SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2344–2370

ANALYTICAL AND NUMERICAL STUDY OF A MODEL OF EROSION AND SEDIMENTATION*

ROBERT EYMARD† AND THIERRY GALLOUËT‡

Abstract. We consider the following problem, arising within a geological model of sedimentation-erosion: For a given vector field g and a given nonnegative function F defined on a one- or two-dimensional domain Ω, find a vector field under the form $\tilde{g} = ug$, with $0 \leq u(x) \leq 1$ for a.e. $x \in \Omega$, such that $\mathrm{div}\tilde{g} + F \geq 0$ and $(u-1)(\mathrm{div}\tilde{g} + F) = 0$ in Ω. We first give a weak formulation of this problem, and we prove a comparison principle on a weak solution of the problem. Thanks to this property, we get the proof of the uniqueness of the weak solution. The existence of a solution results from the proof of the convergence of an original scheme. Numerical examples show the efficiency of this scheme and illustrate its convergence properties.

Key words. hyperbolic inequalities, doubling variable technique, process solutions, finite volume methods, erosion and sedimentation models

AMS subject classifications. 65N12, 65N30, 35R45

DOI. 10.1137/040605874

1. Introduction.

In the framework of the petroleum industry, geological simulations are used more and more in order to get a better knowledge of the history of the sedimentary basins. Among them, the computation of the sedimentation and erosion processes leads to a better knowledge of the geometry of the layers and of their lithological nature (see, for example, [15], [11], or [5]). An unknown of such models is the thickness $H(x,t)$ of the sediments at a point $(x,t) \in \Omega \times (0,T)$, where Ω describes the horizontal extension of the basin (the magnitude of the diameter of Ω can be about several hundreds of kilometers) and T is the age of the basin (between 0 and 10^7 years for example). The simplest model is a diffusion equation

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

where $\Lambda(x)$ is a matrix in the general case, reducing in most of the cases to a scalar function. But the model (1) is not sufficient for actual applications, in particular, because it does not account for the assymetry between the erosion process (due to the action of the weather) and the sedimentation process. Indeed, more realistic models (see [1] or [6] and references therein) are based on the introduction in (1) of a multiplier $\overline{u}(x,t)$ on the fluxes of sediments:

$$H_t(x,t) - \mathrm{div}[\Lambda(x)\overline{u}(x,t)\nabla H(x,t)] = 0 \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{2}$$

in order to satisfy the following constraints on $(\overline{u}, H)$,

$$H_t(x,t) \geq -F(x) \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{3}$$

$$0 \leq \overline{u}(x,t) \leq 1 \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{4}$$

*Received by the editors March 29, 2004; accepted for publication (in revised form) June 1, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sinum/43-6/60587.html
†Université de Marne-la-Vallée, 77454 Marne-la-Vallée, France (Robert.Eymard@univ-mlv.fr).
‡Université de Provence, 13453 Marseille cedex 13, France (gallouet@cmi.univ-mrs.fr).

and

$$(\overline{u}(x,t)-1)\ (H_t(x,t)+F(x))=0 \text{ for a.e. } (x,t)\in\Omega\times(0,T). \tag{5}$$

In (3) and (5), we denote by $F(x)\geq 0$ the maximum erosion rate at point x.

In practical situations, F is estimated by the geological study of the sedimental history, and may be improved by solving an inverse problem (which is quite complicated, by the way), using (2)–(5) as the direct problem. In large parts of the simulation domain, the transport of sediments is due mainly to gravity effects, taken into account by a scalar value for the matrix $\Lambda(x)$. In a same way as above, this scalar value can be estimated by some geological studies or by solving an inverse problem. However, the main mechanism for the transport of sediments is the action of surface water flows. These flows, located in river basins, can be represented by introducing anisotropic values for this matrix $\Lambda(x)$. The determination of realistic values for these parameters is not an easy task and is still a challenging subject of research for the simulation of the sedimentary basins. The function $\overline{u}$ is a complete unknown factor, reducing the flux of sediments in order to respect the constraint (3). Despite these difficulties of data identification, this model is considered interesting enough to be actually implemented in an industrial simulator (see [11], [6]).

Existence and uniqueness for the full problem (2)–(5) is an open problem (some partial results can be found in [10] or [2]). Thus we consider a semidiscretization in time of this system of equations. We define a time step $\delta t>0$, and for an integer n such that $n\delta t<T$, we assume that the function $H^{(n)}$ is an approximation of $H(\cdot,n\delta t)$. We then look for the functions $H^{(n+1)}$ and $\overline{u}^{(n+1)}$, respective approximations of $H(\cdot,(n+1)\delta t)$ and $\overline{u}(\cdot,(n+1)\delta t)$, which are solutions of the system of equations

$$\frac{1}{\delta t}(H^{(n+1)}(x)-H^{(n)}(x))-\operatorname{div}[\Lambda(x)\overline{u}^{(n+1)}(x)\nabla H^{(n)}(x)]=0 \text{ for a.e. } x\in\Omega, \tag{6}$$

under the constraints

$$\frac{1}{\delta t}(H^{(n+1)}(x)-H^{(n)}(x))\geq -F(x) \text{ for a.e. } x\in\Omega, \tag{7}$$

$$0\leq \overline{u}^{(n+1)}(x)\leq 1 \text{ for a.e. } x\in\Omega, \tag{8}$$

and

$$(\overline{u}^{(n+1)}(x)-1)\ \left(\frac{1}{\delta t}(H^{(n+1)}(x)-H^{(n)}(x))+F(x)\right)=0 \text{ for a.e. } x\in\Omega. \tag{9}$$

Denoting by $g(x)=\Lambda(x)\nabla H^{(n)}(x)$ and reporting in (7)–(9) the expression of $\frac{1}{\delta t}(H^{(n+1)}-H^{(n)})$ taken from (6), the unknown function $\overline{u}^{(n+1)}$ is then a solution u of the following system of equations:

$$\begin{aligned}&\operatorname{div}[u(x)g(x)]+F(x)\geq 0 \text{ for a.e. } x\in\Omega,\\ &0\leq u(x)\leq 1 \text{ for a.e. } x\in\Omega,\end{aligned} \tag{10}$$

and

$$(u(x)-1)\ (\operatorname{div}[u(x)g(x)]+F(x))=0 \text{ for a.e. } x\in\Omega. \tag{11}$$

Hence, if we are able to prove that problem (10)–(11) has one and only one solution $\tilde{g} = u(\cdot)g(\cdot)$, the function $H^{(n+1)}$ is then given by the relation $H^{(n+1)}(x) = H^{(n)}(x) + \delta t \, \mathrm{div}\tilde{g}(x)$ for a.e. $x \in \Omega$.

Remark 1.1. If there exist some regions where $g = 0$ and $F = 0$ simultaneously, it is clear that any value in $[0, 1]$ is possible for u. Nevertheless, $\tilde{g}$ is uniquely defined by the value 0 in such a region.

A fully implicit version of this method (namely $\nabla H^{(n)}(x)$ is replaced by $\nabla H^{(n+1)}(x)$ in (6)) in addition to a finite volume space discretization are used in an industrial simulator (see [11], [6]).

The aim of this paper is to focus on both the analytical and the numerical aspects of the subproblem (10)–(11). Although it is not clear that the resolution of this subproblem yields the complete theoretical resolution of the fully coupled problem (2)–(5), we emphasize that it leads to the key points of a correct numerical implementation.

In this paper, the following hypotheses, denoted Hypotheses (H), are assumed.

Hypotheses (H).

1. Ω is a bounded open subset of $\mathbb{R}^d$, $d \in \mathbb{N}^\star = \mathbb{N} \setminus \{0\}$ (in applications, $d = 2$), 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).
2. There exist two functions, $h \in C^1(\overline{\Omega})$ and $\Lambda : \Omega \longrightarrow \mathcal{M}_d$ (the set of bounded, symmetric, definite positive, $d \times d$ matrices) such that the function $g : \Omega \to \mathbb{R}^d$, defined by $g(x) = \Lambda(x)\nabla h(x)$ for all $x \in \Omega$, is Lipschitz continuous on $\overline{\Omega}$ and satisfies $g(x) \cdot \mathbf{n}(x) = 0$ for a.e. $x \in \partial\Omega$.
3. $F \in L^\infty(\Omega)$ is such that $F(x) \geq 0$ for a.e. $x \in \Omega$.

As we see below in section 2, there does not always exist a continuous function $u : \Omega \to \mathbb{R}$ such that (10)–(11) are satisfied, and the regularity of $\tilde{g} = ug$ in the general case is an open problem. Therefore we first look for a weak formulation of problem (10)–(11). 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 (10) by $\xi'(u(x))\varphi(x)$, and we integrate on Ω. We get

(12)
$$\int_\Omega \xi'(u(x))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x = \int_\Omega \xi'(1)\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x + \int_\Omega (\xi'(u(x)) - \xi'(1))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x.$$

The second term of the right-hand side vanishes, using (11), and the first one is nonnegative. This leads to

$$\int_\Omega \xi'(u(x))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x \geq 0. \tag{13}$$

We remark that, for any function ξ which is such that $\xi'(1) \geq 0$ and ξ' is decreasing, we can get (13) from (12) for any function u which only verifies (10). For this reason, we now assume that ξ is convex (in the sense that ξ' is nondecreasing, this terminology is used in the sequel of this paper), and we develop equation (13), integrating by parts. We then derive the following weak sense for a solution to problem (10)–(11).

Definition 1.1 (weak solution to problem (10)–(11)). *Under Hypotheses* (H), *we say that a function* $\tilde{g} \in L^\infty(\Omega)^d$ *is a weak solution to problem* (10)–(11) *if there*

exists $u \in L^\infty(\Omega)$ *such that* $\tilde{g}(x) = u(x)g(x)$ *for a.e.* $x \in \Omega$, *and* u *satisfies the following inequalities:* $0 \le u(x) \le 1$ *for a.e.* $x \in \Omega$ *and*

$$\int_\Omega \big(\xi(u(x))(-g(x)\cdot\nabla\varphi(x)) + [\xi'(u(x))u(x) - \xi(u(x))]\varphi(x)\mathrm{div}g(x) + \xi'(u(x))\varphi(x)F(x)\big)\mathrm{d}x \ge 0 \tag{14}$$
$$\forall \xi \in C^1(\mathbb{R}) \text{ convex such that } \xi'(1) \ge 0 \qquad \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+).$$

The following proposition expresses that any weak solution in the above sense satisfies (10) in a weak sense, and the next one shows that any regular weak solution satisfies (10)–(11), thus completing the justification of Definition 1.1.

PROPOSITION 1.2. *Under Hypotheses* (H), *let* $\tilde{g} : \Omega \to \mathbb{R}^d$ *be a weak solution to problem* (10)–(11) *in the sense of Definition* 1.1. *Then*

$$\int_\Omega (-\tilde{g}(x)\cdot\nabla\varphi(x)\mathrm{d}x + F(x)\varphi(x))\mathrm{d}x \ge 0 \qquad \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+). \tag{15}$$

Proof. Let us assume that $u \in L^\infty(\Omega)$ is such that $\tilde{g}(x) = u(x)g(x)$ and $0 \le u(x) \le 1$ for a.e. $x \in \Omega$, and (14) is satisfied. Let us take $\xi : s \mapsto s$ in (14). We then obtain (15). ∎

PROPOSITION 1.3. *Under Hypotheses* (H), *let* $\tilde{g} : \Omega \to \mathbb{R}^d$ *be a Lipschitz continuous function. Then* $\tilde{g}$ *is a weak solution to problem* (10)–(11) *in the sense of Definition* 1.1 *if and only if there exists a function* $u \in L^\infty(\Omega)$ *with* $\tilde{g}(x) = u(x)g(x)$ *and* $0 \le u(x) \le 1$ *for a.e.* $x \in \Omega$ *such that* (10) *and* (11) *are satisfied by the function* u.

