UNIVERSIDAD DE CONCEPCIÓN

Centro de Investigación en Ingeniería Matemática (CI^{2}MA)

Second-order schemes for conservation laws with discontinuous flux modelling clarifier-thickener units

Raimund Bürger, Kenneth H. KARLSEN,
Hector Torres, John D. Towers

PREPRINT 2009-18

SERIE DE PRE-PUBLICACIONES

SECOND-ORDER SCHEMES FOR CONSERVATION LAWS WITH DISCONTINUOUS FLUX MODELLING CLARIFIER-THICKENER UNITS

RAIMUND BÜRGER[A], KENNETH H. KARLSEN[B], HÉCTOR TORRES[C], AND JOHN D. TOWERS[D]

Abstract. Continuously operated clarifier-thickener units can be modeled by a non-linear, scalar conservation law with a flux that involves two parameters that depend discontinuously on the space variable. This paper presents two numerical schemes for the solution of this equation that have formal second-order accuracy in both the time and space variable. One of the schemes is a standard total variation diminishing (TVD) method, while the other scheme, the so-called flux-TVD (FTVD) scheme, is based on the property that due to the presence of the discontinuous parameters, the flux of the solution (rather than the solution itself) has the TVD property. The FTVD property is enforced by a new nonlocal limiter algorithm. We prove that the FTVD scheme converges to a BV_t solution of the conservation law with discontinuous flux. Numerical examples for both resulting schemes are presented. They produce comparable numerical errors, while the FTVD scheme is supported by convergence analysis. The accuracy of both schemes is superior to that of an available monotone first-order scheme. In the clarifier-thickener application there is interest in modelling sediment compressibility by an additional strongly degenerate diffusion term. Second-order schemes for this extended equation are obtained by combining either the TVD or the FTVD scheme with a Crank-Nicolson discretization of the degenerate diffusion term in a Strang-type operator splitting procedure. Numerical examples illustrate the resulting schemes.

1. Introduction

In a series of papers including [5, 6, 7], we proposed and analyzed difference schemes for conservation laws with discontinuous flux modelling so-called clarifier-thickener (CT) units for the continuous solid-liquid separation of suspensions in engineering applications. Most spatially one-dimensional mathematical models of these units are based on the kinematic sedimentation theory [17], which describes the batch settling of small, equal-sized rigid spheres suspended in a viscous fluid by the conservation law

$$u_t + b(u)_x = 0 \tag{1.1}$$

Date: November 8, 2009.

Key words and phrases. sedimentation, scalar conservation law, discontinuous coefficient, weak solution, entropy solution, uniqueness, difference scheme, convergence, existence.

[A]CI^{2}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.

[B]Centre of Mathematics for Applications (CMA), University of Oslo, P.O. Box 1053, Blindern, N–0316 Oslo, Norway. E-Mail: kennethk@math.uio.no.

[C]Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Universidad de Concepción, Casilla 160-C, Concepción, Chile. E-Mail: htorres@ing-mat.udec.cl.

[D]MiraCosta College, 3333 Manchester Avenue, Cardiff-by-the-Sea, CA 92007-1516, USA.
E-mail:john.towers@cox.net.

for the solids volume fraction u as a function of depth x and time t. The flux $b(u)$, called batch flux density function in the context of CT models, describes material specific properties of the suspension. The extension of this theory to CT units with continuous feed, sediment removal, and clarified liquid overflow leads to an initial value problem for a conservation law of the type

$$\begin{aligned} u_t + f\big(\boldsymbol{\gamma}(x), u\big)_x = 0, \quad x \in \mathbb{R}, \quad t > 0, \\ u(x,0) = u_0(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.2}$$

with a flux $f(\boldsymbol{\gamma}(x), u)$ that depends discontinuously on x via a vector $\boldsymbol{\gamma}(x) = (\gamma_1(x), \gamma_2(x))$ of discontinuous parameters. The flux discontinuities are a consequence of the assumption that within a CT unit, the suspension feed flow is split into upwards- and downwards-directed bulk flows, and of the particular description of vessel outlets. It is the purpose of this paper to introduce second-order accurate finite difference schemes for the approximate solution of (1.2) under specific assumptions of the context of the CT model.

The discontinuous flux makes the well-posedness analysis and numerical simulation of the CT model rather difficult. For example, if we express the discontinuous parameter $\boldsymbol{\gamma}(x)$ as an additional conservation law $\boldsymbol{\gamma}_t = 0$, we obtain a system of conservations laws for the "unknowns" $(\boldsymbol{\gamma}, u)$. The equation $\boldsymbol{\gamma}_t = 0$ introduces linearly degenerate fields with eigenvalues that are zero. Indeed, if $f_u = 0$ at some points $(\boldsymbol{\gamma}, u)$, then the system is non-strictly hyperbolic and it experiences so-called nonlinear resonant behavior. Consequently, one cannot in general expect to bound the total variation of the conserved quantities directly, but only when measured under a certain singular mapping, as was done first in [22] for a related system.

The papers [3, 4, 5, 6] cited above were inspired by previous work on conservation laws with discontinuous flux (cf. these papers for lists of relevant references). This area has enjoyed a lot of interest in recent years due to its intrinsic mathematical difficulties and the large number of its applications including, besides the CT model, two-phase flow in porous media, traffic flow with discontinuous road surface conditions, and shape-from-shading problems (again we refer to [3, 4, 5, 6] for long lists of relevant references). On the other hand, CT models have been studied extensively in the literature by several authors (see, e.g., [1, 9, 18]). Important contributions to the mathematical analysis and the determination of solutions to these first-order models have been made by Diehl, see, e.g., [10, 11, 12].

In many applications, suspensions are flocculated and form compressible sediment layers, which cannot be described by (1.1). A suitable extended model is provided by a sedimentation-consolidation theory (see, e.g., [7]), whose governing equation (for one-dimensional batch settling) is

$$u_t + b(u)_x = A(u)_{xx}, \tag{1.3}$$

where the diffusion term $A(u)_{xx}$ accounts for sediment compressibility. This theory postulates a material-dependent critical concentration (or gel point) u_c such that $A(u) = 0$ for $u \le u_c$ and $A(u) \ge 0$ for $u > u_c$. Thus, (1.3) degenerates into the first-order equation (1.1) when $u \le u_c$, and is therefore called *strongly* degenerate. If we combine this extension of (1.1) to continuously operated CTs (with the degenerate diffusion term describing sediment compressibility), then the resulting model for a CT treating a flocculated suspension is of the type

$$\begin{aligned} u_t + f\big(\boldsymbol{\gamma}(x), u\big)_x = \big(\gamma_1(x) A(u)_x\big)_x, \quad x \in \mathbb{R}, \quad t > 0, \\ u(x,0) = u_0(x), \quad x \in \mathbb{R}, \end{aligned} \tag{1.4}$$

where the strongly degenerating (with respect to u) diffusion term is modulated by the discontinuous parameter $\gamma_1(x)$. In this paper we will mostly consider the

purely hyperbolic CT model (1.2), which can be obtained by taking $A \equiv 0$ in (1.4), but see Sections 2 and 7.

In [6, 8], our interest was focused on the well-posedness analysis for conservation laws with discontinuous flux. Although these papers include numerical experiments, our main interest in numerical schemes, in particular in a suitable adaptation of the first-order Engquist-Osher scheme [13] to account for flux discontinuities, has so far been motivated by providing a constructive proof of existence of a weak solution, or even of an entropy solution, by proving convergence of the scheme. However, the schemes used so far are only first-order accurate in space and time, and due to their poor resolution are not recommended for practical simulations.

It is the purpose of this paper to present, and in part analyze, finite difference schemes that form second-order accurate approximations (both in space and in time) of the CT model. These schemes utilize our previous first-order scheme and a new flux-total variation diminishing (FTVD) method. In more detail, by a truncation error analysis we identify a correction term that formally upgrades our scheme to second-order accuracy. As is well known, the resulting Lax-Wendroff-type scheme produces spurious oscillations near discontinuities. A well-established way to correct this is the application of a limiter function to the solution itself. This results in a TVD scheme. The problem with the application of the TVD methodology, however, lies in the fact that for equations involving a discontinuous flux, it is not ensured that the solution itself satisfies the TVD property; rather, we can only say that the *flux* has the TVD property. This observation leads us to propose here the so-called flux-TVD (FTVD) schemes, which precisely mimic the latter property of the exact solution. The new correction terms introduced by the flux-TVD approach should be as large as possible in order to ensure overall second-order accuracy. This requirement has inspired us to propose a new nonlocal limiter algorithm, which as we prove, indeed diminishes total variation and preserves second-order accuracy wherever possible.

We prove that the FTVD scheme converges to a BV_t weak solution of the CT model. A decisive ingredient of the proof is the application of a so-called singular mapping, which maps the sequence of approximate solution values, which do not necessarily satisfy a spatial TVD property, to a sequence of transformed quantities, which does have this property. A standard compactness argument yields that the transformed sequence has a limit, and applying the inverse of the singular mapping we see that the sequence of solution itself has a limit. This analysis puts the FTVD scheme on a rigorous ground. Regarding the first-order version of the scheme, we know that it satisfies a discrete entropy inequalities, the continuous version of which implies L^1 stability and uniqueness, cf. [6, 8]. For the second-order extension, we have not been able to establish such discrete entropy inequalities, although the numerical results seem to indicate that they are satisfied.

The remainder of this paper is organized as follows: In Section 2 we outline the CT model. Moreover, we state the definition of a BV_t weak solution. In Section 3, we recall from [6] our first-order scheme for the discretization of (1.2), which forms the starting point of our analysis, and identify a correction term that formally upgrades our scheme to second-order accuracy. To avoid that the resulting Lax-Wendroff-type scheme produces spurious oscillations near discontinuities, we utilize limiter functions, including a simple minmod TVD limiter and a novel (nonlocal) flux-TVD limiter. The nonlocal limiter function and some of its properties are further discussed in Section 4. In Section 5, we prove that the flux-TVD scheme converges to a BV_t weak solution of the CT model. In Section 6 we provide several numerical examples illustrating the proposed schemes. While in Sections 3–6, which form the core of this paper, we are concerned with the CT model defined by (1.2),

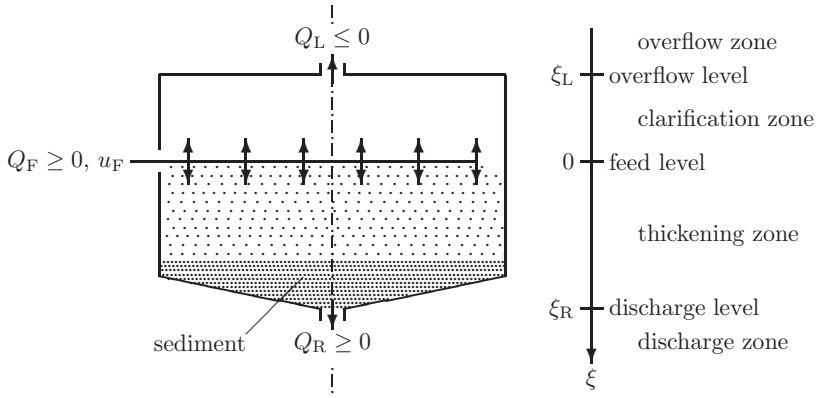

FIGURE 1. Schematic illustration of a clarifier-thickener (CT) unit.

in Section 7 we propose an extension of the TVD and flux-TVD to the version of the CT model that incorporates a strongly degenerate parabolic term modelling sediment compressibility via an operator splitting procedure with a Crank-Nicolson discretization of the parabolic term, and provide some numerical examples for this extension.

2. The clarifier-thickener model

In this section we outline a general CT model. For the case of a varying cross-sectional area, the final model equation slightly differs from the one stated in [7], since by a simple transformation of the spatial variable, we now rewrite the governing PDE in conservative form, while in previous papers [5, 7] we still had the (possibly discontinuous) cross-sectional area function multiplying the time derivative of the solution.

We here derive the complete CT model that also includes the effect of sediment compressibility modeled by a degenerate diffusion term since the particular algebraic form of that term, and the discontiuous parameter that appears in it, are most easily motivated by extensions of expressions that appear in the derivation of the first-order hyperbolic model. However, we are mainly analyzing the special case of a first-order hyperbolic model for which this term is not present. If this term is not present, we speak of an *ideal suspension.*

2.1. The clarifier-thickener unit.

We consider a continuously operated axisymmetric clarifier-thickener (CT) vessel as drawn in Figure 1, and assume that all flow variables depend on depth ξ and time t only. We subdivide the vessel into four different zones: the thickening zone ($0 < \xi < \xi_{\mathrm{R}}$), the clarification zone ($\xi_{\mathrm{L}} < \xi < 0$), the underflow zone ($\xi > \xi_{\mathrm{R}}$) and the overflow zone ($\xi < \xi_{\mathrm{L}}$). The vessel is continuously fed at depth $\xi = 0$, the feed level, with fresh feed suspension at a volume feed rate $Q_{\mathrm{F}}(t) \ge 0$. The concentration of the feed suspension is $u_{\mathrm{F}}(t)$. The prescribed volume underflow rate, at which the thickened sediment is removed from the unit, is $Q_{\mathrm{R}}(t) \ge 0$. Consequently, the overflow rate is $Q_{\mathrm{L}}(t) = Q_{\mathrm{F}}(t) - Q_{\mathrm{R}}(t)$, where we assume that the two control functions $Q_{\mathrm{F}}(t)$ and $Q_{\mathrm{R}}(t)$ are chosen such that $Q_{\mathrm{F}}(t) \ge 0$.

2.2. Derivation of the mathematical model.

The spatially one-dimensional balance equation for $u = u(\xi, t)$ in a vessel with varying cross-sectional area $S(\xi)$

is given by

$$S(\xi)u_t + \big(Q(t)u + S(\xi)u(1-u)v_{\mathrm{r}}\big)_\xi = 0, \tag{2.1}$$

where $Q(t)$ is the controllable volume average flow rate and v_{r} is the solid-fluid relative velocity; see [7] for details. Within the kinematic sedimentation theory [17] for ideal suspensions, v_{r} is assumed to be a function of u only, $v_{\mathrm{r}} = v_{\mathrm{r}}(u)$. In terms of the batch flux density function $b(u)$ we get

$$v_{\mathrm{r}}(u) = \frac{b(u)}{u(1-u)}. \tag{2.2}$$

The function b is usually assumed to be continuous and piecewise twice differentiable with $b \in C^2([0, u_{\max}])$ and $b(u) = 0$ for $u \le 0$ or $u \ge u_{\max}$, where $u_{\max}$ is the maximum solids concentration, $b(u) > 0$ for $0 < u < u_{\max}$, $b'(0) > 0$ and $b'(u_{\max}) \le 0$. A typical example that satisfies these assumptions is

$$b(u) = \begin{cases} v_\infty u(1-u)^C & \text{if } 0 < u < u_{\max}, \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

where $C \ge 1$ and $v_\infty > 0$ is the settling velocity of a single particle in pure fluid.

If we include the effect of sediment compressibility, then (2.2) is replaced by

$$v_{\mathrm{r}} = \frac{b(u)}{u(1-u)}\left(1 - \frac{\sigma_{\mathrm{e}}'(u)}{\Delta\varrho g u}u_x\right), \tag{2.4}$$

where $\Delta\varrho > 0$ denotes the solid-fluid density difference, g the acceleration of gravity, and $\sigma_{\mathrm{e}}(u)$ is the effective solid stress function, which is now the second constitutive function (besides b) characterizing the suspension. This function is assumed to satisfy $\sigma_{\mathrm{e}}(u) \ge 0$ for all u and

$$\sigma_{\mathrm{e}}'(u) := \frac{d\sigma_{\mathrm{e}}(u)}{du} \begin{cases} = 0 & \text{for } u \le u_{\mathrm{c}}, \\ > 0 & \text{for } u > u_{\mathrm{c}}. \end{cases}$$

Clearly, the first-order model based on (2.2) is included as the sub-case of (2.4) produced by setting $u_{\mathrm{c}} = u_{\max}$.

