Université de Rennes 1

UFR Mathématiques
Institut de Recherche Mathématique de Rennes (Irmar)

Master Thesis

Staggered and well-balanced discretization of shallow water equations

Author:
Quan Ba Hong Nguyen

Supervisor:
Prof. Nicolas Seguin

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

UFR Mathématiques

January 30, 2020

Declaration of Authorship

I, Quan Ba Hong Nguyen, declare that this thesis titled, "Staggered and well-balanced discretization of shallow water equations" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

“*The essence of mathematics lies in its freedom.*”

Georg Cantor

UNIVERSITÉ DE RENNES 1

Abstract

Institut de Recherche Mathématique de Rennes (Irmar)
UFR Mathématiques

Master of Science

Staggered and well-balanced discretization of shallow water equations

by Quan Ba Hong Nguyen

In this context, we investigate some staggered well-balanced finite volume scheme for the well-known shallow water equations. For completeness purpose, we first restudy the derivation of Navier-Stokes equations via the mass and momentum balance laws. We also mention some derivations of free surface Euler equations, free surface Bernoulli equations, water waves systems, Saint-Venant systems before using both layer-averaged horizontal surface approximations to derive shallow water equations, both in one (1D) and two dimensions (2D) in the spatial variable.

After deriving the main model, we extend the classical 3-point cell-centered Godunov-type scheme for systems of conservation laws with nonlinear fluxes to 1D shallow water equations and on general admissible meshes. It is proved that the modified Godunov-type scheme for 1D shallow water equations preserves some steady states involving Riemann problems provided the approximate Riemann solvers are suitably chosen to be consistent with some integral forms of the conservation law. In addition, it also preserves entropy provided the approximate Riemann solvers are consistent with some integral forms of the entropy inequality satisfied by shallow water equations. Unfortunately, the modified Godunov-type scheme is complicated in the process of choosing suitable approximate Riemann solvers and thus we need a new approach.

An upwind staggered scheme for 1D shallow water equations developed in Doyen and Gunawan, 2014 (briefly DG-staggered scheme) are then restudied and extended into the whole 1D horizontal fluid domain $\mathbb{R}$ and onto general admissible meshes instead of uniform ones. It is verified again that DG-staggered scheme preserves the water volume, the total mass and, in the flat bottom topography case, the total momentum but not the total horizontal impulse and total energy (i.e., Hamiltonian for shallow water equations). Moreover, DG-staggered scheme also preserves the nonnegativity and positivity of the water height during the survival time of shallow water model provided that the initial water height fulfills these and some CFL-like condition is satisfied. Although DG-staggered scheme is "well-balanced", i.e., preserving some steady states at rest, in the case of the whole horizontal fluid domain being fully wet, it collapses completely for the case when at least one dry area appears in the fluid domain. It is also proved that DG-staggered scheme achieves some balances on kinetic energy, potential and topography energies. At the end of the study of this staggered scheme, when we let both the mesh size and time step tend to zero, then the limit, under some estimates, satisfies the weak formulation and weak form of entropy inequality of 1D shallow water equation. Some idea one high-order staggered upwind schemes are also mentioned.

The main target of this context lies on a new staggered scheme, called "offset equilibrium" and thus briefly denoted as OE-staggered scheme, which is investigated to conquer the dry-wet transition issues where DG-staggered scheme can not overcome. OE-staggered scheme also inherits all the conservation properties satisfied by DG-staggered one, especially the well-balanced property in fully wet domains since they actually coincides in that best-case. The dominance of OE-staggered scheme, compared to DG-staggered one, is prevailed by the fact that it can resolve effortlessly the dry-wet transition issues, provided all necessary notions are readily settled. Some ideas on higher order extension for OE-staggered scheme are also mentioned.

AMS Subject Classifications. 35L65, 65M60, 65M12, 76M12.

Key words. Shallow water equations, steady states, finite volume methods, well-balanced property, positive preserving scheme, entropy preserving scheme.

Contents

List of Abbreviations

PDEs	Partial Differential Equations.
FVM	Finite Volume Methods.
SW	Shallow Water equations (SW), (SW$_1$), (SW$_2$).
1DSW	One-Dimensional Shallow Water equations (1DSW), (1DSW$_1$), (1DSW$_2$).
2DSW	Two-Dimensional Shallow Water equations (2DSW).
NS	Navier-Stokes Equations (NS).
fsE	free surface Euler equations, also full Euler equations (fsE).
hfsE	homogeneous free surface Euler equation (hfsE).
2DhfsE	Two-Dimensional homogeneous free surface Euler equation (2DhfsE).
3DhfsE	Three-Dimensional homogeneous free surface Euler equation (3DhfsE).
fsB	free surface Bernoulli equations (fsB).
L	Laplace equation for velocity potential (L).
ww	water waves system (ww).
w	d-dimensional nonlinear water equation (w).
SV	d-dimensional Saint-Venant system (SV), (SV$_1$), (SV$_2$), (SV$_3$), (SV$_4$).
gss	general steady state (gss).
sss	smooth steady state (sss).
lar	lake at rest steady states (lar$_1$), (lar$_2$), (lar$_3$).
sw	still water (sw).
Eb	Entropy involving bottom topography (Eb).
EIb	Entropy Inequality involving bottom topography (EIb).
EFb	Entropy Flux involving bottom topography (EFb).
ee	entropy equality (EEb), (EE).
e/ef	entropy-entropy flux without the bottom topography (e/ef).
sf	step function (sf).
RH	Rankine-Hugoniot relations (RH).
EI	Entropy Inequality (EI), (EI$_1$).
Ri	continuity of Riemann invariants (Ri).
Cp1DSW	Cauchy problem for One-Dimensional Shallow Water equation (Cp1DSW).
3peccFVS	3-point explicit cell-centered finite volume scheme for (Cp1DSW).
FVS	approximate Finite Volume Solution (FVS).
Rp1DSW	Riemann problem for One-Dimensional Shallow Water equations (Rp1DSW).
GtS	Godunov-type Scheme (GtS), (GtS$_1$).
RpU	Riemann sub-problem (Rp1DU) for the vector $U = (h, hu)$ of water height and momentum.
Rpb	Riemann sub-problem (Rp1Db) for the bottom topography b.
DEI	Discrete Entropy Inequality (DEI).
cDEI	condensed Discrete Entropy Inequality (cDEI).

CFL	Courant-Friedrichs-Lewy condition (CFL_1), (CFL_2), (CFL3), (CFL4).
Rs1DSW	Riemann solution for One-Dimensional Shallow Water equations (Rs1DSW).
Rsss	Riemann-steady state solution (Rsss).
cicl	consistency with the integral form of conservation law (cicl).
sp	stationary property (sp).
ciec	consistency with the integral form of the entropy condition (ciec).
ep	entropy preserving (ep).
dmc	discrete mass conservation equation (dmc).
dmb	discrete momentum balance equation (dmb).
DGsc	Doyan-Gunawan's staggered scheme (DGsc), (DGsc_1).
mc	mass conservation (mc).
Mc	Momentum conservation (Mc).
DGscfb	Doyan-Gunawan's staggered scheme (DGscfb) in the case of flat bottom topography.
hic	horizontal impulse conservation (hic).
dhi	discrete horizontal impulse (dhi).
dte	discrete total energy (dte_1), (dte_2).
rDGsc	restricted Doyan-Gunawan staggered scheme (rDGsc).
rDGscfb	restricted Doyan-Gunawan staggered scheme (rDGscfb) in the case of flat bottom topography.

Physical Constants

Earth-surface gravitational acceleration	$g = 9.80665\ \mathrm{m/s^2}$.
Sea level standard atmospheric pressure	$P_{\mathrm{atm}} = 101325\ \mathrm{Pa}$.

List of Notations

General notations

Ω_t	time-dependent fluid domain defined by (1.2).
H_0	a positive constant reference depth relative to the geoid.
ζ	surface deformation.
b	bottom topography.
u	velocity of the fluid particle inside the time-dependent fluid domain Ω_t.
h	water height defined in Proposition 1.2.
Ω	2D natural phase space.
Ω_b	3D natural phase space involving the bottom topography.
Ω_c	computational horizontal fluid domains.
Φ_b	selected entropy for the shallow water equations (1DSW) involving the bottom topography.
Ψ_b	selected entropy flux for the shallow water equations (1DSW) involving the bottom topography.
$\mathfrak{h}$	water volume (or, total water height).
$\mathcal{Z}$	total mass.
$\mathcal{M}$	total momentum.
$\mathcal{I}$	total horizontal impulse.
$\mathcal{H}_{\mathrm{SV}}$	total energy for Saint-Venant system (SV).
$\mathcal{H}_{\mathrm{SW}}$	total energy for shallow water equations (SW).

Functional spaces for $\Omega \subset \mathbb{R}^{d+1}$

$C(\Omega)$	space of real valued continuous functions on Ω.
$C_c^\infty(\Omega)$	space of all smooth functions with compact support in Ω.
$L^p(\Omega)$	standard Lebesgue spaces on Ω.
$W^{k,p}(\Omega)$	Sobolev spaces.
$H^k(\Omega)$	Sobolev space $W^{2,k}(\Omega)$.
$\dot{H}^{k+1}(\Omega)$	homogeneous Sobolev space on Ω.
$\dot{H}^s\left(\mathbb{R}^d\right)$	Beppo-Levi spaces on $\mathbb{R}^d$.

Dedicated to my beloved father and mother, you are always alive in my heart.

Chapter 1

Introduction to Shallow Water Equations

In this chapter, we introduce the shallow water equations (SW), a system of hyperbolic/parabolic partial differential equations (PDEs) governing fluid flow in coastal regions, estuaries, rivers, channels, and sometimes oceans[1]. The general characteristic of shallow water flows is that the vertical dimension is much smaller than the typical horizontal scale, which allows us to average over the depth to get rid of the vertical dimension. (SW) can be used to predict tides, storm surge levels and coastline reformation due to hurricanes, ocean currents; and to study dredging feasibility. (SW) also arises in atmospheric flows and debris flows.

SW are derived from Navier-Stokes (NS) equation describing the motion of fluids while Navier-Stokes are derived from the equations for conservation of mass and linear momentum. The first section of this chapter is devoted for the derivations of both (NS) and (SW).

Some of recent developments of finite volume schemes for shallow water equations are also mentioned. The main target of the whole context is a one-dimensional shallow water equation (1DSW) involving bottom topography (also called a source term). Under suitable variables, this model can be considered as a quasilinear system of conservation laws. Hence, applying the theory for semilinear systems in Appendix A provides us some primary insights into our main model.

1 Derivations of some models for water

1.1 Derivation of Navier-Stokes equations

On a domain of the fluid, denoted by Ω_t, which will be defined later, consider a control volume $\omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, the *mass balance*[2] (also called a *material balance*) states

[1] See Duchêne, 2019 for the case of shallow free-surface fluid flows, mostly oceans far from the shore, without shoaling effects.

[2] A *mass balance* is an application of conservation of mass to the analysis of physical systems. The general idea of mass conservation is that matter cannot disappear or be created spontaneously. The general idea of mass balance is that the mass entering a system must, by conservation of mass, either leave the system or accumulate within the sytem.

that the rate of change in time of the total mass in ω is equal to the net mass flux[3] across boundary of ω, i.e.,

$$\frac{d}{dt}\int_{\omega}\rho\left(t,x\right)dx=-\int_{\partial\omega}\left(\rho u\right)\left(t,x\right)\cdot\mathbf{n}\left(x\right)dS,$$

where $\rho:\mathbb{R}_{+}\times\Omega_{t}\to\mathbb{R}$ is the fluid/mass density (kg/m^{3}), $u:\mathbb{R}_{+}\times\Omega_{t}\to\mathbb{R}^{d}$ is the fluid velocity (m/s), and $\mathbf{n}\left(x\right)$ is the usual outward unit normal vector on ω.

Applying Gauss's theorem[4] (also called divergence theorem, or Ostrogradsky's theorem) gives

$$\frac{d}{dt}\int_{\omega}\rho\left(t,x\right)dx=-\int_{\omega}\nabla\cdot\left(\rho u\right)\left(t,x\right)dx.$$

Assuming that ρ is smooth, Leibniz integral rule allows us to interchange the differentiation in time and the integral to obtain

$$\int_{\omega}\left(\partial_{t}\rho+\nabla\cdot\left(\rho u\right)\right)\left(t,x\right)dx=0.$$

Since ω is arbitrary, the last equality gives the following conservation of mass[5] equation

$$\partial_{t}\rho+\nabla\cdot\left(\rho u\right)=0,\ \text{ in }\mathbb{R}_{+}\times\Omega_{t}.$$

Next, the *linear momentum balance*[6] over ω states that the rate of change in time of the total momentum in ω is equal to the sum of the net momentum flux[7] across boundary $\partial\omega$, body forces[8] acting on ω, and external contact forces[9] acting on $\partial\omega$, i.e.,

$$\frac{d}{dt}\int_{\omega}\rho u dx=-\int_{\partial\omega}\left(\rho u\right)u\cdot\mathbf{n}dS+\int_{\omega}\rho\mathbf{B}dx+\int_{\partial\omega}\mathbf{T}\mathbf{n}dS,$$

where $\mathbf{B}:\mathbb{R}_{+}\times\Omega_{t}\to\mathbb{R}$ is the body force density per unit mass acting on the fluid (N/kg), and $\mathbf{T}$ is the Cauchy stress tensor[10] (N/m^{2}).

[3] In physics and engineering contexts, *mass flux* is the rate of mass flow per unit area, perfectly overlapping with the momentum density. Mass flux also refers to an alternate form of flux including the molecular mass in Fick's law, or flux including the mass density in Darcy's law.

[4] Gauss's theorem states that the sum of all sources, with sinks regarded as negative sources, gives the net flux out of a region.

[5] The *law of conservation of mass*, also *principle of mass conservation*, states that for any system closed to all transfers of matter and energy, the mass of the system must remain constant over time, i.e., the quantity of mass is conserved over time. However, mass is not generally conserved in open systems.

[6] In Newtonian mechanics, Newton's second law of motion states that a body's rate of change in momentum is equal to the net force acting on it.

[7] A *momentum flux* is defined as the rate of transfer of momentum across a unit volume.

[8] A *body force* is a force acting throughout the volume of a body, e.g., forces due to gravity, electric fields, magnetic field, or fictitious/inertial forces such as centrifugal force, Euler force, Coriolis effect, etc. Body forces contrast with contact forces or surface forces exerted to the surface of an object.

[9] A *contact force* is any force occurring due to contacts. Contact forces are ubiquitous and are responsible for most visible interactions between macroscopic collections of matter.

[10] In continuum mechanics, the *Cauchy stress tensor* $\mathbf{T}$, also *true stress tensor*, is a second order tensor consisting 9 components $\mathbf{T}_{ij}$, $i,j\in\{1,2,3\}$ that completely define the state of stress at a point inside a material in the deformed state, placement, or configuration. The Cauchy stress tensor obeys

Applying Gauss's theorem again yields

$$\frac{d}{dt}\int_{\omega}\rho u dx+\int_{\omega}\nabla\cdot(\rho u\otimes u)\,dx-\int_{\omega}\rho\mathbf{B}dx-\int_{\omega}\nabla\cdot\mathbf{T}dx=0,$$

where the second term is carried as

$$\begin{aligned}\int_{\partial\omega}(\rho u)\,u\cdot\mathbf{n}dS=\int_{\partial\omega}\rho u\sum_{i=1}^{d}u_i\mathbf{n}_i dS&=-\int_{\omega}\sum_{i=1}^{d}\partial_{x_i}(\rho u_i u)\,dx\\&=-\int_{\omega}\sum_{i=1}^{d}(\partial_{x_i}\rho u_i u+\rho\partial_{x_i}u_i u+\rho u_i\partial_{x_i}u)\,dx\\&=-\left(\int_{\omega}\sum_{i=1}^{d}(\partial_{x_i}\rho u_i u_j+\rho\partial_{x_i}u_i u_j+\rho u_i\partial_{x_i}u_j)\,dx\right)_{j=1}^{d}\\&=-\int_{\omega}(\partial_{x_j})^{\top}_{j=\overline{1,d}}(\rho u_i u_j)^{d}_{i,j=1}\,dx\\&=-\int_{\omega}\nabla\cdot(\rho u\otimes u)\,dx.\end{aligned}$$

Assume ρu is smooth, applying Leibniz integral rule again yields

$$\int_{\omega}(\partial_t(\rho u)+\nabla\cdot(\rho u\otimes u)-\rho\mathbf{B}-\nabla\cdot\mathbf{T})\,dx=0.$$

Since ω is arbitrary, the last equality gives the following conservation of momentum equation

$$\partial_t(\rho u)+\nabla\cdot(\rho u\otimes u)-\rho\mathbf{B}-\nabla\cdot\mathbf{T}=0,\ \text{in }\mathbb{R}_+\times\Omega_t.$$

Combining the differential forms of conservation of mass and linear momentum equations gives

$$\begin{cases}\partial_t\rho+\nabla\cdot(\rho u)=0,\\ \partial_t(\rho u)+\nabla\cdot(\rho u\otimes u)=\rho\mathbf{B}+\nabla\cdot\mathbf{T},\end{cases}\text{in }\mathbb{R}_+\times\Omega_t. \tag{1.1}$$

Now, some assumptions about the fluid, density ρ, body forces $\mathbf{B}$ and Cauchy stress tensor T are made to obtain (NS):

a) *Fluid.* The fluid is incompressible, i.e., ρ does not depend on P. However, this does not necessarily mean that ρ is constant. For instance, in ocean modeling, ρ depends on the salinity and temperature of the sea water. Next, both salinity and temperature are assumed to be constant throughout $\mathbb{R}_+\times\Omega$, thus $\rho=\rho_0$. Then (1.1) becomes

$$\begin{cases}\nabla\cdot u=0,\\ \partial_t u+\nabla\cdot(u\otimes u)=\mathbf{B}+\dfrac{1}{\rho_0}\nabla\cdot\mathbf{T},\end{cases}\text{in }\mathbb{R}_+\times\Omega_t.$$

the tensor transformation law under a change in the system of coordinates, which can be graphically represented by the Mohr's circle for stress.

The first equation is the *mass continuity*. Sea water is a Newtonian fluid, which affects the form of T.

b) *Body force.* Since gravity is one body force, so

$$\rho_0 \mathbf{B} = -\rho_0 g \mathbf{e}_z + \rho_0 \mathbf{B}_{\text{others}},$$

where $-g\mathbf{e}_z$ is the acceleration due to gravity (m/s^2) and $\mathbf{e}_z = (0,1)$ if $d=2$ and $\mathbf{e}_z = (0,0,1)$ if $d=3$, and $\mathbf{B}_{\text{others}}$ are other body forces (e.g., Coriolis force in rotating reference frames) which will be neglected.

For a Newtonian fluid,

$$\mathbf{T} = -PI_d + \overline{\mathbf{T}},$$

where P is the pressure and $\overline{\mathbf{T}}$ is a matrix of stress terms.

Therefore, 3D Navier-Stokes equations are given by

$$\begin{cases} \nabla \cdot u = 0, & \text{in } \mathbb{R}_+ \times \Omega_t, \\ \partial_t u + \nabla \cdot (u \otimes u) = -\dfrac{1}{\rho_0} \nabla P - g\mathbf{e}_z + \dfrac{1}{\rho_0} \nabla \cdot \overline{\mathbf{T}}, & \text{in } \mathbb{R}_+ \times \Omega_t, \end{cases} \tag{NS}$$

explicitly,

$$\begin{cases} \partial_x u_x + \partial_y u_y + \partial_z u_z = 0, \\ \partial_t (\rho u_x) + \partial_x \left(\rho u_x^2\right) + \partial_y (\rho u_x u_y) + \partial_z (\rho u_x u_z) = \partial_x (\tau_{xx} - P) + \partial_y \tau_{xy} + \partial_z \tau_{xz}, \\ \partial_t (\rho u_y) + \partial_x (\rho u_x u_y) + \partial_y \left(\rho u_y^2\right) + \partial_z (\rho u_y u_z) = \partial_x \tau_{xy} + \partial_y (\tau_{yy} - P) + \partial_z \tau_{yz}, \\ \partial_t (\rho u_z) + \partial_x (\rho u_x u_z) + \partial_y (\rho u_y u_z) + \partial_z \left(\rho u_z^2\right) = -\rho g + \partial_x \tau_{xz} + \partial_y \tau_{yz} + \partial_z (\tau_{zz} - P). \end{cases}$$

1.2 Derivation of free surface Euler equations

More generally with varied density but more specifically, we turn our attention to the full Euler equations, which are used to predict the evolution of an infinite layer of a fluid (typically water) delimited above by a free surface and below by a moving bottom under the effect of gravity. Borrow the setting in Duchêne, 2019 and Lannes, 2013, assume that no surging waves are allowed, the domain $\Omega_t \subset \mathbb{R}^{d+1}$ occupied by the fluid at time t is denoted by

$$\Omega_t := \left\{ (x,z) \in \mathbb{R}^{d+1}; -H_0 + b(t,x) < z < \zeta(t,x) \right\}, \tag{1.2}$$

where $d \in \{1,2\}$ is the horizontal dimension, x and z are the horizontal and vertical variables, respectively, $H_0 > 0$ is a constant reference depth relative to the geoid, note that $z=0$ corresponds to the still water level. The bottom topography and the surface deformation are also denoted by

$$\begin{aligned} \Gamma_{\text{bot}} &:= \left\{ (x,z) \in \mathbb{R}^{d+1}; z = -H_0 + b(t,x) \right\}, \\ \Gamma_{\text{top}} &:= \left\{ (x,z) \in \mathbb{R}^{d+1}; z = \zeta(t,x) \right\}. \end{aligned}$$

Having in our mind how conservation laws are used to derive Navier-Stokes models for fluids in previous section, we now consider the derivation of a more general one: water waves model. The water waves problem concentrates on describing the motion, under the influence of gravity, of a fluid occupying a domain delimited below by a fixed bottom and above by a free surface separating it from vacuum, e.g., air-water interface. Listed in Lannes, 2013, the following assumptions are made on the fluid and the flow:

(H1) The fluid is homogeneous and inviscid.

(H2) The fluid is incompressible.

(H3) The flow is irrotational.

(H4) The surface and the bottom can be parametrized as graphs above the still water level.

(H5) The fluid particles do not cross the bottom.

(H6) The fluid particles do not cross the surface, i.e., no surging waves are allowed.

(H7) There is no surface tension and the external pressure is constant.

(H8) The fluid is at rest at infinity.

(H9) The water depth is always bounded from below by a nonnegative constant.

To describe these statements mathematically, we reuse the notations in the previous section and introduce the following notation:

- The constant acceleration of gravity is denoted by $-g\mathbf{e}_z$, where g is the gravitational constant defined in Physical Constants table, and $\mathbf{e}_z$ is the unit upward vector in the vertical direction.
- The velocity $u : [0, T_0) \times \Omega_t \to \mathbb{R}^{d+1}$ of the fluid particle located at $(x, z) \in \Omega_t$ is written $u\,(t, x, z)$. Its horizontal and vertical components are denoted by $v\,(t, x, z) \in \mathbb{R}^d$ and $w\,(t, x, z) \in \mathbb{R}$, respectively.
- Introduce the density of the fluid $\rho : [0, T_0) \times \Omega_t \to \mathbb{R}$, this density will be assumed to be constant after the free surface Euler equation is established.
- Introduce the pressure $P : [0, T_0) \times \Omega_t \to \mathbb{R}$ inside the fluid, it is not an unknown but rather the Lagrange multiplier associated with the incompressibility constrant and can be deduced from other unknowns at any time instant by solving the equation obtained when taking the divergence of the second equation in (fsE) derived later.
- Finally, P_{atm} is the prescribed atmospheric pressure at the surface.

We also denote $\nabla_{x,z}$ the $(d+1)$-dimensional gradient operator while $\nabla := \nabla_x$ is the horizontal gradient operator.

After these initial settings, the first two assumptions, (H1) and (H2), imply that the fluid motion is governed by the incompressible Euler equation inside the fluid domain Ω_t. More explicitly, let homogeneity $\rho = \text{const}$ be active later, (H1) reads (Cf. (1.1), inviscid assumption means $\overline{\mathbf{T}} = 0_d$)

$$\begin{cases} \partial_t \rho + \nabla_{x,z} \cdot (\rho u) = 0, & \text{in } \mathbb{R}_+ \times \Omega_t, \\ \partial_t (\rho u) + \nabla_{x,z} \cdot (\rho u \otimes u) = -\nabla_{x,z} P - \rho g \mathbf{e}_z, & \text{in } \mathbb{R}_+ \times \Omega_t. \end{cases} \tag{1.3}$$

To simplify the second equation in (1.3), we need the following identity.

Lemma 1.1. *Let* $\rho \in L^\infty \left(\mathbb{R}_+ \times \Omega_t, \mathbb{R}_+^\star\right)$, $u \in L^\infty \left(\mathbb{R}_+ \times \Omega_t, \mathbb{R}^{d+1}\right)$, *then*

$$\nabla_{x,z} \cdot (\rho u \otimes u) = u \nabla_{x,z} \cdot (\rho u) + \rho \left(u \cdot \nabla_{x,z}\right) u, \ \textit{in } \mathbb{R}_+ \times \Omega_t.$$

Consequently,

$$\begin{cases} \partial_t \rho + \nabla_{x,z} \cdot (\rho u) = 0, \\ \partial_t u + \left(u \cdot \nabla_{x,z}\right) u = -\dfrac{1}{\rho} \nabla_{x,z} P - g \mathbf{e}_z, \end{cases} \textit{in } \mathbb{R}_+ \times \Omega_t.$$

Proof. It is straightforward that

$$\begin{aligned} \nabla_{x,z} \cdot (\rho u \otimes u) &= \left(\sum_{i=1}^{d+1} \partial_{x_i} \left(\rho u_i u_j\right) \right)_{j=1}^{d+1} \\ &= \left(\sum_{i=1}^{d+1} \rho u_i \partial_{x_i} u_j \right)_{j=1}^{d+1} + \left(u_j \sum_{i=1}^{d+1} \partial_{x_i} \left(\rho u_i\right) \right)_{j=1}^{d+1} \\ &= \rho \left(u \cdot \nabla_{x,z}\right) u + u \nabla_{x,z} \cdot (\rho u), \ \text{in } \mathbb{R}_+ \times \Omega_t. \end{aligned}$$

Consequently, subtracting the product of u and first equation in (1.3) by the second one yields the desired result. □

The incompressibility constraint means

$$\frac{d}{dt} \rho \left(t, x(t), z(t)\right) = 0, \ \ \forall t \in \mathbb{R}_+,$$

i.e.[11]

$$\partial_t \rho + u \cdot \nabla_{x,z} \rho = 0, \ \text{in } \mathbb{R}_+ \times \Omega_t, \tag{1.4}$$

where U is the trajectory of the fluid

$$u\left(t, x(t), z(t)\right) = \begin{pmatrix} \dot{x}(t) \\ \dot{z}(t) \end{pmatrix}, \ \ \forall t \in \mathbb{R}_+.$$

[11] Denote by

$$D_t \rho := \partial_t \rho + u \cdot \nabla_{x,z} \rho, \ \text{in } \mathbb{R}_+ \times \Omega_t,$$

the material derivative, also total derivative or Lagrangian derivative, then (1.4) states that the material derivative is zero everywhere in the fluid and for all the time.

Subtracting (1.4) from the conservation of mass equation yields

$$\nabla_{x,z} \cdot u = 0, \text{ in } \mathbb{R}_+ \times \Omega_t.$$

The irrotationality assumption (H3) curl $u = 0$ is useful but not necessary. However, it is commonly made in coastal oceanography because rotational effects are negligible up to the breaking point of the waves for most applications. In the homogeneous setting, since all the forces in the right-hand side of the second equation in (fsE) are potential, the irrotational assumption needs only to be activated initially, and it is automatically propagated by the equations for positive times.

The assumption (H4) means that there exist two functions $b : \mathbb{R}^d \to \mathbb{R}$ and $\zeta : [0, T_0) \times \mathbb{R}^d \to \mathbb{R}$, for some finite time $T_0 > 0$ such that (1.2) holds for all $t \in [0, T_0)$. This excludes overhanging waves, which does not affect our purpose of describing asymptotic dynamics in coastal oceanography. Nevertheless, describing the free surface with a parametrized hypersurface possibly handles with overhanging waves.

Assumption (H5) and (H6) provide boundary conditions to Euler equations. The former is the *impermeability condition* ensuring that no fluid particle shall cross the bottom:

$$\partial_t b - \sqrt{1 + |\nabla b|^2} u \cdot \mathbf{n} = 0, \text{ on } \Gamma_{\text{bot}},$$

or equivalently,

$$\partial_t b = w - \nabla \zeta \cdot v, \text{ on } \Gamma_{\text{bot}},$$

while the latter provide a nonlinear *kinetic boundary condition* at the surface which ensures that no fluid particle shall cross the surface, i.e., fluid particles at the surface are forever "trapped" at the surface:

$$\partial_t \zeta - \sqrt{1 + |\nabla \zeta|^2} u \cdot \mathbf{n} = 0, \text{ on } \Gamma_{\text{top}},$$

or equivalently,

$$\partial_t \zeta = w - \nabla \zeta \cdot v, \text{ on } \Gamma_{\text{top}}.$$

About the derivations of (H5) and (H6), let Γ_t be a hypersurface given implicitly by an equation $\gamma(t, x, z) = 0$, and denoted by $M(t) = (x(t), z(t))$ the position of a fluid particle at time t. It is on the hypersurface Γ_t if and only if $\gamma(t, M(t)) = 0$ and stays on Γ_t for all times if

$$\frac{d}{dt} \gamma(t, M(t)) = 0, \quad \forall t \in [0, T_0),$$

or equivalently,

$$\partial_t \gamma + \frac{d}{dt} M \cdot \nabla_{x,z} \gamma = 0, \quad \forall t \in [0, T_0).$$

Since by definition $\frac{d}{dt}M = u$, we get

$$\partial_t \gamma + u \cdot \nabla_{x,z} \gamma = 0, \ \ \forall t \in [0, T_0),$$

i.e., the material derivative of γ vanishes: $D_t \gamma = 0$ for all $t \in [0, T_0)$. Then taking $\gamma(t, x, z) = z + H_0 - b(t, x)$[12], $\gamma(t, x, z) = z - \zeta(t, x)$ yields the above mathematical description of (H5) and (H6).

Next, the idea of neglecting the surface tension as in (H7) is completely reasonable in coastal oceanography since the typical scale for which surface tensions occurs is 1.6 cm, see Lannes, 2013, Chapter 9 for more details. So we have

$$P = P_{\mathrm{atm}}, \ \text{on } \Gamma_{\mathrm{top}}.$$

Nevertheless, it is worth mentioning that in Duchêne, 2019, the author assume that the pressure jump at the surface is proportional to the mean curvature of the surface, with the constant σ denoting the *surface tension* coefficient, i.e.,

$$P - P_{\mathrm{atm}} = -\sigma \nabla \cdot \left(\frac{\nabla \zeta}{\sqrt{1 + |\nabla \zeta|^2}} \right), \ \text{on } \Gamma_{\mathrm{top}}.$$

Return to (H7), although the external pressure is assumed to be constant, dealing with a nonconstant external pressure does not raise any particular difficulty.

Assumption (H8), called *finite energy* assumptions

$$\forall t \in [0, T_0), \lim_{(x,z) \in \Omega_t, \|(x,z)\|_2 \to \infty} (\zeta(t, x) + \|u(t, x, z)\|) = 0, \ \lim_{\|x\|_2 \to \infty} \rho(x, z) = \underline{\rho}(z)$$

is quite natural as long as one considers infinite domains satisfying (H9):

$$\exists H_{\min} > 0, \quad \text{in } [0, T_0) \times \mathbb{R}^d, \ \ H_0 + \zeta(t, x) - b(t, x) \geq H_{\min}. \tag{1.5}$$

The latter condition excludes beaches, seen as vanishing shorelines, and in particular shoaling effects. Obviously, this is a serious restriction for applications to coastal oceanography but removing this remains an open mathematical problem.

In addition to (H1)-(H9), we impose that (see Duchêne, 2019, p. 3) the density does not vanish on the fluid domain or on the boundaries and enforce the *non-cavitation assumption*, i.e., $\rho > 0$ everywhere in the fluid body for all $t \in [0, T_0)$. This also means that the depth of the layer nowhere vanishes, i.e., the water height satisfies $h(t, x) > 0$ everywhere in Ω_t for all the time $t \in [0, T_0)$.

After translating (H1)-(H9) into mathematical languages, the following *free surface Euler equations*, or *full Euler equations*, are introduced to predict the evolution of the

[12] In Lannes, 2013, p. 3, the author make a mistake in the footnote. He took $\gamma(t, x, z) = z - H_0 + b(t, x)$, which must be $\gamma(t, x, z) = z - H_0 + b(t, x)$.

surface deformation Γ_{top} and the velocity field u inside the layer:

$$\begin{cases} \partial_t \rho + \nabla_{x,z} \cdot (\rho u) = 0, & \text{in } \Omega_t, \\ \partial_t u + (u \cdot \nabla_{x,z})\, u = -\frac{1}{\rho}\nabla_{x,z} P - g\mathbf{e}_z, & \text{in } \Omega_t, \\ \nabla_{x,z} \cdot u = 0, & \text{in } [0, T_0) \times \Omega_t, \\ \partial_t \zeta = w - v \cdot \nabla \zeta, & \text{on } \Gamma_{\text{top}}, \\ \partial_t b = w - v \cdot \nabla b, & \text{on } \Gamma_{\text{bot}}, \\ P - P_{\text{atm}} = -\sigma \nabla \cdot \left(\frac{\nabla \zeta}{\sqrt{1 + |\nabla \zeta|^2}} \right), & \text{on } \Gamma_{\text{top}}. \end{cases} \tag{fsE}$$

The assumptions (H1)-(H9), except the homogeneity in (H1), were made in order to write (fsE), e.g., we neglected the effects of compressibility, viscosity, and friction at the bottom. The motivation for this is the fact that when considering a large body of water with relatively mild behavior, these effects are expected to have almost no contribution on the evolution of the flow. The Coriolis effect is also neglected in the previous section, as well as the curve of the Earth. This assumption is valid provided the body of water being considered is not too large.

To avoid unnecessary complexities, we first discard the surface tension effects, i.e., $\sigma = 0$ on Γ_{top} and thus the last equation in (fsE) becomes

$$P = P_{\text{atm}}, \text{ on } \Gamma_{\text{top}}.$$

However, the surface tension component has very important theoretical consequences for some problems due its strong capacity to modify the high-frequency behavior of the equations, e.g., when looking for traveling waves solutions, or for the well-posedness theory in the bilayer setting. The bottom is also assumed to be time-independent:

$$\partial_t b = 0, \text{ on } \Gamma_{\text{bot}},$$

and the atmospheric pressure at the surface is assumed to be uniform in space,

$$\nabla P_{\text{atm}} = 0, \text{ on } \Gamma_{\text{top}}.$$

It must be emphasize that it is not difficult to add the effects of atmospheric or topographic changes in the models. These effect then can be used to study the generation of waves and their propagation.

In Duchêne, 2019, in addition to (1.5), the author also assumes that there exists $\rho_{\min}$ positive constants such that

$$\rho \geq \rho_\star > 0, \text{ in } [0, T_0) \times \Omega_t,$$

which is quite natural in the oceanographic context, but not if (fsE) is applied to the atmospheric motion.

It is high time to activate the homogeneity assumption in (H1) as promised before. Assume the fluid is homogeneous, i.e., there exists a constant $\rho_0 > 0$ such that

$$\rho \equiv \rho_0 \text{ in } \Omega_t.$$

This assumption needs only to be activated initially in time, since the mass conservation and incompressibility constraint will automatically propagate this for positive times.

At this moment, after removing some germs of unnecessary complexitites, (fsE) becomes the following *homogeneous free surface Euler* equation

$$\begin{cases} \partial_t u + (u \cdot \nabla_{x,z})\, u = -\dfrac{1}{\rho_0} \nabla_{x,z} P - g\mathbf{e}_z, & \text{in } \Omega_t, \\ \nabla_{x,z} \cdot u = 0, & \text{in } \Omega_t, \\ \text{curl } u = 0, & \text{in } \Omega_t, \\ \partial_t \zeta = w - v \cdot \nabla \zeta, & \text{on } \Gamma_{\text{top}}, \\ w = v \cdot \nabla b, & \text{on } \Gamma_{\text{bot}}, \\ P = P_{\text{atm}}, & \text{on } \Gamma_{\text{top}}. \end{cases} \tag{hfsE}$$

In 2D, (hfsE) reads

$$\begin{cases} \partial_t v + v \partial_x v + w \partial_z v = -\dfrac{1}{\rho_0} \partial_x P, & \text{in } \Omega_t, \\ \partial_t w + v \partial_x w + w \partial_z w = -\dfrac{1}{\rho_0} \partial_z P - g, & \text{in } \Omega_t, \\ \partial_x v + \partial_z w = 0, & \text{in } \Omega_t, \\ \partial_z v = \partial_x w, & \text{in } \Omega_t, \\ \partial_t \zeta = w - v \partial_x \zeta, & \text{on } \Gamma_{\text{top}}, \\ w = v \partial_x b, & \text{on } \Gamma_{\text{bot}}, \\ P = P_{\text{atm}}, & \text{on } \Gamma_{\text{top}}. \end{cases} \tag{2DhfsE}$$

In 3D, (hfsE) reads

$$\begin{cases} \partial_t u_x + u_x \partial_x u_x + u_y \partial_y u_x + u_z \partial_z u_x = -\frac{1}{\rho}\partial_x P, & \text{in } \Omega_t, \\ \partial_t u_y + u_x \partial_x u_y + u_y \partial_y u_y + u_z \partial_z u_y = -\frac{1}{\rho}\partial_y P, & \text{in } \Omega_t, \\ \partial_t u_z + u_x \partial_x u_z + u_y \partial_y u_z + u_z \partial_z u_z = -\frac{1}{\rho}\partial_z P - g, & \text{in } \Omega_t, \\ \partial_x u_x + \partial_y u_y + \partial_z u_z = 0, & \text{in } \Omega_t, \\ \partial_y u_z = \partial_z u_y, & \text{in } \Omega_t, \\ \partial_z u_x = \partial_x u_z, & \text{in } \Omega_t, \\ \partial_x u_y = \partial_y u_x, & \text{in } \Omega_t, \\ \partial_t \zeta = u_z - u_x \partial_x \zeta - u_y \partial_y \zeta, & \text{on } \Gamma_{\text{top}}, \\ u_z = u_x \partial_x b + u_y \partial_y b, & \text{on } \Gamma_{\text{bot}}, \\ P = P_{\text{atm}}, & \text{on } \Gamma_{\text{top}}. \end{cases} \tag{3DhfsE}$$

1.3 The free surface Bernoulli equations

The free surface Bernoulli equations are another formulation of the free surface Euler equations based on the representation of the velocity field in terms of a velocity potential.

More precisely, the irrotationality assumption induces

$$u = \nabla_{x,z}\Phi, \text{ in } \Omega_t,$$

where $\Phi(t, x, z) \in \mathbb{R}$ is the *velocity potential*, defined by u up to a time-dependent additive constant.

The *free surface Bernoulli equations* reads

$$\begin{cases} \partial_t \Phi + \frac{1}{2}|\nabla_{x,z}\Phi|^2 + gz = -\frac{1}{\rho_0}(P - P_{\text{atm}}), & \text{in } \Omega_t, \\ \Delta_{x,z}\Phi = 0, & \text{in } \Omega_t, \\ u = \nabla_{x,z}\Phi, & \text{in } \Omega_t, \\ \sqrt{1 + |\nabla b|^2}\partial_{\mathbf{n}}\Phi = 0, & \text{on } \Gamma_{\text{bot}}, \\ \partial_t \zeta - \sqrt{1 + |\nabla \zeta|^2}\partial_{\mathbf{n}}\Phi = 0, & \text{on } \Gamma_{\text{top}}, \\ P = P_{\text{atm}}, & \text{on } \Gamma_{\text{top}}, \end{cases} \tag{fsB}$$

where $\partial_{\mathbf{n}}$ always stands for the upward normal derivative.

Note that the first two equation in the free surface Bernoulli equations are the momentum equation and incompressibility constraint rewritten in terms of the velocity potential. The former is called the Bernoulli equation, which is obtained from an integration in space and should include a time-dependent source term being set to zero by choosing suitable time-dependent additive constant in Φ. The latter is Laplace's equation and thus the velocity potential is harmonic. To prepare for the next section, the trace of the velocity potential is defined as

$$\psi\left(t,x\right):=\Phi\left(t,x,\zeta\left(t,x\right)\right),\quad\text{in }\left[0,T_0\right)\times\mathbb{R}^d.$$

1.4 The Zakharov/Craig-Sulem formulation

This section is devoted to summary the Zakharov/Craig-Sulem formulation for waver waves systems, see e.g., Lannes, 2013, pp. 4–5, Duchêne, 2019, pp. 4–7.

Zakharove remarked in 1968 (see Lannes, 2013 and references therein) that the flow of fluid can be fully defined by the knowledge of the free surface elevation ζ and the trace of the velocity potential at the surface $\psi = \Phi|_{z=\zeta}$. A quarter of century later, in Craig, Sulem, and Sulem, 1992, Craig and Sulem, 1993, the authors gave an elegant formulation of the equations involving the Dirichlet-Neumann operator.

The velocity potential can be recovered from its trace at the surface by solving a boundary problem with a nonhomogeneous Neumann condition at the bottom:

$$\begin{cases}\Delta_{x,z}\Phi=0,\ \text{in }\Omega_t,\\ \Phi|_{z=\zeta}=\psi,\\ \partial_{\mathbf{n}}\Phi|_{z=-H_0+b}=0,\end{cases}$$

or more explicitly,

$$\begin{cases}\Delta_{x,z}\Phi=0, & \text{in }\Omega_t,\\ \Phi=\psi, & \text{on }\Gamma_{\text{top}},\\ \partial_z\Phi-\nabla b\cdot\nabla\Phi=0, & \text{on }\Gamma_{\text{bot}}.\end{cases}\tag{L}$$

Note that the resolution of the Laplace equation with Neumann (at the bottom) and Dirichlet (at the surface) boundary conditions is possible under reasonable regularity assumptions, as stated below, on ζ and ψ if the nonvanishing shoreline assumption (1.5) is satisfied and if the flow at rest at infinity as in in the mathematical description for (H8).

The following result is standard in the theory of elliptic operators.

Proposition 1.1 (Existence and uniqueness result for (L))**.** *Let* $\left(\zeta,b\right)\in W^{2,\infty}\left(\mathbb{R}^d\right)$ *such that* (1.5) *holds. Then for any* $\psi\in\dot{H}^2\left(\mathbb{R}^d\right)$*, there exists a unique* $\Phi\in\dot{H}^2\left(\Omega_t\right)$ *strong solution to* (L).

Now, knowing Φ, the velocity field u is obtained the third equation in (fsB) and the pressure P through the first one. This leads us to find a set of two equations determining ζ and ψ, and thus all the physical quantities relevant to the water waves problem.

Following Craig, Sulem, and Sulem, 1992, and Craig and Sulem, 1993, it is convenient to introduce the Dirichlet-Neumann operator mentioned at the beginning of this section.

Definition 1.1 (Dirichlet-Neumann operator)**.** *Given* $(\zeta, b) \in W^{2,\infty}\left(\mathbb{R}^d\right)$ *such that* (1.5) *holds, the Dirichlet-Neumann operator*[13]

$$\begin{aligned} \mathcal{G}[\zeta, b] : \dot{H}^2\left(\mathbb{R}^d\right) &\to H^{\frac{1}{2}}\left(\mathbb{R}^d\right) \\ \psi &\mapsto \left.\left(\partial_z \Phi - \nabla\zeta \cdot \nabla\Phi\right)\right|_{z=\zeta} \end{aligned}$$

is well-defined and continuous. If moreover, $\zeta, b, \psi \in \dot{H}^{2+s_\star}\left(\mathbb{R}^d\right)$ *with* $s_\star > \frac{d}{2}$*, then* $\mathcal{G}[\zeta, b]\,\psi \in \dot{H}^{s_\star}\left(\mathbb{R}^d\right)$*.*

This operator $\mathcal{G}[\zeta, b]$ is linear w.r.t. ψ but highly nonlinear w.r.t. the surface and bottom parametrizations ζ and b, respectively.

There are many properties of the Dirichlet-Neumann operator listed in Duchêne, 2019, p. 5, and Lannes, 2013. However, we only need its relation with layer-averaged velocity for numerical schemes derived later.

Proposition 1.2 (Relation of Dirichlet-Neumann with layer-averaged velocity)**.** *Let* $(\zeta, b) \in W^{2,\infty}\left(\mathbb{R}^d\right)$ *such that* (1.5) *holds. Then for any* $\psi \in \dot{H}^2\left(\mathbb{R}^d\right)$*, the Dirichlet-Neumann operator satisfies the following identity*

$$\mathcal{G}[\zeta, b]\,\psi = -\nabla \cdot \left(h\bar{v}\right), \quad \textit{in } [0, T_0) \times \mathbb{R}^d,$$

where the water height h and the layer-averaged horizontal velocity are defined by

$$\begin{aligned} h(t, x) &:= H_0 + \zeta(t, x) - b(x), \quad \textit{in } [0, T_0) \times \mathbb{R}^d, \\ \bar{v}(t, x) &:= \frac{1}{h(t, x)} \int_{-H_0 + b(x)}^{\zeta(t,x)} \nabla\Phi(t, x, z)\, dz, \quad \textit{in } [0, T_0) \times \mathbb{R}^d. \end{aligned}$$

In particular, $\mathcal{G}[\zeta, b]\,\psi \in \left(\dot{H}^2\right)'$*.*

Applying chain rule yields

$$\begin{aligned} \left.\left(\partial_t \Phi\right)\right|_{z=\zeta} &= \partial_t \psi - \left.\left(\partial_z \Phi\right)\right|_{z=\zeta} \partial_t \zeta, \\ \left.\left(\nabla \Phi\right)\right|_{z=\zeta} &= \nabla \psi - \left.\left(\partial_z \Phi\right)\right|_{z=\zeta} \nabla \zeta, \\ \left.\left(\partial_z \Phi\right)\right|_{z=\zeta} &= \frac{\mathcal{G}[\zeta, b]\,\psi + \nabla\zeta \cdot \nabla\psi}{1 + |\nabla\zeta|^2}. \end{aligned}$$

The Bernoulli equation in (fsB) and the kinetic boundary condition at the surface can now be rewritten as the following (dimensional) *water waves system* of two scalar evolution equations

$$\begin{cases} \partial_t \zeta - \mathcal{G}[\zeta, b]\,\psi = 0, & \text{in } \Omega_t, \\ \partial_t \psi + g\zeta + \dfrac{1}{2}|\nabla\psi|^2 - \dfrac{\left(\mathcal{G}[\zeta, b]\,\psi + \nabla\zeta \cdot \nabla\psi\right)^2}{2\left(1 + |\nabla\zeta|^2\right)} = 0, & \text{in } \Omega_t. \end{cases} \tag{ww}$$

[13] Also written as

$$\mathcal{G}[\zeta, b] : \psi \mapsto \sqrt{1 + |\nabla\zeta|^2}\, \left.\partial_{\mathbf{n}} \Phi\right|_{z=\zeta}.$$

The first equation in (ww) is an evolution on ζ in terms of ζ, ψ and b only while the second one is an evolution on ψ in terms of the same quantities.

Furthermore, even for mildly regular data, in particular a very rough topography, any sufficient regular solution to the homogeneous free surface Euler equations (hfsE) or free surface Bernoulli equations (fsB) satisfies the water-waves systems (ww). The converse also holds, solving the Laplace problem (L), for Φ as well as the one satisfied by $\partial_t \Phi$ and deducing the pressure P.

1.5 Variational structure of the water waves system

Zakharov pointed out that (ww) has a Hamiltonian structure in the canonical variables (ζ, ψ). Indeed, the Hamiltonian $\mathcal{H}(\zeta, \psi)$ is defined as the total energy, summing up the potential and kinetic energies

$$\mathcal{H}(\zeta,\psi) = K(\zeta,\psi) + E(\zeta,\psi),$$

where the kinetic and potential energies K and E are defined as

$$\begin{aligned} K(\zeta,\psi)(t) &:= \frac{1}{2}\int_{\mathbb{R}^d}\int_{-H_0+b(t,x)}^{\zeta(t,x)} |\nabla_{x,z}\Phi(t,x,z)|^2 dz dx = \frac{1}{2}\int_{\mathbb{R}^d} \psi \mathcal{G}[\zeta,b]\psi dx, \\ E(\zeta,\psi)(t) &:= \frac{1}{2}\int_{\mathbb{R}^d} g\zeta^2(t,x)\,dx. \end{aligned}$$

Hence, the Hamiltonian can be written explicitly as

$$\begin{aligned} \mathcal{H}(\zeta,\psi)(t) &= \frac{1}{2}\int_{\mathbb{R}^d}\left(g\zeta^2 + \psi\mathcal{G}[\zeta,b]\psi\right)(t,x)\,dx \\ &= \int_{\mathbb{R}^d}\int_{-H_0+b(t,x)}^{\zeta(t,x)}\left(gz + \frac{1}{2}|\nabla_{x,z}\Phi(t,x,z)|^2\right)dz - \frac{1}{2}(H_0 - b(t,x))^2 dx, \end{aligned}$$

for all $t \in [0, T_0)$.

Therefore, the Hamiltonian $\mathcal{H}$ is a conserved quantity and the water waves systems (ww) takes the following condensed form

$$\partial_t \begin{pmatrix} \zeta \\ \psi \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} \delta_\zeta \mathcal{H} \\ \delta_\psi \mathcal{H} \end{pmatrix}, \text{ in } \Omega_t,$$

where $\delta_\zeta \mathcal{H}$ and $\delta_\psi \mathcal{H}$ denote the functional derivatives, e.g.,

$$\lim_{\varepsilon\to\infty} \frac{\mathcal{H}(\zeta,\psi+\varepsilon\varphi) - \mathcal{H}(\zeta,\psi)}{\varepsilon} = \int_{\mathbb{R}^d} (\delta_\psi \mathcal{H})\varphi dx, \ \ \forall \varphi \in \mathcal{D}\left(\mathbb{R}^d\right).$$

This Hamiltonian structure may also be correlated with a Lagrangian formalism defined as

$$\mathcal{L}_Z(t) := \int_{\mathbb{R}^d} (\psi\partial_t\zeta)(t,x)\,dx - \mathcal{H}(\zeta,\psi)(t), \ \ \forall t \in [0, T_0).$$

Then the water waves systems (ww) follows from Hamilton's principle

$$\delta \int_{t_0}^{t_1} \mathcal{L}_Z(t)\, dt = 0, \;\; \forall t_1, t_2 \in [0, T_0).$$

Using the kinetic boundary condition at the surface in (ww), the Lagrangian can be rewritten as the difference between the kinetic and the potential energies, i.e.,

$$\begin{aligned}\mathcal{L}_Z(t) = K(\zeta,\psi)(t) - E(\zeta,\psi)(t) &= \frac{1}{2}\int_{\mathbb{R}^d} \psi \mathcal{G}[\zeta,b]\,\psi dx - \frac{1}{2}\int_{\mathbb{R}^d} g\zeta^2(t,x)\,dx \\ &= \int_{\mathbb{R}^d}\int_{-H_0+b(t,x)}^{\zeta(t,x)} \left(\frac{1}{2}|\nabla_{x,z}\Phi(t,x,z)|^2 - gz\right) dz + \frac{1}{2}(H_0 - b(t,x))^2 dx,\end{aligned}$$

for all $t \in [0, T_0)$.

A nontrivial consequence of Hamiltonian structure is that it relates, through Noether's theorem, symmetry groups and conserved quantities of the system.

⋆ **Group symmetries.** If (ζ,ψ) is a solution to (ww), then for any $\theta \in \mathbb{R}$, $\left(\zeta^\theta, \psi^\theta\right)$ also satisfies (ww), where

1. *Variation of base level for the velocity potential*

$$\left(\zeta^\theta,\psi^\theta\right)(t,x) := (\zeta,\psi+\theta)(t,x), \quad \text{in } [0,T_0)\times\mathbb{R}^d.$$

2. *Horizontal translation along the direction υ (in the flat bottom case $b = 0$)*

$$\left(\zeta^\theta,\psi^\theta\right)(t,x) := (\zeta,\psi)(t,x-\theta\upsilon), \quad \text{in } [0,T_0)\times\mathbb{R}^d.$$

3. *Time translation*

$$\left(\zeta^\theta,\psi^\theta\right)(t,x) := (\zeta,\psi)(t-\theta,x), \quad \text{in } [0,T_0)\times\mathbb{R}^d.$$

4. *Galilean boost along the direction υ (in the flat bottom case $b = 0$).*

$$\left(\zeta^\theta,\psi^\theta\right)(t,x) := (\zeta,\psi+\theta\upsilon\cdot x)(t,x-\theta\upsilon t), \quad \text{in } [0,T_0)\times\mathbb{R}^d.$$

5. *Horizontal rotation (in dimension $d = 2$ and for a rotation-invariant bottom, $x^\perp \cdot \nabla b = 0$)*

$$\left(\zeta^\theta,\psi^\theta\right)(t,x) := (\zeta,\psi)(t,R_\theta x), \quad \text{in } [0,T_0)\times\mathbb{R}^d.$$

where R_θ is the rotation matrix of angle θ.

⋆ **Preserved quantities.** Here are some related preserved quantities:

1. *Mass*

$$\frac{d\mathcal{Z}(t)}{dt} = 0, \;\; \forall t \in [0,T_0), \text{ where } \mathcal{Z}(t) := \int_{\mathbb{R}^d} \zeta(t,x)\,dx.$$

2. *Horizontal impulse (in the flat bottom case $b = 0$)*

$$\frac{d\mathcal{I}\left(t\right)}{dt}=0,\ \ \forall t\in\left[0,T_0\right),\ \text{where}\ \mathcal{I}\left(t\right):=\int_{\mathbb{R}^d}\left(\zeta\nabla\psi\right)\left(t,x\right)dx.$$

3. *Total energy*

$$\frac{d\mathcal{H}\left(\zeta,\psi\right)\left(t\right)}{dt}=0,\ \ \forall t\in\left[0,T_0\right).$$

4. *Horizontal coordinate of mass centroid times mass (in the flat bottom case $b = 0$)*

$$\frac{d\mathcal{C}\left(t\right)}{dt}=I\left(t\right),\ \ \forall t\in\left[0,T_0\right),\ \text{where}\ \mathcal{C}\left(t\right):=\int_{\mathbb{R}^d}x\zeta\left(t,x\right)dx.$$

5. *Angular impulse (in dimension $d = 2$ and for a rotation-invariant bottom, $x^{\perp}\cdot\nabla b = 0$)*

$$\frac{d\mathcal{A}\left(t\right)}{dt}=0,\ \ \forall t\in\left[0,T_0\right),\ \text{where}\ \mathcal{A}:=\int_{\mathbb{R}^d}\left(\zeta x^{\perp}\cdot\nabla\psi\right)\left(t,x\right)dx,$$

   where $\left(x,y\right)^{\perp}:=\left(-y,x\right)$ for all $x,y\in\mathbb{R}$.

The horizontal impulse and horizontal momentum are directly related after integration by parts: e.g., in dimension $d = 1$:

$$\begin{aligned}\mathcal{M}_x\left(t\right)&:=\int_{\mathbb{R}^d}\int_{-H_0+b(t,x)}^{\zeta(t,x)}\partial_x\Phi\left(t,x,z\right)dzdx\\&=\mathcal{I}\left(t\right)+\lim_{x\to\infty}\int_{-H_0+b(t,x)}^{\zeta(t,x)}\Phi\left(t,x,z\right)dz-\lim_{x\to-\infty}\int_{-H_0+b(t,x)}^{\zeta(t,x)}\Phi\left(t,x,z\right)dz,\end{aligned}$$

and the latter terms are independent in time as a consequence of Bernoulli equation and proposed boundary conditions. The quantities are preserved in a strong sense: their integrand satisfies a conservation law.

1.6 Derivation of shallow water equations

1.6.1 Layer-averaged horizontal approximations

Integrating the second equation in (hfsE) from $z=-H_0+b\left(t,x\right)$ to $z=\zeta\left(t,x\right)$, with the help of Leibniz integral rule, yields

$$\begin{aligned}0&=\int_{-H_0+b(t,x)}^{\zeta(t,x)}\nabla_{x,z}\cdot u\left(t,x,z\right)dz\\&=\int_{-H_0+b(t,x)}^{\zeta(t,x)}\nabla\cdot v\left(t,x,z\right)dz+w|_{z=\zeta}-w|_{z=-H_0+b}\\&=\nabla\cdot\left(\int_{-H_0+b(t,x)}^{\zeta(t,x)}v\left(t,x,z\right)dz\right)-\nabla\zeta\cdot v|_{z=\zeta}+\nabla b\cdot v|_{z=-H_0+b}+w|_{z=\zeta}-w|_{z=-H_0+b}\end{aligned}$$

$$= \nabla \cdot \left(\int_{-H_0+b(t,x)}^{\zeta(t,x)} v\,(t,x,z)\,dz \right) + \partial_t \zeta, \quad \text{in } [0,T_0) \times \mathbb{R}^d.$$

Extend the layer-averaged horizontal velocity is defined in the Proposition 1.2 to an arbitrary vector-valued function $f \in L^\infty_{t,x} L^1_{z,\text{loc}} \left([0,T_0) \times \mathbb{R}^{d+1}, \mathbb{R}^m \right)$ for some $m \in \mathbb{N}$, the layer-averaged horizontal values of f is defined by

$$\bar{f}\,(t,x) := \frac{1}{h\,(t,x)} \int_{-H_0+b(t,x)}^{\zeta(t,x)} f\,(t,x,z)\,dz, \quad \text{in } [0,T_0) \times \mathbb{R}^d.$$

Under this horizontal layer-averaged notation, the last equality can be rewritten in the following condensed form

$$\partial_t \xi + \nabla \cdot (h\bar{v}) = 0, \quad \text{in } [0,T_0) \times \mathbb{R}^d,$$

or equivalently,

$$\partial_t h + \nabla \cdot (h\bar{v}) = 0, \quad \text{in } [0,T_0) \times \mathbb{R}^d.$$

Next, Lemma 1.1 provides us the following identity in the case of $\rho = \rho_0$:

$$(u \cdot \nabla_{x,z})\,u = \nabla_{x,z} \cdot (u \otimes u) - u \nabla_{x,z} \cdot u = \nabla_{x,z} \cdot (u \otimes u)\,, \text{ in } \Omega_t.$$

Hence, we can use the following version of momentum conservation instead:

$$\partial_t u + \nabla_{x,z} \cdot (u \otimes u) = -\frac{1}{\rho_0} \nabla_{x,z} P - g\mathbf{e}_z, \text{ in } \Omega_t.$$

Integrating the both sides of this momentum equation over the layer yields

$$\begin{aligned}
\int_{-H_0+b(t,x)}^{\zeta(t,x)} (\partial_t u + \nabla_{x,z} \cdot (u \otimes u))\,dz &= -\frac{1}{\rho_0} \int_{-H_0+b(t,x)}^{\zeta(t,x)} \nabla_{x,z} P dz - gh\,(t,x)\,\mathbf{e}_z \\
&= \begin{pmatrix} -\frac{1}{\rho_0} \int_{-H_0+b(t,x)}^{\zeta(t,x)} \nabla P dz \\ -\frac{1}{\rho_0} \int_{-H_0+b(t,x)}^{\zeta(t,x)} \partial_z P dz - gh\,(t,x) \end{pmatrix} \\
&= \begin{pmatrix} -\frac{1}{\rho_0} \nabla \left(h\overline{P} \right) + \frac{1}{\rho_0} \left(P_{\text{atm}} \nabla \zeta - P|_{-H_0+b} \nabla b \right) \\ -\frac{1}{\rho_0} \left(P_{\text{atm}} - P|_{-H_0+b} \right) - gh\,(t,x) \end{pmatrix},
\end{aligned}$$

for all $(t,x) \in [0,T_0) \times \mathbb{R}^d$. Applying Leibniz integral rule, the left-hand side of the last equation is equivalent to

$$\begin{aligned}
&\int_{-H_0+b(t,x)}^{\zeta(t,x)} (\partial_t u + \nabla_{x,z} \cdot (u \otimes u))\,dz \\
&= \int_{-H_0+b(t,x)}^{\zeta(t,x)} \partial_t u dz + \sum_{i=1}^{d} \int_{-H_0+b(t,x)}^{\zeta(t,x)} \partial_{x_i} (u_i u)\,dz + \int_{-H_0+b(t,x)}^{\zeta(t,x)} \partial_z (wu)\,dz
\end{aligned}$$

$$
\begin{aligned}
&= \partial_t \left(\int_{-H_0+b(t,x)}^{\zeta(t,x)} u dz \right) - u|_{z=\zeta} \partial_t \zeta \\
&\quad + \sum_{i=1}^{d} \left(\partial_{x_i} \int_{-H_0+b(t,x)}^{\zeta(t,x)} u_i u dz - (u_i u)|_{z=\zeta} \partial_{x_i} \zeta + (u_i u)|_{z=-H_0+b} \partial_{x_i} b \right) \\
&\quad + (wu)|_{z=\zeta} - (wu)|_{z=-H_0+b} \\
&= \partial_t (h\bar{u}) + \sum_{i=1}^{d} \partial_{x_i} (h\overline{u_i u}) - u|_{z=\zeta} \partial_t \zeta \\
&\quad - \sum_{i=1}^{d} (u_i u)|_{z=\zeta} \partial_{x_i} \zeta + \sum_{i=1}^{d} (u_i u)|_{z=-H_0+b} \partial_{x_i} b + (wu)|_{z=\zeta} - (wu)|_{z=-H_0+b} \\
&= \partial_t (h\bar{u}) + \sum_{i=1}^{d} \partial_{x_i} (h\overline{u_i u}) \\
&\quad - [u (w - v \cdot \nabla \xi)]|_{z=\zeta} - \sum_{i=1}^{d} (u_i u)|_{z=\zeta} \partial_{x_i} \zeta + (wu)|_{z=\zeta} \\
&\quad + [u (w - v \cdot \nabla b)]|_{z=-H_0+b} + \sum_{i=1}^{d} (u_i u)|_{z=-H_0+b} \partial_{x_i} b - (wu)|_{z=-H_0+b} \\
&= \partial_t (h\bar{u}) + \sum_{i=1}^{d} \partial_{x_i} (h\overline{u_i u}) - [u (w - v \cdot \nabla \xi)]|_{z=\zeta} - \sum_{i=1}^{d} (u_i u)|_{z=\zeta} \partial_{x_i} \zeta + (wu)|_{z=\zeta} \\
&\quad + [u (w - v \cdot \nabla b)]|_{z=-H_0+b} + \sum_{i=1}^{d} (u_i u)|_{z=-H_0+b} \partial_{x_i} b - (wu)|_{z=-H_0+b} \\
&= \partial_t (h\bar{u}) + \sum_{i=1}^{d} \partial_{x_i} (h\overline{u_i u}) \\
&\approx \partial_t (h\bar{u}) + \sum_{i=1}^{d} \partial_{x_i} (h\overline{u_i}\bar{u}),
\end{aligned}
$$

where the last line can be demonstrated as follows.

Demonstration. For an arbitrary functions $f, g \in W_x^{1,\infty} L_{z,\text{loc}}^2 \left(\mathbb{R}^{d+1}, \mathbb{R}\right)$ and $p \in W^{1,\infty}\left(\mathbb{R}^d, \mathbb{R}_+\right)$, $q \in W^{1,\infty}\left(\mathbb{R}^d, \mathbb{R}\right)$, we have the following approximation

$$
\nabla \cdot \left(p(x)\, \overline{fg}(x)\right) \approx \nabla \cdot \left(p(x)\, \overline{f}(x)\, \overline{g}(x)\right), \quad \forall x \in \mathbb{R}^d,
$$

where

$$
\overline{f}(x) := \frac{1}{p(x)} \int_{q(x)}^{p(x)+q(x)} f(x, z)\, dz.
$$

To illustrate this, we only need the mean value theorem for definite integral:

$$
\begin{aligned}
\nabla \cdot \left(p(x)\, \overline{fg}(x)\right) &= \nabla \cdot \left(p(x)\, (fg)(x, C_1(x))\right), \\
\nabla \cdot \left(p(x)\, \overline{f}(x)\, \overline{g}(x)\right) &= \nabla \cdot \left(p(x)\, f(x, C_2(x))\, g(x, C_3(x))\right),
\end{aligned}
$$

for some functions $C_i(x) \in (q(x), p(x) + q(x))$, $i \in \{1, 2, 3\}$.

The desired approximation is obtained by assuming $C_1(x) \approx C_2(x) \approx C_3(x)$ for all $x \in \mathbb{R}^d$. □

After the above averaging-approximations and these assumptions, we obtain the following dD nonlinear water equation in conservation form

$$\begin{cases} \partial_t h + \nabla \cdot (h\bar{v}) = 0, \\ \partial_t (h\bar{v}) + \nabla \cdot (h\bar{v} \otimes \bar{v}) = -\frac{1}{\rho_0} \nabla \left(h\overline{P}\right) + \frac{1}{\rho_0} \left(P_{\text{atm}} \nabla \zeta - P|_{-H_0+b} \nabla b\right), \\ \partial_t (h\bar{w}) + \sum_{i=1}^{d} \partial_{x_i} (h\overline{u_i}\bar{w}) = -\frac{1}{\rho_0} \left(P_{\text{atm}} - P|_{-H_0+b}\right) - gh(t,x), \end{cases} \tag{w}$$

for all $(t,x) \in [0,T_0) \times \mathbb{R}^d$.

To simplify our framework further, we now examine the momentum equation for vertical velocity. The main simplification here is the following assumption which is crucial for shallow water theory, see e.g., Duchêne, 2019, pp. 8–19.

Assumption 1.1 (Primary assumption in shallow water theory)**.** *Horizontal scales, i.e., typical horizontal wavelength L, are much larger than vertical scales, i.e., undistubed water height H.*

We also neglect vertical accelerations:

Assumption 1.2 (Boundary layer assumption)**.** *The vertical accelerations are neglected:*

$$\partial_t w \approx 0, \ \ \nabla_{x,z} w \approx \mathbf{0}_{(d+1)\times 1}, \ \ \textit{in} \ \ [0,T_0) \times \Omega_t.$$

Hence, shallow water equation can be considered as a boundary layer. This also explains the term "shallow".

By these assumptions and the corresponding scaling arguments, all of the terms in the z-momentum equation except the pressure derivative and the gravity term are small. Then the z-momentum equation collapses to

$$\partial_z P = -\rho_0 g, \ \text{in } \Omega_t.$$

Integrating this equation, with the help of $P|_{z=\zeta} = P_{\text{atm}}$ yields the following *hydrostatic pressure distribution*

$$P(t,x,z) = \rho_0 g (\zeta(t,x) - z) + P_{\text{atm}}, \ \text{in } \Omega_t,$$

in particular, the pressure at the bottom is given by

$$\begin{aligned} P|_{-H_0+b}(t,x) &= \rho_0 g (\zeta(t,x) + H_0 - b(t,x)) + P_{\text{atm}} \\ &= \rho_0 g h(t,x) + P_{\text{atm}}, \quad \text{in } [0,T_0) \times \mathbb{R}^d. \end{aligned}$$

The pressure gradients are then given by

$$\nabla P(t,x,z) = \rho_0 g \nabla \zeta(t,x), \ \text{in } \Omega_t.$$

Under these assumption, (w) becomes

$$\begin{cases} \partial_t h + \nabla \cdot (h\bar{v}) = 0, & \text{in } [0,T_0) \times \mathbb{R}^d, \\ \partial_t (h\bar{v}) + \nabla \cdot (h\bar{v} \otimes \bar{v}) = -gh\nabla\zeta, & \text{in } [0,T_0) \times \mathbb{R}^d, \\ \partial_t (h\bar{w}) + \displaystyle\sum_{i=1}^{d} \partial_{x_i} (h\overline{u_i w}) = 0, & \text{in } [0,T_0) \times \mathbb{R}^d. \end{cases}$$

By Assumptions 1.1 and 1.2, the last equation can be dropped for the sake of simplicity and note that $\nabla\xi = \nabla h + \nabla b$ for all $(t,x) \in [0,T_0) \times \mathbb{R}^d$ to obtain

$$\begin{cases} \partial_t h + \nabla \cdot (h\bar{v}) = 0, & \text{in } [0,T_0) \times \mathbb{R}^d, \\ \partial_t (h\bar{v}) + \nabla \cdot (h\bar{v} \otimes \bar{v}) = -gh\nabla (h+b), & \text{in } [0,T_0) \times \mathbb{R}^d. \end{cases} \tag{SW}$$

As in the proof Lemma 1.1, we have

$$\begin{aligned} \nabla \cdot (h\bar{v} \otimes \bar{v}) &= \left(\sum_{i=1}^{d} \partial_{x_i} (h\bar{v}_i\bar{v}_j) \right)_{j=1}^{d} = \left(\sum_{i=1}^{d} h\bar{v}_i \partial_{x_i} \bar{v}_j \right)_{j=1}^{d} + \left(\bar{v}_j \sum_{i=1}^{d} \partial_{x_i} (h\bar{v}_i) \right)_{j=1}^{d} \\ &= h(\bar{v} \cdot \nabla)\bar{v} + \bar{v}\nabla \cdot (h\bar{v}), \quad \text{in } [0,T_0) \times \mathbb{R}^d. \end{aligned}$$

Plugging this into (SW) yields[14]

$$\begin{cases} \partial_t h + \nabla \cdot (h\bar{v}) = 0, \\ \partial_t (h\bar{v}) + h(\bar{v} \cdot \nabla)\bar{v} + \bar{v}\nabla \cdot (h\bar{v}) + gh\nabla (h+b) = 0, \end{cases} \tag{SW$_1$}$$

in $[0,T_0) \times \mathbb{R}^d$.

1.6.2 Surface approximations

The Saint-Venant system reads

$$\begin{cases} \partial_t \zeta + \nabla \cdot (h\nabla\psi) = 0, & \text{in } [0,T_0) \times \mathbb{R}^d, \\ \partial_t \psi + g\zeta + \dfrac{1}{2}|\nabla\psi|^2 = 0, & \text{in } [0,T_0) \times \mathbb{R}^d. \end{cases} \tag{SV$_1$}$$

Define

$$v_{\text{top}}(t,x) := \nabla\psi = \nabla\left(\Phi|_{z=\zeta}\right) = (\nabla\Phi)|_{z=\zeta} + (\partial_z\Phi)|_{z=\zeta}\nabla\zeta, \quad \text{in } [0,T_0) \times \mathbb{R}^d,$$

[14] Cf with the following system in Duchêne, 2019, p.20:

$$\begin{cases} \partial_t h + \nabla \cdot (h\bar{v}) = 0, & \text{in } [0,T_0) \times \mathbb{R}^d, \\ \partial_t \bar{v} + g\nabla (h+b) + (\bar{v} \cdot \nabla)\bar{v} = 0, & \text{in } [0,T_0) \times \mathbb{R}^d. \end{cases} \tag{SV}$$

It is obvious that the second equation (SW$_2$) can be obtained from (SV) by summing the product of the first equation of the latter with u and the product of the second one with h. Thus, provided enough regularity for h, $\bar{v}$, (SV) and (SW$_2$) are equivalent. However, in the distribution sense, they are not. Although (SW$_2$) is more complicated than (SV), the latter is referred to our main model. The crucial reason for this is the fact that the second equation (SW$_2$) provides us an approximate conservation of momentum while the first one does not.

i.e., $v_{\text{top}} \equiv v|_{z=\zeta}$, taking the gradient of the second equation in (SV$_1$) yields

$$\begin{cases} \partial_t \zeta + \nabla \cdot (h v_{\text{top}}) = 0, & \text{in } [0, T_0) \times \mathbb{R}^d, \\ \partial_t v_{\text{top}} + g \nabla \zeta + \frac{1}{2} \nabla \left(|v_{\text{top}}|^2 \right) = 0, & \text{in } [0, T_0) \times \mathbb{R}^d. \end{cases} \tag{SV$_2$}$$

Due to the condition curl $u = 0$ in $[0, T_0) \times \Omega_t$, one gets

$$\partial x_i v_{\text{top},j} = \partial x_j v_{\text{top},i}, \quad \forall i \neq j, \quad \text{in } [0, T_0) \times \mathbb{R}^d,$$

and thus

$$\begin{aligned} \frac{1}{2} \nabla \left(|v_{\text{top}}|^2 \right) = \frac{1}{2} \nabla \left(\sum_{j=1}^{d} v_{\text{top},j}^2 \right) &= \left(\sum_{j=1}^{d} v_{\text{top},j} \partial_{x_i} v_{\text{top},j} \right)_{i=1}^{d} \\ &= \left(\sum_{j=1}^{d} v_{\text{top},j} \partial_{x_j} v_{\text{top},i} \right)_{i=1}^{d} = (v_{\text{top}} \cdot \nabla) v_{\text{top}}, \quad \text{in } [0, T_0) \times \mathbb{R}^d. \end{aligned}$$

Hence, (SV$_2$) can be rewritten as

$$\text{(SV}_3\text{)} \quad \begin{cases} \partial_t \zeta + \nabla \cdot (h v_{\text{top}}) = 0, & \text{in } [0, T_0) \times \mathbb{R}^d, \\ \partial_t v_{\text{top}} + g \nabla \zeta + (v_{\text{top}} \cdot \nabla) v_{\text{top}} = 0, & \text{in } [0, T_0) \times \mathbb{R}^d. \end{cases} \tag{SV$_3$}$$

Systems (SW$_1$), (SV$_1$) and (SV$_3$) are the prototypes of hyperbolic quasilinear systems, where the strong hyperbolicity is guaranteed by the non-cavitation assumption $h > 0$.

In the flat-bottom case $b \equiv 0$, these Saint-Venant systems correspond to the isentropic, *compressible* Euler equation for ideal gases with the pressure law $P(\rho) \propto \rho^2$.

Using physical variables, (SV$_3$) reads

$$\begin{cases} \partial_t h + \nabla \cdot (h v_{\text{top}}) = 0, & \text{in } [0, T_0) \times \mathbb{R}^d, \\ \partial_t v_{\text{top}} + g \nabla (h + b) + (v_{\text{top}} \cdot \nabla) v_{\text{top}} = 0, & \text{in } [0, T_0) \times \mathbb{R}^d. \end{cases} \tag{SV$_4$}$$

As before, multiply the first equation in (SV$_4$) with v_{top} and the second one with h then adding together yields the following surface-version shallow water equations[15]

$$\begin{cases} \partial_t h + \nabla \cdot (h v_{\text{top}}) = 0, \\ \partial_t (h v_{\text{top}}) + h (v_{\text{top}} \cdot \nabla) v_{\text{top}} + v_{\text{top}} \nabla \cdot (h v_{\text{top}}) + g h \nabla (h + b) = 0, \end{cases} \tag{SW$_2$}$$

in $[0, T_0) \times \mathbb{R}^d$.

To end this section, it should be mentioned that shallow water equations (SW$_1$) (or (SW$_2$)) has several applications, such as tsunamis prediction, atmospheric flows, storm surges, flows around structures, planetary flows. However, (SW) can not be applied in the cases where 3D effects become essential or waves become too short or too high.

[15] Cf., (SW$_1$) is considered as layer-averaged horizontal shallow water equations.

2 The 1D shallow water equations in dry-wet fluid domain

We now pay our attention to the 1D shallow water equation extended by allowing *dry areas* ($h = 0$) in our fluid domain:

$$\begin{cases} \partial_t h + \partial_x (hu) = 0, & \text{in } [0, T_0) \times \mathbb{R}, \\ \partial_t (hu) + \partial_x \left(hu^2 + \frac{g}{2} h^2 \right) = -gh\partial_x b, & \text{in } [0, T_0) \times \mathbb{R}. \end{cases} \tag{1DSW}$$

Setting

$$U := \begin{pmatrix} h \\ hu \end{pmatrix}, \quad F(U) := \begin{pmatrix} hu \\ hu^2 + \frac{g}{2} h^2 \end{pmatrix}, \text{ and } S(U, b) := - \begin{pmatrix} 0 \\ gh\partial_x b \end{pmatrix},$$

where $S(U, b)$ is called the source term, then (1DSW) can be rewritten in the following condensed form

$$\partial_t U + \partial_x (F(U)) = S(U, b), \text{ in } [0, T_0) \times \mathbb{R}.$$

The residuum R is defined as

$$R(t, x) := -\partial_x (F(U)) + S(U, b),$$

which indicates near-equilibrium flows when it nearly vanishes. The natural phase space, denoted by Ω, is defined by[16]

$$\Omega_b := \left\{ (x, xy)^\top \in \mathbb{R}^2; x > 0 \right\} \cup \left\{ (0, 0)^\top \right\}.$$

The source term S models the force of gravity tangential to a sloped bottom[17]. The residuum can be rewritten as

$$R = -\partial_x \begin{pmatrix} hu \\ hu^2 \end{pmatrix} - gh\partial_x \begin{pmatrix} 0 \\ \zeta \end{pmatrix}.$$

For this kind of problem, where shocks can form in the solution, finite volume methods have proved to be very effective. In this context, all the unknowns of the system are approximated on the different meshes, i.e., staggered meshes. Several finite volume discretizations used to solve numerically nonlinear hyperbolic system of conservation laws have been investigated recently (see, e.g., Herbin, Latché, and Nguyen, 2013, Herbin, Kheriji, and Latché, 2014, Stelling and Duinmeijer, 2003).

[16] The definition of the natural phase space Ω also contains the convention: If (x_0, t_0) satisfies $h(x_0, t_0) = 0$ then $u(x_0, t_0) := 0$. This convention will be used later in the discrete level, see e.g., Convention 3.1. It should be emphasized that the element $(0, 0)$ indicates the dry area.

[17] The bottom topography b is assumed to be of class C_1 for simplicity, but well-balanced schemes, which preserves the lake at rest discretely, use either continuous or discontinuous topography, see Chen and Noelle, 2017, p. 760.

2.1 Dry-wet decomposition of the horizontal fluid domain

Concerning the allowance of dry areas in the fluid domain, we first attempt to decompose the horizontal fluid domain $\mathbb{R}$ during $[0, T_0)$ into dry component and wet one.

Mathematically, for each $t \in [0, T_0)$, consider a nondecreasing sequence $\left(x^t_{i+\frac{1}{2}}\right)_{i\in\mathbb{Z}} \subset \overline{\mathbb{R}}$ of extended reals (if $x^t_{i+\frac{1}{2}} = -\infty$ then $x^t_{j+\frac{1}{2}} = -\infty$ for all $j < i$ and thus can be dropped, similarly for the case $x^t_{i+\frac{1}{2}} = \infty$):

$$x^t_{i-\frac{1}{2}} \le x^t_{i+\frac{1}{2}}, \ \ \forall i \in \mathbb{Z}, \ \ \forall t \in [0, T_0)\,,$$

which separates dry and wet areas in the horizontal fluid domain $\mathbb{R}$.

Denote the i^{th} area (dry or wet) by

$$A^t_i := \left(x^t_{i-\frac{1}{2}}, x^t_{i+\frac{1}{2}}\right), \ \ \forall i \in \mathbb{Z}, \ \ \forall t \in [0, T_0)\,,$$

it is also demanded that

$$\forall i \in \mathbb{Z}, \ \ \forall t \in [0, T_0)\,, h|_{A^t_i}(t, x) > 0 \Rightarrow x^t_{i-\frac{1}{2}} < x^t_{i+\frac{1}{2}},$$

which means that the “1-point” dry areas are allowed in our decomposition whereas the “1-point” wet areas are absolutely not, as it is meaningless to do so.