Proof. Let us assume that $\tilde{g}$ is a weak solution to problem (10)–(11) in the sense of Definition 1.1. Then there exists $u \in L^\infty(\Omega)$ such that $\tilde{g}(x) = u(x)g(x)$ and $0 \le u(x) \le 1$ for a.e. $x \in \Omega$, and (14) is satisfied. Proposition 1.2 shows that (10) is satisfied by the function u for a.e. $x \in \Omega$. In order to prove that (11) is satisfied for a.e. $x \in \Omega$ by the function u, we shall separate the cases $x \in \Omega_0 := \{x \in \Omega, g(x) = 0\}$ and $x \in \Omega \setminus \Omega_0$. Let us take in (14) a test function φ whose support is included in the open set $\Omega \setminus \Omega_0$. Since the function u verifies $u(x) = |\tilde{g}(x)|/|g(x)|$ for a.e. $x \in \Omega \setminus \Omega_0$, u is Lipschitz continuous on the support of φ; we can thus integrate by parts, which produces, from (14), that (13) is satisfied by u. Let us now prove that u verifies (11).

Choosing $\xi : s \mapsto (s-1)^2$, we get that $\int_\Omega (u(x)-1)\varphi(x)(\mathrm{div}[u(x)g(x)]+F(x))\mathrm{d}x \ge 0$ holds. This implies that $(u(x)-1)(\mathrm{div}[u(x)g(x)] + F(x)) \ge 0$ for a.e. $x \in \Omega$ such that $g(x) \ne 0$. But on one hand, $u(x) \le 1$ for a.e. $x \in \Omega$, and on the other hand, (10) is satisfied for a.e. $x \in \Omega$. Therefore, u verifies (11) for a.e. $x \in \Omega \setminus \Omega_0$.

Let us now obtain the same conclusion for a.e. $x \in \Omega_0$. Let $\eta \in C^1(\mathbb{R})$ be a function such that $0 \le \eta(x) \le 1$ for all $x \in \mathbb{R}$, $\eta(0) = 1$ and $\mathrm{support}(\eta) \subset [-1, 1]$. For all $n \in \mathbb{N}^\star$, let us define the Lipschitz continuous function $\varphi_n : x \mapsto \eta(n|g(x)|)$. On one hand, we have that for a.e. $x \in \Omega_0$, $g(x)\cdot\nabla\varphi_n(x) = 0$ holds. On the other hand, for all $x \in \Omega \setminus \Omega_0$, we get that $g(x)\cdot\nabla\varphi_n(x)$ tends to 0 as $n \to \infty$ and remains bounded (indeed, it suffices to consider the cases $|g(x)| \le 1/n$ and $|g(x)| \ge 1/n$ and to use the property $\nabla g_i \in L^\infty(\Omega)^d$, where g_i, $i = 1, \ldots, d$ are the components of g).

We then introduce $\xi : s \to (s-1)^2$ and $\varphi = \varphi_n$ in (14) (this is possible, taking regularizations in $C^1(\overline{\Omega}, \mathbb{R}_+)$ of φ_n). We then get

$$T_1^{(n)} + T_2^{(n)} + T_3^{(n)} \ge 0, \tag{16}$$

with $T_1^{(n)} = \int_\Omega (u(x)-1)^2(-g(x)\cdot\nabla\varphi_n(x))\mathrm{d}x$, $T_2^{(n)} = \int_\Omega (u(x)^2-1)\varphi_n(x)\mathrm{div}g(x)\mathrm{d}x$, and $T_3^{(n)} = 2\int_\Omega (u(x)-1)F(x)\varphi_n(x)\mathrm{d}x$. Thus, thanks to the convergence properties of $g\cdot\nabla\varphi_n$ and to the dominated convergence theorem, we get that $T_1^{(n)}$ tends to 0 as $n\to\infty$.

Since $\varphi_n(x)$ tends to 0 for all $x\in\Omega\setminus\Omega_0$ and to 1 for all $x\in\Omega_0$, we get that $T_2^{(n)}$ tends to $\int_{\Omega_0}(u(x)^2-1)\mathrm{div}g(x)\mathrm{d}x$. Since $g(x)=0$ for all $x\in\Omega_0$, then $\partial_i g(x)=0$ for a.e. $x\in\Omega_0$ and all $i=1,\dots,d$ (this classical property has been shown, for example, in [17]), which produces $\int_{\Omega_0}(u(x)^2-1)\mathrm{div}g(x)\mathrm{d}x=0$.

We finally get that $T_3^{(n)}$ tends to $2\int_{\Omega_0}(u(x)-1)F(x)\mathrm{d}x$ as $n\to\infty$.

We thus get, passing to the limit $n\to\infty$ in (16), $\int_{\Omega_0}(u(x)-1)F(x)\mathrm{d}x\ge 0$, which proves that $u(x)=1$ for a.e. $x\in\Omega_0$ such that $F(x)>0$.

Therefore, for a.e. $x\in\Omega_0$, either $F(x)>0$ and $u(x)=1$, or $F(x)=0$ and $\mathrm{div}(\tilde{g}(x))+F(x)=0$, since $\tilde{g}(x)=0$ for a.e. $x\in\Omega_0$. Thus (11) is satisfied for a.e. $x\in\Omega_0$.

Reciprocally, let us assume that (10) and (11) are satisfied a.e. by the function u. We then get that (13) is satisfied, and therefore equation (14) is satisfied. This proves that $\tilde{g}$ is a weak solution to problem (10)–(11) in the sense of Definition 1.1. ∎

This paper is organized as follows. We first give, in section 2, the analytical expression of the weak solution in the one-dimensional case (the uniqueness result, proved in section 3, indeed holds in this case). In section 3, we first give a characterization of the set $\mathcal{C}(g,F)$ of functions which weakly satisfy (10). We prove a comparison result between a weak process solution to problem (10)–(11) (defined in Definition 3.3) and any element of $\mathcal{C}(g,F)$. This result suffices to prove the uniqueness of the weak solution to problem (10)–(11) in the sense of Definition 1.1. We then present a numerical scheme in section 4. The existence and uniqueness of a discrete solution is itself a nontrivial problem, which we solve by proving the convergence of an iterative method. This scheme is then proven to converge to a weak process solution to problem (10)–(11) in the sense of Definition 3.3. Thanks to the uniqueness result of the weak solution, we deduce the strong convergence result of the numerical scheme to this weak solution. We then give some numerical results in section 5 and conclude with some open problems.

2. Weak solutions in the one-dimensional case. We have the following result.

PROPOSITION 2.1 (expression of the weak solution in the one-dimensional case). *Let $(a,b)\in\mathbb{R}^2$ be such that $a<b$, let $F\in L^\infty((a,b))$ be a nonnegative function, and let $g\in C^0([a,b])$ be a Lipschitz continuous function with $g(a)=g(b)=0$.*

Then, the function $\tilde{g}:[a,b]\to\mathbb{R}$ defined by

$$\tilde{g}(x) = \min_{y\in[x,b]}\left(g^+(y)+\int_x^y F(t)\mathrm{d}t\right) - \min_{y\in[a,x]}\left(g^-(y)+\int_y^x F(t)\mathrm{d}t\right) \quad \forall x\in[a,b], \tag{17}$$

where for all $s\in\mathbb{R}$ we denote $s^+=\max(s,0)$ and $s^-=\max(-s,0)$, is the unique weak solution to problem (10)–(11) *in the sense of Definition* 1.1.

Proof. Let us first remark that $\tilde{g}$ defined as such verifies that for all $x\in[a,b]$, $\tilde{g}^+(x)=\min_{y\in[x,b]}\left(g^+(y)+\int_x^y F(t)\mathrm{d}t\right)$ and $\tilde{g}^-(x)=\min_{y\in[a,x]}(g^-(y)+\int_y^x F(t)\mathrm{d}t)$ with $0\le\tilde{g}^+(x)\le g^+(x)$ and $0\le\tilde{g}^-(x)\le g^-(x)$. Then the function $\tilde{g}^+$ satisfies

$\tilde{g}^+(x) = \min_{y\in[a,b]} G_p(x,y)$ for all $x \in [a,b]$ with

$$G_p(x,y) = g^+(\max(x,y)) + \int_x^{\max(x,y)} F(t)\mathrm{d}t \qquad \forall (x,y) \in [a,b]^2.$$

Similarly, we have $\tilde{g}^-(x) = \min_{y\in[a,b]} G_m(x,y)$ for all $x \in [a,b]$ with

$$G_m(x,y) = g^-(\min(x,y)) + \int_{\min(x,y)}^x F(t)\mathrm{d}t \qquad \forall (x,y) \in [a,b]^2.$$

It is then clear that the functions G_p and G_m are Lipschitz continuous on $[a,b]^2$ with any Lipschitz constant M such that M is a bound of $F + |g'|$ in $L^\infty((a,b))$. Let $(x,\bar{x}) \in [a,b]^2$ be given, and let $(Y,\bar{Y}) \in [a,b]^2$ be such that $\tilde{g}^+(x) = G_p(x,Y)$ and $\tilde{g}^+(\bar{x}) = G_p(\bar{x},\bar{Y})$. Since we have

$$\tilde{g}^+(x) - \tilde{g}^+(\bar{x}) \le G_p(x,\bar{Y}) - G_p(\bar{x},\bar{Y}) \le M|x-\bar{x}|,$$

and, inverting the roles of x and $\bar{x}$,

$$\tilde{g}^+(\bar{x}) - \tilde{g}^+(x) \le G_p(x,Y) - G_p(\bar{x},Y) \le M|x-\bar{x}|,$$

we thus get that $\tilde{g}^+$ is Lipschitz continuous. Since the same proof holds for $\tilde{g}^-$, we thus get that $\tilde{g} = \tilde{g}^+ - \tilde{g}^-$ is Lipschitz continuous as well. We thus define the function $u : [a,b] \to [0,1]$ by $u(x) = 1$ for all $x \in \Omega$ such that $g(x) = 0$ and $u(x) = \tilde{g}(x)/g(x)$ for all $x \in [a,b]$ such that $g(x) \neq 0$. Let us prove that u satisfies (10)–(11) (from Proposition 1.3, since Hypotheses (H) are satisfied, this is sufficient to conclude). Since for all $x \in [a,b]$ such that $g(x) = 0$, $\tilde{g}(x) = 0$ holds, $\tilde{g}'(x) + F(x) \ge 0$ for a.e. $x \in [a,b]$ such that $g(x) = 0$ [17]. Let $x \in [a,b]$ be such that $g(x) > 0$. Then there exists $\alpha > 0$ such that $x+\alpha \le b$ and $g(y) > 0$ for all $y \in (x, x+\alpha)$. For $\bar{x} \in (x, x+\alpha)$, let $\bar{Y} \in [\bar{x}, b]$ be such that $\tilde{g}(\bar{x}) = G_p(\bar{x},\bar{Y})$. We have

$$\tilde{g}(x) - \tilde{g}(\bar{x}) \le G_p(x,\bar{Y}) - G_p(\bar{x},\bar{Y}) = \int_x^{\bar{x}} F(t)\mathrm{d}t. \tag{18}$$

The above inequality proves that $\tilde{g}'(x)+F(x) \ge 0$ for a.e. $x \in [a,b]$ such that $g(x) > 0$. Similarly, we obtain that $\tilde{g}'(x) + F(x) \ge 0$ for a.e. $x \in [a,b]$ such that $g(x) < 0$. This proves that (10) is satisfied. Let $x \in (a,b)$ such that $u(x) < 1$. Let us assume that $g(x) > 0$. Again, there exists $\alpha > 0$ such that $x + \alpha \le b$ and $g(y) > 0$ for all $y \in (x, x+\alpha)$, and again, for all $\bar{x} \in (x, x+\alpha)$, (18) holds. Since we have $0 \le \tilde{g}(x) < g(x)$, there exists $Y \in (x,b)$ such that $\tilde{g}(x) = G_p(x,Y)$. Therefore, for all $\bar{x} \in (x,Y)$, since $Y > \bar{x}$, we get

$$\tilde{g}(x) - \tilde{g}(\bar{x}) \ge G_p(x,Y) - G_p(\bar{x},Y) = \int_x^{\bar{x}} F(t)\mathrm{d}t.$$

Thus, for all $\bar{x} \in (x, \min(Y, x+\alpha))$, we get $\tilde{g}(x) - \tilde{g}(\bar{x}) = \int_x^{\bar{x}} F(t)\mathrm{d}t$, which implies that $\tilde{g}'(x) = -F(x)$ for a.e. $x \in \Omega$ such that $u(x) < 1$ and $g(x) > 0$. The case $u(x) < 1$ and $g(x) < 0$ can be similarly handled. Therefore $\tilde{g}$ is Lipschitz continuous and (10)–(11) are satisfied. Thanks to Proposition 1.3, this completes the proof that $\tilde{g}$ is a weak solution to problem (10)–(11) in the sense of Definition 1.1.

