  (right) for  $k = 0.1, 0.2, \dots, 1$ .

vacuum into account. In fact, this type of solution is the physically relevant alternative in both traffic flow and sedimentation.

The convergence proof for the Av-RS scheme is valid only if it is defined in terms of local type A Riemann solutions. However, the relevance of the convergence result is strengthened by a proposition, which states that if the discretized initial datum is such that locally type A coincides with type B, then the Av-RS scheme preserves this agreement, i.e. the unstable vacuum case is not formed in the numerical approximations. Furthermore, numerical experiments suggest convergence of the Av-RS scheme also if it is defined by means of local type B solutions.

**1.2. Application to sedimentation.** In [4], we presented a model for sedimentation of suspensions with inhomogeneous settling properties. Herein, effects from bulk flows and compression will be ignored and we account for so-called “hindered settling” only. To this end, we let  $\phi$  be the solids volume fraction and denote by  $v_s \geq 0$  the settling velocity of the solids. The conservation of mass yields

$$\phi_t + (v_s \phi)_x = 0. \quad (5)$$

Kynch [25] assumed that  $v_s := v(\phi)$  depends on  $\phi$  only, and thereby imposed the same settling properties on all particles. This simplification is somewhat relaxed if we to each particle associate a scalar quantity, called grey tone, which influences the settling properties. Such a grey tone could for example represent the degree of flocculation, i.e. the extent to which the solids are lumped together in so-called flocs. At a continuum level, we denote the grey tone by  $k$  and redefine the settling velocity as  $v_s := kv(\phi)$ . Values of  $k$  close to 0 thus mean poor settling properties while portions of the suspension with  $k = 1$  settle fast. Since  $k$  is advected with the solids, we obtain  $k_t + v_s k_x = 0$ . Multiplying this equation by  $\phi$  and using (5), we find (1) with

$$V(\phi, k) = kv(\phi). \quad (6)$$

The particular choice of  $v$  used herein to generate figures and to compute numerical solutions, is given by the Richardson-Zaki [37] type of expression:  $v(\phi) = (1 - \phi)^2$ , which yields

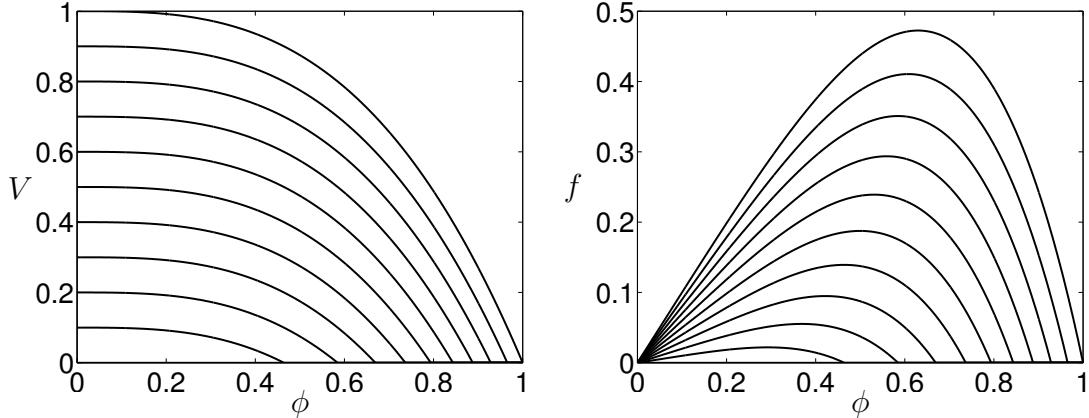


FIGURE 2. The system velocity  $V(\cdot, k)$  of the ARZ model (left) and the corresponding scalar flux  $f(\cdot, k)$  (right) for  $k = 0.1, 0.2, \dots, 1$ .

$\Omega = [0, 1] \times (0, 1]$ . The corresponding flux  $\phi \mapsto f(\phi, k) = k(1 - \phi)^2\phi$  has one inflection point at  $\phi = 2/3$  for every  $k \in (0, 1]$  (see Figure 1).

Note that the sample formula  $v(\phi) = (1 - \phi)^2$  implies  $V_\phi(1, k) = 0$  and the requirement on Lipschitz continuity of  $\Phi(\cdot, k)$  makes the convergence proof for the Av-RS scheme valid only if the initial datum is such that  $\mathcal{M}^0$  is contained in a closed subset of  $[0, 1] \times (0, 1]$ . This extra restriction on the intial data would be superfluous if we instead considered the perturbed velocity  $v(\phi) := (1 + \varepsilon - \phi)^2 - \varepsilon^2$  for some fixed small  $\varepsilon > 0$ .

**1.3. Application to traffic flow: The ARZ model.** Aw and Rascle [1] and Zhang [42] independently proposed a  $2 \times 2$  system as a model of traffic flow. The first equation in this system conserves the mass (of vehicles):

$$\phi_t + (\nu\phi)_x = 0, \quad (7)$$

were we denote by  $\phi$  the density of cars and by  $\nu$  the mean car velocity. Influenced by gas dynamics, some earlier models had included a pressure  $p(\phi)$  via an additional second-order (in space) equation for the momentum. In [1, 42], it was concluded that traffic flow behaves somewhat differently from gas fluids, wherefore this second equation was replaced by the convection equation

$$(\nu + p(\phi))_t + \nu(\nu + p(\phi))_x = 0. \quad (8)$$

The system (7), (8) converts to (1) via the identification  $k = \nu + p(\phi)$  and

$$V(\phi, k) = k - p(\phi). \quad (9)$$

Herein, all figures and numerical examples illustrating the ARZ model are rendered with  $p(\phi) = \phi^3$ . This choice implies  $\Omega = \{\mathbf{u} : 0 \leq \phi^3 \leq k \leq 1\}$  and the associated flux  $\phi \mapsto f(\phi, k) = (k - \phi^3)\phi$  is concave on  $[0, k^{1/3}]$  for every  $k \in [0, 1]$  (Figure 2). We again emphasize that such a concavity is not necessary for the theory, but it simplifies the construction of Riemann solutions.

In analogy with the sample model for sedimentation above, the convergence proof of the Av-RS scheme is not valid on the entire  $\Omega$  when  $p(\phi) = \phi^3$ . There holds  $V_\phi(0, k) = 0$ , and to

obtain Lipschitz continuity of  $\Phi(\cdot, k)$ , the initial datum must be such that  $\mathcal{M}^0$  is contained in a closed subset of  $\Omega \cap \{(\phi, k) : \phi > 0\}$ . Thus, the proof does not capture solutions with vacuum if  $p$  is described by this particular formula. However, for the perturbed expression  $p(\phi) = (\phi + \varepsilon)^3 - \varepsilon^3$ , where the small number  $\varepsilon > 0$  is fixed, the entire (and perturbed) region  $\Omega$  is covered by the theory.

**1.4. Related works.** Temple [39] showed the existence of solutions for the Cauchy problem of (1), but with  $V$  chosen such that  $f(\cdot, k)$  is increasing with one inflection point and  $f(\phi, \cdot)$  is decreasing. The proof was given for initial data with bounded variation and relied on the convergence of approximate solutions obtained by Glimm's method [16]. The system models two-phase polymer flow in a uniformly porous medium and loses strict hyperbolicity along a curve in the phase plane. This loss may cause unbounded total variation for the approximate solution, a complication that was handled in [39] by letting the random choice variable in Glimm's method be random in both space and time. An extension of the results on the  $2 \times 2$  system in [39] to the case of arbitrary large systems was provided by Isaacsson and Temple [21], who also used Glimm's random choice method.

Another special case of system (1) is the well-known Keyfitz-Krantzer system [22]. Existence of bounded weak solutions to that system was given by Lu and Gu [33].

It is known that Godunov's method produces spurious oscillations near states of coinciding wave speeds [30, 32], but such undesirable behaviour can occur even in regions where the system (1) is strictly hyperbolic. This is emphasized by Bressan et al. [5], who showed that the total variation of an approximate solution produced by Godunov's method can become arbitrarily large. For certain initial data, numerical solutions with various solvers may not contain oscillations, see the traffic-flow example by Qiao et al. [36]. However, we are interested in a numerical scheme that can handle initial data that may produce all types of waves.

The necessity to adjust Godunov's method because of its poor accuracy also for strictly hyperbolic systems was further investigated for the Euler equations for a compressible two-fluid mixture; see e.g. Chalons and Coquel [8] and Bachmann et al. [2] and references therein. They kept track of contact discontinuities and suggested a Lagrange-projection method with the random-sampling technique in the spirit of Chalons and Goatin [9, 10]. See also Chalons [7, 11] for a related method built on a macroscopic model of pedestrian flows or the recent work by Bürger et al. [6] on the multi-class Lighthill-Whitham-Richards traffic model. A contact discontinuity, e.g. modelling a gas-fluid interface, can be tracked by a level-set method or a colour function that is evolved by a single transport equation [2, 24]. Further examples of oscillations produced by Godunov's method for the Euler equations are given by Banks [3].

Several authors [18, 23, 29, 34] have proved the existence of weak solutions to systems of the form (4) by using that it is hyperbolic away from vacuum and can be written in triangular form via the idea by Wagner [41] to perform a Lagrangian coordinate transformation and identify  $1/\phi$  as one of the variables. The eigenvalue of the second characteristic field of the triangular system is zero, which means that the contact discontinuity of this field is stationary and can be handled exactly by Godunov's method in contrast to a moving contact discontinuity. Naturally,  $1/\phi$  cannot be used as a variable for solutions with vacuum;

however, in the present work, we also use a sort of Lagrangian approach with respect to the velocity of the contact discontinuity when we utilize that no mass passes this discontinuity.

The traffic flow model by Aw and Rascle [1] and Zhang [42] satisfies (1) and our assumptions on  $V$  (see Section 1.3). Aw and Rascle solved the Riemann problem and discussed in detail the vacuum case and the instabilities that may appear for data near the vacuum. See also [26].

Godvik and Hanche-Olsen [18] proved the existence of a weak entropy solution of the Cauchy problem for the ARZ model with the vacuum case included. They considered a modified problem having a small parameter, used the Glimm scheme and the transformation by Wagner [41]. In this way they cover the vacuum case as a boundary case obtained from the natural entropy solutions for  $\phi > 0$  as the small parameter tends to zero. There is, however, a case in traffic flow where we argue that the physically correct solution of a Riemann problem with  $\phi_R = 0$  is not the (stable) solution obtained as the limit case of  $\phi_R > 0$ ; see Section 3.2.

Chalons and Goatin [9] resolved the problem of spurious oscillations near contact discontinuities that appear when using the Godunov flux for obtaining approximate solutions of the ARZ traffic model. They used a random-sampling strategy in the numerical treatment of contact discontinuities and Godunov's method elsewhere. In this sense, the Av-RS scheme proposed in the present work resembles their scheme and they are both non-diffusive around isolated contact discontinuities. Chalons and Goatin [10] extended the scheme to the case when the phase space is non-convex, in fact a disjoint union of two sets corresponding to free and congested traffic.

Several extensions of the ARZ model have been presented; see e.g. [13, 14, 15, 17]. For example, Moutari and Rascle [34] presented a hybrid macro-microscopic variant of the ARZ model. With a Lagrangian discretization of this hybrid model, the Godunov scheme preserves a bounded and invariant region, which is one ingredient in their proof of convergence to a weak entropy solution when the vacuum case is avoided.

Although not analyzed in their original paper, the acceleration equation of the ARZ model has a relaxation source term. This case was covered by the existence and uniqueness results by Godvik and Hanche-Olsen [18], see above. The extended equation was also analyzed by Li [29], who, in contrast to all references to traffic-flow models mentioned above, did not assume convexity of the flux function  $f(\cdot, k)$  and showed existence of a unique entropy solution of the Cauchy problem. The existence was shown by convergence of a monotone conservative upwind difference scheme. The analysis includes the limit case with zero relaxation, which is (1). The vacuum case is, however, avoided.

**1.5. Outline of this work.** The remainder of this paper is organized as follows. The basic attributes such as eigenvalues, Riemann invariants and Hugoniot curves of System (1) are given in Section 2, where we also define the concept of admissible solution to (1)–(2) and introduce some central notations. In Section 3, we state and analyze the type A solutions to the Riemann problem associated with (1). In particular, we prove the invariance property of  $\mathcal{M}$ . Furthermore, the type B solution is defined and motivated by physical arguments concerning the application to traffic flow. It will be seen in Section 4.1 that Godunov's numerical flux is well defined only if it is based on the type B Riemann solution. In the same section, we also show that a large family of conservative numerical schemes – Godunov's

method included – produce inadequate approximations near contact discontinuities. This result justifies why we turn to alternative schemes in Section 4.2. There, we consider a family of two-step averaging-remap schemes having a common averaging technique in the first step and distinguished only by their different remap strategies in the second step. The Av-RS scheme is a member of this family. We also define another member, whose remap relies on a convex combination (CC) of Riemann invariants. The resulting scheme is called the *Av-CC scheme*. We close Section 4 by stating and proving that the unstable vacuum case is never formed by the Av-RS scheme if it does not appear in the (discrete) initial data. In Section 5 we prove convergence to a limit function for the entire family of averaging-remap schemes, and then show that the limit function obtained with the Av-RS scheme is in fact a weak solution to (1)–(2). The performance of the Godunov, Av-RS and Av-CC schemes are numerically investigated on six sample problems in Section 6.

## 2. PRELIMINARIES

We let  $\mathbf{f}(\mathbf{u}) := V(\mathbf{u})\mathbf{Y}(\mathbf{u})$  and write the Cauchy problem (1)–(2) in the condensed form:

$$\mathbf{Y}(\mathbf{u})_t + \mathbf{f}(\mathbf{u})_x = \mathbf{0}, \quad (x, t) \in \mathbb{R} \times (0, T), \quad (10)$$

$$\mathbf{u}(x, 0) = \mathbf{u}_0(x), \quad x \in \mathbb{R}, \quad (11)$$

for the initial datum  $\mathbf{u}_0(x) \in \Omega$ . Since solutions of quasilinear, first-order systems may develop discontinuities even for smooth  $\mathbf{u}_0$ , the solution  $\mathbf{u}$  is sought in the weak sense:

$$\iint_{\mathbb{R} \times [0, T]} (\mathbf{Y}(\mathbf{u})\varphi_t + \mathbf{f}(\mathbf{u})\varphi_x) dx dt + \int_{\mathbb{R}} \varphi(x, 0)\mathbf{Y}(\mathbf{u}_0(x)) dx = \mathbf{0} \text{ for all } \varphi \in C_0^1(\mathbb{R} \times [0, T]).$$

**2.1. Basic properties of the governing model system.** To derive the basic properties of System (10), we will work with the conserved vector  $\mathbf{y}$ , turn to the form (4) and restrict our attention to the set  $\mathbf{Y}(\Omega')$ , where  $\Omega' := \{\mathbf{u} \in \Omega : \phi > 0, V_\phi(\mathbf{u}) < 0\}$ . The Jacobian matrix of the flux function  $\tilde{\mathbf{f}}(\mathbf{y}) := \tilde{V}(\mathbf{y})\mathbf{y}$  is given by

$$\mathcal{J}(\mathbf{y}) = \begin{bmatrix} \tilde{V} + \phi\tilde{V}_\phi & \phi\tilde{V}_w \\ w\tilde{V}_\phi & \tilde{V} + w\tilde{V}_w \end{bmatrix}$$

and has the eigenvalue-eigenvector pairs

$$\begin{aligned} \lambda_1 &= \tilde{V} + \phi\tilde{V}_\phi + w\tilde{V}_w = f_\phi(\phi, w/\phi), & \mathbf{r}_1 &= (1, w/\phi)^T, \\ \lambda_2 &= \tilde{V} = V(\phi, w/\phi), & \mathbf{r}_2 &= (\tilde{V}_w, -\tilde{V}_\phi)^T. \end{aligned} \quad (12)$$

The eigenvalues are real and satisfy  $\lambda_1 = \tilde{V} + \phi V_\phi(\phi, w/\phi) < \tilde{V} = \lambda_2$ , so the system (4) is strictly hyperbolic on  $\mathbf{Y}(\Omega')$ . Furthermore, the scalar product  $\nabla_{\mathbf{y}}\lambda_1 \cdot \mathbf{r}_1 = f_{\phi\phi}(\phi, w/\phi)$  reveals that for every fixed value of  $k = w/\phi$ , the first characteristic field is genuinely nonlinear away from the inflection points of  $f(\cdot, k)$ , while  $\nabla_{\mathbf{y}}\lambda_2 \cdot \mathbf{r}_2 = 0$  implies that the second field is (globally) linearly degenerate. This linear degeneracy means that  $\lambda_2 = \tilde{V}$  is a Riemann invariant and it leads to an interesting feature of the solution: the only possible 2-waves are contact discontinuities with propagation speed  $\tilde{V} \geq 0$ . Such contact discontinuities are constructed along the contours of  $\tilde{V}$  in the phase plane  $\mathbf{Y}(\Omega')$ . A direct computation yields  $\nabla_{\mathbf{y}}k \cdot \mathbf{r}_1 = 0$ , so the Riemann invariant of the first family is  $k$ . Consequently, the

1-rarefaction curves (the curves in  $\mathbf{Y}(\Omega')$  along which rarefaction waves of the first field are constructed) follow the contours of  $k$ .

In a weak solution, any discontinuity with propagation speed  $s$  separating  $\mathbf{y}_-$  to the left from  $\mathbf{y}_+$  to the right, both states in  $\mathbf{Y}(\Omega')$ , must satisfy the Rankine-Hugoniot condition [27]:

$$s(\mathbf{y}_+ - \mathbf{y}_-) = \tilde{\mathbf{f}}(\mathbf{y}_+) - \tilde{\mathbf{f}}(\mathbf{y}_-). \quad (13)$$

This motivates the notion of Hugoniot curves through a fixed  $\mathbf{y}^\dagger \in \mathbf{Y}(\Omega')$  defined as the curves along which

$$s(\mathbf{y}^\dagger - \mathbf{y}) = \tilde{\mathbf{f}}(\mathbf{y}^\dagger) - \tilde{\mathbf{f}}(\mathbf{y}) \quad (14)$$

holds. In fact, the Hugoniot curves coincide with  $k = \text{constant}$  and  $\tilde{V} = \text{constant}$  (see e.g. Section 3 in [39]), which we have already seen are the 1-rarefaction curves and curves for the 2-discontinuity, respectively. As rarefaction waves and discontinuities follow the same curves in the phase plane, the system belongs to the Temple class [40].

**2.2. Admissible solution.** The system properties derived in Section 2.1 are restricted to the domain  $\mathbf{Y}(\Omega')$  on which  $k$ ,  $\tilde{V}$  and  $\mathcal{J}$  are well defined and the system (4) is strictly hyperbolic. To include the vacuum ( $\phi = 0$ ) and the saturation ( $V = 0$ ) cases in the discussion, we now move back to the original variables contained in the vector  $\mathbf{u}$ . We denote by

$$\mathcal{H}_1(\mathbf{u}^\dagger) := \{\mathbf{u} \in \Omega : k = k^\dagger\} \quad \text{and} \quad \mathcal{H}_2(\mathbf{u}^\dagger) := \{\mathbf{u} \in \Omega : V = V^\dagger\},$$

the Hugoniot/rarefaction curves extended to the whole of  $\Omega$ . Here, the subscripts emphasize the association of the curves with the two characteristic fields. The set  $\mathcal{H}_1(\mathbf{u}^\dagger)$  contains the states that may be connected to  $\mathbf{u}^\dagger$  via shocks or rarefaction waves or combinations of such along the 1-characteristics, whereas  $\mathcal{H}_2(\mathbf{u}^\dagger)$  is the set of states that can be joined to  $\mathbf{u}^\dagger$  across a contact discontinuity along the 2-characteristics. We note that  $\phi \neq \phi^\dagger$  for all  $\mathbf{u} \in \mathcal{H}_1(\mathbf{u}^\dagger) \cup \mathcal{H}_2(\mathbf{u}^\dagger)$  with  $\mathbf{u} \neq \mathbf{u}^\dagger$  and derive from (14) that the slope  $s$  of a possible discontinuity (also visible in the conserved variables) between  $\mathbf{u}^\dagger$  and  $\mathbf{u}$  is given by

$$s = \sigma(\mathbf{u}, \mathbf{u}^\dagger) := \frac{f(\mathbf{u}^\dagger) - f(\mathbf{u})}{\phi^\dagger - \phi}.$$

To single out physically relevant discontinuities, we use the entropy condition by Liu [31].

