Convergence and a posteriori error analysis for energy-stable finite element approximations of degenerate parabolic equations

Clément Cancès, Flore Nabet, Martin Vohralík

CONVERGENCE AND A POSTERIORI ERROR ANALYSIS FOR ENERGY-STABLE FINITE ELEMENT APPROXIMATIONS OF DEGENERATE PARABOLIC EQUATIONS[*]

CLÉMENT CANCÈS[†], FLORE NABET[‡], AND MARTIN VOHRALÍK[§]

Abstract. We propose a finite element scheme for numerical approximation of degenerate parabolic problems in the form of a nonlinear anisotropic Fokker–Planck equation. The scheme is energy-stable, only involves physically motivated quantities in its definition, and is able to handle general unstructured grids. Its convergence is rigorously proven thanks to compactness arguments, under very general assumptions. Although the scheme is based on Lagrange finite elements of degree 1, it is locally conservative after a local postprocess giving rise to an equilibrated flux. This also allows to derive a guaranteed a posteriori error estimate for the approximate solution. Numerical experiments are presented in order to give evidence of a very good behavior of the proposed scheme in various situations involving strong anisotropy and drift terms.

Key words. degenerate parabolic equation, Fokker–Planck equation, energy-stable discretization, local conservation, equilibrated flux, convergence, a posteriori error estimate.

AMS subject classifications. 65M12, 35K65, 65M15, 65M60

1. Introduction.

1.1. Presentation of the problem.

Degenerate parabolic equations appear in many applications from several fields like biology [48], material sciences [42], or porous media flows [5, 50]. In the context of complex porous media flows arising, for instance, in oil engineering, carbon dioxide sequestration, or nuclear waste repositories management, degenerate parabolic problems may moreover be highly heterogeneous and highly anisotropic. This altogether leads to challenging numerical issues that have been addressed in numerous papers (see, e.g., [41, 20, 45, 4, 49, 46, 43, 33, 34, 30, 9, 10, 31, 23]). Since it appears to be the cornerstone of many complex models, we focus on the following nonlinear anisotropic Fokker–Planck equation:

$$
\begin{cases}
\partial_t u - \boldsymbol{\nabla}\cdot(\eta(u)\boldsymbol{\Lambda}\boldsymbol{\nabla}(p(u)+\Psi)) = f(u) & \text{in } (0,t_{\mathrm{f}})\times\Omega =: Q_{t_{\mathrm{f}}},\\
\eta(u)\boldsymbol{\Lambda}\boldsymbol{\nabla}(p(u)+\Psi)\cdot\boldsymbol{n} = 0 & \text{on } (0,t_{\mathrm{f}})\times\Sigma_{\mathrm{N}},\\
p(u) = p_{\mathrm{D}} & \text{on } (0,t_{\mathrm{f}})\times\Sigma_{\mathrm{D}},\\
u_{|_{t=0}} = u_0 & \text{in } \Omega.
\end{cases} \tag{1.1}
$$

In the above problem, Ω is supposed to be a polyhedral connected and bounded open subset of $\mathbb{R}^d$, $d = 2, 3$, with a Lipschitz-continuous boundary $\partial\Omega$ and $\boldsymbol{n}$ is the unit normal vector to $\partial\Omega$ outward to Ω. This is split into two parts Σ_{N} and Σ_{D} on which no-flux and Dirichlet boundary conditions are respectively imposed. In the sequel, we suppose that Σ_{D} has a strictly positive $(d-1)$-dimensional Lebesgue measure. Note

[*] This work was supported by the French National Research Agency ANR (project GeoPor, grant ANR-13-JS01-0007-01). The third author has also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647134 GATIPOR).

[†] Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille (clement.cances@inria.fr).

[‡] CMAP, Ecole polytechnique, CNRS, Université Paris-Saclay, 91128, Palaiseau, France (flore.nabet@polytechnique.edu).

[§] Inria, 2 rue Simone Iff, 75589 Paris, France & Université Paris-Est, CERMICS (ENPC), 77455 Marne-la-Vallée, France (martin.vohralik@inria.fr).