Inserting (2.4) into (2.1) and defining

$$a(u) := \frac{b(u)\sigma_{\mathrm{e}}'(u)}{\Delta\varrho g u}, \quad A(u) := \int_0^u a(s)\,ds, \tag{2.5}$$

we obtain the governing equation

$$\big(S(\xi)u\big)_t + \big(Q(t)u + S(\xi)b(u)\big)_\xi = \big(S(\xi)A(u)_\xi\big)_\xi. \tag{2.6}$$

Since $a(u) = 0$ for $u \le u_{\mathrm{c}}$ and $u = u_{\max}$ and $a(u) > 0$ otherwise, (2.6) is first-order hyperbolic for $u \le u_{\mathrm{c}}$ and second-order parabolic for $u > u_{\mathrm{c}}$, and therefore (2.6) is called strongly degenerate parabolic. The location of the type-change interface $u = u_{\mathrm{c}}$ (denoting the sediment level) is in general unknown beforehand. In accordance with (2.5), we will assume that $A \in \mathrm{Lip}([0,1])$, $A'(u) = 0$ for $u < u_{\mathrm{c}}$, and that $A'(u) > 0$ for $u \in (u_{\mathrm{c}}, 1)$.

In the present model, the volume bulk flows are $Q(\xi, t) = Q_{\mathrm{R}}(t)$ for $\xi > 0$ and $Q(\xi, t) = Q_{\mathrm{L}}(t)$ for $\xi < 0$. This suggests employing (2.6) with $Q(t) = Q_{\mathrm{R}}(t)$ for $0 < \xi < \xi_{\mathrm{R}}$ and $Q(t) = Q_{\mathrm{L}}(t)$ for $\xi_{\mathrm{L}} < \xi < 0$, however, we herein choose the control functions $u_{\mathrm{F}}(t)$, $Q_{\mathrm{L}}(t)$ and $Q_{\mathrm{R}}(t)$ to be time-independent constants. Furthermore, we assume that in the overflow and underflow zones the solid-fluid relative velocity vanishes, $v_{\mathrm{r}} = 0$. Moreover, the cross-sectional area $S(\xi)$ needs to be positive

outside the interval $[\xi_{\mathrm{L}}, \xi_{\mathrm{R}}]$. We assume that $S(\xi) = S_0$ for $\xi < \xi_{\mathrm{L}}$ and $\xi > \xi_{\mathrm{R}}$, where $S_0 > 0$ is a small but positive pipe diameter. We now obtain that

$$S(\xi) u v_{\mathrm{s}}|_{\xi \notin [\xi_{\mathrm{L}}, \xi_{\mathrm{R}}]} = S_0 u v_{\mathrm{s}} = \begin{cases} Q_{\mathrm{L}} u & \text{for } \xi < \xi_{\mathrm{L}}, \\ Q_{\mathrm{R}} u & \text{for } \xi > \xi_{\mathrm{R}}, \end{cases}$$

where v_{s} is the solids phase velocity. The feed mechanism is introduced by adding the singular source term $Q_{\mathrm{F}} u_{\mathrm{F}} \delta(\xi)$ to the right-hand part of the solids continuity equation. We can summarize the resulting PDE as

$$S(\xi) u_t + \tilde{G}(\xi, u)_\xi = \big(\beta_1(\xi) A(u)_\xi\big)_\xi + Q_{\mathrm{F}} u_{\mathrm{F}} \delta(\xi), \quad \xi \in \mathbb{R}, \quad t > 0, \tag{2.7}$$

$$\tilde{G}(\xi, u) = S(\xi) u v_{\mathrm{s}} = \begin{cases} Q_{\mathrm{L}} u & \text{for } \xi < \xi_{\mathrm{L}}, \\ Q_{\mathrm{L}} u + S(\xi) b(u) & \text{for } \xi_{\mathrm{L}} < \xi < 0, \\ Q_{\mathrm{R}} u + S(\xi) b(u) & \text{for } 0 < \xi < \xi_{\mathrm{R}}, \\ Q_{\mathrm{R}} u & \text{for } \xi > \xi_{\mathrm{R}}, \end{cases}$$

$$\beta_1(\xi) := \begin{cases} S(\xi) & \text{if } \xi_{\mathrm{L}} \le \xi \le \xi_{\mathrm{R}}, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, we may express the singular source term in terms of the derivative of the Heaviside function. Adding $-H(\xi) Q_{\mathrm{F}} u_{\mathrm{F}}$ to $\tilde{G}(\xi, u)$ and subtracting the constant term $Q_{\mathrm{L}} u_{\mathrm{F}}$, and starting from a known initial concentration distribution u_0, we obtain the strongly degenerate convection-diffusion problem

$$\begin{aligned} S(\xi) u_t + g\big(\boldsymbol{\beta}(\xi), u\big)_\xi &= \big(\beta_1(\xi) A(u)_\xi\big)_\xi, \quad \xi \in \mathbb{R}, \quad t > 0, \\ u(\xi, 0) &= u_0(\xi), \quad \xi \in \mathbb{R}, \quad u_0(\xi) \in [0, u_{\max}], \end{aligned} \tag{2.8}$$

where we define the flux

$$g\big(\boldsymbol{\beta}(\xi), u\big) := \beta_1(\xi) b(u) + \beta_2(\xi)(u - u_{\mathrm{F}}),$$

$$\boldsymbol{\beta}(\xi) := \big(\beta_1(\xi), \beta_2(\xi)\big), \quad \beta_2(\xi) := \begin{cases} Q_{\mathrm{L}} & \text{for } \xi < 0, \\ Q_{\mathrm{R}} & \text{for } \xi > 0. \end{cases}$$

Our numerical algorithms and their analysis are greatly simplified if we do not have the term $S(\xi)$ multiplying u_t. With the change of variables

$$x = \int_0^\xi S(\eta)\, d\eta, \quad dx/d\xi = S, \quad x_{\mathrm{L}} = x(\xi_{\mathrm{L}}), \quad x_{\mathrm{R}} = x(\xi_{\mathrm{R}}), \tag{2.9}$$

we can rewrite the initial value problem for (2.8) as (1.4), where we define

$$f\big(\boldsymbol{\gamma}(x), u\big) := \gamma_1(x) b(u) + \gamma_2(x)(u - u_{\mathrm{F}}), \tag{2.10}$$

$$\gamma_1(x) := \begin{cases} S(\xi(x)) & \text{for } x \in (x_{\mathrm{L}}, x_{\mathrm{R}}), \\ 0 & \text{for } x \notin (x_{\mathrm{L}}, x_{\mathrm{R}}), \end{cases}, \qquad \gamma_2(x) := \begin{cases} Q_{\mathrm{L}} & \text{for } x < 0, \\ Q_{\mathrm{R}} & \text{for } x > 0. \end{cases}$$

If we consider an ideal suspension not exhibiting sediment compressibility, then (1.4) takes the purely hyperbolic form (1.2), which is the equation that will be mainly concerned with in this paper.

We assume that the function $x \mapsto S(\xi(x))$ is piecewise smooth with a finite number of discontinuities, and for the initial data in (1.2) we assume that u_0 satisfies

$$u_0 \in BV(\mathbb{R}); \;\; u_0(x) \in [0, 1] \text{ for a.e. } x \in \mathbb{R}.$$

By a solution to the hyperbolic problem (1.2), we understand the following.

Definition 2.1 (BV_t weak solution)**.** *A measurable function $u : \Pi_T \to \mathbb{R}$ is a BV_t weak solution of the initial value problem* (1.4) *if it satisfies the following conditions:*

(**D**.1) $u \in (L^\infty \cap BV_t)(\Pi_T)$.
(**D**.2) *For all test functions* $\phi \in \mathcal{D}(\mathbb{R} \times [0,T))$,

$$\iint_{\Pi_T} \Big(u\phi_t + f\big(\boldsymbol{\gamma}(x), u\big)\phi_x \Big)\, dx\, dt + \int_{\mathbb{R}} u_0\phi(x,0)\, dx = 0.$$

The notation BV_t refers to the space of locally integrable functions on Π_T for which u_t (but not u_x) is a locally bounded measure, which is a superset of BV.

3. The difference schemes

3.1. **Algorithm preliminaries.** We start with a positive spatial mesh size $\Delta x > 0$, set $x_j := j\Delta x$, and discretize the parameter vector $\boldsymbol{\gamma}$ and the initial data by $\boldsymbol{\gamma}_{j+1/2} := \boldsymbol{\gamma}(x_{j+1/2}+)$ and $U_j^0 := u_0(x_j+)$ for $j \in \mathbb{Z}$. Here $x_{j+1/2} := x_j + \Delta x/2$, i.e., the midpoint in the interval $[x_j, x_{j+1})$. Let $t_n := n\Delta t$ and let χ^n denote the characteristic function of $[t_n, t_{n+1})$, χ_j the characteristic function of $[x_{j-1/2}, x_{j+1/2})$, and $\chi_{j+1/2}$ the characteristic function of the interval $[x_j, x_{j+1})$. Our difference algorithm will produce an approximation U_j^n associated with the point (x_j, t_n). We then define

$$u^\Delta(x,t) := \sum_{n\ge 0}\sum_{j\in\mathbb{Z}} U_j^n \chi_j(x)\chi^n(t), \quad \boldsymbol{\gamma}^\Delta(x) := \sum_{j\in\mathbb{Z}} \boldsymbol{\gamma}_{j+1/2}\chi_{j+1/2}(x). \tag{3.1}$$

Our algorithm is defined by the simple marching formula

$$U_j^{n+1} = U_j^n - \lambda\Delta_-\big(h_{j+1/2}^n + \hat{F}_{j+1/2}^n\big), \quad \lambda = \frac{\Delta t}{\Delta x}, \quad j \in \mathbb{Z}, \quad n = 0,1,2,\ldots. \tag{3.2}$$

Here $h_{j+1/2}^n := h(\boldsymbol{\gamma}_{j+1/2}, U_{j+1}^n, U_j^n)$, where h is the Engquist-Osher flux [13]:

$$h(\boldsymbol{\gamma}, v, u) := \frac{1}{2}\big(f(\boldsymbol{\gamma}, u) + f(\boldsymbol{\gamma}, v)\big) - \frac{1}{2}\int_u^v \big|f_u(\boldsymbol{\gamma}, w)\big|\, dw, \tag{3.3}$$

and the quantity $\hat{F}_{j+1/2}^n$ is a correction term that is required in order to achieve second-order accuracy. Without those terms, (3.2) is the first-order scheme that we have analyzed in previous papers. The simplicity of the scheme derives in large part from the fact that the discretization of $\boldsymbol{\gamma}$ is staggered with respect to that of the conserved quantity u, making it possible to avoid solving 2×2 Riemann problems that would result otherwise.

Finally, we will assume that λ remains constant as we refine the mesh, so that $\Delta t = \lambda\Delta x$.

3.2. **Truncation error analysis.** In this section we focus on the difference scheme (3.2) for (1.2). We start by defining second-order correction terms $d_{j+1/2}^n$, $e_{j+1/2}^n$ that are appropriate if $\boldsymbol{\gamma}$ is piecewise constant. We are seeking formal second-order accuracy at points (x,t) where the solution u is smooth. At jumps in $\boldsymbol{\gamma}$ the solution will generally be discontinuous, so for the purpose of defining correction terms, we may restrict our attention to points located away from the jumps in $\boldsymbol{\gamma}$. Combined with our (temporary) assumption that $\boldsymbol{\gamma}$ is piecewise constant we see that we can simply use correction terms that are appropriate for a constant $\boldsymbol{\gamma}$ conservation law. Specifically, we use the following Lax-Wendroff type correction terms that are well known to provide for formal second-order accuracy in both space and time (see e.g. [21]):

$$\begin{aligned} d_{j+1/2}^n &= \frac{1}{2}a_{j+1/2}^+\big(1 - \lambda a_{j+1/2}^+\big)\Delta_+ U_j^n, \\ e_{j+1/2}^n &= \frac{1}{2}a_{j+1/2}^-\big(1 + \lambda a_{j+1/2}^-\big)\Delta_+ U_j^n. \end{aligned} \tag{3.4}$$

Here the quantities $a^{\pm}_{j+1/2}$ are the positive and negative wave speeds associated with the cell boundary located at $x_{j+1/2}$:

$$\begin{aligned} a^{+}_{j+1/2} :=& \frac{1}{\Delta_+ U^n_j} \int_{U^n_j}^{U^n_{j+1}} \max\bigl(0, f_u(\boldsymbol{\gamma}_{j+1/2}, w)\bigr)\, dw \\ =& \frac{f(\boldsymbol{\gamma}_{j+1/2}, U^n_{j+1}) - h^n_{j+1/2}}{\Delta_+ U^n_j} \geq 0, \\ a^{-}_{j+1/2} :=& \frac{1}{\Delta_+ U^n_j} \int_{U^n_j}^{U^n_{j+1}} \min\bigl(0, f_u(\boldsymbol{\gamma}_{j+1/2}, w)\bigr)\, dw \\ =& \frac{h^n_{j+1/2} - f(\boldsymbol{\gamma}_{j+1/2}, U^n_j)}{\Delta_+ U^n_j} \leq 0. \end{aligned} \tag{3.5}$$

The scheme discussed thus far is only first-order accurate if $\boldsymbol{\gamma}$ is variable. We now set out to find second-order correction terms that are required when $x \mapsto \boldsymbol{\gamma}$ is piecewise C^2, and start by identifying the truncation error of the first-order scheme. For the moment, we restrict our attention to the case $f_u(\boldsymbol{\gamma}, u) \geq 0$, so the first-order version of the scheme (3.2) simplifies to

$$U^{n+1}_j - U^n_j + \lambda \Delta_- f\bigl(\boldsymbol{\gamma}_{j+1/2}, U^n_j\bigr) = 0. \tag{3.6}$$

Inserting a smooth solution $u(x,t)$ into (3.6) and using u^n_j to denote $u(x_j, t^n)$, we get the following expression for the truncation error at the point (x_j, t^n):

$$\begin{aligned} TE^+ &:= u^{n+1}_j - u^n_j + \lambda \Delta_- f\bigl(\boldsymbol{\gamma}_{j+1/2}, u^n_j\bigr) \\ &= \Delta t \, (u_t)^n_j + \frac{1}{2} \Delta t^2 \, (u_{tt})^n_j + \lambda \Delta_- f\bigl(\boldsymbol{\gamma}_{j+1/2}, u^n_j\bigr) + \mathcal{O}(\Delta^3). \end{aligned} \tag{3.7}$$

Here we are using the abbreviation $\mathcal{O}(\Delta^\nu) = \mathcal{O}(\Delta t^\nu)$, which is also equal to $\mathcal{O}(\Delta x^\nu)$, since $\Delta t = \lambda \Delta x$. From the differential equation (1.2) we have

$$u_t = -f(\boldsymbol{\gamma}, u)_x, \quad u_{tt} = \bigl(f_u(\boldsymbol{\gamma}, u) f(\boldsymbol{\gamma}, u)_x\bigr)_x.$$