**Definition 2.1** (Admissible solution). *Suppose that  $\mathbf{u}_-(t), \mathbf{u}_+(t)$  are such that  $\mathbf{Y}(\mathbf{u}_-) \in \mathbf{Y}(\mathcal{H}_i(\mathbf{u}_+))$ . A discontinuity with a jump from  $\mathbf{u}_-$  to  $\mathbf{u}_+$  is admissible if  $\mathbf{Y}(\mathbf{u}_-) = \mathbf{Y}(\mathbf{u}_+)$  or if it propagates with speed  $s = \sigma(\mathbf{u}_-, \mathbf{u}_+)$  and for almost every  $t$ ,*

$$\sigma(\mathbf{u}_-, \mathbf{u}_+) \leq \sigma(\boldsymbol{\eta}(\tau), \mathbf{u}_+) \quad \text{for all } \tau \in [0, 1], \quad (15)$$

whenever  $\boldsymbol{\eta}$  is a parametrization of  $\mathcal{H}_i(\mathbf{u}_+)$  such that  $\mathbf{Y}(\boldsymbol{\eta}(0)) = \mathbf{Y}(\mathbf{u}_-)$  and  $\mathbf{Y}(\boldsymbol{\eta}(1)) = \mathbf{Y}(\mathbf{u}_+)$ .

A weak, piecewise continuous solution  $\mathbf{u}$  of (10)–(11) with a finite number of discontinuities is an admissible solution if every discontinuity in  $\mathbf{u}$  is admissible.

Definition 2.1 does not offer a condition for uniqueness for general initial data. In particular, discontinuities in vacuum – where differences in the  $k$ -component comprise the jump – are admissible independently of their propagation speed. However, such discontinuities are not observable in the conserved variables. It will be seen in Section 3 that vacuum may

cause non-uniqueness also in regions where  $\phi > 0$ . This is a more severe consequence of the insufficiency of the above definition as it becomes evident also in the mapped solution  $\mathbf{Y}(\mathbf{u})$ . A remedy would be to impose an extra condition for admissibility, but we intentionally leave this out and instead introduce the notion of type A and type B solutions in Section 3.

**2.3. Notations.** The convergence analysis in Section 5 relies on a TVD-property of the numerical solution measured in the coordinate system of Riemann invariants. Therefore, it is convenient to introduce the map  $\mathbf{R}(\mathbf{u}) := (V(\mathbf{u}), k)^T$  from  $\Omega$  to the  $(V, k)$ -plane. We recall the definition of  $\Phi(\cdot, k)$  as the inverse of  $V(\cdot, k)$  and note that  $\mathbf{R}^{-1}(V, k) = (\Phi(V, k), k)^T$  on  $\mathbf{R}(\Omega)$ .

As was mentioned in Section 1.1, we will appeal to an invariance property of the following set

$$\mathcal{M}(\mathbf{u}_1, \mathbf{u}_2) := \{\mathbf{u} \in \Omega : k \in \text{ch}(k_1, k_2), V \in \text{ch}(V_1, V_2)\}, \quad (16)$$

which is defined for  $\mathbf{u}_1, \mathbf{u}_2 \in \Omega$ . Here,  $\text{ch}(b_1, b_2) := [\min(b_1, b_2), \max(b_1, b_2)]$  denotes the convex hull of two finite real numbers  $b_1$  and  $b_2$ . We remark that  $\mathbf{R}(\mathcal{M}(\mathbf{u}_1, \mathbf{u}_2))$  is a rectangle (or a subset of such) in the  $(V, k)$ -plane having  $\mathbf{R}(\mathbf{u}_1)$  and  $\mathbf{R}(\mathbf{u}_2)$  as diagonally opposite corners.

Throughout this work,  $|\cdot|$  denotes the 1-norm on  $\mathbb{R}^d$  and by  $\|\cdot\|_{L^p}$  and T.V.( $\cdot$ ) we mean the  $L^p$ -norm and the total variation, respectively, defined in terms of  $|\cdot|$ .

At several instances we use self-explanatory super- and subscripts on function quantities. The notation is always inherited from the arguments. For example, we write  $V_L$  and  $V_R$  for  $V(\mathbf{u}_L)$  and  $V(\mathbf{u}_R)$ , respectively, and by  $V_j^n$  we mean  $V(\mathbf{u}_j^n)$ .

### 3. THE RIEMANN PROBLEM

The Riemann problem is defined as the Cauchy problem (10)–(11) with

$$\mathbf{u}_0(x) = \begin{cases} \mathbf{u}_L & \text{if } x \leq 0, \\ \mathbf{u}_R & \text{if } x > 0, \end{cases} \quad (17)$$

where  $\mathbf{u}_L, \mathbf{u}_R \in \Omega$ . We give here a rather detailed review on its solution since this information is required for the analysis of the numerical schemes in Section 4.

**3.1. Type A solution.** As pointed out by Liu [31], condition (15) generalizes Oleinik's [35] entropy inequality for scalar equations. This close relationship between Liu's condition on the system side and Oleinik's condition on the scalar side becomes particularly important when  $\mathbf{u}_L$  and  $\mathbf{u}_R$  can be connected via 1-waves only. On  $\mathcal{H}_1(\mathbf{u}_L)$  we have  $k = k_L$  and the system (10) reduces to the scalar equation

$$\phi_t + f(\phi, k_L)_x = 0, \quad (18)$$

for which Oleinik's condition reads

$$\frac{f(\phi_+, k_L) - f(\phi_-, k_L)}{\phi_+ - \phi_-} \leq \frac{f(\phi, k_L) - f(\phi_-, k_L)}{\phi - \phi_-} \quad \text{for all } \phi \text{ between } \phi_- \text{ and } \phi_+. \quad (19)$$

For  $\mathbf{u}_-, \mathbf{u}_+ \in \mathcal{H}_1(\mathbf{u}_L)$ , the Rankine-Hugoniot condition (13) and the entropy condition (15) are equivalent to the corresponding conditions for (18), which are

$$s(\phi_+ - \phi_-) = f(\phi_+, k_L) - f(\phi_-, k_L)$$

and (19), respectively. Consequently,  $\mathbf{u} = (\phi, k_L)^T$  is an admissible solution to (10)–(11), (17) if  $k_L = k_R$  and  $\phi$  is the entropy solution (in the sense of (19)) to (18) with the initial datum

$$\phi(x, 0) = \begin{cases} \phi_L & \text{if } x \leq 0, \\ \phi_R & \text{if } x > 0, \end{cases} \quad x \in \mathbb{R}.$$

Let us now turn to the situation when  $k_L \neq k_R$  and recall from (12) that  $\lambda_1 < \lambda_2$ . The solution contains a contact discontinuity propagating with speed  $V_R$  (in the  $(x, t)$ -plane) and separating  $\mathbf{u}_R$  to the right from an intermediate state  $\mathbf{u}^*(\mathbf{u}_L, \mathbf{u}_R)$  to the left. If  $V(0, k_L) \geq V_R$  (cf. Figure 3), then this intermediate state is defined as the intersecting point of  $\mathcal{H}_1(\mathbf{u}_L)$  and  $\mathcal{H}_2(\mathbf{u}_R)$ , i.e. the solution to

$$\begin{cases} k = k_L, \\ V(\phi, k) = V_R, \end{cases} \quad (20)$$

cf. Figure 3 (upper mid). The existence and uniqueness in  $\Omega$  of such a solution follows since  $V(\cdot, k)$  has a zero in  $[0, 1]$  and is invertible on each straight line  $k = \text{constant}$  in  $\Omega$ . By definition,  $\Phi(V_R, k_L)$  is the solution to  $V(\phi, k_L) = V_R$  and we thus get  $\mathbf{u}^* = (\Phi(V_R, k_L), k_L)^T$ . If  $V(0, k_L) < V_R$ , then there exists no solution to the system (20) (cf. Figure 3, lower mid) and we let  $\mathbf{u}^* = (0, k_L)^T$ .

As  $\mathbf{u}^*(\mathbf{u}_L, \mathbf{u}_R) \in \mathcal{H}_1(\mathbf{u}_L)$ , the connection between  $\mathbf{u}_L$  and  $\mathbf{u}^*(\mathbf{u}_L, \mathbf{u}_R)$  may be constructed via the solution to the scalar conservation law (18), keeping  $k = k_L$  constant. We formulate the overall solution in the following theorem, the proof of which is straightforward once we realize that the possible contact discontinuity with a jump between  $\mathbf{u}^*(\mathbf{u}_L, \mathbf{u}_R)$  and  $\mathbf{u}_R$  is admissible since  $\sigma(\mathbf{u}, \mathbf{u}_R) = V_R$  for all  $\mathbf{u} \in \mathcal{H}_2(\mathbf{u}_R) \setminus \{\mathbf{u}_R\}$ .

**Theorem 3.1.** *Consider  $\mathbf{u}_L, \mathbf{u}_R \in \Omega$  and let  $\phi^*(\mathbf{u}_L, \mathbf{u}_R)$  be the first component of*

$$\mathbf{u}^*(\mathbf{u}_L, \mathbf{u}_R) := \begin{cases} (\Phi(V_R, k_L), k_L)^T & \text{if } V(0, k_L) \geq V_R, \\ (0, k_L)^T & \text{if } V(0, k_L) < V_R. \end{cases} \quad (21)$$

If  $\phi_s$  is the (Oleinik) entropy solution to (18) with initial datum

$$\phi(x, 0) = \begin{cases} \phi_L & \text{if } x \leq 0, \\ \phi^*(\mathbf{u}_L, \mathbf{u}_R) & \text{if } x > 0, \end{cases}$$

then the function

$$\mathbf{u}(x, t) := \begin{cases} (\phi_s(x, t), k_L)^T & \text{if } x \leq V_R t, \\ \mathbf{u}_R & \text{if } x > V_R t, \end{cases} \quad (22)$$

is an admissible solution of the Riemann problem (10)–(11), (17).

**Remark 3.1.** At first sight,  $\phi$  may appear to be non-conserved in the solution (22) if  $V(0, k_L) > V_R$  and  $\phi_L = \phi_R = 0$ . There is no mass in the system at time  $t = 0$ , but the intermediate density is given by  $\phi^* = \Phi(V_R, k_L) > 0$ . However,  $\mathbf{u}_L$  is connected to  $\mathbf{u}^*$  via a 1-shock propagating with speed  $\sigma(\mathbf{u}_L, \mathbf{u}^*) = V(\mathbf{u}^*) = V_R$ . Hence, we get  $\phi > 0$  only on the null-set  $x = V_R t$ .

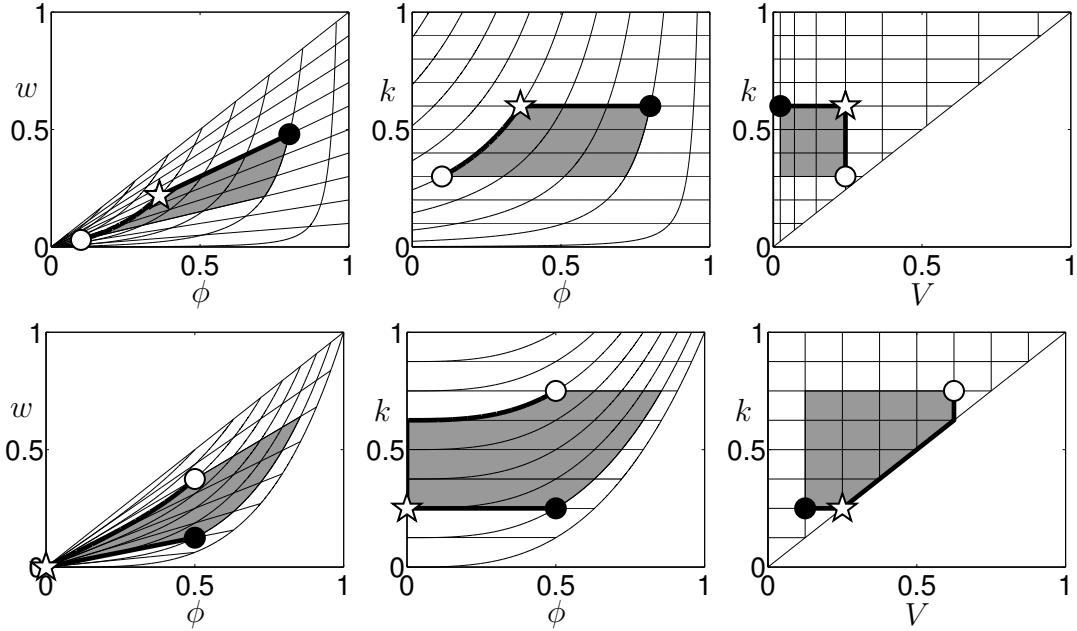


FIGURE 3. The grids show the phase planes  $\mathbf{Y}(\Omega)$  (left column),  $\Omega$  (middle column) and  $\mathbf{R}(\Omega)$  (right column) for the sedimentation model (top row) and the ARZ model (bottom row). The Hugoniot curves (solid lines) and a Riemann solution (thick lines) are shown in each plane. The mappings of the states  $\mathbf{u}_L$  ( $\bullet$ ),  $\mathbf{u}_R$  ( $\circ$ ) and  $\mathbf{u}^*(\mathbf{u}_L, \mathbf{u}_R)$  ( $\star$ ) and the corresponding set  $\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R)$  (shaded region) are also plotted. These Riemann problems are described more carefully in Examples 2 and 3 in Section 6 and the solution profiles at time  $t = 1$  are seen in Figures 10 and 12, respectively.

It is possible to find a function  $V$  and a pair  $\mathbf{u}_L, \mathbf{u}_R \in \Omega$  such that the total variation of the Riemann solution  $\mathbf{u}(\cdot, t)$  is strictly larger than  $\text{T.V.}(\mathbf{u}_0(\cdot))$  for all  $0 < t \leq T$  (see e.g. the solution for the ARZ model in Figure 3). However, considering the Riemann invariants, we derive in Theorem 3.2 an invariance property of both  $\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R)$  and the total variation of the mapped solution  $\mathbf{R}(\mathbf{u})$ . In Figure 3, we have drawn  $\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R)$  and its images under  $\mathbf{Y}$  and  $\mathbf{R}$  in the  $(\phi, k)$ -,  $(\phi, w)$ - and  $(V, k)$ -plane, respectively.

**Theorem 3.2.** *Let  $\mathbf{u}$  be the solution (22) to the Riemann problem (10)–(11), (17). Then  $\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R)$  is an invariant region of  $\mathbf{u}$ , i.e.  $\mathbf{u}(x, t) \in \mathcal{M}(\mathbf{u}_L, \mathbf{u}_R)$  for all  $(x, t) \in \mathbb{R} \times [0, T]$ . Moreover, the spatial total variation of  $\mathbf{u}$  measured in the  $(V, k)$ -plane is constant:*

$$\text{T.V.}(\mathbf{R}(\mathbf{u}(\cdot, t))) = |\mathbf{R}(\mathbf{u}_R) - \mathbf{R}(\mathbf{u}_L)| \quad \text{for all } t \in [0, T]. \quad (23)$$

*Proof.* The invariance property of  $\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R)$  is easiest realized if we follow the construction of  $\mathbf{u}$  in the  $(V, k)$ -plane. For a fixed  $t > 0$  and  $x \leq V_R t$ , there holds  $k = k_L$  and  $V(\mathbf{u})$  is a monotone function of  $x$ , taking values between  $V_L$  and  $V(\mathbf{u}^*) \in \text{ch}(V_L, V_R)$ . The mapped solution  $\mathbf{R}(\mathbf{u})$  is thus kept within  $\mathbf{R}(\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R))$  since  $\mathbf{u} = \mathbf{u}_R$  for  $x > V_R t$ .

To prove (23), we note that  $\mathbf{u}(\cdot, t)$  is constant outside of the interval  $[\lambda_1(\mathbf{u}_L)t, V_R t + 0]$  wherefore we need only consider the total variation on this bounded part of  $\mathbb{R}$ . As the  $V$ -

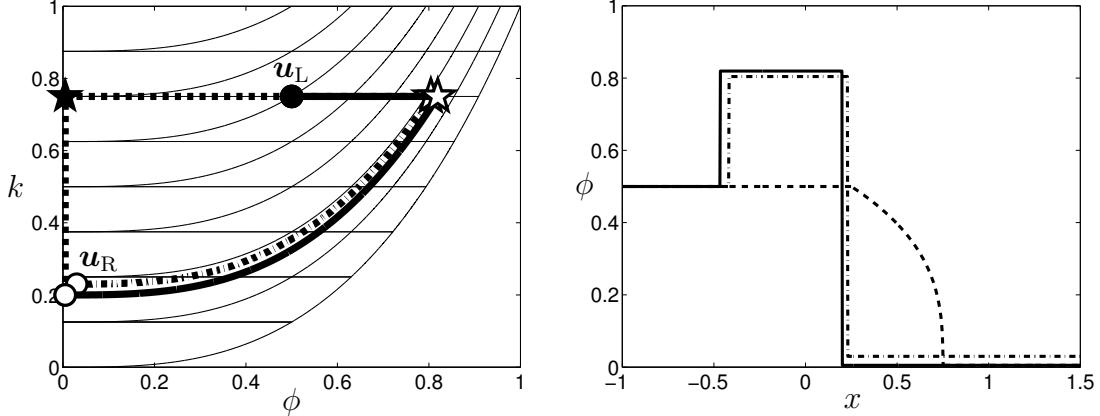


FIGURE 4. Three Riemann solutions at time  $t = 1$  for the ARZ model illustrated by their representations in the  $(\phi, k)$ -plane (left) and by their  $\phi$ -profiles at time  $t = 1$  (right). In all three solutions, the initial datum is chosen such that  $\mathbf{u}_L = (0.5, 0.75)^T$  and  $V_R < V(0, k_L)$ . The solid lines show the type A solution in Theorem 3.1 with  $\mathbf{u}_R = (0, 0.2)^T$ , while the dashed lines show the type B solution in Theorem 3.3 for the same initial datum. The intermediate states (21) and (24) of each type of solution are marked with open ( $\star$ ) and filled ( $\star$ ) stars, respectively. The dash-dot lines show the solution for a perturbed right state  $\mathbf{u}_R = (0, 0.2)^T + 0.03(1, 1)^T$  (the two types of solutions coincide in this case since  $\phi_R > 0$  implies  $(\mathbf{u}_L, \mathbf{u}_R) \in \mathcal{S}$ ).

and  $k$ -components of  $\mathbf{R}(\mathbf{u}(\cdot, t))$  both are monotone, we get

$$\text{T.V.}(V(\mathbf{u}(\cdot, t))) = |V(\mathbf{u}(V_R t + 0, t)) - V(\mathbf{u}(\lambda_1(\mathbf{u}_L)t, t))| = |V_R - V_L|$$

and, analogously,  $\text{T.V.}(k(\mathbf{u}(\cdot, t))) = |k_R - k_L|$ . The identity (23) now follows by summation of the contributions from the two Riemann invariants:

$$\text{T.V.}(\mathbf{R}(\mathbf{u}(\cdot, t))) = \text{T.V.}(V(\mathbf{u}(\cdot, t))) + \text{T.V.}(k(\mathbf{u}(\cdot, t))) = |V_R - V_L| + |k_R - k_L|. \quad \square$$

**3.2. Type B solution.** The Riemann solution given in Theorem 3.1 is stable in the sense that it depends continuously on the initial data. A small perturbation in  $\mathbf{u}_L$  or  $\mathbf{u}_R$  results in just a small alteration of  $\mathbf{u}$  and there will be no abrupt changes in the basic structure of the solution. This is a desired mathematical property, but it is not necessarily correct from a physical point of view. Consider for example a situation in traffic flow where a sequence of vehicles enters an interval of vacuum ( $\phi = 0$ ). The frontal vehicle is not slowed down by any driver ahead and thus travels freely with its own specific velocity. However, a disturbance in the vacuum interval by the addition of a single vehicle ( $\phi > 0$ ) having a slower velocity results in a traffic jam.