that the analysis can still be carried out in the case where $\Sigma_{\mathrm{D}} = \emptyset$ by following the path proposed in [15]. We denote by $t_{\mathrm{f}} > 0$ an arbitrary finite time horizon. The mobility function $\eta : \mathbb{R}_+ \to \mathbb{R}_+$ is supposed to be continuous, non-decreasing, and to satisfy

$$\eta(0) = 0 \qquad \text{and} \qquad \eta(s) > 0, \quad \forall s > 0.$$

Additional assumptions on the behavior of $\eta(s)$ for large s will be stated later on, see (1.7), (1.17), and (1.18). It is extended to the whole $\mathbb{R}$ into an even function. The permeability tensor field $\boldsymbol{\Lambda} : \Omega \to \left[L^\infty(\mathbb{R})\right]^{d\times d}$ is supposed to be symmetric, uniformly elliptic, and bounded: there exists $\lambda_\star, \lambda^\star > 0$ such that

$$\lambda_\star |\boldsymbol{v}|^2 \le \boldsymbol{\Lambda}(\boldsymbol{x})\boldsymbol{v}\cdot\boldsymbol{v} \le \lambda^\star |\boldsymbol{v}|^2, \qquad \forall \boldsymbol{v} \in \mathbb{R}^d, \text{ for a.e. } \boldsymbol{x} \in \Omega. \tag{1.2}$$

The pressure function p is supposed to be absolutely continuous (i.e. $p' \in L^1_{\mathrm{loc}}((0,\infty))$) and increasing on $(0,+\infty)$, and to satisfy

$$\lim_{u\to+\infty} p(u) = +\infty. \tag{1.3}$$

Concerning its behavior near 0, either $\lim_{u\searrow 0} p(u) = -\infty$, or $p(0)$ is finite. In the latter case, the function p is extended on the whole $\mathbb{R}$ into

$$p(u) := 2p(0) - p(-u), \qquad \forall u < 0.$$

We denote by

$$I_p := \begin{cases} (0,+\infty) & \text{if } p(0) = -\infty, \\ \mathbb{R} & \text{otherwise,} \end{cases}$$

and by $\overline{I}_p$ its closure in $\mathbb{R}$. The function p is assumed to belong to $L^1_{\mathrm{loc}}(\overline{I}_p)$, and we can define its inverse $p^{-1} : \mathbb{R} \to I_p$. We additionally assume that $\sqrt{\eta}p' \in L^1_{\mathrm{loc}}(\overline{I}_p)$ is integrable near 0. We define energy density function E by

$$E(u) := \int_1^u (p(a) - p(1))\mathrm{d}a \ge 0, \qquad \forall u \in \overline{I}_p. \tag{1.4}$$

In the case where $p(0)$ is finite, simple calculations allow to check that

$$E(u) = E(|u|) + (p(1) - p(0))(|u| - u) \ge E(|u|), \qquad \forall u \in \mathbb{R}. \tag{1.5}$$

Since p is strictly increasing with $\lim_{u\to+\infty} p(u) = +\infty$, the function E is strictly convex and satisfies

$$\lim_{u\to+\infty} \frac{E(u)}{u} = +\infty. \tag{1.6}$$

We assume moreover that

$$\lim_{u\to+\infty} \frac{E(u)}{p(u)} = +\infty, \qquad \lim_{u\to+\infty} \frac{E(u)}{\eta(u)} = +\infty. \tag{1.7}$$

The first condition in (1.7) means that p has mild-enough growth at infinity. We will use the following inequality proved in [15, Lemmas 3.2 and 3.3]: there exists C_ε depending only on ε and on the nonlinearities p and η such that

$$|u| \le \varepsilon E(u) + C_\varepsilon, \quad p(u) \le \varepsilon E(u) + C_\varepsilon, \quad \text{and } \eta(u) \le \varepsilon E(u) + C_\varepsilon, \quad \forall u \in \overline{I}_p. \tag{1.8}$$

The external potential $\Psi \in W^{1,\infty}(\Omega)$ is supposed to be Lipschitz-continuous on $\overline{\Omega}$. We assume that the Dirichlet condition p_{D} prescribed on $(0,t_{\mathrm{f}}) \times \Sigma_{\mathrm{D}}$ can be extended into $Q_{t_{\mathrm{f}}}$ to a time-and-space Lipschitz-continuous function (still denoted by p_{D}). More precisely, we require that

(1.9)
$$p_{\mathrm{D}} \in W^{1,\infty}(Q_{t_{\mathrm{f}}}), \quad u_{\mathrm{D}} := p^{-1}(p_{\mathrm{D}}) \geq 0 \text{ a.e. in } \Omega \quad \text{and} \quad \|E(u_{\mathrm{D}})\|_{L^\infty(Q_{t_{\mathrm{f}}})} < +\infty.$$