Now, depending on the event that the 0^{th} area is wet or dry, we can index dry and wet areas consecutively as follows:

- *Case* $h|_{A^t_0}(t, x) > 0$: Define the wet and dry areas as

$$W^t_i := A^t_{2i}, \ \ D^t_i := A^t_{2i+1}, \ \ \forall i \in \mathbb{Z}, \ \ \forall t \in [0, T_0)\,.$$

- *Case* $h|_{A^t_0}(t, x) = 0$: Define the wet and dry areas as

$$W^t_i := A^t_{2i+1}, \ \ D^t_i := A^t_{2i}, \ \ \forall i \in \mathbb{Z}, \ \ \forall t \in [0, T_0)\,.$$

Under these settings, the horizontal fluid domain $\mathbb{R}$ can be decomposed as the union of the dry component D^t and the wet one W^t:

$$\mathbb{R} = D^t \cup W^t, \ \ \forall t \in [0, T_0)\,,$$

where

$$D^t := \bigcup_{i=-\infty}^{\infty} D^t_i, \ \ W^t := \bigcup_{i=-\infty}^{\infty} W^t_i, \ \ \forall t \in [0, T_0)\,.$$

By definition, it is straightforward that

$$D^t = \{x \in \mathbb{R}; h(t, x) = 0\}\,, \ \ W^t = \{x \in \mathbb{R}; h(t, x) > 0\}\,, \ \ \forall t \in [0, T_0)\,,$$

and thus our setting makes a sense. We also demand that

$$h|_{D^t} = u|_{D^t} = 0, \text{ in } [0, T_0) \times \mathbb{R},$$

and it should be noted that this is exactly the convention

$$h(t,x) = 0 \Rightarrow u(t,x) = 0, \text{ in } [0, T_0) \times \mathbb{R},$$

which is deliberately embedded in the definition of the natural phase space Ω.

Note that these unions can be both finite, or both infinite. They are both finite provided there exists $i_L < i_R \in \mathbb{Z}$ such that $x_{i_L} = -\infty$ and $x_{i_R} = \infty$.

Having enough settings, we consider the steady states of the shallow water equations (1DSW), after some $T_\star \in (0, T_0)$, called *steady time*, which are the states governed by

$$\begin{cases} \partial_x (hu) = 0, & \text{in } [T_\star, T_0) \times \mathbb{R}, \\ \partial_x \left(hu^2 + \frac{g}{2} h^2 \right) = -gh\partial_x b, & \text{in } [T_\star, T_0) \times \mathbb{R}. \end{cases} \tag{gss}$$

Therefore, the smooth steady states under consideration are given by

$$\begin{cases} (hu)|_{W_i^t} = \overline{C}_i, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ \left(\frac{u^2}{2} + g(h+b) \right)\Big|_{W_i^t} = \overline{\overline{C}}_i, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \end{cases} \tag{sss}$$

where $\left(\overline{C}_i\right)_{i\in\mathbb{Z}}$ and $\left(\overline{\overline{C}}_i\right)_{i\in\mathbb{Z}}$ are two sequences of constants.

Among this full family of steady states, the following steady states, called *lake at rest* (see e.g., Chen and Noelle, 2017, p. 759)

$$\begin{cases} u = 0, & \text{in } [T_\star, T_0) \times \mathbb{R}, \\ h\partial_x (h+b) = 0, & \text{in } [T_\star, T_0) \times \mathbb{R}, \end{cases} \tag{lar$_1$}$$

certainly is of main importance in this context. It is obvious that $h = u = 0$ in the dry areas. Ignoring them, (lar$_1$) can then be rewritten explicitly as

$$\begin{cases} u|_{W_i^t} = 0, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ (h+b)|_{W_i^t} = C_i, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ D_i^t = D_i^{T_\star}, \ W_i^t = W_i^{T_\star}, & \forall i \in \mathbb{Z}, \ \forall t \in [T_\star, T_0), \end{cases} \tag{lar$_2$}$$

or equivalently,

$$\begin{cases} u|_{W_i^t} = 0, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ \zeta|_{W_i^t} = C_i - H_0, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ D_i^t = D_i^{T_\star}, \ W_i^t = W_i^{T_\star}, & \forall i \in \mathbb{Z}, \ \forall t \in [T_\star, T_0). \end{cases} \tag{lar$_3$}$$

In particular, if the horizontal fluid domain is *fully wet*, i.e.,

$$D^t = \emptyset, \ W^t = \mathbb{R}, \ \forall t \in [0, T_0),$$

then $h > 0$ in $[0, T_0) \times \mathbb{R}$ and thus can be dropped in (lar_1). In this fully wet case, (lar_1) becomes the well-known *still water*:

$$\begin{cases} u = 0, & \text{in } [T_\star, T_0) \times \mathbb{R}, \\ h + b = \text{const}, & \text{in } [T_\star, T_0) \times \mathbb{R}. \end{cases} \tag{sw}$$

Thus, the lake at rest residuum (lar_1) combines the dry shore ($h = 0$) with the flat water surface ($\partial_x \zeta = 0$) in a single product, which suggests a natural splitting of the nonconservative product at the wet-dry front.

After the works by Bermudez and Vazquez, 1994 and Greenberg and Leroux, 1996 (see also, e.g. Gosse, 2000, Goutal, 2002), the derivation of well-balanced schemes able to restore the above steady states at rest was a very active research topic. Several strategies have been derived (see, e.g., Audusse et al., 2004, Liang and Marche, 2009, Jin, 2001, Bryson et al., 2011, Gosse, 2000, Greenberg and Leroux, 1996, Buffard, Gallouet, and Hérard, 1998, Noelle et al., 2006, Luna et al., 2009, Parés and Castro, 2004, Noelle, Xing, and Shu, 2007, Castro et al., 2008, Russo and Khe, 2010, Russo and Khe, 2009). The main difficulty occurring from the derivation of well-balanced numerical schemes remains in the discretization of the bottom topography source term to be consistent with the still water (sw). In Audusse et al., 2004, the authors introduced a well-balanced strategy, the so-called hydrostatic reconstruction[18] to enforce the source term discretization to be consistent with the still water (sw). An improvement of the hydrostatic reconstruction as the generalized hydrostatic reconstruction in Castro, Pardo Milanés, and Parés, 2007 or the upwind hydrostatic approach in Berthon and Foucher, 2012 are also worth to be mentioned.

Except for Castro, Pardo Milanés, and Parés, 2007, all the cited references above handle only with the still water steady state but the general steady states (gss) does not receive any considerations. Fortunately, there are some numerical strategies able to restore all the steady states (gss) (see, e.g., Castro, Pardo Milanés, and Parés, 2007, Bouchut and Luna, 2010, Xing, 2014, Xing and Shu, 2006, Xing, Shu, and Noelle, 2011, Xing, Zhung, and Shu, 2010). In Castro, Pardo Milanés, and Parés, 2007, the main idea is a suitable extension of the hydrostatic reconstruction, which produces a scheme able to restore all the steady states (gss). Unfortunately, this scheme fails to preserve the positivity of the water height. In Bouchut and Luna, 2010, a relaxation model is investigated and the resulting numerical scheme is able to preserve the subsonic $\left(|u| < \sqrt{gh}\right)$ steady states. This relaxation method preserve a large family, but not all, of the steady states and it also ensures the positivity of the water height. In addition, this relaxation scheme satisfies discrete entropy inequalities and thus stable.

In general, the preservation of these steady states at rest at the discrete level is considerably valuable since many practical problems are perturbations of such states. With approximate Riemann solvers (e.g., Harten-Lax-van Leer-type Riemann solvers), this preservation is quite entailed Bouchut, 2004, whereas it is straightforward in the staggered framework (on uniform grids w.r.t. each variable).

[18] The hydrostatic reconstruction scheme of Audusse et al., 2004 is considered as *the original HR scheme* in Chen and Noelle, 2017.

The goal of this context is to modify numerical schemes on staggered grids for a precise consideration of these equilibrium states. The lake at rest (lar_1) (also, (lar_2) or (lar_3)) makes it possible to take into account the transitions between dry zones and wet areas, whereas the still water (sw), presuppose water everywhere. We hope, and in face we see, that preserving exactly these states at the discrete level are also more accurate in the transient case.

2.2 Basic algebraic properties of 1D shallow water equations

We recall some properties satisfied by (1DSW) for completeness (see also, e.g., Bouchut, 2004, Bouchut and Luna, 2010, LeFloch and Thanh, 2007). To utilize some properties of general first-order hyperbolic systems (see Appendix A), let us rewrite (1DSW) as the following equivalent form:

$$\begin{cases} \partial_t h + \partial_x (hu) = 0, \\ \partial_t (hu) + \partial_x \left(hu^2 + \frac{g}{2} h^2 \right) + gh\partial_x b = 0, \\ \partial_t b = 0. \end{cases} \tag{1DSW$_1$}$$

2.2.1 Hyperbolicity

Considering the given bottom topography b as a time-dependent unknown, define

$$U_b := \begin{pmatrix} h \\ hu \\ b \end{pmatrix}, \quad F_b(U_b) := \begin{pmatrix} hu \\ hu^2 + \frac{g}{2}h^2 \\ 0 \end{pmatrix}, \quad S_b(U_b) := \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & gh \\ 0 & 0 & 0 \end{pmatrix},$$

and denote Ω_b the *natural phase space* defined by

$$\Omega_b := \left\{ (x, xy, z)^\top \in \mathbb{R}^3; x > 0 \right\} \cup \left\{ (0, 0, z)^\top \in \mathbb{R}^3; z \in \mathbb{R} \right\}.$$

then (1DSW$_1$) can be rewritten in the following condensed form:

$$\partial_t U_b + \partial_x (F_b(U_b)) + S_b(U_b)\partial_x U_b = 0, \text{ in } [0, T_0) \times \mathbb{R}. \tag{1DSW$_2$}$$

The eigenvalues of the matrix

$$\nabla_{U_b} F_b(U_b) + S_b(U_b) = \begin{pmatrix} 0 & 1 & 0 \\ gh - u^2 & 2u & gh \\ 0 & 0 & 0 \end{pmatrix}$$

are given by

$$\lambda_-(U_b) := u - c, \quad \lambda_0(U_b) := 0, \quad \lambda_+(U_b) := u + c,$$

where $c := \sqrt{gh}$ denotes the gravitational speed. And the corresponding eigenvectors are

$$r_{-}(U_b) := \begin{pmatrix} 1 \\ u-c \\ 0 \end{pmatrix}, \quad r_0(U_b) := \begin{pmatrix} c^2 \\ 0 \\ u^2-c^2 \end{pmatrix}, \quad \text{and } r_{+}(U_b) := \begin{pmatrix} 1 \\ u+c \\ 0 \end{pmatrix},$$

forming a basis of $\mathbb{R}^3$ provided that $u \neq \pm c$ and where $h > 0$.

Definition 1.2 (sonic, subsonic, supersonic). *i) The states $U_b \in \Omega_b$ such that $u = \pm c$ are said to be* sonic.

ii) The states $U_b \in \Omega_b$ such that $|u| < c$, resp., $|u| > c$, are said to be subsonic, *resp.,* supersonic.

On wet areas, the system (1DSW$_1$) is then *strictly hyperbolic* on the set of subsonic and supersonic states.

2.2.2 Entropy/entropy flux

We also recall that the *entropy* Φ for the shallow water equations (1DSW$_1$) is the sum of the *kinetic energy*, the *potential energy* and a term stemming from the bottom topography. Explicitly,

$$\Phi_b(U_b) := \Phi_k(U) + \Phi_p(h) + ghb, \quad \forall U_b \in \Omega_b,$$

where

$$\Phi_k(U) := \frac{1}{2}hu^2, \quad \Phi_p(h) := \frac{g}{2}h^2, \quad \forall U \in \Omega,$$

Hence,

$$\Phi_b(U_b) = \frac{1}{2}hu^2 + \frac{g}{2}h^2 + ghb, \quad \forall U_b \in \Omega_b. \tag{Eb}$$

The *entropy inequality* for entropy (weak) solutions reads (see Appendix A)

$$\partial_t(\Phi_b(U_b)) + \partial_x(\Psi_b(U_b)) \leq 0, \text{ a.e. in } [0, T_0) \times \mathbb{R}, \tag{EIb}$$

where the *entropy flux*[19] Ψ_b is defined by

$$\Psi_b(U_b) := \left(\frac{1}{2}hu^2 + gh^2\right)u + ghbu, \quad \forall U_b \in \Omega_b. \tag{EFb}$$

[19] Obviously, the following identity holds

$$\Psi_b(U_b) = \Phi_b(U_b)\,u + \frac{g}{2}h^2u, \text{ in } [0, T_0) \times \mathbb{R}.$$

With the help of a chain rule argument, smooth solutions of (1DSW$_1$) satisfy the following conservation law[20]:

$$\partial_t\left(\Phi_b\left(U_b\right)\right)+\partial_x\left(\Psi_b\left(U_b\right)\right)=0,\ \text{ in }\ [0,T_0)\times\mathbb{R}. \tag{EEb}$$

Proof of the conservation law (EEb). The chain rule will be used in the whole proof. First, the momentum conservation equation in (1DSW) can be expanded as

$$u\partial_t h+h\partial_t u+u^2\partial_x h+2hu\partial_x u+gh\partial_x h=-gh\partial_x b,$$

or equivalently,

$$u\left(\partial_t h+u\partial_x h+h\partial_x u\right)+h\partial_t u+hu\partial_x u+gh\partial_x h=-gh\partial_x b.$$

Combining the last equality with the mass conservation in (1DSW) yields

$$h\partial_t u+hu\partial_x u+gh\partial_x h=-gh\partial_x b.$$

Then the first-order partial derivative in the time variable of the entropy (Eb) is given by

$$\begin{aligned}\partial_t\left(\Phi_b\left(U_b\right)\right)&=\frac{1}{2}u^2\partial_t h+hu\partial_t u+gh\partial_t h+gb\partial_t h\\&=-\frac{1}{2}u^3\partial_x h-\frac{1}{2}hu^2\partial_x u-hu^2\partial_x u-ghu\partial_x h-ghu\partial_x b\\&\quad-ghu\partial_x h-gh^2\partial_x u-gub\partial_x h-ghb\partial_x u\\&=-\left(\frac{1}{2}u^3+2ghu+gub\right)\partial_x h-\left(\frac{3}{2}hu^2+gh^2+ghb\right)\partial_x u-ghu\partial_x b.\end{aligned}$$

And the first-order partial derivative of the entropy flux Ψ_b in the spatial variable is given by

$$\begin{aligned}\partial_x\left(\Psi_b\left(U_b\right)\right)&=\frac{1}{2}u^3\partial_x h+\frac{3}{2}hu^2\partial_x u+2ghu\partial_x h+gh^2\partial_x u+gub\partial_x h+ghu\partial_x b+ghb\partial_x u\\&=\left(\frac{1}{2}u^3+2ghu+gub\right)\partial_x h+\left(\frac{3}{2}hu^2+gh^2+ghb\right)\partial_x u+ghu\partial_x b.\end{aligned}$$

Combining these calculations yields the desired result. □

Remark 1.1 (Another version of entropy equality)**.** *The entropy equality* (EEb) *can be rewritten as*

$$\partial_t\left(\Phi\left(U\right)\right)+\partial_x\left(\Psi\left(U\right)\right)=\nabla\Phi\left(U\right)\cdot S\left(U,b\right)=-ghu\partial_x b,\ \textit{ in }\ [0,T_0)\times\mathbb{R}, \tag{EE}$$

where the pair of entropy-entropy flux without the bottom topography are given by

$$\Phi\left(U\right):=\frac{1}{2}hu^2+\frac{g}{2}h^2,\ \ \Psi\left(U\right):=\left(\frac{1}{2}hu^2+gh^2\right)u,\ \textit{ in }\ [0,T_0)\times\mathbb{R}. \tag{e/ef}$$

which can be proved easily by some simple calculations.

[20] Cf., the entropy inequality (EIb).

The usual entropy inequality associated with (1DSW_1) and used as a selection principle for discontinuous solutions (see e.g., Godlewski and Raviart, 1996), then implies (EIb). Thus, the entropy inequality (EIb) is also proved.

In general, the derivation of entropy preserving schemes is a delicate problem. Some entropy stable schemes have been developed for the shallow-water model or related systems (see, e.g., Berthon, 2006, Bouchut, 2004, Chalons and Coulombel, 2008, Chalons et al., 2010, Bouchut, Klingenberg, and Waagan, 2007, Bouchut and Luna, 2010). In Xing, 2014, a well-balanced discontinuous Galerkin method preserving both the still water at rest and the more general moving water equilibrium. The works Xing and Shu, 2006, Xing, Shu, and Noelle, 2011, and Xing, Zhung, and Shu, 2010 demonstrates the advantage of the *full well-balanced property* over *classical well-balanced* schemes by several numerical examples. In Berthon and Chalons, 2016, the authors exhibit a Godunov-type scheme (see also Harten, Lax, and Leer, 1983) being: i) water height positivity preserving; ii) entropy preserving; iii) fully well-balanced (able to restore all the steady states (sss)).

2.2.3 Characteristic fields and Riemann invariants

We consider the following two cases depending on the positivity of h:

i) *Case where* $h = 0$: By the definition of Ω_b, $h = 0$ implies $u = 0$, and thus $a = 0$. The three eigenvalues coincides

$$\lambda_-\left(0,0,b\right) \equiv \lambda_0\left(0,0,b\right) \equiv \lambda_+\left(0,0,b\right) = 0,$$

and the corresponding eigenvectors are

$$r_-\left(0,0,b\right) \equiv r_0\left(0,0,b\right) \equiv r_+\left(0,0,b\right) = 0_{\mathbb{R}^3}.$$

Thus, it is unnecessary to consider further for this degenerate case.

ii) *Case where* $h > 0$: We claim that

$$\begin{aligned} \nabla_{U_b}\lambda_-\left(U_b\right)\cdot r_-\left(U_b\right) &= -\frac{3c}{2h} < 0, \\ \nabla_{U_b}\lambda_0\left(U_b\right)\cdot r_0\left(U_b\right) &= 0, \\ \nabla_{U_b}\lambda_+\left(U_b\right)\cdot r_+\left(U_b\right) &= +\frac{3c}{2h} > 0. \end{aligned}$$

Indeed, set $k := hu$, then $u = \frac{k}{h}$, and

$$\lambda_\pm\left(U_b\right) = \frac{k}{h} \pm \sqrt{gh}, \quad r_\pm\left(U_b\right) = \begin{pmatrix} 1 \\ \dfrac{k}{h} \pm \sqrt{gh} \\ 0 \end{pmatrix}, \quad r_0\left(U_b\right) = \begin{pmatrix} gh \\ 0 \\ \dfrac{k^2}{h^2} - gh \end{pmatrix}.$$

Hence,

$$\begin{aligned}\nabla_{U_b}\lambda_{\pm}\left(U_b\right)\cdot r_{\pm}\left(U_b\right) &= \begin{pmatrix} -\dfrac{k}{h^2} \pm \frac{1}{2}\sqrt{\dfrac{g}{h}} \\ \dfrac{1}{h} \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \dfrac{k}{h} \pm \sqrt{gh} \\ 0 \end{pmatrix} \\ &= -\frac{k}{h^2} \pm \frac{1}{2}\sqrt{\frac{g}{h}} + \frac{k}{h^2} \pm \sqrt{\frac{g}{h}} = \pm\frac{3}{2}\sqrt{\frac{g}{h}} = \pm\frac{3c}{2h}.\end{aligned}$$

As a consequence, the characteristic fields associated with λ_- and λ_+ are genuinely nonlinear, while the characteristic field associated with λ_0 is linearly degenerate (see e.g., Godlewski and Raviart, 1996 for more details).

The Riemann invariants $\left(I_-^{(l)}\right)_{l=1,2}$, $\left(I_0^{(l)}\right)_{l=1,2}$, and $\left(I_+^{(l)}\right)_{l=1,2}$ respectively associated with λ_-, λ_0, and λ_+ are given by

$$I_{\pm}^{(1)} = z,\ \ I_{\pm}^{(2)} = u \mp 2c,\ \ I_0^{(1)} = hu,\ \ I_0^{(2)} = \frac{u^2}{2} + g\left(h+b\right).$$

Remark 1.2. *The Riemann invariants associated with the stationary characteristic field are exactly the steady states* (sss)*.*

2.2.4 Entropy and admissible discontinuities

Until these discontinuous are involved, we consider the *step function*

$$U_b\left(t,x\right) := \begin{cases} U_{b,1}, & \text{if } x < \sigma t, \\ U_{b,2}, & \text{if } x > \sigma t, \end{cases} \tag{sf}$$

where $U_{b,1} = \left(h_1, h_1u_1, b_1\right)$ and $U_{b,2} = \left(h_2, h_2u_2, b_2\right)$ belong to Ω_b and $\sigma \in \mathbb{R}$ represents the speed of propagation of the discontinuity. We distinguish between the two cases $b_1 = b_2$ and $b_1 \neq b_2$.

i) *Case* $b_1 = b_2$: The system (1DSW$_1$) is made of two conservation laws since $\partial_x b = 0$:

$$\begin{cases} \partial_t h + \partial_x\left(hu\right) = 0, & \text{in } [0,T_0)\times\mathbb{R}, \\ \partial_t\left(hu\right) + \partial_x\left(hu^2 + \dfrac{g}{2}h^2\right) = 0, & \text{in } [0,T_0)\times\mathbb{R}. \end{cases}$$

The step function (sf) is then called a *shock discontinuity* and is said to be *admissible* provided that (1DSW$_2$) and (EIb) are satisfied in the distributional sense, i.e., the *Rankine-Hugoniot relations*

$$-\sigma\left(U_{b,2} - U_{b,1}\right) + F_b\left(U_{b,2}\right) - F_b\left(U_{b,1}\right) = 0, \tag{RH}$$

and the *entropy inequality*

$$-\sigma\left(\Phi_b\left(U_{b,2}\right) - \Phi_b\left(U_{b,1}\right)\right) + \left(\Psi_b\left(U_{b,2}\right) - \Psi_b\left(U_{b,1}\right)\right) \leq 0 \tag{EI}$$

hold true.

ii) *Case* $b_1 \neq b_2$: The system (1DSW$_1$) is no longer conservative and thus the classical Rankine-Hugoniot relations do not make sense anymore. Nevertheless, the rigidity of the bottom topography, i.e., $\partial_t b = 0$ implies that $\sigma = 0$ and thus $\sigma = \lambda_0$. This means that the discontinuity (sf) is associated with a linearly degenerate characteristic field (see Appendix A, or e.g., Godlewski and Raviart, 1996). Classically, the admissibility criterion is defined by the continuity of the Riemann invariants $\left(I_0^{(l)}\right)_{l=1,2}$. More explicitly, (sf) is said to be *admissible* if

$$\begin{cases} I_0^{(1)}\left(U_{b,1}\right) = I_0^{(1)}\left(U_{b,2}\right), \\ I_0^{(2)}\left(U_{b,1}\right) = I_0^{(2)}\left(U_{b,2}\right). \end{cases} \tag{Ri}$$

By definition of $I_0^{(1)}$ and $I_0^{(2)}$, the following *entropy equality* holds:

$$-\sigma\left(\Phi_b\left(U_{b,2}\right) - \Phi_b\left(U_{b,1}\right)\right) + \left(\Psi_b\left(U_{b,2}\right) - \Psi_b\left(U_{b,1}\right)\right) = \Psi_b\left(U_{b,2}\right) - \Psi_b\left(U_{b,1}\right) = 0.$$

Indeed, the first equality holds since $\sigma = 0$ and (Ri) is exactly

$$\begin{cases} h_1 u_1 = h_2 u_2, \\ \dfrac{1}{2}u_1^2 + g\left(h_1 + b_1\right) = \dfrac{1}{2}u_2^2 + g\left(h_2 + b_2\right). \end{cases} \tag{1.6}$$

Plugging (1.6) into

$$\Psi_b\left(U_{b,2}\right) - \Psi_b\left(U_{b,1}\right) = \left(\frac{1}{2}u_2^2 + gh_2 + gb_2\right)h_2u_2 - \left(\frac{1}{2}u_1^2 + gh_1 + gb_1\right)h_1u_1$$

yields

$$\begin{aligned} \Psi_b\left(U_{b,2}\right) - \Psi_b\left(U_{b,1}\right) &= \left(\frac{1}{2}u_2^2 + gh_2 + gb_2\right)h_1u_1 - \left(\frac{1}{2}u_1^2 + gh_1 + gb_1\right)h_1u_1 \\ &= h_1u_1\left[\frac{1}{2}u_2^2 + g\left(h_2 + b_2\right) - \frac{1}{2}u_1^2 + g\left(h_1 + b_1\right)\right] = 0. \end{aligned}$$

The step function (sf) is then called a *contact discontinuity*.

Remark 1.3. *Because of the resonant regime, a shock wave given by the pair of Rankine-Hugoniot relations* (RH) *and the entropy inequality* (EI) *may coincide with a stationary wave defined by the system* (Ri) *of continuity of the Riemann invariants.*

As a consequence of Remark 1.3, we have $\sigma = 0$ within $b_1 = b_2$. In this case ($\sigma = 0$ and $b_1 = b_2 =: b$), the Rankine-Hugoniot relations and the entropy inequality respectively read

$$\begin{cases} h_1u_1 = h_2u_2 =: q, \\ h_1u_1^2 + \dfrac{g}{2}h_1^2 = h_2u_2^2 + \dfrac{g}{2}h_2^2, \end{cases} \quad \text{and } q\left[\left(\frac{u_2^2}{2} + gh_2\right) - \left(\frac{u_1^2}{2} + gh_1\right)\right] \leq 0. \tag{1.7}$$

And (Ri) administers the stationary wave and reads

$$\begin{cases} h_1u_1 = h_2u_2, \\ \dfrac{u_1^2}{2} + gh_1 = \dfrac{u_2^2}{2} + gh_2. \end{cases} \tag{1.8}$$

It is pointed out immediately from the demonstration above that (1.7) and (1.8) cannot be satisfied simultaneously except for $u_1 = u_2 = 0$. In other words, the admissibility criteria suffer from a lack of continuity at point $b_1 = b_2$ for stationary discontinuities. The following definition appears to avoid such an unconsistency (see e.g., Berthon and Chalons, 2016, Section 2 for more details).

Definition 1.3 (Smooth fully steady states)**.** *The* smooth fully steady states *for the shallow water model* (1DSW) *are smooth functions $U : \mathbb{R} \to \Omega$ such that* (sss) *holds.*

This smooth version of fully steady states is compatible with smooth bottom topographies. For discontinuous bottom topography, see e.g., Castro, Pardo Milanés, and Parés, 2007, Noelle, Xing, and Shu, 2007, Russo and Khe, 2010, Russo and Khe, 2009 for more details.

Chapter 2

Explicit Cell-Centered Finite Volume Schemes for 1D Shallow Water in Dry-Wet Fluid Domains

This chapter is devoted to consider a Godunov-type schemes for 1D shallow water equations (1DSW). See Appendix B for the treatment of cell-centered Godunov-type finite volume scheme for systems (A.1) of conservation laws in one space dimension.

1 Explicit 3-point cell-centered finite volume schemes

1.1 Derivation of explicit 3-point cell-centered finite volume schemes

In this chapter, we are interested in numerical schemes for the following Cauchy problem for ($1DSW_1$)

$$\begin{cases} \partial_t U_b + \partial_x \left(F_b\left(U_b\right)\right) + S_b\left(U_b\right)\partial_x U_b = 0, & \text{in } [0,T_0)\times\mathbb{R}, \\ U_b\left(0,x\right) = U_{b,0}\left(x\right) := \left(h_0, q_0, b_0\right)\left(x\right), & \text{in } \mathbb{R}, \end{cases} \tag{Cp1DSW}$$

whose water height and velocity are stored on the same location on an admissible mesh (Cf., staggered mesh, where they are stored on different locations). The exact solution of (Cp1DSW) is denoted by $U_{\mathrm{b}} = (U, \mathrm{b})$ where we have used the notation b for the solution of (Cp1DSW) to distinguish it with the given bottom topography b.

Assume that the initial data $U_{b,0}$ belongs to $\left(L^\infty\left(\mathbb{R}\right)\right)^3$ and define its bounds as (see Assumption B.1)

$$\begin{aligned} U_{b,m} &:= \inf_{x\in\mathbb{R}} \min\left(h_0\left(x\right), q_0\left(x\right), b_0\left(x\right)\right) > -\infty, \\ U_{b,M} &:= \sup_{x\in\mathbb{R}} \max\left(h_0\left(x\right), q_0\left(x\right), b_0\left(x\right)\right) < +\infty. \end{aligned}$$

Let Δt be the time step, $T := \left\lfloor \frac{T_0}{\Delta t} \right\rfloor$, and $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ defined in Definition B.1. For convenience purpose, we denote $[T] := \{0, 1, \ldots, T\}$.

On a 1D admissible mesh $\mathcal{T}$, define

$$U_{b,i}^0 := \frac{1}{\Delta x_i}\int_{C_i} U_{b,0}\left(x\right)dx = \frac{1}{\Delta x_i}\int_{C_i} \left(h_0, q_0, b_0\right)^\top\left(x\right)dx, \ \ \forall i \in \mathbb{Z}.$$

Integrating the PDE in (Cp1DSW) over $\left(t^n, t^{n+1}\right) \times C_i$ yields

$$\int_{C_i} U_{\mathrm{b}}\left(t^{n+1}, x\right) dx - \int_{C_i} U_{\mathrm{b}}\left(t^{n}, x\right) dx + \int_{t^n}^{t^{n+1}} F_b\left(U_{\mathrm{b}}\left(t, x_{i+\frac{1}{2}}\right)\right) dt$$
$$- \int_{t^n}^{t^{n+1}} F_b\left(U_{\mathrm{b}}\left(t, x_{i-\frac{1}{2}}\right)\right) dt + \int_{t^n}^{t^{n+1}} \int_{C_i} S_b\left(U_{\mathrm{b}}\right) \partial_x U_{\mathrm{b}} dx dt = 0, \;\; \forall i \in \mathbb{Z}, \;\; \forall n \in [T-1].$$

Define the following approximation

$$\begin{cases} U_{\mathrm{b},i}^n \approx \dfrac{1}{\Delta x_i} \displaystyle\int_{C_i} U_{\mathrm{b}}\left(t^n, x\right) dx, & \forall i \in \mathbb{Z}, \;\; \forall n \in [T], \\ F_{\mathrm{b},i+\frac{1}{2}}^n \approx \dfrac{1}{\Delta t} \displaystyle\int_{t^n}^{t^{n+1}} F_b\left(U_{\mathrm{b}}\left(t, x_{i+\frac{1}{2}}\right)\right) dt, & \forall i \in \mathbb{Z}, \;\; \forall n \in [T], \\ S_{\mathrm{b},i}^n \approx \dfrac{1}{\Delta t \Delta x_i} \displaystyle\int_{t^n}^{t^{n+1}} \int_{C_i} S_b\left(U_{\mathrm{b}}\right) \partial_x U_{\mathrm{b}} dx dt, & \forall i \in \mathbb{Z}, \;\; \forall n \in [T], \end{cases}$$

then the above integral equality reads

$$\Delta x_i \left(U_{\mathrm{b},i}^{n+1} - U_{\mathrm{b},i}^n\right) + \Delta t \left(F_{\mathrm{b},i+\frac{1}{2}}^n - F_{\mathrm{b},i-\frac{1}{2}}^n\right) + \Delta t \Delta x_i S_{\mathrm{b},i}^n \approx 0, \;\; \forall i \in \mathbb{Z}, \;\; \forall n \in [T-1].$$

Let G_b be a numerical flux (which can not be "monotone" anyway!, see Appendix B) (see Definition B.4) for F_b, i.e., $F_{\mathrm{b},i+\frac{1}{2}}^n$ can be well approximated by

$$F_{\mathrm{b},i+\frac{1}{2}}^n \approx G_b\left(U_{\mathrm{b},i}^n, U_{\mathrm{b},i+1}^n\right), \;\; \forall i \in \mathbb{Z}, \;\; \forall n \in [T].$$

On an admissible mesh $\mathcal{T}$ of $\mathbb{R}$, we now approximate $\partial_x \mathrm{b}$ by, for instance, the forward difference, i.e.,

$$\partial_x \mathrm{b}\left(x\right) \approx \frac{2\left(\mathrm{b}_{i+1} - \mathrm{b}_i\right)}{\Delta x_i + \Delta x_{i+1}}, \;\; \forall x \in C_i, \;\; \forall i \in \mathbb{Z}.$$

Then the approximate source term satisfies

$$\begin{aligned} -\Delta t S_{\mathrm{b},i}^n &\approx -\frac{1}{\Delta x_i} \int_{t^n}^{t^{n+1}} \int_{C_i} \left(0, gh\partial_x \mathrm{b}, 0\right)^\top (t,x)\, dx dt \\ &\approx -\frac{2g\left(\mathrm{b}_{i+1} - \mathrm{b}_i\right)}{\Delta x_i \left(\Delta x_i + \Delta x_{i+1}\right)} \int_{t^n}^{t^{n+1}} \int_{C_i} \left(0, h, 0\right)^\top (t,x)\, dx dt \\ &\approx \left(0, \frac{2g\Delta t h_i^n \left(\mathrm{b}_i - \mathrm{b}_{i+1}\right)}{\Delta x_i + \Delta x_{i+1}}, 0\right)^\top, \;\; \forall i \in \mathbb{Z}, \;\; \forall n \in [T]. \end{aligned}$$

Set

$$S\left(U_b, \bar{U}_b\right) := \left(0, h\left(b - \bar{b}\right), 0\right)^\top, \;\; \forall U_b, \overline{U}_b \in \mathbb{R}^3,$$

its Jacobian is given by

$$J_S\left(U_b, \overline{U}_b\right) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ b - \bar{b} & 0 & h & 0 & 0 & -h \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \;\; \forall U_b, \overline{U}_b \in \mathbb{R}^3.$$

The corresponding 3-point explicit cell-centered finite volume scheme for (Cp1DSW) is given by

$$\begin{cases} U_{\mathrm{b},i}^{0} := \dfrac{1}{\Delta x_i}\displaystyle\int_{C_i} U_{b,0}\left(x\right) dx = \dfrac{1}{\Delta x_i}\displaystyle\int_{C_i} \left(h_0, q_0, b_0\right)^{\top}\left(x\right) dx, \\ U_{\mathrm{b},i}^{n+1} := U_{\mathrm{b},i}^{n} - \nu_i\left[G_b\left(U_{\mathrm{b},i}^{n}, U_{\mathrm{b},i+1}^{n}\right) - G_b\left(U_{\mathrm{b},i-1}^{n}, U_{\mathrm{b},i}^{n}\right)\right] + \dfrac{2g\Delta t}{\Delta x_i + \Delta x_{i+1}} S\left(U_{\mathrm{b},i}^{n}, U_{\mathrm{b},i+1}^{n}\right), \end{cases} \tag{3peccFVS}$$

for all $i \in \mathbb{Z}$, $n \in [T-1]$.

1.2 Properties of explicit 3-point cell-centered finite volume schemes

On an admissible mesh $\mathcal{T}$ of $\mathbb{R}$, assume that $U_{b,0}\left(x\right)$ is equal to some reference state $U_{b,\star}$ for $x \in (-\infty, M_L) \cup (M_R, \infty)$:

$$U_{b,0}\left(x\right) = U_{b,\star}, \quad \forall x \in \left(-\infty, M_L\right) \cup \left(M_R, \infty\right).$$

Denote by i_{M_L} and i_{M_R} the indices indicating which primal cells containing M_L and M_R, i.e.,

$$M_L \in \left[x_{i_{M_L}-\frac{1}{2}}, x_{i_{M_L}+\frac{1}{2}}\right), \quad M_R \in \left(x_{i_{M_R}-\frac{1}{2}}, x_{i_{M_R}+\frac{1}{2}}\right],$$

then

$$U_{b,0}|_{C_i}\left(x\right) = U_{b,\star}, \quad \forall i \text{ s.t. } i < i_{M_L} \text{ or } i > i_{M_R},$$

and consequently

$$U_{\mathrm{b},i}^{0} = U_{b,\star}, \quad \forall i \text{ s.t. } i < i_{M_L} \text{ or } i > i_{M_R}.$$

The finite volume scheme (Cp1DSW) gives us

$$\begin{aligned} U_{\mathrm{b},i}^{1} :&= U_{\mathrm{b},i}^{0} - \nu_i\left[G_b\left(U_{\mathrm{b},i}^{0}, U_{\mathrm{b},i+1}^{0}\right) - G_b\left(U_{\mathrm{b},i-1}^{0}, U_{\mathrm{b},i}^{0}\right)\right] + \frac{2g\Delta t}{\Delta x_i + \Delta x_{i+1}} S\left(U_{\mathrm{b},i}^{0}, U_{\mathrm{b},i+1}^{0}\right) \\ &= U_{b,\star} - \nu_i\left[G_b\left(U_{b,\star}, U_{b,\star}\right) - G_b\left(U_{b,\star}, U_{b,\star}\right)\right] + \frac{2g\Delta t}{\Delta x_i + \Delta x_{i+1}} S\left(U_{b,\star}, U_{b,\star}\right) \\ &= U_{b,\star} + \frac{2g\Delta t}{\Delta x_i + \Delta x_{i+1}}\left(0, h_\star\left(b_\star - b_\star\right), 0\right) \\ &= U_{b,\star}, \quad \forall i \text{ s.t. } i < i_{M_L} - 1 \text{ or } i > i_{M_R} + 1. \end{aligned}$$

By mathematical induction, we can prove that

$$U_{\mathrm{b},i}^{n} = U_{b,\star}, \quad \forall i \text{ s.t. } i < i_{M_L} - n \text{ or } i > i_{M_R} + n.$$

In particular, on a uniform mesh, we can choose

$$i_{M_L} := i_0 - \left\lfloor \frac{M_L}{\Delta x} \right\rfloor, \quad i_{M_R} := i_0 + \left\lfloor \frac{M_R}{\Delta x} \right\rfloor,$$

and thus

$$U^n_{\mathrm{b},i} = U_{b,\star}, \ \ \forall i \text{ s.t. } i < i_0 - \left\lfloor \frac{M_L}{\Delta x} \right\rfloor - n \text{ or } i > i_0 + \left\lfloor \frac{M_R}{\Delta x} \right\rfloor + n.$$

Remark 2.1. *Analogous to Remark B.1, the entropy Φ_0 can be chosen to achieve the following properties:*

$$\Phi_0 \left(U_{b,\star}\right) = 0, \ \ \nabla\Phi_0 \left(U_{b,\star}\right) = 0_{\mathbb{R}^3}.$$

Since Φ_0 is convex, it follows from the last equations that

$$\Phi_0 \left(U_b\right) > 0, \ \textit{for } U_b \neq U_{b,\star}, \ \textit{in } [0, T_0) \times \mathbb{R}.$$

Contrary to Proposition B.1, since (Cp1DSW) contains a source term (outside a conservation form), (3peccFVS) does not preserve the total momentum.

Proposition 2.1 ("Semi-conservation" on uniform meshes)**.** *Assume that the initial data satisfies $U_{b,0} \in \left(L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})\right)^3$ and $U_{b,0} \in [U_{b,m}, U_{b,M}]^3$ for some $U_{b,m}$, $U_{b,M} \in \Omega_b$. Let $\mathcal{T}$ be a uniform mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}^\star_+$ be the time step. Let $\left(U^n_{\mathrm{b},i}\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by the 3-point explicit cell-centered finite volume scheme* (3peccFVS)*, then the following identity holds*

$$\Delta x \sum_{i\in\mathbb{Z}} U^{n+1}_{\mathrm{b},i} = \Delta x \sum_{i\in\mathbb{Z}} U^n_{\mathrm{b},i} + g\Delta t \sum_{i\in\mathbb{Z}} S\left(U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right), \ \ \forall n \in [T],$$

i.e., the total mass and the total bottom topography[1] *are preserved but the total momentum is not:*

$$\begin{cases} \Delta x \displaystyle\sum_{i\in\mathbb{Z}} h^{n+1}_i = \Delta x \sum_{i\in\mathbb{Z}} h^n_i, & \forall n \in [T], \\ \Delta x \displaystyle\sum_{i\in\mathbb{Z}} h^{n+1}_i u^{n+1}_i = \Delta x \sum_{i\in\mathbb{Z}} h^n_i u^n_i + g\Delta t \sum_{i\in\mathbb{Z}} h^n_i \left(\mathrm{b}_i - \mathrm{b}_{i+1}\right), & \forall n \in [T]. \end{cases}$$

Consequently,

$$\begin{cases} \Delta x \displaystyle\sum_{i\in\mathbb{Z}} h^n_i = \Delta x \sum_{i\in\mathbb{Z}} h^0_i, & \forall n \in [T], \\ \Delta x \displaystyle\sum_{i\in\mathbb{Z}} h^n_i u^n_i = \Delta x \sum_{i\in\mathbb{Z}} h^0_i u^0_i + g\Delta t \sum_{i\in\mathbb{Z}} \left(\left(\mathrm{b}_i - \mathrm{b}_{i+1}\right) \sum_{j=0}^{n-1} h^j_i \right), & \forall n \in [T]. \end{cases}$$

This semi-conservation also holds on an arbitrary admissible mesh.

Proposition 2.2 ("Semi-conservation" on admissible meshes)**.** *Assume that the initial data satisfies $U_{b,0} \in \left(L^1 \cap L^\infty(\mathbb{R})\right)^3$ and $U_{b,0} \in [U_{b,m}, U_{b,M}]^3$ for some $U_{b,m}$, $U_{b,M} \in \Omega_b$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}^\star_+$ be the time step. Let $\left(U^n_{\mathrm{b},i}\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by*

[1] This is obvious since the bottom topography is assumed to be time-independent.

the 3-point explicit cell-centered finite volume scheme (3peccFVS), *then the following identity holds*

$$\sum_{i\in\mathbb{Z}} \Delta x_i U_{\mathrm{b},i}^{n+1} = \sum_{i\in\mathbb{Z}} \Delta x_i U_{\mathrm{b},i}^{n} + 2g\Delta t \sum_{i\in\mathbb{Z}} \frac{\Delta x_i}{\Delta x_i + \Delta x_{i+1}} S\left(U_{\mathrm{b},i}^{n}, U_{\mathrm{b},i+1}^{n}\right), \ \forall n \in [T-1],$$

i.e., the "weighted" total mass and "weighted" total bottom topography are preserved but the "weighted" total momentum is not:

$$\begin{cases} \displaystyle\sum_{i\in\mathbb{Z}} \Delta x_i h_i^{n+1} = \sum_{i\in\mathbb{Z}} \Delta x_i h_i^{n}, & \forall n \in [T-1], \\ \displaystyle\sum_{i\in\mathbb{Z}} \Delta x_i h_i^{n+1} u_i^{n+1} = \sum_{i\in\mathbb{Z}} \Delta x_i h_i^{n} u_i^{n} + 2g\Delta t \sum_{i\in\mathbb{Z}} \frac{\Delta x_i}{\Delta x_i + \Delta x_{i+1}} h_i^n \left(\mathrm{b}_i - \mathrm{b}_{i+1}\right), & \forall n \in [T-1]. \end{cases}$$

Consequently,

$$\begin{cases} \displaystyle\sum_{i\in\mathbb{Z}} \Delta x_i h_i^{n} = \sum_{i\in\mathbb{Z}} \Delta x_i h_i^{0}, & \forall n \in [T], \\ \displaystyle\sum_{i\in\mathbb{Z}} \Delta x_i h_i^{n} u_i^{n} = \sum_{i\in\mathbb{Z}} \Delta x_i h_i^{0} u_i^{0} + 2g\Delta t \sum_{i\in\mathbb{Z}} \left(\frac{\Delta x_i \left(\mathrm{b}_i - \mathrm{b}_{i+1}\right)}{\Delta x_i + \Delta x_{i+1}} \sum_{j=0}^{n-1} h_i^n \right), & \forall n \in [T]. \end{cases}$$

The approximate finite volume solution $U_{\mathrm{b},\mathcal{T},\Delta t}$ on $\mathcal{T}$ is defined a.e. on $[0, T_0) \times \mathbb{R}$ piecewisely by

$$U_{\mathrm{b},\mathcal{T},\Delta t}(t,x) := \sum_{n\in[T]} \sum_{i\in\mathbb{Z}} U_{\mathrm{b},i}^{n} \mathbf{1}_{[t^n,t^{n+1})\times\left[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}\right)}(t,x), \ \text{a.e. in } [0,T_0)\times\mathbb{R}. \quad \text{(FVS)}$$

Unfortunately, there is no notion of monotonicity for the case of systems of conservation laws (see Appendix B) and thus the well-known L^∞-estimate for scalar conservation laws failed to extend to shallow water systems. We now turn our attention to another resolution.

2 A Godunov-type scheme for 1D shallow water equations

The solution of the following Riemann problem for 1D shallow water ($1\mathrm{DSW}_2$):

$$\begin{cases} \partial_t U_b + \partial_x \left(F_b\left(U_b\right)\right) + S_b\left(U_b\right)\partial_x U_b = 0, & \text{in } [0,T_0)\times\mathbb{R}, \\ U_b(0,x) = U_{b,0}\left(x; U_{b,L}, U_{b,R}\right) := \begin{cases} U_{b,L}, & \text{if } x<0, \\ U_{b,R}, & \text{if } x \ge 0, \end{cases} & \text{in } \mathbb{R}, \end{cases} \quad \text{(Rp1DSW)}$$

depends only on the states $U_{b,L} := (h_L, h_L u_L, b_L)$, $U_{b,R} := (h_R, h_R u_R, b_R) \in \Omega_b$ and the ratio $\frac{x}{t}$ and can be thus denoted by $U_{\mathrm{b},\mathcal{R}}\left(x/t; U_{b,L}, U_{b,R}\right)$.

Due to the propagation with finite velocity of signals,

$$U_{\mathrm{b},\mathcal{R}}\left(\frac{x}{t}; U_{b,L}, U_{b,R}\right) = \begin{cases} U_{b,L}, & \text{if } \frac{x}{t} < \lambda_{\min}\left(U_{\mathrm{b}}\right), \\ U_{b,R}, & \text{if } \frac{x}{t} > \lambda_{\max}\left(U_{\mathrm{b}}\right), \end{cases} \quad \forall (t,x) \in [0,T_0)\times\mathbb{R}, \tag{2.1}$$

where

$$\lambda_{\min}(U_{\mathrm{b}}) := (\lambda_{-}(U_{\mathrm{b}}))_{-} = (u-c)_{-} = \left(u-\sqrt{gh}\right)_{-},$$
$$\lambda_{\max}(U_{\mathrm{b}}) := (\lambda_{+}(U_{\mathrm{b}}))_{+} = (u+c)_{+} = \left(u+\sqrt{gh}\right)_{+},$$

are the smallest and largest signal velocity, respectively.

2.1 Derivation of a Godunov-type scheme for 1D shallow water equations

To extend the Godunov-type scheme for systems of conservation laws given in Section 3, Appendix B for (1DSW$_2$), given an admissible mesh $\mathcal{T}$ of $\mathbb{R}$, we consider the numerical approximation $U_{\mathrm{b},\mathcal{T},\Delta t}(t^n,x)$ of the discrete time levels t^n, for all $n \in [n_0]$ for some $n_0 \in [T-1]$, to be a piecewise constant function in x, i.e.,

$$U_{\mathrm{b},\mathcal{T},\Delta t}(t^n,x) := \sum_{i\in\mathbb{Z}} U_i^n \mathbf{1}_{\left[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}\right)}(x), \quad \forall x \in \mathbb{R}.$$

To calculate the numerical approximation at the next time level $t^{n+1} = t^n + \Delta t$, we follow the following two steps.

Step 1. We solve exactly the initial-value problem

$$\begin{cases} \partial_t U_b + \partial_x (F_b(U_b)) + S_b(U_b)\,\partial_x U_b = 0, & \text{in } \left[t^n,t^{n+1}\right] \times \mathbb{R}, \\ U_b(t^n,x) = U_{\mathrm{b},\mathcal{T},\Delta t}(t^n,x), & \text{in } \mathbb{R}, \end{cases} \tag{2.2}$$

for $t \in [t^n, t^n+\Delta t] \equiv \left[t^n,t^{n+1}\right]$, and denote its solution by $U^n_{\mathrm{b},\mathcal{T},\Delta t}(t,x)$. Each of discontinuities of $U_{\mathrm{b},\mathcal{T},\Delta t}(t^n,x)$ comprises locally a Riemann's problem. If the following CFL condition holds

$$\Delta t < \frac{\inf_{i\in\mathbb{Z}}\Delta x_i}{2 \sup\limits_{[0,T_0)\times\mathbb{R}} \max\left(\left|u-\sqrt{gh}\right|, \left|u+\sqrt{gh}\right|\right)},$$

or stronger

$$\nu < \frac{\alpha_{\mathcal{T}}}{2 \sup\limits_{[0,T_0)\times\mathbb{R}} \max\left(\left|u-\sqrt{gh}\right|, \left|u+\sqrt{gh}\right|\right)},$$

where $\max\left(\left|u-\sqrt{gh}\right|, \left|u+\sqrt{gh}\right|\right)$ indicates the largest signal speed at the time-space point (t,x). Due to (2.1), there is no interaction between neighboring Riemann problems, and $U^n_{\mathrm{b},\mathcal{T},\Delta t}(t,x)$ can be expressed exactly in terms of the solutions of local Riemann's problems:

$$U^n_{\mathrm{b},\mathcal{T},\Delta t}(t,x) = \sum_{i\in\mathbb{Z}} U_{\mathrm{b},\mathcal{R}}\left(\frac{x-x_{i+\frac{1}{2}}}{t-t^n}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) \mathbf{1}_{(x_i,x_{i+1}]}(x), \text{ in } \left[t^n,t^{n+1}\right] \times \mathbb{R}. \tag{2.3}$$

We then obtain a piecewise constant approximation $U_{\mathrm{b},\mathcal{T},\Delta t}\left(t^{n+1},x\right)$ by averaging the quantity $U^n_{\mathrm{b},\mathcal{T},\Delta t}\left(t^{n+1},x\right)$ over each cells, i.e.,

$$U^{n+1}_{\mathrm{b},i} := \frac{1}{\Delta x_i}\int_{C_i} U^n_{\mathrm{b},\mathcal{T},\Delta t}\left(t^{n+1},x\right)dx, \quad \forall i \in \mathbb{Z}. \tag{2.4}$$

More explicitly, (2.4) can be rewritten in terms of the solutions of the local Riemann's problem as

$$U^{n+1}_{\mathrm{b},i} = \frac{1}{\Delta x_i}\int_0^{\Delta x_i/2} U_{\mathrm{b},\mathcal{R}}\left(\frac{x}{\Delta t}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)dx + \frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^{0} U_{\mathrm{b},\mathcal{R}}\left(\frac{x}{\Delta t}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)dx. \tag{2.5}$$

Next, recall U_b is an exact solution of (Rp1DSW), integrating the first equation in (Rp1DSW) over an arbitrary rectangle $(t_1,t_2)\times(a,b)$ yields the following identity

$$\begin{aligned}\int_a^b U_\mathrm{b}(t_2,x)\,dx - \int_a^b U_\mathrm{b}(t_1,x)\,dx + \int_{t_1}^{t_2} F_b\left(U_\mathrm{b}(t,b)\right)dt \\ - \int_{t_1}^{t_2} F_b\left(U_\mathrm{b}(t,a)\right)dt + \int_{t_1}^{t_2}\int_a^b S_b\left(U_\mathrm{b}\right)\partial_x U_\mathrm{b}\,dxdt = 0,\end{aligned}$$

for all $(t_1,t_2,a,b)\in\mathbb{R}^2_+\times\mathbb{R}^2$ with $t_1\le t_2$, $a\le b$.

Similarly, since $U^n_{\mathrm{b},\mathcal{T},\Delta t}$ is the exact solution of (2.2), the following equality holds:

$$\begin{aligned}&\int_{C_i} U^n_{\mathrm{b},\mathcal{T},\Delta t}\left(t^{n+1},x\right)dx - \int_{C_i} U^n_{\mathrm{b},\mathcal{T},\Delta t}\left(t^{n},x\right)dx \\ &\quad + \int_{t^n}^{t^{n+1}} F_b\left(U^n_{\mathrm{b},\mathcal{T},\Delta t}\left(t,x_{i+\frac{1}{2}}\right)\right)dt - \int_{t^n}^{t^{n+1}} F_b\left(U^n_{\mathrm{b},\mathcal{T},\Delta t}\left(t,x_{i-\frac{1}{2}}\right)\right)dt \\ &\quad + \int_{t^n}^{t^{n+1}}\int_{C_i} S_b\left(U^n_{\mathrm{b},\mathcal{T},\Delta t}\right)\partial_x U^n_{\mathrm{b},\mathcal{T},\Delta t}\,dxdt = 0, \quad \forall i\in\mathbb{Z}.\end{aligned}$$

Plugging (2.3) into the last equality yields

$$\begin{aligned}&\int_{x_{i-1/2}}^{x_i} U_{\mathrm{b},\mathcal{R}}\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta t}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)dx + \int_{x_i}^{x_{i+1/2}} U_{\mathrm{b},\mathcal{R}}\left(\frac{x-x_{i+\frac{1}{2}}}{\Delta t}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)dx \\ &\quad - \int_{x_{i-1/2}}^{x_i} U_{b,0}\left(x-x_{i-\frac{1}{2}}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)dx - \int_{x_i}^{x_{i+1/2}} U_{b,0}\left(x-x_{i+\frac{1}{2}}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)dx \\ &\quad + \int_{t^n}^{t^{n+1}} F_b\left(U_{\mathrm{b},\mathcal{R}}\left(0; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)\right)dt - \int_{t^n}^{t^{n+1}} F_b\left(U_{\mathrm{b},\mathcal{R}}\left(0; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)\right)dt \\ &\quad + \int_{t^n}^{t^{n+1}}\int_{x_{i-1/2}}^{x_{i+1/2}} S_b\left(U^n_{\mathrm{b},\mathcal{T},\Delta t}\right)\partial_x U^n_{\mathrm{b},\mathcal{T},\Delta t}\,dxdt = 0, \quad \forall i\in\mathbb{Z},\end{aligned}$$

where the last term involving source term can be handled as

$$\begin{aligned}&\int_{t^n}^{t^{n+1}}\int_{C_i} S_b\left(U^n_{\mathrm{b},\mathcal{T},\Delta t}\right)\partial_x U^n_{\mathrm{b},\mathcal{T},\Delta t}\,dxdt \\ &= \int_{t^n}^{t^{n+1}}\int_{C_i}\left(0, gh^n_{\mathcal{T},\Delta t}(t,x)\,\partial_x\left(\mathrm{b}^n_{\mathcal{T},\Delta t}(t,x)\right), 0\right)^\top dxdt\end{aligned}$$

$$
\begin{aligned}
&= \int_{t^n}^{t^{n+1}} \int_{x_{i-1/2}}^{x_i} \left(0, gh_{\mathcal{R}}\left(\frac{x-x_{i-\frac{1}{2}}}{t-t^n}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right) \partial_x \left(\mathrm{b}_{\mathcal{R}}\left(\frac{x-x_{i-\frac{1}{2}}}{t-t^n}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)\right), 0\right)^{\top} dxdt \\
&\quad + \int_{t^n}^{t^{n+1}} \int_{x_i}^{x_{i+1/2}} \left(0, gh_{\mathcal{R}}\left(\frac{x-x_{i+\frac{1}{2}}}{t-t^n}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) \partial_x \left(\mathrm{b}_{\mathcal{R}}\left(\frac{x-x_{i+\frac{1}{2}}}{t-t^n}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)\right), 0\right)^{\top} dxdt \\
&= \int_{t^n}^{t^{n+1}} \int_{x_{i-1/2}}^{x_i} \left(0, gh_{\mathcal{R}}\left(\frac{x-x_{i-\frac{1}{2}}}{t-t^n}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right) \partial_x \mathrm{b}\left(t-t^n, x-x_{i-\frac{1}{2}}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right), 0\right)^{\top} dxdt \\
&\quad + \int_{t^n}^{t^{n+1}} \int_{x_i}^{x_{i+1/2}} \left(0, gh_{\mathcal{R}}\left(\frac{x-x_{i+\frac{1}{2}}}{t-t^n}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) \partial_x \mathrm{b}\left(t-t^n, x-x_{i+\frac{1}{2}}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right), 0\right)^{\top} dxdt \\
&= \int_{0}^{\Delta t} \int_{0}^{\Delta x_i/2} \left(0, gh_{\mathcal{R}}\left(\frac{x}{t}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right) \partial_x \mathrm{b}\left(t, x; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right), 0\right)^{\top} dxdt \\
&\quad + \int_{0}^{\Delta t} \int_{-\Delta x_i/2}^{0} \left(0, gh_{\mathcal{R}}\left(\frac{x}{t}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) \partial_x \mathrm{b}\left(t, x; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right), 0\right)^{\top} dxdt \\
&= \int_{0}^{\Delta t} \int_{0}^{\Delta x_i/2} \left(0, gh_{\mathcal{R}}\left(\frac{x}{t}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right) \partial_x b_0\left(x; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right), 0\right)^{\top} dxdt \\
&\quad + \int_{0}^{\Delta t} \int_{-\Delta x_i/2}^{0} \left(0, gh_{\mathcal{R}}\left(\frac{x}{t}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) \partial_x b_0\left(x; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right), 0\right)^{\top} dxdt \\
&= \int_{0}^{\Delta t} \int_{0}^{\Delta x_i/2} \left(0, gh_{\mathcal{R}}\left(\frac{x}{t}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right) \partial_x \mathrm{b}^0_i, 0\right)^{\top} dxdt \\
&\quad + \int_{0}^{\Delta t} \int_{-\Delta x_i/2}^{0} \left(0, gh_{\mathcal{R}}\left(\frac{x}{t}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) \partial_x \mathrm{b}^0_i, 0\right)^{\top} dxdt = 0, \quad \forall i \in \mathbb{Z}.
\end{aligned}
$$

Combining the last two identities yields for all $i \in \mathbb{Z}$:

$$
\begin{aligned}
&\int_{x_{i-1/2}}^{x_i} U_{\mathrm{b},\mathcal{R}}\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta t}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right) dx + \int_{x_i}^{x_{i+1/2}} U_{\mathrm{b},\mathcal{R}}\left(\frac{x-x_{i+\frac{1}{2}}}{\Delta t}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) dx \\
&\qquad - \int_{x_{i-1/2}}^{x_i} U_{b,0}\left(x-x_{i-\frac{1}{2}}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right) dx - \int_{x_i}^{x_{i+1/2}} U_{b,0}\left(x-x_{i+\frac{1}{2}}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) dx \\
&\qquad + \int_{t^n}^{t^{n+1}} F_b\left(U_{\mathrm{b},\mathcal{R}}\left(0; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)\right) dt - \int_{t^n}^{t^{n+1}} F_b\left(U_{\mathrm{b},\mathcal{R}}\left(0; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)\right) dt = 0,
\end{aligned}
$$

for all $i \in \mathbb{Z}$. Due to (2.4), the first two terms in the last approximation are exactly $\Delta x_i U^{n+1}_{\mathrm{b},i}$. Thus, the last identity can be rewritten as

$$
\Delta x_i U^{n+1}_{\mathrm{b},i} = \Delta x_i U^n_{\mathrm{b},i} - \Delta t\left[F_b\left(U_{\mathrm{b},\mathcal{R}}\left(0; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)\right) - F_b\left(U_{\mathrm{b},\mathcal{R}}\left(0; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)\right)\right],
$$

which is equivalent to

$$
U^{n+1}_{\mathrm{b},i} = U^n_{\mathrm{b},i} - \nu_i\left(F_b\left(\widetilde{U}^n_{\mathrm{b},i+\frac{1}{2}}\right) - F_b\left(\widetilde{U}^n_{\mathrm{b},i-\frac{1}{2}}\right)\right), \quad \forall i \in \mathbb{Z}, \tag{GtS}
$$

where

$$
\widetilde{U}^n_{\mathrm{b},i+\frac{1}{2}} := U_{\mathrm{b},\mathcal{R}}\left(0; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right), \quad \forall i \in \mathbb{Z}.
$$

More explicitly, (GtS) can be written as

$$\begin{cases} h_i^{n+1} := h_i^n - \nu_i \left[\left(h_{\mathcal{R}} u_{\mathcal{R}} \right) \left(0; U_{\mathrm{b},i}^n, U_{\mathrm{b},i+1}^n \right) - \left(h_{\mathcal{R}} u_{\mathcal{R}} \right) \left(0; U_{\mathrm{b},i-1}^n, U_{\mathrm{b},i}^n \right) \right], \\ h_i^{n+1} u_i^{n+1} := h_i^n u_i^n - \nu_i \left(h_{\mathcal{R}} u_{\mathcal{R}} + \frac{g}{2} h_{\mathcal{R}}^2 \right) \left(0; U_{\mathrm{b},i}^n, U_{\mathrm{b},i+1}^n \right) \\ \qquad + \nu_i \left(h_{\mathcal{R}} u_{\mathcal{R}} + \frac{g}{2} h_{\mathcal{R}}^2 \right) \left(0; U_{\mathrm{b},i-1}^n, U_{\mathrm{b},i}^n \right) \\ \mathrm{b}_i^{n+1} := \mathrm{b}_i^n = \mathrm{b}_i^0 := \frac{1}{\Delta x_i} \int_{C_i} b_0(x)\, dx, \end{cases} \tag{GtS$_1$}$$

for all $i \in \mathbb{Z}$, $n \in [T-1]$.

2.2 Properties of the Godunov-type scheme (GtS) for 1D shallow water equations

In spite of the terms involving bottom topography, (GtS) can be written in a conservation form, with

$$G_b(U_b, V_b) := F_b\left(U_{\mathrm{b},\mathcal{R}}(0; U_b, V_b) \right), \ \ \forall U_b, V_b \in \Omega_b.$$

The crucial reason for this is the fact that the above integral term involving bottom topography is 0. This may cause loss of serious information on the momentum conservation equation, which can lead to bad approximations for (1DSW$_2$). We will consider another approach later.

The exact solution $U_{\mathrm{b},\mathcal{T},\Delta t}^n(t,x)$ of the Riemann's problem satisfies the entropy condition (EIb). Thus, integrating the entropy inequality

$$\partial_t \left(\Phi_b \left(U_{\mathrm{b},\mathcal{T},\Delta t}^n \right) \right) + \partial_x \left(\Psi_b \left(U_{\mathrm{b},\mathcal{T},\Delta t}^n \right) \right) \le 0, \text{ a.e. in } \left[t^n, t^{n+1} \right] \times \mathbb{R},$$

over $\left[t^n, t^{n+1} \right] \times C_i$ yields

$$\int_{C_i} \Phi_b \left(U_{\mathrm{b},\mathcal{T},\Delta t}^n \left(t^{n+1}, x \right) \right) dx \le \Delta x_i \Phi_b \left(U_{\mathrm{b},i}^n \right) - \Delta t \Psi_b \left(\widetilde{U}_{\mathrm{b},i+\frac{1}{2}}^n \right) + \Delta t \Psi_b \left(\widetilde{U}_{\mathrm{b},i-\frac{1}{2}}^n \right), \ \ \forall i \in \mathbb{Z},$$

where we have used

$$U_{\mathrm{b},\mathcal{T},\Delta t}^n \left(t^n, x \right)\big|_{C_i} = U_{\mathrm{b},\mathcal{T},\Delta t} \left(t^n, x \right)\big|_{C_i} = U_{\mathrm{b},i}^n, \ \ \forall i \in \mathbb{Z}.$$

Unfortunately, Φ_b is not convex since its Hessian matrix

$$\operatorname{Hess}(\Phi_b)(U_b) = \begin{pmatrix} \frac{(hu)^2}{h^3} + g & -\frac{hu}{h^2} & g \\ -\frac{hu}{h^2} & \frac{1}{h} & 0 \\ g & 0 & 0 \end{pmatrix}$$

is not positive semi-definite for all $U_b \in \Omega_b$:

$$\det(\operatorname{Hess}(\Phi_b))(U_b) = -\frac{g^2}{h} < 0, \ \ \forall U_b \in \Omega_b.$$

Thus, we can not use Jensen's inequality for Φ_b. However, this is not actually a problem, we turn our attention to another version of the entropy inequality for (1DSW) mentioned in Remark 1.1, i.e.,

$$\partial_t \left(\Phi\left(U\right)\right) + \partial_x \left(\Psi\left(U\right)\right) \le -ghu\partial_x b, \text{ a.e. } \mathbb{R}_+ \times \mathbb{R}. \tag{EI$_1$}$$

Since the bottom topography is independent in time, the Riemann problem (Rp1DSW) can be separated into the following Riemann sub-problems:

$$\begin{cases} \partial_t U + \partial_x \left(F\left(U\right)\right) = S\left(U, \mathrm{b}\right), \text{ in } [0, T_0) \times \mathbb{R}, \\ U\left(0, x\right) = U_0\left(x; U_L, U_R\right) := \begin{cases} U_L := \left(h_L, h_L u_L\right), & \text{if } x < 0, \\ U_R := \left(h_R, h_R u_R\right), & \text{if } x \ge 0, \end{cases} \text{ in } \mathbb{R}, \end{cases} \tag{Rp1DU}$$

where b is the solution of the following Riemann problem

$$\begin{cases} \partial_t b = 0, \text{ in } [0, T_0) \times \mathbb{R}, \\ b\left(0, x\right) = b_0\left(x; b_L, b_R\right) := \begin{cases} b_L, & \text{if } x < 0, \\ b_R, & \text{if } x \ge 0, \end{cases} \text{ in } \mathbb{R}, \end{cases} \tag{Rp1Db}$$

which gives immediately that $\mathrm{b} \equiv b_0$.

Denote $U^n_{\mathrm{b},\mathcal{T},\Delta t} = \left(U^n_{\mathcal{T},\Delta t}, b_{\mathcal{T}}\right)^\top$, (2.3) reads

$$\begin{cases} U^n_{\mathcal{T},\Delta t}\left(t, x\right) = \displaystyle\sum_{i\in\mathbb{Z}} U_{\mathcal{R}}\left(\frac{x - x_{i+\frac{1}{2}}}{t - t^n}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) \mathbf{1}_{(x_i, x_{i+1}]}\left(x\right), \\ \quad b_{\mathcal{T}}\left(t, x\right) = \displaystyle\sum_{i\in\mathbb{Z}} b_0\left(x - x_{i+\frac{1}{2}}; \mathrm{b}_i, \mathrm{b}_{i+1}\right) \mathbf{1}_{(x_i, x_{i+1}]}\left(x\right), \end{cases} \text{in } \left[t^n, t^{n+1}\right] \times \mathbb{R},$$

Integrating the entropy inequality (EI$_1$) for $U^n_{\mathcal{T},\Delta t}$ and $b_{\mathcal{T}}$ over the rectangle $\left[t^n, t^{n+1}\right] \times C_i$ yields

$$\begin{aligned} \int_{C_i} \Phi\left(U^n_{\mathcal{T},\Delta t}\left(t^{n+1}, x\right)\right) dx \le \Delta x_i \Phi\left(U^n_i\right) - \Delta t \Psi\left(\widetilde{U}^n_{i+\frac{1}{2}}\right) + \Delta t \Psi\left(\widetilde{U}^n_{i-\frac{1}{2}}\right) \\ - \int_{t^n}^{t^{n+1}} \int_{C_i} g\left(hu\right)^n_{\mathcal{T},\Delta t} \partial_x b_{\mathcal{T}} dx dt, \quad \forall i \in \mathbb{Z}. \end{aligned} \tag{2.6}$$

The last term involving bottom topography is handled by integrating by parts as

$$\begin{aligned} & \int_{t^n}^{t^{n+1}} \int_{C_i} g\left(hu\right)^n_{\mathcal{T},\Delta t} \partial_x b_{\mathcal{T}} dx dt \\ = {} & g \int_{t^n}^{t^{n+1}} \left(hu\right)^n_{\mathcal{T},\Delta t} \left(t, x_{i+\frac{1}{2}}\right) b_{\mathcal{T}}\left(x_{i+\frac{1}{2}}\right) dt - g \int_{t^n}^{t^{n+1}} \left(hu\right)^n_{\mathcal{T},\Delta t} \left(t, x_{i-\frac{1}{2}}\right) b_{\mathcal{T}}\left(x_{i-\frac{1}{2}}\right) dt \\ & - g \int_{t^n}^{t^{n+1}} \int_{C_i} \partial_x \left(\left(hu\right)^n_{\mathcal{T},\Delta t}\left(t, x\right)\right) b_{\mathcal{T}}\left(x\right) dx dt \\ = {} & g b_0\left(0; \mathrm{b}_i, \mathrm{b}_{i+1}\right) \int_{t^n}^{t^{n+1}} \left(h_{\mathcal{R}} u_{\mathcal{R}}\right)\left(0; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) dt \end{aligned}$$

$$
\begin{aligned}
&- g b_0\left(0;\mathrm{b}_{i-1},\mathrm{b}_i\right)\int_{t^n}^{t^{n+1}}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)dt\\
&- g\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_i}\partial_x\left(\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(\frac{x-x_{i-\frac{1}{2}}}{t-t^n};U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)\right)b_0\left(x-x_{i-\frac{1}{2}};\mathrm{b}_{i-1},\mathrm{b}_i\right)dxdt\\
&- g\int_{t^n}^{t^{n+1}}\int_{x_i}^{x_{i+\frac{1}{2}}}\partial_x\left(\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(\frac{x-x_{i+\frac{1}{2}}}{t-t^n};U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)\right)b_0\left(x-x_{i+\frac{1}{2}};\mathrm{b}_i,\mathrm{b}_{i+1}\right)dxdt\\
={}& g\Delta t\mathrm{b}_{i+1}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)-g\Delta t\mathrm{b}_i\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)\\
&- g\mathrm{b}_i\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_i}\partial_x\left(\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(\frac{x-x_{i-\frac{1}{2}}}{t-t^n};U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)\right)dxdt\\
&- g\mathrm{b}_i\int_{t^n}^{t^{n+1}}\int_{x_i}^{x_{i+\frac{1}{2}}}\partial_x\left(\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(\frac{x-x_{i+\frac{1}{2}}}{t-t^n};U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)\right)dxdt\\
={}& g\Delta t\left[\mathrm{b}_{i+1}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)-\mathrm{b}_i\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)\right]\\
&- g\mathrm{b}_i\int_{t^n}^{t^{n+1}}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(\frac{x_i-x_{i-\frac{1}{2}}}{t-t^n};U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)dt+g\mathrm{b}_i\int_{t^n}^{t^{n+1}}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)dt\\
&- g\mathrm{b}_i\int_{t^n}^{t^{n+1}}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)dt+g\mathrm{b}_i\int_{t^n}^{t^{n+1}}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(\frac{x_i-x_{i+\frac{1}{2}}}{t-t^n};U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)dt\\
={}& g\Delta t\left[\mathrm{b}_{i+1}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)-\mathrm{b}_i\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)\right]\\
&+ g\Delta t\left[\mathrm{b}_i\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)-\mathrm{b}_i\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)\right]\\
&+ g\mathrm{b}_i\int_{t^n}^{t^{n+1}}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(-\frac{\Delta x_i}{2\left(t-t^n\right)};U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)dt\\
&- g\mathrm{b}_i\int_{t^n}^{t^{n+1}}\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(\frac{\Delta x_i}{2\left(t-t^n\right)};U^n_{\mathrm{b},i-1},U^n_{\mathrm{b},i}\right)dt\\
={}& g\Delta t\left(\mathrm{b}_{i+1}-\mathrm{b}_i\right)\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right)+g\Delta t\mathrm{b}_ih^n_iu^n_i-g\Delta t\mathrm{b}_ih^n_iu^n_i\\
={}& g\Delta t\left(\mathrm{b}_{i+1}-\mathrm{b}_i\right)\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right),\quad\forall n\in[T-1],\quad\forall i\in\mathbb{Z}.
\end{aligned}
$$

Then (2.6) becomes

$$
\begin{aligned}
\int_{C_i}\Phi\left(U^n_{\mathcal{T},\Delta t}\left(t^{n+1},x\right)\right)dx\leq\Delta x_i\Phi\left(U^n_i\right)-\Delta t\left(\Psi\left(\widetilde{U}^n_{i+\frac{1}{2}}\right)-\Psi\left(\widetilde{U}^n_{i-\frac{1}{2}}\right)\right)&\\
-g\Delta t\left(\mathrm{b}_{i+1}-\mathrm{b}_i\right)\widetilde{hu}^n_{i+\frac{1}{2}},\quad\forall i\in\mathbb{Z},&\qquad(2.7)
\end{aligned}
$$

where

$$
\widetilde{hu}^n_{i+\frac{1}{2}}:=\left(h_{\mathcal{R}}u_{\mathcal{R}}\right)\left(0;U^n_{\mathrm{b},i},U^n_{\mathrm{b},i+1}\right),\quad\forall i\in\mathbb{Z}.\qquad(2.8)
$$

The entropy Φ is a convex function since its Hessian matrix

$$
\mathrm{Hess}\left(\Phi\right)\left(U\right)=\begin{pmatrix}\dfrac{(hu)^2}{h^3}+g & -\dfrac{hu}{h^2}\\ -\dfrac{hu}{h^2} & \dfrac{1}{h}\end{pmatrix}.
$$

Applying Jensen's inequality for Φ yields for all function $V \in L^{\infty}\left(\mathbb{R}_{+} \times \mathbb{R}, \mathbb{R}^{2}\right)$:

$$\Phi\left(\frac{1}{\Delta x_i} \int_{C_i} V(t, x)\, dx\right) \leq \frac{1}{\Delta x_i} \int_{C_i} \Phi(V(t, x))\, dx, \quad \forall i \in \mathbb{Z}.$$

Combining (2.7) with the last inequality, with the help of (2.3), yields

$$\Phi\left(U_i^{n+1}\right) \leq \Phi\left(U_i^n\right) - \nu_i\left(\Psi\left(\widetilde{U}_{i+\frac{1}{2}}^n\right) - \Psi\left(\widetilde{U}_{i-\frac{1}{2}}^n\right)\right) - g\nu_i\left(\mathrm{b}_{i+1} - \mathrm{b}_i\right) \widetilde{hu}_{i+\frac{1}{2}}^n, \quad \forall i \in \mathbb{Z}. \tag{DEI}$$

Compared to Definition B.2, the proposed Godunov-type scheme is also said to be *consistent* with the entropy inequality (EI_1). And the corresponding *discrete entropy inequality* (DEI) can be rewritten more condensed as

$$\Phi_i^{n+1} \leq \Phi_i^n - \nu_i\left(\Gamma_{i+\frac{1}{2}}^n - \Gamma_{i-\frac{1}{2}}^n\right) - g\nu_i\left(\mathrm{b}_{i+1} - \mathrm{b}_i\right) \widetilde{hu}_{i+\frac{1}{2}}^n, \quad \forall i \in \mathbb{Z}, \tag{cDEI}$$

where

$$\Phi_i^n := \Phi\left(U_i^n\right), \quad \Gamma_{i+\frac{1}{2}}^n := \Psi\left(\widetilde{U}_{i+\frac{1}{2}}^n\right) = \Psi\left(U_{\mathcal{R}}\left(0; U_{\mathrm{b},i}^n, U_{\mathrm{b},i+1}^n\right)\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T].$$

The corresponding numerical entropy flux is also *consistent with the entropy flux* Ψ in the sense of Definition B.2:

$$\Psi\left(U_{\mathcal{R}}\left(0; U_{\mathrm{b},i}^n, U_{\mathrm{b},i}^n\right)\right) = \Psi\left(U_i^n\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T].$$

Now, on uniform meshes instead of general admissible ones, summing the discrete entropy inequality over all $i \in \mathbb{Z}$ yields the following result.

Proposition 2.3 (Total entropy)**.** *Assume that the initial data* $U_{b,0}$ *satisfies that* $U_{b,0} \in \left(L^1(\mathbb{R}) \cap L^{\infty}(\mathbb{R})\right)^3$ *and* $U_{b,0} \in \left[U_{b,m}, U_{b,M}\right]^3$ *for some* $U_{b,m}$*,* $U_{b,M} \in \Omega_b$*. Let* $\mathcal{T}$ *be a uniform mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_{+}^{\star}$ *be the time step. Let* $\left(U_{\mathrm{b},i}^n\right)_{i \in \mathbb{Z}, n \in [T]}$ *be the discrete finite volume approximate solution generated by the Godunov-type scheme* (GtS)*. Then the total entropy stemming from the Godunov-type scheme* (GtS) *for* (Rp1DSW) *satisfies the following inequality*

$$\sum_{i \in \mathbb{Z}} \Phi_i^{n+1} \leq \sum_{i \in \mathbb{Z}} \Phi_i^n - g\nu \sum_{i \in \mathbb{Z}}\left(\mathrm{b}_{i+1} - \mathrm{b}_i\right) \widetilde{hu}_{i+\frac{1}{2}}^n, \quad \forall n \in [T-1],$$

and thus

$$\sum_{i \in \mathbb{Z}} \Phi_i^n \leq \sum_{i \in \mathbb{Z}} \Phi_i^0 - g\nu \sum_{j=0}^{n-1} \sum_{i \in \mathbb{Z}}\left(\mathrm{b}_{i+1} - \mathrm{b}_i\right) \widetilde{hu}_{i+\frac{1}{2}}^j, \quad \forall n \in [T].$$

In particular, for the flat bottom case, this Godunov-type scheme is total entropy diminishing, i.e.,

$$\sum_{i \in \mathbb{Z}} \Phi_i^{n+1} \leq \sum_{i \in \mathbb{Z}} \Phi_i^n, \quad \forall n \in [T-1].$$

The description (2.2) makes sense only if the local Riemann problems do not interact, i.e., if

$$\Delta t < \frac{\inf_{i \in \mathbb{Z}} \Delta x_i}{2 \sup\limits_{[0,T_0)\times\mathbb{R}} \max\left(\left|u - \sqrt{gh}\right|, \left|u + \sqrt{gh}\right|\right)}, \tag{CFL$_1$}$$

or stronger

$$\nu < \frac{\alpha_{\mathcal{T}}}{2 \sup\limits_{[0,T_0)\times\mathbb{R}} \max\left(\left|u - \sqrt{gh}\right|, \left|u + \sqrt{gh}\right|\right)}. \tag{CFL$_2$}$$

On the other hand, (GtS) remains consistent with (2.3) as long as the waves emerges from $\left(i \pm \frac{1}{2}\right) \Delta x_i$ do not touch $\left(i \mp \frac{1}{2}\right) \Delta x_i$ during the time interval $\left[t^n, t^{n+1}\right]$. This will be the case as long as (CFL$_1$) or (CFL$_2$).

The Rankine-Hugoniot relation (A.7) implies that the function

$$s \mapsto F_b\left(U_{\mathrm{b},\mathcal{R}}\left(s; U_b, V_b\right)\right) - sU_{\mathrm{b},\mathcal{R}}\left(s; U_b, V_b\right), \ \ \forall s \in \mathbb{R}, \tag{2.9}$$

is continuous but only piecewisely differentiable. It follows that the Godunov numerical flux defined by (2.5) is only piecewisely differentiable.

Godunov's finite volume scheme fulfills criterion ii) for upwind schemes (see Definition B.5) thanks to (2.1).

The solution to the Riemann problem (Rp1DSW) has a moderate structure

$$U_{\mathrm{b},\mathcal{R}}\left(\frac{x}{t}; U_{b,L}, U_{b,R}\right) = \begin{cases} U_{\mathrm{b},0} := U_{b,L}, & \text{if } \dfrac{x}{t} < \lambda_{\min}\left(U_{\mathrm{b}}\right), \\ U_{\mathrm{b},1}, & \text{if } \lambda_{\min}\left(U_{\mathrm{b}}\right) < \dfrac{x}{t} < \lambda_{\mathrm{mid}}\left(U_{\mathrm{b}}\right), \\ U_{\mathrm{b},2}, & \text{if } \lambda_{\mathrm{mid}}\left(U_{\mathrm{b}}\right) < \dfrac{x}{t} < \lambda_{\max}\left(U_{\mathrm{b}}\right), \\ U_{\mathrm{b},3} := U_{b,R}, & \text{if } \lambda_{\max}\left(U_{\mathrm{b}}\right) < \dfrac{x}{t}, \end{cases} \tag{Rs1DSW}$$

where λ_{mid} is the "between" signal velocity

$$\lambda_{\min}\left(U_b\right) \leq \lambda_{\mathrm{mid}}\left(U_b\right) \leq \lambda_{\max}\left(U_b\right), \ \ \forall U_b \in \Omega_b,$$

and it is straightforward that

$$\lambda_{\mathrm{mid}}\left(U_b\right) = \left(\lambda_{\min}\left(U_b\right)\right)_+ + \left(\lambda_{\max}\left(U_b\right)\right)_- = \left(u - \sqrt{gh}\right)_+ + \left(u + \sqrt{gh}\right)_-, \ \text{in } [0, T_0) \times \mathbb{R}.$$

Since $F_b\left(U_b\right)$ is nonlinear, the k-wave separating $U_{\mathrm{b},k-1}$ and $U_{\mathrm{b},k}$ is not necessarily a single line having a characteristic speed λ_k. If the k^{th} characteristic field is genuinely nonlinear then the k-wave is either a rarefaction wave: $\lambda_k\left(U_{\mathrm{b},k-1}\right) < \lambda_k\left(U_{\mathrm{b},k}\right)$, or a shock propagating with speed S: $\lambda_k\left(U_{\mathrm{b},k-1}\right) > S > \lambda_k\left(U_{\mathrm{b},k}\right)$. If the k^{th} characteristic field is linearly degenerate then the k-wave is a contact discontinuity propagating with speed $\lambda_k\left(U_{\mathrm{b},k-1}\right) = \lambda_k\left(U_{\mathrm{b},k}\right)$.

