Noname manuscript No.
(will be inserted by the editor)

Implicit monotone difference methods for scalar conservation laws with source terms

Michael Breuß · Andreas Kleefeld

Received: date / Accepted: date

Abstract In this article, a concept of implicit methods for scalar conservation laws in one or more spatial dimensions allowing also for source terms of various types is presented. This material is a significant extension of previous work of the first author [5]. Implicit notions are developed that are centered around a monotonicity criterion. We demonstrate a connection between a numerical scheme and a discrete entropy inequality, which is based on a classical approach by Crandall and Majda. Additionally, three implicit methods are investigated using the developed notions. Next, we conduct a convergence proof which is not based on a classical compactness argument. Finally, the theoretical results are confirmed by various numerical tests.

Keywords Conservation laws · finite difference methods · implicit methods · monotone methods · source term · entropy solution.

Mathematics Subject Classification (2000) 35L65 · 65M06 · 65M12

1 Introduction

This article deals with the entropy solution of hyperbolic conservation laws in the sense of Kružkov. Specifically, we allow the numerical methods to act within the two most general settings, that is (i) smooth fluxes together with non-linear sources and (ii) continuous fluxes and sources depending both on

M. Breuß
Institute for Mathematics, Brandenburg Technical University, Platz der Deutschen Einheit 1, 03046 Cottbus, Germany
Tel.: +49-355-692086
Fax: +49-355-692402
E-mail: breuss@b-tu.de

A. Kleefeld
Forschungszentrum Jülich GmbH, Institute for Advanced Simulation, Jülich Supercomputing Centre, Wilhelm-Johnen-Straße, 52425 Jülich, Germany

space and time. The corresponding analytical existence and uniqueness results for these cases are given within a number of papers of Kružkov and his coworkers, see for example [2,14,15] and the references therein.

This paper represents a significant extension of the work by Breuß [5], where implicit methods are considered for homogeneous scalar equations in one dimension. To our knowledge, the combination of the developed concept of implicit methods for both mentioned general problems together with the application of corresponding schemes on problems belonging to both classes is new. Accordingly, the main contribution of this paper is the extension of the rigorously validated range of applicability of finite difference methods.

The encountered difficulties for the described task have already been discussed in the introduction of Breuß [5]. Summarizing, information that is propagated with infinite speed may take place provided that a flux function of a nonlinear conservation law is not Lipschitz continuous as it is accepted in setting (ii). A detailed one-dimensional example is given by Kružkov and Panov in [15] (see also [5]), where the exact solution is known. This example shows that a rarefaction wave extending to infinity after arbitrarily small time takes place. Additionally, this example has a pole for $u = 0$ and the solution domain is infinite although an initial condition with compact support is given.

Two direct conclusions emerge from this example. At first, the Courant-Friedrichs-Lewy (CFL) number would be effectively zero provided an explicit scheme is used. Additionally, the Kuznetsov approach for convergence [16] is not employable, because it relies on a suitable error estimation which explicitly uses the Lipschitz continuity of the flux and the boundedness of the domain of the solution. At second, a variety of other well-established approaches for the convergence of numerical methods are not applicable. For instance, one approach is based on Helly's theorem which uses the compactness of the function space of bounded variations (BV). This is employed in the convergence proofs of Total Variation Diminishing (TVD) methods. But using the BV concept, this function space is only compact (see LeVeque [17]) provided a fixed compact space-time-domain containing the solution is used. Hence, the compactness property of this function space is unfortunately not applicable in the discussed case. The same is true for explicit monotone methods as it is the case in the fundamental work of Crandall and Majda [7]. They used the properties of this function space to obtain a compactness argument. Especially, in that work the sources are also assumed to be essentially bounded and BV-stable integrable functions depending on space and time. In another important approach introduced by DiPerna [9] measure valued solutions are used, where the compactness of the domain of the solution both in space and time is assumed which has already been discussed in [6].

From this discussion it should be clear that we need to employ implicit schemes such that the convergence strategy is different from the before mentioned ones. Therefore, we use the monotonicity of implicit methods to obtain a discrete comparison principle. This suffices to guarantee the convergence of such methods to the entropy solution in the sense of Kružkov.

Therefore, in this paper the monotonicity property of an implicit scheme is investigated (see [10,17] for the discussion for explicit schemes). Hence, it is indispensable to avoid any derivative of the flux. As we will show, we construct a monotonicity notion that is based on a comparison of data sets using an induction principle.

The application of this monotonicity notion on three implicit variations of well-known monotone explicit schemes is investigated. One would expect, that implicit schemes are generally capable to capture all effects described by a conservation law even for continuous fluxes and general sources, because in the implicit case the numerical characteristics include all the characteristics of the differential equation. However, while our monotonicity investigations of an implicit upwind scheme and an implicit Godunov-type method yield the expected results, the investigation of the implicit variation of the traditional Lax-Friedrichs scheme shows, that the scheme is only monotone even in the full implicit case if the flux is Lipschitz continuous. Furthermore, the restriction on the admissible Lipschitz constant of the flux is not depending on the number of spatial dimensions. This interesting result which is new to our knowledge is explored via a simple experiment using a two-dimensional linear advection equation.

Let us note that, on a broader scope, implicit methods for hyperbolic conservation laws have a long history of interest, see e.g. [1,8] for two important milestones. While implicit discrete formulations require to solve systems of equations, the reason for some interest in implicit schemes arises as their use is sometimes advocated, for instance if the modeled process incorporates different wave speeds that need to be resolved by employing a common time step. This may be the case especially in systems of equations, for which the scalar hyperbolic PDEs traditionally serve as role models in the mathematical sense. As another point of practical interest, many initially time-dependent processes, for instance in gas dynamics (i.e. compressible flows) described by the hyperbolic system of Euler equations, eventually develop steady state solutions which are much better captured using an implicit scheme than an explicit one, cf. [19]. Thus, the theoretical foundation of implicit scheme components which we discuss here is of general interest in several fields.

This article consists of five additional sections. In Section 2, we briefly review the two most general theoretical results on solutions of conservation laws available to our knowledge, namely the existence and uniqueness results established in [2] and [13]. In the next section, we introduce the notions for implicit methods that are centered around monotonicity. The given detailed convergence proof is an extension of the strategy given in Breuß [5]. Section 4 presents the investigation of three numerical methods with respect to their monotonicity. Additionally, for these methods the proofs of convergence towards the entropy solution are given. Finally, we present the results of various numerical tests in Section 5 followed by a short summary and conclusive remarks in Section 6.

2 The setting

Within this section, we define the two mathematical scenarios of interest, i.e. we briefly review the type of problems considered in [2] and [13].

Scenario 1 The Cauchy problem under consideration is

$$\frac{\partial}{\partial t} u\left(\mathbf{x},t\right) + \sum_{l=1}^{d} \frac{\partial}{\partial x_l} f_l\left(u\left(\mathbf{x},t\right)\right) \;=\; q \quad \text{on} \quad \mathbb{R}^d \times (0,T)\;, \tag{1}$$

$$u\left(\mathbf{x},0\right) \;=\; u_0\left(\mathbf{x}\right) \quad \text{on} \quad \mathbb{R}^d\,, \tag{2}$$

where T is a fixed positive number. Concerning the flux functions we generally assume

$$f_l(u) \in C\left(\mathbb{R};\,\mathbb{R}\right)\;, \quad l = 1,\,\ldots,\,d\,. \tag{3}$$

In order to apply the uniqueness theorem given in [2], the fluxes are additionally supposed to satisfy the growth conditions

$$|f_l(u) - f_l(\hat{u})| \le \omega_l(u-\hat{u}) \;\text{ a.e. for }\; u \ge \hat{u} \;\text{ and for }\; l = 1,\ldots,d\,,$$

with the moduli of continuity ω_l featuring

$$\omega_1(0) = \ldots = \omega_d(0) = 0 \quad \text{and} \quad \liminf_{r\to 0} r^{1-d} \prod_{l=1}^{d} \omega_l(r) < \infty\,.$$

Note that these conditions on the fluxes are more general than the usually assumed Lipschitz continuity. The initial condition shall satisfy

$$u_0 \in L^{\infty}_{loc}\left(\mathbb{R}^d;\,\mathbb{R}\right)\;, \tag{4}$$

and for the source term we consider

$$q \equiv q\left(\mathbf{x},t\right) \in L^{1}_{loc}\left(\mathbb{R}^d \times (0,T);\,\mathbb{R}\right)\;, \tag{5}$$

$$q(\cdot,t) \in L^{\infty}\left(\mathbb{R}^d;\,\mathbb{R}\right) \;\text{ for a.e. }\; t \in (0,T) \;\text{ and }\; \int_0^T \|q(\cdot,t)\|_{\infty}\,\mathrm{d}t < \infty\,. \tag{6}$$

Under the conditions (3) – (6), Bénilan and Kružkov [2] proved uniqueness of the entropy solution of (1) – (2).

Because the solution of the Cauchy problem generally develops discontinuities even if u_0 is smooth, it is often considered in its weak form, i.e.

$$\begin{aligned}
&\int_0^{\infty}\int_{\mathbb{R}^d} \left[u\left(\mathbf{x},t\right)\phi_t\left(\mathbf{x},t\right) + \sum_{l=1}^{d} f_l(u\left(\mathbf{x},t\right)) \frac{\partial}{\partial x_l}\phi\left(\mathbf{x},t\right) \right] \mathrm{d}\mathbf{x}\,\mathrm{d}t \\
&= -\int_{\mathbb{R}^d} u_0\left(\mathbf{x}\right)\phi_0\left(\mathbf{x}\right)\,\mathrm{d}\mathbf{x} \\
&- \int_0^{\infty}\int_{\mathbb{R}^d} q\left(\mathbf{x},t\right)\phi\left(\mathbf{x},t\right)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \quad \forall \phi \in C_0^{\infty}\left(\mathbb{R}^{d+1};\,\mathbb{R}\right)\,.
\end{aligned} \tag{7}$$

It is well-known that weak solutions are in general not unique, see for example [17] and the references therein. In order to ensure uniqueness, a so-called entropy condition has to be introduced. The already mentioned entropy condition due to Kružkov [2] which guarantees the uniqueness of a solution of (1) – (2) takes the form

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}^d} \Bigg[|u(\mathbf{x},t) - k| \phi_t(\mathbf{x},t) \\
&\qquad + \sum_{l=1}^{d} \operatorname{sgn}(u(\mathbf{x},t) - k) \left[f_l(u(\mathbf{x},t)) - f_l(k) \right] \frac{\partial}{\partial x_l} \phi(\mathbf{x},t) \Bigg] \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \\
&\geq - \int_{\mathbb{R}^d} |u_0(\mathbf{x}) - k| \phi_0(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&\qquad - \int_0^\infty \int_{\mathbb{R}^d} \operatorname{sgn}[u(\mathbf{x},t) - k] \, q(\mathbf{x},t) \, \phi(\mathbf{x},t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \qquad (8) \\
&\quad \text{for all } \phi \in C_0^\infty\left(\mathbb{R}^{d+1}; \mathbb{R}\right) \text{ with } \phi \geq 0 \text{ and for all } k \in \mathbb{R}.
\end{aligned}$$

Scenario 2 The *Scenario 2* deals with the Cauchy problem

$$\frac{\partial}{\partial t} u(\mathbf{x},t) + \sum_{l=1}^{d} \frac{d}{dx_l} f_l(\mathbf{x},t,u(\mathbf{x},t)) = q \quad \text{on} \quad \mathbb{R}^d \times (0,T), \qquad (9)$$

$$u(\mathbf{x},0) = u_0(\mathbf{x}) \quad \text{on} \quad \mathbb{R}^d, \qquad (10)$$

where T is a fixed positive number and with

$$\frac{d}{dx_l} f_l \equiv f_{l_{x_l}} + f_{l_u} u_{x_l}.$$

In comparison to *Scenario 1*, we impose different assumptions on the fluxes and the source terms. As in (4), there is no particular condition imposed on the initial data. The flux functions are now assumed to satisfy

$$f_l(\mathbf{x},t,u) \in C^1\left(\mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}; \mathbb{R}\right), \quad l = 1, \ldots, d. \qquad (11)$$

As source terms we consider functions

$$q \equiv q(\mathbf{x},t,u(\mathbf{x},t)) \in C^1\left(\mathbb{R}^d \times \mathbb{R}_+ \times \mathbb{R}; \mathbb{R}\right). \qquad (12)$$

Under the conditions (11) and (12), Kružkov [13] proved the uniqueness of the entropy solution of (9) – (10). Comparing the weak formulation of this problem with the weak formulation (7), we have to substitute

$$\int_0^\infty \int_{\mathbb{R}^d} q(\mathbf{x},t,u(\mathbf{x},t)) \, \phi(\mathbf{x},t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \quad \text{for} \quad \int_0^\infty \int_{\mathbb{R}^d} q(\mathbf{x},t) \, \phi(\mathbf{x},t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t. \qquad (13)$$

The assumptions (11) and (12) yield the form of the Kružkov entropy condition as

$$\begin{aligned}
&\int_0^\infty \int_{\mathbb{R}^d} \Bigg[|u(\mathbf{x},t) - k| \phi_t(\mathbf{x},t) \\
&\quad + \sum_{l=1}^{d} \operatorname{sgn}(u(\mathbf{x},t) - k) \left[f_l(\mathbf{x},t,u(\mathbf{x},t)) - f_l(\mathbf{x},t,k) \right] \frac{\partial}{\partial x_l} \phi(\mathbf{x},t) \Bigg] \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \\
&\geq - \int_{\mathbb{R}^d} |u_0(\mathbf{x}) - k| \phi_0(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\
&\quad - \int_0^\infty \int_{\mathbb{R}^d} \sum_{l=1}^{d} \operatorname{sgn}\left[u(\mathbf{x},t) - k \right] \left[q(\mathbf{x},t,u(\mathbf{x},t)) - f_{l_{x_l}}(\mathbf{x},t,k) \right] \phi(\mathbf{x},t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \\
&\quad \text{for all } \phi \in C_0^\infty\left(\mathbb{R}^{d+1}; \mathbb{R}\right) \text{ with } \phi \geq 0 \text{ and for all } k \in \mathbb{R}\,. \qquad (14)
\end{aligned}$$

3 Numerical methods

We first describe the implicit notions, followed by the proofs of the involved Lemmas and Theorems in a separate section. For the sake of brevity, we discuss only *Scenario 1* in detail, since the techniques which have to be used with respect to *Scenario 2* are identical. The proper conceptual extension to *Scenario 2* is described within additional remarks.