Concerning the source term $f : \mathbb{R} \times (0,t_{\mathrm{f}}) \times \Omega \to \mathbb{R}$, we assume that there exist two non-negative functions f_{inj} and f_{out} belonging to $L^\infty(Q_{t_{\mathrm{f}}})$ such that

$$f(u;t,\boldsymbol{x}) = f_{\mathrm{inj}}(t,\boldsymbol{x}) - \eta(u^+) f_{\mathrm{out}}(t,\boldsymbol{x}), \qquad \forall u \in \overline{I}_p, \text{ for a.e. } (t,\boldsymbol{x}) \in Q_{t_{\mathrm{f}}}. \tag{1.10}$$

Here, $u^+ = \max(0,u)$ denotes the positive part of u. Note that $u \mapsto f(u;t,\boldsymbol{x})$ is non-increasing and that $f(u;t,\boldsymbol{x}) \geq 0$ for all $u \leq 0$.

The initial data u_0 is supposed to satisfy

(1.11)
$$u_0(\boldsymbol{x}) \geq 0 \text{ for a.e. } \boldsymbol{x} \in \Omega, \quad u_0 = u_{\mathrm{D}}(0,\cdot) \text{ a.e. on } \Sigma_{\mathrm{D}}, \quad \text{and} \quad \int_\Omega E(u_0)\mathrm{d}\boldsymbol{x} < \infty.$$

1.2. Energy estimate.

In this section, we briefly present the energy estimate that we aim to preserve at the discrete level.

We remain sloppy concerning regularity issues since the calculations presented here are formal and only aim at motivating our approach at the numerical level. The calculations can however be rigorously justified if one regularizes the problem and if one lets tend the regularization parameter to 0. Another way to "regularize" the problem to make our discussion rigorous is to take a finite dimensional Galerkin approximation of the problem, what we will typically do later on in this paper. But for the moment, suppose that all the functions are smooth enough to justify the following calculations.

Multiply the first equation of (1.1) by $p(u) - p_D$ and integrate w.r.t. space over Ω. Then one gets

$$\int_\Omega \partial_t u (p(u) - p_D) + \int_\Omega \eta(u) \boldsymbol{\Lambda} \boldsymbol{\nabla}(p(u) + \Psi) \cdot \boldsymbol{\nabla}(p(u) - p_D) = \int_\Omega f(u)(p(u) - p_D).$$

Using the definition (1.4) of $E(u)$, one gets that

$$\int_\Omega \partial_t u (p(u) - p(1)) = \int_\Omega \partial_t E(u) = \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega E(u),$$

whereas

$$\int_\Omega \partial_t u (p(1) - p_D) = \int_\Omega \partial_t \{u(p(1) - p_D)\} + \int_\Omega u \partial_t p_D.$$

Defining

$$I(u) = u(p(1) - p_D), \tag{1.12}$$

we obtain that

$$\begin{aligned}\int_\Omega \partial_t u (p(u) - p_D) =& \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \{E(u) + I(u)\} + \int_\Omega u \partial_t p_D \\ \geq& \frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega \{E(u) + I(u)\} - \|\partial_t p_D\|_{L^\infty(Q_{t_{\mathrm{f}}})} \int_\Omega |u|.\end{aligned}$$

On the other hand, Young inequality and assumption (1.2) yield

$$\begin{aligned}
&\int_\Omega \eta(u)\mathbf{\Lambda}\boldsymbol{\nabla}(p(u)+\Psi)\cdot\boldsymbol{\nabla}(p(u)-p_D)\\
&\quad= \int_\Omega \eta(u)\mathbf{\Lambda}\boldsymbol{\nabla}(p(u)+\Psi)\cdot\boldsymbol{\nabla}(p(u)+\Psi) - \int_\Omega \eta(u)\mathbf{\Lambda}\boldsymbol{\nabla}(p(u)+\Psi)\cdot\boldsymbol{\nabla}(p_D+\Psi)\\
&\quad\geq \frac{1}{2}\int_\Omega \eta(u)\mathbf{\Lambda}\boldsymbol{\nabla}(p(u)+\Psi)\cdot\boldsymbol{\nabla}(p(u)+\Psi) - \frac{1}{2}\int_\Omega \eta(u)\mathbf{\Lambda}\boldsymbol{\nabla}(p_D+\Psi)\cdot\boldsymbol{\nabla}(p_D+\Psi)\\
&\quad\geq \frac{1}{2}\int_\Omega \eta(u)\mathbf{\Lambda}\boldsymbol{\nabla}(p(u)+\Psi)\cdot\boldsymbol{\nabla}(p(u)+\Psi) - \frac{\lambda^\star}{2}\left\|\boldsymbol{\nabla}(p_D+\Psi)\right\|^2_{L^\infty(Q_{t_\mathrm{f}})}\int_\Omega \eta(u).
\end{aligned}$$