Mathematically, the vacuum case under discussion occurs when  $\mathbf{u}_L$  and  $\mathbf{u}_R$  are chosen such that  $\phi_L > 0$ ,  $\phi_R = 0$  and  $k_L > k_R$ . We have  $V_R = V(0, k_R) < V(0, k_L)$  and the solution (21)–(22) contains a contact discontinuity propagating with speed  $V_R$ . The front that should travel freely is thus hindered by some slow “ghost” vehicle (see Figure 4). We can overcome

this unrealistic behaviour by considering another admissible solution where  $\mathbf{u}_L$  is connected directly to vacuum along the 1-characteristics without taking any concern to  $V_R$  whenever  $\phi_R = 0$ . To be more precise, we redefine the intermediate state so that  $\mathbf{u}^* = (0, k_L)^T$  if  $\phi_R = 0$ , and keep the definition (21) otherwise. Hence, the solution will coincide with the one in Theorem 3.1 if the pair  $(\mathbf{u}_L, \mathbf{u}_R)$  belongs to

$$\mathcal{S} := \{(\mathbf{u}_1, \mathbf{u}_2) \in \Omega^2 : k_1 \leq k_2 \text{ if } \phi_2 = 0\}.$$

In the  $(x, t)$ -plane, the construction via the scalar solution  $\phi_s$  keeping  $k = k_L$  constant now reaches the line  $x = \mathcal{V}(\mathbf{u}_L, \mathbf{u}_R)t$ , where

$$\mathcal{V}(\mathbf{u}_L, \mathbf{u}_R) := \begin{cases} V_R & \text{if } (\mathbf{u}_L, \mathbf{u}_R) \in \mathcal{S}, \\ f_\phi(0, k_L) & \text{if } (\mathbf{u}_L, \mathbf{u}_R) \notin \mathcal{S}. \end{cases}$$

**Theorem 3.3.** *Let  $\mathbf{u}_L, \mathbf{u}_R \in \Omega$  and denote by  $\phi^*(\mathbf{u}_L, \mathbf{u}_R)$  the first component of*

$$\mathbf{u}^*(\mathbf{u}_L, \mathbf{u}_R) := \begin{cases} (\Phi(V_R, k_L), k_L)^T & \text{if } V(0, k_L) \geq V_R \text{ and } \phi_R > 0, \\ (0, k_L)^T & \text{if } V(0, k_L) < V_R \text{ or } \phi_R = 0. \end{cases} \quad (24)$$

*If  $\phi_s$  is defined analogously with Theorem 3.1, then the function*

$$\mathbf{u}(x, t) := \begin{cases} (\phi_s(x, t), k_L)^T & \text{if } x \leq \mathcal{V}(\mathbf{u}_L, \mathbf{u}_R)t, \\ \mathbf{u}_R & \text{if } x > \mathcal{V}(\mathbf{u}_L, \mathbf{u}_R)t, \end{cases} \quad (25)$$

*is an admissible solution of the Riemann problem (10)–(11), (17).*

*Proof.* It suffices to consider the case  $(\mathbf{u}_L, \mathbf{u}_R) \notin \mathcal{S}$ . We have  $\mathbf{u}^* = (0, k_L)^T$  and there is no contact discontinuity in the solution. The admissibility follows since the jump from  $\mathbf{u}^*$  to  $\mathbf{u}_R = (0, k_R)^T$  across  $x = f_\phi(\mathbf{u}^*)t$  satisfies  $\mathbf{Y}(\mathbf{u}^*) = \mathbf{Y}(\mathbf{u}_R)$ .  $\square$

#### 4. NUMERICAL SCHEMES

Throughout this paper, we work with equidistant spatial and temporal grids determined by the step lengths  $\Delta x > 0$  and  $\Delta t > 0$ , respectively. The real axis is divided into cells  $\mathcal{C}_j := [x_{j-1/2}, x_{j+1/2})$ ,  $j \in \mathbb{Z}$  with the boundaries located at  $x_{j+1/2} := j\Delta x$ , while the time axis is discretized at the points  $t_n := n\Delta t$  for the integers  $0 \leq n \leq T/\Delta t =: N$ . The quantities  $\Delta x$  and  $\Delta t$  are chosen such that  $\alpha := \Delta t/\Delta x$  is bounded according to the CFL condition

$$\alpha \max\{\|f_\phi\|_\infty, \|V\|_\infty\} \leq \frac{1}{2}. \quad (26)$$

This inequality prevents information in the numerical solutions from travelling faster than the maximal signal speed of the system (10) (cf. the eigenvalues in (12)). For that reason, the bound (26) is always assumed to be satisfied. In Section 5 we will send  $\Delta x$  and  $\Delta t$  to zero keeping  $\alpha$  constant and these discretization parameters are therefore sometimes referred to by the common notation  $\Delta$ . The initial datum is assumed to be locally integrable and is discretized by means of the cell averages

$$\mathbf{u}_j^0 := \frac{1}{\Delta x} \int_{\mathcal{C}_j} \mathbf{u}_0(x) dx \quad \text{and} \quad \mathbf{y}_j^0 := \frac{1}{\Delta x} \int_{\mathcal{C}_j} \mathbf{Y}(\mathbf{u}_0(x)) dx. \quad (27)$$

**4.1. Some ill-behaved conservative schemes.** This section is devoted to conservative methods, i.e., schemes of the form

$$\mathbf{y}_j^{n+1} = \mathbf{y}_j^n - \alpha(\mathbf{F}_{j+1/2}^n - \mathbf{F}_{j-1/2}^n) \quad (28)$$

for some numerical flux  $\mathbf{F}_{j+1/2}^n$  to be defined. The intention is to point at the problems with such numerical schemes and thereby motivate why we turn to alternative methods in Section 4.2.

Let us begin the discussion with Godunov's method and temporarily assume that  $\phi_j^n, \phi_{j+1}^n > 0$ . If locally near each cell boundary  $x = x_{j+1/2}$  we solve the Riemann problem with initial data  $\tilde{\mathbf{u}}_j^n := \mathbf{Y}^{-1}(\mathbf{y}_j^n)$  and  $\tilde{\mathbf{u}}_{j+1}^n := \mathbf{Y}^{-1}(\mathbf{y}_{j+1}^n)$ , the solution will take a constant value, denote it by  $\tilde{\mathbf{u}}_{j+1/2}^{n,+}$ , along  $x = x_{j+1/2}$  for  $t > 0$ . This feature of the solution is utilized by defining  $\mathbf{F}_{j+1/2}^n := \mathbf{f}(\tilde{\mathbf{u}}_{j+1/2}^{n,+})$ . In (22) or (25) it is seen that  $\tilde{\mathbf{u}}_{j+1/2}^{n,+} = (\tilde{\phi}_{s,j+1/2}^n, \tilde{k}_j^n)^T$ , where  $\tilde{\phi}_{s,j+1/2}^n$  is the solution to the scalar Riemann problem associated with (18) for  $k_L = \tilde{k}_j^n$  and initial datum defined by the pair  $\tilde{\phi}_j^n = \phi_j^n, \phi^*(\tilde{\mathbf{u}}_j^n, \tilde{\mathbf{u}}_{j+1}^n)$ . Including the vacuum cases  $\phi_j^n = 0$  and  $\phi_{j+1}^n = 0$ , we get

$$\mathbf{F}_{j+1/2}^n = \begin{cases} \mathbf{0} & \text{if } \phi_j^n = 0, \\ g(\phi_j^n, \phi^*(\tilde{\mathbf{u}}_j^n, \tilde{\mathbf{u}}_{j+1}^n); \tilde{k}_j^n)(1, \tilde{k}_j^n)^T & \text{if } \phi_j^n > 0, \end{cases} \quad (29)$$

where

$$g(a, b; k) := \begin{cases} \min_{a \leq \phi \leq b} f(\phi, k) & \text{if } a \leq b, \\ \max_{b \leq \phi \leq a} f(\phi, k) & \text{if } a > b \end{cases} \quad (30)$$

is Godunov's flux related to the scalar flux function  $f(\cdot, k)$ . To evaluate the flux (29), we need to compute  $\phi^*(\tilde{\mathbf{u}}_j^n, \tilde{\mathbf{u}}_{j+1}^n)$  if  $\phi_j^n > 0$ . However, this is not possible in the sense of (21) if  $\phi_{j+1}^n = 0$  because then  $\tilde{k}_{j+1}^n$  is not well defined from the conserved state vector  $\mathbf{y}_{j+1}^n$ . The definition (24) of  $\phi^*$ , on the other hand, causes no complications in this respect and the flux (30) is defined for all  $\mathbf{y}_j^n, \mathbf{y}_{j+1}^n \in \Omega$ . Henceforth, when referring to the underlying Riemann solution of Godunov's method, we thus mean the type B solution in Theorem 3.3.

To illustrate the shortcomings of Godunov's method, we return to the Riemann problem (10)–(11), (17) and choose  $\mathbf{u}_L, \mathbf{u}_R \in \Omega'$  such that  $\mathbf{u}_L \neq \mathbf{u}_R$  and  $V_L = V_R$ . The exact solution has one wave only, namely a contact discontinuity propagating with speed  $V_L = V_R$ :

$$\mathbf{u}(x, t) = \begin{cases} \mathbf{u}_L & \text{if } x \leq V_R t, \\ \mathbf{u}_R & \text{if } x > V_R t. \end{cases} \quad (31)$$

Note that  $(\mathbf{u}_L, \mathbf{u}_R) \in \mathcal{S}$  and the type A and B solutions coincide. Moreover, we have  $\mathbf{u} \in \mathcal{M}(\mathbf{u}_L, \mathbf{u}_R) = \{\mathbf{u} : k \in \text{ch}(k_L, k_R), V = V_R\}$ . We let the straight line  $\mathbf{y}_\ell$  passing through  $\mathbf{y}_L = \mathbf{Y}(\mathbf{u}_L)$  and  $\mathbf{y}_R = \mathbf{Y}(\mathbf{u}_R)$  be parametrized as follows:

$$\mathbf{y}_\ell(\tau; \mathbf{y}_L, \mathbf{y}_R) := \tau \mathbf{y}_R + (1 - \tau) \mathbf{y}_L, \quad \tau \in \mathbb{R}. \quad (32)$$

After the first Godunov update, we get for  $j = 1$ :

$$\mathbf{y}_1^1 = \mathbf{y}_R - \alpha(V_R \mathbf{y}_R - V_R \mathbf{y}_L) = \mathbf{y}_\ell(1 - \alpha V_R; \mathbf{y}_L, \mathbf{y}_R).$$

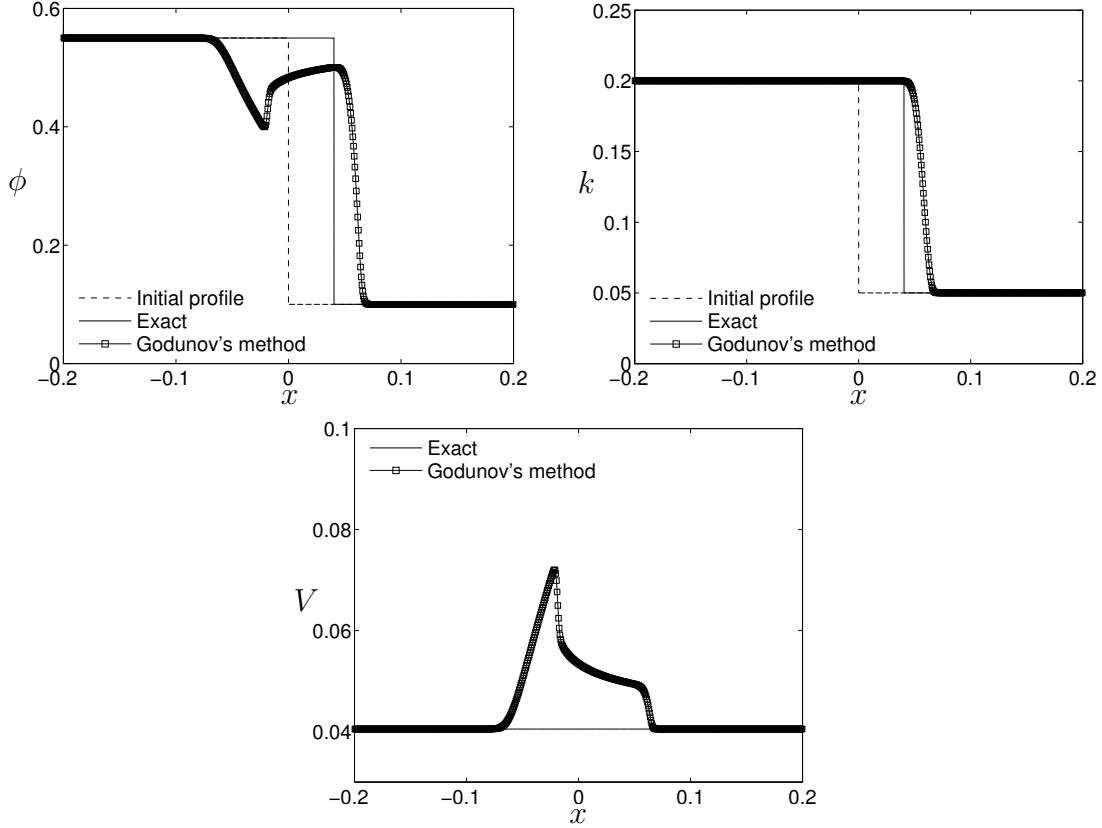


FIGURE 5. A numerical solution at time  $t = 1$  of a Riemann problem associated with the sedimentation model (6) generated with Godunov's method ( $\Delta x = 0.01 \cdot 2^{-4}$ ). The initial datum is determined by  $\mathbf{u}_L = (0.55, 0.2)^T$  and  $\mathbf{u}_R = (0.1, 0.05)^T$ . This means that  $V_L = V_R = 0.0405$  and the exact solution contains an isolated contact discontinuity.

The CFL condition (26) implies  $1 - \alpha V_R \in (0, 1)$ . Unless the 2-Hugoniot curves are straight lines in the  $(\phi, w)$ -plane, we thus cannot assure that  $\tilde{V}_1^1 = V_R$  and the numerical approximation may leave the invariant region  $\mathbf{Y}(\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R))$  of the exact solution. If this happens, we can expect the discontinuity to be smeared and spurious oscillations to develop. Such behaviour is observed in Figure 5, where we show the result from Godunov's method applied to a Riemann problem associated with the sedimentation model (6) (this undesirable feature of Godunov's method when applied to the ARZ model was reported by Chalons and Goatin [9]).

The sample problem above stresses the importance of imposing a requirement that the numerical solution should be preserved within  $\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R)$  if the exact solution has a contact discontinuity. Another natural requirement near any discontinuity is that the  $\phi$ -component of the approximate solution should be bounded by the values determining the jump, e.g. contained in  $\text{ch}(\phi_L, \phi_R)$  in the case of the isolated contact discontinuity in (31). However, as the following proposition states, both requirements cannot in general be met simultaneously

for a large family of schemes. This family includes all conservative schemes based on consistent two-point fluxes (e.g. Godunov's and Roe's methods and the HLL-scheme by Harten et. al. [19]) and the substance of the proposition is that contact discontinuities are inadequately approximated with spurious overshoots either in the  $(\phi, k)$ - or in the  $(V, k)$ -coordinates.

**Proposition 4.1.** *Consider the Riemann problem (10)–(11), (17) and suppose that there is a 2-Hugoniot curve  $V = V^\dagger$ , which is  $C^2$  and not a straight line in  $\mathbf{Y}(\Omega')$ . Then there exist two states  $\mathbf{u}_L, \mathbf{u}_R \in \Omega'$  on this curve such that no conservative scheme in the form (28) with a numerical flux satisfying*

$$\mathbf{F}_{j+1/2}^0 = \begin{cases} \mathbf{f}(\mathbf{u}_L) & \text{if } j < 0, \\ \mathbf{f}(\mathbf{u}_R) & \text{if } j > 0, \end{cases} \quad (33)$$

*produces an approximate solution contained in  $\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R) \cap \{\mathbf{u} : \phi \in \text{ch}(\phi_L, \phi_R)\}$ .*

*Proof.* Since  $V_k = \tilde{V}_w \phi$ , the monotonicity assumption in (3) gives  $\tilde{V}_w > 0$  on  $\mathbf{Y}(\Omega')$  and by the implicit function theorem we can write  $w = W(\phi)$  on  $\tilde{V} = V^\dagger$  for some  $C^2$ -function  $W$ . The non-zero curvature of  $\tilde{V} = V^\dagger$  ensures the existence of  $\mathbf{u}_L$  and  $\mathbf{u}_R$  such that  $W$  is strictly convex or concave on the interval  $\text{ch}(\phi_L, \phi_R)$ . We assume, without loss of generality, the former.

The property (33) implies the following first updates on the cells  $\mathcal{C}_0$  and  $\mathcal{C}_1$ :

$$\mathbf{y}_0^1 = \mathbf{y}_L - \alpha(\mathbf{F}_{1/2}^0 - V^\dagger \mathbf{y}_L), \quad \mathbf{y}_1^1 = \mathbf{y}_R - \alpha(V^\dagger \mathbf{y}_R - \mathbf{F}_{1/2}^0).$$

We recall the definition (32) of  $\mathbf{y}_\ell$  and note that

$$\mathbf{y}_\ell(1/2; \mathbf{y}_0^1, \mathbf{y}_1^1) = \mathbf{y}_\ell(1/2; \mathbf{y}_1^1, \mathbf{y}_0^1) = \mathbf{y}_\ell((1 - \alpha V^\dagger)/2; \mathbf{y}_L, \mathbf{y}_R) =: \mathbf{y}_*. \quad (34)$$

Assume that  $\mathbf{y}_0^1, \mathbf{y}_1^1 \in \mathbf{Y}(\mathcal{M}(\mathbf{u}_L, \mathbf{u}_R) \cap \{\mathbf{u} : \phi \in \text{ch}(\phi_L, \phi_R)\})$ , then  $\tilde{V}_0^1 = \tilde{V}_1^1 = V^\dagger$ . By (34), the straight line parametrized as  $\mathbf{y}_\ell(\cdot; \mathbf{y}_0^1, \mathbf{y}_1^1)$  or  $\mathbf{y}_\ell(\cdot; \mathbf{y}_1^1, \mathbf{y}_0^1)$  share the common point  $\mathbf{y}_*$  with the line  $\mathbf{y}_\ell(\cdot; \mathbf{y}_L, \mathbf{y}_R)$ . The strict convexity of  $W$  implies that these two lines are identical and the only possibilities for the updates are  $(\mathbf{y}_0^1, \mathbf{y}_1^1) = (\mathbf{y}_L, \mathbf{y}_R)$  or  $(\mathbf{y}_0^1, \mathbf{y}_1^1) = (\mathbf{y}_R, \mathbf{y}_L)$ . However, the CFL condition (26) yields  $(1 - \alpha V^\dagger)/2 < 1/2$  and we reach a contradiction in (34).  $\square$

**4.2. Averaging-remap schemes.** Given a sequence  $\{\mathbf{u}_j^n\}_{j \in \mathbb{Z}}$ , we define  $\mathbf{u}_\Delta^n : \mathbb{R} \times [0, \Delta t] \rightarrow \Omega$  as an exact solution of (10)–(11) with the piecewise constant initial datum  $\mathbf{u}_0(x) = \mathbf{u}_j^n$  on  $\mathcal{C}_j$  for all  $j \in \mathbb{Z}$ . Due to the CFL condition (26),  $\mathbf{u}_\Delta^n$  can be understood as the result of gluing together a series of Riemann solutions near each cell interface  $x_{j+1/2}$ , where locally  $\mathbf{u}_L = \mathbf{u}_j^n$  and  $\mathbf{u}_R = \mathbf{u}_{j+1}^n$ . If not otherwise stated, we understand these Riemann solutions in the sense of type A in Theorem 3.1.

