A partial differential inequality in geological models

Robert EYMARD* Thierry GALLOUËT†

March 2, 2006

Abstract

Sedimentation and erosion processes in sedimentary basins can be modeled by a parabolic equation with a limiter on the fluxes and a constraint on the time variation. This limiter happens to satisfy a stationary scalar hyperbolic inequality, within a constraint, for which we prove the existence and the uniqueness of the solution. Actually, this solution is shown to be the maximal element of a convenient convex set of functions. The existence proof is obtained thanks to the use a numerical scheme.

KEYWORDS: hyperbolic inequalities, erosion and sedimentation models.

1 Introduction

Geological models are increasingly used in the framework of the petroleum industry to provide informations on sedimentary basins. Since such models are recent, only few papers present mathematical and numerical studies. We focus in this paper on the sedimentation and erosion model ([1], [3], [10], [14], [19], [9]) given by the following equations:

$$H_t(x,t) - \mathrm{div}[\Lambda(x)\lambda(x,t)\nabla H(x,t)] = 0, \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{1}$$

where the unknowns, λ and H, are limited by the following constraints:

$$H_t(x,t) \geq -F(x), \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{2}$$

$$0 \leq \lambda(x,t) \leq 1, \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{3}$$

and

$$(\lambda(x,t) - 1)\ (H_t(x,t) + F(x)) = 0, \text{ for a.e. } (x,t) \in \Omega \times (0,T). \tag{4}$$

In (1)-(4), the horizontal extension of the basin is modeled by the domain $\Omega \subset \mathbb{R}^2$, the diameter of which can be about several hundreds of kilometers, and T is the age of the basin (between 0 and 10^7 years for example). The unknowns are the thickness of the sediments $H(x,t)$ and the erosion limiter $\lambda(x,t)$, at any point $(x,t) \in \Omega \times (0,T)$. We denote by $F(x) \geq 0$ the maximum erosion rate at any point $x \in \Omega$. We define, for a given $t_0 \in (0,T)$, the functions $u(x) = \lambda(x,t_0)$, $h(x) = H(x,t_0)$ and $g(x) = \Lambda(x)\nabla h(x)$. Then equations (1)-(4) lead to

$$\begin{array}{l} \mathrm{div}[u(x)g(x)] + F(x) \geq 0, \text{ for a.e. } x \in \Omega, \\ 0 \leq u(x) \leq 1, \text{ for a.e. } x \in \Omega, \end{array} \tag{5}$$

and

$$(u(x) - 1)\ (\mathrm{div}[u(x)g(x)] + F(x)) = 0, \text{ for a.e. } x \in \Omega. \tag{6}$$

Assuming that there exists a solution to Problem (5)-(6), denoted by $u = U(g)$ (indeed, the data F is fixed), then the problem (1) can be expressed using this function U:

$$H_t(x,t) - \mathrm{div}[U\left(\Lambda(\cdot)\nabla H(\cdot,t)\right)(x)\Lambda(x)\nabla H(x,t)] = 0. \text{ for a.e. } (x,t) \in \Omega \times (0,T).$$

*Université de Marne-la-Vallée, 5 boulevard Descartes, F-77454 Marne-la-Vallée, France, Robert.Eymard@univ-mlv.fr
†Université de Marseille, 39 rue Joliot Curie, F-13453 Marseille 13, France, gallouet@latp.univ-mrs.fr

Hence the study of Problem (5)-(6) is a key point for the study of the complete problem (we show in [9] that some numerical schemes are based on the resolution of this problem). Note that the continuous framework of problems similar to (5)-(6) has already been studied for example in [18], in which the authors consider stationary hyperbolic inequalities, where the linear hyperbolic operator satisfies some coercivity hypothesis. For Problem (5)-(6), this coercivity condition reads $\mathrm{div}(g) \leq 0$ and leads to the coercivity of the elliptic operator $-\varepsilon\Delta u - \mathrm{div}(gu)$, for any $\varepsilon > 0$. This coercivity condition is not necessary for proving an existence and uniqueness result for Problem (5)-(6) as it is not necessary for proving an existence and uniqueness result for the elliptic problem without coercivity, $-\varepsilon\Delta u - \mathrm{div}(gu) = F$ with (for instance) Dirichlet boundary condition and $F \in H^{-1}(\Omega)$ (see, for instance, [5], for the continuous case, and [7], [8] for convergence of numerical schemes for such an elliptic problem). Note that, for this elliptic problem without coercivity, we have $F \geq 0$ implies $u \geq 0$ (in this paper also, the positivity of u is related to that of F, and is not an additional constraint). It is also interesting to notice that the study of such a linear elliptic problem needs some nonlinear tools to obtain *a priori* estimates, which is the main point of the proof. In the present paper, one proves an existence result for Problem (5)-(6) passing to the limit on numerical schemes. An alternative proof could be to pass to the limit, as $\varepsilon \longrightarrow 0$, on the solution of the problem obtained replacing $\mathrm{div}(gu)$ by $\varepsilon\Delta u + \mathrm{div}(gu)$ which corresponds to a more classical variational inequality but without coercivity condition.
In [9], we proved that, under some regularity hypotheses, there exists one and only one solution ug to Problem (5)-(6) for any given function $g = \Lambda(x)\nabla h(x)$ (the existence of h is explicitly used in the proof of [9], which is not the case in this paper) and that this solution is the weak solution in the sense of Definition 2.1. Numerical applications are provided in [9], in which we also give the analytical solution of Problem (5)-(6) in the 1D case. The proof of the result given in [9] was based on the classical doubling variable technique of Krushkov and on a generalized sense for a solution (Young measure or entropy process solution) for the uniqueness part, and on a finite volume scheme for the existence part. Note that, in [16], the author shows the existence of a solution, in the case where the first inequality of (5) is modified by the introduction of a time derivative of u, by passing to the limit on analytical solutions of a regularized problem, using a similar result of uniqueness (this result is also provided in [17]). Note that, thanks to the time dependent term, this result of uniqueness requires less hypotheses on g and F than that of the stationary case.
Our approach in this paper is somewhat different. We first establish in Section 2 the connection between the weak solution of Problem (5)-(6) and the maximal element of $\mathcal{C}(g, F)$ (which is indeed also the projection of the constant function 1_Ω on $\mathcal{C}(g, F)$, such providing an alternative weak sense for a solution of Problem (5)-(6)). In Section 3, the sense of the maximal element of $\mathcal{C}(g, F)$ is extended (process maximal element) in order to meet the weaker limits of the finite volume scheme. This weaker sense involves a Lagrange multiplier associated to the erosion constraint, the existence of which has to be proven as well as the existence of a process maximal solution (indeed, it is true that the maximal element of $\mathcal{C}(g, F)$ is a process maximal element but the proof of this result will be a consequence of our work). Then, thanks to a uniqueness result, we get that a process maximal element is indeed, if it exists, identical to the maximal element of $\mathcal{C}(g, F)$. In Section 4, we provide a finite volume scheme for the approximation of the solution and also for the approximation of the Lagrange multiplier, thus obtaining, by passing to the limit, the existence of such a process maximal element. Then, one deduces the strong convergence of the approximate solution to the maximal element of $\mathcal{C}(g, F)$. A by product of this proof of convergence is that the maximal element of $\mathcal{C}(g, F)$ is also the unique weak solution of (5)-(6). It also gives the existence of a Lagrange multiplier associated to the erosion constraint (see Proposition 2.13 and Theorem 4.14). Some directions of research are discussed in Section 5.

2 Weak solution

Let us define the following hypotheses, denoted by Hypotheses (H) in this paper (we consider the theoretical problem in any space dimension, but the applications are only considered for $d = 1$ or $d = 2$).

H1. $\Omega \subset \mathbb{R}^d$ (with $d \in \mathbb{N}^\star$) is a bounded open subset, with a Lipschitz continuous boundary $\partial\Omega$ (this gives the existence, for a.e. $x \in \partial\Omega$, of the unit outward vector $\mathbf{n}(x)$ normal to the boundary),

H2. the function $g : \Omega \to \mathbb{R}^d$ is Lipschitz continuous on $\overline{\Omega}$ and satisfies $g(x) \cdot \mathbf{n}(x) = 0$ for a.e. $x \in \partial\Omega$,

H3. $F \in L^\infty(\Omega)$ is such that there exists $F_0 > 0$ with $F(x) \geq F_0$, for a.e. $x \in \Omega$.

As we show in [9], one can find an example of a discontinuous function $u : \Omega \to \mathbb{R}$ solution of (5)-(6). The regularity of $\mathrm{div}(ug)$ in the general case is an open problem. Therefore we first give a weak formulation of Problem (5)-(6), following [9]. For this purpose, let $\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+)$ and let $\xi \in C^1(\mathbb{R})$ be such that $\xi'(1) \geq 0$. We multiply the first inequality of (5) by $\xi'(u(x))\varphi(x)$ and we integrate on Ω. We get

$$\begin{aligned}\int_\Omega \xi'(u(x))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x = \quad & \int_\Omega \xi'(1)\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x + \\ & \int_\Omega (\xi'(u(x)) - \xi'(1))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x.\end{aligned} \tag{7}$$

The second term of the right hand side vanishes, using (6), and the first one is nonnegative. This leads to

$$\int_\Omega \xi'(u(x))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x \geq 0. \tag{8}$$

and we develop equation (8), integrating by parts. We then derive the following weak sense for a solution to Problem (5)-(6).

Definition 2.1 (Weak solution to Problem (5)-(6))
Under hypotheses (H), we say that a function $u \in L^\infty(\Omega)$ is a weak solution to Problem (5)-(6) if u satisfies the following inequalities : $0 \leq u(x) \leq 1$ for a.e. $x \in \Omega$ and

$$\begin{aligned}\int_\Omega \Big(\xi(u(x))(-g(x)\cdot\nabla\varphi(x)) + [\xi'(u(x))u(x) - \xi(u(x))]\varphi(x)\mathrm{div}g(x) + \\ \xi'(u(x))\varphi(x)F(x)\Big)\mathrm{d}x \geq 0, \\ \forall \xi \in C^1(\mathbb{R}) \text{ s.t. } \xi'(1) \geq 0, \ \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+).\end{aligned} \tag{9}$$

Remark 2.2 *If the test function ξ in (9) is such that $\xi'(1) = 0$, then the inequality becomes an equality. Note also that, if ξ is such that $\xi'(1) \geq 0$ and ξ' is nonincreasing, we can get (8) from (7), for any function u which only verifies (5). We show in Proposition 2.5 that one can indeed limit the test functions ξ in (9) to convex ones (in the sense that ξ' is nondecreasing, this terminology is used in this paper).*

We then define the set $\mathcal{C}(g, F)$ of functions which satisfy (5) in a weak sense.

Definition 2.3 *Under hypotheses (H), we define the convex set $\mathcal{C}(g, F)$ of functions $v \in L^\infty(\Omega)$, with $0 \leq v(x) \leq 1$, for a.e. $x \in \Omega$ and*

$$\int_\Omega ([-v(x)g(x)\cdot\nabla\varphi(x)] + \varphi(x)F(x))\,\mathrm{d}x \geq 0, \ \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+). \tag{10}$$

Note that the convexity of $\mathcal{C}(g, F)$ directly results from the linearity of the left hand side of (10) with respect to v. Since $0 \in \mathcal{C}(g, F)$, the set $\mathcal{C}(g, F)$ is therefore nonempty. We also note that $\mathcal{C}(g, F)$ is a closed subset of $L^p(\Omega)$ for all $p \in [1, \infty]$. Let us recall the following proposition, proven in [9] (under a slightly different formulation), which gives a characterization of the functions of $\mathcal{C}(g, F)$.

Proposition 2.4 (Characterization of $\mathcal{C}(g, F)$)
Under hypotheses (H), for all $v \in L^\infty(\Omega)$, the property $v \in \mathcal{C}(g, F)$ (defined in Definition 2.3) holds if and only if the following property holds

$$\begin{aligned}\int_\Omega \Big(\xi(v(x))[-g(x)\cdot\nabla\varphi(x)] + \left[\xi'(v(x))v(x) - \xi(v(x))\right]\varphi(x)\ \mathrm{div}g(x) + \\ \xi'(v(x))\varphi(x)F(x)\Big)\mathrm{d}x \geq 0, \\ \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+), \ \forall \xi \in C^1(\mathbb{R}) \text{ s.t. } \forall \kappa \in [0,1], \ \xi'(\kappa) \geq 0.\end{aligned} \tag{11}$$

Using the characterization 2.4, it is now possible to prove the following properties.

Proposition 2.5 (A convexity property)
Under hypotheses (H), a function $u \in L^\infty(\Omega)$ is a weak solution to Problem (5)-(6) in the sense of Definition 2.1 if and only if u satisfies the following inequalities : $0 \le u(x) \le 1$ for a.e. $x \in \Omega$ and

$$\begin{aligned}\int_\Omega \Big(\xi(u(x))(-g(x)\cdot\nabla\varphi(x)) + [\xi'(u(x))u(x) - \xi(u(x))]\varphi(x)\mathrm{div}g(x) + \\ \xi'(u(x))\varphi(x)F(x)\Big)\mathrm{d}x \ge 0, \\ \forall \xi \in C^1(\mathbb{R}) \text{ convex s.t. } \xi'(1) \ge 0,\ \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+).\end{aligned} \tag{12}$$

Proof. If u satisfies (9), it clearly satisfies (12). Conversely, let us assume that (12) holds. We immediately get $u \in \mathcal{C}(g,F)$, letting $\xi(s) = s$ in (12). Let $\xi \in C^2(\mathbb{R})$ be such that $\xi'(1) \ge 1$. We write $\xi = \xi_0 + \xi_1$, with $\xi_0'(t) = \xi'(1) + \int_1^t \xi''(s)^+ \mathrm{d}s$ and $\xi_1'(t) = -\int_1^t \xi''(s)^- \mathrm{d}s$. Since $\xi_1' \ge 0$, we write (11) with ξ_1, and since ξ_0' is nondecreasing and $\xi_0'(1) \ge 0$, we write (12) with ξ_0. Adding both inequalities leads to (9) for all $\xi \in C^2(\mathbb{R})$. Taking regularizations in $C^2(\mathbb{R})$ of $\xi \in C^1(\mathbb{R})$ permits to get (9) for all $\xi \in C^1(\mathbb{R})$. □

Proposition 2.6 (Maximum of two elements of $\mathcal{C}(g,F)$)
Under hypotheses (H), for all elements u and v of $\mathcal{C}(g,F)$ (this set is defined in Definition 2.3), then the function $w \in L^\infty(\Omega)$ defined, for a.e. $x \in \Omega$ by $w(x) = \max(u(x), v(x))$ satisfies $w \in \mathcal{C}(g,F)$.