3.1 A concept of implicit methods

Since we want to describe numerical methods in d spatial dimensions, we spend some effort on a general notation.

Because we investigate finite difference methods, we have to introduce grid points. For simplicity, we consider grids which are equidistant with respect to the individual d spatial dimensions as well as to time, i.e. we employ grid spacings Δx_l corresponding to the space dimensions $l = 1, \ldots, d$, and Δt corresponding to time.

Since this results in a countable number of grid points, we introduce a *linear numbering* J of the spatial grid points

$$J = \{0, 1, 2, \ldots\} \,.$$

We also define a bijective mapping

$$\begin{aligned}
\tilde{J} \quad : \quad & J \longrightarrow \mathbb{R}^d \\
& i \longrightarrow (i_1 \Delta x_1,\, i_2 \Delta x_2,\, \ldots,\, i_d \Delta x_d)^{\mathrm{T}} \quad \text{with} \quad (i_1,\, i_2,\, \ldots,\, i_d)^{\mathrm{T}} \in \mathbb{Z}^d \,.
\end{aligned}$$

Let us note that the mapping goes formally to $\mathbb{R}^d$, but it just maps to a countable subset of $\mathbb{R}^d$. In order to describe the indices within the stencil of a

numerical method, we define the index $i \pm \delta l$ via

$$i \pm \delta l \xrightarrow{\tilde{J}} (i_1 \Delta x_1, i_2 \Delta x_2, \ldots, (i_l \pm 1)\, \Delta x_l, \ldots, i_d \Delta x_d)^{\mathrm{T}} .$$

Let u_j^k and q_j^k denote the value of the numerical solution and the value of the source term at the point with the index $j \in J$ at the time level $k\Delta t$, respectively. With these notations, we consider *conservative* implicit methods in the form (refer also to [3,17])

$$u_j^{n+1} = u_j^n - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l \left(u_j^{n+1}, u_{j+\delta l}^{n+1} \right) - g_l \left(u_{j-\delta l}^{n+1}, u_j^{n+1} \right) \right\} + \Delta t q_j^{n+1} . \quad (15)$$

We assume that the numerical flux functions g_l introduced in (15) are *consistent*, i.e.

$$g_l(v, v) = f_l(v) \text{ holds for all } v \in \mathbb{R} \text{ and for all } l = 1, \ldots, d .$$

In the case of *Scenario 2*, we simply add arguments $(\mathbf{x}_j, t^{n+1})$ within the fluxes; we will not do this explicitly in the following.

The key to nonlinear stability is the notion of monotonicity.

Definition 1 (Monotonicity) Let two data sequences

$$v^n = \left\{ v_j^n \right\}_{j \in J} \quad \text{and} \quad w^n = \left\{ w_j^n \right\}_{j \in J}$$

be given. Let the investigated consistent and conservative numerical method produce new sequences of data v^{n+1} and w^{n+1} from the given data v^n and w^n, respectively. Then the numerical method is monotone iff the implication

$$v^n \geq w^n \quad \Rightarrow \quad v^{n+1} \geq w^{n+1} \quad (16)$$

holds in the sense of the comparison of components.

It is useful to define H and $\tilde{H}_l$ using $\underline{d} = \{1, \ldots, d\}$ via

$$\begin{aligned} u_j^{n+1} &= H \left(l \in \underline{d}, u_{j-\delta l}^{n+1}, u_j^{n+1}, u_{j+\delta l}^{n+1}, u_j^n \right) \\ &= u_j^n - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l \left(u_j^{n+1}, u_{j+\delta l}^{n+1} \right) - g_l \left(u_{j-\delta l}^{n+1}, u_j^{n+1} \right) \right\} + \Delta t q_j^{n+1} \\ &= u_j^n + \sum_{l=1}^{d} \tilde{H}_l \left(u_{j-\delta l}^{n+1}, u_j^{n+1}, u_{j+\delta l}^{n+1} \right) + \Delta t q_j^{n+1} . \end{aligned} \quad (17)$$

Theorem 1 (Monotonicity of implicit methods) *Let a, b and c be arbitrarily chosen but fixed real numbers. A consistent and conservative implicit method of type (15) is monotone iff for all spatial dimensions $l \in \{1, \ldots, d\}$ holds*

$$\tilde{H}_l (a + \Delta a, b, c) \geq \tilde{H}_l (a, b, c) \quad \forall\, \Delta a \geq 0 , \quad (18)$$

$$\tilde{H}_l (a, b, c + \Delta c) \geq \tilde{H}_l (a, b, c) \quad \forall\, \Delta c \geq 0 . \quad (19)$$

Note that we have omitted the condition

$$H\left(l \in \underline{d},\, u_{j-\delta l}^{n+1},\, u_j^{n+1},\, u_{j+\delta l}^{n+1},\, s + \Delta s\right) \geq H\left(l \in \underline{d},\, u_{j-\delta l}^{n+1},\, u_j^{n+1},\, u_{j+\delta l}^{n+1},\, s\right)$$

for all $j \in J$ and all $\Delta s \geq 0$, since this condition is redundant. This is due to the form of the method (15). Additionally, note that the monotonicity property does not depend on the exact nature of the source terms, i.e. both *Scenario 1* and *Scenario 2* are included within the range of applicability of Theorem 1.

Theorem 2 (*L_∞*-Stability) *Let an implicit method of the form (15) be given, which is also conservative and monotone. Then the numerical solution is L_∞-stable over any finite time interval $[0, T]$.*

The following definition is useful for proving convergence towards the entropy solution.

Definition 2 (Consistency with the entropy condition) An implicit numerical scheme of type (15) is consistent with the entropy condition of Kružkov, if there exist for all $l = 1, \ldots, d$ numerical entropy fluxes G_l which satisfy for all $k \in \mathbb{R}$ the following assertions:

1. Consistency with the entropy flux of Kružkov

$$G_l(v, v; k) = F_l(v; k)\ \forall v \text{ with } F_l(v; k) = \operatorname{sgn}(v - k)\left[f_l(v) - f_l(k)\right]. \quad (20)$$

2. Validity of a discrete entropy inequality

$$\begin{aligned} &\frac{U\left(u_j^{n+1}; k\right) - U\left(u_j^n; k\right)}{\Delta t} \\ &\leq -\sum_{l=1}^{d} \frac{G_l\left(u_j^{n+1}, u_{j+\delta l}^{n+1}; k\right) - G_l\left(u_{j-\delta l}^{n+1}, u_j^{n+1}; k\right)}{\Delta x_l} \\ &\quad + \operatorname{sgn}\left[u_j^{n+1} - k\right] q_j^{n+1}, \end{aligned} \quad (21)$$

   where $U(v; k) = |v - k|$ is chosen due to Kružkov.

In the sequel, we define

$$a \vee b := \max(a, b) \quad \text{and} \quad a \wedge b := \min(a, b). \quad (22)$$

The important connection between the numerical entropy fluxes G_l and the numerical flux functions g_l is now established which is based on a variation of a procedure employed by Crandall and Majda [7].

Lemma 1 *Let a consistent and conservative numerical scheme of type (15) be given with numerical flux functions g_l, $l = 1, \ldots, d$. Then the numerical entropy fluxes defined by*

$$G_l(v, w; k) := g_l(v \vee k, w \vee k) - g_l(v \wedge k, w \wedge k) \quad (23)$$

are consistent with the entropy fluxes of Kružkov.

One can now prove the following result, partly by a variation of the procedure given in [7]. We introduce the source term within the proof.

Theorem 3 *Let an implicit scheme of the form (15) be given, which is also consistent, conservative, and monotone. Then the scheme is also consistent with the entropy condition of Kružkov.*

Under the same assumptions, we prove convergence of the corresponding numerical approximation to the entropy solution. We want to do this later when we concretely investigate numerical schemes.

3.2 Proofs

We first want to prove Theorem 1. The idea of the proof can be sketched as follows. Let two sequences w^n and w^{n+1} be given, where w^{n+1} results from an application of a considered method on w^n. Then, a positive change in a given value w_j^n inspires a positive change in w_j^{n+1}. Secondly, a positive change in w_j^{n+1} inspires positive changes in $w_{j\pm\delta l}^{n+1}$ for all l, thus creating no oscillations. Thirdly, concerning an arbitrary index i, positive changes in $w_{i\pm\delta l}^{n+1}$ result in positive changes in w_i^{n+1}. Since the index j used in the second argument is chosen arbitrarily, this is the same argument as the third one for $j \in \{i \pm \delta l;\, l = 1, \ldots, d\}$. If and only if these conditions are fulfilled by a considered method, the method is monotone.

In order to give the proof of Theorem 1 a convenient structure, we first give the following Lemma.

Lemma 2 *Let a consistent and conservative implicit method of the form (15) be given, which satisfies the conditions (18) and (19). Furthermore, let two sequences $v^n = \left\{v_j^n\right\}_{j\in J}$ and $w^n = \left\{w_j^n\right\}_{j\in J}$ be given. Then from*

$$\exists\, i \in J:\ v_i^n > w_i^n \quad \textit{and} \quad \forall\, j \in \mathbf{J}\,(j \neq i):\ v_j^n = w_j^n$$

follows $v^{n+1} \geq w^{n+1}$ in the sense of the comparison of components.

Proof (of Lemma 2)
By the assumption of the Lemma there exists an index $i \in J$ so that $v_i^n > w_i^n$ holds. Without restriction of generality we choose $i = 0$. The proof of the assertion follows by induction over suitable subsets of J.

Let us introduce these subsets. Therefore, let J_m denote a subset of J containing m elements with

$$\begin{aligned}&\forall\, m_0 \in J_m\, \exists\, m_1 \in J_m\,(m_0 \neq m_1):\\ &\Big[\{m_0\} \cap \{p \in J_m\ ;\ p = m_1 \pm \delta l,\, l = 1, \ldots, d\}\Big] \neq \emptyset\end{aligned}$$

for $m \geq 2$, thus the elements of J_m are indices of neighboring points.

Beginning of the induction: $m = 1$

As indicated, we choose without restriction of generality $J_1 = \{0\}$. The statement is true because of the form of the method (15), so that

$$\begin{aligned} &H\left(l \in \underline{d},\, w_{-\delta l}^{n+1},\, w_0^{n+1},\, w_{\delta l}^{n+1},\, s + \Delta s\right) \\ &\geq H\left(l \in \underline{d},\, w_{-\delta l}^{n+1},\, w_0^{n+1},\, w_{\delta l}^{n+1},\, s\right) \quad \forall\, \Delta s \geq 0 \quad \text{holds.} \end{aligned}$$

Assumption:
The statement is true for arbitrary but fixed $m > 1$.
Induction step: $m \mapsto m + 1$
Let the statement be true for the subsets $\left\{v_i^{n+1}\right\}_{i \in J_m}$ and $\left\{w_i^{n+1}\right\}_{i \in J_m}$ of the sequences v^{n+1} and w^{n+1}. In particular, it holds

$$\begin{aligned} v_{\tilde{m}}^{n+1} &\geq w_{\tilde{m}}^{n+1} \quad \text{for an index} \quad \tilde{m} \in J_m \\ \text{with } &\left[\{i \in J;\, i = \tilde{m} \pm \delta l,\, l = 1, \ldots, d\} \cap (J \setminus J_m)\right] \neq \emptyset \end{aligned}$$

which is otherwise chosen arbitrarily, i.e. we consider an index $\tilde{m}$ corresponding to a grid point with at least one neighbor having an index not in J_m.