If we substitute these relationships into (3.7), then the truncation error becomes

$$\begin{aligned} TE^+ = &-\Delta t \bigl(f(\boldsymbol{\gamma}, u)_x\bigr)^n_j + \frac{1}{2} \Delta t^2 \bigl((f_u(\boldsymbol{\gamma}, u) f(\boldsymbol{\gamma}, u)_x)_x\bigr)^n_j \\ &+ \lambda \Delta_- f\bigl(\boldsymbol{\gamma}_{j+1/2}, u^n_j\bigr) + \mathcal{O}(\Delta^3). \end{aligned} \tag{3.8}$$

Abbreviating $f_u(\boldsymbol{\gamma}, u) := f_u$, $f(\boldsymbol{\gamma}, u)_x := f_x$, etc., for the last term in (3.8), we obtain by a straightforward but lengthy calculation

$$\Delta_- f\bigl(\boldsymbol{\gamma}_{j+1/2}, u^n_j\bigr) = \Delta x \, (f_x)^n_j - \frac{1}{2} \Delta x^2 \bigl((f_u u_x)_x\bigr)^n_j + \mathcal{O}(\Delta^3).$$

Inserting this expression into (3.8) we obtain

$$TE^+ = \Delta x^2 \lambda \left[\frac{1}{2} \lambda \, (f_u f_x)_x - \frac{1}{2} \, (f_u u_x)_x \right]^n_j + \mathcal{O}(\Delta^3). \tag{3.9}$$

Substituting $f_x = f_u u_x + f_{\boldsymbol{\gamma}} \boldsymbol{\gamma}_x$ into (3.9), where we define $f_{\boldsymbol{\gamma}} := \nabla_{\boldsymbol{\gamma}} f$, and suppressing the dependence on the point (x_j, t^n) gives

$$\begin{aligned} TE^+ &= \Delta x^2 \lambda \left[\frac{1}{2} \lambda f_u f_u u_x + \frac{1}{2} \lambda f_u f_{\boldsymbol{\gamma}} \boldsymbol{\gamma}_x - \frac{1}{2} f_u u_x \right]_x + \mathcal{O}(\Delta^3) \\ &= -\Delta x^2 \lambda \left[\frac{1}{2} f_u (1 - \lambda f_u) u_x - \frac{1}{2} \lambda f_u f_{\boldsymbol{\gamma}} \boldsymbol{\gamma}_x \right]_x + \mathcal{O}(\Delta^3). \end{aligned} \tag{3.10}$$

Similarly, when $f_u \leq 0$, the first-order scheme reduces to

$$U^{n+1}_j - U^n_j + \lambda \Delta_+ f\bigl(\boldsymbol{\gamma}_{j-1/2}, U^n_j\bigr) = 0,$$

and we arrive at the following formula for the truncation error:

$$TE^- = \Delta x^2 \lambda \left[\frac{1}{2} f_u (1 + \lambda f_u) u_x + \frac{1}{2} \lambda f_u f_{\boldsymbol{\gamma}} \boldsymbol{\gamma}_x \right]_x + \mathcal{O}(\Delta^3). \tag{3.11}$$

So, when $\boldsymbol{\gamma}$ is piecewise smooth (not piecewise constant), we see from (3.10) and (3.11) that appropriate second-order correction terms are the following modified versions of (3.4):

(3.12)

$$\begin{aligned}
F^n_{j+1/2} &:= D^n_{j+1/2} - E^n_{j+1/2}, \\
D^n_{j+1/2} &:= \frac{1}{2} a^+_{j+1/2} \big(1 - \lambda a^+_{j+1/2}\big) \Delta_+ U^n_j - \frac{1}{2} \lambda a^+_{j+1/2} f_{\boldsymbol{\gamma}} \big(\boldsymbol{\gamma}_{j+1/2}, U^n_{j+1/2}\big) \Delta_+ \boldsymbol{\gamma}_j \\
&= d^n_{j+1/2} - \frac{1}{2} \lambda a^+_{j+1/2} f_{\boldsymbol{\gamma}} \big(\boldsymbol{\gamma}_{j+1/2}, U^n_{j+1/2}\big) \Delta_+ \boldsymbol{\gamma}_j, \\
E^n_{j+1/2} &:= \frac{1}{2} a^-_{j+1/2} \big(1 + \lambda a^-_{j+1/2}\big) \Delta_+ U^n_j + \frac{1}{2} \lambda a^-_{j+1/2} f_{\boldsymbol{\gamma}} \big(\boldsymbol{\gamma}_{j+1/2}, U^n_{j+1/2}\big) \Delta_+ \boldsymbol{\gamma}_j \\
&= e^n_{j+1/2} + \frac{1}{2} \lambda a^-_{j+1/2} f_{\boldsymbol{\gamma}} \big(\boldsymbol{\gamma}_{j+1/2}, U^n_{j+1/2}\big) \Delta_+ \boldsymbol{\gamma}_j.
\end{aligned}$$

For the values $f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_{j+1/2}, U^n_{j+1/2})$ appearing in (3.12), we use the approximation

$$f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_{j+1/2}, U^n_{j+1/2}) \approx \frac{1}{2} \big(f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_{j+1/2}, U^n_j) + f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}_{j+1/2}, U^n_{j+1}) \big). \tag{3.13}$$

Even without the jumps in $\boldsymbol{\gamma}$, the solution will generally develop discontinuities. If we use the correction terms above without further processing, the solution will develop spurious oscillations near these discontinuities. To damp out the oscillations, we apply so-called flux limiters, resulting in the flux-limited quantities $\hat{F}_{j+1/2}$.

3.3. **A simple minmod TVD scheme.** In the constant $\boldsymbol{\gamma}$ case, the actual solution of the conservation law will be TVD, meaning that its total spatial variation decreases (or at least does not increase) in time. There are any number of ways to apply flux limiters in this situation so that the approximations U^n_j are also TVD. A simple limiter that enforces the TVD property when $\boldsymbol{\gamma}$ is constant is the following:

$$\begin{aligned}
\hat{F}^n_{j+1/2} &= \hat{D}^n_{j+1/2} - \hat{E}^n_{j+1/2}, \\
\hat{D}^n_{j+1/2} &= \operatorname{minmod}\big(D^n_{j+1/2}, 2D^n_{j-1/2}\big), \\
\hat{E}^n_{j+1/2} &= \operatorname{minmod}\big(E^n_{j+1/2}, 2E^n_{j+3/2}\big),
\end{aligned} \tag{3.14}$$

where we recall that the m-variable minmod function is defined by

$$\operatorname{minmod}(p_1, \ldots, p_m) = \begin{cases} \min\{p_1, \ldots, p_m\} & \text{if } p_1 \geq 0, \ldots, p_m \geq 0, \\ \max\{p_1, \ldots, p_m\} & \text{if } p_1 \leq 0, \ldots, p_m \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Of course, when $\boldsymbol{\gamma}$ is not constant, the actual solution u is not TVD, and thus our algorithm should not attempt to impose a TVD requirement on the conserved quantity U^n_j. Fortunately, the TVD limiter (3.14) only forces U^n_j to be TVD when $\boldsymbol{\gamma}$ is constant, and in practice turns out to be a reasonable approach to dampening oscillations even in the variable $\boldsymbol{\gamma}$ context considered here. Moreover, it is consistent with formal second accuracy away from extrema of u. Although we are unable to put the resulting algorithm on a firm theoretical footing, for the most part it is very robust. The one negative aspect that we have observed is a small amount of overshoot in certain cases when a shock collides with a stationary discontinuity at a jump in $\boldsymbol{\gamma}$, see Figure 2.

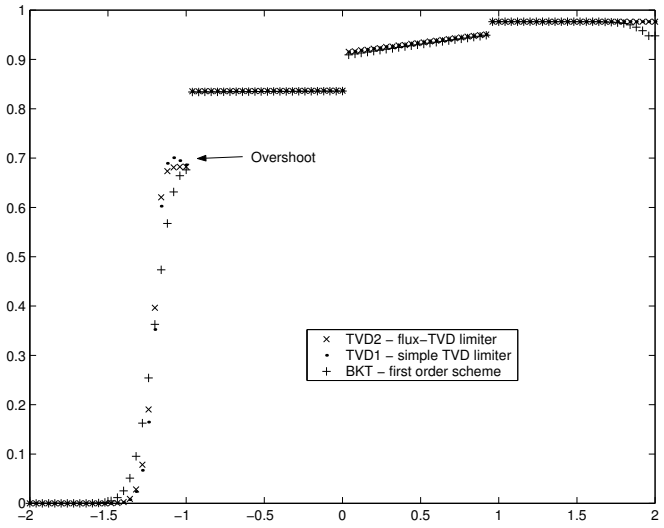

FIGURE 2. Same setup as Figure 3. Second-order schemes, with simple TVD and flux-TVD limiters. Shows overshoot produced by TVD1 but not TVD2. This is the same problem as shown in Figure 5 of [6]. Here $\Delta x = 1/25$, $\Delta t = 1/400$, 1020 time steps.

3.4. **A flux-TVD (FTVD) scheme.** We wish to eliminate the non-physical overshoot observed with the simple TVD limiter (3.14), and also put the resulting difference scheme on a firm theoretical basis. In the constant $\boldsymbol{\gamma}$ setting, the TVD concept originated by requiring that the numerical approximations satisfy a property (TVD) that is also satisfied by the actual solution. In the variable $\boldsymbol{\gamma}$ setting, the actual solution is not TVD, so we should enforce some other regularity property. For a conservation law having a flux with a discontinuous spatial dependency, it is natural to expect not the conserved variable, but the flux, to be TVD; see [23]. Consequently, we require that the first-order numerical flux also be TVD, i.e.,

$$\sum_{j\in\mathbb{Z}} \bigl|\Delta_+ h^{n+1}_{j-1/2}\bigr| \le \sum_{j\in\mathbb{Z}} \bigl|\Delta_+ h^{n}_{j-1/2}\bigr|, \quad n = 0, 1, \ldots.$$

We call this property flux-TVD, or FTVD. We will see (Lemmas 5.1 and 5.3) that under an appropriate CFL condition, the FTVD property (along with a bound on the solution) holds if

$$\bigl|\Delta_+ \hat{F}^n_{j+1/2}\bigr| \le \bigl|\Delta_+ h^n_{j+1/2}\bigr|, \quad j \in \mathbb{Z}, \quad n = 0, 1, 2, \ldots. \tag{3.15}$$

Wherever the solution is smooth, the quantity on the left side of (3.15) is $\mathcal{O}(\Delta^2)$, while the quantity on the right side is $\mathcal{O}(\Delta)$, making it seem plausible that we can satisfy these inequalities without sending $F^n_{j+1/2}$ all the way to zero, which would just give us the first-order scheme. It is reasonable to also impose the condition

$$0 \le \hat{F}^n_{j+1/2}/F^n_{j+1/2} \le 1, \quad j \in \mathbb{Z}, \quad n = 0, 1, 2, \ldots. \tag{3.16}$$

in addition to (3.15), so that after we have applied the correction terms, the numerical flux lies somewhere between the first-order flux and the pre-limiter version of the second-order flux, i.e.,

$$h^n_{j+1/2} + \hat{F}^n_{j+1/2} \in \mathrm{co}\bigl(h^n_{j+1/2}, h^n_{j+1/2} + F^n_{j+1/2}\bigr), \quad j \in \mathbb{Z}, \quad n = 0, 1, 2, \ldots.$$

We can view (3.15), (3.16) as a system of inequalities, and ask if it is possible to find a solution that keeps the ratio $\hat{F}^n_{j+1/2}/F^n_{j+1/2}$ appearing in (3.16) close enough

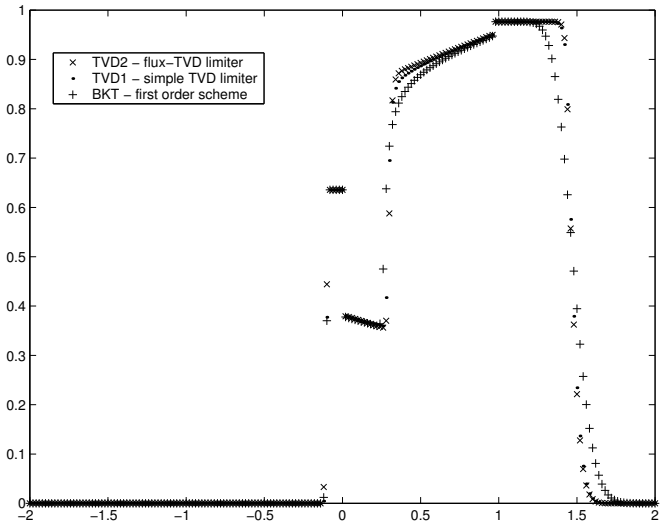

FIGURE 3. Comparison of the nonlocal flux-TVD limiter and the simple TVD limiter (3.14). This is the same problem as shown Figure 5 of [6]. Here $\Delta x = .02$, $\Delta t = .00125$, 800 time steps.

to unity that we still have formal second-order accuracy. This leads us to propose the nonlocal limiter algorithm that we describe in the next section. Via this algorithm we are in fact able to solve the system of inequalities (3.15), (3.16) in a manner that is compatible with formal second-order accuracy. Although the algorithm is nonlocal in nature, computationally it is (at least with our implementation) only slightly slower than the simpler TVD limiter (3.14). A nonlocal limiter seems to be unavoidable here—we believe that there is no FTVD limiter that depends on only some fixed finite number of the quantities $F^n_{j+1/2}$ and is consistent with formal second-order accuracy.

For the case of piecewise constant $\boldsymbol{\gamma}$, the results produced by the two algorithms (TVD and FTVD) usually differ by only a small amount; see Figure 3. However, we have observed one situation where there is a discernable difference—the case of a shock impinging on a discontinuity in $\boldsymbol{\gamma}$. As mentioned above, the simple TVD limiter sometimes allows overshoots by a small amount in this situation. We have not observed any such overshoot with the flux-TVD limiter. See Figure 3.

3.5. **A refinement of the FTVD scheme.** At a steady sonic rarefaction, both the EO scheme and the Godunov scheme are slightly overcompressive, leading to a so-called dogleg feature in the solution. This feature vanishes as the mesh size tends to zero, but it is distracting. The second order scheme above does not correct this behavior. One way to improve the situation is to replace the corrections (3.4) by

$$d^n_{j+1/2} = \frac{1}{2} a^+_{j+1/2}\big(p_{j+1/2} - \lambda a^+_{j+1/2}\big)\Delta_+ U^n_j,$$
$$e^n_{j+1/2} = \frac{1}{2} a^-_{j+1/2}\big(q_{j+1/2} + \lambda a^-_{j+1/2}\big)\Delta_+ U^n_j,$$

where

$$p_{j+1/2} = \frac{a^+_{j+1/2}}{a^+_{j+1/2} - a^-_{j+1/2}}, \quad q_{j+1/2} = \frac{-a^-_{j+1/2}}{a^+_{j+1/2} - a^-_{j+1/2}}.$$

This only changes the scheme near sonic points. The result is that the dogleg feature diminishes noticeably.

4. The nonlocal limiter algorithm

In this section we describe our method for solving the system of inequalities (3.15), (3.16), keeping in mind that we are also trying to maximize the ratio $\hat{F}^n_{j+1/2}/F^n_{j+1/2}$ to maintain formal second-order accuracy wherever possible.

