Fully discrete entropy inequality for the hydrostatic reconstruction scheme

François Bouchut

LAMA, CNRS & Université Paris-Est

in collaboration with

E. Audusse, M.-O. Bristeau, J. Sainte-Marie

Orléans, November 20, 2018

Outline

▷ For a system of conservation laws

$$\partial_t U + \sum_j \partial_j F_j(U) = S \tag{1}$$

with Cauchy data, solutions in the distribution sense are not unique.

▷ In order to select a unique solution, one has to specify entropy conditions. One way to do it is to ask for an entropy inequality

$$\partial_t \eta(U) + \sum_j \partial_j G_j(U) \leq \eta'(U) S, \tag{2}$$

where $\eta = \eta(U)$ is an entropy, i.e. verifies

$$\eta' F_j' = G_j' \tag{3}$$

for some entropy fluxes $G_j(U)$. Moreover, η has to be convex.

▷ The inequality (2) is an equality for smooth solutions to (1) (multiply by $\eta'(U)$). For discontinuous solutions one requires that the inequality (2) holds in the sense of distributions. This means equivalently to ask for Rankine Hugoniot conditions across the discontinuities

$$(\eta(U_+) - \eta(U_-))\nu_t + \sum_j \Big(G_j(U_+) - G_j(U_-) \Big) \nu_j \leq 0 \tag{4}$$

▷ Example : the shallow water system

$$\text{(5)} \qquad \begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + gh^2/2) + ghz_x = 0, \end{cases}$$

with $z(x)$ given, with entropy inequality (energy dissipation)

$$\text{(6)} \qquad \partial_t(hu^2/2 + gh^2/2) + \partial_x((hu^2/2 + gh^2)u) + ghuz_x \leq 0,$$

or equivalently

$$\text{(7)} \qquad \partial_t(hu^2/2 + gh^2/2 + ghz) + \partial_x((hu^2/2 + gh^2 + ghz)u) \leq 0.$$

The topography z can be considered as a source (if z is smooth) or as an additional unknown verifying $\partial_t z = 0$. The entropy $hu^2/2 + gh^2/2$ is a convex function of the conservative variables $U = (h, hu)$.

- A class of specific partially conservative systems :

$$\begin{aligned} &\partial_t U + \partial_x \left(F(U,Z)\right) + B(U,Z)\partial_x Z = 0, \\ &\partial_t Z = 0, \end{aligned} \tag{8}$$

where $U(t,x) \in \mathbb{R}^p$ is the unknown and $Z(x) \in \mathbb{R}^r$ is given. The nonlinearities $F(U,Z) \in \mathbb{R}^p$ and $B(U,Z) \in \mathbb{M}_{p,r}$ are supposed to be smooth. The system (8) is a quasilinear system $\partial_t \widetilde{U} + A\partial_x \widetilde{U} = 0$ on the vector $\widetilde{U} = (U,Z)$, with

$$A = \begin{pmatrix} F_U & F_Z + B \\ 0 & 0 \end{pmatrix}. \tag{9}$$

- Example : The shallow water problem with fixed topography.
- Example : $Z(x) = x$. Then

$$\partial_t U + \partial_x \left(F(U,x)\right) + B(U,x) = 0 \tag{10}$$

is a system of conservation laws with source. The interaction between the conservative term and the source is the key issue.

- The formulation (8) enables to use schemes designed for quasilinear systems that treat both terms at the same level. It allows well-balancing and stability of computations.

For the quasiconservative system $\partial_t U + \partial_x (F(U,Z)) + B(U,Z)\partial_x Z = 0$, a first-order finite volume scheme is a scheme of the form

$$U_i^{n+1} - U_i + \frac{\Delta t}{\Delta x_i}\left(F_{i+1/2-} - F_{i-1/2+}\right) = 0, \tag{11}$$

where U_i^n is an approximation of the solution $U(t,x)$,

$$U_i^n \simeq \frac{1}{\Delta x_i}\int_{x_{i-1/2}}^{x_{i+1/2}} U(t_n,x)dx, \tag{12}$$