Proof. Let $u, v \in \mathcal{C}(g,F)$ be given. We notice that, if u, v are regular enough, say $u, v \in W^{1,1}(\Omega)$, then the conclusion of the proposition is straightforward, since $\mathrm{div}(\max(u,v)g) + F = \mathrm{div}(ug) + F$ a.e. on the set $\{u \ge v\}$. For the general case, we use the doubling variable technique of Krushkov (see Proposition 5.1). We consider, for a given $\varepsilon > 0$, the function $S_\varepsilon \in C^1(\mathbb{R})$ defined by

$$\begin{array}{ll} S_\varepsilon(s) = 0, & \forall s \in (-\infty, 0], \\ S_\varepsilon(s) = s^2(3\varepsilon - 2s)/\varepsilon^3, & \forall s \in [0, \varepsilon], \\ S_\varepsilon(s) = 1, & \forall s \in [\varepsilon, +\infty). \end{array} \tag{13}$$

We define $\xi_\varepsilon(s) = \int_0^s S_\varepsilon(\tau)d\tau$ and we set, for all $(a,b) \in \mathbb{R}^2$, $\eta_\varepsilon(a,b) = a + \xi_\varepsilon(b-a)$. Then this function η_ε satisfies $\partial_1\eta_\varepsilon(a,b) = 1 - S_\varepsilon(b-a) \ge 0$ for all $(a,b) \in \mathbb{R}^2$, and $\partial_2\eta_\varepsilon(a,b) = S_\varepsilon(b-a) \ge 0$ for all $(a,b) \in \mathbb{R}^2$. Thanks to Proposition 2.4, the hypotheses of Proposition 5.1 are therefore satisfied by η_ε, u and v. We then obtain, for all $\phi \in C^1(\mathbb{R}^d, \mathbb{R}_+)$,

$$\begin{aligned}\int_\Omega \Big(\eta_\varepsilon(u(x),v(x))\,[-g(x)\cdot\nabla\phi(x)] + \\ \Big(\partial_1\eta_\varepsilon(u(x),v(x))u(x) + \partial_2\eta_\varepsilon(u(x),v(x))v(x) - \eta_\varepsilon(u(x),v(x))\Big)\phi(x)\mathrm{div}g(x) + \\ \Big(\partial_1\eta_\varepsilon(u(x),v(x)) + \partial_2\eta_\varepsilon(u(x),v(x))\Big)F(x)\phi(x)\Big)\mathrm{d}x \ge 0.\end{aligned} \tag{14}$$

We remark that, for all $(a,b) \in \mathbb{R}^2$,

$$a\partial_1\eta_\varepsilon(a,b) + b\partial_2\eta_\varepsilon(a,b) - \eta_\varepsilon(a,b) = a + (b-a)S_\varepsilon(b-a) - \eta_\varepsilon(a,b)$$

leads to

$$\lim_{\varepsilon\to 0}(a\partial_1\eta_\varepsilon(a,b) + b\partial_2\eta_\varepsilon(a,b) - \eta_\varepsilon(a,b)) = 0,$$

and we also remark that

$$\partial_1\eta_\varepsilon(a,b) + \partial_2\eta_\varepsilon(a,b) = 1.$$

Thus, using the Lebesgue dominated convergence theorem, we can let $\varepsilon \to 0$ in (14). This leads to

$$\int_\Omega \Big((u(x) + (v(x)-u(x))^+)\,[-g(x)\cdot\nabla\phi(x)] + \phi(x)F(x)\Big)\,\mathrm{d}x \quad \ge 0. \tag{15}$$

Since $w(x) = \max(u(x), v(x)) = u(x) + (v(x)-u(x))^+$ for a.e. $x \in \Omega$, we thus get that (10) is satisfied by w, which proves that, since $0 \le w(x) \le 1$ for a.e. $x \in \Omega$, $w \in \mathcal{C}(g,F)$, and thus concludes the proof of the proposition. □

Remark 2.7 *A similar property holds for the minimum of two elements of $\mathcal{C}(g,F)$ (it suffices to consider $\eta_\varepsilon(a,b) = b - \xi_\varepsilon(b-a)$).*

Proposition 2.8 (Maximal element of $\mathcal{C}(g,F)$)
Under hypotheses (H), there exists $u \in \mathcal{C}(g,F)$ such that, for all $v \in \mathcal{C}(g,F)$, $v(x) \leq u(x)$ for a.e. $x \in \Omega$. This element is therefore unique, and is called the maximal element of $\mathcal{C}(g,F)$. An immediate consequence is that this maximal element is equal to the projection in $L^2(\Omega)$ on $\mathcal{C}(g,F)$ of the function 1_Ω (defined by $1_\Omega(x) = 1$ for $x \in \Omega$).

Proof. Since $L^1(\Omega)$ is separable and $\mathcal{C}(g,F)$ is a closed subset of $L^1(\Omega)$, then there exists a sequence $(v_n)_{n\in\mathbb{N}}$ of elements of $\mathcal{C}(g,F)$, dense in $\mathcal{C}(g,F)$. For all $n \in \mathbb{N}$, we define $u_n \in L^\infty(\Omega)$ by $u_n = \max_{m=0,\dots,n} v_m$. We get, from Proposition 2.6, that $u_n \in \mathcal{C}(g,F)$. Since the sequence $(u_n)_{n\in\mathbb{N}}$ is nondecreasing and is bounded by 1, it converges in $L^1(\Omega)$ to some $u \in L^\infty(\Omega)$ such that $0 \leq u(x) \leq 1$ for a.e. $x \in \Omega$. We then get, since $\mathcal{C}(g,F)$ is a closed subset of $L^1(\Omega)$, that $u \in \mathcal{C}(g,F)$. Let $v \in \mathcal{C}(g,F)$ and let $\varepsilon > 0$. There exists $n \in \mathbb{N}$ such that $\|v - v_n\|_{L^1(\Omega)} \leq \varepsilon$. This implies $\|(v-v_n)^+\|_{L^1(\Omega)} \leq \varepsilon$. Since $v_n(x) \leq u(x)$ for a.e. $x \in \Omega$, we then get that $\|(v-u)^+\|_{L^1(\Omega)} \leq \varepsilon$. Letting $\varepsilon \to 0$ gives $\|(v-u)^+\|_{L^1(\Omega)} = 0$, which concludes the proof of the proposition. □
We can now prove, using Hypothesis (H3), that any weak solution to Problem (5)-(6) in the sense of Definition 2.1 is the maximal element of $\mathcal{C}(g,F)$.

Proposition 2.9 (The weak solution is maximal)
Under hypotheses (H), let u be a weak solution to Problem (5)-(6) in the sense of Definition 2.1. Then u is the maximal element of $\mathcal{C}(g,F)$.

Proof. The proof is divided in two steps. In the first step, we show that, for all $v \in \mathcal{C}(g,F)$, we have $g \cdot \nabla \mathrm{sign}^+(v-u) = 0$ (defining $\mathrm{sign}^+ : \mathbb{R} \to \mathbb{R}$ by $\mathrm{sign}^+(s) = 1$ for all $s > 0$ and $\mathrm{sign}^+(s) = 0$ otherwise). We then remark, in the second step, that if $\mathrm{sign}^+(v-u) > 0$, then $u < 1$, and that $\mathrm{div}(ug) + F \leq 0$ where $u < 1$. Then the inequality $\int_\Omega \mathrm{sign}^+(v-u)(\mathrm{div}(ug)+F)\mathrm{d}x \leq 0$ suffices to obtain, thanks to Hypothesis (H3) (namely the fact that $F > 0$ a.e.), that $v \leq u$ almost everywhere in Ω.

Step 1

We again consider the function S_ε defined by (13), and we set for all $(a,b) \in \mathbb{R}^2$, $\eta_\varepsilon(a,b) = \int_0^{b-a} S_\varepsilon(s)\mathrm{d}s$. Then this function η_ε satisfies $\partial_1\eta_\varepsilon(1,b) = -S_\varepsilon(b-1) = 0$ for all $b \leq 1$, and $\partial_2\eta_\varepsilon(a,b) = S_\varepsilon(b-a) \geq 0$ for all $(a,b) \in \mathbb{R}^2$. Hence (9) holds with $\xi(a) = \eta_\varepsilon(a,b)$ for all $b \in [0,1]$, and (11) holds with $\xi(b) = \eta_\varepsilon(a,b)$ for all $a \in [0,1]$. Since the hypotheses of Proposition 5.1 are then verified by u, v, η_ε for any $v \in \mathcal{C}(g,F)$, we get, for a given $\phi \in C^1(\mathbb{R}^d, \mathbb{R}_+)$,

$$\begin{aligned}\int_\Omega \Big(&\eta_\varepsilon(u(x),v(x))\left[-g(x)\cdot\nabla\phi(x) - \phi(x)\mathrm{div}g(x)\right] + \\ &\Big(\partial_1\eta_\varepsilon(u(x),v(x))u(x) + \partial_2\eta_\varepsilon(u(x),v(x))v(x)\Big)\phi(x)\mathrm{div}g(x) + \\ &\Big(\partial_1\eta_\varepsilon(u(x),v(x)) + \partial_2\eta_\varepsilon(u(x),v(x))\Big)F(x)\phi(x)\Big)\mathrm{d}x \geq 0.\end{aligned} \tag{16}$$

Since, for all $(a,b) \in \mathbb{R}^2$,

$$a\partial_1\eta_\varepsilon(a,b) + b\partial_2\eta_\varepsilon(a,b) - \eta_\varepsilon(a,b) = (b-a)S_\varepsilon(b-a) - \eta_\varepsilon(a,b)$$

leads to

$$\lim_{\varepsilon\to 0}(a\partial_1\eta_\varepsilon(a,b) + b\partial_2\eta_\varepsilon(a,b) - \eta_\varepsilon(a,b)) = 0,$$

and since

$$\partial_1\eta_\varepsilon(a,b) + \partial_2\eta_\varepsilon(a,b) = 0,$$

we get, letting $\varepsilon \to 0$ in (16) thanks to the dominated convergence theorem,

$$-\int_\Omega (v(x)-u(x))^+ g(x)\cdot\nabla\phi(x)\mathrm{d}x \geq 0,\ \forall\phi \in C^1(\mathbb{R}^d,\mathbb{R}_+). \tag{17}$$

Remark 2.10 *Inequality (17) is also proven in [9], in which the hypothesis $g = \nabla h$, not assumed in this paper, permits to conclude to the uniqueness of ug.*

We can then apply Proposition 5.2, which leads to

$$\int_\Omega \Big(\text{sign}^+(v(x)-u(x))\left[-g(x)\cdot\nabla\phi(x)-\phi(x)\text{div}g(x)\right]\Big)\text{d}x = 0,\ \forall \phi \in C^1(\mathbb{R}^d,\mathbb{R}), \tag{18}$$

where the application $\text{sign}^+ : \mathbb{R}\to\mathbb{R}$ is defined by $\text{sign}^+(s)=1$ for all $s>0$ and by $\text{sign}^+(s)=0$ for all $s\le 0$.

Step 2

Since $S_\varepsilon(0)=0$, let us now introduce the function $\xi_\varepsilon : a \mapsto \int_0^a S_\varepsilon(1-s)\text{d}s$ in (9). We get

$$\begin{aligned}\int_\Omega \Big(\xi_\varepsilon(u(x))(-g(x)\cdot\nabla\varphi(x)) + [\xi_\varepsilon'(u(x))u(x)-\xi(u(x))]\varphi(x)\text{div}g(x) + \\ \xi_\varepsilon'(u(x))\varphi(x)F(x)\Big)\text{d}x \ge 0, \\ \forall \varphi \in C^1(\overline{\Omega},\mathbb{R}_+).\end{aligned} \tag{19}$$

We have that, for all $a\in[0,1]$, $\xi_\varepsilon(a)\to -a$ as $\varepsilon\to 0$, and that, for all $a\in[0,1[$, $\xi_\varepsilon'(a)a-\xi(a)\to 0$ and $\xi_\varepsilon'(a)\to -1$ as $\varepsilon\to 0$. We then get, thanks to the dominated convergence theorem, letting $\varepsilon\to 0$ in (19),

$$\begin{aligned}\int_\Omega u(x)g(x)\cdot\nabla\varphi(x)\text{d}x + \int_{\{x\in\Omega,u(x)=1\}}\varphi(x)\text{div}g(x)\text{d}x - \int_{\{x\in\Omega,u(x)<1\}}\varphi(x)F(x)\text{d}x \ge 0, \\ \forall\varphi\in C^1(\overline{\Omega},\mathbb{R}_+).\end{aligned} \tag{20}$$

(In fact, this inequality is an equality, see Remark 2.2).
We introduce a sequence of mollifiers in $\mathbb{R}^d$. Let $\rho\in C_c^\infty(\mathbb{R}^d,\mathbb{R}_+)$ (the set of smooth functions with a compact support) be such that

$$\{x\in\mathbb{R}^d;\ \rho(x)\neq 0\}\subset\{x\in\mathbb{R}^d;\ |x|\le 1\}, \tag{21}$$

and

$$\int_{\mathbb{R}^d}\rho(x)\text{d}x = 1. \tag{22}$$

For $n\in\mathbb{N}^\star$, we define

$$\rho_n(x) = n^d\rho(nx),\ \forall x\in\mathbb{R}^d. \tag{23}$$

Let $n\in\mathbb{N}$. We set $\varphi_n(x)=\int_\Omega \rho_n(x-y)\ \text{sign}^+(v(y)-u(y))\text{d}y$ in (20). Defining $T_1^{(n)}$, $T_2^{(n)}$ and $T_3^{(n)}$ by

$$T_1^{(n)} = \int_\Omega\int_\Omega u(x)g(x)\cdot\nabla\rho_n(x-y)\ \text{sign}^+(v(y)-u(y))\text{d}y\text{d}x,$$

$$T_2^{(n)} = \int_{\{x\in\Omega,u(x)=1\}}\int_\Omega \rho_n(x-y)\ \text{sign}^+(v(y)-u(y))\text{d}y\text{div}g(x)\text{d}x,$$

and

$$T_3^{(n)} = -\int_{\{x\in\Omega,u(x)<1\}}\int_\Omega \rho_n(x-y)\ \text{sign}^+(v(y)-u(y))F(x)\text{d}y\text{d}x,$$

we get

$$T_1^{(n)}+T_2^{(n)}+T_3^{(n)}\ge 0. \tag{24}$$

We then have, $T_1^{(n)}=T_4^{(n)}+T_5^{(n)}+T_6^{(n)}$ defining $T_4^{(n)}$, $T_5^{(n)}$ and $T_6^{(n)}$ by

$$T_4^{(n)} = \int_\Omega\int_\Omega u(y)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\ \text{sign}^+(v(y)-u(y))\text{d}y\text{d}x,$$

$$T_5^{(n)} = \int_\Omega\int_\Omega (u(x)-u(y))(g(x)-g(y))\cdot\nabla\rho_n(x-y)\ \text{sign}^+(v(y)-u(y))\text{d}y\text{d}x,$$

$$T_6^{(n)} = \int_\Omega\int_\Omega u(x)g(y)\cdot\nabla\rho_n(x-y)\ \text{sign}^+(v(y)-u(y))\text{d}y\text{d}x.$$