4.1. **Description of the nonlocal limiter algorithm.**

We can simplify the notation somewhat, and also discuss the limiter algorithm more generically, by setting

$$z_i := F^n_{i+1/2}, \quad \theta_i := \left|\Delta_+ h^n_{i+1/2}\right|, \quad \hat{z}_i := \hat{F}^n_{i+1/2},$$

and then restating the system of inequalities (3.15), (3.16) in the form

$$|\hat{z}_{i+1} - \hat{z}_i| \le \theta_i, \tag{4.1}$$

$$0 \le \hat{z}_i / z_i \le 1. \tag{4.2}$$

The unknowns are $\hat{z}_i$, and the data are z_i, $\theta_i \ge 0$. The z_i are assumed to vanish for sufficiently large values of the index i. Specifically, there are indices i_*, i^* such that

$$i \le i_* \Rightarrow z_i = 0, \quad i \ge i^* \Rightarrow z_i = 0.$$

That this assumption is valid for our scheme is evident from the assumption that u_0 has compact support. Even when the parabolic terms are present, the initial data has a finite range of influence (for both the actual and numerical solutions). Thus we may always assume that U^n_j and $F^n_{j+1/2}$ vanish for sufficiently large j.

The nonlocal limiter algorithm.

Preprocessor: For $i = i_*$ increasing to $i = i^* - 1$:
If $z_{i+1} z_i < 0$ and $|z_{i+1} - z_i| > \theta_i$, then

$$z_i \leftarrow \operatorname{sgn}(z_i) \min\bigl\{|z_i|, \theta_i/2\bigr\},$$

$$z_{i+1} \leftarrow \operatorname{sgn}(z_{i+1}) \min\bigl\{|z_{i+1}|, \theta_i/2\bigr\}.$$

Forward sweep: For $i = i_*$ increasing to $i = i^* - 1$:
If $|z_{i+1}| > |z_i|$, then

$$z_{i+1} \leftarrow z_i + \operatorname{sgn}(z_{i+1} - z_i) \min\bigl\{|z_{i+1} - z_i|, \theta_i\bigr\}.$$

Backward sweep: For $i = i^*$ decreasing to $i = i_* + 1$:
If $|z_{i-1}| > |z_i|$, then

$$z_{i-1} \leftarrow z_i + \operatorname{sgn}(z_{i-1} - z_i) \min\bigl\{|z_{i-1} - z_i|, \theta_{i-1}\bigr\}.$$

Here the left arrow $\leftarrow$ is the replacement operator. The algorithm can be written compactly as

$$\hat{Z} = \Phi(Z, \Theta) = \Phi^-\bigl(\Phi^+(\tilde{Z}, \Theta), \Theta\bigr), \quad \tilde{Z} = \mathrm{Pre}(Z, \Theta),$$

where Φ^+ and Φ^- represent the forward and backward sweeps, Pre represents the preprocessor step, and

$$\hat{Z} = \{\hat{z}_i\}, \quad \tilde{Z} = \{\tilde{z}_i\}, \quad Z = \{z_i\}, \quad \Theta = \{\theta_i\}.$$

The operation of the limiter algorithm is best understood by first considering the case where all of the z_i are nonnegative. In this case, the preprocessor step leaves the data z_i unchanged. The forward sweep visits each point z_i in the order of increasing i. If $z_{i-1} \ge z_i$, nothing happens to z_i on the forward sweep. If $z_{i-1} < z_i$, the constraint $|z_i - z_{i-1}| \le \theta_{i-1}$ is checked. Nothing happens to z_i if the constraint is satisfied, but if it is violated, then z_i is moved toward z_{i-1} (decreased) by exactly enough to satisfy the constraint. The points z_{i_*} and z_{i^*} are clamped at

zero, so they never change. On the backward sweep, each point z_i is visited, this time in the order of decreasing i. Nothing happens to z_i if $z_i \le z_{i+1}$. Otherwise the constraint $|z_i - z_{i+1}| \le \theta_i$ is checked, and if there is a violation, then z_i is decreased just enough to satisfy the constraint. The algorithm behaves in an analogous way when all of the z_i are nonpositive. At any contiguous pair of data (z_i, z_{i+1}) where the z_i and z_{i+1} have opposite signs, the effect of preprocessor step is to satisfy the inequalities without changing the sign of either z_i or z_{i+1}. Afterwards, neither the forward sweep nor the backward sweep will cause a constraint violation at that particular pair (z_i, z_{i+1}). Thus the algorithm operates more or less independently on intervals where the z_i do not change sign.

Remark 4.1. The preprocessor part of the algorithm that we have proposed is not the only reasonable way to deal with sign changes in the data $\{z_i\}$. The preprocessor above is simple and and is consistent with second-order accuracy wherever $f_x \neq 0$. In some situations, it is sufficient (and simpler) to set both z_i and z_{i+1} to zero at a sign change. At least in the case where $\boldsymbol{\gamma}$ is piecewise constant, this simpler strategy does not add an additional class of points where formal second-order accuracy is lost.

4.2. **Properties of the nonlocal limiter.**

Lemma 4.1. *The output of the nonlocal limiter algorithm solves the system of inequalities* (4.1), (4.2).

Proof. We use the notation $\hat{Z}, \bar{Z}, \tilde{Z}$ for the outputs of the three portions of the algorithm. It is easy to check by induction on i (increasing i for the preprocessor and the forward sweep, decreasing i for the backward sweep) that

$$0 \le \tilde{z}_i / z_i \le 1, \quad 0 \le \bar{z}_i / \tilde{z}_i \le 1, \quad 0 \le \hat{z}_i / \bar{z}_i \le 1.$$

Combining these three inequalities, we have inequality (4.1), i.e.,

$$0 \le \hat{z}_i / z_i \le 1. \tag{4.3}$$

Next, we claim that as a result of the preprocessor step Pre, wherever there is a sign change in $\tilde{Z}$, the constraint (4.2) is satisfied, i.e., $|\tilde{z}_{i+1} - \tilde{z}_i| \le \theta_i$. Indeed consider the operation of Pre on a pair (z_i, z_{i+1}) where $z_i z_{i+1} < 0$. It is clear that after Pre operates on this pair, (4.2) is satisfied (for this pair). Since Pre moves from left to right, it may or may not also operate on the pair (z_{i+1}, z_{i+2}). If it does not, then the constraint (4.2) obviously remains satisfied for the pair (z_i, z_{i+1}). If it does operate on the pair (z_{i+1}, z_{i+2}), then $|z_{i+1}|$ decreases (or at least does not increase), thus moving z_{i+1} closer to z_i (since they have opposite signs), making it clear that the constraint remains satisfied for the pair (z_i, z_{i+1}). Thus our claim is proved by induction.

We have seen that at any pair $(\tilde{z}_i, \tilde{z}_{i+1})$ where there is a sign change, (4.2) is satisfied. We claim that (4.2) remains satisfied at this pair after both the forward and backward sweeps. To see this, it suffices to observe that neither sweep increases the absolute value of $\tilde{z}_i$ or $\tilde{z}_{i+1}$, and thus $|\tilde{z}_i - \tilde{z}_{i+1}|$ does not increase after either sweep.

From our observations about the effect of the preprocessor, along with the definitions of the forward and backward sweeps, it is clear that

$$|\bar{z}_{i+1}| \ge |\bar{z}_i| \Rightarrow |\bar{z}_{i+1} - \bar{z}_i| \le \theta_i, \tag{4.4}$$

$$|\hat{z}_{i-1}| \ge |\hat{z}_i| \Rightarrow |\hat{z}_{i-1} - \hat{z}_i| \le \theta_{i-1}. \tag{4.5}$$

Now, suppose that $|\hat{z}_{i+1}| \ge |\hat{z}_i|$. It follows from the definition of the backward sweep that $\hat{z}_i = \bar{z}_i$. Then since $|\hat{z}_{i+1}| \le |\bar{z}_{i+1}|$,

$$|\bar{z}_{i+1}| \ge |\bar{z}_i|.$$

By (4.4), $|\bar{z}_{i+1} - \bar{z}_i| \le \theta_i$. and since $\hat{z}_{i+1}$ lies between $\hat{z}_i$ and $\bar{z}_{i+1}$,

$$|\hat{z}_{i+1} - \hat{z}_i| \le \theta_i. \tag{4.6}$$

The proof that $\hat{z}_i$ solves the inequalities is completed by combining (4.3), (4.5), and (4.6). □

Next, we demonstrate that the limiter Φ is consistent with formal second-order accuracy. This consistency property does not rely on the fact that the function u is a solution of a PDE, and so we suppress the dependence on t. For a fixed mesh size $\Delta = \Delta x$, and a smooth function $u(x)$, we define $u_j := u(x_j)$, $\boldsymbol{\gamma}_{j+1/2} := \boldsymbol{\gamma}(x_{j+1/2})$, and

$$h^{\Delta}_{j+1/2} := h(\boldsymbol{\gamma}_{j+1/2}, u_{j+1}, u_j), \quad h^{\Delta} := \left\{h^{\Delta}_{j+1/2}\right\}_{j\in\mathbb{Z}},$$
$$\left|\Delta_+ h^{\Delta}\right| := \left\{\left|\Delta_+ h_{j+1/2}\right|\right\}_{j\in\mathbb{Z}}, \quad F^{\Delta}_{j+1/2} := F_{j+1/2}, \quad F^{\Delta} = \left\{F^{\Delta}_{j+1/2}\right\}_{j\in\mathbb{Z}}.$$

Here the flux $h_{j+1/2}$ and the flux corrections $F_{j+1/2}$ are defined by (3.3), (3.4), (3.5), (3.12), and (3.13). Finally, for $\xi \in \mathbb{R}$ we define $B_r(\xi) := \{x : |x - \xi| < r\}$.

Lemma 4.2. *Let $x \mapsto u(x)$ and $x \mapsto \boldsymbol{\gamma}(x)$ be C^2 in a neighborhood of the point ζ where*

$$f\big(\boldsymbol{\gamma}(\zeta), u(\zeta)\big)_x \neq 0.$$

Assume that $u(\pm x) = u_{\pm\infty}$ for x sufficiently large, so that the limiter Φ is well-defined on the flux corrections $F^{\Delta}_{j+1/2} = F_{j+1/2}$. Let

$$\hat{F}^{\Delta} = \Phi\big(F^{\Delta}, \left|\Delta_+ h^{\Delta}\right|\big).$$

Then there is a mesh size $\Delta_0(\zeta) > 0$ and a $\delta(\zeta) > 0$ such that for $\Delta \le \Delta_0$, we have

$$\hat{F}^{\Delta}_{j+1/2} = F^{\Delta}_{j+1/2} \ \text{ for all } x_j \in B_{\delta}(\zeta).$$

Remark 4.2. The condition $f(\boldsymbol{\gamma}(\zeta), u(\zeta))_x \neq 0$ is analogous to the well-known fact in the constant $\boldsymbol{\gamma}$ setting that a TVD scheme can be at most first-order accurate at a nonsonic extremum.

Proof of Lemma 4.2. Choose $\delta > 0$ and $\varepsilon > 0$ so that $u, \boldsymbol{\gamma} \in C^2(B_{3\delta}(\zeta))$ and for $x \in B_{3\delta}(\zeta)$

$$\left|f\big(\boldsymbol{\gamma}(x), u(x)\big)_x\right| > 2\varepsilon. \tag{4.7}$$

Due to our regularity assumptions concerning the flux $f(\boldsymbol{\gamma}, u)$, and the easily verified fact that both partial derivatives $h_u(\boldsymbol{\gamma}, v, u)$ and $h_v(\boldsymbol{\gamma}, v, u)$ are Lipschitz continuous with respect to all of $u, v, \boldsymbol{\gamma}$, it is a straightforward exercise to show that for $x \in B_{3\delta}(\zeta)$

$$\Delta_+ h^{\Delta}_{j+1/2} = f(\boldsymbol{\gamma}(x_j), u(x_j))_x \Delta + \mathcal{O}\left(\Delta^2\right). \tag{4.8}$$

Next, we claim that it is possible to choose $\Delta_0 > 0$ such that the following conditions hold for $\Delta < \Delta_0$ and $x_j \in B_{2\delta}(\zeta)$:

$$\left|\Delta_+ F^{\Delta}_{j+1/2}\right| \le \left|\Delta_+ h^{\Delta}_{j+1/2}\right|, \tag{4.9}$$

$$\left|F^{\Delta}_{j+1/2}\right| < \varepsilon\delta/2, \tag{4.10}$$

$$\left|h^{\Delta}_{j+1/2}\right| > \Delta\varepsilon. \tag{4.11}$$

To verify (4.9), note that because of (4.8) and (4.7), the right side of (4.9) is $\mathcal{O}(\Delta)$. At the same time, the left side is $\mathcal{O}\left(\Delta^2\right)$. For (4.10), the left side is $\mathcal{O}(\Delta)$, while the right side is fixed (with respect to Δ). Finally, by combining the assumption $|f(\boldsymbol{\gamma}(x), u(x))_x| > 2\varepsilon$ and (4.8), it is clear that we will have (4.11) for sufficiently small $\Delta > 0$.

Let $x^- := \zeta - \delta$ and $x^+ := \zeta + \delta$. The immediate objective is to prove that for $\Delta \le \Delta_0$,

$$\bar{F}^{\Delta}_{j+1/2} = F^{\Delta}_{j+1/2}, \quad \forall x_j \in (x^-, \zeta + 2\delta), \tag{4.12}$$

where $\bar{F}^{\Delta}_{j+1/2}$ is the output of the forward sweep of the limiter Φ.

By way of contradiction, suppose that (4.12) fails, and choose $x_J \in (x^-, \zeta + 2\delta)$ and $\Delta_1 \le \Delta_0$ such that

$$\bar{F}^{\Delta_1}_{J+1/2} \ne F^{\Delta_1}_{J+1/2}.$$

Because of assumption (4.9), it must be that the preprocessor has not modified any of the flux corrections $F^{\Delta}_{j+1/2}$ with $x_j \in [x^- - \delta, x_J)$, and that the forward pass has modified all of them. For the forward pass to have modified them all, it must be that $F^{\Delta}_{j-1/2} F^{\Delta}_{j+1/2} \ge 0$ for $x_j \in [x^- - \delta, x_J)$. Without loss of generality, assume that $F^{\Delta}_{j+1/2} \ge 0$ in the area of interest. Since the forward pass modified all of the $F^{\Delta}_{j+1/2}$ for $x_j \in [x^- - \delta, x_J)$, we have

$$\left|\Delta_+ \bar{F}^{\Delta_1}_{j+1/2}\right| = \left|\Delta_+ h^{\Delta}_{j+1/2}\right| \quad \text{for } x^- - \delta \le x_j < x_J. \tag{4.13}$$

Since all of the flux corrections in the area of interest have been modified by the forward pass of the limiter,

$$\bar{F}^{\Delta_1}_{j-1/2} \le \bar{F}^{\Delta_1}_{j+1/2} \tag{4.14}$$

for $x^- - \delta \le x_j < x_J$. Let $P := \max\{p \in \mathbb{Z}^+ : p \le \delta/\Delta_1\}$.

Summing (4.13) over p, we get telescoping due to (4.14), and find that

$$\bar{F}^{\Delta_1}_{J+1/2} - \bar{F}^{\Delta_1}_{J+1/2-P} = \sum_{p=0}^{P-1} \left|\Delta_+ h_{J+1/2-p-1}\right| \ge P\varepsilon\Delta_1 \ge \delta\varepsilon.$$

On the other hand, it follows from (4.10) and (4.14) that

$$\bar{F}^{\Delta_1}_{J+1/2} - \bar{F}^{\Delta_1}_{J+1/2-P} \le \bar{F}^{\Delta_1}_{J+1/2} = \left|\bar{F}^{\Delta_1}_{J+1/2}\right| < \delta\varepsilon/2,$$

which gives the desired contradiction.