It is undeniable from (2.3) that, due to averaging, the Godunov scheme does not use of all information contained in the exact solution of the Riemann problem.

Thus, the exact solution $U_{\mathrm{b},\mathcal{R}}\left(x/t; U_{b,L}, U_{b,R}\right)$ of the Riemann problem in (2.4) by an approximation $V_{\mathrm{b}}\left(x/t; U_{b,L}, U_{b,R}\right)$. This approximation V_{b} can have a much simpler structure provided it does not violate the essential properties of conservation and entropy inequality.

Let us now define an extended definition of a steady state solution in agreement with the smooth steady states defined by (sss).

Definition 2.1 (Riemann-steady state solution)**.** *The states $U_{b,L}$ and $U_{b,R}$ define a* steady state solution *if and only if the following relations are satisfied:*

$$\begin{cases} \left\| U_{b,R} - U_{b,L} \right\| = \mathcal{O}\left(\Delta x\right), \\ \left(u_L^2 - c_L^2\right)\left(u_R^2 - c_R^2\right) > 0, \\ h_L u_L = h_R u_R, \\ \dfrac{u_L^2}{2} + g\left(h_L + b_L\right) = \dfrac{u_R^2}{2} + g\left(h_R + b_R\right). \end{cases} \tag{Rsss}$$

Note that the natural condition $\left(u_L^2 - c_L^2\right)\left(u_R^2 - c_R^2\right) > 0$ imposes that both states are either subsonic or supersonic.

The following theorem, analogous to Theorem B.6, indicates that this type of approximation is consistent.

Theorem 2.1 (Modified Harten-Lax for 1D shallow water on an admissible mesh)**.** *Assume $V_{\mathrm{b},i}^n \in \Omega_b$ for all $i \in \mathbb{Z}$, and let $U_{b,L}$ and $U_{b,R}$ belong to Ω_b. Let $V_{\mathrm{b}}\left(x/t; U_{b,L}, U_{b,R}\right)$ be an approximation to the solution of the Riemann problem that fulfills the following conditions:*

i) (Consistency with the integral form of the conservation law) *Assume the following integral consistency condition:*

$$\begin{aligned} \frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; U_{b,L}, U_{b,R}\right) dx = \frac{1}{2}\left(U_{b,L} + U_{b,R}\right) - \nu_i\left(F_b\left(U_{b,R}\right) - F_b\left(U_{b,L}\right)\right) \\ + \Delta t \overline{S}_b\left(\Delta t, \Delta x_i; U_{b,L}, U_{b,R}\right), \end{aligned} \tag{cicl}$$

for $i \in \mathbb{Z}$ and for (CFL$_1$) *(or stronger* (CFL$_2$)*), where $\overline{S}_b\left(\Delta t, \Delta x_i; U_{b,L}, U_{b,R}\right)$ denotes an approximation the "vanishing source term" (Cf.,* (GtS)*):*

$$\lim_{\substack{U_{b,L}\to U_b,\ U_{b,R}\to U_b \\ \Delta t\to 0,\ \Delta x_i \to 0}} \overline{S}_b\left(\Delta t, \Delta x_i; U_{b,L}, U_{b,R}\right) = -S_b\left(U_b\right)\partial_x U_b = 0.$$

Then the updated formula (2.3) *is consistent with* (Rp1DSW) *and*

$$V_{\mathrm{b},i}^{n+1} = V_{\mathrm{b},i}^n - \nu_i\left(F_{\mathrm{b},i+\frac{1}{2}}^n - F_{\mathrm{b},i-\frac{1}{2}}^n\right) + \Delta t S_{\mathrm{b},i}^n, \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T-1], \tag{2.10}$$

with the corresponding numerical flux

$$\begin{aligned} F_{\mathrm{b},i+\frac{1}{2}}^n := \frac{1}{2}\left(F_b\left(V_{\mathrm{b},i}^n\right) + F_b\left(V_{\mathrm{b},i+1}^n\right)\right) - \frac{1}{4\nu_i}\left(V_{\mathrm{b},i+1}^n - V_{\mathrm{b},i}^n\right) \\ + \frac{1}{2\Delta t}\left[\int_0^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V_{\mathrm{b},i}^n, V_{\mathrm{b},i+1}^n\right) dx - \int_{-\Delta x_i/2}^0 V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V_{\mathrm{b},i}^n, V_{\mathrm{b},i+1}^n\right) dx\right], \end{aligned}$$

$$S^n_{\mathrm{b},i} := \frac{1}{2}\left(\overline{S}_b\left(\Delta t, \Delta x_i; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) + \overline{S}_b\left(\Delta t, \Delta x_i; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right)\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T].$$

ii) *For all* $(t, x) \in \mathbb{R}_+ \times \mathbb{R}$, *assume* $V_{\mathrm{b}}\left(x/t; U_{b,L}, U_{b,R}\right)$ *belongs to* Ω_b. *Then* $V^{n+1}_{\mathrm{b},i}$ *stays in* Ω_b *for all* $i \in \mathbb{Z}$.

iii) *If* $U_{b,L}$ *and* $U_{b,R}$ *define a Riemann-steady state according to Definition* 2.1, *assume the following stationary property:*

$$V_{\mathrm{b}}\left(\frac{x}{t}; U_{b,L}, U_{b,R}\right) = V_{b,0}\left(x\right), \;\; \textit{in } \mathbb{R}_+ \times \mathbb{R}. \tag{sp}$$

Then, if $\left(V^n_{\mathrm{b},i}\right)_{i\in\mathbb{Z}}$ *define a Riemann-steady state, we get* $V^{n+1}_{\mathrm{b},i} = V^n_{\mathrm{b},i}$ *for all* $i \in \mathbb{Z}$.

iv) (Consistency with the integral form of the entropy condition) *Write* $V_{\mathrm{b}} = (V, \mathrm{b})$. *Assume the following relation holds:*

$$\frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^{\Delta x_i/2} \Phi\left(V\left(\frac{x}{\Delta t}; U_{b,L}, U_{b,R}\right)\right) dx \leq \frac{1}{2}\left(\Phi\left(U_L\right) + \Phi\left(U_R\right)\right) - \nu_i\left(\Psi\left(U_R\right) - \Psi\left(U_L\right)\right)$$
$$+ \Delta t \sigma\left(\Delta t, \Delta x_i; U_{b,L}, U_{b,R}\right), \tag{ciec}$$

where

$$\lim_{\substack{U_{b,L}\to U_b,\; U_{b,R}\to U_b\\ \Delta t\to 0,\; \Delta x_i \to 0}} \sigma\left(\Delta t, \Delta x_i; U_{b,L}, U_{b,R}\right) = -\left(S_b\left(F_b\left(U_b\right)\right)\partial_x U_b\right)\cdot(0,1,0) = 0.$$

Then, the numerical scheme defined by (2.3) *is entropy preserving:*

$$\Phi\left(V^{n+1}_i\right) \leq \Phi\left(V^n_i\right) - \nu_i\left(\Psi^n_{i+\frac{1}{2}} - \Psi^n_{i-\frac{1}{2}}\right) + \Delta t \sigma^n_i, \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T-1], \tag{ep}$$

with the corresponding numerical entropy flux

$$\begin{aligned}
\Psi^n_{i+\frac{1}{2}} :=\; & \frac{1}{2}\left(\Psi\left(V^n_i\right) + \Psi\left(V^n_{i+1}\right)\right) - \frac{1}{4}\left(\Phi\left(V^n_{i+1}\right) - \Phi\left(V^n_i\right)\right)\\
& + \frac{1}{2\Delta t}\int_0^{\Delta x_i/2} \Phi\left(V\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right)\right) dx\\
& - \frac{1}{2\Delta t}\int_{-\Delta x_i/2}^{0} \Phi\left(V\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right)\right) dx,\\
\sigma^n_i :=\; & \frac{1}{2}\left(\sigma\left(\Delta t, \Delta x_i; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) + \sigma\left(\Delta t, \Delta x_i; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right)\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T].
\end{aligned}$$

Proof. i) By the update formula (2.3) (or (2.4)) for V_{b} instead of $U_{b,\mathcal{R}}$:

$$\begin{aligned}
V^{n+1}_{\mathrm{b},i} :=\; & \frac{1}{\Delta x_i}\int_0^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx + \frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx\\
=\; & \frac{\nu_i}{2}\left[\frac{1}{\Delta t}\int_0^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx - \frac{1}{\Delta t}\int_0^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx\right]\\
& + \frac{\nu_i}{2}\left[-\frac{1}{\Delta t}\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx + \frac{1}{\Delta t}\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \frac{1}{2\Delta x_i}\int_{-\Delta x_i/2}^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx + \frac{1}{2\Delta x_i}\int_{-\Delta x_i/2}^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx \\
= {} & \frac{\nu_i}{2}\left[\frac{1}{\Delta t}\int_{0}^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx - \frac{1}{\Delta t}\int_{0}^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx\right] \\
&+ \frac{\nu_i}{2}\left[-\frac{1}{\Delta t}\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx + \frac{1}{\Delta t}\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx\right] \\
&+ \frac{1}{4}\left(V^n_{\mathrm{b},i-1} + V^n_{\mathrm{b},i}\right) - \frac{\nu_i}{2}\left(F_b\left(V^n_{\mathrm{b},i}\right) - F_b\left(V^n_{\mathrm{b},i-1}\right)\right) + \frac{\Delta t}{2}\overline{S}_b\left(\Delta t, \Delta x_i; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) \\
&+ \frac{1}{4}\left(V^n_{\mathrm{b},i} + V^n_{\mathrm{b},i+1}\right) - \frac{\nu_i}{2}\left(F_b\left(V^n_{\mathrm{b},i+1}\right) - F_b\left(V^n_{\mathrm{b},i}\right)\right) + \frac{\Delta t}{2}\overline{S}_b\left(\Delta t, \Delta x_i; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) \\
= {} & V^n_{\mathrm{b},i} - \frac{\nu_i}{2}\left(F_b\left(V^n_{\mathrm{b},i}\right) + F_b\left(V^n_{\mathrm{b},i+1}\right)\right) + \frac{\nu_i}{2}\left(F_b\left(V^n_{\mathrm{b},i-1}\right) + F_b\left(V^n_{\mathrm{b},i}\right)\right) \\
&- \frac{\nu_i}{2}\left[\frac{1}{\Delta t}\int_{0}^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx - \frac{1}{\Delta t}\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx\right] \\
&+ \frac{\nu_i}{2}\left[\frac{1}{\Delta t}\int_{0}^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx - \frac{1}{\Delta t}\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx\right] \\
&+ \frac{1}{4}\left(V^n_{\mathrm{b},i+1} - V^n_{\mathrm{b},i}\right) - \frac{1}{4}\left(V^n_{\mathrm{b},i} - V^n_{\mathrm{b},i-1}\right) \\
&+ \frac{\Delta t}{2}\left(\overline{S}_b\left(\Delta t, \Delta x_i; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) + \overline{S}_b\left(\Delta t, \Delta x_i; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right)\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T].
\end{aligned}
$$

Combining this with (cicl) yields exactly (2.10).

ii) First of all,

$$
\begin{aligned}
V^n_{\mathrm{b},\mathcal{T},\Delta t}(t,x) &:= \sum_{i\in\mathbb{Z}} V_{\mathrm{b}}\left(\frac{x - x_{i+\frac{1}{2}}}{t - t^n}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right)\mathbf{1}_{(x_i, x_{i+1}]}(x), \text{ in } \left[t^n, t^{n+1}\right] \times \mathbb{R}, \\
V^{n+1}_{\mathrm{b},i} &:= \frac{1}{\Delta x_i}\int_{C_i} V^n_{\mathrm{b},\mathcal{T},\Delta t}\left(t^{n+1}, x\right) dx \\
&= \frac{1}{\Delta x_i}\int_{0}^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i-1}, V^n_{\mathrm{b},i}\right) dx + \frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) dx, \\
V_{\mathrm{b},\mathcal{T},\Delta t}(t^n, x) &:= \sum_{i\in\mathbb{Z}} V^n_{\mathrm{b},i}\mathbf{1}_{\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right)}(x), \quad \forall x \in \mathbb{R}.
\end{aligned}
$$

Assume $V_{\mathrm{b}}\left(x/t; U_{b,L}, U_{b,R}\right) \in \Omega_b$ for all $(t,x) \in \mathbb{R}_+ \times \mathbb{R}$, i.e., the approximate Riemann solver V_{b} preserves the water height nonnegativity. The last second equality implies that $V^{n+1}_{\mathrm{b},i} \in \Omega_b$.

iii) Suppose that $U_{b,L}$ and $U_{b,R}$ define a Riemann-steady state, i.e., (Rsss) holds, the stationary property (sp) holds, and $\left(V^n_{\mathrm{b},i}\right)_{i\in\mathbb{Z}}$, with $V^n_{\mathrm{b},i} = (k^n_i, k^n_i v^n_i, \mathrm{b}_i)$, define a Riemann-steady state, i.e., for all $i \in \mathbb{Z}$,

$$
V_{\mathrm{b}}\left(\frac{x}{t}; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) = V_{b,0}\left(x; V^n_{\mathrm{b},i}, V^n_{\mathrm{b},i+1}\right) := \begin{cases} V^n_{\mathrm{b},i}, & \text{if } x < 0, \\ V^n_{\mathrm{b},i+1}, & \text{if } x \geq 0, \end{cases}
$$

and

$$\begin{cases} \left|h_{i+1}^{n}-h_{i}^{n}\right|+\left|h_{i+1}^{n}u_{i+1}^{n}-h_{i}^{n}u_{i}^{n}\right|=\mathcal{O}\left(\Delta x\right), \\ \left(\left(u_{i}^{n}\right)^{2}-gh_{i}^{n}\right)\left(\left(u_{i+1}^{n}\right)^{2}-gh_{i+1}^{n}\right)>0, \\ h_{i}^{n}u_{i}^{n}=h_{i+1}^{n}u_{i+1}^{n}, \\ \dfrac{\left(u_{i}^{n}\right)^{2}}{2}+g\left(h_{i}^{n}+b_{i}\right)=\dfrac{\left(u_{i+1}^{n}\right)^{2}}{2}+g\left(h_{i+1}^{n}+b_{i+1}\right), \end{cases}$$

then the relation in the proof i) and (sp) imply

$$\begin{aligned} V_{\mathrm{b},i}^{n+1}:&=\frac{1}{\Delta x_{i}}\int_{0}^{\Delta x_{i}/2}V_{\mathrm{b}}\left(\frac{x}{\Delta t};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)dx+\frac{1}{\Delta x_{i}}\int_{-\Delta x_{i}/2}^{0}V_{\mathrm{b}}\left(\frac{x}{\Delta t};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)dx \\ &=\frac{1}{\Delta x_{i}}\frac{\Delta x_{i}}{2}V_{\mathrm{b},i}^{n}+\frac{1}{\Delta x_{i}}\frac{\Delta x_{i}}{2}V_{\mathrm{b},i}^{n}=V_{\mathrm{b},i}^{n},\quad\forall i\in\mathbb{Z}. \end{aligned}$$

iv) Since $U\mapsto\Phi\left(U\right)$ is convex, we have, with the help of (ep),

$$\begin{aligned} &\Phi\left(V_{i}^{n+1}\right) \\ &=\Phi\left(\frac{1}{\Delta x_{i}}\int_{0}^{\Delta x_{i}/2}V\left(\frac{x}{\Delta t};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)dx+\frac{1}{\Delta x_{i}}\int_{-\Delta x_{i}/2}^{0}V\left(\frac{x}{\Delta t};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)dx\right) \\ &\leq\Phi\left(\frac{1}{\Delta x_{i}}\int_{0}^{\Delta x_{i}/2}V\left(\frac{x}{\Delta t};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)dx\right)+\Phi\left(\frac{1}{\Delta x_{i}}\int_{-\Delta x_{i}/2}^{0}V\left(\frac{x}{\Delta t};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)dx\right) \\ &\leq\frac{1}{\Delta x_{i}}\int_{0}^{\Delta x_{i}/2}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)\right)dx+\frac{1}{\Delta x_{i}}\int_{-\Delta x_{i}/2}^{0}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)\right)dx \\ &\leq\frac{1}{2}\left[\frac{1}{\Delta x_{i}}\int_{0}^{\Delta x_{i}/2}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)\right)dx-\frac{1}{\Delta x_{i}}\int_{0}^{\Delta x_{i}/2}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)\right)dx\right] \\ &\quad+\frac{1}{2}\left[-\frac{1}{\Delta x_{i}}\int_{-\Delta x_{i}/2}^{0}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)\right)dx+\frac{1}{\Delta x_{i}}\int_{-\Delta x_{i}/2}^{0}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)\right)dx\right] \\ &\quad+\frac{1}{4}\left(\Phi\left(V_{i-1}^{n}\right)+\Phi\left(V_{i}^{n}\right)\right)-\frac{\nu_{i}}{2}\left(\Psi\left(V_{i}^{n}\right)-\Psi\left(V_{i-1}^{n}\right)\right)+\frac{\Delta t}{2}\sigma\left(\Delta t,\Delta x_{i};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right) \\ &\quad+\frac{1}{4}\left(\Phi\left(V_{i}^{n}\right)+\Phi\left(V_{i+1}^{n}\right)\right)-\frac{\nu_{i}}{2}\left(\Psi\left(V_{i+1}^{n}\right)-\Psi\left(V_{i}^{n}\right)\right)+\frac{\Delta t}{2}\sigma\left(\Delta t,\Delta x_{i};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right) \\ &=\Phi\left(V_{i}^{n}\right)-\frac{\nu_{i}}{2}\left(\Psi\left(V_{i}^{n}\right)+\Psi\left(V_{i+1}^{n}\right)\right)+\frac{\nu_{i}}{2}\left(\Psi\left(V_{i-1}^{n}\right)+\Psi\left(V_{i}^{n}\right)\right) \\ &\quad-\frac{1}{2}\left[\frac{1}{\Delta x_{i}}\int_{0}^{\Delta x_{i}/2}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)\right)dx-\frac{1}{\Delta x_{i}}\int_{-\Delta x_{i}/2}^{0}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)\right)dx\right] \\ &\quad+\frac{1}{2}\left[\frac{1}{\Delta x_{i}}\int_{0}^{\Delta x_{i}/2}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)\right)dx-\frac{1}{\Delta x_{i}}\int_{-\Delta x_{i}/2}^{0}\Phi\left(V\left(\frac{x}{\Delta t};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)\right)dx+\right] \\ &\quad+\frac{1}{4}\left(\Phi\left(V_{i+1}^{n}\right)-\Phi\left(V_{i}^{n}\right)\right)-\frac{1}{4}\left(\Phi\left(V_{i}^{n}\right)-\Phi\left(V_{i-1}^{n}\right)\right) \\ &\quad+\frac{\Delta t}{2}\left(\sigma\left(\Delta t,\Delta x_{i};V_{\mathrm{b},i-1}^{n},V_{\mathrm{b},i}^{n}\right)+\sigma\left(\Delta t,\Delta x_{i};V_{\mathrm{b},i}^{n},V_{\mathrm{b},i+1}^{n}\right)\right), \end{aligned}$$

where the second inequality is obtained by using the following property of convex function: “If f is a convex function of one real variable, and $f\left(0\right)\leq0$ then f is superadditive on the positive reals.” (note that Φ is convex, $\Phi\left(0\right)=0$, $\Phi\left(h,hu\right)=\Phi\left(h,-hu\right)$). Under the setting in iv), the last inequality is exactly the required discrete entropy inequality (ep). This completes our proof. □

Given $U_{b,L}$, $U_{b,R} \in \Omega_b$ arbitrary, the approximate solution $V_{\mathrm{b}} = (k, kv, \mathrm{b})$ satisfies the following approximate Riemann problem

$$\begin{cases} \partial_t U_b + \partial_x \left(F_b \left(U_b \right) \right) + \mathcal{S}_b \left(t, x; U_{b,L}, U_{b,R} \right) = 0, & \text{in } [0, T_0) \times \mathbb{R}, \\ U_b \left(0, x \right) = U_{b,0} \left(x; U_{b,L}, U_{b,R} \right), & \text{in } \mathbb{R}, \end{cases} \tag{aRp1DSW}$$

where $\mathcal{S}_b \left(t, x; U_{b,L}, U_{b,R} \right) = \left(\mathcal{S} \left(t, x; U_{b,L}, U_{b,R} \right), 0 \right)^\top : [0, T_0) \times \mathbb{R} \to \mathbb{R}^3$ is an approximate source term satisfying

$$\mathcal{S}_b \left(t, x; U_{b,L}, U_{b,R} \right) \approx \left(S_b \left(U_{b,\mathcal{R}} \right) \partial_x U_{b,\mathcal{R}} \right) \left(\frac{x}{t}; U_{b,L}, U_{b,R} \right), \text{ in } [0, T_0) \times \mathbb{R},$$

especially the third component of $\mathcal{S}_b$ vanishes:

$$\mathcal{S}_b \left(t, x; U_{b,L}, U_{b,R} \right) \cdot (0, 0, 1) = 0, \text{ in } [0, T_0) \times \mathbb{R}.$$

Integrating the PDE solved by V_{b} over the rectangle $(0, \Delta t) \times (-\Delta x_i/2, 0)$ yields

$$\begin{aligned} &\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}} \left(\frac{x}{\Delta t}; U_{b,L}, U_{b,R} \right) dx - \int_{-\Delta x_i/2}^{0} U_{b,0} \left(x; U_{b,L}, U_{b,R} \right) dx \\ &\quad + \int_0^{\Delta t} F_b \left(V_{\mathrm{b}} \left(0; U_{b,L}, U_{b,R} \right) \right) dt - \int_0^{\Delta t} F_b \left(V_{\mathrm{b}} \left(-\frac{\Delta x_i}{2t}; U_{b,L}, U_{b,R} \right) \right) dt \\ &\quad + \int_0^{\Delta t} \int_{-\Delta x_i/2}^{0} \mathcal{S}_b \left(t, x; U_{b,L}, U_{b,R} \right) dx dt = 0, \quad \forall i \in \mathbb{Z}, \end{aligned}$$

i.e.,

$$\begin{aligned} &\int_{-\Delta x_i/2}^{0} V_{\mathrm{b}} \left(\frac{x}{\Delta t}; U_{b,L}, U_{b,R} \right) dx - \frac{\Delta x_i}{2} U_{b,L} + \Delta t \left[F_b \left(V_{\mathrm{b}} \left(0; U_{b,L}, U_{b,R} \right) \right) - F_b \left(U_{b,L} \right) \right] \\ &\quad + \int_0^{\Delta t} \int_{-\Delta x_i/2}^{0} \mathcal{S}_b \left(t, x; U_{b,L}, U_{b,R} \right) dx dt, \quad \forall i \in \mathbb{Z}, \end{aligned}$$

or

$$\begin{aligned} F_b \left(V_{\mathrm{b}} \left(0; U_{b,L}, U_{b,R} \right) \right) = F_b \left(U_{b,L} \right) + \frac{\Delta x_i}{2 \Delta t} U_{b,L} - \frac{1}{\Delta t} \int_{-\Delta x_i/2}^{0} V_{\mathrm{b}} \left(\frac{x}{\Delta t}; U_{b,L}, U_{b,R} \right) dx \\ - \frac{1}{\Delta t} \int_0^{\Delta t} \int_{-\Delta x_i/2}^{0} \mathcal{S}_b \left(t, x; U_{b,L}, U_{b,R} \right) dx dt, \quad \forall i \in \mathbb{Z}, \end{aligned}$$

where we have used the following assumption on V_{b}:

$$V_{\mathrm{b}} \left(\frac{x}{t}; U_{b,L}, U_{b,R} \right) = \begin{cases} U_{b,L}, & \text{if } \frac{x}{t} < \lambda_{\min} \left(U_b \right), \\ U_{b,R}, & \text{if } \frac{x}{t} > \lambda_{\min} \left(U_b \right), \end{cases} \text{ in } [0, T_0) \times \mathbb{R}, \tag{2.11}$$

which is analogous to (2.1), to deduce $V_{\mathrm{b}} \left(-\Delta x_i / 2t; U_{b,L}, U_{b,R} \right) = U_{b,L}$, $\forall t \in (0, \Delta t)$, provided the mentioned CFL condition holds.

Similarly, integrating the PDE in (aRp1DSW) over the rectangle $(0,\Delta t)\times(0,\Delta x_i/2)$ yields

$$\int_0^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t};U_{b,L},U_{b,R}\right)dx-\frac{\Delta x_i}{2}U_{b,R}+\Delta t\left[F_b\left(U_{b,R}\right)-F_b\left(V_{\mathrm{b}}\left(0;U_{b,L},U_{b,R}\right)\right)\right]$$
$$+\int_0^{\Delta t}\int_0^{\Delta x_i/2}\mathcal{S}_b\left(t,x;U_{b,L},U_{b,R}\right)dxdt=0,\ \ \forall i\in\mathbb{Z},$$

or

$$F_b\left(V_{\mathrm{b}}\left(0;U_{b,L},U_{b,R}\right)\right)=F_b\left(U_{b,R}\right)-\frac{\Delta x_i}{2\Delta t}U_{b,R}+\frac{1}{\Delta t}\int_0^{\Delta x_i/2}V_{\mathrm{b}}\left(\frac{x}{\Delta t};U_{b,L},U_{b,R}\right)dx$$
$$+\frac{1}{\Delta t}\int_0^{\Delta t}\int_0^{\Delta x_i/2}\mathcal{S}_b\left(t,x;U_{b,L},U_{b,R}\right)dxdt,\ \ \forall i\in\mathbb{Z}.$$

Subtracting two formulas for $F_b\left(V_{\mathrm{b}}\left(0;U_{b,L},U_{b,R}\right)\right)$, we obtain

$$\frac{1}{\Delta t}\int_{-\Delta x_i/2}^{\Delta x_i/2}V_{\mathrm{b}}\left(\frac{x}{\Delta t};U_{b,L},U_{b,R}\right)dx=\frac{\Delta x_i}{2\Delta t}\left(U_{b,L}+U_{b,R}\right)-\left(F_b\left(U_{b,R}\right)-F_b\left(U_{b,L}\right)\right)$$
$$-\frac{1}{\Delta t}\int_0^{\Delta t}\int_{-\Delta x_i/2}^{\Delta x_i/2}\mathcal{S}_b\left(t,x;U_{b,L},U_{b,R}\right)dxdt,\ \ \forall i\in\mathbb{Z}.$$

Setting

$$\overline{\mathcal{S}}_b\left(\Delta t,\Delta x_i;U_{b,L},U_{b,R}\right):=\frac{1}{\Delta t\Delta x_i}\int_0^{\Delta t}\int_{-\Delta x_i/2}^{\Delta x_i/2}\mathcal{S}_b\left(t,x;U_{b,L},U_{b,R}\right)dxdt,\ \ \forall i\in\mathbb{Z},$$

then the last integral identity is exactly (cicl) with the approximate source term $\overline{\mathcal{S}}_b\left(\Delta t,\Delta x_i;U_{b,L},U_{b,R}\right)$.

Here is a short program for the approximate solution V_{b} imitated that for the exact solution $U_{b,\mathcal{R}}$.

1. Denote by $V_{\mathrm{b}}\left(x/t;U_{b,L},U_{b,R}\right)$ the exact solution of the approximate Riemann's problem (aRp1DSW).

2. Assume that V_{b} satisfies (2.11).

3. Given an admissible mesh $\mathcal{T}$ of $\mathbb{R}$ and some time step $\Delta t\in\mathbb{R}_+^\star$, consider the numerical approximation $V_{\mathrm{b},\mathcal{T},\Delta t}\left(t^n,x\right)$ of the discrete time levels t^n, for all $n\in[n_0]$, for some $n_0\in[T-1]$, to be a piecewise constant function in x, i.e.,

$$V_{\mathrm{b},\mathcal{T},\Delta t}\left(t^n,x\right):=\sum_{i\in\mathbb{Z}}V_i^n\mathbf{1}_{\left[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}\right)}(x),\ \ \forall x\in\mathbb{R}.$$

4. Solve exactly the initial-value problem

$$\begin{cases}\partial_tU_b+\partial_x\left(F_b\left(U_b\right)\right)+\mathcal{S}_b^n\left(t,x\right)=0, & \text{in } \left[t^n,t^{n+1}\right]\times\mathbb{R},\\ U_b\left(t^n,x\right)=V_{\mathrm{b},\mathcal{T},\Delta t}\left(t^n,x\right), & \text{in } \mathbb{R},\end{cases}$$

for all $t \in \left[t^n, t^{n+1}\right]$, where

$$\mathcal{S}_b^n(t,x) := \sum_{i\in\mathbb{Z}} \mathcal{S}_b\left(t,x;V_{\mathrm{b},i}^n, V_{\mathrm{b},i+1}^n\right)\mathbf{1}_{(x_i,x_{i+1}]}(x), \text{ in } \left[t^n, t^{n+1}\right]\times\mathbb{R},$$

and denote its solution by $V_{\mathrm{b},\mathcal{T},\Delta t}^n(t,x)$.

Each of discontinuities in $V_{\mathrm{b},\mathcal{T},\Delta t}(t^n,x)$ comprises locally a Riemann's problem. If one of the CFL conditions (CFL$_1$) and (CFL$_2$) holds. Due to (2.11), there is no interaction between neighboring Riemann problems, and $V_{\mathrm{b},\mathcal{T},\Delta t}^n(t,x)$ can be expressed exactly in terms of the solutions of local Riemann's problems:

$$V_{\mathrm{b},\mathcal{T},\Delta t}^n(t,x) = \sum_{i\in\mathbb{Z}} V_{\mathrm{b}}\left(\frac{x-x_{i+\frac{1}{2}}}{t-t^n}; V_{\mathrm{b},i}^n, V_{\mathrm{b},i+1}^n\right)\mathbf{1}_{(x_i,x_{i+1}]}(x), \text{ in } \left[t^n,t^{n+1}\right]\times\mathbb{R}.$$

We then obtain a piecewise constant approximation $V_{\mathrm{b},\mathcal{T},\Delta t}\left(t^{n+1},x\right)$ by averaging the quantity $V_{\mathrm{b},\mathcal{T},\Delta t}^n\left(t^{n+1},x\right)$ over each cells, i.e.,

$$V_{\mathrm{b},i}^{n+1} := \frac{1}{\Delta x_i}\int_{C_i} V_{\mathrm{b},\mathcal{T},\Delta t}^n\left(t^{n+1},x\right)dx, \ \ \forall i\in\mathbb{Z},$$

which can be rewritten in terms of the solutions of the local approximate Riemann's problem as

$$V_{\mathrm{b},i}^{n+1} = \frac{1}{\Delta x_i}\int_0^{\Delta x_i/2} V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V_{\mathrm{b},i-1}^n, V_{\mathrm{b},i}^n\right)dx + \frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^0 V_{\mathrm{b}}\left(\frac{x}{\Delta t}; V_{\mathrm{b},i}^n, V_{\mathrm{b},i+1}^n\right)dx,$$

for all $i\in\mathbb{Z}$.

5. Integrating the following equation

$$\partial_t V_{\mathrm{b},\mathcal{T},\Delta t}^n + \partial_x\left(F_b\left(V_{\mathrm{b},\mathcal{T},\Delta t}^n\right)\right) + \mathcal{S}_b^n(t,x) = 0, \text{ in } \left[t^n,t^{n+1}\right]\times\mathbb{R},$$

over the rectangle $\left[t^n,t^{n+1}\right]\times C_i$ yields

$$\int_{C_i} V_{\mathrm{b},\mathcal{T},\Delta t}^n\left(t^{n+1},x\right)dx - \int_{C_i} V_{\mathrm{b},\mathcal{T},\Delta t}^n\left(t^n,x\right)dx + \int_{t^n}^{t^{n+1}} F_b\left(V_{\mathrm{b},\mathcal{T},\Delta t}^n\left(t,x_{i+\frac{1}{2}}\right)\right)dt$$
$$-\int_{t^n}^{t^{n+1}} F_b\left(V_{\mathrm{b},\mathcal{T},\Delta t}^n\left(t,x_{i-\frac{1}{2}}\right)\right)dt + \int_{t^n}^{t^{n+1}}\int_{C_i}\mathcal{S}_b^n(t,x)\,dxdt = 0, \ \ \forall i\in\mathbb{Z},$$

And then

$$\int_{x_{i-\frac{1}{2}}}^{x_i} V_{\mathrm{b}}\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta t}; V_{\mathrm{b},i-1}^n, V_{\mathrm{b},i}^n\right)dx + \int_{x_i}^{x_{i+\frac{1}{2}}} V_{\mathrm{b}}\left(\frac{x-x_{i+\frac{1}{2}}}{\Delta t}; V_{\mathrm{b},i}^n, V_{\mathrm{b},i+1}^n\right)dx$$
$$-\Delta x_i V_{\mathrm{b},i}^n + \Delta t\left[F_b\left(V_{\mathrm{b}}\left(0;V_{\mathrm{b},i}^n,V_{\mathrm{b},i+1}^n\right)\right) - F_b\left(V_{\mathrm{b}}\left(0;V_{\mathrm{b},i-1}^n,V_{\mathrm{b},i}^n\right)\right)\right]$$
$$+\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_i}\mathcal{S}_b^n\left(t,x;V_{\mathrm{b},i-1}^n,V_{\mathrm{b},i}^n\right)dxdt$$
$$+\int_{t^n}^{t^{n+1}}\int_{x_i}^{x_{i+\frac{1}{2}}}\mathcal{S}_b^n\left(t,x;V_{\mathrm{b},i}^n,V_{\mathrm{b},i+1}^n\right)dxdt = 0, \ \ \forall i\in\mathbb{Z},$$

equivalently,

$$\begin{aligned}\Delta x_i V_{\mathrm{b},i}^{n+1} = \Delta x_i V_{\mathrm{b},i}^{n} - \Delta t\left[F_b\left(V_{\mathrm{b}}\left(0; V_{\mathrm{b},i}^{n}, V_{\mathrm{b},i+1}^{n}\right)\right) - F_b\left(V_{\mathrm{b}}\left(0; V_{\mathrm{b},i-1}^{n}, V_{\mathrm{b},i}^{n}\right)\right)\right] \\ + \frac{\Delta t \Delta x_i}{2}\left(\overline{\mathcal{S}}_{\mathrm{b},i}^{n,-} + \overline{\mathcal{S}}_{\mathrm{b},i}^{n,+}\right), \quad \forall i \in \mathbb{Z},\end{aligned}$$

where

$$\begin{cases} \overline{\mathcal{S}}_{\mathrm{b},i}^{n,-} := \dfrac{2}{\Delta t \Delta x_i}\displaystyle\int_{t^n}^{t^{n+1}}\int_{x_{i-\frac{1}{2}}}^{x_i} \mathcal{S}_b^n\left(t, x; V_{\mathrm{b},i-1}^{n}, V_{\mathrm{b},i}^{n}\right) dxdt, \\ \overline{\mathcal{S}}_{\mathrm{b},i}^{n,+} := \dfrac{2}{\Delta t \Delta x_i}\displaystyle\int_{t^n}^{t^{n+1}}\int_{x_i}^{x_{i+\frac{1}{2}}} \mathcal{S}_b^n\left(t, x; V_{\mathrm{b},i}^{n}, V_{\mathrm{b},i+1}^{n}\right) dxdt, \end{cases} \quad \forall i \in \mathbb{Z}.$$

Hence,

$$V_{\mathrm{b},i}^{n+1} = V_{\mathrm{b},i}^{n} - \nu_i\left[F_b\left(V_{\mathrm{b}}\left(0; V_{\mathrm{b},i}^{n}, V_{\mathrm{b},i+1}^{n}\right)\right) - F_b\left(V_{\mathrm{b}}\left(0; V_{\mathrm{b},i-1}^{n}, V_{\mathrm{b},i}^{n}\right)\right)\right] + \Delta t \overline{\mathcal{S}}_{\mathrm{b},i}^{n},$$

where

$$\overline{\mathcal{S}}_{\mathrm{b},i}^{n} := \frac{1}{2}\left(\overline{\mathcal{S}}_{\mathrm{b},i}^{n,-} + \overline{\mathcal{S}}_{\mathrm{b},i}^{n,+}\right) = \frac{1}{\Delta t \Delta x_i}\int_{t^n}^{t^{n+1}}\int_{C_i} \mathcal{S}_b^n\left(t, x\right) dxdt, \quad \forall i \in \mathbb{Z}.$$

Hence we obtain the approximate Godunov-type scheme

$$V_{\mathrm{b},i}^{n+1} = V_{\mathrm{b},i}^{n} - \nu_i\left(F_b\left(\widetilde{V}_{\mathrm{b},i+\frac{1}{2}}^{n}\right) - F_b\left(\widetilde{V}_{\mathrm{b},i-\frac{1}{2}}^{n}\right)\right) + \Delta t \overline{\mathcal{S}}_{\mathrm{b},i}^{n}, \quad \forall i \in \mathbb{Z}, \tag{aGtS}$$

where

$$\widetilde{V}_{\mathrm{b},i+\frac{1}{2}}^{n} := V_b\left(0; V_{\mathrm{b},i}^{n}, V_{\mathrm{b},i+1}^{n}\right), \quad \forall i \in \mathbb{Z},$$

2.3 Properties of the approximate Godunov-type scheme (aGtS) for 1D shallow water equations

Let V satisfy the approximate PDE for (1DSW), i.e.,

$$\partial_t V + \partial_x\left(F\left(V\right)\right) + \mathcal{S} = 0, \text{ in } [0, T_0) \times \mathbb{R}, \tag{a1DSW}$$

where the source term $\mathcal{S}$ is defined by

$$\mathcal{S}\left(t, x\right) := \sum_{n=0}^{T}\sum_{i \in \mathbb{Z}} \mathcal{S}\left(t, x; V_{\mathrm{b},i}^{n}, V_{\mathrm{b},i+1}^{n}\right) \mathbf{1}_{[t^n, t^{n+1}) \times (x_i, x_{i+1}]}\left(t, x\right), \text{ in } [0, T_0) \times \mathbb{R}.$$

Similar to (EI_1), it is easy to verify the following approximate entropy inequality for (a1DSW):

$$\partial_t\left(\Phi\left(V\right)\right) + \partial_x\left(\Psi\left(V\right)\right) \leq -\nabla\Phi\left(V\right) \cdot \mathcal{S}, \text{ in } [0, T_0) \times \mathbb{R}. \tag{aEI}$$

Integrating (aEI) for $V^n_{\mathcal{T},\Delta t}$ over the rectangle $\left[t^n, t^{n+1}\right] \times C_i$ yields

$$\begin{aligned}\int_{C_i} \Phi\left(V^n_{\mathcal{T},\Delta t}\left(t^{n+1}, x\right)\right) dx \leq \Delta x_i \Phi\left(V^n_i\right) - \Delta t\left(\Psi\left(\widetilde{V}^n_{i+\frac{1}{2}}\right) - \Psi\left(\widetilde{V}^n_{i-\frac{1}{2}}\right)\right)\\ - \int_{t^n}^{t^{n+1}} \int_{C_i} \left(\nabla\Phi\left(V^n_{\mathcal{T},\Delta t}\right) \cdot \mathcal{S}\right)(t, x)\, dxdt, \quad \forall i \in \mathbb{Z}.\end{aligned}$$

Chapter 3

A Staggered Upwind Scheme for 1D Shallow Water Equations

In this chapter, we restudy an explicit staggered finite volume scheme for the shallow water equations (1DSW). This scheme is identical to the one investigated in Herbin, Latché, and Nguyen, 2013 when the topography is flat, e.g., $b \equiv 0$. After reintroducing the numerical scheme, a number of its properties will be proved:

- preservation of the water height nonnegativity, i.e., $h(t, x) \geq 0$.
- preservation of some particular discrete steady states (well-balanced property),
- consistency with the entropy inequality.

The preservation of the water height nonnegativity $h(t, x) \geq 0$ is physically relevant and is crucial for the stability. If negative quantities occur in computation, the numerical scheme will fail in general. The nonnegativity of the water height is obtained for this staggered scheme under a CFL-like condition (2.5).

The staggered scheme in this chapter does not satisfy a discrete entropy inequality[1]. Nevertheless, in Doyen and Gunawan, 2014, Sec. 3, pp. 231–232, the authors proved that it is consistent with the global entropy inequality, guaranteeing that the discontinuities computed by this scheme are admissible discontinuities.

We start this chapter by recalling the staggered scheme investigated in Doyen and Gunawan, 2014 being adapted on any admissible meshes.

1 Doyan-Gunawan's staggered scheme in the horizontal fluid domain

This section is devoted to discuss the adaptation of the upwind scheme on *staggered meshes*. While we present the framework in the 1D case, the method can be adapted to higher dimensions. We will use a separated chapter to discuss this. First of all, let us describe the notion of staggered meshes (Cf., admissible meshes of $\mathbb{R}$ defined in Definition (B.1)).

[1] A continuous entropy inequality is given by (Eb).

1.1 Staggered meshes

First, we keep consider (1DSW) on the horizontal fluid domain $\mathbb{R}$ during $[0, T_0)$ described in the previous chapter. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$. The set of points $\left(x_{i+\frac{1}{2}}\right)_{i\in\mathbb{Z}}$ defines a subdivision of the horizontal fluid domain $\mathbb{R}$ which is called the *primal mesh*. And $x_{i+\frac{1}{2}}$'s are referred to as the *interfaces* (or *edges, vertices*) of the primal mesh. Set

$$\Delta x_{i+\frac{1}{2}} := x_{i+1} - x_i = \frac{\Delta x_i + \Delta x_{i+1}}{2}, \text{ and } \nu_{i+\frac{1}{2}} := \frac{\Delta t}{\Delta x_{i+\frac{1}{2}}}, \quad \forall i \in \mathbb{Z}.$$

The points $(x_i)_{i\in\mathbb{Z}}$ are the mid-points of the cells C_i's and also the centers of the primal cells and realize the *dual mesh* whose dual cells $C_{i+\frac{1}{2}}$'s are defined by $C_{i+\frac{1}{2}} := (x_i, x_{i+1})$ for all $i \in \mathbb{Z}$.

The water height h and the bottom topography b are discretized at the center of the primal cells of $\mathcal{T}$.

- The approximation of the water height h at point x_i and at time t^n is denoted by h_i^n, for all $i \in \mathbb{Z}$, $n \in [T]$.
- The approximation of the time-independent bottom topography b at point x_i is denoted by b_i, for all $i \in \mathbb{Z}$, $n \in [T]$.

The velocity u is discretized at the interfaces between the primal cells (also, the nodes of the primal cells). Note also that these points $x_{i+\frac{1}{2}}$'s are the centers of the dual cells in general if $\mathcal{T}$ is not uniform.

- The approximation of u at point $x_{i+\frac{1}{2}}$ and at time t^n is denoted by $u^n_{i+\frac{1}{2}}$, for all $i \in \mathbb{Z}$, $n \in [T]$.

1.2 An explicit upwind scheme

The mass conservation equation in (1DSW) is discretized with an explicit upwind scheme on the dual mesh, which thus requires an approximation of the fluxes hu on the primal mesh:

$$h_i^{n+1} := h_i^n - \nu_i \left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T-1], \tag{dmc}$$

where

$$F^n_{i+\frac{1}{2}} := h^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}}, \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T],$$

and where $h^n_{i+\frac{1}{2}}$ is calculated by an upwind shift according to the sign of $u^n_{i+\frac{1}{2}}$:

$$h^n_{i+\frac{1}{2}} = \begin{cases} h_i^n, & \text{if } u^n_{i+\frac{1}{2}} \geq 0, \\ h^n_{i+1}, & \text{otherwise,} \end{cases} \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T].$$

Hence, the numerical flux $F^n_{i+\frac{1}{2}}$ can be rewritten as

$$F^n_{i+\frac{1}{2}} = h^n_i \left(u^n_{i+\frac{1}{2}}\right)_+ + h^n_{i+1}\left(u^n_{i+\frac{1}{2}}\right)_-, \quad \forall i \in \mathbb{Z}, \ \forall n \in [T] .$$

Analogous to Proposition 2.1, 2.2, and Proposition B.1, the explicit upwind scheme (dmc) preserve the total height water, i.e., the water volume (this is also the best advantage of Finite Volume Methods).

Proposition 3.1 (Conservation of water volume)**.** *Assume that the initial data satisfies $h_0 \in L^\infty(\mathbb{R})$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}^\star_+$ be the time step. Let $\left(u^n_{i+\frac{1}{2}}\right)_{i\in\mathbb{Z},n\in[T]}$ be an arbitrary sequence of reals and then let $(h^n_i)_{i\in\mathbb{Z},n\in[T]}$ be the discrete finite volume approximate water height generated by* (dmc)*. Then the explicit upwind scheme* (dmc) *preserves the water volume in the discrete level, with the 1D discrete water volume at the time $t = t^n$ by*

$$\mathfrak{h}^n := \sum_{i\in\mathbb{Z}} \Delta x_i h^n_i, \quad \forall n \in [T] .$$

Proof. We claim that $\mathfrak{h}^{n+1} = \mathfrak{h}^n$ for all $n \in [T-1]$. Indeed,

$$\mathfrak{h}^{n+1} := \sum_{i\in\mathbb{Z}} \Delta x_i h^{n+1}_i = \sum_{i\in\mathbb{Z}} \left[\Delta x_i h^n_i - \Delta t \left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\right)\right] = \sum_{i\in\mathbb{Z}} \Delta x_i h^n_i = \mathfrak{h}^n,$$

for all $n \in [T-1]$. As a consequence, the initial discrete water volume is preserved under (dmc):

$$\mathfrak{h}^n = \mathfrak{h}^0, \quad \forall n \in [T] .$$

This completes our proof. □

The momentum balance equation in (1DSW) is discretized on the primal mesh with explicit upwind fluxes for the convection term and implicit centered fluxes for the pressure term and bottom topography term on the dual mesh:

$$\begin{aligned}\bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} &= \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - \nu_{i+\frac{1}{2}} \left[G^n_{i+1} - G^n_i + \frac{g}{2}\left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2\right)\right] \\ &\quad - g\nu_{i+\frac{1}{2}} \bar{h}^{n+1}_{i+\frac{1}{2}} (b_{i+1} - b_i), \qquad \forall i \in \mathbb{Z}, \ \forall n \in [T-1], \end{aligned} \tag{dmb}$$

where

$$\begin{aligned} \bar{h}^n_{i+\frac{1}{2}} &:= \frac{1}{2}\left(h^n_i + h^n_{i+1}\right), && \forall i \in \mathbb{Z}, \ \forall n \in [T], \\ \overline{F}^n_i &:= \frac{1}{2}\left(F^n_{i-\frac{1}{2}} + F^n_{i+\frac{1}{2}}\right), && \forall i \in \mathbb{Z}, \ \forall n \in [T], \\ G^n_i &:= u^n_i \overline{F}^n_i = \frac{1}{2} u^n_i \left(F^n_{i-\frac{1}{2}} + F^n_{i+\frac{1}{2}}\right), && \forall i \in \mathbb{Z}, \ \forall n \in [T], \end{aligned}$$

where $(u_i^n)_{i\in\mathbb{Z},n\in[T]}$ is calculated by an upwind shift according to the sign of $\overline{F}_i^n$:

$$u_i^n := \begin{cases} u_{i-\frac{1}{2}}^n, & \text{if } \overline{F}_i^n \geq 0, \\ u_{i+\frac{1}{2}}^n, & \text{if } \overline{F}_i^n < 0, \end{cases} \quad \forall i \in \mathbb{Z}, \ \ \forall n \in [T],$$

and thus

$$G_i^n = u_{i-\frac{1}{2}}^n \left(\overline{F}_i^n\right)_+ + u_{i+\frac{1}{2}}^n \left(\overline{F}_i^n\right)_-, \ \ \forall i \in \mathbb{Z}, \ \ \forall n \in [T].$$

The computation of the discrete unknowns at each time step is completely explicit. First the discrete water heights $\left(h_i^{n+1}\right)_{i\in\mathbb{Z}}$ are computed with (dmc), then the discrete velocities $\left(u_{i+\frac{1}{2}}^{n+1}\right)_{i\in\mathbb{Z}}$ are computed with (dmb). Here is an important convention, which is a discrete version of the definition of Ω.

Convention 3.1 (Zero velocity of water in dry areas in the horizontal fluid domain $\mathbb{R}$)**.** *For all $i \in \mathbb{Z}$, $n \in [T]$, if $\bar{h}_{i+\frac{1}{2}}^n = 0$, by convention, $u_{i+\frac{1}{2}}^n$ is then set to zero.*

This convention is highly reasonable with the reality in the following obvious sense: "If there is no water somewhere, there must be also no velocity of water there!". Thanks to this convention, now we can define the initial velocity as

$$u_0(x) := \begin{cases} \dfrac{q_0(x)}{h_0(x)}, & \text{if } h_0(x) \neq 0, \\ 0, & \text{else}, \end{cases} \quad \forall x \in \mathbb{R},$$

and the discrete initial velocity as

$$u_{i+\frac{1}{2}}^0 := \frac{1}{\Delta x_{i+\frac{1}{2}}} \int_{C_{i+\frac{1}{2}}} u_0(x)\, dx, \ \ \forall i \in \mathbb{Z}.$$

To end this subsection, the staggered upwind scheme just established reads

$$\begin{cases} h_i^{n+1} := h_i^n - \nu_i \left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right), \\ \bar{h}_{i+\frac{1}{2}}^{n+1} u_{i+\frac{1}{2}}^{n+1} := \bar{h}_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n - \nu_{i+\frac{1}{2}} \left[G_{i+1}^n - G_i^n + \dfrac{g}{2}\left(\left(h_{i+1}^{n+1}\right)^2 - \left(h_i^{n+1}\right)^2\right)\right] \\ \qquad - g\nu_{i+\frac{1}{2}} \bar{h}_{i+\frac{1}{2}}^{n+1} \left(b_{i+1} - b_i\right), \end{cases} \tag{DGsc}$$

for all $i \in \mathbb{Z}$, $n \in [T-1]$.

In the flat bottom topography case, (DGsc) becomes

$$\begin{cases} h_i^{n+1} := h_i^n - \nu_i \left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right), \\ \bar{h}_{i+\frac{1}{2}}^{n+1} u_{i+\frac{1}{2}}^{n+1} := \bar{h}_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n - \nu_{i+\frac{1}{2}} \left[G_{i+1}^n - G_i^n + \dfrac{g}{2}\left(\left(h_{i+1}^{n+1}\right)^2 - \left(h_i^{n+1}\right)^2\right)\right], \end{cases} \tag{DGscfb}$$

for all $i \in \mathbb{Z}$, $n \in [T-1]$.

1.3 Conservation properties of DG staggered upwind schemes on the horizontal fluid domain

As for (3peccFVS) studied in the previous chapter, we have the following result involving some reference state for (DGscfb).

Lemma 3.1 (Reference states for (DGscfb))**.** *Assume that the initial data satisfies* $(h_0, q_0) \in (L^\infty(\mathbb{R}))^2$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step, and assume that the bottom topography is flat, i.e.,* $b = 0$ *in* $\mathbb{R}$ *and* (h_0, u_0)[2] *is equal to some reference state* $(h_\star, u_\star) \in \Omega$ *for all* $x \in (-\infty, M_L) \cup (M_R, \infty)$ *for some reals* $M_L \le M_R$*, i.e.,*

$$h_0(x) = h_\star, \;\; u_0(x) = u_\star, \;\; \forall x \in (-\infty, M_L) \cup (M_R, \infty).$$

Denote by i_{M_L} *and* i_{M_R} *the indices indicating which primal cells containing* M_L *and* M_R*, i.e.,*

$$M_L \in \left[x_{i_{M_L}-\frac{1}{2}}, x_{i_{M_L}+\frac{1}{2}}\right), \;\; M_R \in \left(x_{i_{M_R}-\frac{1}{2}}, x_{i_{M_R}+\frac{1}{2}}\right].$$

Let $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by the DG-staggered upwind scheme (DGscfb). Then it satisfies*

$$\begin{aligned} h_i^n &= h_\star, \;\; \forall i \in \mathbb{Z} \textit{ s.t. } \left(i < i_{M_L} - n\right) \vee \left(i > i_{M_R} + n\right), \\ u_{i+\frac{1}{2}}^n &= u_\star, \;\; \forall i \in \mathbb{Z} \textit{ s.t. } \left(i < i_{M_L} - n - 1\right) \vee \left(i > i_{M_R} + n\right). \end{aligned}$$

Proof. It is straightforward from the initial condition that

$$h_0|_{C_i}(x) = h_\star, \, u_0|_{C_i}(x) = u_\star, \;\; \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L}\right) \vee \left(i > i_{M_R}\right),$$

and consequently,

$$\begin{aligned} h_i^0 &= h_\star, \;\; \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L}\right) \vee \left(i > i_{M_R}\right), \\ u_{i+\frac{1}{2}}^0 &= u_\star, \;\; \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - 1\right) \vee \left(i > i_{M_R}\right). \end{aligned}$$

On one hand, the discrete mass conservation equation (dmc) in staggered upwind scheme (DGscfb) gives us at the initial time step $n = 0$:

$$\begin{aligned} h_i^1 :&= h_i^0 - \nu_i \left(h_{i+\frac{1}{2}}^0 u_{i+\frac{1}{2}}^0 - h_{i-\frac{1}{2}}^0 u_{i-\frac{1}{2}}^0\right) \\ &= h_i^0 - \nu_i \left[h_i^0 \left(u_{i+\frac{1}{2}}^0\right)_+ + h_{i+1}^0 \left(u_{i+\frac{1}{2}}^0\right)_- - h_{i-1}^0 \left(u_{i-\frac{1}{2}}^0\right)_+ - h_i^0 \left(u_{i-\frac{1}{2}}^0\right)_-\right] \\ &= h_\star - \nu_i \left(h_\star (u_\star)_+ + h_\star (u_\star)_- - h_\star (u_\star)_+ - h_\star (u_\star)_-\right) \\ &= h_\star, \;\; \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - 1\right) \vee \left(i > i_{M_R} + 1\right). \end{aligned}$$

[2] By Convention 3.1, we can write (h_0, u_0) instead of (h_0, q_0), where $q_0 =: h_0 u_0$ (not $q_0 := h_0 u_0$!).

On the other hand, the discrete momentum balance equation (dmb) with $b = 0$ in (DGscfb) gives us at the initial time step $n = 0$ for all $i \in \mathbb{Z}$:

$$\begin{aligned}\frac{1}{2}\left(h_i^1 + h_{i+1}^1\right) u_{i+\frac{1}{2}}^1 = \frac{1}{2}\left(h_i^0 + h_{i+1}^0\right) u_{i+\frac{1}{2}}^0 - \frac{g}{2}\nu_{i+\frac{1}{2}}\left(\left(h_{i+1}^1\right)^2 - \left(h_i^1\right)^2\right) \\ - \frac{1}{2}\nu_{i+\frac{1}{2}}\left[u_{i+1}^0\left(h_{i+\frac{1}{2}}^0 u_{i+\frac{1}{2}}^0 + h_{i+\frac{3}{2}}^0 u_{i+\frac{3}{2}}^0\right) - u_i^0\left(h_{i-\frac{1}{2}}^0 u_{i-\frac{1}{2}}^0 + h_{i+\frac{1}{2}}^0 u_{i+\frac{1}{2}}^0\right)\right],\end{aligned}$$

thus

$$\begin{aligned}h_\star u_{i+\frac{1}{2}}^1 &= h_\star u_\star - \frac{g}{2}\nu_{i+\frac{1}{2}}\left(h_\star^2 - h_\star^2\right) - \frac{1}{2}\nu_{i+\frac{1}{2}}\left[u_\star\left(h_\star u_\star + h_\star u_\star\right) - u_\star\left(h_\star u_\star + h_\star u_\star\right)\right] \\ &= h_\star u_\star, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - 2\right) \vee \left(i > i_{M_R} + 1\right).\end{aligned}$$

If $h_\star = 0$, then

$$h_{i+\frac{1}{2}}^1 := \frac{1}{2}\left(h_i^1 + h_{i+1}^1\right) = h_\star = 0, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - 2\right) \vee \left(i > i_{M_R} + 1\right),$$

and thus, by Convention (3.1), we have

$$u_{i+\frac{1}{2}}^1 = 0 = u_\star, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - 2\right) \vee \left(i > i_{M_R} + 1\right).$$

Hence, in both cases $h_\star = 0$ and $h_\star \neq 0$, we have

$$u_{i+\frac{1}{2}}^1 = u_\star, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - 2\right) \vee \left(i > i_{M_R} + 1\right).$$

Now suppose that for some $n \in [T-1]$,

$$\begin{aligned}h_i^n &= h_\star, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - n\right) \vee \left(i > i_{M_R} + n\right), \\ u_{i+\frac{1}{2}}^n &= u_\star, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - n - 1\right) \vee \left(i > i_{M_R} + n\right),\end{aligned}$$

we prove this for $n+1$. Indeed, similarly to the initial time step, the discrete mass conservation equation (dmc) in (DGscfb) gives us

$$\begin{aligned}h_i^{n+1} &= h_i^n - \nu_i\left(h_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n - h_{i-\frac{1}{2}}^n u_{i-\frac{1}{2}}^n\right) \\ &= h_i^n - \nu_i\left[h_i^n\left(u_{i+\frac{1}{2}}^n\right)_+ + h_{i+1}^n\left(u_{i+\frac{1}{2}}^n\right)_- - h_{i-1}^n\left(u_{i-\frac{1}{2}}^n\right)_+ - h_i^n\left(u_{i-\frac{1}{2}}^n\right)_-\right] \\ &= h_\star - \nu_i\left[h_\star(u_\star)_+ + h_\star(u_\star)_- - h_\star(u_\star)_+ - h_\star(u_\star)_-\right] \\ &= h_\star, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - n - 1\right) \vee \left(i > i_{M_R} + n + 1\right).\end{aligned}$$

Furthermore, the discrete momentum balance equation (dmb) with $b = 0$ in (DGscfb) gives us for all $i \in \mathbb{Z}$:

$$\begin{aligned}\frac{1}{2}\left(h_i^{n+1} + h_{i+1}^{n+1}\right) u_{i+\frac{1}{2}}^{n+1} = \frac{1}{2}\left(h_i^n + h_{i+1}^n\right) u_{i+\frac{1}{2}}^n - \frac{g}{2}\nu_{i+\frac{1}{2}}\left(\left(h_{i+1}^{n+1}\right)^2 - \left(h_i^{n+1}\right)^2\right) \\ - \frac{1}{2}\nu_{i+\frac{1}{2}}\left[u_{i+1}^n\left(h_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n + h_{i+\frac{3}{2}}^n u_{i+\frac{3}{2}}^n\right) - u_i^n\left(h_{i-\frac{1}{2}}^n u_{i-\frac{1}{2}}^n + h_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n\right)\right],\end{aligned}$$

thus

$$\begin{aligned} h_\star u_{i+\frac{1}{2}}^{n+1} &= h_\star u_\star - \frac{g}{2}\nu_{i+\frac{1}{2}}\left(h_\star^2 - h_\star^2\right) - \frac{1}{2}\nu_{i+\frac{1}{2}}\left[u_\star\left(h_\star u_\star + h_\star u_\star\right) - u_\star\left(h_\star u_\star + h_\star u_\star\right)\right] \\ &= h_\star u_\star, \ \ \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - n - 2\right) \vee \left(i > i_{M_R} + n + 1\right). \end{aligned}$$

Consider the cases $h_\star = 0$ and $h_\star \neq 0$ as in the initial time step, we obtain

$$u_{i+\frac{1}{2}}^{n+1} = u_\star, \ \ \forall i \in \mathbb{Z} \text{ s.t. } \left(i < i_{M_L} - n - 2\right) \vee \left(i > i_{M_R} + n + 1\right).$$

The desired result then follows by the principle of mathematical induction. □

Recall that the variational structure of the water waves system (ww) preserves some physical quantities in the continuous level, same for Saint-Venant system (SV). The following result demonstrates that the staggered upwind scheme (DGsc) preserves some of those physical quantities in the discrete level.

Proposition 3.2 (Preserved quantities under (DGsc))**.** *Assume that the initial data satisfies* $(h_0, q_0) \in \left(L^\infty(\mathbb{R})\right)^2$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by* (DGsc)*. The staggered upwind scheme* (DGsc) *preserves the total mass in the discrete level, with a suitable notion of 1D discrete total mass. Moreover, for the flat bottom topography case,* (DGscfb) *also preserves the total momentum in the discrete level, with a suitable notion of 1D discrete total momentum.*

Proof. Recall that some quantities preserved by the water wave system are listed in Chapter 1, Section 1.5, we now define their corresponding discrete counterparts and prove that (DGsc) preserves them in the discrete level.

1. *Total mass.* Recall that in Chapter 1, Section 1.5, the 1D total mass is defined by

$$\mathcal{Z}(t) := \int_{\mathbb{R}} \zeta(t,x)\,dx = \int_{\mathbb{R}} \left(h(t,x) + b(x) - H_0\right)dx, \ \ \forall t \in [0, T_0),$$

this quantity is preserved in time:

$$\frac{d\mathcal{Z}(t)}{dt} = 0, \ \ \forall t \in [0, T_0). \tag{mc}$$

Now we prove (mc) in the discrete level.

First, the corresponding 1D discrete total mass is defined by

$$\mathcal{Z}^n := \sum_{i\in\mathbb{Z}} \Delta x_i \left(h_i^n + b_i - H_0\right), \ \ \forall n \in [T].$$

We claim that $\mathcal{Z}^{n+1} = \mathcal{Z}^n$ for all $n \in [T-1]$. Indeed, for all $n \in [T-1]$,

$$\mathcal{Z}^{n+1} = \sum_{i\in\mathbb{Z}} \Delta x_i h_i^{n+1} + \sum_{i\in\mathbb{Z}} \Delta x_i \left(b_i - H_0\right) = \sum_{i\in\mathbb{Z}} \Delta x_i h_i^n + \sum_{i\in\mathbb{Z}} \Delta x_i \left(b_i - H_0\right) = \mathcal{Z}^n,$$

where the second equality is deduced from Proposition (3.1). As a consequence, the initial total mass is preserved in time:

$$\mathcal{Z}^n = \mathcal{Z}^0, \ \ \forall n \in [T],$$

i.e., the discrete version of (mc).

2. *Total momentum (in the flat bottom topography case $b = 0$).* Recall that the 1D total momentum is defined by

$$\mathcal{M}(t) := \int_{\mathbb{R}} (hu)(t,x)\, dx, \ \ \forall t \in [0, T_0),$$

this quantity is preserved in time:

$$\frac{d\mathcal{M}(t)}{dt} = 0, \ \ \forall t \in [0, T_0). \tag{Mc}$$

We also prove (Mc) in the discrete level.

The corresponding 1D discrete total momentum is defined by

$$\mathcal{M}^n := \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}}, \ \ \forall n \in [T],$$

we claim that $\mathcal{M}^{n+1} = \mathcal{M}^n$ for all $n \in [T-1]$ provided that the bottom topography is flat. Assume $b = 0$, then $b_i = 0$ for all $i \in \mathbb{Z}$ and then

$$\begin{aligned}
\mathcal{M}^{n+1} &= \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} \bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} \\
&= \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} \left[\bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - \nu_{i+\frac{1}{2}} \left[G^n_{i+1} - G^n_i + \frac{g}{2}\left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2\right)\right]\right] \\
&= \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - \Delta t \sum_{i\in\mathbb{Z}} \left(G^n_{i+1} - G^n_i\right) + \frac{g\Delta t}{2} \sum_{i\in\mathbb{Z}} \left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2\right) \\
&= \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} = \mathcal{M}^n, \ \ \forall n \in [T-1].
\end{aligned}$$

Consequently, the initial total momentum is preserved in time:

$$\mathcal{M}^n = \mathcal{M}^0, \ \ \forall n \in [T].$$

i.e., the discrete version of (Mc).

This completes our proof. □

Remark 3.1 (Non-conservation of the total horizontal impulse and total energy)**.** *Recall that the 1D horizontal impulse is defined by*

$$\mathcal{I}(t) := \int_{\mathbb{R}} (\zeta\nabla\psi)(t,x)\, dx = \int_{\mathbb{R}} (h(t,x) + b(x) - H_0)\, u(t,x)\, dx, \ \ \forall t \in [0, T_0),$$

this quantity is preserved in time for Saint-Venant system (SV) in the flat bottom topography case $b = 0$:

$$\frac{d\mathcal{I}(t)}{dt} = 0, \quad \forall t \in [0, T_0) . \tag{hic}$$

Unfortunately, it is hard to check whether (DGscfb) preserve the horizontal impulse in the discrete level or not. The most difficulty here is to define a suitable notion of 1D discrete horizontal impulse to make use of both equations in (DGscfb). For instance, consider the following 1D discrete horizontal impulse

$$\mathcal{I}^n := \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} \left(\bar{h}^n_{i+\frac{1}{2}} + b_i - H_0\right) u^n_{i+\frac{1}{2}}, \quad \forall n \in [T] , \tag{dhi}$$

we claim that $\mathcal{I}^{n+1}$ can be different to $\mathcal{I}^n$ in general.

Indeed, assume $b = 0$, inserting the discrete mass conservation equation (dmc) into the discrete momentum balance equation (dmb) yields for all $i \in \mathbb{Z}$, $n \in [T]$:

$$\begin{aligned} &\frac{1}{2}\left(h^n_i + h^n_{i+1}\right) u^{n+1}_{i+\frac{1}{2}} - \frac{1}{2}\left[\nu_i\left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\right) + \nu_{i+1}\left(F^n_{i+\frac{3}{2}} - F^n_{i+\frac{1}{2}}\right)\right] u^{n+1}_{i+\frac{1}{2}} \\ &\quad = \frac{1}{2}\left(h^n_i + h^n_{i+1}\right) u^n_{i+\frac{1}{2}} - \nu_{i+\frac{1}{2}}\left[G^n_{i+1} - G^n_i + \frac{g}{2}\left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2\right)\right] . \end{aligned}$$

Assume that for some $n_0 \in [T]$ the horizontal fluid domain $\mathbb{R}$ is fully wet at $t = t^{n_0}$ in the discrete level, i.e., $h^{n_0}_i > 0$ for all $i \in \mathbb{Z}$, and the last identity at $n = n_0$ reads

$$\begin{aligned} u^{n_0+1}_{i+\frac{1}{2}} &= u^{n_0}_{i+\frac{1}{2}} - \frac{2}{h^{n_0}_i + h^{n_0}_{i+1}} \nu_{i+\frac{1}{2}}\left[G^{n_0}_{i+1} - G^{n_0}_i + \frac{g}{2}\left(\left(h^{n_0+1}_{i+1}\right)^2 - \left(h^{n_0+1}_i\right)^2\right)\right] \\ &\quad + \frac{1}{h^{n_0}_i + h^{n_0}_{i+1}}\left[\nu_i\left(F^{n_0}_{i+\frac{1}{2}} - F^{n_0}_{i-\frac{1}{2}}\right) + \nu_{i+1}\left(F^{n_0}_{i+\frac{3}{2}} - F^{n_0}_{i+\frac{1}{2}}\right)\right] u^{n_0+1}_{i+\frac{1}{2}}, \quad \forall i \in \mathbb{Z}, \end{aligned}$$

or equivalently,

$$\begin{aligned} \Delta x_{i+\frac{1}{2}} u^{n_0+1}_{i+\frac{1}{2}} &= \Delta x_{i+\frac{1}{2}} u^{n_0}_{i+\frac{1}{2}} - \frac{2\Delta t}{h^{n_0}_i + h^{n_0}_{i+1}}\left[G^{n_0}_{i+1} - G^{n_0}_i + \frac{g}{2}\left(\left(h^{n_0+1}_{i+1}\right)^2 - \left(h^{n_0+1}_i\right)^2\right)\right] \\ &+ \frac{1}{h^{n_0}_i + h^{n_0}_{i+1}} \Delta x_{i+\frac{1}{2}}\left[\nu_i\left(F^{n_0}_{i+\frac{1}{2}} - F^{n_0}_{i-\frac{1}{2}}\right) + \nu_{i+1}\left(F^{n_0}_{i+\frac{3}{2}} - F^{n_0}_{i+\frac{1}{2}}\right)\right] u^{n_0+1}_{i+\frac{1}{2}}, \quad \forall i \in \mathbb{Z}. \end{aligned}$$

Hence,

$$\begin{aligned} &\mathcal{I}^{n_0+1} \\ &= \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}}\left(\bar{h}^{n_0+1}_{i+\frac{1}{2}} - H_0\right) u^{n_0+1}_{i+\frac{1}{2}} \\ &= \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} \bar{h}^{n_0+1}_{i+\frac{1}{2}} u^{n_0+1}_{i+\frac{1}{2}} - H_0 \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} u^{n_0+1}_{i+\frac{1}{2}} \\ &= \mathcal{M}^{n_0+1} - H_0 \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} u^{n_0}_{i+\frac{1}{2}} \\ &\quad + 2H_0\Delta t \sum_{i\in\mathbb{Z}} \frac{1}{h^{n_0}_i + h^{n_0}_{i+1}}\left[G^{n_0}_{i+1} - G^{n_0}_i + \frac{g}{2}\left(\left(h^{n_0+1}_{i+1}\right)^2 - \left(h^{n_0+1}_i\right)^2\right)\right] \end{aligned}$$

$$
\begin{aligned}
&- H_0 \sum_{i\in\mathbb{Z}} \frac{1}{h_i^{n_0}+h_{i+1}^{n_0}} \Delta x_{i+\frac{1}{2}} \left[\nu_i \left(F_{i+\frac{1}{2}}^{n_0} - F_{i-\frac{1}{2}}^{n_0}\right) + \nu_{i+1}\left(F_{i+\frac{3}{2}}^{n_0} - F_{i+\frac{1}{2}}^{n_0}\right)\right] u_{i+\frac{1}{2}}^{n_0+1} \\
= {}& \mathcal{M}^{n_0} - H_0 \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}} u_{i+\frac{1}{2}}^{n_0} \\
&+ 2H_0\Delta t \sum_{i\in\mathbb{Z}} \frac{1}{h_i^{n_0}+h_{i+1}^{n_0}} \left[G_{i+1}^{n_0} - G_i^{n_0} + \frac{g}{2}\left(\left(h_{i+1}^{n_0+1}\right)^2 - \left(h_i^{n_0+1}\right)^2\right)\right] \\
&- H_0 \sum_{i\in\mathbb{Z}} \frac{1}{h_i^{n_0}+h_{i+1}^{n_0}} \Delta x_{i+\frac{1}{2}} \left[\nu_i \left(F_{i+\frac{1}{2}}^{n_0} - F_{i-\frac{1}{2}}^{n_0}\right) + \nu_{i+1}\left(F_{i+\frac{3}{2}}^{n_0} - F_{i+\frac{1}{2}}^{n_0}\right)\right] u_{i+\frac{1}{2}}^{n_0+1} \\
= {}& \mathcal{I}^{n_0} + 2H_0\Delta t \sum_{i\in\mathbb{Z}} \frac{1}{h_i^{n_0}+h_{i+1}^{n_0}} \left[G_{i+1}^{n_0} - G_i^{n_0} + \frac{g}{2}\left(\left(h_{i+1}^{n_0+1}\right)^2 - \left(h_i^{n_0+1}\right)^2\right)\right] \\
&- H_0 \sum_{i\in\mathbb{Z}} \frac{1}{h_i^{n_0}+h_{i+1}^{n_0}} \Delta x_{i+\frac{1}{2}} \left[\nu_i \left(F_{i+\frac{1}{2}}^{n_0} - F_{i-\frac{1}{2}}^{n_0}\right) + \nu_{i+1}\left(F_{i+\frac{3}{2}}^{n_0} - F_{i+\frac{1}{2}}^{n_0}\right)\right] u_{i+\frac{1}{2}}^{n_0+1}.
\end{aligned}
$$

Even if we assume that the water is at rest at $t = t^{n_0+1}$ in the discrete level, i.e., $u_{i+\frac{1}{2}}^{n_0+1} = 0$ for all $i \in \mathbb{Z}$, then the last identity becomes

$$\mathcal{I}^{n_0+1} = \mathcal{I}^{n_0} + 2H_0\Delta t \sum_{i\in\mathbb{Z}} \frac{1}{h_i^{n_0}+h_{i+1}^{n_0}} \left[G_{i+1}^{n_0} - G_i^{n_0} + \frac{g}{2}\left(\left(h_{i+1}^{n_0+1}\right)^2 - \left(h_i^{n_0+1}\right)^2\right)\right],$$

and thus the remaining sum (equal to $\mathcal{I}^{n_0+1} - \mathcal{I}^{n_0}$) is still very complicated to handle. In a worse case, the water is not at rest at $t = t^{n_0+1}$ in the discrete level, the two remaining sums prevent us from proving that (DGscfb) preserves the total horizontal impulse in the discrete level. We also start to believe that there are some counterexamples to prove that the horizontal impulse is not preserved under (DGscfb) in the discrete level, or at least with this notion of 1D discrete horizontal impulse. Actually, other notions with different choices of indices maybe result in these calculations some nightmares though.

Furthermore, it is very far from clear whether (DGscfb) does preserve the total energy in the discrete level or not. Recall that the total energy for dD Saint-Venant, where $d = 1$ or $d = 2$, is defined in Duchêne, 2019, pp. 20–21 by

$$\mathcal{H}_{\mathrm{SV}}(\zeta,\psi)(t) := \frac{1}{2}\int_{\mathbb{R}^d} \left(\zeta^2(t,x) + (H_0 + \zeta(t,x) - b(x))\left|\nabla\psi(t,x)\right|^2\right) dx, \quad \forall t \in [0, T_0),$$

the total energy for (1DSW) is defined similarly by

$$\mathcal{H}_{\mathrm{SW}}(h,u)(t) := \frac{1}{2}\int_{\mathbb{R}} \left((h(t,x) + b(x) - H_0)^2 + \left(hu^2\right)(t,x)\right) dx, \quad \forall t \in [0, T_0).$$

The corresponding 1D discrete total energy can be defined for all $n \in [T]$ by

$$\mathcal{H}_{\mathrm{SW},\alpha,\beta,\gamma,\delta}^{n} := \frac{1}{2}\sum_{i\in\mathbb{Z}} \left[\Delta x_{i+\alpha}\left(h_{i+\alpha}^n + b_i - H_0\right)^2 + \Delta x_{i+\beta} h_{i+\gamma}^n \left(u_{i+\delta}^n\right)^2\right], \qquad (\mathrm{dte}_1)$$

alternatively,

$$\overline{\mathcal{H}}^n_{\mathrm{SW},\alpha,\beta,\gamma,\delta} := \frac{1}{2}\sum_{i\in\mathbb{Z}}\left[\Delta x_{i+\alpha}\left(\bar{h}^n_{i+\frac{1}{2}} + b_i - H_0\right)^2 + \Delta x_{i+\beta}\bar{h}^n_{i+\frac{1}{2}}\left(u^n_{i+\gamma}\right)^2\right], \qquad (\mathrm{dte}_2)$$

where the indices $\alpha, \beta, \gamma, \delta \in \left\{0, \frac{1}{2}\right\}$. The problem of choosing "suitable" indices in the notion of the 1D discrete total energy to utilize (DGscfb) *is immensely delicate. No matter how "suitable" one chooses these indices, the remaining sums being equal to $\mathcal{H}^{n+1}_{\mathrm{SW}} - \mathcal{H}^n_{\mathrm{SW}}$ (resp., $\mathfrak{H}^{n+1}_{\mathrm{SW}} - \mathfrak{H}^n_{\mathrm{SW}}$) generated by* (DGsc) *or* (DGscfb) *in the flat bottom case is massive!*

These calculations motivate the following non-conservative result.

Proposition 3.3 (Non-preserved quantities under (DGscfb))**.** *Assume that the initial data satisfies $(h_0, q_0) \in (L^\infty(\mathbb{R}))^2$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}^\star_+$ be the time step. Let $\left(h^n_i, u^n_{i+\frac{1}{2}}\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by* (DGscfb)*. The staggered upwind scheme* (DGscfb)*, in the flat bottom topography case, does not preserve the total horizontal impulse and also the total energy in the discrete level with the choices of the discrete horizontal impulse* (dhi) *and the discrete total energy* (dte_1) *or* (dte_2)*.*

Proof. It suffices to build a counterexample for the horizontal impulse and total energy.

For the sake of simplicity, we consider (DGscfb) on a uniform mesh, i.e.,

$$\Delta x_i = \Delta x_{i+\frac{1}{2}} = \Delta x, \quad \nu_i = \nu, \quad \forall i \in \mathbb{Z}.$$

Assume $b = 0$, we consider (DGscfb) with the following initial conditions

$$h_0 \text{ is a mollifier s.t. } \mathrm{Supp}\,(h_0) \subset \overline{C_0}, \int_{\mathbb{R}} h_0\,(x)\,dx = 1,$$

thus

$$h^0_0 = \int_{C_0} h_0\,(x)\,dx = 1, \quad h^0_i = 0, \quad \forall i \in \mathbb{Z}^\star,$$

and thus

$$\bar{h}^0_{-\frac{1}{2}} = \bar{h}^0_{\frac{1}{2}} = \frac{1}{2}, \quad \bar{h}^0_{i+\frac{1}{2}} = 0, \quad \forall i \in \mathbb{Z}^\star \setminus \{-1\}.$$

Convention (3.1) yields that

$$u^0_{i+\frac{1}{2}} = 0, \quad \forall i \in \mathbb{Z}^\star \setminus \{-1\}.$$

Put $\overline{u} := u^0_{-\frac{1}{2}}$ and $\overline{\overline{u}} := u^0_{\frac{1}{2}}$, the initial 1D discrete horizontal impulse is given by

$$\begin{aligned}\mathcal{I}^0 &= \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}}\left(\bar{h}^0_{i+\frac{1}{2}} - H_0\right)u^0_{i+\frac{1}{2}} = \Delta x\left(\bar{h}^0_{-\frac{1}{2}} - H_0\right)u^0_{-\frac{1}{2}} + \Delta x\left(\bar{h}^0_{\frac{1}{2}} - H_0\right)u^0_{\frac{1}{2}} \\ &= \Delta x\left(\frac{1}{2} - H_0\right)\left(\overline{u} + \overline{\overline{u}}\right).\end{aligned}$$

It suffices to compute $\mathcal{I}^1$ and prove that $\mathcal{I}^0 \neq \mathcal{I}^1$.