Thanks to an integrate by parts with respect to x, we get

$$T_4^{(n)} = -\int_\Omega \int_\Omega u(y)\rho_n(x-y)\ \mathrm{sign}^+(v(y)-u(y))\mathrm{div}(g(x))\mathrm{d}y\mathrm{d}x,$$

and therefore $T_4^{(n)} = T_7^{(n)} + T_8^{(n)}$, with

$$T_7^{(n)} = -\int_\Omega \int_\Omega u(x)\rho_n(x-y)\ \mathrm{sign}^+(v(y)-u(y))\mathrm{div}(g(y))\mathrm{d}y\mathrm{d}x,$$

$$T_8^{(n)} = -\int_\Omega \int_\Omega \rho_n(x-y)\ \mathrm{sign}^+(v(y)-u(y))(u(y)\mathrm{div}(g(x)) - u(x)\mathrm{div}(g(y)))\mathrm{d}y\mathrm{d}x.$$

We now remark that, letting $\phi(x) = \rho_n(y-x)$ in (18) (recall that the gradient of $\rho_n(y-x)$ with respect to x is equal to $-\nabla\rho_n(y-x)$), multiplying by $u(y)$ and integrating on Ω, we get $T_6^{(n)} + T_7^{(n)} = 0$. Since we easily obtain

$$\lim_{n\to+\infty} T_5^{(n)} = 0,$$

and

$$\lim_{n\to+\infty} T_8^{(n)} = 0,$$

we thus get

$$\lim_{n\to+\infty} T_1^{(n)} = 0.$$

Since, for a.e. $x \in \Omega$, $v(x) > u(x)$ implies $u(x) < 1$, we have

$$\lim_{n\to+\infty} T_3^{(n)} = -\int_{\{x\in\Omega, u(x)<1\}} \mathrm{sign}^+(v(x)-u(x))F(x)\mathrm{d}x = -\int_{\{x\in\Omega, v(x)>u(x)\}} F(x)\mathrm{d}x,$$

and

$$\lim_{n\to+\infty} T_2^{(n)} = \int_{\{x\in\Omega, u(x)=1\}} \mathrm{sign}^+(v(x)-u(x))\mathrm{div}g(x)\mathrm{d}x = 0,$$

we get passing to the limit $n \to \infty$ in (24),

$$-\int_{\{x\in\Omega, v(x)>u(x)\}} F(x)\mathrm{d}x \geq 0.$$

Thanks to Hypothesis (H3), this implies that $v(x) \leq u(x)$ for a.e. $x \in \Omega$ and thus concludes the proof. □

Using Theorem 4.14, which expresses the convergence of a numerical scheme, we can state the following proposition.

Proposition 2.11 (The maximal element is a weak solution)
Under hypotheses (H), the maximal element of $\mathcal{C}(g,F)$ is a weak solution of Problem (5)-(6) in the sense of Definition 2.1.

Therefore, we conclude from Propositions 2.9 and 2.11 the following theorem.

Theorem 2.12 (The weak solution is unique and is the maximal element)
Under hypotheses (H), the maximal element of $\mathcal{C}(g,F)$ is the unique weak solution of Problem (5)-(6) in the sense of Definition 2.1.

We present now a characterization of the maximal element of $\mathcal{C}(g,F)$ using a Lagrange multiplier. This characterization will be useful to define is Section 3 the notion of "process maximal element" of $\mathcal{C}(g,F)$.

Since the maximal element of $\mathcal{C}(g,F)$ is the projection in $L^2(\Omega)$ of the function 1_Ω on $\mathcal{C}(g,F)$ (this is proven in Proposition 2.8), it is the unique solution of the following problem, which is a minimization problem under constraints :

$$\begin{array}{l} u \in C(g,F), \\ J(u) \leq J(v),\ \forall v \in C(g,F), \end{array} \tag{25}$$

where $J(v) : \int_\Omega (1-v(x))^2\,\mathrm{d}x$ for all $v \in C(g,F)$.

The set $C(g,F)$ is defined with the three constraints, $u \geq 0$, $u \leq 1$ and $\mathrm{div}(ug) + F \geq 0$. Thanks to the hypothesis $F \geq F_0$ a.e, with $F_0 > 0$, we can prove that the first constraint is not active. Indeed, for $\underline{u} > 0$ small enough, the function $\underline{u}1_\Omega$ belongs to $C(g,F)$ which gives that the maximal element u of $C(g,F)$ (which is the solution of (25)) satisfies $u \geq \underline{u}$ a.e.. The two other constraints are possibly active and we can prove the existence of Lagrange multipliers linked with these constraints (thus stating a generalization of the Kuhn-Tucker theorem in this infinite dimension case): let u be the solution of (25), a consequence of Theorem 4.14 is the existence of a function $\mu \in L^2(\Omega)$, $\mu \geq 0$ a.e., and of a finite nonnegative measure on $\overline{\Omega}$, denoted by ν, such that:

$$\int_\Omega (1-u(x))\phi(x)\mathrm{d}x + \int_\Omega \mu(x)\mathrm{div}(\phi g)(x)\mathrm{d}x - \int_\Omega \phi(x)d\nu(x) = 0,\ \forall \phi \in C^1(\overline{\Omega},\mathbb{R}). \tag{26}$$

Furthermore, one formally gets that $\mu(\mathrm{div}(ug)+F) = 0$ and $\nu(1-u) = 0$ (this can be only formal since the regularity proven for u, μ and ν does not suffices to give a precise sense to these quantities), which means that the multiplier is nonzero only when the corresponding constraint is active.

Then, taking $\phi = u - v$ in (26), one obtains, again formally since the regularity of this function is not sufficient:

$$\int_\Omega (1-u(x))(u(x)-v(x))\mathrm{d}x - \int_\Omega \mu(x)(\mathrm{div}(vg)(x)+F(x))\mathrm{d}x - \int_\Omega (1-v(x))d\nu(x) = 0, \quad \forall v \in C^1(\overline{\Omega},\mathbb{R}). \tag{27}$$

Considering only the functions v such that $v \leq 1$, (27) leads to:

$$\int_\Omega (1-u(x))(u(x)-v(x))\mathrm{d}x - \int_\Omega \mu(x)(\mathrm{div}(vg)(x)+F(x))\mathrm{d}x \geq 0,\ \forall v \in C^1(\overline{\Omega},\mathbb{R}), v \leq 1. \tag{28}$$

Although (27) is only formally obtained from (26), the discrete counterpart of the former, i.e. equation (55), is rigorously deduced in Section 4 from (52), the discrete counterpart of the latter. Then, passing to the limit (Theorem 4.14) and using a uniqueness result (Proposition 3.2 in Section 3), we prove that (28) gives a characterization of the solution to (25), leading to the following proposition.

Proposition 2.13 (Characterization of the maximal element) *Under hypotheses (H), u is the unique solution of (25), i.e. the maximal element of $\mathcal{C}(g,F)$, if and only if $u \in \mathcal{C}(g,F)$ and there exists $\mu \in L^2(\Omega)$, $\mu \geq 0$ a.e., such that (28) holds.*

3 Uniqueness of the process maximal element

Since we consider below the convergence of numerical schemes, on which we only prove an $L^\infty(\Omega)$ estimate, we have to introduce, for technical reasons, a definition of "maximal element" of $\mathcal{C}(g,F)$ in a weaker sense that the one given in Proposition 2.8. This new notion is called process maximal element (this notion is an extension of that introduced in [11], related to the notion of Young measure, first used by [4] in the nonlinear scalar hyperbolic framework). When this process maximal element reduces to a classical function (which is shown in Proposition 3.2), the definition below gives (28).

Definition 3.1 (Process maximal element of $\mathcal{C}(g,F)$)
Under hypotheses (H), we say that a function $u \in L^\infty(\Omega\times(0,1))$ is a process maximal element of $\mathcal{C}(g,F)$ if the function $\bar{u} : x \mapsto \int_0^1 u(x,\alpha)\mathrm{d}\alpha$ is such that $\bar{u} \in \mathcal{C}(g,F)$ and there exists $\mu \in L^2(\Omega)$ such that the pair (u,μ) satisfies the following inequalities : $0 \leq u(x,\alpha) \leq 1$ and $0 \leq \mu(x)$ for a.e. $(x,\alpha) \in \Omega\times(0,1)$ and

$$\int_\Omega \left[\int_0^1 (1-u(x,\alpha))(u(x,\alpha)-\varphi(x))\mathrm{d}\alpha - \mu(x)(\mathrm{div}(\varphi(x)g(x))+F(x))\right]\mathrm{d}x \geq 0, \quad \forall \varphi \in C^1(\overline{\Omega},[0,1]). \tag{29}$$

We now state the uniqueness of such a process maximal element, indeed equal to the projection of 1_Ω in $L^2(\Omega)$ on $\mathcal{C}(g,F)$.

Proposition 3.2 (Uniqueness of the process maximal element) *Under hypotheses (H), let u be a process maximal element of $\mathcal{C}(g,F)$ in the sense of Definition (3.1). Then the following inequality holds:*

$$\int_\Omega \int_0^1 (1-u(x,\alpha))(u(x,\alpha)-v(x))\mathrm{d}\alpha\mathrm{d}x \geq 0,\ \forall v \in \mathcal{C}(g,F). \tag{30}$$

As an immediate consequence, we get that the function $\bar{u}$ defined in Definition 3.1 is such that $\bar{u}(x) = u(x,\alpha)$ for a.e. $(x,\alpha) \in \Omega \times (0,1)$ and $\bar{u}$ is the unique maximal element of $\mathcal{C}(g,F)$.

Proof. We again use the sequence $(\rho_n)_{n\in\mathbb{N}}$ of mollifiers in $\mathbb{R}^d$ defined by (21)-(23). We then introduce the functions $v_n(y) = \int_\Omega v(x)\rho_n(x-y)\mathrm{d}x$ in (29) and the functions $\mu_n(x) = \int_\Omega \mu(y)\rho_n(x-y)\mathrm{d}y$ in (10). We then get

$$\int_\Omega \left[\int_0^1 (1-u(y,\alpha))(u(y,\alpha)-v_n(y))\mathrm{d}\alpha - \mu(y)(\mathrm{div}(gv_n)(y)+F(y))\right]\mathrm{d}y \geq 0,$$

and

$$\int_\Omega \left([-v(x)g(x)\cdot\nabla\mu_n(x)] + \mu_n(x)F(x)\right)\mathrm{d}x \geq 0.$$

We sum the two above inequalities. Defining $T_9^{(n)}$, $T_{10}^{(n)}$ and $T_{11}^{(n)}$ by

$$\begin{aligned}
T_9^{(n)} &= \int_\Omega\int_0^1 (1-u(y,\alpha))(u(y,\alpha)-v_n(y))\mathrm{d}\alpha\mathrm{d}y\\
T_{10}^{(n)} &= +\int_\Omega\int_\Omega \mu(y)v(x)\left(g(y)\cdot\nabla\rho_n(x-y)+\rho_n(x-y)\mathrm{div}g(y)+g(x)\cdot\nabla\rho_n(x-y)\right)\mathrm{d}x\mathrm{d}y\\
T_{11}^{(n)} &= \int_\Omega [-\mu(y)F(y)+\mu_n(y)F(y)]\,\mathrm{d}y,
\end{aligned}$$

we get

$$T_9^{(n)} + T_{10}^{(n)} + T_{11}^{(n)} \geq 0. \tag{31}$$

We have $T_{10}^{(n)} = T_{12}^{(n)} + T_{13}^{(n)} + T_{14}^{(n)}$, with

$$\begin{aligned}
T_{12}^{(n)} &= -\int_\Omega\int_\Omega \mu(y)(v(x)-v(y))(g(x)-g(y))\cdot\nabla\rho_n(x-y)\mathrm{d}x\mathrm{d}y\\
T_{13}^{(n)} &= -\int_\Omega\int_\Omega \mu(y)v(y)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\mathrm{d}x\mathrm{d}y\\
T_{14}^{(n)} &= -\int_\Omega\int_\Omega \mu(y)v(x)\rho_n(x-y)\mathrm{div}g(y)\mathrm{d}x\mathrm{d}y.
\end{aligned}$$

Thanks to the fact that $(x,y) \mapsto (g(x)-g(y))\cdot\nabla\rho_n(x-y)$ vanishes for $|x-y| > 1/n$ and belongs to $L^1(\Omega)$ since g is regular, we can apply the theorem of continuity in means applied to the function v. We thus get that

$$\lim_{n\to\infty} T_{12}^{(n)} = 0.$$

We have, thanks to an integrate by parts with respect to x, that $T_{13}^{(n)} = T_{15}^{(n)} + T_{16}^{(n)}$ with

$$T_{15}^{(n)} = \int_\Omega\int_\Omega \mu(y)v(y)\rho_n(x-y)\mathrm{div}g(x)\mathrm{d}x\mathrm{d}y,$$

and

$$T_{16}^{(n)} = -\int_\Omega \mu(y)v(y)\int_{\partial\Omega}\rho_n(x-y)(g(x)-g(y))\cdot\mathbf{n}(x)\mathrm{d}\gamma(x)\mathrm{d}y.$$