If  $\mathbf{y}_j^n = \mathbf{Y}(\mathbf{u}_j^n)$  and  $\mathbf{y}_j^{n+1}$  is interpreted as the cell average

$$\mathbf{y}_j^{n+1} = \frac{1}{\Delta x} \int_{\mathcal{C}_j} \mathbf{Y}(\mathbf{u}_\Delta^n(x, \Delta t)) \, dx,$$

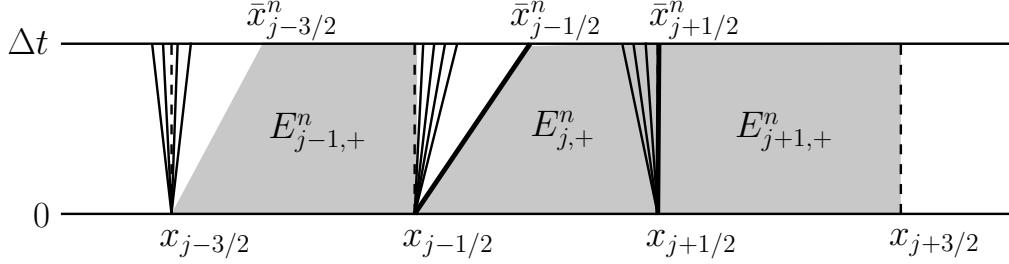


FIGURE 6. Contours of  $\phi_\Delta^n$  on the three rectangles  $\mathcal{C}_i \times [0, \Delta t]$ ,  $i = j-1, j, j+1$ . The shaded regions are the sets  $E_{i,+}^n$  in the partitions of each rectangle. In this example, there is no contact discontinuity in  $\mathcal{C}_{j-1} \times [0, \Delta t]$ , i.e.  $k_{j-2}^n = k_{j-1}^n$ , and saturation has been reached in  $\mathcal{C}_{j+1}$  ( $V_{j+1}^n = 0$ ).

then Godunov's method can be derived from the conservation law (10) divided by  $\Delta x$  and integrated over the rectangle  $\mathcal{C}_j \times [0, \Delta t]$ :

$$\mathbf{0} = \frac{1}{\Delta x} \iint_{\mathcal{C}_j \times [0, \Delta t]} (\mathbf{Y}(\mathbf{u}_\Delta^n)_t + \mathbf{f}(\mathbf{u}_\Delta^n)_x) \, dx \, dt. \quad (35)$$

After applying Green's formula on regions where  $\mathbf{u}_\Delta^n$  is smooth and the Rankine-Hugoniot condition (13) across discontinuities, we find (28) with the numerical flux  $\mathbf{F}_{j+1/2}^n = \mathbf{f}(\mathbf{u}_{j+1/2}^{n,+})$ , where  $\mathbf{u}_{j+1/2}^{n,+} := \mathbf{u}_\Delta^n(x_{j+1/2}, t)$ ,  $t \in (0, \Delta t]$ .

The averaging-remap schemes of this section are given in two sequential steps, where the first step means a slight modification of Godunov's approach by restricting the domain of integration in (35) to regions on which  $k_\Delta^n$  is constant. To be more precise, we divide the rectangle  $\mathcal{C}_j \times [0, \Delta t]$  into the two sets

$$\begin{aligned} E_{j,-}^n &:= \{(x, t) \in \mathcal{C}_j \times [0, \Delta t] : x_{j-1/2} < x < x_{j-1/2} + V_j^n t\}, \\ E_{j,+}^n &:= \{(x, t) \in \mathcal{C}_j \times [0, \Delta t] : x_{j-1/2} + V_j^n t < x < x_{j+1/2}\}. \end{aligned}$$

We avoid taking averages over contact discontinuities (which follow the straight lines  $x = x_{j-1/2} + V_j^n t$  if they are present) by defining the first step in terms of the two state vectors:

$$\begin{aligned} \mathbf{u}_{j,-}^{n+1} &:= \begin{cases} \mathbf{u}_{j-1}^n & \text{if } V_j^n = 0, \\ \frac{1}{\overline{\Delta x}_{j,-}^n} \int_{x_{j-1/2}}^{\bar{x}_{j-1/2}^n} \mathbf{u}_\Delta^n(x, \Delta t) \, dx & \text{if } V_j^n > 0, \end{cases} \\ \mathbf{u}_{j,+}^{n+1} &:= \frac{1}{\overline{\Delta x}_{j,+}^n} \int_{\bar{x}_{j-1/2}^n}^{x_{j+1/2}} \mathbf{u}_\Delta^n(x, \Delta t) \, dx, \end{aligned} \quad (36)$$

where  $\bar{x}_{j-1/2}^n := x_{j-1/2} + \Delta t V_j^n$ ,  $\overline{\Delta x}_{j,-}^n := \bar{x}_{j-1/2}^n - x_{j-1/2}$  and  $\overline{\Delta x}_{j,+}^n := x_{j+1/2} - \bar{x}_{j-1/2}^n$ . To bring back the approximations to the original mesh, a family of remap strategies can be used. This remapping is the second step, and its specific definition is what distinguishes the numerical methods discussed henceforth.

4.2.1. *Step 1: averaging.* The updating formula that maps  $\{\mathbf{u}_j^n\}_j$  to  $\{\mathbf{u}_{j,\pm}^{n+1}\}_j$  becomes rather simple because  $V(\mathbf{u}_\Delta^n)$  is continuous in a neighbourhood of each straight line  $\gamma_{j-1/2}^n(t)$  defined by  $x = x_{j-1/2} + V_j^n t$ . The net flux of  $\mathbf{u}_\Delta^n$  relative to the speed  $V_j^n$  thereby vanishes along  $\gamma_{j-1/2}^n(t)$ , i.e.

$$\int_0^{\Delta t} (V(\mathbf{u}_\Delta^n) \phi_\Delta^n - V_j^n \phi_\Delta^n) \Big|_{\gamma_{j-1/2,\pm}^n} dt = 0, \quad (37)$$

where  $\gamma_{j-1/2,\pm}^n$  denote the spatial one-sided limits of  $\gamma_{j-1/2}^n(t)$ .

Recall the definition of the scalar Godunov flux  $g$  from (30). To shorten the notation, we let  $g_{j-1/2}^n := g(\phi_{j-1}^n, \phi^\star(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n); k_{j-1}^n)$ . If  $V_j^n > 0$ , then we use that  $k_\Delta^n = k_{j-1}^n$  on  $E_{j-1,+}^n \cup E_{j,-}^n$ , integrate the first equation of System (10) over  $E_{j,-}^n$ , divide by  $\overline{\Delta x}_{j,-}^n$  and apply the left limit of (37) to obtain

$$\mathbf{u}_{j,-}^{n+1} = \begin{cases} \mathbf{u}_{j-1}^n & \text{if } V_j^n = 0, \\ \left( \frac{\Delta t}{\overline{\Delta x}_{j,-}^n} g_{j-1/2}^n, k_{j-1}^n \right)^T & \text{if } V_j^n > 0. \end{cases} \quad (38)$$

Analogous arguments over  $E_{j,+}^n$  yield

$$\mathbf{u}_{j,+}^{n+1} = \left( \frac{\Delta x}{\overline{\Delta x}_{j,+}^n} \phi_j^n - \frac{\Delta t}{\overline{\Delta x}_{j,+}^n} g_{j+1/2}^n, k_j^n \right)^T \quad (39)$$

if  $\phi_j^n$  is understood as the cell average of  $\phi_\Delta^n(\cdot, 0)$  over  $\mathcal{C}_j$ .

4.2.2. *Step 2: remap.* We perform the remapping from  $\{[x_{j-1/2}, \bar{x}_{j-1/2}^n), [\bar{x}_{j-1/2}^n, x_{j+1/2})\}$  to  $\mathcal{C}_j$  while bounding the Riemann invariants  $V$  and  $k$  by their values at  $\mathbf{u}_{j,\pm}^{n+1}$ . To be more precise, we require that

$$\mathbf{u}_j^{n+1} \in \mathcal{M}(\mathbf{u}_{j,-}^{n+1}, \mathbf{u}_{j,+}^{n+1}). \quad (40)$$

The condition (40) yields a whole family of remapping techniques and as will be seen in Section 5, they all give rise to schemes producing approximations that converge to some limit function. However, this limit need not be an admissible solution of (10). In fact, we prove convergence to a weak solution only for one particular choice of remap satisfying (40), namely the random sampling technique that yields the Av-RS scheme.

**Definition 4.1** (Av-RS scheme). *Let  $\mathbf{a} = \{a_n\}_{n=1}^\infty$  be a uniformly distributed sequence in  $(0, 1)$ . The Av-RS scheme is given by (38)–(39) and the following sampling remap:*

$$\mathbf{u}_j^{n+1} := \begin{cases} \mathbf{u}_{j,-}^{n+1} & \text{if } a_{n+1} \in (0, \alpha V_j^n), \\ \mathbf{u}_{j,+}^{n+1} & \text{if } a_{n+1} \in [\alpha V_j^n, 1). \end{cases} \quad (41)$$

For the actual computations behind the numerical examples in Section 6, we use the van der Corput sequence defined by [12]

$$a_n := \sum_{i=0}^I b_i 2^{-(i+1)},$$

where  $n = \sum_{i=0}^I b_i 2^i$ ,  $b_i \in \{0, 1\}$  is the binary expansion of  $n \in \mathbb{N}$ .

In Section 6, we also examine numerically the performance of a second remap, which is constructed via a convex combination of the state vectors  $\mathbf{u}_{j,-}^{n+1}$  and  $\mathbf{u}_{j,+}^{n+1}$  in their  $(V, k)$ -coordinates. The complete scheme is defined as follows.

**Definition 4.2** (Av-CC scheme). *The scheme obtained by (38)–(39) together with the remap*

$$\mathbf{u}_j^{n+1} := \mathbf{R}^{-1}(\alpha V_j^n \mathbf{R}(\mathbf{u}_{j,-}^{n+1}) + (1 - \alpha V_j^n) \mathbf{R}(\mathbf{u}_{j,+}^{n+1})), \quad (42)$$

*is called the Av-CC scheme.*

In the the CC-remap (42), it is understood that  $\mathbf{R}^{-1}$  is defined for the convex combination, at which it is evaluated. We recall that for the ARZ model (9) and the sedimentation model (6), the set  $\mathbf{R}(\Omega)$  is convex and the CC-remap is applicable on every pair  $(\mathbf{u}_{j,-}^{n+1}, \mathbf{u}_{j,+}^{n+1}) \in \Omega^2$ .

**4.2.3. Local type B Riemann solutions.** By replacing  $V_j^n$  with  $\mathcal{V}_{j-1/2}^n := \mathcal{V}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)$  and interpreting  $\mathbf{u}^*$  in the sense (24) at every instance in Section 4.2, we obtain numerical schemes based on  $\mathbf{u}_\Delta^n$  defined in terms of the unstable type B Riemann solution in Theorem 3.3. Recall that the type A and type B solutions coincide for initial data in  $\mathcal{S}$ . In the following proposition, it is claimed that  $\mathcal{S}$  is in a sense an invariant region under the Av-RS scheme. As an immediate consequence, if the initial data satisfy the requirements of this proposition, it is insignificant which type of Riemann solution we choose in the definition of  $\mathbf{u}_\Delta^n$  (for the Av-RS scheme).

**Proposition 4.2.** *Consider the Av-RS scheme in Definition 4.1. If  $(\mathbf{u}_{j-1}^0, \mathbf{u}_j^0) \in \mathcal{S}$  for all  $j$ , then  $(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) \in \mathcal{S}$  for all  $j$  and  $0 \leq n \leq N$ .*

*Proof.* We will argue by induction and assume therefore that the assertion holds for a fixed  $0 \leq n < N$ , i.e.  $(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) \in \mathcal{S}$  for all  $j$ . Suppose that  $\phi_j^{n+1} = 0$  for some  $j$  and recall from (41) that  $\mathbf{u}_j^{n+1}$  equals either  $\mathbf{u}_{j,-}^{n+1}$  or  $\mathbf{u}_{j,+}^{n+1}$ , depending on the value of  $a_{n+1}$ .

Consider first the case  $\mathbf{u}_j^{n+1} = \mathbf{u}_{j,+}^{n+1}$ , then  $a_{n+1} \geq \alpha V_j^n$  and  $\mathbf{u}_j^n = \mathbf{u}_j^{n+1}$ . This implies that  $\phi_j^n = 0$  and by assumption  $k_j^n \geq k_{j-1}^n$ . Using the monotonicity of  $V$ , we get  $V_j^n \geq V(0, k_{j-1}^n) \geq V_{j-1}^n$  and thus  $a_{n+1} \geq \alpha V_{j-1}^n$ . This means that  $\mathbf{u}_{j-1}^{n+1} = \mathbf{u}_{j-1,+}^{n+1}$  and

$$k_{j-1}^{n+1} = k_{j-1,+}^{n+1} = k_{j-1}^n \leq k_j^n = k_j^{n+1}.$$

If the sampling step gave  $\mathbf{u}_j^{n+1} = \mathbf{u}_{j,-}^{n+1}$  instead, then  $V_j^n > 0$ . Since  $\phi_{j,-}^{n+1} = 0$  is the average of  $\phi_\Delta^n$  over  $[x_{j-1/2}, \bar{x}_{j-1}^n]$ , there holds  $\phi^*(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) = 0$ . Recall that  $\phi_\Delta^n$  is constructed such that  $\phi_{j-1}^n$  is connected to  $\phi^*(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)$  along the 1-characteristics with  $k = k_{j-1}^n$  and that  $f_\phi(\phi^*(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n), k_{j-1}^n) = f_\phi(0, k_{j-1}^n) \geq 0$ . We note that  $f_\phi(0, k_{j-1}^n) = V(0, k_{j-1}^n)$  so if  $f_\phi(0, k_{j-1}^n) = 0$ , then  $V_{j-1}^n = 0$  because  $0 \leq V_{j-1}^n = V(\phi_{j-1}^n, k_{j-1}^n) \leq V(0, k_{j-1}^n) = 0$ . Hence, the sampling step in the cell  $\mathcal{C}_{j-1}$  gave  $\mathbf{u}_{j-1}^{n+1} = \mathbf{u}_{j-1,+}^{n+1}$  and we have

$$k_{j-1}^{n+1} = k_{j-1,+}^{n+1} = k_{j-1}^n = k_{j,-}^{n+1} = k_j^{n+1}.$$

On the other hand, if  $f_\phi(0, k_{j-1}^n) > 0$ , then there holds  $\phi_{j-1}^n = 0$  because otherwise positive wave speeds (due to  $f_\phi(0, k_{j-1}^n) = 0$  and  $\phi^*(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) = 0$ ) would disturb the zero average

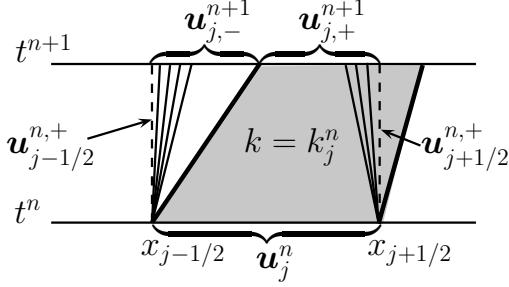


FIGURE 7. The rectangle  $\mathcal{C}_j \times [t_n, t_{n+1})$  with some of the state vectors. The shaded region shows where  $k_\Delta = k_j^n$ .

$\phi_{j,-}^{n+1}$ . Thus,  $k_{j-2}^n \leq k_{j-1}^n$  follows by assumption. If  $\mathbf{u}_{j-1}^{n+1} = \mathbf{u}_{j-1,+}^{n+1}$ , then

$$k_{j-1}^{n+1} = k_{j-1,+}^{n+1} = k_{j-1}^n = k_{j,-}^{n+1} = k_j^{n+1},$$

while  $\mathbf{u}_{j-1}^{n+1} = \mathbf{u}_{j-1,-}^{n+1}$  implies

$$k_{j-1}^{n+1} = k_{j-1,-}^{n+1} = k_{j-2}^n \leq k_{j-1}^n = k_{j,-}^{n+1} = k_j^{n+1}.$$

□

## 5. CONVERGENCE

We define the function  $\mathbf{u}_\Delta : \mathbb{R} \times [0, T) \rightarrow \Omega$  by  $\mathbf{u}_\Delta(x, t) := \mathbf{u}_\Delta^n(x, t - t_n)$  on each time strip  $\mathbb{R} \times [t_n, t_{n+1})$ . The purpose of this section is to prove convergence of  $\mathbf{u}_\Delta$  to a weak solution of the Cauchy problem (10)–(11) as the grid is refined, i.e. as  $\Delta \rightarrow 0$ . To this end, we need the invariance property of  $\mathcal{M}$  introduced in (16), and therefore understand  $\mathbf{u}_\Delta^n$  in the stable sense of type A Riemann solutions in Theorem 3.1. Convergence (along a subsequence) to a limit function  $\mathbf{u}$  is reached when  $\mathbf{u}_\Delta$  is generated by any of the members in the family of schemes (38)–(40). The proof is given via the following piecewise constant representation of the numerical approximations:

$$\hat{\mathbf{u}}_\Delta := \sum_{n=0}^{N-1} \sum_{j \in \mathbb{Z}} \chi_{\mathcal{C}_j \times [t_n, t_{n+1})} \mathbf{u}_j^n, \quad (43)$$

where  $\chi_{\mathcal{C}_j \times [t_n, t_{n+1})}$  is the characteristic function of  $\mathcal{C}_j \times [t_n, t_{n+1})$ . As was previously announced, to show that  $\mathbf{u}$  is a weak solution we restrict our attention to the Av-RS scheme (Definition 4.1).

To clarify the notation regarding  $\mathcal{M}$ , we consider a sequence  $\{\mathbf{u}_j\}_{j \in \mathcal{I}}$  in  $\Omega$  indexed by some set  $\mathcal{I}$  and observe the following identity:

$$\bigcup_{i,j \in \mathcal{I}} \mathcal{M}(\mathbf{u}_i, \mathbf{u}_j) = \left\{ \mathbf{u} : \inf_{j \in \mathcal{I}} \{k_j\} \leq k \leq \sup_{j \in \mathcal{I}} \{k_j\}, \inf_{j \in \mathcal{I}} \{V_j\} \leq V(\mathbf{u}) \leq \sup_{j \in \mathcal{I}} \{V_j\} \right\}.$$

**5.1. Convergence to a limit function.** The constant value  $\mathbf{u}_{j+1/2}^{n,+}$  that  $\mathbf{u}_\Delta^n$  takes on the cell boundary  $x = x_{j+1/2}$  for  $t > 0$  will play an important role when finding an upper bound on the total variation of the numerical solution. One property of  $\mathbf{u}_{j+1/2}^{n,+}$  that will become useful is the following immediate consequence of Theorem 3.2:

$$\mathbf{u}_{j+1/2}^{n,+} \in \mathcal{M}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n), \quad (44)$$

and another is part of Lemma 5.1 below. We have in Figure 7 drawn some of the state vectors related to the cell  $\mathcal{C}_j$ .