A symmetric argument applied to the backward sweep of the nonlocal limiter algorithm, operating on $\{\bar{F}^{\Delta}_{j+1/2}\}$, proves that for some $0 < \tilde{\Delta}_0 \le \Delta_0$,

$$\hat{F}^{\Delta}_{j,k} = F^{\Delta}_{j,k}, \quad \forall x_j \in (x^-, x^+),$$

for $\Delta \le \tilde{\Delta}_0$. Replacing Δ_0 by $\tilde{\Delta}_0$ completes the proof. □

5. Convergence of the second-order scheme

In this section we analyze the flux-TVD scheme

$$U^{n+1}_j = U^n_j - \lambda\Delta_-\big(h^n_{j+1/2} + \hat{F}^n_{j+1/2}\big). \tag{5.1}$$

We assume the nonlocal limiter has been applied to the flux corrections $F^n_{j+1/2}$, i.e., we are focusing on the FTVD algorithm. We analyzed the first-order version of this scheme,

$$U^{n+1}_j = U^n_j - \lambda\Delta_- h^n_{j+1/2}, \tag{5.2}$$

that results by deleting the second-order corrections in [6], where $\boldsymbol{\gamma}$ was piecewise constant, and [5], where we dealt with the more general case of piecewise smooth $\boldsymbol{\gamma}$. Wherever possible in the analysis that follows, we will rely on results from our analysis in [5] and [6]. In this section we will assume that the following CFL condition is satisfied:

$$\lambda\Big(\max\{-q_{\mathrm{L}}, q_{\mathrm{R}}\} + \|\gamma_1 b'\|\Big) \le \frac{1}{4}, \tag{5.3}$$

where we define

$$\|\gamma_1 b'\| := \max\big\{|\gamma_1(x)b'(u)| : x \in [x_{\mathrm{L}}, x_{\mathrm{R}}],\, u \in [0,1]\big\}.$$

In [6], we imposed essentially the same CFL condition, but with $1/2$ on the right-hand side. The halving of the allowable time step implied by this new CFL condition (5.3) is required to prove Lemma 5.1 guaranteeing that the computed solutions remain in the interval $[0,1]$. This halving of the time step to achieve a bound on the solution is also common when designing second-order TVD schemes for the case of constant $\boldsymbol{\gamma}$. In practice one finds that it is often not necessary to impose the reduced time step.

Our theorem concerning convergence is the following.

Theorem 5.1 (Convergence of the FTVD scheme)**.** *Let u^Δ be defined by* (3.1), (3.2), (3.3), (3.4), (3.5), (3.12), (3.13). *Assume that the flux corrections $\hat{F}^n_{j+1/2}$ are produced by applying the limiter algorithm of Section* 4 *to the flux corrections $F^n_{j+1/2}$. Let $\Delta \to 0$ with λ constant and the CFL condition* (5.3) *satisfied. Then u^Δ converges along a subsequence in $L^1_{\mathrm{loc}}(\Pi_T)$ and boundedly a.e. in Π_T to a BV_t weak solution of the CT model* (1.2).

The proof of Theorem 5.1 amounts to checking that Lemmas 5.1 through 5.7, along with the relevant portion of Theorem 5.1, of [5] remain valid in the present context. We start with two lemmas that replace Lemma 5.1 of [5].

Lemma 5.1. *Under the CFL condition* (5.3) *we get a uniform bound on U^n_j, specifically $U^n_j \in [0,1]$.*

Proof. Let V^n_j denote the result of applying the *first-order* version of the scheme to U^n, with the time step doubled, i.e.,

$$V^{n+1}_j = U^n_j - 2\lambda\Delta_- h^n_{j+1/2}. \tag{5.4}$$

The proof Lemma 5.1 of [5], or the proof of Lemma 3.1 of [6], gives us $0 \le V^n_j \le 1$, assuming that we impose the more restrictive CFL condition (5.3) to account for doubling the time step. Now let U^{n+1}_j be the result of applying our second-order scheme

$$U^{n+1}_j = U^n_j - \lambda\Delta_-\big(h^n_{j+1/2} + \hat{F}^n_{j+1/2}\big). \tag{5.5}$$

Comparing (5.4) and (5.5), we find after some algebra that the following relationship holds:

$$\frac{U^{n+1}_j - U^n_j}{V^{n+1}_j - U^n_j} = \frac{1}{2}\left[1 + \frac{\Delta_+\hat{F}^n_{j-1/2}}{\Delta_+ h^n_{j-1/2}}\right].$$

Because of the conditions (3.15), (3.16) enforced on $\hat{F}^n_{j-1/2}$ by the flux-TVD limiter we find that

$$0 \le \frac{U^{n+1}_j - U^n_j}{V^{n+1}_j - U^n_j} \le 1,$$

and from this relationship (along with $V^n_j \in [0,1]$) it follows that $0 \le U^n_j \le 1$. □

In Section 3, we stated that the flux limiter (3.14) was designed to enforce a TVD condition on the first-order numerical flux $h^n_{j+1/2}$. The following lemma shows that our limiter performs as advertised.

Lemma 5.2. *The flux-TVD property is satisfied, i.e.,*

$$\sum_{j\in\mathbb{Z}}\big|h^{n+1}_{j+1/2} - h^{n+1}_{j-1/2}\big| \le \sum_{j\in\mathbb{Z}}\big|h^n_{j+1/2} - h^n_{j-1/2}\big|, \quad n = 0,1,2,\ldots.$$

Proof. We start from the relationship

$$h_{j+1/2}^{n+1} = h_{j+1/2}^{n} - \eta_{j+1}^{n+1/2}\big(U_{j+1}^{n+1} - U_{j+1}^{n}\big) + \zeta_{j}^{n+1/2}\big(U_{j}^{n+1} - U_{j}^{n}\big),$$

where

$$\begin{aligned} \eta_{j+1}^{n+1/2} &:= -\int_0^1 h_v\left(\boldsymbol{\gamma}_{j+1/2}, U_{j+1}^{n} + \theta\left(U_{j+1}^{n+1} - U_{j+1}^{n}\right)\right) d\theta \geq 0, \\ \zeta_{j}^{n+1/2} &:= \int_0^1 h_u\left(\boldsymbol{\gamma}_{j+1/2}, U_{j}^{n} + \theta\left(U_{j}^{n+1} - U_{j}^{n}\right)\right) d\theta \geq 0. \end{aligned} \tag{5.6}$$

Now using the definition (3.2) of the scheme to substitute for $U_{j+1}^{n+1} - U_{j+1}^{n}$ and $U_{j}^{n+1} - U_{j}^{n}$, we get (after some algebra)

$$h_{j+1/2}^{n+1} = h_{j+1/2}^{n} + P_{j+1/2}^{n}\Delta_+ h_{j+1/2}^{n} - Q_{j-1/2}^{n}\Delta_- h_{j+1/2}^{n},$$

where

$$P_{j+1/2}^{n} = \lambda\eta_{j+1}^{n+1/2}\left[1 + \frac{\Delta_+\hat{F}_{j+1/2}^{n}}{\Delta_+ h_{j+1/2}^{n}}\right], \quad Q_{j-1/2}^{n} = \lambda\zeta_{j}^{n+1/2}\left[1 + \frac{\Delta_+\hat{F}_{j-1/2}^{n}}{\Delta_+ h_{j-1/2}^{n}}\right].$$

By Harten's lemma [16], we will have the flux-TVD property if

$$P_{j+1/2}^{n} \geq 0, \quad Q_{j-1/2}^{n} \geq 0, \qquad P_{j-1/2}^{n} + Q_{j-1/2}^{n} \leq 1. \tag{5.7}$$

In more detail, the second condition in (5.7) is

$$\lambda\eta_{j}^{n+1/2}\left[1 + \frac{\Delta_+\hat{F}_{j-1/2}^{n}}{\Delta_+ h_{j-1/2}^{n}}\right] + \lambda\zeta_{j}^{n+1/2}\left[1 + \frac{\Delta_+\hat{F}_{j-1/2}^{n}}{\Delta_+ h_{j-1/2}^{n}}\right] \leq 1. \tag{5.8}$$

From (5.6) and the CFL condition (5.3), we obtain

$$\begin{aligned} 0 &\leq \lambda\eta_{j}^{n+1/2} + \lambda\zeta_{j}^{n+1/2} \\ &\leq \int_0^1 \left|f_u\big(\boldsymbol{\gamma}_{j-1/2}, U_j^n + \theta(U_j^{n+1} - U_j^n)\big)\right| d\theta \\ &\quad + \int_0^1 \left|f_u\big(\boldsymbol{\gamma}_{j+1/2}, U_j^n + \theta(U_j^{n+1} - U_j^n)\big)\right| d\theta \leq \frac{1}{2}. \end{aligned} \tag{5.9}$$

It is immediate from (3.15) that

$$0 \leq 1 + \frac{\Delta_+\hat{F}_{j+1/2}^{n}}{\Delta_+ h_{j+1/2}^{n}} \leq 2. \tag{5.10}$$

Combining (5.8), (5.9) and (5.10), we see that both conditions in (5.7) are satisfied. □

For our first-order scheme (5.2), we derived a discrete time continuity estimate (Lemma 5.1 of [5]) using the fact that the scheme was both conservative and monotone. In the process of making the scheme second-order accurate, we have sacrificed the monotonicity property, and so the proof of time continuity requires a different approach. The flux-TVD property is the ingredient that allows us to maintain time continuity in the absence of monotonicity.

Lemma 5.3. *There exists a constant C, independent of Δ and n, such that*

$$\Delta x \sum_{j\in\mathbb{Z}} \left|U_j^{n+1} - U_j^{n}\right| \leq \Delta x \sum_{j\in\mathbb{Z}} \left|U_j^{1} - U_j^{0}\right| \leq C\Delta t.$$

Proof. Starting from the marching formula (5.1), we take absolute values, apply the triangle inequality, and then sum over j. This yields

$$\sum_{j\in\mathbb{Z}} \big|U_j^{n+1} - U_j^n\big| \le \lambda \sum_{j\in\mathbb{Z}} \big|\Delta_- h_{j+1/2}^n\big| + \lambda \sum_{j\in\mathbb{Z}} \big|\hat{F}_{j+1/2}^n\big|. \tag{5.11}$$

By the flux-TVD property, the first of these sums satisfies

$$\lambda \sum_{j\in\mathbb{Z}} \big|\Delta_- h_{j+1/2}^n\big| \le \lambda \sum_{j\in\mathbb{Z}} \big|\Delta_- h_{j+1/2}^0\big|.$$

Referring to (3.15), we see that also

$$\lambda \sum_{j\in\mathbb{Z}} \big|\hat{F}_{j+1/2}^n\big| \le \lambda \sum_{j\in\mathbb{Z}} \big|\Delta_- h_{j+1/2}^0\big|.$$

Proceeding as in Lemma 5.1 of [5] we can show that

$$\sum_{j\in\mathbb{Z}} \big|\Delta_- h_{j+1/2}^0\big| = \mathcal{O}(1),$$

and thus

$$\sum_{j\in\mathbb{Z}} \big|U_j^{n+1} - U_j^n\big| = \mathcal{O}(1).$$

Multiplying both sides of this estimate by Δx completes the proof. □

To continue with our analysis, we introduce the the so-called singular mappingΨ, defined by

$$\Psi(\boldsymbol{\gamma}, u) := \int_0^u |f_u(\boldsymbol{\gamma}, w)|\, dw,$$

and let

$$z^\Delta(x,t) := \Psi\big(\boldsymbol{\gamma}(x), u^\Delta(x,t)\big).$$

As in [5], to prove that the difference scheme converges, we establish compactness for the transformed quantity z^Δ, the critical ingredient being a bound on its total variation. We then derive compactness for u^Δ by appealing to the monotonicity and continuity of the mapping $u \mapsto \Psi(\boldsymbol{\gamma}, u)$.

Thus our goal now is to show that z^Δ has bounded variation. For this it suffices to invoke Lemmas 2 through 7 of [5], making modifications where necessary to account for the addition of the second-order correction terms. In what follows, we use the notation Δ_+^u and Δ_-^u for spatial difference operators with respect to u only, keeping γ fixed, e.g.,

$$\Delta_+^u f\big(\gamma_j, U_j^n\big) = f\big(\gamma_j, U_{j+1}^n\big) - f\big(\gamma_j, U_j^n\big).$$

Also, we use the notation $\mathcal{O}\left(\Delta\gamma_j\right)$ to mean terms which sum (over j) to $\mathcal{O}\left(|\gamma|_{BV}\right)$. Finally, we will use the Kružkov entropy-entropy flux pair indexed by c:

$$q(u) := |u-c|, \quad \eta(\gamma, u) := \operatorname{sgn}(u-c)\big(f(\gamma,u) - f(\gamma,c)\big),$$

where $\operatorname{sgn}(w) = w/|w|$ if $w \neq 0$ and $\operatorname{sgn}(0) = 0$.

The following is basically Lemma 5.2 of [5], modified to accommodate the second-order correction terms.

Lemma 5.4. *For each $c \in \mathbb{R}$, the following inequality holds:*

$$\begin{aligned} q\big(U_j^{n+1}\big) \le{}& q\big(U_j^n\big) - \lambda \Delta_-^u H\big(\gamma_{j+1/2}, U_{j+1}^n, U_j^n\big) \\ &+ \lambda\big|\Delta_+ h_{j-1/2}^n\big| + \lambda\mathcal{O}\left(\Delta\gamma_j\right), \quad j \in \mathbb{Z}, \quad n = 0, 1, 2 \ldots, \end{aligned} \tag{5.12}$$

where the EO numerical entropy flux is given by

$$H(\gamma, v, u) = \frac{1}{2}\big(\eta(\gamma, u) + \eta(\gamma, v)\big) - \frac{1}{2}\int_u^v \operatorname{sgn}(w - c)\big|f_u(\gamma, w)\big|\, dw.$$

Proof. Let $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$. With

$$\rho_j^{n+1} = U_j^n - \lambda \Delta_-^u h\left(\gamma_{j+1/2}, U_{j+1}^n, U_j^n\right),$$

the following discrete entropy inequality holds:

$$q\left(\rho_j^{n+1}\right) \le q\left(U_j^n\right) - \lambda \Delta_-^u H\left(\gamma_{j+1/2}, U_{j+1}^n, U_j^n\right),$$

since H can be written in the form

$$H(\gamma, v, u) = h(\gamma, v \vee c, u \vee c) - h(\gamma, v \wedge c, u \wedge c).$$

Then we obtain the inequality

$$q(U_j^{n+1}) \le q(U_j^n) - \lambda \Delta_-^u H\left(\gamma_{j+1/2}, U_{j+1}^n, U_j^n\right) - q\left(\rho_j^{n+1}\right) + q(U_j^{n+1}).$$

It remains to show that $q(\rho_j^{n+1}) - q(U_j^{n+1}) = \lambda \mathcal{O}\left(\Delta\gamma_j\right) + \lambda\big|\Delta_+ h_{j-1/2}^n\big|$:

$$\begin{aligned}
\big|q(\rho_j^{n+1}) - q(U_j^{n+1})\big| &\le \big|\rho_j^{n+1} - U_j^{n+1}\big| \\
&= \lambda\big|\Delta_- h(\gamma_{j+1/2}, U_{j+1}^n, U_j^n) \\
&\quad - \Delta_-^u h(\gamma_{j+1/2}, U_{j+1}^n, U_j^n) + \Delta_- \hat{F}_{j+1/2}^n\big| \\
&\le \lambda(2\|f_\gamma\| + L_{u\gamma})\left|\gamma_{j+1/2} - \gamma_{j-1/2}\right| + \lambda\big|\Delta_- h_{j+1/2}^n\big| \\
&= \lambda\mathcal{O}\left(\Delta\gamma_j\right) + \lambda\big|\Delta_- h_{j+1/2}^n\big|,
\end{aligned}$$

where $L_{u\gamma}$ denotes the Lipschitz constant of f_u. Here we have used the proof of Lemma 5.2, which ensures that inequality (3.15) holds. □

It is now possible to repeat the proofs of Lemmas 3 through 7 of [5], the only change being the contribution of the term $\lambda|\Delta_- h_{j+1/2}^n|$ appearing in (5.12). In order for the proofs of those lemmas to remain valid, we must have

$$\sum_{j\in\mathbb{Z}} \big|\Delta_- h_{j+1/2}^n\big| = \mathcal{O}\left(1\right),$$

independently of n and Δ. But this follows directly than our flux-TVD property, which we established in Lemma 5.2, along with the relationship

$$\sum_{j\in\mathbb{Z}} \big|\Delta_- h_{j+1/2}^0\big| = \mathcal{O}\left(1\right),$$

which we established in the proof of Lemma 5.3.