Without restriction on generality, let us choose a particular index l_m corresponding to the situation

$$\tilde{m} \in J_m \quad \text{and} \quad \tilde{m} + \delta l_m \notin J_m\,.$$

Since by construction the sequences v^{n+1} and w^{n+1} are identical outside the considered subsets, it holds

$$\tilde{H}_{l_m}\left(v_{\tilde{m}}^{n+1},\, w_{\tilde{m}+\delta l_m}^{n+1},\, w_{\tilde{m}+2\delta l_m}^{n+1}\right) \;\geq\; \tilde{H}_{l_m}\left(w_{\tilde{m}}^{n+1},\, w_{\tilde{m}+\delta l_m}^{n+1},\, w_{\tilde{m}+2\delta l_m}^{n+1}\right)$$

by (18). If the index $\tilde{m} + 2\delta l_m$ is already in J_m, we estimate

$$\tilde{H}_{l_m}\left(v_{\tilde{m}}^{n+1},\, w_{\tilde{m}+\delta l_m}^{n+1},\, v_{\tilde{m}+2\delta l_m}^{n+1}\right) \geq \tilde{H}_{l_m}\left(w_{\tilde{m}}^{n+1},\, w_{\tilde{m}+\delta l_m}^{n+1},\, w_{\tilde{m}+2\delta l_m}^{n+1}\right)$$

by also using (19). The case $\tilde{m} \in J_m$ and $\tilde{m} - \delta l_m \notin J_m$ can be handled analogously.

By defining

$$J_{m+1} := J_m \cup \{\tilde{m} + \delta l_m\} \quad \text{or} \quad J_{m+1} := J_m \cup \{\tilde{m} - \delta l_m\}$$

corresponding to the situation under consideration, it follows $v_i^{n+1} \geq w_i^{n+1}$ for all $i \in J_{m+1}$. Since $\tilde{m}$ and l_m were chosen arbitrarily within the framework of the construction, the procedure is well-defined and the proof is finished.

Proof (of Theorem 1)
Let again two sequences v^n, w^n be given, which are mapped on sequences v^{n+1} and w^{n+1} by application of the considered consistent and conservative numerical method, respectively.
"$\Rightarrow$":
Let the method be monotone in the sense of Definition 1. Let $v^n \geq w^n$ hold in

the sense of comparison of components. By the assumed monotonicity of the scheme follows $v^{n+1} \geq w^{n+1}$. It remains to verify the validity of the conditions (18) and (19).

To condition (18):
Let $l \in \{1, \ldots, d\}$ be chosen arbitrarily but fixed. Accordingly, let an arbitrarily chosen but fixed index i and a corresponding set of values

$$\{a, b, c\} \subset w^{n+1} \quad \text{be given with} \quad \left(w_{i-\delta l}^{n+1},\, w_i^{n+1},\, w_{i+\delta l}^{n+1}\right) = (a, b, c)\ .$$

Assume that for $\Delta a \geq 0$ it does not hold in general

$$\tilde{H}_l\,(a + \Delta a, b, c) \quad \geq \quad \tilde{H}_l\,(a, b, c)\ .$$

Then there exist two tuples (a_1, b_1, c) and (a_2, b_2, c) with $a_1 > a_2$ and

$$\tilde{H}_l\,(a_1, b_1, c) < \tilde{H}_l\,(a_2, b_2, c)\ . \tag{24}$$

Since we investigate the general situation, we may well assume equality of the remainder of the sequences under consideration, thus the only resulting change by application of the method originates from (24). By (15) it follows that $b_1 < b_2$ has in general to be valid. On the other hand there is $(a_1, b_1) \geq (a_2, b_2)$ in the sense of comparison of components by the assumed monotonicity of the method, and so the assumption is wrong and the validity of (18) is verified.

To condition (19):
The proof can be done analogously.

"$\Leftarrow$":
Next, the validity of the monotonicity condition (16) under the assumptions (18) and (19) is proven. Therefore, we define the set

$$\hat{J}^n := \{i \in J\ ;\ v_i^n > w_i^n,\, v_i^n \in v^n,\, w_i^n \in w^n\}\ .$$

There are only a few possibilities for the composition of $\hat{J}^n$: It may consist of the empty set or a finite or infinite subset of the index set J containing the indices of all spatial grid points. Since we have to take into account all these cases, we define

$$\hat{J}_m^n := \left\{\hat{J}^n\ ;\ \sharp\left(\hat{J}^n\right) = m\right\}\ .$$

The proof of the assertion follows by induction over $m \geq 1$ concerning these sets. Note that the case $m = 0$ is trivial.

Beginning of the induction: $\hat{J}^n = \hat{J}_1^n$.
Let i be the index in the arbitrarily chosen but fixed index set $\hat{J}_1^n$. Then the validity of the monotonicity condition follows by application of Lemma 2.

Assumption: The assertion holds for all subsets of $\hat{J}^n = \hat{J}_m^n$ for an arbitrarily chosen but fixed number $m > 1$.

Induction step: $m \mapsto m + 1$

Now we consider $\hat{J}^n_{m+1}$ with $\hat{J}^n_m \subset \hat{J}^n_{m+1}$. We define two particular indices m_1, m_2 with

$$m_1 \in \hat{J}^n_m \quad \text{and} \quad m_2 \in \left(\hat{J}^n_{m+1} \setminus \hat{J}^n_m \right) .$$

Thereby, the index m_1 is chosen arbitrarily but fixed. By the assumption of the induction, the scheme is monotone with respect to positive changes in values corresponding to the index set $\hat{J}^n_m$. This means in particular that a positive change in $v^n_{m_1}$ together with positive changes in other values corresponding to $\hat{J}^n_m$ leads to non-negative changes in the sequence v^{n+1}.

Now a simultaneous positive change in $v^n_{m_1}$ and $v^n_{m_2}$ is considered while in the background there are arbitrary but fixed positive changes in the values corresponding to $\hat{J}^n_{m+1} \setminus \{m_1, m_2\}$.

Let the data resulting from positive changes in v^n_i, $i \in \hat{J}^n_m \setminus \{m_1, m_2\}$, be denoted by $\bar{v}^{n+1}$, i.e. $\bar{v}^{n+1} \geq w^{n+1}$ holds by the assumption of the induction step.

Moreover, let Δ^1_j be a change in $\bar{v}^{n+1}_j$ induced by a positive change in $v^n_{m_1}$. Thus Δ^1_j is always non-negative by the assumption of the induction. Analogously, let Δ^2_j a change in $\bar{v}^{n+1}_j$ induced by a positive change in $v^n_{m_2}$. The change Δ^2_j is also non-negative which follows analogously to the proof of Lemma 2.

There are two possibilities to investigate for the mutual effects of such changes in data corresponding to an arbitrary but fixed index $\tilde{i}$ and an accordingly arranged index $l_i \in \{1, \ldots, d\}$:

$$\begin{aligned} & \tilde{H}_{l_i} \left(\bar{v}^{n+1}_{\tilde{i}-\delta l_i} + \Delta^1_{\tilde{i}-\delta l_i}, \bar{v}^{n+1}_{\tilde{i}}, \bar{v}^{n+1}_{\tilde{i}+\delta l_i} + \Delta^2_{\tilde{i}+\delta l_i} \right) \\ \overset{(18),(19)}{\geq} & \tilde{H}_{l_i} \left(\bar{v}^{n+1}_{\tilde{i}-\delta l_i}, \bar{v}^{n+1}_{\tilde{i}}, \bar{v}^{n+1}_{\tilde{i}+\delta l_i} \right) \end{aligned}$$

and

$$\begin{aligned} \tilde{H}_{l_i} & \left(\bar{v}^{n+1}_{\tilde{i}-\delta l_i} + \Delta^2_{\tilde{i}-\delta l_i}, \bar{v}^{n+1}_{\tilde{i}}, \bar{v}^{n+1}_{\tilde{i}+\delta l_i} + \Delta^1_{\tilde{i}+\delta l_i} \right) \\ \overset{(18),(19)}{\geq} & \tilde{H}_{l_i} \left(\bar{v}^{n+1}_{\tilde{i}-\delta l_i}, \bar{v}^{n+1}_{\tilde{i}}, \bar{v}^{n+1}_{\tilde{i}+\delta l_i} \right) . \end{aligned}$$

Note the arbitrary choice of m_1 and m_2 by a simultaneous change in the data corresponding to the index set $\hat{J}^n_m \setminus \{m_1, m_2\}$. Since there are also no limitations concerning the choices of $\hat{J}^n_m$ and l_i, the procedure is well defined and the proof is finished.

Proof (of Theorem 2)
Let a sequence $u^0 \in L_\infty$ be given. We then identify the finite values

$$a := \inf_{j \in J} u^0_j \quad \text{and} \quad b := \sup_{j \in J} u^0_j .$$

Since the source terms are pointwise bounded over the time interval $(0,T)$ — see assumptions (6) and (12), respectively — they are in both scenarios of interest especially bounded by a finite number M with

$$\int_0^T \|q\|_\infty \,\mathrm{d}t \,<\, M\,.$$

Consequently, by the assumed monotonicity follows that the numerical solution obtained via given data u_0 is bounded for all n with $n\Delta t < T$ by $a^n \le u^n \le b^n$ with

$$a_j^n := a - M\,(>-\infty)\;\forall j \in J \quad \text{and} \quad b_j^n := b + M\,(<\infty)\;\forall j \in J\,.$$

Proof (of Lemma 1)
Because the numerical scheme is consistent and conservative, the statement

$$G_l(v,v;k) = g_l(v\vee k, v\vee k) - g_l(v\wedge k, v\wedge k) = \operatorname{sgn}(v-k)[f_l(v)-f_l(k)]$$

holds by (22) for all $l = 1, \ldots, d$ and all $k \in \mathbb{R}$.

Proof (of Theorem 3)
Since the method is assumed to be consistent and conservative, there exist numerical flux functions g_l, $l = 1, \ldots, d$, so that one can construct numerical entropy fluxes G_l by applying Lemma 1. Thereby, the consistency with the entropy fluxes due to Kružkov is given. It is left to show the validity of a discrete entropy inequality. Therefore, let $k \in \mathbb{R}$ be chosen arbitrarily but fixed. By using the definition of G_l, we derive

$$\begin{aligned}
&-\sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l}\left\{ G_l\left(u_j^{n+1},\, u_{j+\delta l}^{n+1}; k\right) - G_l\left(u_{j-\delta l}^{n+1},\, u_j^{n+1}; k\right)\right\} \\
&= H\left(l\in\underline{d},\, u_{j-\delta l}^{n+1}\vee k,\, u_j^{n+1}\vee k,\, u_{j+\delta l}^{n+1}\vee k,\, u_j^{n}\vee k\right) \\
&\quad -H\left(l\in\underline{d},\, u_{j-\delta l}^{n+1}\wedge k,\, u_j^{n+1}\wedge k,\, u_{j+\delta l}^{n+1}\wedge k,\, u_j^{n}\wedge k\right) - \left|u_j^n - k\right|\,. \quad (25)
\end{aligned}$$

Now we estimate the terms involving H by using the monotonicity properties of the method. It is necessary to employ a diversion of the cases $u_j^{n+1} \ge k$ and $u_j^{n+1} < k$.