**Lemma 5.1.** *Consider the averaging step (38)–(39). There holds*

$$\mathbf{u}_{j,\pm}^{n+1} \in \mathcal{M}(\mathbf{u}_{j\pm 1/2}^{n,+}, \mathbf{u}_j^n) \subseteq \mathcal{M}(\mathbf{u}_{j\pm 1}^n, \mathbf{u}_j^n). \quad (45)$$

*Proof.* The inclusion in (45) follows from (44). Turning to the other part of the claim, namely  $\mathbf{u}_{j,\pm}^{n+1} \in \mathcal{M}(\mathbf{u}_{j\pm 1/2}^{n,+}, \mathbf{u}_j^n)$ , we see from the updating formulas (38)–(39) that  $k_{j,\pm}^{n+1} = k_{j\pm 1/2}^{n,+} \in \text{ch}(k_{j\pm 1/2}^{n,+}, k_j^n)$ . It remains to prove  $V_{j,\pm}^{n+1} \in \text{ch}(V_{j\pm 1/2}^{n,+}, V_j^n)$ . To this end, consider first the left state  $\mathbf{u}_{j,-}^{n+1}$ . If  $V_j^n = 0$ , then  $V_{j,-}^{n+1} = V_{j-1}^n = V_{j-1/2}^{n,+}$ . On the other hand, if  $V_j^n > 0$ , then by (36)  $\mathbf{u}_{j,-}^{n+1}$  is an average of states on the convex set  $\{\mathbf{u} \in \Omega : k = k_{j-1}^n, \phi \in \text{ch}(\phi_{j-1/2}^{n,+}, \phi^\star(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n))\}$ . We have thus shown that  $V_{j,-}^{n+1} \in \text{ch}(V_{j-1/2}^{n,+}, V(\mathbf{u}^\star(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)))$  and now claim that this set is contained in  $\text{ch}(V_{j-1/2}^{n,+}, V_j^n)$ . To see this, recall the definition (21) of  $\mathbf{u}^\star$ . If  $\phi^\star(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) > 0$ , then  $V(\mathbf{u}^\star(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)) = V_j^n$  while  $\phi^\star(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) = 0$  only if  $V(\mathbf{u}^\star(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n)) = V(0, k_{j-1}^n) < V_j^n$ .

Analogous arguments can be used for the right state  $\mathbf{u}_{j,+}^{n+1}$ , but with the averaging made over states along  $k = k_j^n$  between  $\phi_j^n$  and  $\phi_{j+1/2}^{n,+}$ . Indeed, we need not involve an intermediate state in this case.  $\square$

**Corollary 5.1.** *For a fixed mesh width  $\Delta > 0$ , let*

$$\mathcal{M}_\Delta^0 := \bigcup_{i,j \in \mathbb{Z}} \mathcal{M}(\mathbf{u}_j^0, \mathbf{u}_i^0).$$

*If the remap satisfies (40), then  $\mathcal{M}_\Delta^0$  is an invariant region under the resulting numerical scheme, i.e.  $\mathbf{u}_j^n \in \mathcal{M}_\Delta^0$  for all  $j \in \mathbb{Z}$  and  $n \geq 0$ .*

*Proof.* Suppose that, for a fixed  $n \geq 0$ ,  $\mathbf{u}_j^n \in \mathcal{M}_\Delta^0$  for all  $j$ . Then  $\mathcal{M}(\mathbf{u}_j^n, \mathbf{u}_i^n) \subseteq \mathcal{M}_\Delta^0$  for every pair  $i, j$  and the restriction (40) on the remap together with Lemma 5.1 yields

$$\mathbf{u}_j^{n+1} \in \mathcal{M}(\mathbf{u}_{j,-}^{n+1}, \mathbf{u}_{j,+}^{n+1}) \subseteq \mathcal{M}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) \cup \mathcal{M}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n) \cup \mathcal{M}(\mathbf{u}_{j-1}^n, \mathbf{u}_{j+1}^n) \subseteq \mathcal{M}_\Delta^0.$$

The assertion follows from induction over  $n$  since  $\mathbf{u}_j^0 \in \mathcal{M}_\Delta^0$  for all  $j$ .  $\square$

**Lemma 5.2.** *Let  $\Delta > 0$  and suppose that  $\mathbf{R}$  is Lipschitz continuous on  $\mathcal{M}_\Delta^0$ . If  $\text{T.V.}(\mathbf{u}_0) < \infty$  and the remap satisfies (40), then for all  $0 \leq n \leq N - 1$*

$$\sup_{j \in \mathbb{Z}} |\mathbf{u}_j^n| < \infty, \quad (46)$$

$$\sum_{j \in \mathbb{Z}} |\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j-1}^n)| \leq \|\mathbf{R}\|_{\text{Lip}} \text{T.V.}(\mathbf{u}_0), \quad (47)$$

$$\sum_{j \in \mathbb{Z}} |\mathbf{R}(\mathbf{u}_j^{n+1}) - \mathbf{R}(\mathbf{u}_j^n)| \leq 2\|\mathbf{R}\|_{\text{Lip}} \text{T.V.}(\mathbf{u}_0). \quad (48)$$

*Proof.* The first bound (46) follows from the invariance of  $\mathcal{M}_\Delta^0$  stated in Corollary 5.1.

Assume that  $n \geq 1$ . The restriction (40) combined with (45) yields

$$\mathbf{u}_j^n \in \mathcal{M}(\mathbf{u}_{j,-}^n, \mathbf{u}_{j,+}^n) \subseteq \mathcal{M}(\mathbf{u}_{j-1/2}^{n-1,+}, \mathbf{u}_j^{n-1}) \cup \mathcal{M}(\mathbf{u}_{j+1/2}^{n-1,+}, \mathbf{u}_j^{n-1}) \cup \mathcal{M}(\mathbf{u}_{j-1/2}^{n-1,+}, \mathbf{u}_{j+1/2}^{n-1,+}), \quad (49)$$

which is, together with (44), the key ingredient of the proofs of Lemmas 3.1 and 3.2 in [28]. These Lemmas state the following TVD property in the  $(V, k)$ -plane:

$$\sum_{j \in \mathbb{Z}} |\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j-1}^n)| \leq \sum_{j \in \mathbb{Z}} |\mathbf{R}(\mathbf{u}_j^{n-1}) - \mathbf{R}(\mathbf{u}_{j-1}^{n-1})|. \quad (50)$$

Induction over  $n$ , the invariance of  $\mathcal{M}_\Delta^0$  and the Lipschitz continuity of  $\mathbf{R}$  yield (47). Due to the notational differences with [28], we review here the proof of (50) and consider first the Riemann invariant  $k$ . From (49) we derive

$$|k_{j-1/2}^{n-1,+} - k_j^n| + |k_j^n - k_{j+1/2}^{n-1,+}| \leq |k_{j-1/2}^{n-1,+} - k_j^{n-1}| + |k_j^{n-1} - k_{j+1/2}^{n-1,+}|,$$

while (44) gives

$$|k_j^{n-1} - k_{j-1/2}^{n-1,+}| + |k_{j-1/2}^{n-1,+} - k_{j-1}^{n-1}| = |k_j^{n-1} - k_{j-1}^{n-1}|.$$

Since the corresponding bounds hold also for the second Riemann invariant,  $V$ , we get

$$\begin{aligned} & |\mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+}) - \mathbf{R}(\mathbf{u}_j^n)| + |\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j+1/2}^{n-1,+})| \\ & \leq |\mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+}) - \mathbf{R}(\mathbf{u}_j^{n-1})| + |\mathbf{R}(\mathbf{u}_j^{n-1}) - \mathbf{R}(\mathbf{u}_{j+1/2}^{n-1,+})| \end{aligned}$$

and

$$|\mathbf{R}(\mathbf{u}_j^{n-1}) - \mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+})| + |\mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+}) - \mathbf{R}(\mathbf{u}_{j-1}^{n-1})| = |\mathbf{R}(\mathbf{u}_j^{n-1}) - \mathbf{R}(\mathbf{u}_{j-1}^{n-1})|.$$

The proof of (50) in [28] is then completed by

$$\begin{aligned}
\sum_{j \in \mathbb{Z}} |\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j-1}^n)| &\leq \sum_{j \in \mathbb{Z}} (|\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+})| + |\mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+}) - \mathbf{R}(\mathbf{u}_{j-1}^n)|) \\
&= \sum_{j \in \mathbb{Z}} (|\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+})| + |\mathbf{R}(\mathbf{u}_{j+1/2}^{n-1,+}) - \mathbf{R}(\mathbf{u}_j^n)|) \\
&\leq \sum_{j \in \mathbb{Z}} (|\mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+}) - \mathbf{R}(\mathbf{u}_j^{n-1})| + |\mathbf{R}(\mathbf{u}_j^{n-1}) - \mathbf{R}(\mathbf{u}_{j+1/2}^{n-1,+})|) \\
&= \sum_{j \in \mathbb{Z}} (|\mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+}) - \mathbf{R}(\mathbf{u}_j^{n-1})| + |\mathbf{R}(\mathbf{u}_{j-1}^{n-1}) - \mathbf{R}(\mathbf{u}_{j-1/2}^{n-1,+})|) \\
&= \sum_{j \in \mathbb{Z}} |\mathbf{R}(\mathbf{u}_j^{n-1}) - \mathbf{R}(\mathbf{u}_{j-1}^{n-1})|.
\end{aligned}$$

To show (48), we use that

$$\mathbf{u}_j^n, \mathbf{u}_j^{n+1} \in \mathcal{M}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) \cup \mathcal{M}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n) \cup \mathcal{M}(\mathbf{u}_{j-1}^n, \mathbf{u}_{j+1}^n)$$

(see the proof of Corollary 5.1 for  $\mathbf{u}_j^{n+1}$ ) and derive

$$\begin{aligned}
|k_j^{n+1} - k_j^n| &\leq \max(|k_{j-1}^n - k_j^n|, |k_j^n - k_{j+1}^n|, |k_{j-1}^n - k_{j+1}^n|) \\
&\leq |k_{j-1}^n - k_j^n| + |k_j^n - k_{j+1}^n|.
\end{aligned}$$

Adding the corresponding bound for the  $V$ -components, we obtain

$$|\mathbf{R}(\mathbf{u}_j^{n+1}) - \mathbf{R}(\mathbf{u}_j^n)| \leq |\mathbf{R}(\mathbf{u}_{j-1}^n) - \mathbf{R}(\mathbf{u}_j^n)| + |\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j+1}^n)|,$$

and (48) follows from (47) after summation over  $j$ .  $\square$

**Theorem 5.1.** *Suppose that  $\mathbf{u}_0 \in L^1(\mathbb{R})$ ,  $\text{T.V.}(\mathbf{u}_0) < \infty$  and that  $\mathbf{R}$  and  $\mathbf{R}^{-1}$  are Lipschitz continuous on  $\mathcal{M}^0 := \cup_{\Delta>0} \mathcal{M}_\Delta^0$  and  $\mathbf{R}(\mathcal{M}^0)$ , respectively. If the remap satisfies (40), then there exists a subsequence of  $\{\mathbf{u}_\Delta\}$  that converges in  $L^1_{\text{loc}}(\mathbb{R} \times [0, T])$  to a limit function  $\mathbf{u}$ .*

*Proof.* We give the proof via the convergence along a subsequence for the piecewise constant function  $\hat{\mathbf{u}}_\Delta$  defined in (43), and then show that  $\|\mathbf{u}_\Delta - \hat{\mathbf{u}}_\Delta\|_{L^1} \rightarrow 0$  as  $\Delta \rightarrow 0$ .

The bound (46) immediately gives  $\|\hat{\mathbf{u}}_\Delta\|_\infty < \infty$  while (47) and the Lipschitz continuity of  $\mathbf{R}^{-1}$  yield

$$\text{T.V.}(\hat{\mathbf{u}}_\Delta(\cdot, t_n)) = \sum_{j \in \mathbb{Z}} |\mathbf{u}_j^n - \mathbf{u}_{j-1}^n| \leq \|\mathbf{R}^{-1}\|_{\text{Lip}} \sum_{j \in \mathbb{Z}} |\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j-1}^n)| < \infty.$$

For a general  $t \in [0, T]$ , we let  $n$  be the largest integer such that  $t \geq t_n$  and use the identity  $\hat{\mathbf{u}}_\Delta(\cdot, t) = \hat{\mathbf{u}}_\Delta(\cdot, t_n)$ . Using the Lipschitz continuity of  $\mathbf{R}^{-1}$  once more, but this time combined with (48), we can derive

$$\|\hat{\mathbf{u}}_\Delta(\cdot, t_1) - \hat{\mathbf{u}}_\Delta(\cdot, t_2)\|_{L^1} \leq C_1 |t_1 - t_2| + \mathcal{O}(\Delta) \text{ for all } t_1, t_2 \in [0, T],$$

see e.g. the proof of Theorem 3.8 in [20]. Here, the constant  $C_1$  is independent of  $\Delta$ ,  $t_1$  and  $t_2$ . Standard compactness arguments now give the existence of a subsequence of  $\{\hat{\mathbf{u}}_\Delta\}$  that converges in  $L^1_{\text{loc}}(\mathbb{R} \times [0, T])$ .

To see that the difference between the two functions  $\mathbf{u}_\Delta$  and  $\hat{\mathbf{u}}_\Delta$  vanishes in  $L^1$  as  $\Delta \rightarrow 0$ , fix  $t \in [0, T)$  and again let  $n$  be the largest integer such that  $t \geq t_n$ . On each cell  $\mathcal{C}_j$ , we have  $\hat{\mathbf{u}}_\Delta(x, t) = \mathbf{u}_j^n$  and

$$\mathbf{u}_\Delta(x, t) \in \mathcal{M}(\mathbf{u}_{j-1}^n, \mathbf{u}_j^n) \cup \mathcal{M}(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n)$$

(according to Theorem 3.2). This implies

$$|\mathbf{R}(\mathbf{u}_\Delta(x, t)) - \mathbf{R}(\hat{\mathbf{u}}_\Delta(x, t))| \leq |\mathbf{R}(\mathbf{u}_{j+1}^n) - \mathbf{R}(\mathbf{u}_j^n)| + |\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j-1}^n)|$$

for all  $x \in \mathcal{C}_j$ . Thus, by (47):

$$\begin{aligned} \|\mathbf{u}_\Delta(\cdot, t) - \hat{\mathbf{u}}_\Delta(\cdot, t)\|_{L^1} &\leq \|\mathbf{R}^{-1}\|_{\text{Lip}} \sum_{j \in \mathbb{Z}} \int_{\mathcal{C}_j} |\mathbf{R}(\mathbf{u}_\Delta(x, t)) - \mathbf{R}(\hat{\mathbf{u}}_\Delta(x, t))| \, dx \\ &\leq \|\mathbf{R}^{-1}\|_{\text{Lip}} \Delta x \sum_{j \in \mathbb{Z}} (|\mathbf{R}(\mathbf{u}_{j+1}^n) - \mathbf{R}(\mathbf{u}_j^n)| + |\mathbf{R}(\mathbf{u}_j^n) - \mathbf{R}(\mathbf{u}_{j-1}^n)|) \leq C_2 \Delta x, \end{aligned}$$

where  $C_2 = 2\|\mathbf{R}^{-1}\|_{\text{Lip}}\|\mathbf{R}\|_{\text{Lip}}\text{T.V.}(\mathbf{u}_0)$ . This bound is uniform on  $[0, T)$  so

$$\|\mathbf{u}_\Delta - \hat{\mathbf{u}}_\Delta\|_{L^1} = \int_0^T \|\mathbf{u}_\Delta(\cdot, t) - \hat{\mathbf{u}}_\Delta(\cdot, t)\|_{L^1} \, dt \leq T C_2 \Delta x.$$

□

**5.2. Weak solution.** To show that the limit function in Theorem 5.1 is in fact a weak solution of (10), we will soon restrict the discussion to the Av-RS scheme given in Definition 4.1. However, let us dwell at the more general scheme (38)–(40) in the next Lemma, which states an  $L^1$ -bound on the jump in  $\mathbf{u}_\Delta$  across  $t = t_n$ . This is where we have a transition from the evolved exact solution  $\mathbf{u}_\Delta^{n-1}$  in one strip  $\mathbb{R} \times [t_{n-1}, t_n]$  to initial data for  $\mathbf{u}_\Delta^n$  in the next strip  $\mathbb{R} \times [t_n, t_{n+1}]$ . The handling of these transitions is the most delicate part of the proof of the main result stated in Theorem 5.2.

**Lemma 5.3.** *Let  $\Delta > 0$  and suppose that  $\mathbf{R}$  and  $\mathbf{R}^{-1}$  are Lipschitz continuous on  $\mathcal{M}_\Delta^0$  and  $\mathbf{R}(\mathcal{M}_\Delta^0)$ , respectively. If  $\mathbf{u}_0 \in L^1(\mathbb{R})$ ,  $\text{T.V.}(\mathbf{u}_0) < \infty$  and the remap satisfies (40), then*

$$\|\mathbf{u}_\Delta(\cdot, t_n + 0) - \mathbf{u}_\Delta(\cdot, t_n - 0)\|_{L^1} \leq 2\|\mathbf{R}^{-1}\|_{\text{Lip}}\|\mathbf{R}\|_{\text{Lip}}\text{T.V.}(\mathbf{u}_0)\Delta x$$

for all integers  $1 \leq n \leq N - 1$ .

*Proof.* Consider first a single cell  $\mathcal{C}_j$  and recall that  $\mathbf{u}_\Delta(x, t_n + 0) = \mathbf{u}_j^n$  on  $\mathcal{C}_j$ . By (40) and (45) we thus get

$$\mathbf{u}_\Delta(x, t_n + 0) \in \mathcal{M}(\mathbf{u}_{j-1}^{n-1}, \mathbf{u}_j^{n-1}) \cup \mathcal{M}(\mathbf{u}_j^{n-1}, \mathbf{u}_{j+1}^{n-1}) \cup \mathcal{M}(\mathbf{u}_{j-1}^{n-1}, \mathbf{u}_{j+1}^{n-1}),$$

while Theorem 3.2 yields

$$\mathbf{u}_\Delta(x, t_n - 0) \in \mathcal{M}(\mathbf{u}_{j-1}^{n-1}, \mathbf{u}_j^{n-1}) \cup \mathcal{M}(\mathbf{u}_j^{n-1}, \mathbf{u}_{j+1}^{n-1}).$$

Using analogous arguments as in the proof of (48), we obtain

$$|\mathbf{R}(\mathbf{u}_\Delta(x, t_n + 0)) - \mathbf{R}(\mathbf{u}_\Delta(x, t_n - 0))| \leq |\mathbf{R}(\mathbf{u}_{j-1}^{n-1}) - \mathbf{R}(\mathbf{u}_j^{n-1})| + |\mathbf{R}(\mathbf{u}_j^{n-1}) - \mathbf{R}(\mathbf{u}_{j+1}^{n-1})|,$$

and the claim follows from the Lipschitz continuity of  $\mathbf{R}^{-1}$  and the bound (47):

$$\begin{aligned} & \|\mathbf{u}_\Delta(\cdot, t_n + 0) - \mathbf{u}_\Delta(\cdot, t_n - 0)\|_{L^1} \\ &= \sum_{j \in \mathbb{Z}} \int_{C_j} |(\mathbf{u}_\Delta(x, t_n + 0) - \mathbf{u}_\Delta(x, t_n - 0))| dx \\ &\leq \|\mathbf{R}^{-1}\|_{\text{Lip}} \sum_{j \in \mathbb{Z}} \int_{C_j} |\mathbf{R}(\mathbf{u}_\Delta(x, t_n + 0)) - \mathbf{R}(\mathbf{u}_\Delta(x, t_n - 0))| dx \\ &\leq 2\|\mathbf{R}^{-1}\|_{\text{Lip}} \|\mathbf{R}\|_{\text{Lip}} \text{T.V.}(\mathbf{u}_0) \Delta x. \end{aligned}$$

□

We now confine our attention to the Av-RS scheme (38)–(39), (41) and adjust the arguments made in [38, Chapter 19, Section C], where convergence of the Glimm scheme was proved. We emphasize the dependence on the random sequence  $\mathbf{a} \in \mathcal{A} := \prod_{n=1}^\infty (0, 1)$  by the subscript  $\mathbf{u}_{\Delta, \mathbf{a}}$ . However, at some instances we leave this dependence in the discrete variables implicitly understood to avoid to get entangled in indices. Analogously with [38], we introduce

$$\begin{aligned} \mathbf{J}_n(\mathbf{a}, \Delta, \varphi) &:= \int_{\mathbb{R}} \varphi(x, t_n) (\mathbf{y}_{\Delta, \mathbf{a}}(x, t_n + 0) - \mathbf{y}_{\Delta, \mathbf{a}}(x, t_n - 0)) dx, \\ \mathbf{J}(\mathbf{a}, \Delta, \varphi) &:= \sum_{n=1}^{N-1} \mathbf{J}_n(\mathbf{a}, \Delta, \varphi) \end{aligned}$$

for bounded functions  $\varphi$ , continuous in time and with compact support on  $\mathbb{R} \times [0, T]$ . Here, the conserved counterpart of  $\mathbf{u}_{\Delta, \mathbf{a}}$  is naturally defined as  $\mathbf{y}_{\Delta, \mathbf{a}} := \mathbf{Y}(\mathbf{u}_{\Delta, \mathbf{a}})$ .

