J Sci Comput (2011) 48:274–295
DOI 10.1007/s10915-010-9425-7

On the Convergence and Well-Balanced Property of Path-Conservative Numerical Schemes for Systems of Balance Laws

María Luz Muñoz-Ruiz · Carlos Parés

Received: 29 January 2010 / Revised: 24 August 2010 / Accepted: 4 October 2010 /
Published online: 19 October 2010

Abstract This paper deals with the numerical approximation of one-dimensional hyperbolic systems of balance laws. We consider these systems as a particular case of hyperbolic systems in nonconservative form, for which we use the theory introduced by Dal Maso, LeFloch and Murat (J. Math. Pures Appl. 74:483, 1995) in order to define the concept of weak solutions. This theory is based on the prescription of a family of paths in the phases space. We also consider path-conservative schemes, that were introduced in Parés (SIAM J. Numer. Anal. 44:300, 2006). The first goal is to prove a Lax-Wendroff type convergence theorem. In Castro et al. (J. Comput. Phys. 227:8107, 2008) it was shown that, for general nonconservative systems a rather strong convergence assumption is needed to prove such a result. Here, we prove that the same hypotheses used in the classical Lax-Wendroff theorem are enough to ensure the convergence in the particular case of systems of balance laws, as the numerical results shown in Castro et al. (J. Comput. Phys. 227:8107, 2008) seemed to suggest. Next, we study the relationship between the well-balanced properties of path-conservative schemes applied to systems of balance laws and the family of paths.

Keywords Hyperbolic systems of balance laws · Hyperbolic nonconservative systems · Path-conservative schemes · Convergence · Well-balanced schemes

1 Introduction

This paper is concerned with the numerical approximation of Cauchy problems for one-dimensional systems of conservation laws with source terms or balance laws

$$U_t + F(U)_x = S(U)\sigma_x, \tag{1.1}$$

M.L. Muñoz-Ruiz (✉)
Dept. Matemática Aplicada, Universidad de Málaga, 29071 Málaga, Spain
e-mail: munoz@anamat.cie.uma.es

C. Parés
Dept. Análisis Matemático, Universidad de Málaga, 29071 Málaga, Spain
e-mail: pares@anamat.cie.uma.es

where $U(x,t) \in \Omega$, being Ω an open convex set of $\mathbb{R}^N$, $\sigma(x)$ is a known function from $\mathbb{R}$ to $\mathbb{R}$, F is a regular function from Ω to $\mathbb{R}^N$, and S is also a function from Ω to $\mathbb{R}^N$.

It is well known that standard methods that solve correctly systems of conservation laws can fail in solving (1.1), specially when approaching equilibria or near to equilibria solutions. In the context of shallow water equations Bermúdez and Vázquez-Cendón introduced in [6] the condition called *C-property*: a scheme is said to satisfy this condition if it solves correctly the steady-state solutions corresponding to water at rest. This idea of constructing numerical schemes that preserve some equilibria, which are called in general *well-balanced* schemes, has been studied by many authors. The design of numerical schemes with good properties for problems of the form (1.1) is a very active front of research, as PDE systems of this nature arise in many flow models: see [8] for a review on this topic.

Here, the strategy to derive well-balanced stable numerical methods for (1.1) will consist in writing first the system in the more general form

$$W_t + \mathcal{A}(W)W_x = 0 \quad x \in \mathbb{R},\ t > 0. \tag{1.2}$$

The main advantages of this approach is that it allows one to discretize the system as a whole, making thus easier to balance correctly the numerical flux and the source terms.

The major difficulty of solving general systems of the form (1.2) comes from the presence of the nonconservative product $\mathcal{A}(W)W_x$, which makes difficult even the definition of the weak solutions. After the theory developed by Dal Maso, LeFloch and Murat [12] a notion of weak solution can be introduced for which the equality (1.2) is satisfied in the sense of Borel measures. This notion is based on a family of paths in the phases space, whose selection is important, as it determines the speed of propagation of the discontinuities. The choice of an appropriate family of paths may be a difficult task. Although in physical applications it should depend on the problem under consideration and it could be suggested by an argument of regularization, it is natural from the mathematical point of view to impose this family to satisfy some conditions (see [23]). It turns out that, in the particular case of systems of balance laws, these conditions are enough to determine the family of paths, as least for close enough states, as it will be seen is Sect. 2.

The introduction of a family of paths not only gives a way to properly define the concept of weak solutions of nonconservative systems: a theoretical framework for the design of finite difference methods for systems of the form (1.2) has been introduced in [25]. This framework is based on the concept of *path-conservative* numerical scheme, which is a generalization of the usual concept of conservative method for conservation laws.

A basic requirement concerning the convergence of a numerical scheme is the following: if the approximations produced by the scheme converge to some function as the mesh is refined, then this function should be a weak solution of the system. In the particular case of systems of conservation laws, the classical theorem of Lax-Wendroff [21] gives the answer to this problem: conservative numerical methods have this property. Moreover, in [19] or [13] some negative results have been shown concerning the failure of the convergence of nonconservative schemes to weak solutions of conservative problems. These convergence difficulties have also been discussed recently in [2]. In [26] a Lax-Wendroff type theorem has been proved for a class of consistent numerical schemes for solving scalar conservation laws with a source term. In [10] it has been proved that, for general nonconservative systems, if the approximations produced by a path-conservative numerical scheme converges *uniformly in the sense of graphs*, then the limit is a weak solution according to the same family of paths. Unfortunately, this notion of convergence is too strong to expect the solutions produced by a finite-difference type method to converge in this sense and, in practice, the limit of the

approximations satisfies a perturbed problem in which an error source term supported on the discontinuities appears.

The first goal of this paper is to prove that, for systems of balance laws (1.1) with smooth enough σ, a Lax-Wendroff type result can be proved for path-conservative numerical schemes assuming the same hypotheses of convergence of the classical theorem for conservative problems. Together with the hypothesis concerning the convergence of the approximations, some minimal assumptions both on the chosen family of paths and on the numerical scheme are required, so that if σ is constant the usual notion of weak solution and a conservative numerical scheme are recovered.

In general, the assumptions on the family of paths required to prove the Lax-Wendroff theorem are not enough to ensure the well-balanced property. The second goal of this article is to present some general results showing that, in order to obtain well-balanced schemes, some further assumptions have to be imposed to the family of paths, concerning the relationship between the paths and the integral curves of the linearly degenerate field of the system associated to the null eigenvalue.

The organization is as follows: in Sect. 2 we present the reformulation of (1.1) as a system of the form (1.2) and discuss the definition of weak solutions and the choice of paths. In Sect. 3 the notion of path-conservative numerical schemes is recalled and applied to the particular case of a system of balance laws. Moreover, the schemes that reduce to a conservative method when σ is constant are characterized. Next, in Sect. 4 the Lax-Wendroff type result is proved. Some general results concerning the well-balanced property of these schemes are presented in Sect. 5. Finally, some examples of path-conservative schemes for systems of balance laws are recalled in the last section.

2 Weak Solutions

We consider system (1.1). In order to write the system in form (1.2) we add the trivially satisfied equation

$$\sigma_t = 0,$$

as it has been done previously by several authors: see for instance [16, 18, 22]. The augmented system is equivalent then to (1.2) with

$$W = \begin{bmatrix} U \\ \sigma \end{bmatrix}, \quad \mathcal{A}(W) = \begin{bmatrix} J(U) & -S(U) \\ 0 & 0 \end{bmatrix}, \tag{2.1}$$

being $J(U) \in \mathcal{M}_N(\mathbb{R})$ the Jacobian matrix of $F(U)$,

$$J(U) = \frac{\partial F}{\partial U}(U).$$

Notice that, once the problem has been written in this form, the actual value of σ is given by the initial conditions.