the indices n and i refer to time t_n and cell $C_i = (x_{i-1/2}, x_{i+1/2})$ of size $\Delta x_i = x_{i+1/2} - x_{i-1/2}$, where $t_{n+1} - t_n = \Delta t$ and $\cdots < x_{i-1/2} < x_{i+1/2} < \ldots$ is the grid. The interface terms are defined by

$$F_{i+1/2-} = \mathcal{F}_l(U_i, U_{i+1}, Z_i, Z_{i+1}), \qquad F_{i+1/2+} = \mathcal{F}_r(U_i, U_{i+1}, Z_i, Z_{i+1}), \tag{13}$$

and $\mathcal{F}_l(U_l, U_r, Z_l, Z_r)$, $\mathcal{F}_r(U_l, U_r, Z_l, Z_r)$ are the numerical flux functions, to be chosen satisfying some accuracy and stability properties.

▷ For a conservative system ($B(U,Z) \equiv 0$), one would require conservativity, $\mathcal{F}_l \equiv \mathcal{F}_r$.

▷ The consistency of a numerical method is the minimal accuracy property we can ask for. Here we consider smooth consistency : if the sequence U_i^n converges to a smooth function, it must be a classical solution to the equation considered.
This condition does not ensure weak consistency for discontinuous solutions (no Rankine-Hugoniot condition is ensured)
▷ For our partially conservative problem

$$\partial_t U + \partial_x (F(U,Z)) + B(U,Z)\partial_x Z = 0, \qquad Z = Z(x), \tag{14}$$

we require moreover that when $Z = cst$, the scheme becomes conservative. This ensures the Rankine-Hugoniot condition for continuous Z.
▷ Then the consistency+asymptotic conservativity conditions resume as

$$\mathcal{F}_l(U,U,Z,Z) = \mathcal{F}_r(U,U,Z,Z) = F(U,Z), \tag{15}$$

$$\mathcal{F}_r(U_l,U_r,Z_l,Z_r) - \mathcal{F}_l(U_l,U_r,Z_l,Z_r) = -B(U,Z)(Z_r - Z_l) + o(Z_r - Z_l) \quad \text{as } U_l, U_r \to U \text{ and } Z_l, Z_r \to Z. \tag{16}$$

This formulation via a quasilinear problem enables to distribute the source term at the interfaces $x_{i+1/2}$.

▷ Stability properties are usually required for a finite volume scheme. They can be of two types :

- Preservation of some invariants domains. This means that some natural bounds are preserved during the evolution, like nonnegativity of density, volume fraction between 0 and 1...
- Existence of a discrete entropy inequality

$$\eta(U_i^{n+1}) - \eta(U_i) + \frac{\Delta t}{\Delta x_i}\left(G_{i+1/2} - G_{i-1/2}\right) \leq 0 \tag{17}$$

for some numerical entropy flux $G_{i+1/2} = \mathcal{G}(U_i, U_{i+1})$. It selects suitable solutions, and provides the a priori bound $\sum_i \Delta x_i \eta(U_i^n) \leq \sum_i \Delta x_i \eta(U_i^0)$.

- A variant of this property is the existence of a semi-discrete entropy inequality, that is valid only in the limit $\Delta t \to 0$. It does not give any a priori bound.

▷ These stability properties can only hold under a CFL condition

$$\Delta t\, a_{i+1/2} \leq \min(\Delta x_i, \Delta x_{i+1}), \tag{18}$$

where $a_{i+1/2}$ is a suitable approximation of the wave speed.

The shallow water problem with topography writes

$$\begin{cases} \partial_t h + \partial_x(hu) = 0, \\ \partial_t(hu) + \partial_x(hu^2 + p(h)) + hg\partial_x z = 0, \end{cases} \tag{19}$$

where $h(t,x) \geq 0$ is the water height, $u(t,x) \in \mathbb{R}$ is the velocity, and $z(x)$ is the topography. The "pressure" is $p(h) = gh^2/2$, $g > 0$.