Applying Lemma 3.1 for the uniform mesh $\mathcal{T}$, $h_\star = u_\star = 0$, $M_L = x_{-\frac{1}{2}}$, $M_R = x_{\frac{1}{2}}$, $i_{M_L} = 0$, $i_{M_R} = 0$ yields that with the current initial condition, (DGscfb) generates the sequence $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ satisfying

$$\begin{aligned} h_i^n &= 0, \quad \forall i \in \mathbb{Z} \text{ s.t. } \ (i < -n) \vee (i > n)\,, \\ u_{i+\frac{1}{2}}^n &= 0, \quad \forall i \in \mathbb{Z} \text{ s.t. } \ (i < -n-1) \vee (i > n)\,. \end{aligned}$$

In particular, for $n = 1$,

$$\begin{aligned} h_i^1 &= 0, \quad \forall i \in \mathbb{Z} \text{ s.t. } \ (i < -1) \vee (i > 1)\,, \\ u_{i+\frac{1}{2}}^1 &= 0, \quad \forall i \in \mathbb{Z} \text{ s.t. } \ (i < -2) \vee (i > 1)\,. \end{aligned}$$

Hence, we only need to compute h_{-1}^1, h_0^1, h_1^1 and $u_{-\frac{3}{2}}^1$, $u_{-\frac{1}{2}}^1$, $u_{\frac{1}{2}}^1$, $u_{\frac{3}{2}}^1$. To do this, on one hand, (dmc) at $n = 0$ gives us for all $i \in \mathbb{Z}$:

$$h_i^1 := h_i^0 - \nu\left[h_i^0\left(u_{i+\frac{1}{2}}^0\right)_+ + h_{i+1}^0\left(u_{i+\frac{1}{2}}^0\right)_- - h_{i-1}^0\left(u_{i-\frac{1}{2}}^0\right)_+ - h_i^0\left(u_{i-\frac{1}{2}}^0\right)_-\right].$$

Hence,

$$\begin{aligned} h_{-1}^1 :&= h_{-1}^0 - \nu\left[h_{-1}^0\left(u_{-\frac{1}{2}}^0\right)_+ + h_0^0\left(u_{-\frac{1}{2}}^0\right)_- - h_{-2}^0\left(u_{-\frac{3}{2}}^0\right)_+ - h_{-1}^0\left(u_{-\frac{3}{2}}^0\right)_-\right] = -\nu\overline{u}_-, \\ h_0^1 :&= h_0^0 - \nu\left[h_0^0\left(u_{\frac{1}{2}}^0\right)_+ + h_1^0\left(u_{\frac{1}{2}}^0\right)_- - h_{-1}^0\left(u_{-\frac{1}{2}}^0\right)_+ - h_0^0\left(u_{-\frac{1}{2}}^0\right)_-\right] = 1 - \nu\left(\overline{\overline{u}}_+ - \overline{u}_-\right), \\ h_1^1 :&= h_1^0 - \nu\left[h_1^0\left(u_{\frac{3}{2}}^0\right)_+ + h_2^0\left(u_{\frac{3}{2}}^0\right)_- - h_0^0\left(u_{\frac{1}{2}}^0\right)_+ - h_1^0\left(u_{\frac{1}{2}}^0\right)_-\right] = \nu\overline{\overline{u}}_+. \end{aligned}$$

On the other hand, (dmb) with $b = 0$ at $n = 0$ gives us for all $i \in \mathbb{Z}$:

$$\begin{aligned} \frac{1}{2}\left(h_{-2}^1 + h_{-1}^1\right)u_{-\frac{3}{2}}^1 &= \frac{1}{2}\left(h_{-2}^0 + h_{-1}^0\right)u_{-\frac{3}{2}}^0 - \frac{g}{2}\nu\left[\left(h_{-1}^1\right)^2 - \left(h_{-2}^1\right)^2\right] \\ &\quad - \frac{1}{2}\nu\left[u_{-1}^0\left(h_{-\frac{3}{2}}^0 u_{-\frac{3}{2}}^0 + h_{-\frac{1}{2}}^0 u_{-\frac{1}{2}}^0\right) - u_{-2}^0\left(h_{-\frac{5}{2}}^0 u_{-\frac{5}{2}}^0 + h_{-\frac{3}{2}}^0 u_{-\frac{3}{2}}^0\right)\right], \\ \frac{1}{2}\left(h_{-1}^1 + h_0^1\right)u_{-\frac{1}{2}}^1 &= \frac{1}{2}\left(h_{-1}^0 + h_0^0\right)u_{-\frac{1}{2}}^0 - \frac{g}{2}\nu\left[\left(h_0^1\right)^2 - \left(h_{-1}^1\right)^2\right] \\ &\quad - \frac{1}{2}\nu\left[u_0^0\left(h_{-\frac{1}{2}}^0 u_{-\frac{1}{2}}^0 + h_{\frac{1}{2}}^0 u_{\frac{1}{2}}^0\right) - u_{-1}^0\left(h_{-\frac{3}{2}}^0 u_{-\frac{3}{2}}^0 + h_{-\frac{1}{2}}^0 u_{-\frac{1}{2}}^0\right)\right], \\ \frac{1}{2}\left(h_0^1 + h_1^1\right)u_{\frac{1}{2}}^1 &= \frac{1}{2}\left(h_0^0 + h_1^0\right)u_{\frac{1}{2}}^0 - \frac{g}{2}\nu\left[\left(h_1^1\right)^2 - \left(h_0^1\right)^2\right] \\ &\quad - \frac{1}{2}\nu\left[u_1^0\left(h_{\frac{1}{2}}^0 u_{\frac{1}{2}}^0 + h_{\frac{3}{2}}^0 u_{\frac{3}{2}}^0\right) - u_0^0\left(h_{-\frac{1}{2}}^0 u_{-\frac{1}{2}}^0 + h_{\frac{1}{2}}^0 u_{\frac{1}{2}}^0\right)\right], \\ \frac{1}{2}\left(h_1^1 + h_2^1\right)u_{\frac{3}{2}}^1 &= \frac{1}{2}\left(h_1^0 + h_2^0\right)u_{\frac{3}{2}}^0 - \frac{g}{2}\nu\left[\left(h_2^1\right)^2 - \left(h_1^1\right)^2\right] \\ &\quad - \frac{1}{2}\nu\left[u_2^0\left(h_{\frac{3}{2}}^0 u_{\frac{3}{2}}^0 + h_{\frac{5}{2}}^0 u_{\frac{5}{2}}^0\right) - u_1^0\left(h_{\frac{1}{2}}^0 u_{\frac{1}{2}}^0 + h_{\frac{3}{2}}^0 u_{\frac{3}{2}}^0\right)\right], \end{aligned}$$

or equivalently,

$$-\frac{\nu\overline{u}_-}{2}u_{-\frac{3}{2}}^1 = -\frac{g}{2}\nu^3(\overline{u}_-)^2 - \frac{1}{2}\nu\overline{uu}_-,$$

$$
\begin{aligned}
\frac{1-\nu\overline{\overline{u}}_+}{2}u^1_{-\frac{1}{2}} &= \frac{1}{2}\overline{u} - \frac{g}{2}\nu\left[\left(1-\nu\overline{\overline{u}}_+ + \nu\overline{u}_-\right)^2 - (\nu\overline{u}_-)^2\right] \\
&\quad - \frac{1}{2}\nu\left[\overline{u}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_+ + \overline{\overline{u}}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_- - \overline{uu}_-\right], \\
\frac{1+\nu\overline{u}_-}{2}u^1_{\frac{1}{2}} &= \frac{1}{2}\overline{\overline{u}} - \frac{g}{2}\nu\left[\left(\nu\overline{\overline{u}}_+\right)^2 - \left(1-\nu\overline{\overline{u}}_+ + \nu\overline{u}_-\right)^2\right] \\
&\quad - \frac{1}{2}\nu\left[\overline{\overline{uu}}_+ - \overline{u}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_- - \overline{\overline{u}}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_+\right], \\
\frac{\nu\overline{\overline{u}}_+}{2}u^1_{\frac{3}{2}} &= \frac{g}{2}\nu^3\left(\overline{\overline{u}}_+\right)^2 + \frac{1}{2}\nu\overline{\overline{uu}}_+.
\end{aligned}
$$

Now assume that $\left(\nu, \overline{u}, \overline{\overline{u}}\right) \in \left(\mathbb{R}^\star_+\right)^3$ satisfies

$$1-\nu\overline{\overline{u}}_+ \neq 0, \;\; 1+v\overline{u}_- \neq 0,$$

so that the coefficients of $u^1_{-\frac{1}{2}}$ and $u^1_{\frac{1}{2}}$ are nonzero. Then

$$
\begin{aligned}
u^1_{-\frac{3}{2}} &= g\nu^2\overline{u}_- + \overline{u}, \\
u^1_{-\frac{1}{2}} &= \frac{\overline{u}}{1-\nu\overline{\overline{u}}_+} - g\nu\left(1-\nu\overline{\overline{u}}_+ + 2\nu\overline{u}_-\right) - \frac{\nu}{1-\nu\overline{\overline{u}}_+}\left[\overline{u}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_+ + \overline{\overline{u}}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_- - \overline{uu}_-\right], \\
u^1_{\frac{1}{2}} &= \frac{\overline{\overline{u}}}{1+\nu\overline{u}_-} + g\nu\left(1-2\nu\overline{\overline{u}}_+ + \nu\overline{u}_-\right) - \frac{\nu}{1+\nu\overline{u}_-}\left[\overline{\overline{uu}}_+ - \overline{u}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_- - \overline{\overline{u}}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_+\right], \\
u^1_{\frac{3}{2}} &= g\nu^2\overline{\overline{u}}_+ + \overline{\overline{u}}.
\end{aligned}
$$

Now we have enough data to compute the 1D discrete horizontal impulse at $t = t^1$:

$$
\begin{aligned}
\mathcal{I}^1 :&= \sum_{i\in\mathbb{Z}} \Delta x_{i+\frac{1}{2}}\left(\bar{h}^1_{i+\frac{1}{2}} - H_0\right)u^1_{i+\frac{1}{2}} \\
&= \Delta x_{-\frac{3}{2}}\left(\bar{h}^1_{-\frac{3}{2}} - H_0\right)u^1_{-\frac{3}{2}} + \Delta x_{-\frac{1}{2}}\left(\bar{h}^1_{-\frac{1}{2}} - H_0\right)u^1_{-\frac{1}{2}} \\
&\quad + \Delta x_{\frac{1}{2}}\left(\bar{h}^1_{\frac{1}{2}} - H_0\right)u^1_{\frac{1}{2}} + \Delta x_{\frac{3}{2}}\left(\bar{h}^1_{\frac{3}{2}} - H_0\right)u^1_{\frac{3}{2}} \\
&= \Delta x\left(-\frac{\nu\overline{u}_-}{2} - H_0\right)u^1_{-\frac{3}{2}} + \Delta x\left(\frac{1-\nu\overline{\overline{u}}_+}{2} - H_0\right)u^1_{-\frac{1}{2}} \\
&\quad + \Delta x\left(\frac{1+\nu\overline{u}_-}{2} - H_0\right)u^1_{\frac{1}{2}} + \Delta x\left(\frac{\nu\overline{\overline{u}}_+}{2} - H_0\right)u^1_{\frac{3}{2}} \\
&= \frac{\Delta x}{2}\left(-\nu\overline{u}_- u^1_{-\frac{3}{2}} + \left(1-\nu\overline{\overline{u}}_+\right)u^1_{-\frac{1}{2}} + \left(1+\nu\overline{u}_-\right)u^1_{\frac{1}{2}} + \nu\overline{\overline{u}}_+ u^1_{\frac{3}{2}}\right) \\
&\quad - H_0\Delta x\left(u^1_{-\frac{3}{2}} + u^1_{-\frac{1}{2}} + u^1_{\frac{1}{2}} + u^1_{\frac{3}{2}}\right) \\
&= \frac{\Delta x}{2}\begin{pmatrix} -g\nu^3(\overline{u}_-)^2 - \nu\overline{uu}_- + \overline{u} - g\nu\left[\left(1-\nu\overline{\overline{u}}_+ + \nu\overline{u}_-\right)^2 - (\nu\overline{u}_-)^2\right] \\ -\nu\left[\overline{u}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_+ + \overline{\overline{u}}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_- - \overline{uu}_-\right] \\ +\overline{\overline{u}} - g\nu\left[\left(\nu\overline{\overline{u}}_+\right)^2 - \left(1-\nu\overline{\overline{u}}_+ + \nu\overline{u}_-\right)^2\right] \\ -\nu\left[\overline{\overline{uu}}_+ - \overline{u}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_- - \overline{\overline{u}}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_+\right] + g\nu^3\left(\overline{\overline{u}}_+\right)^2 + \nu\overline{\overline{uu}}_+ \end{pmatrix}
\end{aligned}
$$

$$
- H_0\Delta x \begin{pmatrix} g\nu^2\overline{u}_- + \overline{u} + \frac{\overline{u}}{1-\nu\overline{\overline{u}}_+} - g\nu\left(1 - \nu\overline{\overline{u}}_+ + 2\nu\overline{u}_-\right) \\ -\frac{\nu}{1-\nu\overline{\overline{u}}_+}\left[\overline{u}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_+ + \overline{\overline{u}}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_- - \overline{uu}_-\right] \\ +g\nu^2\overline{\overline{u}}_+ + \overline{\overline{u}} + \frac{\overline{\overline{u}}}{1+\nu\overline{u}_-} + g\nu\left(1 - 2\nu\overline{\overline{u}}_+ + \nu\overline{u}_-\right) \\ -\frac{\nu}{1+\nu\overline{u}_-}\left[\overline{\overline{u}\,\overline{u}}_+ - \overline{u}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_- - \overline{\overline{u}}\left(\overline{u}_- + \overline{\overline{u}}_+\right)_+\right] \end{pmatrix}.
$$

Assume even more that $\overline{u} = \overline{\overline{u}} = \alpha > 0$, then the initial discrete horizontal impulse becomes

$$
\mathcal{I}^0 = \alpha\Delta x\left(1 - 2H_0\right),
$$

and the next discrete horizontal impulse $\mathcal{I}^1$ can be rewritten as

$$
\begin{aligned}
\mathcal{I}^1 = \frac{\Delta x}{2} & \begin{pmatrix} \alpha - g\nu(1-\nu\alpha)^2 - \nu\alpha^2 + \alpha - g\nu\left[(\nu\alpha)^2 - (1-\nu\alpha)^2\right] \\ -\nu\left(\alpha^2 - \alpha^2\right) + g\nu^3\alpha^2 + \nu\alpha^2 \end{pmatrix} \\
& - H_0\Delta x \begin{pmatrix} \alpha + \dfrac{\alpha}{1-\nu\alpha} - g\nu\left(1-\nu\alpha\right) - \dfrac{\nu\alpha^2}{1-\nu\alpha} + g\nu^2\alpha + \alpha \\ +\alpha + g\nu\left(1 - 2\nu\alpha\right) - \nu\left(\alpha^2 - \alpha^2\right) \end{pmatrix} \\
& = \alpha\Delta x\left(1 - 4H_0\right).
\end{aligned}
$$

Hence, $\mathcal{I}_0 \neq \mathcal{I}^1$ provided $H_0 \neq 0$. One can choose another triple $\left(\mathcal{T}, \overline{u}, \overline{\overline{u}}\right)$ to show this. We just need to keep in mind that the remaining being equal to $\mathcal{I}^1 - \mathcal{I}^0$ is complicated enough to be far from zero.

This counterexample also works with the discrete total energy. But the amount of calculations is massive, so its proof is omitted here. And our proof can end now. □

Applying the Finite Volume method for mass conservation equation in (1DSW) with the corresponding numerical flux $g^n_{i+\frac{1}{2}} : \mathbb{R}^2_+ \to \mathbb{R}$, for all $i \in \mathbb{Z}$, $n \in [T]$, defined piecewisely on each double dual cells by

$$
g^n_{i+\frac{1}{2}}\left(h,k\right) := u^n_{i+\frac{1}{2}} \cdot \begin{cases} h, & \text{if } u^n_{i+\frac{1}{2}} \geq 0, \\ k, & \text{if } u^n_{i+\frac{1}{2}} < 0, \end{cases} = h\left(u^n_{i+\frac{1}{2}}\right)_+ + k\left(u^n_{i+\frac{1}{2}}\right)_-, \quad \forall\left(h,k\right) \in \mathbb{R}^2_+,
$$

and their total accumulation is defined by[3]

$$
g^n\left(h,k,x\right) := \sum_{i\in\mathbb{Z}} g^n_{i+\frac{1}{2}}\left(h,k\right)\mathbf{1}_{\overline{C_{i+\frac{1}{2}}}}(x), \quad \forall\left(h,k,x\right) \in \mathbb{R}^2_+ \times \mathbb{R},
$$

in particular,

$$
g^n\left(h^n_i, h^n_{i+1}, x_{i+\frac{1}{2}}\right) = g^n_{i+\frac{1}{2}}\left(h^n_i, h^n_{i+1}\right) = h^n_{i+\frac{1}{2}}u^n_{i+\frac{1}{2}} = F^n_{i+\frac{1}{2}}, \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T].
$$

This numerical flux g^n is "staggered-consistent" (Cf., Definition B.3) as follows,

$$
g^n\left(h^n_i, h^n_i, x_{i-\frac{1}{2}}\right) = g^n_{i-\frac{1}{2}}\left(h^n_i, h^n_i\right) = h^n_i u^n_{i-\frac{1}{2}},
$$

[3] The remaining variable x in g^n is used for tracking the "current position" of the numerical flux on the horizontal flux domain $\mathbb{R}$.

$$g^n\left(h_i^n, h_i^n, x_{i+\frac{1}{2}}\right) = g_{i+\frac{1}{2}}^n\left(h_i^n, h_i^n\right) = h_i^n u_{i+\frac{1}{2}}^n,$$

$$g^n\left(h_i^n, h_i^n, x_i\right) = g_{i-\frac{1}{2}}^n\left(h_i^n, h_i^n\right) + g_{i+\frac{1}{2}}^n\left(h_i^n, h_i^n\right) = h_i^n u_{i-\frac{1}{2}}^n + h_i^n u_{i+\frac{1}{2}}^n = 2h_i^n \bar{u}_i^n,$$

where

$$\bar{u}_i^n := \frac{1}{2}\left(u_{i-\frac{1}{2}}^n + u_{i+\frac{1}{2}}^n\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T],$$

and monotone in the sense of Definition (B.3) $g^n(\nearrow, \searrow, x)$, i.e., g^n is non-decreasing w.r.t. its first variable and non-increasing w.r.t. its second variable since its first-order partial derivatives are given by

$$\begin{aligned}
\partial_1 g^n(h,k,x) &= \sum_{i\in\mathbb{Z}} \left(u_{i+\frac{1}{2}}^n\right)_+ \mathbf{1}_{\overline{C_{i+\frac{1}{2}}}}(x), && \text{a.e. in } \mathbb{R}_+^2 \times \mathbb{R},\\
\partial_2 g^n(h,k,x) &= \sum_{i\in\mathbb{Z}} \left(u_{i+\frac{1}{2}}^n\right)_- \mathbf{1}_{\overline{C_{i+\frac{1}{2}}}}(x), && \text{a.e. in } \mathbb{R}_+^2 \times \mathbb{R},\\
\partial_3 g^n(h,k,x) &= 0, && \text{a.e. in } \mathbb{R}_+^2 \times \mathbb{R}.
\end{aligned}$$

The following result illustrates that if the initial water height is nonnegative and the time step is small enough, then the water height remains nonnegative later, which physically makes a sense.

Proposition 3.4 (Nonnegativity conservation of water height)**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty(\mathbb{R}; \mathbb{R}_+) \times L^\infty(\mathbb{R})$. *Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by* (DGsc).

If the following Courant-Friedrichs-Lewy-like (CFL-like) condition holds

$$\Delta t \le \frac{\inf_{i\in\mathbb{Z}} \Delta x_i}{\sup\limits_{i\in\mathbb{Z}, n\in[T]} \left(\left(u_{i+\frac{1}{2}}^n\right)_+ - \left(u_{i-\frac{1}{2}}^n\right)_-\right)}, \tag{CFL3}$$

or stronger

$$\nu \le \frac{\alpha_{\mathcal{T}}}{\sup\limits_{i\in\mathbb{Z}, n\in[T]} \left(\left(u_{i+\frac{1}{2}}^n\right)_+ - \left(u_{i-\frac{1}{2}}^n\right)_-\right)}, \tag{CFL4}$$

then the DG-staggered upwind scheme (DGsc) *is monotone in the following sense:*

Let $H_i^n : \mathbb{R}_+^3 \to \mathbb{R}$ *be defined by*

$$H_i^n(h,k,l) := k - \nu_i \left[g^n\left(k, l, x_{i+\frac{1}{2}}\right) - g^n\left(h, k, x_{i-\frac{1}{2}}\right)\right], \ \forall (h,k,l) \in \mathbb{R}_+^3,$$

so that

$$\begin{aligned}
H_i^n\left(h_{i-1}^n, h_i^n, h_{i+1}^n\right) &= h_i^n - \nu_i\left[g^n\left(h_i^n, h_{i+1}^n, x_{i+\frac{1}{2}}\right) - g^n\left(h_{i-1}^n, h_i^n, x_{i-\frac{1}{2}}\right)\right]\\
&= h_i^n - \nu_i\left[g_{i+\frac{1}{2}}^n\left(h_i^n, h_{i+1}^n\right) - g_{i-\frac{1}{2}}^n\left(h_{i-1}^n, h_i^n\right)\right]\\
&= h_i^n - \nu_i\left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right) = h_i^{n+1}, \ \forall i \in \mathbb{Z}, \ \forall n \in [T],
\end{aligned}$$

then H_i^n is non-decreasing w.r.t. its all three variables $H_i^n(\nearrow, \nearrow, \nearrow)$.

As a consequence, the DG-staggered upwind scheme (DGsc) preserves the nonnegativity of water height, i.e.: If the initial water height is nonnegative everywhere in the horizontal fluid domain $\mathbb{R}$, i.e., $h_0 \geq 0$ a.e. in $\mathbb{R}$, then it remains nonnegative in the discrete level, i.e.,

$$h_{\mathcal{T},\Delta t}(t,x) \geq 0, \ \ a.e. \ \ in \ \ [0,T_0) \times \mathbb{R},$$

where

$$h_{\mathcal{T},\Delta t}(t,x) := \sum_{i\in\mathbb{Z}} \sum_{n\in[T-1]} h_i^n \mathbf{1}_{[t^n,t^{n+1})\times C_i}(t,x), \ \ in \ \ [0,T_0)\times\mathbb{R}.$$

Proof. The first-order partial derivatives are given by

$$\begin{aligned}
\partial_1 H_i^n(h,k,l) &= \nu_i \partial_1 g^n\left(h,k,x_{i-\frac{1}{2}}\right) = \nu_i \left(u_{i-\frac{1}{2}}^n\right)_+ \geq 0,\\
\partial_2 H_i^n(h,k,l) &= 1 - \nu_i\left[\partial_1 g^n\left(k,l,x_{i+\frac{1}{2}}\right) - \partial_2 g^n\left(h,k,x_{i-\frac{1}{2}}\right)\right]\\
&= 1 - \nu_i\left[\left(u_{i+\frac{1}{2}}^n\right)_+ - \left(u_{i-\frac{1}{2}}^n\right)_-\right],\\
\partial_3 H_i^n(h,k,l) &= -\nu_i \partial_2 g^n\left(h,k,x_{i+\frac{1}{2}}\right) = -\nu_i\left(u_{i+\frac{1}{2}}^n\right)_- \geq 0.
\end{aligned}$$

The CFL condition (CFL3) (or (CFL4)) guarantees that

$$\partial_2 H_i^n(h,k,l) \geq 0, \ \ \text{a.e. in } \mathbb{R}_+^3.$$

Thus, $H_i^n(\nearrow, \nearrow, \nearrow)$ for all $i\in\mathbb{Z}$, $n\in[T]$.

Now assume that the initial data $h_0 \geq 0$ a.e. in $\mathbb{R}$, let us remark that if $h_i^n \geq 0$ for all $i \in \mathbb{Z}$, and for some $n \in [T-1]$, then

$$h_i^{n+1} = H_i^n\left(h_{i-1}^n, h_i^n, h_{i+1}^n\right) \geq H_i^n(0,0,0) = 0, \ \ \forall i \in \mathbb{Z}.$$

Thus, the desired result follows from the principle of mathematical induction. □

Alternative proof for nonnegativity conservation of water height. The discrete mass conservation equation (dmc) can be rewritten as

$$\begin{aligned}
h_i^{n+1} &= h_i^n - \nu_i\left[h_i^n\left(u_{i+\frac{1}{2}}^n\right)_+ + h_{i+1}^n\left(u_{i+\frac{1}{2}}^n\right)_- - h_{i-1}^n\left(u_{i-\frac{1}{2}}^n\right)_+ - h_i^n\left(u_{i-\frac{1}{2}}^n\right)_-\right]\\
&= \nu_i\left(u_{i-\frac{1}{2}}^n\right)_+ h_{i-1}^n + \left[1 - \nu_i\left(\left(u_{i+\frac{1}{2}}^n\right)_+ - \left(u_{i-\frac{1}{2}}^n\right)_-\right)\right] h_i^n - \nu_i\left(u_{i+\frac{1}{2}}^n\right)_- h_{i+1}^n,
\end{aligned}$$

for all $i\in\mathbb{Z}$, $n\in[T]$. The CFL-like condition (CFL3) (or (CFL4)) guarantees the coefficients of h_i^n is nonnegative. Thus, h_i^{n+1} is a linear combination[4] of h_{i-1}^n, h_i^n, and

[4] However, this is not a *convex* linear combination since the sum of three coefficients is different from 1:

$$\nu_i\left(u_{i-\frac{1}{2}}^n\right)_+ + 1 - \nu_i\left(u_{i+\frac{1}{2}}^n\right)_+ + \nu_i\left(u_{i-\frac{1}{2}}^n\right)_- - \nu_i\left(u_{i+\frac{1}{2}}^n\right)_- = 1 - \nu_i\left(u_{i+\frac{1}{2}}^n - u_{i-\frac{1}{2}}^n\right),$$

h_{i+1}^n with nonnegative coefficients. The desired result then follows by the principle of mathematical induction. □

Moreover, if, in addition, the initial water height is positive everywhere and the CFL-like condition is strict, then (dmc) preserves the positivity of water height instead of nonnegativity only.

Proposition 3.5 (Positivity conservation of water height)**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty\left(\mathbb{R}; \mathbb{R}_+^\star\right) \times L^\infty(\mathbb{R})$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by* (DGsc)*.*

If the following strict Courant-Friedrichs-Lewy-like (CFL-like) condition holds

$$\Delta t < \frac{\inf_{i\in\mathbb{Z}} \Delta x_i}{\sup\limits_{i\in\mathbb{Z}, n\in[T]} \left(\left(u_{i+\frac{1}{2}}^n\right)_+ - \left(u_{i-\frac{1}{2}}^n\right)_-\right)}, \tag{sCFL$_3$}$$

or stronger

$$\nu < \frac{\alpha_{\mathcal{T}}}{\sup\limits_{i\in\mathbb{Z}, n\in[T]} \left(\left(u_{i+\frac{1}{2}}^n\right)_+ - \left(u_{i-\frac{1}{2}}^n\right)_-\right)}. \tag{sCFL$_4$}$$

Then the DG-staggered upwind scheme (DGsc) *preserves the positivity of water height, i.e.: If the initial water height is positive everywhere in the horizontal fluid domain* $\mathbb{R}$*, i.e.,* $h_0 > 0$ *a.e. in* $\mathbb{R}$*, then it remains positive in the discrete level, i.e.,*

$$h_{\mathcal{T},\Delta t}(t,x) > 0, \text{ a.e. in } [0, T_0) \times \mathbb{R},$$

where

$$h_{\mathcal{T},\Delta t}(t,x) := \sum_{i\in\mathbb{Z}} \sum_{n\in[T-1]} h_i^n \mathbf{1}_{[t^n,t^{n+1})\times C_i}(t,x), \text{ in } [0, T_0) \times \mathbb{R}.$$

Proof. Since $h_0 > 0$ a.e. in $\mathbb{R}$, $h_i^0 > 0$ for all $i \in \mathbb{Z}$. Suppose that for some $n \in [T-1]$, $h_i^n > 0$ for all $i \in \mathbb{Z}$. We now prove that $h_i^{n+1} > 0$ for all $i \in \mathbb{Z}$. Indeed, using the same representation of (dmc) of the alternative proof for nonnegativity conservation of water height of Proposition (3.4) yields

$$\begin{aligned} h_i^{n+1} &= \nu_i \left(u_{i-\frac{1}{2}}^n\right)_+ h_{i-1}^n + \left[1 - \nu_i \left(\left(u_{i+\frac{1}{2}}^n\right)_+ - \left(u_{i-\frac{1}{2}}^n\right)_-\right)\right] h_i^n - \nu_i \left(u_{i+\frac{1}{2}}^n\right)_- h_{i+1}^n \\ &\geq \left[1 - \nu_i \left(\left(u_{i+\frac{1}{2}}^n\right)_+ - \left(u_{i-\frac{1}{2}}^n\right)_-\right)\right] h_i^n > 0, \quad \forall i \in \mathbb{Z}. \end{aligned}$$

The desired result then follows by the principle of mathematical induction. □

Remark 3.2 (L^∞-estimate fails for staggered schemes)**.** *Unlike the* L^∞*-estimate for cell-centered finite volume scheme stated in Proposition* B.2*, the staggered numerical*

for all $i \in \mathbb{Z}$, $n \in [T]$.

scheme (DGsc) does not bound the water height by the same bound used for the initial water height, i.e.,

$$h_0 \in [h_m, h_M] \;\; a.e. \;\; in \; \mathbb{R} \not\Rightarrow h_{\mathcal{T},\Delta t}(t,x) \in [h_m, h_M] \;\; a.e. \;\; in \;\; [0, T_0) \times \mathbb{R},$$

in the spirit of the proof of Proposition B.2. This is mainly because

$$\begin{aligned} H_i^n(\kappa,\kappa,\kappa) &= \kappa - \nu_i \left[g^n\left(\kappa,\kappa,x_{i+\frac{1}{2}}\right) - g^n\left(\kappa,\kappa,x_{i-\frac{1}{2}}\right)\right] \\ &= \kappa - \nu_i \left(g^n_{i+\frac{1}{2}}(\kappa,\kappa) - g^n_{i-\frac{1}{2}}(\kappa,\kappa)\right) \\ &= \kappa - \kappa\nu_i \left(u^n_{i+\frac{1}{2}} - u^n_{i-\frac{1}{2}}\right), \;\; \forall \kappa \in \mathbb{R}_+. \end{aligned}$$

Thus, in general, $H_i^n(\kappa,\kappa,\kappa) \neq \kappa$ provided $u^n_{i+\frac{1}{2}} \neq u^n_{i-\frac{1}{2}}$.

For staggered schemes such as (DGsc), we have only $H_i^n(0,0,0) = 0$ for all $i \in \mathbb{Z}$, $n \in [T]$, in order to preserve the nonnegativity of water height, but difficult to bound it during the observed time with the same bounds on the initial water height, in the discrete level, especially for staggered framework. Actually, we can not expect such L^∞ estimate for discrete solutions to exist since it fails for the continuous solution of (1DSW) in the first place!

Recall the still water (sw) defined previously, the following proposition indicates the well-balanced property of the upwind scheme (DGsc) in fully wet horizontal fluid domains.

Proposition 3.6 (Well-balanced property for fully wet horizontal fluid domain $\mathbb{R}$)**.** *Assume that the initial data satisfies $(h_0, u_0) \in L^\infty\left(\mathbb{R};\mathbb{R}_+^\star\right) \times L^\infty(\mathbb{R})$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $\left(h_i^n, u^n_{i+\frac{1}{2}}\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by (DGsc). Assume that the strict CFL-like condition (sCFL$_3$) or (sCFL$_4$) holds.*

Then the positivity of the initial water height and the strict CFL-like condition imply that the horizontal fluid domain $\mathbb{R}$ is fully wet during the survival time of (1DSW), i.e., $D^t = \emptyset$ and $W^t = \mathbb{R}$ for all $t \in [0, T_0)$. The still water steady states (sw) are preserved under the DG-staggered upwind scheme (DGsc), i.e., if for some $n \in [T-1]$, $u^n_{i+\frac{1}{2}} = 0$ and $h_i^n + b_i = C$ for all $i \in \mathbb{Z}$ for some constant C, then $u^{n+1}_{i+\frac{1}{2}} = 0$ and $h_i^{n+1} + b_i = C$ for all $i \in \mathbb{Z}$.

Proof. The upwind scheme (DGsc) can be rewritten more "conveniently" as

$$\begin{cases} h_i^{n+1} = h_i^n - \nu_i \left(h^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - h^n_{i-\frac{1}{2}} u^n_{i-\frac{1}{2}}\right), \\ \bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} = \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - \dfrac{1}{2}\nu_{i+\frac{1}{2}} u^n_{i+1}\left(h^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} + h^n_{i+\frac{3}{2}} u^n_{i+\frac{3}{2}}\right) \\ \qquad\qquad + \dfrac{1}{2}\nu_{i+\frac{1}{2}} u^n_i \left(h^n_{i-\frac{1}{2}} u^n_{i-\frac{1}{2}} + h^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}}\right) \\ \qquad\qquad - \dfrac{g}{2}\nu_{i+\frac{1}{2}} \left(h_i^{n+1} + h_{i+1}^{n+1}\right) \left[\left(h_{i+1}^{n+1} + b_{i+1}\right) - \left(h_i^{n+1} + b_i\right)\right], \end{cases} \tag{DGsc$_1$}$$

for all $i \in \mathbb{Z}$, $n \in [T-1]$.

Suppose that, for some $n_0 \in [T-1]$, $u_{i+\frac{1}{2}}^{n_0} = 0$ and $h_i^{n_0} + b_i = C$ for all $i \in \mathbb{Z}$ and for some constant C, then the last couple of equations becomes

$$\begin{cases} h_i^{n_0+1} = h_i^{n_0}, \\ \left(h_i^{n_0+1} + h_{i+1}^{n_0+1}\right) u_{i+\frac{1}{2}}^{n_0+1} = -g\nu_{i+\frac{1}{2}} \left(h_i^{n_0+1} + h_{i+1}^{n_0+1}\right) \left[\left(h_{i+1}^{n_0+1} + b_{i+1}\right) - \left(h_i^{n_0+1} + b_i\right)\right], \end{cases}$$

for all $i \in \mathbb{Z}$. As a direct consequence,

$$h_i^{n_0+1} + b_i = h_i^{n_0} + b_i = C, \quad \forall i \in \mathbb{Z}.$$

Combining the two last identities yields

$$\left(h_i^{n_0} + h_{i+1}^{n_0}\right) u_{i+\frac{1}{2}}^{n_0+1} = -g\nu_{i+\frac{1}{2}} \left(h_i^{n_0+1} + h_{i+1}^{n_0+1}\right) \left[\left(h_{i+1}^{n_0} + b_{i+1}\right) - \left(h_i^{n_0} + b_i\right)\right] = 0,$$

for $i \in \mathbb{Z}$. For each $i \in \mathbb{Z}$, if $h_i^{n_0} + h_{i+1}^{n_0} > 0$, the last identity implies $u_{i+\frac{1}{2}}^{n_0+1} = 0$. Otherwise, if $h_i^{n_0} + h_{i+1}^{n_0} = 0$, Proposition 3.4 implies that $h_i^{n_0} = h_{i+1}^{n_0} = 0$, and thus $h_i^{n_0+1} = h_{i+1}^{n_0+1} = 0$. Convention 3.1 then implies $u_{i+\frac{1}{2}}^{n_0+1} = 0$. This demonstrates the well-balanced property corresponded to (sw) in the discrete level. □

Hence, we have just established the well-balanced property which is the discrete version of (sw) for (DGsc) in fully wet horizontal fluid domains. Unfortunately, if the dry shore, or dry areas are allowed in our horizontal fluid domain, this kind of well-balanced property collapses completely, especially in the dry-wet transitions (i.e., the small areas between the dry and wet areas).

For the sake of convenience, analogous to the dry and wet areas described in Chapter 1, Section 2.1 in the continuous level, we define the following sets of dry- and wet-indices in the discrete level:

$$I_D^n := \{i \in \mathbb{Z}; h_i^n = 0\}, \quad I_W^n := \{i \in \mathbb{Z}; h_i^n > 0\}, \quad \forall n \in [T],$$

and the nonnegativity conservation of water height stated in Proposition 3.4 implies that

$$\mathbb{Z} = I_D^n \cup I_W^n, \quad \forall n \in [T].$$

As another direct consequence, Convention 3.1 reads

$$u_{i+\frac{1}{2}}^n := 0, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^n\right) \wedge \left(i+1 \in I_D^n\right), \quad \forall n \in [T].$$

Under these new notations, the following Proposition can illustrate the “ill-balanced” mentioned in the previous paragraph more precisely.

Proposition 3.7 (Ill-balanced property for dry-wet horizontal fluid domain $\mathbb{R}$)**.** *Assume that the initial data satisfies $(h_0, u_0) \in L^\infty(\mathbb{R}; \mathbb{R}_+) \times L^\infty(\mathbb{R})$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by (DGsc). Assume that the CFL-like condition (CFL3) or (CFL4) holds.*

Assume that the horizontal fluid domain $\mathbb{R}$ *contains at least one non-degenerate dry area (not 1-point dry area type) but not fully dry*[5] *during the survival time of* (1DSW)*, i.e.,*

$$\operatorname{card}\left(D^t\right) = \operatorname{card}\left(W^t\right) = c = 2^{\aleph_0} > \aleph_0, \quad \forall t \in [0, T_0)\,,$$

the still water steady states (sw) *are not preserved under the DG-staggered upwind scheme* (DGsc) *in general.*

Proof. Establishing a counterexample, it suffices to show a case for which the main knot in the proof of the Proposition 3.6 collapses.

The assumption that the horizontal fluid domain $\mathbb{R}$ contains at least one non-degenerate dry area but not fully dry allows the following case to happen (for suitable initial settings, e.g., $\mathcal{T}$, h_0, and also $u^n_{i+\frac{1}{2}}$'s): for some $n_0 \in [T]$, there exists an index $i_0 \in \mathbb{Z}$ such that

$$\left(i_0 \in I_D^{n_0}\right) \wedge \left[\left(i_0 - 1 \in I_W^{n_0}\right) \vee \left(i_0 + 1 \in I_W^{n_0}\right)\right].$$

Roughly speaking, this means that the i_0th primal cell C_{i_0} is completely dry while one of its neighbored cells is partially wet (i.e., that cell is not fully wet (some part of that cell is dry and some other part is wet) but still guarantees that the averaged water height in that cell must be positive instead of zero). Without loss of generality, we assume the primal cell C_{i+1} is partially wet.

In this case, (DGsc$_1$) at $(i, n) = (i_0, n_0)$ reads

$$\begin{cases} h_{i_0}^{n_0+1} = -\,\nu_{i_0}\left[h_{i_0+1}^{n_0}\left(u_{i_0+\frac{1}{2}}^{n_0}\right)_- - h_{i_0-1}^{n_0}\left(u_{i_0-\frac{1}{2}}^{n_0}\right)_+\right], \\ \left(h_{i_0}^{n_0+1} + h_{i_0+1}^{n_0+1}\right) u_{i_0+\frac{1}{2}}^{n_0+1} = h_{i_0+1}^{n_0} u_{i_0+\frac{1}{2}}^{n_0} - \nu_{i_0+\frac{1}{2}} u_{i_0+1}^{n_0}\left[h_{i_0+1}^{n_0}\left(u_{i_0+\frac{1}{2}}^{n_0}\right)_- + h_{i_0+\frac{3}{2}}^{n_0} u_{i_0+\frac{3}{2}}^{n_0}\right] \\ \qquad + \nu_{i_0+\frac{1}{2}} u_{i_0}^{n_0}\left[h_{i_0-1}^{n_0}\left(u_{i_0-\frac{1}{2}}^{n_0}\right)_+ + h_{i_0+1}^{n_0}\left(u_{i_0+\frac{1}{2}}^{n_0}\right)_-\right] \\ \qquad - g\nu_{i_0+\frac{1}{2}}\left(h_{i_0}^{n_0+1} + h_{i_0+1}^{n_0+1}\right)\left[\left(h_{i_0+1}^{n_0+1} + b_{i_0+1}\right) - \left(h_{i_0}^{n_0+1} + b_{i_0}\right)\right]. \end{cases}$$

Activating another assumption that at the discrete time $t = t^{n_0}$, all the water in the horizontal fluid domain $\mathbb{R}$ is completely at rest[6], i.e.,

$$u_{i+\frac{1}{2}}^{n_0} = 0, \quad \forall i \in \mathbb{Z},$$

[5] The former of this assumption prevents the dry component D^t from consisting of only (countable union) 1-point dry areas (in that case, we have $\operatorname{card}\left(D^t\right) = \operatorname{card}\left(\mathbb{N}\right) = \aleph_0$ instead of cardinality $\mathfrak{c}$ of continuum) while the latter prevents the considered horizontal fluid domain $\mathbb{R}$ from being *fully dry*. This is both physically and logically reasonable because if the whole horizontal fluid domain is dry, there is no water to consider, to model! So why do we need the shallow water equations (1DSW) in this worst-case scenario in the first place?

[6] Roughly speaking, $t^{n_0} \in [T_\star, T_0)$, where $T_\star$ is the steady time defined in Chapter 1, Section 2.1.

and consequently $u_i^{n_0} = 0$ for all $i \in \mathbb{Z}$. Under this assumption, (DGsc$_1$) at $(i,n) = (i_0, n_0)$ reduces to

$$\begin{cases} h_{i_0}^{n_0+1} = 0, \\ \left(h_{i_0}^{n_0+1} + h_{i_0+1}^{n_0+1}\right) u_{i_0+\frac{1}{2}}^{n_0+1} = -g\nu_{i_0+\frac{1}{2}} \left(h_{i_0}^{n_0+1} + h_{i_0+1}^{n_0+1}\right) \left[\left(h_{i_0+1}^{n_0+1} + b_{i_0+1}\right) - \left(h_{i_0}^{n_0+1} + b_{i_0}\right)\right], \end{cases}$$

or equivalently,

$$\begin{cases} h_{i_0}^{n_0+1} = 0, \\ h_{i_0+1}^{n_0+1} u_{i_0+\frac{1}{2}}^{n_0+1} = -g\nu_{i_0+\frac{1}{2}} h_{i_0+1}^{n_0+1} \left(h_{i_0+1}^{n_0+1} + b_{i_0+1} - b_{i_0}\right). \end{cases}$$

We claim that there is still water in the primal cell C_{i_0+1} at the time t^{n_0+1}, i.e., $h_{i_0+1}^{n_0+1} > 0$. Indeed, consider (DGsc$_1$) at $(i,n) = (i_0+1, n_0)$, the corresponding discrete mass conservation equation gives us

$$h_{i_0+1}^{n_0+1} = h_{i_0+1}^{n_0} - \nu_{i_0+1} \left(h_{i_0+\frac{3}{2}}^{n_0} u_{i_0+\frac{3}{2}}^{n_0} - h_{i_0+\frac{1}{2}}^{n_0} u_{i_0+\frac{1}{2}}^{n_0}\right) = h_{i_0+1}^{n_0} > 0,$$

where the last inequality is implied by $i_0 + 1 \in I_W^{n_0}$. Dividing the last equation by $h_{i_0+1}^{n_0+1} > 0$ yields

$$u_{i_0+\frac{1}{2}}^{n_0+1} = -g\nu_{i_0+\frac{1}{2}} \left(h_{i_0+1}^{n_0+1} + b_{i_0+1} - b_{i_0}\right).$$

So if the bottom at the primal cell C_{i_0} is strictly higher than the water surface at the primal cell C_{i_0+1} at the time $t = t^{n_0}$, i.e., $b_{i_0} > h_{i_0+1}^{n_0+1} + b_{i_0+1}$ and simultaneously, then the last identity yields $u_{i_0+\frac{1}{2}}^{n_0+1} > 0$. This destroys the well-balanced property in the dry-wet transition (i_0, i_0+1) and thus completes our counterexample. □

Remark 3.3. *Source terms are often added to the momentum equation to model friction phenomena, e.g., the Manning friction term is*

$$\mathfrak{M}(t,x) := \frac{C\left(u\,|u|\right)(t,x)}{h^{\frac{1}{3}}(t,x)}, \quad in \ [0, T_0) \times \mathbb{R},$$

where C is a given coefficient. In the staggered framework, the Manning friction term can be approximated with

$$\mathfrak{M}_{i+\frac{1}{2}}^n := \frac{C u_{i+\frac{1}{2}}^{n+1} \left|u_{i+\frac{1}{2}}^n\right|}{\left(h_{i+\frac{1}{2}}^{n+1}\right)^{\frac{1}{3}}}, \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T].$$

1.4 Passing to the limit in the DG staggered upwind scheme in the horizontal fluid domain

This subsection is devoted to show that if a sequence of solutions is controlled in suitable norms and converges to a limit, this limit necessarily satisfies a weak formulation of (1DSW), e.g., (EEb), (EE), and (EIb).

The following proposition depicts a balance on the kinetic energy and a balance on the potential and topography energy.

Proposition 3.8 (Balance on kinetic energy, potential energy and topography energy)**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty\left(\mathbb{R}; \mathbb{R}_+^\star\right) \times L^\infty(\mathbb{R})$. *Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by* (DGsc)*. Assume that the strict CFL-like condition* (sCFL_3) *or* (sCFL_4) *holds.*

The discrete solution of the scheme (DGsc) *satisfies, for all* $i \in \mathbb{Z}$ *and all* $n \in [T-1]$*, the balance*

$$\begin{aligned}\frac{1}{2\nu_{i+\frac{1}{2}}}\left(\bar{h}_{i+\frac{1}{2}}^{n+1}\left(u_{i+\frac{1}{2}}^{n+1}\right)^2 - \bar{h}_{i+\frac{1}{2}}^{n}\left(u_{i+\frac{1}{2}}^{n}\right)^2\right) + \frac{1}{2}\left(\overline{F}_{i+1}^n\left(u_{i+1}^n\right)^2 - \overline{F}_i^n\left(u_i^n\right)^2\right) \\ + \frac{g}{2}\left(\left(h_{i+1}^{n+1}\right)^2 - \left(h_i^{n+1}\right)^2\right)u_{i+\frac{1}{2}}^{n+1} + g\bar{h}_{i+\frac{1}{2}}^{n+1}u_{i+\frac{1}{2}}^{n+1}\left(b_{i+1}-b_i\right) = -R_{i+\frac{1}{2}}^{n+1},\end{aligned} \tag{BKE}$$

with

$$\begin{aligned}R_{i+\frac{1}{2}}^{n+1} := {} & \frac{1}{2\nu_{i+\frac{1}{2}}}\bar{h}_{i+\frac{1}{2}}^{n+1}\left(u_{i+\frac{1}{2}}^{n+1} - u_{i+\frac{1}{2}}^{n}\right)^2 \\ & + \frac{1}{2}\left[\left(\overline{F}_{i+1}^n\right)_-\left(u_{i+\frac{3}{2}}^n - u_{i+\frac{1}{2}}^n\right)^2 - \left(\overline{F}_i^n\right)_+\left(u_{i+\frac{1}{2}}^n - u_{i-\frac{1}{2}}^n\right)^2\right] \\ & - \left[\left(\overline{F}_{i+1}^n\right)_-\left(u_{i+\frac{3}{2}}^n - u_{i+\frac{1}{2}}^n\right) + \left(\overline{F}_i^n\right)_+\left(u_{i+\frac{1}{2}}^n - u_{i-\frac{1}{2}}^n\right)\right]\left(u_{i+\frac{1}{2}}^{n+1} - u_{i+\frac{1}{2}}^n\right),\end{aligned} \tag{BKEr}$$

and, for all $i \in \mathbb{Z}$ *and all* $n \in [T-1]$*, the balance*

$$\begin{aligned}\frac{1}{\nu_i}\left(\frac{g}{2}\left(h_i^{n+1}\right)^2 - \frac{g}{2}\left(h_i^n\right)^2\right) + \left(\frac{g}{2}\left(h_{i+\frac{1}{2}}^n\right)^2 u_{i+\frac{1}{2}}^n - \frac{g}{2}\left(h_{i-\frac{1}{2}}^n\right)^2 u_{i-\frac{1}{2}}^n\right) \\ + \frac{g}{2}\left(h_i^n\right)^2\left(u_{i+\frac{1}{2}}^n - u_{i-\frac{1}{2}}^n\right) = -R_i^{n+1},\end{aligned} \tag{BPETE}$$

with

$$\begin{aligned}R_i^{n+1} := {} & \frac{g}{2\nu_i}\left(h_i^{n+1} - h_i^n\right)^2 + \frac{1}{2}\left(h_{i+1}^n - h_i^n\right)^2\left(u_{i+\frac{1}{2}}^n\right)_- \\ & - \frac{1}{2}\left(h_{i-1}^n - h_i^n\right)^2\left(u_{i-\frac{1}{2}}^n\right)_+ + g\left(h_i^{n+1} - h_i^n\right)\left(\overline{F}_{i+\frac{1}{2}}^n - \overline{F}_{i-\frac{1}{2}}^n\right).\end{aligned} \tag{BPETEr}$$

Some ideas for the proof of these balances can be found in Herbin, Latché, and Nguyen, 2013.

Now, let a sequence of discretizations $\left(\mathcal{T}^{(m)}, \Delta t^{(m)}\right)_{m\in\mathbb{N}}$ be given

$$\mathcal{T}^{(m)} := \left\{C_i^{(m)}; i \in \mathbb{Z}\right\}, \text{ where } C_i^{(m)} := \left(x_{i-\frac{1}{2}}^{(m)}, x_{i+\frac{1}{2}}^{(m)}\right), \quad \forall i \in \mathbb{Z}, \quad \forall m \in \mathbb{N}.$$

The size mesh $\Delta x^{(m)}$ of the mesh $\mathcal{T}^{(m)}$ by

$$\Delta x^{(m)} := \sup_{i\in\mathbb{Z}} \Delta x_i^{(m)}, \text{ where } \Delta x_i^{(m)} := x_{i+\frac{1}{2}}^{(m)} - x_{i+\frac{1}{2}}^{(m)}, \quad \forall i \in \mathbb{Z}, \quad \forall m \in \mathbb{N}.$$

Let $\left(h^{(m)}, u^{(m)}\right)$ be the solution given by (DGsc) with the admissible $\mathcal{T}^{(m)}$ and the time step $\Delta t^{(m)}$, i.e.,

$$h^{(m)} := h_{\mathcal{T}^{(m)},\Delta t^{(m)}}, \ \ u^{(m)} := u_{\mathcal{T}^{(m)},\Delta t^{(m)}}, \ \text{a.e. in } [0,T_0)\times\mathbb{R}, \ \ \forall m\in\mathbb{N},$$

where

$$h_{\mathcal{T}^{(m)},\Delta t^{(m)}}(t,x) : = \sum_{i\in\mathbb{Z}}\sum_{n\in[T-1]} \left(h^{(m)}\right)_i^n \mathbf{1}_{\left[n\Delta t^{(m)},(n+1)\Delta t^{(m)}\right)\times C_i^{(m)}}(t,x), \ \text{in } [0,T_0)\times\mathbb{R},$$
$$u_{\mathcal{T}^{(m)},\Delta t^{(m)}}(t,x) : = \sum_{i\in\mathbb{Z}}\sum_{n\in[T-1]} \left(u^{(m)}\right)_{i+\frac{1}{2}}^n \mathbf{1}_{\left[n\Delta t^{(m)},(n+1)\Delta t^{(m)}\right)\times C_{i+\frac{1}{2}}^{(m)}}(t,x), \ \text{in } [0,T_0)\times\mathbb{R},$$

and where

$$C_{i+\frac{1}{2}}^{(m)} := \left(x_i^{(m)}, x_{i+1}^{(m)}\right), \ \ \forall i\in\mathbb{Z}, \ \ \forall m\in\mathbb{N}.$$

For discrete functions k and v defined on the primal and dual meshes, respectively, we define a discrete $L^1\left([0,T_0)\,;BV\left(\mathbb{R}\right)\right)$ norm by

$$\|k\|_{\mathcal{T},x,BV} : = \sum_{n\in[T]}\Delta t\sum_{i\in\mathbb{Z}}\left|k_{i+1}^n - k_i^n\right|,$$
$$\|v\|_{\mathcal{T},x,BV} : = \sum_{n\in[T]}\Delta t\sum_{i\in\mathbb{Z}}\left|v_{i+\frac{1}{2}}^n - v_{i-\frac{1}{2}}^n\right|,$$

and a discrete $L^1\left(\mathbb{R};BV\left([0,T_0)\right)\right)$ norm by

$$\|k\|_{\mathcal{T},t,BV} : = \sum_{i\in\mathbb{Z}}\Delta x_i\sum_{n\in[T-1]}\left|k_i^{n+1} - k_i^n\right|,$$
$$\|v\|_{\mathcal{T},t,BV} : = \sum_{i\in\mathbb{Z}}\Delta x_{i+\frac{1}{2}}\sum_{n\in[T-1]}\left|v_{i+\frac{1}{2}}^{n+1} - v_{i+\frac{1}{2}}^n\right|.$$

For the consistency results derived later, we have to assume that a sequence of discrete solutions $\left(h^{(m)}, u^{(m)}\right)_{m\in\mathbb{N}}$ satisfies $h^{(m)} > 0$ (Proposition (3.5) guarantees the positivity conservation of water height provided the mentioned strict CFL-like condition holds and the initial water height is positive everywhere in the horizontal fluid domain $\mathbb{R}$) and is uniformly bounded in $L^\infty([0,T_0)\times\mathbb{R})^2$, i.e.,

$$0 < \left(h^{(m)}\right)_i^n \le C, \ \ \forall i\in\mathbb{Z}, \ \ \forall n\in\left[T^{(m)}\right], \ \text{where } T^{(m)} := \left\lfloor\frac{T_0}{\Delta t^{(m)}}\right\rfloor, \ \ \forall m\in\mathbb{N}, \quad (3.1)$$

and

$$\left|\left(u^{(m)}\right)_{i+\frac{1}{2}}^n\right| \le C, \ \ \forall i\in\mathbb{Z}, \ \ \forall n\in\left[T^{(m)}\right], \ \ \forall m\in\mathbb{N}, \quad (3.2)$$

where C is a positive real number. Note that, by definition of the initial conditions of the scheme, these inequalities imply that the function h_0 and u_0 belong to $L^\infty\left(\mathbb{R}\right)$. We also have to assume that a sequence of discrete solutions satisfies the following uniform bounds w.r.t. the discrete BV-norms:

$$\left\|h^{(m)}\right\|_{\mathcal{T},x,BV}+\left\|u^{(m)}\right\|_{\mathcal{T},x,BV}\leq C,\ \ \forall m\in\mathbb{N},\tag{3.3}$$

and

$$\left\|h^{(m)}\right\|_{\mathcal{T},t,BV}\leq C,\ \ \forall m\in\mathbb{N}.\tag{3.4}$$

A weak solution (also, integral solution, see Definition A.2) to the continuous problem (1DSW) satisfies

$$\begin{cases}-\displaystyle\int_0^{T_0}\int_{\mathbb{R}}\left(h\partial_t\varphi+hu\partial_x\varphi\right)dxdt-\int_{\mathbb{R}}h_0\left(x\right)\varphi\left(0,x\right)dx=0,\\ -\displaystyle\int_0^{T_0}\int_{\mathbb{R}}\left[hu\partial_t\varphi+\left(hu^2+\frac{g}{2}h^2\right)\partial_x\varphi\right]dxdt-\int_{\mathbb{R}}\left(h_0u_0\right)\left(x\right)\varphi\left(0,x\right)dx=0,\end{cases}\tag{WF1DSW}$$

for all $\varphi\in C_c^\infty\left([0,T_0)\times\mathbb{R}\right)$.

This weak formulation of (1DSW) allow to derive the Rankine-Hugoniot relations (RH). Hence if (RH) is satisfied by the limit of a sequence of solutions to the discrete problem, loosely speaking, *the staggered upwind scheme* (DGsc) *computes correct shocks* (i.e., shocks where the jumps of the unknowns and of the fluxes are linked to the shock speed by Rankine-Hugoniot conditions) (see Chapter 1, Subsection 2.2.4).

Theorem 3.1 (Consistency of (DGsc))**.** *Assume that the initial data satisfies* $(h_0,u_0)\in L^\infty\left(\mathbb{R};\mathbb{R}_+^\star\right)\times L^\infty\left(\mathbb{R}\right)$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t\in\mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n,u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z},n\in[T]}$ *be the discrete finite volume approximate solution generated by* (DGsc)*. Assume that the strict CFL-like condition* (sCFL$_3$) *or* (sCFL$_4$) *holds.*

Assume that the bottom topography is flat. Let $\left(\mathcal{T}^{(m)},\Delta t^{(m)}\right)_{m\in\mathbb{N}}$ *be a sequence of discretizations such that both the time step* $\Delta t^{(m)}$ *and the size* $\Delta x^{(m)}$ *of the mesh* $\mathcal{T}^{(m)}$ *tend to zero as* $m\to\infty$*, and let* $\left(h^{(m)},u^{(m)}\right)_{m\in\mathbb{N}}$ *be the corresponding sequence of solutions. If this sequence satisfies the estimates* (3.1)-(3.3) *and converges in* $L^p\left([0,T_0)\times\mathbb{R};\mathbb{R}_+^\star\right)\times L^p\left([0,T_0)\times\mathbb{R}\right)$*, for* $1\leq p<\infty$*, to* $\left(\bar{h},\bar{u}\right)\in L^\infty\left([0,T_0)\times\mathbb{R};\mathbb{R}_+^\star\right)\times L^\infty\left([0,T_0)\times\mathbb{R}\right)$*. Then the limit* $\left(\bar{h},\bar{u}\right)$ *satisfies* (WF1DSW)*.*

Proof. For the main ideas of the proof, see Herbin, Latché, and Nguyen, 2013, p. 93 and references therein. □

We now turn to the entropy inequality (EIb), especially its integral form

$$\int_0^{T_0}\int_{\mathbb{R}}\left(-\Phi_b\left(U_b\right)\partial_t\varphi-\Psi_b\left(U_b\right)\partial_x\varphi\right)dxdt-\int_{\mathbb{R}}\Phi_b\left(U_{b,0}\right)\varphi\left(0,x\right)dx\leq 0,\tag{WEI}$$

for all $\varphi\in C_c^\infty\left([0,T_0)\times\mathbb{R};\mathbb{R}_+\right)$.

Now, we need to introduce the following additional condition for a regular sequence of discretizations:

$$\lim_{m\to\infty} \frac{\Delta t^{(m)}}{\inf\limits_{i\in\mathbb{Z}} \Delta x_i^{(m)}} = 0. \tag{lCFL}$$

Note that (lCFL) is slightly more restrictive than a standard CFL condition (CFL_1). We are now ready to state the final result for (DGsc).

Theorem 3.2 (Entropy consistency)**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty\left(\mathbb{R};\mathbb{R}_+^\star\right) \times L^\infty(\mathbb{R})$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by* (DGsc)*. Assume that the strict CFL-like condition* (sCFL_3) *or* (sCFL_4) *holds.*

Assume that the bottom topography is flat. Let $\left(\mathcal{T}^{(m)}, \Delta t^{(m)}\right)_{m\in\mathbb{N}}$ *be a sequence of discretizations such that both the time step* $\Delta t^{(m)}$ *and the size* $\Delta x^{(m)}$ *of the mesh* $\mathcal{T}^{(m)}$ *tend to zero as* $m\to\infty$*, and let* $\left(h^{(m)}, u^{(m)}\right)_{m\in\mathbb{N}}$ *be the corresponding sequence of solutions. If this sequence satisfies the estimates* (3.1)-(3.3) *and converges in* $L^p\left([0,T_0)\times\mathbb{R};\mathbb{R}_+^\star\right) \times L^p\left([0,T_0)\times\mathbb{R}\right)$*, for* $1 \le p < \infty$*, to* $\left(\bar{h}, \bar{u}\right) \in L^\infty\left([0,T_0)\times\mathbb{R};\mathbb{R}_+^\star\right) \times L^\infty\left([0,T_0)\times\mathbb{R}\right)$*. If we suppose in addition that the considered sequence of discretization satisfies* (lCFL)*, the sequence of solutions satisfies* (3.4)*, then the limit* $\left(\bar{h}, \bar{u}\right)$ *satisfies the entropy condition* (WEI)*.*

Proof. For the main ideas of the proof, see e.g., Herbin, Latché, and Nguyen, 2013, p. 93 and references therein. □

Remark 3.4 (On BV-stability assumptions)**.** *The mentioned proof of Theorem* (3.1) *shows that the scheme proposed in Herbin, Latché, and Nguyen, 2013 is consistent under a BV-stability assumption much weaker than* (3.3)*, namely,*

$$\lim_{m\to\infty} \Delta x^{(m)} \left(\left\| h^{(m)} \right\|_{\mathcal{T},x,BV} + \left\| u^{(m)} \right\|_{\mathcal{T},x,BV} \right) = 0.$$

The proof of Theorem (3.2)*, i.e., the limit of convergent sequences is an entropy solution, is completely different, since the stability condition which is used in the proof actually is*

$$\lim_{m\to\infty} \frac{\Delta t^{(m)}}{\inf\limits_{i\in\mathbb{Z}} \Delta x_i^{(m)}} \left\| h^{(m)} \right\|_{\mathcal{T},t,BV} = 0.$$

1.5 Higher-order staggered upwind schemes on the horizontal fluid domain

In the spirit of Appendix B, Subsection 1.6, a 5-point staggered upwind scheme is proposed by introducing “slopes” on all primal and dual cells and limiting the slopes so that the resulted scheme remains stable in order to reduce the diffusion of the 3-point staggered upwind scheme (DGsc). We develop this 5-point staggered upwind scheme by modifying the “MUSCL” technique.

Firstly, for the mass conservation equation in (1DSW), reconstructing a slope on each primal cell enables to compute interface values on each side of each node of the dual mesh. Then these values are used to compute the new numerical fluxes. For the sake of convenience, we add a bonus “5” in the superscripts to all discrete approximations below to indicate that they are generated by a new 5-point staggered upwind scheme so that we can distinguish them with the 3-point staggered upwind scheme (DGsc), with added “3” in the superscripts, being derived before.

More precisely, for each $n \in [T]$, the 3-point numerical fluxes are given by

$$\begin{aligned}
g^{3,n}_{i+\frac{1}{2}}(h,k) &:= h\left(u^{3,n}_{i+\frac{1}{2}}\right)_+ + k\left(u^{3,n}_{i+\frac{1}{2}}\right)_-, \quad \forall (h,k) \in \mathbb{R}^2_+,\\
g^{3,n}(h,k,x) &:= \sum_{i\in\mathbb{Z}} g^{3,n}_{i+\frac{1}{2}}(h,k)\,\mathbf{1}_{\overline{C_{i+\frac{1}{2}}}}(x), \quad \forall (h,k,x) \in \mathbb{R}^2_+ \times \mathbb{R},\\
g^{3,n}\left(h^{3,n}_i, h^{3,n}_{i+1}, x_{i+\frac{1}{2}}\right) &= g^{3,n}_{i+\frac{1}{2}}\left(h^{3,n}_i, h^{3,n}_{i+1}\right) = h^{3,n}_{i+\frac{1}{2}} u^{3,n}_{i+\frac{1}{2}} = F^{3,n}_{i+\frac{1}{2}},\\
g^{3,n}\left(h^{3,n}_i, h^{3,n}_i, x_{i-\frac{1}{2}}\right) &= g^{3,n}_{i-\frac{1}{2}}\left(h^{3,n}_i, h^{3,n}_i\right) = h^{3,n}_i u^{3,n}_{i-\frac{1}{2}},\\
g^{3,n}\left(h^{3,n}_i, h^{3,n}_i, x_{i+\frac{1}{2}}\right) &= g^{3,n}_{i+\frac{1}{2}}\left(h^{3,n}_i, h^{3,n}_i\right) = h^{3,n}_i u^{3,n}_{i+\frac{1}{2}},\\
g^{3,n}\left(h^{3,n}_i, h^{3,n}_i, x_i\right) &= g^{3,n}_{i-\frac{1}{2}}\left(h^{3,n}_i, h^{3,n}_i\right) + g^{3,n}_{i+\frac{1}{2}}\left(h^{3,n}_i, h^{3,n}_i\right) = 2h^{3,n}_i \bar{u}^{3,n}_i,\\
\Gamma^{3,n}_i(u,v) &:= u\left(\overline{F}^{3,n}_i\right)_+ + v\left(\overline{F}^{3,n}_i\right)_-, \quad \forall (u,v) \in \mathbb{R}^2,\\
\Gamma^{3,n}(u,v,x) &:= \sum_{i\in\mathbb{Z}} \Gamma^{3,n}_i(u,v)\,\mathbf{1}_{\overline{C_i}}(x), \quad \forall (u,v,x) \in \mathbb{R}^3,\\
\Gamma^{3,n}\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i+\frac{1}{2}}, x_i\right) &= \Gamma^{3,n}_i\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i+\frac{1}{2}}\right) = u^{3,n}_{i-\frac{1}{2}}\left(\overline{F}^{3,n}_i\right)_+ + u^{3,n}_{i+\frac{1}{2}}\left(\overline{F}^{3,n}_i\right)_- = G^{3,n}_i,\\
\Gamma^{3,n}\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i-\frac{1}{2}}, x_{i-1}\right) &= \Gamma^{3,n}_{i-1}\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i-\frac{1}{2}}\right) = u^{3,n}_{i-\frac{1}{2}}\overline{F}^{3,n}_{i-1},\\
\Gamma^{3,n}\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i-\frac{1}{2}}, x_i\right) &= \Gamma^{3,n}_i\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i-\frac{1}{2}}\right) = u^{3,n}_{i-\frac{1}{2}}\overline{F}^{3,n}_i,\\
\Gamma^{3,n}\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i-\frac{1}{2}}, x_{i-\frac{1}{2}}\right) &= \Gamma^{3,n}_{i-1}\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i-\frac{1}{2}}\right) + \Gamma^{3,n}_i\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i-\frac{1}{2}}\right)\\
&= u^{3,n}_{i-\frac{1}{2}}\overline{F}^{3,n}_{i-1} + u^{3,n}_{i-\frac{1}{2}}\overline{F}^{3,n}_i = u^{3,n}_{i-\frac{1}{2}}\left(F^{3,n}_{i-\frac{3}{2}} + 2F^{3,n}_{i-\frac{1}{2}} + F^{3,n}_{i+\frac{1}{2}}\right),
\end{aligned}$$

for all $i \in \mathbb{Z}$, $n \in [T]$, and the Doyen-Gunawan’s staggered upwind scheme (DGsc) can be rewritten with added “3” as

$$\left\{\begin{aligned}
h^{3,n+1}_i &:= h^{3,n}_i - \nu_i\left(F^{3,n}_{i+\frac{1}{2}} - F^{3,n}_{i-\frac{1}{2}}\right),\\
\bar{h}^{3,n+1}_{i+\frac{1}{2}} u^{3,n+1}_{i+\frac{1}{2}} &:= \bar{h}^{3,n}_{i+\frac{1}{2}} u^{3,n}_{i+\frac{1}{2}} - \nu_{i+\frac{1}{2}}\left[G^{3,n}_{i+1} - G^{3,n}_i + \frac{g}{2}\left(\left(h^{3,n+1}_{i+1}\right)^2 - \left(h^{3,n+1}_i\right)^2\right)\right]\\
&\qquad - g\nu_{i+\frac{1}{2}}\bar{h}^{3,n+1}_{i+\frac{1}{2}}\left(b_{i+1} - b_i\right),
\end{aligned}\right.$$

for all $i \in \mathbb{Z}$, $n \in [T-1]$.

We now construct a new 5-point staggered upwind scheme.

- *Computation of the slopes*:

$$\widetilde{k}_i^n := \frac{h_{i+1}^{5,n} - h_i^{5,n}}{\Delta x_i + \frac{1}{2}\left(\Delta x_{i-1} + \Delta x_{i+1}\right)} = \frac{h_{i+1}^{5,n} - h_i^{5,n}}{\Delta x_{i-\frac{1}{2}} + \Delta x_{i+\frac{1}{2}}}, \quad \forall i \in \mathbb{Z}.$$

- *Limitation of the slopes*:

$$k_i^n := \alpha_i^n \widetilde{k}_i^n, \quad \forall i \in \mathbb{Z},$$

where α_i^n is the largest number in $[0,1]$ such that

$$\begin{cases} h_i^{5,n} + \dfrac{\Delta x_i}{2}\alpha_i^n \widetilde{k}_i^n \in \left[h_i^{5,n} \perp h_{i+1}^{5,n}, h_i^{5,n} \top h_{i+1}^{5,n}\right], \\ h_i^{5,n} - \dfrac{\Delta x_i}{2}\alpha_i^n \widetilde{k}_i^n \in \left[h_{i-1}^{5,n} \perp h_i^{5,n}, h_{i-1}^{5,n} \top h_i^{5,n}\right]. \end{cases}$$

In practice, other formulas giving smaller values of α_i^n may be needed for stability purposes.

- *Computation of $h_i^{5,n+1}$ for all $i \in \mathbb{Z}$*: First, we have to define a new numerical flux,

$$\begin{aligned} g_{i+\frac{1}{2}}^{5,n}(h,k) &:= h\left(u_{i+\frac{1}{2}}^{5,n}\right)_+ + k\left(u_{i+\frac{1}{2}}^{5,n}\right)_-, \quad \forall (h,k) \in \mathbb{R}_+^2, \\ g^{5,n}(h,k,x) &:= \sum_{i\in\mathbb{Z}} g_{i+\frac{1}{2}}^{5,n}(h,k)\, \mathbf{1}_{\overline{C_{i+\frac{1}{2}}}}(x), \quad \forall (h,k,x) \in \mathbb{R}_+^2 \times \mathbb{R}, \\ g^{5,n}\left(h_i^{5,n}, h_{i+1}^{5,n}, x_{i+\frac{1}{2}}\right) &= g_{i+\frac{1}{2}}^{5,n}\left(h_i^{5,n}, h_{i+1}^{5,n}\right) = h_{i+\frac{1}{2}}^{5,n} u_{i+\frac{1}{2}}^{5,n} = F_{i+\frac{1}{2}}^{5,n}, \end{aligned}$$

for all $i \in \mathbb{Z}$, $n \in [T]$.

One replace

$$F_{i+\frac{1}{2}}^{3,n} := g^{3,n}\left(h_i^{3,n}, h_{i+1}^{3,n}, x_{i+\frac{1}{2}}\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T],$$

in (dmc) by

$$\begin{aligned} F_{i+\frac{1}{2}}^{5,n} &:= \widetilde{g}^{5,n}\left(h_{i-1}^{5,n}, h_i^{5,n}, h_{i+1}^{5,n}, h_{i+2}^{5,n}, x_{i+\frac{1}{2}}\right) \\ &:= g^{5,n}\left(h_i^{5,n} + \frac{\Delta x_i}{2}k_i^n, h_{i+1}^{5,n} - \frac{\Delta x_{i+1}}{2}k_{i+1}^n, x_{i+\frac{1}{2}}\right) \\ &= g_{i+\frac{1}{2}}^{5,n}\left(h_i^{5,n} + \frac{\Delta x_i}{2}k_i^n, h_{i+1}^{5,n} - \frac{\Delta x_{i+1}}{2}k_{i+1}^n\right) \\ &= \left(h_i^{5,n} + \frac{\Delta x_i}{2}k_i^n\right)\left(u_{i+\frac{1}{2}}^{5,n}\right)_+ + \left(h_{i+1}^{5,n} - \frac{\Delta x_{i+1}}{2}k_{i+1}^n\right)\left(u_{i+\frac{1}{2}}^{5,n}\right)_-, \end{aligned}$$

for all $i \in \mathbb{Z}$, $n \in [T]$.

The new discrete mass conservation equation reads

$$h_i^{5,n+1} := h_i^{5,n} - \nu_i\left(F_{i+\frac{1}{2}}^{5,n} - F_{i-\frac{1}{2}}^{5,n}\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T-1].$$

Secondly, for the momentum balance equation in (1DSW), reconstructing a slope on each dual cell enables to compute centered values on each side of each node of the primal mesh. More precisely, for each $n \in [T]$,

- *Computation of the slopes*:

$$\widetilde{v}^n_{i+\frac{1}{2}} := \frac{u^{5,n}_{i+\frac{3}{2}} - u^{5,n}_{i-\frac{1}{2}}}{\Delta x_{i-\frac{1}{2}} + \frac{1}{2}\left(\Delta x_{i+\frac{1}{2}} + \Delta x_{i+\frac{3}{2}}\right)}, \quad \forall i \in \mathbb{Z}.$$

- *Limitation of the slopes*:

$$v^n_{i+\frac{1}{2}} := \beta^n_{i+\frac{1}{2}} \widetilde{v}^n_{i+\frac{1}{2}}, \quad \forall i \in \mathbb{Z},$$

  where $\beta^n_{i+\frac{1}{2}}$ is the largest number in $[0,1]$ such that

$$\begin{cases} u^{5,n}_{i+\frac{1}{2}} + \dfrac{1}{2}\Delta x_{i+\frac{1}{2}} \beta^n_{i+\frac{1}{2}} \widetilde{v}^n_{i+\frac{1}{2}} \in \left[u^{5,n}_{i+\frac{1}{2}} \perp u^{5,n}_{i+\frac{3}{2}}, u^{5,n}_{i+\frac{1}{2}} \top u^{5,n}_{i+\frac{3}{2}}\right], \\ u^{5,n}_{i+\frac{1}{2}} - \dfrac{1}{2}\Delta x_{i+\frac{1}{2}} \beta^n_{i+\frac{1}{2}} \widetilde{v}^n_{i+\frac{1}{2}} \in \left[u^{5,n}_{i-\frac{1}{2}} \perp u^{5,n}_{i+\frac{1}{2}}, u^{5,n}_{i-\frac{1}{2}} \top u^{5,n}_{i+\frac{1}{2}}\right]. \end{cases}$$

  In practice, other formulas giving smaller values of $\beta^n_{i+\frac{1}{2}}$ may be needed for stability purposes.

- *Computation of $u^{5,n+1}_{i+\frac{1}{2}}$ for all $i \in \mathbb{Z}$*: First, we have to define a new numerical flux,

$$\begin{aligned} \overline{F}^{5,n}_i &:= \frac{1}{2}\left(F^{5,n}_{i-\frac{1}{2}} + F^{5,n}_{i+\frac{1}{2}}\right), \\ \Gamma^{5,n}_i(u,v) &:= u\left(\overline{F}^{5,n}_i\right)_+ + v\left(\overline{F}^{5,n}_i\right)_-, \quad \forall (u,v) \in \mathbb{R}^2, \\ \Gamma^{5,n}(u,v,x) &:= \sum_{i\in\mathbb{Z}} \Gamma^{5,n}_i(u,v)\mathbf{1}_{\overline{C_i}}(x), \quad \forall (u,v,x) \in \mathbb{R}^3, \\ \Gamma^{5,n}\left(u^{5,n}_{i-\frac{1}{2}}, u^{5,n}_{i+\frac{1}{2}}, x_i\right) &= \Gamma^{5,n}_i\left(u^{5,n}_{i-\frac{1}{2}}, u^{5,n}_{i+\frac{1}{2}}\right) = u^{5,n}_{i-\frac{1}{2}}\left(\overline{F}^{5,n}_i\right)_+ + u^{5,n}_{i+\frac{1}{2}}\left(\overline{F}^{5,n}_i\right)_- \\ &= \Gamma^{5,n}_i\left(u^{5,n}_{i-\frac{1}{2}}, u^{5,n}_{i+\frac{1}{2}}\right), \end{aligned}$$

  for all $i \in \mathbb{Z}$, $n \in [T]$.

  One replace

$$G^{3,n}_i = \Gamma^{3,n}\left(u^{3,n}_{i-\frac{1}{2}}, u^{3,n}_{i+\frac{1}{2}}, x_i\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in [T],$$

  in (dmb) by

$$\begin{aligned} G^{5,n}_i &:= \widetilde{\Gamma}^{5,n}\left(u^{5,n}_{i-\frac{3}{2}}, u^{5,n}_{i-\frac{1}{2}}, u^{5,n}_{i+\frac{1}{2}}, u^{5,n}_{i+\frac{3}{2}}, x_i\right) \\ &:= \Gamma^{5,n}\left(u^{5,n}_{i-\frac{1}{2}} + \frac{1}{2}\Delta x_{i-\frac{1}{2}} v^n_{i-\frac{1}{2}}, u^{5,n}_{i+\frac{1}{2}} - \frac{1}{2}\Delta x_{i+\frac{1}{2}} v^n_{i+\frac{1}{2}}, x_i\right) \end{aligned}$$

$$
\begin{aligned}
&= \Gamma_i^{5,n}\left(u_{i-\frac{1}{2}}^{5,n}+\frac{1}{2}\Delta x_{i-\frac{1}{2}}v_{i-\frac{1}{2}}^{n}, u_{i+\frac{1}{2}}^{5,n}-\frac{1}{2}\Delta x_{i+\frac{1}{2}}v_{i+\frac{1}{2}}^{n}\right)\\
&= \left(u_{i-\frac{1}{2}}^{5,n}+\frac{1}{2}\Delta x_{i-\frac{1}{2}}v_{i-\frac{1}{2}}^{n}\right)\left(\overline{F}_i^{5,n}\right)_+ + \left(u_{i+\frac{1}{2}}^{5,n}-\frac{1}{2}\Delta x_{i+\frac{1}{2}}v_{i+\frac{1}{2}}^{n}\right)\left(\overline{F}_i^{5,n}\right)_-,
\end{aligned}
$$

for all $i\in\mathbb{Z}$, $n\in[T]$.

The new discrete momentum balance equation reads

$$
\begin{aligned}
\bar{h}_{i+\frac{1}{2}}^{5,n+1}u_{i+\frac{1}{2}}^{5,n+1} := \bar{h}_{i+\frac{1}{2}}^{5,n}u_{i+\frac{1}{2}}^{5,n} - \nu_{i+\frac{1}{2}}\left[G_{i+1}^{5,n}-G_i^{5,n}+\frac{g}{2}\left(\left(h_{i+1}^{5,n+1}\right)^2-\left(h_i^{5,n+1}\right)^2\right)\right]\\
- g\nu_{i+\frac{1}{2}}\bar{h}_{i+\frac{1}{2}}^{5,n+1}\left(b_{i+1}-b_i\right), \qquad \forall i\in\mathbb{Z},\ \forall n\in[T-1].
\end{aligned}
$$

The constructed 5-point scheme

$$
\left\{
\begin{aligned}
h_i^{5,n+1} &:= h_i^{5,n}-\nu_i\left(F_{i+\frac{1}{2}}^{5,n}-F_{i-\frac{1}{2}}^{5,n}\right),\\
\bar{h}_{i+\frac{1}{2}}^{5,n+1}u_{i+\frac{1}{2}}^{5,n+1} &:= \bar{h}_{i+\frac{1}{2}}^{5,n}u_{i+\frac{1}{2}}^{5,n}-\nu_{i+\frac{1}{2}}\left[G_{i+1}^{5,n}-G_i^{5,n}+\frac{g}{2}\left(\left(h_{i+1}^{5,n+1}\right)^2-\left(h_i^{5,n+1}\right)^2\right)\right]\\
&\quad - g\nu_{i+\frac{1}{2}}\bar{h}_{i+\frac{1}{2}}^{5,n+1}\left(b_{i+1}-b_i\right),
\end{aligned}
\right.
\tag{5pDGsc}
$$

is hoped to be less diffusive than (DGsc) and it remains stable thanks to the limitation of the slope. Indeed, if

$$
\alpha_i^n=\beta_i^n=1,\ \ \forall i\in\mathbb{Z},\ \ \forall n\in[T],
$$

i.e., the limitation of the slope is inactive, the space diffusion term disappears from this new 5-point staggered upwind scheme, while the time “antidiffusion” term remains. Hence a higher order scheme for the time discretization is used, e.g., Runge-Kutta, or Heun method for the discretization of the time derivative.

The MUSCL scheme may be written as

$$
\left\{
\begin{aligned}
\frac{h^{5,n+1}-h^{5,n}}{\Delta t} &= \overline{H}_1\left(h^{5,n}\right), & \forall n\in[T-1],\\
\frac{\left(\bar{h}u\right)^{5,n+1}-\left(\bar{h}u\right)^{5,n}}{\Delta t} &= \overline{H}_2\left(\left(\bar{h}u\right)^{5,n}\right), & \forall n\in[T-1],
\end{aligned}
\right.
$$

where

$$
h^{5,n}=\left(h_i^n\right)_{i\in\mathbb{Z}},\ \ \left(\bar{h}u\right)^{5,n}=\left(\bar{h}_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}},\ \ \forall n\in[T],
$$

which may be seen as the explicit Euler discretization of

$$
\left\{
\begin{aligned}
\partial_t h^5 &= \overline{H}_1\left(h^5\right), & \forall n\in[T-1],\\
\partial_t\left(\bar{h}u\right)^5 &= \overline{H}_2\left(\left(\bar{h}u\right)^5\right), & \forall n\in[T-1],
\end{aligned}
\right.
$$

The RK2 time discretization yields the following scheme

$$\begin{cases} \dfrac{h^{5,n+1}-h^{5,n}}{\Delta t} = \dfrac{1}{2}\overline{H}_1\left(h^{5,n}\right)+\dfrac{1}{2}\overline{H}_1\left(h^{5,n}+\Delta t\overline{H}_1\left(h^{5,n}\right)\right), \\ \dfrac{\left(\bar{h}u\right)^{5,n+1}-\left(\bar{h}u\right)^{5,n}}{\Delta t} = \dfrac{1}{2}\overline{H}_2\left(\left(\bar{h}u\right)^{5,n}\right)+\dfrac{1}{2}\overline{H}_2\left(\left(\bar{h}u\right)^{5,n}+\Delta t\overline{H}_2\left(\left(\bar{h}u\right)^{5,n}\right)\right), \end{cases} \tag{RK2-5pDGsc}$$

for all $n \in [T-1]$. Such a second-order discretization in time allows larger time steps, without loss of stability. The scheme (RK2-5pDGsc) is proposed here for future purposes, there is no section devoted to study properties of this high-order scheme.