Let us suppose that matrix $J(U)$ has N real distinct eigenvalues

$$\lambda_1(U) < \cdots < \lambda_N(U)$$

with associated eigenvectors

$$R_1(U), \ldots, R_N(U).$$

If we assume that these eigenvalues do not vanish, then system (1.2), (2.1) is strictly hyperbolic, since $\mathcal{A}(W)$ has $N+1$ real distinct eigenvalues

$$\widetilde{\lambda}_1(W),\ldots,\widetilde{\lambda}_N(W),\widetilde{\lambda}_{N+1}(W)$$

given by

$$\widetilde{\lambda}_i(W)=\lambda_i(U),\quad i=1,\ldots,N;\qquad \widetilde{\lambda}_{N+1}(W)=0.$$

These eigenvalues have associated eigenvectors

$$\widetilde{R}_1(W),\ldots,\widetilde{R}_N(W),\widetilde{R}_{N+1}(W)$$

given by

$$\widetilde{R}_i(W)=\begin{bmatrix} R_i(U)\\ 0\end{bmatrix},\quad i=1,\ldots,N;\qquad \widetilde{R}_{N+1}(W)=\begin{bmatrix} J(U)^{-1}S(U)\\ 1\end{bmatrix}.$$

The $(N+1)$th characteristic field is linearly degenerate and we will suppose, for the sake of simplicity, that it is the only one.

The integral curves of the first N characteristic fields are given by the solution of the o.d.e. systems

$$\frac{dU}{ds}=R_i(U),\qquad \frac{d\sigma}{ds}=0,\quad i=1,\ldots,N, \tag{2.2}$$

and those of the $(N+1)$th field by the solutions of

$$\frac{dU}{d\sigma}=J(U)^{-1}S(U). \tag{2.3}$$

These last integral curves are strongly related to the stationary solutions of the system: in fact, stationary solutions are obtained by parametrizing in terms of x these curves. Indeed, taking x as a parameter in (2.3), we obtain:

$$J(U)\frac{dU}{dx}=S(U)\frac{d\sigma}{dx},$$

or equivalently

$$F(U)_x=S(U)\sigma_x,$$

which is the system satisfied by a stationary solution.

In the case in which some of the eigenvalues of $J(U)$ vanishes, system (1.2), (2.1) becomes nonstrictly hyperbolic. For these resonant problems, the definition and the analysis of weak solutions are much more difficult. In particular, Riemann problems may have more than one entropy solution (see [4, 14]). In this work, we will only consider the strictly hyperbolic case.

Notice that, when σ is discontinuous, the source term $S(U)\sigma_x$ does not make sense in the distributional framework. Even if here we are mainly interested in problems in which σ is continuous, it is interesting to consider the more general case for two reasons. On the one hand, in the numerical schemes considered here, piecewise constant approximations of σ will be considered. Therefore, numerical schemes based on approximate or exact Riemann solvers have to deal with the difficulty related to the lack of sense of the source term. On the

other hand, as we shall see, the well-balanced property of the numerical schemes is strongly related to their ability to handle with the contact discontinuities added to the system when σ is allowed to be discontinuous.

Let us thus consider the general case in which σ is discontinuous. In order to define the weak solutions of (1.2), (2.1) according to the theory developed in [12] a family of paths $\widetilde{\Phi}$ in $\Omega \times \mathbb{R}$ has to be chosen, i.e. a locally Lipschitz map

$$\widetilde{\Phi}\colon [0,1] \times (\Omega \times \mathbb{R}) \times (\Omega \times \mathbb{R}) \to (\Omega \times \mathbb{R})$$

satisfying some natural properties, as:

$$\widetilde{\Phi}(0; W_L, W_R) = W_L, \quad \widetilde{\Phi}(1; W_L, W_R) = W_R,$$

or

$$\widetilde{\Phi}(s; W, W) = W, \quad \forall s \in [0,1].$$

We will use the notation

$$W_L = \begin{bmatrix} U_L \\ \sigma_L \end{bmatrix}, \quad W_R = \begin{bmatrix} U_R \\ \sigma_R \end{bmatrix}$$

for the states W_L and W_R and

$$\widetilde{\Phi}(s; W_L, W_R) = \begin{bmatrix} \Phi(s; W_L, W_R) \\ \Phi_{N+1}(s; W_L, W_R) \end{bmatrix},$$

where

$$\Phi(s; W_L, W_R) = \begin{bmatrix} \Phi_1(s; W_L, W_R) \\ \vdots \\ \Phi_N(s; W_L, W_R) \end{bmatrix},$$

for the paths connecting both states.

The family of paths $\widetilde{\Phi}$ allows one to give a sense to the nonconservative product $\mathcal{A}(W)W_x$ as a Borel measure for $W \in (L^\infty(\mathbb{R} \times \mathbb{R}^+) \cap BV(\mathbb{R} \times \mathbb{R}^+))^{N+1}$, denoted by $[\mathcal{A}(W)W_x]_\Phi$. Given a time t, the Borel measure related to the nonconservative product is defined as follows:

$$\begin{aligned} \left\langle \left[\mathcal{A}(W(\cdot,t)) \frac{\partial W}{\partial x}(\cdot,t) \right]_{\widetilde{\Phi}}, \varphi \right\rangle &= \int_{\mathbb{R}} \mathcal{A}(W(x,t)) \frac{\partial W}{\partial x}(x,t)\varphi(x)\,dx \\ &+ \sum_m \left(\int_0^1 \mathcal{A}(\widetilde{\Phi}(s; W_m^-, W_m^+)) \frac{\partial \widetilde{\Phi}}{\partial s}(s; W_m^-, W_m^+)\,ds \right) \varphi(x_m(t)), \quad \forall \varphi \in \mathcal{C}_0(\mathbb{R}). \end{aligned} \tag{2.4}$$

In the above equality, the expression W_x appearing in the first integral represents the pointwise derivative of $W(\cdot,t)$; $x_m(t)$ are the locations of the discontinuities of W at time t; W_m^- and W_m^+ are respectively the limits of W to the left and to the right at the mth discontinuity at time t; and $\mathcal{C}_0(\mathbb{R})$ is the set of continuous maps with compact support.

A function $W \in (L^\infty(\mathbb{R}\times\mathbb{R}^+)\cap BV(\mathbb{R}\times\mathbb{R}^+))^{N+1}$ is said to be a weak solution of (1.2), (2.1) if it satisfies the equality

$$\frac{\partial W}{\partial t} + \left[\mathcal{A}(W)\frac{\partial W}{\partial x}\right]_{\widetilde{\Phi}} = 0. \tag{2.5}$$

Across a discontinuity, a weak solution must satisfy the generalized Rankine-Hugoniot condition:

$$\int_0^1 \left(\xi Id - \mathcal{A}(\widetilde{\Phi}(s; W^-, W^+))\frac{\partial \widetilde{\Phi}}{\partial s}(s; W^-, W^+)\, ds = 0, \right. \tag{2.6}$$

where ξ is the speed of propagation of the discontinuity, Id is the identity matrix, and W^- and W^+ are the left and right limits of the solution at the discontinuity.

Notice that, in this particular case, by using the structure of matrix $\mathcal{A}$, it can be easily seen that the $(N+1)$th component of the right-hand side of (2.4) vanishes and, for differentiable test functions, the first N components reduce to:

$$-\int_{\mathbb{R}} F(U(x,t))\varphi_x(x)\, dx - \int_{\mathbb{R}} S(U(x))\sigma_x(x)\varphi(x)\, dx - \sum_m S_{\widetilde{\Phi}}(W_m^-, W_m^+)\varphi(x_m(t)), \tag{2.7}$$

where

$$S_{\widetilde{\Phi}}\colon (\Omega\times\mathbb{R})\times(\Omega\times\mathbb{R}) \to \mathbb{R}^N$$

is defined by:

$$S_{\widetilde{\Phi}}(W_L, W_R) = \int_0^1 S(\Phi(s; W_L, W_R))\frac{\partial \Phi_{N+1}}{\partial s}(s; W_L, W_R))\, ds. \tag{2.8}$$

In the same way, the generalized Rankine-Hugoniot condition (2.6) reduces to:

$$\begin{cases} \xi(U^+ - U^-) = F(U^+) - F(U^-) - S_{\widetilde{\Phi}}(W^-, W^+), \\ \xi(\sigma^+ - \sigma^-) = 0. \end{cases} \tag{2.9}$$

According to the second equality in (2.9), the discontinuities appearing in weak solutions have to be either stationary ($\xi = 0$) or develop in regions where σ is continuous.

