Implicit kinetic schemes for the Saint-Venant system

Chourouk El Hassanieh, Mathieu Rigal, Jacques Sainte-Marie

Implicit kinetic schemes for the Saint-Venant system

Chourouk El Hassanieh, Mathieu Rigal, Jacques Sainte-Marie

March 28, 2023

Abstract

Explicit kinetic schemes applied to the nonlinear shallow water equations have been extensively studied in the past. The novelty of this paper is to investigate an implicit version of such methods in order to improve their stability properties. In the case of a flat bathymetry we obtain a fully implicit kinetic solver satisfying a discrete entropy inequality and keeping the water height non negative without any restriction on the time step. Remarkably, a simplified version of this nonlinear implicit scheme allows to express the update explicitly which we implement in practice. An extension to the 2D case is also discussed. The case of varying bottoms is then dealt with through an iterative solver combined with the hydrostatic reconstruction technique. We show that this scheme preserves the water height positivity under a CFL condition and satisfies a discrete entropy inequality without error term, which is an improvement over its explicit version. Finally we perform some numerical validations underlying the advantages and the computational cost of our strategy.

Keywords: Shallow water equations, well-balanced schemes, hydrostatic reconstruction, kinetic solver, fully discrete entropy inequality

AMS classification 65M12, 74S10, 76M12, 35L65

Contents

1 Introduction

Mathematical models for free surface flows are widely studied but their analysis and numerical approximation remain a challenging issue. The incompressible Navier-Stokes system with free surface being very difficult to study, it is often replaced by the classical Saint-Venant system [22, 14] that is a hyperbolic system of conservation laws approximating various geophysical flows, such as rivers, coastal areas, and oceans when completed with a Coriolis term, and granular flows when completed with friction terms.

The derivation of an efficient, robust and stable numerical scheme for the Saint-Venant system has received an extensive coverage, we refer the reader to [8, 17, 15, 23] and references therein. One of the challenges involves the construction of a well-balanced scheme i.e. preserving some characteristic stationary solutions. In a recent work, some of the authors have proposed a numerical scheme for the Saint-Venant system with topography (1) satisfying a fully discrete entropy inequality [3]. The proposed numerical scheme is based on a kinetic solver [6, 20, 7, 4, 16, 11] coupled with the hydrostatic reconstruction technique introduced in [2] for the numerical treatment of the topography source term. Based on the results obtained in [3], Bouchut and Lhebrard have proved the convergence of the the kinetic hydrostatic reconstruction scheme for the Saint-Venant system (1), see [9].

Finite volume approaches for the approximation of conservation laws have to deal with a CFL constraint that can be very restrictive for some applications where large time scales and significant wave velocities have to be considered. This is for instance the case in the low Froude regime, where the surface gravity waves travel at a much larger velocity than the fluid particles. Moreover, the explicit in time discretization induces a non-negative term in the discrete energy balance that cannot be always controlled by the dissipation coming from the upwinding of the numerical fluxes, see [3]. The novelties of this paper are:

- to propose a fully implicit kinetic scheme in 1d and 2d for the Saint-Venant system satisfying a fully discrete entropy inequality without any restriction on the time step,
- to evaluate the practical interest of implicit schemes for the Saint-Venant system in the sense that the CFL constraint for explicit schemes is replaced in the context of implicit schemes by the computational costs due to the computation of the numerical fluxes. Indeed in the explicit setting, the numerical fluxes at an interface depend on the value of the variables at the two neighbouring vertices whereas in the implicit context, the numerical fluxes depend on the value of the variables at all vertices (the stencil encompasses the whole computational domain).

Notice that an implicit scheme often requires to invert an operator – here a matrix – at each time step. However, the kinetic scheme gives a very favorable context in which we have an explicit expression of the inverse of the operator. Hence, one can hardly imagine a truly implicit scheme for the Saint-Venant system with a lower computational cost than a kinetic solver.

The aim of this paper is to propose an implicit – in time – version of the kinetic scheme given in [3] and to study its properties. More precisely, we prove some stability properties and most importantly, we derive a fully discrete entropy inequality without any error term.

This paper is organized as follows. First, we recall the formulation of the Saint-Venant system, its kinetic description and the framework of its numerical approximation in the context of a kinetic solver. Then, the implicit kinetic scheme for the Saint-Venant system without topography is proposed and studied in 1d and in the 2d case. An iterative version of the implicit scheme is proposed in Section 5 where the topography can be taken into account through the hydrostatic reconstruction technique [4]. Finally, numerical examples are given to evaluate the interest of our approach.

2 The Saint-Venant system and its kinetic interpretation

2.1 The Saint-Venant system

The classical Saint Venant system for shallow water describes the height of water $h(t,x) \geq 0$, and the water velocity $u(t,x) \in \mathbb{R}$ (x denotes a coordinate in the horizontal direction) in the direction parallel to the bottom. It assumes a slowly varying topography $z(x)$, and reads

$$\begin{aligned} &\partial_t h + \partial_x(hu) = 0, \\ &\partial_t(hu) + \partial_x(hu^2 + g\frac{h^2}{2}) + gh\partial_x z = 0, \end{aligned} \tag{1}$$

where $g > 0$ is the gravity constant. This system is completed with an entropy (energy) inequality

$$\partial_t\Big(h\frac{u^2}{2} + g\frac{h^2}{2} + ghz\Big) + \partial_x\Big(\big(h\frac{u^2}{2} + gh^2 + ghz\big)u\Big) \leq 0. \tag{2}$$

We shall denote $U = (h, hu)^T$ and

$$\eta(U) = h\frac{u^2}{2} + g\frac{h^2}{2}, \qquad G(U) = \big(h\frac{u^2}{2} + gh^2\big)u, \tag{3}$$

the entropy and entropy fluxes without topography.

2.2 Kinetic interpretation of the Saint-Venant system

The reader can refer to [3] and references therein for a complete presentation of the description of the Saint-Venant system.

The classical kinetic Maxwellian (see e.g. [20]) is given by

$$M(U,\xi) = \frac{1}{g\pi}\Big(2gh - (\xi - u)^2\Big)_+^{1/2}, \tag{4}$$

where $\xi \in \mathbb{R}$ and $x_+ \equiv \max(0,x)$ for any $x \in \mathbb{R}$. It satisfies the following moment relations,

$$\begin{aligned} \int_{\mathbb{R}} M(U,\xi)\,d\xi = h, \qquad \int_{\mathbb{R}} \xi M(U,\xi)\,d\xi = hu, \\ \int_{\mathbb{R}} \xi^2 M(U,\xi)\,d\xi = hu^2 + g\frac{h^2}{2}. \end{aligned} \tag{5}$$

These definitions allow us to obtain a *kinetic representation* of the Saint-Venant system.

Lemma 2.1 *If the topography $z(x)$ is Lipschitz continuous, the pair of functions (h, hu) is a weak solution to the Saint-Venant system (1) if and only if $M(U,\xi)$ satisfies the kinetic equation*

$$\partial_t M + \xi\partial_x M - g(\partial_x z)\partial_\xi M = Q, \tag{6}$$

for some "collision term" $Q(t,x,\xi)$ that satisfies, for a.e. (t,x),

$$\int_{\mathbb{R}} Q d\xi = \int_{\mathbb{R}} \xi Q d\xi = 0. \tag{7}$$

Proof. If (6) and (7) are satisfied, we can multiply (6) by $(1,\xi)^T$, and integrate with respect to ξ. Using (5) and (7) and integrating by parts the term in $\partial_\xi M$, we obtain (1). Conversely, if (h, hu) is a weak solution to (1), just define Q by (6); it will satisfy (7) according to the same computations. □

The standard way to use Lemma 2.1 is to write a kinetic relaxation equation [18, 19, 12, 6, 7], like

$$\partial_t f + \xi \partial_x f - g(\partial_x z)\partial_\xi f = \frac{M - f}{\epsilon}, \tag{8}$$

where $f(t,x,\xi) \geq 0$, $M = M(U,\xi)$ with $U(t,x) = \int (1,\xi)^T f(t,x,\xi) d\xi$, and $\epsilon > 0$ is a relaxation time. In the limit $\epsilon \to 0$ we recover formally the formulation (6), (7). We refer to [6] for general considerations on such kinetic relaxation models without topography, the case with topography being introduced in [20]. Note that the notion of *kinetic representation* as (6), (7) differs from the so called *kinetic formulations* where a large set of entropies is involved, see [21]. For systems of conservation laws, these kinetic formulations include non-advective terms that prevent from writing down simple approximations. In general, kinetic relaxation approximations can be compatible with just a single entropy. Nevertheless this is enough for proving the convergence as $\epsilon \to 0$, see [5].

The interest of the particular form (4) lies in its link with a kinetic entropy. Consider the kinetic entropy,

$$H(f,\xi,z) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{6} f^3 + gzf, \tag{9}$$

where $f \geq 0$, $\xi \in \mathbb{R}$ and $z \in \mathbb{R}$, and its version without topography

$$H_0(f,\xi) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{6} f^3. \tag{10}$$

Then one can check the relations

$$\int_{\mathbb{R}} H\big(M(U,\xi),\xi,z\big)\, d\xi = \eta(U) + ghz, \tag{11}$$

$$\int_{\mathbb{R}} \xi H\big(M(U,\xi),\xi,z\big)\, d\xi = G(U) + ghzu. \tag{12}$$

One has the following entropy relations.

Lemma 2.2 *(i) For any $f(\xi) \geq 0$, setting $h = \int f(\xi)d\xi$, $hu = \int \xi f(\xi) d\xi$ (assumed finite), one has*

$$\eta(U) = \int_{\mathbb{R}} H_0\big(M(U,\xi),\xi\big)\, d\xi \leq \int_{\mathbb{R}} H_0\big(f(\xi),\xi\big)\, d\xi. \tag{13}$$

(ii) The kinetic entropy inequality

$$\partial_t H(M,\xi,z) + \partial_x\big(\xi H(M,\xi,z)\big) - g(\partial_x z)\partial_\xi H(M,\xi,z) = H'(M,\xi,z)Q,$$

holds, leading to the macroscopic inequality

$$\partial_t \int_{\mathbb{R}} H(M,\xi,z) d\xi + \partial_x \int_{\mathbb{R}} \xi H(M,\xi,z) d\xi \leq 0.$$

Proof of Lemma 2.2. The property *(i)* is proved in [20]. For proving *(ii)*, we multiply (6) by $H'(M,\xi,z)$ and an integration in ξ of the obtained equation gives the result. □

2.3 Kinetic scheme for the Saint-Venant system

For numerical purposes it is usual to replace the right-hand side in the kinetic relaxation equation (8) by a time discrete projection to the Maxwellian state. When space discretization is present it leads to flux-vector splitting schemes, see [7] for the case without topography, [20] for the case with topography, and [4] for the 2d case on unstructured meshes.

We would like to approximate the solution $U(t,x)$, $x \in \mathbb{R}$, $t \geq 0$ of the system (1) by discrete values U_i^n, $i \in \mathbb{Z}$, $n \in \mathbb{N}$. In order to do so, we consider a grid of points $x_{i+1/2}$, $i \in \mathbb{Z}$,

$$\ldots < x_{i-1/2} < x_{i+1/2} < x_{i+3/2} < \ldots,$$

and we define the cells (or finite volumes) and their lengths

$$C_i =]x_{i-1/2}, x_{i+1/2}[, \qquad \Delta x_i = x_{i+1/2} - x_{i-1/2}.$$

We consider discrete times t^n with $t^{n+1} = t^n + \Delta t^n$, and we define the piecewise constant functions $U^n(x)$ corresponding to time t^n and $z(x)$ as

$$U^n(x) = U_i^n, \quad z(x) = z_i, \quad \text{for } x_{i-1/2} < x < x_{i+1/2}. \tag{14}$$

A finite volume scheme for solving (1) is a formula of the form

$$U_i^{n+1} = U_i^n - \sigma_i (F_{i+1/2-} - F_{i-1/2+}), \tag{15}$$

where $\sigma_i = \Delta t^n / \Delta x_i$, telling how to compute the values U_i^{n+1} knowing U_i^n and discretized values z_i of the topography. Here we consider first-order explicit three points schemes where

$$F_{i+1/2-} = \mathcal{F}_l(U_i^{n+p}, U_{i+1}^{n+p}, z_{i+1} - z_i), \qquad F_{i+1/2+} = \mathcal{F}_r(U_i^{n+p}, U_{i+1}^{n+p}, z_{i+1} - z_i), \tag{16}$$

with $p = 0, 1$. The value $p = 0$ classically corresponds a first order explicit time scheme for solving (1) whereas $p = 1$ means an implicit time scheme. In this paper, we focus on the case $p = 1$. The functions $\mathcal{F}_{l/r}(U_l, U_r, \Delta z) \in \mathbb{R}^2$ are the numerical fluxes, see [8].

Indeed the method used in [20] in order to solve (1) can be viewed as solving

$$\partial_t f + \xi \partial_x f - g(\partial_x z) \partial_\xi f = 0 \tag{17}$$

for the unknown $f(t,x,\xi)$, over the time interval (t^n, t^{n+1}), with initial data

$$f(t^n, x, \xi) = M(U^n(x), \xi). \tag{18}$$

Defining the update as

$$U_i^{n+1} = \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f(t^{n+1-}, x, \xi)\, dx d\xi, \tag{19}$$

and

$$f_i^{n+1}(\xi) = \frac{1}{\Delta x_i} \int_{x_{i-1/2}}^{x_{i+1/2}} f(t^{n+1-}, x, \xi)\, dx, \tag{20}$$

the formula (19) can then be written

$$U_i^{n+1} = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^{n+1}(\xi)\, d\xi. \tag{21}$$

This formula can in fact be written under the form (15), (16) for some numerical fluxes $\mathcal{F}_{l/r}$.

3 An implicit kinetic scheme

In this section we consider the problem (1) without topography, and the kinetic scheme (17)-(21). First we present and study the discrete implicit kinetic then we detail the macroscopic scheme obtained from the kinetic discretization.

3.1 Implicit scheme without topography

Without topography, the kinetic scheme is a *flux vector splitting* scheme [7]. The update (20) of the solution of (17),(18) simplifies to the discrete kinetic scheme

$$f_i^{n+1-} = M_i - \sigma\xi\left(\mathbb{1}_{\xi<0} f_{i+1}^{n+1-} + \mathbb{1}_{\xi>0} f_i^{n+1-} - \mathbb{1}_{\xi<0} f_i^{n+1-} - \mathbb{1}_{\xi>0} f_{i-1}^{n+1-}\right), \tag{22}$$

with $\sigma = \Delta t^n/\Delta x$. We rewrite the previous equations under the form

$$\left\{\begin{array}{l} -\sigma\mathbb{1}_{\xi>0}\xi f_{i-1}^{n+1-} + (1+\sigma|\xi|) f_i^{n+1-} + \sigma\mathbb{1}_{\xi<0}\xi f_{i+1}^{n+1-} = M_i \\ (1+\sigma|\xi|) f_1^{n+1-} + \sigma\mathbb{1}_{\xi<0}\xi f_2^{n+1-} = M_1 + \sigma\mathbb{1}_{\xi>0}\xi M_0^{n+1}, \\ -\sigma\mathbb{1}_{\xi>0}\xi f_{P-1}^{n+1-} + (1+\sigma|\xi|) f_P^{n+1-} = M_P - \sigma\mathbb{1}_{\xi<0}\xi M_{P+1}^{n+1}, \end{array}\right. \tag{23}$$

The quantities $M_0^{n+1} = M(U_0^{n+1},\xi)$ and $M_{P+1}^{n+1} = M(U_{P+1}^{n+1},\xi)$ appearing the last two lines of (23) account for the imposed boundary conditions. In a first step, we assume that M_0^{n+1} and M_{P+1}^{n+1} are two known kinetic Maxwellian, their expressions will be discussed in more details in the paragraph devoted to the practical computation of the implicit variables, see paragraph 3.4.

With obvious notations, the system (23) consists in finding $f^{n+1} = \{f_i^{n+1-}\}_{i\in\{1,\dots,P\}}$ satisfying

$$(\mathbf{I} + \sigma\mathbf{L}) f^{n+1} = M + \sigma B^{n+1}, \tag{24}$$

where $\mathbf{I}$ is the identity matrix of length P and the three vectors f^{n+1}, M and B^{n+1} of $\mathbb{R}^P$ are defined by

$$f^{n+1} = \begin{pmatrix} f_1^{n+1} \\ \vdots \\ f_i^{n+1} \\ \vdots \\ f_P^{n+1} \end{pmatrix}, \quad M = \begin{pmatrix} M_1 \\ \vdots \\ M_i \\ \vdots \\ M_P \end{pmatrix} \quad \text{and} \quad B^{n+1} = \begin{pmatrix} \mathbb{1}_{\xi>0}\xi M_0^{n+1}, \\ 0 \\ \vdots \\ 0 \\ -\mathbb{1}_{\xi<0}\xi M_{P+1}^{n+1}, \end{pmatrix}. \tag{25}$$

The practical computation of the densities vector f^{n+1} will be discussed in paragraph 3.2. Hereafter, we focus on the properties of the numerical scheme (23) and the two following results hold.

Lemma 3.1 *The matrix $\mathbf{I} + \sigma\mathbf{L}$ defined by (24)*

(i) is invertible for any σ and ξ,

(ii) its inverse $(\mathbf{I} + \sigma\mathbf{L})^{-1}$ has only positive coefficients.