2 Doyan-Gunawan's staggered scheme on computational horizontal fluid domains

This section is devoted to discuss the staggered upwind scheme (DGsc) proposed in Doyen and Gunawan, 2014 on a computational domain $\Omega_c := [0, L]$ during the time $[0, T_0)$, where L the length of the restricted horizontal fluid domain Ω_c whose water is at rest at its both boundaries, i.e., the following homogeneous Dirichlet boundary conditions are imposed at the boundary of Ω_c:

$$u\,(t,x)=0, \text{ on } [0,T_0)\times\partial\Omega_c.$$

In addition, the following initial condition is imposed at $t=0$:

$$(h,hu)\,(0,x)=(h_0,q_0)\,(x)\,, \text{ in } \Omega_c.$$

Let $\mathcal{T}_c$ be an admissible mesh in the sense of Definition (B.1) which is "perfectly" restricted on the computational domain Ω_c, i.e.,

$$\mathcal{T}_c := \{C_i; i\in[N_x]^\star\}\,, \text{ where } x_{\frac{1}{2}}=0,\;\; x_{N_x+\frac{1}{2}}=L,\;\; \sum_{i=1}^{N_x}\Delta x_i = x_{N_x+\frac{1}{2}}-x_{\frac{1}{2}}=L,$$

where $[N_x]^\star := [N_x]\setminus\{0\}$.

To define the two numerical fluxes at the boundaries of Ω_c, let $\overline{h}, \overline{\overline{h}} \in L^\infty\left([0,T_0),\mathbb{R}_+\right)$ such that $\overline{h}, \overline{\overline{h}} \in [h_m, h_M]$ a.e. in $\mathbb{R}_+$. Define

$$\overline{h}^n := \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\overline{h}\,(t)\,dt,\;\; \overline{\overline{h}}^n := \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}}\overline{\overline{h}}\,(t)\,dt,\;\; \forall n\in[T]\,,$$

then the staggered upwind scheme (DGsc) restricted in Ω_c reads

$$\begin{cases} h_i^{n+1} := h_i^n - \nu_i \left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}} \right), \;\; \forall i \in [N_x]^\star, \\ \bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} := \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - \nu_{i+\frac{1}{2}} \left[G^n_{i+1} - G^n_i + \frac{g}{2} \left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2 \right) \right] \\ \qquad\qquad - g \nu_{i+\frac{1}{2}} \bar{h}^{n+1}_{i+\frac{1}{2}} \left(b_{i+1} - b_i \right), \;\; \forall i \in [N_x], \\ F^n_{\frac{1}{2}} := g^n \left(h^n_0, h^n_1, x_{\frac{1}{2}} \right), \text{ where } h^n_0 := \overline{h}^n, \\ F^n_{N_x+\frac{1}{2}} := g^n \left(h^n_{N_x}, h^n_{N_x+1}, x_{N_x+\frac{1}{2}} \right), \text{ where } h^n_{N_x+1} := \overline{\overline{h}}^n, \end{cases} \tag{rDGsc}$$

for all $n \in [T-1]$.

In the flat bottom topography case, this staggered upwind scheme reads

$$\begin{cases} h_i^{n+1} : = h_i^n - \nu_i \left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}} \right), & \forall i \in [N_x]^\star, \\ \bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} : = \bar{h}^n_{i+\frac{1}{2}} u^n_{i+\frac{1}{2}} - \nu_{i+\frac{1}{2}} \left[G^n_{i+1} - G^n_i + \frac{g}{2} \left(\left(h^{n+1}_{i+1}\right)^2 - \left(h^{n+1}_i\right)^2 \right) \right], & \forall i \in [N_x], \\ F^n_{\frac{1}{2}} : = g^n \left(h^n_0, h^n_1, x_{\frac{1}{2}} \right), & \\ F^n_{N_x+\frac{1}{2}} : = g^n \left(h^n_{N_x}, h^n_{N_x+1}, x_{N_x+\frac{1}{2}} \right), & \end{cases} \tag{rDGscfb}$$

for all $n \in [T-1]$.

The two numerical fluxes introduced at the boundaries $\partial\Omega_c$ are used to treat mainly nonhomogeneous Dirichlet condition. But in the current situation, the velocity is assumed to vanish at the boundaries $\partial\Omega_c$. The discrete boundary conditions are

$$u^n_{\frac{1}{2}} = u^n_{N_x+\frac{1}{2}} = 0, \;\; \forall n \in [T].$$

We still use these two boundary-numerical fluxes in our framework for the sake of generality.

The Convention (3.1) also is restricted straightforwardly in Ω_c as follows.

Convention 3.2 (Zero velocity of water in dry areas in computational horizontal fluid domains Ω_c)**.** *For all $i \in [N_x+1]$, $n \in [T]$, if $\overline{h}^n_{i+\frac{1}{2}} = 0$, by convention, $u^n_{i+\frac{1}{2}}$ is then set to zero.*

2.1 Conservation properties of DG staggered upwind schemes on computational horizontal fluid domains

Analogous to Lemma 3.1, we have the following result involving some reference state for (rDGscfb), with some suitable choices of the numerical fluxes on the boundaries of the computational domain Ω_c.

Lemma 3.2 (Reference states at rest for (rDGscfb))**.** *Assume that the initial data satisfies $(h_0, u_0) \in \left(L^\infty(\Omega_c)\right)^2$. Let $\mathcal{T}$ be an admissible mesh of Ω_c and $\Delta t \in \mathbb{R}^\star_+$ be the time step, and assume the bottom topography is flat, i.e., $b = 0$ in $\mathbb{R}$ and (h_0, u_0) is* *equal to some reference state at rest*[7] $(h_\star, 0) \in \Omega$ *for all* $x \in ((-\infty, M_L) \cup (M_L, \infty)) \cap \Omega_c$ *for some positive reals* $M_L \leq M_R$*, i.e.,*

$$h_0\left(x\right) = h_\star,\ \ u_0\left(x\right) = 0,\ \ \forall x \in ((-\infty, M_L) \cup (M_L, \infty)) \cap \Omega_c.$$

Choose

$$\overline{h}\left(t\right) = \overline{\overline{h}}\left(t\right) = h_\star,\ \ \forall t \in [0, T_0)\,,$$

and denote by i_{M_L} *and* i_{M_R} *the indices indicating which primal cells containing* M_L *and* M_R*, i.e.,*

$$M_L \in \left[x_{i_{M_L}-\frac{1}{2}}, x_{i_{M_L}+\frac{1}{2}}\right),\ \ M_R \in \left(x_{i_{M_R}-\frac{1}{2}}, x_{i_{M_R}+\frac{1}{2}}\right].$$

Let $\left(h_i^n\right)_{i\in[N_x]^\star, n\in[T]}$ *and* $\left(u_{i+\frac{1}{2}}^n\right)_{i\in[N_x], n\in[T]}$ *be the discrete finite volume approximate water height and velocity, respectively, generated by the restricted DG-staggered upwind scheme* (rDGscfb)*. Then they satisfy*

$$\begin{aligned} h_i^n &= h_\star,\ \ \forall i \in [N_x+1]\ \ s.t.\ \ (i < i_{M_L} - n) \vee (i > i_{M_R} + n)\,, \\ u_{i+\frac{1}{2}}^n &= 0,\ \ \forall i \in [N_x+1]\ \ s.t.\ \ (i < i_{M_L} - n - 1) \vee (i > i_{M_R} + n)\,. \end{aligned}$$

Note that when n is large enough, e.g., $n \geq \max\left(i_{M_L} - 1, N_x + 1 - i_{M_R}\right)$, the reference state $(h_\star, 0)$ will spread outside Ω_c and thus Lemma (3.2) is not useful anymore (Cf., Lemma (3.1)).

Analogous to Proposition 3.1 and Proposition 3.2, the restricted staggered upwind scheme (rDGsc) also possesses the same conservation properties with some modifications on the boundaries of the computational horizontal fluid domain.

Proposition 3.9 (Preserved quantities under (rDGsc))**.** *Assume that the initial data satisfies* $(h_0, u_0) \in \left(L^\infty\left(\Omega_c\right)\right)^2$*. Let* $\mathcal{T}$ *be an admissible mesh of* Ω_c *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n\right)_{i\in[N_x]^\star, n\in[T]}$ *and* $\left(u_{i+\frac{1}{2}}^n\right)_{i\in[N_x], n\in[T]}$ *be the discrete finite volume approximate water height and velocity, respectively, generated by* (rDGsc)*.*

Then the restricted staggered upwind scheme (rDGsc) *preserves the water volume, the total mass in the discrete level, with suitable 1D notions of 1D discrete water volume and total mass. Moreover, for the flat bottom topography case,* (rDGscfb) *also preserves the total momentum, provided* $\overline{h} \equiv \pm\overline{\overline{h}}$*, in the discrete level, with a suitable notion of 1D discrete total momentum.*

Proof. We prove the conservation properties in turns.

1. *Water volume (i.e., total height water).* Define the 1D restricted discrete water volume at the time $t = t^n$ by

$$\mathfrak{h}_c^n := \sum_{i=1}^{N_x} \Delta x_i h_i^n,\ \ \forall n \in [T]\,,$$

[7] Since $u_{\frac{1}{2}}^n = u_{N_x+\frac{1}{2}}^n = 0$ for all $n \in [T]$, the reference velocity $u_\star$ must be zero.

we claim that $\mathfrak{h}_c^{n+1} = \mathfrak{h}_c^n$ for all $n \in [T-1]$. Indeed,

$$\begin{aligned}\mathfrak{h}_c^{n+1} :&= \sum_{i=1}^{N_x} \Delta x_i h_i^{n+1} = \sum_{i=1}^{N_x} \left[\Delta x_i h_i^n - \Delta t\left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right)\right] \\ &= \sum_{i=1}^{N_x} \Delta x_i h_i^n - \Delta t\left(F_{N_x+\frac{1}{2}}^n - F_{\frac{1}{2}}^n\right) = \mathfrak{h}_c^n - \Delta t\left(h_{N_x+\frac{1}{2}}^n u_{N_x+\frac{1}{2}}^n - h_{\frac{1}{2}}^n u_{\frac{1}{2}}^n\right) = \mathfrak{h}_c^n,\end{aligned}$$

for all $n \in [T-1]$. As a consequence, the initial 1D restricted discrete water volume is preserved,

$$\mathfrak{h}_c^n = \mathfrak{h}_c^0, \quad \forall n \in [T].$$

2. *Total mass.* The 1D restricted discrete total mass at time $t = t^n$ is defined by

$$\mathscr{Z}_c^n := \sum_{i=1}^{N_x} \Delta x_i \left(h_i^n + b_i - H_0\right), \quad \forall n \in [T],$$

we claim that $\mathscr{Z}_c^{n+1} = \mathscr{Z}_c^n$ for all $n \in [T-1]$. Indeed,

$$\begin{aligned}\mathscr{Z}_c^{n+1} &= \sum_{i=1}^{N_x} \Delta x_i \left(h_i^{n+1} + b_i - H_0\right) \\ &= \sum_{i=1}^{N_x} \Delta x_i \left(h_i^n - \nu_i\left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right) + b_i - H_0\right) \\ &= \sum_{i=1}^{N_x} \Delta x_i \left(h_i^n + b_i - H_0\right) - \Delta t \sum_{i=1}^{N_x}\left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right) \\ &= \mathscr{Z}_c^n - \Delta t\left(F_{N_x+\frac{1}{2}}^n - F_{\frac{1}{2}}^n\right) = \mathscr{Z}_c^n, \quad \forall n \in [T-1].\end{aligned}$$

As a consequence, the initial 1D restricted discrete total mass is preserved,

$$\mathscr{Z}_c^n = \mathscr{Z}_c^0, \quad \forall n \in [T].$$

3. *Total momentum.* The 1D restricted discrete total momentum is defined by

$$\mathscr{M}_c^n := \sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} \bar{h}_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n, \quad \forall n \in [T],$$

we claim that $\mathscr{M}_c^{n+1} = \mathscr{M}_c^n$ for all $n \in [T-1]$ provided that the bottom topography is flat. Assume $b = 0$, (rDGsc) becomes

$$\begin{cases} h_i^{n+1} := h_i^n - \nu_i\left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right), & \forall i \in [N_x]^\star, \\ \bar{h}_{i+\frac{1}{2}}^{n+1} u_{i+\frac{1}{2}}^{n+1} := \bar{h}_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n - \nu_{i+\frac{1}{2}}\left[G_{i+1}^n - G_i^n + \frac{g}{2}\left(\left(h_{i+1}^{n+1}\right)^2 - \left(h_i^{n+1}\right)^2\right)\right], & \forall i \in [N_x], \\ F_{\frac{1}{2}}^n := g^n\left(h_0^n, h_1^n, x_{\frac{1}{2}}\right), \text{ where } h_0^n := \overline{h}^n, & \\ F_{N_x+\frac{1}{2}}^n := g^n\left(h_{N_x}^n, h_{N_x+1}^n, x_{N_x+\frac{1}{2}}\right), \text{ where } h_{N_x+1}^n := \overline{\overline{h}}^n, & \end{cases}$$

and then

$$
\begin{aligned}
\mathcal{M}_c^{n+1} &= \sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} \bar{h}_{i+\frac{1}{2}}^{n+1} u_{i+\frac{1}{2}}^{n+1} \\
&= \sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} \left[\bar{h}_{i+\frac{1}{2}}^{n} u_{i+\frac{1}{2}}^{n} - \nu_i \left[G_{i+1}^n - G_i^n + \frac{g}{2} \left(\left(h_{i+1}^{n+1}\right)^2 - \left(h_i^{n+1}\right)^2 \right) \right] \right] \\
&= \sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} \bar{h}_{i+\frac{1}{2}}^{n} u_{i+\frac{1}{2}}^{n} - \Delta t \sum_{i=0}^{N_x} \left[G_{i+1}^n - G_i^n + \frac{g}{2} \left(\left(h_{i+1}^{n+1}\right)^2 - \left(h_i^{n+1}\right)^2 \right) \right] \\
&= \mathcal{M}_c^n - \Delta t \left[G_{N_x+1}^n - G_0^n + \frac{g}{2} \left(\left(h_{N_x+1}^{n+1}\right)^2 - \left(h_0^{n+1}\right)^2 \right) \right] \\
&= \mathcal{M}_c^n - \frac{g}{2} \Delta t \left(\left(\overline{\overline{h}}^{n+1} \right)^2 - \left(\overline{h}^{n+1} \right)^2 \right), \quad \forall n \in [T-1] \, .
\end{aligned}
$$

Hence, if we choose $\overline{h} \equiv \pm \overline{\overline{h}}$, then $\mathcal{M}_c^{n+1} = \mathcal{M}_c^n$ for all $n \in [T-1]$. As a consequence, the initial 1D discrete total momentum is preserved,

$$
\mathcal{M}_c^n = \mathcal{M}_c^0, \quad \forall n \in [T] \, .
$$

The proof is completed. □

Analogous to Remark 3.1 and Proposition 3.3, the restricted DG-staggered upwind scheme (rDGsc), even in the flat bottom case $b = 0$, does not preserve the horizontal impulse $\mathcal{I}(t)$ and the total energy $\mathcal{H}_{\mathrm{SW}}(h, u)(t)$ in the discrete level.

Proposition 3.10 (Non-preserved quantities under (rDGscfb))**.** *Assume that the initial data satisfies* $(h_0, u_0) \in \left(L^\infty(\Omega_c)\right)^2$. *Let* $\mathcal{T}$ *be an admissible mesh of* Ω_c *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n\right)_{i \in [N_x]^\star, n \in [T]}$ *and* $\left(u_{i+\frac{1}{2}}^n\right)_{i \in [N_x], n \in [T]}$ *be the discrete finite volume approximate water height and velocity, respectively, generated by* (rDGsc).

The restricted upwind scheme (rDGscfb)*, in the flat bottom topography case, does not preserve the total horizontal impulse and also the total energy in the discrete level with the following choices of the discrete horizontal impulse and the discrete total energy:*

$$
\begin{aligned}
\mathcal{I}_c^n &:= \sum_{i=0}^{N_x} \Delta x_{i+\frac{1}{2}} \left(\bar{h}_{i+\frac{1}{2}}^n + b_i - H_0 \right) u_{i+\frac{1}{2}}^n, \\
\mathcal{H}_{\mathrm{SW},c,\alpha,\beta,\gamma,\delta}^n &:= \frac{1}{2} \sum_{i=0}^{N_x} \left[\Delta x_{i+\alpha} \left(h_{i+\alpha}^n + b_i - H_0 \right)^2 + \Delta x_{i+\beta} h_{i+\gamma}^n \left(u_{i+\delta}^n \right)^2 \right], \\
\overline{\mathcal{H}}_{\mathrm{SW},c,\alpha,\beta,\gamma,\delta}^n &:= \frac{1}{2} \sum_{i=0}^{N_x} \left[\Delta x_{i+\alpha} \left(\bar{h}_{i+\frac{1}{2}}^n + b_i - H_0 \right)^2 + \Delta x_{i+\beta} \bar{h}_{i+\frac{1}{2}}^n \left(u_{i+\gamma}^n \right)^2 \right],
\end{aligned}
$$

for all $n \in [T]$*, where the indices* $\alpha, \beta, \gamma, \delta \in \left\{0, \frac{1}{2}\right\}$.

For the sake of convenience, we define two new edges of the dual cells beyond the boundaries $\partial\Omega_c$, which are symmetry to their nearest edges relative to the boundaries, respectively, i.e.,

$$x_0 := -x_1, \;\; x_{N_x+1} := 2L - x_{N_x}.$$

Next, the restricted numerical flux $g_c^n : \mathbb{R}_+^2 \times \mathbb{R} \to \mathbb{R}$ is defined by

$$g_c^n\left(h,k,x\right) := \sum_{i=0}^{N_x} g_{i+\frac{1}{2}}^n\left(h,k\right)\mathbf{1}_{\overline{C_{i+\frac{1}{2}}}}(x), \;\; \forall\left(h,k,x\right) \in \mathbb{R}_+^2 \times \Omega_c,$$

in particular,

$$g_c^n\left(h_i^n, h_{i+1}^n, x_{i+\frac{1}{2}}\right) = g_{i+\frac{1}{2}}^n\left(h_i^n, h_{i+1}^n\right) = h_{i+\frac{1}{2}}^n u_{i+\frac{1}{2}}^n = F_{i+\frac{1}{2}}^n, \;\; \forall i \in \left[N_x\right], \;\; \forall n \in \left[T\right].$$

This numerical flux g_c^n is “staggered-consistent” (Cf., Definition (B.3)) as follows,

$$\begin{aligned}
g_c^n\left(h_i^n, h_i^n, x_{i-\frac{1}{2}}\right) &= g_{i-\frac{1}{2}}^n\left(h_i^n, h_i^n\right) = h_i^n u_{i-\frac{1}{2}}^n,\\
g_c^n\left(h_i^n, h_i^n, x_{i+\frac{1}{2}}\right) &= g_{i+\frac{1}{2}}^n\left(h_i^n, h_i^n\right) = h_i^n u_{i+\frac{1}{2}}^n,\\
g_c^n\left(h_i^n, h_i^n, x_i\right) &= g_{i-\frac{1}{2}}^n\left(h_i^n, h_i^n\right) + g_{i+\frac{1}{2}}^n\left(h_i^n, h_i^n\right) = h_i^n u_{i-\frac{1}{2}}^n + h_i^n u_{i+\frac{1}{2}}^n = 2h_i^n \bar{u}_i^n,
\end{aligned}$$

for all $n \in [T]$, and monotone in the sense of Definition B.3 $g_c^n(\nearrow, \searrow, x)$, i.e., g_c^n is non-decreasing w.r.t. its first variable and non-increasing w.r.t. its second variable since its first-order partial derivatives are given by

$$\begin{aligned}
\partial_1 g_c^n\left(h,k,x\right) &= \sum_{i=0}^{N_x}\left(u_{i+\frac{1}{2}}^n\right)_+ \mathbf{1}_{\overline{C_{i+\frac{1}{2}}}}(x), && \text{a.e. in } \mathbb{R}_+^2 \times \Omega_c,\\
\partial_2 g_c^n\left(h,k,x\right) &= \sum_{i=0}^{N_x}\left(u_{i+\frac{1}{2}}^n\right)_- \mathbf{1}_{\overline{C_{i+\frac{1}{2}}}}(x), && \text{a.e. in } \mathbb{R}_+^2 \times \Omega_c,\\
\partial_3 g_c^n\left(h,k,x\right) &= 0, && \text{a.e. in } \mathbb{R}_+^2 \times \Omega_c.
\end{aligned}$$

Analogous to Proposition 3.4, the restricted upwind scheme (rDGsc) also preserves the nonnegativity of water height provided the initial water height is nonnegative and the time step is small enough.

Proposition 3.11 (Nonnegativity conservation of water height). *Assume that the initial data satisfies $(h_0, u_0) \in L^\infty\left(\Omega_c; \mathbb{R}_+\right) \times L^\infty\left(\Omega_c\right)$. Let $\mathcal{T}$ be an admissible mesh of Ω_c and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $\left(h_i^n\right)_{i\in[N_x]^\star, n\in[T]}$ and $\left(u_{i+\frac{1}{2}}^n\right)_{i\in[N_x], n\in[T]}$ be the discrete finite volume approximate water height and velocity, respectively, generated by* (rDGsc)*.*

If the numerical fluxes on the boundaries $\partial\Omega_c$ are chosen to be nonnegative, i.e.,

$$\overline{h}\left(t\right) \geq 0, \;\; \overline{\overline{h}}\left(t\right) \geq 0, \;\; \forall t \in \left[0, T_0\right),$$

and the following restricted Courant-Friedrichs-Lewy-like (CFL-like) condition holds

$$\Delta t \leq \frac{\inf_{i\in\mathbb{Z}}\Delta x_i}{\max\limits_{i\in[N_x]^\star,n\in[T]}\left(\left(u^n_{i+\frac{1}{2}}\right)_+ - \left(u^n_{i-\frac{1}{2}}\right)_-\right)}, \qquad (\text{rCFL}_3)$$

or stronger

$$\nu \leq \frac{\alpha_{\mathcal{T}}}{\max\limits_{i\in[N_x]^\star,n\in[T]}\left(\left(u^n_{i+\frac{1}{2}}\right)_+ - \left(u^n_{i-\frac{1}{2}}\right)_-\right)}, \qquad (\text{rCFL}_4)$$

then the restricted DG-staggered upwind scheme (rDGsc) is monotone in the following sense:

Let $H^n_{c,i}:\mathbb{R}^3_+\to\mathbb{R}$ be defined by

$$H^n_{c,i}(h,k,l) := k-\nu_i\left[g^n_c\left(k,l,x_{i+\frac{1}{2}}\right)-g^n_c\left(h,k,x_{i-\frac{1}{2}}\right)\right],\ \ \forall(h,k,l)\in\mathbb{R}^3_+,$$

so that

$$\begin{aligned} H^n_{c,i}\left(h^n_{i-1},h^n_i,h^n_{i+1}\right) &= h^n_i-\nu_i\left[g^n_c\left(h^n_i,h^n_{i+1},x_{i+\frac{1}{2}}\right)-g^n_c\left(h^n_{i-1},h^n_i,x_{i-\frac{1}{2}}\right)\right] \\ &= h^n_i-\nu_i\left[g^n_{i+\frac{1}{2}}\left(h^n_i,h^n_{i+1}\right)-g^n_{i-\frac{1}{2}}\left(h^n_{i-1},h^n_i\right)\right] \\ &= h^n_i-\nu_i\left(F^n_{i+\frac{1}{2}}-F^n_{i-\frac{1}{2}}\right)=h^{n+1}_i,\ \ \forall i\in[N_x]^\star,\ \ \forall n\in[T], \end{aligned}$$

then $H^n_{c,i}$ is non-decreasing w.r.t. its all three variables $H^n_{c,i}(\nearrow,\nearrow,\nearrow)$.

As a consequence, the restricted DG-staggered upwind scheme (rDGsc) preserves the nonnegativity of water height, i.e.: If the initial water height is nonnegative everywhere in the computational horizontal fluid domain Ω_c, i.e., $h_0\geq 0$ a.e. in Ω_c, then it remains nonnegative in the discrete level, i.e.,

$$h_{c,\mathcal{T},\Delta t}(t,x)\geq 0,\ \ a.e.\ \ in\ \ [0,T_0)\times\Omega_c,$$

where

$$h_{c,\mathcal{T},\Delta t}(t,x) := \sum_{i=1}^{N_x}\sum_{n\in[T-1]} h^n_i\mathbf{1}_{[t^n,t^{n+1})\times C_i}(t,x),\ \ in\ \ [0,T_0)\times\Omega_c.$$

Analogous to Proposition 3.5, (rDGscfb) also preserves the positivity of water height provided, in addition, the numerical fluxes on the boundaries $\partial\Omega_c$ is positive.

Proposition 3.12 (Positivity conservation of water height)**.** *Assume that the initial data satisfies $(h_0,u_0)\in L^\infty\left(\Omega_c;\mathbb{R}^\star_+\right)\times L^\infty\left(\Omega_c\right)$. Let $\mathcal{T}$ be an admissible mesh of Ω_c and $\Delta t\in\mathbb{R}^\star_+$ be the time step. Let $(h^n_i)_{i\in[N_x]^\star,n\in[T]}$ and $\left(u^n_{i+\frac{1}{2}}\right)_{i\in[N_x],n\in[T]}$ be the discrete finite volume approximate water height and velocity, respectively, generated by (rDGsc).*

If the numerical fluxes on the boundaries $\partial\Omega_c$ are chosen to be positive, i.e.,

$$\overline{h}(t) > 0, \ \ \overline{\overline{h}}(t) > 0, \ \ \forall t \in [0, T_0),$$

and the following strict restricted CFL-like condition holds

$$\Delta t < \frac{\inf_{i\in\mathbb{Z}} \Delta x_i}{\sup\limits_{i\in[N_x]^\star, n\in[T]} \left(\left(u^n_{i+\frac{1}{2}}\right)_+ - \left(u^n_{i-\frac{1}{2}}\right)_-\right)}, \tag{srCFL3}$$

or stronger

$$\nu < \frac{\alpha_{\mathcal{T}}}{\sup\limits_{i\in[N_x]^\star, n\in[T]} \left(\left(u^n_{i+\frac{1}{2}}\right)_+ - \left(u^n_{i-\frac{1}{2}}\right)_-\right)}. \tag{srCFL4}$$

Then the restricted DG-staggered upwind scheme (rDGsc) preserves the positivity of water height, i.e.: If the initial water height is positive everywhere in the computational horizontal fluid domain Ω_c, i.e., $h_0 > 0$ a.e. in Ω_c, then it remains positive in the discrete level, i.e.,

$$h_{c,\mathcal{T},\Delta t}(t, x) > 0, \ \ a.e. \ \ in \ \ [0, T_0) \times \Omega_c.$$

Remark 3.5 (L^∞-estimate fails for staggered schemes)**.** *The restricted staggered upwind scheme (rDGsc) also does not bound the water height by the same bound used for the initial water height, i.e.,*

$$h_0 \in [h_m, h_M] \ \ a.e. \ \ in \ \Omega_c \not\Rightarrow h_{c,\mathcal{T},\Delta t}(t, x) \in [h_m, h_M] \ \ a.e. \ \ in \ \ [0, T_0) \times \Omega_c.$$

Analogous to Proposition 3.6, (rDGsc) also possesses a well-balanced property for any restricted horizontal fluid domain being fully wet. For sake of convenience, we define the restricted dry- and wet-components of Ω_c as

$$D^t_c := D^t \cap \Omega_c, \ \ W^t_c := W^t \cap \Omega_c, \ \ \forall t \in [0, T_0).$$

Thus,

$$D^t_c \cap W^t_c = \emptyset, \ \ D^t_c \cup W^t_c = \Omega_c, \ \ \forall t \in [0, T_0).$$

Proposition 3.13 (Well-balanced property for a fully wet restricted horizontal fluid domain Ω_c)**.** *Assume that the initial data satisfies $(h_0, u_0) \in L^\infty\left(\Omega_c; \mathbb{R}^\star_+\right) \times L^\infty(\Omega_c)$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}^\star_+$ be the time step. Let $(h^n_i)_{i\in[N_x]^\star, n\in[T]}$ and $\left(u^n_{i+\frac{1}{2}}\right)_{i\in[N_x], n\in[T]}$ be the discrete finite volume approximate water height and velocity, respectively, generated by (rDGsc). Assume that the strict restricted CFL-like condition (srCFL3) or (srCFL4) holds. The numerical fluxes on the boundaries $\partial\Omega_c$ are also chosen to be positive.*

Then the positivity of the initial water height and the strict restricted CFL-like condition imply that the computational horizontal fluid domain Ω_c is fully wet during the survival time of (1DSW), i.e., $D^t_c = \emptyset$ and $W^t_c = \mathbb{R}$ for all $t \in [0, T_0)$, the *still water steady states* (sw) *are preserved under the restricted DG-staggered upwind scheme* (rDGsc), *i.e., if for some* $n \in [T-1]$, $u^n_{i+\frac{1}{2}} = 0$ *and* $h^n_i + b_i = C$ *for all* $i \in [N_x]^\star$ *for some constant* C, *then* $u^{n+1}_{i+\frac{1}{2}} = 0$ *and* $h^{n+1}_i + b_i = C$ *for all* $i \in [N_x]^\star$.

Proposition 3.13 indicates the well-balanced property which is the discrete version of (sw) for (rDGsc) in the case of the restricted horizontal fluid domain Ω_c being fully wet. Unfortunately, as in the case of horizontal fluid domain $\mathbb{R}$, if the dry shore, or dry areas are allowed in Ω_c, this kind of well-balanced property collapses completely, especially in the dry-wet transitions.

Analogous to the sets of dry- and wet-indices in the discrete level defined before, their restricted counterparts are defined as

$$I^n_{c,D} := I^n_D \cap [N_x]^\star, \;\; I^n_{c,W} := I^n_W \cap [N_x]^\star, \;\; \forall n \in [T],$$

and the nonnegativity conservation of water height stated in Proposition (3.7) implies that

$$[N_x]^\star = I^n_{c,D} \cap I^n_{c,W}, \;\; \forall n \in [T].$$

As another direct consequence, combining the discrete boundary condition and Convention 3.2 yields

$$u^n_{i+\frac{1}{2}} := 0, \;\; \forall i \text{ s.t. } (i=0) \vee (i=N_x) \vee \left(\left(i \in I^n_{c,D}\right) \wedge \left(i+1 \in I^n_{c,D}\right)\right), \;\; \forall n \in [T].$$

Analogous to Proposition 3.7, the following Proposition also demonstrates the "ill-balanced" property in the case of Ω_c being dry and wet in different parts at the same time.

Proposition 3.14 (Ill-balanced property for dry-wet restricted horizontal fluid domain Ω_c)**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty(\Omega_c; \mathbb{R}_+) \times L^\infty(\Omega_c)$. *Let* $\mathcal{T}$ *be an admissible mesh of* Ω_c *and* $\Delta t \in \mathbb{R}^\star_+$ *be the time step. Let* $(h^n_i)_{i\in[N_x]^\star, n\in[T]}$ *and* $\left(u^n_{i+\frac{1}{2}}\right)_{i\in[N_x], n\in[T]}$ *be the discrete finite volume approximate water height and velocity, respectively, generated by* (rDGsc). *Assume that the restricted CFL-like condition* (rCFL$_3$) *or* (rCFL$_4$) *holds. The numerical fluxes on the boundaries are chosen to be nonnegative.*

Assume that the restricted horizontal fluid domain Ω_c *contains at least one nondegenerate dry area (not 1-point dry area type) but not fully dry during the survival time of* (1DSW), *i.e.,*

$$\operatorname{card}\left(D^{t_0}_c\right) = \mathfrak{c}, \;\; W^{t_0}_c \neq \emptyset, \;\; \textit{for some } t_0 \in [0, T_0),$$

the still water steady states (sw) *are not preserved under the restricted DG-staggered upwind scheme* (rDGsc) *in general.*

At each time step, the Courant number is defined by

$$\widetilde{\nu} := \max_{i\in[N_x]^\star} \left(\frac{\nu_i}{2h^n_i} \left| F^n_{i+\frac{1}{2}} + F^n_{i-\frac{1}{2}} \right| + \nu_i \sqrt{g h^n_i} \right),$$

and for uniform meshes, it reads

$$\widetilde{\nu} = \nu \max_{i\in[N_x]^\star} \left(\frac{1}{2h_i^n} \left| F^n_{i+\frac{1}{2}} + F^n_{i-\frac{1}{2}} \right| + \sqrt{gh_i^n} \right).$$

The numerical simulations in Doyen and Gunawan, 2014, pp. 232–234 show that the staggered scheme is stable under the CFL condition $\widetilde{\nu} < 1$.

2.2 Passing to the limit in the DG staggered upwind scheme in the computational horizontal fluid domain

Analogous to Proposition (3.8), the following proposition illustrates a balance on the kinetic energy and a balance on the and topography energy.

Proposition 3.15 (Balance on kinetic energy, potential energy and topography energy)**.** *Assume that the initial data satisfies $(h_0, u_0) \in L^\infty\left(\Omega_c; \mathbb{R}^\star_+\right) \times L^\infty\left(\Omega_c\right)$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}^\star_+$ be the time step. Let $(h_i^n)_{i\in[N_x]^\star, n\in[T]}$ and $\left(u^n_{i+\frac{1}{2}}\right)_{i\in[N_x], n\in[T]}$ be the discrete finite volume approximate water height and velocity, respectively, generated by (rDGsc). Assume that the strict restricted CFL-like condition (srCFL3) or (srCFL4) holds. The numerical fluxes on the boundaries $\partial\Omega_c$ are also chosen to be positive.*

The discrete solution of the staggered upwind scheme (rDGsc) satisfies, for all $i \in [N_x]^\star$ and all $n \in [T-1]$, the balance

$$\begin{aligned}\frac{1}{2\nu_{i+\frac{1}{2}}} \left(\bar{h}^{n+1}_{i+\frac{1}{2}} \left(u^{n+1}_{i+\frac{1}{2}}\right)^2 - \bar{h}^n_{i+\frac{1}{2}} \left(u^n_{i+\frac{1}{2}}\right)^2 \right) + \frac{1}{2} \left(\overline{F}^n_{i+1} (u^n_{i+1})^2 - \overline{F}^n_i (u^n_i)^2 \right) \\ + \frac{g}{2} \left((h^{n+1}_{i+1})^2 - (h^{n+1}_i)^2 \right) u^{n+1}_{i+\frac{1}{2}} + g\bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} (b_{i+1} - b_i) = -R^{n+1}_{i+\frac{1}{2}}, \qquad \text{(rBKE)}\end{aligned}$$

with

$$\begin{aligned}R^{n+1}_{i+\frac{1}{2}} := {} & \frac{1}{2\nu_{i+\frac{1}{2}}} \bar{h}^{n+1}_{i+\frac{1}{2}} \left(u^{n+1}_{i+\frac{1}{2}} - u^n_{i+\frac{1}{2}} \right)^2 \\ & + \frac{1}{2} \left[(\overline{F}^n_{i+1})_- \left(u^n_{i+\frac{3}{2}} - u^n_{i+\frac{1}{2}} \right)^2 - (\overline{F}^n_i)_+ \left(u^n_{i+\frac{1}{2}} - u^n_{i-\frac{1}{2}} \right)^2 \right] \qquad \text{(rBKEr)} \\ & - \left[(\overline{F}^n_{i+1})_- \left(u^n_{i+\frac{3}{2}} - u^n_{i+\frac{1}{2}} \right) + (\overline{F}^n_i)_+ \left(u^n_{i+\frac{1}{2}} - u^n_{i-\frac{1}{2}} \right) \right] \left(u^{n+1}_{i+\frac{1}{2}} - u^n_{i+\frac{1}{2}} \right),\end{aligned}$$

and, for all $i \in [N_x]^\star$ and all $n \in [T-1]$, the balance

$$\begin{aligned}\frac{1}{\nu_i} \left(\frac{g}{2} (h^{n+1}_i)^2 - \frac{g}{2} (h^n_i)^2 \right) + \left(\frac{g}{2} \left(h^n_{i+\frac{1}{2}} \right)^2 u^n_{i+\frac{1}{2}} - \frac{g}{2} \left(h^n_{i-\frac{1}{2}} \right)^2 u^n_{i-\frac{1}{2}} \right) \\ + \frac{g}{2} (h^n_i)^2 \left(u^n_{i+\frac{1}{2}} - u^n_{i-\frac{1}{2}} \right) = -R^{n+1}_i, \qquad \text{(rBPETE)}\end{aligned}$$

with

$$R^{n+1}_i := \frac{g}{2\nu_i} \left(h^{n+1}_i - h^n_i \right)^2 + \frac{1}{2} \left(h^n_{i+1} - h^n_i \right)^2 \left(u^n_{i+\frac{1}{2}} \right)_-$$

$$-\frac{1}{2}\left(h_{i-1}^{n}-h_{i}^{n}\right)^{2}\left(u_{i-\frac{1}{2}}^{n}\right)_{+}+g\left(h_{i}^{n+1}-h_{i}^{n}\right)\left(\overline{F}_{i+\frac{1}{2}}^{n}-\overline{F}_{i-\frac{1}{2}}^{n}\right). \qquad \text{(rBPETEr)}$$

Now, let a sequence of discretizations $\left(\mathcal{T}_c^{(m)},\Delta t^{(m)}\right)_{m\in\mathbb{N}}$ be given, where for each $m\in\mathbb{N}$, $\mathcal{T}_c^{(m)}$ is an admissible mesh in the sense of Definition B.1 which is "perfectly" restricted on the computational horizontal fluid domain Ω_c, i.e.,

$$\mathcal{T}_c^{(m)}:=\left\{C_i^{(m)};i\in\left[N_x^{(m)}\right]^{\star}\right\},\ \ \forall m\in\mathbb{N},\ \text{where}\ N_x^{(m)}\to\infty\ \text{as}\ m\to\infty,$$

where

$$C_i^{(m)}:=\left(x_{i-\frac{1}{2}}^{(m)},x_{i+\frac{1}{2}}^{(m)}\right),\quad \Delta x_i^{(m)}:=x_{i+\frac{1}{2}}^{(m)}-x_{i-\frac{1}{2}}^{(m)},\ \ \forall i\in\left[N_x^{(m)}\right]^{\star},\ \ \forall m\in\mathbb{N},$$

and

$$x_{\frac{1}{2}}^{(m)}=0,\ \ x_{N_x^{(m)}+\frac{1}{2}}^{(m)}=L,\ \ \sum_{i=1}^{N_x^{(m)}}\Delta x_i^{(m)}=x_{N_x^{(m)}+\frac{1}{2}}^{(m)}-x_{\frac{1}{2}}^{(m)}=L,\ \ \forall m\in\mathbb{N}.$$

The discrete numerical fluxes at the boundaries $\partial\Omega_c$ are given by

$$\left(h^{(m)}\right)_0^n:=\left(\overline{h}^{(m)}\right)^n:=\frac{1}{\Delta t^{(m)}}\int_{n\Delta t^{(m)}}^{(n+1)\Delta t^{(m)}}\overline{h}\left(t\right)dt,\ \ \forall n\in\left[T-1\right],\ \ \forall m\in\mathbb{N},$$

$$\left(h^{(m)}\right)_{N_x+1}^n:=\left(\overline{\overline{h}}^{(m)}\right)^n:=\frac{1}{\Delta t^{(m)}}\int_{n\Delta t^{(m)}}^{(n+1)\Delta t^{(m)}}\overline{\overline{h}}\left(t\right)dt,\ \ \forall n\in\left[T-1\right],\ \ \forall m\in\mathbb{N}.$$

The size mesh $\Delta x^{(m)}$ of the mesh $\mathcal{T}_c^{(m)}$ by

$$\Delta x^{(m)}:=\sup_{i\in\left[N_x^{(m)}\right]^{\star}}\Delta x_i^{(m)},\ \ \forall m\in\mathbb{N}.$$

Let $\left(h^{(m)},u^{(m)}\right)$ be the solution given by (rDGsc) with the admissible $\mathcal{T}_c^{(m)}$ and the time step $\Delta t^{(m)}$, i.e.,

$$h^{(m)}:=h_{\mathcal{T}_c^{(m)},\Delta t^{(m)}},\ \ u^{(m)}:=u_{\mathcal{T}_c^{(m)},\Delta t^{(m)}},\ \ \forall m\in\mathbb{N},$$

where

$$h_{\mathcal{T}_c^{(m)},\Delta t^{(m)}}\left(t,x\right):=\sum_{i=1}^{N_x}\sum_{n\in[T-1]}\left(h^{(m)}\right)_i^n\mathbf{1}_{\left[n\Delta t^{(m)},(n+1)\Delta t^{(m)}\right)\times C_i^{(m)}}\left(t,x\right),\ \ \text{in}\ \left[0,T_0\right)\times\mathbb{R},$$

$$u_{\mathcal{T}_c^{(m)},\Delta t^{(m)}}\left(t,x\right):=\sum_{i=0}^{N_x}\sum_{n\in[T-1]}\left(u^{(m)}\right)_{i+\frac{1}{2}}^n\mathbf{1}_{\left[n\Delta t^{(m)},(n+1)\Delta t^{(m)}\right)\times C_{i+\frac{1}{2}}^{(m)}}\left(t,x\right),\ \ \text{in}\ \left[0,T_0\right)\times\mathbb{R},$$

and where

$$C_{i+\frac{1}{2}}^{(m)}:=\left(x_i^{(m)},x_{i+1}^{(m)}\right),\ \ \forall i\in\left[N_x^{(m)}\right],\ \ \forall m\in\mathbb{N}.$$

For discrete functions k and v defined on the primal and dual meshes, respectively, we define a discrete $L^1\left([0,T_0);BV\left(\Omega_c\right)\right)$ norm by

$$\|k\|_{\mathcal{T}_c,x,BV} := \sum_{n\in[T]} \Delta t \sum_{i=1}^{N_x^{(m)}-1} \left| \left(k^{(m)}\right)_{i+1}^n - \left(k^{(m)}\right)_i^n \right|,$$

$$\|v\|_{\mathcal{T}_c,x,BV} := \sum_{n\in[T]} \Delta t \sum_{i=1}^{N_x^{(m)}} \left| \left(v^{(m)}\right)_{i+\frac{1}{2}}^n - \left(v^{(m)}\right)_{i-\frac{1}{2}}^n \right|,$$

For the consistency results analogous to Theorem (3.1) and Theorem (3.2), we have to assume that a sequence of discrete solutions $\left(h^{(m)},u^{(m)}\right)_{m\in\mathbb{N}}$ satisfies $h^{(m)}>0$ (Proposition 3.12) guarantees the positivity conservation of water height provided the mentioned strict CFL-like condition holds and the initial water height is positive everywhere in the computational horizontal fluid domain Ω_c) and is uniformly bounded in $L^\infty([0,T_0)\times\Omega_c)^2$, i.e.,

$$0<\left(h^{(m)}\right)_i^n \le C,\ \ \forall i\in\left[N_x^{(m)}\right]^\star,\ \ \forall n\in\left[T^{(m)}\right],\ \text{where } T^{(m)} := \left\lfloor \frac{T_0}{\Delta t^{(m)}} \right\rfloor,\ \ \forall m\in\mathbb{N}, \tag{3.5}$$

and

$$\left|\left(u^{(m)}\right)_{i+\frac{1}{2}}^n\right| \le C,\ \ \forall i\in\left[N_x^{(m)}\right]^\star,\ \ \forall n\in\left[T^{(m)}\right],\ \ \forall m\in\mathbb{N}, \tag{3.6}$$

where C is a positive real number. These inequalities imply that the function h_0 and u_0 belong to $L^\infty\left(\Omega_c\right)$. We also have to assume that a sequence of discrete solutions satisfies the following uniform bounds w.r.t. the discrete BV-norms:

$$\left\|h^{(m)}\right\|_{\mathcal{T}_c,x,BV} + \left\|u^{(m)}\right\|_{\mathcal{T}_c,x,BV} \le C,\ \ \forall m\in\mathbb{N}, \tag{3.7}$$

and

$$\left\|h^{(m)}\right\|_{\mathcal{T}_c,t,BV} \le C,\ \ \forall m\in\mathbb{N}. \tag{3.8}$$

A weak solution to the continuous problem (1DSW) restricted on Ω_c satisfies

$$\begin{cases} -\displaystyle\int_0^{T_0}\int_{\Omega_c} \left(h\partial_t\varphi + hu\partial_x\varphi\right) dxdt - \int_{\Omega_c} h_0\left(x\right)\varphi\left(0,x\right)dx = 0, \\ -\displaystyle\int_0^{T_0}\int_{\Omega_c} \left[hu\partial_t\varphi + \left(hu^2+\frac{g}{2}h^2\right)\partial_x\varphi\right] dxdt - \int_{\Omega_c} \left(h_0u_0\right)\left(x\right)\varphi\left(0,x\right)dx = 0, \end{cases} \tag{rWF1DSW}$$

for all $\varphi\in C_c^\infty\left([0,T_0)\times\Omega_c\right)$.

Note that (rWF1DSW) is insufficient to define a weak solution to (1DSW) restricted in Ω_c, since it does not imply anything about the boundary conditions. However, (rWF1DSW) allows to derive the Rankine-Hugoniot conditions. Hence if (RH) is satisfied by the limit of a sequence of solutions to the discrete problem, roughly speaking, *the restricted staggered upwind scheme* (rDGsc) *computes correct shocks.*

Theorem 3.3 (Consistency of (rDGsc))**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty\left(\Omega_c; \mathbb{R}_+^\star\right) \times L^\infty\left(\Omega_c\right)$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $(h_i^n)_{i\in[N_x]^\star, n\in[T]}$ *and* $\left(u_{i+\frac{1}{2}}^n\right)_{i\in[N_x], n\in[T]}$ *be the discrete finite volume approximate water height and velocity, respectively, generated by* (rDGsc)*. Assume that the strict restricted CFL-like condition* (srCFL3) *or* (srCFL4) *holds. The numerical fluxes on the boundaries* $\partial\Omega_c$ *are also chosen to be positive.*

Assume that the bottom topography is flat. Let $\left(\mathcal{T}_c^{(m)}, \Delta t^{(m)}\right)_{m\in\mathbb{N}}$ *be a sequence of discretizations such that both the time step* $\Delta t^{(m)}$ *and the size* $\Delta x^{(m)}$ *of the mesh* $\mathcal{T}^{(m)}$ *tend to zero as* $m \to \infty$*, and let* $\left(h^{(m)}, u^{(m)}\right)_{m\in\mathbb{N}}$ *be the corresponding sequence of solutions. If this sequence satisfies the estimates* (3.5)*-*(3.7) *and converges in* $L^p\left([0,T_0) \times \Omega_c; \mathbb{R}_+^\star\right) \times L^p\left([0,T_0) \times \Omega_c\right)$*, for* $1 \le p < \infty$*, to* $\left(\bar{h}, \bar{u}\right) \in L^\infty\left([0,T_0) \times \Omega_c; \mathbb{R}_+^\star\right) \times L^\infty\left([0,T_0) \times \Omega_c\right)$*. Then the limit* $\left(\bar{h}, \bar{u}\right)$ *satisfies* (rWF1DSW)*.*

We now turn to the entropy inequality (EIb), especially its integral form restricted in the computational horizontal fluid domain Ω_c:

$$\int_0^{T_0} \int_{\Omega_c} \left(-\Phi_b\left(U_b\right)\partial_t\varphi - \Psi_b\left(U_b\right)\partial_x\varphi\right) dxdt - \int_{\Omega_c} \Phi_b\left(U_{b,0}\right)\varphi\left(0,x\right) dx \le 0, \qquad \text{(rWEI)}$$

for all $\varphi \in C_c^\infty\left([0,T_0) \times \Omega_c; \mathbb{R}_+\right)$.

Now, we need to introduce the following additional condition for a regular sequence of discretizatons:

$$\lim_{m\to\infty} \frac{\Delta t^{(m)}}{\min\limits_{i\in[N_x]^\star} \Delta x_i^{(m)}} = 0. \qquad \text{(rlCFL)}$$

We are now ready to state the final result for (rDGsc).

Theorem 3.4 (Entropy consistency)**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty\left(\Omega_c; \mathbb{R}_+^\star\right) \times L^\infty\left(\Omega_c\right)$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $(h_i^n)_{i\in[N_x]^\star, n\in[T]}$ *and* $\left(u_{i+\frac{1}{2}}^n\right)_{i\in[N_x], n\in[T]}$ *be the discrete finite volume approximate water height and velocity, respectively, generated by* (rDGsc)*. Assume that the strict restricted CFL-like condition* (srCFL3) *or* (srCFL4) *holds. The numerical fluxes on the boundaries* $\partial\Omega_c$ *are also chosen to be positive.*

Assume that the bottom topography is flat. Let $\left(\mathcal{T}_c^{(m)}, \Delta t^{(m)}\right)_{m\in\mathbb{N}}$ *be a sequence of discretizations such that both the time step* $\Delta t^{(m)}$ *and the size* $\Delta x^{(m)}$ *of the mesh* $\mathcal{T}^{(m)}$ *tend to zero as* $m \to \infty$*, and let* $\left(h^{(m)}, u^{(m)}\right)_{m\in\mathbb{N}}$ *be the corresponding sequence of solutions. If this sequence satisfies the estimates* (3.5)*-*(3.7) *and converges in* $L^p\left([0,T_0) \times \Omega_c; \mathbb{R}_+^\star\right) \times L^p\left([0,T_0) \times \Omega_c\right)$*, for* $1 \le p < \infty$*, to* $\left(\bar{h}, \bar{u}\right) \in L^\infty\left([0,T_0) \times \Omega_c; \mathbb{R}_+^\star\right) \times L^\infty\left([0,T_0) \times \Omega_c\right)$*. If we suppose in addition that the considered sequence of discretization satisfies* (rlCFL)*, the sequence of solutions satisfies* (3.8)*, then the limit* $\left(\bar{h}, \bar{u}\right)$ *satisfies the restricted entropy condition* (rWEI)*.*

Remark 3.6 (On BV-stability assumptions)**.** *The mentioned proof of Theorem* (3.3) *shows that the scheme proposed in Herbin, Latché, and Nguyen, 2013 is consistent* *under a BV-stability assumption much weaker than (3.7), namely,*

$$\lim_{m\to\infty} \Delta x^{(m)} \left(\left\| h^{(m)} \right\|_{\mathcal{T}_c,x,BV} + \left\| u^{(m)} \right\|_{\mathcal{T}_c,x,BV} \right) = 0.$$

The proof of Theorem 3.4, i.e., the limit of convergent sequences is an entropy solution, is completely different, since the stability condition which is used in the proof is

$$\lim_{m\to\infty} \frac{\Delta t^{(m)}}{\min\limits_{i\in[N_x]^\star} \Delta x_i^{(m)}} \left\| h^{(m)} \right\|_{\mathcal{T}_c,x,BV} = 0.$$

2.3 Higher-order staggered upwind schemes on the computational horizontal fluid domain

Everything in Chapter 3, Subsection 1.5 can be restricted in Ω_c by replacing $i \in \mathbb{Z}$ by $i \in [N_x]^\star$ with some suitable choices of the numerical fluxes $\overline{h}, \overline{\overline{h}}$ on the boundaries $\partial\Omega_c$. It should be emphasized that higher-order Runge-Kutta method can be used to derive other staggered upwind schemes possessing better properties in time-discretization.

Chapter 4

New Offset Equilibrium Staggered Schemes

This chapter is devoted to study a new *offset equilibrium* staggered scheme for the 1D shallow water equations (1DSW). It suffices to modify in the upwind decenter scheme (DGsc) only the discretization of velocity (unlike collocated schemes).

In the case where the horizontal fluid domain $\mathbb{R}$ includes transitions between zones dry and wet areas, we can draw inspiration from the scheme developed in Chen and Noelle, 2017.

1 Description of a new offset equilibrium scheme in the horizontal fluid domain

In this section, we discuss a modification of the staggered upwind scheme (DGsc) proposed in Doyen and Gunawan, 2014 on staggered meshes. While the framework is presented in the 1D case, the method can be adapted to higher dimensions.

Reuse the settings described in Chapter 3, Section 1.1. We also reuse the discrete mass conservation equation (dmc). As an immediate consequence, by Proposition 3.1, our new scheme will preserve the water volume.

The difference between this new scheme and the staggered upwind scheme (DGsc) lies in the discrete momentum balance equation.

Define for all $i \in \mathbb{Z}$, $n \in [T]$,

$$\begin{cases} w_i^n := h_i^n + b_i, \\ b_{i+\frac{1}{2}}^n := \min\left(\max\left(b_i, b_{i+1}\right), \min\left(w_i^n, w_{i+1}^n\right)\right), \\ h_{i+\frac{1}{2}-}^n := \min\left(w_i^n - b_{i+\frac{1}{2}}^n, h_i^n\right), \\ h_{i+\frac{1}{2}+}^n := \min\left(w_{i+1}^n - b_{i+\frac{1}{2}}^n, h_{i+1}^n\right), \end{cases} \tag{4.1}$$

from which we deduce the discretization of the equation in velocity with an implicit discretization of the pressure term:

$$\bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} := \bar{h}^{n}_{i+\frac{1}{2}} u^{n}_{i+\frac{1}{2}} - \nu_{i+\frac{1}{2}} \left[G^n_{i+1} - G^n_i + \frac{g}{2} \left(h^{n+1}_i + h^{n+1}_{i+1} \right) \left(h^{n+1}_{i+\frac{1}{2}+} - h^{n+1}_{i+\frac{1}{2}-} \right) \right], \tag{OEdmb}$$

for all $i \in \mathbb{Z}$, $n \in [N-1]$. The Convention 3.1 is also reused for this new scheme.

Therefore, the new offset equilibrium staggered scheme just established reads

$$\begin{cases} h^{n+1}_i := h^n_i - \nu_i \left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}} \right), \\ \bar{h}^{n+1}_{i+\frac{1}{2}} u^{n+1}_{i+\frac{1}{2}} := \bar{h}^{n}_{i+\frac{1}{2}} u^{n}_{i+\frac{1}{2}} - \nu_{i+\frac{1}{2}} \left[G^n_{i+1} - G^n_i + \frac{g}{2} \left(h^{n+1}_i + h^{n+1}_{i+1} \right) \left(h^{n+1}_{i+\frac{1}{2}+} - h^{n+1}_{i+\frac{1}{2}-} \right) \right], \end{cases} \tag{OEsc}$$

for all $i \in \mathbb{Z}$, $n \in [T-1]$.

Follow Chen and Noelle, 2017, pp. 760–761, there will be two main types of interfaces, depending on how the water covers the bottom to the left and right sides of the interface.

- *Fully wet interface*: where the water level on each side is higher than the higher side of the bottom topography

$$\min\left(w_i, w_{i+1}\right) > \max\left(b_i, b_{i+1}\right).$$

- *Partially wet interface*: where the water level on one side is equal to or below the topography on the other side:

$$\min\left(w_i, w_{i+1}\right) \leq \max\left(b_i, b_{i+1}\right).$$

Since (OEsc) shares the discrete mass conservation equation (dmc) in common with (DGsc), (OEsc) also inherits Proposition 3.4, (3.5) in the horizontal fluid domain $\mathbb{R}$. Hence, suppose that the initial water height is nonnegative, i.e., $h_0 \geq 0$ a.e. in $\mathbb{R}$ and the CFL-like condition (CFL3) or (CFL4) holds, then the water height remains nonnegative in the discrete level during $[0, T_0)$.

As a consequence, in the flat bottom topography case $b = 0$, (4.1) becomes

$$\begin{cases} w^n_i &:= h^n_i, \\ b^n_{i+\frac{1}{2}} &:= \min\left(0, \min\left(w^n_i, w^n_{i+1}\right)\right), \\ h^n_{i+\frac{1}{2}-} &:= \min\left(w^n_i - b^n_{i+\frac{1}{2}}, h^n_i\right), \\ h^n_{i+\frac{1}{2}+} &:= \min\left(w^n_{i+1} - b^n_{i+\frac{1}{2}}, h^n_{i+1}\right), \end{cases}$$

or equivalently,

$$\begin{cases} w_i^n := h_i^n, \\ b_{i+\frac{1}{2}}^n := \min\left(0, \min\left(h_i^n, h_{i+1}^n\right)\right) = 0, \\ h_{i+\frac{1}{2}-}^n := \min\left(h_i^n - b_{i+\frac{1}{2}}^n, h_i^n\right) = h_i^n, \\ h_{i+\frac{1}{2}+}^n := \min\left(h_{i+1}^n - b_{i+\frac{1}{2}}^n, h_{i+1}^n\right) = h_{i+1}^n, \end{cases}$$

and (OEsc) becomes exactly (DGscfb).

1.1 Conservation properties of a new offset equilibrium scheme on the horizontal fluid domain

Analogous to Proposition 3.2, the following result demonstrates that the offset equilibrium staggered upwind scheme (OEsc) preserves some physical quantities in the discrete level as (DGsc) does.

Proposition 4.1 (Preserved quantities under (OEsc))**.** *Assume that the initial data satisfies $(h_0, q_0) \in (L^\infty(\mathbb{R}))^2$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by* (OEsc)*.*

The offset equilibrium staggered upwind scheme (OEsc) *preserves the water volume, the total mass in the discrete level. Moreover, for the flat bottom topography case,* (OEsc) *also preserves the total momentum in the discrete level.*

Analogous to Proposition 3.3, (OEsc) does not preserve the total horizontal impulse and the total energy.

Proposition 4.2 (Non-preserved quantities under (OEsc))**.** *Assume that the initial data satisfies $(h_0, q_0) \in (L^\infty(\mathbb{R}))^2$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by* (OEsc)*.*

The offset equilibrium staggered upwind scheme (OEsc)*, in the flat bottom topography case, does not preserve the total horizontal impulse and also the total energy in the discrete level with the choices of the discrete horizontal impulse* (dhi) *and the discrete total energy* (dte_1) *or* (dte_2)*.*

Again, since (OEsc) and (DGsc) share in common the discrete mass conservation (dmc), Proposition (3.4) and Proposition (3.5) also hold for (OEsc).

Proposition 4.3 (Nonnegativity and positivity conservation of water height)**.** *Assume that the initial data satisfies $(h_0, u_0) \in L^\infty(\mathbb{R}; \mathbb{R}_+) \times L^\infty(\mathbb{R})$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by* (OEsc)*.*

If the CFL-like condition (CFL3) *or* (CFL4) *holds then* (OEsc) *is a monotone scheme. As a consequence, if the initial water height is nonnegative everywhere in the horizontal fluid domain $\mathbb{R}$ then it remains nonnegative in the discrete level.*

If the strict CFL-like condition (sCFL$_3$) *or* (sCFL$_4$) *holds and if the initial water height is positive everywhere in the horizontal fluid domain* $\mathbb{R}$*, then it remains positive in the discrete level.*

It should be emphasize that in the case of the horizontal fluid domain being fully wet, (OEsc) and (DGsc) coincides. Thus, Proposition 3.6 also holds for (OEsc) in this best-case.

Proposition 4.4 (Well-balanced property for fully wet horizontal fluid domain $\mathbb{R}$)**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty\left(\mathbb{R}; \mathbb{R}_+^\star\right) \times L^\infty(\mathbb{R})$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by* (OEsc)*. Assume that the strict CFL-like condition* (sCFL$_3$) *or* (sCFL$_4$) *holds.*

Then the posivitity of the initial water height and the strict CFL-like condition imply that the horizontal fluid domain $\mathbb{R}$ *is fully wet during the survival time of* (1DSW)*, i.e.,* $D^t = \emptyset$ *and* $W^t = \mathbb{R}$ *for all* $t \in [0, T_0)$*, the still water steady states* (sw) *are preserved under the offset equilibrium staggered upwind scheme* (OEsc) *in the discrete level, i.e., if for some* $n \in [T-1]$*,* $u_{i+\frac{1}{2}}^n = 0$ *and* $h_i^n + b_i = C$ *for all* $i \in \mathbb{Z}$ *for some constant* C*, then* $u_{i+\frac{1}{2}}^{n+1} = 0$ *and* $h_i^{n+1} + b_i = C$ *for all* $i \in \mathbb{Z}$*.*

Proof. Suppose that, for some $n_0 \in [T-1]$, $u_{i+\frac{1}{2}}^{n_0} = 0$ and $h_i^{n_0} + b_i = C$ for all $i \in \mathbb{Z}$ and for some constant C, then (4.1) become (note the assumption fully wet and Proposition 4.3 imply that $h_i^n > 0$ generated by (OEsc) for all $i \in \mathbb{Z}$ and $n \in [T]$)

$$\begin{cases} w_i^{n_0} = h_i^{n_0} + b_i = C \geq b_i, \\ b_{i+\frac{1}{2}}^{n_0} = \min\left(\max\left(b_i, b_{i+1}\right), C\right) = \max\left(b_i, b_{i+1}\right), \\ h_{i+\frac{1}{2}-}^{n_0} = \min\left(w_i^{n_0} - \max\left(b_i, b_{i+1}\right), h_i^{n_0}\right) = w_i^{n_0} - \max\left(b_i, b_{i+1}\right), \\ h_{i+\frac{1}{2}+}^{n_0} = \min\left(w_{i+1}^{n_0} - \max\left(b_i, b_{i+1}\right), h_{i+1}^{n_0}\right) = w_{i+1}^{n_0} - \max\left(b_i, b_{i+1}\right), \end{cases}$$

and then (OEsc) at $n = n_0$ becomes

$$\begin{aligned} h_i^{n_0+1} &:= h_i^{n_0} - \nu_i\left(F_{i+\frac{1}{2}}^{n_0} - F_{i-\frac{1}{2}}^{n_0}\right) = h_i^{n_0} - \nu_i\left(h_{i+\frac{1}{2}}^{n_0} u_{i+\frac{1}{2}}^{n_0} - h_{i-\frac{1}{2}}^{n_0} u_{i-\frac{1}{2}}^{n_0}\right) = h_i^{n_0}, \\ \bar{h}_{i+\frac{1}{2}}^{n_0+1} u_{i+\frac{1}{2}}^{n_0+1} &:= \bar{h}_{i+\frac{1}{2}}^{n_0} u_{i+\frac{1}{2}}^{n_0} - \nu_{i+\frac{1}{2}}\left[G_{i+1}^{n_0} - G_i^{n_0} + \frac{g}{2}\left(h_i^{n_0+1} + h_{i+1}^{n_0+1}\right)\left(h_{i+\frac{1}{2}+}^{n_0+1} - h_{i+\frac{1}{2}-}^{n_0+1}\right)\right] \\ &= \bar{h}_{i+\frac{1}{2}}^{n_0} u_{i+\frac{1}{2}}^{n_0} - \nu_{i+\frac{1}{2}}\left(u_{i+1}^{n_0} \overline{F}_{i+1}^{n_0} - u_i^{n_0} \overline{F}_i^{n_0}\right) \\ &\quad - g\nu_{i+\frac{1}{2}} \bar{h}_{i+\frac{1}{2}}^{n_0+1}\left(w_{i+1}^{n_0+1} - \max\left(b_i, b_{i+1}\right) - w_i^{n_0+1} + \max\left(b_i, b_{i+1}\right)\right) \\ &= 0, \quad \forall i \in \mathbb{Z}. \end{aligned}$$

On one hand, as a direct consequence from the mass conservation equation of (OEsc) at $n = n_0$,

$$h_i^{n_0+1} + b_i = h_i^{n_0} + b_i = C, \quad \forall i \in \mathbb{Z}.$$

On the other hand, as a direct consequence from the momentum balance equation of (OEsc) at $n = n_0$ and Convention (3.1), we have

$$u_{i+\frac{1}{2}}^{n_0+1} = 0, \ \ \forall i \in \mathbb{Z}.$$

Therefore, the well-balanced property corresponded to (sw) holds for (OEsc) in the case of the horizontal fluid domain being fully wet, in the discrete level. □

Now we have to face with the ill-balanced issue encountered by (DGsc). But we have to be very careful about the notion of the lake at rest steady state (lar_1) in the discrete level. It is not easy as that of (sw). In addition, if the initial water height is positive, Proposition 4.3 implies that the water height remains positive in the whole horizontal fluid domain $\mathbb{R}$ during $[0, T_0)$ in the discrete level, i.e.,

$$h_i^n > 0, \ \ \forall i \in \mathbb{Z}, \ \ \forall n \in [T] \, .$$

This indicates immediately that we can not define the discrete version of lake at rest (lar_1) involving something like $h_i^n = 0$ to illustrate the cell C_i being dry. This kind of idealistic notion is impossible due to the positivity conservation of (OEsc)! Hence, as the usual spirit of mathematical analysis, we need to define some approximate notion for this. The next subsection is devoted to derive some approximate notions which are mainly adapted from Chapter 1, Section 2.1.

1.2 Approximate dry-wet decomposition of the horizontal fluid domain

First of all, let $\varepsilon_0 > 0$ be some tolerance. Here ε_0 can be chosen ideally as the machine epsilon, e.g., double precision for 64-bit, $\varepsilon_0 := 2^{-51} \approx 2.22 \times 10^{-16}$. This is reasonable for both theoretical purposes and simulation: When the water height calculated by some pieces of programming scripts is smaller than ε_0 (but still positive!) then the machine will automatically set it to zero.

To distinguish the notions developed in this subsection with those of Chapter 1, Subsection 2.1, we add "ε_0" into the superscripts of all those notions.

For each $t \in [0, T_0)$, consider a nondecreasing sequence $\left(x_{i+\frac{1}{2}}^{\varepsilon_0,t}\right)_{i\in\mathbb{Z}} \subset \overline{\mathbb{R}}$ of extended reals (if $x_{i+\frac{1}{2}}^{\varepsilon_0,t} = -\infty$ then $x_{j+\frac{1}{2}}^{\varepsilon_0,t} = -\infty$ for all $j < i$ and thus can be dropped, similarly for the case $x_{i+\frac{1}{2}}^{\varepsilon_0,t} = \infty$):

$$x_{i-\frac{1}{2}}^{\varepsilon_0,t} \leq x_{i+\frac{1}{2}}^{\varepsilon_0,t}, \ \ \forall i \in \mathbb{Z}, \ \ \forall t \in [0, T_0) \, ,$$

which separates the ε_0-dry and ε_0-wet areas in the horizontal fluid domain $\mathbb{R}$.

Denote the i^{th} ε_0-area (ε_0-dry or ε_0-wet) by

$$A_i^{\varepsilon_0,t} := \left(x_{i-\frac{1}{2}}^{\varepsilon_0,t}, x_{i+\frac{1}{2}}^{\varepsilon_0,t}\right), \ \ \forall i \in \mathbb{Z}, \ \ \forall t \in [0, T_0) \, ,$$

it is also demanded that

$$\forall i \in \mathbb{Z},\ \ \forall t \in [0, T_0)\,,\ \ h|_{A_i^{\varepsilon_0,t}} > 0 \Rightarrow x_{i-\frac{1}{2}}^{\varepsilon_0,t} < x_{i+\frac{1}{2}}^{\varepsilon_0,t},$$

which means that the "1-point" ε_0-dry areas are allowed in our decomposition whereas the "1-point" ε_0-wet areas are absolutely not, as it is meaningless to do so.

Now, depending on the event that the 0th area is wet or dry, we can index ε_0-dry and ε_0-wet areas consecutively as follows:

- *Case* $h|_{A_0^{\varepsilon_0,t}} > \varepsilon_0$: Define the ε_0-dry and ε_0-wet areas as

$$W_i^{\varepsilon_0,t} := A_{2i}^{\varepsilon_0,t},\ \ D_i^{\varepsilon_0,t} := A_{2i+1}^{\varepsilon_0,t},\ \ \forall i \in \mathbb{Z},\ \ \forall t \in [0, T_0)\,.$$

- *Case* $h|_{A_0^{\varepsilon_0,t}} \leq \varepsilon_0$. Define the ε_0-dry and ε_0-wet areas as

$$W_i^{\varepsilon_0,t} := A_{2i+1}^{\varepsilon_0,t},\ \ D_i^{\varepsilon_0,t} := A_{2i}^{\varepsilon_0,t},\ \ \forall i \in \mathbb{Z},\ \ \forall t \in [0, T_0)\,.$$

Under these settings, the horizontal fluid domain $\mathbb{R}$ can be decomposed as the union of the ε_0-dry component $D_{\varepsilon_0,t}$ and the wet one $W^{\varepsilon_0,t}$:

$$\mathbb{R} = D^{\varepsilon_0,t} \cup W^{\varepsilon_0,t},\ \ \forall t \in [0, T_0)\,,$$

where

$$D^{\varepsilon_0,t} := \bigcup_{i=-\infty}^{\infty} D_i^{\varepsilon_0,t},\ \ W^{\varepsilon_0,t} := \bigcup_{i=-\infty}^{\infty} W_i^{\varepsilon_0,t},\ \ \forall t \in [0, T_0)\,.$$

By definition, it is straightforward that

$$D^{\varepsilon_0,t} = \{x \in \mathbb{R}; h\,(t,x) \leq \varepsilon_0\}\,,\ \ W^{\varepsilon_0,t} = \{x \in \mathbb{R}; h\,(t,x) > \varepsilon_0\}\,,\ \ \forall t \in [0, T_0)\,,$$

and thus our ε_0-setting makes a sense. We also demand that the velocity of water is set to zero in the ε_t-dry component $D^{\varepsilon_0,t}$, i.e.,

$$u|_{D^{\varepsilon_0,t}} := 0,\ \text{ in }\ [0, T_0) \times \mathbb{R},$$

which is equivalent to the following convention.

Convention 4.1 (Zero velocity of water in ε_0-dry areas in the horizontal fluid domain $\mathbb{R}$)**.** *In the horizontal fluid domain* $\mathbb{R}$*, during* $[0, T_0)$*, if* $h\,(t,x) \leq \varepsilon_0$*, then* $u\,(t,x)$ *is set to zero.*

This also means that the natural phase space Ω is replaced by the following ε_0-natural phase space Ω^{ε_0}:

$$\Omega^{\varepsilon_0} := \left\{U = (h, hu)^\top \subset \mathbb{R}^2; h > \varepsilon_0\right\} \cup \left\{(h, 0)^\top \subset \mathbb{R}^2; h \leq \varepsilon_0\right\}.$$

Note that these unions can be bot finite, or both infinite. They are both finite provided there exists $i_L < i_R \in \mathbb{Z}$ such that $x_{i_L} = -\infty$ and $x_{i_R} = \infty$.

Having enough ε_0-settings, we consider the approximate steady states of the shallow water equations (1DSW), after the steady time $T_\star \in (0, T_0)$ as before. Among the full family of (exact) stead states (sss), we are interested in the lake at rest (lar_1) and the still water (sw) steady states only.

The approximate, or more precisely, ε_0-lake at rest steady state is defined as

$$\begin{cases} u|_{W_i^{\varepsilon_0,t}} = 0, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ (h+b)|_{W_i^{\varepsilon_0,t}} = C_i, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ D_i^{\varepsilon_0,t} = D_i^{\varepsilon_0,T_\star}, \ W_i^{\varepsilon_0,t} = W_i^{\varepsilon_0,T_\star}, & \forall i \in \mathbb{Z}, \ \forall t \in [T_\star, T_0), \end{cases} \tag{alar_1}$$

or equivalently,

$$\begin{cases} u|_{W_i^{\varepsilon_0,t}} = 0, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ w|_{W_i^{\varepsilon_0,t}} = C_i, & \forall i \in \mathbb{Z}, \text{ in } [T_\star, T_0) \times \mathbb{R}, \\ D_i^{\varepsilon_0,t} = D_i^{\varepsilon_0,T_\star}, \ W_i^{\varepsilon_0,t} = W_i^{\varepsilon_0,T_\star}, & \forall i \in \mathbb{Z}, \ \forall t \in [T_\star, T_0). \end{cases} \tag{alar_2}$$

The ε_0-lake at rest steady state (alar_1) is exactly what we need for our main purpose.

1.3 Approximate dry-wet decomposition of the horizontal fluid domain in the discrete level

Analogous to Convention 4.1, we have the following convention for (OEsc) in the discrete level.

Convention 4.2 (Zero velocity of water in ε_0-dry areas in the horizontal fluid domain $\mathbb{R}$)**.** *For all $i \in \mathbb{Z}$, $n \in [T]$, if $\bar{h}^n_{i+\frac{1}{2}} \le \varepsilon_0$, by convention, $u^n_{i+\frac{1}{2}}$ is then set to zero.*

For the sake of convenience, we define the following sets of ε_0-dry and ε_0-wet indices in the discrete level:

$$I_D^{\varepsilon_0,n} := \{i \in \mathbb{Z}; 0 \le h_i^n \le \varepsilon_0\}, \quad I_W^{\varepsilon_0,n} := \{i \in \mathbb{Z}; h_i^n > \varepsilon_0\}, \quad \forall n \in [T],$$

and the nonnegativity conservation of water height stated in Proposition (4.3) implies that

$$\mathbb{Z} = I_D^{\varepsilon_0,n} \cup I_W^{\varepsilon_0,n}, \quad \forall n \in [T],$$

provided the initial water height is nonnegative.

As another direct consequence, Convention 4.2 reads

$$u^n_{i+\frac{1}{2}} := 0, \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^{\varepsilon_0,n}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0,n}\right), \quad \forall n \in [T].$$

It should be noted that the condition $\left(i \in I_D^{\varepsilon_0,n}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0,n}\right)$ guarantees that $\bar{h}^n_{i+\frac{1}{2}} \le \varepsilon_0$ but does not detect all indices i such that $\bar{h}^n_{i+\frac{1}{2}} \le \varepsilon_0$, compared to this approximate version the exact one in previous chapter.

We also define these notions in the discrete level. Depending on the event that the 0^{th} area is wet or dry, we can decompose the set $\mathbb{Z}$ of integers into discrete ε_0-dry and ε_0-wet areas for each $n \in [T]$, as in Algorithm 1. Note that the "infinity loops" $i \leftarrow 1 : \infty$ and $i \leftarrow -1 : -\infty$ are meaningless in programming. Here the author wants to specify (OEsc) in the whole horizontal fluid domain $\mathbb{R}$. When (OEsc) is restricted in some computational horizontal fluid domain Ω_c as before, these loops become finite and thus programmatically reasonable again.

Algorithm 1 Decomposition of $\mathbb{Z}$ into discrete ε_0-dry and ε_0-wet areas.

```
Inputs: The discrete water height $(h_i^n)_{i\in\mathbb{Z},n\in[T]}$ generated by (OEsc).
$i_D \leftarrow 0$; $i_W \leftarrow 0$;    ▷ Counters for the number of $\varepsilon_0$-dry and $\varepsilon_0$-wet areas
for $n \leftarrow 0$ to $T$ do
    if $h_0^n > \varepsilon_0$ then
        $I_{W,i_W}^{\varepsilon_0,n} \leftarrow \{0\}$;    ▷ Initialize the $0^{\text{th}}$ $\varepsilon_0$-wet set of indices
    else    ▷ $0 \le h_0^n \le \varepsilon_0$
        $I_{D,i_D}^{\varepsilon_0,n} \leftarrow \{0\}$;    ▷ Initialize the $0^{\text{th}}$ $\varepsilon_0$-dry set of indices
    end if
    for $i \leftarrow 1 \to \infty$ do    ▷ or $i \leftarrow 1 \to N_R$ for some $N_R \in \mathbb{Z}_+$ large enough
        if $i-1 \in I_{W,i_W}^{\varepsilon_0,n}$ then
            if $h_i^n \le \varepsilon_0$ then
                $i_D \leftarrow i_D + 1$;
                $I_{D,i_D}^{\varepsilon_0,n} \leftarrow \{i\}$;
            else    ▷ $h_i^n > \varepsilon_0$
                $I_{W,i_W}^{\varepsilon_0,n} \leftarrow I_{W,i_W}^{\varepsilon_0,n} \cup \{i\}$;
            end if
        else    ▷ $i-1 \in I_{D,i_D}^{\varepsilon_0,n}$ for the current $\varepsilon_0$-dry counter $i_D$
            if $h_i^n > \varepsilon_0$ then
                $i_W \leftarrow i_W + 1$;
                $I_{W,i_W}^{\varepsilon_0,n} \leftarrow \{i\}$;
            else    ▷ $h_i^n \le \varepsilon_0$
                $I_{D,i_D}^{\varepsilon_0,n} \leftarrow I_{D,i_D}^{\varepsilon_0,n} \cup \{i\}$;
            end if
        end if
    end for
    $i_D \leftarrow 0$; $i_W \leftarrow 0$;    ▷ Reset counters for the number of $\varepsilon_0$-dry and $\varepsilon_0$-wet areas
    for $i \leftarrow -1 \to -\infty$ do    ▷ or $i \leftarrow -1 \to N_L$ for some $N_L \in \mathbb{Z}_-$ small enough
        if $i+1 \in I_{W,i_W}^{\varepsilon_0,n}$ then
            if $h_i^n \le \varepsilon_0$ then
                $i_D \leftarrow i_D - 1$;
                $I_{D,i_D}^{\varepsilon_0,n} \leftarrow \{i\}$;
            else    ▷ $h_i^n > \varepsilon_0$
                $I_{W,i_W}^{\varepsilon_0,n} \leftarrow I_{W,i_W}^{\varepsilon_0,n} \cup \{i\}$;
            end if
        else    ▷ $i+1 \in I_{D,i_D}^{\varepsilon_0,n}$ for the current $\varepsilon_0$-dry counter $i_D$
            if $h_i^n > \varepsilon_0$ then
                $i_W \leftarrow i_W - 1$;
                $I_{W,i_W}^{\varepsilon_0,n} \leftarrow \{i\}$;
            else    ▷ $h_i^n \le \varepsilon_0$
                $I_{D,i_D}^{\varepsilon_0,n} \leftarrow I_{D,i_D}^{\varepsilon_0,n} \cup \{i\}$;
            end if
        end if
    end for
end for
Outputs: The decomposition of $\mathbb{Z}$ into discrete $\varepsilon_0$-dry and $\varepsilon_0$-wet areas, i.e.,
    $$\mathbb{Z} = \left(\bigcup_{i=-\infty}^{\infty} I_{D,i}^{\varepsilon_0,n}\right) \cup \left(\bigcup_{i=-\infty}^{\infty} I_{W,i}^{\varepsilon_0,n}\right), \quad \forall n \in [T].$$
```

As an immediate consequence,

$$I_D^{\varepsilon_0,n} = \bigcup_{i=-\infty}^{\infty} I_{D,i}^{\varepsilon_0,n}, \quad I_W^{\varepsilon_0,n} = \bigcup_{i=-\infty}^{\infty} I_{W,i}^{\varepsilon_0,n}, \quad \forall n \in [T] .$$

Now, in the case of the horizontal fluid domain $\mathbb{R}$ being fully wet, the discrete still water steady state is described as in Proposition 4.4 is stated as follows: For any $n \in [T-1]$,

$$\begin{cases} u_{i+\frac{1}{2}}^n = 0, & \forall i \in \mathbb{Z}, \\ h_i^n + b_i = C, & \forall i \in \mathbb{Z}, \end{cases} \Rightarrow \begin{cases} u_{i+\frac{1}{2}}^{n+1} = 0, & \forall i \in \mathbb{Z}, \\ h_i^{n+1} + b_i = C, & \forall i \in \mathbb{Z}. \end{cases} \tag{dsw}$$

Thus, Proposition 4.4 can be restated as:

Proposition 4.5 (Well-balanced property for fully wet horizontal fluid domain $\mathbb{R}$)**.** *Assume that the initial data satisfies $(h_0, u_0) \in L^\infty(\mathbb{R}; \mathbb{R}_+^\star) \times L^\infty(\mathbb{R})$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by (OEsc). Assume that the strict CFL-like condition (sCFL$_3$) or (sCFL$_4$) holds.*

Then the positivity of the initial water heihgt and the strict CFL-like condition imply that the horizontal fluid domain $\mathbb{R}$ is fully wet for all $t \in [0, T_0)$ during the survival time of (1DSW), the offset equilibrium staggered upwind scheme (OEsc) preserves the discrete still water steady state (dsw) for all $n \in [T-1]$.