As it occurs in the conservative case, not any discontinuity is admissible. Therefore, a notion of entropy solution has to be assumed. Let us suppose for instance that there exists an entropy pair, i.e. a pair or smooth functions $(\eta(U), G(U,\sigma))$ such that η is convex and

$$\partial_U G(U,\sigma) = \nabla\eta(U)\cdot J(U), \quad \partial_\sigma G(U,\sigma) = -\nabla\eta(U)\cdot S(U),$$

for every U and σ. Then a weak solution is said to be an entropy solution if it satisfies the inequality

$$\frac{\partial \eta(U)}{\partial t} + \frac{\partial G(U,\sigma)}{\partial x} \le 0$$

in the distributions sense.

When an entropy pair is not available, the classical shock conditions of Lax can be considered to define the notion of entropy solution.

Even if any arbitrary choice of the family of paths allows us to give a consistent mathematical meaning to the weak solutions, it is natural to impose some requirements. Following [23], we impose the family of paths $\widetilde{\Phi}$ to satisfy the following assumptions:

H1 *If W_L and W_R are such that $\sigma_L = \sigma_R = \bar{\sigma}$, then*

$$\Phi_{N+1}(s; W_L, W_R) = \bar{\sigma} \quad \forall s \in [0,1]. \tag{2.10}$$

H2 *Given two states W_L and W_R belonging to the same integral curve γ of the linearly degenerate field, the path $\widetilde{\Phi}(\cdot; W_L, W_R)$ is a parametrization of the arc of γ linking W_L and W_R.*

H3 *Let us denote by $\widetilde{\mathcal{RP}} \subset (\Omega \times \mathbb{R}) \times (\Omega \times \mathbb{R})$ the set of pairs (W_L, W_R) such that the Riemann problem*

$$\begin{cases} W_t + \mathcal{A}(W)\, W_x = 0, \\ W(x,0) = \begin{cases} W_L & \text{if } x < 0, \\ W_R & \text{if } x > 0, \end{cases} \end{cases} \tag{2.11}$$

has a unique self-similar weak solution composed by at most $N+1$ simple waves (i.e. entropy shocks, contact discontinuities or rarefaction waves). Let $W^{\pm}_{L,R}$ denote the limits to the right and to the left of $x = 0$ of the solution of (2.11). Then, given $(W_L, W_R) \in \widetilde{\mathcal{RP}}$, the curve described by the path $\widetilde{\Phi}(\cdot; W_L, W_R)$ is equal to the union of the paths linking the pairs $(W_L, W^-_{L,R})$, $(W^-_{L,R}, W^+_{L,R})$, and $(W^+_{L,R}, W_R)$.

Proposition 2.1 *If we assume that the concept of weak solutions of (1.2), (2.1) is defined on the basis of a family of paths satisfying hypotheses (H1)–(H3), then*:

(i) *If*

$$W = \begin{bmatrix} U \\ \sigma \end{bmatrix}$$

is a weak solution of (1.2), (2.1) with σ constant, $\sigma(x) = \bar{\sigma}$, then U is a weak solution, in the usual sense, of the conservative problem

$$U_t + F(U)_x = 0. \tag{2.12}$$

(ii) *Given two states W_L and W_R belonging to the same integral curve of the linearly degenerate field, the stationary contact discontinuity given by*

$$W(x,t) = \begin{cases} W_L & \text{if } x < 0, \\ W_R & \text{if } x > 0, \end{cases} \tag{2.13}$$

is a weak solution of (1.2), (2.1).

(iii) *If (W_L, W_R) belongs to $\widetilde{\mathcal{RP}}$, then*:

$$\int_0^1 \mathcal{A}(\widetilde{\Phi}(s; W_L, W_R)) \frac{\partial \widetilde{\Phi}}{\partial s}(s; W_L, W_R)\, ds = \begin{bmatrix} F(U^-_{L,R}) - F(U_L) + F(U_R) - F(U^+_{L,R}) \\ 0 \end{bmatrix}, \tag{2.14}$$

where $U^{\pm}_{L,R}$ represent the first N components of $W^{\pm}_{L,R}$.

Proof To prove (i) observe that (H1) implies

$$S_{\widetilde{\Phi}}(W_L, W_R) = 0 \quad \text{if } \sigma_L = \sigma_R. \tag{2.15}$$

Therefore, when σ is constant, (2.7) reduces to:

$$-\int_{\mathbb{R}} F(U(x,t))\varphi_x(x)\,dx,$$

and the usual definition of weak solution is recovered.

Next, to prove (ii) notice that (H2) implies that, given two states W_L and W_R belonging to the same integral curve of the linearly degenerate field, one has:

$$\mathcal{A}(\widetilde{\Phi}(s; W_L, W_R))\frac{\partial \widetilde{\Phi}}{\partial s}(s; W_L, W_R)\,ds = 0, \tag{2.16}$$

and then (2.13) trivially satisfies (2.5).

Let us finally prove (iii). Notice first that (i), (ii) imply that the simple waves of the system are: (1) contact discontinuities standing on a discontinuity of σ and linking two states that belong to the same integral curve of the $(N+1)$th characteristic field; (2) rarefaction waves associated to the first N characteristic fields; and (3) entropy shocks also associated to the first N characteristic fields evolving in regions where σ is continuous. Due to the hypothesis of strict hyperbolicity of (1.2), (2.1) these shocks cannot be stationary and due to (2.15), they satisfy the usual Rankine-Hugoniot conditions

$$\xi(U^+ - U^-) = F(U^+) - F(U^-).$$

Therefore, given $(W_L, W_R) \in \widetilde{\mathcal{RP}}$, the only possible discontinuity at $x = 0$ for $t > 0$ would be a stationary contact discontinuity linking two states $W^-_{L,R} = [U^-_{L,R}, \sigma_L]$ and $W^+_{L,R} = [U^+_{L,R}, \sigma_R]$ belonging to the same integral curve of the linearly degenerate field. Taking into account (H3) we have:

$$\begin{aligned}
&\int_0^1 \mathcal{A}(\widetilde{\Phi}(s; W_L, W_R))\frac{\partial \widetilde{\Phi}}{\partial s}(s; W_L, W_R)\,ds \\
&\quad= \int_0^1 \mathcal{A}(\widetilde{\Phi}(s; W_L, W^-_{L,R}))\frac{\partial \widetilde{\Phi}}{\partial s}(s; W_L, W^-_{L,R})\,ds \\
&\qquad+ \int_0^1 \mathcal{A}(\widetilde{\Phi}(s; W^-_{L,R}, W^+_{L,R}))\frac{\partial \widetilde{\Phi}}{\partial s}(s; W^-_{L,R}, W^+_{L,R})\,ds \\
&\qquad+ \int_0^1 \mathcal{A}(\widetilde{\Phi}(s; W^+_{L,R}, W_R))\frac{\partial \widetilde{\Phi}}{\partial s}(s; W^+_{L,R}, W_R)\,ds.
\end{aligned} \tag{2.17}$$

The second summand in the right-hand side of the (2.17) vanishes due to (2.16) and the first and the third ones reduce respectively to:

$$\begin{bmatrix} F(U^-_{L,R}) - F(U_L) \\ 0 \end{bmatrix}, \quad \begin{bmatrix} F(U_R) - F(U^+_{L,R}) \\ 0 \end{bmatrix},$$

due to (2.15). Then, (2.14) is obtained from (2.17) and (iii) is also proved. □

It is worth noticing that, as a consequence of the previous proposition, the solutions of the Riemann problems (2.11) are the same for every family of paths satisfying (H1)–(H3), as the families of simple curves are independent of the family of paths. Moreover, the meaning of the nonconservative products across a discontinuity linking two states belonging to $\widetilde{\mathcal{RP}}$ is also the same for every family of paths satisfying (H1)–(H3), as the weight of the corresponding Dirac mass (2.14) is independent of the paths.

We stress the fact that assumptions (H1)–(H3) are not necessary to have a consistent definition of the weak solutions. For instance, for the particular case of the shallow water system with a bottom step, while the definition of weak solution given in [3] is equivalent to the one associated to a family of paths satisfying these assumptions, this is not the case for the definitions in [7] or [29]. In particular, the statement (ii) in the proposition above is not valid for the definition of weak solutions in the two last references.

We will assume in the sequel that a family of paths satisfying (H1)–(H3) has been chosen to define the weak solutions of (1.2), (2.1). In practice, such a family can be constructed at least in the class $\widetilde{\mathcal{RP}}$ as follows: given (W_L, W_R) in this class, first the corresponding Riemann problem (2.11) is solved (observe that it is possible to solve this problem before defining the family of paths, as the simple curves are already known). Then, the path linking $(W_L, W_R) \in \widetilde{\mathcal{RP}}$ is a parametrization in $[0, 1]$ of the curve composed by:

- the straight segment linking W_L and $W_{L,R}^-$;
- the arc of the integral curve of the linearly degenerate field linking $W_{L,R}^-$ and $W_{L,R}^+$;
- the straight segment linking $W_{L,R}^+$ and W_R.