- The system is of the form (8) with $U = (h, hu)$, $Z = z$,

$$F(U,z) = F(U) = (hu, hu^2 + p(h)), \qquad B(U,z) = B(U) = (0, gh). \tag{20}$$

- The system has an entropy $\widetilde{\eta}(U,z) = hu^2/2 + gh^2/2 + hgz \equiv \eta(U) + hgz$ with entropy flux $\widetilde{G}(U,z) = (\widetilde{\eta} + p)u \equiv G(U) + hgzu$.
- The steady states are characterized by

$$hu = cst, \qquad u^2/2 + g(h+z) = cst. \tag{21}$$

The steady states at rest are those for which $u = 0$ and $h + z = cst$. They can be characterized at the discrete level by the local relations

$$u_l = u_r = 0, \qquad h_l + z_l - h_r - z_r = 0. \tag{22}$$

Numerical difficulties :

- Keep the water height h nonnegative,
- Compute dry areas where $h = 0$,
- Preserve the total amount of water,
- Preserve the steady states at rest (well-balanced property)
- Satisfy a discrete entropy inequality,
- Produce stable computations (no oscillations) for all data, including transcritical cases.

The hydrostatic reconstruction scheme proposed by Audusse, Bouchut, Bristeau, Klein, Perthame 04 satisfies all these properties (only semi-discrete entropy inequality), and is computationally cheap. It writes as

$$\begin{aligned} \mathcal{F}_l(U_l, U_r, z_l, z_r) &= \mathcal{F}(U_l^*, U_r^*) + \begin{pmatrix} 0 \\ p(h_l) - p(h_l^*) \end{pmatrix}, \\ \mathcal{F}_r(U_l, U_r, z_l, z_r) &= \mathcal{F}(U_l^*, U_r^*) + \begin{pmatrix} 0 \\ p(h_r) - p(h_r^*) \end{pmatrix}, \end{aligned} \tag{23}$$

where $\mathcal{F}(U_l, U_r)$ is a (good) consistent numerical flux for the shallow water problem without source, and the reconstructed states U_l^*, U_r^* are defined by

$$\begin{aligned} U_l^* = (h_l^*, h_l^* u_l), \qquad & U_r^* = (h_r^*, h_r^* u_r), \\ h_l^* = \max(0, h_l + z_l - z^*), \qquad & h_r^* = \max(0, h_r + z_r - z^*) \\ z^* = \max(z_l, z_r). & \end{aligned} \tag{24}$$

▷ This scheme can be easily adapted to problems of the same type, like Savage Hutter models.

▷ Other schemes have been proposed. One that has a similar efficiency is the *F-wave method* Leveque et al., or the equivalent Roe method developed by Pares, Castro et al..

▷ The semi-discrete entropy inequality can be written under the form

$$\frac{d}{dt}\widetilde{\eta}_i + \frac{1}{\Delta x_i}\left(\widetilde{\mathcal{G}}_{i+1/2} - \widetilde{\mathcal{G}}_{i-1/2}\right) \leq 0, \tag{25}$$

with $\widetilde{\mathcal{G}}_{i+1/2} = \widetilde{\mathcal{G}}(U_i, U_{i+1}, z_{i+1} - z_i)$,

$$\widetilde{\mathcal{G}}(U_l, U_r, \Delta z) = \mathcal{G}(U_l^*, U_r^*) + \mathcal{F}^0(U_l^*, U_r^*) g z^* \tag{26}$$

where we recall that $\widetilde{\eta}(U, z) = \eta(U) + hgz$, $\widetilde{G}(U, z) = G(U) + hgzu$, and where $\mathcal{G}$ is the numerical entropy flux for the problem without source, and $\mathcal{F}^0$ the mass component (first componant) of the numerical flux without source $\mathcal{F}$.

A counter-result :

Proposition (ABBS 2016) The hydrostatic reconstruction scheme does not verify a fully discrete entropy inequality, whatever small is the ratio $\Delta t/\Delta x$.

The proof is based on an argument of strict convexity of the dissipation, and on the fact that the semi-discrete dissipation vanishes for data such that

$$u_l = u_r \neq 0, \qquad h_l + z_l = h_r + z_r, \qquad z_r - z_l \neq 0. \tag{27}$$

▷ Consider a differential system

$$\frac{dU}{dt} = F(U), \tag{28}$$

verifying an entropy inequality

$$\frac{d}{dt}\eta(U) \leq 0, \qquad \text{i.e. } \eta'(U)F(U) \leq 0 \quad \text{for all } U, \tag{29}$$

with η convex.