Similarly, the discrete lake at rest steady state is described as

$$\begin{cases} \begin{cases} u_{i+\frac{1}{2}}^n = 0, & \forall i \in \mathbb{Z}, \\ h_j^n + b_j = C_i, \quad \forall j \in I_{W,i}^n, & \forall i \in \mathbb{Z}, \end{cases} \Rightarrow \begin{cases} u_{i+\frac{1}{2}}^{n+1} = 0, & \forall i \in \mathbb{Z}, \\ h_j^{n+1} + b_j = C_i, \quad \forall j \in I_{W,i}^{n+1}, & \forall i \in \mathbb{Z}, \end{cases} \\ I_{W,i}^n = I_{W,i}^{n+1}, \quad I_{D,i}^n = I_{D,i}^{n+1}, \quad \forall i \in \mathbb{Z}. \end{cases} \tag{dlar}$$

The following result indicates that (OEsc) does not fulfill mathematically this discrete version of lake at rest steady states in general.

Proposition 4.6 (Lake at rest non-conservation property for dry-wet horizontal fluid domain $\mathbb{R}$)**.** *Assume that the initial data satisfies $(h_0, u_0) \in L^\infty(\mathbb{R}; \mathbb{R}_+) \times L^\infty(\mathbb{R})$. Let $\mathcal{T}$ be an admissible mesh of $\mathbb{R}$ and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ be the discrete finite volume approximate solution generated by (OEsc). Assume that the CFL-like condition (CFL3) or (CFL4) holds.*

Assume that the horizontal fluid domain $\mathbb{R}$ contains at least one non-degenerate dry area but not fully dry for all $t \in [0, T_0)$, the offset equilibrium staggered upwind scheme (OEsc) does not preserve the discrete lake at rest steady state (dlar) in general.

Demonstration. Inspired by the following demonstration, a counterexample should be found!

Suppose for some $n_0 \in [T-1]$,

$$\begin{cases} u^{n_0}_{i+\frac{1}{2}} = 0, & \forall i \in \mathbb{Z}, \\ h^{n_0}_j + b_j = C_i, \quad \forall j \in I^{n_0}_{W,i}, & \forall i \in \mathbb{Z}, \end{cases}$$

then plugging the first equation into (dmc) at $n = n_0$ yields the water height at $t = t^{n_0}$ and $t = t^{n_0+1}$ are the same,

$$h^{n_0+1}_i = h^{n_0}_i, \quad \forall i \in \mathbb{Z}.$$

As a direct consequence, due to the rigidity of the bottom topography, the water level at $t = t^{n_0}$ and $t = t^{n_0+1}$ are the same,

$$w^{n_0+1}_i = w^{n_0}_i, \quad \forall i \in \mathbb{Z}.$$

And (4.1) becomes (note that the assumption dry-wet and Proposition 4.3 imply that $h^n_i \geq 0$ generated by (OEsc) for all $i \in \mathbb{Z}$ and $n \in [T]$)

$$\begin{cases} w^{n_0}_j = C_i, & \forall j \in I^{n_0}_{W,i}, \quad \forall i \in \mathbb{Z}, \\ b^{n_0}_{j+\frac{1}{2}} = \min\left(\max\left(b_j, b_{j+1}\right), \min\left(C_i, w^{n_0}_{j+1}\right)\right), & \forall j \in I^{n_0}_{W,i}, \quad \forall i \in \mathbb{Z}, \\ h^{n_0}_{j+\frac{1}{2}-} = \min\left(C_i - b^{n_0}_{j+\frac{1}{2}}, C_i - b_j\right), & \forall j \in I^{n_0}_{W,i}, \quad \forall i \in \mathbb{Z}, \\ h^{n_0}_{j+\frac{1}{2}+} = \min\left(w^{n_0}_{j+1} - b^{n_0}_{j+\frac{1}{2}}, h^{n_0}_{j+1}\right), & \forall j \in I^{n_0}_{W,i}, \quad \forall i \in \mathbb{Z}, \end{cases} \tag{4.2}$$

Since the horizontal fluid domain is both wet and dry, there exists an index i_0 such that $I^{n_0}_{W,i_0}$ has either a maximum element or a minimum element. Without loss of generality (w.l.o.g.), assume the former holds, we denote by j_0 the maximum element of $I^{n_0}_{W,i_0}$:

$$j_0 := \max\left\{j; j \in I^{n_0}_{W,i_0}\right\},$$

then by definition of j_0, $j_0 + 1 \in I^{n_0}_D$. This couple of indices (j_0, j_0+1) is called to form an wet-dry front, which is defined later. Now, assume the water level at the primal cell C_{j_0} at the time $t = t^{n_0}$ is higher than the bottom topography in the primal cell C_{j_0+1}[1], i.e.,

$$w^{n_0}_{j_0} := h^{n_0}_{j_0} + b_{j_0} = C_{i_0} > b_{j_0+1} = w^{n_0}_{j_0+1},$$

[1] Although this is unreasonable compared to reality: If the water in somewhere is higher than the nearby dry area without any "wall" to keep the water stay in that place, then it must flow into the nearby area. Thanks the gravity!

then (4.2) becomes

$$\begin{cases} w_{j_0}^{n_0} = C_{i_0}, w_{j_0+1}^{n_0} = b_{j_0+1}, \\ b_{j_0+\frac{1}{2}}^{n_0} = \min\left(\max\left(b_{j_0}, b_{j_0+1}\right), \min\left(C_{i_0}, b_{j_0+1}\right)\right) = \min\left(\max\left(b_{j_0}, b_{j_0+1}\right), b_{j_0+1}\right) = b_{j_0+1}, \\ h_{j_0+\frac{1}{2}-}^{n_0} = \min\left(C_{i_0} - b_{j_0+1}, C_{i_0} - b_{j_0}\right), \\ h_{j_0+\frac{1}{2}+}^{n_0} = \min\left(w_{j_0+1}^{n_0} - b_{j_0+1}, 0\right) = 0, \end{cases}$$

and thus the discrete momentum balance equation of (OEsc) at $(i,n) = (j_0, n_0)$ reads

$$\begin{aligned} \frac{1}{2}\left(h_{j_0}^{n_0+1} + h_{j_0+1}^{n_0+1}\right) u_{j_0+\frac{1}{2}}^{n_0+1} := \frac{1}{2}\left(h_{j_0}^{n_0} + h_{j_0+1}^{n_0}\right) u_{j_0+\frac{1}{2}}^{n_0} &- \nu_{j_0+\frac{1}{2}}\left(u_{j_0+1}^{n_0}\overline{F}_{j_0+1}^{n_0} - u_{j_0}^{n_0}\overline{F}_{j_0}^{n_0}\right) \\ &- \frac{g}{2}\nu_{j_0+\frac{1}{2}}\left(h_{j_0}^{n_0} + h_{j_0+1}^{n_0}\right)\left(h_{j_0+\frac{1}{2}+}^{n_0+1} - h_{j_0+\frac{1}{2}-}^{n_0+1}\right) \\ = &- \frac{g}{2}\nu_{j_0+\frac{1}{2}} h_{j_0}^{n_0}\left(h_{j_0+\frac{1}{2}+}^{n_0+1} - h_{j_0+\frac{1}{2}-}^{n_0+1}\right). \end{aligned}$$

Plugging the first equation $w_{j_0}^{n_0+1} = w_{j_0}^{n_0}$ into (4.1) at $(i,n) = (j_0+1, n_0+1)$ yields

$$\begin{cases} w_{j_0}^{n_0+1} = w_{j_0}^{n_0}, \quad w_{j_0+1}^{n_0+1} = w_{j_0+1}^{n_0}, \\ b_{j_0+\frac{1}{2}}^{n_0+1} = \min\left(\max\left(b_{j_0}, b_{j_0+1}\right), \min\left(w_{j_0}^{n_0}, w_{j_0+1}^{n_0}\right)\right) = b_{j_0+\frac{1}{2}}^{n_0}, \\ h_{j_0+\frac{1}{2}-}^{n_0+1} = \min\left(w_{j_0}^{n_0} - b_{j_0+\frac{1}{2}}^{n_0}, h_{j_0}^{n_0}\right) = h_{j_0+\frac{1}{2}-}^{n_0}, \\ h_{j_0+\frac{1}{2}+}^{n_0+1} = \min\left(w_{j_0+1}^{n_0} - b_{j_0+\frac{1}{2}}^{n_0}, h_{j_0+1}^{n_0}\right) = h_{j_0+\frac{1}{2}+}^{n_0}, \end{cases}$$

Thus, the discrete momentum balance equation of (OEsc) at $(i,n) = (j_0, n_0)$ becomes

$$u_{j_0+\frac{1}{2}}^{n_0+1} = g\nu_{j_0+\frac{1}{2}} \min\left(C_{i_0} - b_{j_0+1}, C_{i_0} - b_{j_0}\right).$$

Hence, $u_{j_0+\frac{1}{2}}^{n_0+1}$ can be nonzero in general. More explicitly,

- *Case* $b_{j_0+1} < b_{j_0}$: If the bottom topography at the primal cell C_{j_0+1} is lower than that at the previous primal cell, then $b_{j_0+1} < b_{j_0} < C_{i_0}$ since $h_{j_0}^{n_0} > 0$. The velocity at the interface $j_0 + \frac{1}{2}$ at $t = t^{n_0+1}$ is positive.
- *Case* $b_{j_0} < b_{j_0+1} < C_{i_0}$: If the bottom topography at the primal cell C_{j_0+1} is higher than that at the previous primal cell but still lower than the water level there, then the velocity at the interface $j_0 + \frac{1}{2}$ at $t = t^{n_0+1}$ is also positive!

Thus, (OEsc) does not preserve mathematically the discrete lake at rest (dlar) in general. □

Another demonstration of Proposition 4.6. It suffices to build a counterexample. For the sake of simplicity, we consider (OEsc) on a uniform mesh, with the flat bottom topography. We consider (OEsc) with the following initial conditions

$$h_0(x) := h_{0,0}(x) + 2h_{0,1}(x), \quad \forall x \in \mathbb{R},$$

where

$$\alpha_0 \text{ is a mollifier s.t. } \operatorname{Supp}(\alpha_0) \subset \overline{C_0}, \quad \int_{\mathbb{R}} \alpha_0(x)\,dx = 1,$$
$$\alpha_1 \text{ is a mollifier s.t. } \operatorname{Supp}(\alpha_1) \subset \overline{C_1}, \quad \int_{\mathbb{R}} \alpha_1(x)\,dx = 1,$$

and thus

$$h_i^0 = \int_{C_i} (\alpha_0(x) + 2\alpha_1(x))\,dx = \delta_{0,i} + 2\delta_{1,i}, \quad \forall i \in \mathbb{Z},$$

or equivalently

$$h_0^0 = 1, h_1^0 = 2, h_i^0 = 0, \quad \forall i \in \mathbb{Z} \backslash \{0,1\}.$$

Now we assume that the water is at rest at the initial time $t = 0$, i.e., $u^0_{i+\frac{1}{2}} = 0$ for all $i \in \mathbb{Z}$. This situation can be imagined as follows: With the flat bottom topography $b = 0$, the water level and the water height coincide in all primal cells of the given admissible mesh $\mathcal{T}$. Moreover, the water level at the cell C_0 is equal to 1 and the water level at the cell C_1 is equal to 2 - higher than that at adjacent cells, while other cells contains no water. This violates the discrete lake at rest steady state

$$w_0^0 = 1 \neq w_1^0 = 2, \text{ while } \left(0 \in I^0_{W,0}\right) \wedge \left(1 \in I^0_{W,0}\right).$$

This counterexample completes our proof. □

Under the notions of discrete ε_0-dry and ε_0-dry areas just defined, the discrete ε_0-lake at rest steady state is defined as follows,

$$\begin{cases} \begin{cases} u^n_{i+\frac{1}{2}} = 0, & \forall i \in \mathbb{Z}, \\ h_j^n + b_j = C_i, \quad \forall j \in I^{\varepsilon_0,n}_{W,i}, & \forall i \in \mathbb{Z}, \end{cases} \Rightarrow \begin{cases} u^{n+1}_{i+\frac{1}{2}} = 0, & \forall i \in \mathbb{Z}, \\ h_j^{n+1} + b_j = C_i, \quad \forall j \in I^{\varepsilon_0,n+1}_{W,i}, & \forall i \in \mathbb{Z}, \end{cases} \\ I^{\varepsilon_0,n}_{W,i} = I^{\varepsilon_0,n+1}_{W,i}, \quad I^{\varepsilon_0,n}_{D,i} = I^{\varepsilon_0,n+1}_{D,i}, \quad \forall i \in \mathbb{Z}. \end{cases} \tag{adlar}$$

Sadly, similarly to Proposition 4.6, (OEsc) does not fulfill mathematically this discrete approximate version of lake at rest steady states in general.

Proposition 4.7 (ε_0-Lake at rest non-conservation property for dry-wet horizontal fluid domain $\mathbb{R}$). *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty(\mathbb{R}; \mathbb{R}_+) \times L^\infty(\mathbb{R})$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n, u^n_{i+\frac{1}{2}}\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by* (OEsc)*. Assume that the CFL-like condition* (CFL3) *or* (CFL4) *holds.*

Assume that the horizontal fluid domain $\mathbb{R}$ *contains at least one non-degenerate dry area but not fully dry for all* $t \in [0, T_0)$*, the offset equilibrium staggered upwind scheme* (OEsc) *does not preserve the discrete* ε_0*-lake at rest steady state* (adlar) *in general.*

Motivated by these failures, we need to treat this situation more carefully. Idealistically, we should add some constraints which are more “reasonable” compared to the reality. We should start at exactly where the conservation properties have collapsed in the proof of Proposition 4.6: the set of couples of indices indicating wet-dry transitions in the whole horizontal fluid domain $\mathbb{R}$.

1.4 Well-balanced of a new offset equilibrium staggered upwind scheme in the dry-wet case

We borrow Chen and Noelle, 2017, Definition 4.2, p. 771 about two classes of equilibria and adapt it in our framework.

Definition 4.1 (Equilibrium). *i) Given a constant water level w_{eq}, the* still water equilibrium *is defined by*

$$\begin{cases} u\left(t,x\right) = 0, & \text{in } \left[T_\star, T_0\right) \times \mathbb{R}, \\ h\left(t,x\right) + b\left(x\right) = w_{eq}, & \text{in } \left[T_\star, T_0\right) \times \mathbb{R}. \end{cases}$$

The discrete still water equilibrium *during $\left[t^{N_\star}, t^T\right]$ is defined by*

$$\begin{cases} u^n_{i+\frac{1}{2}} = 0, & \forall i \in \mathbb{Z}, \ \forall n \in [T] \setminus [N_\star - 1], \\ h^n_i + b_i = w_{eq}, & \forall i \in \mathbb{Z}, \ \forall n \in [T] \setminus [N_\star - 1], \end{cases}$$

where $N_\star$ can be understood as the discrete steady time corresponding to the steady time $T_\star$.

ii) The lake at rest equilibrium is defined by (lar$_2$). Moreover, near a wet-dry interface, the dry part of the bottom topography b should not be lower than the adjacent water level.

The locally discrete lake at rest equilibrium *at $t = t^n$ is defined as follows.*

a) either $x_{i+\frac{1}{2}}$ is an interior interface (the still water case, or "wet-wet front")

$$\left(u^n_{i+\frac{1}{2}} = 0\right) \wedge \left(h^n_i > 0\right) \wedge \left(h^n_{i+1} > 0\right) \wedge \left(h^n_i + b_i = h^n_{i+1} + b_{i+1}\right), \quad \text{(w-wf)}$$

b) or a dry-wet front

$$\left(u^n_{i+\frac{1}{2}} = 0\right) \wedge \left(h^n_i = 0\right) \wedge \left(h^n_{i+1} > 0\right) \wedge \left(b_i \geq h^n_{i+1} + b_{i+1}\right), \quad \text{(d-wf)}$$

c) or a dry-wet front

$$\left(u^n_{i+\frac{1}{2}} = 0\right) \wedge \left(h^n_i > 0\right) \wedge \left(h^n_{i+1} = 0\right) \wedge \left(h^n_i + b_i \leq b_{i+1}\right), \quad \text{(w-df)}$$

d) or "dry-dry" front

$$h^n_i = h^n_{i+1} = 0 \Rightarrow u^n_{i+\frac{1}{2}} := 0. \quad \text{(d-df)}$$

The locally discrete lake at rest equilibrium *during $\left[t^{N_\star}, t^T\right]$ is a state that one of the cases (w-wf), (w-df), (d-wf) and (d-df) holds for each $n \in [T] \setminus [N_\star - 1]$.*

The globally discrete lake at rest equilibrium *at $t = t^n$ (resp., during $\left[t^{N_\star}, t^T\right]$) is a state that is a locally discrete lake at rest equilibrium at $t = t^n$ (resp., during $\left[t^{N_\star}, t^T\right]$) on all dual cells $C_{i+\frac{1}{2}}$'s of the given admissible mesh $\mathcal{T}$.*

iii) A general staggered scheme being able to generate a discrete solution of water height and velocity $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ for (1DSW) *is said to be* well-balanced for dry-wet horizontal fluid domain $\mathbb{R}$ *if and only if it preserves the globally discrete lake at rest equilibrium after some discrete steady time.*

More precisely, if $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}}$ fulfills the globally discrete lake at rest equilibrium at $t = t^n$ for some $n \in [T-1]$, then $\left(h_i^{n+1}, u_{i+\frac{1}{2}}^{n+1}\right)_{i\in\mathbb{Z}}$ fulfills the globally discrete lake at rest equilibrium at $t = t^{n+1}$.

As an immediate consequence, $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ fulfills the globally discrete lake at rest equilibrium during $\left[t^{N_\star}, t^T\right]$ where $N_\star$ is the smallest nonnegative integer such that $\left(h_i^{N_\star}, u_{i+\frac{1}{2}}^{N_\star}\right)_{i\in\mathbb{Z}}$ fulfills the globally discrete lake at rest equilibrium at $t = t^{N_\star}$, i.e., $N_\star$ corresponds to $T_\star$ in the discrete level.

Remark 4.1. *The four cases* (w-wf), (w-df), (d-wf) *and* (d-df) *can be rewritten in terms of water levels and the sets of dry- and wet-indices as follows,*

$$\begin{cases} \left(u_{i+\frac{1}{2}}^n = 0\right) \wedge \left(w_i^n = w_{i+1}^n\right), & \forall i \in \mathbb{Z} \text{ s.t. } (i \in I_W^n) \wedge (i+1 \in I_W^n), \\ \left(u_{i+\frac{1}{2}}^n = 0\right) \wedge \left(b_i \geq w_{i+1}^n\right), & \forall i \in \mathbb{Z} \text{ s.t. } (i \in I_D^n) \wedge (i+1 \in I_W^n), \\ \left(u_{i+\frac{1}{2}}^n = 0\right) \wedge \left(w_i^n \leq b_{i+1}\right), & \forall i \in \mathbb{Z} \text{ s.t. } (i \in I_W^n) \wedge (i+1 \in I_D^n), \\ u_{i+\frac{1}{2}}^n := 0, & \forall i \in \mathbb{Z} \text{ s.t. } (i \in I_D^n) \wedge (i+1 \in I_D^n). \end{cases} \tag{DLAR}$$

With the help of the notions of sets of dry- and wet-indices, the labels "wet-wet", "dry-wet", "wet-dry" and "dry-dry" can be seen clearer as "WW", "DW", "WD" and "DD", respectively.

We also modify the lake at rest steady state in Definition 4.1 for all ε_0-dry/wet notions defined previously. The main motivation for this is the following case: When the initial water height is positive but some parts of the bottom topography are"high enough" to become "almost dry" when the water start to be at rest. More precisely, "high enough" here means that those parts of the bottom topography are higher than the water level at rest in adjacent wet areas. Mathematically, suppose such a part of the bottom topography is located at the cell C_i, and at least one of the couples of indices $(i-1, i)$ and $(i, i+1)$ forms a dry-wet front or a wet-dry front. W.l.o.g., suppose the latter holds, then this case can be described mathematically as $b_i > w_{i+1}$. However, the positivity conservation of water height of (DGsc) and (OEsc) only allows the discrete water height tend to zero but can not be equal to zero exactly, which can be interpreted as "almost dry" or "a little bit wet". This kind of situation is quite reasonable to the reality when the shallow water model is considered to be closed. For instance, there is no sunshine affecting these "almost dry" areas so that some "naughty" water particles still stuck there in spite of the event that the whole horizontal fluid domain $\mathbb{R}$ has already reached the lake at rest steady state.

There is also a case that, for instance, in the cell $C_i = \left(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right)$, the situation there reads for some $\varepsilon \in \left(0, \frac{\Delta x_i}{2}\right)$:

$$\begin{cases} b(x) < w(x), & \forall x \in \left[x_{i-\frac{1}{2}}, x_{i-\frac{1}{2}} + \varepsilon\right) \cup \left(x_{i+\frac{1}{2}} - \varepsilon, x_{i+\frac{1}{2}}\right], \\ b\left(x_{i-\frac{1}{2}} + \varepsilon\right) = w\left(x_{i-\frac{1}{2}} + \varepsilon\right), & \text{i.e., } h\left(x_{i-\frac{1}{2}} + \varepsilon\right) = 0, \\ b\left(x_{i+\frac{1}{2}} - \varepsilon\right) = w\left(x_{i+\frac{1}{2}} - \varepsilon\right), & \text{i.e., } h\left(x_{i+\frac{1}{2}} - \varepsilon\right) = 0, \\ b(x) \geq \min\left(b\left(x_{i-\frac{1}{2}} + \varepsilon\right), b\left(x_{i+\frac{1}{2}} - \varepsilon\right)\right), & \forall x \in \left(x_{i-\frac{1}{2}} + \varepsilon, x_{i+\frac{1}{2}} - \varepsilon\right). \end{cases}$$

It should be noticed that the part of bottom topography in the space interval $\left(x_{i-\frac{1}{2}} + \varepsilon, x_{i+\frac{1}{2}} - \varepsilon\right)$ is absolutely over the water surface. When this kind of situations happen, we still treat this cell as “almost dry” provided the average amount of water contained in this cell is small enough, e.g., $b_i + \varepsilon_0 \geq w_{i+1}$ if the next cell C_{i+1} is completely wet. This discussion can be mathematically formulated in the following definition.

Definition 4.2 (Approximate equilibrium). *i) Near a ε_0 wet-dry interface, the dry part of the bottom topography b is required again to be not lower than the adjacent water level.*

The locally discrete ε_0-lake at rest equilibrium *at $t = t^n$ is defined as follows.*

a) either $x_{i+\frac{1}{2}}$ is an ε_0-interior interface, i.e., a “ε_0-wet-wet” front,

$$\left(u^n_{i+\frac{1}{2}} = 0\right) \wedge \left(h^n_i > \varepsilon_0\right) \wedge \left(h^n_{i+1} > \varepsilon_0\right) \wedge \left(h^n_i + b_i = h^n_{i+1} + b_{i+1}\right), \tag{aw-wf}$$

b) or a ε_0-dry-wet front

$$\left(u^n_{i+\frac{1}{2}} = 0\right) \wedge \left(h^n_i \leq \varepsilon_0\right) \wedge \left(h^n_{i+1} > \varepsilon_0\right) \wedge \left(b_i + \varepsilon_0 \geq h^n_{i+1} + b_{i+1}\right), \tag{ad-wf}$$

c) or a ε_0-wet-dry front

$$\left(u^n_{i+\frac{1}{2}} = 0\right) \wedge \left(h^n_i > \varepsilon_0\right) \wedge \left(h^n_{i+1} \leq \varepsilon_0\right) \wedge \left(h^n_i + b_i \leq b_{i+1} + \varepsilon_0\right), \tag{aw-df}$$

d) or “ε_0-dry-dry” front

$$\left(h^n_i \leq \varepsilon_0\right) \wedge \left(h^n_{i+1} \leq \varepsilon_0\right) \Rightarrow u^n_{i+\frac{1}{2}} := 0. \tag{ad-df}$$

The locally discrete ε_0-lake at rest equilibrium *during $\left[t^{N_\star}, t^T\right]$ is a state that one of the cases* (aw-wf), (aw-df), (ad-wf) *and* (ad-df) *holds for each $n \in [T] \setminus [N_\star - 1]$.*

The globally discrete ε_0-lake at rest equilibrium *at $t = t^n$ (resp., during $\left[t^{N_\star}, t^T\right]$) is a state that is a locally discrete ε_0-lake at rest equilibrium at $t = t^n$ (resp., during $\left[t^{N_\star}, t^T\right]$) on all dual cells $C_{i+\frac{1}{2}}$'s of the given admissible mesh $\mathcal{T}$.*

ii) A general staggered scheme being able to generate a discrete solution of water height and velocity $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *for* (1DSW) *is said to be* ε_0-well-balanced for dry-wet horizontal fluid domain $\mathbb{R}$ *if and only if it preserves the globally discrete* ε_0*-lake at rest equilibrium after some discrete steady time.*

More precisely, if $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}}$ *fulfills the globally discrete* ε_0*-lake at rest equilibrium at* $t = t^n$ *for some* $n \in [T-1]$*, then* $\left(h_i^{n+1}, u_{i+\frac{1}{2}}^{n+1}\right)_{i\in\mathbb{Z}}$ *fulfills the globally discrete* ε_0*-lake at rest equilibrium at* $t = t^{n+1}$*.*

As an immediate consequence, $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *fulfills the globally discrete* ε_0*-lake at rest equilibrium during* $\left[t^{N_\star}, t^T\right]$ *where* $N_\star$ *is the smallest nonnegative integer such that* $\left(h_i^{N_\star}, u_{i+\frac{1}{2}}^{N_\star}\right)_{i\in\mathbb{Z}}$ *fulfills the globally discrete* ε_0*-lake at rest equilibrium at* $t = t^{N_\star}$*, i.e.,* $N_\star$ *corresponds to* $T_\star$ *in the discrete level.*

Remark 4.2. *The four approximate cases* (aw-wf)*,* (aw-df)*,* (ad-wf) *and* (ad-df) *can be rewritten in terms of water levels and the sets of* ε_0*-dry and wet-indices as follows,*

$$\begin{cases} \left(u_{i+\frac{1}{2}}^n = 0\right) \wedge \left(w_i^n = w_{i+1}^n\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_W^{\varepsilon_0,n}\right) \wedge \left(i+1 \in I_W^{\varepsilon_0,n}\right), \\ \left(u_{i+\frac{1}{2}}^n = 0\right) \wedge \left(b_i + \varepsilon_0 \geq w_{i+1}^n\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^{\varepsilon_0,n}\right) \wedge \left(i+1 \in I_W^{\varepsilon_0,n}\right), \\ \left(u_{i+\frac{1}{2}}^n = 0\right) \wedge \left(w_i^n \leq b_{i+1} + \varepsilon_0\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_W^{\varepsilon_0,n}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0,n}\right), \\ \qquad u_{i+\frac{1}{2}}^n := 0, & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^{\varepsilon_0,n}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0,n}\right). \end{cases} \tag{aDLAR}$$

Now, we need to turn all intentions to the main issue which (DGsc) collapses completely: the case of the horizontal fluid domain $\mathbb{R}$ being dry-wet.

Theorem 4.1 (Well-balanced property for dry-wet horizontal fluid domain $\mathbb{R}$)**.** *Assume that the initial data satisfies* $(h_0, u_0) \in L^\infty(\mathbb{R}; \mathbb{R}_+) \times L^\infty(\mathbb{R})$*. Let* $\mathcal{T}$ *be an admissible mesh of* $\mathbb{R}$ *and* $\Delta t \in \mathbb{R}_+^\star$ *be the time step. Let* $\left(h_i^n, u_{i+\frac{1}{2}}^n\right)_{i\in\mathbb{Z}, n\in[T]}$ *be the discrete finite volume approximate solution generated by* (OEsc)*. Assume that the CFL-like condition* (CFL3) *or* (CFL4) *holds.*

Let ε_0 *be an arbitrary nonnegative real. Assume that the horizontal fluid domain* $\mathbb{R}$ *contains at least one non-degenerate dry area (not 1-point dry area type) but not fully dry, the offset equilibrium staggered scheme* (OEsc) *is* ε_0*-well-balanced for* (1DSW) *in the dry-wet horizontal fluid domain* $\mathbb{R}$ *in the discrete level.*

Proof. Assume that for some $n_0 \in [T-1]$, $\left(h_i^{n_0}, u_{i+\frac{1}{2}}^{n_0}\right)_{i\in\mathbb{Z}}$ fulfills the globally discrete ε_0-lake at rest equilibrium at $t = t^{n_0}$, i.e.,

$$\begin{cases} \left(u_{i+\frac{1}{2}}^{n_0} = 0\right) \wedge \left(w_i^{n_0} = w_{i+1}^{n_0}\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_W^{\varepsilon_0,n_0}\right) \wedge \left(i+1 \in I_W^{\varepsilon_0,n_0}\right), \\ \left(u_{i+\frac{1}{2}}^{n_0} = 0\right) \wedge \left(b_i + \varepsilon_0 \geq w_{i+1}^{n_0}\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^{\varepsilon_0,n_0}\right) \wedge \left(i+1 \in I_W^{\varepsilon_0,n_0}\right), \\ \left(u_{i+\frac{1}{2}}^{n_0} = 0\right) \wedge \left(w_i^{n_0} \leq b_{i+1} + \varepsilon_0\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_W^{\varepsilon_0,n_0}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0,n_0}\right), \\ \qquad u_{i+\frac{1}{2}}^{n_0} := 0, & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^{\varepsilon_0,n_0}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0,n_0}\right). \end{cases}$$

Note that since the discrete velocity at $t = t^{n_0}$ vanishes on all interfaces on all these cases, the discrete mass conservation equation (dmc) of (OEsc) implies that the discrete water height at $t = t^{n_0+1}$ is the same as that at $t = t^{n_0}$, i.e.,

$$h_i^{n_0+1} = h_i^{n_0}, \quad \forall i \in \mathbb{Z},$$

and thus

$$I_{D,i}^{\varepsilon_0,n_0} = I_{D,i}^{\varepsilon_0,n_0+1}, \quad I_{W,i}^{\varepsilon_0,n_0} = I_{W,i}^{\varepsilon_0,n_0+1}, \quad \forall i \in \mathbb{Z}.$$

Consequently, by the rigidity of the bottom topography again, the discrete water level at $t = t^{n_0+1}$ is also the same as that at $t = t^{n_0}$, i.e.,

$$w_i^{n_0+1} = w_i^{n_0}, \quad \forall i \in \mathbb{Z}.$$

As another consequence, (4.1) at $n = n_0$ and (4.1) at $n = n_0 + 1$ coincides, i.e.,

$$\begin{cases} w_i^{n_0+1} = w_i^{n_0}, & \forall i \in \mathbb{Z}, \\ b_{i+\frac{1}{2}}^{n_0+1} = b_{i+\frac{1}{2}}^{n_0}, & \forall i \in \mathbb{Z}, \\ h_{i+\frac{1}{2}-}^{n_0+1} = h_{i+\frac{1}{2}-}^{n_0}, & \forall i \in \mathbb{Z}, \\ h_{i+\frac{1}{2}-}^{n_0+1} = h_{i+\frac{1}{2}-}^{n_0}, & \forall i \in \mathbb{Z}. \end{cases}$$

The vanishing of discrete velocity at $t = t^{n_0}$ also simplifies the discrete momentum balance equation of (OEsc) to

$$\left(h_i^{n_0+1} + h_{i+1}^{n_0+1}\right) u_{i+\frac{1}{2}}^{n_0+1} = -g\nu_{i+\frac{1}{2}} \left(h_i^{n_0+1} + h_{i+1}^{n_0+1}\right) \left(h_{i+\frac{1}{2}+}^{n_0+1} - h_{i+\frac{1}{2}-}^{n_0+1}\right), \quad \forall i \in \mathbb{Z},$$

and thus

$$\left(h_i^{n_0} + h_{i+1}^{n_0}\right) u_{i+\frac{1}{2}}^{n_0+1} = -g\nu_{i+\frac{1}{2}} \left(h_i^{n_0} + h_{i+1}^{n_0}\right) \left(h_{i+\frac{1}{2}+}^{n_0} - h_{i+\frac{1}{2}-}^{n_0}\right), \quad \forall i \in \mathbb{Z}.$$

If $(i, i+1)$ is ε_0-dry-dry then $h_i^{n_0} = h_{i+1}^{n_0} = 0$ and thus $h_i^{n_0+1} = h_{i+1}^{n_0+1} = 0$. Convention 4.2 then implies $u_{i+\frac{1}{2}}^{n_0+1} := 0$. Otherwise, if $(i, i+1)$ belongs to the remaining three cases, then $h_i^{n_0} + h_{i+1}^{n_0} > 0$ and thus can be dropped in the last equation to obtain

$$u_{i+\frac{1}{2}}^{n_0+1} = -g\nu_{i+\frac{1}{2}} \left(h_{i+\frac{1}{2}+}^{n_0} - h_{i+\frac{1}{2}-}^{n_0} \right), \quad \forall i \in \mathbb{Z}.$$

To prove $u_{i+\frac{1}{2}}^{n_0+1} = 0$, it suffices to prove

$$h_{i+\frac{1}{2}+}^{n_0} = h_{i+\frac{1}{2}-}^{n_0}, \quad \forall i \in \mathbb{Z}. \tag{4.3}$$

in the first three cases of dry/wet transitions.

Depending on the couple of indices $(i, i+1)$, we consider the following three cases (the last one is handled immediately by Convention 4.2).

a) ε_0-*"wet-wet" front*: Suppose that $(i, i+1)$ is "wet-wet", then (4.1) at $n = n_0$ becomes

$$\begin{cases} w_i^{n_0} = w_{i+1}^{n_0}, \\ b_{i+\frac{1}{2}}^{n_0} = \max\left(b_i, b_{i+1}\right), \\ h_{i+\frac{1}{2}-}^{n_0} = \min\left(w_i^{n_0} - \max\left(b_i, b_{i+1}\right), h_i^{n_0}\right) = w_i^{n_0} - \max\left(b_i, b_{i+1}\right), \\ h_{i+\frac{1}{2}+}^{n_0} = \min\left(w_{i+1}^{n_0} - \max\left(b_i, b_{i+1}\right), h_{i+1}^{n_0}\right) = w_{i+1}^{n_0} - \max\left(b_i, b_{i+1}\right), \end{cases}$$

then

$$h_{i+\frac{1}{2}+}^{n_0} = w_{i+1}^{n_0} - \max\left(b_i, b_{i+1}\right) = w_i^{n_0} - \max\left(b_i, b_{i+1}\right) = h_{i+\frac{1}{2}-}^{n_0}, \quad \forall i \in \mathbb{Z}.$$

i.e., (4.3) holds. Thus, the discrete velocity at the interface $x_{i+\frac{1}{2}}$ at $t = t^{n_0+1}$ vanishes, i.e., $u_{i+\frac{1}{2}}^{n_0+1} = 0$. Moreover, $w_i^{n_0+1} = w_i^{n_0} = w_{i+1}^{n_0} = w_{i+1}^{n_0+1}$. Therefore, we obtain in this case,

$$\left(u_{i+\frac{1}{2}}^{n_0+1} = 0\right) \wedge \left(w_i^{n_0+1} = w_{i+1}^{n_0+1}\right), \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_W^{\varepsilon_0, n_0+1}\right) \wedge \left(i+1 \in I_W^{\varepsilon_0, n_0+1}\right).$$

b) ε_0-*dry-wet front*: Suppose that $(i, i+1)$ is "dry-wet", then (4.1) at $n = n_0$ becomes

$$\begin{cases} w_i^{n_0} = b_i \geq w_{i+1}^{n_0} = h_{i+1}^{n_0} + b_{i+1}, \\ b_{i+\frac{1}{2}}^{n_0} = \min\left(b_i, w_{i+1}^{n_0}\right) = w_{i+1}^{n_0}, \\ h_{i+\frac{1}{2}-}^{n_0} = \min\left(w_i^{n_0} - w_{i+1}^{n_0}, 0\right) = 0, \\ h_{i+\frac{1}{2}+}^{n_0} = \min\left(w_{i+1}^{n_0} - w_{i+1}^{n_0}, h_{i+1}^{n_0}\right) = 0, \end{cases}$$

then (4.3) holds. Thus, $u_{i+\frac{1}{2}}^{n_0+1} = 0$. Moreover, $b_i + \varepsilon_0 \geq w_{i+1}^{n_0} = w_{i+1}^{n_0+1}$. Therefore, we obtain in this case,

$$\left(u_{i+\frac{1}{2}}^{n_0+1} = 0\right) \wedge \left(b_i + \varepsilon_0 \geq w_{i+1}^{n_0+1}\right), \quad \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^{\varepsilon_0, n_0+1}\right) \wedge \left(i+1 \in I_W^{\varepsilon_0, n_0+1}\right).$$

c) ε_0-*wet-dry front*: Suppose that $(i, i+1)$ is "wet-dry", then (4.1) at $n = n_0$ becomes

$$\begin{cases} w_i^{n_0} = h_i^{n_0} + b_i \le w_{i+1}^{n_0} = b_{i+1}, \\ b_{i+\frac{1}{2}}^{n_0} = \min\left(b_{i+1}, w_i^{n_0}\right) = w_i^{n_0}, \\ h_{i+\frac{1}{2}-}^{n_0} = \min\left(w_i^{n_0} - w_i^{n_0}, h_i^{n}\right) = 0, \\ h_{i+\frac{1}{2}+}^{n_0} = \min\left(w_{i+1}^{n_0} - w_i^{n_0}, 0\right) = 0, \end{cases}$$

then (4.3) holds. Thus, $u_{i+\frac{1}{2}}^{n_0+1} = 0$. Moreover, $b_{i+1} + \varepsilon_0 \ge w_i^{n_0} = w_i^{n_0+1}$. Therefore, we obtain in this case,

$$\left(u_{i+\frac{1}{2}}^{n_0+1} = 0\right) \wedge \left(w_i^{n_0+1} \le b_{i+1} + \varepsilon_0\right), \ \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_W^{\varepsilon_0, n_0+1}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0, n_0+1}\right).$$

Summing up these cases, we obtain

$$\begin{cases} \left(u_{i+\frac{1}{2}}^{n_0+1} = 0\right) \wedge \left(w_i^{n_0+1} = w_{i+1}^{n_0+1}\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_W^{\varepsilon_0, n_0+1}\right) \wedge \left(i+1 \in I_W^{\varepsilon_0, n_0+1}\right), \\ \left(u_{i+\frac{1}{2}}^{n_0+1} = 0\right) \wedge \left(b_i + \varepsilon_0 \ge w_{i+1}^{n_0+1}\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^{\varepsilon_0, n_0+1}\right) \wedge \left(i+1 \in I_W^{\varepsilon_0, n_0+1}\right), \\ \left(u_{i+\frac{1}{2}}^{n_0+1} = 0\right) \wedge \left(w_i^{n_0+1} \le b_{i+1} + \varepsilon_0\right), & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_W^{\varepsilon_0, n_0+1}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0, n_0+1}\right), \\ \quad u_{i+\frac{1}{2}}^{n_0+1} := 0, & \forall i \in \mathbb{Z} \text{ s.t. } \left(i \in I_D^{\varepsilon_0, n_0+1}\right) \wedge \left(i+1 \in I_D^{\varepsilon_0, n_0+1}\right), \end{cases}$$

i.e., $\left(h_i^{n_0+1}, u_{i+\frac{1}{2}}^{n_0+1}\right)_{i\in\mathbb{Z}}$ fulfills the globally discrete ε_0-lake at rest equilibrium at $t = t^{n_0}$. By Definition 4.2, (OEsc) is ε_0-well-balanced for the horizontal fluid domain $\mathbb{R}$ being both dry and wet. □

1.5 Higher-order staggered upwind schemes on the horizontal fluid domain

Similar to Chapter 3, Section 1.5, the 5-point offset equilibrium staggered upwind scheme is defined as

$$\begin{cases} w_i^{5,n} := h_i^{5,n} + b_i, \\ b_{i+\frac{1}{2}}^{5,n} := \min\left(\max\left(b_i, b_{i+1}\right), \min\left(w_i^{5,n}, w_{i+1}^{5,n}\right)\right), \\ h_{i+\frac{1}{2}-}^{5,n} := \min\left(w_i^{5,n} - b_{i+\frac{1}{2}}^{5,n}, h_i^{5,n}\right), \\ h_{i+\frac{1}{2}+}^{5,n} := \min\left(w_{i+1}^{5,n} - b_{i+\frac{1}{2}}^{5,n}, h_{i+1}^{5,n}\right), \end{cases}$$

for all $i \in \mathbb{Z}$, $n \in [T]$, and

$$\begin{cases} h_i^{5,n+1} := h_i^{5,n} - \nu_i \left(F_{i+\frac{1}{2}}^{5,n} - F_{i-\frac{1}{2}}^{5,n} \right), \\ \bar{h}_{i+\frac{1}{2}}^{5,n+1} u_{i+\frac{1}{2}}^{5,n+1} := \bar{h}_{i+\frac{1}{2}}^{5,n} u_{i+\frac{1}{2}}^{5,n} - \nu_{i+\frac{1}{2}} \left[G_{i+1}^{5,n} - G_i^{5,n} + g \bar{h}_{i+\frac{1}{2}}^{5,n+1} \left(h_{i+\frac{1}{2}+}^{5,n+1} - h_{i+\frac{1}{2}-}^{5,n+1} \right) \right], \end{cases} \tag{5pOEsc}$$

for all $i \in \mathbb{Z}$, $n \in [T-1]$. The constructed 5-point scheme (5pOEsc) is hoped to be less diffusive than (OEsc) and it remains stable thanks to the limitation of the slope. A higher order scheme for the time discretization is used, e.g., Runge-Kutta, or Heun method for the discretization of the time derivative.

The MUSCL scheme may be written as

$$\begin{cases} \dfrac{h^{5,n+1} - h^{5,n}}{\Delta t} = \overline{H}_1^{\mathrm{OE}} \left(h^{5,n} \right), & \forall n \in [T-1], \\ \dfrac{\left(\bar{h}u\right)^{5,n+1} - \left(\bar{h}u\right)^{5,n}}{\Delta t} = \overline{H}_2^{\mathrm{OE}} \left(\left(\bar{h}u\right)^{5,n} \right), & \forall n \in [T-1], \end{cases}$$

which may be seen as the explicit Euler discretization of

$$\begin{cases} \partial_t h^5 = \overline{H}_1^{\mathrm{OE}} \left(h^5 \right), & \forall n \in [T-1], \\ \partial_t \left(\left(\bar{h}u\right)^5 \right) = \overline{H}_2^{\mathrm{OE}} \left(\left(\bar{h}u\right)^5 \right), & \forall n \in [T-1]. \end{cases}$$

The RK2 time discretization yields the following scheme

$$\begin{cases} \dfrac{h^{5,n+1} - h^{5,n}}{\Delta t} = \dfrac{1}{2} \overline{H}_1^{\mathrm{OE}} \left(h^{5,n} \right) + \dfrac{1}{2} \overline{H}_1^{\mathrm{OE}} \left(h^{5,n} + \Delta t \overline{H}_1^{\mathrm{OE}} \left(h^{5,n} \right) \right), \\ \dfrac{\left(\bar{h}u\right)^{5,n+1} - \left(\bar{h}u\right)^{5,n}}{\Delta t} = \dfrac{1}{2} \overline{H}_2^{\mathrm{OE}} \left(\left(\bar{h}u\right)^{5,n} \right) + \dfrac{1}{2} \overline{H}_2^{\mathrm{OE}} \left(\left(\bar{h}u\right)^{5,n} + \Delta t \overline{H}_2^{\mathrm{OE}} \left(\left(\bar{h}u\right)^{5,n} \right) \right), \end{cases} \tag{RK2-5pOEsc}$$

for all $n \in [T-1]$. Such a second-order discretization in time allows larger time steps, without loss of stability. The 5-point scheme (RK2-5pOEsc) is proposed here for future purposes, the properties of this high-order scheme are not investigated in the scope of this context.

2 An offset equilibrium scheme in the computational horizontal fluid domain

The framework on the whole horizontal fluid domain $\mathbb{R}$ presented in the previous section can be restricted straightforwardly into a computational horizontal fluid domain $\Omega_c = [L_1, L_2]$ for some reals $L_1 < L_2$. See Chapter 3, Section 2 for the main ideas.

Chapter 5

2D Shallow Water Equations

1 Properties of 2D shallow water equation

We can extend our framework to the following 2D shallow equation, which is obtained by let $d = 2$ in (SW$_2$):

$$\begin{cases} \partial_t h + \partial_x (hu) + \partial_y (hu) = 0, & \text{in } [0, T_0) \times \mathbb{R}^2, \\ \partial_t (hu) + \partial_x \left(hu^2 + \frac{g}{2} h^2 \right) + \partial_y (huv) = -gh\partial_x b, & \text{in } [0, T_0) \times \mathbb{R}^2, \\ \partial_t (hv) + \partial_x (huv) + \partial_y \left(hu^2 + \frac{g}{2} h^2 \right) = -gh\partial_y b, & \text{in } [0, T_0) \times \mathbb{R}^2. \end{cases} \tag{2DSW}$$

Appendix A

Systems of Conservation Laws in One Space Dimension

This appendix is based heavily on the course[1] *Hyperbolic equations* taught by Prof. Florian Méhats, the course *Numerical Transport* taught by Prof. Nicolas Seguin, and also Eymard, Gallouët, and Herbin, 2003, Evans, 2010, Benzoni-Gavage and Serre, 2007, Harten, Lax, and Leer, 1983, with notations adapted to the main style of this context.

In these appendices, we will use capital letters (e.g., U, F, G, etc.) for vector-valued function, i.e., the case of systems of conservation laws, and normal letters (u, f, g[2], etc.) for scalar functions, i.e., the case of scalar conservation laws.

1 Introduction

Consider a PDE of the form

$$\partial_t U + \partial_x \left(F\left(U\right)\right) = 0, \text{ in } \mathbb{R}_+ \times \mathbb{R}^m,$$

and then

$$\frac{d}{dt}\int_A^B U\left(t,x\right)dx + F\left(U\left(t,B\right)\right) - F\left(U\left(t,A\right)\right) = 0, \ \forall t \in \mathbb{R}_+, \ \forall A \in \mathbb{R}^m, \ \forall B \in \mathbb{R}^m.$$

The evolution of the quantity $\int_A^B U\left(t,x\right)dx$ only depends on the fluxes at points A and B:

$$\int_A^B U\left(T,x\right)dx = \int_A^B U\left(0,x\right)dx - \int_0^T \left[F\left(U\left(t,B\right)\right) - F\left(U\left(t,A\right)\right)\right]dt,$$

for all $T \in \mathbb{R}_+^\star$, $A \in \mathbb{R}^m$, $B \in \mathbb{R}^m$.

[1] Master 2 in Fundamental Mathematics and Application Program, 2018-2019, Université de Rennes 1, France.

[2] In appendices, g is used to denote the *numerical flux* of finite volume schemes, not the gravitational constant as in previous chapters.

Consider the following initial-value problem (or Cauchy problem) for a systems of conservation laws in one space dimension:

$$\begin{cases} \partial_t U + \partial_x (F(U)) = 0, & \forall (t,x) \in \mathbb{R}_+ \times \mathbb{R}, \\ U(0,x) = U_0(x), & \forall x \in \mathbb{R}, \end{cases} \tag{A.1}$$

where $F \in C^1(\mathbb{R}^m, \mathbb{R}^m)$ and $U_0 \in L^\infty(\mathbb{R}, \mathbb{R}^m)$ are given and $U : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}^m$ is the unknown, $U = U(t,x)$. The Euclidean space $\mathbb{R}^m$ is called the *state space* and the smooth flux vector field F is written as

$$F = F(z) = (F_1(z), \dots, F_m(z)), \ \forall z \in \mathbb{R}^m.$$

Definition A.1 (Classical solution)**.** *Let $F \in C^1(\mathbb{R}^m, \mathbb{R}^m)$ and $U_0 \in C^1(\mathbb{R}, \mathbb{R}^m)$. A* classical solution *to* (A.1) *is a function $U \in C^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R})$ such that*

$$\begin{cases} \partial_t U + J_F(U) \cdot \partial_x U = 0, & \textit{in } \mathbb{R}_+ \times \mathbb{R}, \\ U(0,x) = U_0(x), & \forall x \in \mathbb{R}. \end{cases}$$

In the simplest linear case, i.e., $F(X) = C \odot X$ for all $X \in \mathbb{R}^m$, for some $C = (C_1, \dots, C_m) \in \mathbb{R}^m$, where $\odot$ denotes *Hadamard product* (also *Schur product, entrywise product*) for matrices, there exists a unique classical solution provided $U_0 \in C^1(\mathbb{R}, \mathbb{R}^m)$. It is

$$U(t,x) = (U_{0,1}(x - C_1 t), \dots, U_{0,m}(x - C_m t)), \text{ in } \mathbb{R}_+ \times \mathbb{R}.$$

However, in the nonlinear case, the existence of such a solution depends on the initial data U_0. The following proposition will illustrate this point for the scalar $(m = 1)$ conservation law:

$$\begin{cases} \partial_t u + \partial_x (f(u)) = 0, & \text{in } \mathbb{R}_+ \times \mathbb{R}, \\ u(0,x) = u_0(x), & \forall x \in \mathbb{R}. \end{cases} \tag{A.2}$$

Proposition A.1 (Nonexistence of classical solution in the nonlinear case)**.** *Let $f \in C^1(\mathbb{R}, \mathbb{R})$ be a nonlinear function, i.e., such that there exist s_1, $s_2 \in \mathbb{R}$ satisfying $f'(s_1) \neq f'(s_2)$, then there exists $u_0 \in C_c^\infty(\mathbb{R}, \mathbb{R})$ such that* (A.2) *has no classical solution.*

1.1 Method of characteristics for scalar conservation laws

If u is a classical solution to (A.2), then the PDE in (A.2) is equivalent to

$$\partial_t u + f'(u)\, \partial_x u = 0, \text{ in } \mathbb{R}_+ \times \mathbb{R}.$$

The characteristics are the solution $X(t,x)$ of the following ODE:

$$\begin{cases} \partial_t X(t,x) = f'(u(t, X(t,x))), & \text{in } \mathbb{R}_+ \times \mathbb{R}, \\ X(0,x) = x, & \forall x \in \mathbb{R}. \end{cases} \tag{A.3}$$

The usefulness of characteristics is indicated in the following proposition.

Proposition A.2. *Let u be a classical solution of* (A.2)*. Then u is constant along the characteristics $(t, X(t,x))$, where X is defined by* (A.3)*.*

Proof. It is straightforward that

$$\begin{aligned}\frac{d}{dt}\left(u\left(t, X\left(t,x\right)\right)\right) &= \partial_t u\left(t, X\left(t,x\right)\right) + \partial_t X\left(t,x\right)\partial_x u\left(t, X\left(t,x\right)\right)\\ &= \left(\partial_t u + f'\left(u\right)\partial_x u\right)\left(t, X\left(t,x\right)\right) = 0, \text{ in } \mathbb{R}_+ \times \mathbb{R}.\end{aligned}$$

The proof is accomplished. □

Hence, $u(t, X(t,x))$ is a constant of t, depending only on x, and thus

$$u\left(t, X\left(t,x\right)\right) = u\left(0, X\left(0,x\right)\right) = u\left(0,x\right) = u_0\left(x\right), \text{ in } \mathbb{R}_+ \times \mathbb{R}.$$

The characteristics are global and are written explicitly as

$$X\left(t,x\right) = f'\left(u_0\left(x\right)\right)t + x, \text{ in } \mathbb{R}_+ \times \mathbb{R}.$$

In the (x,t) coordinates, the characteristic $X(t,x)$ is the equation of a straight line starting from the point $(x,0)$.

Denote $X_t(x) := X(t,x)$, Proposition A.2 gives us

$$u\left(t, X_t\left(x\right)\right) = u_0\left(x\right), \text{ in } \mathbb{R}_+ \times \mathbb{R}.$$

If $\mathcal{X} : x \mapsto X_t(x)$ is a C^1-diffeomorphism on $\mathbb{R}$, then we necessarily have the following expression of the classical solution u:

$$u\left(t,x\right) = u_0\left(X_t^{-1}\left(x\right)\right), \text{ in } \mathbb{R}_+ \times \mathbb{R}.$$

We now assume that

i) $f \in C^2(\mathbb{R},\mathbb{R})$ so that $f' \in C^1(\mathbb{R},\mathbb{R})$,

ii) $u_0 \in \left(C^1 \cap L^\infty\right)(\mathbb{R},\mathbb{R})$ with $u_0' \in L^\infty(\mathbb{R},\mathbb{R})$, and $X_0 = \mathrm{Id}$.

Then the function $f'(u_0(x))t$ is a bounded function of x, for all $t \in \mathbb{R}_+$, and thus $X_t(x) \to \pm\infty$ as $x \to \pm\infty$. Therefore, given $t \in \mathbb{R}_+^\star$ arbitrarily, X_t is a C^1 diffeomorphism if and only if $X_t'(x) > 0$, for all $x \in \mathbb{R}$. That means,

$$\left(f' \circ u_0\right)'\left(x\right)t + 1 > 0, \ \ \forall x \in \mathbb{R},$$

which is equivalent to

$$\left(f' \circ u_0\right)'\left(x\right) > -\frac{1}{t}, \ \ \forall x \in \mathbb{R}.$$

Let $\alpha := \inf_{x\in\mathbb{R}}\left(f' \circ u_0\right)'(x)$. We consider the following cases depending on the positivity of α:

i) *Case* $\alpha \geq 0$: In this case, the last inequality holds for all $t \in \mathbb{R}_+^\star$, and thus X_t is a C^1 diffeomorphism for all $t \in \mathbb{R}_+^\star$. Therefore, u is globally defined.

ii) *Case* $\alpha < 0$: Then for $t > -\frac{1}{\alpha}$, we have $\alpha < -\frac{1}{t}$. So there exists $x_0 \in \mathbb{R}$ such that $\left(f' \circ u_0\right)'(x_0) < -\frac{1}{t}$, and thus X_t is not a diffeomorphism anymore.

Lemma A.1. *Define*

$$T^{\star} := \begin{cases} +\infty, & \text{if } \alpha \geq 0, \\ -\dfrac{1}{\alpha}, & \text{if } \alpha < 0, \end{cases}$$

then (A.2) *admits a unique classical solution which is maximal on* $[0, T_0^{\star})$.

Proof. We have proved that, for $0 \leq t < T^{\star}$, $x \mapsto X_t(x)$ is a C^1 diffeomorphism. Hence, it is clear that on $[0, T_0^{\star})$, the unique solution of the problem is given by

$$u(t,x) = u_0\left(X_t^{-1}(x)\right), \text{ in } [0, T^{\star}) \times \mathbb{R}.$$

We not check that it is a solution. Recall from the previous paragraph that

$$\forall (t,x) \in [0, T^{\star}) \times \mathbb{R}, \quad \begin{cases} u(t, X_t(x)) = u_0(x), \\ X_t(x) = f'(u_0(x))\, t + x, \end{cases}$$

differentiating the former w.r.t. the time variable t yields

$$\begin{aligned} 0 &= \partial_t u(t, X_t(x)) + f'(u_0(x))\, \partial_x u(t, X_t(x)) \\ &= \partial_t u(t, X_t(x)) + f'(u(t, X_t(x)))\, \partial_x u(t, X_t(x)) \\ &= \left(\partial_t u + f'(u)\, \partial_x u\right)(t, X_t(x)), \text{ in } [0, T^{\star}) \times \mathbb{R}. \end{aligned}$$

Let $y \in \mathbb{R}$, $t \in [0, T_0^{\star})$, and $x := X_t^{-1}(y)$, then the last equation reads exactly the PDE in (A.2). □

Property of maximality. Let $\alpha < 0$ and $T^{\star} = -\frac{1}{\alpha}$, $\alpha := \inf_{x \in \mathbb{R}} \left(f' \circ u_0\right)'(x) < 0$. There exists a real x such that $\left(f' \circ u_0\right)'(x) < 0$, and thus there exist two reals $x_1 < x_2$ such that $\left(f' \circ u_0\right)'(x_1) > \left(f' \circ u_0\right)'(x_2)$. The two characteristics $X_t(x_1)$ and $X_t(x_2)$ intersect each other at time, denoted by $T(x_1, x_2)$, and their common value is denoted by $X(x_1, x_2)$.

If u is defined and C^1 until $T(x_1, x_2)$, then u is constant along the characteristics. Hence,

$$u(0, x_1) = u(0, x_2) = u\left(T(x_1, x_2), X(x_1, x_2)\right),$$

this implies $u_0(x_1) = u_0(x_2)$, which contradicts with the definition of x_1, x_2. Therefore, u cannot be solution until $T(x_1, x_2)$.

We now claim that

$$\inf_{x_1 < x_2 \text{ s.t. } (f' \circ u_0)(x_1) > (f' \circ u_0)(x_2)} T(x_1, x_2) = T^{\star}.$$

Assume $\alpha = \inf \left(f' \circ u_0\right)' < 0$, let x_1^n such that $\left(f' \circ u_0\right)'(x_1^n) \to \alpha$.

The equation for intersection of the two characteristics is given by

$$\left(f' \circ u_0\right)(x_1)\, t + x_1 = \left(f' \circ u_0\right)(x_2)\, t + x_2.$$

Solving this for t yields[3]

$$T(x_1, x_2) = -\frac{x_2 - x_1}{(f' \circ u_0)(x_2) - (f' \circ u_0)(x_1)}.$$

There exist two sequences $x_1^n < x_2^n$ such that

$$\frac{(f' \circ u_0)(x_2^n) - (f' \circ u_0)(x_1^n)}{x_2^n - x_1^n} \to \alpha.$$

Therefore,

$$T(x_1^n, x_2^n) \to -\frac{1}{\alpha} = T^\star.$$

Characteristics for the flux $f(x,u)$. It should be emphasized that if f depends on x and u (rather than only u, the characteristics are no longer straight lines.

Indeed, we consider the Cauchy problem (A.2) obtained by replacing $f(u)$ by $f(x,u)$, i.e.,

$$\begin{cases} \partial_t u + \partial_x (f(x,u)) = 0, & \text{in } \mathbb{R}_+ \times \mathbb{R}, \\ u(0,x) = u_0(x), & \forall x \in \mathbb{R}, \end{cases}$$

with $f \in C^1(\mathbb{R}^2, \mathbb{R})$ and $u_0 \in C^1(\mathbb{R}, \mathbb{R})$. Suppose $u \in C^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R})$ is a classical solution to this Cauchy problem, i.e.,

$$\begin{cases} \partial_t u + \partial_1 f(x,u) + \partial_2 f(x,u)\, \partial_x u = 0, & \text{in } \mathbb{R}_+ \times \mathbb{R}, \\ u(0,x) = u_0(x), & \forall x \in \mathbb{R}. \end{cases}$$

Suppose for the contrary that all the characteristics are straight lines, i.e.,

$$X(t,x) = a(x)t + b(x), \ \forall (t,x) \in \mathbb{R}_+ \times \mathbb{R},$$

where $a, b \in C(\mathbb{R}, \mathbb{R})$. Then

$$\begin{aligned} \frac{d}{dt} u(t, X(t,x)) &= \partial_t u(t, X(t,x)) + \partial_t X(t,x)\, \partial_x u(t, X(t,x)) \\ &= -\partial_1 f(X(t,x), u(t, X(t,x))) + \partial_t X(t,x)\, \partial_x u(t, X(t,x)) \\ &\quad - \partial_2 f(X(t,x), u(t, X(t,x)))\, \partial_x u(t, X(t,x)) \\ &= -\partial_1 f(a(x)t + b(x), u(t, a(x)t + b(x))) + a(x)\, \partial_x u(t, X(t,x)) \\ &\quad - \partial_2 f(a(x)t + b(x), u(t, a(x)t + b(x)))\, \partial_x u(t, a(x)t + b(x)), \end{aligned}$$

for all $(t,x) \in \mathbb{R}_+ \times \mathbb{R}$. Notice that the last expression will be nonzero where $\partial_x u = 0$ but $\partial_1 f \neq 0$. This contradicts implies that the characteristics in this case cannot be straight lines. □

[3] Thus, these characteristics intersect at the point whose x-coordinate are given by

$$\begin{aligned} X(x_1, x_2) &= (f' \circ u_0)(x_1)\, T(x_1, x_2) + x_1 = (f' \circ u_0)(x_2)\, T(x_1, x_2) + x_2 \\ &= \frac{x_1 (f' \circ u_0)(x_2) - x_2 (f' \circ u_0)(x_1)}{(f' \circ u_0)(x_2) - (f' \circ u_0)(x_1)}. \end{aligned}$$

Example A.1 (1D Shallow water equations without bottom topography)**.** *The 1D* shallow water equations without bottom topography *are*

$$\begin{cases} \partial_t h + \partial_x (hu) = 0, \\ \partial_t (hu) + \partial_x \left(hu^2 + \frac{g}{2} h^2 \right) = 0, \end{cases} \tag{A.4}$$

in $\mathbb{R}_+ \times \mathbb{R}$*, where* u *is the* horizontal velocity *and* $h > 0$ *is the* water height. *The first equation in* (A.4) *is the* conservation of mass*, and the second one is the* conservation of momentum.

Putting $q := hu$*,* (A.4) *can be rewritten into standard conservation law form:*

$$\begin{cases} \partial_t h + \partial_x q = 0, \\ \partial_t q + \partial_x \left(\frac{q^2}{h} + \frac{g}{2} h^2 \right) = 0. \end{cases} \tag{A.5}$$

Here

$$F(z) := \left(z_2, \frac{z_2^2}{z_1} + \frac{z_1^2}{2} \right), \quad \forall z = (z_1, z_2), \; z_1 > 0.$$

Remark A.1. *If the domain contains dry areas (i.e., where* $h = 0$*),* (1DSW) *can not be rewritten in the form* (A.5)*.*

1.2 Integral solutions

The following definition provides us a notion of weak solution for the initial-value problem (A.1), when the classical solutions do not exist.

Definition A.2 (Integral solution)**.** *A function* $U \in L^\infty \left(\mathbb{R}_+^\star \times \mathbb{R}, \mathbb{R}^m \right)$ *is an* integral solution *of the initial-value problem* (A.1) *provided the following equality holds for all* test functions $V \in C_c^\infty (\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}^m)$*:*

$$\int_0^\infty \int_{\mathbb{R}} (U \cdot \partial_t V + F(U) \cdot \partial_x V) \, dx dt + \int_{\mathbb{R}} U_0 \cdot V|_{t=0} dx = 0. \tag{A.6}$$

Remark A.2.
1. *If* $U \in C^1 (\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}^m) \cap L^\infty \left(\mathbb{R}_+^\star \times \mathbb{R}, \mathbb{R}^m \right)$ *then* U *is a integral solution of* (A.1) *if and only if* U *is a classical solution.*
2. *In Definition* A.2*, the test function* V *is required to belong to* $C_c^\infty \left(\mathbb{R}_+^\star \times \mathbb{R}, \mathbb{R}^m \right)$*, so that* V *may be nonzero at time* $t = 0$*.*

It can be proved that there exists at least one integral solution to (A.1). In the linear case, i.e., $F(X) = C \odot X$, for all $X \in \mathbb{R}^m$, for some $C \in \mathbb{R}^m$, this solution is unique and given by

$$U(t, x) = (U_{0,1}(x - C_1 t), \ldots, U_{0,m}(x - C_m t)), \quad \text{for a.e. } (t, x) \in \mathbb{R}_+ \times \mathbb{R}.$$

However, the uniqueness of this weak solution in the general nonlinear case is no longer true. Hence the concept of entropy solution, for which an existence and uniqueness result is known, will be introduced later.

Condition (A.6) is equivalent to requiring that for all rectangles $(t_1, t_2) \times (a, b)$ the relation obtained by integrating the PDE in (A.1) over the rectangle should hold:

$$\int_a^b U\left(t_2, x\right) dx - \int_a^b U\left(t_1, x\right) dx + \int_{t_1}^{t_2} F\left(U\left(t, b\right)\right) dt - \int_{t_1}^{t_2} F\left(U\left(t, a\right)\right) dt = 0.$$

Let $\widetilde{\Omega} \subset \mathbb{R}_+ \times \mathbb{R}$ be some region cut by a smooth curve C into a left-hand part $\widetilde{\Omega}_L$ and a right-hand part $\widetilde{\Omega}_R$. Suppose the curve C is represented parametrically as

$$C = \left\{(t, x) \in \mathbb{R}_+^\star \times \mathbb{R}; x = s_C\left(t\right)\right\},$$

for some smooth function $s_C : \mathbb{R}_+ \to \mathbb{R}$.

The well-known *Rankine-Hugoniot jump condition* reads

$$F\left(U_L\right) - F\left(U_R\right) = s_C'\left(U_L - U_R\right), \tag{A.7}$$

in $\widetilde{\Omega}$ along the curve C.

Notation A.1 (Jump, speed of a curve)**.**

$$\begin{aligned} [[U]] &:= U_L - U_R = \text{jump in } U \text{across the curve } C, \\ [[F\left(U\right)]] &:= F\left(U_L\right) - F\left(U_R\right) = \text{jump in } F\left(U\right), \\ \sigma_C &:= {s_C}' = \text{speed of the curve } C. \end{aligned}$$

Under this notation, (A.7) can be rewritten in the following condensed form:

$$[[F\left(U\right)]] = \sigma_C\left[[U]\right],$$

along the discontinuity curve C.

1.3 Traveling waves, hyperbolic systems

Consider the class of *semilinear systems* having the nondivergence form:

$$\partial_t U + B\left(U\right) \partial_x U = 0, \text{ in } \mathbb{R}_+^\star \times \mathbb{R}, \tag{A.8}$$

where $B : \mathbb{R}^m \to \mathbb{M}^{m\times m}$. For smooth functions, this system is equivalent to the conservation law in (A.1) provided $B = J_F = \left(\partial_{z_j} F_i\right)_{i,j=1}^m$.

Definition A.3 (Strict hyperbolicity)**.** *The system* (A.8) *is called* strictly hyperbolic *if for each $z \in \mathbb{R}^m$, the eigenvalues of $B\left(z\right)$ are real and distinct.*

Notation A.2 (Eigenvalues, left eigenvectors, right eigenvectors)**.** *i) The real and distinct eigenvalues of $B\left(z\right)$, in increasing order, are denoted by*

$$\lambda_1\left(z\right) < \cdots < \lambda_m\left(z\right), \ \ \forall z \in \mathbb{R}^m.$$

ii) Let $\left\{l_k\left(z\right)\right\}_{k=1}^m$ and $\left\{r_k\left(z\right)\right\}_{k=1}^m$ be left eigenvectors *and* right eigenvectors *of $B\left(z\right)$, respectively.*

The vectors $\{l_k(z)\}_{k=1}^m$ (or $\{r_k(z)\}_{k=1}^m$) span $\mathbb{R}^m$ for each $z \in \mathbb{R}^m$ provided the strict hyperbolicity condition holds. Additionally,

$$l_p(z) \cdot r_q(z) = 0, \text{ if } p \neq q, \ \forall z \in \mathbb{R}^m.$$

Next, the notion of strict hyperbolicity is independent of coordinates.

Theorem A.1 (Invariance of hyperbolicity under change of coordinates)**.** *Let U be a smooth solution of the strictly hyperbolic system (A.8). Assume also $\Phi : \mathbb{R}^m \to \mathbb{R}^m$ is a smooth diffeomorphism, with inverse Ψ. Then $\widetilde{U} := \Phi(U)$ solves the strictly hyperbolic system*

$$\partial_t \widetilde{U} + \widetilde{B}\left(\widetilde{U}\right)\partial_x \widetilde{U} = 0, \ \text{in } \mathbb{R}_+^\star \times \mathbb{R},$$

for

$$\widetilde{B}(\widetilde{z}) := \nabla\Phi(\Psi(\widetilde{z}))\, B(\Psi(\widetilde{z}))\, D\Psi(\widetilde{z}), \ \forall \widetilde{z} \in \mathbb{R}^m.$$

The following theorem indicates the smooth dependence of the eigenvalues, left and right eigenvectors on the parameter $z \in \mathbb{R}^m$.

Theorem A.2 (Dependence of eigenvalues and eigenvectors on parameters)**.** *Assume the matrix function B is smooth, strictly hyperbolic.*

i) Then the eigenvalues $\lambda_k(z)$ depend smoothly on $z \in \mathbb{R}^m$, for all $k = 1, \ldots, m$.

ii) Furthermore, the left eigenvectors $l_k(z)$ and the right eigenvectors $r_k(z)$ can be selected to depend smoothly on $z \in \mathbb{R}^m$ and satisfy the normalization

$$|l_k(z)| = |r_k(z)| = 1, \ \forall k \in \{1, \ldots, m\}.$$

Thanks to the proof of this theorem in Evans, 2010, pp. 618–619 Not only the eigenvalues and eigenspaces of B are globally and smoothly defined, but also these eigenspaces are provided with an orientation.

2 Riemann's problem

We consider the following *Riemann's problem*

$$\begin{cases} \partial_t U + \partial_x (F(U)) = 0, \text{ in } \mathbb{R}_+^\star \times \mathbb{R}, \\ U(0,x) = U_0(x) := \begin{cases} U_L, & \text{if } x < 0, \\ U_R, & \text{if } x > 0, \end{cases} \text{ in } \mathbb{R}. \end{cases} \tag{A.9}$$

The given vectors U_L and U_R are called the left and right *initial states*, respectively.

2.1 Simple waves

The *simple waves* are solutions of (A.9) having the structure

$$U\left(t,x\right)=V\left(W\left(t,x\right)\right),\ \text{ in } \mathbb{R}_{+}^{\star}\times\mathbb{R}, \tag{A.10}$$

where $v:\mathbb{R}^{m}\rightarrow\mathbb{R}^{m}$, $v=\left(v^{1},\ldots,v^{m}\right)$, and $w:\mathbb{R}_{+}\times\mathbb{R}\rightarrow\mathbb{R}$.

The function U defined by (A.10) is caleld a *k-simple wave* if

$$\partial_{t}W+\lambda_{k}\left(V\left(W\right)\right)\partial_{x}W=0, \tag{A.11}$$

$$V'\left(s\right)=r_{k}\left(V\left(s\right)\right). \tag{A.12}$$

Definition A.4 (k^{th}-rarefaction curve)**.** *Given a fixed state $z_0\in\mathbb{R}^m$, the* k^{th}-rarefaction curve *$R_k\left(z_0\right)$ is defined to be the path in $\mathbb{R}^m$ of the solution of the ODE* (A.12) *which passes through z_0.*

Definition A.5 (Genuine nonlinearity, linear degeneracy)**.** *i) The pair $\left(\lambda_k\left(z\right),r_k\left(z\right)\right)$ is called* genuinely nonlinear *provided*

$$D\lambda_k\left(z\right)\cdot r_k\left(z\right)\neq0,\ \ \forall z\in\mathbb{R}^m.$$

ii) The pair $\left(\lambda_k\left(z\right),r_k\left(z\right)\right)$ is called linearly degenerate *if*

$$D\lambda_k\left(z\right)\cdot r_k\left(z\right)=0,\ \ \forall z\in\mathbb{R}^m.$$

Notation A.3. *If the pair $\left(\lambda_k\left(z\right),r_k\left(z\right)\right)$ is genuinely nonlinear, write*

$$R_k^{+}\left(z_0\right):=\left\{z\in R_k\left(z_0\right);\lambda_k\left(z\right)>\lambda_k\left(z_0\right)\right\},$$
$$R_k^{-}\left(z_0\right):=\left\{z\in R_k\left(z_0\right);\lambda_k\left(z\right)<\lambda_k\left(z_0\right)\right\}.$$

Then

$$R_k\left(z_0\right)=R_k^{-}\left(z_0\right)\cup\left\{z_0\right\}\cup R_k^{+}\left(z_0\right).$$

2.2 Rarefaction waves

Theorem A.3 (Existence of k-rarefaction waves)**.** *Suppose that for some $k\in\{1,\ldots,m\}$,*

i) the pair $\left(\lambda_k,r_k\right)$ is genuinely nonlinear, and

ii) $U_R\in R_k^{+}\left(U_L\right)$.

Then there exists a continuous integral solution U of Riemann's problem (A.9)*, which is a k-simple wave constant along lines through the origin.*

This solution U is called a (centered) k-rarefaction wave.

More explicitly, let $W_L,W_R\in\mathbb{R}$ such that $U_L=V\left(W_L\right)$, $U_R=V\left(W_R\right)$. For the case $W_L<W_R$, the centered k-rarefaction wave is given by $U\left(t,x\right)=V\left(W\left(t,x\right)\right)$,

where V solves (A.12) and passes through U_L and W is given by

$$W\left(t,x\right)=\begin{cases} W_L, \text{ if } \dfrac{x}{t}<F_k{}'\left(W_L\right), \\ \left(F_k{}'\right)^{-1}\left(\dfrac{x}{t}\right), \text{ if } F_k{}'\left(W_L\right)<\dfrac{x}{t}<F_k{}'\left(W_R\right), \\ W_R, \text{ if } \dfrac{x}{t}>F_k{}'\left(W_R\right), \end{cases} \text{ in } \mathbb{R}_+^\star\times\mathbb{R}. \tag{A.13}$$

The case $W_L>W_R$ is handled similarly, since F_k is concave in that case.

2.3 Shock waves, contact discontinuities

2.3.1 Shock set

The necessary condition for a shock wave joining the states U_L and U_R to exist is the following Rankine-Hugoniot condition holds for some $\sigma\in\mathbb{R}$:

$$F\left(U_L\right)-F\left(U_R\right)=\sigma\left(U_L-U_R\right). \tag{A.14}$$

Definition A.6 (Shock set)**.** *Given a fixed state $z_0\in\mathbb{R}^m$, the* shock set *is defined by*

$$S\left(z_0\right):=\left\{z\in\mathbb{R}^m;F\left(z\right)-F\left(z_0\right)=\sigma\left(z-z_0\right)\ \textit{for a constant } \sigma=\sigma\left(z,z_0\right)\right\}.$$

Theorem A.4 (Structure of the shock set)**.** *Fix $z_0\in\mathbb{R}^m$. In some neighborhood of z_0, $S\left(z_0\right)$ consists of the union of m smooth curves $S_k\left(z_0\right)$, for $k\in\{1,\dots,m\}$, with the following properties:*

i) The curve $S_k\left(z_0\right)$ passes through z_0, with tangent $r_k\left(z_0\right)$.

ii) $\displaystyle\lim_{z\to z_0,z\in S_k(z_0)}\sigma\left(z,z_0\right)=\lambda_k\left(z_0\right)$.

iii) $\sigma\left(z,z_0\right)=\frac{1}{2}\left(\lambda_k\left(z\right)+\lambda_k\left(z_0\right)\right)+\mathcal{O}\left(\left|z-z_0\right|^2\right)$*, as $z\to z_0$ with $z\in S_k\left(z_0\right)$.*

The last statement of Theorem A.4 demonstrates that the curves $R_k\left(z_0\right)$ and $S_k\left(z_0\right)$ agree at least to first order at z_0.

The following theorem asserts that in the linearly degenerate case these curves coincide.

Theorem A.5 (Linear degeneracy)**.** *Suppose for some $k\in\{1,\dots,m\}$ that the pair $\left(\lambda_k,r_k\right)$ is linearly degenerate. Then for each $z_0\in\mathbb{R}^m$,*

i) $R_k\left(z_0\right)=S_k\left(z_0\right)$, and

ii) $\sigma\left(z,z_0\right)=\lambda_k\left(z\right)=\lambda_k\left(z_0\right)$, for all $z\in S_k\left(z_0\right)$.

2.3.2 Contact discontinuities

Suppose that $\left(\lambda_k,r_k\right)$ is linearly degenerate and $U_R\in S_k\left(U_L\right)$, an integral solution of (A.9) is then defined by

$$U\left(t,x\right):=\begin{cases} U_L, \text{ if } x<\sigma t, \\ U_R, \text{ if } x>\sigma t, \end{cases} \tag{A.15}$$

where

$$\sigma = \sigma\left(U_R, U_L\right) = \lambda_k\left(U_L\right) = \lambda_k\left(U_R\right).$$

Since $\lambda_k\left(U_L\right) = \lambda_k\left(U_R\right) = \sigma$, the projected characteristics to the left and right are parallel to the line of discontinuity. This situation is physically interpreted by saying that the fluid particles do not cross the discontinuity. And the line $x = \sigma t$ is called a *k-contact discontinuity*.

2.3.3 Shock waves

Consider the case that (λ_k, r_k) is genuinely nonlinear and $U_R \in S_k\left(U_L\right)$ and the following integral solution

$$U\left(t,x\right) := \begin{cases} U_L, & \text{if } x < \sigma\left(U_R, U_L\right)t, \\ U_R, & \text{if } x > \sigma\left(U_R, U_L\right)t. \end{cases} \tag{A.16}$$

Definition A.7 (Admissible shock waves). *Assume the pair (λ_k, r_k) is genuinely nonlinear at U_l. The pair (U_L, U_R) is said to be* admissible *provided $U_R \in S_k\left(U_L\right)$ and*

$$\begin{aligned} U_R &\in S_k\left(U_L\right), \\ \lambda_k\left(U_R\right) < \sigma&\left(U_R, U_L\right) < \lambda_k\left(U_L\right). \end{aligned} \tag{A.17}$$

*If (U_L, U_R) is admissible, the solution U is defined by (A.16) a k-*shock wave.

The condition (A.17) is referred as the *Lax entropy condition.*

Definition A.8. *If the pair (λ_k, r_k) is genuinely nonlinear, let us define*

$$\begin{aligned} S_k^+\left(z_0\right) &:= \left\{z \in S_k\left(z_0\right); \lambda_k\left(z_0\right) < \sigma\left(z, z_0\right) < \lambda_k\left(z\right)\right\}, \\ S_k^-\left(z_0\right) &:= \left\{z \in S_k\left(z_0\right); \lambda_k\left(z\right) < \sigma\left(z, z_0\right) < \lambda_k\left(z_0\right)\right\}. \end{aligned}$$

Then

$$S_k\left(z_0\right) = S_k^-\left(z_0\right) \cup \left\{z_0\right\} \cup S_k^+\left(z_0\right) \text{ near } z_0.$$

The pair (U_L, U_R) is admissible if and only if $U_R \in S_k^-\left(U_L\right)$.

2.4 Local solution of Riemann's problem

The physically relevant parts of the rarefaction and shock waves will be glued together.

Definition A.9. *i) If the pair (λ_k, r_k) is genuinely nonlinear, write*

$$T_k\left(z_0\right) := R_k^-\left(z_0\right) \cup \left\{z_0\right\} \cup R_k^+\left(z_0\right).$$

ii) If the pair (λ_k, r_k) is linearly degenerate, set

$$T_k\left(z_0\right) := R_k\left(z_0\right) = S_k\left(z_0\right).$$

The curve $T_k(z_0)$ is C^1 by Theorem A.4. Nearby states U_L and U_R can be joined by a k-rarefaction wave, a shock wave or a contact discontinuity provided $U_L \in T_k(U_R)$.

Theorem A.6 (Local solution of Riemann's problem). *Assume for each $k \in \{1, \ldots, m\}$ that the pair (λ_k, r_k) is either genuinely nonlinear or else linearly degenerate. Suppose further the left state U_L is given. Then for each right state U_R sufficiently close to U_L there exists an integral solution U of Riemann's problem, which is constant on lines through the origin.*

3 Systems of two conservation laws

For $m = 2$, consider a pair of conservation laws:

$$\begin{cases} \partial_t U_1 + \partial_x (F_1(U_1, U_2)) = 0, \text{ in } \mathbb{R}_+^\star \times \mathbb{R}, \\ \partial_t U_2 + \partial_x (F_2(U_1, U_2)) = 0, \text{ in } \mathbb{R}_+^\star \times \mathbb{R}, \\ U_1(0, x) = U_{0,1}(x), \quad U_2(0, x) = U_{0,2}(x), \quad \forall x \in \mathbb{R}. \end{cases} \tag{A.18}$$

Here $F = (F_1, F_2)$, $U_0 = (U_{0,1}, U_{0,2})$, and $U = (U_1, U_2)$.

3.1 Riemann invariants

Definition A.10 (Riemann invariants). *The function $W_i : \mathbb{R}^2 \to \mathbb{R}$ is an i^{th}*-Riemann invariant *provided*

$$DW_i(z) \;\textit{ is parallel to } l_j(z), \quad \forall z \in \mathbb{R}^2, \quad i \neq j. \tag{A.19}$$

For $m = 2$, (A.19) is equivalent in $\mathbb{R}^2$ to the statement

$$DW_i(z) \cdot r_i(z) = 0, \quad i = 1, 2, \quad \forall z \in \mathbb{R}^2, \tag{A.20}$$

i.e.,

$$W_i \text{ is constant along the rarefaction curve } R_i, \quad i = 1, 2. \tag{A.21}$$

In particular, any smooth function W_i satisfying (A.21) satisfies also (A.20), (A.19) and so is an i^{th}-Riemann invariant.