To finish this section, let us remark that, when σ is continuous, every family of paths satisfying (H1) defines the same concept of weak solution: observe that (2.7) in this case reduces to

$$-\int_{\mathbb{R}} F(U(x,t))\varphi_x(x)\,dx - \int_{\mathbb{R}} S(U(x))\sigma_x(x)\varphi(x)\,dx, \tag{2.18}$$

independently of the chosen family, so that the weak solution of (1.1) with initial condition

$$U(x,0) = U^0(x) \tag{2.19}$$

can be defined by means of the following variational formulation:

$$\int_0^\infty \int_{-\infty}^\infty U(x,t)\,\varphi_t(x,t)\,dx\,dt + \int_0^\infty \int_{-\infty}^\infty F(U(x,t))\,\varphi_x(x,t)\,dx\,dt$$
$$+ \int_0^\infty \int_{-\infty}^\infty S(U(x,t))\sigma_x(x)\varphi(x,t)\,dx\,dt + \int_{-\infty}^{+\infty} U^0(x)\varphi(x,0)\,dx = 0, \tag{2.20}$$

where φ is any $\mathcal{C}_0^1(\mathbb{R}\times[0,\infty))$ test function.

3 Numerical Schemes

We consider problem (1.1) with initial condition (2.19) or, equivalently, problem (1.2), (2.1) with initial condition

$$W(x,0) = W^0(x), \tag{3.1}$$

where

$$W^0 = \begin{bmatrix} U^0 \\ \sigma \end{bmatrix}. \tag{3.2}$$

We will use the notation

$$W_i^n = \begin{bmatrix} U_i^n \\ \sigma_i \end{bmatrix},$$

where

$$\sigma_i = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \sigma(x)\, dx,$$

for the approximation of the cell averages of a solution of (1.2), (2.1) obtained by a numerical scheme.

We consider here numerical schemes which are path-conservative in the sense introduced in [25], i.e. numerical schemes that can be written as follows:

$$W_i^{n+1} = W_i^n - \frac{\Delta t}{\Delta x}\big(\widetilde{D}_{i-1/2}^{n,+} + \widetilde{D}_{i+1/2}^{n,-}\big), \tag{3.3}$$

where

$$\widetilde{D}_{i+1/2}^{n,\pm} = \widetilde{D}^{\pm}\big(W_{i-q}^n, \ldots, W_{i+p}^n\big),$$

with $\widetilde{D}^-$ and $\widetilde{D}^+$ continuous functions satisfying

$$\widetilde{D}^{\pm}(W, \ldots, W) = 0 \quad \forall W \in \Omega \times \mathbb{R} \tag{3.4}$$

and

$$\begin{aligned} &\widetilde{D}^-(W_{-q}, \ldots, W_p) + \widetilde{D}^+(W_{-q}, \ldots, W_p) \\ &\quad = \int_0^1 \mathcal{A}\big(\widetilde{\Phi}(s; W_0, W_1)\big) \frac{\partial \widetilde{\Phi}}{\partial s}(s; W_0, W_1)\, ds, \end{aligned} \tag{3.5}$$

for every $W_i \in \Omega \times \mathbb{R}$, $i = -q, \ldots, p$.

Due to the structure of W and $\mathcal{A}(W)$ in system (1.2), (2.1), it is natural to choose $\widetilde{D}^{\pm}$ such that:

$$\widetilde{D}^{\pm} = \begin{bmatrix} D^{\pm} \\ 0 \end{bmatrix}$$

and thus:

$$\widetilde{D}_{i+1/2}^{n,\pm} = \begin{bmatrix} D_{i+1/2}^{n,\pm} \\ 0 \end{bmatrix}.$$

Dropping the last component in (3.3), the scheme is reduced to

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}\big(D_{i-1/2}^{n,+} + D_{i+1/2}^{n,-}\big), \tag{3.6}$$

where $D^{n,\pm}_{i+1/2} = D^{\pm}(W^n_{i-q}, \ldots, W^n_{i+p})$, being D^- and D^+ two continuous functions. Conditions (3.4) and (3.5) reduce then to

$$D^{\pm}(W, \ldots, W) = 0 \quad \forall W \in \Omega \times \mathbb{R} \tag{3.7}$$

and

$$\begin{aligned} &D^-(W_{-q}, \ldots, W_p) + D^+(W_{-q}, \ldots, W_p) \\ &\quad = F(U_1) - F(U_0) - S_{\widetilde{\Phi}}(W_0, W_1), \end{aligned} \tag{3.8}$$

for $W_i \in \Omega \times \mathbb{R},\ i = -q, \ldots, p$.

According to the previous section, the more appropriate choice for the family of paths would be given by a family satisfying (H1)–(H3). The following proposition, whose proof is trivial by (3.5) and (2.14), characterizes the numerical schemes that are path-conservative for such a family of paths:

Proposition 3.1 *Let $\widetilde{\Phi}$ be a family of paths satisfying* (H1)–(H3). *A numerical scheme of the form* (3.6)–(3.8) *is $\widetilde{\Phi}$-conservative if and only if, given $\{W_{-q}, \ldots, W_p\}$ such that $(W_0, W_1) \in \widetilde{\mathcal{RP}}$ one has*:

$$\begin{aligned} &D^-\big(W_{-q}, \ldots, W_p\big) + D^+\big(W_{-q}, \ldots, W_p\big) \\ &\quad = F(U_1) - F(U^+_{1/2}) + F(U^-_{1/2}) - F(U_0), \end{aligned} \tag{3.9}$$

where

$$W^-_{1/2} = \begin{bmatrix} U^-_{1/2} \\ \sigma_0 \end{bmatrix} \quad \textit{and} \quad W^+_{1/2} = \begin{bmatrix} U^+_{1/2} \\ \sigma_1 \end{bmatrix}$$

are, respectively, the left and right limits at $x = 0$ of the solution of the Riemann problem associated to W_0 and W_1.

Unfortunately, the design of a numerical scheme which is path-conservative for a family satisfying (H1)–(H3) can be difficult or very costly in practice since the construction of the path $\widetilde{\Phi}$ linking two states involves the resolution of a Riemann problem. Consequently, we will consider numerical schemes that are path-conservative for an arbitrary family of paths $\widetilde{\Psi}$ and we will analyze their properties.

In the sequel we will focus on numerical schemes with $q = 0$ and $p = 1$, i.e.

$$\widetilde{D}^{n,\pm}_{i+1/2} = \widetilde{D}^{\pm}(W^n_i, W^n_{i+1}).$$

In order to be able to prove a Lax-Wendroff type theorem, the numerical schemes should reduce to a conservative scheme when σ is constant. Let us characterize these schemes:

Proposition 3.2 *A $\widetilde{\Psi}$-conservative scheme reduces, when σ is constant, to a conservative scheme whose numerical flux is independent of σ if and only if*

$$\frac{\partial D^{\pm}}{\partial \sigma}\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) = 0 \tag{3.10}$$

and

$$S_{\widetilde{\Psi}}\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) = 0 \tag{3.11}$$

hold for every U_0, U_1, *and* σ.

Proof Let us suppose that there exists a numerical flux $G(U_0, U_1)$ consistent with F such that:

$$D^+\left(\begin{bmatrix} U_{-1} \\ \sigma \end{bmatrix}, \begin{bmatrix} U_0 \\ \sigma \end{bmatrix}\right) + D^-\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) = G(U_0, U_1) - G(U_{-1}, U_0) \tag{3.12}$$

for every U_{-1}, U_0, U_1, and σ.

Choosing $U_{-1} = U_0$ we deduce:

$$D^-\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) = G(U_0, U_1) - F(U_0),$$

and choosing $U_0 = U_1$:

$$D^+\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) = F(U_1) - G(U_0, U_1).$$

So, (3.10) holds. Adding the last two equalities we obtain

$$D^-\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) + D^+\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) = F(U_1) - F(U_0), \tag{3.13}$$

which, compared to (3.8), gives (3.11).