(a) *Case* $u_j^{n+1} \ge k$:

$$\begin{aligned}
&H\left(l\in\underline{d},\, u_{j-\delta l}^{n+1}\vee k,\, u_j^{n+1}\vee k,\, u_{j+\delta l}^{n+1}\vee k,\, u_j^{n}\vee k\right) \\
&\overset{\text{(a)}}{=} u_j^n \vee k - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l}\left\{ g_l\left(u_j^{n+1},\, u_{j+\delta l}^{n+1}\vee k\right) - g_l\left(u_{j-\delta l}^{n+1}\vee k,\, u_j^{n+1}\right)\right\} \\
&\quad +\Delta t q_j^{n+1}
\end{aligned}$$

$$\geq u_j^n - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l\left(u_j^{n+1},\, u_{j+\delta l}^{n+1}\right) - g_l\left(u_{j-\delta l}^{n+1},\, u_j^{n+1}\right) \right\} + \Delta t q_j^{n+1}$$

$$= u_j^{n+1} \overset{\text{(a)}}{=} u_j^{n+1} \vee k\,.$$

(b) *Case* $u_j^{n+1} < k$:

$$H\left(l \in \underline{d},\, u_{j-\delta l}^{n+1} \vee k,\, u_j^{n+1} \vee k,\, u_{j+\delta l}^{n+1} \vee k,\, u_j^n \vee k\right)$$

$$\overset{\text{(b)}}{=} u_j^n \vee k - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l\left(k,\, u_{j+\delta l}^{n+1} \vee k\right) - g_l\left(u_{j-\delta l}^{n+1} \vee k,\, k\right) \right\} + \Delta t q_j^{n+1}$$

$$\geq k - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l\left(k,\, k\right) - g_l\left(k,\, k\right) \right\} + \Delta t q_j^{n+1}$$

$$= k + \Delta t q_j^{n+1} \overset{\text{(b)}}{=} u_j^{n+1} \vee k + \Delta t q_j^{n+1}\,.$$

(c) *Case* $u_j^{n+1} \geq k$:

$$H\left(l \in \underline{d},\, u_{j-\delta l}^{n+1} \wedge k,\, u_j^{n+1} \wedge k,\, u_{j+\delta l}^{n+1} \wedge k,\, u_j^n \wedge k\right)$$

$$\overset{\text{(c)}}{=} u_j^n \wedge k - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l\left(k,\, u_{j+\delta l}^{n+1} \wedge k\right) - g_l\left(u_{j-\delta l}^{n+1} \wedge k,\, k\right) \right\} + \Delta t q_j^{n+1}$$

$$\leq k - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l\left(k,\, k\right) - g_l\left(k,\, k\right) \right\} + \Delta t q_j^{n+1}$$

$$= k + \Delta t q_j^{n+1} \overset{\text{(c)}}{=} u_j^{n+1} \wedge k + \Delta t q_j^{n+1}\,.$$

(d) *Case* $u_j^{n+1} < k$:

$$H\left(l \in \underline{d},\, u_{j-\delta l}^{n+1} \wedge k,\, u_j^{n+1} \wedge k,\, u_{j+\delta l}^{n+1} \wedge k,\, u_j^n \wedge k\right)$$

$$\overset{\text{(d)}}{=} u_j^n \wedge k - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l\left(u_j^{n+1},\, u_{j+\delta l}^{n+1} \wedge k\right) - g_l\left(u_{j-\delta l}^{n+1} \wedge k,\, u_j^{n+1}\right) \right\}$$
$$+ \Delta t q_j^{n+1}$$

$$\leq u_j^n - \sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ g_l\left(u_j^{n+1},\, u_{j+\delta l}^{n+1}\right) - g_l\left(u_{j-\delta l}^{n+1},\, u_j^{n+1}\right) \right\} + \Delta t q_j^{n+1}$$

$$= u_j^{n+1} \overset{\text{(d)}}{=} u_j^{n+1} \wedge k\,.$$

By combining all these cases, we obtain from (25) the inequality

$$-\sum_{l=1}^{d} \frac{\Delta t}{\Delta x_l} \left\{ G_l\left(u_j^{n+1},\, u_{j+\delta l}^{n+1}; k\right) - G_l\left(u_{j-\delta l}^{n+1},\, u_j^{n+1}; k\right) \right\}$$

$$
\begin{aligned}
&\quad +\operatorname{sgn}\left[u_j^{n+1}-k\right]\Delta t q_j^{n+1}\\
&\geq u_j^{n+1}\vee k-u_j^{n+1}\wedge k-\operatorname{sgn}\left[u_j^{n+1}-k\right]\Delta t q_j^{n+1}\\
&\quad +\operatorname{sgn}\left[u_j^{n+1}-k\right]\Delta t q_j^{n+1}-\left|u_j^{n}-k\right|\\
&=\left|u_j^{n+1}-k\right|-\left|u_j^{n}-k\right|\,.
\end{aligned}
$$

By construction, the procedure is well defined. Division by Δt gives the desired discrete entropy inequality.

In the case of *Scenario 2*, the validity of the corresponding discrete entropy inequality can be proven in the same way, resulting essentially from the monotonicity of the method. The difference between *Scenario 1* and *Scenario 2* is made up by substituting

$$
\sum_{l=1}^{d}\operatorname{sgn}\left[u_j^{n+1}-k\right]\left[q_j^{n+1}-f_{l_{x_l}}(j,n+1)\right]\quad\text{for}\quad\operatorname{sgn}\left[u_j^{n+1}-k\right]q_j^{n+1}
$$

$$
\text{with}\quad f_{l_{x_l}}(j,n+1):=f_{l_{x_l}}\left(\tilde{J}(j),(n+1)\Delta t,u_j^{n+1}\right)\,.
$$

4 Implicit numerical methods

This section contains the theoretical investigation of a few selected implicit methods. These are: (1) An implicit upwind scheme, (2) an implicit version of the Lax-Friedrichs scheme and (3) an implicit Godunov-type method.

4.1 An implicit upwind method

The implicit formulation of the upwind method reads

$$
u_j^{n+1}=u_j^{n}-\sum_{l=1}^{d}\frac{\Delta t}{\Delta x_l}\left\{f_l\left(u_j^{n+1}\right)-f_l\left(u_{j-\delta l}^{n+1}\right)\right\}+\Delta t q_j^{n+1}\,. \tag{26}
$$

We now employ the developed implicit notion of monotonicity.
To condition (18):

$$
\begin{aligned}
&\tilde{H}_l\left(a+\Delta a,\,b,\,c\right)-\tilde{H}_l\left(a,\,b,\,c\right)\\
&=\left[-\frac{\Delta t}{\Delta x_l}\left[f_l(b)-f_l(a+\Delta a)\right]\right]-\left[-\frac{\Delta t}{\Delta x_l}\left[f_l(b)-f_l(a)\right]\right]\\
&=\frac{\Delta t}{\Delta x_l}\left[f_l(a+\Delta a)-f_l(a)\right]\,.
\end{aligned}
$$

The condition (18) is fulfilled if f_l grows monotonically for all $l=1,\ldots,d$.
To condition (19):

$$
\begin{aligned}
&\tilde{H}_l\left(a,b,c+\Delta c\right)-\tilde{H}_l\left(a,b,c\right)\\
&=\left[-\frac{\Delta t}{\Delta x_l}\left[f_l(b)-f_l(a)\right]\right]-\left[-\frac{\Delta t}{\Delta x_l}\left[f_l(b)-f_l(a)\right]\right]=0\ (\geq 0)\,.
\end{aligned}
$$

Thus, the condition (19) is always fulfilled and the implicit upwind scheme is monotone if all the fluxes f_l grow monotonically. This is a nice property of the developed notions, since also the implicit scheme respects the direction of the flow. Note that the f_l do not need to be Lipschitz continuous to ensure the monotonicity of the scheme.

4.2 The implicit Lax-Friedrichs method

We investigate the implicit Lax-Friedrichs scheme

$$u_j^{n+1} = u_j^n + \sum_{l=1}^{d} \left\{ \frac{1}{2} \left[u_{j-\delta l}^{n+1} - 2u_j^{n+1} + u_{j+\delta l}^{n+1} \right] - \frac{\Delta t}{2\Delta x_l} \left[f_l(u_{j+\delta l}^{n+1}) - f_l(u_{j-\delta l}^{n+1}) \right] \right\} .$$

To condition (18):

$$\tilde{H}_l(a + \Delta a, b, c) - \tilde{H}_l(a, b, c) = \frac{1}{2}\Delta a + \frac{\Delta t}{2\Delta x_l} \left[f_l(a + \Delta a) - f_l(a) \right] . \qquad (27)$$

This expression is not positive or equal to zero without additional requirements.
To condition (19):

$$\tilde{H}_l(a, b, c + \Delta c) - \tilde{H}_l(a, b, c) = \frac{1}{2}\Delta c - \frac{\Delta t}{2\Delta x_l} \left[f_l(c + \Delta c) - f_l(c) \right] . \qquad (28)$$

Again this expression is not automatically positive or equal to zero. The requirements (27) and (28) can be combined to

$$\frac{|f_l(x + \Delta x) - f_l(x)|}{\Delta x_l} \leq \frac{\Delta x_l}{\Delta t} \quad \forall\, l = 1, \ldots, d \text{ and } \forall\, \Delta x \geq 0 .$$

Therefore, the implicit Lax-Friedrichs scheme is monotone only for Lipschitz-continuous flux functions with Lipschitz constants $L_l \leq (\Delta x_l / \Delta t)$. Note that this can also be read as a condition on the time step size which does not depend on the dimension, since each single one of the $2l$ conditions (18) and (19) has to be satisfied and no coupling is involved. This is quite surprising (a) because it is normally suggested that the numerical characteristics include the whole domain in the case of implicit methods, and (b) since no dimensional influence on the monotonicity property is obtained. In order to illuminate point (a), we briefly review the discussion of the situation for the case of the linear advection equation without sources in one dimension which is done in [5] in much more detail. With respect to point (b), we demonstrate numerically a similar behavior in two dimensions in order to illustrate the noted missing dimensional dependence of the implicit monotonicity criterion.

In the case of a linear flux $f(u) = vu$, the nonlinear system defined by the implicit Lax-Friedrichs scheme degenerates to a linear system with $\lambda = \Delta t/\Delta x$ given through

$$\left[-\frac{1}{2} - v\frac{\lambda}{2}\right] u_{j-1}^{n+1} + 2u_j^{n+1} + \left[-\frac{1}{2} + v\frac{\lambda}{2}\right] u_{j+1}^{n+1} = u_j^n \,. \tag{29}$$

We investigate the structure of the tridiagonal matrix $A = (a_{ij})$ defined by (29). Therefore, let v be positive with $v > (1/\lambda)$ so that the formal monotonicity property of the scheme is lost. Then the entries in the lower diagonal $a_{i+1,i}$ always take on negative values while the entries in the upper diagonal $a_{i,i+1}$ are always positive.

We at first eliminate the entries in the lower diagonal $a_{i+1,i}$. The diagonal entries of the matrix have to be modified accordingly, i.e. the diagonal entry in the i-th row is modified via

$$a_{ii}^{new} = a_{ii}^{old} - \frac{a_{i,i-1}}{a_{i-1,i-1}} a_{i-1,i} \,.$$

Thereby, note that we always have the situation

$$a_{i,i-1} < 0,\, a_{i-1,i-1} > 0 \,, \text{ and } \; a_{i-1,i} > 0 \,,$$

so that $a_{ii}^{new} > a_{ii}^{old}$ is always satisfied. Since the right hand side (b_i) of the investigated system incorporating the given data is modified via

$$b_i = u_i^n - \frac{a_{i,i-1}}{a_{i-1,i-1}} b_{i-1} \,,$$

data sets with $u_k^n \geq 0 \; \forall k$ imply only positive possible changes in the values b_i. In particular, the values in the upper diagonal $a_{i,i+1}$ remain unchanged and positive.

We now investigate what happens at a jump in given data u_k^n from values 0 to 1 when backward elimination is applied in order to solve the system. Therefore, we fix $u_j^n := 0 \, \forall j < i$ and $u_j^n := 1 \, \forall j \geq i$. By the described procedure, it is clear that the corresponding entries on the right hand side also show a jump from 0 to 1 after the modification due to elimination of the lower diagonal since $b_{i-1} = u_{i-1}^n = 0$, so that no positive update in b_i takes place. Backward elimination results in

$$u_{i-1}^{n+1} = \underbrace{\frac{1}{a_{i-1,i-1}^{new}}}_{>0} \Big(\underbrace{u_{i-1}^n}_{=0} - \underbrace{a_{i-1,i}}_{>0} \underbrace{u_i^n}_{=1} \Big) < 0 \,,$$

so that the monotonicity is violated, as expected. The violation of the monotonicity property can also be observed at jumps from high to lower values within given data.

Concerning the two-dimensional situation, we consider the linear advection equation

$$\frac{\partial}{\partial t} u(x,y,t) + \frac{\partial}{\partial x} \left(v u(x,y,t) \right) + \frac{\partial}{\partial y} \left(v u(x,y,t) \right) = 0$$

with grid parameters $\Delta x = \Delta y = 0.1$ and the initial condition

$$u(x,y,0) = \begin{cases} 1 \quad \text{for} \quad (x,y) \in [0,1] \times [0,1] \ , \\ 0 \ \text{else.} \end{cases}$$

The monotonicity condition yields that the chosen time step size $\Delta t = 0.1$ is the largest one allowed for $v = 1.0$ in order to preserve the monotonicity of the scheme, the same as would be in the one-dimensional case. See Fig. 1 for a visualization of the monotone and monotonicity-violating property of the method. Figure 2 gives a more detailed picture of the latter case.