We get, using $|g(x)-g(y)| \leq C_1|y-x|$ (where C_1 only depends on g), $0 \leq v(x) \leq 1$ and the Cauchy-Schwarz inequality,

$$\left(T_{16}^{(n)}\right)^2 \leq C_1^2\int_\Omega\int_{\partial\Omega}|y-x|\rho_n(x-y)\mathrm{d}\gamma(x)\mathrm{d}y\int_\Omega\int_{\partial\Omega}\mu(y)^2|y-x|\rho_n(x-y)\mathrm{d}\gamma(x)\mathrm{d}y,$$

which gives

$$\left(T_{16}^{(n)}\right)^2 \leq C_1^2\frac{1}{n}\int_{\partial\Omega}\left(\int_\Omega\rho_n(x-y)\mathrm{d}y\right)\mathrm{d}\gamma(x)\int_\Omega\mu(y)^2\left(\int_{\partial\Omega}|y-x|\rho_n(x-y)\mathrm{d}\gamma(x)\right)\mathrm{d}y.$$

We have on one hand $\int_\Omega \rho_n(x-y)\mathrm{d}y \leq 1$, and on the other hand, for all $y \in \Omega$, $\int_{\partial\Omega} |y-x|\rho_n(x-y)\mathrm{d}\gamma(x) \leq n^d \frac{1}{n}\gamma\left(B(y,\frac{1}{n}) \cap \partial\Omega\right) \leq C_2$ (where C_2 only depends on d and Ω). We thus get

$$\left(T_{16}^{(n)}\right)^2 \leq C_1^2 \frac{1}{n}\mathrm{meas}(\partial\Omega)C_2 \int_\Omega \mu(y)^2 \mathrm{d}y,$$

and therefore

$$\lim_{n\to\infty} T_{16}^{(n)} = 0$$

We then get that

$$\lim_{n\to\infty} T_{15}^{(n)} = -\lim_{n\to\infty} T_{14}^{(n)} = \int_\Omega \mu(y)v(y)\mathrm{div}g(y)\mathrm{d}y.$$

Gathering the above results gives (30). Applying (30) when v is the projection of 1_Ω on $\mathcal{C}(g,F)$ gives

$$\int_\Omega \int_0^1 (1-u(y,\alpha))(u(y,\alpha)-v(y))\mathrm{d}\alpha\mathrm{d}y \geq 0,$$

and the characterization of this projection gives, since $\bar{u} \in \mathcal{C}(g,F)$ which implies that

$$\int_\Omega \int_0^1 (1-v(y))(v(y)-u(y,\alpha))\mathrm{d}\alpha\mathrm{d}y \geq 0.$$

The sum of the two above inequalities then gives

$$-\int_\Omega \int_0^1 (v(y)-u(y,\alpha))^2\mathrm{d}\alpha\mathrm{d}y \geq 0,$$

which gives the conclusion. □

4 Passing to the limit on numerical schemes

We now start the study of the convergence of numerical schemes, which are based, in the industrial framework, on finite volume methods. Let us first define the notion of admissible mesh, following [12].

Definition 4.1 (Admissible meshes) *An admissible finite volume mesh of Ω, denoted by $\mathcal{T}$, is given by a finite family of disjoint polygonal (one uses here the two space dimensions terms, for the setting of the general space dimension) connected subsets of $\mathbb{R}^d$ such that Ω is the union of the closure of the elements of $\mathcal{T}$ (which are called control volumes in the following) and such that the common "interface" of any pair of neighboring control volumes is included in a hyperplane of $\mathbb{R}^d$ (this is not necessary but is introduced in order to simplify the formulation). We denote by* $\mathrm{size}(\mathcal{T}) = \sup\{\mathrm{diam}(K), K \in \mathcal{T}\}$, *by m_K the measure of K, for all $K \in \mathcal{T}$, by $\mathcal{N}_K$ the subset of $\mathcal{T}$ of all the control volumes having a common interface with K. We then denote by $\mathcal{E}$ one set of pairs of neighbors $(K,L) \in \mathcal{T}^2$, such that, if $(K,L) \in \mathcal{E}$, $(L,K) \notin \mathcal{E}$, and for all $K \in \mathcal{T}$ and $L \in \mathcal{N}_K$, $(K,L) \in \mathcal{E}$ or $(L,K) \in \mathcal{E}$. For $K \in \mathcal{T}$ and $L \in \mathcal{N}_K$, we denote by m_{KL} the measure of the common interface between K and L. We measure the regularity of the mesh by means of the following expression:* $\mathrm{regul}(\mathcal{T}) = \max\{\sum_{L\in\mathcal{N}_K} m_{KL}\mathrm{diam}(K)/m_K,\ K \in \mathcal{T}\}$.

Let $\mathcal{T}$ be an admissible mesh of Ω. Let $g_\mathcal{T} := (g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that

$$g_{K,L} = -g_{L,K},\ \forall K \in \mathcal{T},\ \forall L \in \mathcal{N}_K \tag{32}$$

and

$$\sum_{L\in\mathcal{N}_K} g_{K,L} = \int_K \mathrm{div}g(x)\mathrm{d}x := G_K,\ \forall K \in \mathcal{T}. \tag{33}$$

Denoting by

$$F_K = \int_K F(x)\mathrm{d}x, \tag{34}$$

the finite volume scheme, in order to approximate Problem (5)-(6), is given by

$$\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K = 0 \text{ and } u_K \leq 1 \text{ or} \tag{35}$$

$$\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K \geq 0 \text{ and } u_K = 1. \tag{36}$$

We define the function $u_\mathcal{T}$ by

$$u_\mathcal{T}(x) = u_K,\ \forall x \in K,\ \forall K \in \mathcal{T}. \tag{37}$$

We then define the following value, which measures the consistency of the approximation $g_\mathcal{T}$ of the fluxes by means of a discrete $L^2(\Omega)^d$ norm, and which is expected to tend to 0 with size($\mathcal{T}$):

$$\text{cons}(g_\mathcal{T}) = \sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}_K}\frac{\text{diam}(K)}{m_{KL}}\left(g_{K,L} - \bar{g}_{K,L}\right)^2, \tag{38}$$

where

$$\bar{g}_{K,L} = \int_{K|L} g(x)\cdot\mathbf{n}_{K,L}\mathrm{d}s(x),\ \forall K \in \mathcal{T},\ \forall L \in \mathcal{N}_K. \tag{39}$$

Different choices are possible for $g_\mathcal{T}$. We can propose, for example:

- The choice $g_{K,L} = \bar{g}_{K,L}$, for all $K \in \mathcal{T}$ and $L \in \mathcal{N}_K$, is the simplest one which satisfies that cons($g_\mathcal{T}$) tends to 0 as size($\mathcal{T}$) tends to 0. Unfortunately, it demands in the general case to know an analytical expression of g.
- In the framework of the coupled problem given in the introduction to this paper, the field $g = \Lambda\nabla h$ is not analytically known, and it must be approximated. This can be achieved, using for example the finite volume method (see [12] for the isotropic case and [13] for the general case). The notion of admissible meshes must then be restricted to the case where there exists, for all $K \in \mathcal{T}$, a point x_K in the control volume K such that, for a pair of two neighboring grid blocks K and L, the line (x_K, x_L) is orthogonal to the interface $\bar{K}\cap\bar{L}$ between these grid blocks. Let us recall the scheme in the isotropic case: one defines $\tau_{KL} = \int_{\bar{K}\cap\bar{L}}\Lambda(x)\mathrm{d}s(x)/d(x_K, x_L)$, where we denote by $\mathrm{d}s(x)$ the $d-1$ Lebesgue measure at point $x \in \bar{K}\cap\bar{L}$. One can then compute the family $(h_K)_{K\in\mathcal{T}}$ of reals such that (33) holds under the condition

  $$g_{K,L} = \tau_{KL}(h_L - h_K),\ \forall K \in \mathcal{T},\ \forall L \in \mathcal{N}_K, \tag{40}$$

  in addition to such a relation as $\sum_{K\in\mathcal{T}} m_K h_K = 0$ (this corresponds to the discrete solution of a homogeneous Neumann problem). One can then prove that, under Hypotheses (H), cons($g_\mathcal{T}$) tends to 0 as size($\mathcal{T}$) tends to 0 (see [12] and [20]).
- In the same way, one can compute a mixed finite element approximate for $g_{K,L}$ which also satisfies that cons($g_\mathcal{T}$) tends to 0 as size($\mathcal{T}$) tends to 0 (see [6]).

We then have the following property, which is available under Hypotheses (H) of this paper and in particular (H3), which was not proven under the hypotheses made in [9].

Proposition 4.2 (A positivity property) *Under Hypotheses (H), let $\mathcal{T}$ be an admissible mesh of Ω and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that (32) and (33) are satisfied and let $(F_K)_{K\in\mathcal{T}}$ be defined by (34). Let $(u_K)_{K\in\mathcal{T}}$ be a solution to System (35)-(36). Then, the property $u_K > 0$, for all $K \in \mathcal{T}$, holds.*

Proof. Let us prove Proposition 4.2 by contradiction. Let us assume that the set $\mathcal{T}_- = \{K \in \mathcal{T};\ u_K \leq 0\}$ is not empty. Then, if $K \in \mathcal{T}_-$, one has $u_K < 1$, and therefore

$$\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K = 0,\ \forall K \in \mathcal{T}_-. \tag{41}$$

Summing (41) for $K \in \mathcal{T}_-$ leads to

$$\sum_{K\in\mathcal{T}_-}\sum_{L\in\mathcal{N}_K\setminus\mathcal{T}_-}(g^+_{K,L}u_L - g^-_{K,L}u_K) + \sum_{K\in\mathcal{T}_-}F_K = 0. \tag{42}$$

Since $u_K \leq 0$ for $K \in \mathcal{T}_-$ and $u_L > 0$ for $L \notin \mathcal{T}_-$, (42) gives $F_K = 0$ for all $K \in \mathcal{T}_-$, which is in contradiction with Hypothesis (H3). □

Definition 4.3 (The discrete convex set) *Under Hypotheses (H), let $\mathcal{T}$ be an admissible mesh of Ω and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that (32) and (33) are satisfied and let $(F_K)_{K\in\mathcal{T}}$ be defined by (34). We then define $\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$ (this set of functions is a natural discretization of $\mathcal{C}(g,F)$) as the set of all $(v_K)_{K\in\mathcal{T}}$ such that, for all $K\in\mathcal{T}$, the following inequalities hold:*

$$0\le v_K\le 1 \text{ and } \sum_{L\in\mathcal{N}_K}(g^+_{K,L}v_L-g^-_{K,L}v_K)+F_K\ge 0,\ \forall K\in\mathcal{T}. \tag{43}$$

Note that the set $\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$ is closed and nonempty since $(0)_{K\in\mathcal{T}}\in\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$. We can then prove the existence of at least one solution to System (35)-(36).

Proposition 4.4 (Property of the maximal element of $\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$) *Under Hypotheses (H), let $\mathcal{T}$ be an admissible mesh of Ω and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that (32) and (33) are satisfied and let $(F_K)_{K\in\mathcal{T}}$ be defined by (34). Let us denote by $(u_K)_{K\in\mathcal{T}}$ the family defined, for all $K\in\mathcal{T}$, by $u_K=\sup_{v\in\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})} v_K$. Then $(u_K)_{K\in\mathcal{T}}$ is a solution to System (35)-(36).*

Proof. Let us first denote by $(v_K)_{K\in\mathcal{T}}$ and $(w_K)_{K\in\mathcal{T}}$ two elements of $\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$. Then

$$(\max(v_K,w_K))_{K\in\mathcal{T}}\in\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T}) \tag{44}$$

Indeed, we have, for all $K\in\mathcal{T}$,

$$v_K\sum_{L\in\mathcal{N}_K}g^-_{K,L}\le\sum_{L\in\mathcal{N}_K}g^+_{K,L}v_L+F_K\le\sum_{L\in\mathcal{N}_K}g^+_{K,L}\max(v_L,w_L)+F_K,$$

and

$$w_K\sum_{L\in\mathcal{N}_K}g^-_{K,L}\le\sum_{L\in\mathcal{N}_K}g^+_{K,L}w_L+F_K\le\sum_{L\in\mathcal{N}_K}g^+_{K,L}\max(v_L,w_L)+F_K.$$

Therefore, since $\max\left(v_K\sum_{L\in\mathcal{N}_K}g^-_{K,L},w_K\sum_{L\in\mathcal{N}_K}g^-_{K,L}\right)=\max(v_K,w_K)\sum_{L\in\mathcal{N}_K}g^-_{K,L}$, we get

$$\max(v_K,w_K)\sum_{L\in\mathcal{N}_K}g^-_{K,L}\le\sum_{L\in\mathcal{N}_K}g^+_{K,L}\max(v_L,w_L)+F_K,$$

which proves (44), since $\max(v_K,w_K)\le[0,1]$. We now consider the family $(u_K)_{K\in\mathcal{T}}$ defined, for all $K\in\mathcal{T}$, by $u_K=\sup_{v\in\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})}v_K$. For all $n\in\mathbb{N}^\star$, it is possible to find, for all $L\in\mathcal{T}$, an element $(v^L_K)_{K\in\mathcal{T}}\in\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$ such that $u_L\le v^L_L+\frac{1}{n}$. Using (44), we get that the family $(w^{(n)}_K)_{K\in\mathcal{T}}$ defined, for all $K\in\mathcal{T}$ by $w^{(n)}_K=\max_{L\in\mathcal{T}}v^L_K$ is such that $(w^{(n)}_K)_{K\in\mathcal{T}}\in\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$ and $w^{(n)}_K\le u_K\le w^{(n)}_K+\frac{1}{n}$ for all $K\in\mathcal{T}$. Passing to the limit $n\to\infty$ in (43), with $v_K=w^{(n)}_K$ for all $K\in\mathcal{T}$, gives that $(u_K)_{K\in\mathcal{T}}\in\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$. Let us assume that there exists some $K\in\mathcal{T}$ such that $u_K<1$ and $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L-g^-_{K,L}u_K)+F_K>0$. Then there exists $\varepsilon>0$ such that $u_K+\varepsilon<1$ and $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L-g^-_{K,L}(u_K+\varepsilon))+F_K>0$. Let us denote $(\tilde{u}_M)_{M\in\mathcal{T}}$ the family defined by $\tilde{u}_K=u_K+\varepsilon$ and $\tilde{u}_M=u_M$ for all $M\in\mathcal{T}$ such that $M\ne K$. We then have, for all $M\in\mathcal{T}$ such that $M\ne K$,

$$\begin{aligned}\sum_{L\in\mathcal{N}_M}(g^+_{M,L}\tilde{u}_L-g^-_{K,L}\tilde{u}_M)+F_M=\sum_{L\in\mathcal{N}_M}(g^+_{M,L}\tilde{u}_L-g^-_{K,L}u_M)+F_M &\ge\\ \sum_{L\in\mathcal{N}_M}(g^+_{M,L}u_L-g^-_{K,L}u_M)+F_M &\ge 0.\end{aligned}$$

This completes the proof that $(\tilde{u}_M)_{M\in\mathcal{T}}\in\mathcal{C}(g_{\mathcal{T}},F,\mathcal{T})$, which is in contradiction with the definition of $(u_K)_{K\in\mathcal{T}}$. Therefore, for all $K\in\mathcal{T}$, $u_K=1$ or $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L-g^-_{K,L}u_K)+F_K=0$, which completes the proof that $(u_K)_{K\in\mathcal{T}}$ is a solution to System (35)-(36). □

Proposition 4.5 (A monotony property) *Under Hypotheses (H), let $\mathcal{T}$ be an admissible mesh of Ω and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that (32) and (33) are satisfied and let $(F_K)_{K\in\mathcal{T}}$ be defined by (34). Let $(u_K)_{K\in\mathcal{T}}$ be a solution to System (35)-(36).*

Then, for all family of reals $(w_K, s_K)_{K\in\mathcal{T}}$ such that $s_K \geq 0$, for all $K \in \mathcal{T}$, and such that

$$\begin{aligned} \sum_{L\in\mathcal{N}_K} (g^+_{K,L} w_L - g^-_{K,L} w_K) &= -s_K, \quad \text{for all } K \in \mathcal{T} \text{ s.t. } u_K < 1, \\ w_K &= s_K, \quad \text{for all } K \in \mathcal{T} \text{ s.t. } u_K = 1, \end{aligned} \tag{45}$$

the property $w_K \geq 0$, for all $K \in \mathcal{T}$, holds.