Proposition 3.2 *The numerical scheme (23) satisfies the following properties*

(i) the discretization (23) is consistent with (1),

(ii) the system (23) – or equivalently the system (24) – admits an unique solution and the solution satisfies

$$f_i^{n+1-} = f_i^{n+1-}(\xi) \geq 0, \quad \forall 1 \leq i \leq P,\ \forall \xi \in \mathbb{R}.$$

Proposition 3.3 *Since the system (24) admits a unique solution of positive quantities, it defines an implicit kinetic scheme. Moreover, the numerical scheme defined by (24) satisfies the fully discrete entropy equality*

$$\begin{aligned} H_0(f_i^{n+1-}) = H_0(M_i) &- \sigma\Big(H_{0,i+1/2}^{n+1-} - H_{0,i-1/2}^{n+1-}\Big) - \Psi(f_i^{n+1-}, M_i) \\ &+ \sigma\xi\Big(\mathbb{1}_{\xi<0}\Psi(f_i^{n+1-}, f_{i+1}^{n+1-}) - \mathbb{1}_{\xi>0}\Psi(f_i^{n+1-}, f_{i-1}^{n+1-})\Big) \end{aligned} \tag{26}$$

where $H^{n+1-}_{0,i+1/2}$, $H^{n+1-}_{0,i-1/2}$ *are given by*

$$H^{n+1-}_{0,i+1/2} = \xi \mathbb{1}_{\xi<0} H_0(f^{n+1-}_{i+1}) + \xi \mathbb{1}_{\xi>0} H_0(f^{n+1-}_{i}), \tag{27}$$

$$H^{n+1-}_{0,i-1/2} = \xi \mathbb{1}_{\xi<0} H_0(f^{n+1-}_{i}) + \xi \mathbb{1}_{\xi>0} H_0(f^{n+1-}_{i-1}), \tag{28}$$

and where the function Ψ *is defined by*

$$\Psi : \mathbb{R}^2 \ni (a,b) \longmapsto \frac{g^2\pi^2}{6}(b+2a)(b-a)^2. \tag{29}$$

Since Ψ *is positive on* $\mathbb{R}^2_+$*, the last two terms of equality (26) define a nonpositive dissipative term.*

Notice that the results obtained in the two propositions 3.2 and 3.3 do not require any CFL condition. A consequence of Proposition 3.3 is that, when using the classical Maxwellian (4), the macroscopic scheme associated to (22) will satisfy a discrete entropy inequality that always dissipates the energy. In fact since the Maxwellian (4) minimizes the functional (13) we have the following upper bound on the macroscopic entropy $\eta(U_i^{n+1})$

$$\eta(U_i^{n+1}) = \int_{\mathbb{R}} H_0(M(U_i^{n+1},\xi),\xi)d\xi \leq \int_{\mathbb{R}} H_0(f_i^{n+1-}(\xi),\xi)d\xi.$$

We then use equality (26) yielding

$$\eta(U_i^{n+1}) \leq \eta(U_i^n) - \sigma\left(\int_{\mathbb{R}} H^{n+1-}_{0,i+1/2}(\xi)d\xi - \int_{\mathbb{R}} H^{n+1-}_{0,i-1/2}(\xi)d\xi\right) + \int_{\mathbb{R}} D_i(\xi)d\xi, \tag{30}$$

where D_i is the negative dissipation term corresponding to the last three lines of (26).

Proof of Lemma 3.1. The matrix $\mathbf{I}+\sigma\mathbf{L}$ writes

$$\mathbf{I}+\sigma\mathbf{L} = \begin{pmatrix} 1+\sigma_1|\xi| & \sigma_1\xi\mathbb{1}_{\xi\leq 0} & 0 & \dots & 0 \\ -\sigma_2\xi\mathbb{1}_{\xi\geq 0} & 1+\sigma_2|\xi| & \sigma_2\xi\mathbb{1}_{\xi\leq 0} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\sigma_{P-1}\xi\mathbb{1}_{\xi\geq 0} & 1+\sigma_{P-1}|\xi| & \sigma_{P-1}\xi\mathbb{1}_{\xi\leq 0} \\ 0 & \dots & 0 & -\sigma_P\xi\mathbb{1}_{\xi\geq 0} & 1+\sigma_P|\xi| \end{pmatrix},$$

and it is easy to see that the matrix $\mathbf{I}+\sigma\mathbf{L}$ is strictly diagonally dominant and hence invertible. Moreover the matrix $\mathbf{\Lambda} = \mathbf{I}+\sigma\mathbf{L}$ is such that

$$\mathbf{\Lambda}_{i,i} > 0, \quad \text{and} \quad \mathbf{\Lambda}_{i,j} \leq 0, \text{ when } i \neq j,$$

meaning $\mathbf{I}+\sigma\mathbf{L}$ is a monotone matrix and hence the solution of (24) satisfies

$$f_i^{n+1-} = \left((\mathbf{I}+\sigma\mathbf{L})^{-1}(M+\sigma B^{n+1,k})\right)_i \geq 0, \quad \forall i,$$

proving the result.

Denoting $\mathbf{L}^d$ (resp. $\mathbf{L}^{nd}$) the diagonal (resp. non diagonal) part of $\mathbf{L}$ we can write

$$\mathbf{I}+\sigma\mathbf{L} = (\mathbf{I}+\sigma\mathbf{L}^d)\left(\mathbf{I}-(\mathbf{I}+\sigma\mathbf{L}^d)^{-1}(-\sigma\mathbf{L}^{nd})\right),$$

where all the entries of the matrix

$$\mathbf{J} = (\mathbf{I}+\sigma\mathbf{L}^d)^{-1}(-\sigma\mathbf{L}^{nd}),$$

are non negative and less than 1. And hence, we can write

$$(\mathbf{I}+\sigma\mathbf{L})^{-1} = (\mathbf{I}-\mathbf{J})^{-1}(\mathbf{I}+\sigma\mathbf{L}^d)^{-1} = \sum_{k=0}^{\infty}\mathbf{J}^k(\mathbf{I}+\sigma\mathbf{L}^d)^{-1},$$

proving all the entries of $(\mathbf{I}+\sigma\mathbf{L})^{-1}$ are non negative. □

Proof of prop. 3.2. *(i)* The four terms in parentheses in (22) are conservative, and are classically consistent with $\xi\partial_x f$ in (17).
(ii) This is a direct consequence of Lemma 3.1. □

The proof of Proposition 3.3 makes use of the following Lemma which will also be useful later.

Lemma 3.4 *The following idendity holds for any real pair (a,b) and for any $\xi\in\mathbb{R}$*

$$H(b,\xi)=H(a,\xi)+\partial_f H(a,\xi)(b-a)+\Psi(a,b), \tag{31}$$

with the function Ψ defined in (29). Especially, we recover the convexity of $H(\cdot,\xi)$ on $\mathbb{R}_+$ thanks to the positivity of Ψ on $\mathbb{R}_+^2$.

Proof of Lemma 3.4. For any (a,b) in $\mathbb{R}^2$ there holds

$$\begin{aligned}\partial_f H(a)(b-a)&=\frac{\xi^2}{2}b+\frac{g^2\pi^2}{2}a^2b-\frac{\xi^2}{2}a-\frac{g^2\pi^2}{2}a^3\\&=H(b)+\frac{g^2\pi^2}{2}a^2b-\frac{g^2\pi^2}{6}b^3-H(a)-\frac{g^2\pi^2}{2}a^3+\frac{g^2\pi^2}{6}a^3\\&=H(b)-H(a)-\frac{g^2\pi^2}{6}\Big(b^3-a^3-3a^2(b-a)\Big)\ ,\end{aligned}$$

and equality (31) is recovered using the formula

$$b^3-a^3-3a^2(b-a)=(b+2a)(b-a)^2.$$

□

Proof of prop. 3.3. The proof follows similar lines as what was done in the case of the fully explicit version of the kinetic scheme in [3]. Instead of multiplying Equation (22) by $\partial_f H_0(f_i^n)$, we multiply it with $\partial_f H_0(f_i^{n+1-})$, which leads to

$$\begin{aligned}\partial_f H_0(f_i^{n+1-})(f_i^{n+1-}-M_i)=&-\sigma\xi\mathbb{1}_{\xi<0}\partial_f H_0(f_i^{n+1-})(f_{i+1}^{n+1-}-f_i^{n+1-})\\&+\sigma\xi\mathbb{1}_{\xi>0}\partial_f H_0(f_i^{n+1-})(f_{i-1}^{n+1-}-f_i^{n+1-})\ .\end{aligned} \tag{32}$$

In (32) we recognize three terms of the form $\partial_f H(a)(b-a)$ with $a=f_i^{n+1-}$ and $b\in\{f_{i-1}^{n+1-},M_i,f_{i+1}^{n+1-}\}$. Taking advantage of Lemma 3.4 we can write

$$\begin{aligned}&H(f_i^{n+1-})-H(M_i)+\Psi(f_i^{n+1-},M_i)=\\&\quad-\sigma\xi\mathbb{1}_{\xi<0}\Big(H(f_{i+1}^{n+1-})-H(f_i^{n+1-})-\Psi(f_i^{n+1-},f_{i+1}^{n+1-})\Big)\\&\quad+\sigma\xi\mathbb{1}_{\xi>0}\Big(H(f_{i-1}^{n+1-})-H(f_i^{n+1-})-\Psi(f_i^{n+1-},f_{i-1}^{n+1-})\Big),\end{aligned}$$

and we conclude by grouping the expressions. □

3.2 Practical computation of f^{n+1-}

When dealing with implicit schemes, one has often to invert an operator and the key point of the numerical scheme (24) is the computation of the inverse of the matrix $\mathbf{I}+\sigma\mathbf{L}$. In our case, it will be possible to compute analytically this inverse thanks to the triangular structure of the matrix, which is due to the upwinding of the fluxes in (22). More precisely we decompose $(\mathbf{I}+\sigma\mathbf{L})^{-1}$ as the contributions of the left- and right-going information, which gives

$$(\mathbf{I}+\sigma\mathbf{L})^{-1}=(\mathbf{I}+\sigma\mathbf{L}^+)^{-1}\mathbb{1}_{\xi\le0}+(\mathbf{I}+\sigma\mathbf{L}^-)^{-1}\mathbb{1}_{\xi\ge0}, \tag{33}$$

with the upwinding matrices $\mathbf{L}^+$ and $\mathbf{L}^-$ corresponging to

$$\mathbf{L}^+ = \begin{pmatrix} -\xi & \xi & 0 & \dots & 0 \\ 0 & -\xi & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \xi \\ 0 & \dots & & 0 & -\xi \end{pmatrix}, \quad \mathbf{L}^- = \begin{pmatrix} \xi & 0 & & \dots & 0 \\ -\xi & \ddots & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\xi & \xi \end{pmatrix}.$$

Introducing $\mathbf{J}^+$ and $\mathbf{J}^-$ the matrices of $\mathbb{R}^{P\times P}$ defined as

$$(\mathbf{J}^+)_{i,j} = \begin{cases} 1 & \text{if } i = j-1 \\ 0 & \text{otherwise} \end{cases}, \quad (\mathbf{J}^-)_{i,j} = \begin{cases} 1 & \text{if } i = j+1 \\ 0 & \text{otherwise} \end{cases},$$

we can write

$$\begin{cases} (\mathbf{I}+\sigma\mathbf{L}^-)^{-1} = \big((1+\sigma\xi)\mathbf{I} - \sigma\xi\mathbf{J}^-\big)^{-1} = \dfrac{1}{1+\sigma\xi}\left(\mathbf{I} - \dfrac{\sigma\xi}{1+\sigma\xi}\mathbf{J}^-\right)^{-1} \\ (\mathbf{I}+\sigma\mathbf{L}^+)^{-1} = \big((1-\sigma\xi)\mathbf{I} + \sigma\xi\mathbf{J}^+\big)^{-1} = \dfrac{1}{1-\sigma\xi}\left(\mathbf{I} + \dfrac{\sigma\xi}{1-\sigma\xi}\mathbf{J}^+\right)^{-1} \end{cases}.$$

The above inverses can be computed through geometric sums since $\mathbf{J}_P^+$ and $\mathbf{J}_P^-$ have a spectral radius equal to zero. More specifically these two matrices are nilpotent, which implies that the geometric sums in question contain a finite number of nonzero terms and are given below

$$(\mathbf{I}+\sigma\mathbf{L}^-)^{-1} = \sum_{k=0}^{P} \frac{(\sigma\xi)^k}{(1+\sigma\xi)^{k+1}}(\mathbf{J}^-)^k,$$
$$(\mathbf{I}+\sigma\mathbf{L}^+)^{-1} = \sum_{k=0}^{P} \frac{(-\sigma\xi)^k}{(1-\sigma\xi)^{k+1}}(\mathbf{J}^+)^k.$$

To conclude we give the analytic expression of the inverse:

$$(\mathbf{I}+\sigma\mathbf{L}^-)^{-1}_{i,j} = \begin{cases} \dfrac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} & \text{if } i \geq j \\ 0 & \text{else} \end{cases}, \tag{34}$$

$$(\mathbf{I}+\sigma\mathbf{L}^+)^{-1}_{i,j} = \begin{cases} \dfrac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} & \text{if } i \leq j \\ 0 & \text{else} \end{cases}. \tag{35}$$

Especially we recover the properties enumerated in Lemma 3.1, since we see that all the coefficients of the inverse (33) are comprised between zero and one respectively when $\xi \geq 0$ and $\xi \leq 0$.

3.3 Macroscopic implicit scheme

We now turn towards obtaining an explicit writing of the macroscopic update associated to (22). Since the right hand side of (24) is made of Maxwellians, we will see that this amounts to compute the integral of

$$\mathbb{1}_{\pm\xi>0}\frac{(\pm\sigma\xi)^k}{(1\pm\sigma\xi)^{k+1}}M(U,\xi)$$

against 1, ξ and ξ^2 for $0 \leq k \leq P-1$. This hardly seems possible with the classical Maxwellian proposed in (4). Instead in this section we will use the simpler equilibrium function given by

$$M(U,\xi) = \frac{h}{2\sqrt{3}c}\mathbb{1}_{|\xi-u|\leq\sqrt{3}c}, \quad c = \sqrt{\frac{gh}{2}}, \tag{36}$$

and referred to as the index Maxwellian. This is the simplest choice we can make, and it will enable us to obtain analytic expressions for the aforementionned integrals. Furthermore it satisfies all the moment relations (5), and we make the following remark.

Remark 3.5 *We recall that the half-disk Maxwellian $M(U,\xi)$ defined by (4) has some optimal properties presented in Lemma 2.2, which allow to obtain the discrete entropy inequality (30) at the macroscopic scale. Other choices of Maxwellian are possible such as (36), but the previous discrete entropy inequality is not granted to hold anymore. A general possibility is to choose $M(U,\xi)$ of the form*

$$M(U,\xi) = \frac{h}{c}\chi\Big(\frac{\xi-u}{c}\Big).$$

To satisfy the moment relations (5), it is then sufficient for χ to be an even function verifying

$$\int_{\mathbb{R}} \chi(z)dz = \int_{\mathbb{R}} z^2\chi(z)dz = 1.$$

Furthermore we ask χ to be nonnegative with compact support, and possible choices are for instance

$$\chi_1(z) = \frac{1}{2\sqrt{3}}\mathbb{1}_{|z|\leq\sqrt{3}}, \quad or \quad \chi_2(z) = \frac{3}{20\sqrt{5}}z^2 + \frac{3}{4\sqrt{5}}\mathbb{1}_{|z|\leq\sqrt{5}}. \tag{37}$$

The definition (36) of the index Maxwellian corresponds to the first shape function χ_1 in (37), whereas the definition (4) of the half-disk Maxwellian corresponds to the third choice below

$$\chi_3(z) = \frac{1}{\pi}\sqrt{1-\frac{z^2}{4}}\mathbb{1}_{|z|\leq 2}.$$

We proceed in two steps to compute our scheme with boundary conditions. The strategy consists to dissociate the contribution of the information coming from the interior of the computational domain, and the one coming from the exterior as below

$$\begin{cases} U^{\rm int} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix}1\\ \xi\end{pmatrix} (\mathbf{I}+\sigma\mathbf{L})^{-1} M\, d\xi \\ U^{\rm ext} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix}1\\ \xi\end{pmatrix} (\mathbf{I}+\sigma\mathbf{L})^{-1}\sigma B^{n+1}\, d\xi \end{cases}. \tag{38}$$

The final update is then set as $U^{n+1} = U^{\rm int} + U^{\rm ext}$, which coincides with definition (21). We postpone the details about the computation of B^{n+1} to the next section, and assume that it is known for now. First for $U^{\rm int}$, we have

$$U^{\rm int} = \int_{\mathbb{R}} \mathbb{1}_{\xi\leq 0}\begin{pmatrix}1\\ \xi\end{pmatrix}(\mathbf{I}+\sigma\mathbf{L}^+)^{-1}Md\xi + \int_{\mathbb{R}} \mathbb{1}_{\xi\geq 0}\begin{pmatrix}1\\ \xi\end{pmatrix}(\mathbf{I}+\sigma\mathbf{L}^-)^{-1}Md\xi.$$

Plugging the analytic expressions (34) and (35) in the above integrals, we can express the i-th component of $U^{\rm int}$ as

$$\begin{aligned} U_i^{\rm int} &= \int_{\xi<0}\sum_{j=1}^{P}\begin{pmatrix}1\\ \xi\end{pmatrix}(\mathbf{I}+\sigma\mathbf{L}^+)^{-1}_{i,j}M_j d\xi + \int_{\xi>0}\sum_{j=1}^{P}\begin{pmatrix}1\\ \xi\end{pmatrix}(\mathbf{I}+\sigma\mathbf{L}^-)^{-1}_{i,j}M_j d\xi \\ &= \int_{\xi<0}\sum_{j=i}^{P}\begin{pmatrix}1\\ \xi\end{pmatrix}\frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}}M_j d\xi + \int_{\xi>0}\sum_{j=1}^{i}\begin{pmatrix}1\\ \xi\end{pmatrix}\frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}}M_j d\xi. \end{aligned} \tag{39}$$

A detailed expression of the quantities appearing in relation (39) is given in Appendix A. Similarly for the exterior contribution we have

$$U^{\rm ext} = \sigma\int_{\xi<0}\begin{pmatrix}1\\ \xi\end{pmatrix}(\mathbf{I}+\sigma\mathbf{L}^+)^{-1}B^{n+1}d\xi + \sigma\int_{\xi>0}\begin{pmatrix}1\\ \xi\end{pmatrix}(\mathbf{I}+\sigma\mathbf{L}^-)^{-1}B^{n+1}d\xi.$$

Using definition (25) and equalities (34)–(35), the i-th component of U^{ext} is

$$U_i^{\text{ext}} = \int_{\xi<0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}} M_{P+1}^{n+1} d\xi + \int_{\xi>0} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^i}{(1+\sigma\xi)^i} M_0^{n+1} d\xi. \tag{40}$$

We can reuse the primitive obtained in Appendix A to express the water height h_i^{ext}. However for the flux $(hu)_i^{\text{ext}}$ we need to write a primitive for

$$\int_{\pm\xi>0} \xi \frac{(\pm\sigma\xi)^k}{(1\pm\sigma\xi)^k} M(U,\xi) d\xi,$$

which is given in Appendix B. In the end we obtain a fully explicit writing of the implicit update at the macroscopic level.