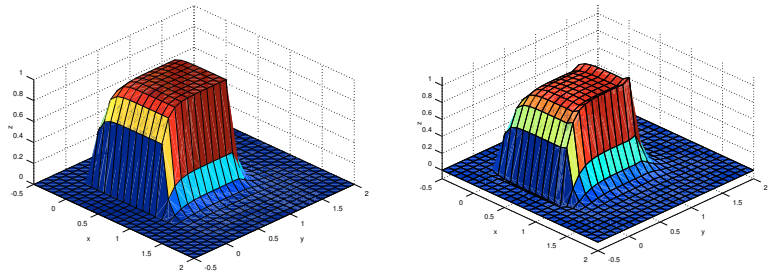

Fig. 1 Numerical solutions of the linear two-dimensional problem after one time step with $v = 1$ (left) satisfying the monotonicity condition and with $v = 1.5$ (right), resulting in a monotonicity violation as in the one-dimensional case. The same behavior also occurs for velocities $1 < v < 1.5$, resulting in much less amplitudes of the violations.

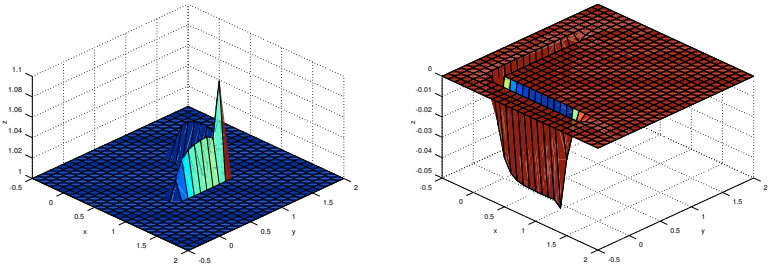

Fig. 2 Plots showing in detail the monotonicity violation in the case $\Delta t = \Delta x = \Delta y = 0.1$ and $v = 1.5$, obtained after one time step with the implicit Lax-Friedrichs scheme: The maximum of 1.0 and the numerical solution (left) and the minimum of 0.0 and the numerical solution (right).

4.3 An implicit Godunov-type method

In the scalar case, a closed form of the exact solution of a Riemann-problem was described by Osher [20]. Using this, a numerical scheme can be defined via the d numerical flux functions

$$g_l^G(v,w) \;=\; \begin{cases} \min_{v\le u\le w} f_l(u) : v \le w\,, \\ \max_{w\le u\le v} f_l(u) : v > w\,. \end{cases}$$

Since the relative values of the test variables have to be compared within the scheme, diversions by cases have to be employed.
To condition (18):
Generally, for $l = 1, \ldots, d$,

$$\tilde{H}_l(a+\Delta a, b, c) - \tilde{H}_l(a,b,c) = \frac{\Delta t}{\Delta x_l}\left[g_l^G(a+\Delta a, b) - g_l^G(a,b)\right]$$

holds. Since only the values b, a and $a+\Delta a$ are of importance, it is necessary to investigate three cases for each $l \in \underline{d}$.

1. Case: $b \le a \le a+\Delta a$

$$\frac{\Delta t}{\Delta x_l}\left[g_l^G(a+\Delta a,b) - g_l^G(a,b)\right] = \frac{\Delta t}{\Delta x_l}\left[\max_{b\le u\le a+\Delta a} f_l(u) - \max_{b\le u\le a} f_l(u)\right] \;\ge 0$$

2. Case: $a \le b \le a+\Delta a$

$$\frac{\Delta t}{\Delta x_l}\left[g_l^G(a+\Delta a,b) - g_l^G(a,b)\right] = \frac{\Delta t}{\Delta x_l}\left[\max_{b\le u\le a+\Delta a} f_l(u) - \min_{a\le u\le b} f_l(u)\right] \;\ge 0$$

3. Case: $a \le a+\Delta a \le b$

$$\frac{\Delta t}{\Delta x_l}\left[g_l^G(a+\Delta a,b) - g_l^G(a,b)\right] = \frac{\Delta t}{\Delta x_l}\left[\min_{a+\Delta a\le u\le b} f_l(u) - \min_{a\le u\le b} f_l(u)\right] \;\ge 0$$

Thus, the validity of the condition (18) is guaranteed without any additional condition on the flux function. This can be verified analogously for condition (19), so that the investigated Godunov-type scheme is monotone for general continuous flux functions.

4.4 Convergence

Within this section, we prove convergence of the mentioned schemes under the assumption that the conditions for monotonicity are fulfilled. We do this in some detail for the implicit upwind method, since this is demonstrated in the easiest fashion, and we refer to the differences concerning the proofs of convergence with respect to the other methods afterwards. The same holds true with respect to the type of sources employed in *Scenario 2*. Since part of the convergence proof is technically identical to the proofs in the one-dimensional case without sources described in [5], we refer to that work for more details.

The basic idea of the convergence proofs is the following. Corresponding to sequences $\Delta x_l^k \downarrow 0$ for $k \to \infty$, $l \in \underline{d}$, we construct a monotonically growing sequence of discrete initial data. Then by the monotonicity of the method we get a monotonically growing sequence of numerical solutions. Since we multiply the initial function u_0 with an arbitrarily chosen but fixed test function with compact support, we only have to consider u_0 over a finite domain. Because of the assumption $u_0 \in L_\infty$ and since we have L_∞-Stability, the corresponding function sequence is integrable and bounded from above. Then we can use the well-known theorem of monotone convergence of Beppo Levi to show convergence (almost everywhere) to a limit function. More formally, we state the following

Theorem 4 *Let $u_0(\mathbf{x})$ be in $L_\infty^{loc}\left(\mathbb{R}^d;\mathbb{R}\right)$. Consider a sequence of nested grids indexed by $k = 1, 2, \ldots$, with mesh parameters $\Delta t_k \downarrow 0$ and $\Delta x_l^k \downarrow 0$, $l = 1, \ldots, d$, as $k \to \infty$, and let $u_k(\mathbf{x}, t)$ denote the step function obtained via the numerical approximation by a consistent, conservative and monotone scheme in the form of the discussed methods. Then $u_k(\mathbf{x}, t)$ converges to the unique entropy solution of the given conservation law as $k \to \infty$.*

Proof At first, the convergence to a weak solution of the conservation law is established, followed by the verification that this weak solution is the entropy solution. For brevity of the notation, we omit the arguments $(\mathbf{x}, t)$ when appropriate.

We employ sequences $\Delta t_k \downarrow 0$ and $\Delta x_l^k \downarrow 0$, assuming that the resulting grids are nested in order to compare data sets of values, i.e. refined grids always inherit cell borders.

The most important technical detail is the special discretization of the initial condition $u_0 \in L_\infty^{loc}\left(\mathbb{R}^d;\mathbb{R}\right)$. After a suitable modification on a set of Lebesgue measure zero, the initial condition is discretized on cell $j \in J$, i.e. for

$$\mathbf{x} \in \left((j_1 - 1)\Delta x_1^0, j_1 \Delta x_1^0\right] \times \ldots \times \left((j_d - 1)\Delta x_d^0, j_d \Delta x_d^0\right] ,$$

by

$$u_j^0 := \inf_{\mathbf{x} \text{ in cell } j} u_0(\mathbf{x}) . \tag{30}$$

Corresponding to the initial data we also define a piecewise continuous function

$$u_k(\mathbf{x}, 0) := u_j^0, \ \mathbf{x} \text{ in cell } j . \tag{31}$$

It is a simple matter of classical analysis to verify that the discretization (30) together with (31) gives on any compact spatial domain a monotonically growing function sequence with

$$\lim_{k\to\infty} u_k(\mathbf{x}, 0) \ = \ u_0(\mathbf{x}) \text{ almost everywhere} \tag{32}$$

by application of the theorem of monotone convergence. In the classical fashion using point values, we extract discrete test elements ϕ_j^0 out of a given test function $\phi \in C_0^\infty\left(\mathbb{R}^{d+1};\mathbb{R}\right)$. Additionally, we define for $n \geq 1$ the step function

$$u_k(x, t) = u_j^n, \ \mathbf{x} \text{ in cell } j, \ t^{n-1} < t \leq t^n .$$

In the following, let the test function ϕ be chosen arbitrarily but fixed.

Multiplication of the implicit upwind scheme (26) with $\Delta t^k \prod_{l=1}^{d} \Delta x_l^k$ as well as with the discrete test element ϕ_j^{n+1}, summation over the spatial indices $j \in J$ and the temporal indices $n \geq 0$, and finally summation by parts yields

$$\begin{aligned}
&\Delta t^k \prod_{l=1}^{d} \Delta x_l^k \left\{ \sum_{j\in J} \sum_{n\geq 0} \left[u_j^n \frac{\phi_j^{n+1} - \phi_j^n}{\Delta t^k} + \sum_{l=1}^{d} f_l(u_j^{n+1}) \frac{\phi_{j+\delta l}^{n+1} - \phi_j^{n+1}}{\Delta x_l^k} \right] \right\} \\
&= -\prod_{l=1}^{d} \Delta x_l^k \sum_{j\in J} u_j^0 \phi_j^0 + \Delta t^k \prod_{l=1}^{d} \Delta x_l^k \sum_{j\in J} q_j^{n+1} \phi_j^{n+1} \,. \qquad (33)
\end{aligned}$$

By the definition of the introduced step functions, (33) is equivalent to

$$\begin{aligned}
&\int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \left[u_k(\mathbf{x},t) \frac{\phi_k(\mathbf{x},t+\Delta t^k) - \phi_k(\mathbf{x},t)}{\Delta t^k} \right. \\
&\left. + \sum_{l=1}^{d} f_l(u_k(\mathbf{x},t+\Delta t^k)) \frac{\phi_k(\mathbf{x}+\Delta x_l^k, t+\Delta t^k) - \phi_k(\mathbf{x},t+\Delta t^k)}{\Delta x_l^k} \right] \mathrm{d}\mathbf{x}\,\mathrm{d}t \\
&= -\int_{\mathbb{R}^d} u_k(\mathbf{x},0)\phi_k(\mathbf{x},0)\,\mathrm{d}\mathbf{x} \\
&+ \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} q_k(\mathbf{x},t+\Delta t^k)\phi_k(\mathbf{x},t+\Delta t^k)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \,. \qquad (34)
\end{aligned}$$

We now prove convergence of (34) to the form which implies that u is a weak solution of the original problem, see (7).

We first investigate the right hand side of (34). Set $\tilde{\Delta} := \max_{l\in \underline{d}} \Delta x_l^0$ and let

$$K := \left\{ (\mathbf{x},t) \,|\, \exists (\mathbf{y},t) \in \mathrm{supp}(\phi) : \, t = 0 \;\&\; y_l - \tilde{\Delta} \leq x_l \leq y_l + \tilde{\Delta} \;\forall\, l \in \underline{d} \right\} .$$

By construction, K is compact and gives the largest possible spatial domain where non-zero discrete initial data may occur. Adding zeroes, we now cast the problem into a more suitable form, namely

$$\begin{aligned}
\int_{\mathbb{R}^d} u_k(\mathbf{x},0)\phi_k(\mathbf{x},0)\,\mathrm{d}\mathbf{x} &= \int_K u_0(\mathbf{x})\phi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} \\
&+ \int_K u_k(\mathbf{x},0)\left[\phi_k(\mathbf{x},0) - \phi(\mathbf{x},0)\right]\mathrm{d}\mathbf{x} \\
&+ \int_K \left[u_k(\mathbf{x},0) - u_0(\mathbf{x})\right]\phi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} \,. \qquad (35)
\end{aligned}$$

Because of $u_0 \in L^\infty(\mathbb{R}^d;\mathbb{R})$ and by our construction, we can estimate the absolute value of the second right hand side term in (35) by the help of a constant $M_u < \infty$:

$$\left| \int_K u_k(\mathbf{x},0)\left[\phi_k(\mathbf{x},0) - \phi(\mathbf{x},0)\right]\mathrm{d}\mathbf{x} \right| \leq M_u |K| \sup_{x\in K} \left|\phi_k(\mathbf{x},0) - \phi(\mathbf{x},0)\right| . \qquad (36)$$

Since ϕ is a smooth test function, it is a simple but technical exercise to show

$$\|\phi_k(\mathbf{x},0) - \phi(\mathbf{x},0)\|_\infty \to 0 \quad \text{for} \quad k \to \infty\,. \tag{37}$$

By (36) and (37), the investigated term tends to zero with $k \to \infty$. Since ϕ is continuous and since $u_k(\mathbf{x},0)$ approaches $u_0(\mathbf{x})$ from below by construction, we can estimate the absolute of the third right hand side term in (35) with the help of a constant $M_\phi < \infty$ by

$$\left| \int_K \left[u_k(\mathbf{x},0) - u_0(\mathbf{x})\right] \phi(\mathbf{x},0)\,\mathrm{d}\mathbf{x} \right| \leq M_\phi \int_K u_0(\mathbf{x}) - u_k(\mathbf{x},0)\,\mathrm{d}\mathbf{x}\,.$$

The theorem of monotone convergence implies that

$$\int_K u_0(\mathbf{x}) - u_k(\mathbf{x},0)\,\mathrm{d}\mathbf{x}$$

vanishes in the limit for $k \to \infty$, i.e. the corresponding term in (35) goes to zero for $k \to \infty$. To condense these results, we obtain

$$\lim_{k\to\infty} \int_{\mathbb{R}^d} u_k(\mathbf{x},0)\phi_k(\mathbf{x},0)\,\mathrm{d}\mathbf{x} = \int_{\mathbb{R}^d} u_0(\mathbf{x})\phi(\mathbf{x},0)\,\mathrm{d}\mathbf{x}\,.$$

It remains to show

$$\int_{\mathbb{R}_+}\int_{\mathbb{R}^d} q_k(\mathbf{x},t+\Delta t^k)\phi_k(\mathbf{x},t+\Delta t^k)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \overset{k\to\infty}{\longrightarrow} \int_{\mathbb{R}_+}\int_{\mathbb{R}^d} q(\mathbf{x},t)\phi(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t\,.$$

This result can easily be achieved by analogously introducing a compact domain $S \subset \mathbb{R}^d$ including the support of ϕ in space and time, setting for $n \geq 1$ ($n = 0$ is not relevant since $q(\cdot,0) \equiv 0$)

$$q_j^n := \inf_{(\mathbf{x},t) \text{ with } \mathbf{x} \text{ in cell } j \text{ and } t \text{ in } (t^n - \Delta t^0, t^n]} q(\mathbf{x},t)$$

and using a similar manipulation as for the terms involving u_0.