Since, within the hypotheses of the above proposition, Hypotheses (H) are satisfied (in particular, $g = h'$ with $h : x \mapsto \int_a^x g(t)\mathrm{d}t$), we can apply Proposition 3.5, which

implies the uniqueness of the weak solution to problem (10)–(11) in the sense of Definition 1.1. □

Let us take two simple examples (one can find some examples inspired by geological problems in [6]). We consider a one-dimensional case (see Figure 1 below), with $\Omega = (-1,1)$, $g : x \mapsto x^3 - x$, and $F : x \mapsto 1/2$. In this case, it is easy to verify that the function $\tilde{g}$ defined by (17) is such that $\tilde{g} = ug$, where the function u is such that $u : x \mapsto 1$ for all $x \in (-1, -\sqrt{1/2}) \cup (\sqrt{1/2}, 1)$ and $u : x \mapsto 1/(2(1-x^2))$ for all $x \in (-\sqrt{1/2}, \sqrt{1/2})$. We thus obtain that the function u is continuous over Ω, but this is not always the case.

Indeed, let us consider the case $\Omega = (-1,1)$, $g : x \mapsto x^3 - x$ for all $x \in [-1,0]$, $g : x \mapsto \frac{1}{2}(x^3 - x)$ for all $x \in [0,1]$, and $F : x \mapsto 1/2$. In such a case, g is only Lipschitz continuous, and the function $\tilde{g} = ug$ given by (17) is such that $u : x \mapsto 1$ for all $x \in (-1, -\sqrt{1/2}) \cup (0,1)$ and $u : x \mapsto 1/(2(1-x^2))$ for all $x \in (-\sqrt{1/2}, 0)$. This function u is therefore discontinuous in 0, although the function $\tilde{g} = ug$ remains Lipschitz continuous.

3. Uniqueness results.

3.1. Properties of the set of functions which satisfy (10). We consider in this section the set $\mathcal{C}(g,F)$ of functions which satisfy (10) in the sense of distributions. We shall prove below that the weak solution $\tilde{g}$ to problem (10)–(11) in the sense of Definition 1.1 is the projection of g in $L^2(\Omega)^d$ on $\mathcal{C}(g,F)$, and it is an extremal point of $\mathcal{C}(g,F)$ in the sense that $|\tilde{g}| \geq |\gamma|$ for all $\gamma \in \mathcal{C}(g,F)$ (see Proposition 3.5). The proof of this property is obtained thanks to the characterization of $\mathcal{C}(g,F)$ given by Proposition 3.2.

DEFINITION 3.1 (the set $\mathcal{C}(g,F)$). *Under Hypotheses* (H), *we define the set* $\mathcal{C}(g,F)$ *of functions* $\gamma \in L^2(\Omega)^d$ *such that there exists* $v \in L^\infty(\Omega)$, *with* $\gamma(x) = v(x)g(x)$ *and* $0 \leq v(x) \leq 1$, *for a.e.* $x \in \Omega$ *and*

$$\int_{\mathbb{R}_+} \int_\Omega \left([-\gamma(x) \cdot \nabla\varphi(x)] + \varphi(x)F(x)\right) \mathrm{d}x \geq 0 \qquad \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+). \tag{19}$$

Remark 3.1 (some properties of $\mathcal{C}(g,F)$). The set $\mathcal{C}(g,F)$ is nonempty (because $0 \in \mathcal{C}(g,F)$), convex (since the left-hand side of (19) is linear with respect to γ), and closed (in $L^2(\Omega)^d$).

Remark 3.2 (weak solutions and $\mathcal{C}(g,F)$). Thanks to Proposition 1.2, any weak solution to problem (10)–(11) in the sense of Definition 1.1 belongs to $\mathcal{C}(g,F)$.

We have the following proposition, which gives a characterization of the functions of $\mathcal{C}(g,F)$.

PROPOSITION 3.2 (characterization of $\mathcal{C}(g,F)$). *Under Hypotheses* (H), *let* $v \in L^\infty(\Omega)$, *such that* $0 \leq v(x) \leq 1$ *for a.e.* $x \in \Omega$, *and let* $\gamma(x) = v(x)g(x)$. *Then* $\gamma \in \mathcal{C}(g,F)$ *(defined in Definition* 3.1*) holds if and only if the following property holds:*

$$\begin{aligned} \int_\Omega \big(\xi(v(x))[-g(x) \cdot \nabla\varphi(x)] + [\xi'(v(x))v(x) - \xi(v(x))]\,\varphi(x)\ \mathrm{div} g(x) \\ + \xi'(v(x))\varphi(x)F(x)\big)\mathrm{d}x \geq 0 \\ \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+),\ \forall \xi \in C^1(\mathbb{R})\ \textit{s.t.}\ \forall \kappa \in [0,1],\ \xi'(\kappa) \geq 0. \end{aligned} \tag{20}$$

Proof. Under the hypotheses of the above proposition, let us assume that $\gamma \in \mathcal{C}(g,F)$. 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

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

and

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

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

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

We then define the functions $v_n(y) = \int_\Omega v(x)\rho_n(x-y)\mathrm{d}x$. Let $\psi \in C^1(\overline{\Omega}, \mathbb{R}_+)$ be given. For a given $y \in \Omega$, we introduce the function $\varphi : x \to \xi'(v_n(y))\psi(y)\rho_n(y-x) \in C^1(\overline{\Omega}, \mathbb{R}_+)$ in (19), and we integrate with respect to y. We thus get $T_4^{(n)} + T_5^{(n)} \geq 0$ with

$$T_4^{(n)} = -\int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)v(x)g(x) \cdot \nabla\rho_n(y-x)\mathrm{d}x\mathrm{d}y \tag{24}$$

and

$$T_5^{(n)} = \int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)F(x)\rho_n(y-x)\mathrm{d}x\mathrm{d}y. \tag{25}$$

The limit of the last term, as $n \to \infty$, satisfies

$$\lim_{n\to\infty} T_5^{(n)} = \int_\Omega (\xi'(v(y))\psi(y)F(y))\,\mathrm{d}y.$$

We then turn to the study of $T_4^{(n)}$ as $n \to \infty$. We have $T_4^{(n)} = T_6^{(n)} + T_7^{(n)} + T_8^{(n)}$ with

$$T_6^{(n)} = \int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)v(x)g(y) \cdot \nabla\rho_n(y-x)\mathrm{d}x\mathrm{d}y,$$

$$T_7^{(n)} = \int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)v(y)(g(x)-g(y)) \cdot \nabla\rho_n(y-x)\mathrm{d}x\mathrm{d}y,$$

and

$$T_8^{(n)} = \int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)(v(x)-v(y))(g(x)-g(y)) \cdot \nabla\rho_n(y-x)\mathrm{d}x\mathrm{d}y.$$

We then have

$$T_6^{(n)} = \int_\Omega \xi'(v_n(y))\psi(y)g(y) \cdot \nabla v_n(y)\mathrm{d}y = \int_\Omega \psi(y)g(y) \cdot \nabla\xi(v_n)(y)\mathrm{d}y,$$

which delivers, thanks to an integration by parts with respect to y,

$$T_6^{(n)} = -\int_\Omega \xi(v_n(y))\mathrm{div}[\psi(y)g(y)]\mathrm{d}y.$$

This leads to

$$\lim_{n\to\infty} T_6^{(n)} = \int_\Omega \xi(v(y))\mathrm{div}[\psi(y)g(y)]\mathrm{d}y.$$

We also have, thanks to an integration by parts with respect to x,

$$T_7^{(n)} = -\int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)v(y)\rho_n(y-x)\mathrm{div}g(x)\mathrm{d}x\mathrm{d}y,$$

which produces

$$\lim_{n\to\infty} T_7^{(n)} = \int_\Omega \xi'(v(y))v(y)\psi(y)\mathrm{div}g(y)\mathrm{d}y.$$

Finally, we get

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

thanks to the continuity in means of v and to the fact that $x \mapsto (g(x)-g(y))\cdot\nabla\rho_n(y-x)$ belongs to $L^1(\Omega)$. Then (20) is obtained by gathering all the results obtained above by passing to the limit $n \to \infty$.

Conversely, it suffices to choose the function $\xi : s \mapsto s$ in (20), for obtaining (19). □

3.2. Weak process solutions. Since we consider below the convergence of numerical schemes, on which the only estimate that we obtain in this case is an $L^\infty(\Omega)$ estimate, we must therefore consider weaker solutions than that defined in Definition 1.1, namely, weak process solutions. This notion of a weak process solution, introduced in [7], is related to the notion of Young measure first used by [3] in the nonlinear scalar hyperbolic framework. Young measures are extensively used in optimal control, nonconvex variational problems, phase transitions, microstructure problems, ... (see, e.g., [14], [16]).

The uniqueness result proven below leads to the uniqueness of such a weak process solution and to the fact that any weak process solution is indeed a weak solution. We then obtain the uniqueness of the weak solution to problem (10)–(11) in the sense of Definition 1.1. Moreover, this result is mainly used in the study of the numerical scheme in order to prove its strong convergence.

DEFINITION 3.3 (weak process solutions to problem (10)–(11)). *Under Hypotheses* (H), *we say that a function $\hat{g}$ is a weak process solution to problem* (10)–(11) *if there exists $u \in L^\infty(\Omega\times(0,1))$ such that $\hat{g} : (x,\alpha) \mapsto u(x,\alpha)g(x)$ for a.e. $(x,\alpha) \in \Omega\times(0,1)$. And u satisfies the following inequalities: $0 \le u(x,\alpha) \le 1$ for a.e. $(x,\alpha) \in \Omega\times(0,1)$ and*

(26)
$$\int_\Omega \int_0^1 \big(\xi(u(x,\alpha))(-g(x)\cdot\nabla\varphi(x)) + [\xi'(u(x,\alpha))u(x,\alpha) - \xi(u(x,\alpha))]\varphi(x)\mathrm{div}g(x) + \xi'(u(x,\alpha))\varphi(x)F(x)\big)\mathrm{d}\alpha\mathrm{d}x \ge 0$$
$$\forall \xi \in C^1(\mathbb{R}),\ \textit{convex s.t. } \xi'(1) \ge 0,\ \forall \varphi \in C^1(\overline{\Omega},\mathbb{R}_+).$$

We first prove the following property, which at the same time, gives some elements to conclude to the uniqueness of the weak process solution but also helps to prove that this solution is an extremal point of $\mathcal{C}(g,F)$.

PROPOSITION 3.4 (comparison of a weak process solution and an element of $\mathcal{C}(g,F)$). *Under Hypotheses* (H), *let $\gamma \in \mathcal{C}(g,F)$ be given, where $\mathcal{C}(g,F)$ is defined*

in Definition 3.1, *and let* $v \in L^\infty(\Omega)$, *such that* $\gamma(x) = v(x)g(x)$ *and* $0 \le v(x) \le 1$ *for a.e.* $x \in \Omega$. *Let* $\hat{g}$ *be a weak process solution to problem* (10)–(11) *in the sense of Definition* (3.3). *Let* $u \in L^\infty(\Omega \times (0,1))$ *be such that* $0 \le u(x,\alpha) \le 1$ *and* $\hat{g} : (x,\alpha) \mapsto u(x,\alpha)g(x)$ *for a.e.* $(x,\alpha) \in \Omega \times (0,1)$ *and such that* u *satisfies* (27). *Then the following inequality holds:*

$$\int_\Omega \int_0^1 (v(x) - u(x,\alpha))^+ \left[-g(x) \cdot \nabla\varphi(x)\right] \mathrm{d}\alpha \mathrm{d}x \ge 0 \qquad \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+). \tag{27}$$

Proof. This proof uses the method of doubling variables (first introduced by Krushkov [12]) adapted to weak process solutions [8].