▷ One can resolve it by an implicit Euler scheme

$$U^{n+1} = U^n + \Delta t F(U^{n+1}), \tag{30}$$

or by an explicit Euler scheme

$$U^{n+1} = U^n + \Delta t F(U^n). \tag{31}$$

We say that the scheme is entropy satisfying if $\eta(U^{n+1}) \leq \eta(U^n)$ for Δt small enough.

Proposition.

▷ Implicit Euler is entropy satisfying if and only if the system is entropy satisfying, and in this case there is no need of any restriction on the time step.

▷ If explicit Euler is entropy satisfying then the system is entropy satisfying, but the converse is wrong.

The only nontrivial assertion is that (the system is entropy satisfying) implies that (implicit Euler is entropy satisfying). One just has to write according to the convexity of η

$$\begin{aligned} \eta(U^{n+1}) &\leq \eta(U^n) + \eta'(U^{n+1})(U^{n+1} - U^n) \\ &= \eta(U^n) + \Delta t \eta'(U^{n+1}) F(U^{n+1}) \\ &\leq \eta(U^n). \end{aligned} \tag{32}$$

To understand the explicit Euler scheme, one can write

$$\begin{aligned} \eta(U^{n+1}) &\simeq \eta(U^n) + \eta'(U^n)(U^{n+1} - U^n) + \eta''(U^n)\frac{(U^{n+1} - U^n)^2}{2} \\ &= \eta(U^n) + \Delta t \eta'(U^n) F(U^n) + \frac{\Delta t^2}{2}\eta''(U^n) F(U^n)^2 \end{aligned} \tag{33}$$

We observe that the linear term is nonpositive which is good, but the second-order term is nonnegative !

If the system is strictly entropy satisfying and Δt is small enough, then the second-order term will be dominated by the first-order one, and the explicit scheme will be entropic.

Theorem[ABBS 2016] Under the CFL condition

$$\sigma_i v_m \leq \beta < 1, \tag{34}$$

where v_m is an upper bound for the propagation speed, i.e. $|u_i| + \sqrt{2gh_i} \leq v_m$, β is fixed, and $\sigma_i = \Delta t/\Delta x_i$, the hydrostatic reconstruction scheme, with particular homogeneous fluxes defined by a kinetic method, verifies the fully discrete time-space entropy inequality

$$\widetilde{\eta}_i^{n+1} - \widetilde{\eta}_i + \sigma_i(\widetilde{\mathcal{G}}_{i+1/2} - \widetilde{\mathcal{G}}_{i-1/2}) \leq C_\beta(\sigma_i v_m)^2 \left(g(h_i - h_{i+1/2-})^2 + g(h_i - h_{i-1/2+})^2 \right). \tag{35}$$

In particular, the right-hand side is upper bounded by $C(\sigma_i v_m)^2 g \Delta z^2$.

Note that the error term vanishes when $z = cst$ (no topo), or when $\sigma_i \to 0$ (semi-discrete limit). Moreover for a Lipschitz topography it gives rise to an error term that tends to 0 strongly when the time and space steps tend to 0. We have thus strong consistency with the limit entropy inequality, even if the limit solution is discontinuous.

▷ The principle of the kinetic scheme with topography is given in [Perthame, Simeoni 2001]. One solves

$$\partial_t f + \xi \partial_x f - g(\partial_x z)\partial_\xi f = 0 \tag{36}$$

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

$$f(t^n,x,\xi) = M(U^n(x),\xi), \tag{37}$$

where $U^n(x)$ is constant by cells with values U_i^n, and we compute the updated values by

$$U_i^{n+1} = \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} f_i^{n+1-}(\xi) d\xi. \tag{38}$$

Here M is the "Maxwellian"

$$M(h,u,\xi) = \frac{1}{g\pi}\left(2gh - (\xi-u)^2\right)_+^{1/2} \tag{39}$$

that verifies the moment relations

$$\begin{aligned} \int_{\mathbb{R}} \begin{pmatrix} 1 \\ \xi \end{pmatrix} M(U,\xi)\, d\xi &= U, \\ \int_{\mathbb{R}} \xi^2 M(U,\xi)\, d\xi &= hu^2 + g\frac{h^2}{2}. \end{aligned} \tag{40}$$

This gives rise to a consistant and entropy satisfying scheme, but it involves integrals in ξ which are not computable simply.