6. Numerical results

6.1. **Examples 1 and 2: ideal suspension in a cylindrical unit.** Consider a suspension characterized by the function $b(u)$ given by (2.3) with $v_\infty = 1.0 \times 10^{-4}$m/s, $C = 5$ and $u_{\max} = 1$ (as in [7]). In this example, we assume that the effect of sediment compressibility is absent ($A \equiv 0$). In Examples 1 and 2, we consider a cylindrical CT with $x_{\mathrm{L}} = -1\,\mathrm{m}$ and $x_{\mathrm{R}} = 1\,\mathrm{m}$ with (nominal) interior cross-sectional area $S = 1\,\mathrm{m}$. This vessel is assumed to initially contain no solids ($u_0 \equiv 0$), is operated with a feed suspension of concentration $u_{\mathrm{F}} = 0.3$ in Example 1 and $u_{\mathrm{F}} = 0.5$ in Example 2, and the relevant flow velocities are $q_{\mathrm{L}} = Q_{\mathrm{L}}/S = -1.0 \times 10^{-5}\,\mathrm{m/s}$ and $q_{\mathrm{R}} = Q_{\mathrm{R}}/S = 2.5 \times 10^{-6}\,\mathrm{m/s}$. Note that in Examples 1 and 2 it is not necessary to distinguish between the ξ and x variables.

Figures 4 and 5 show the numerical solution of the continuous fill-up of the CT calculated by the first-order scheme described in [7] (BKT), the scheme described herein that uses the simple TVD limiter (TVD1) described in Section 3.3, and the

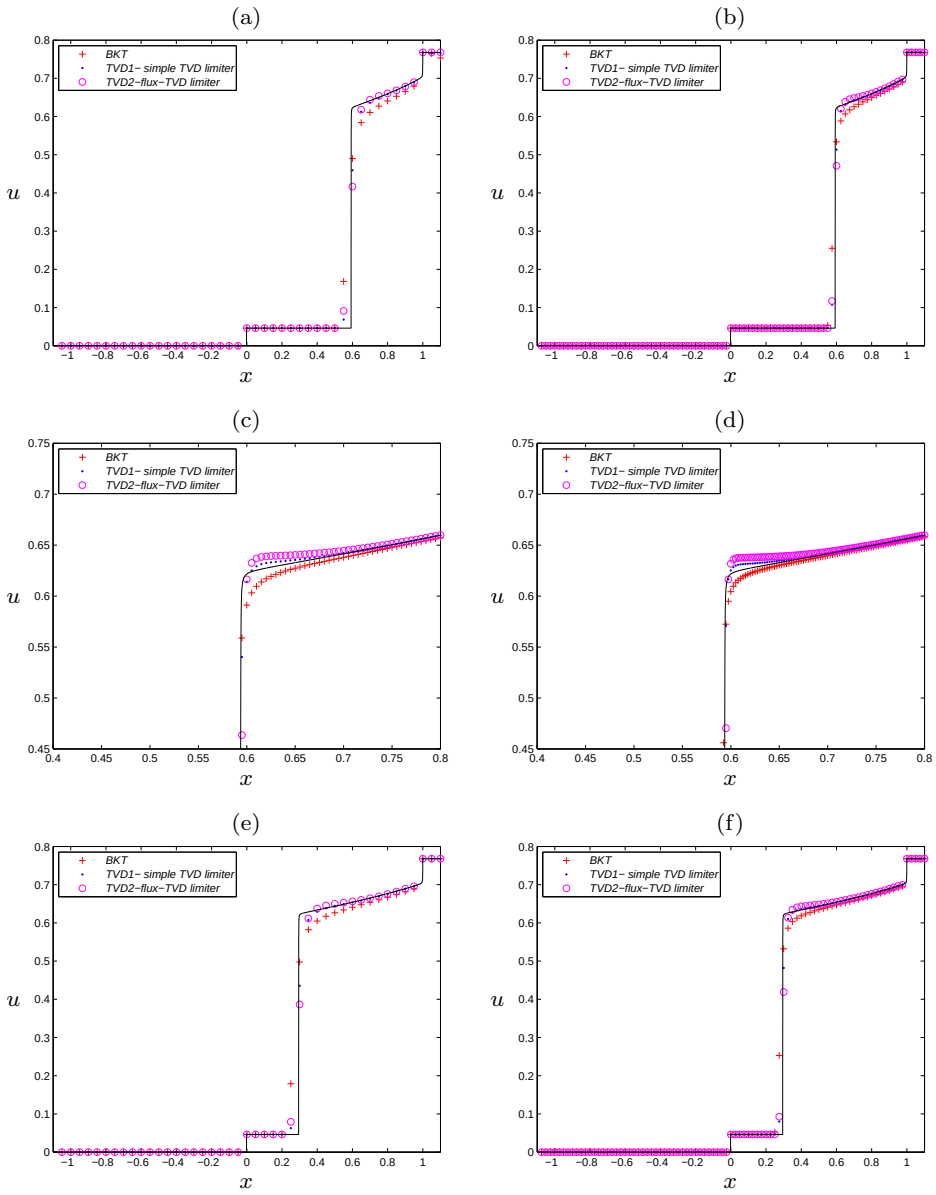

FIGURE 4. Example 1: numerical solution at (a)–(d) $t = 150000\,\mathrm{s}$ with (a) $J = 20$, (b) $J = 40$, (c) $J = 200$ (enlarged view around $x = 0.6$), (d) $J = 400$ (enlarged view around $x = 0.6$), and at (e, f) $t = 250000\,\mathrm{s}$ with (e) $J = 20$ and (f) $J = 40$. The solid line is the reference solution.

scheme that involves the non-local limiter (TVD2), which is outlined in Sections 3.4 and 3.5. All calculations were performed with $\lambda = 2000\,\mathrm{s/m}$, and errors were compared against a reference solution calculated by the first-order scheme presented in [2] with $J = 10000$, where $J = 1/\Delta x$ (in meters). Table 1 shows approximate L^1 errors (errors measured over the finite interval $[-1.1, 1.1]$).

Example 2 has been designed to illustrate the effect of the overshoot. Figure 6 shows the numerical solution at $t = 272760\,\mathrm{s}$ for three different spatial discretization. The time has been chosen such that the "overshoot" mentioned in Section 3.3

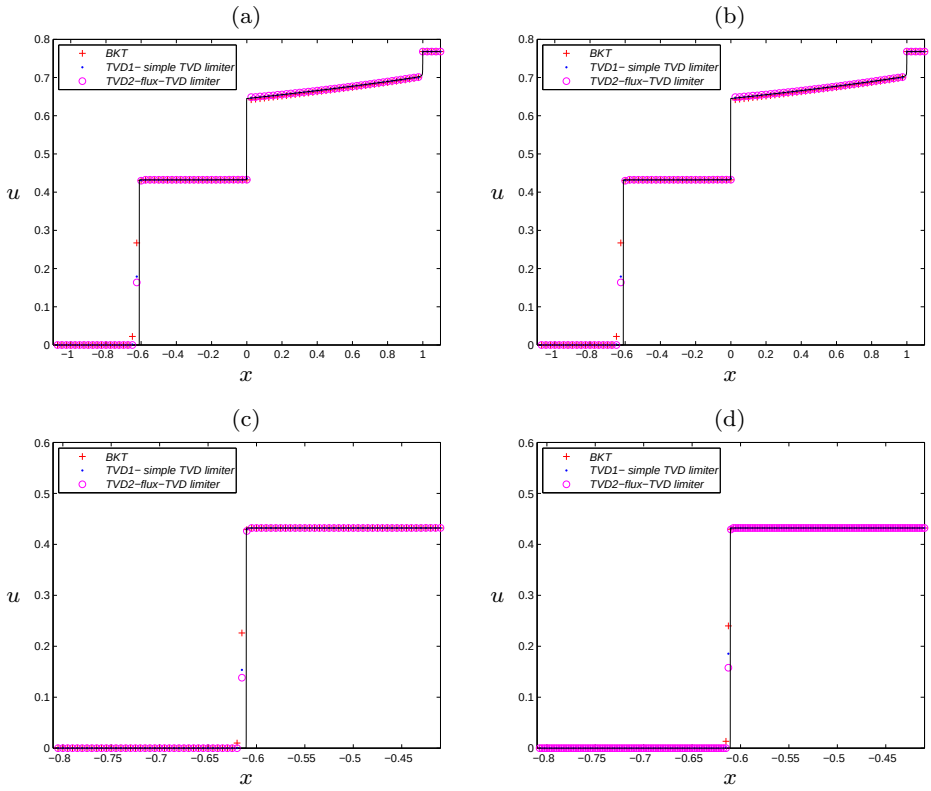

FIGURE 5. Example 1: numerical solution at $t = 500000\,\mathrm{s}$ with (a) $J = 20$, (b) $J = 40$, (c) $J = 200$ (enlarged view around $x = \dots$), (d) $J = 400$ (enlarged view around $x = -0.61$). The solid line is the reference solution.

and shown in Figure 2 becomes visible. As the enlarged views, Figures 6 (b) to (d) illustrate, this phenomenon diminishes as $\Delta x \to 0$.

6.2. **Example 3: ideal suspension in a unit with varying cross-sectional area.** In Example 3 we consider a vessel whose non-constant cross-sectional area is given by

$$S(\xi) = \begin{cases} 0.04\,\mathrm{m}^2 & \text{for } \xi < \xi_\mathrm{R} := -1\,\mathrm{m}, \\ 1\,\mathrm{m}^2 & \text{for } \xi_\mathrm{L} \le \xi < -0.5\,\mathrm{m}, \\ 0.75\,\mathrm{m}^2 & \text{for } -0.5\,\mathrm{m} \le \xi < 0\,\mathrm{m}, \\ 1\,\mathrm{m}^2 & \text{for } 0\,\mathrm{m} \le \xi < 0.5\,\mathrm{m}, \\ (\alpha + \beta\xi)^2 & \text{for } 0.5\,\mathrm{m} \le \xi \le \xi_\mathrm{R} := 1\,\mathrm{m}, \\ S_1 & \text{for } \xi > \xi_\mathrm{R}, \end{cases}$$

where we define the parameters

$$\alpha := \frac{5 - \sqrt{2}}{2}\,\mathrm{m}, \quad \beta := \sqrt{3} - 3, \quad S_1 = \frac{(\sqrt{3} - 1)^2}{4}\,\mathrm{m}^2.$$

In Example 3, we consider the same model functions as in Example 1; in particular, we assume that there is no sediment compressibility. We assume that $u_0 \equiv 0$, and that the vessel is filled up with feed suspension of concentration $u_\mathrm{F} = 0.5$. The volume bulk flows are $Q_\mathrm{L} = -1.0 \times 10^{-5}\,\mathrm{m}^3/\mathrm{s}$ and $Q_\mathrm{R} = 2.5 \times 10^{-6}\,\mathrm{m}^3/\mathrm{s}$. Figure 7

$J = \frac{1}{\Delta x}$	$t = 150000\,\mathrm{s}$		$t = 250000\,\mathrm{s}$		$t = 500000\,\mathrm{s}$	
	approx. L^1 error	conv. rate	approx. L^1 error	conv. rate	approx. L^1 error	conv. rate
	First-order scheme BKT					
10	5.43e-2		5.77e-2		5.20e-2	
20	2.96e-2	0.875	3.25e-2	0.830	2.78e-2	0.903
40	1.67e-2	0.823	1.85e-2	0.810	1.55e-2	0.845
100	8.11e-3	0.790	8.84e-3	0.808	6.76e-3	0.905
200	4.42e-3	0.877	4.83e-3	0.870	3.61e-3	0.906
400	2.31e-3	0.932	2.51e-3	0.946	1.82e-3	0.984
	Simple TVD scheme TVD1					
10	3.93e-2		3.89e-2		3.71e-2	
20	1.85e-2	1.090	1.86e-2	1.065	1.87e-2	0.989
40	8.85e-3	1.060	9.12e-3	1.028	1.01e-2	0.884
100	3.97e-3	0.875	3.85e-3	0.943	4.46e-3	0.894
200	1.94e-3	1.033	2.23e-3	0.787	2.42e-3	0.883
400	1.03e-3	0.917	1.14e-3	0.964	1.24e-3	0.969
	Nonlocal TVD scheme TVD2					
10	4.02e-2		3.92e-2		3.88e-2	
20	1.96e-2	1.039	2.04e-2	0.945	1.93e-2	1.005
40	9.98e-3	0.970	1.09e-2	0.902	1.00e-2	0.945
100	4.37e-3	0.902	4.87e-3	0.878	4.58e-3	0.857
200	2.56e-3	0.774	2.98e-3	0.712	2.42e-3	0.918
400	1.58e-3	0.691	2.14e-3	0.474	1.21e-3	1.002

TABLE 1. Example 1: approximate L^1 errors.

shows the numerical solution for this case at three selected times obtained by the BKT, TVD1 and TVD2 schemes. Note that the equi-distant spatial discretization $\Delta x = 1\,\mathrm{m}^3/J$ corresponds to the x variable obtained from (2.9), while the numerical results shown in Figures 7 and 8 are referred to the original (physical) ξ variable, and therefore are non-equidistant.

6.3. **Observations and conclusions.** A general observation visible in both all test cases is that the newly introduced schemes, TVD1 and TVD2, are significantly more accurate than their first-order counterpart, the first-order BKT scheme introduced in [6]. Clearly, due to the appearance of discontinuities in the solution, the measured order of convergence for these schemes is lower than the theoretically possible value of two. It seems that both TVD schemes, TVD1 and TVD2, have comparable accuracy.

7. A note on second-order degenerate parabolic equations

7.1. **Operator splitting and Crank-Nicolson scheme.** For the more complete model (1.4) that includes a degenerate diffusion term we propose a Strang-type operator splitting scheme. To describe it, let U^n denote the approximate solution at time level n, and we describe the marching algorithm (3.2) in operator notation via $U^{n+1} = \mathcal{H}(\Delta t)U^n$. Then the proposed operator splitting scheme for (1.4) is

$$U^{n+1} = \big[\mathcal{H}(\Delta t/2) \circ \mathcal{P}(\Delta t) \circ \mathcal{H}(\Delta t/2)\big]U^n, \quad n = 0, 1, 2, \ldots.$$

Here $\mathcal{P}(\Delta t)$ represents a second-order scheme for the purely diffusive problem $u_t = (\gamma_1(x)A(u)_x)_x$ written as $U^{n+1} = \mathcal{P}(\Delta t)U^n$.