Let us assume the hypotheses of the proposition. Let $\eta \in C^1(\mathbb{R}^2, \mathbb{R})$ be given such that $\eta(\cdot, b)$ is convex for all $b \in (-\infty, 1]$. We also assume that $\partial_1\eta$, the derivative of η with respect to its first argument, is such that $\partial_1\eta(1,b) \ge 0$ for all $b \in [0,1]$, and that $\partial_2\eta$, the derivative of η with respect to its second argument, is such that $\partial_2\eta(a,b) \ge 0$ for all $a, b \in [0,1]$. 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 introduce $\xi(\cdot) = \eta(\cdot, v(y))$ and $\varphi = \psi(\cdot, y)$ in (27) for $y \in \Omega$, and we integrate the result on Ω. This produces

$$\begin{aligned} &\int_\Omega\int_\Omega\int_0^1 \Big(\eta(u(x,\alpha), v(y)) \left[-g(x) \cdot \nabla_x\psi(x,y)\right] \\ &+ \left[\partial_1\eta(u(x,\alpha), v(y))u(x,\alpha) - \eta(u(x,\alpha), v(y))\right] \psi(x,y)\mathrm{div}g(x) \\ &\qquad + \partial_1\eta(u(x,\alpha), v(y))\psi(x,y)F(x)\Big)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y \ge 0. \end{aligned} \tag{28}$$

We now consider (20) for v, with $\xi(\cdot) = \eta(u(x,\alpha), \cdot)$ and $\varphi = \psi(x, \cdot)$, and we integrate the result on $\Omega \times (0,1)$. We thus get

$$\begin{aligned} &\int_\Omega\int_\Omega\int_0^1 \Big(\eta(u(x,\alpha), v(y)) \left[-g(y) \cdot \nabla_y\psi(x,y)\right] \\ &+ \left[\partial_2\eta(u(x,\alpha), v(y))v(y) - \eta(u(x,\alpha), v(y))\right] \psi(x,y)\mathrm{div}g(y) \\ &\qquad + \partial_2\eta(u(x,\alpha), v(y))\psi(x,y)F(y)\Big)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y \ge 0. \end{aligned} \tag{29}$$

We now add (28) and (29). This delivers

$$T_9 + T_{10} + T_{11} \ge 0, \tag{30}$$

where

$$\begin{aligned} T_9 = -\int_\Omega\int_\Omega\int_0^1 \quad &\eta(u(x,\alpha), v(y)) \\ &\times \big(g(x) \cdot \nabla_x\psi(x,y) + g(y) \cdot \nabla_y\psi(x,y)\big)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{31}$$

$$\begin{aligned} T_{10} = \int_\Omega\int_\Omega\int_0^1 \Big(&\big(\partial_1\eta(u(x,\alpha), v(y))u(x,\alpha) - \eta(u(x,\alpha), v(y))\big)\psi(x,y)\mathrm{div}g(x) \\ &+ \big(\partial_2\eta(u(x,\alpha), v(y))v(y,\beta) - \eta(u(x,\alpha), v(y))\big)\psi(x,y)\mathrm{div}g(y)\Big)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{32}$$

and

$$T_{11} = \int_\Omega \int_\Omega \int_0^1 \big(\partial_1\eta(u(x,\alpha),v(y))F(x) + \partial_2\eta(u(x,\alpha),v(y))F(y)\big)\psi(x,y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y. \tag{33}$$

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 (28) and (29), which gives $\psi \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}_+)$. We thus get, from (30),

$$T_9^{(n)} + T_{10}^{(n)} + T_{11}^{(n)} \geq 0, \tag{34}$$

with

$$\begin{aligned} T_9^{(n)} &= -\int_\Omega \int_\Omega \int_0^1 \eta(u(x,\alpha),v(y)) \\ &\quad \times \Big(\rho_n(x-y)g(x)\cdot\nabla\phi(x) + \phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\Big)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{35}$$

$$\begin{aligned} T_{10}^{(n)} &= \int_\Omega \int_\Omega \int_0^1 \Big([\partial_1\eta(u(x,\alpha),v(y))u(x,\alpha) - \eta(u(x,\alpha),v(y))]\,\mathrm{div}g(x) \\ &\quad + [\partial_2\eta(u(x,\alpha),v(y))v(y) - \eta(u(x,\alpha),v(y))]\,\mathrm{div}g(y)\Big)\phi(x)\rho_n(x-y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{36}$$

$$\begin{aligned} T_{11}^{(n)} = \int_\Omega \int_\Omega \int_0^1 &\big(\partial_1\eta(u(x,\alpha),v(y))F(x) + \partial_2\eta(u(x,\alpha),v(y))F(y)\big) \\ &\times \phi(x)\rho_n(x-y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{37}$$

We have $T_9^{(n)} = T_{12}^{(n)} + T_{13}^{(n)} + T_{14}^{(n)}$, with

$$T_{12}^{(n)} = -\int_\Omega \int_\Omega \int_0^1 \eta(u(x,\alpha),v(y))\rho_n(x-y)g(x)\cdot\nabla\phi(x)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \tag{38}$$

$$\begin{aligned} T_{13}^{(n)} = -\int_\Omega \int_\Omega \int_0^1 &\eta(u(x,\alpha),v(x)) \\ &\times \phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{39}$$

$$\begin{aligned} T_{14}^{(n)} = -\int_\Omega \int_\Omega \int_0^1 &\Big(\eta(u(x,\alpha),v(y)) - \eta(u(x,\alpha),v(x))\Big) \\ &\times \phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{40}$$

The limit of $T_{12}^{(n)}$ as $n \longrightarrow \infty$ is given by

$$\lim_{n\to\infty} T_{12}^{(n)} = -\int_\Omega \int_0^1 \eta(u(x,\alpha),v(x))g(x)\cdot\nabla\phi(x)\mathrm{d}\alpha\mathrm{d}x.$$

Thanks to an integration by parts with respect to y and to Hypotheses (H), we get $T_{13}^{(n)} = T_{15}^{(n)} + T_{16}^{(n)}$, where

$$T_{15}^{(n)} = \int_\Omega \int_{\partial\Omega} \int_0^1 \eta(u(x,\alpha),v(x))\phi(x)\rho_n(x-y)g(x)\cdot\mathbf{n}(y)\mathrm{d}\alpha\mathrm{d}y\mathrm{d}x \tag{41}$$

and

$$T_{16}^{(n)} = \int_\Omega \int_\Omega \int_0^1 \eta(u(x,\alpha),v(x))\phi(x)\rho_n(x-y)\mathrm{div}g(y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y. \tag{42}$$

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

$$\lim_{n\to\infty} \int_\Omega \int_0^1 \eta(u(x,\alpha),v(x))\phi(x)\rho_n(x-y)g(x)\cdot\mathbf{n}(y)\mathrm{d}\alpha\mathrm{d}x = 0,$$

which produces

$$\lim_{n\to\infty} T_{15}^{(n)} = 0,$$

and therefore

$$\lim_{n\to\infty} T_{13}^{(n)} = \lim_{n\to\infty} T_{16}^{(n)} = \int_\Omega \int_0^1 \eta(u(x,\alpha),v(x))\phi(x)\mathrm{div}g(x)\mathrm{d}\alpha\mathrm{d}x.$$

Thanks to the theorem of continuity in means applied to the function v and 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

$$\lim_{n\to\infty} T_{14}^{(n)} = 0.$$

We have, again using the Lebesgue dominated convergence theorem,

$$\lim_{n\to\infty} T_{10}^{(n)} = \int_\Omega \int_0^1 \big(\partial_1\eta(u(x,\alpha),v(x))u(x,\alpha) + \partial_2\eta(u(x,\alpha),v(x))v(x) - 2\eta(u(x,\alpha),v(x))\big)\phi(x)\mathrm{div}g(x)\mathrm{d}\alpha\mathrm{d}x$$

and

$$\lim_{n\to\infty} T_{11}^{(n)} = \int_\Omega \int_0^1 \left(\partial_1\eta(u(x,\alpha),v(x)) + \partial_2\eta(u(x,\alpha),v(x))\right) F(x)\phi(x)\mathrm{d}\alpha\mathrm{d}x.$$

We thus get, passing to the limit $n\to\infty$ in (34),

$$\begin{aligned}\int_\Omega \int_0^1 \Big(&\eta(u(x,\alpha),v(x))\left[-g(x)\cdot\nabla\phi(x)\right]\\ &+\big(\partial_1\eta(u(x,\alpha),v(x))u(x,\alpha) + \partial_2\eta(u(x,\alpha),v(x))v(x) - \eta(u(x,\alpha),v(x))\big)\phi(x)\mathrm{div}g(x)\\ &+\big(\partial_1\eta(u(x,\alpha),v(x)) + \partial_2\eta(u(x,\alpha),v(x))\big)F(x)\phi(x)\Big)\mathrm{d}\alpha\mathrm{d}x \geq 0.\end{aligned} \tag{43}$$

We now consider, for a given $\varepsilon > 0$, the function $S_\varepsilon \in C^1(\mathbb{R})$ defined by

$$\begin{aligned} &S_\varepsilon(s) = 0 && \forall s \in (-\infty, 0],\\ &S_\varepsilon(s) = s^2(3\varepsilon - 2s)/\varepsilon^3 && \forall s \in [0,\varepsilon],\\ &S_\varepsilon(s) = 1 && \forall s \in [\varepsilon, +\infty).\end{aligned} \tag{44}$$

We define $\xi_\varepsilon(s) = \int_0^s S_\varepsilon(\tau)d\tau$, and we set, for all $(a,b) \in \mathbb{R}^2$, $\eta_\varepsilon(a,b) = \xi_\varepsilon(b-a)$. We then easily get that this function η_ε satisfies $\partial_1\eta_\varepsilon(1,b) = -S_\varepsilon(b-1) = 0 \geq 0$ for all

$b \leq 1$, $\eta_\varepsilon(\cdot, b)$ is convex for all $b \leq 1$, and $\partial_2 \eta_\varepsilon(a,b) = S_\varepsilon(b-a) \geq 0$ for all $(a,b) \in \mathbb{R}^2$. We can then use this function in (44). We remark that, for all $(a,b) \in \mathbb{R}^2$,

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

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

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

Thus, using the Lebesgue dominated convergence theorem, we can let $\varepsilon \to 0$ in (44) which produces

$$\int_\Omega \int_0^1 (v(x) - u(x,\alpha))^+ \left[-g(x) \cdot \nabla \phi(x)\right] \mathrm{d}\alpha \mathrm{d}x \geq 0, \tag{45}$$

which is (27) and thus concludes the proof of the proposition. ∎

The above result is now used to yield the uniqueness of the weak process solution, and thus to obtain that this weak process solution is in fact a weak solution. Note that, in the proof of all the above propositions, the hypothesis that g can be written under the form $g(x) = \Lambda(x)\nabla h(x)$ for all $x \in \Omega$ is not used (g being Lipschitz continuous is sufficient). A uniqueness result for the weak solution could then be obtained assuming that $F > 0$ a.e. in addition to g being Lipschitz continuous, but the uniqueness result for the weak process solution remains an open problem under such hypotheses. The proof of the uniqueness result given below explicitly uses the hypothesis $g(x) = \Lambda(x)\nabla h(x)$ for all $x \in \Omega$, which fortunately holds in the physical problem.