Let us first remark that Proposition 4.5 suffices to prove that the matrix of the linear system (45) is invertible, since, in the case $s_K = 0$, for all $K \in \mathcal{T}$, for any family $(w_K)_{K\in\mathcal{T}}$ satisfying (45), then $(-w_K)_{K\in\mathcal{T}}$ also satisfies (45), which proves that $w_K = 0$, for all $K \in \mathcal{T}$. We therefore state the following corollary.

Proposition 4.6 *Under the hypotheses of Proposition 4.5, for all family $(s_K)_{K\in\mathcal{T}}$ of reals, there exists one and only one family of reals $(w_K)_{K\in\mathcal{T}}$ such that (45) holds.*

Proof. of Proposition 4.5. Let us assume the hypotheses of Proposition 4.5, and let $(w_K, s_K)_{K\in\mathcal{T}}$ be a family of reals such that $s_K \geq 0$, for all $K \in \mathcal{T}$, and such that (45) holds. Let us assume that the set $\mathcal{T}_- = \{K \in \mathcal{T};\ w_K < 0\}$ is not empty. Then, if $K \in \mathcal{T}_-$, one has $u_K < 1$, since $w_K = s_K \geq 0$ for $K \in \mathcal{T}$ such that $u_K = 1$. We therefore have

$$\sum_{L\in\mathcal{N}_K} (g^+_{K,L} w_L - g^-_{K,L} w_K) + s_K = 0,\ \forall K \in \mathcal{T}_-. \tag{46}$$

Summing (46) for $K \in \mathcal{T}_-$ leads to

$$\sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K\setminus\mathcal{T}_-} (g^+_{K,L} w_L - g^-_{K,L} w_K) + \sum_{K\in\mathcal{T}_-} s_K = 0. \tag{47}$$

Since $w_K < 0$ for $K \in \mathcal{T}_-$ and $w_L \geq 0$ for $L \notin \mathcal{T}_-$, (47) gives $s_K = 0$ for all $K \in \mathcal{T}_-$ and

$$\forall K \in \mathcal{T}_-,\ \forall L \in \mathcal{N}_K \setminus \mathcal{T}_-,\ g^-_{K,L} = 0. \tag{48}$$

Since, for all $K \in \mathcal{T}_-$, we have $u_K < 1$, we therefore have

$$\sum_{L\in\mathcal{N}_K} (g^+_{K,L} u_L - g^-_{K,L} u_K) + F_K = 0,\ \forall K \in \mathcal{T}_-.$$

Summing the above equation for $K \in \mathcal{T}_-$ leads to

$$\sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K\setminus\mathcal{T}_-} (g^+_{K,L} u_L - g^-_{K,L} u_K) + \sum_{K\in\mathcal{T}_-} F_K = 0,$$

and, using (48), we get

$$\sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K\setminus\mathcal{T}_-} g^+_{K,L} u_L + \sum_{K\in\mathcal{T}_-} F_K = 0.$$

which is impossible, since $u_L > 0$ for all $L \in \mathcal{N}_K \setminus \mathcal{T}_-$ and $F_K > 0$. This contradiction proves that $\mathcal{T}_-$ is empty, which concludes the proof of the proposition. □

Proposition 4.7 (Property of the discrete solution) *Under Hypotheses (H), let $\mathcal{T}$ be an admissible mesh of Ω and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that (32) and (33) are satisfied and let $(F_K)_{K\in\mathcal{T}}$ be defined by (34). Then any solution $(u_K)_{K\in\mathcal{T}}$ to System (35)-(36) is such that, for all $(v_K)_{K\in\mathcal{T}} \in \mathcal{C}(g_\mathcal{T}, F, \mathcal{T})$, then $0 \leq v_K \leq u_K$ for all $K \in \mathcal{T}$. Since $0 \leq u_K$ for all $K \in \mathcal{T}$, (35)-(36) imply that $(u_K)_{K\in\mathcal{T}} \in \mathcal{C}(g_\mathcal{T}, F, \mathcal{T})$, and therefore, as an immediate consequence, there is one and only one solution $(u_K)_{K\in\mathcal{T}}$ to System (35)-(36).*

Remark 4.8 *The above proposition easily gives that the solution $(u_K)_{K\in\mathcal{T}}$ to System (35)-(36) is the projection in $L^2(\Omega)$ of the function 1_Ω on $\mathcal{C}(g_\mathcal{T}, F, \mathcal{T})$.*

Proof. Let $(u_K)_{K\in\mathcal{T}}$ be a solution to System (35)-(36) and let $(v_K)_{K\in\mathcal{T}} \in \mathcal{C}(g_\mathcal{T}, F, \mathcal{T})$. We get from (35)-(36)

$$\sum_{L\in\mathcal{N}_K} (g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K = 0,\ \forall K \in \mathcal{T} \text{ s.t. } u_K < 1. \tag{49}$$

On the other hand, since $(v_K)_{K\in\mathcal{T}} \in \mathcal{C}(g_\mathcal{T}, F, \mathcal{T})$, we have

$$\sum_{L\in\mathcal{N}_K} (g^+_{K,L}v_L - g^-_{K,L}v_K) + F_K \geq 0,\ \forall K \in \mathcal{T}. \tag{50}$$

Subtracting (50) to (49) gives, setting $w_K = u_K - v_K$ for all $K \in \mathcal{T}$,

$$\sum_{L\in\mathcal{N}_K} (g^+_{K,L}w_L - g^-_{K,L}w_K) \leq 0,\ \forall K \in \mathcal{T} \text{ s.t. } u_K < 1,$$

and

$$w_K \geq 0,\ \forall K \in \mathcal{T} \text{ s.t. } u_K = 1.$$

We can therefore apply Proposition 4.5, which proves that $w_K \geq 0$ for all $K \in \mathcal{T}$, which concludes the proof of the proposition. □

The following property is proven in [9].

Proposition 4.9 (Weak bounded variation inequality) *Under Hypotheses (H), let $\mathcal{T}$ be an admissible mesh of Ω and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that (32) and (33) are satisfied and let $(F_K)_{K\in\mathcal{T}}$ be defined by (34). Let $(u_K)_{K\in\mathcal{T}}$ be the solution to System (35)-(36). Then there exists $C > 0$, which only depends on d, Ω, g, F and not on $\mathcal{T}$, such that*

$$\sum_{(K,L)\in\mathcal{E}} |g_{K,L}|(u_K - u_L)^2 \leq C. \tag{51}$$

The two following propositions concern the Lagrange multipliers.

Proposition 4.10 (Existence of the discrete Lagrange multipliers)
Under Hypotheses (H), let $\mathcal{T}$ be an admissible mesh of Ω and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that (32) and (33) are satisfied and let $(F_K)_{K\in\mathcal{T}}$ be defined by (34). Let $(u_K)_{K\in\mathcal{T}}$ be the solution to System (35)-(36). Then there exists $(\mu_K)_{K\in\mathcal{T}}$ and $(\nu_K)_{K\in\mathcal{T}}$ such that:

$$\sum_{K\in\mathcal{T}} \Big(m_K(u_K - 1)v_K - \mu_K\big(\sum_{L\in\mathcal{N}_K} (g^+_{K,L}v_L - g^-_{K,L}v_K)\big) + \nu_K v_K = 0 \Big),\ \forall (v_K)_{K\in\mathcal{T}} \in \mathbb{R}^\mathcal{T}, \tag{52}$$

$$\mu_K \geq 0 \text{ and } \mu_K\big(\sum_{L\in\mathcal{N}_K} (g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K\big) = 0,\ \forall K \in \mathcal{T}, \tag{53}$$

$$\nu_K \geq 0 \text{ and } \nu_K(1 - u_K) = 0,\ \forall K \in \mathcal{T}. \tag{54}$$

Furthermore, one has:

$$\begin{aligned}\sum_{K\in\mathcal{T}} \Big(m_K(u_K - 1)(u_K - \varphi_K) + \mu_K\big(\sum_{L\in\mathcal{N}_K} (g^+_{K,L}\varphi_L - g^-_{K,L}\varphi_K) + F_K\big) \\ + \nu_K(1 - \varphi_K) = 0 \Big),\ \forall (\varphi_K)_{K\in\mathcal{T}} \in \mathbb{R}^\mathcal{T}.\end{aligned} \tag{55}$$

One defines the functions $\mu_\mathcal{T}$ and $\nu_\mathcal{T}$ in $L^\infty(\Omega)$ by:

$$\mu_\mathcal{T} = \mu_K \text{ and } \nu_\mathcal{T} = \frac{\nu_K}{m_K} \text{ a.e. on } K, \text{ for all } K \in \mathcal{T}. \tag{56}$$

Proof.
Since $(u_K)_{K\in\mathcal{T}}$ is the solution to System (35)-(36), it is also the projection in $L^2(\Omega)$ of the function 1_Ω on $\mathcal{C}(g_\mathcal{T}, F, \mathcal{T})$ (see Remark 4.8). Then, $(u_K)_{K\in\mathcal{T}}$ is the solution of the following problem:

$$\begin{aligned}&u = (u_K)_{K\in\mathcal{T}} \in \mathcal{C}(g_\mathcal{T}, F, \mathcal{T}),\\ &J(u) \leq J(v),\ \ \forall v = (v_K)_{K\in\mathcal{T}} \in \mathcal{C}(g_\mathcal{T}, F, \mathcal{T}),\end{aligned} \tag{57}$$

with $J(v) = \sum_{K\in\mathcal{T}} m_K(v_K - 1)^2$ for $v = (v_K)_{K\in\mathcal{T}}$.

Problem (57) is the minimization, in a finite dimension space, of the differentiable function J under affine constraints, which is a classical case of the Kuhn-Tucker theorem. Since the constraint $u \geq 0$ is not active (we already know that $u_K > 0$ for all $K \in \mathcal{T}$, see Proposition 4.2), the Kuhn-Tucker theorem gives the existence of $(\mu_K)_{K\in\mathcal{T}}$ and $(\nu_K)_{K\in\mathcal{T}}$ satisfying (52)-(54).

In order to obtain (55), for $(\varphi_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, one takes, in (52), $v = (v_K)_{K\in\mathcal{T}}$, with $v_K = u_K - \varphi_K$ for all $K \in \mathcal{T}$. Using (53) and (54) leads to (55).
□

Proposition 4.11 (Estimates on the discrete Lagrange multipliers)
Under Hypotheses (H), let $\mathcal{T}$ be an admissible mesh of Ω and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real values such that (32) and (33) are satisfied and let $(F_K)_{K\in\mathcal{T}}$ be defined by (34). Let $(u_K)_{K\in\mathcal{T}}$ be the solution to System (35)-(36) and let $\mu_\mathcal{T}$ and $\nu_\mathcal{T}$ satisfying (52)-(54), (55) and (56). Then

$$\|\mu_\mathcal{T}\|_{L^2(\Omega)} = \Big(\sum m_K \mu_K^2\Big)^{\frac{1}{2}} \leq \frac{(2\,\mathrm{meas}(\Omega))^{1/2}}{F_0}, \tag{58}$$

$$\|\nu_\mathcal{T}\|_{L^1(\Omega)} = \sum_{K\in\mathcal{T}} \nu_K \leq \mathrm{meas}(\Omega) \tag{59}$$

and there exists C_3, only depending on Ω, g and F_0, such that and

$$\sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}_K} g_{K,L}^-(\mu_K - \mu_L)^2 = \sum_{(K,L)\in\mathcal{E}} |g_{K,L}|(\mu_K - \mu_L)^2 \leq C_3. \tag{60}$$

Remark 4.12 *The proof of this proposition uses in particular Assumption (H3).*

Proof.
We first take $\varphi_K = 0$, for all $K \in \mathcal{T}$, in (55). Since $0 \leq u_K \leq 1$, $\mu_K \geq 0$ and $\nu_K \geq 0$ for all $K \in \mathcal{T}$, this gives (59) and

$$F_0 \sum_{K\in\mathcal{T}} m_K \mu_K \leq \sum_{K\in\mathcal{T}} \mu_K F_K \leq \mathrm{meas}(\Omega). \tag{61}$$

Inequality (61) gives an L^1-estimate on $\mu_\mathcal{T}$. In order to obtain (58) (which is an L^2-estimate on $\mu_\mathcal{T}$), we use (53) which gives $\sum_{K\in\mathcal{T}} \mu_K^2 (\sum_{L\in\mathcal{N}_K} (g_{K,L}^+ u_L - g_{K,L}^- u_K) + F_K) = 0$ and then:

$$F_0 \sum_{K\in\mathcal{T}} m_K \mu_K^2 \leq \sum_{K\in\mathcal{T}} F_K \mu_K^2 = -\sum_{K\in\mathcal{T}} \mu_K^2 \sum_{L\in\mathcal{N}_K} (g_{K,L}^+ u_L - g_{K,L}^- u_K). \tag{62}$$

Changing the order of summation in (62), this inequality reads:

$$F_0 \sum_{K\in\mathcal{T}} m_K \mu_K^2 \leq \sum_{K\in\mathcal{T}} F_K \mu_K^2 = \sum_{K\in\mathcal{T}} u_K \sum_{L\in\mathcal{N}_K} g_{K,L}^- (\mu_K^2 - \mu_L^2). \tag{63}$$

We take now $v_K = u_K \mu_K$ for all $K \in \mathcal{T}$ in (52). It gives, using $\nu_K \mu_K u_K \geq 0$ for all $K \in \mathcal{T}$:

$$-\sum_{K\in\mathcal{T}} \mu_K \sum_{L\in\mathcal{N}_K} (g_{K,L}^+ u_L \mu_L - g_{K,L}^- u_K \mu_K) \leq \sum_{K\in\mathcal{T}} m_K (1 - u_K) u_K \mu_K. \tag{64}$$

Changing, here also, the order of summation in (64), and using (61), this inequality leads to:

$$\sum_{K\in\mathcal{T}} u_K \sum_{L\in\mathcal{N}_K} g_{K,L}^- \frac{(\mu_K - \mu_L)^2}{2} + \sum_{K\in\mathcal{T}} u_K \sum_{L\in\mathcal{N}_K} g_{K,L}^- \frac{(\mu_K^2 - \mu_L^2)}{2} \leq \frac{\mathrm{meas}(\Omega)}{F_0},$$

and then, with (63):

$$\sum_{K\in\mathcal{T}} u_K \sum_{L\in\mathcal{N}_K} g_{K,L}^- \frac{(\mu_K - \mu_L)^2}{2} + \frac{1}{2} \sum_{K\in\mathcal{T}} F_K \mu_K^2 \leq \frac{\mathrm{meas}(\Omega)}{F_0}. \tag{65}$$

Inequality (65) gives, in particular, (58).