3.4 Boundary conditions

In this paragraph we discuss how to enforce the boundary conditions. These are represented by the exterior contribution U^{ext} introduced in (38), and accordingly we need to specify the ghost values U_0^{n+1} and U_{P+1}^{n+1} appearing in the definition (25) of B^{n+1}. The problem we are facing is that these ghost quantities depend on the neighboring values in cells C_1 and C_P at time t^{n+1}, and which are themselves unknown. Hence we have an implicit problem where the relation between the ghost and border terms can be nonlinear depending on the type of boundary conditions. In practice we will avoid this issue by substituting B^{n+1} with B^n in the definition of U^{ext}. Doing so can be interpreted as a first order approximation in time since we have

$$U_0^{n+1} = U_0^n + O(\Delta t), \quad U_{P+1}^{n+1} = U_{P+1}^n + O(\Delta t).$$

The benefit is that we can more easily determine the ghost quantities U_0^n, U_{P+1}^n at time t^n based on U_1^n, U_P^n following the procedure described hereafter and similar to that of Bristeau and Coussin in [10]. We will focus on fluvial flows where the material velocity of particles $|u|$ is smaller than the celerity of surface gravity waves $\sqrt{gh}$. In particular low Froude flows enter this regime. Since in this case the eigenvalues $u-\sqrt{gh}$ and $u+\sqrt{gh}$ have opposite sign, at each boundary we have exactly one wave entering the domain and one wave leaving it. Hence we dispose of a single degree of freedom to set the ghost values, which generally consists in enforcing either a given water height or a discharge. The ghost state is then fully determined by asking the outward-going Riemann invariant to remain constant through the interface.

Given water height. First we treat the case where the water height is enforced at the boundary of the domain. We denote by $h_{g,l}$ the value attributed to the left ghost cell, and $h_{g,r}$ the one attributed to the right ghost cell. Together with the condition on the outgoing Riemann invariant, we get the following nonlinear systems

$$\begin{cases} h_0^n = h_{g,l} \\ u_0^n - 2\sqrt{gh_0^n} = u_1^n - 2\sqrt{gh_1^n} \end{cases}, \quad \begin{cases} h_{P+1}^n = h_{g,r} \\ u_{P+1}^n + 2\sqrt{gh_{P+1}^n} = u_P^n + 2\sqrt{gh_P^n} \end{cases}.$$

They can be solved explicitly and we get

$$U_0^n = h_{g,l} \begin{pmatrix} 1 \\ u_1^n - 2(\sqrt{gh_1^n} - \sqrt{gh_{g,l}}) \end{pmatrix},$$
$$U_{P+1}^n = h_{g,r} \begin{pmatrix} 1 \\ u_P^n + 2(\sqrt{gh_P^n} - \sqrt{gh_{g,r}}) \end{pmatrix}.$$

Given flux. Another possibility is to enforce the discharge at the boundary, and we denote by $q_{g,l}$ and $q_{g,r}$ the left and right ghost values. This time around, the constraint on the Riemann invariant will enable to determine the ghost water height. Indeed we have the systems

$$\begin{cases} q_0^n = q_{g,l} \\ u_0^n - 2\sqrt{gh_0^n} = u_1^n - 2\sqrt{gh_1^n} \end{cases}, \quad \begin{cases} q_{P+1}^n = q_{g,r} \\ u_{P+1}^n + 2\sqrt{gh_{P+1}^n} = u_P^n + 2\sqrt{gh_P^n} \end{cases}, \tag{41}$$

and the equalities satisfied by the Riemann invariants amount to finding the real roots of the third order polynomials in $\sqrt{h_0^n}$ and $\sqrt{h_{P+1}^n}$ below

$$-2\sqrt{g}(h_0^n)^{3/2} - (u_1^n - 2\sqrt{gh_1^n})h_0^n + q_{g,l} = 0,$$
$$2\sqrt{g}(h_{P+1}^n)^{3/2} - (u_P^n + 2\sqrt{gh_P^n})h_{P+1}^n + q_{g,r} = 0.$$

Note that in this case, our approach differs from that of Bristeau and Coussin in [10], where the ghost value is chosen such that the resulting numerical flux at the interface coincides with the boundary discharge. Instead we do not enforce any value at the interface but directly in the ghost cell, which can be seen as a first order simplification in space. Obtaining the ghost water heights h_0^n and h_{P+1}^n requires to study the roots of a third degree polynomial, and we were able to ensure the existence of at least one non-negative real root when the Froude number is less than one and under the respective conditions

$$q_{g,l} \geq \frac{1}{27g^2}(u_1^n - 2\sqrt{gh_1^n})^3, \quad q_{g,r} \leq \frac{1}{27g^2}(u_P^n + 2\sqrt{gh_P^n})^3.$$

If the previous constraints do not hold, this means that the systems in (41) cannot be satisfied and one can assume that the equations (41) can be replaced by h_0^n and/or h_{P+1}^n given. When more than one non-negative real root exist, we choose the smaller one.

We comment on the fact that nothing prevents us from mixing the boundary conditions, for instance we can enforce a water height on the left boundary, and a discharge on the right. A common practice for chanel flows is to enforce the water height at the inlet and the flux at the outlet.

Remark 3.6 *When substituting B^{n+1} with B^n in the implicit kinetic scheme (24), the corresponding update can be reformulated as $(\overline{\mathbf{I}+\sigma\mathbf{L}})\overline{f}^{n+1} = \overline{M}^n$ with*

$$\overline{\mathbf{I}+\sigma\mathbf{L}} = \left(\begin{array}{c|c|c} 1 & & \\ \hline -\sigma\xi\mathbb{1}_{\xi>0} & & \vdots \\ 0 & \mathbf{I}+\sigma\mathbf{L} & 0 \\ \vdots & & \sigma\xi\mathbb{1}_{\xi<0} \\ \hline & & 1 \end{array}\right), \qquad \overline{f}^{n+1} = \begin{pmatrix} f_0^{n+1} \\ f^{n+1} \\ f_{P+1}^{n+1} \end{pmatrix}, \qquad \overline{M}^n = \begin{pmatrix} M_0^n \\ M^n \\ M_{P+1}^n \end{pmatrix}$$

As a consequence the maximum principle $\|\overline{f}^{n+1}(\xi)\|_\infty \leq \|\overline{M}^n(\xi)\|_\infty$ holds for any ξ in $\mathbb{R}$ during the transport step. In fact we can verify that matrix $(\overline{\mathbf{I}+\sigma\mathbf{L}})$ is monotone, and following the argument involved in Lemma 5.1 from [1] we can write

$$0 \leq \overline{f}^{n+1} = (\overline{\mathbf{I}+\sigma\mathbf{L}})^{-1}\overline{M}^n \leq (\overline{\mathbf{I}+\sigma\mathbf{L}})^{-1}(\|\overline{M}^n\|_\infty \mathbf{1}),$$

with $\mathbf{1}$ the vector from $\mathbb{R}^{P+2}$ whose entries are all equal to one. Using equality $(\overline{\mathbf{I}+\sigma\mathbf{L}})^{-1}\mathbf{1} = \mathbf{1}$ allows to conclude. Note however that there is no such principle at the macroscopic scale, similarly to the continuous Saint-Venant system.

3.5 Implementation and computational costs

It is important to try and keep a reasonable algorithmic complexity so that the implicit method presented in the previous lines remains usable in practice. We discuss here how to improve its computational cost by a substantial margin. In Appendix A, we show that the i-th component of vectors h^{int} and $(hu)^{\text{int}}$ have the form

$$\left\{\begin{array}{l} h_i^{\text{int}} = \dfrac{1}{2\sigma\sqrt{3}}\left(\displaystyle\sum_{j=i}^{P}\sqrt{\frac{2h_j^n}{g}}(Ah)_{i,j} + \sum_{j=1}^{i}\sqrt{\frac{2h_j^n}{g}}(Bh)_{i,j}\right) \\ (hu)_i^{\text{int}} = \dfrac{1}{2\sigma^2\sqrt{3}}\left(-\displaystyle\sum_{j=i}^{P}\sqrt{\frac{2h_j^n}{g}}(Ahu)_{i,j} + \sum_{j=1}^{i}\sqrt{\frac{2h_j^n}{g}}(Bhu)_{i,j}\right) \end{array}\right., \tag{42}$$

where Ah, Ahu are dense upper triangular matrices, and Bh, Bhu are dense lower triangular matrices. Therefore computing h^{int} and $(hu)^{\text{int}}$ through (42) is analog to performing a matrix-vector product which has a quadratic complexity $O(P^2)$, and we cannot hope to do better than that. However the coefficients (86)–(89) of the above matrices involve a summation, and at a first glance the cost to assemble them is seemingly cubic. This is quite expensive and can render the method pretty much inefficient. However this complexity can be reduced to a quadratic cost by computing the coefficients in the correct order. More specifically we show that all the matrices above can be defined through a recurrence relation allowing to compute each coefficient from a previous one in $O(1)$ operation. In fact, denoting $y = x/(1+x)$ and $z = \ln|1+x|$, the matrix Ah is given by

$$\begin{pmatrix} \big[z\big]_{-\min(0,b_1)\sigma}^{-\min(0,a_1)\sigma} & \big[z-y\big]_{-\min(0,b_2)\sigma}^{-\min(0,a_2)\sigma} & \cdots & \cdots & \big[z-\sum_{l=1}^{P-1} y^l/l\big]_{-\min(0,b_P)\sigma}^{-\min(0,a_P)\sigma} \\ 0 & \big[z\big]_{-\min(0,b_2)\sigma}^{-\min(0,a_2)\sigma} & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \big[z-y\big]_{-\min(0,b_P)\sigma}^{-\min(0,a_P)\sigma} \\ 0 & \cdots & \cdots & 0 & \big[z\big]_{-\min(0,b_P)\sigma}^{-\min(0,a_P)\sigma} \end{pmatrix},$$

where $a_j^n = u_j^n - \sqrt{3}\,c_j^n$ and $b_j^n = u_j^n + \sqrt{3}\,c_j^n$. This corresponds to the recursive definition below

$$(Ah)_{i,j} = \begin{cases} 0 & \text{if } j < i \\ \big[\ln(|1+x|)\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } i = j \\ (Ah)_{i+1,j} - \frac{1}{j-i}\big[y^{j-i}\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } j > i \end{cases}. \tag{43}$$

Likewise, the lower triangular matrix Bh is given by

$$\begin{pmatrix} \big[z\big]_{\max(0,a_1^n)\sigma}^{\max(0,b_1^n)\sigma} & 0 & \cdots & \cdots & 0 \\ \big[z-y\big]_{\max(0,a_1^n)\sigma}^{\max(0,b_1^n)\sigma} & \ddots & & & \\ \vdots & & \ddots & & \\ \vdots & & & \big[z\big]_{\max(0,a_{P-1}^n)\sigma}^{\max(0,b_{P-1}^n)\sigma} & 0 \\ \big[z-\sum_{l=1}^{P-1} y^l/l\big]_{\max(0,a_1^n)\sigma}^{\max(0,b_1^n)\sigma} & \cdots & \cdots & \big[z-y\big]_{\max(0,a_{P-1}^n)\sigma}^{\max(0,b_{P-1}^n)\sigma} & \big[z\big]_{\max(0,a_P^n)\sigma}^{\max(0,b_P^n)\sigma} \end{pmatrix},$$

and can be defined by the following recurrence formula

$$(Bh)_{i,j} = \begin{cases} 0 & \text{if } i < j \\ \big[\ln(|1+x|)\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i = j \\ (Bh)_{i-1,j} - \frac{1}{i-j}\big[y^{i-j}\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i > j \end{cases}. \tag{44}$$

Hence it is more efficient to assemble matrices Ah and Bh column wise, starting from the diagonal coefficient and moving towards the first or last row. This way we only have to substract one term to the previous coefficient so as to get the next one, and the cost of this operation is in $O(1)$. Since there are $P(P+1)/2$ coefficients to compute in total, the assembly of Ah and Bh following this strategy requires $O(P^2)$ steps.

A similar conclusion is achieved for Ahu and Bhu, although the recurrence relation is less straightforward to obtain. We first remark that, introducing $(l)_{i,j} = i - j + 1$ the relations (88) and (89) become

$$(Ahu)_{i,j} = \mathbb{1}_{j\geq i}\Big[-(l)_{j,i}\ln|1+x| + x + \sum_{k=1}^{j-i} k\,\frac{y^{(l)_{j,i}-k}}{(l)_{j,i}-k}\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma},$$

$$(Bhu)_{i,j} = \mathbb{1}_{i\geq j}\Big[-(l)_{i,j}\ln|1+x| + x + \sum_{k=1}^{i-j} k\,\frac{y^{(l)_{i,j}-k}}{(l)_{i,j}-k}\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma}.$$

Performing the change of index $r = (l)_{j,i} - k$ for matrix Ahu and $s = (l)_{i,j} - k$ for matrix Bhu we find

$$(Ahu)_{i,j} = \Big[-(l)_{i,j}\ln|1+x| + x + (l)_{i,j}\sum_{r=1}^{j-i}\frac{y^r}{r} - \sum_{r=1}^{j-i} y^r\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma},$$

$$(Bhu)_{i,j} = \Big[-(l)_{i,j}\ln|1+x| + x + (l)_{i,j}\sum_{s=1}^{i-j}\frac{y^s}{s} - \sum_{s=1}^{i-j} y^s\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma}.$$

Next we introduce the matrices defined column wise in a recursive manner

$$(UA)_{i,j} = \begin{cases} 0 & \text{if } j \leq i \\ (UA)_{i+1,j} + \big[y^{j-i}\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } j > i \end{cases},$$

$$(VA)_{i,j} = \begin{cases} 0 & \text{if } j \leq i \\ (VA)_{i+1,j} + \Big[\frac{y^{j-i}}{j-i}\Big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & \text{if } j > i \end{cases}.$$

Then we can write that

$$(Ahu)_{i,j} = \begin{cases} 0 & j < i \\ \big[x - \ln|1+x|\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & j = i \\ (l)_{j,i}(VA)_{i,j} - (UA)_{i,j} + \big[x - (l)_{j,i}\ln|1+x|\big]_{-\min(0,b_j)\sigma}^{-\min(0,a_j)\sigma} & j > i \end{cases}. \tag{45}$$

Similarly we introduce

$$(UB)_{i,j} = \begin{cases} 0 & \text{if } i \leq j \\ (UB)_{i-1,j} + \big[y^{i-j}\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i > j \end{cases},$$

$$(VB)_{i,j} = \begin{cases} 0 & \text{if } i \leq j \\ (VB)_{i-1,j} + \Big[\frac{y^{i-j}}{i-j}\Big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & \text{if } i > j \end{cases},$$

so that we have

$$(Bhu)_{i,j} = \begin{cases} 0 & i < j \\ \big[x - \ln|1+x|\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & i = j \\ (l)_{i,j}(VB)_{i,j} - (UB)_{i,j} + \big[x - (l)_{i,j}\ln|1+x|\big]_{\max(0,a_j)\sigma}^{\max(0,b_j)\sigma} & i > j \end{cases}. \tag{46}$$

To conclude, through relations (45) and (46) we are also able to assemble matrices Ahu and Bhu with a quadratic cost with respect to the number of cells, which means that the overall method has a $O(P^2)$ complexity.

We have considered here the specific case of a kinetic solver and one can imagine that an implicit scheme for another finite volume solver can lead to reduced numerical costs. But it is worth noticing that since the explicit expression of the inverse of the matrix $\mathbf{I}+\sigma\mathbf{L}$ is accessible in the kinetic context, one can hardly find a more efficient implicit technique.

Obviously the proposed implicit scheme is not constrained by any CFL condition associated with an explicit scheme, nevertheless it is important to compare the computational costs of the explicit and implicit strategies in the context of a kinetic solver.

Explicit scheme. Let Δt^n be the time step allowing to satisfy the CFL constraint. In order to obtain the expression of U^{n+1} from U^n, approximately $4P$ numerical fluxes have to be computed (2 numerical fluxes at each interface for each variable h and hu). The explicit kinetic scheme is fully detailed in [4, 3].

Implicit scheme. The CFL constraint being relaxed, we can consider a time step $\Delta t^n_{imp} \gg \Delta t^n$. The results obtained in this paragraph shows that the update from U^{n+1} from U^n requires approximately P^2 numerical fluxes to compute.

We conclude that the implicit strategy is less expensive when

$$\frac{\Delta t^n_{imp}}{\Delta t^n} \gg \frac{P^2}{P} = P. \tag{47}$$

Note however that the computational cost is not the only factor to account for, and one should also consider the efficiency of the scheme, that is to say the relation between the error and the computational time. Generally, taking a very coarse resolution in time results in poorly accurate results, in which case it is not desirable to have (47). Nevertheless there are some cases where the fast dynamics do not play an important role such as in the low Froude regime. Then it might be advantageous to consider large time steps. We will see through the upcoming numerical results from Section 6 that the interest of the implicit kinetic scheme is rather limited when it comes to efficiency, at least for the considered test cases. Hence the explicit strategy is preferable to the implicit one, unless we account for the greater stability offered by the latter in terms of discrete entropy inequality.

4 The 2d case

With obvious notations, we consider the 2d Saint-Venant system written under the form

$$\frac{\partial h}{\partial t} + \nabla_{x,y} \cdot (h\mathbf{u}) = 0, \tag{48}$$

$$\frac{\partial (h\mathbf{u})}{\partial t} + \nabla_{x,y} \cdot (h\mathbf{u} \otimes \mathbf{u}) + \nabla_{x,y}\big(\frac{g}{2}h^2\big) = -gh\nabla_{x,y} z_{2d}, \tag{49}$$

with $\mathbf{u} = (u, v)^T$. The kinetic interpretation of the 2d Saint-Venant system (48)-(49) is a straightforward extension of Lemma 2.1 and has been studied in [4, 1].

To build the 2d Gibbs equilibrium, we define the function

$$\chi_{2d}(z_1, z_2) = \frac{1}{4\pi} \mathbb{1}_{z_1^2 + z_2^2 \leq 4}. \tag{50}$$

This choice corresponds to the 2d version of the kinetic Maxwellian used in 1d (see [1, Remark 4.2]) and we have

$$M_{2d} = M(U_{2d}, \xi, \gamma) = \frac{h}{c^2} \chi_{2d}\left(\frac{\xi - u}{c}, \frac{\gamma - v}{c}\right), \tag{51}$$

with $c = \sqrt{\frac{g}{2}h}$, $(\xi, \gamma) \in \mathbb{R}^2$ and

$$U_{2d} = (h, hu, hv)^T. \tag{52}$$

In other words, we have $M_{2d} = \frac{1}{2g\pi} \mathbb{1}_{(\xi-u)^2 + (\gamma-v)^2 \leq 2gh}$ and the following lemma holds.