The assumptions we made on the nonlinearities ensure that

$$\begin{aligned}
&|I(u)| \leq C(1+E(u)), &&|u| \leq C(1+E(u)+I(u)),\\
&\eta(u) \leq C(1+E(u)+I(u)), \quad\text{and} &&f(u)(p(u)-p_D) \leq C(1+E(u)+I(u)).
\end{aligned}$$

Therefore, we obtain that

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega \{E(u)+I(u)\} + \frac{\lambda_\star}{2}\int_\Omega \eta(u)\left|\boldsymbol{\nabla}(p(u)+\Psi)\right|^2 \leq C\left(1+\int_\Omega \{E(u)+I(u)\}\right). \tag{1.13}$$

We infer from Gronwall Lemma, from the Lipschitz continuity of Ψ, and from the inequality

$$\eta(u) \leq C(1+E(u))$$

that

$$\sup_{t\in[0,t_\mathrm{f}]}\int_\Omega E(u) \leq C \quad\text{and}\quad \iint_{Q_{t_\mathrm{f}}} \eta(u)\left|\boldsymbol{\nabla}p(u)\right|^2 \leq C. \tag{1.14}$$

In many interesting contexts, as for example for porous media flows, estimate (1.13) and its consequence (1.14) have a strong meaning since they encode the stability of the system in terms of physically motivated quantities. And as it will appear later on, these estimates are sufficient to give a mathematical sense to the notion of solution.

1.3. Weak solutions and well-posedness of the continuous problem. In order to define properly the notion of weak solution to the problem, we introduce the so-called Kirchhoff transform γ and semi-Kirchhoff transform ξ defined respectively by

$$\gamma(u) := \int_0^u \eta(a)p'(a)\mathrm{d}a, \qquad \xi(u) := \int_0^u \sqrt{\eta(a)}p'(a)\mathrm{d}a, \qquad \forall u \in \overline{I}_p. \tag{1.15}$$

The introduction of ξ is motivated by the relation $\eta(u)|\boldsymbol{\nabla}p(u)|^2 = |\boldsymbol{\nabla}\xi(u)|^2$, while

$$\eta(u)\boldsymbol{\nabla}p(u) = \boldsymbol{\nabla}\gamma(u),$$

as soon as $p(u)$ is regular enough to justify the calculations. Moreover, one can check that the following chain-rule property holds: if u is such that $\eta(u) \in L^\infty((0,t_\mathrm{f});L^1(\Omega))$ and $\boldsymbol{\nabla}\xi(u) \in [L^2(Q_{t_\mathrm{f}})]^d$, then

$$\sqrt{\eta(u)}\boldsymbol{\nabla}\xi(u) = \boldsymbol{\nabla}\gamma(u) \in L^2((0,t_\mathrm{f});L^1(\Omega)). \tag{1.16}$$

For technical reasons that will appear later on, we assume that there exists $\alpha > 0$ such that

$$1 + |\xi(u)| \geq \alpha\sqrt{|u|}, \qquad \forall u \in \overline{I}_p, \tag{1.17}$$

and we also assume that

$$\sqrt{\eta \circ \xi^{-1}} \text{ is uniformly continuous on } \xi(\overline{I}_p). \tag{1.18}$$

This last assumption ensures the existence of a nondecreasing continuous modulus of continuity $\varpi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\varpi(0) = 0$ such that

$$\left|\sqrt{\eta \circ \xi^{-1}}(x) - \sqrt{\eta \circ \xi^{-1}}(y)\right| \leq \varpi(|x-y|), \qquad \forall x, y \in \xi(\overline{I}_p).$$

Our assumptions are for instance fulfilled in the case where $\eta(u) = u$ and $p(u) = \log(u)$, for which $\xi(u) = 2\sqrt{u}$, corresponding to a linear convection diffusion equation, but also in the case where $\eta(u) = u$ and $p(u) = \dfrac{m}{m-1}u^{m-1}$, $m > 1$, for which $\xi(u) = \dfrac{2m}{2m-1}|u|^{m-3/2}u$, corresponding to the case of the porous medium equation. Note however that (1.17)–(1.18) do not longer hold in the fast diffusion case $m < 1$ for this choice of nonlinearities η, p.

Although the physical meaning of the Kirchhoff transforms is often unclear, their introduction is needed to give a proper sense to the solutions of (1.1).