It remains to prove (60). To obtain this bound, we first remark that Inequality (65) gives:

$$\sum_{K\in\mathcal{T}} u_K \sum_{L\in\mathcal{N}_K} g_{K,L}^- \frac{(\mu_K-\mu_L)^2}{2} \le \frac{\mathrm{meas}(\Omega)}{F_0}. \tag{66}$$

Since $(u_K)_{K\in\mathcal{T}}$ is the maximal element of $\mathcal{C}(g_\mathcal{T}, F, \mathcal{T})$, we can easily find a strictly positive lower bound for $(u_K)_{K\in\mathcal{T}}$, denoted by $\underline{u}$, which only depends on g and F_0 and thus provides (60) with $C_3 = 2\,\mathrm{meas}(\Omega)/\underline{u}F_0$. Indeed, if $\|\mathrm{div}g\|_\infty = 0$, then $u_K = \underline{u}$ holds for all $K \in \mathcal{T}$, with $\underline{u} = 1$. Otherwise, setting $\underline{u} = \min(1, F_0/\|\mathrm{div}g\|_\infty)$, we get that the constant family $(\underline{u})_{K\in\mathcal{T}}$ satisfies $(\underline{u})_{K\in\mathcal{T}} \in \mathcal{C}(g_\mathcal{T}, F, \mathcal{T})$, from which one deduces $u_K \ge \underline{u}$ for all $K \in \mathcal{T}$.
□

We can now state the convergence of the scheme to the solution.

Proposition 4.13 (Convergence of the scheme to a process solution)
Under hypotheses (H), let $(\mathcal{T}^{(m)}, g_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ be a sequence such that, for all $m \in \mathbb{N}$, $\mathcal{T}^{(m)}$ is an admissible mesh of Ω in the sense of Definition 4.1, and $g_{\mathcal{T}^{(m)}}$ is a family of reals such that (32)-(33) are satisfied. We assume that $\lim_{m\to\infty} \mathrm{size}(\mathcal{T}^{(m)}) = 0$, that there exists $R > 0$ s.t $\mathrm{regul}(\mathcal{T}^{(m)}) \le R$ for all $m \in \mathbb{N}$, and that $\lim_{m\to\infty} cons(g_{\mathcal{T}^{(m)}}) = 0$. For all $m \in \mathbb{N}$, we denote by $u_{\mathcal{T}^{(m)}}$ and $(\mu_{\mathcal{T}^{(m)}}, \nu_{\mathcal{T}^{(m)}})$ the respective solutions to System (35)-(36) and to (52)-(56) for $\mathcal{T} = \mathcal{T}^{(m)}$ and $g_\mathcal{T} = g_{\mathcal{T}^{(m)}}$. Then, from the sequence $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$, one can extract a subsequence, again denoted $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$, such that the corresponding sequences $(u_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ and $(\mu_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ are such that

1. *$(u_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ converges in $L^\infty(\Omega)$ for the nonlinear weak-$\star$ sense to some function u with $0 \le u(x,\alpha) \le 1$ for a.e. $x \in \Omega$ and a.e. $\alpha \in (0,1)$ (see [11] or [12]),*
2. *$(\mu_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ weakly converges in $L^2(\Omega)$ to some function μ with $0 \le \mu(x)$ for a.e. $x \in \Omega$,*
3. *the pair (u,μ) is such that (29) holds,*
4. *u is such that*

$$\int_\Omega \int_0^1 \Big(\xi(u(x,\alpha))(-g(x)\cdot\nabla\varphi(x)) + [\xi'(u(x,\alpha))u(x,\alpha) - \xi(u(x,\alpha))]\varphi(x)\mathrm{div}g(x) + \xi'(u(x,\alpha))\varphi(x)F(x)\Big)\mathrm{d}\alpha\mathrm{d}x \ge 0 \quad \forall \xi \in C^1(\mathbb{R}),\ \text{convex s.t. } \xi'(1) \ge 0,\ \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+). \tag{67}$$

and therefore $\bar{u} : x \mapsto \int_0^1 u(x,\alpha)\mathrm{d}\alpha$ is such that $\bar{u} \in \mathcal{C}(g,F)$.

Thanks to the uniqueness theorem 3.2, we therefore deduce that all the sequence $(u_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ converges in $L^p(\Omega)$ for all $p \in [1,+\infty)$ to the maximal element of $\mathcal{C}(g,F)$ as $m \to \infty$, which is, thanks to (67) and to Proposition 2.5, a weak solution of Problem (5)-(6) in the sense of Definition 2.1.

Proof. Using the property (35) satisfied by $u_{\mathcal{T}^{(m)}}$, we can deduce the existence of a subsequence, again denoted $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$, such that the corresponding sequence $(u_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ converges in the nonlinear weak-$\star$ sense to some function $u \in L^\infty(\Omega\times(0,1))$, $(\mu_\mathcal{T})_{m\in\mathbb{N}}$ weakly converges to μ in $L^2(\Omega)$. Let $\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+)$ with $0 \le \varphi(x) \le 1$ for all $x \in \Omega$. Let $m \in \mathbb{N}$, and let $(\mathcal{T}^{(m)})$ the corresponding admissible mesh of the subsequence. For the simplicity of the notation, we do not mention the index m until we consider some convergence properties as $m \to \infty$. We take $\varphi_K = \frac{1}{m_K}\int_K \varphi(x)\mathrm{d}x$ in (55). We get, thanks to the positivity of $\nu_\mathcal{T}$, $T_{17} - T_{18} - T_{19} \ge 0$, with

$$\begin{aligned} T_{17} &= \sum_{K\in\mathcal{T}} m_K(1-u_K)(u_K-\varphi_K) \\ T_{18} &= \sum_{K\in\mathcal{T}} \mu_K \left(\sum_{L\in\mathcal{N}_K} (g_{K,L}^+\varphi_L - g_{K,L}^-\varphi_K)\right) \\ T_{19} &= \sum_{K\in\mathcal{T}} \mu_K F_K. \end{aligned}$$

We get, from the nonlinear weak convergence of $(u_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$, that

$$\limsup_{m\to\infty} T_{17}^{(m)} = \int_\Omega\int_0^1 (1-u(x,\alpha))(u(x,\alpha)-\varphi(x))\mathrm{d}\alpha\mathrm{d}x.$$

We remark that, replacing g by $-g$, we can apply Proposition 5.3 to T_{18}, thanks to (60). We then get

$$\lim_{m\to\infty} T_{18}^{(m)} = \int_\Omega \mu(x)\mathrm{div}(\varphi(x)g(x))\mathrm{d}x.$$

Since we easily obtain that

$$\lim_{m\to\infty} T_{19}^{(m)} = \int_\Omega \mu(x)F(x)\mathrm{d}x,$$

we get that (29) is satisfied.

Let us now prove (67). Let $\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+)$ be given and let $\xi \in C^1(\mathbb{R})$ be a convex function such that $\xi'(1) \geq 0$.
We get from (36), using $\xi'(u_K) = \xi'(1) + \xi'(u_K) - \xi'(1)$, that

$$\xi'(u_K)\left(\sum_{L\in\mathcal{N}_K}(g_{K,L}^+ u_L - g_{K,L}^- u_K) + F_K\right) \geq 0,\ \forall K \in \mathcal{T}. \tag{68}$$

We can then multiply (68) by φ_K, where we denote by $\varphi_K = \frac{1}{m_K}\int_K \varphi(x)\mathrm{d}x$, and we sum on $K \in \mathcal{T}$. We get $T_{20}^{(m)} + T_{21}^{(m)} + T_{22}^{(m)} \geq 0$, with

$$T_{20}^{(m)} = \sum_{K\in\mathcal{T}} \xi'(u_K)u_K\varphi_K \sum_{L\in\mathcal{N}_K} g_{K,L},$$

$$T_{21}^{(m)} = \sum_{K\in\mathcal{T}} \xi'(u_K)\varphi_K \sum_{L\in\mathcal{N}_K} g_{K,L}^+(u_L - u_K),$$

and

$$T_{22}^{(m)} = \sum_{K\in\mathcal{T}} \xi'(u_K)\varphi_K F_K.$$

Since $\sum_{L\in\mathcal{N}_K} g_{K,L} = \int_K \mathrm{div}g(x)\mathrm{d}x$, we thus get that

$$\lim_{m\to\infty} T_{20}^{(m)} = \int_\Omega \int_0^1 \xi'(u(x,\alpha))u(x,\alpha)\varphi(x)\mathrm{div}g(x)\mathrm{d}\alpha\mathrm{d}x.$$

On the other hand, thanks to the convexity of ξ, we have

$$T_{21}^{(m)} \leq T_{23}^{(m)} := \sum_{K\in\mathcal{T}} \varphi_K \sum_{L\in\mathcal{N}_K} g_{K,L}^+(\xi(u_L) - \xi(u_K)).$$

Since $|\xi(u_L) - \xi(u_K)| \leq |u_L - u_K| \max_{s\in[0,1]} |\xi'(s)|$, thanks to Proposition 4.9, we can apply Proposition 5.3 to T_{23}. This shows that

$$\lim_{m\to\infty} T_{23}^{(m)} = -\int_\Omega \int_0^1 \xi(u(x),\alpha)\mathrm{div}(\varphi(x)g(x))\mathrm{d}\alpha\mathrm{d}x.$$

Finally, we easily get

$$\lim_{m\to\infty} T_{22}^{(m)} = \int_\Omega \int_0^1 \xi'(u(x,\alpha))\varphi(x)F(x)\mathrm{d}\alpha\mathrm{d}x.$$

We then get (67), letting $m \to \infty$ in $T_{20}^{(m)} + T_{23}^{(m)} + T_{22}^{(m)} \geq 0$. Letting $\xi(s) = s$ in (67) proves that $\bar{u}(x) = \int_0^1 u(x,\alpha)\mathrm{d}\alpha$ is in $\mathcal{C}(g,F)$. We can then apply the uniqueness result Proposition 3.2. We thus classically get that the convergence is strong, and therefore we get that (67) gives (12). This concludes the proof of Proposition 4.13 and completes the proof of Proposition 2.11.
□
We can now state the concluding result.

Theorem 4.14 (Convergence of the scheme to the unique weak solution of the problem) *Let $R > 0$. Under hypotheses (H), for an admissible mesh $\mathcal{T}$ of Ω, in the sense of Definition 4.1, and for $g_\mathcal{T}$ satisfying (32)-(33), let $u_\mathcal{T}$ be the unique solution to System (35)-(36). Let u be the unique weak solution of Problem (5)-(6) in the sense of Definition 2.1. Then, $u_\mathcal{T} \longrightarrow u$ in $L^p(\Omega)$, for all $p \in [1,\infty[$, as $\mathrm{size}(\mathcal{T}) \longrightarrow 0$ and $cons(g_\mathcal{T}) \longrightarrow 0$, with $\mathrm{regul}(\mathcal{T}) \leq R$.*
Furthermore, there exist $\mu \in L^2(\Omega)$ and a finite nonnegative measure ν, such that (26) and (28) hold.

As remarked above, this theorem allows to prove Propositions 2.11 and 2.13.

5 Conclusion

The strong convergence of the scheme has been practically observed (see [9]). However, much work now remains to be completed. In particular, the regularity which is necessary for the function g cannot be easily expected in the coupled problem given in the introduction of this paper. Different ways can be chosen for solving this problem: one can directly study the time dependent problem and its approximation (see [2] for some attempts in direction of the resolution of the continuous problem), or one can look for an extension of the results given here, assuming less regularity for the function g.

References

[1] R.S. Anderson and N.F. Humphrey. Interaction of weathering and transport processes in the evolution of arid landscapes. *Quantitative Dynamics Stratigraphy, T.A. Cross ed.*, pages 349–361, 1989.

[2] S.N. Antontsev, G. Gagneux, and G. Vallet. On some stratigraphic control problems. *J. of Appl. Mech. and Tech. Phy.*, 44(6):821–828, 2003.

[3] R. Burger, C. Liu, and W.L. Wendland. Existence and stability for mathematical models of sedimentation-consolidation processes in several space dimensions. *J. Math. Anal. Appl.*, 264:288–310, 2001.

[4] R. DiPerna. Measure-valued solutions to conservation laws. *Arch. Rat. Mech. Anal.*, 88:223–270, 1985.

[5] J. Droniou. Solving convection-diffusion equations with mixed, neumann and fourier boundary conditions and measures as data, by a duality method. *Adv. Differential Equations*, 5(10-12):1341–1396, 2000.