Lemma 4.1 *If the topography $z_{2d}(x, y)$ is Lipschitz continuous, the pair of functions $(h, h\mathbf{u})$ is a weak solution to the Saint-Venant system (48)-(49) if and only if $M_{2d}(U, \xi)$ satisfies the kinetic equation*

$$\partial_t M_{2d} + \begin{pmatrix} \xi \\ \gamma \end{pmatrix} \cdot \nabla_{x,y} M_{2d} - g\nabla_{x,y} z_{2d} \cdot \nabla_{\xi,\gamma} M_{2d} = Q_{2d}, \tag{53}$$

for some "collision term" $Q_{2d}(t, x, y, \xi, \gamma)$ that satisfies, for a.e. (t, x, y),

$$\int_{\mathbb{R}^2} Q_{2d} d\xi d\gamma = \int_{\mathbb{R}^2} \xi Q_{2d} d\xi d\gamma = \int_{\mathbb{R}^2} \gamma Q_{2d} d\xi d\gamma = 0. \tag{54}$$

Proof of Lemma 4.1. The proof relies on simple computations. Classically, the integral of Eq. (53) over $\mathbb{R}^2$ gives Eq. (48) whereas the integral over $\mathbb{R}^2$ of Eq. (53) multiplied by $(\xi, \gamma)^T$ ives Eq. (49). □

Let us consider a cartesian mesh of a 2d domain $\Omega = (0, L_x) \times (O, L_y)$, the vertices are denoted $P_{i,j}$ for $0 \leq i \leq P+1$, $0 \leq j \leq L+1$. The coordinates of $P_{i,j}$ are $(x_i, y_j)^T$ with

$$x_i = i\Delta x, \qquad y_j = j\Delta y,$$

and $\Delta_x = L_x/(P+1)$, $\Delta_y = L_y/(L+1)$. Without loss of generality, we consider $L_x = L_y$ and $P = L$ hence $\Delta_x = \Delta_y$. We use the following notations (see Fig. 1):

- $K_{i,j}$, set of subscripts of nodes $P_{k,l}$ surrounding $P_{i,j}$,
- $|C_{i,j}|$, area of $C_{i,j}$,
- $\partial C_{i,j}$, boundary of $C_{i,j}$

We define the piecewise constant functions $U^n(x, y)$ and $z_{2d}(x, y)$ on cells $C_{i,j}$ corresponding to time t^n as

$$U^n(x, y) = U^n_{i,j}, \quad z_{2d}(x, y) = z_{2d,i,j}, \quad \text{for } (x, y) \in C_{i,j}, \tag{55}$$

with $U^n_{2d,i,j} = (h^n_{i,j}, q^n_{x,i,j}, q^n_{y,i,j})^T$ i.e.

$$U^n_{2d,i,j} \approx \frac{1}{|C_{i,j}|} \int_{C_{i,j}} U_{2d}(t^n, x, y)dxdy, \quad z_{2d,i,j} \approx \frac{1}{|C_{i,j}|} \int_{C_{i,j}} z_{2d}(x, y)dxdy,$$

with U_{2d} defined by (52).

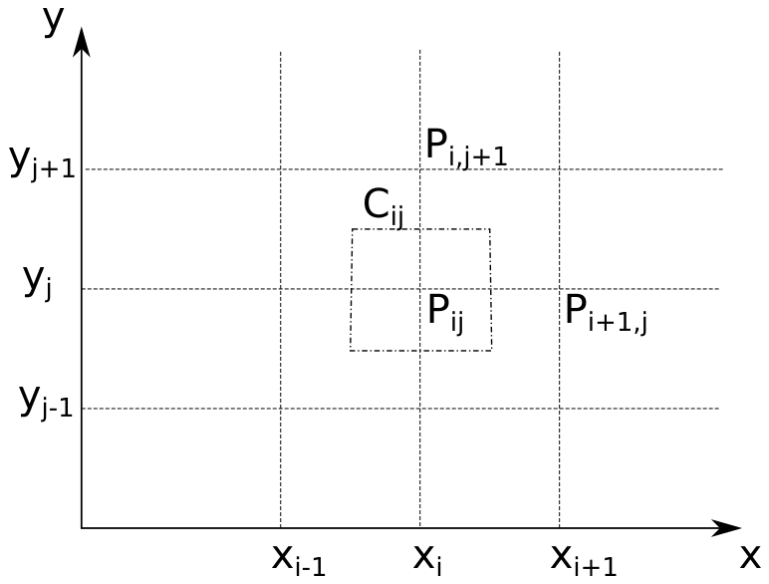

Figure 1: The vertices $\{P_{i,j}\}$ and the dual cell $C_{i,j}$.

Let $C_{i,j}$ be a dual cell of the structured mesh defined by the vertices $\{P_{i,j}\}$, see Fig. 1. In the case of a flat topography, the integral over $C_{i,j}$ of the convective part of the kinetic equation (6) gives

$$\int_{C_{i,j}} \left(\frac{\partial M_{2d}}{\partial t} + \begin{pmatrix} \xi \\ \gamma \end{pmatrix} . \nabla_{x,y} M_{2d} \right) dxdy \approx |C_{i,j}| \frac{\partial M_{2d,i,j}}{\partial t} + \sum_{(k,l) \in K_{i,j}} \int_{\partial C_{i,j}} M_{i,j,k,l} dl, \tag{56}$$

with $M_{2d,i,j} = M(U_{2d,i,j}, \xi, \gamma)$ and the upwinding formula

$$M_{i,j,k,l} = M_{2d,i,j} \zeta_{k,l} \mathbb{1}_{\zeta_{k,l} \geq 0} + M_{2d,k,l} \zeta_{k,l} \mathbb{1}_{\zeta_{k,l} \leq 0},$$

where $\zeta_{k,l} = (\xi,\gamma)^T \cdot \mathbf{n}_{k,l}$, $\mathbf{n}_{k,l}$ for $(k,l) \in K_{i,j}$ being the outward normal to the contour $\partial C_{i,j}$. The implicit Euler scheme applied to the kinetic interpretation (56) gives the kinetic scheme

$$f_{2d,i,j}^{n+1} = M_{2d,i,j}^{n} - \frac{\Delta t^n}{|C_{i,j}|} \sum_{(k,l)\in K_{i,j}} \Big(f_{2d,i,j}^{n+1} \zeta_{k,l} \mathbb{1}_{\zeta_{k,l}\geq 0} + f_{2d,k,l}^{n+1} \zeta_{k,l} \mathbb{1}_{\zeta_{k,l}\leq 0} \Big). \tag{57}$$

Denoting

$$f_{2d} = (f_{2d,1,1}, f_{2d,2,1}, \ldots, f_{2d,P,1}, f_{2d,2,1}, \ldots)^T,$$

the kinetic scheme (57) also writes

$$\Big(\mathbf{I}_{P^2} + \frac{\Delta t^n}{\Delta x} \mathbf{L}_{P^2} \Big) f_{2d}^{n+1} = M_{2d} + \sigma B_{2d}^{n+1}, \tag{58}$$

where we have used the particular geometry of the mesh and with I_{P^2} is the identity matrix of length P^2, B_{2d}^{n+1} accounts for the boundary conditions and the block matrix $\mathbf{L}_{P^2}$ is defined by

$$\mathbf{L}_{P^2} = \begin{pmatrix} D_{\xi,\gamma} & N_\gamma^+ & 0 & \ldots & \ldots \\ N_\gamma^- & D_{\xi,\gamma} & N_\gamma^+ & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & N_\gamma^- & D_{\xi,\gamma} & N_\gamma^+ \\ \ldots & \ldots & 0 & N_\gamma^- & D_{\xi,\gamma} \end{pmatrix},$$

where $D_{\xi,\gamma}, N_\gamma^\pm$ are $P \times P$ matrices defined by $N_\gamma^+ = -\gamma \mathbb{1}_{\gamma\geq 0} I_P$, $N_\gamma^- = \gamma \mathbb{1}_{\gamma\leq 0} I_P$ and

$$D_{\xi,\gamma} = \begin{pmatrix} |\xi|+|\gamma| & \xi \mathbb{1}_{\xi\leq 0} & 0 & \ldots & \ldots \\ -\xi \mathbb{1}_{\xi\geq 0} & |\xi|+|\gamma| & \xi \mathbb{1}_{\xi\leq 0} & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & -\xi \mathbb{1}_{\xi\geq 0} & |\xi|+|\gamma| & \xi \mathbb{1}_{\xi\leq 0} \\ \ldots & \ldots & 0 & -\xi \mathbb{1}_{\xi\geq 0} & |\xi|+|\gamma| \end{pmatrix}.$$

Since the matrix

$$\mathbf{I}_{P^2} + \frac{\Delta t^n}{\Delta x} \mathbf{L}_{P^2},$$

has the same structure as the matrix $\mathbf{I} + \sigma \mathbf{L}$ studied in Lemma 3.1, the results of Prop. 3.2 and 3.3 are valid.

We do not give the explicit formula neither for the inverse of the matrix $\mathbf{I}_{P^2} + \frac{\Delta t^n}{\Delta x} \mathbf{L}_{P^2}$ nor for the numerical fluxes at the macroscopic level.

5 An iterative resolution scheme

The kinetic scheme (24) requires to solve a linear system and in the previous section, we have seen that it was possible to have an analytic expression for the inverse of the matrix

$$\mathbf{I} + \sigma \mathbf{L}.$$

For the numerical approximation of PDEs e.g. in finite elements methods when the linear system to solve is large an iterative strategy is singled out compared to a direct inversion of the matrix. We propose to follow the same idea here, with mainly two benefits. First it will allow us to use the half disk Maxwellian (4), for which we recall the integrals (38) could not be computed analytically in the case of the fully implicit kinetic scheme. This is important as it will enable to prove some discrete entropy inequality at the macroscopic scale thanks to (13), while having an explicit writting of the update. The second advantage lies in the possibility to couple the iterative strategy with the

hydrostatic reconstruction to obtain a well balanced treatment for varying bottoms, which we will discuss in the next section.

More precisely using a Gauss-Jacobi type decomposition, let us rewrite

$$\mathbf{I} + \sigma\mathbf{L} = \mathbf{D} - \mathbf{N},$$

where $\mathbf{D}$ and $\mathbf{N}$ are two matrices from $\mathbb{R}^{P\times P}$ with $\mathbf{D}$ is invertible. Then the scheme (24) also writes

$$f^{n+1} = \mathbf{D}^{-1}\mathbf{N}f^{n+1} + \mathbf{D}^{-1}(M + \sigma B^{n+1}),$$

and if it converges, the sequence $\{f^{n+1,k}\}_{k\in\mathbb{N}}$ defined by

$$\mathbf{D}f^{n+1,k+1} = \mathbf{N}f^{n+1,k} + M + \sigma B^{n+1},$$

converges towards the solution of (24).

5.1 Case without topography

In this section, we study this iterative strategy with the particular choice

$$\mathbf{D} = (1+\alpha)\mathbf{I}, \quad \text{and} \quad \mathbf{N} = \alpha\mathbf{I} - \sigma\mathbf{L},$$

when the bathymetry is flat and where $\alpha \in \mathbb{R}_+$ is a relaxation parameter. When developped, this iterative process reads:

$$\begin{cases} f^{n+1,0} = M \\ (1+\alpha)f^{n+1,k+1} = (\alpha\mathbf{I} - \sigma\mathbf{L})f^{n+1,k} + M + \sigma B^{n+1,k} \\ \forall 1 \le i \le P,\ U_i^{n+1,k} = \displaystyle\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^{n+1,k}(\xi)\,\mathrm{d}\xi \end{cases}, \tag{59}$$

with $B^{n+1,k}$ the boundary condition associated with the macroscopic state $U^{n+1,k}$ as explained in Section 3.4. The following Proposition highlights the main compromise linked with such an iterative approach, which is the requirement for a CFL condition in order for the method to converge.

Proposition 5.1 *Assume that $B^{n+1,k}$ remains constant equal to B^n for any k in $\mathbb{N}$. Then (59) defines an arithmetico-geometric sequence which converges if the CFL condition $\sigma|\xi| < 1 + 2\alpha$ holds for all ξ belonging to* $\operatorname{supp} M \cup \operatorname{supp} B^n$.

Proof. By recurrence, we can show that for any $k \in \mathbb{N}$ the support of $f^{n+1,k}$ is included in $\operatorname{supp} M \cup \operatorname{supp} B^n$, which is why we restrict to velocities ξ belonging to this set. Consider f the solution of

$$f = \mathbf{D}^{-1}\mathbf{N}f + \mathbf{D}^{-1}(M + \sigma B^n)\ .$$

The sequence $(g^k)_k$ defined by $g^k = f^{n+1,k} - f$ satisfies $g^{k+1} = \mathbf{D}^{-1}\mathbf{N}g^k$ and converges to zero as soon as the spectral radius of $\mathbf{D}^{-1}\mathbf{N}$ is strictly less than one. Since $\mathbf{D}^{-1}\mathbf{N}$ is a triangular matrix, its eigenvalues are given by its diagonal coefficients, all equal to $(1+\alpha)^{-1}(\alpha - \sigma|\xi|)$. Under the assumption $\sigma|\xi| < 1 + 2\alpha$, this quantity is strictly less than one in absolute value, which concludes the proof. □

Remark 5.2 *As we did in Section 3.4 for the fully implicit scheme, we can replace $B^{n+1,k}$ by B^n in the iterative process (59). In fact this constitutes a first order approximation in time since we have $f^{n+1,k} = M + \mathcal{O}(\Delta t)$. Under this simplification, one can drop the assumption $B^{n+1,k} = B^n$ from Proposition 5.1.*

In practice, we wish to apply an iterative method directly at the macroscopic level. An issue with (59) is that the distribution involved in the kinetic flux (i.e. the term in factor of $\sigma\mathbf{L}$) is not a vector of Maxwellians, which prevents us to write the recurrence relation at the macroscopic level since there is no general expression for the numerical flux. To bypass this issue, we propose the following

modification of (59), where we replace all occurrences of $f^{n+1,k}$ on the right hand side by a vector of Maxwellians $M^{n+1,k}$, which defines a new sequence $(g^{n+1,k}(\xi))_{k\in\mathbb{N}}$ as

$$\left\{\begin{array}{l} g^{n+1,0}(\xi) = M \\ (1+\alpha)g^{n+1,k+1}(\xi) = (\alpha\mathbf{I} - \sigma^k\mathbf{L})M^{n+1,k} + M + \sigma^k B^{n+1,k} \\ M^{n+1,k+1} = g^{n+1,k+1} + \Delta t^k Q^{n+1,k+1} \end{array}\right. . \tag{60}$$

This new iterative process is alternating two stages, the first one being the usual transport step, while the second one is a projection step onto the set of Maxwellians yielding $M^{n+1,k+1}$. In this sense (60) is an iterative BGK splitting approach where the projection step doesn't modify the macroscopic quantities of interest since the term $Q^{n+1,k}$ is a vector of collision operators each one satisfying the conservation constraints (7). Note that the time stepping Δt^k is made dependent on k as the support of $M^{n+1,k}$ can now change from iteration to iteration. It is important to remark that this iterative scheme differs from (59) and we cannot apply the result of Proposition 5.1. The practical implementation of scheme (60) is based on its macroscopic version given for all $1 \leq i \leq P$ by

$$(1+\alpha)U_i^{n+1,k+1} = \alpha U_i^{n+1,k} + U_i - \sigma\Big(\mathcal{F}(U_i^{n+1,k}, U_{i+1}^{n+1,k}) - \mathcal{F}(U_{i-1}^{n+1,k}, U_i^{n+1,k})\Big), \tag{61}$$

where the numerical flux F is defined as

$$\mathcal{F}(U_L, U_R) = \int_{\mathbb{R}} \xi \begin{pmatrix} 1 \\ \xi \end{pmatrix} \Big(\mathbb{1}_{\xi>0} M(U_L,\xi) + \mathbb{1}_{\xi<0} M(U_R,\xi)\Big)\,\mathrm{d}\xi, \tag{62}$$

and where the vectors $U_0^{n+1,k}, U_{P+1}^{n+1,k}$ appearing for $i \in \{1, P\}$ are respectively functions of $U_1^{n+1,k}$ and $U_P^{n+1,k}$ since the boundary conditions are imposed though a ghost cell strategy fully described in [10, 1]. Notice that if the sequence $\{U^{n+1,k}\}_{k\in\mathbb{N}} \subset (\mathbb{R}^2)^P$ from (61) converges in $(\mathbb{R}^2)^P$, its limit U^{n+1} then satisfies

$$\forall 1 \leq i \leq P, \quad U_i^{n+1} = U_i^n - \sigma\Big(\mathcal{F}(U_i^{n+1}, U_{i+1}^{n+1}) - \mathcal{F}(U_{i-1}^{n+1}, U_i^{n+1})\Big)$$

by continuity of the numerical flux (62). Besides we want to remark that the iterative method described here could have been applied with any other numerical flux at the macroscopic level. However, using a numerical flux different from the kinetic one would have made it very difficult (if possible at all) to prove the forthcoming properties, whereas using (62) gives us a favorable setting to perform the proofs. These properties include the preservation of the water height positivity under a CFL condition, and the existence of a discrete entropy equality with dissipation.

Proposition 5.3 *Assume that the water height vectors h^n and $h^{n+1,k}$ are positive. Then the update $g^{n+1,k+1}$ defined in the iterative scheme (60) is positive if for all $1 \leq i \leq P$ the CFL condition $\sigma^k|\xi| \leq \alpha + M_i/M_i^{n+1,k}$ holds for any ξ belonging to $\operatorname{supp} M^{n+1,k}$. As a direct consequence, the water height vector $h^{n+1,k+1}$ from scheme (61) is positive under these assumptions.*

We postpone the proof of Proposition 5.3 to the next section, where it is generalized to the case with varying bottom in Proposition 5.6.