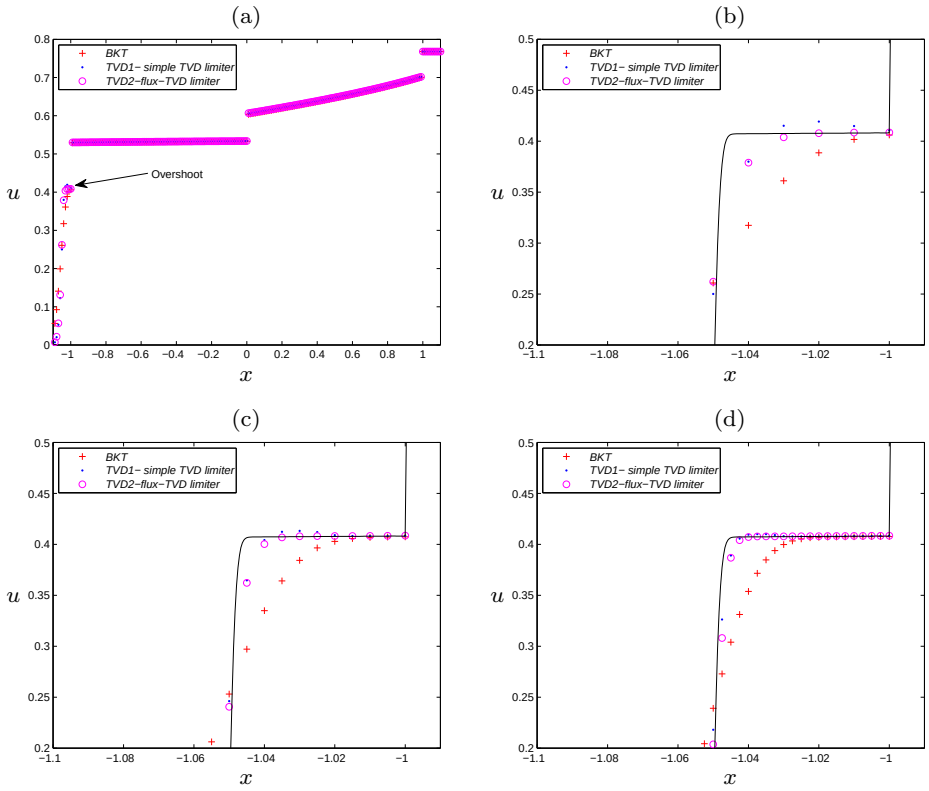

FIGURE 6. Example 2: numerical solution at $t = 272760\,\mathrm{s}$ (a) with $J = 100$ and (b, c, d) enlarged views around $x = -1$ for (b) $J = 100$, (c) $J = 200$ and (d) $J = 400$. The solid line is the reference solution.

For the parabolic portion of the scheme, we can use the Crank-Nicolson scheme, which has second-order accuracy in both space and time. Specifically, the operator $\mathcal{P}(\Delta t)$ is defined by

$$U_j^{n+1} = U_j^n + \frac{\mu}{2}\left[\Delta_+\left(s_{j-1/2}\Delta_- A_j^n\right) + \Delta_+\left(s_{j-1/2}\Delta_- A_j^{n+1}\right)\right], \quad \mu = \frac{\Delta t}{\Delta x^2}. \tag{7.1}$$

Here $s_{j-1/2}$ denotes our discretization of the parameter $\gamma_1(x)$.

The Crank-Nicolson scheme is stable with linear stability analysis. For our nonlinear problem, we generally need a very strong type of stability, both from a practical and theoretical point of view. It seems that it is impossible to get this type of strong stability for implicit schemes of accuracy greater than one [15]. On the other hand, we know from [7] that the solution u is continuous in the regions where the parabolic operator is in effect, and thus we may not require such strong stability in order to keep the numerical approximation well-behaved.

We briefly describe the implementation of the Crank-Nicolson scheme. To simplify the notation, we write $U_j = U_j^n$, $V_j = U_j^{n+1}$. We start by writing the single step of (7.1) in the form

$$V_j = U_j + \frac{1}{2}\mu\Delta_+\big(s_{j-1/2}\Delta_- A(U_j)\big) + \frac{1}{2}\mu\Delta_+\big(s_{j-1/2}\Delta_- A(V_j)\big).$$

We can rewrite this nonlinear system of equations as

$$\mathcal{E}_j(V)V_{j-1} + \mathcal{F}_j(V)V_j + \mathcal{G}_j(V)V_{j+1} = \mathcal{R}_j, \quad V := \{V_j\}_{j\in\mathbb{Z}}, \tag{7.2}$$

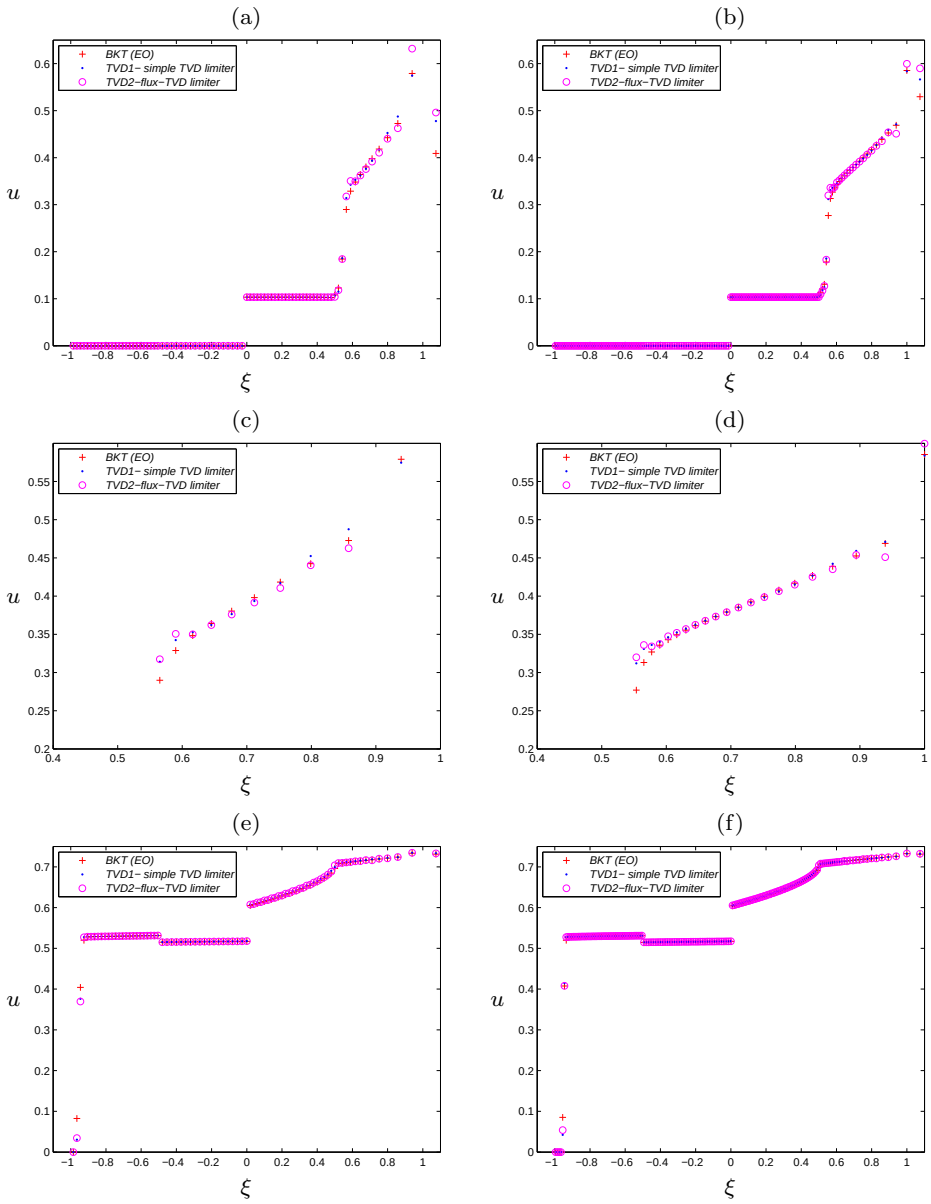

FIGURE 7. Example 3: numerical solution at (a)–(d) $t = 25000\,\mathrm{s}$ with (a) $J = 50$, (b) $J = 100$, (c, d) enlarged views around $\xi = 0.6$ for (c) $J = 50$ and (d) $J = 100$; (e), (f) numerical solution at $t = 200000\,\mathrm{s}$ with (e) $J = 50$, (f) $J = 100$.

where

$$\mathcal{E}_j(V) := \begin{cases} -\dfrac{1}{2}\mu s_{j-1/2}\dfrac{\Delta_- A(V_j)}{\Delta_- V_j} & \text{if } \Delta_- V_j \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathcal{G}_j(V) := \begin{cases} -\dfrac{1}{2}\mu s_{j+1/2}\dfrac{\Delta_+ A(V_j)}{\Delta_+ V_j} & \text{if } \Delta_+ V_j \neq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad \mathcal{F}_j(V) = 1 - \mathcal{E}_j(V) - \mathcal{G}_j(V),$$

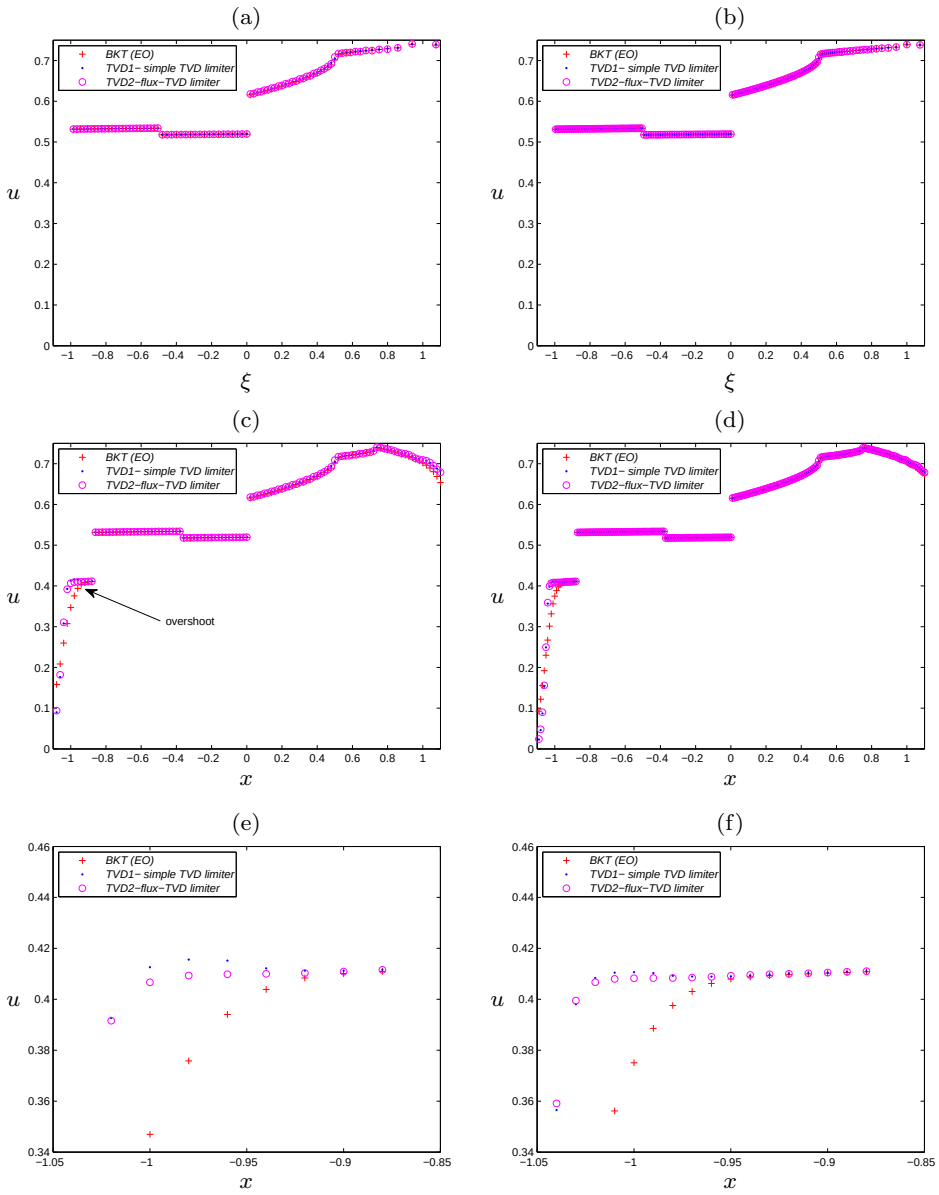

FIGURE 8. Example 3: numerical solution at $t = 225000\,\mathrm{s}$ with (a) $J = 50$, (b) $J = 100$, (c, d) the solutions of (a) and (b), respectively, referred to x instead of ξ, (e, f) enlarged views around $x = -1$ for (e) $J = 50$ and (f) $J = 100$.

and the right-hand side $\mathcal{R}_j$ is defined by

$$\mathcal{R}_j = U_j + \frac{1}{2}\mu\Delta_+\big(s_{j-1/2}\Delta_- A(U_j)\big).$$

To solve the nonlinear system (7.2), we set $V^0 = U$, and proceed via iteration, at each step solving the tridiagonal linear system

$$\mathcal{E}_j(V^k)V_{j-1}^{k+1} + \mathcal{F}_j(V^k)V_j^{k+1} + \mathcal{G}_j(V^k)V_{j+1}^{k+1} = \mathcal{R}_j.$$

Our experience is that these iterations converge rather quickly.

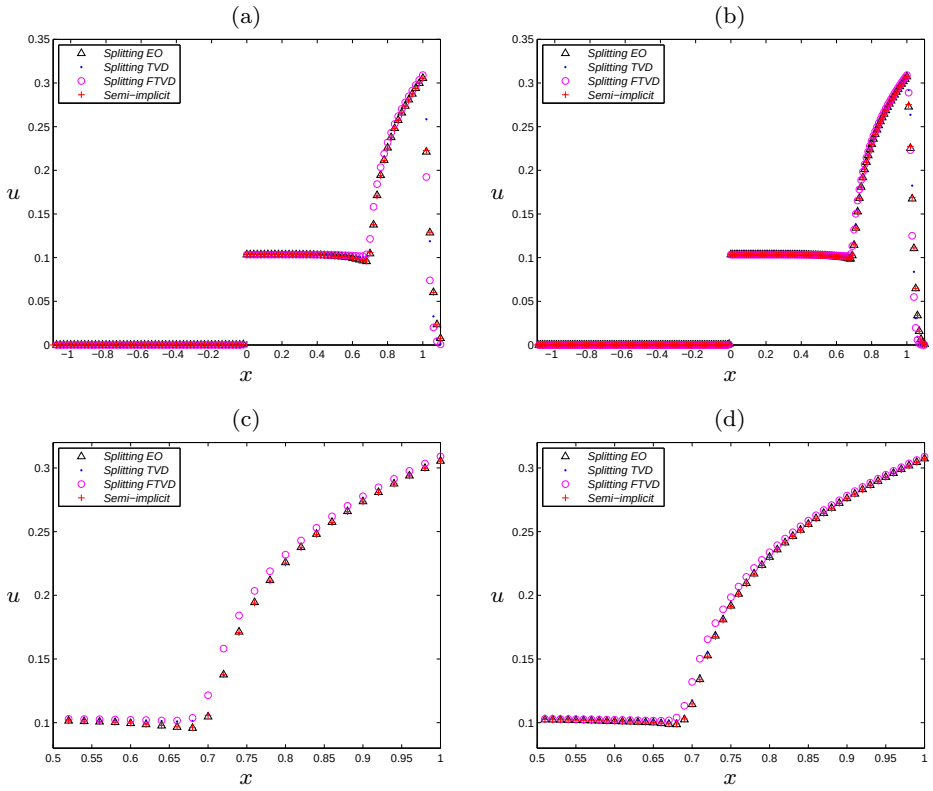

FIGURE 9. Example 4: numerical solution at $t = 25000\,\mathrm{s}$ with (a) $J = 50$, (b) $J = 100$, (c, d) enlarged views around $\xi = 0.7$ for (c) $J = 50$ and (d) $J = 100$.