It should be emphasized that for $m > 2$, Riemann invariants do not exist in general.

Define $W : \mathbb{R}^2 \to \mathbb{R}^2$ by setting

$$W(z) = W(z_1, z_2) = (W_1(z_1, z_2), W_2(z_1, z_2)).$$

The inverse mapping is

$$Z(w) = Z(w_1, w_2) = (Z_1(w_1, w_2), Z_2(w_1, w_2)).$$

Now set

$$V\left(t,x\right):=W\left(U\left(t,x\right)\right),\ \text{in } \mathbb{R}_{+}^{\star}\times\mathbb{R}.$$

The following theorem indicate which system of PDE $V=(V_1,V_2)$ solves.

Theorem A.7 (Conservation laws and Riemann invariants)**.** *The functions V_1, V_2 solve the system*

$$\begin{cases}\partial_t V_1+\lambda_2\left(U\right)\partial_x V_1=0,\\ \partial_t V_2+\lambda_2\left(U\right)\partial_x V_2=0,\end{cases}\quad \textit{in } \mathbb{R}_{+}^{\star}\times\mathbb{R}. \tag{A.22}$$

It is emphasized in Evans, 2010, p. 636 that although the system (A.22) is not in conservation law form, it is in many ways rather simpler than (A.18).

Remark A.3. *i)* Methods of characteristics: *The system of PDE* (A.22) *can be interpreted by introducing the ODE*

$$\dot{x}_i\left(s\right)=\lambda_j\left(s,U\left(x_i\left(s\right)\right)\right),\ \ \forall s\in\mathbb{R}_{+}^{\star},\ \ i=1,2,\ \ j\neq i. \tag{A.23}$$

Then it is deduced from (A.22) *that*

$$V_i \textit{ is constant along the characteristic curve } \left(s,x_i\left(s\right)\right),\ \ \forall s\in\mathbb{R}_{+}^{\star},\ \ i=1,2.$$

ii) The genuine nonlinearity condition reads

$$D\lambda_i\left(z\right)\cdot r_i\left(z\right)\neq 0,\ \ \forall z\in\mathbb{R}^2,\ \ i=1,2.$$

which is equivalent to

$$\frac{\partial\lambda_i}{\partial w_j}\neq 0,\ \ \forall w\in\mathbb{R}^2,\ \ i\neq j.$$

Example A.2 (Barotropic compressible gas dynamics)**.** *Consider the Euler's equations for compressible gas dynamics in the special case that the internal energy e is constant. The relevant PDE are*

$$\begin{cases}\partial_t\rho+\partial_x\left(\rho u\right)=0,\\ \partial_t\left(\rho u\right)+\partial_x\left(\rho u^2+p\right)=0,\end{cases} \tag{A.24}$$

where we assume

$$p=p\left(\rho\right) \tag{A.25}$$

for some smooth function $p:\mathbb{R}\to\mathbb{R}$. Formula (A.25) *is called a* barotropic equation of state. *Assume the strict hyperbolicity condition $p'>0$.*

Setting $U=(U_1,U_2)=(\rho,\rho u)$, (A.24)-(A.25) *can be rewritten to read*

$$\partial_t U+\partial_x\left(F\left(U\right)\right)=0,$$

for

$$F=(F_1,F_2)=\left(z_2,\frac{z_2^2}{z_1}+p\left(z_1\right)\right),$$

and $z=(z_1,z_2)$, *provided* $z_1>0$. *Then*

$$J_F=\left(\begin{array}{cc} 0 & 1 \\ -\left(\frac{z_2}{z_1}\right)^2+p'\left(z_1\right) & \frac{2z_2}{z_1}\end{array}\right).$$

Its eigenvalues are given by

$$\lambda_1\left(z_1,z_2\right):=\frac{z_2}{z_1}-\sqrt{p'\left(z_1\right)},\ \ \lambda_2\left(z_1,z_2\right):=\frac{z_2}{z_1}+\sqrt{p'\left(z_1\right)},$$

and eigenvectors are chosen as

$$r_1\left(z_1,z_2\right):=\left(\begin{array}{c} 1 \\ \frac{z_2}{z_1}-\sqrt{p'\left(z_1\right)}\end{array}\right),\ \ r_2\left(z_1,z_2\right):=\left(\begin{array}{c} 1 \\ \frac{z_2}{z_1}+\sqrt{p'\left(z_1\right)}\end{array}\right).$$

In physical notation,

$$\lambda_1=u-\sigma,\ \ \lambda_2=u+\sigma,$$
$$r_1=\left(\begin{array}{c} 1 \\ u-\sigma\end{array}\right),\ r_2=\left(\begin{array}{c} 1 \\ u+\sigma\end{array}\right),$$

where $\sigma:=\sqrt{p'\left(\rho\right)}$ *is the* sound speed.

The ODE (A.23) *reads*

$$\dot{x}_1\left(t\right)=V\left(t,x_1\left(t\right)\right)+\sigma\left(t,x_1\left(t\right)\right), \tag{A.26}$$
$$\dot{x}_2\left(t\right)=V\left(t,x_2\left(t\right)\right)-\sigma\left(t,x_2\left(t\right)\right), \tag{A.27}$$

where $\sigma\left(t,x\right):=\sqrt{p'\left(\rho\left(t,x\right)\right)}$, $t\geq 0$. *And then the Riemann invariant* $V_1=W_1\left(U\right)$ *is constant along the trajectories of* (A.26) *and* $V_2=W_2\left(U\right)$ *is constant along trajectories of* (A.27).

To compute W_1 *and* W_2 *directly, we first transform* (A.24) *in nondivergence form:*

$$\partial_t\rho+\rho\partial_x u+u\partial_x\rho=0, \tag{A.28}$$
$$u\partial_t\rho+\rho\partial_t u+u^2\partial_x\rho+2\rho u\partial_x u+\partial_x p=0. \tag{A.29}$$

Multiplying (A.28) *by* $\sigma^2=p'\left(\rho\right)$ *and recalling* (A.25) *gives us*

$$\partial_t p+u\partial_x p+\sigma^2\rho\partial_x u=0. \tag{A.30}$$

Combining (A.28), (A.29) *yields*

$$\rho\partial_t u+\rho u\partial_x u+\partial_x p=0. \tag{A.31}$$

Multiplying (A.31) *by* σ *and then adding to and subtracting from* (A.30) *yields*

$$\begin{cases}\partial_t p + (u+\sigma)\,\partial_x p + \rho\sigma\left[\partial_t u + (u+\sigma)\,\partial_x u\right] = 0,\\ \partial_t p + (u-\sigma)\,\partial_x p - \rho\sigma\left[\partial_t u + (u-\sigma)\,\partial_x u\right] = 0.\end{cases} \tag{A.32}$$

It is deduced from (A.32) *that*

$$\begin{cases}\frac{d}{dt}\left[p\left(t,x_1\left(t\right)\right)\right] + \rho\left(t,x_1\left(t\right)\right)\sigma\left(t,x_1\left(t\right)\right)\frac{d}{dt}\left[u\left(t,x_1\left(t\right)\right)\right] = 0,\\ \frac{d}{dt}\left[p\left(t,x_2\left(t\right)\right)\right] - \rho\left(t,x_2\left(t\right)\right)\sigma\left(t,x_2\left(t\right)\right)\frac{d}{dt}\left[u\left(t,x_2\left(t\right)\right)\right] = 0.\end{cases}$$

As $\frac{dp}{dt} = \sigma^2\frac{d\rho}{dt}$, *provided* $\rho > 0$,

$$\frac{\sigma}{\rho}\frac{d\rho}{dt} \pm \frac{du}{dt} = 0 \text{ along the trajectories of (A.26) and (A.27).} \tag{A.33}$$

Think of the Riemann invariants as functions of ρ *and* u. *Then since* $V_1 = W_1\left(\rho,u\right)$ *is constant along the curve determined by* $x_1\left(\cdot\right)$, *we have*

$$\begin{aligned}0 &= \frac{d}{dt}\left[W_1\left(\rho\left(t,x_1\left(t\right)\right),V\left(t,x_1\left(t\right)\right)\right)\right]\\ &= \frac{\partial W_1}{\partial \rho}\frac{d}{dt}\left[\rho\left(t,x_1\left(t\right)\right)\right] + \frac{\partial W_1}{\partial V}\frac{d}{dt}\left[V\left(t,x_1\left(t\right)\right)\right].\end{aligned}$$

This is consistent with (A.33) *if*

$$\frac{\partial W_1}{\partial \rho} = \frac{\sigma\left(\rho\right)}{\rho},\quad \frac{\partial W_1}{\partial u} = 1.$$

Similarly,

$$\frac{\partial W_2}{\partial \rho} = \frac{\sigma\left(\rho\right)}{\rho},\quad \frac{\partial W_2}{\partial u} = -1.$$

Integrating these implies that the Riemann invariants are, up to additive constants,

$$W_1 = \int_1^\rho \frac{\sigma\left(s\right)}{s}ds + u,\quad W_2 = \int_1^\rho \frac{\sigma\left(s\right)}{s}ds - u.$$

Finally, we claim that W_1, W_2, *taken now as functions of* $z = \left(z_1,z_2\right)$, *satisfy the definition of Riemann invariants. Indeed, in the variables* z, W_1 *and* W_2 *read, for all* $\left(z_1,z_2\right) \in \mathbb{R}^2$,

$$W_1\left(z_1,z_2\right) = \int_1^{z_1} \frac{\sigma\left(s\right)}{s}ds + \frac{z_2}{z_1},\quad W_2\left(z_1,z_2\right) = \int_1^{z_1} \frac{\sigma\left(s\right)}{s}ds - \frac{z_2}{z_1},$$

and

$$DW_1\left(z_1,z_2\right) = \begin{pmatrix}\frac{\sqrt{p'(z_1)}}{z_1} - \frac{z_2}{z_1^2}\\ \frac{1}{z_1}\end{pmatrix},\quad DW_2\left(z_1,z_2\right) = \begin{pmatrix}\frac{\sqrt{p'(z_1)}}{z_1} + \frac{z_2}{z_1^2}\\ -\frac{1}{z_1}\end{pmatrix}.$$

Thus, for all $(z_1, z_2) \in \mathbb{R}^2$,

$$(DW_1 \cdot r_1)(z_1, z_2) = \begin{pmatrix} \frac{\sqrt{p'(z_1)}}{z_1} - \frac{z_2}{z_1^2} \\ \frac{1}{z_1} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \frac{z_2}{z_1} - \sqrt{p'(z_1)} \end{pmatrix} = 0,$$

$$(DW_2 \cdot r_2)(z_1, z_2) = \begin{pmatrix} \frac{\sqrt{p'(z_1)}}{z_1} + \frac{z_2}{z_1^2} \\ -\frac{1}{z_1} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ \frac{z_2}{z_1} + \sqrt{p'(z_1)} \end{pmatrix} = 0.$$

i.e., (A.20) *holds and thus* W_1, W_2 *are Riemann invariants.*

3.2 Nonexistence of smooth solutions

The following theorem illustrates the usefulness of Riemann invariants in establishing a criterion for the nonexistence of a smooth solution.

Theorem A.8 (Riemann invariants and blow-up)**.** *Assume g is smooth, with compact support. Suppose also the genuine nonlinearity condition*

$$\frac{\partial \lambda_i}{\partial w_j} > 0, \ \ \textit{in } \mathbb{R}^2, \ \ i = 1, 2, \ \ i \neq j,$$

holds. Then the initial-value problem (A.18) *cannot have a smooth solution U existing for all times $t \geq 0$ if either $\partial_x V_1 < 0$ or $\partial_x V_2 < 0$ somewhere on $\{t = 0\} \times \mathbb{R}$.*

4 Entropy criteria

Recall that Lax's entropy condition

$$\lambda_k(U_R) < \sigma(U_R, U_L) < \lambda_k(U_L), \ \text{for some } k \in \{1, \dots, m\}$$

as the selection criteria for admissible shock waves.

One general principle is that physically and mathematically correct solutions should arise as the limit of solutions to the regularized system

$$\partial_t U^\varepsilon + \partial_x(F(U^\varepsilon)) - \varepsilon \partial_x^2 U^\varepsilon = 0, \ \text{in } \mathbb{R}_+^\star \times \mathbb{R}, \tag{A.34}$$

as $\varepsilon \to 0$.

4.1 Vanishing viscosity, traveling waves

Theorem A.9 (Existence of traveling waves for genuinely nonlinear systems)**.** *Assume the pair (λ_k, r_k) is genuinely nonlinear for $k = 1, \dots, m$. Let U_R be selected sufficiently close to U_L. Then there exists a traveling wave solution*

$$U^\varepsilon(t, x) = V\left(\frac{x - \sigma t}{\varepsilon}\right), \ \textit{in } \mathbb{R}_+^\star \times \mathbb{R},$$

of (A.34) *connecting* U_L *to* U_R *if and only if*

$$U_R \in S_k^- (U_L) \text{ for some } k \in \{1, \ldots, m\},$$

Suppose $U_R \in S_k (U_L)$ for some $\{1, \ldots, m\}$ and the following *Liu's entropy criterion* holds:

$$\begin{cases} \sigma (z, U_L) > \sigma (U_R, U_L) \text{ for each } z \text{ lying} \\ \text{ on the curve } S_k (U_L) \text{ between } U_R \text{ and } U_L. \end{cases} \tag{A.35}$$

Remark A.4. *Liu's entropy criterion is automatic provided* (λ_k, r_k) *is genuinely nonlinear,* $U_R \in S_k^- (U_L)$*, and* U_R *is sufficient close to* U_L*.*

Theorem A.10 (Existence of traveling waves)**.** *Let* U_R *be selected sufficiently close to* U_L*. Then there exists a traveling wave solution*

$$U^\varepsilon (t, x) = V \left(\frac{x - \sigma t}{\varepsilon} \right), \text{ in } \mathbb{R}_+^\star \times \mathbb{R},$$

of (A.34) *connecting* U_L *to* U_R*, where* V *solving*

$$\ddot{V} = -\sigma \dot{V} + J_F (V) \dot{V},$$
$$\lim_{s \to -\infty} V = U_L, \quad \lim_{s \to +\infty} V = U_R, \quad \lim_{s \to \pm\infty} \dot{V} = 0,$$

if and only if the Liu's entropy condition (A.35) *is satisfied.*

Both Lax's and Liu's entropy criteria provide restrictions on possible left- and right-hand states joined by a shock wave (or a traveling wave for viscous approximation).

4.2 Entropy/entropy-flux pairs

Definition A.11 (Entropy/entropy-flux)**.** *Two smooth funcitons* $\Phi, \Psi : \mathbb{R}^m \to \mathbb{R}$ *comprise an* entropy/entropy-flux pair *for the conservation law*

$$\partial_t U + \partial_x (F (U)) = 0, \tag{A.36}$$

provided Φ *is convex and*

$$\nabla \Phi (z) J_F (z) = D \Psi (z), \quad \forall z \in \mathbb{R}^m. \tag{A.37}$$

Proposition A.3 (Scalar conservation law for entropy/entropy-flux pair)**.** *Suppose that* U *is a smooth solution of* (A.36) *then* $\Phi (U)$ *satisfies a scalar conservation law, with flux* $\Psi (U)$*, i.e.,*

$$\partial_t (\Phi (U)) + \partial_x (\Psi (U)) = 0, \text{ in } \mathbb{R}_+^\star \times \mathbb{R}. \tag{A.38}$$

Proof. It is straightforward by the calculation

$$\begin{aligned} \partial_t (\Phi (U)) + \partial_x (\Psi (U)) &= \nabla \Phi (U) \cdot \partial_t U + D \Psi (U) \cdot \partial_x U \\ &= (-\nabla \Phi (U) J_F (U) + D \Psi (U)) \cdot \partial_x U = 0, \end{aligned}$$

by (A.28). □

Nevertheless, integral solutions of (A.36) will not be smooth enough, owning to shocks and other irregularities. To overcome this obstacle, (A.38) is replaced with the following *entropy inequality*:

$$\partial_t \left(\Phi\left(U\right)\right) + \partial_x \left(\Psi\left(U\right)\right) \leq 0, \text{ a.e. } (t,x) \in \mathbb{R}_+ \times \mathbb{R}, \tag{A.39}$$

which can be understood as

$$\int_0^\infty \int_{\mathbb{R}} \left(\Phi\left(U\right)\partial_t \varphi + \Psi\left(U\right)\partial_x \varphi\right) dxdt \geq 0, \ \ \forall \varphi \in C_c^\infty\left(\mathbb{R}_+^\star \times \mathbb{R}, \mathbb{R}_+\right), \tag{A.40}$$

or

$$\int_0^\infty \int_{\mathbb{R}} \left(\Phi\left(U\right)\partial_t \varphi + \Psi\left(U\right)\partial_x \varphi\right) dxdt + \int_{\mathbb{R}} \Phi\left(U_0\right)\varphi|_{t=0} \geq 0, \ \ \forall \varphi \in C_c^\infty\left(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}_+\right). \tag{A.41}$$

Condition (A.40) is equivalent to requiring that for all rectangles $(t_1, t_2) \times (a, b)$ the inequality obtained by integrating (A.39) over the rectangle should hold:

$$\int_a^b \Phi\left(U\left(t_2, x\right)\right) dx - \int_a^b \Phi\left(U\left(t_1, x\right)\right) dx + \int_{t_1}^{t_2} \Psi\left(U\left(t, b\right)\right) dt - \int_{t_1}^{t_2} \Psi\left(U\left(t, a\right)\right) dt \leq 0.$$

If U is piecewise smooth with discontinuities, then (A.38) holds pointwise in the smooth regions, while across a discontinuity

$$\Psi\left(U_R\right) - \Psi\left(U_L\right) - \sigma\left(U_R, U_L\right)\left(\Phi\left(U_R\right) - \Phi\left(U_L\right)\right) \leq 0. \tag{A.42}$$

The last four inequalities are called *entropy conditions*.

Definition A.12 (Entropy solution)**.** *The function U is called an* entropy solution *of* (A.1) *provided U is an integral solution and U satisfies the entropy inequality* (A.39)*, or* (A.41)*, for each entropy/entropy-flux pair* (Φ, Ψ)*.*

Consider the following approximating viscous problems

$$\begin{cases} \partial_t U^\varepsilon + \partial_x \left(F\left(U^\varepsilon\right)\right) - \varepsilon \partial_x^2 U^\varepsilon = 0, & \text{in } \mathbb{R}_+ \times \mathbb{R}, \\ U^\varepsilon\left(0, x\right) = U_0\left(x\right), & \text{in } \mathbb{R}. \end{cases} \tag{A.43}$$

Assume that U^ε is a smooth solution of (A.43), converging to 0 as $|x| \to \infty$ sufficiently rapidly, we also suppose that $\{U^\varepsilon\}_{0<\varepsilon\leq 1}$ is uniformly bounded in L^∞ and furthermore[4]

$$U^\varepsilon \to U \text{ a.e. as } \varepsilon \to 0,$$

for some limit function U.

Theorem A.11 (Entropy and vanishing viscosity)**.** *The function U is an entropy solution of the conservation law* (A.1)*.*

[4] The verification of this a.e. convergence is extremely difficult in practice.

Example A.3 (Entropy flux for scalar conservation law)**.** *In the case of a scalar conservation law (i.e., $m = 1$), for any convex Φ, a corresponding flux function Ψ is given by*

$$\Psi(z) := \int_{z_0}^{z} \Phi'(w)\, F'(w)\, dw, \ \ \forall z \in \mathbb{R}.$$

4.3 Uniqueness for scalar conservation laws

Consider the initial-value problem for a scalar conservation law

$$\begin{cases} \partial_t u + \partial_x (f(u)) = 0, & \text{in } \mathbb{R}_+ \times \mathbb{R}, \\ u(0,x) = u_0(x), & \text{in } \mathbb{R}. \end{cases} \tag{A.44}$$

The unknown $u = u(t,x)$ is real-valued and $f : \mathbb{R} \to \mathbb{R}$ is a given smooth flux function.

Definition A.13 (Entropy/entropy-flux pair)**.** *Two smooth functions Φ, $\Psi : \mathbb{R} \to \mathbb{R}$ comprise an* entropy/entropy-flux pair *for the conservation law*

$$\partial_t u + \partial_x (f(u)) = 0,$$

provided Φ is convex and

$$\Phi'(z)\, f'(z) = \Psi'(z), \ \ \forall z \in \mathbb{R}. \tag{A.45}$$

Example A.3 indicates that for each convex Φ there exists a corresponding flux Ψ.

The entropy condition for u reads

$$\partial_t (\Phi(u)) + \partial_x (\Psi(u)) \le 0, \ \text{in } \mathbb{R}_+ \times \mathbb{R},$$

for each entropy/entropy-flux pair Φ, Ψ, i.e.,

$$\int_0^\infty \int_{\mathbb{R}} (\Phi(u)\, \partial_t \varphi + \Psi(u)\, \partial_x \varphi)\, dxdt \ge 0, \ \ \forall \varphi \in C_c^\infty(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}_+). \tag{A.46}$$

Definition A.14 (Entropy solution)**.** *A function $u \in C\left(\mathbb{R}_+^\star, L^1(\mathbb{R})\right) \cap L^\infty(\mathbb{R}_+ \times \mathbb{R})$ is called an* entropy solution *of* (A.44) *provided u satisfies the inequalities* (A.46) *for each entropy/entropy-flux pair (Φ, Ψ) and $u(t,\cdot) \to u_0$ in $L^1(\mathbb{R})$ as $t \downarrow 0$.*

Proposition A.4. *An entropy solution of* (A.44) *is an integral solution of* (A.44)*.*

Proof. Taking $\Phi(z) = \pm z$, $\Psi(z) = \pm f(z)$ in (A.46) gives us

$$\int_0^\infty \int_{\mathbb{R}} (u \partial_t v + f(u)\, \partial_x v)\, dxdt = 0, \ \ \forall v \in C_c^\infty(\mathbb{R}_+ \times \mathbb{R}), \ \ v \ge 0,$$

and thus

$$\int_0^\infty \int_{\mathbb{R}} (u \partial_t v + f(u)\, \partial_x v)\, dxdt = 0, \ \ \forall v \in C_c^1(\mathbb{R}_+ \times \mathbb{R}).$$

It is classic[5] to prove then that

$$\int_0^\infty \int_{\mathbb{R}} \left(u\partial_t v + f(u)\,\partial_x v\right) dxdt + \int_{\mathbb{R}} u_0 v|_{t=0} dx = 0, \quad \forall v \in C_c^1\left(\mathbb{R}_+ \times \mathbb{R}\right),$$

since $u(t,\cdot) \to u_0$ in $L^1(\mathbb{R})$. The proof is accomplished. □

Theorem A.12 (Uniqueness of entropy solutions for a single conservation law)**.** *Let $f \in C^1(\mathbb{R},\mathbb{R})$, $u_0 \in L^\infty(\mathbb{R})$, then there exists - up to a set of measure zero - at most one entropy solution of* (A.44)*.*

The basic idea of the proof for this theorem is "doubling variables" techinique, see Evans, 2010, pp. 650–653 for more details.

We also introduce another way to "install" the notion of entropy into nonlinear scalar conservation laws in the following subsections.

4.4 Lyapunov functional for scalar conservation laws

Consider the *Lyapunov functional* $\int_{\mathbb{R}} \Phi(u)\,dx$, where Φ is a convex smooth function $\mathbb{R} \to \mathbb{R}$. We have the following equivalent definition of entropy weak solution.

Definition A.15 (Entropy weak solution)**.** *Let $f \in C^1(\mathbb{R},\mathbb{R})$ and $u_0 \in L^\infty(\mathbb{R})$, a function $u \in L^\infty\left(\mathbb{R}_+^\star \times \mathbb{R}\right)$ is an* entropy (weak) solution *(EWS) of the Cauchy problem* (A.44) *if*

$$\int_0^\infty \int_{\mathbb{R}} \left(\Phi(u)\,\partial_t\varphi + \Psi(u)\,\partial_x\varphi\right) dxdt + \int_{\mathbb{R}} \Phi(u_0)\,\varphi(0,x)\,dx \geq 0, \tag{A.47}$$

for all function $\Phi \in C^2(\mathbb{R},\mathbb{R})$, $\Phi'' > 0$ and $\Psi \in C^1(\mathbb{R},\mathbb{R})$ such that $\Psi' = \Phi' f'$, and all test functions $\varphi \in C_c^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}_+)$.

In Lyapunov theory of stability, (A.47) is equivalent to

$$\partial_t(\Phi(u)) + \partial_x(\Psi(u)) \leq 0, \text{ in } \mathbb{R}_+ \times \mathbb{R},$$

and to

$$\frac{d}{dt}\int_{\mathbb{R}} \Phi(u(t,x))\,dx \leq 0.$$

As Proposition A.4, an EWS in the sense of Definition A.15 is also an integral solution.

Lemma A.2 (Integral representation of entropy)**.** *For all $\Phi \in C^2(\mathbb{R})$, $\Phi'' \geq 0$, $a \in \mathbb{R}$, $b \in \mathbb{R}$, $a < b$, there exists a constant $C_\Phi \in \mathbb{R}$ such that*

$$\Phi(x) = C_\Phi + \frac{1}{2}\int_a^b \Phi''(\kappa)\,|x-\kappa|\,d\kappa, \quad \forall x \in (a,b),$$

where $C = C(a,b,\Phi)$.

[5] From the lecture *Hyperbolic equations* taught by Prof. Florian Méhats.

To prove this lemma, it should be noted that

$$\frac{\partial}{\partial \kappa^2} |x-\kappa| = 2\delta_x(\kappa), \quad \forall x \in \mathbb{R}, \quad \forall \kappa \in \mathbb{R}.$$

Krushkov used a characterization of entropy solutions which is given in the following proposition.

Proposition A.5 (Entropy weak solution using "Krushkov's entropies"). *Let $f \in C^1(\mathbb{R},\mathbb{R})$ and $u_0 \in L^\infty(\mathbb{R})$, u is the unique entropy weak solution to (A.2) if and only if $u \in L^\infty(\mathbb{R}_+^\star \times \mathbb{R})$ is such that*

$$\int_0^\infty \int_{\mathbb{R}} \left(|u-\kappa|\,\partial_t\varphi + \Psi(u,\kappa)\,\partial_x\varphi\right) dxdt + \int_{\mathbb{R}} |u_0-\kappa|\,\varphi(0,x)\,dx \geq 0,$$

for all $\kappa \in \mathbb{R}$, $\varphi \in C_c^1(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}_+)$, where

$$\Psi(x,\kappa) := \operatorname{sgn}(x-\kappa)\left(f(x)-f(\kappa)\right) = f(x\top\kappa) - f(x\perp\kappa),$$

with $a\top b = \max(a,b)$, and $a\perp b = \min(a,b)$.

The result of existence of an entropy weak solution was proved by passing to the limit on the solutions of an appropriate numerical scheme, and can also be obtained by passing to the limit on finite volume approximations of the solution.

Theorem A.13 (Comparison of entropy solutions). *Let $u_0, v_0 \in L^\infty(\mathbb{R})$, and $m, M \in \mathbb{R}$ such that $m \leq u_0, v_0 \leq M$ a.e. Let u and v belong $L^\infty(\mathbb{R}_+^\star \times \mathbb{R})$ some entropy solutions associated with u_0 and v_0, respectively. Then, for all $R > 0$, $T > 0$, the following inequality holds:*

$$\int_{-R}^{R} |u(T,x) - v(T,x)|\,dx \leq \int_{-R-T\mathrm{Lip}(f;[m,M])}^{R+T\mathrm{Lip}(f;[m,M])} |u_0(x) - v_0(x)|\,dx.$$

Corollary A.1. *i) Uniqueness of entropy solutions, Theorem A.12 holds.*

ii) If $u_0 - v_0 \in L^1(\mathbb{R})$, then

$$\|(u-v)(T,\cdot)\|_{L^1(\mathbb{R})} \leq \|u_0 - v_0\|_{L^1(\mathbb{R})}.$$

iii) Any smooth solution is an entropy solution.

iv) If $m \leq u_0 \leq M$ a.e. then $m \leq u \leq M$ a.e.

v) The absolute value $|\cdot|$ can be replaced by $(\cdot)_+$, i.e.,

$$\int_{-R}^{R} (u(T,x) - v(T,x))_+ dx \leq \int_{-R-T\mathrm{Lip}(f)}^{R+T\mathrm{Lip}(f)} (u_0(x) - v_0(x))_+ dx.$$

v) (Monotonicity) *If $u_0 \leq v_0$ a.e., then for all $T > 0$, $u(T,\cdot) \leq v(T,\cdot)$ a.e.*

Definition A.16 (Total variation, $BV(\Omega)$). *Let $p \in \mathbb{N}^\star$ and let Ω be an open subset of $\mathbb{R}^p$. A function $f \in L^1_{\mathrm{loc}}(\Omega)$ has a bounded variation, that is $f \in BV(\Omega)$, if*

$|f|_{BV(\Omega)} < \infty$ *where the* total variation *of f is defined by*

$$|f|_{BV(\Omega)} := \sup\left\{\int_{\Omega} f \operatorname{div}\varphi dx; \varphi \in C_c^1\left(\Omega, \mathbb{R}^p\right), \|\varphi\|_{L^\infty(\Omega)} \leq 1\right\}.$$

In 1D, the total variation reads

$$|f|_{BV(\mathbb{R})} = \sup\left\{\sum_{i=1}^{N} |f\left(x_{i+1}\right) - f\left(x_i\right)|; x_0 < x_1 < \cdots < x_{N+1},\ N \in \mathbb{N}\right\}.$$

Remark A.5. $|\cdot|_{BV(\Omega)}$ *is a semi-norm, i.e.,* $|f|_{BV(\Omega)} = 0 \Leftrightarrow f = const$ *a.e.*

The interpretation of this notion is a way to control the oscillations, without smoothness.

Remark A.6. *1. If $v : \mathbb{R} \to \mathbb{R}$ is piecewise constant, i.e., there exists an increasing sequence $(x_i)_{i\in\mathbb{Z}}$ with $\mathbb{R} = \bigcup_{i\in\mathbb{Z}} [x_i, x_{i+1}]$ and a sequence $(v_i)_{i\in\mathbb{Z}}$ such that $v|_{(x_i,x_{i+1})} = v_i$, then $|v|_{BV(\mathbb{R})} = \sum_{i\in\mathbb{Z}} |v_{i+1} - v_i|$.*

2. If $v \in C^1(\mathbb{R}, \mathbb{R})$ then $|v|_{BV(\mathbb{R})} = \|v_x\|_{L^1(\mathbb{R})}$.

3. $BV(\mathbb{R}) \subset L^\infty(\mathbb{R})$; furthermore, if $u \in BV(\mathbb{R}) \cap L^1(\mathbb{R})$ then $\|u\|_{L^\infty(\mathbb{R})} \leq |u|_{BV(\mathbb{R})}$.

4. Let $u \in BV(\mathbb{R})$ and let $\left(x_{i+\frac{1}{2}}\right)_{i\in\mathbb{Z}}$ be an increasing sequence of real values such that $\mathbb{R} = \bigcup_{i\in\mathbb{Z}} \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right]$. For $i \in \mathbb{Z}$, let $C_i := \left(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right)$ and u_i be the mean value of u over C_i (or even a subinterval of C_i). Then

$$\sum_{i\in\mathbb{Z}} |u_{i+1} - u_i| \leq |u|_{BV(\mathbb{R})}.$$

Theorem A.14 (Helly's theorem)**.** *If Ω is an open bounded subset of $\mathbb{R}^n$ and if $(f_n)_{n\in\mathbb{N}} \subset L^1_{\text{loc}}(\Omega)$ is a family, such that there exist two positive constants $C_1, C_2 > 0$ satisfying*

$$\|f_n\|_{L^\infty(\Omega)} \leq C_1,\ \ |f_n|_{BV(\Omega)} \leq C_2,\ \ \forall n \in \mathbb{N}.$$

Then there exists a subsequence of $(f_{n_k})_{k\in\mathbb{N}}$ which converges to $f \in L^1_{\text{loc}}(\Omega)$. Moreover,

$$\|f\|_{L^\infty(\Omega)} \leq C_1,\ \ |f|_{BV(\Omega)} \leq C_2.$$

The entropy weak solution to (A.2) satisfies the following L^∞ and BV stability properties:

Proposition A.6 (L^∞ and BV stability properties)**.** *Let $f \in C^1(\mathbb{R}, \mathbb{R})$ and $u_0 \in L^\infty(\mathbb{R})$. Let u be the entropy weak solution to (A.2). Then, $u \in C\left(\mathbb{R}_+, L^1_{\text{loc}}(\mathbb{R})\right)$. Furthermore, the following estimates hold:*

i) $\|u(t,\cdot)\|_{L^\infty(\mathbb{R})} \leq \|u_0\|_{L^\infty(\mathbb{R})}$, $\forall t \in \mathbb{R}_+$.

ii) If $u_0 \in BV(\mathbb{R})$*, then*

$$|u(t,\cdot)|_{BV(\mathbb{R})} \leq |u_0|_{BV(\mathbb{R})}, \ \ \forall t \in \mathbb{R}_+.$$

Appendix B

Cell-Centered Finite Volume Schemes for Systems of Conservation Laws

The appendix is based mainly on the course[1] *Numerical transport* taught by Prof. Nicolas Seguin, and also Eymard, Gallouët, and Herbin, 2003, Harten, Lax, and Leer, 1983, with notations adapted to the main style of this context.

1 Numerical solutions for scalar conservation laws

The following properties are assumed to be satisfied by the data of (A.1).

Assumption B.1. *The flux vector field F belongs to $C^1(\mathbb{R}^m, \mathbb{R}^m)$, the initial data U_0 belongs to $(L^\infty(\mathbb{R}))^m$ and U_m, $U_M \in \mathbb{R}$ are such that $U_0 \in [U_m, U_M]^m$, a.e. on $\mathbb{R}$.*

For systems (A.1) of conservation laws, the following assumption is proposed by Lax to introduce the notation of entropy weak solution.

Assumption B.2 (Lax's assumption)**.** *We consider only systems of conservation laws* (A.1) *that possess a pair of entropy/entropy-flux in the sense of Definition A.11.*

Let us propose a definition of the *admissible meshes* for the finite volume schemes.

Definition B.1 (One-dimensional admissible mesh)**.** *An* admissible mesh $\mathcal{T}$ *of* $\mathbb{R}$ *is given by an increasing sequence of real values* $\left(x_{i+\frac{1}{2}}\right)_{i\in\mathbb{Z}}$*, such that*

$$\mathbb{R} = \bigcup\nolimits_{i\in\mathbb{Z}} \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right].$$

The mesh $\mathcal{T}$ is the set $\mathcal{T} := \{C_i; i \in \mathbb{Z}\}$ of subsets of $\mathbb{R}$ defined by $C_i := \left(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right)$ for all $i \in \mathbb{Z}$. The length of C_i is denoted by Δx_i, so that $\Delta x_i := x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ for all $i \in \mathbb{Z}$. Define the mid-point of the i^{th} cell C_i as $x_i := \frac{1}{2}\left(x_{i-\frac{1}{2}} + x_{i+\frac{1}{2}}\right)$. It is assumed that

$$\Delta x = \text{size}(\mathcal{T}) := \sup_{i\in\mathbb{Z}} \Delta x_i < +\infty,$$

[1] Master 2 in Fundamental Mathematics and Application Program, 2018–2019, Université de Rennes 1, France.

and that, for some $\alpha_{\mathcal{T}} \in \mathbb{R}_+^\star$, $\alpha_{\mathcal{T}} \Delta x \leq \inf_{i \in \mathbb{Z}} \Delta x_i$.

In particular, when $\Delta x_i = const$ *for all* $i \in \mathbb{Z}$, $\mathcal{T}$ *is called a* uniform mesh.

On a 1D admissible mesh $\mathcal{T}$, define

$$U_i^0 := \frac{1}{\Delta x_i} \int_{C_i} U_0\left(x\right) dx, \ \ \forall i \in \mathbb{Z}.$$

Integrating the PDE in (A.1) over $\left(t^n, t^{n+1}\right) \times C_i$ yields

$$\int_{C_i} U\left(t^{n+1}, x\right) dx - \int_{C_i} U\left(t^n, x\right) dx \\ + \int_{t^n}^{t^{n+1}} F\left(U\left(t, x_{i+\frac{1}{2}}\right)\right) dt - \int_{t^n}^{t^{n+1}} F\left(U\left(t, x_{i-\frac{1}{2}}\right)\right) dt = 0.$$

If U_i^n approximates $\frac{1}{\Delta x_i} \int_{C_i} U\left(t^n, x\right) dx$, one has

$$\Delta x_i \left(U_i^{n+1} - U_i^n\right) + \Delta t \left(F_{i+\frac{1}{2}}^n - F_{i-\frac{1}{2}}^n\right) = 0, \ \ \forall i \in \mathbb{Z}, \ \ \forall n \in \mathbb{N},$$

where $F_{i+\frac{1}{2}}^n$ should approximate $\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} F\left(U\left(t, x_{i+\frac{1}{2}}\right)\right) dt.$

We assume that $F_{i+\frac{1}{2}}^n$ can be well approximated by

$$F_{i+\frac{1}{2}}^n = G\left(U_i^n, U_{i+1}^n\right), \ \ \forall i \in \mathbb{Z}, \ \ \forall n \in \mathbb{N},$$

where $G\left(U, V\right)$ is called a *numerical flux*.

The numerical approximations to the entropy weak solution of (A.1) that can be obtained by 3-point explicit cell-centered finite volume schemes in conservation form:

$$\begin{cases} U_i^0 := \dfrac{1}{\Delta x_i} \displaystyle\int_{C_i} U_0\left(x\right) dx, & \forall i \in \mathbb{Z}, \\ U_i^{n+1} := U_i^n - \nu_i \left[G\left(U_i^n, U_{i+1}^n\right) - G\left(U_{i-1}^n, U_i^n\right)\right], & \forall i \in \mathbb{Z}, \ \ \forall n \in \mathbb{N}, \end{cases} \tag{B.1}$$

where $\nu_i := \frac{\Delta t}{\Delta x_i}$ for all $i \in \mathbb{Z}$.

On uniform meshes, assume that $U_0\left(x\right)$ is equal to some reference state $U_\star$ for $|x|$ large:

$$U_0\left(x\right) = U_\star, \ \text{for} \ \ |x| > M,$$

then

$$U_i^n = U_\star, \ \text{for} \ \ |i| \, \Delta x > M + n\Delta x.$$

Remark B.1. *It is deduced immediately from Definition A.11 that the entropy* Φ *can be altered by adding an arbitrary inhomogeneous affine mapping, i.e.,* $A \cdot X + b$ *for some* $A \in \mathbb{R}^m$ *and* $b \in \mathbb{R}$. *Adding such an affine mapping to* Φ *will not alter its* *convexity, but achieves the following properties:*

$$\Phi\left(U_\star\right)=0,\quad \nabla\Phi\left(U_\star\right)=0_{\mathbb{R}^m}.$$

Since Φ is convex, it follows from the last equations that

$$\Phi\left(U\right)>0,\ \textit{for } U\neq U_\star.$$

If fact, if U is strictly convex,

$$\Phi\left(U\right)\geq c\left\|U-U_\star\right\|_2^2\ \textit{for some constant } c>0.$$

Demonstration of Remark B.1. These statements are obvious for the scalar case. For the vector case, given a reference state $U_\star$, an entropy Φ, we consider a function $\widetilde{\Phi}:\mathbb{R}^m\to\mathbb{R}$ defined by

$$\widetilde{\Phi}\left(X\right):=\Phi\left(X\right)+A\cdot X+b,\quad \forall X\in\mathbb{R}^m,\ \text{for some } A\in\mathbb{R}^m,\quad b\in\mathbb{R}.$$

Write $U_\star=\left(U_{\star,i}\right)_{i=1}^m$, then the system $\widetilde{\Phi}\left(U_\star\right)=\nabla\widetilde{\Phi}\left(U_\star\right)=0$ reads

$$\begin{cases}\Phi\left(U_\star\right)+A\cdot U_\star+b=0,\\ \partial_i\Phi\left(U_\star\right)+A_i=0,\quad \forall i\in\left\{1,\ldots,m\right\}.\end{cases}$$

Solving this system yields

$$\begin{cases}b=-\Phi\left(U_\star\right)+\nabla\Phi\left(U_\star\right)\cdot U_\star,\\ A=-\nabla\Phi\left(U_\star\right),\end{cases}$$

and thus

$$\widetilde{\Phi}\left(X\right)=\Phi\left(X\right)-\Phi\left(U_\star\right)-\nabla\Phi\left(U_\star\right)\cdot\left(X-U_\star\right),\quad \forall X\in\mathbb{R}^m.$$

The rest of Remark B.1 is straightforward. □

Next, note that a common expression of $F^n_{i+\frac{1}{2}}$ is used for both equations i and $i+1$ in (B.1), the scheme (B.1) thus satisfies the property of conservativity, common to all finite volume schemes.

Proposition B.1 (Conservation on uniform meshes)**.** *If $U_0\in\left(L^1\left(\mathbb{R}\right)\right)^m$, the total mass is preserved, i.e.,*

$$\Delta x\sum_{i\in\mathbb{Z}}U_i^n=\Delta x\sum_{i\in\mathbb{Z}}U_i^0,\quad \forall n\in\mathbb{N}.$$

In particular, for scalar case $m=1$, if $u_0\geq 0$, then

$$\Delta x\sum_{i\in\mathbb{Z}}u_i^0=\Delta x\sum_{i\in\mathbb{Z}}u_i^n=\left\|u_0\right\|_{L^1(\mathbb{R})},\quad \forall n\in\mathbb{N}.$$

Proof. Since the updates on U_i^{n+1} and U_{i+1}^{n+1} share the same numerical flux $G\left(U_i^n,U_{i+1}^n\right)$, it is easily deduced from the recursion in (B.1) the desired result. □

Definition B.2 (Consistency with entropy inequality). *The finite volume scheme* (B.1) *is said to be* consistent *with the entropy inequality* (A.39) *if the following* discrete entropy inequality *is satisfied:*

$$\Phi_i^{n+1} \leq \Phi_i^n - \nu_i \left(\Gamma_{i+\frac{1}{2}}^n - \Gamma_{i-\frac{1}{2}}^n\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in \mathbb{N},$$

where

$$\Phi_i^n := \Phi\left(U_i^n\right), \quad \Gamma_{i+\frac{1}{2}}^n := \Gamma\left(U_i^n, U_{i+1}^n\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in \mathbb{N},$$

where $\Gamma\left(U, V\right)$ *is a* numerical entropy flux*, consistent with the entropy flux:*

$$\Gamma\left(X, X\right) = \Psi\left(X\right), \quad \forall X \in \mathbb{R}^m.$$

Now, on uniform meshes (so $\nu_i = \nu$ for all $i \in \mathbb{Z}$), summing the discrete entropy inequality over all $i \in \mathbb{Z}$ yields

$$\sum_{i \in \mathbb{Z}} \Phi_i^{n+1} \leq \sum_{i \in \mathbb{Z}} \Phi_i^n, \quad \forall n \in \mathbb{N}.$$

In other words: *total entropy is a decreasing function of time.* The finite volume scheme consistent with the entropy inequality is then said to be *total entropy diminishing.* In particular,

$$\sum_{i \in \mathbb{Z}} \Phi_i^n \leq \sum_{i \in \mathbb{Z}} \Phi_i^0, \quad \forall n \in \mathbb{N}.$$

This is an a priori inequality for solutions of the finite volume scheme (B.1), analogous to the energy inequality for linear symmetric hyperbolic differential and difference equations. Recall that $\Phi\left(U\right) > 0$ for all $U \neq U_\star$, this is an a priori estimate for the solutions of the finite volume scheme (B.1), and indicates that the scheme is stable. Nevertheless, the last inequality on total entropy is not strong enough to prove the pointwise boundedness of solutions of (B.1), or the existence of convergent subsequences.

In the case of a so-called $2p+1$ point scheme with $p \in \mathbb{N}^\star$, the numerical flux G may be written as

$$F_{i+\frac{1}{2}}^n = G_{i+\frac{1}{2}}^n \left(U_{i-p+1}^n, \ldots, U_{i+p}^n\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in \mathbb{N}, \tag{B.2}$$

where $G_{i+\frac{1}{2}}^n$ is the numerical flux function at point $x_{i+\frac{1}{2}}$ and time t_n, which determines the scheme. The fact that the numerical flux may depend on the interface and the time is important in some applications, e.g., in the case of boundary faces or interfaces coupling different domains.

In particular, for $p = 1$, the numerical flux for a 3-point scheme reads

$$F_{i+\frac{1}{2}}^n = G_{i+\frac{1}{2}}^n \left(U_i^n, U_{i+1}^n\right), \quad \forall i \in \mathbb{Z}, \quad \forall n \in \mathbb{N}.$$

The property of consistency for the finite volume scheme (B.1)-(B.2) with $2p+1$ points, is ensured by the following condition:

$$G(X, \ldots, X) = F(X), \ \ \forall X \in \mathbb{R}^m.$$

1.1 Monotone flux schemes

In this section, we propose a sufficient condition on the numerical flux G to prove the convergence of the approximate finite volume solution $U_{\mathcal{T},\Delta t}$ defined a.e. on $\mathbb{R}_+ \times \mathbb{R}$ from the discrete unknowns U_i^n, $i \in \mathbb{Z}$, $n \in \mathbb{Z}$ which are computed in (B.1):

$$U_{\mathcal{T},\Delta t}(t,x) := \sum_{i\in\mathbb{Z}} \sum_{n\in\mathbb{N}} U_i^n \mathbf{1}_{[t^n,t^{n+1})\times\left[x_{i-\frac{1}{2}},x_{i+\frac{1}{2}}\right)}(t,x), \ \text{a.e.} \ (t,x) \in \mathbb{R}_+ \times \mathbb{R}, \tag{B.3}$$

and for uniform meshes, simply denoted by $U_{\Delta x,\Delta t}(t,x)$, to the entropy weak solution $u \in L^\infty(\mathbb{R}_+ \times \mathbb{R})$ in the scalar case $m = 1$, and to an entropy solution $U \in L^\infty(\mathbb{R}_+ \times \mathbb{R}, \mathbb{R}^m)$ if we are lucky enough!

An interesting class of schemes if the class of 3-point schemes with a monotone flux, which is defined as follows.

Definition B.3 (Consistency, monotone flux for scalar conservation laws)**.** *Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a Lipschitz continuous function (possibly only locally Lipschitz continuous).*

i) The numerical flux $\mathbf{g}$ *is* consistent *if*

$$\mathbf{g}(x,x) = f(x), \ \ \forall x \in \mathbb{R}.$$

ii) Under Assumption B.1, the finite volume scheme (B.1) is said to be a monotone flux scheme *if the function g, only depending on f, u_m, and u_M, satisfies the following assumptions:*

a) $g : \mathbb{R}^2 \to \mathbb{R}$ is locally Lipschitz continuous.

b) g is consistent on $[u_m, u_M]$, i.e.,

$$\mathbf{g}(x,x) = f(x), \ \ \forall x \in [u_m, u_M],$$

c) g is non-decreasing w.r.t. its first variable and non-increasing w.r.t. the second one:

$$\partial_1 \mathbf{g}(x,y) \geq 0, \ \ \partial_2 \mathbf{g}(x,y) \leq 0, \ \ \forall x,y \in \mathbb{R}.$$

The numerical flux g is also said to be monotone.

Monotone flux schemes are consistent with the finite volume sense, i.e., they are L^∞-stable under the Courant-Friedrichs-Lewy condition of the type $\Delta t \leq C_1 \Delta x$, where C_1 depends only on G and U_0. Moreover, they are also consistent with the entropy inequalities under a condition of the type $\Delta t \leq C_2 \Delta x$, where C_2 only depends on G and U_0.

Remark B.2. *For scalar conservation laws, under a Courant-Friedrichs-Lewy condition, a monotone flux scheme is a monotone scheme, i.e., the scheme can be written under the form*

$$u_i^{n+1} = H_i\left(u_{i-1}^n, u_i^n, u_{i+1}^n\right), \ \ \forall i \in \mathbb{Z}, \ \ \forall n \in \mathbb{N},$$

with H_i nondecreasing w.r.t. its 3 arguments. On uniform meshes, we can take $H_i = H$, for all $i \in \mathbb{Z}$.

Unfortunately, there is no notion of monotonicity for the case of systems of conservation laws. We propose the following "failed definition" of monotonicity for systems of conservation laws in order to see the gaps between the scalar case and the system one.

Definition B.4 ("Consistency, monotone flux schemes for systems of conservation laws")**.** *Let $G : \mathbb{R}^{2m} \to \mathbb{R}^m$ be a Lipschitz continuous function (possibly only locally Lipschitz continuous).*

i) The numerical flux G is consistent *if*

$$G\left(X, X\right) = F\left(X\right), \ \ \forall X \in \mathbb{R}^m.$$

ii) Under Assumption B.1, the finite volume scheme (B.1) is said to be a monotone flux scheme *if the function G, only depending on F, U_m and U_M, satisfies the following assumptions:*

a) G is locally Lipschitz continuous from $\mathbb{R}^{2m}$ to $\mathbb{R}^m$,

b) G is consistent on $\left[U_m, U_M\right]^m$, i.e.,

$$G\left(X, X\right) = F\left(X\right), \ \ \forall X \in \left[U_m, U_M\right]^m,$$

c) G is non-decreasing w.r.t. its first m variables and non-increasing w.r.t. the rest[2] *in the following sense:*

$$\begin{cases} \partial_i G_j\left(U, V\right) \geq 0, & \forall i \in \{1, \dots, m\}, \\ \partial_i G_j\left(U, V\right) \leq 0, & \forall i \in \{m+1, \dots, 2m\}, \end{cases}$$

for all $j \in \{1, \dots, m\}$, $U \in \mathbb{R}^m$, $V \in \mathbb{R}^m$.

The numerical flux G is also said to be monotone*.*

Example B.1 (Vector-generalizations of some monotone flux schemes)**.** Under the assumption B.1, here are some numerical flux functions G, which are generalized for the vector case, for which the finite volume scheme (B.2) is a monotone flux scheme in the scalar case[3].

If the flux F is in the simplest linear form with $\nabla F\left(X\right) \in \mathbb{R}_+^m$, i.e., $F\left(X\right) = A \odot U$, for all $X \in \mathbb{R}^m$, with $A \in \mathbb{R}_+^m$, then the following classical numerical schemes is obtained with their corresponding numerical fluxes:

[2] Briefly: $G\left(\nearrow^m, \searrow^m\right)$.

[3] Some schemes are no longer monotone to the vector case when generalized as in this example.

i) *Upwind scheme* (also, *upstream*): The upwind numerical flux is defined by

$$G(U,V) := F(U) = A \odot U, \ \ \forall U, V \in [U_m, U_M]^m.$$

ii) *Downwind scheme* (also, *downstream*): The downwind numerical flux is defined by

$$G(U,V) := F(V) = A \odot V, \ \ \forall U, V \in [U_m, U_M]^m.$$

iii) *Central scheme* The central scheme numerical flux is defined by

$$G(U,V) := \frac{1}{2} F(U+V) = \frac{1}{2} A \odot (U+V), \ \ \forall U, V \in [U_m, U_M]^m.$$

Otherwise, if the flux F is nonlinear, then the following numerical schemes are commonly used:

i) *Flux splitting scheme*: For the scalar case, assume $f = f_1 + f_2$, with $f_1, f_2 \in C^1(\mathbb{R}, \mathbb{R})$, $f_1'(s) \geq 0$ and $f_2'(s) \leq 0$ for all $s \in [U_m, U_M]$ (such a decomposition for f is always possible) and the splitting numerical flux is taken as

$$\mathbf{g}(u,v) = f_1(u) + f_2(u), \ \forall a, b \in \mathbb{R}.$$

For the vector case, assume $F = F_1 + F_2$, with $F_1, F_2 \in C^1(\mathbb{R}^m, \mathbb{R}^m)$,

$${F_1}'(X) \in \mathbb{R}^m_+, \ \ {F_2}'(X) \in \mathbb{R}^m_-, \ \ \forall X \in [U_m, U_M]^m,$$

and the splitting numerical flux is taken by

$$G(U,V) = F_1(U) + F_2(V), \ \ \forall U, V \in [U_m, U_M]^m.$$

If $F' \in \mathbb{R}^m_+$, taking $F_1 = F$ and $F_2 = 0$, the flux splitting scheme becomes the upwind scheme, i.e., $G(U,V) = F(U)$.

ii) *Rusanov*, or *modified Lax-Friedrichs scheme* For the scalar case, the Rusanov numerical flux is given by

$$\mathbf{g}(u,v) = \frac{f(u) + f(v)}{2} + \frac{D}{2}(u - v),$$

with $D \in \mathbb{R}$ such that $D \geq \max\{|f'(s)|; s \in [U_m, U_M]\}$. In this modified version of the Lax-Friedrichs scheme, the coefficient D only depends on f, U_m and U_M, while the original Lax-Friedrichs scheme consists in taking $D = \frac{h}{k}$ in the uniform mesh case. Thus, the original one is monotone under the condition

$$\nu^{-1} \geq \max\left\{\left|f'(s)\right|; s \in [U_m, U_M]\right\}.$$

Nevertheless, the convergence of the original Lax-Friedrichs scheme necessarily requires an inverse CFL condition while the modified one does not. Note also that the modified Lax-Friedrichs scheme consists in a particular flux splitting scheme with

$$f_1(s) = \frac{1}{2}f(s) + Ds, \;\; f_2(s) = \frac{1}{2}f(s) - Ds, \;\; \forall s \in [U_m, U_M].$$

For the vector case, the Rusanov numerical flux is defined by

$$G(U,V) := \frac{1}{2}(F(U) + F(V)) + \frac{1}{2}A \odot (U - V), \;\; \forall U, V \in [U_m, U_M]^m.$$

where $A = (A_1, \ldots, A_m) \in \mathbb{R}^m$, and

$$A_i \geq \max_{\kappa \in [U_m, U_M]^m} |\nabla F_i(\kappa)|, \;\; i \in \{1, \ldots, m\}.$$

iii) *Godunov* For the scalar case, the Godunov numerical flux is defined by

$$\mathbf{g}(u,v) = \begin{cases} \min\limits_{s \in [u,v]} f(s), & \text{if } u \leq v, \\ \max\limits_{s \in [v,u]} f(s), & \text{if } v \leq u. \end{cases}$$

For the vector case, the Godunov numerical flux is defined by

$$G_i(U,V) := \begin{cases} \min\limits_{\kappa \in [U \perp V, U \top V]} F_i(\kappa), & \text{if } \sum\limits_{i=1}^m U_i < \sum\limits_{i=1}^m V_i, \\ \max\limits_{\kappa \in [U \perp V, U \top V]} F_i(\kappa), & \text{if } \sum\limits_{i=1}^m U_i \geq \sum\limits_{i=1}^m V_i, \end{cases} \quad \forall U, V \in [U_m, U_M]^m,$$

or also

$$G_i(U,V) := \begin{cases} \min\limits_{\kappa \in [U \perp V, U \top V]} F_i(\kappa), & \text{if } U_1 < V_1, \\ \max\limits_{\kappa \in [U \perp V, U \top V]} F_i(\kappa), & \text{if } U_1 \geq V_1, \end{cases} \quad \forall U, V \in [U_m, U_M]^m,$$

where $\kappa \in [U \perp V, U \top V]$ means in entrywise:

$$\kappa_i \in [U_i \perp V_i, U_i \top V_i], \;\; \forall i \in \{1, \ldots, m\}.$$

iv) *Engquist-Osher* For the scalar case, the Engquist-Osher numerical flux is defined by

$$\mathbf{g}(u,v) = \frac{1}{2}(f(u) + f(v)) - \frac{1}{2}\int_u^v |f'(s)|\, ds, \;\; \forall u, v \in [U_m, U_M].$$

For the vector case, the Engquist-Osher numerical flux is defined by

$$G(U,V) := \frac{1}{2}(F(U) + F(V)) - \frac{1}{2}\int_U^V \begin{pmatrix} |\nabla F_1(\kappa)| \\ \vdots \\ |\nabla F_m(\kappa)| \end{pmatrix} d\kappa, \;\; \forall U, V \in [U_m, U_M]^m.$$

v) *Roe* For the scalar case, the Roe numerical case is defined by

$$\mathbf{g}(u,v) = \frac{1}{2}\left(f(u) + f(v)\right) - \frac{a(u,v)}{2}, \; \forall u, v \in [U_m, U_M],$$

where

$$a(u,v) = \begin{cases} \dfrac{f(v) - f(u)}{v-u}, & \text{if } u \neq v, \\ f'(u), & \text{else.} \end{cases}$$

For the scalar case, the Roe numerical flux is defined by

$$G(U,V) := \frac{1}{2}\left(F(U) + F(V)\right) - \frac{1}{2}A(U,V), \; \forall U, V \in [U_m, U_M]^m,$$

where

$$A(U,V) = \begin{cases} \dfrac{F(V) - F(U)}{|V-U|}\operatorname{sgn}\left(V_{i_0} - U_{i_0}\right), \text{ with } i_0 = \min\{i; U_i \neq V_i\}, & \text{if } U \neq V, \\ \nabla F(U), & \text{if } U = V. \end{cases}$$

Remark B.3. *In the case of a nondecreasing (resp., nonincreasing) vector field F w.r.t. all m variables, the Godunov scheme reduces to the upwind (resp., downwind) finite volume scheme, sometimes called "upwind finite difference" (resp., "downwind finite difference") scheme.*

1.2 L^∞-stability for monotone flux scheme

The following L^∞-estimate holds for the scalar case.

Proposition B.2 (L^∞-estimate for scalar conservation laws)**.** *Under Assumption B.1, let $\mathcal{T}$ be an admissible mesh and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $u_{\mathcal{T},\Delta t}$ be the finite volume approximate solution defined by (B.3) and assume that the scheme is a monotone flux scheme. Let g_1 and g_2 be the Lipschitz constants of g on $[U_m, U_M]^2$ w.r.t. its two arguments.*

Under the Courant-Friedrichs-Lewy (CFL) condition

$$\Delta t \leq \frac{\inf_{i\in\mathbb{Z}} \Delta x_i}{g_1 + g_2}, \tag{B.4}$$

(note that taking

$$\nu \leq \frac{\alpha_{\mathcal{T}}}{g_1 + g_2},$$

implies (B.4)), the approximate solution $u_{\mathcal{T},\Delta t}$ satisfies

$$U_m \leq u_{\mathcal{T},\Delta t}(t,x) \leq U_M, \; \text{a.e. } (t,x) \in \mathbb{R}_+ \times \mathbb{R}.$$

Proof. See Eymard, Gallouët, and Herbin, 2003, p. 136 or the Lecture "Numerical Transport" of Prof. Nicolas Seguin. □

Unfortunately, for systems of conservation laws, the corresponding vector field H is not non-decreasing w.r.t. all of its variables anymore, which is shown in the following quick demonstration.

Demonstration of failure of monotonicity notion for systems of conservation laws. For each $i \in \mathbb{Z}$, H_i can be written explicitly as

$$H_i\left(U,V,W\right) = V - \nu_i\left[G\left(V,W\right) - G\left(U,V\right)\right], \ \ \forall U,V,W \in \mathbb{R}^m.$$

The first-order derivatives of H_i are given by, for all $j \in \{1,\dots,m\}$,

$$\begin{aligned}
&\partial_k H_{i,j}\left(U,V,W\right) = \nu_i \partial_k G_j\left(U,V\right) \geq 0, \ \ \forall k \in \left\{1,\dots,m\right\}, \\
&\partial_k H_{i,j}\left(U,V,W\right) = \delta_{kj} - \nu_i\left[\partial_{k-m} G_j\left(V,W\right) - \partial_k G_j\left(U,V\right)\right], \ \ \forall k \in \left\{m+1,\dots,2m\right\}, \\
&\partial_k H_{i,j}\left(U,V,W\right) = -\nu_i \partial_{k-m} G_j\left(V,W\right) \geq 0, \ \ \forall k \in \left\{2m+1,\dots,3m\right\}.
\end{aligned}$$

Now, we notice for the partial derivatives $\partial_k H_{i,j}$ when $k \in \{m+1,\dots,2m\}$, $i \in \mathbb{Z}$, $j \in \{1,\dots,m\}$ and $k \neq j$:

$$\partial_k H_{i,j}\left(U,V,W\right) = -\nu_i\left[\underbrace{\partial_{k-m} G_j\left(V,W\right)}_{\geq 0} - \underbrace{\partial_k G_j\left(U,V\right)}_{\leq 0}\right] \leq 0, \ \ \forall U,V,W \in \mathbb{R}^m.$$

Hence, the “notion of monotonicity for systems of conservation laws” collapses completely. □

1.3 Discrete entropy inequalities

Proposition B.3 (Discrete entropy inequalities for scalar conservation laws)**.** *Under Assumption B.1, let $\mathcal{T}$ be an admissible mesh and $\Delta t \in \mathbb{R}_+^\star$ be the time step.*

Let $u_{\mathcal{T},\Delta t}$ be the finite volume approximation solution defined by (B.3) and assume that the scheme is a monotone flux scheme. Let g_1 and g_2 be the Lipschitz constants of g on $\left[U_m, U_M\right]^2$ w.r.t. its two arguments. Under the CFL condition (B.4), the following inequation holds:

$$\begin{aligned}
\frac{1}{\nu_i}\left(\left|u_i^{n+1} - \kappa\right| - \left|u_i^n - \kappa\right|\right) + \mathbf{g}\left(u_i^n \top \kappa, u_{i+1}^n \top \kappa\right) - \mathbf{g}\left(u_i^n \bot \kappa, u_{i+1}^n \bot \kappa\right) \\
- \mathbf{g}\left(u_{i-1}^n \top \kappa, u_i^n \top \kappa\right) + \mathbf{g}\left(u_{i-1}^n \bot \kappa, u_i^n \bot \kappa\right) \leq 0,
\end{aligned} \tag{B.5}$$

for all $i \in \mathbb{Z}$, $n \in \mathbb{N}$, $\kappa \in \mathbb{R}$.

Proof. See Eymard, Gallouët, and Herbin, 2003, pp. 136–137. □

For the vector case, a generalization of discrete entropy inequalities fails completely since there is no notion of monotonicity of the corresponding vector field H in order that the proof for scalar case can be extended for systems of conservation laws.

1.4 Convergence of the upwind scheme in the general case

For simplicity, we first consider the scalar case of a nondecreasing function f and of the classical upwind scheme. The following result is a "weak BV" estimate.

Proposition B.4 (Weak BV estimate for upwind scheme for scalar conservation laws)**.** *Under Assumption B.1, assume that f is nondecreasing. Let $\zeta\,(0,1)$ be a given value. Let $\mathcal{T}$ be an admissible mesh and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Suppose the following CFL condition holds:*

$$\Delta t \leq (1-\zeta)\,\frac{\inf_{i\in\mathbb{Z}}\Delta x_i}{\mathrm{Lip}\,(f;[U_m,U_M])}. \tag{B.6}$$

(The condition

$$\nu \leq \frac{\alpha_{\mathcal{T}}\,(1-\zeta)}{\mathrm{Lip}\,(f;[U_m,U_M])} \tag{B.7}$$

implies (B.6)). Let $(U_i^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$ be given by the finite volume scheme (B.1) and $\mathbf{g}\,(u,v) = f\,(u)$. Let $R \in \mathbb{R}_+^\star$ and $T \in \mathbb{R}_+^\star$ and assume $h < R$ and $k < T$. Let $i_0 \in \mathbb{Z}$, $i_1 \in \mathbb{Z}$ and $N \in \mathbb{N}$ be such that $-R \in \overline{K}_{i_0}$, $R \in \overline{K}_{i_1}$, and $T \in \left(t^N, t^{N+1}\right]$. Then there exists $C \in \mathbb{R}_+^\star$, only depending on R, T, u_0, $\alpha_{\mathcal{T}}$, f and ζ, such that

$$\sum_{i=i_0}^{i_1}\sum_{n=0}^{N}\Delta t\left|f\left(u_i^n\right) - f\left(u_{i-1}^n\right)\right| \leq C\Delta x^{-\frac{1}{2}}.$$

Proof. See Eymard, Gallouët, and Herbin, 2003, pp. 137–139. □

Theorem B.1 (Convergence for upwind scheme for scalar conservation laws)**.** *Assume Assumption B.1 and f nondecreasing. Let $\zeta \in (0,1)$ and $\alpha > 0$ be given. For an admissible mesh $\mathcal{T}$ of $\mathbb{R}$ and a time step $\Delta t \in \mathbb{R}_+^\star$, assume the CFL condition (B.6) holds ((B.7) is a sufficient condition, note that ζ and α do not depend on $\mathcal{T}$), let $u_{\mathcal{T},\Delta t}$ be the finite volume approximate solution defined by (B.1) and $\mathbf{g}\,(u,v) = f\,(u)$.*

Then the function $u_{\mathcal{T},\Delta t}$ converges to the unique entropy weak solution U of (A.2) in $L^1_{\mathrm{loc}}\,(\mathbb{R}_+\times\mathbb{R})$ as $\mathrm{size}\,(\mathcal{T})$ tends to 0.

Proof. See Eymard, Gallouët, and Herbin, 2003, pp. 139–141. □

Remark B.4. *Proposition B.4 and Theorem B.1 only consider the case $f' \geq 0$ and the upwind scheme. It is quite easy to generalize the result for any $f \in C^1\,(\mathbb{R},\mathbb{R})$ and any monotone flux scheme. We can also consider other schemes, e.g., some 5-points schemes. For a given scheme, the proof of convergence of the approximate solution towards the entropy weak solution contains 2 steps:*

1. *prove an L^∞-estimate on the approximate solutions,*
2. *prove a "weak BV" estimate and some "discrete entropy inequality" in order to have the following property:*

If $(u_k)_{k\in\mathbb{N}}$ is a sequence of approximate solutions converging in the nonlinear weak-$\star$-sense, then

$$\lim_{k\to\mathbb{N}}\int_0^\infty\int_{\mathbb{R}}\left(|u_k(t,x)-\kappa|\,\partial_t\varphi+\left(f(u_k\top\kappa)-f(u_k\bot\kappa)\right)\partial_x\varphi\right)dxdt \\ +\int_{\mathbb{R}}|u_0-\kappa|\,\varphi(0,x)\,dx\geq 0,\ \ \forall\varphi\in C_c^1(\mathbb{R}_+\times\mathbb{R},\mathbb{R}_+),\ \ \forall\kappa\in\mathbb{R}.$$

1.5 Convergence proof using BV

This section devotes to the classical proof of convergence (for only 3-point schemes) requiring regularization of u_0 in $BV(\mathbb{R})$. This proof consists in using Helly's compactness theorem, which is a direct consequence of Kolmogorov's theorem.

Theorem B.2 (Kolmogorov compactness lemma)**.** *Let ω be an open bounded set of $\mathbb{R}^N$, $N\geq 1$, $1\leq q<\infty$ and $A\subset L^q(\omega)$. Then, A is relatively compact in $L^q(\omega)$ if and only if there exists $\{p(u)\,;u\in A\}\subset L^q\left(R^N\right)$ such that*

i) $p(u)=u$ a.e. on ω, for all $u\in A$,

ii) $\{p(u)\,;u\in A\}$ is bounded in $L^q\left(\mathbb{R}^N\right)$,

iii) $\|p(u)(\cdot+\eta)-p(u)\|_{L^q(R^N)}\to 0$ as $\eta\to 0$, uniformly w.r.t. $u\in A$.

Proposition B.5 (Consequence of Helly's theorem)**.** *Let $\mathcal{A}\subset L^\infty\left(\mathbb{R}^2\right)$. Assume that there exists $C\in\mathbb{R}_+$ and, for all $T>0$, there exists $C_T\in\mathbb{R}_+$ such that*

$$\|v\|_{L^\infty(\mathbb{R}^2)}\leq C,\ \ \forall v\in\mathcal{A},$$

and,

$$|v|_{BV((-T,T)\times\mathbb{R})}\leq C_T,\ \ \forall v\in\mathcal{A},\ \ \forall T>0.$$

Then for any sequence $(v_n)_{n\in\mathbb{N}}$ of elements of $\mathcal{A}$, there exists a subsequence, denoted by $(v_{n_k})_{n\in\mathbb{N}}$, and there exists $v\in L^\infty\left(\mathbb{R}^2\right)$, with

$$\|v\|_{L^\infty(\mathbb{R}^2)}\leq C,\ \ |v|_{BV((-T,T)\times\mathbb{R})}\leq C_T,\ \ \forall T>0.$$

such that $v_{n_k}\to v$ in $L^1_{\text{loc}}\left(\mathbb{R}^2\right)$ as $k\to\infty$, i.e.,

$$\int_{\varpi}|v_{n_k}-v|\,dx\to 0 \text{ as } n\to\infty,\ \ \forall\varpi\subset\mathbb{R}^2 \text{ compact.}$$

To use Proposition B.5, one first proves the following BV stability estimate for the approximate solution.

Proposition B.6 (Discrete space BV estimate for scalar conservation laws)**.** *Under Assumption B.1, assume that $u_0\in BV(\mathbb{R})$, let $\mathcal{T}$ be an admissible mesh and $\Delta t\in\mathbb{R}_+^\star$ be the time step. Let $(u_i^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$ be given by (B.1) and assume that the scheme is a monotone flux scheme. Let g_1 and g_2 be the Lipschitz constants of g on $[U_m,U_M]^2$*

w.r.t. its two arguments. Then, under the CFL condition (B.4)*, the following inequality holds:*

$$\sum_{i\in\mathbb{Z}} \left|u_{i+1}^{n+1} - u_i^{n+1}\right| \leq \sum_{i\in\mathbb{Z}} \left|u_{i+1}^{n} - u_i^{n}\right|, \ \ \forall n \in \mathbb{N}.$$

One then says the finite volume scheme (B.1) is *total variation diminishing* (TVD).

Proof. See Eymard, Gallouët, and Herbin, 2003, p. 142. □

Corollary B.1 (Discrete BV estimate)**.** *Under Assumption B.1, let $u_0 \in BV(\mathbb{R})$, let $\mathcal{T}$ be an admissible mesh and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Let $u_{\mathcal{T},\Delta t}$ be the finite volume approximation solution defined by* (B.1)*-*(B.3) *and assume that the scheme is a monotone flux scheme. Let g_1 and g_2 be the Lipschitz constants of g on $[U_m, U_M]^2$ w.r.t. its two arguments and assume that Δ satisfies the CFL condition* (B.4)*. Let*

$$u_{\mathcal{T},\Delta t}(t,x) = u_i^0 \ \text{a.e.} \ \ (t,x) \in \mathbb{R}_- \times C_i, \ \ \forall i \in \mathbb{Z},$$

(hence $u_{\mathcal{T},\Delta t}$ is defined a.e. on $\mathbb{R}^2$). Then, for any $T > 0$, there exists $C \in \mathbb{R}_+^\star$, only depending on u_0, g and T such that:

$$|u_{\mathcal{T},\Delta t}|_{BV((-T,T)\times\mathbb{R})} \leq C.$$

Proof. See Eymard, Gallouët, and Herbin, 2003, pp. 142–143. □

For the scalar case, consider a sequence of admissible meshes and time steps verifying the CFL condition, and the associated sequence of approximate solutions, which is prolonged on $\mathbb{R}_- \times \mathbb{R}$ as in Corollary B.1. By Proposition B.2 and Corollary B.1, the sequence of approximate solutions fulfills the hypotheses of Proposition B.5. Therefore, it is possible to extract a subsequence which converges in $L^1_{\text{loc}}(\mathbb{R}_+ \times \mathbb{R})$ to a function $u \in L^\infty\left(\mathbb{R}_+^\star \times \mathbb{R}\right)$. It remains to prove that the function u is the unique weak entropy solution of (A.2). To prove this, one may use the discrete entropy inequalities (B.5) and the strong BV estimate (B.6) or the following classical Lax-Wendroff theorem.

Theorem B.3 (Lax-Wendroff)**.** *Under Assumption B.1, let $\alpha > 0$ be given and let $(\mathcal{T}_k)_{k\in\mathbb{N}}$ be a sequence of admissible meshes (note that, for all $k \in \mathbb{N}$, the mesh $\mathcal{T}_k$ satisfies the hypothesis of Definition B.1 where $\mathcal{T} = \mathcal{T}_k$ and α is independent of k). Let $(\Delta t_k)_{k\in\mathbb{N}}$ be a sequence of (positive) time steps. Assume that*

$$\text{size}(\mathcal{T}_k) \to 0 \ \textit{and} \ \Delta t_k \to 0 \ \textit{as} \ k \to \infty.$$

For $k \in \mathbb{N}$, setting $\mathcal{T} = \mathcal{T}_k$ and $\Delta t = \Delta t_k$, let $u_k = u_{\mathcal{T},\Delta t}$ be the solution of (B.1)*-*(B.3) *and some $g : \mathbb{R}^2 \to \mathbb{R}$, only depending on f and u_0, locally Lipschitz continuous and such that $\mathbf{g}(s,s) = f(s)$ for all $s \in \mathbb{R}$.*

Assume that $(u_k)_{k\in\mathbb{N}}$ is bounded in $L^\infty(\mathbb{R}_+ \times \mathbb{R})$ and that $u_k \to u$ a.e. on $\mathbb{R}_+ \times \mathbb{R}$. Then, u is a weak solution to (A.2)*.*

Furthermore, assume that for any $\kappa \in \mathbb{R}$ there exists some locally Lipschitz continuous function $\Psi_\kappa : \mathbb{R}^2 \to \mathbb{R}$, only depending on f, u_0 and κ, such that

$$\Psi_\kappa(s, s) = f(s \top \kappa) - f(s \perp \kappa), \quad \forall s \in \mathbb{R},$$

and such that for all $m \in \mathbb{N}$,

$$\frac{1}{\Delta t}\left(\left|u_i^{n+1} - \kappa\right| - \left|u_i^n - \kappa\right|\right) + \frac{1}{\Delta x_i}\left(\Psi_\kappa\left(u_i^n, u_{i+1}^n\right) - \Psi_\kappa\left(u_{i-1}^n, u_i^n\right)\right) \leq 0,$$

for all $i \in \mathbb{Z}$, $n \in \mathbb{N}$, where $\{u_i^n; i \in \mathbb{Z},\ n \in \mathbb{N}\}$ is the solution to (B.1) for $\mathcal{T} = \mathcal{T}_k$ and $\Delta t = \Delta t_k$. Then, u is the entropy weak solution to (A.2).

Proof. See Eymard, Gallouët, and Herbin, 2003, pp. 143–145. □

Remark B.5. *Lax-Wendroff theorem (B.3) still holds with $(2p+1)$-point schemes for all $p > 1$.*

1.6 Higher order schemes

The approximate solution obtained from a 3-point monotone flux scheme converges to the entropy weak solution of (A.2) as the mesh size tends to 0 and under a CFL condition. However, 3-point schemes are known to be diffusive, so that the approximate solution is not very precise near the discontinuities. To reduce the diffusion, a 5-point scheme is proposed by introducing "slopes" on each discretization cell and limiting the slopes so that the scheme remains stable. A classical approach is the "MUSCL" (Monotonic Upwind Scheme for Conservation Laws) technique.

Reconstructing a slope on each cell enables to compute interface values on each side of an interface $x_{i+\frac{1}{2}}$. Then these values are used to compute the fluxes.

For each $n \in \mathbb{N}$,

- *Computation of the slopes*:

$$\widetilde{p}_i^n := \frac{u_{i+1}^n - u_{i-1}^n}{\Delta x_i + \frac{1}{2}\left(\Delta x_{i-1} + \Delta x_{i+1}\right)}, \quad \forall i \in \mathbb{Z}.$$

- *Limitation of the slopes*:

$$p_i^n := \alpha_i^n \widetilde{p}_i^n, \quad \forall i \in \mathbb{Z},$$

  where α_i^n is the largest number in $[0, 1]$ such that

$$\begin{cases} u_i^n + \dfrac{\Delta x_i}{2}\alpha_i^n \widetilde{p}_i^n \in \left[u_i^n \perp u_{i+1}^n, u_i^n \top u_{i+1}^n\right], \\ u_i^n - \dfrac{\Delta x_i}{2}\alpha_i^n \widetilde{p}_i^n \in \left[u_{i-1}^n \perp u_i^n, u_{i-1}^n \top u_i^n\right]. \end{cases}$$

  In practice, other formulas giving smaller values of α_i^n are sometimes needed for stability reasons.

- *Computation of u_i^{n+1} for $i \in \mathbb{Z}$:* One replaces $\mathbf{g}\left(u_i^n, u_{i+1}^n\right)$ in (B.1) by

$$\bar{\mathbf{g}}\left(u_{i-1}^n, u_i^n, u_{i+1}^n, u_{i+2}^n\right) := \mathbf{g}\left(u_i^n + \frac{\Delta x_i}{2} p_i^n, u_{i+1}^n - \frac{\Delta x_{i+1}}{2} p_{i+1}^n\right).$$

The constructed 5-point scheme is less diffusive than the original one and it remains stable thanks to the limitation of the slope. Indeed, if $\alpha_i^n = 1$, i.e., the limitation of the slope is inactive, the space diffusion term disappears from this new scheme, while the time "antidiffusion" term remains. Hence a higher order scheme for the time discretization is used, e.g., Runge-Kutta, or Heun method for the discretization of the time derivative.

The MUSCL scheme may be written as

$$\frac{U^{n+1} - U^n}{\Delta t} = \overline{H}\left(U^n\right), \quad \forall n \in \mathbb{N},$$

where $U^n = \left(u_i^n\right)_{i\in\mathbb{Z}}$, which may be seen as the explicit Euler discretization of

$$\partial_t U = \overline{H}\left(U\right).$$

The RK2 time discretization yields the following scheme;

$$\frac{U^{n+1} - U^n}{\Delta t} = \frac{1}{2}\overline{H}\left(U^n\right) + \frac{1}{2}\overline{H}\left(U^n + \Delta t \overline{H}\left(U^n\right)\right), \quad \forall n \in \mathbb{N}.$$

Such a second-order discretization in time allows larger time steps, without loss of stability. Results of convergence are possible with these new schemes.

1.7 Boundary conditions

A general convergence result will be applied to understand the sense of the boundary condition, described at $x = L$ in a simplified scalar case.

Consider the unknown $u : \mathbb{R}_+ \times (0, L) \to \mathbb{R}$, the flux $f \in C^1(\mathbb{R}, \mathbb{R})$ (or $f : \mathbb{R} \to \mathbb{R}$ Lipschitz continuous) and the initial datum is $u_0 \in L^\infty\left((0, L), \mathbb{R}\right)$. Let A, $B \in \mathbb{R}$ such that $A \le u_0 \le B$ a.e., we are interested in the following problem

$$\begin{cases} \partial_t u + \partial_x\left(f\left(u\right)\right) = 0, & \text{in } \mathbb{R}_+ \times (0, L), \\ u\left(0, x\right) = u_0\left(x\right), & \forall x \in (0, L), \end{cases} \tag{B.8}$$

and some boundary conditions prescribed later.

Let $\Delta x = \frac{L}{N}$, with $N \in \mathbb{N}^\star$, be the mesh size and Δt be the time step, assumed to be constant for simplicity. The discrete unknowns are now the values $u_i^n \in \mathbb{R}$ for $i \in \{1, \dots, N\}$ and $n \in \mathbb{N}$. The approximate solution is defined a.e. in $(0, L) \times \mathbb{R}$ by

$$u_{\Delta x, \Delta t}\left(t, x\right) := \sum_{i=1}^{N} \sum_{n\in\mathbb{N}} u_i^n \mathbf{1}_{[t^n, t^{n+1}) \times [(i-1)\Delta x, i\Delta x)}\left(t, x\right), \text{ in } \mathbb{R}_+ \times (0, L). \tag{B.9}$$

The discretization of the initial condition leads to

$$u_i^0 := \frac{1}{\Delta x}\int_{(i-1)\Delta x}^{i\Delta x} u_0\left(x\right)dx, \ \ \forall i \in \left\{1, \ldots, N\right\}.$$

For the computation of u_i^n for $n > 0$, one uses an explicit 3-point scheme:

$$u_i^{n+1} = u_i^n - \nu\left(f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n\right), \ \ \forall i \in \left\{1, \ldots, N\right\}, \ \ \forall n \in \mathbb{N}.$$

For $i \in \{1, \ldots, N-1\}$, one takes

$$f_{i+\frac{1}{2}}^n = \mathbf{g}\left(u_i^n, u_{i+1}^n\right), \ \ \forall i \in \left\{1, \ldots, N\right\}, \ \ \forall n \in \mathbb{N},$$

where $g : [A, B]^2 \to \mathbb{R}$ is the monotone numerical flux.