PROPOSITION 3.5 (uniqueness of the weak process solution). *Under Hypotheses* (H), *there exists at most one weak process solution $\hat{g}$ to problem* (10)–(11) *in the sense of Definition* 3.3. *Moreover, if $\hat{u} \in L^\infty(\Omega \times (0,1))$ is such that $0 \leq u(x,\alpha) \leq 1$ and $\hat{g} : (x,\alpha) \mapsto u(x,\alpha)g(x)$ for a.e. $(x,\alpha) \in \Omega \times (0,1)$ and if u satisfies* (27), *then $u(x,\alpha)$ does not depend on α on a.e. $x \in \Omega$ such that $g(x) \neq 0$ ($g(x) = 0$ and $F(x) > 0$). Then the function $\tilde{g}$ defined by $\tilde{g}(x) = u(x,\alpha)g(x)$ for a.e. $x \in \Omega$ and $\alpha \in (0,1)$ is the unique weak solution to problem* (10)–(11) *in the sense of Definition* 1.1. *Moreover, this function $\tilde{g}$ is an extremal point of $\mathcal{C}(g,F)$ in the sense that $|\tilde{g}| \geq |\gamma|$ for all $\gamma \in \mathcal{C}(g,F)$ (the set $\mathcal{C}(g,F)$ is defined in Definition* 3.1*), and it is also the projection in $L^2(\Omega)^d$ of g on the convex set $\mathcal{C}(g,F)$.*

Proof. Let us assume that $\hat{g}$ is a weak process solution to problem (10)–(11) in the sense of Definition 3.3. Let $u \in L^\infty(\Omega \times (0,1))$ correspond to $\hat{g}$ in Definition 3.3. We again denote $\Omega_0 = \{x \in \Omega,\ g(x) = 0\}$ and we remark that (27), proven in Proposition 3.4, gives for all $\gamma \in \mathcal{C}(g,F)$, letting $v \in L^\infty(\Omega)$ be such that $\gamma(x) = v(x)g(x)$ and $0 \leq v(x) \leq 1$ for a.e. $x \in \Omega$, that

$$\int_{\Omega \setminus \Omega_0} \int_0^1 (v(x) - u(x,\alpha))^+ \left[-g(x) \cdot \nabla \varphi(x)\right] \mathrm{d}\alpha \mathrm{d}x \geq 0.$$

Thanks to Hypotheses (H), we can define the nonnegative function φ by $\varphi(x) = h(x) - \min_{y \in \Omega} h(y)$ for all $x \in \Omega$, where $h \in C^1(\overline{\Omega})$ is such that $g(x) = \Lambda(x)\nabla h(x)$ for all

$x \in \Omega$. We then get that, for all $x \in \Omega \setminus \Omega_0$, $-g(x) \cdot \nabla\varphi(x) = -\Lambda(x)\nabla h(x).\nabla h(x) < 0$. This produces

$$(v(x) - u(x,\alpha))^+ = 0 \text{ for a.e. } (x,\alpha) \in \Omega \setminus \Omega_0 \times (0,1). \tag{46}$$

We then remark that the function $\gamma : x \mapsto \int_0^1 u(x,\alpha)\mathrm{d}\alpha g(x)$ belongs to the convex set $\mathcal{C}(g,F)$. Therefore, setting $v = \int_0^1 u(\cdot,\alpha)\mathrm{d}\alpha$ in (46), we get that for a.e. $x \in \Omega\setminus\Omega_0$, $\int_0^1(\int_0^1 u(x,\beta)d\beta - u(x,\alpha))^+\mathrm{d}\alpha = 0$, which proves that $u(x,\alpha)$ does not depend on α for a.e. $x \in \Omega \setminus \Omega_0$. We define $T(u) \in L^\infty(\Omega)$ by $T(u)(x) = u(x,\alpha)$ for a.e. $x \in \Omega \setminus \Omega_0$ and $\alpha \in (0,1)$ and by $T(u)(x) = 1$ for a.e. $x \in \Omega_0$. We then get that the function $\tilde{g} : \Omega \to \mathbb{R}^d$ such that $\tilde{g} = T(u)g$ is such that $\tilde{g}(x) = \hat{g}(x,\alpha)$ for a.e. $x \in \Omega$ and $\alpha \in (0,1)$.

Let us assume that $\hat{g}$ and $\hat{\hat{g}}$ are two weak process solutions to problem (10)–(11) in the sense of Definition 3.3. Let u and $\hat{u}$ be some elements of $L^\infty(\Omega \times (0,1))$ which correspond to $\hat{g}$ and $\hat{\hat{g}}$, respectively, in Definition 3.3. We then get, setting $v = T(\hat{u})$ in (46), that $(T(\hat{u})(x) - T(u)(x))^+ = 0$ for a.e. $x \in \Omega \setminus \Omega_0$ and, inverting the roles of u and $\hat{u}$, $(T(u)(x) - T(\hat{u})(x))^+ = 0$. This suffices to prove that $T(\hat{u})(x) = T(u)(x)$ for a.e. $x \in \Omega \setminus \Omega_0$, which completes the proof of uniqueness of the weak process solution.

Let us prove that the function $\tilde{g} = T(u)g$ is a weak solution to problem (10)-(11) in the sense of Definition 1.1. We introduce in (27) the functions $\xi : s \to (s-1)^2$ and, for all $n \in \mathbb{N}^\star$, $\varphi = \varphi_n$, as defined in the proof of Proposition 1.3. The same analysis as that which is done in the proof of Proposition 1.3 delivers that, passing to the limit $n \to \infty$, $\int_{\Omega_0}\int_0^1 (u(x,\alpha) - 1)\mathrm{d}\alpha F(x)\mathrm{d}x \geq 0$. This proves that $u(x,\alpha) = 1 = u(x)$ for a.e. $\alpha \in (0,1)$ and a.e. $x \in \Omega_0$ such that $F(x) > 0$. Since all the terms of (27) under the symbols $\int$ vanish a.e. on $\{x \in \Omega,\ g(x) = 0 \text{ and } F(x) = 0\}$, we get that (27) with u implies (14) with $T(u)$. Thus the function $\tilde{g}$ is a weak solution to problem (10)–(11) in the sense of Definition 1.1. Since it is obvious that any weak solution is a weak process solution, we thus deduce, from the uniqueness of the weak process solution, that of this weak solution.

Let us now show that $\tilde{g}$ is an extremal point of $\mathcal{C}(g,F)$. Let $\gamma \in \mathcal{C}(g,F)$, and let $v \in L^\infty(\Omega)$ such that $\gamma(x) = v(x)g(x)$ and $0 \leq v(x) \leq 1$ for a.e. $x \in \Omega$. Thanks to (46), we get that, for a.e. $x \in \Omega \setminus \Omega_0$, $v(x) \leq T(u)(x)$. This proves that, for a.e. $x \in \Omega$, $|\gamma(x)| \leq |\tilde{g}(x)|$. This property implies that $\int_\Omega (g(x) - \tilde{g}(x)) \cdot (\tilde{g}(x) - \gamma(x))\mathrm{d}x = \int_\Omega |g(x)|^2(1 - T(u)(x))(T(u)(x) - v(x))\mathrm{d}x \geq 0$ for all $\gamma \in \mathcal{C}(g,F)$, which shows that $\tilde{g}$ is the projection of g on $\mathcal{C}(g,F)$ in $L^2(\Omega)^d$. □

4. Passing to the limit in numerical schemes. We now start the study of the convergence of a numerical scheme, which is based on finite volume methods. Such methods proved their efficiency for various nonlinear problems such as, for instance, nonlinear degenerate problems (see, e.g., [9], [13] and references therein) and nonlinear hyperbolic problems (see [8], but there exists a huge literature on this subject). The main additional difficulty of the present problem is due to the introduction of the limiter $\overline{u}$ in (2) in order to satisfy the constraints (3)–(5) (recall that (2)–(5) lead to (10) and (11) using a time discretization). The "equation" on this unknown $\overline{u}$ seems to lead to a new type of problem which is unexpectedly not really related to variational inequalities but has some similarity with a scalar conservation law, leading to a nonlinear hyperbolic equation. From the numerical point of view, this similarity may be viewed in the upwinding choice for u in (50) (more precisely, the choice of u_K or u_L, on the interface between the control volumes K and L, depends on the sign of $g_{K,L}$). This upwinding is crucial, for instance, in order to have a solution

u taking values in $[0,1]$ (which is a constraint given by (10)). Numerical simulations using a centered choice of u often lead to troubles (such as oscillations) and the simulation has to stop (this is also true in the industrial framework). Another similarity with scalar conservation laws appears in the choice of the convex function ξ in Definition 1.1.

Let us first define the notion of admissible mesh of $\mathbb{R}^d$ (this definition is inspired by [8]).

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 $K \in \mathcal{T}$, and 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:*

$$\mathrm{regul}(\mathcal{T}) := \max\left\{\sum_{L\in\mathcal{N}_K} m_{KL}\mathrm{diam}(K)/m_K,\ K \in \mathcal{T}\right\}.$$

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 numbers such that

$$g_{K,L} = -g_{L,K} \qquad \forall K \in \mathcal{T},\ \forall L \in \mathcal{N}_K \tag{47}$$

and

$$\sum_{L\in\mathcal{N}_K} g_{K,L} = \int_K \mathrm{div} g(x)\mathrm{d}x := G_K \qquad \forall K \in \mathcal{T}. \tag{48}$$

Denoting

$$F_K = \int_K F(x)\mathrm{d}x, \tag{49}$$

the finite volume scheme, in order to approximate problem (10)–(11), is given by

$$\begin{aligned} &\sum_{L\in\mathcal{N}_K} (g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K = 0 \text{ and } u_K \le 1 \text{ or} \\ &\sum_{L\in\mathcal{N}_K} (g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K \ge 0 \text{ and } u_K = 1. \end{aligned} \tag{50}$$

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

$$u_{\mathcal{T}}(x) = u_K \qquad \forall x \in K,\ \forall K \in \mathcal{T}. \tag{51}$$

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($\mathcal{T}$):

$$\text{cons}(g_{\mathcal{T}}) := \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}_K} \frac{\text{diam}(K)}{m_{KL}} \left(g_{K,L} - \bar{g}_{K,L}\right)^2, \tag{52}$$

where

$$\bar{g}_{K,L} = \int_{K|L} g(x)\cdot \mathbf{n}_{K,L} \mathrm{d}s(x) \qquad \forall K\in\mathcal{T},\ \forall L\in\mathcal{N}_K. \tag{53}$$

Different choices are possible for $g_{\mathcal{T}}$. We can propose the following, 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 the knowledge 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, assuming that Λ is scalar (this is the case in some of the geological applications), using for example the finite volume method (see [8]). 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. One then 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 (48) holds under the condition

  $$g_{K,L} = \tau_{KL}(h_L - h_K) \qquad \forall K\in\mathcal{T},\ \forall L\in\mathcal{N}_K \tag{54}$$

  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 [8] and [18]).
- 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 [4]).

In order to compute a solution of (47)–(50), we shall now describe an algorithm, denoted by Algorithm (A) below.

Algorithm (A).

Initialization: $u_K^{(0)} = 1$ and $p_K^{(0)} = 1$ for all $K\in\mathcal{T}$.

Iterations: Let $n\in\mathbb{N}^\star$. Assume that $u_K^{(n-1)}$ and $p_K^{(n-1)}$ are known for all $K\in\mathcal{T}$.

1. Computation of $\{p_K^{(n)},\ K\in\mathcal{T}\}$:

$$\begin{aligned} &\text{If } \sum_{L\in\mathcal{N}_K} (g_{K,L}^+ u_L^{(n-1)} - g_{K,L}^- u_K^{(n-1)}) + F_K < 0, \text{ then } p_K^{(n)} = 0.\\ &\text{If } \sum_{L\in\mathcal{N}_K} (g_{K,L}^+ u_L^{(n-1)} - g_{K,L}^- u_K^{(n-1)}) + F_K \geq 0, \text{ then } p_K^{(n)} = p_K^{(n-1)}. \end{aligned} \tag{55}$$

2. Computation of $\{u_K^{(n)},\ K\in\mathcal{T}\}$, solution to the following linear system:

$$\begin{aligned} \sum_{L\in\mathcal{N}_K} (g_{K,L}^+ u_L^{(n)} - g_{K,L}^- u_K^{(n)}) &= -F_K \qquad \forall K\in\mathcal{T} \text{ s.t. } p_K^{(n)} = 0,\\ u_K^{(n)} &= 1 \qquad \forall K\in\mathcal{T} \text{ s.t. } p_K^{(n)} = 1. \end{aligned} \tag{56}$$

The following proposition gives a monotonicity property of Algorithm (A).