Conversely, if (3.10) and (3.11) are satisfied, we can define a consistent numerical flux either by

$$G(U_0, U_1) = D^-\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) + F(U_0), \tag{3.14}$$

or by

$$G(U_0, U_1) = -D^+\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) + F(U_1). \tag{3.15}$$

Both definitions coincide and are independent of σ due to (3.13), (3.10), and (3.11). Moreover, (3.12) is trivially satisfied. □

Remark 3.3 The reciprocal implication in the previous proposition is also valid for numerical schemes with arbitrary values of p and q.

Clearly, (3.11) is satisfied if the family of paths satisfies (H1). We will suppose in the sequel that this hypothesis is fulfilled.

Proposition 3.4 *A $\widetilde{\Psi}$-conservative numerical scheme satisfying (3.10) can be written in the form*

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}\big(G_{i+1/2}^n - G_{i-1/2}^n\big) + \frac{\Delta t}{\Delta x}\big(H_{i-1/2}^{n,+} + H_{i+1/2}^{n,-}\big), \tag{3.16}$$

where

$$G_{i+1/2}^n = G\big(U_i^n, U_{i+1}^n\big),$$

being G a continuous function such that

$$G(U,U) = F(U), \tag{3.17}$$

and

$$H_{i+1/2}^{n,\pm} = H^{\pm}\big(W_i^n, W_{i+1}^n\big),$$

being H^- and H^+ two continuous functions such that

$$H^{\pm}\big(W, W\big) = 0, \tag{3.18}$$

$$H^-\big(W_0, W_1\big) + H^+\big(W_0, W_1\big) = S_{\widetilde{\Psi}}\big(W_0, W_1\big) \tag{3.19}$$

and

$$H^{\pm}\left(\begin{bmatrix} U_0 \\ \sigma \end{bmatrix}, \begin{bmatrix} U_1 \\ \sigma \end{bmatrix}\right) = 0. \tag{3.20}$$

Conversely, a numerical scheme (3.16)–(3.20) can be written as a $\widetilde{\Psi}$-conservative numerical scheme satisfying (3.10).

Proof Let us define $G(U_0, U_1)$ by (3.14) or (3.15). Next, we define

$$H^-(W_0, W_1) = -D^-(W_0, W_1) + G(U_0, U_1) - F(U_0) \tag{3.21}$$

and

$$H^+(W_0, W_1) = -D^+(W_0, W_1) - G(U_0, U_1) + F(U_1). \tag{3.22}$$

We can easily check that conditions (3.17)–(3.19) are satisfied. Condition (3.20) is easily deduced from the definition of G and from (3.10).

The proof of the reciprocal implication is straightforward.

□

Remark 3.5 The form (3.16) is that of an Upwind Interface Source method as presented in [26]. In this reference, the functions $H^{\pm}$ are supposed to be of the form:

$$H^{\pm}(W_L, W_R) = \bar{H}^{\pm}(U_L, U_R, \sigma_R - \sigma_L),$$

where $\bar{H}^{\pm}$ are functions satisfying

$$\bar{H}^{\pm}(U_L, U_R, 0) = 0,$$

for every U_L, U_R. Notice that, in this case, (3.18) and (3.20) are automatically satisfied. Let us finally remark that (3.19) implies the consistency of the scheme in the sense defined in [26].

4 A Convergence Result

In this section we prove a Lax-Wendroff convergence theorem for the particular case of systems of balance laws, provided that the path-numerical scheme used to solve it reduces to a conservative scheme when σ is constant.

Assume that the mesh ratio $\Delta t/\Delta x$ is a fixed constant λ as Δx, Δt tend to 0.

Set $h = \Delta x$ and introduce the piecewise constant function defined a.e. in $\mathbb{R} \times [0, \infty)$ by

$$U_h(x,t) = U_i^n, \quad (x,t) \in (x_{i-1/2}, x_{i+1/2}) \times [t^n, t^{n+1}).$$

For the discretization of the initial condition we choose the cell averages

$$U_i^0 = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} U_0(x)\,dx. \tag{4.1}$$

We will also use the notation

$$W_h = \begin{bmatrix} U_h \\ \sigma \end{bmatrix}.$$

Theorem 4.1 *Let $F \in \mathcal{C}^2(\Omega)^N$, $S \in \mathcal{C}^1(\Omega)^N$ and $\sigma \in W^{1,1}(\mathbb{R})$. Let $\widetilde{\Psi}$ be any family of paths satisfying* (H1). *For $h > 0$, let U_h be the numerical approximation obtained from a $\widetilde{\Psi}$-conservative numerical scheme of the form* (3.16)–(3.20) *and $U_h(x,0)$ given by* (4.1).

Suppose that

$$\|U_h\|_{L^\infty(\mathbb{R}\times[0,\infty))^N} \le C \tag{4.2}$$

and that there exists a function $U \in (L^\infty(\mathbb{R} \times [0,\infty)) \cap BV(\mathbb{R} \times [0,\infty)))^N$ such that

$$U_h \xrightarrow[h\to 0]{} U \quad \text{in } L^1_{loc}(\mathbb{R} \times [0,\infty))^N. \tag{4.3}$$

Then U is a weak solution of the problem (1.1) *with initial condition* (2.19).

Proof The proof is an adaptation of the classical Lax-Wendroff convergence theorem for hyperbolic systems of conservation laws [21].

We want to prove that (2.20) is satisfied for every $\varphi \in \mathcal{C}_0^1(\mathbb{R} \times [0,\infty))$.

Let $\varphi \in \mathcal{C}_0^1(\mathbb{R} \times [0,\infty))$ be a test function and set

$$\varphi_i^n = \varphi(x_i, t^n).$$

By multiplying (3.16) by $\Delta x\, \varphi_i^n$ and summing over i and n we have:

$$\Delta x \sum_{n=0}^{\infty} \sum_{i=-\infty}^{\infty} \big(U_i^{n+1} - U_i^n\big)\varphi_i^n + \Delta t \sum_{n=0}^{\infty} \sum_{i=-\infty}^{\infty} \big(G_{i+1/2}^n - G_{i-1/2}^n\big)\varphi_i^n$$

$$- \Delta t \sum_{n=0}^{\infty} \sum_{i=-\infty}^{\infty} \big(H_{i-1/2}^{n,+} + H_{i+1/2}^{n,-}\big)\varphi_i^n = 0.$$

Notice that all these sums are finite since φ has a compact support.

Applying summation by parts we obtain:

$$\Delta x \sum_{n=1}^{\infty}\sum_{i=-\infty}^{\infty} U_i^n\big(\varphi_i^n-\varphi_i^{n-1}\big)+\Delta t\sum_{n=0}^{\infty}\sum_{i=-\infty}^{\infty} G_{i+1/2}^n\big(\varphi_{i+1}^n-\varphi_i^n\big)$$
$$+\Delta t\sum_{n=0}^{\infty}\sum_{i=-\infty}^{\infty}\big(H_{i+1/2}^{n,+}\varphi_{i+1}^n+H_{i+1/2}^{n,-}\varphi_i^n\big)+\Delta x\sum_{i=-\infty}^{\infty}U_i^0\varphi_i^0=0.$$

Now set

$$\varphi_{i+1/2}^n=\varphi(x_{i+1/2},t^n).$$

It is straightforward to verify that

$$\varphi_j^n=\varphi_{i+1/2}^n+O(\Delta x),\quad j=i,i+1,$$

and then

$$\Delta x \sum_{n=1}^{\infty}\sum_{i=-\infty}^{\infty} U_i^n\big(\varphi_i^n-\varphi_i^{n-1}\big)+\Delta t\sum_{n=0}^{\infty}\sum_{i=-\infty}^{\infty} G_{i+1/2}^n\big(\varphi_{i+1}^n-\varphi_i^n\big)$$
$$+\Delta t\sum_{n=0}^{\infty}\sum_{i=-\infty}^{\infty}\big(H_{i+1/2}^{n,-}+H_{i+1/2}^{n,+}\big)\big(\varphi_{i+1/2}^n+O(\Delta x)\big)+\Delta x\sum_{i=-\infty}^{\infty}U_i^0\varphi_i^0=0. \quad (4.4)$$