We replace the Perthame-Simeoni scheme by a fully time discrete formula

$$
\begin{aligned}
f_i^{n+1-} = M_i - \sigma_i \Big(& \xi \mathbf{1}_{\xi<0} M_{i+1/2+} + \xi \mathbf{1}_{\xi>0} M_{i+1/2-} + \delta M_{i+1/2-} \\
& - \xi \mathbf{1}_{\xi>0} M_{i-1/2-} - \xi \mathbf{1}_{\xi<0} M_{i-1/2+} - \delta M_{i-1/2+} \Big),
\end{aligned} \tag{41}
$$

with $M_i = M(U_i, \xi)$, $M_{i+1/2\pm} = M(U_{i+1/2\pm}, \xi)$, $f_i^{n+1-} = f_i^{n+1-}(\xi)$, and

$$
\delta M_{i+1/2-} = (\xi - u_i)(M_i - M_{i+1/2-}), \qquad \delta M_{i-1/2+} = (\xi - u_i)(M_i - M_{i-1/2+}). \tag{42}
$$

Theorem [ABBS 2016] When we take the integral in ξ of (41), we obtain the hydrostatic reconstruction scheme (associated to certain kinetic homogeneous fluxes). Moreover the scheme verifies a kinetic entropy inequality

$$
\begin{aligned}
& H(f_i^{n+1-}, z_i) \\
\leq\ & H(M_i, z_i) - \sigma_i \Big(\widetilde{H}_{i+1/2-} - \widetilde{H}_{i-1/2+} \Big) \\
& - \nu_\beta \, \sigma_i |\xi| \frac{g^2 \pi^2}{6} \Big(\mathbf{1}_{\xi<0} \, (M_{i+1/2+} + M_{i+1/2-})(M_{i+1/2+} - M_{i+1/2-})^2 \\
& + \mathbf{1}_{\xi>0} \, (M_{i-1/2-} + M_{i-1/2+})(M_{i-1/2+} - M_{i-1/2-})^2 \Big) \\
& + C_\beta (\sigma_i v_m)^2 \frac{g^2 \pi^2}{6} M_i \Big((M_i - M_{i+1/2-})^2 + (M_i - M_{i-1/2+})^2 \Big),
\end{aligned} \tag{43}
$$

where $\widetilde{H}_{i+1/2-}$, $\widetilde{H}_{i-1/2+}$ are some "kinetic entropy fluxes", $\nu_\beta > 0$ is a dissipation constant, and $C_\beta \geq 0$ is a constant giving an error term.

Here, $H(f,\xi,z)$ is the kinetic entropy

$$H(f,\xi,z) = H_0(f,\xi) + gzf, \tag{44}$$

$$H_0(f,\xi) = \frac{\xi^2}{2} f + \frac{g^2\pi^2}{2} f^3. \tag{45}$$

The Maxwellian M verifies the kinetic entropy minimization principle :
Theorem[Bouchut 1999]

- For any $h \geq 0$, $u \in \mathbb{R}$, $f \geq 0$, $\xi \in \mathbb{R}$ we have

$$H_0(f,\xi) \geq H_0(M(U,\xi),\xi) + \eta'(U) \begin{pmatrix} 1 \\ \xi \end{pmatrix} (f - M(U,\xi)). \tag{46}$$

- For any function $f(\xi) \geq 0$, setting $h = \int f(\xi) d\xi$, $hu = \int \xi f(\xi) d\xi$, we have

$$\eta(U) = \int_{\mathbb{R}} H_0(M(U,\xi),\xi) d\xi \leq \int_{\mathbb{R}} H_0(f(\xi),\xi) d\xi. \tag{47}$$

The second inequality (47) can be simply obtained by integrating (46) with respect to ξ.

▷ The proof of the theorem giving the kinetic entropy inequality with dissipation and error terms is obtained by estimating the second-order term (in Δt^2) with respect to the first-order term (in Δt) when the latter is nonzero. The residual that controls the error when the dissipation term vanishes yields the final error term.

▷ Note that one can neglect the dissipation term in ν_β when integrating with respect to ξ in order to obtain the entropy inequality of the HR scheme.

▷ We need this dissipation term in order to prove the convergence of the scheme [Bouchut,Lhébrard 2017] !