Proposition 5.4 *Let us denote $\Xi = \operatorname{supp} M^{n+1,k}$. The kinetic entropy of the iterative scheme (60) satisfies the following equality*

$$\begin{aligned} H(M_i^{n+1,k+1}) = \frac{H(M_i) + \alpha H(M_i^{n+1,k})}{1+\alpha} - \frac{\sigma^k|\xi|}{1+\alpha}\Big(H_{i+1/2}^{n+1,k} - H_{i-1/2}^{n+1,k}\Big) \\ + \Delta t^k \partial_f H(M_i^{n+1,k+1}) Q_i^{n+1,k+1} + D_i^{n+1,k+1}, \end{aligned} \tag{63}$$

with $Q_i^{n+1,k+1} = (g_i^{n+1,k+1} - M_i^{n+1,k+1})/\Delta t^k$ a collision operator verifying the conservation constraints (7). The interfacial kinetic entropies $H_{i\pm1/2}^{n+1,k}$ are given by

$$H_{i-1/2}^{n+1,k} = \mathbb{1}_{\xi>0} H(M_{i-1}^{n+1,k}, \xi) + \mathbb{1}_{\xi<0} H(M_i^{n+1,k}, \xi),$$

$$H_{i+1/2}^{n+1,k} = \mathbb{1}_{\xi>0} H(M_i^{n+1,k},\xi) + \mathbb{1}_{\xi<0} H(M_{i+1}^{n+1,k},\xi),$$

and the term $D_i^{n+1,k+1}$ *is given by*

$$\begin{aligned} D_i^{n+1,k+1} = & -\frac{1}{1+\alpha}\Psi(M_i^{n+1,k+1},M_i) - \frac{\alpha-\sigma^k|\xi|\mathbb{1}_\Xi}{1+\alpha}\Psi(M_i^{n+1,k+1},M_i^{n+1,k}) \\ & - \frac{\sigma^k|\xi|\mathbb{1}_\Xi}{1+\alpha}\Psi(M_i^{n+1,k+1},M_{i\pm1}^{n+1,k}), \end{aligned}$$

where we recall that the function Ψ *defined in (29) is positive on* $\mathbb{R}_+^2$ *and where* $i\pm1 = i-\operatorname{sgn}\xi$*. As a consequence, if for any integer* k *the CFL condition*

$$\forall \xi \in \operatorname{supp} M^{n+1,k},\ \sigma^k|\xi| \le \alpha \tag{64}$$

holds, then $D_i^{n+1,k+1}$ *is a dissipation term with negative sign and at each iteration the kinetic entropy is dissipated up to terms that are macroscopically zero, that is to say there exists a kinetic entropy flux* $\widetilde{H}_{i+1/2}^{n+1,k}$*, a negative dissipation* $\widetilde{D}_i^{n+1,k+1}$ *and a term* $\widetilde{Z}_i^{n+1,k+1}(\xi)$ *whose integral over* $\xi\in\mathbb{R}$ *is zero such that*

$$H(M_i^{n+1,k+1},\xi) = H(M_i,\xi) - \sigma^k|\xi|\Big(\widetilde{H}_{i+1/2}^{n+1,k} - \widetilde{H}_{i-1/2}^{n+1,k}\Big) + \widetilde{D}_i^{n+1,k+1} + \widetilde{Z}_i^{n+1,k+1}. \tag{65}$$

Before giving the proof we have the remark below.

Remark 5.5 *Even when the CFL condition (64) is not satisfied, we can ensure that the scheme (60) satisfies a discrete entropy inequality from some rank* k *assuming the convergence of the method. In fact, multiplying equality (63) by* $1+\alpha$ *it is possible to write*

$$\begin{aligned} & H(M_i^{n+1,k+1},\xi) = \\ & \quad H(M_i,\xi) - \sigma^k\xi\Big(H_{i+1/2}^{n+1,k} - H_{i-1/2}^{n+1,k}\Big) + (1+\alpha)\Delta t^k\,\partial_f H(M_i^{n+1,k+1},\xi)Q_i^{n+1,k+1} \\ & \quad - \Psi(M_i^{n+1,k+1},M_i) - \sigma^k|\xi|\mathbb{1}_\Xi\Psi(M_i^{n+1,k+1},M_{i\pm1}^{n+1,k}) \\ & \quad + \alpha\Big(H(M_i^{n+1,k},\xi) - H(M_i^{n+1,k+1},\xi)\Big) - (\alpha-\sigma^k|\xi|\mathbb{1}_\Xi)\Psi(M_i^{n+1,k+1},M_i^{n+1,k}). \end{aligned} \tag{66}$$

When $M_i^{n+1,k}$ *is a half-disk Maxwellian (4), the term*

$$(1+\alpha)\Delta t^k\,\partial_f H(M_i^{n+1,k},\xi)Q_i^{n+1,k+1},$$

appearing on the right hand side of (66) does not cause any issue as it vanishes upon integration over $\xi\in\mathbb{R}$*. This is intrinsically related to the form of the half-disk Maxwellian* $M_i^{n+1,k+1}$ *which makes* $\partial_f H(M_i^{n+1,k+1},\xi)$ *linear in* ξ *over the support of* $M_i^{n+1,k}$*, i.e. when* $M_i^{n+1,k}>0$ *one has*

$$\begin{aligned} \partial_f H(M_i^{n+1,k+1},\xi) &= \frac{\xi^2}{2} + \frac{g^2\pi^2}{2}\left(\frac{1}{g\pi}\sqrt{2gh_i^{n+1,k+1} - (\xi-u_i^{n+1,k+1})^2}\right)^2 \\ &= gh_i^{n+1,k+1} + u_i^{n+1,k+1}\xi - \frac{(u_i^{n+1,k+1})^2}{2}\ . \end{aligned}$$

Furthermore we remind that $Q_i^{n+1,k+1}$ *satisfies the conservation constraints (7), meaning that its integral against* $(1,\xi)^T$ *vanishes. Therefore in (66) the only problematic terms are contained in the last line, as their sign can be positive and they have no reason to cancel after integration. Nevertheless, by regularity of* $H(\cdot,\xi)$ *and by definition (29) of* Ψ*, these terms write as a* $\mathcal{O}(M_i^{n+1,k+1} - M_i^{n+1,k})$ *and vanish as* $k\to\infty$ *assuming the method converges. As a consequence, from some rank* k *these two terms become negligible compared to* $-\Psi(M_i^{n+1,k+1},M_i)<0$ *which remains bounded away from zero, and we recover a dissipation with negative sign. (Note that if* $M^{n+1,k}$ *was converging to* M *as* $k\to\infty$*, it would imply that* M *solves the fixed point problem and thus* $M^{n+1,k}=M$ *for all* k*; putting aside this trivial case, this is why* $\Psi(M_i^{n+1,k+1},M_i)$ *remains bounded away from zero).*

Proof of prop. 5.4. We first prove equality (63). For this we write the subiteration (60) as

$$M_i^{n+1,k+1} = \frac{1}{1+\alpha}\Big(M_i + (\alpha - \sigma^k|\xi|\mathbb{1}_\Xi)M_i^{n+1,k} + \sigma^k|\xi|\mathbb{1}_\Xi M_{i\pm1}^{n+1,k}\Big) + \Delta t^k Q_i^{n+1,k+1},$$

with $i \pm 1 = i - \operatorname{sgn}\xi$ and $Q_i^{n+1,k+1}$ a collision operator. Applying Lemma 3.4 for $a = M_i^{n+1,k+1}$ and $b = M_{i\pm1}^{n+1,k}, M_i, M_i^{n+1,k}$, we respectively get:

$$\begin{aligned} H(M_{i\pm1}^{n+1,k}) = H(M_i^{n+1,k+1}) &+ \partial_f H(M_i^{n+1,k+1})(M_{i\pm1}^{n+1,k} - M_i^{n+1,k+1}) \\ &+ \Psi(M_i^{n+1,k+1}, M_{i\pm1}^{n+1,k}) \end{aligned} \tag{67}$$

$$\begin{aligned} H(M_i^{n+1,k}) = H(M_i^{n+1,k+1}) &+ \partial_f H(M_i^{n+1,k+1})(M_i^{n+1,k} - M_i^{n+1,k+1}) \\ &+ \Psi(M_i^{n+1,k+1}, M_i^{n+1,k}) \end{aligned} \tag{68}$$

$$\begin{aligned} H(M_i) = H(M_i^{n+1,k+1}) &+ \partial_f H(M_i^{n+1,k+1})(M_i - M_i^{n+1,k+1}) \\ &+ \Psi(M_i^{n+1,k+1}, M_i) \end{aligned} \tag{69}$$

Performing the linear combination

$$\frac{1}{1+\alpha}\Big((69) + (\alpha - \sigma^k|\xi|\mathbb{1}_\Xi)(68)) + \sigma^k|\xi|\mathbb{1}_\Xi(67)\Big)$$

we obtain

$$\begin{aligned} &\frac{1}{1+\alpha}\Big(H(M_i) + (\alpha - \sigma^k|\xi|\mathbb{1}_\Xi)H(M_i^{n+1,k}) + \sigma^k|\xi|\mathbb{1}_\Xi H(M_{i\pm1}^{n+1,k})\Big) = \\ &\qquad H(M_i^{n+1,k+1}) - \Delta t^k \partial_f H(M_i^{n+1,k+1})Q_i^{n+1,k+1} + \frac{1}{1+\alpha}\Psi(M_i^{n+1,k+1}, M_i) \\ &\qquad + \frac{\alpha - \sigma^k|\xi|\mathbb{1}_\Xi}{1+\alpha}\Psi(M_i^{n+1,k+1}, M_i^{n+1,k}) + \frac{\sigma^k|\xi|\mathbb{1}_\Xi}{1+\alpha}\Psi(M_i^{n+1,k+1}, M_{i\pm1}^{n+1,k}) \end{aligned}$$

which corresponds to equality (63) after rearranging the terms.

Next we proceed by induction to show that the kinetic entropy is dissipated at every iteration assuming the CFL condition (64) holds for any integer k. The key argument is that under this CFL condition, the term $D_i^{n+1,k+1}$ defines a convex combination of negative quantities, and is thus negative. The initialization is obvious since we have $M_i^{n+1,0} = M_i$, so we focus on the recurrence. We want to show that (65) holds at some rank $k \geq 1$ assuming that it is satisfied at rank $k-1$. Under this assumption we can recast (63) as

$$\begin{aligned} &H(M_i^{n+1,k+1}) = \\ &\quad \frac{1}{1+\alpha}\Big(H(M_i) + \alpha\Big(H(M_i) - \sigma^k|\xi|\Big(\widetilde{H}_{i+1/2}^{n+1,k-1} - \widetilde{H}_{i-1/2}^{n+1,k-1}\Big) + \widetilde{D}_i^{n+1,k} + \widetilde{Z}_i^{n+1,k}\Big)\Big) \\ &\quad - \frac{\sigma^k|\xi|}{1+\alpha}\Big(H_{i+1/2}^{n+1,k} - H_{i-1/2}^{n+1,k}\Big) + \Delta t^k \partial_f H(M_i^{n+1,k+1})Q_i^{n+1,k+1} + D_i^{n+1,k+1}, \end{aligned}$$

with $\widetilde{Z}_i^{n+1,k}$ and $\partial_f H(M_i^{n+1,k+1})Q_i^{n+1,k+1}$ macroscopically zero as per Remark 5.5. Therefore we have

$$\begin{aligned} &H(M_i^{n+1,k+1}) = \\ &\quad H(M_i) - \frac{\sigma^k|\xi|}{1+\alpha}\Big(H_{i+1/2}^{n+1,k} + \alpha\widetilde{H}_{i+1/2}^{n+1,k-1} - H_{i-1/2}^{n+1,k} - \alpha\widetilde{H}_{i-1/2}^{n+1,k-1}\Big) \\ &\quad + \frac{\alpha}{1+\alpha}\widetilde{Z}_i^{n+1,k} + \Delta t^k \partial_f H(M_i^{n+1,k+1})Q_i^{n+1,k+1} + \frac{\alpha}{1+\alpha}\widetilde{D}_i^{n+1,k} + D_i^{n+1,k+1} \end{aligned}$$

and the proof is complete by setting

$$\widetilde{H}_{i+1/2}^{n+1,k} = \frac{1}{1+\alpha}\Big(H_{i+1/2}^{n+1,k} + \alpha\widetilde{H}_{i+1/2}^{n+1,k-1}\Big), \quad \widetilde{D}_i^{n+1,k+1} = \frac{\alpha}{1+\alpha}\widetilde{D}_i^{n+1,k} + D_i^{n+1,k+1},$$

$$\widetilde{Z}_i^{n+1,k+1} = \frac{\alpha}{1+\alpha}\widetilde{Z}_i^{n+1,k} + \Delta t^k \partial_f H(M_i^{n+1,k+1})Q_i^{n+1,k+1}.$$

□

5.2 Case with topography

To deal with varying bathymetries in a well balanced way, the hydrostatic reconstruction technique introduced by Audusse et al. in [2] can be used. It is based on the reconstruction of the water height according to a procedure that we briefly recall. Let $U_i = (h_i, h_i u_i)^T \in \mathbb{R}^2$ denote the vector of quantities of interest over cell $1 \leq i \leq P$, with P the number of interior cells and with ghost cells corresponding to indices 0 and $P+1$. The reconstructed states are vectors from $\mathbb{R}^2$ defined on the left and right neighborhood of each cell interface as follows:

$$\forall 1 \leq i \leq P, \quad U_{i+1/2-} = \begin{pmatrix} h_{i+1/2-} \\ h_{i+1/2-} u_i \end{pmatrix}, \quad U_{i-1/2+} = \begin{pmatrix} h_{i-1/2+} \\ h_{i-1/2+} u_i \end{pmatrix} . \tag{70}$$

The reconstructed interfacial water heights are given by

$$h_{i-1/2+} = (h_i + z_i - z_{i-1/2})_+, \quad h_{i+1/2-} = (h_i + z_i - z_{i+1/2})_+, \tag{71}$$

with the interfacial bathymetry variation $z_{i+1/2} = \max(z_i, z_{i+1})$. The truly implicit kinetic scheme we are considering reads as below

$$U_i^{n+1} = U_i^n - \sigma\big(F_{i+1/2-}^{n+1} - F_{i-1/2+}^{n+1}\big), \tag{72}$$

with $\sigma = \Delta t/\Delta x$ and the numerical fluxes decomposed as:

$$F_{i+1/2-}^{n+1} = \mathcal{F}(U_{i+1/2-}^{n+1}, U_{i+1/2+}^{n+1}) + \frac{g}{2}\begin{pmatrix} 0 \\ (h_i^{n+1})^2 - (h_{i+1/2-}^{n+1})^2 \end{pmatrix}$$
$$F_{i-1/2+}^{n+1} = \mathcal{F}(U_{i-1/2-}^{n+1}, U_{i-1/2+}^{n+1}) + \frac{g}{2}\begin{pmatrix} 0 \\ (h_i^{n+1})^2 - (h_{i-1/2+}^{n+1})^2 \end{pmatrix}$$

We recall that in our case the upwinding of the numerical flux $\mathcal{F}$ is induced at the kinetic level according to definition (62).

Because the update (72) is nonlinear, it is not possible to solve it analytically. Instead we will consider an iterative process with a relaxation parameter $\alpha > 0$ similarly to section 5.1. At the kinetic level, this process consists in introducing for any real ξ the sequence $(f^{n+1,k}(\xi))_{k\in\mathbb{N}} \subset \mathbb{R}_+^P$ initialized with $f^{n+1,0}(\xi) = M(U^n, \xi)$ and defined recursively as:

$$\begin{aligned} &(1+\alpha) f_i^{n+1,k+1} = \\ &\quad M_i + \alpha M_i^{n+1,k} - \sigma^k \xi \Big(\mathbb{1}_{\xi<0}(M_{i+1/2+}^{n+1,k} - M_{i-1/2+}^{n+1,k}) + \mathbb{1}_{\xi>0}(M_{i+1/2-}^{n+1,k} - M_{i-1/2-}^{n+1,k}) \Big) \\ &\quad + \sigma^k(\xi - u_i^{n+1,k})(M_{i+1/2-}^{n+1,k} - M_i^{n+1,k}) - \sigma^k(\xi - u_i^{n+1,k})(M_{i-1/2+}^{n+1,k} - M_i^{n+1,k}), \end{aligned} \tag{73}$$

where the last line of (73) corresponds to the kinetic interpretation of the topography source term, see [3]. In the above we used the notation

$$M_\square^{n+1,k} = M(U_\square^{n+1,k}, \xi), \quad h^{n+1,k} = \int_{\mathbb{R}} f^{n+1,k}(\xi)\,\mathrm{d}\xi, \quad (hu)^{n+1,k} = \int_{\mathbb{R}} \xi\, f^{n+1,k}(\xi)\,\mathrm{d}\xi,$$

where the square symbol "$\square$" in subscript can be replaced by i (centered value) or $i \pm 1/2\mp$ (reconstructed interfacial value). Making use of the relations

$$\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\xi - u_i)(M_i - M_{i+1/2-})\,\mathrm{d}\xi = \begin{pmatrix} 0 \\ \frac{g}{2}(h_i^2 - h_{i+1/2-}^2) \end{pmatrix}$$
$$\int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} (\xi - u_i)(M_i - M_{i-1/2+})\,\mathrm{d}\xi = \begin{pmatrix} 0 \\ \frac{g}{2}(h_i^2 - h_{i-1/2+}^2) \end{pmatrix}$$

given in [3], the macroscopic version of scheme (73) obtained by integrating the update against the vector $(1, \xi)^T$ reads

$$(1+\alpha) U_i^{n+1,k} = U_i + \alpha U_i^{n+1,k} - \sigma^k (F_{i+1/2-}^{n+1,k} - F_{i-1/2+}^{n+1,k}). \tag{74}$$

If the iterative process (74) converges, we recover the implicit scheme (72) by setting the macroscopic update as $U^{n+1} = \lim_{k\to\infty} U^{n+1,k}$ and $\sigma = \lim_{k\to\infty} \sigma^k$. In practice we will not be able to compute this limit, hence we will set $U^{n+1} = U^{n+1,k}$ and $\sigma = \sigma^k$ for k large enough, hoping that $U^{n+1,k} \approx U^{n+1,\infty}$. One should also notice that when the bathymetry is flat the hydrostatic reconstruction becomes transparent, and the scheme (74) coincides with (61).

Finally we propose an estimate for the entropy associated with the scheme (73), as well as a CFL condition to ensure its positivity.

Proposition 5.6 *The following properties are satisfied by the scheme (73):*

(i) Assume that the water height vectors h^n and $h^{n+1,k}$ are positive. Then the update $g^{n+1,k+1}$ defined in the iterative scheme (73) is positive if for all $1 \le i \le P$ the CFL condition $\sigma^k|\xi| \le \alpha + M_i/M_i^{n+1,k}$ holds for any ξ belonging to $\operatorname{supp} M^{n+1,k}$. As a direct consequence, the water height vector $h^{n+1,k+1}$ from scheme (74) is positive under these assumptions.