Concerning the left hand side of (35), adding zeroes and using the attributes of test functions together with the L_∞-stability of u_k yields that we finally have to show

$$\lim_{k\to\infty} \int_S \left|u(\mathbf{x},t) - u_k(\mathbf{x},t)\right| \left|\phi_t(\mathbf{x},t)\right|\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \overset{k\to\infty}{\longrightarrow} 0 \tag{38}$$

and also for all $l \in \underline{d}$

$$\lim_{k\to\infty} \int_S \left|f_l(u(\mathbf{x},t)) - f_l(u_k(\mathbf{x},t+\Delta t^k))\right| \left|\frac{\partial}{\partial x_l}\phi(\mathbf{x},t)\right|\,\mathrm{d}\mathbf{x}\mathrm{d}t \overset{k\to\infty}{\longrightarrow} 0 \tag{39}$$

in order to prove convergence to a weak solution. Since ϕ_t is continuous on S, we can estimate $|\phi_t|$ in (38) by a constant $M_t < \infty$. Since $u_k(\mathbf{x},t)$ grows monotonically with $k \to \infty$ in the sense of pointwise comparison, and since it is positive and bounded from above because of $u_0 \in L_\infty(S)$ and the monotonicity of the method, the function sequence $(u_k(\mathbf{x},t))_{k\in\mathbf{N}}$ converges almost

everywhere to an integrable limit function on S by the theorem of monotone convergence due to Levi. We set

$$u(\mathbf{x},t) := \lim_{k\to\infty} u_k(\mathbf{x},t)\,.$$

Introducing exactly this limit function as the function $u(\mathbf{x},t)$ used up to now, the corresponding term in (38) becomes zero in the limit:

$$\begin{aligned}&\lim_{k\to\infty}\int_S \big|u(\mathbf{x},t)-u_k(\mathbf{x},t)\big|\,|\phi_t(\mathbf{x},t)|\;\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &\le M_t\int_S u(\mathbf{x},t)-\lim_{k\to\infty}u_k(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t=0\,.\end{aligned}$$

Note that the pointwise convergence $u_k \to u$ almost everywhere is now established and can be used in the following proofs. For proving (39), we need some further simple manipulations. We use again the continuity of the derivatives of ϕ to introduce constants $M_x^l < \infty$ to obtain

$$\begin{aligned}&\int_S \big|f_l(u(\mathbf{x},t))-f_l(u_k(\mathbf{x},t+\Delta t^k))\big|\left|\frac{\partial}{\partial x_l}\phi(\mathbf{x},t)\right|\;\mathrm{d}\mathbf{x}\,\mathrm{d}t\le\\ &M_x^l\int_S |f_l(u_k(\mathbf{x},t))-f_l(u(\mathbf{x},t))|\;\mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &+M_x^l\int_S \big|f_l(u_k(\mathbf{x},t+\Delta t^k))-f_l(u_k(\mathbf{x},t))\big|\;\mathrm{d}\mathbf{x}\,\mathrm{d}t\end{aligned}\tag{40}$$

for all $l \in \underline{d}$. We now discuss the first right hand side term in (40). Since by construction u_k and u are in $L_\infty(S)$, we can estimate every

$$\big|f_l(u_k(\mathbf{x},t+\Delta t^k))-f_l(u_k(\mathbf{x},t))\big|$$

over S from above by a constant $M_f^l < \infty$ because of the continuity of the f_l on the compact set of possible values. Then the functions

$$M_f^l(\mathbf{x},t) := \begin{cases} M_f^l\;, & (\mathbf{x},t)\in S\\ 0\;, & \text{otherwise}\end{cases},$$

are in $L_1(\mathbb{R}^d\times\mathbf{R}_+;\mathbb{R})$ and dominate $|f_l(u_k(\mathbf{x},t))-f_l(u(\mathbf{x},t))|$ for all $l\in\underline{d}$ and all k. Because of the established pointwise convergence $u_k\to u$ a.e., we can apply the theorem of dominated convergence by Lebesgue to obtain for all l

$$\lim_{k\to\infty} M_x\int_S |f_l(u_k(\mathbf{x},t))-f_l(u(\mathbf{x},t))|\;\mathrm{d}\mathbf{x}\,\mathrm{d}t=0\,.\tag{41}$$

Now we discuss the second right hand side term in (40). Since by construction u_k is a step function with finite values on the compact domain S, u_k is in $L_1(S)$. Since the f_l are continuous, also $f_l\circ u_k$ are in $L_1(S)$. By the continuity in the mean of L_1-functions, there exist $\delta_l(\epsilon)$ for all $\epsilon>0$ with

$$\int_S \big|f_l(u_k(\mathbf{x},t+\Delta t^k))-f_l(u_k(\mathbf{x},t))\big|\;\mathrm{d}\mathbf{x}\,\mathrm{d}t\;<\;\epsilon\,,$$

if $\Delta t^k < \delta_l(\epsilon)$. Since $\Delta t^k \downarrow 0$ for $k \to \infty$, ϵ can be chosen arbitrarily small, i.e.

$$M_x \int_S \left| f_l(u_k(\mathbf{x}, t + \Delta t^k)) - f_l(u_k(\mathbf{x}, t)) \right| \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \; \to \; 0 \quad \text{for} \quad k \to \infty \tag{42}$$

holds for all $l \in \underline{d}$. By (41) and (42) the assertion in (39) is proven. Since the test element ϕ was chosen arbitrarily, convergence to a weak solution is established.

We have now to show that exactly this weak solution is the unique entropy solution in the sense of Kružkov. Therefore, we derive in a similar fashion as in the derivation of (33) the weak form of the discrete entropy condition (21) connected with the implicit upwind scheme using Lemma 1 and Theorem 2. It reads

$$\begin{aligned}
&-\Delta t^k \prod_{l=1}^{d} \Delta x_l^k \sum_{j \in J} |u_j^0 - k| \phi_j^0 \\
&-\Delta t^k \prod_{l=1}^{d} \Delta x_l^k \sum_{j \in J} \sum_{n \geq 0} \operatorname{sgn}\left(u_j^{n+1} - k\right) q_j^{n+1} \phi_j^{n+1} \\
\leq\; &\Delta t^k \prod_{l=1}^{d} \Delta x_l^k \sum_{j \in J} \sum_{n \geq 0} \left[|u_j^n - k| \frac{\phi_j^{n+1} - \phi_j^n}{\Delta t^k} \right. \\
&\left. + \operatorname{sgn}\left(u_j^{n+1} - k\right) \sum_{l=1}^{d} \left\{ \left[f_l(u_j^{n+1}) - f_l(k) \right] \frac{\phi_{j+1}^{n+1} - \phi_j^{n+1}}{\Delta x_l^k} \right\} \right] .
\end{aligned} \tag{43}$$

Using the established convergence $u_k \to u$ a.e. of the function sequence generated by the numerical method for $\Delta t^k \downarrow 0$ and $\Delta x_l^k \downarrow 0$ for all $l \in \underline{d}$, we now prove convergence of (43) towards the form of the entropy condition due to Kružkov (8). Therefore, we have to consider arbitrarily chosen but fixed test elements composed of a test function ϕ with $\phi \geq 0$, $\phi \in C_0^\infty(\mathbb{R}^{d+1}; \mathbb{R})$, and a test number $k \in \mathbb{R}$.

Using the same notation and applying a similar procedure as in the case of the convergence proof to a weak solution, we first want to prove

$$\lim_{k \to \infty} M_\phi \int_K \left| |u_k(\mathbf{x}, 0) - k| - |u_0(\mathbf{x}) - k| \right| \mathrm{d}\mathbf{x} = 0 \, . \tag{44}$$

Since k is fixed and $u_k(\mathbf{x}, 0)$ and u_0 bounded, one can find a constant function over the compact interval K which dominates the integrand for all k. Then (44) follows by the use of the already established convergence $u_k(x, 0) \to u_0(x)$ a.e. and the theorem of dominated convergence by Lebesgue. We also have to treat

$$\lim_{k \to \infty} \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \operatorname{sgn}\left(u_k(\mathbf{x}, t + \Delta t^k) - k\right) q_k(\mathbf{x}, t + \Delta t^k) \phi(\mathbf{x}, t + \Delta t^k) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t \, .$$

Therefore, we expand the factor $\phi(\mathbf{x}, t + \Delta t^k)$ by adding zeroes in the form

$$\phi(\mathbf{x}, t + \Delta t^k) = \phi(\mathbf{x}, t + \Delta t^k) - \phi(\mathbf{x}, t) + \phi(\mathbf{x}, t) \, .$$

Convergence of the integrals involving the factor $\phi(\mathbf{x}, t + \Delta t^k) - \phi(\mathbf{x}, t)$ tend to zero. This follows by estimating sgn, u_k and q_k from above and using the usual properties of test functions. In a similar fashion, we expand the factor $\operatorname{sgn}\big(u_k(\mathbf{x}, t + \Delta t^k) - k\big)$, adding zero in the form $-\operatorname{sgn}(u_k(\mathbf{x}, t) - k) + \operatorname{sgn}(u_k(\mathbf{x}, t) - k)$. The proof that the integrals involving the expression of the form $\operatorname{sgn}\big(u_k(\mathbf{x}, t + \Delta t^k) - k\big) - \operatorname{sgn}(u_k(\mathbf{x}, t) - k)$ vanish follows from the continuity in the mean of L_1-functions. Again similarly, we expand in the form $q_k(\mathbf{x}, t + \Delta t^k) = q_k(\mathbf{x}, t + \Delta t^k) - q(\mathbf{x}, t + \Delta t^k) + q(\mathbf{x}, t + \Delta t^k)$ and use the concept of monotone convergence due to Beppo Levi to obtain convergence to zero of the integrals involving $q_k(\mathbf{x}, t + \Delta t^k) - q(\mathbf{x}, t + \Delta t^k)$. Lastly, the proof of convergence of $q(\mathbf{x}, t + \Delta t^k)$ to $q(\mathbf{x}, t)$ under the integral follows from the continuity in the mean of L_1-functions. The technical details only require to take all expansions obtained by taking suitable zeroes into account and eliminating all integrals which involve discrete notions. The other terms left to investigate are

$$\int_S |u_k(\mathbf{x}, t) - k|\, \phi_t(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \quad \text{and}$$
$$\int_S \sum_{l=1}^{d} \operatorname{sgn}\big[u_k(\mathbf{x}, t + \Delta t^k) - k\big]\, \big[f_l(u_k(\mathbf{x}, t + \Delta t^k)) - f_l(k)\big]\, \frac{\partial}{\partial x_l} \phi(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t\,.$$

The procedure is the same in both cases. Since the occurring derivatives of ϕ are continuous, we can estimate these over the compact domain S by finite constants. Since k is a fixed value (and so is $f_l(k)$ for all $l \in \underline{d}$), since $u_k(\mathbf{x}, t)$ is bounded and because the f_l are continuous over the bounded interval of possible values of u_k (due to the established L_∞-stability), we can also give constants which estimate all the expressions involving u_k from above. Using the product of these finite constants as dominating function over S as well as $u_k \to u$ a.e., we employ the theorem of dominated convergence to receive the desired result for the implicit upwind scheme.