[6] J. Droniou, R. Eymard, D. Hilhorst, and X. D. Zhou. Convergence of a finite volume - mixed finite element method for a system of a hyperbolic and an elliptic equations. *IMA Journal of Numerical Analysis*, 23:07–538, 2003.

[7] J. Droniou and T. Gallouët. Finite volume methods for right-hand side in H^{-1}. *Math. Mod. Anal. Num.*, 4:705–724, 2002.

[8] J. Droniou, T. Gallouët, and R. Herbin. A finite volume scheme for noncoercive elliptic equation with measure data. *SIAM J. Numer. Anal.*, 41(6):1997–2031, 2003.

[9] R. Eymard and T. Gallouët. Analytical and numerical study of a hyperbolic inequality arising in a model of erosion and sedimentation. *accepted for publication in SIAM J. on Num. Anal.*, 2005.

[10] R. Eymard, T. Gallouët, D. Granjeon, R. Masson, and Q.H. Tran. Multi-lithology stratigraphic model under maximum erosion rate constraint. *Internat. J. Numer. Methods Engrg.*, 60(2):527–548, 2004.

[11] R. Eymard, T. Gallouët, and R. Herbin. Existence and uniqueness of the entropy solution to a nonlinear hyperbolic equation. *Chi. Ann. of Math*, 16(B1):1–14, 1995.

[12] R. Eymard, T. Gallouët, and R. Herbin. The finite volume method. *Handbook of Numerical Analysis, Ph. Ciarlet J.L. Lions eds*, 7:715–1022, 2000.

[13] R. Eymard, T. Gallouët, and R. Herbin. A finite volume scheme for anisotropic diffusion problems. *Accepted for publication in Comptes Rendus de l'Académie des Sciences*, 2004.

[14] D. Granjeon, P. Joseph, and B. Doligez. Using a 3-d stratigraphic model to optimize reservoir description. *Hart's Petroleum Engineer International*, pages 51–58.

[15] S.N. Krushkov. First order quasilinear equations with several space variables. *Math. USSR. Sb.*, 10:217–243, 1970.

[16] L. Lévi. The singular limit of a bilateral obstacle problem for a class of degenerate parabolic-hyperbolic operators. *Adv. in Appl. Math.*, 35:34–57, 2005.

[17] L. Lévi, E. Rouvre, and G. Vallet. Weak entropy solutions for degenerate parabolic-hyperbolic inequalities. *Appl. Math. Letters*, 18:497–504, 2005.

[18] F. Mignot and J.P. Puel. Inéquations variationnelles et quasivariationnelles hyperboliques du premier ordre. *J. Math. pures et appl.*, 55:353–378, 1976.

[19] J.C. Rivenaes. Impact of sediment transport efficiency on large scale sequence architecture: results from stratigraphic computer simulation. *Basin Research*, 4:133–146, 1992.

[20] M.H. Vignal. Convergence of a finite volume scheme for a system of an elliptic equation and a hyperbolic equation. *Modél. Math. Anal. Numér.*, 30(7):841–872, 1996.

Appendix: technical results

The following result, extracted from [9], is based on Krushkov's doubling variable technique [15]. The proof is given for the sake of completeness.

Proposition 5.1 (Doubling variable result)
Under hypotheses (H), let us assume that there exist $u, v \in L^\infty(\Omega)$ *with* $0 \leq u(x) \leq 1$ *and* $0 \leq v(x) \leq 1$ *for a.e.* $x \in \Omega$ *and* $\eta \in C^1(\mathbb{R}^2, \mathbb{R})$ *such that:*

$$\begin{aligned}\int_\Omega \Big(\eta(u(x),b)(-g(x)\cdot\nabla\varphi(x)) + [\partial_1\eta(u(x),b)u(x) - \eta(u(x),b)]\varphi(x)\mathrm{div}g(x) + \\ \partial_1\eta(u(x),b)\varphi(x)F(x)\Big)\mathrm{d}x \geq 0, \\ \forall b \in [0,1],\ \forall\varphi \in C^1(\overline{\Omega},\mathbb{R}_+),\end{aligned} \tag{69}$$

and

$$\begin{aligned}\int_\Omega \Big(\eta(a,v(y))(-g(y)\cdot\nabla\varphi(y)) + [\partial_2\eta(a,v(y))v(y) - \eta(a,u(y))]\varphi(y)\mathrm{div}g(y) + \\ \partial_2\eta(a,v(y))\varphi(y)F(y)\Big)\mathrm{d}y \geq 0, \\ \forall a \in [0,1],\ \forall\varphi \in C^1(\overline{\Omega},\mathbb{R}_+).\end{aligned} \tag{70}$$

Then the following inequality holds:

$$\begin{aligned}&\int_\Omega \Big(\eta(u(x),v(x))\,[-g(x)\cdot\nabla\phi(x)] + \\ &\Big(\partial_1\eta(u(x),v(x))u(x) + \partial_2\eta(u(x),v(x))v(x) - \eta(u(x),v(x))\Big)\phi(x)\mathrm{div}g(x) + \\ &\qquad\Big(\partial_1\eta(u(x),v(x)) + \partial_2\eta(u(x),v(x))\Big)F(x)\phi(x)\Big)\mathrm{d}x \geq 0, \\ &\forall\varphi \in C^1(\overline{\Omega},\mathbb{R}_+).\end{aligned} \tag{71}$$

Proof. Let us assume the hypotheses of the proposition. Let $\psi \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}_+)$ be given. Then, for all $x \in \Omega$, we have $\psi(x,\cdot) \in C^1(\overline{\Omega},\mathbb{R}_+)$ and for all $y \in \Omega$, $\psi(\cdot,y) \in C^1(\overline{\Omega},\mathbb{R}_+)$. We write (69) with $b = v(y)$ and and $\varphi = \psi(\cdot,y)$, for a.e. $y \in \Omega$, and we integrate the result on Ω. This produces

$$\begin{aligned}&\int_\Omega\int_\Omega \Big(\eta(u(x),v(y))\,[-g(x)\cdot\nabla_x\psi(x,y)] + \\ &[\partial_1\eta(u(x),v(y))u(x) - \eta(u(x),v(y))]\,\psi(x,y)\mathrm{div}g(x) + \\ &\qquad\partial_1\eta(u(x),v(y))\psi(x,y)F(x)\Big)\mathrm{d}x\mathrm{d}y \geq 0.\end{aligned} \tag{72}$$

We now consider (70) with $a = u(x)$ and $\varphi = \psi(x,\cdot)$ for a.e. $x \in \Omega$, and we integrate the result on Ω. We thus get

$$\begin{aligned}&\int_\Omega\int_\Omega \Big(\eta(u(x),v(y))\,[-g(y)\cdot\nabla_y\psi(x,y)] + \\ &[\partial_2\eta(u(x),v(y))v(y) - \eta(u(x),v(y))]\,\psi(x,y)\mathrm{div}g(y) + \\ &\qquad\partial_2\eta(u(x,\alpha),v(y))\psi(x,y)F(y)\Big)\mathrm{d}x\mathrm{d}y \geq 0.\end{aligned} \tag{73}$$

We now add (72) and (73). Defining T_{24}, T_{25} and T_{26} by

$$T_{24} = -\int_\Omega\int_\Omega\int_0^1 \eta(u(x),v(y))\Big(g(x)\cdot\nabla_x\psi(x,y) + g(y)\cdot\nabla_y\psi(x,y)\Big)\mathrm{d}x\mathrm{d}y, \tag{74}$$

$$T_{25} = \int_\Omega\int_\Omega \Big(\Big(\partial_1\eta(u(x),v(y))u(x) - \eta(u(x),v(y))\Big)\psi(x,y)\mathrm{div}g(x) + \Big(\partial_2\eta(u(x),v(y))v(y) - \eta(u(x),v(y))\Big)\psi(x,y)\mathrm{div}g(y)\Big)\mathrm{d}x\mathrm{d}y \tag{75}$$

and

$$T_{26} = \int_\Omega\int_\Omega \Big(\partial_1\eta(u(x),v(y))F(x) + \partial_2\eta(u(x),v(y))F(y)\Big)\psi(x,y)\mathrm{d}x\mathrm{d}y. \tag{76}$$

we get

$$T_{24} + T_{25} + T_{26} \geq 0. \tag{77}$$

We again use the sequence of mollifiers in $\mathbb{R}$ and $\mathbb{R}^d$, defined by (21)-(23). Let $\phi \in C^1(\mathbb{R}^d, \mathbb{R}_+)$ and $n \in \mathbb{N}^\star$ be given. We then take $\psi(x,y) = \phi(x)\rho_n(x-y)$ in (72) and (73), which gives $\psi \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}_+)$. We thus get, from (77):

$$T_{24}^{(n)} + T_{25}^{(n)} + T_{26}^{(n)} \geq 0, \tag{78}$$

with

$$T_{24}^{(n)} = -\int_\Omega\int_\Omega \eta(u(x),v(y))\Big(\rho_n(x-y)g(x)\cdot\nabla\phi(x) + \phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\Big)\mathrm{d}x\mathrm{d}y, \tag{79}$$

$$T_{25}^{(n)} = \int_\Omega\int_\Omega \Big(\left[\partial_1\eta(u(x),v(y))u(x) - \eta(u(x),v(y))\right]\mathrm{div}g(x) + \left[\partial_2\eta(u(x),v(y))v(y) - \eta(u(x),v(y))\right]\mathrm{div}g(y)\Big)\phi(x)\rho_n(x-y)\mathrm{d}x\mathrm{d}y, \tag{80}$$

$$T_{26}^{(n)} = \int_\Omega\int_\Omega \Big(\partial_1\eta(u(x),v(y))F(x) + \partial_2\eta(u(x),v(y))F(y)\Big)\phi(x)\rho_n(x-y)\mathrm{d}x\mathrm{d}y. \tag{81}$$

We have $T_{24}^{(n)} = T_{27}^{(n)} + T_{28}^{(n)} + T_{29}^{(n)}$, with

$$T_{27}^{(n)} = -\int_\Omega\int_\Omega \eta(u(x),v(y))\rho_n(x-y)g(x)\cdot\nabla\phi(x)\mathrm{d}x\mathrm{d}y, \tag{82}$$

$$T_{28}^{(n)} = -\int_\Omega\int_\Omega \eta(u(x),v(x))\phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\mathrm{d}x\mathrm{d}y, \tag{83}$$

$$T_{29}^{(n)} = -\int_\Omega\int_\Omega \Big(\eta(u(x),v(y)) - \eta(u(x),v(x))\Big)\phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\mathrm{d}x\mathrm{d}y. \tag{84}$$

The limit of $T_{27}^{(n)}$ as $n \longrightarrow \infty$ is given by

$$\lim_{n\to\infty} T_{27}^{(n)} = -\int_\Omega \eta(u(x),v(x))g(x)\cdot\nabla\phi(x)\mathrm{d}x.$$

Thanks to an integration by parts with respect to y and to Hypotheses (H), we get $T_{28}^{(n)} = T_{30}^{(n)} + T_{31}^{(n)}$ where

$$T_{30}^{(n)} = \int_\Omega\int_{\partial\Omega} \eta(u(x),v(x))\phi(x)\rho_n(x-y)g(x)\cdot\mathbf{n}(y)\mathrm{d}y\mathrm{d}x, \tag{85}$$

and

$$T_{31}^{(n)} = \int_\Omega\int_\Omega \eta(u(x),v(x))\phi(x)\rho_n(x-y)\mathrm{div}g(y)\mathrm{d}x\mathrm{d}y. \tag{86}$$

We have, for a.e. $y \in \partial\Omega$,

$$\lim_{n\to\infty}\int_\Omega \eta(u(x),v(x))\phi(x)\rho_n(x-y)g(x)\cdot\mathbf{n}(y)\mathrm{d}x = 0,$$

which produces

$$\lim_{n\to\infty} T_{30}^{(n)} = 0,$$

and therefore

$$\lim_{n\to\infty} T_{28}^{(n)} = \lim_{n\to\infty} T_{31}^{(n)} = \int_\Omega \eta(u(x),v(x))\phi(x)\mathrm{div}g(x)\mathrm{d}x.$$

Thanks to the theorem of continuity in means applied to the function v, thanks to the fact that $(x,y)\mapsto (g(x)-g(y))\cdot\nabla\rho_n(x-y)$ vanishes for $|x-y|>1/n$ and belongs to $L^1(\Omega)$ since g is regular, we get

$$\lim_{n\to\infty} T_{29}^{(n)} = 0.$$

We have, again using the Lebesgue dominated convergence theorem,

$$\lim_{n\to\infty} T_{25}^{(n)} = \int_\Omega \Big(\partial_1\eta(u(x),v(x))u(x) + \partial_2\eta(u(x),v(x))v(x) - 2\eta(u(x),v(x))\Big)\phi(x)\mathrm{div}g(x)\mathrm{d}x$$

and

$$\lim_{n\to\infty} T_{26}^{(n)} = \int_\Omega \left(\partial_1\eta(u(x),v(x)) + \partial_2\eta(u(x),v(x))\right)F(x)\phi(x)\mathrm{d}x.$$

We thus get (71), passing to the limit $n\to\infty$ in (78). □
We have the following technical result.