Definition 1 (Weak solution). *A measurable function u is said to be a weak solution to the problem (1.1) if u and $\eta(u)$ belong to $L^\infty((0,t_{\mathrm{f}}); L^1(\Omega))$, if $\xi(u)$ belongs to $L^2((0,t_{\mathrm{f}}); H^1(\Omega))$ with $\xi(u) = \xi(u_{\mathrm{D}})$ a.e. on $(0,t_{\mathrm{f}}) \times \Sigma_{\mathrm{D}}$, and if*

$$\iint_{Q_{t_{\mathrm{f}}}} u\partial_t\varphi + \int_\Omega u_0\varphi(0,\cdot) - \iint_{Q_{t_{\mathrm{f}}}} \left(\boldsymbol{\nabla}\gamma(u) + \eta(u)\boldsymbol{\nabla}\Psi\right)\cdot\boldsymbol{\Lambda}\boldsymbol{\nabla}\varphi + \iint_{Q_{t_{\mathrm{f}}}} f(u)\varphi = 0, \tag{1.19}$$

for all $\varphi \in C_{\mathrm{c}}^\infty([0,t_{\mathrm{f}}); \overline{\Omega})$ with $\varphi(t,\boldsymbol{x}) = 0$ for all $(t,\boldsymbol{x}) \in [0,t_{\mathrm{f}}) \times \Sigma_{\mathrm{D}}$. Note that from (1.16), one then has $\gamma(u) \in L^2((0,t_{\mathrm{f}}); W^{1,1}(\Omega))$.

The weak formulation (1.19) is satisfied for test functions $\varphi \in C_{\mathrm{c}}^\infty([0,t_{\mathrm{f}}); \overline{\Omega})$ with $\varphi = 0$ on $[0,t_{\mathrm{f}}) \times \Sigma_{\mathrm{D}}$. Thanks to a straightforward density argument, this can be extended to merely C^1 test functions. Since u and the flux $\boldsymbol{\Lambda}\left(\boldsymbol{\nabla}\gamma(u) + \eta(u)\boldsymbol{\nabla}\Psi\right)$ are in L^1, Egoroff's theorem (see, e.g., [51, Chapter XI.3]) implies that one can still extend the weak formulation to test functions in the Banach space

$$X := \Big\{\varphi \in C([0,t_{\mathrm{f}}]; \overline{\Omega}) \,\Big|\, \varphi(t_{\mathrm{f}},\cdot) = 0, \varphi_{|_{[0,t_{\mathrm{f}}]\times\Sigma_{\mathrm{D}}}} = 0, \\ \partial_t\varphi \in L^1((0,t_{\mathrm{f}}); L^\infty(\Omega)), \boldsymbol{\nabla}\varphi \in [L^\infty(Q_{t_{\mathrm{f}}})]^d\Big\}. \tag{1.20}$$

As a consequence of the convergence result for our numerical scheme (cf. Theorem 2.4 below), there exists (at least) one weak solution to the continuous problem (1.1) in the sense of the above definition. The question of the uniqueness of the weak solution is more intricate and still open in general up to our knowledge. Let us just notice that if $\eta(u)$ belongs to $L^\infty(Q_{t_{\mathrm{f}}})$, then the flux $\boldsymbol{\nabla}\gamma(u) + \eta(u)\boldsymbol{\nabla}\Psi$ belongs to $[L^2(Q_{t_{\mathrm{f}}})]^d$. This allows to use Otto's uniqueness result [47] provided $\eta \circ \gamma^{-1} \in C^{0,1/2}$.

Remark 1.1 (Degenerate elliptic-parabolic problems). *The extension to the case of degenerate elliptic-parabolic problems where p is no longer a function but a maximal monotone graph allowed to be multivalued can be easily performed as soon as η is bounded and $p(0)$ is either infinite or single-valued. This allows in particular to treat the case of the so-called Richards equation (see e.g. [32, 1]). In order to keep the presentation reasonably simple, we avoid this difficulty here.*

Remark 1.2 (Time dependent external potential). *The aimed application when we designed our scheme was complex porous media flows, where the external potential $\Psi(\boldsymbol{x})$ is typically the gravitational potential and is constant along time as in our presentation. In some other settings, one can be interested in considering a time dependent external potential $\Psi(t, \boldsymbol{x})$, that could for instance represent an electrostatic potential. Our purpose can easily be extended to this framework provided $\Psi \in W^{1,\infty}(Q_{t_\mathrm{f}})$ is also Lipschitz continuous w.r.t. time. Here again, this difficulty is avoided in order to keep the presentation as simple as possible.*

1.4. Goal and positioning of the paper. In this paper, we propose an extensive numerical analysis—stability, existence of solutions for a fixed mesh, convergence as the discretization parameters tend to 0, and a posteriori error analysis including distinction of error components and design of adaptive stopping criteria for the iterative linearization and algebraic resolution—for a numerical scheme designed to approximate the solutions of (1.1).