PROPOSITION 4.2 (a monotonicity property of Algorithm (A)). *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 numbers such that* (47) *and* (48) *are satisfied. Let $n\in\mathbb{N}^\star$ be given such that there exists a family $\{(p_K^{(k)},u_K^{(k)}),\ K\in\mathcal{T},\ k=0,\dots,n-1\}$ such that* (55) *and* (56) *hold in addition to $u_K^{(k)}\geq 0$ for all $K\in\mathcal{T}$, $k=0,\dots,n-1$. Let $(p_K^{(n)})_{K\in\mathcal{T}}$ be given by* (55).

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

$$\begin{aligned}\sum_{L\in\mathcal{N}_K}(g_{K,L}^+w_L-g_{K,L}^-w_K)&=-s_K \qquad \forall K\in\mathcal{T}\ \text{s.t. } p_K^{(n)}=0,\\ w_K&=s_K \qquad \forall K\in\mathcal{T}\ \text{s.t. } p_K^{(n)}=1,\end{aligned}\tag{57}$$

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

Let us first remark that Proposition 4.2 suffices to prove that the matrix of the linear system (57) 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 (57), then $(-w_K)_{K\in\mathcal{T}}$ also satisfies (57), which proves that $w_K=0$ for all $K\in\mathcal{T}$. We therefore state the following corollary.

COROLLARY 4.3. *Under the hypotheses of Proposition* 4.2, *for all families $(s_K)_{K\in\mathcal{T}}$ of reals, there exists one and only one family of reals $(w_K)_{K\in\mathcal{T}}$ such that* (57) *holds.*

Proof of Proposition 4.2. Let us assume the hypotheses of Proposition 4.2, 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 (57) 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 $p_K^{(n)}=0$, since $w_K=s_K\geq 0$ for $K\in\mathcal{T}$ such that $p_K^{(n)}=1$. We therefore have

$$\sum_{L\in\mathcal{N}_K}(g_{K,L}^+w_L-g_{K,L}^-w_K)+s_K=0 \qquad \forall K\in\mathcal{T}_-.\tag{58}$$

Summing (58) for $K\in\mathcal{T}_-$ leads to

$$\sum_{K\in\mathcal{T}_-}\sum_{L\in\mathcal{N}_K\setminus\mathcal{T}_-}(g_{K,L}^+w_L-g_{K,L}^-w_K)+\sum_{K\in\mathcal{T}_-}s_K=0.\tag{59}$$

Since $w_K<0$ for $K\in\mathcal{T}_-$ and $w_L\geq 0$ for $L\notin\mathcal{T}_-$, (59) gives $s_K=0$ for all $K\in\mathcal{T}_-$ and $g_{K,L}\geq 0$ for all (K,L) such that $K\in\mathcal{T}_-$ and $L\in\mathcal{N}_K\setminus\mathcal{T}_-$. Let $k<n$ be the greatest integer such that there exists $K\in\mathcal{T}_-$ with $p_K^{(k)}=1$ and $p_K^{(k+1)}=0$ (such a k exists since $p_K^{(0)}=1$ for all $K\in\mathcal{T}$). We then have, for all $K\in\mathcal{T}_-$, $p_K^{(k+1)}=0$ (otherwise this would be in contradiction with the choice of k), and therefore one has $\sum_{L\in\mathcal{N}_K}(g_{K,L}^+u_L^{(k)}-g_{K,L}^-u_K^{(k)})+F_K\leq 0$.

For $K\in\mathcal{T}_-$ such that $p_K^{(k)}=1$ and $p_K^{(k+1)}=0$, one has $\sum_{L\in\mathcal{N}_K}(g_{K,L}^+u_L^{(k)}-g_{K,L}^-u_K^{(k)})+F_K<0$. We thus get

$$\sum_{K\in\mathcal{T}_-}\sum_{L\in\mathcal{N}_K,\,L\notin\mathcal{T}_-}(g_{K,L}^+u_L^{(k)}-g_{K,L}^-u_K^{(k)})+\sum_{K\in\mathcal{T}_-}F_K<0.$$

On the other hand, since $u_L^{(k)}\geq 0$ and since $g_{K,L}\geq 0$ for all (K,L) such that $K\in\mathcal{T}_-$

and $L \in \mathcal{N}_K \setminus \mathcal{T}_-$ and $F_K \geq 0$, we can write

$$\begin{aligned} 0 &\leq \sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K,\, L\notin\mathcal{T}_-} g^+_{K,L} u^{(k)}_L \\ &\leq \sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K,\, L\notin\mathcal{T}_-} (g^+_{K,L} u^{(k)}_L - g^-_{K,L} u^{(k)}_K) + \sum_{K\in\mathcal{T}_-} F_K < 0, \end{aligned}$$

which is impossible. This contradiction proves that $\mathcal{T}_-$ is empty, which concludes the proof of the proposition. ∎

We can now prove the following proposition, which states that Algorithm (A) is well defined and leads to a solution of (50) for some $n \leq \mathrm{card}(\mathcal{T})$.

PROPOSITION 4.4 (convergence of an algorithm and existence of a 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 numbers such that* (47) *and* (48) *are satisfied. Then the following hold.*

1. *There exists a unique family $\{(p^{(n)}_K, u^{(n)}_K),\ K \in \mathcal{T},\ n \in \mathbb{N}\}$ solution of Algorithm* (A).
2. *For all $K \in \mathcal{T}$ and all $n \in \mathbb{N}$, one has $u^{(n)}_K \geq 0$.*
3. *For all $K \in \mathcal{T}$, the sequence $(u^{(n)}_K)_{n\in\mathbb{N}}$ is nonincreasing.*
4. *There exists $n \leq$ card($\mathcal{T}$) such that, setting $u_K = u^{(n)}_K$ for all $K \in \mathcal{T}$, the family $\{u_K,\ K \in \mathcal{T}\}$ is such that $u^{(p)}_K = u_K$ for all $K \in \mathcal{T}$ and $p \geq n$. This family is therefore a solution of* (49) *and* (50) *such that*

$$0 \leq u_K \leq 1 \qquad \forall K \in \mathcal{T}. \tag{60}$$

Proof. The family $\{(p^{(0)}_K, u^{(0)}_K),\ K \in \mathcal{T}\}$ is uniquely defined and satisfies $u^{(0)}_K \geq 0$ for all $K \in \mathcal{T}$.

Let us prove the first two items of the above proposition by induction. Let $n \in \mathbb{N}^\star$; we assume that there exists a family $\{(p^{(k)}_K, u^{(k)}_K),\ K \in \mathcal{T},\ k = 0, \ldots, n-1\}$ such that (55) and (56) hold in addition to $u^{(k)}_K \geq 0$ for all $K \in \mathcal{T}$, $k = 0, \ldots, n-1$. Let $(p^{(n)}_K)_{K\in\mathcal{T}}$ be given by (55). We can then apply Proposition 4.2 and Corollary 4.3, setting $s_K = 1$ for all $K \in \mathcal{T}$ such that $p^{(n)}_K = 1$ and $s_K = F_K \geq 0$ for all $K \in \mathcal{T}$ such that $p^{(n)}_K = 0$. We thus immediately get the existence and the uniqueness of $u^{(n)}_K \geq 0$ for all $K \in \mathcal{T}$ such that (56) holds. This suffices to prove the first two items at the level n.

We can now prove that $u^{(n)}_K \leq u^{(n-1)}_K$ for all $K \in \mathcal{T}$. Indeed let us consider $w_K = u^{(n-1)}_K - u^{(n)}_K$ for all $K \in \mathcal{T}$. We have, for all $K \in \mathcal{T}$ such that $p^{(n)}_K = 0$, $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u^{(n)}_L - g^-_{K,L}u^{(n)}_K) + F_K = 0$ and $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u^{(n-1)}_L - g^-_{K,L}u^{(n-1)}_K) + F_K \leq 0$, which gives, by subtraction

$$\sum_{L\in\mathcal{N}_K} (g^+_{K,L} w_L - g^-_{K,L} w_K) := -s_K,$$

with $s_K \geq 0$. For all $K \in \mathcal{T}$ such that $p^{(n)}_K = 1$, we have

$$w_K = s_K := 0.$$

We can then apply Proposition 4.2, and we get that $0 \leq w_K$ for all $K \in \mathcal{T}$, which is the third item of the proposition. Let us prove the last item.

The definition of the algorithm gives $p_K^{(n)} = p_K^{(n-1)}$ or $p_K^{(n)} = 0$ for all K and all $n \in \mathbb{N}^\star$. Then, setting $A_n = \{K \in \mathcal{T};\ p_K^{(n)} = 0\}$, one has $\mathrm{card}(A_n) \geq \mathrm{card}(A_{n-1})$ for all $n \in \mathbb{N}^\star$. Since $\mathrm{card}(A_0) = 0$, there exists $n \leq \mathrm{card}(\mathcal{T}) + 1$ such that $\mathrm{card}(A_n) = \mathrm{card}(A_{n-1})$. For this value of n one has $p_K^{(n)} = p_K^{(n-1)}$ for all $K \in \mathcal{T}$. If $p_K^{(n-1)} = 1$, one has $u_K^{(n-1)} = 1$ and $\sum_{L\in\mathcal{N}_K}(g_{K,L}^+ u_L^{(n-1)} - g_{K,L}^- u_K^{(n-1)}) + F_K \geq 0$ (since $\sum_{L\in\mathcal{N}_K}(g_{K,L}^+ u_L^{(n-1)} - g_{K,L}^- u_K^{(n-1)}) + F_K < 0$ gives $p_K^{(n)} = 0$).

If $p_K^{(n-1)} = 0$, one has $\sum_{L\in\mathcal{N}_K}(g_{K,L}^+ u_L^{(n-1)} - g_{K,L}^- u_K^{(n-1)}) + F_K = 0$ and $u_K^{(n-1)} \leq 1$ thanks to the fact that the sequence $(u_K^{(n)})_{n\in\mathbb{N}}$ is nonincreasing and $u_K^{(0)} = 1$.

Therefore, setting $u_K = u_K^{(n-1)}$ for all $K \in \mathcal{T}$, the family $\{u_K,\ K \in \mathcal{T}\}$ is a solution of (49) and (50). It is also obvious to see that $u_K^{(p)} = u_K$ for all $K \in \mathcal{T}$ and for all $p \geq n-1$.

This concludes the proof of Proposition 4.4. □

Remark 4.1. Under Hypotheses (H), assuming that Λ is a scalar function and following a method similar to the proof of uniqueness of Proposition 3.5, it is possible to prove that there exists a unique solution to (49) and (50), with the choice (54) for the discrete fluxes.

We then have the following proposition.