**Corollary 5.2.** *Let  $\mathbf{a} \in \mathcal{A}$ ,  $\Delta > 0$  and suppose that  $\varphi$  is a bounded function, continuous in time and with compact support on  $\mathbb{R} \times [0, T]$ . If  $\mathbf{u}_0 \in L^1(\mathbb{R})$ ,  $\text{T.V.}(\mathbf{u}_0) < \infty$  and  $\mathbf{R}$  and  $\mathbf{R}^{-1}$  are Lipschitz continuous on  $\mathcal{M}_\Delta^0$  and  $\mathbf{R}(\mathcal{M}_\Delta^0)$ , respectively, then there exist two constants  $C_1$  and  $C_2$  independent of  $\mathbf{a}, \Delta$  and  $\varphi$  such that*

$$|\mathbf{J}_n(\mathbf{a}, \Delta, \varphi)| \leq C_1 \|\varphi\|_\infty \Delta x \quad \text{and} \quad |\mathbf{J}(\mathbf{a}, \Delta, \varphi)| \leq C_2 \|\varphi\|_\infty. \quad (51)$$

*Proof.* The first bound in (51) follows from Lemma 5.3 since  $\mathbf{Y}$  is Lipschitz continuous with  $\|\mathbf{Y}\|_{\text{Lip}} \leq 1$  on  $\Omega$ :

$$\begin{aligned} |\mathbf{J}_n(\mathbf{a}, \Delta, \varphi)| &\leq \|\varphi\|_\infty \|\mathbf{y}_{\Delta, \mathbf{a}}(\cdot, t_n + 0) - \mathbf{y}_{\Delta, \mathbf{a}}(\cdot, t_n - 0)\|_{L^1} \\ &\leq \|\varphi\|_\infty \|\mathbf{u}_{\Delta, \mathbf{a}}(\cdot, t_n + 0) - \mathbf{u}_{\Delta, \mathbf{a}}(\cdot, t_n - 0)\|_{L^1}, \end{aligned}$$

while the second bound can be derived from the first

$$|\mathbf{J}(\mathbf{a}, \Delta, \varphi)| \leq \sum_{n=1}^{N-1} |\mathbf{J}_n(\mathbf{a}, \Delta, \varphi)| \leq C_1 \|\varphi\|_\infty N \Delta x \leq C_1 \frac{T}{\alpha} \|\varphi\|_\infty.$$

□

The requirement on  $\mathbf{a}$  to be uniformly distributed in  $(0, 1)$  becomes relevant in the following lemma, in which we consider the space  $L^2(\mathcal{A})$  equipped with an unweighted inner product.

**Lemma 5.4.** Let  $\Delta > 0$  and suppose that the function  $\varphi$  is a bounded function, continuous in time, constant on each cell  $\mathcal{C}_j$  and with compact support on  $\mathbb{R} \times [0, T]$ . Then, for  $1 \leq n_1, n_2 \leq N - 1$ , the components of  $\mathbf{J}_{n_1}$  and  $\mathbf{J}_{n_2}$  are pairwise orthogonal in  $L^2(\mathcal{A})$ , i.e.

$$\int_{\mathcal{A}} \mathbf{J}_{n_1}(\mathbf{a}, \Delta, \varphi) \mathbf{J}_{n_2}(\mathbf{a}, \Delta, \varphi) d\mathbf{a} = \mathbf{0} \quad \text{if } n_1 \neq n_2, \quad (52)$$

where  $\mathbf{J}_{n_1} \mathbf{J}_{n_2}$  means the componentwise product.

*Proof.* Denote by  $\varphi_j^n$  the value of  $\varphi$  on  $\mathcal{C}_j$  at  $t = t_n$  and note that  $\mathbf{y}_{\Delta, \mathbf{a}}(\cdot, t)$  is independent of  $a_n$  if  $t < t_n$ . We first consider the simple integral

$$\begin{aligned} \int_{(0,1)} \mathbf{J}_n da_n &= \int_{(0,1)} \sum_{j \in \mathbb{Z}} \varphi_j^n \int_{\mathcal{C}_j} (\mathbf{y}_{\Delta, \mathbf{a}}(x, t_n + 0) - \mathbf{y}_{\Delta, \mathbf{a}}(x, t_n - 0)) dx da_n \\ &= \sum_{j \in \mathbb{Z}} \varphi_j^n \left( \int_{\mathcal{C}_j} \int_{(0,1)} \mathbf{y}_{\Delta, \mathbf{a}}(x, t_n + 0) da_n dx - \int_{\mathcal{C}_j} \mathbf{y}_{\Delta, \mathbf{a}}(x, t_n - 0) dx \right). \end{aligned}$$

Each term in this series vanishes because the expected value of the mass is preserved by the remap step according to the calculation:

$$\begin{aligned} \int_{\mathcal{C}_j} \int_{(0,1)} \mathbf{y}_{\Delta, \mathbf{a}}(x, t_n + 0) da_n dx &= \int_{\mathcal{C}_j} \left( \alpha V_j^{n-1} \mathbf{Y}(\mathbf{u}_{j,-}^n) + (1 - \alpha V_j^{n-1}) \mathbf{Y}(\mathbf{u}_{j,+}^n) \right) dx \\ &= \overline{\Delta x}_{j,-}^{n-1} \mathbf{Y}(\mathbf{u}_{j,-}^n) + \overline{\Delta x}_{j,+}^{n-1} \mathbf{Y}(\mathbf{u}_{j,+}^n) = \int_{\mathcal{C}_j} \mathbf{y}_{\Delta, \mathbf{a}}(x, t_n - 0) dx, \end{aligned} \quad (53)$$

and we obtain

$$\int_{(0,1)} \mathbf{J}_n da_n = \mathbf{0}. \quad (54)$$

In the last step of (53), we have used that  $k_{\Delta, \mathbf{a}}(\cdot, t_n - 0) \equiv k_{j,-}^n$  on  $(x_{j-1/2}, \bar{x}_{j-1/2}^{n-1})$  if  $V_j^n > 0$  to derive

$$k_{j,-}^n \phi_{j,-}^n = \frac{1}{\overline{\Delta x}_{j,-}^{n-1}} \int_{x_{j-1/2}}^{\bar{x}_{j-1/2}^{n-1}} k_{\Delta, \mathbf{a}}(x, t_n - 0) \phi_{\Delta, \mathbf{a}}(x, t_n - 0) dx,$$

which gives

$$\mathbf{Y}(\mathbf{u}_{j,-}^n) = \frac{1}{\overline{\Delta x}_{j,-}^{n-1}} \int_{x_{j-1/2}}^{\bar{x}_{j-1/2}^{n-1}} \mathbf{y}_{\Delta, \mathbf{a}}(x, t_n - 0) dx.$$

Analogous arguments on  $(\bar{x}_{j-1/2}^{n-1}, x_{j+1/2})$  yield the corresponding identity for  $\mathbf{Y}(\mathbf{u}_{j,+}^n)$ .

Now, turn to the inner product in (52) and assume that  $n_1 < n_2$ . The orthogonality then follows from (54) and the independence of  $J_{n_1}$  on  $a_{n_2}$ :

$$\begin{aligned} \int_{\mathcal{A}} \mathbf{J}_{n_1} \mathbf{J}_{n_2} d\mathbf{a} &= \int_{\mathcal{A} \setminus (0,1)} \left( \int_{(0,1)} \mathbf{J}_{n_1} \mathbf{J}_{n_2} da_{n_2} \right) \prod_{n \neq n_2} da_n \\ &= \int_{\mathcal{A} \setminus (0,1)} \mathbf{J}_{n_1} \left( \int_{(0,1)} \mathbf{J}_{n_2} da_{n_2} \right) \prod_{n \neq n_2} da_n = \mathbf{0}. \end{aligned}$$

□

The proof of the main result, Theorem 5.2 below, is a straightforward adaptation of the arguments in [38] but for completeness of the current presentation, we give here some of the details. Note that admissibility (see Definition 2.1) is not claimed for the weak solution in this theorem.

**Theorem 5.2.** *Suppose that  $\mathbf{u}_0 \in L^1(\mathbb{R})$ ,  $\text{T.V.}(\mathbf{u}_0) < \infty$  and that  $\mathbf{R}$  and  $\mathbf{R}^{-1}$  are Lipschitz continuous on  $\mathcal{M}^0$  and  $\mathbf{R}(\mathcal{M}^0)$ , respectively. Then there exists a sequence  $\Delta_i \rightarrow 0$  such that  $\mathbf{u}_{\mathbf{a}} := \lim_{i \rightarrow \infty} \mathbf{u}_{\Delta_i, \mathbf{a}}$  is a weak solution to (10)–(11) for almost every  $\mathbf{a} \in \mathcal{A}$ .*

*Proof.* Let  $\varphi$  be a smooth function with compact support in  $\mathbb{R} \times [0, T]$ . By construction,  $\mathbf{u}_{\Delta, \mathbf{a}}$  is a weak solution of (10) on each strip  $\mathbb{R} \times [t_n, t_{n+1}]$ . Hence,

$$\begin{aligned} & \iint_{\mathbb{R} \times [t_n, t_{n+1}]} (\mathbf{y}_{\Delta, \mathbf{a}} \varphi_t + \mathbf{f}(\mathbf{u}_{\Delta, \mathbf{a}}) \varphi_x) dx dt + \int_{\mathbb{R}} \varphi(x, t_n) \mathbf{y}_{\Delta, \mathbf{a}}(x, t_n + 0) dx \\ & - \int_{\mathbb{R}} \varphi(x, t_{n+1}) \mathbf{y}_{\Delta, \mathbf{a}}(x, t_{n+1} - 0) dx = \mathbf{0}. \end{aligned} \quad (55)$$

Summation of (55) over  $n = 0, \dots, N - 1$  yields

$$\iint_{\mathbb{R} \times [0, T]} (\mathbf{y}_{\Delta, \mathbf{a}} \varphi_t + \mathbf{f}(\mathbf{u}_{\Delta, \mathbf{a}}) \varphi_x) dx dt + \int_{\mathbb{R}} \varphi(x, 0) \mathbf{y}_{\Delta, \mathbf{a}}(x, 0) dx + \mathbf{J}(\mathbf{a}, \Delta, \varphi) = \mathbf{0}. \quad (56)$$

By the discretization of  $\mathbf{u}_0$  in (27) and the continuity of  $\mathbf{Y}$ , we obtain

$$\lim_{\Delta \searrow 0} \int_{\mathbb{R}} \varphi(x, 0) \mathbf{y}_{\Delta, \mathbf{a}}(x, 0) dx = \lim_{\Delta \searrow 0} \int_{\mathbb{R}} \varphi(x, 0) \mathbf{Y}(\mathbf{u}_{\Delta}^0(x, 0)) dx = \int_{\mathbb{R}} \varphi(x, 0) \mathbf{Y}(\mathbf{u}_0(x)) dx.$$

It remains to control the last term on the left hand side of (56), namely  $\mathbf{J}(\mathbf{a}, \Delta, \varphi)$ . With Corollary 5.2 and Lemma 5.4 at hand, the existence of a null set  $\mathcal{N}$  and a sequence  $\Delta_i$  such that  $\mathbf{J}(\mathbf{a}, \Delta_i, \varphi) \rightarrow \mathbf{0}$  for any  $\mathbf{a} \in \mathcal{A} \setminus \mathcal{N}$  follows from a direct translation of the proof of Theorem 19.14 in [38].  $\square$

## 6. NUMERICAL EXAMPLES

We study here the performance of Godunov's method, the Av-RS and Av-CC schemes on a selection of sample problems. The numerical solutions are presented by their associated  $\hat{\phi}_{\Delta}$ - and  $\tilde{k}_{\Delta}$ -profiles at fixed time instances  $t \in [0, T]$  with  $T$  sufficiently large. The initial data of Examples 1 to 5 are chosen such that for all three schemes, for all used values of  $\Delta x$  and at each time instance  $t_n$ , the local Riemann data satisfy  $(\mathbf{u}_j^n, \mathbf{u}_{j-1}^n) \in \mathcal{S}$  for all  $j$  and the type A solutions agree with those of type B. In Example 6, the schemes are defined in terms of local type B solutions.

The deviation of the numerical approximations from the exact solution is measured by

$$\text{err}_{L^1}^t(\Delta) := \|\mathbf{Y}(\hat{\mathbf{u}}_{\Delta}(\cdot, t)) - \mathbf{y}(\cdot, t)\|_{L^1(I)} / \|\mathbf{y}(\cdot, t)\|_{L^1(I)},$$

where  $\mathbf{Y}(\hat{\mathbf{u}}_{\Delta})$  is defined as  $\mathbf{y}_j^n$  on  $\mathcal{C}_j \times [t_n, t_{n+1}]$  for Godunov's method. In each sample problem, the bounded interval  $I \subset \mathbb{R}$  is chosen such that the numerical solutions are identical with the exact solution at the boundaries of  $I$  up to the time  $t$  and for all  $\Delta$  under consideration. In this way, the fluxes of the numerical solutions remain identical with the exact fluxes

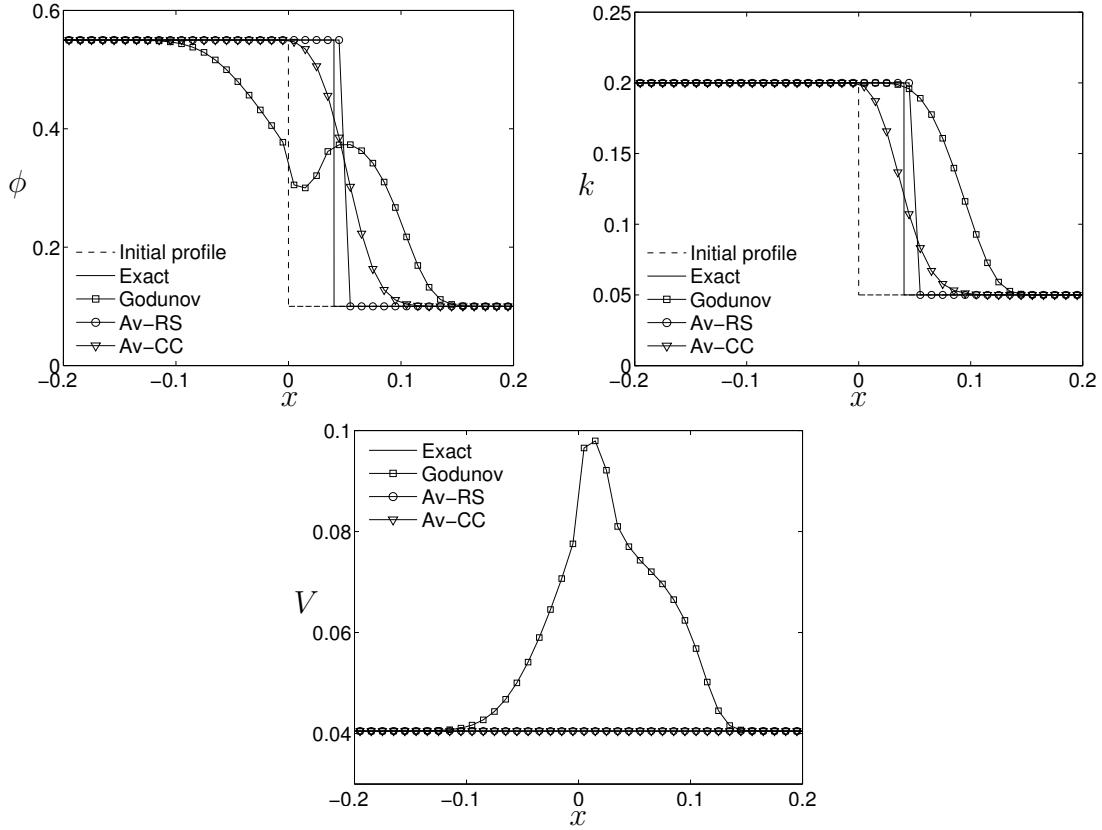


FIGURE 8. Example 1: Solutions at time  $t = 1$ . The exact solution contains an isolated contact discontinuity (cf. Figure 5) and the numerical solutions are computed with  $\Delta x = 0.01$ . The velocities  $V$  evaluated for the approximations computed with the Av-RS and Av-CC schemes are constant and agree with the exact velocity  $V = 0.0405$ .

at the boundaries of  $I$  throughout the computations, and we can study the conservativity of the Av-RS and Av-CC schemes simply by measuring the relative mass error

$$\text{err}_{\text{mass}}^t(\Delta) := \left| \int_I \mathbf{Y}(\hat{\mathbf{u}}_\Delta(x, t)) \, dx - \int_I \mathbf{y}(x, t) \, dx \right| / \left| \int_I \mathbf{y}(x, t) \, dx \right|.$$

**6.1. Example 1: An isolated contact discontinuity (sedimentation).** In Section 4.1 we addressed the inability of Godunov's method to adequately capture contact discontinuities. This shortcoming of the method was illustrated in Figure 5 where we considered the Riemann problem associated with the sedimentation model (6) and with the initial datum determined by the states  $\mathbf{u}_L = (0.55, 0.2)^T$  and  $\mathbf{u}_R = (0.1, 0.05)^T$ . Recall that such a setup gives  $V_L = V_R$  and the exact solution has an isolated contact discontinuity as its only wave. Godunov's method does not carry this simple structure over to the numerical solution, which exhibits a distinct oscillating behaviour. In Figure 8 it is seen that the numerical solutions

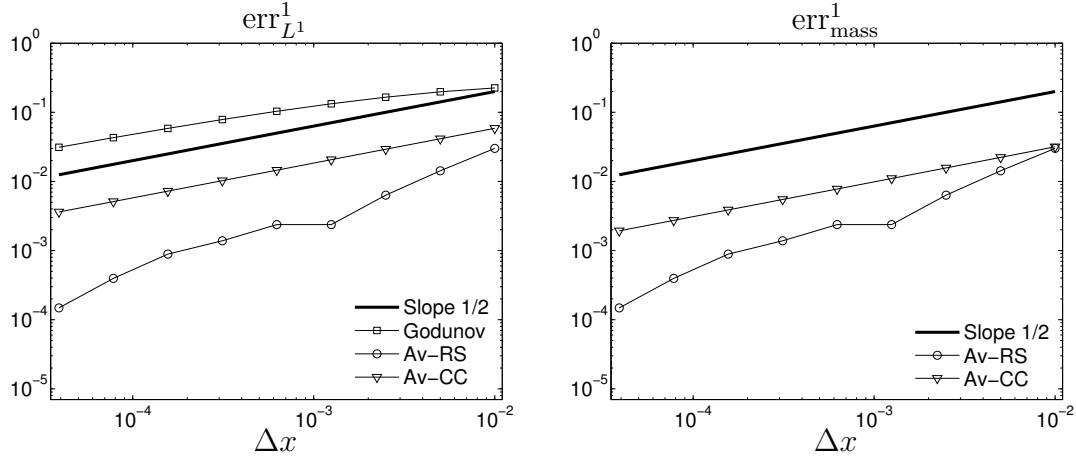


FIGURE 9. Example 1: The errors  $\text{err}_{L^1}^1$  (left) and  $\text{err}_{\text{mass}}^1$  (right) for the numerical solutions.

generated with the Av-RS and Av-CC schemes show more conformity with the exact solution. The anti-diffusive property of the Av-RS scheme is clearly demonstrated since the discontinuity is sharply resolved without any smearing.