(ii) The kinetic entropy of the iterative process (73) verifies the following kinetic entropy inequality

$$
\begin{aligned}
&H(M_i^{n+1,k+1}, z_i) \le \\
&\quad H(M_i, z_i) - \sigma^k\Big(\widetilde{G}_{i+1/2-}^{n+1,k} - \widetilde{G}_{i-1/2+}^{n+1,k}\Big) + (1+\alpha)\Delta t^k \partial_f H(M_i^{n+1,k}, z_i) Q_i^{n+1,k+1} \\
&\quad + \alpha\Big(H(M_i^{n+1,k}, z_i) - H(M_i^{n+1,k+1}, z_i)\Big) + (1+\alpha)\Psi(M_i^{n+1,k}, M_i^{n+1,k+1}) \\
&\quad - \Psi(M_i^{n+1,k}, M_i),
\end{aligned}
\tag{75}
$$

with $Q_i^{n+1,k+1} = (M_i^{n+1,k+1} - f_i^{n+1,k+1})/\Delta t^k$ a collision term satisfying the conservation constraints (7), and where

$$
\begin{aligned}
&\widetilde{G}_{i+1/2-}^{n+1,k} = \\
&\quad \xi \mathbb{1}_{\xi<0} H(M_{i+1/2+}^{n+1,k}, z_{i+1/2}) + \xi \mathbb{1}_{\xi>0} H(M_{i+1/2-}^{n+1,k}, z_{i+1/2}) \\
&\quad + \xi H(M_i^{n+1,k}, z_i) - \xi H(M_{i+1/2-}^{n+1,k}, z_{i+1/2}) \\
&\quad + \Big(\nabla\eta(U_i^{n+1,k})^T \begin{pmatrix} 1 \\ \xi \end{pmatrix} + g z_i\Big)(\xi M_{i+1/2-}^{n+1,k} - \xi M_i^{n+1,k} + (\xi - u_i^{n+1,k})(M_i^{n+1,k} - M_{i+1/2-}^{n+1,k}))\ ,
\end{aligned}
\tag{76}
$$

$$
\begin{aligned}
&\widetilde{G}_{i-1/2+}^{n+1,k} = \\
&\quad \xi \mathbb{1}_{\xi<0} H(M_{i-1/2+}^{n+1,k}, z_{i-1/2}) + \xi \mathbb{1}_{\xi>0} H(M_{i-1/2-}^{n+1,k}, z_{i-1/2}) \\
&\quad + \xi H(M_i^{n+1,k}, z_i) - \xi H(M_{i-1/2+}^{n+1,k}, z_{i-1/2}) \\
&\quad + \Big(\nabla\eta(U_i^{n+1,k})^T \begin{pmatrix} 1 \\ \xi \end{pmatrix} + g z_i\Big)(\xi M_{i-1/2+}^{n+1,k} - \xi M_i^{n+1,k} + (\xi - u_i^{n+1,k})(M_i^{n+1,k} - M_{i-1/2+}^{n+1,k}))\ .
\end{aligned}
\tag{77}
$$

We recall that the expression of the function Ψ was given in (29) and that the entropy is $\eta(U) = \frac{hu^2}{2} + \frac{g}{2}h^2$.

Before giving the proof we make the following remark.

Remark 5.7 *In inequality (75) the difference $\widetilde{G}_{i+1/2-}^{n+1,k} - \widetilde{G}_{i-1/2+}^{n+1,k}$ is non conservative at the kinetic level, but becomes conservative when it is integrated over $\xi \in \mathbb{R}$. This is due do the fact that the last two lines of (76) and (77) are macroscopically zero, see [3] Proposition 3.1. Furthermore, we reiterate the comments made in Remark 5.5 which are to say that in (75) the term*

$$(1+\alpha)\Delta t^k \partial_f H(M_i^{n+1,k}, z_i) Q_i^{n+1,k+1}$$

is macroscopically zero for the half-disk Maxwellian (4). Besides, the quantity

$$\alpha\Big(H(M_i^{n+1,k}, z_i) - H(M_i^{n+1,k+1}, z_i)\Big) - \Psi(M_i^{n+1,k}, M_i) + (1+\alpha)\Psi(M_i^{n+1,k}, M_i^{n+1,k+1})$$

will eventually become negative for k large enough similarly to the argument from Remark 5.5. Integrating inequality (75) over $\xi \in \mathbb{R}$, this implies that there exists $K \in \mathbb{N}$ such that for any $k \geq K$ the fully discrete entropy inequality

$$\eta(U_i^{n+1,k+1}) \leq \eta(U_i^n) - \sigma^k \left(\int_{\mathbb{R}} \xi \, H_{i+1/2}^{n+1,k}(\xi) \, \mathrm{d}\xi - \int_{\mathbb{R}} \xi \, H_{i-1/2}^{n+1,k}(\xi) \, \mathrm{d}\xi \right) \tag{78}$$

is satisfied at the macroscopic level. Summing inequality (78) over every cell $1 \leq i \leq P$ we obtain the dissipation of the total energy up to boundary fluxes

$$\frac{1}{\Delta t^k} \sum_{i=1}^{P} \Big(\eta(U_i^{n+1,k+1}) - \eta(U_i^n) \Big) + \frac{1}{\Delta x} \left(\int_{\mathbb{R}} \xi \, H_{P+1/2}^{n+1,k}(\xi) \, \mathrm{d}\xi - \int_{\mathbb{R}} \xi \, H_{1/2}^{n+1,k}(\xi) \, \mathrm{d}\xi \right) \leq 0. \tag{79}$$

In addition to the usual tolerance criterion where the iterations are stopped whenever two successive iterates are sufficiently close to each other, we can use (79) as a complementary condition to ensure the dissipation of total energy.

Proof. The proof makes use of the kinetic writing (73) of scheme (74).

(i) Remarking that the quantity $\sigma^k(\xi - u_i^{n+1,k})(M_{i+1/2-}^{n+1,k} - M_{i-1/2+}^{n+1,k})$ appearing in the last line of (73) defines an odd function of $\xi - u_i^{n+1,k}$, its integral over $\xi \in \mathbb{R}$ vanishes and we have at the macroscopic level

$$(1+\alpha) h_i^{n+1,k+1} = \int_{\mathbb{R}} \Big(M_i + \alpha M_i^{n+1,k} - \sigma^k \xi (M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k}) \Big) \, \mathrm{d}\xi \ .$$

Thus it is enough to prove the positivity of the integrand, whose developed form is

$$\begin{aligned} M_i + \alpha M_i^{n+1,k} - \sigma^k \xi \Big(\mathbb{1}_{\xi>0} M_{i+1/2-}^{n+1,k} - \mathbb{1}_{\xi<0} M_{i-1/2+}^{n+1,k} \Big) \\ + \sigma^k \xi \Big(\mathbb{1}_{\xi>0} M_{i-1/2-}^{n+1,k} - \mathbb{1}_{\xi<0} M_{i+1/2+}^{n+1,k} \Big). \end{aligned}$$

By definition of the water height reconstruction (71), we have the inequalities $h_{i+1/2-}^{n+1,k} \leq h_i^{n+1,k}$ and $h_{i-1/2+}^{n+1,k} \leq h_i^{n+1,k}$. As a consequence $M_{i+1/2-}^{n+1,k} \leq M_i^{n+1,k}$ and $M_{i-1/2+}^{n+1,k} \leq M_i^{n+1,k}$, which allows us to bound the integrand from below by

$$M_i + \alpha M_i^{n+1,k} - \sigma^k |\xi| M_i^{n+1,k} \ .$$

If ξ does not belong to $\operatorname{supp} M^{n+1,k}$ this quantity equals M_i which is positive. Otherwise, it is made positive under the condition $\sigma^k |\xi| \leq \alpha + M_i^0 / M_i^{n+1,k}$ which gives the desired result.

(ii) We start to rewrite (73) as

$$\begin{aligned} &(1+\alpha)(M_i^{n+1,k+1} - M_i^{n+1,k}) = \\ &\quad (M_i - M_i^{n+1,k}) - \sigma^k \xi (M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k}) + \sigma^k (\xi - u_i^{n+1,k})(M_{i+1/2-}^{n+1,k} - M_{i-1/2+}^{n+1,k}) \\ &\quad + (1+\alpha) \Delta t^k Q_i^{n+1,k+1} \ . \end{aligned} \tag{80}$$

The strategy is to multiply (80) by $\partial_f H(M_i^{n+1,k}, z_i)$ and to write

$$\begin{aligned} \partial_f H(M_i^{n+1,k}, z_i) \Big[(1+\alpha)(M_i^{n+1,k+1} - M_i^{n+1,k}) - (M_i - M_i^{n+1,k}) - (1+\alpha) \Delta t^k Q_i^{n+1,k+1} \Big] = \\ - \sigma^k \partial_f H(M_i^{n+1,k}, z_i) \Big[\xi (M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k}) + \delta M_{i+1/2-}^{n+1,k} - \delta M_{i-1/2+}^{n+1,k} \Big] \ , \end{aligned} \tag{81}$$

where we defined

$$\delta M_{i+1/2-}^{n+1,k} = (\xi - u_i^{n+1,k})(M_i^{n+1,k} - M_{i+1/2-}^{n+1,k})$$

$$\delta M_{i-1/2+}^{n+1,k} = (\xi - u_i^{n+1,k})(M_i^{n+1,k} - M_{i-1/2+}^{n+1,k}).$$

We use formula (31) in the left hand side of (81) to get

$$\begin{aligned}
&\partial_f H(M_i^{n+1,k}, z_i)\Big[(1+\alpha)(M_i^{n+1,k+1} - M_i^{n+1,k}) - (M_i - M_i^{n+1,k}) - (1+\alpha)\Delta t^k Q_i^{n+1,k+1}\Big] = \\
&\quad (1+\alpha)\Big(H(M_i^{n+1,k+1}, z_i) - H(M_i^{n+1,k}, z_i) - \Psi(M_i^{n+1,k}, M_i^{n+1,k+1})\Big) \\
&\quad - \Big(H(M_i, z_i) - H(M_i^{n+1,k}, z_i) - \Psi(M_i^{n+1,k}, M_i)\Big) \\
&\quad - (1+\alpha)\Delta t^k \partial_f H(M_i^{n+1,k}, z) Q_i^{n+1,k+1} \ . \qquad (82)
\end{aligned}$$

Furthermore, an upper bound on the right hand side of (81) is obtained by applying Proposition 3.1 from [3] which directly yields

$$\begin{aligned}
&- \partial_f H(M_i^{n+1,k}, z_i)\Big[\xi(M_{i+1/2}^{n+1,k} - M_{i-1/2}^{n+1,k}) + \delta M_{i+1/2-}^{n+1,k} - \delta M_{i-1/2+}^{n+1,k}\Big] \\
&\qquad \leq \widetilde{G}_{i-1/2+}^{n+1,k} - \widetilde{G}_{i+1/2-}^{n+1,k} \ , \qquad (83)
\end{aligned}$$

with $\widetilde{G}_{i+1/2-}^{n+1,k}$ and $\widetilde{G}_{i-1/2+}^{n+1,k}$ defined by (76) and (77). Injecting equality (82) and inequality (83) into (81) we obtain the desired kinetic entropy inequality (75).

□

6 Numerical examples

6.1 The one dimensional case

We start by evaluating the qualitative properties and the efficiency related to the fully implicit and iterative kinetic schemes in the one dimensional case.

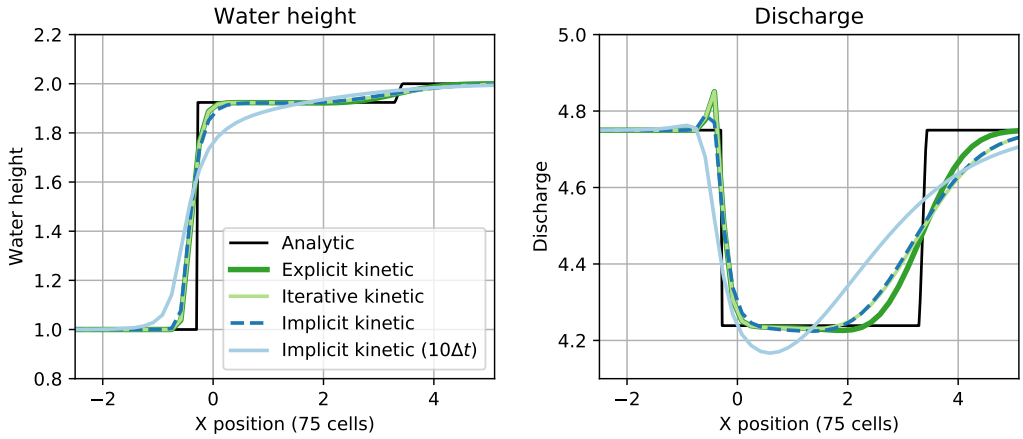

Figure 2: Slow moving shock approximated by various kinetic schemes, including explicit, implicit and iterative strategies. The initial condition is given by a Riemann data with discontinuity at position $x = 0$.

Slow moving shock. To assess the efficiency and interest of the implicit scheme (23), we perform a numerical test involving a Riemann problem with a slowly moving shock over a flat bottom. This configuration is achieved for a nearly transcritical flow where the material velocity u is positive and satisfies $u - \sqrt{gh} \approx 0$ and $u + \sqrt{gh} \gg 1$. Hence the maximum eigenvalue severely constrains the time

step, however a small time step might not be necessary to accurately resolve the slow shock. In Figure 2 we compare several schemes with an explicit time step Δt_{exp} given by the usual CFL condition, as well as the implicit kinetic scheme using a time step $\Delta t_{\text{imp}} = 10\Delta t_{\text{exp}}$. We set $\alpha = 1$ for the iterative scheme (61). We notice that in the discharge profile, an oscillation appears downwind of the shock, which is quite pronounced for the explicit and iterative kinetic schemes, and less so for the fully implicit ones. As expected the implicit scheme using Δt_{imp} strongly diffuses the fast travelling rarefaction. On the other hand the slow shock seems to be slightly less impacted by the large time steps, however it is still less diffused when using Δt_{exp}. Despite requiring ten times less iterations to reach the final time, the use of large time increments for the implicit kinetic scheme only results in arround two percents faster computations compared to the explicit strategy which is due to the high quadratic cost of the implicit method. We believe that it is not possible to lower this cost when it comes to unconditionally stable methods, because the associated stencil has to cover the entire computational domain.

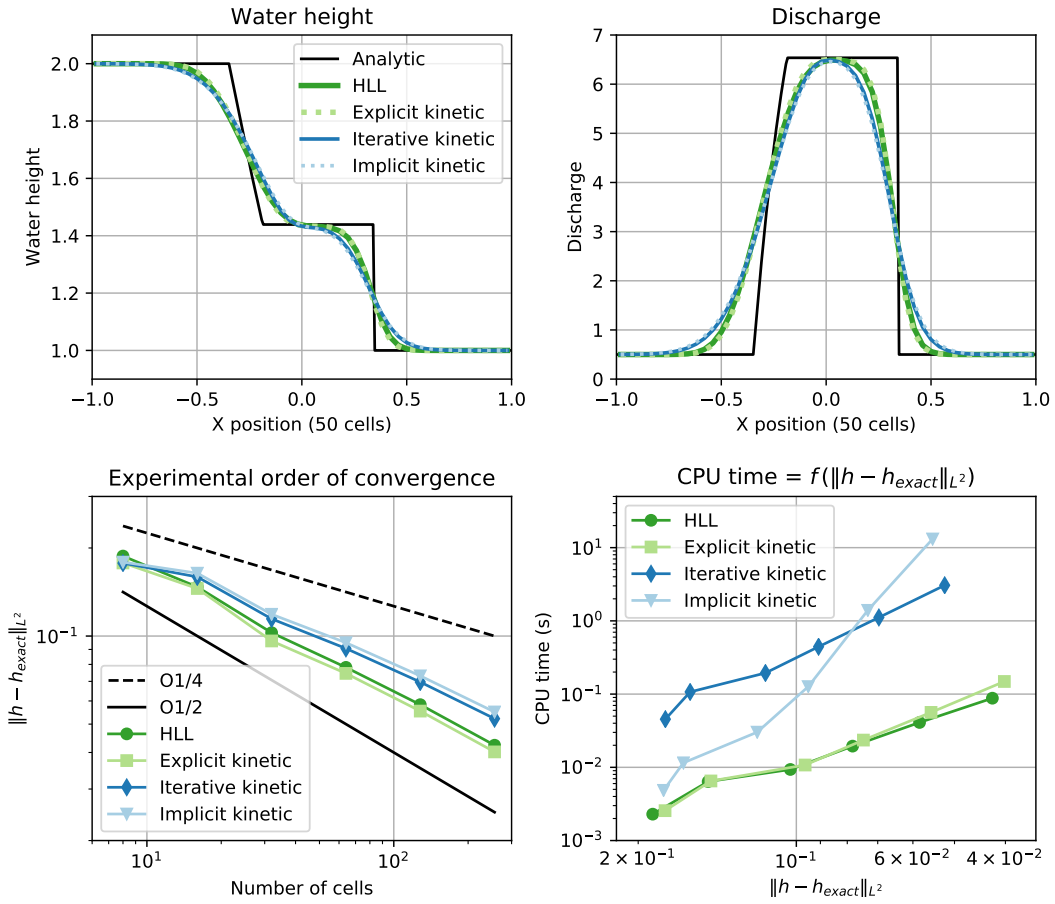

Figure 3: Comparing implicit, iterative and explicit kinetic solvers on a Riemann problem.

Riemann problem. We compare the fully implicit kinetic scheme and iterative kinetic scheme to explicit methods. The testcase is given by the Riemann problem with initial data $U^0(x) = \mathbb{1}_{x<0} U_L + \mathbb{1}_{x>0} U_R$ where we define

$$U_L = \begin{pmatrix} 2 \\ 1/2 \end{pmatrix}, \quad U_R = \begin{pmatrix} 1 \\ 1/2 \end{pmatrix}.$$

The solutions consists in a 1-rarefaction and a 2-shock. The iterative kinetic scheme uses the half-disk Maxwellian, and we choose the parameters $\alpha = 1$ and $\varepsilon_{\text{tol}} = 10^{-9}$ for the stopping criterion. All the

schemes use an explicit time step, and the results are given in Figure 3. Three aspects have to be considered, namely the accuracy, the computational cost and the stability. In the plotted curves, we see that in terms of efficiency both iterative and implicit kinetic schemes are at their disadvantage. Especially, the quadratic complexity of the fully implicit version results in a steeper slope of the efficiency curve. However this is only one part of the picture, and we know from Proposition 5.4 and Remark 5.5 that the iterative kinetic scheme (61) satisfies a discrete entropy inequality without restriction on the time step, assuming enough iterations are performed. Concretely the greater stability comes with a higher level of diffusion which is noticeable in the first two plots of Figure 3. This increased diffusion remains within acceptable margin, and is the price to pay to have better stability properties.