If we consider the functions defined a.e. by

$$\varphi_h(x,t)=\varphi_i^n,\quad (x,t)\in(x_{i-1/2},x_{i+1/2})\times[t^n,t^{n+1}),$$
$$\psi_h(x,t)=\varphi_{i+1/2}^n,\quad (x,t)\in(x_i,x_{i+1})\times[t^n,t^{n+1}),$$
$$G_h(x,t)=G_{i+1/2}^n,\quad (x,t)\in(x_i,x_{i+1})\times[t^n,t^{n+1})$$

and

$$H_h(x,t)=\frac{1}{\Delta x}\big(H_{i+1/2}^{n,-}+H_{i+1/2}^{n,+}\big)=\frac{1}{\Delta x}S_{\widetilde{\Psi}}\big(\widetilde{W}_i^n,\widetilde{W}_{i+1}^n\big),$$

for $(x,t)\in(x_i,x_{i+1})\times[t^n,t^{n+1})$, then (4.4) can be rewritten as

$$\int_{\Delta t}^{\infty}\int_{-\infty}^{\infty}U_h(x,t)\frac{\varphi_h(x,t)-\varphi_h(x,t-\Delta t)}{\Delta t}\,dx\,dt$$
$$+\int_0^{\infty}\int_{-\infty}^{\infty}G_h(x,t)\frac{\varphi_h(x+\Delta x/2,t)-\varphi_h(x-\Delta x/2,t)}{\Delta x}\,dx\,dt$$
$$+\int_0^{\infty}\int_{-\infty}^{\infty}H_h(x,t)\big(\psi_h(x,t)+O(h)\big)\,dx\,dt$$
$$+\int_0^{\infty}\int_{-\infty}^{\infty}U^0(x)\varphi_h(x,0)dx=0. \quad (4.5)$$

We want to prove that (2.20) can be obtained by passing to the limit as h tends to 0 in (4.5).

It is not difficult to check that the first, the second and the last term in (4.5) converge to the first, the second and the last term in (2.20), respectively. It remains to prove that the third term in (4.5) converges to the third one in (2.20).

It suffices to study the convergence of

$$\int_0^\infty \int_{-\infty}^\infty H_h(x,t)\psi_h(x,t)\,dx\,dt.$$

In order to do that, we use the decomposition

$$\begin{aligned}
&\int_0^\infty \int_{-\infty}^\infty (H_h(x,t)\,\psi_h(x,t) - S(U(x,t))\sigma_x(x)\varphi(x,t))\,dx\,dt \\
&\quad = \int_0^\infty \int_{-\infty}^\infty H_h(x,t)\big(\psi_h(x,t) - \varphi(x,t)\big)\,dx\,dt \\
&\qquad + \int_0^\infty \int_{-\infty}^\infty (H_h(x,t) - S(U(x,t))\sigma_x(x))\varphi(x,t)\,dx\,dt. \qquad (4.6)
\end{aligned}$$

The first summand in (4.6) trivially converges to zero because H_h is bounded and ψ_h converges uniformly to φ as h tends to 0.

To prove the convergence to zero of the second summand we want to prove that H_h converges to $S(U)\sigma_x$ in $L^1_{\mathrm{loc}}(\mathbb{R}\times[0,\infty))^N$.

For $(x,t)\in(x_i,x_{i+1})\times[t^n,t^{n+1})$ we have:

$$H_h(x,t) = \frac{1}{\Delta x} S_{\widetilde{\Psi}}\big(W_h(x-\Delta x/2,t),\, W_h(x+\Delta x/2,t)\big)$$

and we can use the decomposition

$$\begin{aligned}
&H_h(x,t) - S(U(x,t))\sigma_x(x) \\
&\quad = \int_0^1 S\big(\Psi(s;\, W_h(x-\Delta x/2,t),\, W_h(x+\Delta x/2,t))\big)\left(\frac{1}{\Delta x}\frac{\partial \Psi_{N+1}}{\partial s}(s) - \sigma_x(x)\right) ds \\
&\qquad + \int_0^1 \big(S\big(\Psi(s;\, W_h(x-\Delta x/2,t),\, W_h(x+\Delta x/2,t))\big) - S(U(x,t))\big)\sigma_x(x)\,ds.
\end{aligned} \qquad (4.7)$$

The second summand converges to zero because of the convergence of $W_h(x-\Delta x/2,t)$ and $W_h(x+\Delta x/2,t)$ to $W(x,t)$ in $L^1_{\mathrm{loc}}(\mathbb{R}\times[0,\infty))$, the regularity of S and $\widetilde{\Psi}$, and the Lebesgue dominated convergence theorem.

To prove the convergence to zero of the first summand in (4.7) we just need to prove the convergence to zero of

$$\begin{aligned}
&\int_0^1 \left(\frac{1}{\Delta x}\frac{\partial \Psi_{N+1}}{\partial s}\big(s;\, W_h(x-\Delta x/2,t),\, W_h(x+\Delta x/2,t)\big) - \sigma_x(x)\right) ds \\
&\quad = \frac{\sigma(x+\Delta x/2) - \sigma(x-\Delta x/2)}{\Delta x} - \sigma_x(x),
\end{aligned}$$

what is given by the hypothesis $\sigma\in W^{1,1}(\mathbb{R})$.

The Lebesgue dominated convergence theorem assures the convergence of H_h to $S(U)\sigma_x$ in $L^1_{\mathrm{loc}}(\mathbb{R}\times[0,\infty))$ and allows us to finish the proof of the theorem. □

5 Well-Balancing

We have seen in the previous section that, in order to prove the convergence to the weak solutions of the problem when σ is smooth, only the property (H1) has to be imposed to the family of paths. In this section we will see that further assumptions have to be imposed if we want to obtain well-balanced path-conservative numerical schemes.

A numerical scheme (3.3) is said to be well-balanced for a stationary solution $W(x)$ if it solves exactly that solution, i.e. if, when the numerical scheme is applied to the initial conditions

$$W_i^0 = W(x_i) \quad \forall i,$$

then

$$W_i^n = W_i^0 \quad \forall i, n.$$

It has been seen in Sect. 2 that every stationary solution of the problem may be understood as a parametrization of an integral curve γ of the linearly degenerate field, i.e. an integral curve of the o.d.e. system (2.3). This fact motivates the following definition: a numerical scheme (3.3) with $q = 0$ and $p = 1$ is said to be well-balanced for an integral curve γ of the linearly degenerate field if

$$\widetilde{D}^{\pm}(W_0, W_1) = 0 \quad \forall W_0,\ W_1 \in \gamma. \tag{5.1}$$

Clearly, a numerical scheme which is well-balanced for γ is well-balanced for any stationary solution constructed by parameterizing an arc of γ with x. Moreover, remember from Sect. 2 that the contact discontinuities of the problem also link two states belonging to an integral curve of the linearly degenerate field. Therefore, a numerical scheme which is exactly well-balanced for a curve γ solves exactly any contact discontinuity linking two states belonging to γ. And for a numerical scheme (3.3) with $q = 0$ and $p = 1$, the reciprocal implication is also true (see [25]), so that such a numerical scheme is exactly well-balanced for an integral curve γ if and only if it solves exactly every contact discontinuity linking two states belonging to γ.

For schemes of the form (3.6)–(3.8), equality (5.1) reduces to

$$D^{\pm}(W_0, W_1) = 0 \quad \forall W_0,\ W_1 \in \gamma. \tag{5.2}$$

In particular, a numerical scheme of the form (3.16) is well-balanced for an integral curve γ of the linearly degenerate field if and only if

$$H^{-}(W_0, W_1) = G(U_0, U_1) - F(U_0) \quad \forall W_0,\ W_1 \in \gamma \tag{5.3}$$

and

$$H^{+}(W_0, W_1) = F(U_1) - G(U_0, U_1) \quad \forall W_0,\ W_1 \in \gamma. \tag{5.4}$$

Let us consider now a numerical scheme (3.3) which is path-conservative for a family of paths $\widetilde{\Psi}$. Then, the following result can be easily proved by using (3.5) and (5.1):

Proposition 5.1 *A necessary condition for a $\widetilde{\Psi}$-conservative scheme to be well-balanced for $\gamma \in \Gamma$ is that*

$$\int_0^1 \mathcal{A}\big(\widetilde{\Psi}(s; W_0, W_1)\big) \frac{\partial \widetilde{\Psi}}{\partial s}(s; W_0, W_1)\, ds = 0, \quad \forall W_0,\ W_1 \in \gamma. \tag{5.5}$$

In particular, if the family of paths satisfies the following property:

(P_γ) For every W_L and W_R belonging to γ, $\widetilde{\Psi}(\cdot; W_L, W_R)$ is a parametrization of the arc of γ connecting W_L and W_R,

then condition (5.5) is fulfilled. Observe that, if the family of paths satisfies (H2), then (5.5) holds for every integral curve γ.