To complete the scheme, one has to define $f_{\frac{1}{2}}^n$ and $f_{N+\frac{1}{2}}^n$.

Let $\overline{u}, \overline{\overline{u}} \in L^\infty\left(\mathbb{R}_+\right)$ be such that $A \leq \overline{u}, \overline{\overline{u}} \leq B$, a.e. on $\mathbb{R}_+$, let g_0, $g_L : [A, B]^2 \to \mathbb{R}$ be monotone, and define

$$f_{\frac{1}{2}}^n = g_0\left(\overline{u}^n, u_1^n\right), \ \text{with } \overline{u}^n = \frac{1}{\Delta t}\int_{n\Delta t}^{(n+1)\Delta t} \overline{u}\left(t\right)dt, \tag{B.10}$$

$$f_{N+\frac{1}{2}}^n = g_L\left(u_N^n, \overline{\overline{u}}^n\right), \ \text{with } \overline{\overline{u}}^n = \frac{1}{\Delta t}\int_{n\Delta t}^{(n+1)\Delta t} \overline{\overline{u}}\left(t\right)dt. \tag{B.11}$$

Then the following convergence theorem holds.

Theorem B.4 (Convergence theorem)**.** *Let $f \in C^1\left(\mathbb{R}, \mathbb{R}\right)$ (or $f : \mathbb{R} \to \mathbb{R}$ Lipschitz continuous). Let $u_0 \in L^\infty\left((0, L)\right)$, $\overline{u}, \overline{\overline{u}} \in L^\infty\left(R_+\right)$ and A, $B \in \mathbb{R}$ be such that $A \leq u_0 \leq B$ a.e. on $(0, L)$, $A \leq \overline{u}, \overline{\overline{u}} \leq B$ a.e. on $\mathbb{R}_+$. Let g_0, $g_L : [A, B]^2 \to \mathbb{R}$ be monotone. Let L be a common Lipschitz constant for g, g_0 and g_L on $[A, B]^2$:*

$$L := \max\left(\text{Lip}\left(g, [A, B]^2\right), \text{Lip}\left(g_1, [A, B]^2\right), \text{Lip}\left(g_L, [A, B]^2\right)\right),$$

and let $\zeta > 0$. Then, if

$$\nu \leq \frac{1-\zeta}{L},$$

the equation (B.9) to (B.11) define an approximation solution $u_{\Delta x, \Delta t}$ which takes its values in $[A, B]$ and converges towards the unique solution of (B.12) in $L_{\text{loc}}^p\left([0, L] \times \mathbb{R}_+\right)$ for any $1 \leq p < \infty$, as $\Delta x \to 0$:

$$\begin{aligned} &u \in L^\infty\left((0, L) \times \mathbb{R}_+^\star\right), \ \forall \kappa \in [A, B], \ \forall \varphi \in C_c^1\left([0, L] \times \mathbb{R}_+, \mathbb{R}_+\right): \\ &\int_0^\infty \int_0^L \left[(u-\kappa)_\pm \partial_t \varphi + \text{sgn}_\pm\left(u-\kappa\right)\left(f\left(u\right) - f\left(\kappa\right)\right)\partial_x \varphi\right]dxdt \\ &\qquad + M\int_0^\infty \left(\overline{u}\left(t\right) - \kappa\right)_\pm \varphi\left(t, 0\right)dt + M\int_0^\infty \left(\overline{\overline{u}}\left(t\right) - \kappa\right)_\pm \varphi\left(t, L\right)dt \\ &\qquad + \int_0^L \left(u_0 - \kappa\right)\varphi\left(0, x\right)dx \geq 0, \end{aligned} \tag{B.12}$$

In (B.12)*, M is any bound for $|f'|$ on $[A,B]$ and the solution of* (B.12) *is independent of the choice of M. The definition of* $\operatorname{sgn}_{\pm}$ *is:*

$$\operatorname{sgn}_{+}(s) = \begin{cases} 1, & \textit{if } s > 0, \\ 0, & \textit{if } s < 0, \end{cases} \qquad \operatorname{sgn}_{-}(s) = \begin{cases} 0, & \textit{if } s > 0, \\ -1, & \textit{if } s < 0. \end{cases}$$

Remark B.6. *1. This convergence result also holds if the numerical flux g depends on i and n, provided that L is a common Lipschitz constant for all these functions.*

2. If the two entropies $(u-\kappa)_{\pm}$ are replaced by the sole entropy $|u-\kappa|$, one has an existence result since

$$|u-\kappa| = (u-\kappa)_{+} + (u-\kappa)_{-},$$

but no uniqueness result[4]*.*

3. This convergence result can be generalized to the multidimensional case.

If the solution u of (B.12) is regular enough, e.g., $u \in C^1([0,L]\times\mathbb{R}_+)$, u satisfies $u(t,0) = \overline{u}(t)$ and $u(t,L) = \overline{\overline{u}}(t)$ in the weak sense. This condition is very simple if f is monotone:

i) If $f' > 0$, then $u(\cdot,0) = \overline{u}$ and u does not depend on $\overline{\overline{u}}$.

ii) If $f' < 0$, then $u(\cdot,L) = \overline{\overline{u}}$ and u does not depend on $\overline{u}$.

2 Upwind schemes for linear fluxes

2.1 Derivation of upwind schemes for linear fluxes

This section is devoted to extend the first-order-accurate Courant-Isaacson-Rees scheme, the simplest upwind differencing scheme for constant-coefficient systems of conservation laws.

Consider (A.1) with the simplest linear flux[5] F, i.e., there exists a vector $A \in \mathbb{R}^m$ such that $F(X) = A \odot X$, for all $X \in \mathbb{R}^m$. The upwind numerical flux G is defined by

$$G_j(U,V) := \begin{cases} A_j U_j, & \text{if } A_j \geq 0, \\ A_j V_j, & \text{if } A_j < 0, \end{cases} \quad \forall j \in \{1,\ldots,m\}, \ \forall U,V \in \mathbb{R}^m,$$

and thus the upwind finite volume scheme (B.1) then reads

$$U_{i,j}^{n+1} := U_{i,j}^{n} - \nu_i A_j \cdot \begin{cases} U_{i,j}^{n} - U_{i-1,j}^{n}, & \text{if } A_j \geq 0, \\ U_{i+1,j}^{n} - U_{i,j}^{n}, & \text{if } A_j < 0, \end{cases} \tag{B.13}$$

for all $j \in \{1,\ldots,m\}$, $i \in \mathbb{Z}$, $n \in \mathbb{N}$.

[4] There is a counterexample to uniqueness, see references cited in Eymard, Gallouët, and Herbin, 2003, p. 149.

[5] Cf., Example B.1.

Under the well-known notations

$$\forall x \in \mathbb{R}, \quad \begin{cases} x_- := \min(x, 0) = \dfrac{1}{2}(x - |x|), \\ x_+ := \max(x, 0) = \dfrac{1}{2}(x + |x|), \end{cases}$$

the scheme (B.13) can be rewritten as

$$U_{i,j}^{n+1} := U_{i,j}^n - \nu_i \left[(A_j)_+ \left(U_{i,j}^n - U_{i-1,j}^n \right) + (A_j)_- \left(U_{i+1,j}^n - U_{i,j}^n \right) \right], \tag{B.14}$$

which can be rewritten one more time as[6]

$$U_{i,j}^{n+1} := \nu_i (A_j)_+ U_{i-1,j}^n + (1 - \nu_i |A_j|) U_{i,j}^n - \nu_i (A_j)_- U_{i+1,j}^n, \tag{B.15}$$

for all $j \in \{1, \ldots, m\}$, $i \in \mathbb{Z}$, $n \in \mathbb{N}$.

2.2 Properties of upwind schemes for linear fluxes

Under the Courant-Friedrichs-Lewy condition

$$\Delta t \leq \frac{\inf_{i \in \mathbb{Z}} \Delta x_i}{\max_{j \in \{1,\ldots,m\}} |A_j|}, \tag{B.16}$$

or stronger

$$\nu \leq \frac{\alpha_{\mathcal{T}}}{\max_{j \in \{1,\ldots,m\}} |A_j|}, \tag{B.17}$$

all coefficients of $U_{i,j}^n$ on the right in (B.14) are positive. It is straightforward that this upwind scheme is monotone.

Proposition B.7 (L^∞-stability of upwind schemes)**.** *Under Assumption B.1, let $\mathcal{T}$ be an admissible mesh and $\Delta t \in \mathbb{R}_+^\star$ be the time step. Under the CFL condition (B.16) or (B.17), the scheme (B.14) is stable in the maximum norm, i.e.,*

$$\max_{i \in \mathbb{Z}, j \in \{1,\ldots,m\}} \left| U_{i,j}^{n+1} \right| \leq \max_{i \in \mathbb{Z}, j \in \{1,\ldots,m\}} \left| U_{i,j}^n \right|.$$

Proof. Under the CFL condition (B.16) and (B.17), (B.15) indicates that $U_{i,j}^{n+1}$ is a convex combination of $U_{i,j}^n$ and $U_{i\pm1,j}^n$. This implies immediately the desired result. □

Furthermore, (B.14) can be rewritten as

$$U_{i,j}^{n+1} = U_{i,j}^n - \frac{\nu_i}{2} A_j \left(U_{i+1,j}^n - U_{i-1,j}^n \right) + \frac{\nu_i}{2} |A_j| \left(U_{i+1,j}^n - 2U_{i,j}^n + U_{i-1,j}^n \right), \tag{B.18}$$

[6] There is a mistake in the equation numbered (2.2c) in Harten, Lax, and Leer, 1983, p. 38: the term $\lambda(1 - |a|) v_j^n$ should be corrected as $(1 - \lambda |a|) v_j^n$.

which shows that, on uniform meshes, solutions of (B.14) can be thought of as approximation solutions of[7]

$$\partial_t U_j + A_j \partial_x U_j = \frac{\Delta x}{2} \left| A_j \right| \partial_x^2 U, \tag{B.19}$$

to second-order accuracy. It is observed that the viscosity term in (B.19) vanishes for $A_j = 0$, which allows perfectly resolved stationary shocks but may also result in admitting entropy violating discontinuities, see Harten, Lax, and Leer, 1983 for more details.

We now extend, but *on uniform meshes only*, our framework to systems of equations with constant coefficients, i.e., with the flux $F(X) = AX$ for all $X \in \mathbb{R}^m$, for some $A \in \mathrm{Mat}(m; \mathbb{R})$:

$$\partial_t U + A \partial_x U = 0, \text{ in } \mathbb{R}_+ \times \mathbb{R}, \quad A = \text{const matrix.} \tag{B.20}$$

Due to hyperbolicity, (B.20) can be diagonalized by a similarity transformation

$$\begin{aligned} V = P^{-1} U, \;\; P^{-1} A P = \Lambda = \mathrm{diag}\,(A_i)_{i=1}^m, \\ \partial_t V + \Lambda \partial_x V = 0, \text{ in } \mathbb{R}_+ \times \mathbb{R}. \end{aligned} \tag{B.21}$$

The components of V are called *characteristics variables* and (B.20) is a system of decoupled characteristic equations.

The Courant-Isaacson-Rees scheme can be extended to systems by applying the scalar scheme (B.18) to each of the decoupled scalar characteristic equations. In matrix form and on uniform meshes, this can be written as

$$V_i^{n+1} = V_i^n - \frac{\nu}{2} \Lambda \left(V_{i+1}^n - V_{i-1}^n \right) + \frac{\nu}{2} \left| \Lambda \right| \left(V_{i+1}^n - 2V_i^n + V_{i-1}^n \right), \;\; \forall i \in \mathbb{Z}, \;\; \forall n \in \mathbb{N}, \tag{B.22}$$

where $|\Lambda| := \mathrm{diag}\,(|A_i|)_{i=1}^m$.

In the original variables, the scheme (B.22) takes the form

$$U_i^{n+1} = U_i^n - \frac{\nu}{2} A \left(U_{i+1}^n - U_{i-1}^n \right) + \frac{\nu}{2} \left| A \right| \left(U_{i+1}^n - 2U_i^n + U_{i-1}^n \right), \;\; \forall i \in \mathbb{Z}, \;\; \forall n \in \mathbb{N}, \tag{B.23}$$

where $|A| := P \left| \Lambda \right| P^{-1}$. The stability condition for (B.22) and (B.23) clearly is

$$\nu \leq \frac{1}{\max\limits_{j \in \{1, \ldots, m\}} |A_j|}.$$

In general, define the matrix $\chi(A)$ by

$$\chi(A) = P \chi(\Lambda) P^{-1}, \text{ where } \chi(\Lambda) := \mathrm{diag}\,(\chi(A_i))_{i=1}^m. \tag{B.24}$$

Under the assumption of a full set of eigenvectors, we can compute $\chi(A)$ by $\chi(A) =$

[7] There is a mistake in the equation numbered (2.3) in Harten, Lax, and Leer, 1983, p. 39: The right-hand side term $\frac{1}{2} \Delta x \left| a \right| (1 - \lambda \left| a \right|) w_{xx}$ should be corrected as $\frac{\Delta x}{2} \left| a \right| w_{xx}$.

$P(A)$, where $P(x)$ is the Lagrangian interpolation polynomial such that $P(A_i) = \chi(A_i)$, $i = 1, \ldots, m$.

The linear system (B.20) can be regarded as a system of conservation laws (A.1) where the flux depends linearly on U:

$$F(U) = AU, \text{ for some } A \in \operatorname{Mat}(m; \mathbb{R}).$$

The upwind scheme (B.23) is in conservation form (B.1) where

$$G(U, V) := A_+ U + A_- V, \quad \forall U, V \in \mathbb{R}^m, \tag{B.25}$$

where A^+ and A^- are the positive and negative parts of A, defined by the mentioned functional calculus

$$A_+ = \chi_+(A), \quad A_- = \chi_-(A),$$

where $\chi_\pm(x) = x_\pm$, for all $x \in \mathbb{R}$.

Since

$$\chi_+(x) + \chi_-(x) = x, \quad \chi_+(x) - \chi_-(x) = |x|, \quad \forall x \in \mathbb{R},$$

we can write

$$A_+ = \frac{1}{2}(A + |A|), \quad A_- = \frac{1}{2}(A - |A|), \quad \forall A \in \operatorname{Mat}(m; \mathbb{R}).$$

Definition B.5 (Upwind scheme)**.** *A finite volume scheme in conservation form* (B.1) *is said to be an* upwind *(also,* upstream*) scheme if:*

i) For U and V nearby states, (B.25) *is a linear approximation to the numerical flux $G(U, V)$.*

ii) When all signal speeds associated with the numerical flux $G(U, V)$ are positive,

$$G(U, V) = F(U), \quad \forall U, V \in \mathbb{R}^m.$$

When all signal speeds are negative,

$$G(U, V) = F(V), \quad \forall U, V \in \mathbb{R}^m.$$

The first statement in Definition B.5 in analytic terms: Suppose U and V are near some reference state $U_\star$, then it is required that

$$\begin{aligned} G(U, V) &= G(U_\star, U_\star) + A_+(U_\star)(U - U_\star) + A_-(U_\star)(V - U_\star) \\ &\quad + o(\|U - U_\star\|_2 + \|V - U_\star\|_2). \\ &= F(U_\star) + A_+(U_\star)(U - U_\star) + A_-(U_\star)(V - U_\star) \\ &\quad + o(\|U - U_\star\|_2 + \|V - U_\star\|_2), \end{aligned} \tag{B.26}$$

where $A_+ (U_\star)$ and $A_- (U_\star)$ are entrywisely piecewise constant matrices, i.e.,

$$\partial_i A_+ (X) = \partial_i A_- (X) = 0 \text{ a.e. } X \in \mathbb{R}^m, \ \ \forall i \{1, \dots, m\} .$$

A natural choice for the reference state $U_\star$ is $\frac{1}{2} (U + V)$. Plugging this choice into (B.26) with the help of the identity $A_+ - A_- = |A|$ and noting that

$$F \left(\frac{1}{2} (U + V) \right) = \frac{1}{2} (F (U) + F (V)) + o (\|U - V\|_2) ,$$

yields

$$G (U, V) = \frac{1}{2} (F (U) + F (V)) - \frac{1}{2} \left| A \left(\frac{1}{2} (U + V) \right) \right| (V - U) + o (\|U - V\|_2) . \tag{B.27}$$

We can write any numerical flux in the following form

$$\begin{aligned} G (U, V) &= \frac{1}{2} (G (U, U) + G (V, V)) - \frac{1}{2} d (U, V) \\ &= \frac{1}{2} (F (U) + F (V)) - \frac{1}{2} d (U, V) , \ \ \forall U, V \in \mathbb{R}^m. \end{aligned} \tag{B.28}$$

For consistency, we need

$$d (U, U) = 0, \ \ \forall U \in \mathbb{R}^m.$$

The upwind condition (B.27) then can be expressed as

$$d (U, V) = \left| A \left(\frac{1}{2} (U + V) \right) \right| (V - U) + o (\|U - V\|_2) .$$

Formula (B.22) indicates that linear upwind schemes contain a large dose of artificial viscosity, except for those components where A_k is small, especially $A_k = 0$. This also holds for all upwind schemes for nonlinear conservation laws: when all characterisitc speeds are nonzero, each component is handled like a scheme with a massive amount of artificial viscosity, smudging discontinuities. The event that one of the characteristic speeds may be zero reveals in the way each scheme resolves a stationary shock, centered transonic rarefaction wave, and stationary contact discontinuity.

The most critical difference in performance occurs in resolving a stationary shock (A.42):

$$U_0 (x; U_L, U_R) = \begin{cases} U_L, & \text{if } x < 0, \\ U_R, & \text{if } x > 0, \end{cases} \ \ F (U_L) = F (U_R) , \ \ \Psi (U_R) < \Psi (U_L) . \tag{B.29}$$

The lack of numerical dissipation permits us to design schemes that perfectly resolve stationary shocks, i.e., (B.29) is a steady solution of the numerical scheme. The condition for that is

$$d (U_L, U_R) = 0 \text{ if } F (U_L) = F (U_R) , \ \ \Psi (U_R) < \Psi (U_L) .$$

Otherwise,

$$U_0\left(x; U_L, U_R\right) = \begin{cases} U_L, & \text{if } x < 0, \\ U_R, & \text{if } x > 0, \end{cases} \quad \Psi\left(U_R\right) > \Psi\left(U_L\right)$$

is not an admissible discontinuity and should not be a steady solution of the finite volume scheme (B.1), i.e., it is required that

$$d\left(U_L, U_R\right) \neq 0 \text{ if } F\left(U_L\right) = F\left(U_R\right), \;\; \Psi\left(U_R\right) > \Psi\left(U_L\right). \tag{B.30}$$

The hazard that an upwind scheme selects a nonphysical solution will occur only for stationary or near-stationary discontinuities. Alternatively, there is enough numerical viscosity in (B.19) to enforce the selection of a physically relevant solution.

Introduced in Harten, Lax, and Leer, 1983, p. 41, there are two options to design an upwind scheme for solving problems with discontinuous solution:

1. to switch direction of differencing in a way effectively introducing nonlinear dissipation at the expenditure of slight spread of the shock,
2. to fulfill (B.29)-(B.30) and thus obtain perfect resolution of a stationary shock, but to add a mechanism for examining the admissibility of the discontinuity.

Now various forms of $d\left(U, V\right)$ in (B.28). The most straighforward way to generate this distance is

$$d\left(U, V\right) := \left|\mathcal{A}\right|\left(U, V\right)\left(V - U\right), \;\; \forall U, V \in \mathbb{R}^m, \tag{B.31}$$

where $\left|\mathcal{A}\right|\left(U, V\right)$ is a matrix function of U and V that has nonnegative eigenvalues, and such that

$$\left|\mathcal{A}\right|\left(U, U\right) = \left|A\left(U\right)\right|, \;\; \forall U \in \mathbb{R}^m,$$

where $\left|A\left(U\right)\right|$ is defined by (B.24).

The simplest forms of $\left|\mathcal{A}\right|\left(U, V\right)$ are

$$\left|\mathcal{A}\right|\left(U, V\right) := \left|A\left(\frac{1}{2}\left(U + V\right)\right)\right|, \;\; \forall U, V \in \mathbb{R}^m,$$

or

$$\left|\mathcal{A}\right|\left(U, V\right) = \frac{1}{2}\left(\left|A\left(U\right)\right| + \left|A\left(V\right)\right|\right), \;\; \forall U, V \in \mathbb{R}^m. \tag{B.32}$$

In Leer, 1997, van Leer used (B.32) and introduced some nonlinear dissipation smearing stationary shocks but excluding nonphysical discontinuities.

Huang has suggested the following form of $\left|\mathcal{A}\right|\left(U, V\right)$ in his work Huang, 1981:

$$d\left(U, V\right) := \operatorname{sgn}\left(A\left(\frac{1}{2}\left(U + V\right)\right)\right)\left(F\left(V\right) - F\left(U\right)\right), \;\; \forall U, V \in \mathbb{R}^m,$$

where $\operatorname{sgn}\left(A\right)$ is defined by (B.24).

Roe has designed another scheme of the form (B.31), where the matrix function $\mathcal{A}(U,V)$ is required to have the following properties:

i) For all U, $V \in \mathbb{R}^m$,

$$F(V) - F(U) = \mathcal{A}(U,V)(V-U). \tag{B.33}$$

ii) $\mathcal{A}(U,V)$ has real eigenvalues and a complete set of eigenvectors.

iii) For all $U \in \mathbb{R}^m$,

$$\mathcal{A}(U,U) = A(U).$$

Theorem B.5. *Suppose the initial-value problem for hyperbolic systems of conservation laws* (A.1) *has an entropy function. Then* (A.1) *has a Roe-type linearization* (B.33).

Proof. See Harten, Lax, and Leer, 1983, pp. 42–43. □

A similar result also holds for system of conservation laws with multidimensional space variable provided there is an entropy. Having constructed $\mathcal{A}$, its absolute value $|\mathcal{A}|$ is defined by (B.24). Then we set

$$d(U,V) := |\mathcal{A}(U,V)|(V-U), \ \forall U,V \in \mathbb{R}^m.$$

When $U = U_L$ and $V = U_R$ correspond to a stationary discontinuity (B.29)-(B.30), then it is deduced from $F(U_L) - F(U_R) = 0$ and (B.33) that $U_R - U_L$ is a null vector of $\mathcal{A}(U_L,U_R)$, and consequently in the null space of $|\mathcal{A}(U_L,U_R)|$. Thus $d(U_L,U_R) = 0$, whether or not the entropy condition $\Psi(U_R) < \Psi(U_L)$ is fulfilled. Therefore the corresponding upwind finite scheme may admit nonphysical solutions.

Another way to construct $d(U,V)$ is

$$d(U,V) := \int_U^V |A(W)|\,dW,$$

where the integration is carried out on a path in state-space connecting U and V. See Harten, Lax, and Leer, 1983, p. 44 fore more details.

3 General Godunov-type schemes

3.1 Derivation of general Godunov-type schemes via Riemann's problems

Godunov has used the exact solutions of local Riemann's problems to obtain an upwind finite volume scheme.

The solution of the following Riemann's problem (without source terms)

$$\begin{cases} \partial_t U + \partial_x \left(F\left(U\right)\right) = 0, & \text{in } \mathbb{R}_+ \times \mathbb{R}, \\ U\left(0,x\right) = U_0\left(x; U_L, U_R\right) := \begin{cases} U_L, & \text{if } x < 0, \\ U_R, & \text{if } x > 0, \end{cases} & \text{in } \mathbb{R}, \end{cases} \tag{B.34}$$

depends only on the states U_L and U_R and the ratio $\frac{x}{t}$ and thus can be denoted by $U_{\mathcal{R}}\left(\frac{x}{t}; U_L, U_R\right)$. Because of the propagation with finite velocity of signals,

$$U_{\mathcal{R}}\left(\frac{x}{t}; U_L, U_R\right) = \begin{cases} U_L, & \text{if } \dfrac{x}{t} < \lambda_{\min}, \\ U_R, & \text{if } \dfrac{x}{t} \geq \lambda_{\max}, \end{cases} \tag{B.35}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest signal velocity, respectively.

To derive his scheme, given an admissible mesh $\mathcal{T}$ of $\mathbb{R}$, Godunov considered the numerical approximation $U_{\mathcal{T},\Delta t}\left(t^n, x\right)$ of the discrete time levels t^n, for all $n \in \mathbb{N}$, to be a piecewise constant function in x, i.e.,

$$U_{\mathcal{T},\Delta t}\left(t^n, x\right) := \sum_{i \in \mathbb{Z}} U_i^n \mathbf{1}_{\left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right)}\left(x\right), \quad \forall x \in \mathbb{R}.$$

To calculate the numerical approximation at the next time level $t^{n+1} = t^n + \Delta t$, we follow the following two steps.

Step 1. We solve exactly the initial-value problem

$$\begin{cases} \partial_t U + \partial_x \left(F\left(U\right)\right) = 0, & \text{in } \left[t^n, t^{n+1}\right] \times \mathbb{R}, \\ U\left(t^n, x\right) = U_{\mathcal{T},\Delta t}\left(t^n, x\right), & \text{in } \mathbb{R}, \end{cases} \tag{B.36}$$

for $t \in \left[t^n, t^n + \Delta t\right] \equiv \left[t^n, t^{n+1}\right]$, and denote its solution by $U^n_{\mathcal{T},\Delta t}\left(t, x\right)$. Each of discontinuities in $U_{\mathcal{T},\Delta t}\left(t^n, x\right)$ comprises locally a Riemann's problem. If the following CFL condition holds

$$\Delta t < \frac{\inf_{i \in \mathbb{Z}} \Delta x_i}{2 \max_{j \in \{1,\ldots,m\}} \left|\lambda_j\right|},$$

or stronger

$$\nu < \frac{\alpha_{\mathcal{T}}}{2 \max_{j \in \{1,\ldots,m\}} \left|\lambda_j\right|},$$

where $\max_i \left|\lambda_i\right|$ indicates the largest signal speed. Due to (B.35), there is no interaction between neighboring Riemann's problems, and $U^n_{\mathcal{T},\Delta t}\left(t, x\right)$ can be expressed exactly in terms of the solutions of local Riemann's problems[8]

$$U^n_{\mathcal{T},\Delta t}\left(t, x\right) := \sum_{i \in \mathbb{Z}} U_{\mathcal{R}}\left(\frac{x - x_{i+\frac{1}{2}}}{t - t^n}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right) \mathbf{1}_{\left(x_i, x_{i+1}\right]}\left(x\right), \text{ in } \left[t^n, t^{n+1}\right] \times \mathbb{R}. \tag{B.37}$$

[8] Here, I have corrected the formula of approximation solution $u_n\left(x, t\right)$ in Harten, Lax, and Leer, 1983, p. 45 in my notations.

Godunov obtained a piecewise constant approximation $U_{\mathcal{T},\Delta t}\left(t^{n+1},x\right)$ by averaging the quantity $U^n_{\mathcal{T},\Delta t}\left(t^{n+1},x\right)$ over each cells, i.e.,

$$U_i^{n+1} := \frac{1}{\Delta x_i}\int_{C_i} U^n_{\mathcal{T},\Delta t}\left(t^{n+1},x\right)dx, \quad \forall i \in \mathbb{Z}. \tag{B.38}$$

More explicitly, (B.38) can be rewritten in terms of the solutions of the local Riemann's problem as

$$U_i^{n+1} = \frac{1}{\Delta x_i}\int_0^{\Delta x_i/2} U_{\mathcal{R}}\left(\frac{x}{\Delta t}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)dx + \frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^{0} U_{\mathcal{R}}\left(\frac{x}{\Delta t}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)dx. \tag{B.39}$$

Next, if U is an exact solution of (B.34), integrating the first equation in (B.34) over an arbitrary rectangle $(t_1,t_2)\times(a,b)$ yields the following identity

$$\int_a^b U\left(t_2,x\right)dx - \int_a^b U\left(t_1,x\right)dx + \int_{t_1}^{t_2} F\left(U\left(t,b\right)\right)dt - \int_{t_1}^{t_2} F\left(U\left(t,a\right)\right)dt = 0,$$

for all $(t_1,t_2,a,b)\in\mathbb{R}^2_+\times\mathbb{R}^2$ with $t_1\leq t_2$, $a\leq b$.

Similarly, since $U^n_{\mathcal{T},\Delta t}$ is the exact solution of (B.36), the following equality holds:

$$\int_{x_{i-1/2}}^{x_{i+1/2}} U^n_{\mathcal{T},\Delta t}\left(t^{n+1},x\right)dx - \int_{x_{i-1/2}}^{x_{i+1/2}} U^n_{\mathcal{T},\Delta t}\left(t^{n},x\right)dx \\ + \int_{t^n}^{t^{n+1}} F\left(U^n_{\mathcal{T},\Delta t}\left(t, x_{i+\frac{1}{2}}\right)\right)dt - \int_{t^n}^{t^{n+1}} F\left(U^n_{\mathcal{T},\Delta t}\left(t, x_{i-\frac{1}{2}}\right)\right)dt = 0.$$

Plugging (B.37) into the last equality yields

$$\int_{x_{i-1/2}}^{x_i} U_{\mathcal{R}}\left(\frac{x-x_{i-\frac{1}{2}}}{\Delta t}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)dx + \int_{x_i}^{x_{i+1/2}} U_{\mathcal{R}}\left(\frac{x-x_{i+\frac{1}{2}}}{\Delta t}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)dx \\ - \int_{x_{i-1/2}}^{x_i} U_0\left(x-x_{i-\frac{1}{2}}; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)dx - \int_{x_i}^{x_{i+1/2}} U_0\left(x-x_{i+\frac{1}{2}}; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)dx \\ + \int_{t^n}^{t^{n+1}} F\left(U_{\mathcal{R}}\left(0; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)\right)dt - \int_{t^n}^{t^{n+1}} F\left(U_{\mathcal{R}}\left(0; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)\right)dt = 0.$$

Due to (B.39), the first two terms in the last equality are exactly $\Delta x_i U_i^{n+1}$. Thus, it can be rewritten as

$$\Delta x_i U_i^{n+1} = \frac{\Delta x_i}{2}U_i^n + \frac{\Delta x_i}{2}U_i^n - \Delta t F\left(U_{\mathcal{R}}\left(0; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right)\right) + \Delta t F\left(U_{\mathcal{R}}\left(0; U^n_{\mathrm{b},i-1}, U^n_{\mathrm{b},i}\right)\right),$$

which is equivalent to

$$U_i^{n+1} = U_i^n - \nu_i\left(F\left(\widetilde{U}^n_{i+\frac{1}{2}}\right) - F\left(\widetilde{U}^n_{i-\frac{1}{2}}\right)\right), \quad \forall i\in\mathbb{Z}, \tag{B.40}$$

where

$$\widetilde{U}^n_{i+\frac{1}{2}} := U_{\mathcal{R}}\left(0; U^n_{\mathrm{b},i}, U^n_{\mathrm{b},i+1}\right), \quad \forall i\in\mathbb{Z}.$$

This shows that (B.40) is in conservation form, with

$$G(U,V) := F\left(U_{\mathcal{R}}(0;U,V)\right), \quad \forall U,V \in \mathbb{R}^m. \tag{B.41}$$

3.2 Properties of general Godunov-type schemes

The exact solution $U^n_{\mathcal{T},\Delta t}(t,x)$ of the Riemann's problem satisfies the entropy condition (A.39). Thus, integrating the entropy inequality

$$\partial_t\left(\Phi\left(U^n_{\mathcal{T},\Delta t}\right)\right) + \partial_x\left(\Psi\left(U^n_{\mathcal{T},\Delta t}\right)\right) \leq 0, \text{ a.e. in } \left[t^n, t^{n+1}\right]\times\mathbb{R},$$

over $\left[t^n,t^{n+1}\right]\times C_i$ yields

$$\int_{C_i}\Phi\left(U^n_{\mathcal{T},\Delta t}\left(t^{n+1},x\right)\right)dx \leq \Delta x_i\Phi\left(U_i^n\right) - \Delta t\Psi\left(\widetilde{U}^n_{i+\frac{1}{2}}\right) + \Delta t\Psi\left(\widetilde{U}^n_{i-\frac{1}{2}}\right), \quad \forall i\in\mathbb{Z},$$

where we have used

$$U^n_{\mathcal{T},\Delta t}\left(t^n,x\right) = U_{\mathcal{T},\Delta t}\left(t^n,x\right) = U_i^n, \quad \forall x\in C_i, \quad \forall i\in\mathbb{Z}.$$

Since Φ is a convex function, applying Jensen's inequality for Φ yields for all function $V\in L^\infty\left(\mathbb{R}_+\times\mathbb{R},\mathbb{R}^m\right)$:

$$\Phi\left(\frac{1}{\Delta x_i}\int_{C_i}V(t,x)\,dx\right) \leq \frac{1}{\Delta x_i}\int_{C_i}\Phi\left(V(t,x)\right)dx, \quad \forall i\in\mathbb{Z}.$$

Combining the last two inequalities, with the help of (B.38), implies that

$$\Phi\left(U_i^{n+1}\right) \leq \Phi\left(U_i^n\right) - \nu_i\left(\Psi\left(\widetilde{U}_{i+\frac{1}{2}}\right) - \Psi\left(\widetilde{U}_{i-\frac{1}{2}}\right)\right), \quad \forall i\in\mathbb{Z},$$

i.e., Godunov's scheme is consistent with the entropy inequality (A.39) (see Definition B.2).

The description (B.37) makes sense only if the local Riemann problems do not interact, i.e., if

$$\Delta t < \frac{\inf_{i\in\mathbb{Z}}\Delta x_i}{2\max_{j\in\{1,\dots,m\}}|\lambda_j|}, \text{ or } \nu < \frac{\alpha_{\mathcal{T}}}{2\max_{j\in\{1,\dots,m\}}|\lambda_j|}.$$

On the other hand, (B.40) remains consistent with (B.38) as long as the waves emerged from $\left(i\pm\frac{1}{2}\right)\Delta x_i$ do not touch $\left(i\mp\frac{1}{2}\right)\Delta x_i$ during the time interval $\left[t^n,t^{n+1}\right]$. This will be the case as long as

$$\Delta t < \frac{\inf_{i\in\mathbb{Z}}\Delta x_i}{\max_{j\in\{1,\dots,m\}}|\lambda_j|}, \text{ or } \nu < \frac{\alpha_{\mathcal{T}}}{\max_{j\in\{1,\dots,m\}}|\lambda_j|}.$$

The Rankine-Hugoniot relation (A.7) implies that[9] the function

$$s \mapsto F\left(U_{\mathcal{R}}\left(s; U, V\right)\right) - sU_{\mathcal{R}}\left(s; U, V\right), \quad \forall s \in \mathbb{R},$$

is continuous but only piecewisely differentiable[10]. It follows that the Godunov numerical flux $G\left(U, V\right)$ defined by (B.41) is only piecewisely differentiable.

Godunov's finite volume scheme fulfills criterion ii) for upwind schemes (see Definition B.5) thanks to (B.35). To verify the first criterion of Definition B.5, it suffices to show that Godunov's finite volume scheme, when applied to linear equations, reduces to (B.23).

Consider (B.20), the solution of the Riemann problem is composed of constant states separated by a fan of m characteristic lines:

$$U_{\mathcal{R}}\left(\frac{x}{t}; U_L, U_R\right) = \begin{cases} U_0 := U_L, \text{ if } \dfrac{x}{t} < \lambda_1, \\ \vdots \\ U_k, \text{ if } \lambda_k < \dfrac{x}{t} < \lambda_{k+1}, \\ \vdots \\ U_m := U_R, \text{ if } \lambda_m < \dfrac{x}{t}. \end{cases} \tag{B.42}$$

The intermediate state U_k can be calculated from the representation of $U_R - U_L$ in terms of the right eigenvectors R_k of A in the following way

$$U_R - U_L = \sum_{i=1}^{m} J_i R_i,$$

$$U_k : = U_L + \sum_{i=1}^{k} J_i R_i, \quad \forall k \in \left\{1, \ldots, m\right\}. \tag{B.43}$$

By (B.24), (B.43) can be written as

$$U_k = U_L + \sigma_k\left(A\right)\left(U_R - U_L\right), \quad \forall k \in \left\{1, \ldots, m\right\}, \tag{B.44}$$

where σ_k is defined by

$$\sigma_k\left(x\right) = \begin{cases} 1, & \text{if } x < \lambda_k, \\ 0, & \text{if } x > \lambda_k. \end{cases}$$

Let k_0 be an integer such that $\lambda_{k_0} < 0 < \lambda_{k_0+1}$. Then by (B.44)

$$U_{\mathcal{R}}\left(0; U_L, U_R\right) = U_{k_0} = U_L + \sigma_{k_0}\left(A\right)\left(U_R - U_L\right) = \left(I - \sigma_{k_0}\left(A\right)\right)U_L + \sigma_{k_0}\left(A\right)U_R.$$

Substituting this into (B.41) gives

$$G\left(U, V\right) = A\left(I - \sigma_{k_0}\left(A\right)\right)U + A\sigma_{k_0}\left(A\right)V, \quad \forall U, V \in \mathbb{R}^m.$$

[9] Here, only the curves of the form $\frac{x}{t} = C$, i.e., straight lines through the origin (the speeds of these curves are constant) are considered, and thus (A.7) ensures that the function $s \mapsto F\left(U\left(s; U_L, U_R\right)\right) - sU\left(s; U_L, U_R\right)$ is continuous at these curves.

[10] This function is not differentiable at the mentioned lines.

From the functional calculus (B.24) and $A_+ = \chi_+(A)$, $A_- = \chi_-(A)$, we deduce that

$$A\left(I - \sigma_{k_0}(A)\right) = A_+, \quad A\sigma_{k_0}(A) = A_-,$$

so that $G(U,V)$ is in full agreement with (B.25). Therefore in the linear case, Godunov's finite volume scheme reduces to the upwind scheme (B.23).

The solution to the Riemann problem (B.34) has a moderate structure. For the constant coefficient case, the solution to (B.20) depends on $\frac{x}{t}$ and consists of constant states U_L, U_R, $U_1, \ldots, U_{m-1}$, separated by a fan of waves.

Unlike in the constant coefficient case, the k-wave separating U_{k-1} and U_k is not necessarily a single line having a characteristic speed λ_k. If the k^{th} characteristic field is genuinely nonlinear then the k-wave is either a rarefaction wave: $\lambda_k(U_{k-1}) < \lambda_k(U_k)$, or a shock propagating with speed S: $\lambda_k(U_{k-1}) > S > \lambda_k(U_k)$. If the k^{th} characteristic field is linearly degenerate then the k-wave is a contact discontinuity propagating with speed $\lambda_k(U_{k-1}) = \lambda_k(U_k)$.

It is indisputable from (B.38) that, due to averaging, the Godunov scheme does not use of all information contained in the exact solution of the Riemann problem. Thus, the exact solution $U_{\mathcal{R}}\left(\frac{x}{t}; U_L, U_R\right)$ of the Riemann problem in (B.39) by an approximation $V\left(\frac{x}{t}; U_L, U_R\right)$. This approximation V can have a much simpler structure provided it does not violate the essential properties of conservation and entropy inequality. The following theorem demonstrates that this type of approximation is consistent.

Theorem B.6 (Modified Harten-Lax for an admissible mesh). *Let $V\left(\frac{x}{t}; U_L, U_R\right)$ be an approximation to the solution of the Riemann problem that fulfills the following conditions:*

i) consistency with the integral form of the conservation law in the sense that[11]

$$\int_{-\Delta x_i/2}^{\Delta x_i/2} V\left(\frac{x}{\Delta t}; U_L, U_R\right) dx = \frac{\Delta x_i}{2}\left(U_L + U_R\right) - \Delta t\left(F\left(U_R\right) - F\left(U_L\right)\right), \tag{B.45}$$

for all $i \in \mathbb{Z}$ and for

$$\frac{\inf_{i\in\mathbb{Z}} \Delta x_i}{2} > \Delta t \max_{k\in\{1,\ldots,m\}} \left|\lambda_k\right|. \tag{B.46}$$

ii) consistency with the integral form of the entropy condition in the sense that

$$\int_{-\Delta x_i/2}^{\Delta x_i/2} \Phi\left(V\left(\frac{x}{t}; U_L, U_R\right)\right) dx \leq \frac{\Delta x_i}{2}\left(\Phi_L + \Phi_R\right) - \Delta t\left(\Psi\left(U_R\right) - \Psi\left(U_L\right)\right), \tag{B.47}$$

for all $i \in \mathbb{Z}$ and for (B.46).

[11] The integrand $w\left(\frac{x}{t}; u_L, u_R\right)$ in the formula numbered (3.7a) in Harten, Lax, and Leer, 1983 should be corrected as $w\left(\frac{x}{\Delta t}; u_L, u_R\right)$.

Using the approximation V to the Riemann problem, we can defined a Godunov-type *scheme as follows:*

$$U_i^{n+1} := \frac{1}{\Delta x_i}\int_0^{\Delta x_i/2} V\left(\frac{x}{t}; U_{\mathrm{b},i-1}^n, U_{\mathrm{b},i}^n\right) dx + \frac{1}{\Delta x_i}\int_{-\Delta x_i/2}^{0} V\left(\frac{x}{t}; U_{\mathrm{b},i}^n, U_{\mathrm{b},i+1}^n\right) dx. \tag{B.48}$$

If conditions (B.45) *and* (B.47) *are satisfied, the scheme* (B.48) *is in conservation form consistent with* (B.34)*, and satisfies the entropy inequality.*

Obviously, Godunov's scheme is of Godunov type.

Proposition B.8. *All schemes of Godunov type can be expressed in conservation form.*

Proof. Applying the integral conservation law in Section A.1.2. to the approximate solution V of the Riemann problem over the rectangle $\left(-\frac{\Delta x_i}{2}, 0\right) \times (0, \Delta t)$:

$$\int_{-\Delta x_i/2}^{0} V\left(\frac{x}{\Delta t}; U_L, U_R\right) dx - \int_{-\Delta x_i/2}^{0} V_0\left(x; U_L, U_R\right) dx \\ + \int_0^{\Delta t} F\left(V\left(0; U_L, U_R\right)\right) dt - \int_0^{\Delta t} F\left(V\left(-\frac{\Delta x_i}{2t}; U_L, U_R\right)\right) dt = 0,$$

i.e.,

$$\int_{-\Delta x_i/2}^{0} V\left(\frac{x}{\Delta t}; U_L, U_R\right) dx - \frac{\Delta x_i}{2} U_L + \Delta t\left(F\left(V\left(0; U_L, U_R\right)\right) - F\left(U_L\right)\right) = 0,$$

where we have used (B.46) to deduce $V\left(-\frac{\Delta x_i}{2\Delta t}; U_L, U_R\right) = U_L$. □

This gives

$$F\left(V\left(0; U_L, U_R\right)\right) = -\frac{1}{\Delta t}\int_{-\Delta x_i/2}^{0} V\left(\frac{x}{\Delta t}; U_L, U_R\right) dx + \frac{\Delta x_i}{2\Delta t} U_L + F\left(U_L\right), \quad \forall i \in \mathbb{Z}. \tag{B.49}$$

Similarly, applying the integral conservation law over the rectangle $\left(0, \frac{\Delta x_i}{\Delta t}\right)$, we obtain

$$F\left(V\left(0; U_L, U_R\right)\right) = \frac{1}{\Delta t}\int_0^{\Delta x_i/2} V\left(\frac{x}{\Delta t}; U_L, U_R\right) dx - \frac{\Delta x_i}{2\Delta t} U_R + F\left(U_R\right), \quad \forall i \in \mathbb{Z}.$$

The equality of the last two identities is exactly (B.45).

With the help of these identities, the Godunov-type scheme is settled in conservation form (B.1).

If all signal speeds are nonnegative, i.e. $\lambda_{\min} \geq 0$, then $V\left(s; U_L, U_R\right) = U_L$ for $s < 0$ and thus $F\left(V\left(0; U_L, U_R\right)\right) = F\left(U_L\right)$. Similarly, if all signal speeds are nonpositive, i.e., $\lambda_{\max} \leq 0$, then $V\left(s; U_L, U_R\right) = U_R$ for $s > 0$ and thus $F\left(V\left(0; U_L, U_R\right)\right) = F\left(U_R\right)$. This is the second property of upwind schemes stated in Definition B.5, which is thus satisfied by all Godunov-type schemes.

Moreover, the numerical flux of a Godunov-type scheme can be incorporated all physical insights being able to settle into the approximate solution of the Riemann problem. Furthermore, in Harten and Lax, 1981, the authors specified a Godunov-type scheme (B.48) can be used just as easily on a grid varying in time, by adjusting the intervals of integration on the right-hand side of (B.48).

3.3 Roe's approximate Riemann solver

Roe's approximate Riemann solver is based on a linearization of the type (B.33) also satisfying two properties followed it. Roe approximated solutions of the Riemann problem (B.34) by exact solutions of the Riemann problem for the following linear hyperbolic equation with constant coefficients:

$$\begin{cases} \partial_t V + \mathcal{A}\left(U_L, U_R\right) \partial_x V = 0, & \forall\left(t, x\right) \in \mathbb{R}_+ \times \mathbb{R}, \\ V\left(0, x\right) = V_0\left(x; U_L, U_R\right) := \begin{cases} U_L, & \text{if } x < 0, \\ U_R, & \text{if } x \geq 0, \end{cases} & \forall x \in \mathbb{R}. \end{cases}$$

Here $\mathcal{A}\left(U_L, U_R\right)$ is a matrix satisfying (B.33) and three properties listed there. Combining (B.33) and (B.43) yields

$$F\left(U_R\right) - F\left(U_L\right) = A\left(U_L, U_R\right)\left(U_R - U_L\right) = \sum_{i=1}^{m} \left(\lambda_i J_i R_i\right)\left(U_L, U_R\right),$$

where $\lambda_i\left(U_L, U_R\right)$ are the eigenvalues of $\mathcal{A}\left(U_L, U_R\right)$, $R_i\left(U_L, U_R\right)$ the corresponding right eigenvectors, and $J_i\left(U_L, U_R\right)$ the coefficients in the resolution (B.43), i.e.,

$$U_R - U_L = \sum_{i=1}^{m} \left(J_i R_i\right)\left(U_L, U_R\right).$$

The approximate Riemann solver is given by (B.42), with U_k defined by (B.43).

The numerical flux associated with an approximate Riemann solver is given by (B.49). Substituting (B.43) into (B.49) yields

$$F\left(W\left(0; U_L, U_R\right)\right) = F\left(U_L\right) + \sum_{i=1}^{m} \left(\left(\lambda_i\right)_- J_i R_i\right)\left(U_L, U_R\right),$$

and thus

$$\begin{aligned} F\left(W\left(0; U_L, U_R\right)\right) &= \frac{1}{2}\left(F\left(U_L\right) + F\left(U_R\right)\right) \\ &\quad - \frac{1}{2}\sum_{i=1}^{m} \left(\lambda_i J_i R_i\right)\left(U_L, U_R\right) + \sum_{i=1}^{m} \left(\left(\lambda_i\right)_- J_i R_i\right)\left(U_L, U_R\right) \\ &= \frac{1}{2}\left(F\left(U_L\right) + F\left(U_R\right)\right) - \frac{1}{2}\sum_{i=1}^{m} \left(\left|\lambda_i\right| J_i R_i\right)\left(U_L, U_R\right) \\ &= \frac{1}{2}\left(F\left(U_L\right) + F\left(U_R\right)\right) - \frac{1}{2}\left|A\left(U_L, U_R\right)\right|\left(U_R - U_L\right). \end{aligned}$$

By (B.41), this reads

$$G\left(U_L,U_R\right)=\frac{1}{2}\left(F\left(U_L\right)+F\left(U_R\right)\right)-\frac{1}{2}\left|A\left(U_L,U_R\right)\right|\left(U_R-U_L\right),$$

which is exactly Roe's scheme defined before (around (B.33)). As mentioned, Roe's scheme admits nonphysical, i.e., entropy-violating, stationary discontinuities. Harten and Hyman have modified Roe's scheme to eliminate such entropy-violating discontinuities but still entropy preserving.

Roe's approximate Riemann solver incorporate a great amount of detail: $m-1$ intermediate states.

Next, we consider some much simpler approximate Riemann solvers where these $m-1$ details are lumped together.

3.4 1-intermediate-state approximate Riemann solver

Denote by λ_L and λ_R lower and upper bounds, respectively, for $\lambda_{\min}$ and $\lambda_{\max}$, the approximate Riemann solver is then defined by

$$V_{\mathcal{R},1}\left(\frac{x}{t};U_L,U_R\right):=\begin{cases}U_L, & \text{if } \frac{x}{t}<\lambda_L,\\ U_{LR}, & \text{if } \lambda_L<\frac{x}{t}<\lambda_R,\\ U_R, & \text{if } \frac{x}{t}>\lambda_R,\end{cases}\tag{B.50}$$

where the state U_{LR} is determined from the conservation law (B.45):

$$\begin{aligned}\left(\frac{\Delta x_i}{2}+\Delta t\lambda_L\right)U_L+\Delta t\left(\lambda_R-\lambda_L\right)U_{LR}+\left(\frac{\Delta x_i}{2}-\Delta t\lambda_R\right)U_R&\\=\frac{\Delta x_i}{2}\left(U_L+U_R\right)-\Delta t\left(F\left(U_R\right)-F\left(U_L\right)\right)&,\end{aligned}$$

which gives us

$$U_{LR}=\frac{\lambda_R U_R-\lambda_L U_L}{\lambda_R-\lambda_L}-\frac{F\left(U_R\right)-F\left(U_L\right)}{\lambda_R-\lambda_L}.$$

Substituting $V_{\mathcal{R},1}$ into (B.49) yields that the associated numerical flux is given by

$$F\left(V_{\mathcal{R},1}\left(0;U_L,U_R\right)\right)=\begin{cases}F\left(U_L\right), & \text{if } 0<\lambda_L,\\ \dfrac{\lambda_R F\left(U_L\right)-\lambda_L F\left(U_R\right)}{\lambda_R-\lambda_L}+\dfrac{\lambda_R\lambda_L}{\lambda_R-\lambda_L}\left(U_R-U_L\right), & \text{if } \lambda_L<0<\lambda_R,\\ F\left(U_R\right), & \text{if } \lambda_R<0,\end{cases}$$

which can be combined into a single formula

$$\begin{aligned}F\left(V_{\mathcal{R},1}\left(0;U_L,U_R\right)\right)=&\frac{\left(\lambda_R\right)_{-}-\left(\lambda_L\right)_{-}}{\lambda_R-\lambda_L}F\left(U_R\right)+\frac{\left(\lambda_R\right)_{+}-\left(\lambda_L\right)_{+}}{\lambda_R-\lambda_L}F\left(U_L\right)\\&-\frac{1}{2}\frac{\lambda_R\left|\lambda_L\right|-\lambda_L\left|\lambda_R\right|}{\lambda_R-\lambda_L}\left(U_R-U_L\right).\end{aligned}\tag{B.51}$$

Since U_{LR} was chosen to fulfill the conservation law (B.45), U_{LR} is the mean value of the exact solution over the interval $(\lambda_L \Delta t, \lambda_R \Delta t)$. Therefore, by Jensen's inequality, (B.50) satisfies the entropy inequality (B.47).

Suppose that U_L and U_R can be connected with a shock of the first or the m^{th} family, in these cases the exact solution is given by

$$U(t,x) = \begin{cases} U_L, & \text{if } \dfrac{x}{t} < S, \\ U_R, & \text{if } \dfrac{x}{t} > S, \end{cases} \tag{B.52}$$

where S is the speed of propagation of the shock. Suppose $\lambda_L = S$ or $\lambda_R = S$, depending on whether the shock belongs to the first or the m^{th} family. Then (B.50) is the exact solution. Indeed, suppose that $\lambda_L = S$ (the case $\lambda_R = S$ is handled similarly), it suffices to prove that $U_{LR} = U_R$. Calculating

$$U_{LR} - U_R = \frac{F(U_R) - F(U_L) - \lambda_L (U_R - U_L)}{\lambda_R - \lambda_L} = \frac{F(U_R) - F(U_L) - S(U_R - U_L)}{\lambda_R - S} = 0,$$

and thus (B.50) is equal to (B.52).

3.5 2-intermediate-state approximate Riemann solver

We consider the class of approximate Riemann solvers whose U_L and U_R are linked through two intermediate states. These states are chosen to satisfy the following properties:

a) The conservation laws are satisfied.

b) if the exact solution of the Riemann problem links U_L and U_R through a single shock (or contact discontinuity) of any of the m families of waves, then so does the approximate Riemann solver.

c) The entropy inequality is satisfied.

Reuse the velocities λ_L and λ_R defined in the previous subsection. Define a velocity λ as follows. Let U be an entropy function, set

$$\lambda := \frac{(\nabla\Phi(U_R) - \nabla\Phi(U_L)) \cdot (F(U_R) - F(U_L))}{(\nabla\Phi(U_R) - \nabla\Phi(U_L)) \cdot (U_R - U_L)}.$$

The following lemma indicates the well-definedness of λ and lists some of its crucial properties.

Lemma B.1. *i) The denominator of λ is positive if $U_L \neq U_R$.*

ii) λ is uniformly bounded.

iii) If U_L and U_R satisfy the Rankine-Hugoniot condition

$$F(U_R) - F(U_L) = S(U_R - U_L), \tag{B.53}$$

then $\lambda = S$.

Now, we recalled from Harten, Lax, and Leer, 1983, pp. 51–54 two approximate Riemann solvers $V_{\mathcal{R},2}\left(\frac{x}{t}; U_L, U_R\right)$ with two intermediate states $U_L^\star$ and $U_R^\star$ separated by the line $\frac{x}{t} = \lambda$, i.e., $V_{\mathcal{R},2}$ is of the form

$$V_{\mathcal{R},2}\left(\frac{x}{t}; U_L, U_R\right) = \begin{cases} U_L, & \text{if } \dfrac{x}{t} < \lambda_L, \\ U_L^\star, & \text{if } \lambda_L < \dfrac{x}{t} < \lambda, \\ U_R^\star, & \text{if } \lambda < \dfrac{x}{t} < \lambda_R, \\ U_R, & \text{if } \lambda_R < \dfrac{x}{t}. \end{cases} \tag{B.54}$$

The flux across the line $x = st$ for equation (A.1) is defined as

$$F_s(U) := F(U) - sU, \quad \forall s \in \mathbb{R}, \quad \forall U \in \mathbb{R}^m.$$

A numerical flux, denoted by $G_\lambda(U_L, U_R)$, across the line $x = \lambda t$ is now introduced and required to be consistent with the exact flux, i.e.,

$$G_\lambda(U, U) = F_\lambda(U) = F(U) - \lambda U, \quad \forall U \in \mathbb{R}^m. \tag{B.55}$$

Approximate conservation laws for the triangle regions bounded by $t = \Delta t$, $x = \lambda t$, and $x = \lambda_L t$ or $x = \lambda_R t$, respectively:

$$(\lambda - \lambda_L) U_L^\star + G_\lambda(U_L, U_R) - F_{\lambda_L}(U_L) = 0, \tag{B.56}$$

$$(\lambda_R - \lambda) U_R^\star + F_{\lambda_R}(U_R) - G_\lambda(U_L, U_R) = 0, \tag{B.57}$$

which determines $U_L^\star$ and $U_R^\star$ as

$$U_L^\star = \frac{G_\lambda(U_L, U_R) - F_{\lambda_L}(U_L)}{\lambda_L - \lambda}, \quad U_R^\star = \frac{G_\lambda(U_L, U_R) - F_{\lambda_R}(U_R)}{\lambda_R - \lambda}.$$

Since (B.56)-(B.57) are conservation laws, the resulting scheme (B.54) satisfies the consistency relation (B.45). Thus requirement a) is fulfilled.

Next, the exact resolution of single shocks and contact discontinuities. A shock or contact discontinuity is characterized by the Rankine-Hugoniot condition (B.53) and the entropy condition (A.42). Using the notation of fluxes across lines, these can be rewritten as

$$F_S(U_L) = F_S(U_R), \quad \Psi_S(U_L) \geq \Psi_S(U_R), \tag{B.58}$$

where we have used Lemma B.1 iii) and

$$\Psi_S(U) = \Psi(U) - \lambda S, \quad \forall U \in \mathbb{R}^m.$$

Denote by $\widetilde{G}_\lambda$ a vector-valued function satisfying:

$$F_\lambda(U_L) = F_\lambda(U_R) \Rightarrow \widetilde{G}_\lambda(U_L, U_R) = F_\lambda(U_L) = F_\lambda(U_R). \tag{B.59}$$

Set

$$G_\lambda\left(U_L,U_R\right):=\widetilde{G}_\lambda\left(U_L,U_R\right)-\beta\left(U_L,U_R\right)\left(U_R-U_L\right), \tag{B.60}$$

where β has the following property:

$$\text{(B.58) holds } \Rightarrow \beta\left(U_L,U_R\right)=0.$$

Now the numerical flux is taken by (B.60). It follows from (B.59) that this numerical flux is consistent with the exact flux, i.e., (B.55) holds. Moreover, thanks to (B.58), (B.60) and Lemma B.1 iii), single shocks and contact discontinuities are resolved exactly.

We now choose $\widetilde{G}_\lambda$ and β. Define

$$\delta_R:=\frac{\lambda_R-\lambda}{\lambda_R-\lambda_L},\ \ \delta_L:=\frac{\lambda-\lambda_L}{\lambda_R-\lambda_L},$$

then

$$\delta_L\geq 0,\ \ \delta_R\geq 0,\ \ \delta_L+\delta_R=1.$$

Then we define

$$\begin{aligned}\widetilde{G}^a_\lambda\left(U_L,U_R\right):&=\delta_R F_\lambda\left(U_L\right)+\delta_L F_\lambda\left(U_R\right),\\ \widetilde{G}^b_\lambda\left(U_L,U_R\right):&=F_\lambda\left(U_{LR}\right)-F\left(\delta_L U_L+\delta_R U_R\right)+\delta_L F\left(U_L\right)+\delta_R F\left(U_R\right).\end{aligned}$$

Claim: G^a_λ satisfies (B.59).

Indeed, assume $F_\lambda\left(U_L\right)=F_\lambda\left(U_R\right)$, then

$$G^a_\lambda\left(U_L,U_R\right)=\frac{\lambda_R-\lambda}{\lambda_R-\lambda_L}F_\lambda\left(U_L\right)+\frac{\lambda-\lambda_L}{\lambda_R-\lambda_L}F_\lambda\left(U_L\right)-\beta\left(U_L,U_R\right)\left(U_R-U_L\right)=F_\lambda\left(U_L\right).$$

Note that

$$\begin{aligned}U_{LR}&=\frac{\lambda_R U_R-\lambda_L U_L-F\left(U_R\right)+F\left(U_L\right)}{\lambda_R-\lambda_L}=\frac{\lambda_R U_R-\lambda_L U_L-\lambda\left(U_R-U_L\right)}{\lambda_R-\lambda_L}\\ &=\frac{\lambda-\lambda_L}{\lambda_R-\lambda_L}U_L+\frac{\lambda_R-\lambda}{\lambda_R-\lambda_L}U_R=\delta_L U_L+\delta_R U_R,\end{aligned}$$

we then have

$$\begin{aligned}\widetilde{G}^b_\lambda\left(U_L,U_R\right)&=F_\lambda\left(\delta_L U_L+\delta_R U_R\right)-F\left(\delta_L U_L+\delta_R U_R\right)+\delta_L F\left(U_L\right)+\delta_R F\left(U_R\right)\\ &=-\lambda\left(\delta_L U_L+\delta_R U_R\right)+\delta_L F\left(U_L\right)+\delta_R F\left(U_R\right)\\ &=\delta_R F_\lambda\left(U_L\right)+\delta_L F_\lambda\left(U_R\right)=G^a_\lambda\left(U_L,U_R\right)=F_\lambda\left(U_L\right).\end{aligned}$$

We define β as follows:

$$\beta:=C_1\beta_1+C_2\beta_2,$$

where

$$\beta_1(U_L,U_R) := \frac{\left[\Psi_\lambda(U_R) - \Psi_\lambda(U_L) - \frac{1}{2}(U_L+U_R)\cdot(F_\lambda(U_R)-F_\lambda(U_L))\right]_+}{\|U_R-U_L\|_2^2},$$

$$\beta_2(U_L,U_R) := \frac{\|F_\lambda(U_R)-F_\lambda(U_L)\|_2^2}{(\lambda_R-\lambda_L)\|U_R-U_L\|_2^2},$$

then β_1 is a bounded function. By construction, $\beta_1 = 0$ when the shock condition (B.58) is satisfied and $\beta_2 = 0$ when the Rankine-Hugoniot condition (B.53) alone is satisfied. Thus, β fulfills b). Additionally, the positive constants C_1 and C_2 in definition of β can be chosen in order that the entropy condition is satisfied.

Substituting (B.54) into (B.49) yields the numerical flux associated with the above Godunov-type scheme as

$$F(V_{\mathcal{R},2}(0;U_L,U_R)) = \frac{1}{2}\left[\begin{array}{l} F(U_L)+F(U_R)+\gamma_L(F_\lambda(U_L)-G_\lambda(U_L,U_R)) \\ +\gamma_R(G_\lambda(U_L,U_R)-F_\lambda(U_R))-|\lambda|(U_R-U_L) \end{array}\right], \tag{B.61}$$

where

$$\gamma_L := \frac{|\lambda|-|\lambda_L|}{\lambda-\lambda_L}, \quad \gamma_R := \frac{|\lambda_R|-|\lambda|}{\lambda_R-\lambda}.$$

Claim: *If $\lambda_L \geq 0$, then $F(V_{R,2}(0;U_L,U_R)) = F(U_L)$.*

Indeed, assume $\lambda_L \geq 0$, then $\lambda \geq 0$, $\lambda_R \geq 0$ and thus $\gamma_L = \gamma_R = 1$. Plugging these into (B.61) yields

$$F(V_{\mathcal{R},2}(0;U_L,U_R)) = \frac{1}{2}\left[F(U_L)+F(U_R)+F_\lambda(U_L)-F_\lambda(U_R)-\lambda(U_R-U_L)\right] = F(U_L).$$

Claim: *If $\lambda_R \leq 0$, then then $F(V_{R,2}(0;U_L,U_R)) = F(U_R)$.*

Indeed, assume $\lambda_R \leq 0$, then $\gamma_L = \gamma_R = 1$ and thus

$$F(V_{\mathcal{R},2}(0;U_L,U_R)) = \frac{1}{2}\left[F(U_L)+F(U_R)-F_\lambda(U_L)+F_\lambda(U_R)+\lambda(U_R-U_L)\right] = F(U_R).$$

Claim: *If $\lambda = 0$, then $F(V_{\mathcal{R},2}(0;U_L,U_R)) = F_\lambda(U_L,U_R)$.*

Indeed, assume $\lambda = 0$, we have $\lambda_L \leq 0$, $\lambda_R \geq 0$. So $\gamma_L = -1$, $\gamma_R = 1$ and thus

$$\begin{aligned} F(V_{R,2}(0;U_L,U_R)) &= \frac{1}{2}\left[\begin{array}{l} F(U_L)+F(U_R)-(F_0(U_L)-G_0(U_L,U_R)) \\ +(G_0(U_L,U_R)-F_0(U_R)) \end{array}\right] \\ &= \frac{1}{2}\left[\begin{array}{l} F(U_L)+F(U_R)-(F(U_L)-G_0(U_L,U_R)) \\ +(G_0(U_L,U_R)-F(U_R)) \end{array}\right] = G_0(U_L,U_R). \end{aligned}$$

Substituting (B.60) for $F_\lambda(U_L,U_R)$ in (B.61) yields

$$F(V_{\mathcal{R},2}(0;U_L,U_R)) = \frac{1}{2}\left[\begin{array}{l} F(U_L)+F(U_R)+\gamma_L F_\lambda(U_L)-\gamma_R F_\lambda(U_R)+(\gamma_R-\gamma_L)\widetilde{G}_\lambda(U_L,U_R) \\ -(\gamma_R-\gamma_L)\beta(U_L,U_R)(U_R-U_L)-|\lambda|(U_R-U_L) \end{array}\right].$$

By definitions of γ_L, γ_R,

$$\frac{1}{2}\left(\gamma_R - \gamma_L\right) = \begin{cases} 0, & \text{if } \lambda_L > 0 \text{ or } \lambda_R < 0, \\ \dfrac{|\lambda_L|}{|\lambda_L| + |\lambda|}, & \text{if } \lambda_L < 0 < \lambda < \lambda_R, \\ \dfrac{|\lambda_R|}{|\lambda_R| + |\lambda|}, & \text{if } \lambda_L < \lambda < 0 < \lambda_R, \end{cases}$$

a nonnegative quantity. This indicates that β enters the finite volume scheme as an artificial viscosity. Unlike classical artificial viscosity, this β is zero across a shock and is positive across an incipient rarefaction wave. The scheme (B.61) is not upwind. The decrease of entropy guarantees the L^2 stability of the scheme. Schemes of type (B.61) are especially suitable for computation on a *moving mesh*, i.e., each meshpoint moves with velocity λ.

Finally, any scheme in conservation form (B.1) with a numerical flux $G\left(U, V\right)$ that yields perfect resolution of discontinuities but also admits entropy-violating ones may be corrected by modifying its numerical flux as

$$\widetilde{G}\left(U, V\right) := G\left(U, V\right) - C_1 \beta\left(U, V\right)\left(V - U\right), \quad \forall U, V \in \mathbb{R}^m.$$

Alphabetical Index

Bibliography

Audusse, Emmanuel et al. (2004). “A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows”. In: *SIAM J. Sci. Comput.* 25.6, pp. 2050–2065. ISSN: 1064-8275. DOI: 10.1137/S1064827503431090. URL: https://doi.org/10.1137/S1064827503431090.

Benzoni-Gavage, Sylvie and Denis Serre (2007). *Multidimensional hyperbolic partial differential equations.* Oxford Mathematical Monographs. First-order systems and applications. The Clarendon Press, Oxford University Press, Oxford, pp. xxvi+508. ISBN: 978-0-19-921123-4; 0-19-921123-X.

Bermudez, Alfredo and Ma. Elena Vazquez (1994). “Upwind methods for hyperbolic conservation laws with source terms”. In: *Comput. & Fluids* 23.8, pp. 1049–1071. ISSN: 0045-7930. DOI: 10.1016/0045-7930(94)90004-3. URL: https://doi.org/10.1016/0045-7930(94)90004-3.

Berthon, Christophe (2006). “Numerical approximations of the 10-moment Gaussian closure”. In: *Math. Comp.* 75.256, pp. 1809–1831. ISSN: 0025-5718. DOI: 10.1090/S0025-5718-06-01860-6. URL: https://doi.org/10.1090/S0025-5718-06-01860-6.

Berthon, Christophe and Christophe Chalons (2016). “A fully well-balanced, positive and entropy-satisfying Godunov-type method for the shallow-water equations”. In: *Math. Comp.* 85.299, pp. 1281–1307. ISSN: 0025-5718. DOI: 10.1090/mcom3045. URL: https://doi.org/10.1090/mcom3045.

Berthon, Christophe and Françoise Foucher (2012). “Efficient well-balanced hydrostatic upwind schemes for shallow-water equations”. In: *J. Comput. Phys.* 231.15, pp. 4993–5015. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2012.02.031. URL: https://doi.org/10.1016/j.jcp.2012.02.031.

Bouchut, François (2004). *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources.* Frontiers in Mathematics. Birkhäuser Verlag, Basel, pp. viii+135. ISBN: 3-7643-6665-6. DOI: 10.1007/b93802. URL: https://doi.org/10.1007/b93802.

Bouchut, François, Christian Klingenberg, and Knut Waagan (2007). “A multiwave approximate Riemann solver for ideal MHD based on relaxation. I. Theoretical framework”. In: *Numer. Math.* 108.1, pp. 7–42. ISSN: 0029-599X. DOI: 10.1007/s00211-007-0108-8. URL: https://doi.org/10.1007/s00211-007-0108-8.

Bouchut, François and Tomás Morales de Luna (2010). “A subsonic-well-balanced reconstruction scheme for shallow water flows”. In: *SIAM J. Numer. Anal.* 48.5, pp. 1733–1758. ISSN: 0036-1429. DOI: 10.1137/090758416. URL: https://doi.org/10.1137/090758416.

Bryson, Steve et al. (2011). "Well-balanced positivity preserving central-upwind scheme on triangular grids for the Saint-Venant system". In: *ESAIM Math. Model. Numer. Anal.* 45.3, pp. 423–446. ISSN: 0764-583X. DOI: 10.1051/m2an/2010060. URL: https://doi.org/10.1051/m2an/2010060.

Buffard, Thierry, Thierry Gallouet, and Jean-Marc Hérard (1998). "Un schéma simple pour les équations de Saint-Venant". In: *C. R. Acad. Sci. Paris Sér. I Math.* 326.3, pp. 385–390. ISSN: 0764-4442. DOI: 10.1016/S0764-4442(97)83000-5. URL: https://doi.org/10.1016/S0764-4442(97)83000-5.

Castro, Manuel et al. (2008). "Well-balanced high order extensions of Godunov's method for semilinear balance laws". In: *SIAM J. Numer. Anal.* 46.2, pp. 1012–1039. ISSN: 0036-1429. DOI: 10.1137/060674879. URL: https://doi.org/10.1137/060674879.

Castro, Manuel J., Alberto Pardo Milanés, and Carlos Parés (2007). "Well-balanced numerical schemes based on a generalized hydrostatic reconstruction technique". In: *Math. Models Methods Appl. Sci.* 17.12, pp. 2055–2113. ISSN: 0218-2025. DOI: 10.1142/S021820250700256X. URL: https://doi.org/10.1142/S021820250700256X.

Chalons, Christophe and Jean-François Coulombel (2008). "Relaxation approximation of the Euler equations". In: *J. Math. Anal. Appl.* 348.2, pp. 872–893. ISSN: 0022-247X. DOI: 10.1016/j.jmaa.2008.07.034. URL: https://doi.org/10.1016/j.jmaa.2008.07.034.

Chalons, Christophe et al. (2010). "Godunov-type schemes for hyperbolic systems with parameter-dependent source. The case of Euler system with friction". In: *Math. Models Methods Appl. Sci.* 20.11, pp. 2109–2166. ISSN: 0218-2025. DOI: 10.1142/S021820251000488X. URL: https://doi.org/10.1142/S021820251000488X.

Chen, Guoxian and Sebastian Noelle (2017). "A new hydrostatic reconstruction scheme based on subcell reconstructions". In: *SIAM J. Numer. Anal.* 55.2, pp. 758–784. ISSN: 0036-1429. DOI: 10.1137/15M1053074. URL: https://doi.org/10.1137/15M1053074.

Craig, W. and C. Sulem (1993). "Numerical simulation of gravity waves". In: *J. Comput. Phys.* 108.1, pp. 73–83. ISSN: 0021-9991. DOI: 10.1006/jcph.1993.1164. URL: https://doi.org/10.1006/jcph.1993.1164.

Craig, W., C. Sulem, and P.-L. Sulem (1992). "Nonlinear modulation of gravity waves: a rigorous approach". In: *Nonlinearity* 5.2, pp. 497–522. ISSN: 0951-7715. URL: http://stacks.iop.org/0951-7715/5/497.

Doyen, D. and P. H. Gunawan (2014). "An explicit staggered finite volume scheme for the shallow water equations". In: *Finite volumes for complex applications VII. Methods and theoretical aspects*. Vol. 77. Springer Proc. Math. Stat. Springer, Cham, pp. 227–235. DOI: 10.1007/978-3-319-05684-5_21. URL: https://doi.org/10.1007/978-3-319-05684-5_21.

Duchêne, Vincent (2019). *Shallow-water models for water waves*. URL: https://perso.univ-rennes1.fr/vincent.duchene/CoursM2.pdf.

Evans, Lawrence C. (2010). *Partial differential equations*. Second. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, pp. xxii+749. ISBN: 978-0-8218-4974-3. DOI: 10.1090/gsm/019. URL: https://doi.org/10.1090/gsm/019.

Eymard, Robert, Thierry Gallouët, and Raphaèle Herbin (2003). *Finite volume methods*, p. 253. URL: https://old.i2m.univ-amu.fr/~herbin/PUBLI/bookevol.pdf.

Godlewski, Edwige and Pierre-Arnaud Raviart (1996). *Numerical approximation of hyperbolic systems of conservation laws*. Vol. 118. Applied Mathematical Sciences. Springer-Verlag, New York, pp. viii+509. ISBN: 0-387-94529-6. DOI: 10.1007/978-1-4612-0713-9. URL: https://doi.org/10.1007/978-1-4612-0713-9.

Gosse, L. (2000). "A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms". In: *Comput. Math. Appl.* 39.9-10, pp. 135–159. ISSN: 0898-1221. DOI: 10.1016/S0898-1221(00)00093-6. URL: https://doi.org/10.1016/S0898-1221(00)00093-6.

Goutal, N. (2002). "A finite volume solver for 1D shallow-water equations applied to an actual river". In: *Numer. Meth. Fluids* 38, pp. 1–19. DOI: 10.1002/fld.201. URL: https://doi.org/10.1002/fld.201.

Greenberg, J. M. and A. Y. Leroux (1996). "A well-balanced scheme for the numerical processing of source terms in hyperbolic equations". In: *SIAM J. Numer. Anal.* 33.1, pp. 1–16. ISSN: 0036-1429. DOI: 10.1137/0733001. URL: https://doi.org/10.1137/0733001.

Harten, Amiram and Peter D. Lax (1981). "A random choice finite difference scheme for hyperbolic conservation laws". In: *SIAM J. Numer. Anal.* 18.2, pp. 289–315. ISSN: 0036-1429. DOI: 10.1137/0718021. URL: https://doi.org/10.1137/0718021.

Harten, Amiram, Peter D. Lax, and Bram van Leer (1983). "On upstream differencing and Godunov-type schemes for hyperbolic conservation laws". In: *SIAM Rev.* 25.1, pp. 35–61. ISSN: 0036-1445. DOI: 10.1137/1025002. URL: https://doi.org/10.1137/1025002.

Herbin, R., W. Kheriji, and J.-C. Latché (2014). "On some implicit and semi-implicit staggered schemes for the shallow water and Euler equations". In: *ESAIM Math. Model. Numer. Anal.* 48.6, pp. 1807–1857. ISSN: 0764-583X. DOI: 10.1051/m2an/2014021. URL: https://doi.org/10.1051/m2an/2014021.

Herbin, R., J.-C. Latché, and T. T. Nguyen (2013). "Explicit staggered schemes for the compressible Euler equations". In: *Applied mathematics in Savoie—AMIS 2012: Multiphase flow in industrial and environmental engineering*. Vol. 40. ESAIM Proc. EDP Sci., Les Ulis, pp. 83–102. DOI: 10.1051/proc/201340006. URL: https://doi.org/10.1051/proc/201340006.

Huang, Lan Chieh (1981). "Pseudo-unsteady difference schemes for discontinuous solutions of steady-state, one-dimensional fluid dynamics problems". In: *J. Comput. Phys.* 42.1, pp. 195–211. ISSN: 0021-9991. DOI: 10.1016/0021-9991(81)90239-4. URL: https://doi.org/10.1016/0021-9991(81)90239-4.

Jin, Shi (2001). "A steady-state capturing method for hyperbolic systems with geometrical source terms". In: *M2AN Math. Model. Numer. Anal.* 35.4, pp. 631–645. ISSN: 0764-583X. DOI: 10.1051/m2an:2001130. URL: https://doi.org/10.1051/m2an:2001130.

Lannes, David (2013). *The water waves problem.* Vol. 188. Mathematical Surveys and Monographs. Mathematical analysis and asymptotics. American Mathematical Society, Providence, RI, pp. xx+321. ISBN: 978-0-8218-9470-5. DOI: 10.1090/surv/188. URL: https://doi.org/10.1090/surv/188.

Leer, Bram van (1997). "Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method [J. Comput. Phys. **32** (1979), no. 1, 101–136]". In: *J. Comput. Phys.* 135.2. With an introduction by Ch. Hirsch, Commemoration of the 30th anniversary {of J. Comput. Phys.}, pp. 227–248. ISSN: 0021-9991. DOI: 10.1006/jcph.1997.5757. URL: https://doi.org/10.1006/jcph.1997.5757.

LeFloch, Philippe G. and Mai Duc Thanh (2007). "The Riemann problem for the shallow water equations with discontinuous topography". In: *Commun. Math. Sci.* 5.4, pp. 865–885. ISSN: 1539-6746. URL: http://projecteuclid.org/euclid.cms/1199377555.

Liang, Q. and F. Marche (2009). "Numerical resolution of well-balanced shallow water equations with complex source terms". In: *Advances in Water Resources* 32.6, pp. 873–884. DOI: 10.1016/j.advwatres.2009.02.010. URL: https://doi.org/10.1016/j.advwatres.2009.02.010.

Luna, T. Morales de et al. (2009). "On a shallow water model for the simulation of turbidity currents". In: *Commun. Comput. Phys.* 6.4, pp. 848–882. ISSN: 1815-2406. DOI: 10.4208/cicp.2009.v6.p848. URL: https://doi.org/10.4208/cicp.2009.v6.p848.

Noelle, Sebastian, Yulong Xing, and Chi-Wang Shu (2007). "High-order well-balanced finite volume WENO schemes for shallow water equation with moving water". In: *J. Comput. Phys.* 226.1, pp. 29–58. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2007.03.031. URL: https://doi.org/10.1016/j.jcp.2007.03.031.

Noelle, Sebastian et al. (2006). "Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows". In: *J. Comput. Phys.* 213.2, pp. 474–499. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2005.08.019. URL: https://doi.org/10.1016/j.jcp.2005.08.019.

Parés, Carlos and Manuel Castro (2004). "On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow-water systems". In: *M2AN Math. Model. Numer. Anal.* 38.5, pp. 821–852. ISSN: 0764-583X. DOI: 10.1051/m2an:2004041. URL: https://doi.org/10.1051/m2an:2004041.

Russo, G. and A. Khe (2010). "High order well-balanced schemes based on numerical reconstruction of the equilibrium variables". In: *Proceedings "WASCOM 2009" 15th Conference on Waves and Stability in Continuous Media.* World Sci. Publ., Hackensack, NJ, pp. 230–241. DOI: 10.1142/9789814317429_0032. URL: https://doi.org/10.1142/9789814317429_0032.

Russo, Giovanni and Alexander Khe (2009). "High order well balanced schemes for systems of balance laws". In: *Hyperbolic problems: theory, numerics and applications.* Vol. 67. Proc. Sympos. Appl. Math. Amer. Math. Soc., Providence, RI, pp. 919–928. DOI: 10.1090/psapm/067.2/2605287. URL: https://doi.org/10.1090/psapm/067.2/2605287.

Stelling, G. S. and S. P. A. Duinmeijer (2003). "A staggered conservative scheme for every Froude number in rapidly varied shallow water flows". In: *Internat. J. Numer.*

Methods Fluids 43.12, pp. 1329–1354. ISSN: 0271-2091. DOI: 10.1002/fld.537. URL: https://doi.org/10.1002/fld.537.

Xing, Yulong (2014). “Exactly well-balanced discontinuous Galerkin methods for the shallow water equations with moving water equilibrium”. In: *J. Comput. Phys.* 257.part A, pp. 536–553. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2013.10.010. URL: https://doi.org/10.1016/j.jcp.2013.10.010.

Xing, Yulong and Chi-Wang Shu (2006). “High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms”. In: *J. Comput. Phys.* 214.2, pp. 567–598. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2005.10.005. URL: https://doi.org/10.1016/j.jcp.2005.10.005.

Xing, Yulong, Chi-Wang Shu, and Sebastian Noelle (2011). “On the advantage of well-balanced schemes for moving-water equilibria of the shallow water equations”. In: *J. Sci. Comput.* 48.1-3, pp. 339–349. ISSN: 0885-7474. DOI: 10.1007/s10915-010-9377-y. URL: https://doi.org/10.1007/s10915-010-9377-y.

Xing, Yulong, X. Zhung, and Chi-Wang Shu (2010). “Positivity-preserving high order well-balanced discontinuous Galerkin methods for the shallow water equations”. In: *Adv. Water Resour* 33, pp. 1476–1493. ISSN: 0885-7474.