It is now well understood that preserving at the discrete level the energy stability is of great importance for the accuracy in the long-time regime [18, 19, 35, 36, 6, 37, 2] or in some other asymptotic regime [7]. All these works are based on finite volumes with Two-Point Flux Approximation (TPFA) [40, 29], and fail to extend to the anisotropic setting. In this paper, we provide a scheme for which the calculations of Section 1.2 can be transposed to the discrete setting, i.e., our scheme is energy-stable, and that allows to go beyond the monotone setting of TPFA finite volume schemes. It is of a finite element form, in contrast to the so-called nonlinear VAG scheme proposed in [15] or the energy-diminishing DDFV scheme [12, 13]. This leads to a simple writing and implementation compared to [15, 12, 13], whilst preserving the crucial features, namely:

- strong theoretical foundations based on rigorously-proved theorems are provided;
- the discrete counterpart of the physically-motivated energy $\displaystyle\int_\Omega E(u)$ is controlled (the scheme is energy-stable);
- the scheme does not involve the Kirchhoff transforms ξ and γ;
- the scheme is locally conservative after a local postprocess;
- the scheme numerically appears to be second-order accurate, in opposition to the upstream mobility scheme [38, 14, 1] that is merely of order 1. Moreover, the scheme appears to be extremely robust w.r.t. the anisotropy ratio, in opposition to those proposed in [38, 14, 1];
- the scheme handles general unstructured grids;
- the computational cost at fixed grid is affordable and the convergence of the Newton linearization appears to be often reasonably fast.

We refer to [11] for a general presentation of the ideas of energy-stable numerical methods for drift-diffusion problems.

2. Definition of the numerical scheme and main results.

2.1. Spatial discretization. The scheme we propose is based on conforming $\mathbb{P}_1$ finite elements with mass lumping. More precisely, let $\mathcal{T}$ be a conforming simplicial discretization of Ω with size and regularity respectively defined by

$$h_{\mathcal{T}} := \max_{T\in\mathcal{T}} h_T, \qquad \theta_{\mathcal{T}} := \max_{T\in\mathcal{T}} \theta_T = \max_{T\in\mathcal{T}} \frac{h_T}{\rho_T}, \tag{2.1}$$

where h_T denotes the diameter of the simplex $T \in \mathcal{T}$, and ρ_T denotes the diameter of the largest sphere included in $T \in \mathcal{T}$.

We also denote by $\mathcal{V}_{\mathcal{T}}$ and $\mathcal{E}_{\mathcal{T}}$ the set of the vertices and $(d-1)$-dimensional faces, respectively. Furthermore, since we have a (non-homogeneous) Dirichlet boundary condition on a part Σ_{D} of the boundary, we have to distinguish the vertices on Σ_{D}. We decompose the set of vertices $\mathcal{V}_{\mathcal{T}}$ into interior vertices $\mathcal{V}_{\mathcal{T}}^{\mathrm{int}}$ belonging to Ω, and exterior vertices $\mathcal{V}_{\mathcal{T}}^{\mathrm{ext}}$ belonging to the boundary $\partial\Omega$. Furthermore, we assume that a $(d-1)$-dimensional face included in the boundary $\partial\Omega$ lies entirely either in Σ_{N} or in Σ_{D}. Then, we note $\mathcal{V}_{\mathcal{T}}^{\mathrm{ext,N}}$ (resp. $\mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}$) the mesh vertices which belong to some Neumann (resp. Dirichlet) boundary face and we remark that $\mathcal{V}_{\mathcal{T}}^{\mathrm{ext,N}} \cap \mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}} \neq \emptyset$ if $\Sigma_{\mathrm{D}} \neq \partial\Omega$.