Proposition 5.2
Under hypotheses (H), let us assume that there exist $w\in L^\infty(\Omega)$ such that:

$$-\int_\Omega w(x)g(x)\cdot\nabla\varphi(x)\mathrm{d}x \geq 0,\ \forall\varphi\in C^1(\overline{\Omega},\mathbb{R}_+). \tag{87}$$

Then

$$\int_\Omega w(x)g(x)\cdot\nabla\varphi(x)\mathrm{d}x = 0,\ \forall\varphi\in C^1(\overline{\Omega},\mathbb{R}), \tag{88}$$

and, defining the function $\mathrm{sign}^+(s)$ by $\mathrm{sign}^+(s)=1$ for all $s>0$ and $\mathrm{sign}^+(s)=0$ for all $s\leq 0$,

$$\int_\Omega \mathrm{sign}^+(w(x))\mathrm{div}(g(x)\varphi(x))\mathrm{d}x = 0,\ \forall\varphi\in C^1(\overline{\Omega},\mathbb{R}). \tag{89}$$

Proof. We first remark that, for any $\varphi\in C^1(\overline{\Omega},\mathbb{R}_+)$, then the function $\psi = \|\varphi\|_{L^\infty(\Omega)} - \varphi$ is such that $\psi\in C^1(\overline{\Omega},\mathbb{R}_+)$. Inequality (87) applied to ψ provides $\int_\Omega w(x)g(x)\cdot\nabla\psi(x)\mathrm{d}x = -\int_\Omega w(x)g(x)\cdot\nabla\varphi(x)\mathrm{d}x \geq 0$, which gives (88) for all $\varphi\in C^1(\overline{\Omega},\mathbb{R}_+)$. For all $\varphi\in C^1(\overline{\Omega},\mathbb{R})$, it suffices to consider (88) written with the function $\psi = \varphi - \min_{x\in\overline{\Omega}}\varphi(x)$, which is such that $\psi\in C^1(\overline{\Omega},\mathbb{R}_+)$. We now prove the following relation: for all $f\in C^1(\mathbb{R},\mathbb{R})$ such that f' is Lipschitz continuous,

$$\int_\Omega f(w(x))\mathrm{div}(g(x)\varphi(x))\mathrm{d}x = \int_\Omega f'(w(x))w(x)\varphi(x)\mathrm{div}g(x)\mathrm{d}x,\ \forall\varphi\in C^1(\overline{\Omega},\mathbb{R}). \tag{90}$$

In order to prove (90), we again use the sequence of mollifiers in $\mathbb{R}$ and $\mathbb{R}^d$, defined by (21)-(23). Let $\varphi\in C^1(\mathbb{R}^d,\mathbb{R})$ and $n\in\mathbb{N}^\star$ be given. We define the function $w_n(x) = \int_\Omega \rho_n(x-y)w(y)\mathrm{d}y$. We define the term $T_{32}^{(n)}$ by

$$T_{32}^{(n)} = \int_\Omega f(w_n(x))\mathrm{div}(g(x)\varphi(x))\mathrm{d}x.$$

We then have

$$\lim_{n\to\infty} T_{32}^{(n)} = \int_\Omega f(w(x))\mathrm{div}(g(x)\varphi(x))\mathrm{d}x.$$

We then write

$$T_{32}^{(n)} = -\int_\Omega \varphi(x)g(x)\cdot\nabla f(w_n(x))\mathrm{d}x = -\int_\Omega \varphi(x)f'(w_n(x))g(x)\cdot\nabla w_n(x)\mathrm{d}x.$$

Hence we get

$$T_{32}^{(n)} = -\int_\Omega \varphi(x)g(x)\cdot\nabla f(w_n(x))\mathrm{d}x = -\int_\Omega\int_\Omega \varphi(x)f'(w_n(x))g(x)\cdot\nabla\rho_n(x-y)w(y)\mathrm{d}y\mathrm{d}x.$$

We remark that (88) gives

$$\int_\Omega \int_\Omega \varphi(x) f'(w_n(x)) g(y) \cdot \nabla \rho_n(x-y) w(y) \mathrm{d}y \mathrm{d}x = 0,$$

and therefore, defining $T_{33}^{(n)}$, $T_{34}^{(n)}$ and $T_{35}^{(n)}$ by

$$T_{33}^{(n)} = -\int_\Omega \int_\Omega \varphi(y) f'(w_n(y))(g(x)-g(y)) \cdot \nabla \rho_n(x-y) w(y) \mathrm{d}y \mathrm{d}x,$$

$$T_{34}^{(n)} = -\int_\Omega \int_\Omega f'(w_n(x))(\varphi(x)-\varphi(y))(g(x)-g(y)) \cdot \nabla \rho_n(x-y) w(y) \mathrm{d}y \mathrm{d}x,$$

and

$$T_{35}^{(n)} = -\int_\Omega \int_\Omega \varphi(y)(f'(w_n(x))-f'(w_n(y)))(g(x)-g(y)) \cdot \nabla \rho_n(x-y) w(y) \mathrm{d}y \mathrm{d}x,$$

we get $T_{32}^{(n)} = T_{33}^{(n)} + T_{34}^{(n)} + T_{35}^{(n)}$. Thanks to an integrate by parts with respect to x, we get

$$T_{33}^{(n)} = \int_\Omega \int_\Omega \varphi(y) f'(w_n(y)) \rho_n(x-y) w(y) \mathrm{div} g(x) \mathrm{d}y \mathrm{d}x,$$

which proves that

$$\lim_{n\to\infty} T_{33}^{(n)} = \int_\Omega \varphi(x) f'(w(x)) w(x) \mathrm{div} g(x) \mathrm{d}x.$$

Thanks to the facts that $w \in L^\infty(\Omega)$, and $(x,y) \mapsto (g(x)-g(y)) \cdot \nabla \rho_n(x-y)$ vanishes for $|x-y| > 1/n$ and belongs to $L^1(\Omega)$ since g is regular, we get that

$$\lim_{n\to\infty} T_{34}^{(n)} = 0.$$

Let us turn to the study of $T_{35}^{(n)}$. We have, using the fact that f' is Lipschitz continuous with the constant L, and that g is Lipschitz continuous with the constant L_g,

$$|T_{35}^{(n)}| \leq L\ L_g\ \|\varphi\|_{L^\infty(\Omega)} \|w\|_{L^\infty(\Omega)} \int_\Omega \int_\Omega\ |w_n(x)-w_n(y)|\ |x-y|\ |\nabla \rho_n(x-y)| \mathrm{d}y \mathrm{d}x.$$

Prolonging the function w by 0 at the exterior of Ω, we have

$$\begin{aligned} |w_n(x)-w_n(y)| &= \left| \int_{\mathbb{R}^d} \rho_n(x-z) w(z) \mathrm{d}z - \int_{\mathbb{R}^d} \rho_n(y-z) w(z) \mathrm{d}z \right| \\ &\leq \int_{\mathbb{R}^d} \rho_n(z)\, |w(x+z)-w(y+z)|\, \mathrm{d}z, \end{aligned}$$

and therefore we get that $|T_{35}^{(n)}| \leq L\ L_g\ \|\varphi\|_{L^\infty(\Omega)} \|w\|_{L^\infty(\Omega)} T_{36}^{(n)}$, defining $T_{36}^{(n)}$ by

$$\begin{aligned} T_{36}^{(n)} &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \rho_n(z)\, |w(x+z)-w(y+z)|\ \ |x-y|\ |\nabla \rho_n(x-y)| \mathrm{d}y \mathrm{d}x \mathrm{d}z \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \rho_n(z)\, \left|w(x')-w(y')\right|\ \ |x'-y'|\ |\nabla \rho_n(x'-y')| \mathrm{d}y' \mathrm{d}x' \mathrm{d}z \\ &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |w(x)-w(y)|\ \ |x-y|\ |\nabla \rho_n(x-y)| \mathrm{d}y \mathrm{d}x. \end{aligned}$$

This proves that

$$\lim_{n\to\infty} T_{35}^{(n)} = \lim_{n\to\infty} T_{36}^{(n)} = 0,$$

and concludes the proof of (90). We now take in (90) $f = S_\varepsilon$, for $\varepsilon > 0$, where S_ε is defined by (13). We thus get

$$\int_\Omega S_\varepsilon(w(x)) \mathrm{div}(g(x)\varphi(x)) \mathrm{d}x = \int_\Omega S_\varepsilon'(w(x)) w(x) \varphi(x) \mathrm{div} g(x) \mathrm{d}x.$$

Letting $\varepsilon \to 0$ in the above equation provides (89) thanks to the dominated convergence theorem, since $|S_\varepsilon'(a)\ a|$ remains bounded and tends to 0 for all $a \in \mathbb{R}$ as $\varepsilon \to 0$. $\square$

The following result is used twice in the course of the proof of convergence of the scheme.

Proposition 5.3 (A convergence property) *Under hypotheses (H), let $(\mathcal{T}^{(m)}, g_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ be a sequence such that, for all $m \in \mathbb{N}$, $\mathcal{T}^{(m)}$ is an admissible mesh of Ω in the sense of Definition 4.1, and $g_{\mathcal{T}^{(m)}}$ is a family of reals such that (32)-(33) are satisfied. We assume that $\lim_{m\to\infty} \mathrm{size}(\mathcal{T}^{(m)}) = 0$, that there exists $R > 0$ s.t $\mathrm{regul}(\mathcal{T}^{(m)}) \leq R$ for all $m \in \mathbb{N}$, and that $\lim_{m\to\infty} cons(g_{\mathcal{T}^{(m)}}) = 0$. We assume that $(v^{(m)})_{m\in\mathbb{N}}$ is a sequence of functions such that $v^{(m)}$ is, for all $K \in \mathcal{T}^{(m)}$ a constant $v_K^{(m)}$, such that there exists C_4 with $\|v^{(m)}\|_{L^2(\Omega)} \leq C_4$ for all $m \in \mathbb{N}$, that the sequence $(v^{(m)})_{m\in\mathbb{N}}$ converges to $v \in L^\infty(\Omega)$ for the weak topology of $L^2(\Omega)$ and*

$$\sum_{(K,L)\in\mathcal{E}^{(m)}} |g_{K,L}|(v_K^{(m)} - v_L^{(m)})^2 \leq C_4, \ \forall m \in \mathbb{N}. \tag{91}$$

Let $\varphi \in C^1(\overline{\Omega}, \mathbb{R})$ be given. Then the term $T_{37}^{(m)}$, defined for all $m \in \mathbb{N}$ by

$$T_{37}^{(m)} = \sum_{(K,L)\in\mathcal{E}} (v_L - v_K)(\varphi_K g_{K,L}^+ - \varphi_L g_{K,L}^-),$$

where we denote for all $K \in \mathcal{T}^{(m)}$ by $\varphi_K = \frac{1}{m_K}\int_K \varphi(x)\mathrm{d}x$, is such that

$$\lim_{m\to\infty} T_{37}^{(m)} = -\int_\Omega \xi(u(x))\mathrm{div}(\varphi(x)g(x))\mathrm{d}x. \tag{92}$$

Proof. In the following proof, we designate by C_i various real values which can depend on d, Ω, g, F, R, φ and C_4 but not on m, and we drop the index m when this does not make any ambiguity. Let $m \in \mathbb{N}$ be given. Let us compare $T_{37}^{(m)}$ with $T_{38}^{(m)}$ defined by

$$T_{38}^{(m)} = -\sum_{K\in\mathcal{T}} v_K \int_K \mathrm{div}(\varphi(x)g(x))\mathrm{d}x.$$

We have, on one hand, that

$$\lim_{m\to\infty} T_{38}^{(m)} = -\int_\Omega v(x)\mathrm{div}(\varphi(x)g(x))\mathrm{d}x,$$

and on the other hand, we have

$$T_{38}^{(m)} = \sum_{(K,L)\in\mathcal{E}} (v_L - v_K)\int_{K|L} \varphi(x)g(x)\cdot\mathbf{n}_{K,L}\mathrm{d}s(x).$$

Thus we get that

$$T_{37}^{(m)} - T_{38}^{(m)} = T_{39}^{(m)} + T_{40}^{(m)} + T_{41}^{(m)},$$

with

$$T_{39}^{(m)} = \sum_{(K,L)\in\mathcal{E}} (v_L - v_K)\left(\varphi_K g_{K,L}^+ - \varphi_L g_{K,L}^- - \frac{g_{K,L}}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)\right),$$

$$T_{40}^{(m)} = \sum_{(K,L)\in\mathcal{E}} (v_L - v_K)\,(g_{K,L} - \bar{g}_{K,L})\left(\frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)\right),$$

and

$$T_{41}^{(m)} = \sum_{(K,L)\in\mathcal{E}} (v_L - v_K)\left(\int_{K|L}(\frac{\bar{g}_{K,L}}{m_{KL}} - g(x)\cdot\mathbf{n}_{K,L})\varphi(x)\mathrm{d}s(x)\right).$$

(recall that $\bar{g}_{K,L}$ is defined by (39)). Using $|\varphi_K - \frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)| \leq \mathrm{diam}(K)C_5$ and $|\varphi_L - \frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)| \leq \mathrm{diam}(L)C_5$, we get thanks to the Cauchy-Schwarz inequality,

$$|T_{39}^{(m)}|^2 \leq C_6 \quad \left(\sum_{(K,L)\in\mathcal{E}} |g_{K,L}|(v_K - v_L)^2\right)\left(\sum_{(K,L)\in\mathcal{E}} |g_{K,L}|(\mathrm{diam}(K)^2 + \mathrm{diam}(L)^2)\right).$$

Using (91) and

$$\sum_{(K,L)\in\mathcal{E}} |g_{K,L}|(\mathrm{diam}(K)^2 + \mathrm{diam}(L)^2) \leq C_7 \mathrm{size}(\mathcal{T}),$$

we thus get that

$$\lim_{m\to\infty} |T_{39}^{(m)}| = 0.$$

We now turn to the study of $T_{40}^{(m)}$. Since we have

$$T_{40}^{(m)} = -\sum_{K\in\mathcal{T}} v_K \sum_{L\in\mathcal{N}_K} (g_{K,L} - \bar{g}_{K,L}) \left(\frac{1}{m_{KL}} \int_{K|L} \varphi(x) \mathrm{d}s(x) \right),$$

we get, using the property (33),

$$T_{40}^{(m)} = -\sum_{K\in\mathcal{T}} v_K \sum_{L\in\mathcal{N}_K} (g_{K,L} - \bar{g}_{K,L}) \left(\frac{1}{m_{KL}} \int_{K|L} \varphi(x) \mathrm{d}s(x) - \varphi_K \right).$$

Thus, thanks to the Cauchy-Schwarz inequality and using (38), we get the existence of C_8, only depending on d, C_4 and R such that

$$(T_{40}^{(m)})^2 \leq C_8 \ \mathrm{cons}(g_{\mathcal{T}}).$$

Thus

$$\lim_{m\to\infty} |T_{40}^{(m)}| = 0.$$

We conclude with the study of $T_{41}^{(m)}$. Since

$$T_{41}^{(m)} = -\sum_{K\in\mathcal{T}} v_K \sum_{L\in\mathcal{N}_K} \left(\int_{K|L} (\frac{\bar{g}_{K,L}}{m_{KL}} - g(x)\cdot\mathbf{n}_{K,L})(\varphi(x) - \varphi_K) \mathrm{d}s(x) \right),$$

and since $\int_{K|L} (\frac{\bar{g}_{K,L}}{m_{KL}} - g(x)\cdot\mathbf{n}_{K,L})(\varphi(x) - \varphi_K)\mathrm{d}s(x) \leq C_9 m_{KL} \mathrm{diam}(K)^2$, we easily get

$$\lim_{m\to\infty} |T_{41}^{(m)}| = 0.$$

Gathering these results gives (92). □