Condition (5.5) is not sufficient to obtain a well-balanced scheme. Indeed, consider the numerical scheme defined by:

$$D^-(W_0, W_1) = F(U^-_{1/2}) - F(U_0) - K(W^-_{1/2}, W^+_{1/2}),$$

$$D^+(W_0, W_1) = F(U_1) - F(U^+_{1/2}) + K(W^-_{1/2}, W^+_{1/2}),$$

where $W^-_{1/2}$ and $W^+_{1/2}$ are, respectively, the left and right limits of the solution of the Riemann problem that has W_0 and W_1 as initial condition, and K is any continuous function from $(\Omega \times \mathbb{R}) \times (\Omega \times \mathbb{R})$ into Ω such that

$$K(W, W) = 0 \quad \forall W \in \Omega \times \mathbb{R}$$

and

$$K(W_L, W_R) \neq 0$$

for a pair of states W_L and W_R belonging to the same integral curve of the linearly degenerate field. This numerical scheme is path-conservative for a family of paths $\widetilde{\Phi}$ satisfying (H1)–(H3) (see Proposition 3.1) and thus (5.5) is satisfied for every integral curve of the linearly degenerate field. Nevertheless, it is not well-balanced. To see it, apply the numerical scheme to the initial condition:

$$W_i^0 = \begin{cases} W_L & \text{if } i \leq 0, \\ W_R & \text{if } i > 0. \end{cases}$$

Nevertheless, as it will be seen in next section, in most cases property (P_γ) is enough to ensure the well-balanced property of a numerical scheme.

6 Examples

We present in this section some families of path-conservative schemes that reduce to conservative methods when σ is constant and we briefly discuss their well-balanced properties.

6.1 Godunov Method

The extension of the classical Godunov method [15] to systems of balance laws is given by

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}\big(F(U^{n,-}_{i+1/2}) - F(U^{n,+}_{i-1/2})\big),$$

where:

$$W^{n,-}_{i+1/2} = \begin{bmatrix} U^{n,-}_{i+1/2} \\ \sigma_i \end{bmatrix}, \qquad W^{n,+}_{i+1/2} = \begin{bmatrix} U^{n,+}_{i+1/2} \\ \sigma_{i+1} \end{bmatrix},$$

are the limits to the left and to the right of $x = 0$ of the solution of the Riemann problem with initial conditions $W_i^n = [U_i^n, \sigma_i]^T$ and $W_{i+1}^n = [U_{i+1}^n, \sigma_{i+1}]^T$. This scheme is path-conservative for any family of paths satisfying (H1)–(H3) and is exactly well-balanced for any integral curve γ of the linearly degenerate field: see [23] for details.

6.2 Roe Methods

A Roe scheme which is path-conservative for a family of paths $\widetilde{\Psi}$ can be written in the form (3.16) with the choices:

$$G_{i+1/2}^n = \frac{1}{2}\big(F(U_i^n) + F(U_{i+1}^n)\big) - \frac{1}{2}\left|J_{i+1/2}^n\right|(U_{i+1}^n - U_i^n), \tag{6.1}$$

$$H_{i+1/2}^{n,\pm} = \mathcal{P}_{i+1/2}^{n,\pm} \cdot S_{i+1/2}^n(\sigma_{i+1} - \sigma_i), \tag{6.2}$$

where $J_{i+1/2}^n$ is a Roe matrix for the homogeneous problem, i.e. a matrix satisfying

$$J_{i+1/2}^n(U_{i+1}^n - U_i^n) = F(U_{i+1}^n) - F(U_i^n),$$

$S_{i+1/2}^n$ is a vector satisfying

$$S_{i+1/2}^n(\sigma_{i+1} - \sigma_i) = S_{\widetilde{\Phi}}(W_i^n, W_{i+1}^n),$$

and

$$\mathcal{P}_{i+1/2}^{n,\pm} = \frac{1}{2}\left(Id \pm \left|J_{i+1/2}^n\right|(J_{i+1/2}^n)^{-1}\right).$$

Here, Id represents the identity matrix and

$$|J_{i+1/2}^n| = K_{i+1/2}^n \left|L_{i+1/2}^n\right| (K_{i+1/2}^n)^{-1},$$

where $|L_{i+1/2}^n|$ is the diagonal matrix whose coefficients are the absolute value of the eigenvalues of $J_{i+1/2}^n$ and $K_{i+1/2}^n$ a $N \times N$ matrix whose columns are associated eigenvectors.

This scheme is obtained by applying to system (1.2), (2.1) a Roe scheme [27] based on a generalized Roe linearization in the sense introduced in [30] (see [16, 24]). Several extensions of Roe method for systems of balance laws have been used in the literature: see for instance [6, 17, 28]. See also [1].

A Roe scheme is well-balanced for a curve $\gamma \in \Gamma$ if the family of paths satisfies the property (P_γ). In particular, if the family of paths satisfies (H2), the scheme is well-balanced for every γ.

If the family of straight segments is chosen, the numerical scheme is still exactly well-balanced for the integral curves γ which are straight lines in the phases space: this is the case for water at rest solutions for both the shallow water and the two-layer shallow water systems. Moreover, all the smooth stationary solutions are solved up to second order (see [24] for details).

6.3 Lax-Friedrichs Scheme

A path-conservative extension of the classical Lax-Friedrichs scheme [20] based on a Roe linearization is given by (3.16) with the choices:

$$G_{i+1/2}^n = \frac{1}{2}\left(F(U_i^n) + F(U_{i+1}^n) - \frac{\Delta x}{\Delta t}(U_{i+1}^n - U_i^n)\right),$$

$$H^{n,\pm}_{i+1/2} = \frac{1}{2}\left(Id \pm \frac{\Delta x}{\Delta t}(J^n_{i+1/2})^{-1}\right) S^n_{i+1/2}(\sigma_{i+1} - \sigma_i),$$

where $J^n_{i+1/2}$ and $S^n_{i+1/2}$ are defined as above. This scheme is also well-balanced for an integral curve γ of the linearly degenerate field if (P_γ) is satisfied. Nevertheless, when the family of straight segments is chosen, the property of solving with second order accuracy all the smooth stationary solutions is lost.

Notice that the numerical scheme is not well defined when the Roe matrix $J^n_{i+1/2}$ is singular. This difficulty has been discussed in [11]. In this reference, a family of well-balanced numerical schemes including this Lax-Friedrichs scheme as a particular case, as well as path-conservative extensions of the Lax-Wendroff, FORCE, and GFORCE schemes have also been introduced.

6.4 Generalized Hydrostatic Reconstruction

In [9] the Hydrostatic Reconstruction technique, introduced for the shallow water system in [5], has been generalized to obtain a numerical scheme which is exactly well-balanced for a set Γ_0 of integral curves of the linearly degenerate field. The expression of the numerical scheme is again (3.16) with:

$$G^n_{i+1/2} = G(U^{n,-}_{i+1/2}, U^{n,+}_{i+1/2}),$$

$$H^{n,\pm}_{i+1/2} = \int_0^1 S\big(P^{n,\pm}_{i+1/2}(s)\big)\frac{\partial p^{n,\pm}_{i+1/2,N+1}(s)}{\partial s}\,ds,$$

where G is any consistent numerical flux for the homogeneous problem, $U^{n,\pm}_{i+1/2}$ are two chosen intermediate states, and

$$s \in [0,1] \mapsto \widetilde{P}^{n,-}_{i+1/2}(s) = \begin{bmatrix} P^{n,-}_{i+1/2}(s) \\ p^{n,-}_{i+1/2,N+1}(s) \end{bmatrix} = \begin{bmatrix} p^{n,-}_{1,i+1/2}(s) \\ \vdots \\ p^{n,-}_{i+1/2,N}(s) \\ p^{n,-}_{i+1/2,N+1}(s) \end{bmatrix}, \tag{6.3}$$

$$s \in [0,1] \mapsto \widetilde{P}^{n,+}_{i+1/2}(s) = \begin{bmatrix} P^{n,+}_{i+1/2}(s) \\ p^{n,+}_{i+1/2,N+1}(s) \end{bmatrix} = \begin{bmatrix} p^{n,+}_{1,i+1/2}(s) \\ \vdots \\ p^{n,+}_{i+1/2,N}(s) \\ p^{n,+}_{i+1/2,N+1}(s) \end{bmatrix}, \tag{6.4}$$

are two paths linking respectively the pairs of states $(W^n_i, W^{n,-}_{i+1/2})$ and $(W^{n,+}_{i+1/2}, W^n_{i+1})$, where

$$W^{n,\pm}_{i+1/2} = \begin{bmatrix} U^{n,\pm}_{i+1/2} \\ \sigma_{i+1/2} \end{bmatrix},$$

being $\sigma_{i+1/2}$ a chosen intermediary value of σ. In order to be well-balanced for the curves of the set Γ_0 these paths have to be such that