In the case of the implicit Lax-Friedrichs method, we have to assume Lipschitz continuity with a Lipschitz constant $L \leq 1/\lambda$ of the flux functions so that the method is monotone. In comparison to the implicit upwind method, the difference in the corresponding weak forms are made up from

$$-\frac{\Delta t_k}{2} \int_S u_k(\mathbf{x}, t + \Delta t^k) \hat{\phi}\, \mathrm{d}\mathbf{x}\, \mathrm{d}t \text{ and } -\frac{\Delta t_k}{2} \int_S \big|u_k(\mathbf{x}, t + \Delta t^k) - k\big|\, \hat{\phi}\, \mathrm{d}\mathbf{x}\, \mathrm{d}t\,,$$

respectively. Thereby, $\hat{\phi}$ converges in the L_∞-Norm to $\partial^2_{x_l^2} \phi(\mathbf{x}, t)$ which is continuous since $\phi \in C_0^\infty(\mathbb{R}^{d+1}; \mathbb{R})$. Thus, the corresponding term can be estimated from above by a constant over the compact domain S. Since k is fixed and $u_k(\mathbf{x}, t)$ is bounded as usual, both expressions vanish with $\Delta t^k \downarrow 0$.

With respect to the described implicit Godunov-type method, we use a similar procedure as in the case of the implicit Lax-Friedrichs methods, namely to write down the differences in the weak forms to the case of the implicit

upwind method. These are made up from

$$\Big\{ \big[g_l^G(u_j^{n+1} \vee k, u_{j+\delta l}^{n+1} \vee k) - g_l^G(u_j^{n+1} \wedge k, u_{j+\delta l}^{n+1} \wedge k) \big] \\ -\operatorname{sgn}\big(u_j^{n+1} - k\big) \big[f_l(u_j^{n+1}) - f_l(k) \big] \Big\} \frac{\phi_{j+\delta l}^{n+1} - \phi_j^{n+1}}{\Delta x_l^k} \tag{45}$$

$$\text{and} \quad \big[g_l^G(u_j^{n+1}, u_{j+1}^{n+1}) - f_l(u_j^{n+1}) \big] \frac{\phi_{j+\delta l}^{n+1} - \phi_j^{n+1}}{\Delta x_l^k} \,. \tag{46}$$

Since g_G is continuous in the components and $u_k(\mathbf{x}, t) \in L_1(S)$, the $g_l^G \circ u_k$ are also in $L_1(S)$. After introducing step functions as usual, the expressions incorporating g_l^G from (45) give values $f_l(\xi_l)$ with

$$\begin{aligned} \xi_l &\in \big[u_k(\mathbf{x}, t + \Delta t^k), u_k(\mathbf{x} + \Delta x_l^k, t + \Delta t^k) \big] \\ \text{or } \xi_l &\in \big[u_k(\mathbf{x} + \Delta x_l^k, t + \Delta t^k), u_k(\mathbf{x}, t + \Delta t^k) \big] \,, \end{aligned}$$

respectively. The integrals over the terms corresponding to (45) then go to zero with $k \to \infty$ because of the continuity in the mean of $g_l^G \circ u_k$. The idea for proving convergence to zero concerning the integral of the expressions corresponding to (46) is the same.

Concerning *Scenario 2*, the described strategy is fully transferable by employing accordingly the notions developed in section 3.

5 Numerical tests

In order to show the applicability of the developed notions, we investigate numerically a number of test cases which were employed within the literature in various contexts. In contrast to the cited examples from [21,12] we do not employ any further manipulations of the problem or on the numerical side, we simply rely on the straightforward application of the implicit Godunov-type method in all cases.

We remark that the notions we developed reduce in the one-dimensional case without sources to the notions described within [4,5]. In that works, especially the applicability of implicit schemes with respect to a conservation law given in [15] was shown where the solution features a rarefaction wave extending in an arbitrarily small time step to infinity. Thus, the applicability of the described concept is already established in the case without sources, where the flux is merely continuous and where a meaningful CFL-condition does not exist. With the numerical tests documented in this section, we focus on the theoretical extensions developed in this paper.

First, we employ a one-dimensional conservation law featuring as a particular problem a point source depending on space and time. This test case was used in [21] to show experimentally convergence to the entropy solution. In contrast to the scheme used in their work, for our scheme convergence to the entropy solution is guaranteed.

At second, we consider a couple of model problems featuring spatial dependent sources having the form of the derivative of certain functions. These model problems were used in [12] in order to show numerically convergence to steady state solutions featuring various difficulties which is in contrast to the first unsteady example. Moreover, since we use the described Godunov-type scheme, we do not rely on a CFL-like condition as in [12] which greatly restricts the time step size, an annoying aspect in steady state calculations. Note also that we simply employ the described implicit method without any further improvements as done with the explicit method employed in [12] for which convergence was not guaranteed.

While the first two examples could be identified as belonging to *Scenario 1*, the third test case refers to *Scenario 2*. It consists of a one-dimensional model problem featuring a parameter dependent source term depending also in a nonlinear way on the solution. This model problem was used by LeVeque and Yee [18] to illuminate numerical difficulties in the case of stiffness.

As fourth and last example, we show numerical results of a two-dimensional problem used in [11] which exhibits all principal difficulties encountered when dealing with hyperbolic equations. As in the second example, the implicitness of our methods is advantageous in order to calculate the steady state solution.

In all examples, nonlinear systems of equations of the form $F(x) = 0$ with $F(x) : \mathbb{R}^n \to \mathbb{R}^m$ for some $n, m \geq 1$ arise which have been solved numerically with an iterative solver. Precisely, we used the Matlab subroutine *fsolve* of the optimization toolbox (www.mathworks.com/help/optim/ug/fsolve.html). By default the Powell's dogleg algorithm (a trust-region method) is used (see www.mathworks.com/help/optim/ug/equation-solving-algorithms.html for a detailed description of how this method are defined and how they work).

5.1 Example 1

The one-dimensional scalar conservation law under consideration is

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}u(x,t) = \sin(\pi t)\delta(x-0.1)\,, \quad x \in (0,1)\,,\ t > 0\,,$$

$$\text{with} \quad u_0(x) = u(x,0) = 0\ \forall x \in (0,1) \quad \text{and} \quad u(0,t) = 0\ \forall t \geq 0\,.$$

The exact solution is given in [21] and reads

$$u(x,t) \ = \ \begin{cases} u_0(x-t) & : x < 0.1 \text{ or } x \geq 0.1 + t\,, \\ \sin\left(\pi(0.1+t-x)\right) + u_0(x-t) : & 0.1 \leq x < 0.1 + t\,. \end{cases}$$

By Fig. 3, we can compare the exact and numerical solutions obtained with the implicit upwind method in the same situations as displayed in [21], using also exactly the same grid parameters. They used $\Delta x = \Delta t = 1/20$ and $\Delta x = \Delta t = 1/40$ in the three moments $t = 1/4$, $t = 1/2$, and $t = 1$.

Relating to the method used in [21], our scheme is overall much more viscous. This is as expected since Santos and Oliveira especially sought a good

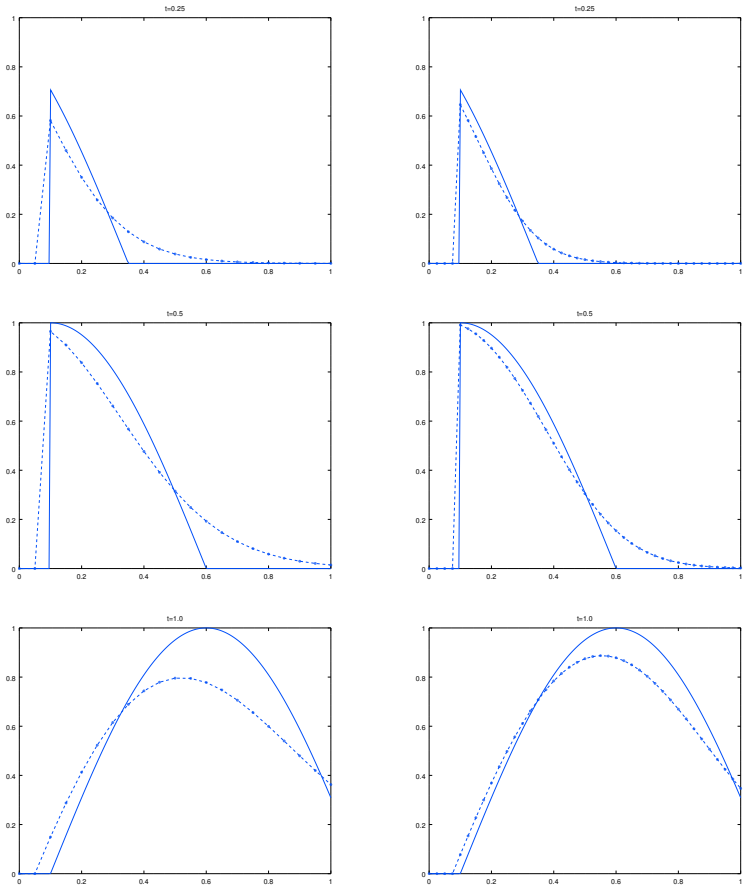

Fig. 3 The columns show the numerical solutions (dashed lines) in comparison with the exact solution (continuous lines). The situations displayed in the right column are obtained using a grid twice as fine as in the left column.

accuracy of their method. We also observe experimentally convergence to the correct solution by our method, documented by the bottom pictures within Fig. 4 showing results of analogous computations with a more refined grid.

5.2 Example 2

The conservation law generally under consideration is

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}\left(\frac{1}{2}u(x,t)^2\right) = q_x(x)\,, \quad x \in \mathbb{R}\,,\ t > 0\,,$$

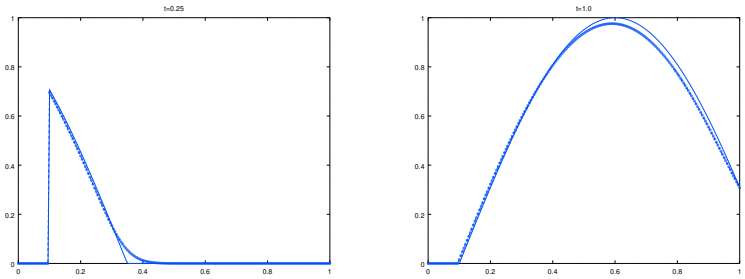

Fig. 4 The first and third state from Fig. 3 revisited, this time obtained via numerical approximation using a grid ten times as fine as in the left column of Fig. 3.

which is used featuring different sources and initial conditions resulting in various difficulties. The source terms in use are

$$q(x) = \begin{cases} 0\,, & x < -1\,, \\ \cos^2(\pi x/2)\,, & -1 \le x \le 1\,, \\ 0\,, & 1 < x\,, \end{cases} \quad \& \, q(x) = \begin{cases} 0\,, & x < -1\,, \\ -\cos^2(\pi x/2)\,, & -1 \le x \le 1\,, \\ 0\,, & 1 < x\,. \end{cases}$$

The initial conditions in use are

$$u(x,0^+) = 0, \qquad -\infty < x < \infty\,, \qquad u(x,0^+) = \begin{cases} 0\,, & x < -1\,, \\ 1\,, & -1 \le x \le 1\,, \\ 0\,, & 1 < x\,, \end{cases} \text{ and}$$

$$u(x,0^+) = \begin{cases} 0\,, & x < -1\,, \\ -1\,, & -1 \le x \le 1\,, \\ 0\,, & 1 < x\,. \end{cases}$$

The following four experiments are analogous to the ones in [12], using exactly the same grid parameters as initially in [12] where later on spatial regridding was used in order to obtain sharp shock profiles. We use the implicit Lax-Friedrichs method with $\delta x = 0.025$ and $\delta t = 0.0125$.

In the Figs. 5 and 6, we show in all test cases from top to bottom the numerical solutions obtained by using our method at times $t = 0.2$, 0.5, 1.0 and 3.0 (line featuring small circles) together with the stationary solution (continuous line). Thereby, different experiments correspond to different columns of pictures.

Concerning the first experiment, the numerical solution is almost identical to the exact one except at the point $x = 0.0$ where a grid point is located exactly on the shock front. With respect to the other experiments, the numerical solutions exhibit slightly smeared shocks while they are otherwise quite accurate. Comparing with the numerical results shown in [12], not employing a regridding procedure results in slightly smeared shocks. Further numerical experiments have shown that we can employ much larger time steps — usually of about 20–30 times the one used for the presented experiments — without degrading our numerical solution.