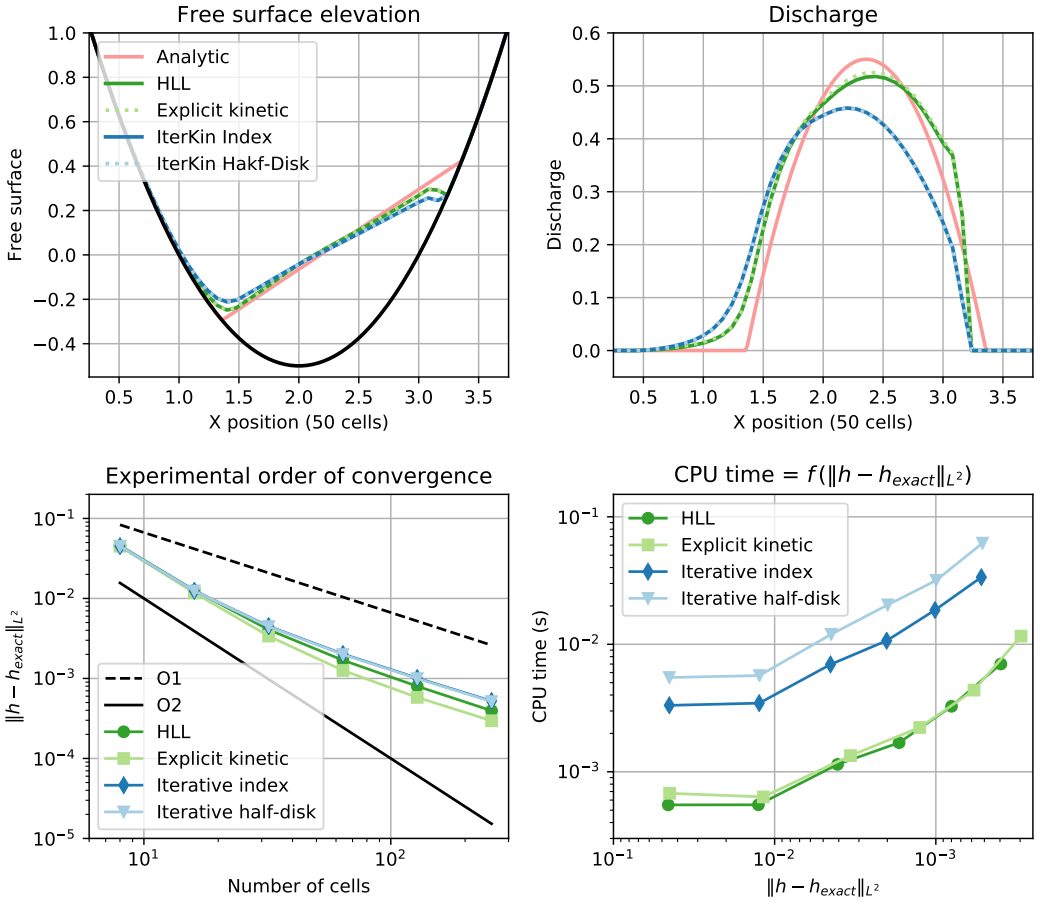

Figure 4: Parabolic bowl approximated by explicit and iterative kinetic schemes. First row: elevation and discharge at time 0.75, second row: convergence and efficiency curves. The stopping criteria used in the two kinetic iterative schemes combines the standard tolerance condition with tolerance $\varepsilon = 10^{-9}$ and the entropy condition (79).

Parabolic bowl. Next we consider Thacker's testcase, also known as the parabolic bowl testcase, taken from [13]. We plot the numerical solution at time 0.75 in Figure 4. This testcase is relevant as it provides us with a non trivial analytical solution enabling to plot convergence curves, and it is known to be challenging numerically, as it presents a varying bottom together with an evolving wet/dry front and a discontinuous velocity profile. It is interesting to note that the different choice of Maxwellian used in the two iterative kinetic schemes has very little impact on the approximation. In both cases we obtain a convergence with first order accuracy, and unsurprisingly the numerical cost is higher than for fully explicit methods due to the number of subiterations required to update the solution. One should also note that the use of the half-disk Maxwellian is slightly more expensive than the simpler

index Maxwellian. Besides, in this testcase the iterative kinetic scheme with index Maxwellian was always able to fulfill the entropy condition (79) after some iterations, which we only proved rigorously for the half-disk Maxwellian. Hence despite using the wrong Maxwellian, it seems that the iterative kinetic scheme in question still has better stability properties than fully explicit methods. This will be further corroborated with the next testcase.

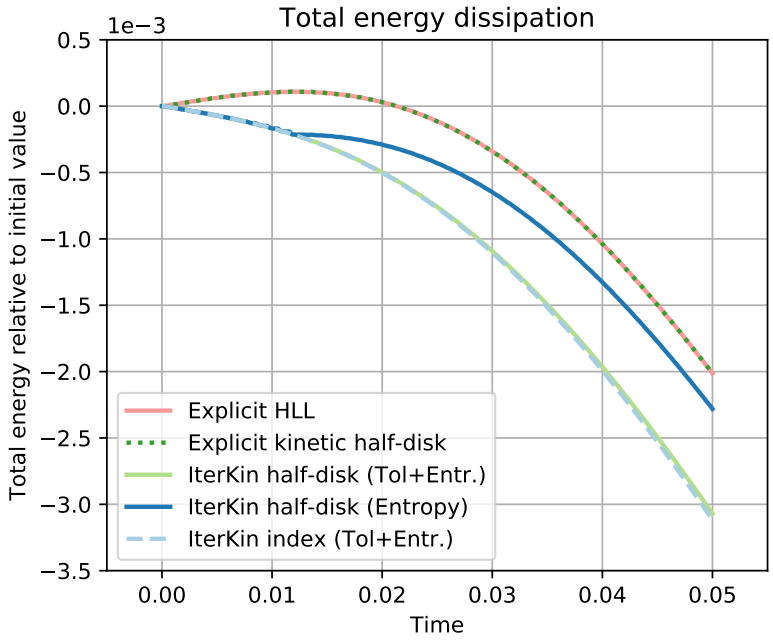

Figure 5: Evolution of the relative total energy obtained for various explicit and iterative kinetic schemes.

Total energy dissipation. Given the efficiency curves shown in Figures 3 and 4, the explicit strategy seems preferable in terms of the computational cost at a prescribed accuracy. However we have to stress that among all the considered methods, the iterative kinetic scheme with half-disk Maxwellian is the only one for which we can prove existence of a fully discrete entropy inequality for a large enough but finite number of iterations. We remind that on the opposite, the explicit kinetic scheme with hydrostatic reconstruction does not satisfy a discrete entropy inequality without quadratic error term, however restrictive the CFL condition is, which is the result from Proposition 3.8 in [3]. Therefore the iterative scheme can be considered an improvement over this aspect, and we illustrate this through a numerical test where the explicit strategy increases the total energy, unlike the iterative method.

More precisely we measure the variation of total energy in a configuration with a varying bottom, and where the initial condition is given by a flat free surface and a constant nonzero velocity. Periodic boundary conditions are used, and the results can be seen in Figure 5. Interestingly all the iterative methods manage to dissipate the total energy, even the scheme using the index Maxwellian, for which there is no proof of discrete entropy inequality. On the contrary, the explicit kinetic scheme with half-disk Maxwellian increases the energy in the first few time steps, after what it decreases. The same goes for the explicit HLL scheme, and as a result these two explicit methods might not converge to the entropy solution. For comparison we also added in dark blue the iterative kinetic scheme with $\alpha = 0$ and whose subiterations stop as soon as the entropy condition (79) is verified. We can see that after some time this scheme becomes less dissipative than iterative kinetic methods using the standard tolerance condition.

6.2 The two dimensional case

The iterative kinetic scheme (60) and its version with hydrostatic reconstruction (73) can be easily extended to the two dimensional case. We believe that the results obtained in Section 5 the 1D setting carry to the higher dimension. We leave this study for later work, and perform a numerical experiment consisting of the 2D parabolic bowl [13] with a cartesian mesh. The results are displayed in Figures 6 and 7. We see that when increasing the tolerance value to $\varepsilon_{\text{tol}} = 10^{-5}$, the experimental order of convergence of the iterative scheme decreases. On the other hand, the smaller ε_{tol} is, the more iterations are needed to reach the stopping criteria which translates to an increase in computational time. We also note that we needed to decrease the CFL constant form 1/2 to 1/10 to converge with a tolerance of $\varepsilon_{\text{tol}} = 10^{-9}$.

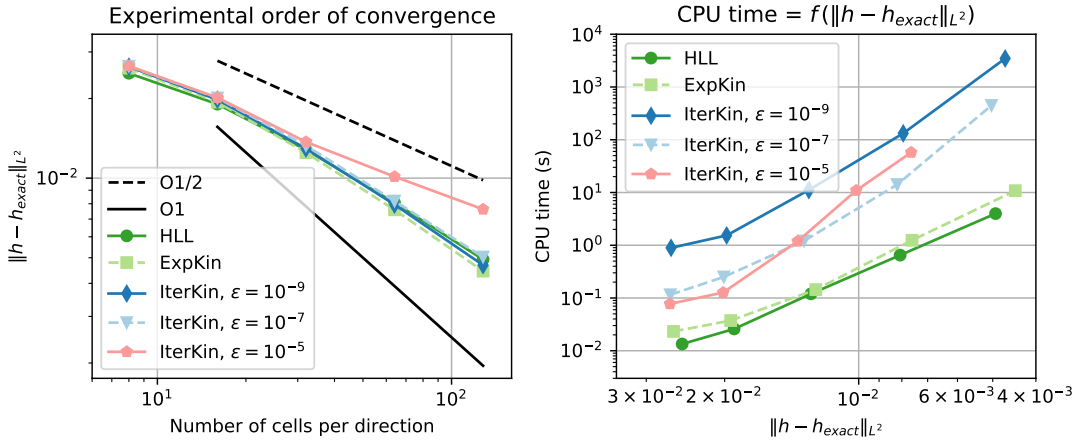

Figure 6: Convergence and efficiency curves obtained with the 2D parabolic bowl testcase. Different tolerances ε are compared for the iterative kinetic scheme. A CFL constant of 1/2 was used, except for the case $\varepsilon = 10^{-9}$ where we set it to 1/10.

Acknowledgments

The authors wish to express their warmest thanks to François Bouchut for many fruitful discussions. Antonin Leprevost and Bilal Al Taki have contributed preliminary versions of the work. This project has been supported by Région Île-de-France.

References

[1] Sebastien Allgeyer, Marie-Odile Bristeau, David Froger, Raouf Hamouda, V. Jauzein, Anne Mangeney, Jacques Sainte-Marie, Fabien Souillé, and Martin Valée, *Numerical approximation of the 3d hydrostatic Navier-Stokes system with free surface*, ESAIM, Math. Model. Numer. Anal. **53** (2019), no. 6, 1981–2024 (English).

[2] Emmanuel Audusse, François Bouchut, Marie-Odile Bristeau, Rupert Klein, and Benoît Perthame, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows*, SIAM J. Sci. Comput. **25** (2004), no. 6, 2050–2065 (English).

[3] Emmanuel Audusse, François Bouchut, Marie-Odile Bristeau, and Jacques Sainte-Marie, *Kinetic entropy inequality and hydrostatic reconstruction scheme for the Saint-Venant system*, Math. Comput. **85** (2016), no. 302, 2815–2837 (English).

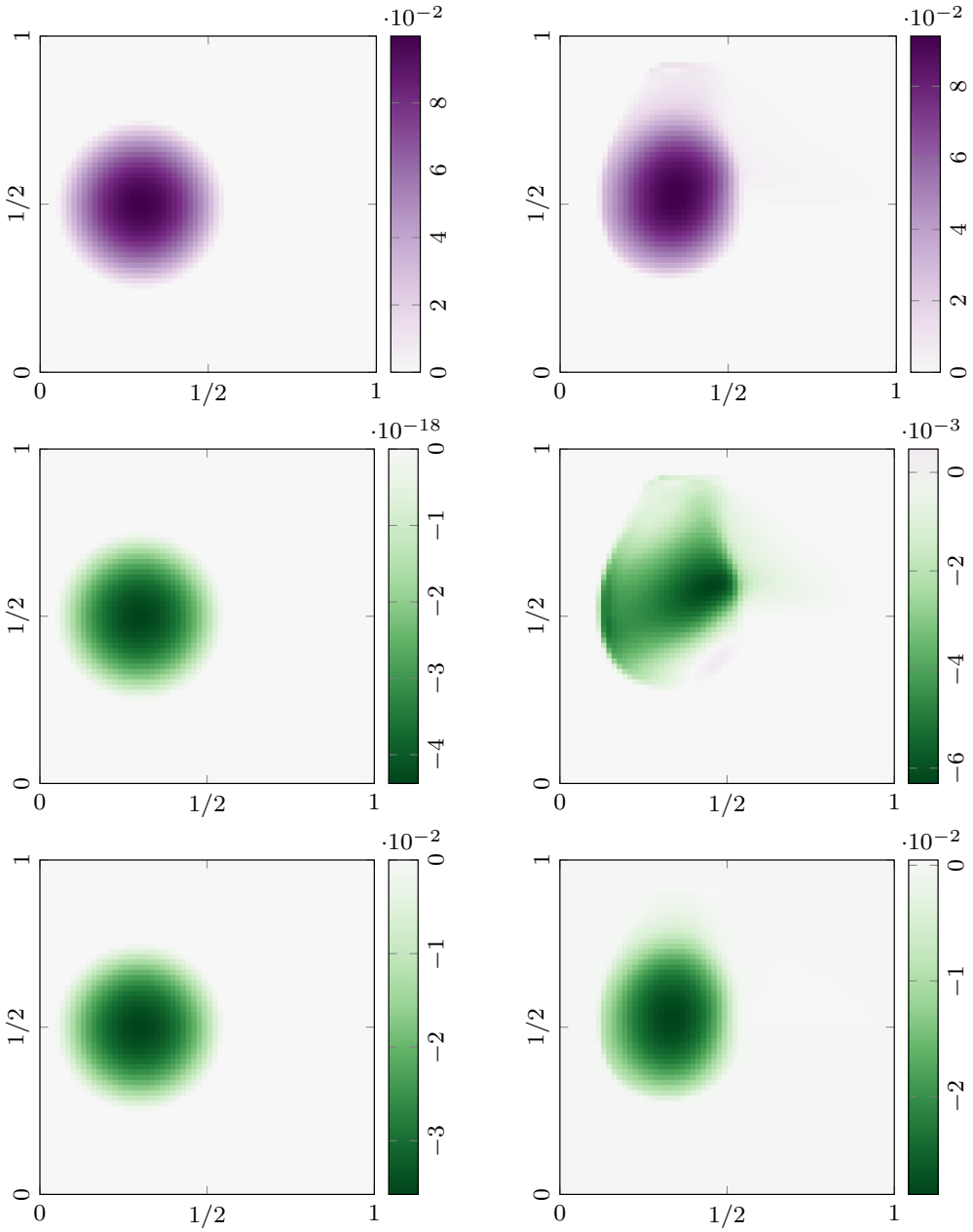

Figure 7: On the left: analytic solution to the 2D parabolic bowl. On the right: approximation obtained by the iterative kinetic scheme with $\varepsilon = 10^{-7}$ over a 64×64 mesh. The first row gives the water height, the second the discharge in the x-axis and the third the discharge in the y-axis. Note that the time of the plot coincides to when the analitycal discharge q cancels.

[4] Emmanuel Audusse and Marie-Odile Bristeau, *A well-balanced positivity preserving "second-order" scheme for shallow water flows on unstructured meshes*, J. Comput. Phys. **206** (2005), no. 1, 311–333 (English).

[5] F. Berthelin and F. Bouchut, *Relaxation to isentropic gas dynamics for a BGK system with single kinetic entropy*, Methods Appl. Anal. **9** (2002), no. 2, 313–327 (English).

[6] F. Bouchut, *Construction of BGK models with a family of kinetic entropies for a given system of conservation laws*, J. Stat. Phys. **95** (1999), no. 1-2, 113–170 (English).

[7] ———, *Entropy satisfying flux vector splittings and kinetic BGK models*, Numer. Math. **94** (2003), no. 4, 623–672 (English).

[8] François Bouchut, *Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources.*, Front. Math., Basel: Birkhäuser, 2004 (English).

[9] François Bouchut and Xavier Lhébrard, *Convergence of the kinetic hydrostatic reconstruction scheme for the Saint Venant system with topography*, Math. Comput. **90** (2021), no. 329, 1119–1153 (English).

[10] Marie-Odile Bristeau and Benoit Coussin, *Boundary Conditions for the Shallow Water Equations solved by Kinetic Schemes*, Research Report RR-4282, INRIA, 2001, Projet M3N.

[11] Marie-Odile Bristeau, Nicole Goutal, and Jacques Sainte-Marie, *Numerical simulations of a non-hydrostatic shallow water model*, Comput. Fluids **47** (2011), no. 1, 51–64 (English).

[12] F. Coron and B. Perthame, *Numerical passage from kinetic to fluid equations*, SIAM J. Numer. Anal. **28** (1991), no. 1, 26–42 (English).

[13] Olivier Delestre, Carine Lucas, Pierre-Antoine Ksinant, Frédéric Darboux, Christian Laguerre, T.-N.-Tuoi Vo, François James, and Stéphane Cordier, *SWASHES: a compilation of shallow water analytic solutions for hydraulic and environmental studies*, Int. J. Numer. Methods Fluids **72** (2013), no. 3, 269–300 (English).

[14] J.-F. Gerbeau and B. Perthame, *Derivation of viscous Saint-Venant system for laminar shallow water; numerical validation*, Discrete Contin. Dyn. Syst., Ser. B **1** (2001), no. 1, 89–102 (English).