We denote by V_h the usual conforming $\mathbb{P}_1$ finite element space corresponding to the mesh $\mathcal{T}$, that is

$$V_h := \left\{ v_h \in C(\overline{\Omega}) \;\middle|\; v_h|_T \text{ is affine} \, \forall T \in \mathcal{T} \right\}.$$

In order to take into account the Dirichlet boundary condition we also introduce the spaces V_h^0 in which the test functions will be chosen, that is

$$V_h^0 := \{ v_h \in V_h : v_h = 0 \text{ on } \Sigma_{\mathrm{D}} \}.$$

The discrete solutions u_h^n will take values in the spaces $V_h^{\mathrm{D},n}$ $(n \geq 0)$ that incorporates two important constraints. First, the values at the boundary vertices are prescribed by the boundary data u_D. Second, we have to make sure that $p(u_h(t_n, \boldsymbol{a}))$ makes sense, i.e., $u_h(t_n, \boldsymbol{a}) \in I_p$, for any $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}$ and any $n \geq 1$, as well as $E(u_h(0, \boldsymbol{a}))$, hence $u_h(0, \boldsymbol{a}) \in \overline{I}_p$. This motivates the definition of the spaces

$$\begin{aligned} V_h^{\mathrm{D},n} &:= \left\{ v_h \in V_h \cap C(\overline{\Omega}; I_p) : v_h(\boldsymbol{a}) = u_{\mathrm{D}}(t_n, \boldsymbol{a}) \text{ for } \boldsymbol{a} \in \Sigma_{\mathrm{D}} \right\}, \qquad n \geq 1, \\ V_h^{D,0} &:= \left\{ v_h \in V_h \cap C(\overline{\Omega}; \overline{I}_p) : v_h(\boldsymbol{a}) = u_{\mathrm{D}}(0, \boldsymbol{a}) \text{ for } \boldsymbol{a} \in \Sigma_{\mathrm{D}} \right\}. \end{aligned}$$

Here t_n, $1 \leq n \leq N$, are the discrete times defined in Section 2.2 below.

In what follows, we denote by $(\phi_{\boldsymbol{a}})_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}}$ the basis of V_h made of the shape functions corresponding to $\mathcal{T}$:

$$\phi_{\boldsymbol{a}} \in V_h, \quad \phi_{\boldsymbol{a}}(\boldsymbol{a}') = \delta_{\boldsymbol{a}}^{\boldsymbol{a}'}, \qquad \forall \boldsymbol{a}, \boldsymbol{a}' \in \mathcal{V}_{\mathcal{T}}. \tag{2.2}$$

For any vertex $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}$, we also define the patch $\omega_{\boldsymbol{a}}$ that is the set of all the simplices sharing the given vertex $\boldsymbol{a}$ (see Fig. 1) and we denoted by $h_{\omega_{\boldsymbol{a}}}$ its diameter. In order to deal with the mass-lumping procedure, we introduce the so-called dual barycentric (or Donald) mesh. For all $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}$, we define the polyhedral open subset $s_{\boldsymbol{a}}$ of Ω whose boundary $\partial s_{\boldsymbol{a}}$ is defined by the hyperplanes joining

- the centers of mass $\boldsymbol{x}_T$ and $\boldsymbol{x}_e$ of the triangles and edges sharing $\boldsymbol{a}$ as a vertex if $d = 2$;
- the centers of mass $\boldsymbol{x}_T$, $\boldsymbol{x}_F$, and $\boldsymbol{x}_e$ of the tetrahedra, faces, and edges sharing $\boldsymbol{a}$ as a vertex if $d = 3$.

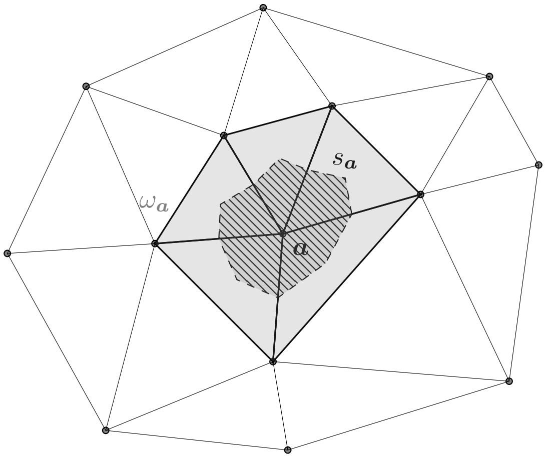


Fig. 1: Definition of $s_{\boldsymbol{a}}$ and $\omega_{\boldsymbol{a}}$

This construction is made so that

$$\bigcup_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} \overline{s_{\boldsymbol{a}}} = \overline{\Omega}, \qquad s_{\boldsymbol{a}} \cap s'_{\boldsymbol{a}} = \emptyset \text{ if } \boldsymbol{a} \neq \boldsymbol{a}', \qquad |s_{\boldsymbol{a}}| := \int_{s_{\boldsymbol{a}}} \mathrm{d}\boldsymbol{x} = \int_{\omega_{\boldsymbol{a}}} \phi_{\boldsymbol{a}} = \frac{|\omega_{\boldsymbol{a}}|}{(d+1)}. \tag{2.3}$$