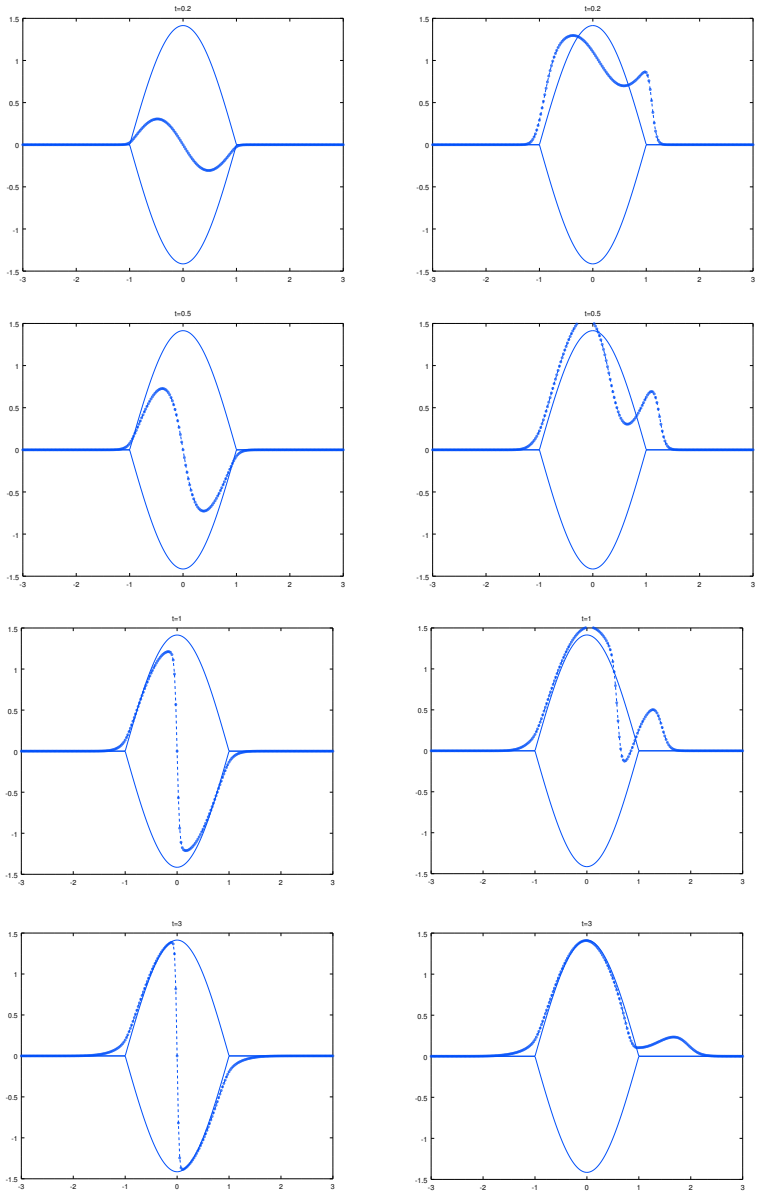

Fig. 5 Experiments one (left) and two (right).

5.3 Example 3

The conservation law under consideration is

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}u(x,t) = -\mu u(x,t)\big(u(x,t)-1\big)\left(u(x,t)-\frac{1}{2}\right)$$

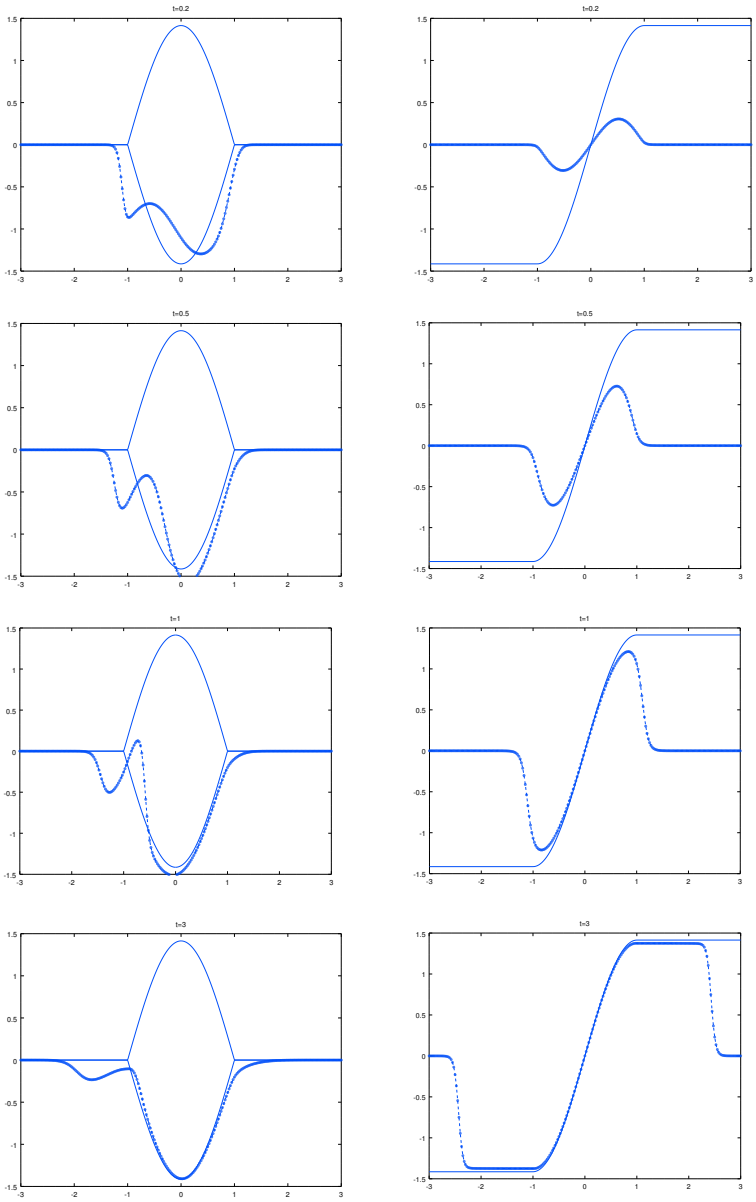

Fig. 6 Experiments three (left) and four (right).

which exhibits a nonlinear source term with an increasingly stiff behavior for μ growing large.

Our numerical investigations are completely analogous to the ones in [18]. The experiments consist of the numerical solution of a Riemann problem whose

exact solution features a shock front moving from $x = 0.3$ to $x = 0.6$ after a couple of time steps, see Fig. 7. For small and medium μ, the numerical solution shows the correct behavior incorporating numerical viscosity; note the sharpening and slight translation of the shock approximation for $\mu = 100$, an effect of the increasing stiffness of the source term. For $\mu = 1000$ the usual problem is faced, see [18] for details. This experiment shows that although the discussed methods are generally capable to deal with non-linear sources, they are not recommended without modification for stiff problems even though they are fully implicit. Of course, grid refinement results in the approximation of the correct solution as expected.

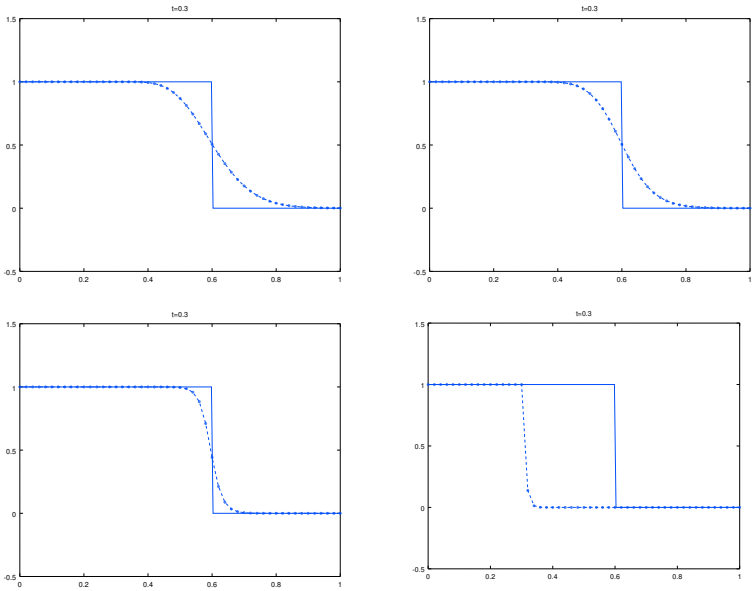

Fig. 7 Numerical solutions of a conservation law exhibiting a nonlinear source term with an increasingly stiff behavior for a large parameter μ (dashed lines) compared with the exact solution (continuous lines) with a step function as the initial condition. The pictures correspond from left to right and top to bottom to the choices $\mu = 1, 10, 100, 1000$.

The reason why we show the four graphs in Fig. 7 is to directly compare them with [18, Figure 2]. We are able to obtain well-behaved solutions whereas the method in [18] uses limiter to avoid oscillations close to the discontinuity at $x = 0.6$.

6 Summary and conclusive remarks

In this paper, we have introduced a new concept for implicit methods for scalar conservation laws in one or more spatial dimensions which may also include

source terms of different type. We developed implicit notions that are centered around a monotonicity criterion and show the relation between a numerical scheme and a discrete entropy inequality. We investigate in detail three implicit methods and give a convergence proof. Hence, we extend the rigorously verified range of applicability of those three implicit numerical methods. By numerical experiments we have shown the validity and usefulness of our theoretical results.

References

1. Beam, R.M., Warming, R.F.: An implicit finite-difference algorithm for hyperbolic systems in conservation-law form. Journal of Computational Physics **22**(1), 87–110 (1976)
2. Bénilan, P., Kružkov, S.: Conservation laws with continuous flux functions. Nonlinear Differential Equations and Applications NoDEA **3**(4), 395–419 (1996)
3. Bouchut, F.: Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws. Frontiers in Mathematics. Birkhäuser Verlag, Basel (2004)
4. Breuß, M.: Numerical methods for conservation laws in non-standard-situations. Doctoral thesis, University of Hamburg (2001)
5. Breuß, M.: The implicit upwind method for 1-d scalar conservation laws with continuous fluxes. SIAM J. Numer. Anal. **43**(3), 970–986 (2005)
6. Coquel, F., Floch, P.L.: Convergence of finite difference schemes for conservation laws in several space dimensions: A general theory. SIAM Journal on Numerical Analysis **30**(3), 675–700 (1993)
7. Crandall, M.G., Majda, A.: Monotone difference approximations for scalar conservation laws. Mathematics of Computation **34**(149), 1–21 (1980)
8. Dey, S.K.: An implicit numerical solver for nonlinear hyperbolic partial differential equations. Computers & Mathematics with Applications **11**(1), 79–91 (1985)
9. DiPerna, R.J.: Measure-valued solutions to conservation laws. Archive for Rational Mechanics and Analysis **88**(3), 223–270 (1985)
10. Godlewski, E., Raviart, P.A.: Hyperbolic systems of conservation laws. Collection Mathématiques et Applications de la SMAI. Ellipses, Paris (1991)
11. Grahs, T., Meister, A., Sonar, T.: Image processing for numerical approximations of conservation laws: Nonlinear anisotropic artificial dissipation. SIAM Journal on Scientific Computing **23**(5), 1439–1455 (2002)
12. Greenberg, J.M., LeRoux, A.Y., Baraille, R., Noussair, A.: Analysis and approximation of conservation laws with source terms. SIAM Journal on Numerical Analysis **34**(5), 1980–2007 (1997)
13. Kružkov, S.N.: First order quasilinear equations in several independent variables. Mathematics of the USSR-Sbornik **10**(2), 217–243 (1970)
14. Kružkov, S.N., Hildebrand, F.: The cauchy problem for quasilinear first order equations in the case the domain of dependence on initial data is infinite. Moscow Univ. Math. Bull. **29**(5), 75–81 (1974)
15. Kružkov, S.N., Panov, E.Y.: Conservative quasilinear first-order laws with an infinite domain of dependence on the initial data. Soviet Math. Doklady **42**, 316–321 (1991)
16. Kuznetsov, N.N.: Accuracy of some approximate methods for computing the weak solutions of a first order quasi-linear equation. USSR Comp. Math. and Math. Phys. **16**(6), 105–119 (1976)
17. LeVeque, R.J.: Numerical Methods for Conservation Laws. Lectures in Mathematics ETH Zürich, Department of Mathematics Research Institute of Mathematics. Birkhäuser Verlag, Basel (1992)
18. LeVeque, R.J., Yee, H.C.: A study of numerical methods for hyperbolic conservation laws with stiff source terms. Journal of Computational Physics **86**(1), 187–210 (1990)
19. Meister, A.: Comparison of different Krylov subspace methods embedded in an implicit finite volume scheme for the computation of viscous and inviscid flow fields on unstructured grids. Journal of Computational Physics **140**(2), 311–345 (1998)

20. Osher, S.: Riemann solvers, the entropy condition, and difference approximations. SIAM Journal on Numerical Analysis **21**(2), 217–235 (1984)
21. Santos, J., de Oliveira, P.: A converging finite volume scheme for hyperbolic conservation laws with source terms. Journal of Computational and Applied Mathematics **111**(1–2), 239–251 (1999)