Figure 9 (left) suggests convergence of all three schemes, the errors introduced by the Av-RS and Av-CC schemes being many times smaller than the error for Godunov's method. The Av-RS scheme is the most accurate one for this sample problem. The mass errors for the Av-RS and Av-CC schemes are shown in Figure 9 (right) and indicate that these schemes approximate to being conservative for decreasing  $\Delta$ .

**6.2. Example 2: A mixed 1-wave and a contact discontinuity (sedimentation).** A more complicated solution to the Riemann problem associated with (6) is obtained if  $\mathbf{u}_L = (0.8, 0.6)^T$  and  $\mathbf{u}_R = (0.1, 0.3)^T$ . The exact solution, which is shown in the phase planes in Figure 3 (top row), contains waves along both characteristic fields. Since  $f(\cdot, k_L)$  has an inflection point at  $\phi_{\text{infl}} = 2/3$  and  $\phi^*(\mathbf{u}_L, \mathbf{u}_R) \approx 0.36 < \phi_{\text{infl}} < \phi_L$ , the admissible 1-wave is composed of both a shock and a rarefaction. In Figure 10 we see that the three numerical schemes produce similar results near this wave, while the 2-discontinuity is approximated similarly to the isolated discontinuity in Example 1.

The errors seen in Figure 11 again suggest that all three schemes converge and that the non-conservative nature of the Av-RS and Av-CC schemes has small impact for fine meshes.

**6.3. Example 3: Formation of vacuum (traffic flow).** We now turn to the Riemann problem shown in Figure 3 (bottom row) associated with the ARZ traffic flow model (9), where  $\mathbf{u}_L = (0.5, 0.2)^T$  and  $\mathbf{u}_R = (0.5, 0.75)^T$ . This example illustrates a road with a fraction of fast vehicles in front of a slower fraction. Initially, the vehicles are uniformly distributed on the road but as the faster fraction leaves the slow vehicles behind, vacuum is formed. As is seen in Figure 12, the Av-RS scheme captures this behaviour well, while Godunov's method and the Av-CC scheme have obvious difficulties to reproduce the intermediate vacuum state. Observe that this sample problem is not captured in the convergence analysis in Section 5

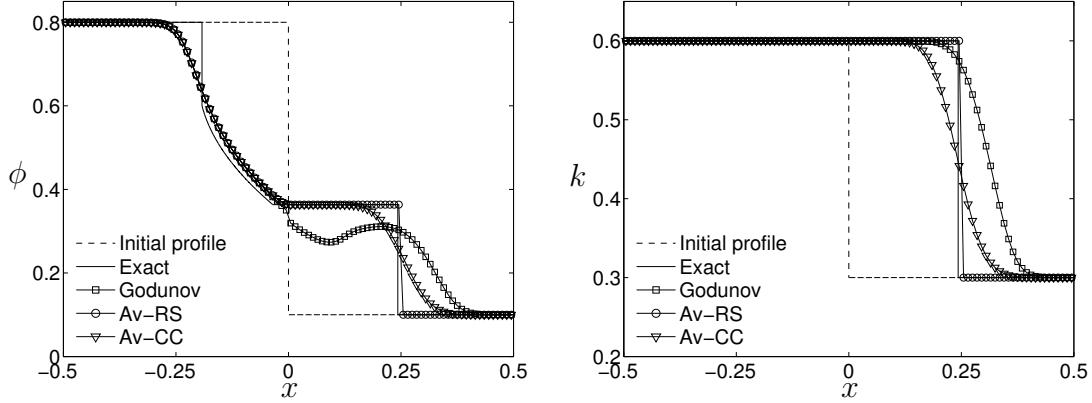


FIGURE 10. Example 2: Solutions at time  $t = 1$ . The exact solution contains a mixed shock- and rarefaction wave along the 1-field and a contact discontinuity along the 2-field. The phase plane representations of this solution are shown in Figure 3 (top row). The numerical solutions are computed with  $\Delta x = 0.01$ .

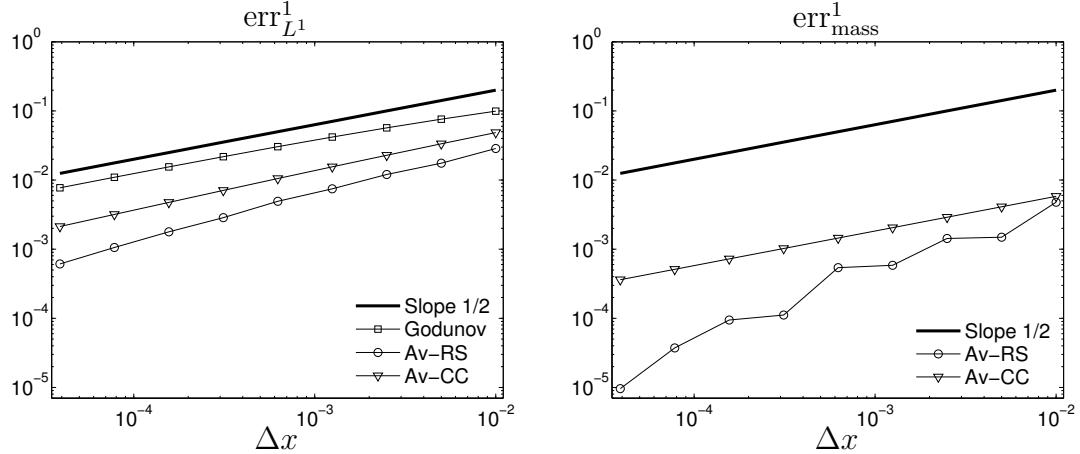


FIGURE 11. Example 2: The errors  $\text{err}_{L^1}^1$  (left) and  $\text{err}_{\text{mass}}^1$  (right) for the numerical solutions.

since  $\mathbf{R}^{-1}$  is not Lipschitz continuous on the entire  $\mathcal{M}^0$  (recall the discussion in the last paragraph of Section 1.3). Nevertheless, the errors plotted in Figure 13 suggest convergence.

**6.4. Example 4: Collision of a contact discontinuity with a 1-shock (traffic flow).** Thus far in this section, the numerical schemes have exclusively been applied on Riemann problems. The simplicity with non-interacting 1- and 2-waves has made the construction of exact solutions possible, and thereby enabled quantitative examinations of the errors. We now approach more general problems by considering the ARZ model (9) with initial datum

$$\mathbf{u}_0(x) = \begin{cases} \mathbf{u}_L & \text{if } x < -0.05, \\ \mathbf{u}_M & \text{if } -0.05 < x < 0.05, \\ \mathbf{u}_R & \text{if } x > 0.05, \end{cases}$$

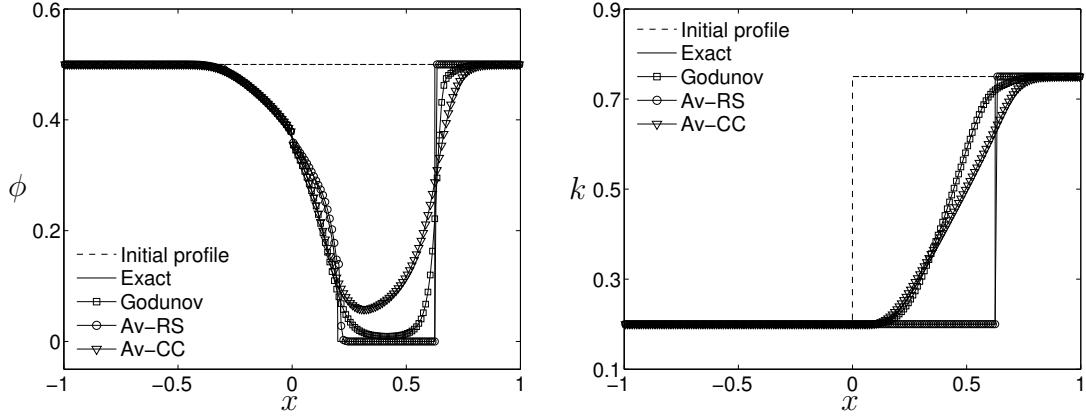


FIGURE 12. Example 3: Solutions at time  $t = 1$ . The exact solution has a 1-rarefaction separated by vacuum from a contact discontinuity along the 2-field. The phase plane representations of this solution are shown in Figure 3 (bottom row). The numerical solutions are computed with  $\Delta x = 0.01$ .

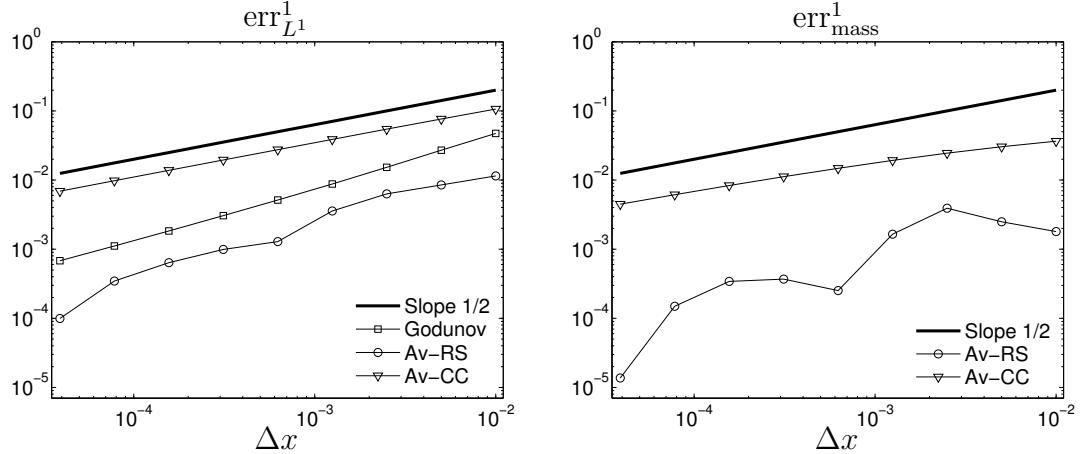


FIGURE 13. Example 3: The errors  $\text{err}_{L^1}^1$  (left) and  $\text{err}_{\text{mass}}^1$  (right) for the numerical solutions.

where  $\mathbf{u}_L = (\Phi(0.2, V(\mathbf{u}_M)), 0.2)^T$ ,  $\mathbf{u}_M = (0.2, 0.1)^T$  and  $\mathbf{u}_R = (0.4, 0.1)^T$ . The exact solution to this problem contains interacting waves, and yet it is elementary enough to be constructed by the results of Section 3. Due to the piecewise constant nature of  $\mathbf{u}_0$ , we can derive the initial part of the solution by solving two Riemann problems defined by the pairs  $(\mathbf{u}_L, \mathbf{u}_M)$  and  $(\mathbf{u}_M, \mathbf{u}_R)$  around  $x = -0.05$  and  $x = 0.05$ , respectively. Note that  $V_L = V_M$  and  $k_M = k_R$ , which implies that  $\mathbf{u}_L$  can be connected to  $\mathbf{u}_M$  via a contact discontinuity while  $\mathbf{u}_M$  can be connected to  $\mathbf{u}_R$  via an 1-wave. In fact,  $\phi_M$  and  $\phi_R$  are chosen such that this 1-wave, emanating from  $x = 0.05$ , is a shock with negative propagation speed  $s = (f(\mathbf{u}_R) - f(\mathbf{u}_M)) / (\phi_R - \phi_M) = -0.02$ . Simultaneously, the contact discontinuity emanating from  $x = -0.05$  travels with positive speed  $V_M \approx 0.09$  and there will be a collision between the two waves at time  $t_{\text{col}} = 0.1 / (V_M - s) \approx 0.9$ . The continuation of the solution after this time

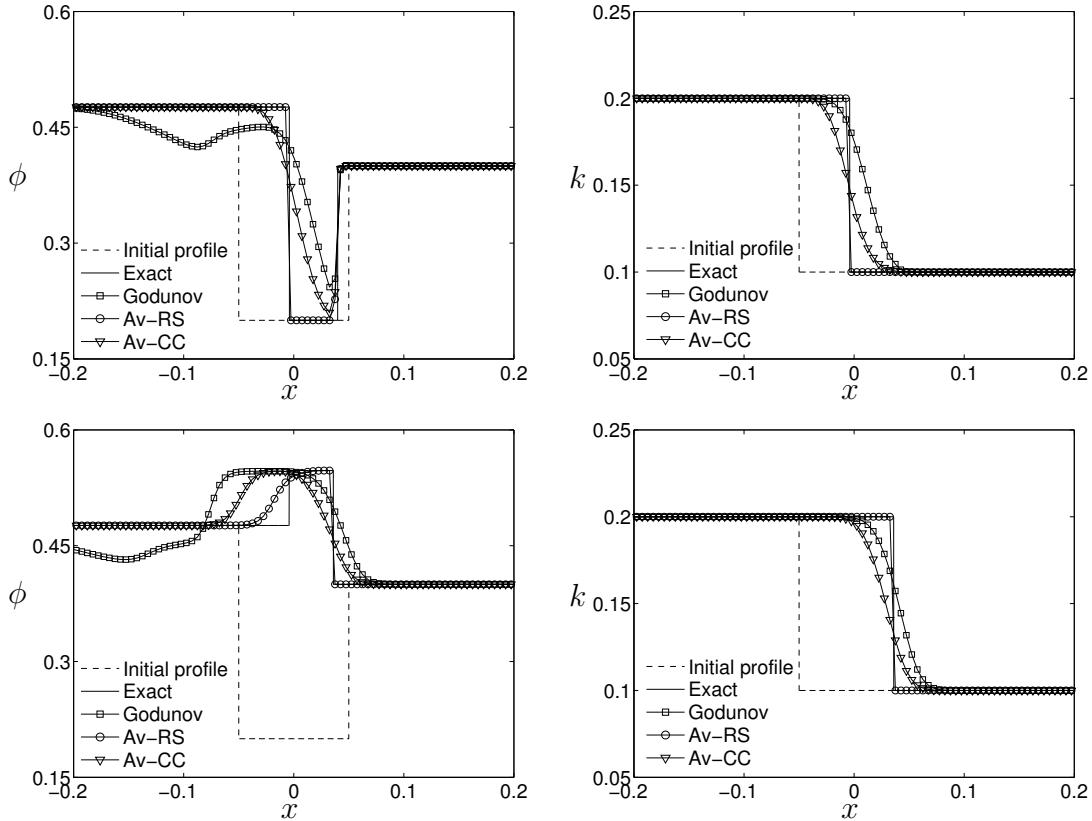


FIGURE 14. Example 4: Solutions at time  $t = 0.5$  (top row) and at  $t = 1$  (bottom row). The exact solution contains a left-moving 1-shock and a right-moving contact discontinuity colliding at  $t \approx 0.9$ . The numerical solutions are computed on the spatial interval  $[-0.45, 0.45]$  with  $\Delta x = 0.005$ .

can be constructed as the solution to the Riemann problem defined by the pair  $(\mathbf{u}_L, \mathbf{u}_R)$  around the location of the collision:  $x = -0.05 + V_M t_{\text{col}} \approx 0.03$ .

In Figure 14, we show the exact and numerical solutions before (top row) and after (bottom row) the collision in the exact solution. The numerical solutions have been computed on  $I = [-0.45, 0.45]$  to ensure constant states  $\mathbf{u}_L$  and  $\mathbf{u}_R$  at each respective boundary, whereas the profiles are plotted over a zoomed spatial interval. The errors  $\text{err}_{L1}^1$  and  $\text{err}_{\text{mass}}^1$  (after the collision) seen in Figure 15 are defined in terms of this larger interval.

**6.5. Example 5: Continuously decreasing initial data (sedimentation).** Let us return to the sedimentation model (6) and consider the continuous, piecewise affine initial datum decreasing from  $\mathbf{u}_L = (0.5, 0.1)^T$  to  $\mathbf{u}_R = (0.1, 0.01)^T$  according to:

$$\mathbf{u}_0(x) = \begin{cases} \mathbf{u}_L & \text{if } x < -0.05, \\ \mathbf{u}_L + 10(\mathbf{u}_R - \mathbf{u}_L)(x + 0.05) & \text{if } -0.05 < x < 0.05, \\ \mathbf{u}_R & \text{if } x > 0.05. \end{cases}$$

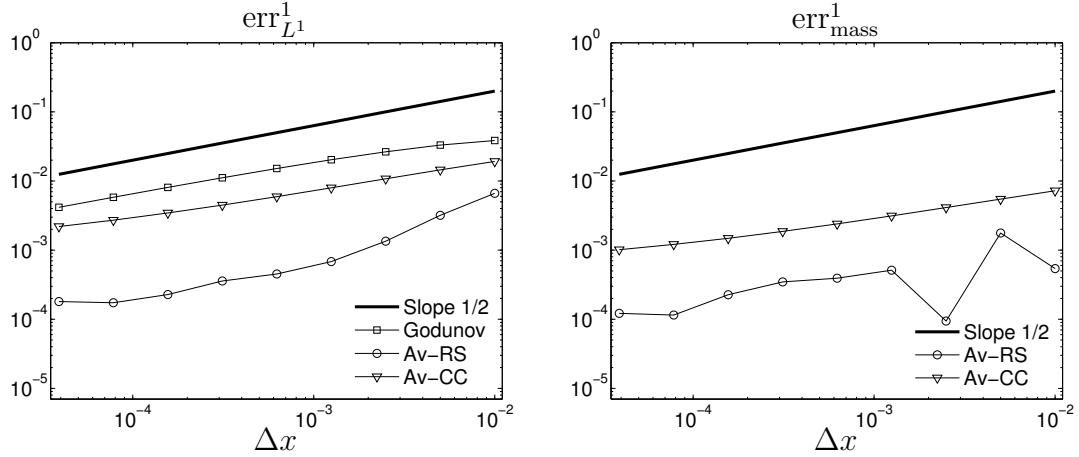


FIGURE 15. Example 4: The errors  $\text{err}_{L_1}^1$  (left) and  $\text{err}_{\text{mass}}^1$  (right) for the numerical solutions.

The theory in Section 3 does not cover this problem and in the absence of an exact admissible solution, we compute a reference solution on a fine mesh using the Av-RS scheme with  $\Delta x = 0.01 \cdot 2^{-12}$ . This reference solution is plotted in Figure 16 together with the numerical approximations of all three schemes for  $\Delta x = 0.0025$ . It is seen that the Av-RS scheme introduces spurious oscillations in the  $\phi$ -component and appears to be inferior to the two other schemes for this problem. A possible cause for these oscillations is found if we consider the local Riemann solutions associated with the cell interfaces, and which form the function  $\mathbf{u}_\Delta^n$ . The piecewise constant approximation  $k_\Delta^n(\cdot, 0)$  of a continuously varying  $k$ -component in the exact solution at  $t = t_n$  leads to contact discontinuities in several neighbouring local Riemann solutions. Each such discontinuity involves an intermediate state  $\mathbf{u}^*(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n)$  whose  $\phi$ -component need not be bounded by the local data  $\phi_j^n$  and  $\phi_{j+1}^n$  (note that  $k^*(\mathbf{u}_j^n, \mathbf{u}_{j+1}^n) = k_j^n$  and there are no oscillations in the  $k$ -component). The diffusive nature of Godunov's method and the Av-CC scheme dampens the impact of these “overshoots”, an effect which is not obtained with the anti-diffusive RS remap.