PROPOSITION 4.5 (weak bounded variation inequality). *Under Hypotheses* (H), *let $\mathcal{T}$ be an admissible mesh of Ω in the sense of Definition* 4.1, *and let $g_\mathcal{T}$ be a family of reals which satisfies* (47) and (48). *Let $(u_K)_{K\in\mathcal{T}}$ be a solution of scheme* (49) and (50) *such that* (60) *holds. Then there exists $C > 0$, which only depends on d, Ω, g, F and not on $\mathcal{T}$, such that*

$$\sum_{(K,L)\in\mathcal{E}} |g_{K,L}|(u_K - u_L)^2 \leq C. \tag{61}$$

Proof. We multiply (50) by $(1 - u_K)$; we sum on K. We get $T_{17} + T_{18} = 0$ with

$$T_{17} = \sum_{K\in\mathcal{T}} (1 - u_K) \sum_{L\in\mathcal{N}_K} (g_{K,L}^+ u_L - g_{K,L}^- u_K)$$

and

$$T_{18} = \sum_{K\in\mathcal{T}} (1 - u_K) F_K.$$

We have $T_{17} = T_{19} + T_{20}$, with

$$T_{19} = \sum_{K\in\mathcal{T}} (1 - u_K) \sum_{L\in\mathcal{N}_K} g_{K,L}^+ (u_L - u_K)$$

and, using (48),

$$T_{20} = \sum_{K\in\mathcal{T}} (1 - u_K) u_K G_K.$$

We develop T_{19}: we get

$$\begin{aligned} T_{19} = \ & \frac{1}{2} \sum_{K\in\mathcal{T}} (1 - u_K)^2 \sum_{L\in\mathcal{N}_K} g_{K,L}^+ + \frac{1}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}_K} g_{K,L}^+ (u_L - u_K)^2 \\ & - \frac{1}{2} \sum_{K\in\mathcal{T}} (1 - u_L)^2 \sum_{L\in\mathcal{N}_K} g_{K,L}^+. \end{aligned}$$

Since $g^+_{K,L} = g^-_{L,K}$, we get

$$T_{19} = \frac{1}{2}\sum_{K\in\mathcal{T}}(1-u_K)^2 G_K + \frac{1}{2}\sum_{(K,L)\in\mathcal{E}}|g_{K,L}|(u_L-u_K)^2.$$

Gathering the previous results, we get the conclusion. ∎

We can now state the convergence of the scheme to a weak process solution. This convergence result is obtained in the sense of the nonlinear weak-$\star$ convergence, defined in [7], which is a convenient way to understand the convergence towards a Young measure. Indeed, a bounded sequence $(u_n)_{n\in\mathbb{N}}$ of $L^\infty(\Omega)$ converges in the nonlinear weak-$\star$ sense to some function $u \in L^\infty(\Omega\times(0,1))$ if, for all $\xi \in C^0(\mathbb{R})$, the sequence $(\xi(u_n))_{n\in\mathbb{N}}$ converges for the weak-$\star$ topology of $L^\infty(\Omega)$ to the function $x \mapsto \int_0^1 \xi(u(x,\alpha))\mathrm{d}\alpha$ (the notation $\mathrm{d}\alpha$ stands here for the Lebesgue measure on $(0,1)$). A main compactness result is that from a bounded sequence of $L^\infty(\Omega)$, it is possible to extract a subsequence converging in the nonlinear weak-$\star$ sense (see [7] or [8] for more details).

PROPOSITION 4.6 (convergence of the scheme to a weak 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* (47) and (48) *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* $m\in\mathbb{N}$ *(see Definition* 4.1 *for the definitions of* size *and* regul*), and that* $\lim_{m\to\infty}\mathrm{cons}(g_{\mathcal{T}^{(m)}})=0$. *For all* $m\in\mathbb{N}$, *we denote by* $u_{\mathcal{T}^{(m)}}$ *a solution of scheme* (49)–(50) *such that* (60) *holds. 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 sequence* $(u_{\mathcal{T}^{(m)}}g)_{m\in\mathbb{N}}$ *converges in the nonlinear weak-*$\star$ *sense (see above for the sense of this convergence) to a weak process solution of problem* (10)–(11) *in the sense of Definition* 1.1.

Proof. Using the property (60) 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 $u\in L^\infty(\Omega\times(0,1))$. We shall now prove that u is the weak process solution of problem (10)–(11) in the sense of Definition 1.1. Let $\varphi\in C^1(\overline{\Omega},\mathbb{R}_+)$, and let $\xi\in C^1(\mathbb{R})$ be a convex function with $\xi'(1)\ge 0$. Let $m\in\mathbb{N}$, and let $(\mathcal{T}^{(m)})$ be the corresponding admissible mesh of the subsequence. For simplicity, we do not mention the index m until we consider some convergence properties as $m\to\infty$. We get from (50), using $\xi'(u_K)=\xi'(1)+\xi'(u_K)-\xi'(1)$, that

$$\xi'(u_K)\left(\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L - g^-_{K,L}u_K)+F_K\right)\ge 0 \qquad \forall K\in\mathcal{T}. \tag{62}$$

We can then multiply (62) by φ_K, where we denote $\varphi_K = \frac{1}{m_K}\int_K\varphi(x)\mathrm{d}x$, and we sum on $K\in\mathcal{T}$. We get $T_{21}+T_{22}\ge 0$, with

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

and

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

We have $T_{21} = T_{23} + T_{24}$ with

$$T_{23} = \sum_{K\in\mathcal{T}} \xi'(u_K) u_K \varphi_K \sum_{L\in\mathcal{N}_K} g_{K,L}$$

and

$$T_{24} = \sum_{K\in\mathcal{T}} \xi'(u_K) \varphi_K \sum_{L\in\mathcal{N}_K} g^+_{K,L}(u_L - u_K).$$

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

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

On the other hand, we have

$$T_{24} \le T_{25} := \sum_{K\in\mathcal{T}} \varphi_K \sum_{L\in\mathcal{N}_K} g^+_{K,L}(\xi(u_L) - \xi(u_K)).$$

Gathering by edges, we get

$$T_{25} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K))(\varphi_K g^+_{K,L} - \varphi_L g^-_{K,L}).$$

Let us compare T_{25} with T_{26} defined by

$$T_{26} = -\sum_{K\in\mathcal{T}} \xi(u_K) \int_K \mathrm{div}(\varphi(x) g(x)) \mathrm{d}x.$$

We have, on one hand, that

$$\lim_{m\to\infty} T_{26}^{(m)} = -\int_\Omega \int_0^1 \xi(u(x,\alpha)) \mathrm{div}(\varphi(x) g(x)) \mathrm{d}\alpha \mathrm{d}x,$$

and on the other hand, we have

$$T_{26} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K)) \int_{K|L} \varphi(x) g(x) \cdot \mathbf{n}_{K,L} \mathrm{d}s(x).$$

Thus we get that

$$T_{25} - T_{26} = T_{27} + T_{28} + T_{29},$$

with

$$T_{27} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K)) \left(\varphi_K g^+_{K,L} - \varphi_L g^-_{K,L} - \frac{g_{K,L}}{m_{KL}} \int_{K|L} \varphi(x) \mathrm{d}s(x) \right),$$

$$T_{28} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K)) \left(g_{K,L} - \bar{g}_{K,L} \right) \left(\frac{1}{m_{KL}} \int_{K|L} \varphi(x) \mathrm{d}s(x) \right),$$

$$T_{29} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K)) \left(\int_{K|L} (\frac{\bar{g}_{K,L}}{m_{KL}} - g(x) \cdot \mathbf{n}_{K,L}) \varphi(x) \mathrm{d}s(x) \right)$$

(recall that $\bar{g}_{K,L}$ is defined by (53)). In the following, we designate by C_i various real numbers which can depend on $d, \Omega, g, F, \varphi, \xi$ but not on $\mathcal{T}$. Using $|\xi(u_K) - \xi(u_L)| \le$

$C_1\,|u_K - u_L|$ and the Cauchy–Schwarz inequality,

$$\left|\varphi_K - \frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)\right| \le \mathrm{diam}(K)C_2\,,$$

and

$$\left|\varphi_L - \frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)\right| \le \mathrm{diam}(L)C_2\,,$$

we get

$$|T_{27}|^2 \le C_3\left(\sum_{(K,L)\in\mathcal{E}}|g_{K,L}|(u_K-u_L)^2\right)\left(\sum_{(K,L)\in\mathcal{E}}|g_{K,L}|(\mathrm{diam}(K)^2+\mathrm{diam}(L)^2)\right).$$

Using (61) and

$$\sum_{(K,L)\in\mathcal{E}}|g_{K,L}|(\mathrm{diam}(K)^2+\mathrm{diam}(L)^2) \le C_4\,\mathrm{size}(\mathcal{T}),$$

we thus get that

$$\lim_{m\to\infty}|T_{27}^{(m)}| = 0.$$

We now turn to the study of T_{28}. Since we have

$$T_{28} = -\sum_{K\in\mathcal{T}}\xi(u_K)\sum_{L\in\mathcal{N}_K}(g_{K,L}-\bar{g}_{K,L})\left(\frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)\right),$$

we get, using the property (48),

$$T_{28} = -\sum_{K\in\mathcal{T}}\xi(u_K)\sum_{L\in\mathcal{N}_K}(g_{K,L}-\bar{g}_{K,L})\left(\frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)-\varphi_K\right).$$

Thus, thanks to the Cauchy–Schwarz inequality and using (52), we get

$$T_{28}^2 \le C_5\,\mathrm{cons}(g_{\mathcal{T}}).$$

Thus

$$\lim_{m\to\infty}|T_{28}^{(m)}| = 0.$$

We conclude with the study of T_{29}. Since

$$T_{29} = -\sum_{K\in\mathcal{T}}\xi(u_K)\sum_{L\in\mathcal{N}_K}\left(\int_{K|L}\left(\frac{\bar{g}_{K,L}}{m_{KL}}-g(x)\cdot\mathbf{n}_{K,L}\right)(\varphi(x)-\varphi_K)\mathrm{d}s(x)\right)$$

and since $\int_{K|L}(\frac{\bar{g}_{K,L}}{m_{KL}}-g(x)\cdot\mathbf{n}_{K,L})(\varphi(x)-\varphi_K)\mathrm{d}s(x) \le C_6\,m_{KL}\mathrm{diam}(K)^2$, we easily get

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

Gathering these results gives

$$\lim_{m\to\infty} T_{25}^{(m)} = -\int_\Omega \int_0^1 \xi(u(x,\alpha))\mathrm{div}(\varphi(x)g(x))\mathrm{d}\alpha\mathrm{d}x.$$

Finally, we easily get

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

Gathering the previous results, we get $T_{23} + T_{25} + T_{22} \leq 0$. Passing to the limit $m \to \infty$ in this inequality, we get

$$\begin{aligned}
&+\int_\Omega \int_0^1 u(x,\alpha)\xi'(u(x,\alpha))\varphi(x)\mathrm{div}g(x)\mathrm{d}\alpha\mathrm{d}x\\
&-\int_\Omega \int_0^1 \xi(u(x,\alpha))\mathrm{div}(\varphi(x)g(x))\mathrm{d}\alpha\mathrm{d}x\\
&+\int_\Omega \int_0^1 \xi'(u(x,\alpha))\varphi(x)F(x)\mathrm{d}\alpha\mathrm{d}x \geq 0,
\end{aligned}$$

which is exactly Definition 3.3. ∎

Thanks to the uniqueness result, we now classically conclude with the following convergence theorem (similar proofs can be found in [8]).

THEOREM 4.7 (strong convergence of the scheme to a weak solution). *Under Hypotheses* (H), *let* $\mathcal{T}$ *be an admissible mesh of* Ω *in the sense of Definition* 4.1, *and let* $g_\mathcal{T}$ *be a family of reals such that* (47) *and* (48) *are satisfied. Then the function* $u_\mathcal{T} g$, *where* $u_\mathcal{T}$ *is a solution of scheme* (49)–(50) *such that* (60) *holds, converges in* $L^p(\Omega)^d$ *for all* $p \in [1,\infty)$ *to* $\tilde{g}$, *the unique weak solution to problem* (10)–(11) *in the sense of Definition* 1.1, *as* $\mathrm{size}(\mathcal{T})$ *tends to* 0, $\mathrm{cons}(g_\mathcal{T})$ *tends to* 0, *and* $\mathrm{regul}(\mathcal{T})$ *remains bounded (see Definition* 4.1 *for the definitions of* $\mathrm{size}(\mathcal{T})$ *and* $\mathrm{regul}(\mathcal{T})$, *and see* (52) *for the definition of* $\mathrm{cons}(g_\mathcal{T})$*).*

Proof. Under Hypotheses (H), let $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$ be a sequence of admissible meshes of Ω in the sense of Definition 4.1 such that $\lim_{m\to\infty} \mathrm{size}(\mathcal{T}^{(m)}) = 0$. For all $m \in \mathbb{N}$, we denote by $u_{\mathcal{T}^{(m)}}$ a solution of scheme (47)–(50) such that (60) holds. Using Proposition 4.6, 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 sequence $(u_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ converges in the nonlinear weak-$\star$ sense to a weak process solution u of problem (10)–(11) in the sense of Definition 1.1. We then get that the limit of $\int_\Omega g(x)^2(u_{\mathcal{T}^{(m)}}(x) - \int_0^1 u(x,\alpha)\mathrm{d}\alpha)^2\mathrm{d}x$ as $m \to \infty$ is equal to $\int_\Omega g(x)^2(\int_0^1 u(x,\alpha)^2\mathrm{d}\alpha - 2(\int_0^1 u(x,\alpha)\mathrm{d}\alpha)^2 + (\int_0^1 u(x,\alpha)\mathrm{d}\alpha)^2)\mathrm{d}x = 0$, using Proposition 3.5 which stands that $\tilde{g}(x) = u(x,\alpha)g(x)$, for a.e. $x \in \Omega$ and $\alpha \in (0,1)$. This proves that $(u_{\mathcal{T}^{(m)}} g)_{m\in\mathbb{N}}$ converges to $\tilde{g}$ in $L^2(\Omega)^d$. The uniqueness of $\tilde{g}$ gives the conclusion of the theorem. ∎