[15] Laurent Gosse, *Computing qualitatively correct approximations of balance laws. Exponential-fit, well-balanced and asymptotic-preserving*, SIMAI Springer Ser., vol. 2, Milano: Springer, 2013 (English).

[16] N. Goutal and J. Sainte-Marie, *A kinetic interpretation of the section-averaged Saint-Venant system for natural river hydraulics*, Int. J. Numer. Methods Fluids **67** (2011), no. 7, 914–938 (English).

[17] Shi Jin and Lorenzo Pareschi, *Asymptotic-preserving (ap) schemes for multiscale kinetic equations: a unified approach*, 2001.

[18] B. Perthame, *Boltzmann type schemes for gas dynamics and the entropy property*, SIAM J. Numer. Anal. **27** (1990), no. 6, 1405–1421 (English).

[19] ———, *Second-order Boltzmann schemes for compressible Euler equations in one and two space dimensions*, SIAM J. Numer. Anal. **29** (1992), no. 1, 1–19 (English).

[20] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo **38** (2001), no. 4, 201–231 (English).

[21] Benoît Perthame, *Kinetic formulation of conservation laws*, Oxf. Lect. Ser. Math. Appl., vol. 21, Oxford: Oxford University Press, 2002 (English).

[22] Saint-Venant, *Théorie du mouvement non permanent des eaux, avec application aux crues des rivières et à l'introduction des marées dans leur lit.*, C. R. Acad. Sci., Paris **73** (1871), 147–154 (French).

[23] Yulong Xing and Chi-Wang Shu, *A survey of high order schemes for the shallow water equations*, J. Math. Study **47** (2014), no. 3, 221–249 (English).

A Expression of the numerical fluxes

The optimal choice for the Maxwellian is given by (4). Unfortunately the explicit expression for the numerical fluxes appearing in (39) is hardly possible with the choice (4) and the use of approximate quadrature formula for the integrals in (4) will degrade the accuracy of the scheme and increase the computational costs. Hence, we we choose M defined by the first expression in (37) and relation (39) becomes

$$U_i^{\text{int}} = \frac{1}{2\sqrt{3}}\Big(\int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} d\xi + \int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} d\xi \Big),$$

with $a_j^n = u_j^n - \sqrt{3}\, c_j^n$ and $b_j^n = u_j^n + \sqrt{3}\, c_j^n$. The expressions of h_i^{int} and $(hu)_i^{\text{int}}$ are given by

$$\begin{aligned} h_i^{\text{int}} =& \frac{1}{2\sqrt{3}}\Big(\sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \underbrace{\int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} d\xi}_{(Ah)_{i,j}} \\ &+ \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \underbrace{\int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} d\xi}_{(Bh)_{i,j}} \Big) \end{aligned} \tag{84}$$

$$\begin{aligned} (hu)_i^{\text{int}} =& \frac{1}{2\sqrt{3}}\Big(\sum_{j=i}^{P} -\frac{1}{\sigma} \sqrt{\frac{2h_j^n}{g}} \underbrace{\int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \frac{(-\sigma\xi)^{j-i+1}}{(1-\sigma\xi)^{j-i+1}} d\xi}_{(Ahu)_{i,j}} \\ &+ \sum_{j=1}^{i} \frac{1}{\sigma} \sqrt{\frac{2h_j^n}{g}} \underbrace{\int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \frac{(\sigma\xi)^{i-j+1}}{(1+\sigma\xi)^{i-j+1}} d\xi}_{(Bhu)_{i,j}} \Big) \end{aligned} \tag{85}$$

Now we need to compute analytically the integrals of both expressions using the following lemmas.

Lemma A.1 *If we denote $y = 1 - \frac{1}{1+x}$ for all $x \in \mathbb{R}\backslash\{-1\}$ and $C \in \mathbb{R}$ we have the following primitive:*

$$\int \frac{x^k}{(1+x)^{k+1}} dx = \ln(|1+x|) - \sum_{l=1}^{k} \frac{y^l}{l} + C.$$

Lemma A.2 *Using the same notation as in the previous lemma, we have*

$$\int \frac{x^k}{(1+x)^k} dx = -k \ln(|1+x|) + x + \sum_{l=1}^{k-1} l \frac{y^{k-l}}{k-l} + C'.$$

Proof of Lemma A.1. We have

$$I = \int \frac{x^k}{(1+x)^{k+1}} dx = \int \frac{x^k}{(1+x)^k} \frac{1}{1+x} dx = \int (1 - \frac{x}{1+x})^k \frac{1}{1+x}$$

.
We pose $y = 1 - \frac{1}{1+x}$

$$I = \int y^k (1-y) \frac{dy}{(1-y)^2} = \int \frac{y^k - 1}{1-y} + \frac{1}{1-y} dy$$

Now we use the formula $y^k - 1 = (y-1)(y^{k-1} + y^{k-2} + \ldots + y + 1)$. And we obtain

$$\begin{aligned} I &= -\int \sum_{l=0}^{k-1} y^l dy - \ln(|1-y|) + C && C \in \mathbb{R} \\ &= \ln(|1+x|) - \sum_{l=1}^{k} \frac{y^l}{l} + C' && C' \in \mathbb{R} \end{aligned}$$

□

Proof of Lemma A.2. We already have denoted $y = \frac{x}{1+x} = 1 - \frac{1}{1+x}$

$$I = \int (\frac{x}{1+x})^k dx = \int \frac{y^k dy}{(1-y)^2} = \int \Big(\frac{y^k - 1}{(1-y)^2} + \frac{1}{(1-y)^2} \Big) dy$$

where the formula $y^k - 1 = (y-1)(y^{k-1} + y^{k-2} + \ldots + y + 1)$ has been used. Hence

$$\begin{aligned} I &= -\int \sum_{l=0}^{k-1} \frac{y^l}{1-y} dy + x + C = -\int \sum_{l=0}^{k-1} \frac{y^l - 1}{1-y} dy - \int \frac{1}{1-y} \sum_{l=0}^{k-1} dy + x + C \\ &= \int \sum_{l=1}^{k-1} \frac{y^l - 1}{y-1} dy + k \ln(|1-y|) + x + C' = \int \sum_{l=1}^{k-1} \sum_{p=0}^{l-1} y^p dy - k \ln(|1+x|) + x + C' \\ &= \sum_{l=1}^{k-1} l \int y^{k-1-l} dy - k \ln(|1+x|) + x + C' = \sum_{l=1}^{k-1} l \frac{y^{k-l}}{k-l} - k \ln(|1+x|) + x + C'', \end{aligned}$$

with $(C, C', C'') \in \mathbb{R}^3$. □

We are now able to compute the quantities $Ah_{i,j}, Bh_{i,j}, Ahu_{i,j}, Bhu_{i,j}$

$$\begin{aligned} Ah_{i,j} &= \int_{\min(0,a_j^n)}^{\min(0,b_j^n)} \frac{(-\sigma\xi)^{j-i}}{(1-\sigma\xi)^{j-i+1}} d\xi = -\frac{1}{\sigma} \int_{-\min(0,a_j^n)\sigma}^{-\min(0,b_j^n)\sigma} \frac{(x)^{j-i}}{(1+x)^{j-i+1}} dx \\ &= \frac{1}{\sigma} \Big[\ln(|1+x|) - \sum_{l=1}^{j-i} \frac{y^l}{l} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma}. \end{aligned} \tag{86}$$

$$\begin{aligned} Bh_{i,j} &= \int_{\max(0,a_j^n)}^{\max(0,b_j^n)} \frac{(\sigma\xi)^{i-j}}{(1+\sigma\xi)^{i-j+1}} d\xi = \frac{1}{\sigma} \int_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \frac{(x)^{i-j}}{(1+x)^{i-j+1}} dx \\ &= \frac{1}{\sigma} \Big[\ln(|1+x|) - \sum_{l=1}^{i-j} \frac{y^l}{l} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \end{aligned} \tag{87}$$

And similarly we obtain the formulas for Ahu and Bhu under the form

$$Ahu_{i,j} = \frac{1}{\sigma} \Big[-(j-i+1) \ln(|1+x|) + x + \sum_{l=1}^{j-i} l \frac{y^{j-i+1-l}}{j-i+1-l} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma} \tag{88}$$

$$Bhu_{i,j} = \frac{1}{\sigma}\Big[-(i-j+1)\ln(|1+x|) + x + \sum_{l=1}^{i-j} l \frac{y^{i-j+1-l}}{i-j+1-l}\Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \tag{89}$$

To conclude this paragraph, we give the final expression of U_i^{int}

$$h_i^{\text{int}} = \frac{1}{2\sigma\sqrt{3}}\Big(\sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \Big[\ln(|1+x|) - \sum_{l=1}^{j-i} \frac{y^l}{l} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma} \tag{90}$$

$$+ \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \Big[\ln(|1+x|) - \sum_{l=1}^{i-j} \frac{y^l}{l} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \Big) \tag{91}$$

$$(hu)_i^{\text{int}} = \frac{1}{2\sigma^2\sqrt{3}}\Big(-\sum_{j=i}^{P} \sqrt{\frac{2h_j^n}{g}} \Big[-(j-i+1)\ln(|1+x|) + x + \sum_{k=1}^{j-i} (j-i+1-k)\frac{y^k}{k} \Big]_{-\min(0,b_j^n)\sigma}^{-\min(0,a_j^n)\sigma} \tag{92}$$

$$+ \sum_{j=1}^{i} \sqrt{\frac{2h_j^n}{g}} \Big[-(i-j+1)\ln(|1+x|) + x + \sum_{k=1}^{i-j} (i-j+1-k)\frac{y^k}{k} \Big]_{\max(0,a_j^n)\sigma}^{\max(0,b_j^n)\sigma} \Big) \tag{93}$$

B Computations of the fluxes involving the boundary conditions

We assume the ghost quantities U_0^{n+1} and U_{P+1}^{n+1} at time t^{n+1} to be known. The exterior contribution given in (38) also writes

$$U_i^{\text{ext}} = \int_{\mathbb{R}^-} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}} M_{P+1}^{n+1} d\xi + \int_{\mathbb{R}^+} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^i}{(1+\sigma\xi)^i} M_0^{n+1} d\xi.$$

Using computations similar to what has been proposed in Appendix A, we get

$$U_i^{\text{ext}} = \frac{1}{2\sqrt{3}} \Bigg[\sqrt{\frac{2h_{P+1}^{n+1}}{g}} \int_{\min(0,a_{P+1}^{n+1})}^{\min(0,b_{P+1}^{n+1})} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}} d\xi$$

$$+ \sqrt{\frac{2h_0^{n+1}}{g}} \int_{\max(0,a_0^{n+1})}^{\max(0,b_0^{n+1})} \begin{pmatrix} 1 \\ \xi \end{pmatrix} \frac{(\sigma\xi)^i}{(1+\sigma\xi)^i} d\xi \Bigg],$$

or equivalently

$$h_i^{\text{ext}} = \frac{1}{2\sqrt{3}} \Bigg[\sqrt{\frac{2h_{P+1}^{n+1}}{g}} \int_{min(0,a_{P+1}^{n+1})}^{min(0,b_{P+1}^{n+1})} \frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}} d\xi$$

$$+ \sqrt{\frac{2h_0^{n+1}}{g}} \int_{max(0,a_0^{n+1})}^{max(0,b_0^{n+1})} \frac{(\sigma\xi)^i}{(1+\sigma\xi)^i} d\xi \Bigg],$$

$$(hu)_i^{\text{ext}} = \frac{1}{2\sqrt{3}} \Bigg[\sqrt{\frac{2h_{P+1}^{n+1}}{g}} \int_{min(0,a_{P+1}^{n+1})}^{min(0,b_{P+1}^{n+1})} \xi \frac{(-\sigma\xi)^{P-i+1}}{(1-\sigma\xi)^{P-i+1}} d\xi$$

$$+ \sqrt{\frac{2h_0^{n+1}}{g}} \int_{max(0,a_0^{n+1})}^{max(0,b_0^{n+1})} \xi \frac{(\sigma\xi)^i}{(1+\sigma\xi)^i} d\xi \Bigg].$$

As explained in Section 3.4, in practice we will replace the unknown values of U_0^{n+1}, U_{P+1}^{n+1} with that of U_0^n, U_{P+1}^n. The expression of h_i^{ext} can then be established by the mean of Lemma A.2. We have now to find an analytic expression for the quantity $\int \frac{x^{k+1}}{(1+x)^k} dx$ in order to obtain the final expression $(hu)_i^{\text{ext}}$. The following lemma holds.

Lemma B.1 *Let $k \in \mathbb{N}^*$. If we denote $y = 1 - \frac{1}{1+x}$ for all $x \in \mathbb{R}\backslash\{-1\}$ and $C \in \mathbb{R}$ we have the following expression*

$$\int \frac{x^{k+1}}{(1+x)^k} dx = \Big(-\sum_{r=1}^{k-2} (k-r-1) \frac{k-r}{2} \frac{y^r}{r} \Big) \mathbb{1}_{k\geq 3} - \Big(\frac{k(k-1)}{2} \ln|1-y| \Big) \mathbb{1}_{k\geq 2} \\ - \frac{k+1}{(1-y)} + \frac{1}{2(1-y)^2} - \Big(\sum_{q=1}^{k-1} (k-q) \frac{y^q}{q} \Big) \mathbb{1}_{k\geq 2} - k \ln|1-y| + C$$

Proof. We begin by performing the change of variable $y = 1 - \frac{1}{1+x}$

$$\int \frac{x^{k+1}}{(1+x)^k} dx = \int y^k \Big(\frac{1}{1-y} - 1 \Big) \frac{\mathrm{d}y}{(1-y)^2} = \int \frac{y^k}{(1-y)^3} \, \mathrm{d}y - \int \frac{y^k}{(1-y)^2} \, \mathrm{d}y \; .$$

Making use of $y^k - 1 = (y-1)(y^{k-1} + y^{k-2} + \cdots + 1)$ as before, we remark the following relation for $k \geq 1$

$$\frac{y^k}{1-y} = \frac{y^k - 1}{1-y} + \frac{1}{1-y} = -\sum_{p=0}^{k-1} y^p + \frac{1}{1-y}$$

Dividing this by $1-y$ leads to

$$\frac{y^k}{(1-y)^2} = -\sum_{p=0}^{k-1} \frac{y^p}{1-y} + \frac{1}{(1-y)^2} = -\sum_{p=0}^{k-1} \Big(\frac{y^p - 1}{1-y} + \frac{1}{1-y} \Big) + \frac{1}{(1-y)^2} \\ = \Big(\sum_{p=1}^{k-1} \sum_{q=0}^{p-1} y^q \Big) \mathbb{1}_{k\geq 2} - \frac{k}{1-y} + \frac{1}{(1-y)^2}$$

Iterating this one more time we find

$$\frac{y^k}{(1-y)^3} = \Big(\sum_{p=1}^{k-1} \sum_{q=0}^{p-1} \frac{y^q}{1-y} \Big) \mathbb{1}_{k\geq 2} - \frac{k}{(1-y)^2} + \frac{1}{(1-y)^3} \\ = \Big(\sum_{p=1}^{k-1} \sum_{q=0}^{p-1} \frac{y^q - 1}{1-y} + \frac{1}{1-y} \Big) \mathbb{1}_{k\geq 2} - \frac{k}{(1-y)^2} + \frac{1}{(1-y)^3} \\ = \Big(-\sum_{p=2}^{k-1} \sum_{q=1}^{p-1} \sum_{r=0}^{q-1} y^r \Big) \mathbb{1}_{k\geq 3} + \frac{k(k-1)}{2(1-y)} \mathbb{1}_{k\geq 2} - \frac{k}{(1-y)^2} + \frac{1}{(1-y)^3}$$

As a consequence we get the following primitives up to a constant

$$\int \frac{y^k}{(1-y)^2} \, \mathrm{d}y = \Big(\sum_{p=1}^{k-1} \sum_{q=1}^{p} \frac{y^q}{q} \Big) \mathbb{1}_{k\geq 2} + k \ln|1-y| + \frac{1}{(1-y)}$$

$$\int \frac{y^k}{(1-y)^3} \, \mathrm{d}y = \Big(-\sum_{p=2}^{k-1} \sum_{q=1}^{p-1} \sum_{r=1}^{q} \frac{y^r}{r} \Big) \mathbb{1}_{k\geq 3} - \frac{k(k-1)}{2} \ln|1-y| \mathbb{1}_{k\geq 2} \\ - \frac{k}{(1-y)} + \frac{1}{2(1-y)^2}$$

Finally, we simplify the double and triple sums

$$\sum_{p=1}^{k-1} \sum_{q=1}^{p} \frac{y^q}{q} = \sum_{q=1}^{k-1} \sum_{p=q}^{k-1} \frac{y^q}{q} = \sum_{q=1}^{k-1} (k-q) \frac{y^q}{q}$$

From this we deduce that

$$\begin{aligned}\sum_{p=2}^{k-1}\sum_{q=1}^{p-1}\sum_{r=1}^{q}\frac{y^r}{r} &= \sum_{p=2}^{k-1}\sum_{r=1}^{p-1}(p-r)\frac{y^r}{r}\\ &= \sum_{p=1}^{k-2}\sum_{r=1}^{p}(p-r+1)\frac{y^r}{r} = \sum_{r=1}^{k-2}\sum_{p=r}^{k-2}(p-r+1)\frac{y^r}{r}\\ &= \sum_{r=1}^{k-2}\Big(\frac{(k-r-1)(k+r-2)}{2}+(k-r-1)(1-r)\Big)\frac{y^r}{r}\\ &= \sum_{r=1}^{k-2}(k-r-1)\frac{k-r}{2}\frac{y^r}{r}\end{aligned}$$

As a conclusion we have the expression

$$\begin{aligned}\int\frac{x^{k+1}}{(1+x)^k}dx = \Big(-\sum_{r=1}^{k-2}(k-r-1)\frac{k-r}{2}\frac{y^r}{r}\Big)\mathbb{1}_{k\geq 3} - \Big(\frac{k(k-1)}{2}\ln|1-y|\Big)\mathbb{1}_{k\geq 2}\\ -\frac{k+1}{(1-y)}+\frac{1}{2(1-y)^2} - \Big(\sum_{q=1}^{k-1}(k-q)\frac{y^q}{q}\Big)\mathbb{1}_{k\geq 2} - k\ln|1-y| + C\end{aligned}$$

where $C\in\mathbb{R}$ and with $y = x/(x+1)$. $\square$