In the purely linear setting, the Crank-Nicolson scheme is unconditionally stable, meaning that no condition on the time step is required. Our experience is that this is not true for the nonlinear degenerate problems that we are considering, mostly due to the presence of the discontinuous parameter γ_1. Nevertheless, we find that by using Crank-Nicolson the allowable time step size for our overall scheme is dictated by the hyperbolic portion of the problem rather than the parabolic portion.

Since each of the parabolic and hyperbolic operators has formal second-order accuracy in both space and time, we will maintain overall second order accuracy with the Strang splitting [20]. This is a well-known result, see, e.g., [14, 19].

7.2. **Examples 4 and 5: flocculated suspension.** Next, we include the strongly degenerate diffusion term by considering the effective solid stress function $\sigma_e(u)$ defined by the commonly used formula

$$\sigma_e(u) = \begin{cases} 0 & \text{for } u \le u_c, \\ \sigma_0\big((u/u_c)^k - 1\big) & \text{for } u > u_c, \end{cases} \tag{7.3}$$

where we use the parameters $\sigma_0 = 1\,\mathrm{Pa}$, $u_c = 0.1$ and $k = 6$ along with $\Delta\varrho = 1500\,\mathrm{kg/m^3}$ and $g = 9.81\,\mathrm{m/s^2}$ [7]. The vessel and control variables are the same as in Example 1, and we again set $u_0 \equiv 0$. Note that the derivative $\sigma_e'(u)$ of $\sigma_e(u)$ defined in (7.3) is in general discontinuous at $u = u_c$.

Figures 9 and 10 show the numerical solution of the continuous fill-up of the CT (operating at this state) calculated by the semi-implicit scheme described in

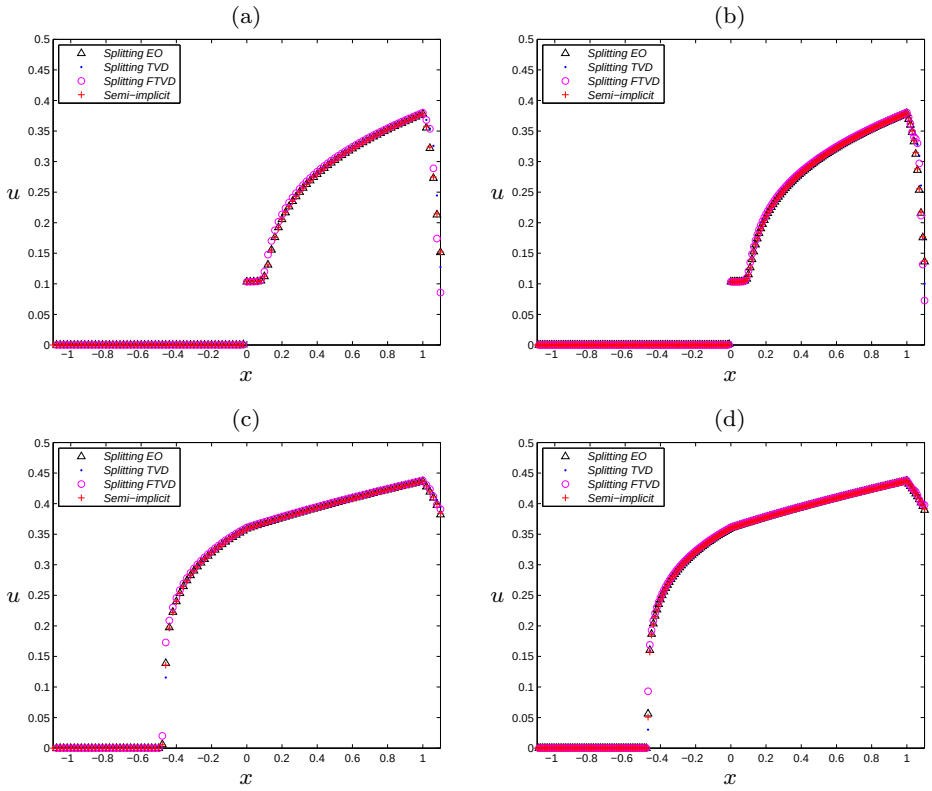

FIGURE 10. Example 4: numerical solution at (a, b) $t = 50000\,\mathrm{s}$ and (c, d) $t = 100000\,\mathrm{s}$ with (a, c) $J = 50$ and (b, d) $J = 100$.

[7] (BKT), the operator splitting scheme described herein (BKTS), the the operator splitting scheme including the simple TVD limiter (TVD1), and the operator splitting scheme involving the non-local limiter (TVD2). All calculations were performed with $\lambda = 2000\,\mathrm{s/m}$.

Finally in Example 5, we include the same effective solid stress function as in Example 1, and the control variables and the function $b(u)$ are the same as in Example 3. Figure 11 shows the numerical solution for this case at three selected times obtained by the BKT, BKTS, TVD1 and TVD2 schemes.

ACKNOWLEDGEMENTS

RB acknowledges support by Conicyt (Chile) through Fondecyt project 1090456, Fondap in Applied Mathematics, project 15000001, BASAL project CMM, Universidad de Chile and Centro de Investigación en Ingeniería Matemática ($\mathrm{CI}^2\mathrm{MA}$), Universidad de Concepción, and project AMIRA P996/INNOVA 08CM01-17 "Instrumentación y Control de Espesadores". The work of KHK was supported by the Research Council of Norway through an Outstanding Young Investigators Award.

REFERENCES

[1] N.G. Barton, C.-H. Li, and S.J. Spencer. Control of a surface of discontinuity in continuous thickeners. *J. Austral. Math. Soc. Ser. B*, 33:269–289, 1992.

[2] R. Bürger, A. García, K.H. Karlsen, and J.D. Towers. A family of numerical schemes for kinematic flows with discontinuous flux. *J. Engrg. Math.*, 60:387–425, 2008.

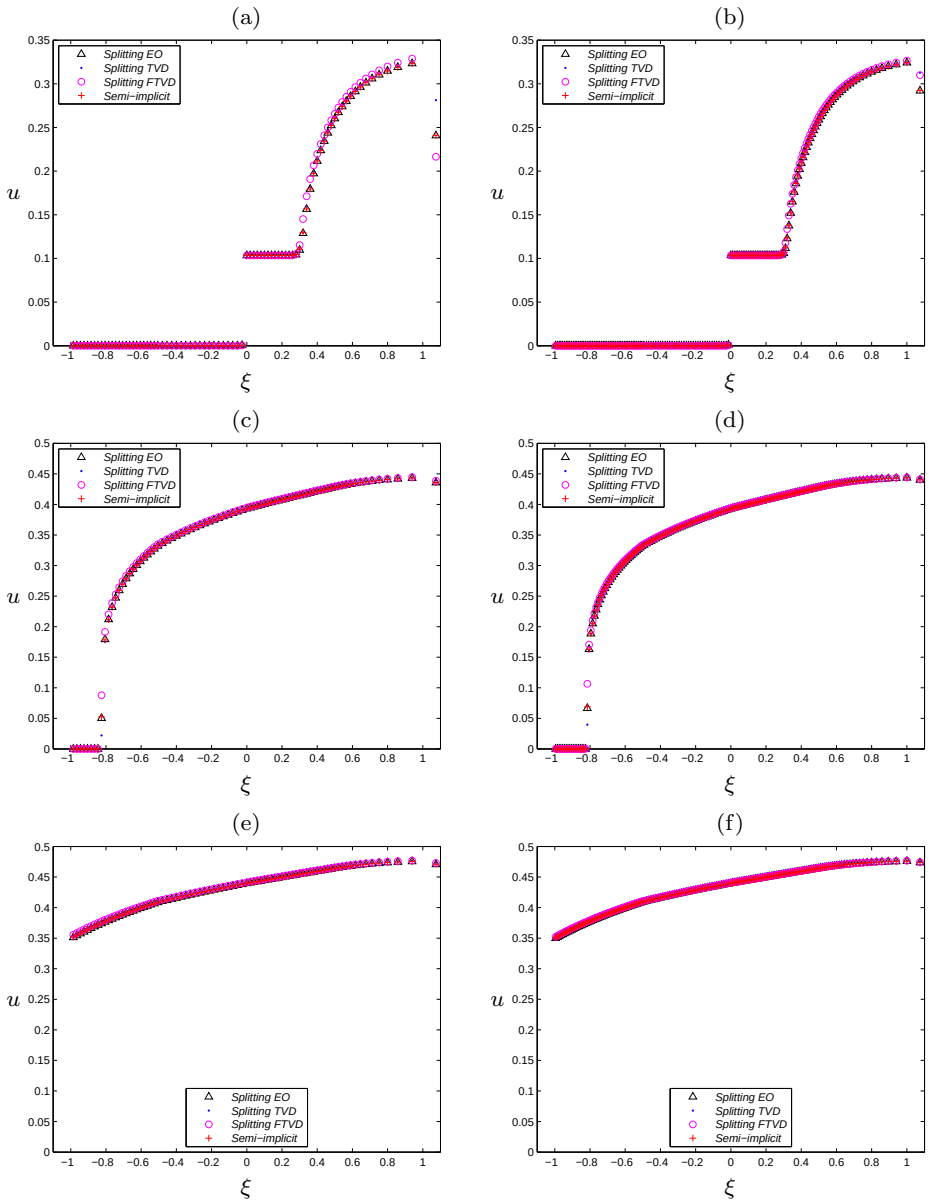

FIGURE 11. Example 5: numerical solution at (a, b) $t = 25000\,\mathrm{s}$, (c, d) $t = 100000\,\mathrm{s}$ and (e,f) $t = 150000\,\mathrm{s}$ with (a, c, e) $J = 50$ and (b, d, f) $J = 100$.

[3] R. Bürger, K.H. Karlsen, C. Klingenberg, and N.H. Risebro. A front tracking approach to a model of continuous sedimentation in ideal clarifier-thickener units. *Nonlin. Anal. Real World Appl.*, 4:457–481, 2003.

[4] R. Bürger, K.H. Karlsen, and N.H. Risebro. A relaxation scheme for continuous sedimentation in ideal clarifier-thickener units. *Comput. Math. Applic.*, 50:993–1009, 2005.

[5] R. Bürger, K.H. Karlsen, N.H. Risebro, and J.D. Towers. Monotone difference approximations for the simulation of clarifier-thickener units. *Comput. Visual. Sci.*, 6:83–91, 2004.

[6] R. Bürger, K.H. Karlsen, N.H. Risebro, and J.D. Towers. Well-posedness in BV_t and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units. *Numer. Math.*, 97:25–65, 2004.

[7] R. Bürger, K.H. Karlsen, and J.D. Towers. A mathematical model of continuous sedimentation of flocculated suspensions in clarifier-thickener units. *SIAM J. Appl. Math.*, 65:882–940, 2005.
[8] R. Bürger, K.H. Karlsen, and J.D. Towers. An Engquist-Osher type for conservation laws with discontinuous flux adapted to flux connections. *SIAM J. Numer. Anal.*, 47:1684–1712, 2009.
[9] J.P. Chancelier, M. Cohen de Lara, and F. Pacard. Analysis of a conservation PDE with discontinuous flux: a model of settler. *SIAM J. Appl. Math.*, 54:954–995, 1994.
[10] S. Diehl. Dynamic and steady-state behaviour of continuous sedimentation. *SIAM J. Appl. Math.*, 57:991–1018, 1997.
[11] S. Diehl. Operating charts for continuous sedimentation II: Step responses. *J. Engrg. Math.*, 53:139–185, 2005.
[12] S. Diehl. A uniqueness condition for nonlinear convection-diffusion equations with discontinuous coefficients. *J. Hyperbolic Differ. Equ.*, 6:127–159, 2009.
[13] B. Engquist and S. Osher. One-sided difference approximations for nonlinear conservation laws. *Math. Comp.*, 36:321–351, 1981.
[14] E. Godlewski and P. Raviart. Numerical Approximation of Hyperbolic Systems of Conservation Laws. Springer-Verlag, New York, 1996.
[15] S. Gottlieb, C. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43:89–112, 2001.
[16] A. Harten. High resolution schemes for hyperbolic conservation laws. *J. Comput. Phys.*, 49:357–393, 1983.
[17] G.J. Kynch. A theory of sedimentation. *Trans. Farad. Soc.*, 48:166–176, 1952.
[18] O. Lev, E. Rubin, and M. Sheintuch. Steady state analysis of a continuous clarifier-thickener system. *AIChE J.*, 32:1516–1525, 1986.
[19] R.J. LeVeque. Finite Difference Methods for Ordinary and Partial Differential Equations. SIAM, Philadelphia, PA, 2007.
[20] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.
[21] P.K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21:995–1011, 1984.
[22] B. Temple. Global solution of the Cauchy problem for a class of 2×2 nonstrictly hyperbolic conservation laws. *Adv. Appl. Math.*, 3:335–375, 1982.
[23] J.D. Towers. A difference scheme for conservation laws with a discontinuous flux: the nonconvex case. *SIAM J. Numer. Anal.*, 39:1197–1218, 2001.

Centro de Investigación en Ingeniería Matemática (CI^{2}MA)

PRE-PUBLICACIONES 2009

2009-07 Rais Ahmad, Fabián Flores-Bazán, Syed S. Irfan: *On completely generalized multi-valued co-variational inequalities involving strongly accretive operators*

2009-08 Gabriel N. Gatica, Ricardo Oyarzúa, Francisco J. Sayas: *Analysis of fully-mixed finite element methods for the Stokes-Darcy coupled problem*

2009-09 Raimund Bürger, Rosa Donat, Pep Mulet, Carlos A. Vega: *Hyperbolicity analysis of polydisperse sedimentation models via a secular equation for the flux Jacobian*

2009-10 Fabián Flores-Bazán, Elvira Hernández: *Unifying and scalarizing vector optimization problems: a theoretical approach and optimality conditions*

2009-11 Raimund Bürger, Ricardo Ruiz-Baier, Kai Schneider: *Adaptive multiresolution methods for the simulation of waves in excitable media*

2009-12 Alfredo Bermúdez, Luis Hervella-Nieto, Andres Prieto, Rodolfo Rodríguez: *Perfectly matched layers for time-harmonic second order elliptic problems*

2009-13 Ricardo Durán, Rodolfo Rodríguez, Frank Sanhueza: *Computation of the vibration modes of a Reissner-Mindlin laminated plate*

2009-14 Gabriel N. Gatica, Antonio Marquez, Manuel A. Sanchez: *Analysis of a velocity-pressure-pseudostress formulation for the stationary Stokes equations*

2009-15 Raimund Bürger, Kenneth H. KARLSEN, John D. Towers: *On some difference schemes and entropy conditions for a class of multi-species kinematic flow models with discontinuous flux*

2009-16 Gabriel N. Gatica, George C. Hsiao, Francisco J. Sayas: *Relaxing the hypotheses of the Bielak-MacCamy BEM-FEM coupling*

2009-17 Iuliu S. Pop, Florin A. Radu, Mauricio Sepúlveda, Octavio Vera: *Error estimates for the finite volume discretization for the porous medium equation*

2009-18 Raimund Bürger, Kenneth H. KARLSEN, Hector Torres, John D. Towers: *Second-order schemes for conservation laws with discontinuous flux modelling clarifier-thickener units*

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 (CI²MA)
Universidad de Concepción

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