5. Numerical results.

5.1. One-dimensional example. We again consider the following data, studied in section 2: $\Omega = (-1,1)$, $g : x \mapsto x^3 - x$, and $F : x \mapsto 1/2$. We recall that the weak solution is the function $\tilde{g}$ given by $\tilde{g} = ug$, where the function u is such that $u : x \mapsto 1$ for all $x \in (-1,-\sqrt{1/2}) \cup (\sqrt{1/2},1)$ and $u : x \mapsto 1/(2(1-x^2))$ for all $x \in (-\sqrt{1/2},\sqrt{1/2})$. We use Algorithm (A) to solve the nonlinear system (49)–(50)

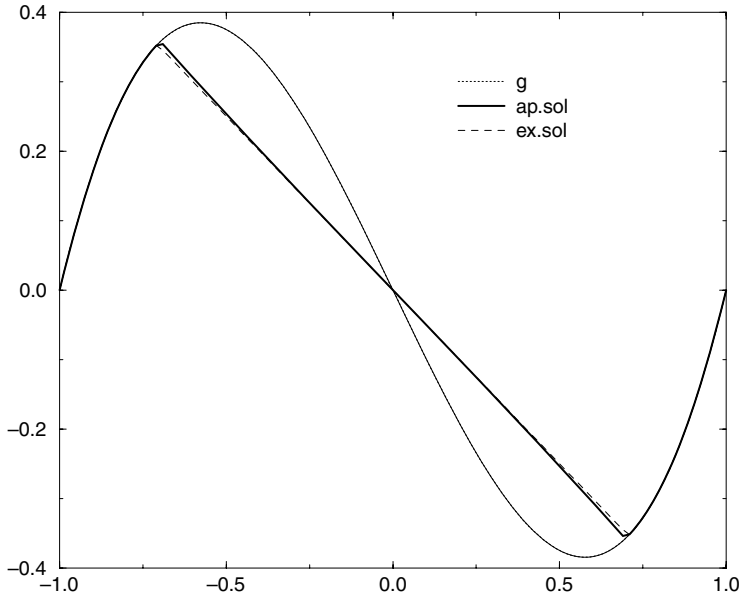

FIG. 1. *Approximate solution (ap.sol) and exact solution (ex.sol) with* 100 *control volumes.*

with $g_{K,L} = \bar{g}_{K,L}$ ($\bar{g}_{K,L}$ is defined in (53)). We get, with 100 uniform control volumes, the results given in Figure 1. The exact solution $\tilde{g}$ is represented by the dashed line (and denoted by "ex.sol." in the legend). The approximate solution of (49)–(50) is $u_{\mathcal{T}}$. Figure 1 gives, with the solid line, the product of $u_{\mathcal{T}}$ with the exact function g (and this product is denoted by "ap.sol." in the legend). The dashed line and the solid line are very close to one another. The last line, namely the grey dotted one, represents the exact function g.

It is interesting to remark that Algorithm (A) converges for a significantly smaller number of iterations than card($\mathcal{T}$). The table below gives, for different numbers of control volumes, the number of iterations until $p_K^{(n)} = p_K^{(n+1)}$ for all $K \in \mathcal{T}$.

Number of control volumes	Number of iterations	$\|\tilde{g} - u_{\mathcal{T}} g\|_{L^1(\Omega)}$
10	3	0.031757
50	9	0.006969
100	17	0.003488
500	76	0.000699
1000	151	0.000348
5000	748	0.000070
10000	1496	0.000035
50000	7473	0.000007

We observe that this number behaves as 1/size($\mathcal{T}$), whereas the error in $L^1(\Omega)$ behaves as size($\mathcal{T}$).

5.2. Two-dimensional examples. We use the coupled finite volume scheme (48)–(54) in order to compute $g_{\mathcal{T}}$. We consider the following data: $\Omega = (0,1)^2$, $\Lambda(x) = \mathrm{I}_d$, and $F(x) = 1/100$ for a.e. $x \in \Omega$, $g = \nabla h$, where h is a solution of

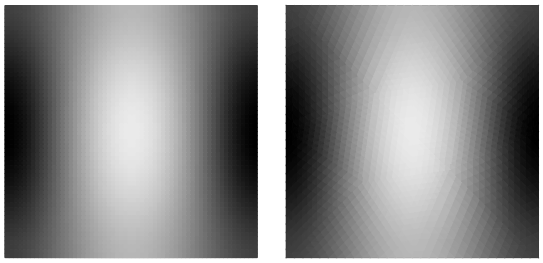

FIG. 2. *Value of h from* 0 *(black) to* 0.00111 *(white): rectangular* 60×60 *mesh (left) and triangular mesh with* 3650 *triangles (right).*

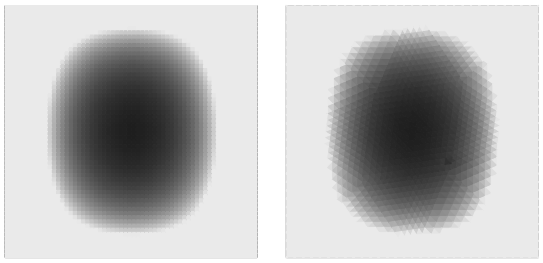

FIG. 3. *Value of u from* 0.48 *(black) to* 1 *(white): rectangular* 60×60 *mesh (left) and triangular mesh with* 3650 *triangles (right).*

the homogeneous Neumann problem

$$-\Delta h(x,y) = y(1-y)(-x^2+x-1/6) \qquad \forall (x,y) \in (0,1)^2,$$

$$\nabla h \cdot \mathbf{n} = 0 \text{ on } \partial\Omega.$$

These data have been chosen since they represent a kind of generalization in two dimensions of the one-dimensional case presented above. Two meshes have been tested. With a rectangular 60×60 mesh, the convergence of Algorithm (A) is obtained after 10 iterations; with a triangular mesh with 3650 triangles, 14 iterations are necessary to converge. The results obtained after the resolution of h by the finite volume method are presented in Figure 2. The corresponding values of the function u such that ug is the weak solution are given in Figure 3, and the values of g_x, g_y, $\tilde{g}_x$, $\tilde{g}_y$ which are the components of g and $\tilde{g}$ are given in Figures 4 and 5 for the rectangular mesh.

These results show the efficiency of the numerical method. In particular, we can remark that the approximate solution obtained with the rectangular mesh is very close to the approximate solution obtained with the triangular mesh.

The following table gives, for rectangular meshes, the number of iterations needed by Algorithm (A) for convergence.

Number of control volumes	Number of iterations
10×10	3
50×50	9
100×100	16
150×150	23
200×200	30

We again observe that this number behaves as $1/\text{size}(\mathcal{T})$.

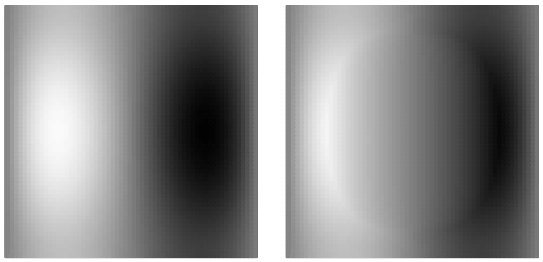

FIG. 4. *Value of g_x (left) and of $\tilde{g}_x$ (right) from* -0.00342 *(black) to* 0.00342 *(white).*

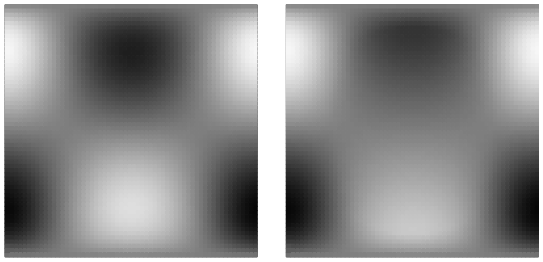

FIG. 5. *Value of g_y (left) and of $\tilde{g}_y$ (right) from* -0.00094 *(black) to* 0.00094 *(white).*

6. Conclusions. We have been able to prove the existence and the uniqueness of the weak solution to problem (10)–(11) in the sense of Definition 1.1, and we have proved the convergence of a numerical scheme, under Hypotheses (H). At this time, we have not yet derived an error estimate although we can guess that it will be possible to follow the same steps as that of a scalar nonlinear hyperbolic problem, since the basis of proof of the uniqueness theorem is the doubling variable technique of Krushkov. It is, however, probable that the error estimate that we shall obtain will be not sharp. Moreover, the mathematical problem is not directly formulated as a function of h but on g. We have only briefly mentioned in remarks that some of the results of this paper can be obtained without the assumption $g = \Lambda \nabla h$. However, this is not the case for all of them. Finally, much work remains to be done in order to handle the complete problem (2)–(5).

REFERENCES

[1] R. S. ANDERSON AND N. F. HUMPHREY, *Interaction of weathering and transport processes in the evolution of arid landscapes*, in Quantitative Dynamics Stratigraphy, T. A. Cross, ed., Prentice–Hall, Englewood Cliffs, NJ, 1989, pp. 349–361.

[2] S. N. ANTONTSEV, G. GAGNEUX, AND G. VALLET, *On some stratigraphic control problems*, J. Appl. Mech. Tech. Phys., 44 (2003), pp. 821–828.

[3] R. DIPERNA, *Measure-valued solutions to conservation laws*, Arch. Ration. Mech. Anal., 88 (1985), pp. 223–270.

[4] J. DRONIOU, R. EYMARD, D. HILHORST, AND X. D. ZHOU, *Convergence of a finite-volume mixed finite-element method for a system of an elliptic-hyperbolic system*, IMA J. Numer. Anal., 23 (2003), pp. 507–538.

[5] R. EYMARD, T. GALLOUËT, V. GERVAIS, AND R. MASSON, *Convergence of a numerical scheme for stratigraphic modeling*, SIAM J. Numer. Anal., 43 (2005), pp. 474–501.

[6] 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 (2004), pp. 527–548.

[7] R. Eymard, T. Gallouët, and R. Herbin, *Existence and uniqueness of the entropy solution to a nonlinear hyperbolic equation*, Chinese Ann. Math. Ser. B, 16 (1995), pp. 1–14.
[8] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume method*, in Handbook of Numerical Analysis, Vol. VII, North-Holland, Amsterdam, 2000, pp. 715–1022.
[9] R. Eymard, T. Gallouët, R. Herbin, and A. Michel, *Convergence of a finite volume scheme for nonlinear degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 41–82.
[10] G. Gagneux and G. Vallet, *Sur des problèmes d'asservissements stratigraphiques*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 715–739.
[11] D. Granjeon, P. Joseph, and B. Doligez, *Using a 3-D stratigraphic model to optimize reservoir description*, Hart's Petroleum Eng. Internat., November (1998), pp. 51–58.
[12] S. N. Krushkov, *First order quasilinear equations with several space variables*, Mat. Sb., 10 (1970), pp. 217–243 (in Russian).
[13] A. Michel, *A finite volume scheme for two-phase immiscible flow in porous media*, SIAM J. Numer. Anal., 41 (2003), pp. 1301–1317.
[14] P. Pedregal, *Optimization, relaxation and Young measures*, Bull. Amer. Math. Soc. (N.S.), 36 (1999), pp. 27–58.
[15] J. C. Rivenaes, *Impact of sediment transport efficiency on large scale sequence architecture: Results from stratigraphic computer simulation*, Basin Res., 4 (1992), pp. 133–146.
[16] T. Roubiček, *Nonlinear Partial Differential Equations with Applications*, Internat. Ser. Numer. Math., Birkhäuser, Basel, Boston, Berlin, 2005.
[17] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.
[18] 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 (1996), pp. 841–872.