- If W_i^n belongs to a curve $\gamma \in \Gamma_0$ then $s \in [0,1] \mapsto \widetilde{P}_{i+1/2}^{n,-}(s)$ is a parametrization of an arc of γ.
- If W_{i+1}^n belongs to a curve $\gamma \in \Gamma_0$ then $s \in [0,1] \mapsto \widetilde{P}_{i+1/2}^{n,+}(s)$ is a parametrization of an arc of γ.
- If $\sigma_i = \sigma_{i+1} = \sigma$ then $\sigma_{i+1/2} = \sigma$.
- If $\sigma_i = \sigma_{i+1/2}$ then $W_{i+1/2}^{n,-} = W_i^n$.
- If $\sigma_{i+1/2} = \sigma_{i+1}$ then $W_{i+1/2}^{n,+} = W_i^n$.
- If both the states W_i^n and W_{i+1}^n belong to a curve $\gamma \in \Gamma_0$, then $W_{i+1/2}^{n,-} = W_{i+1/2}^{n,+}$.

This numerical scheme is path-conservative with respect to the family of paths defined as follows: the path linking W_i^n and W_{i+1}^n is the union of the path (6.3), the straight segment linking $W_{i+1/2}^{n,-}$ and $W_{i+1/2}^{n,+}$, and finally the path (6.4).

If it is possible to choose the functions $\widetilde{P}_{i+1/2}^{n,\pm}(s)$ in such a way that, for every pair of states W_i^n and W_{i+1}^n, the paths (6.3) and (6.4) are parametrizations of the integral curves of the linearly degenerate field passing by W_i^n and W_i^{n+1}, then the numerical scheme is exactly well-balanced for every integral curve of the linearly degenerate field: see [9] for details. In this reference, this technique was applied to obtain a first order numerical scheme for the shallow water system which is exactly well-balanced for every smooth stationary solution. It was also applied to obtain high-order numerical schemes which are exactly well-balanced for water at rest solutions for both the one and the two layer shallow-water system.

References

1. Abgrall, R., Karni, S.: Two-layer shallow water system: a relaxation approach. SIAM J. Sci. Comput. **31**, 1603–1627 (2009)
2. Abgrall, R., Karni, S.: A comment on the computation of non-conservative products. J. Comput. Phys. **229**, 2759–2763 (2010)
3. Alcrudo, F., Benkhaldoun, F.: Exact solutions to the Riemann problem of the shallow water equations with a bottom step. Comput. Fluids **30**, 643–671 (2001)
4. Andrianov, N., Warnecke, G.: On the solution to the Riemann problem for the compressible duct flow. SIAM J. Appl. Math. **64**, 878–901 (2004)
5. Audusse, E., Bouchut, F., Bristeau, M.O., Klein, R., Perthame, B.: A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. Sci. Comput. **25**, 2050–2065 (2004)
6. Bermúdez, A., Vázquez, M.E.: Upwind methods for hyperbolic conservation laws with source terms. Comput. Fluids **23**, 1049–1071 (1994)
7. Bernetti, R., Titarev, V.A., Toro, E.F.: Exact solution of the Riemann problem for the shallow water equations with discontinuous bottom geometry. J. Comput. Phys. **227**, 3212–3243 (2008)
8. Bouchut, F.: Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources. Birkhäuser, Basel (2004)
9. Castro, M.J., Pardo, A., Parés, C.: Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique. Math. Mod. Meth. Appl. Sci. **17**, 2055–2113 (2007)
10. Castro, M.J., LeFloch, P.G., Muñoz-Ruiz, M.L., Parés, C.: Why many theories of shock waves are necessary: convergence error in formally path-consistent schemes. J. Comput. Phys. **227**, 8107–8129 (2008)
11. Castro, M.J., Pardo, A., Parés, C., Toro, E.F.: On some fast well-balanced first order solvers for nonconservative systems. Math. Comput. **79**, 1427–1472 (2010)
12. Dal Maso, G., LeFloch, P.G., Murat, F.: Definition and weak stability of nonconservative products. J. Math. Pures Appl. **74**, 483–548 (1995)
13. De Vuyst, F.: Schémas non-conservatifs et schémas cinétiques pour la simulation numérique d'écoulements hypersoniques non visqueux en déséquilibre thermochimique. Thése de Doctorat de l'Université Paris VI (1994)
14. Goatin, P., LeFloch, P.G.: The Riemann problem for a class of resonant hyperbolic systems of balance laws. Ann. Inst. H. Poincaré Anal. Non Lin. **21**, 881–902 (2004)
15. Godunov, S.K.: A finite difference method for the computation of discontinuous solutions of the equations of fluid dynamics. Mat. Sb. **47**, 357–393 (1959)

16. Gosse, L.: A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms. Comput. Math. Appl. **39**, 135–159 (2000)
17. Karni, S.: Computations of shallow water flows in channels. In: The Proceedings of Numhyp 2009, Conference on Numerical Approximations of Hyperbolic Systems with Source Terms and Applications (2009) (to appear)
18. Greenberg, J.M., LeRoux, A.Y.: A well-balanced scheme for the numerical processing of source terms in hyperbolic equations. SIAM J. Numer. Anal. **33**, 1–16 (1996)
19. Hou, T., LeFloch, P.G.: Why nonconservative schemes converge to wrong solutions: error analysis. Math. Comput. **62**, 497–530 (1994)
20. Lax, P.D.: Weak solutions of nonlinear hyperbolic equations and their numerical computation. Commun. Pure Appl. Math. **7**, 159–193 (1954)
21. Lax, P.D., Wendroff, B.: Systems of conservation laws. Commun. Pure Appl. Math. **13**, 217–237 (1960)
22. LeFloch, P.G.: Shock Waves for Nonlinear Hyperbolic Systems in Nonconservative Form. Institute for Mathematics and its Applications, Minneapolis (1989). Preprint 593
23. Muñoz-Ruiz, M.L., Parés, C.: Godunov method for nonconservative hyperbolic systems. ESAIM: M2AN **41**, 169–185 (2007)
24. Parés, C., Castro, M.J.: On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow water systems. ESAIM: M2AN **38**, 821–852 (2004)
25. Parés, C.: Numerical methods for nonconservative hyperbolic systems: a theoretical framework. SIAM J. Numer. Anal. **44**, 300–321 (2006)
26. Perthame, B., Simeoni, C.: Convergence of the upwind interface source method for hyperbolic conservation laws. In: Thou, Tadmor (eds.) Hyperbolic Problems: Theory, Numerics, Applications. Proceedings of the Ninth International Conference on Hyperbolic Problems, pp. 61–78. Springer, Berlin (2003)
27. Roe, P.L.: Approximate Riemann solvers, paremeter vectors, and difference schemes. J. Comput. Phys. **43**, 357–372 (1981)
28. Roe, P.L.: Upwinding difference schemes for hyperbolic conservation laws with source terms. In: Carasso, Raviart, Serre (eds.) Nonlinear Hyperbolic Problems, Lect. Notes Math., pp. 41–51. Springer, Berlin (1986)
29. Rosatti, G., Begnudelli, L.: The Riemann problem for the one-dimensional free-surface shallow water equations with a bed step: theoretical analysis and numerical simulations. J. Comput. Phys. **229**, 760–787 (2010)
30. Toumi, I.: A weak formulation of Roe's approximate Riemann solver. J. Comput. Phys. **102**, 360–373 (1992)