**6.6. Example 6: The unstable vacuum case (traffic flow).** In the last example, we once again consider a Riemann problem for the ARZ model (9) but now choose the pair  $(\mathbf{u}_L, \mathbf{u}_R)$  outside of  $\mathcal{S}$  by letting  $\mathbf{u}_L = (0.5, 0.75)^T$  and  $\mathbf{u}_R = (0, 0.2)^T$ . Since  $(\mathbf{u}_L, \mathbf{u}_R) \notin \mathcal{S}$ , Theorems 3.1 and 3.3 define different solutions, both shown in Figure 4. We are interested in the type B solution and therefore define the first step of the Av-RS and Av-CC schemes in terms of  $\mathbf{u}_\Delta^n$  constructed via the Riemann solution in Theorem 3.3 (recall from Section 4.1 that Godunov's method is also based on this construction).

As is seen in Figure 17, Godunov's method captures the behaviour of the type B solution but with a dislocation of the discontinuity in the  $k$ -component. This deviation from the exact solution is explained by the numerical solution being close to vacuum and is not as severe in the conserved variables. The Av-RS scheme approximates the solution well and produces satisfactory results in both components. The Av-CC scheme, on the other hand, fails to produce the type B solution and instead generates an approximation closer to the

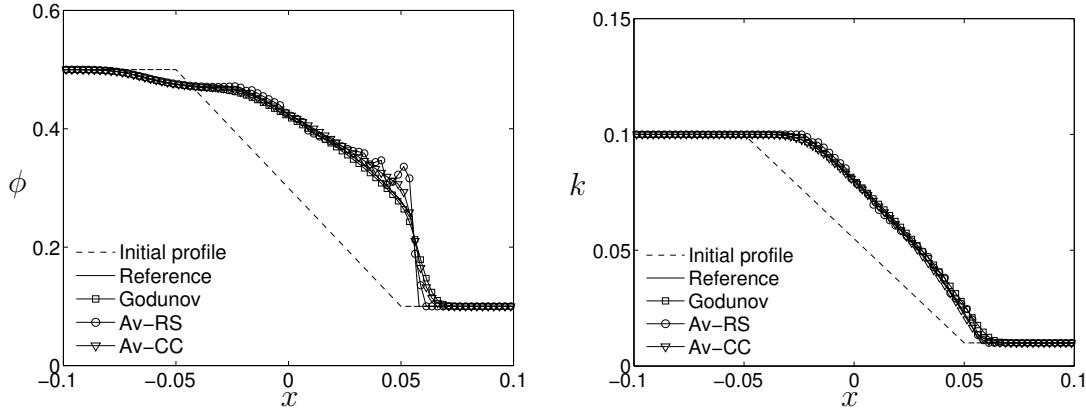


FIGURE 16. Example 5: Solutions at time  $t = 1$ . The reference solution is generated with the Av-RS scheme for  $\Delta x = 0.01 \cdot 2^{-12}$ , while the coarser approximations are computed with  $\Delta x = 0.0025$ .

type A solution in Theorem 3.1. For this reason, we have not included the Av-CC scheme in the convergence plot in Figure 17.

## 7. CONCLUDING REMARKS

It has been the prime purpose of this work to present and in part analyze a method that removes the spurious oscillations that arise when the Godunov scheme is applied to systems of the form (1). These oscillations are well known in the special case of the ARZ traffic model and the same difficulties arose when the authors attempted to apply that scheme to the sedimentation model. Herein, it is demonstrated that a random-sampling method provides a relatively simple working numerical scheme, the Av-RS scheme, that is supported by a convergence analysis. We prove convergence to a weak solution and leave the admissibility of this solution (i.e. that every discontinuity satisfies an entropy condition) as an open problem. Numerical solutions illustrate that different types of waves and their interactions are handled acceptably by the Av-RS scheme, at least in a better way than Godunov's method. In fact, while a frequent objection against random sampling methods is their lack of conservation, it turns out that the numerical errors produced by the statistically conservative scheme are actually consistently smaller than those of the conservative and deterministic Godunov scheme for the sample problems investigated. We remark that the Av-RS scheme may introduce oscillations in regions where the  $k$ -component of the solution is not piecewise constant but varies continuously with respect to the spatial variable (see Section 6, Example 5). These oscillations are, however, of a different nature with less regularity and smaller magnitude than those introduced by Godunov's method. Similar problems have also been reported for Glimm's method (see e.g. [2]).

The convergence proof provides the existence of a weak solution of the problem at hand, and explicitly includes the vacuum state, which is avoided in several previous works (see Section 1.4). No convexity condition is imposed on the scalar flux function  $f$ . The explicit

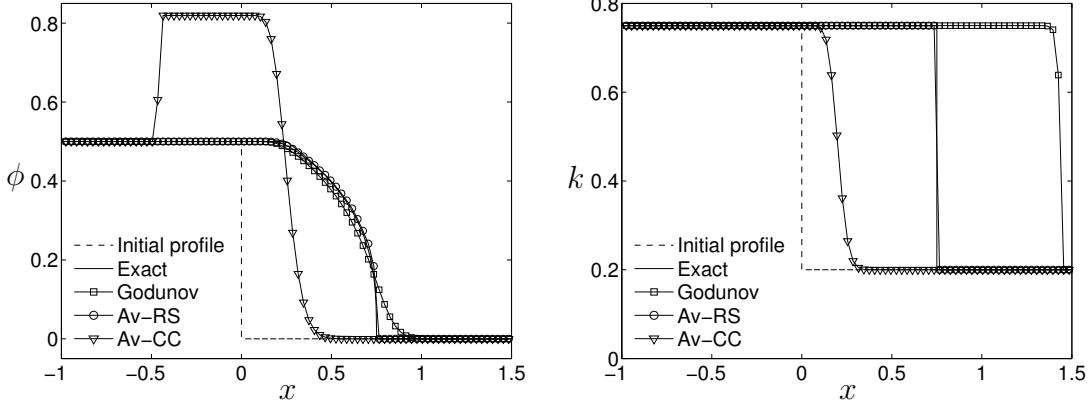


FIGURE 17. Example 6: Solutions at time  $t = 1$ . The exact, type B solution has a 1-rarefaction fronted by vacuum. The phase plane representation of this solution is shown in Figure 4. The numerical solutions are computed with  $\Delta x = 0.01$  but for clarity, they are plotted for a coarser resolution.

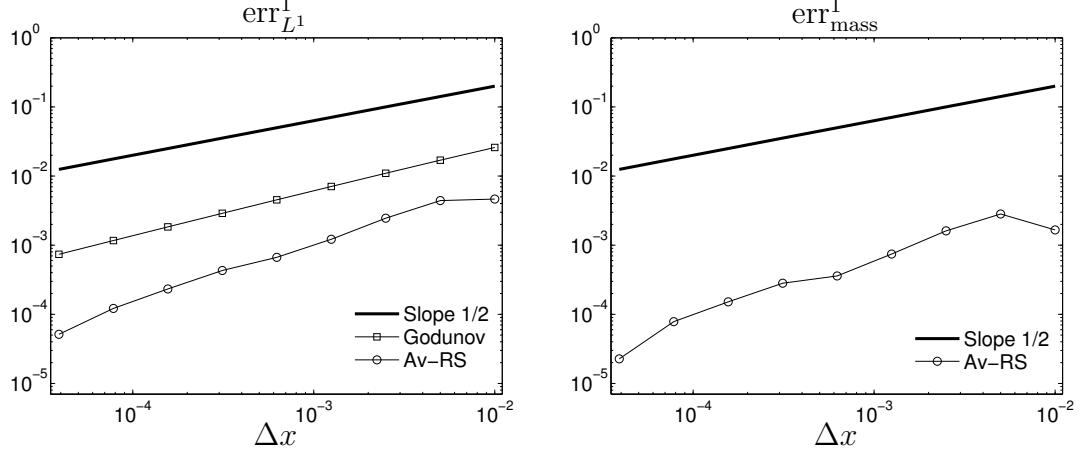


FIGURE 18. Example 6: The errors  $\text{err}_{L^1}^1$  (left) and  $\text{err}_{\text{mass}}^1$  (right) for the numerical solutions.

treatment of the vacuum state is captured in the distinction between the different types of concepts of Riemann solutions (types A and B).

Finally, we mention that the author's original interest in the system (1)–(2) is motivated by the applications discussed. While the new Av-RS scheme can readily be utilized for simulations of vehicular traffic in the framework of the ARZ model, possibly on a closed circuit with easily implemented periodic boundary conditions, the sedimentation model is based on considering  $x$  as a vertical coordinate and one must provide either zero-flux boundary conditions for batch settling, or extend the model to a clarifier-thickener setup with superimposed linear flux and a definition of  $V(\phi, k)$  that changes discontinuously with spatial position. The extension of the scheme to that setup is topic of current research.

## ACKNOWLEDGEMENTS

FB and RB acknowledge support by BASAL project CMM, Universidad de Chile and Centro de Investigación en Ingeniería Matemática (CIMNA), Universidad de Concepción; and Centro CRHIAM Proyecto Conicyt Fondecyt 15130015. In addition, FB is supported by Fondecyt project 11130397, and RB is supported by Fondecyt project 1130154; Conicyt project Anillo ACT1118 (ANANUM); and Red Doctoral REDOC.CTA, MINEDUC project UCO1202.

## REFERENCES

- [1] A. Aw and M. Rascle. Resurrection of “second order” models of traffic flow. *SIAM J. Appl. Math.*, 60(3):916–938, 2000.
- [2] M. Bachmann, P. Helluy, J. Jung, H. Mathis, and S. Müller. Random sampling remap for compressible two-phase flows. *Comput. & Fluids*, 86:275–283, 2013.
- [3] J. W. Banks. A note on the convergence of Godunov type methods for shock reflection problems. *Comput. Math. Appl.*, 66(1):19–23, 2013.
- [4] F. Betancourt, R. Bürger, S. Diehl, and S. Farås. Modeling and controlling clarifier-thickeners fed by suspensions with time-dependent properties. *Minerals Eng.*, 62:91–101, 2014.
- [5] A. Bressan, H. K. Jenssen, and P. Baiti. An instability of the Godunov scheme. *Comm. Pure Appl. Math.*, 59(11):1604–1638, 2006.
- [6] R. Bürger, C. Chalons, and L. M. Villada. Antidiffusive and random-sampling Lagrangian-remap schemes for the multiclass Lighthill–Whitham–Richards traffic model. *SIAM J. Sci. Comput.*, 35(6):B1341–B1368, 2013.
- [7] C. Chalons. Numerical approximation of a macroscopic model of pedestrian flows. *SIAM J. Sci. Comput.*, 29(2):539–555, 2007.
- [8] C. Chalons and F. Coquel. Computing material fronts with a Lagrange–Projection approach. In *Ser. Contemp. Appl. Math. CAM*, volume 1, pages 346–356. World Sci. Publishing, Singapore, 2012. (Proceedings of international Conference HYP2010).
- [9] C. Chalons and P. Goatin. Transport-equilibrium schemes for computing contact discontinuities in traffic flow modeling. *Commun. Math. Sci.*, 5(3):533–551, 2007.
- [10] C. Chalons and P. Goatin. Godunov scheme and sampling technique for computing phase transitions in traffic flow modeling. *Interfaces Free Bound.*, 10(2):197–221, 2008.
- [11] C. Chalons, P. Goatin, and N. Seguin. General constrained conservation laws. Application to pedestrian flow modeling. *Netw. Heterog. Media*, 8(2):433–463, 2013.
- [12] P. Colella. Glimm’s method for gas dynamics. *SIAM J. Sci. and Stat. Comput.*, 3(1):76–110, 1982.
- [13] R. M. Colombo. A  $2 \times 2$  hyperbolic traffic flow model. *Math. Comput. Model.*, 35(5–6):683–688, 2002.
- [14] S. Fan, M. Herty, and B. Seibold. Comparative model accuracy of a data-fitted generalized Aw–Rascle–Zhang model. *Netw. Heterog. Media*, 9(2):239–268, 2014.
- [15] M. Garavello and P. Goatin. The Aw–Rascle traffic model with locally constrained flow. *J. Math. Anal. Appl.*, 378(2):634–648, 2011.
- [16] J. Glimm. Solutions in the large for nonlinear hyperbolic systems of equations. *Comm. Pure Appl. Math.*, 18(4):697–715, 1965.
- [17] P. Goatin. The Aw–Rascle vehicular traffic flow model with phase transitions. *Math. Comput. Model.*, 44(3–4):287–303, 2006.
- [18] M. Godvik and H. Hanche-Olsen. Existence of solutions for the Aw–Rascle traffic flow model with vacuum. *J. Hyperbolic Differ. Equ.*, 05(01):45–63, 2008.
- [19] A. Harten, P. D. Lax, and B. van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *SIAM Rev.*, 25(1):35–61, 1983.
- [20] H. Holden and N. H. Risebro. *Front Tracking for Hyperbolic Conservation Laws*. Springer Berlin Heidelberg, 2011.

- [21] E. L. Isaacson and J. B. Temple. Analysis of a singular hyperbolic system of conservation laws. *J. Differential Equations*, 65(2):250–268, 1986.
- [22] B. L. Keyfitz and H. C. Kranzer. A system of non-strictly hyperbolic conservation laws arising in elasticity theory. *Arch. Ration. Mech. Anal.*, 72(3):219–241, 1980.
- [23] C. Klingenberg and N. H. Risebro. Stability of a resonant system of conservation laws modeling polymer flow with gravitation. *J. Differential Equations*, 170:344–380, 2001.
- [24] S. Kokh and F. Lagoutière. An anti-diffusive numerical scheme for the simulation of interfaces between compressible fluids by means of a five-equation model. *J. Comput. Phys.*, 229(8):2773–2809, 2010.
- [25] G. J. Kynch. A theory of sedimentation. *Trans. Faraday Soc.*, 48:166–176, 1952.
- [26] J.-P. Lebacque, S. Mammar, and H. Haj-Salem. The Aw–Rascle and Zhang’s model: Vacuum problems, existence and regularity of the solutions of the Riemann problem. *Transport. Res. Part B: Methodological*, 41(7):710–721, 2007.
- [27] R. J. LeVeque. *Numerical Methods for Conservation Laws*. Birkhäuser Verlag, 1992.
- [28] R. J. LeVeque and B. Temple. Stability of Godunov’s method for a class of  $2 \times 2$  systems of conservation laws. *Trans. Amer. Math. Soc.*, 288(1):115–123, 1985.
- [29] T. Li. Global solutions of nonconcave hyperbolic conservation laws with relaxation arising from traffic flow. *J. Differential Equations*, 190(1):131–149, 2003.
- [30] L. Lin, J. B. Temple, and J. Wang. A comparison of convergence rates for Godunov’s method and Glimm’s method in resonant nonlinear systems of conservation laws. *SIAM J. Numer. Anal.*, 32(3):824–840, 1995.
- [31] T.-P. Liu. The entropy condition and the admissibility of shocks. *J. Math. Anal. Appl.*, 53(1):78–88, 1976.
- [32] L. Longwei, B. Temple, and W. Jinghua. Suppression of oscillations in Godunov’s method for a resonant non-strictly hyperbolic system. *SIAM J. Numer. Anal.*, 32(3):841–864, 1995.
- [33] Y.-G. Lu and F. Gu. Existence of global bounded weak solutions to a Keyfitz-Kranzer system. *Commun. Math. Sci.*, 10(4):1133–1142, 2012.
- [34] S. Moutari and M. Rascle. A hybrid Lagrangian model based on the Aw–Rascle traffic flow model. *SIAM J. Appl. Math.*, 68(2):413–436, 2007.
- [35] O. A. Oleinik. Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation. *Uspekhi Mat. Nauk*, 14:165–170, 1959. Amer. Math. Soc. Transl. Ser. 2, 33, (1964), pp. 285–290.
- [36] D. L. Qiao, P. Zhang, S. C. Wong, and K. Choi. Discontinuous Galerkin finite element scheme for a conserved higher-order traffic flow model by exploring Riemann solvers. *Appl. Math. Comput.*, 244:567–576, 2014.
- [37] J. F. Richardson and W. N. Zaki. Sedimentation and fluidization: part I. *Trans. Inst. Chem. Eng.*, 32:35–53, 1954.
- [38] J. Smoller. *Shock Waves and Reaction-Diffusion Equations*. Springer-Verlag, 1983.
- [39] B. Temple. Global solution of the Cauchy problem for a class of  $2 \times 2$  nonstrictly hyperbolic conservation laws. *Adv. in Appl. Math.*, 3(3):335–375, 1982.
- [40] B. Temple. Systems of conservation laws with invariant submanifolds. *Trans. Amer. Math. Soc.*, 280(2):781–781, 1983.
- [41] D. H. Wagner. Equivalence of the Euler and Lagrangian equations of gas dynamics for weak solutions. *J. Differential Equations*, 68(1):118–136, 1987.
- [42] H. M. Zhang. A non-equilibrium traffic model devoid of gas-like behavior. *Transport. Res. Part B: Methodological*, 36(3):275–290, 2002.

# Centro de Investigación en Ingeniería Matemática (CI<sup>2</sup>MA)

## PRE-PUBLICACIONES 2015

- 2015-05 ERNESTO CÁCERES, GABRIEL N. GATICA: *A mixed virtual element method for the pseudostress-velocity formulation of the Stokes problem*
- 2015-06 WEIFENG QIU, MANUEL SOLANO: *High order approximation of mixed boundary value problems in curved domains by extensions from polygonal subdomains*
- 2015-07 ELIGIO COLMENARES, GABRIEL N. GATICA, RICARDO OYARZÚA: *Analysis of an augmented mixed-primal formulation for the stationary Boussinesq problem*
- 2015-08 JULIO ARACENA, EDUARDO PALMA, LILIAN SALINAS: *Enumeration and extension of non-equivalent deterministic update schedules in Boolean networks*
- 2015-09 JESSIKA CAMAÑO, GABRIEL N. GATICA, RICARDO OYARZÚA, GIORDANO TIERRA: *An augmented mixed finite element method for the Navier-Stokes equations with variable viscosity*
- 2015-10 MARIO ÁLVAREZ, GABRIEL N. GATICA, RICARDO RUIZ-BAIER: *A mixed-primal finite element approximation of a steady sedimentation-consolidation system*
- 2015-11 SEBASTIANO BOSCARINO, RAIMUND BÜRGER, PEP MULET, GIOVANNI RUSSO, LUIS M. VILLADA: *On linearly implicit IMEX Runge-Kutta methods for degenerate convection-diffusion problems modeling polydisperse sedimentation*
- 2015-12 RAIMUND BÜRGER, CHRISTOPHE CHALONS, LUIS M. VILLADA: *On second-order antididffusive Lagrangian-remap schemes for multispecies kinematic flow models*
- 2015-13 RAIMUND BÜRGER, SUDARSHAN K. KENETTINKARA, SARVESH KUMAR, RICARDO RUIZ-BAIER: *Finite volume element-discontinuous Galerkin approximation of viscous two-phase flow in heterogeneous porous media*
- 2015-14 GABRIEL N. GATICA, LUIS F. GATICA, FILANDER A. SEQUEIRA: *A priori and a posteriori error analyses of a pseudostress-based mixed formulation for linear elasticity*
- 2015-15 ANAHI GAJARDO, NICOLAS OLLINGER, RODRIGO TORRES: *Some undecidable problems about the trace-subshift associated to a Turing machine*
- 2015-16 FERNANDO BETANCOURT, RAIMUND BÜRGER, CHRISTOPHE CHALONS, STEFAN DIEHL, SEBASTIAN FARÅS: *A random sampling approach for a family of Temple-class systems of conservation laws*

Para obtener copias de las Pre-Publicaciones, escribir o llamar a: DIRECTOR, CENTRO DE INVESTIGACIÓN EN INGENIERÍA MATEMÁTICA, UNIVERSIDAD DE CONCEPCIÓN, CASILLA 160-C, CONCEPCIÓN, CHILE, TEL.: 41-2661324, o bien, visitar la página web del centro: <http://www.ci2ma.udec.cl>



**CENTRO DE INVESTIGACIÓN EN  
INGENIERÍA MATEMÁTICA (CI<sup>2</sup>MA)**  
**Universidad de Concepción**



Casilla 160-C, Concepción, Chile  
Tel.: 56-41-2661324/2661554/2661316  
<http://www.ci2ma.udec.cl>