It allows to define the functional space $X_h \subset L^\infty(\Omega) \cap BV(\Omega)$ by

$$X_h := \{v \in L^\infty(\Omega) \mid v|_{s_{\boldsymbol{a}}} \text{ is constant for all } \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}\}.$$

In view of the analysis below, we also need the functional space $\widetilde{X}_h \subset L^\infty(\Omega) \cap BV(\Omega)$ by

$$\widetilde{X}_h := \{v \in L^\infty(\Omega) \mid v|_T \text{ is constant for all } T \in \mathcal{T}\}.$$

The linear mappings $\pi_0 : C(\overline{\Omega}) \to X_h$ and $\pi_1 : C(\overline{\Omega}) \to V_h$ are prescribed by the point values as

$$\pi_\ell v(\boldsymbol{a}) = v(\boldsymbol{a}), \qquad \forall \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}},\ \forall v \in C(\overline{\Omega}),\ \ell \in \{0,1\}. \tag{2.4}$$

In the sequel, we denote by $\mathcal{V}_T \subset \mathcal{V}_{\mathcal{T}}$ the set of the $(d+1)$ vertices of the simplex $T \in \mathcal{T}$. With this notation the linear mapping $\widetilde{\pi}_0 : C(\overline{\Omega}) \to \widetilde{X}_h$ is defined by

$$\widetilde{\pi}_0 v(\boldsymbol{x}_T) = \frac{1}{d+1} \sum_{\boldsymbol{a}\in\mathcal{V}_T} v(\boldsymbol{a}) \qquad \forall T \in \mathcal{T},\ \forall v \in C(\overline{\Omega}). \tag{2.5}$$

In what follows, we will often use the fundamental property of piecewise constant reconstructions, i.e.,

$$\pi_0(g(u)) = g(\pi_0 u), \qquad \forall u \in C(\overline{\Omega}; I),\ \forall g \in C(I;\mathbb{R}); \tag{2.6}$$

as well as the following property which results from the definition of the reconstruction operators $\pi_0, \widetilde{\pi}_0$, and π_1:

$$\int_T \pi_0 v = \int_T \widetilde{\pi}_0 v = \int_T \pi_1 v, \qquad \forall v \in C(\overline{\Omega}),\ \forall T \in \mathcal{T}. \tag{2.7}$$

The approximate permeability tensor field $\mathbf{\Lambda}_h : \Omega \to \mathbb{R}^{d\times d}$ is defined almost everywhere by

$$\mathbf{\Lambda}_h(\boldsymbol{x}) := \mathbf{\Lambda}_T := \frac{1}{|T|}\int_T \mathbf{\Lambda} \quad \text{if } \boldsymbol{x} \in T. \tag{2.8}$$

Therefore, with this definition, $\mathbf{\Lambda}_h$ satisfies

$$\mathbf{\Lambda}_h \xrightarrow[h_{\mathcal{T}}\to 0]{} \mathbf{\Lambda} \text{ a.e. in } \Omega \quad \text{and} \quad \|\mathbf{\Lambda}_h\|_{L^\infty(\Omega)} \leq \|\mathbf{\Lambda}\|_{L^\infty(\Omega)}\,. \tag{2.9}$$

The vertices of any simplex $T \in \mathcal{T}$ are ordered and denoted by $\boldsymbol{a}_0^T, \ldots, \boldsymbol{a}_d^T$ or simply $\boldsymbol{a}_0, \ldots, \boldsymbol{a}_d$ if there is no ambiguity. Then we define the matrix $\mathbf{A}_T := (\alpha_{i,j}^T)_{1\leq i,j\leq d} \in \mathbb{R}^{d\times d}$ by

$$\alpha_{i,j}^T := \alpha_{j,i}^T := \int_T \mathbf{\Lambda}_h \boldsymbol{\nabla}\phi_{\boldsymbol{a}_i}\cdot\boldsymbol{\nabla}\phi_{\boldsymbol{a}_j}.$$

Consequently, for any $u_h, v_h \in V_h$,

$$\int_T \mathbf{\Lambda}_h \boldsymbol{\nabla} u_h\cdot\boldsymbol{\nabla} v_h = \begin{pmatrix} v_{\boldsymbol{a}_1} - v_{\boldsymbol{a}_0} \\ \vdots \\ v_{\boldsymbol{a}_d} - v_{\boldsymbol{a}_0}\end{pmatrix} \cdot \mathbf{A}_T \begin{pmatrix} u_{\boldsymbol{a}_1} - u_{\boldsymbol{a}_0} \\ \vdots \\ u_{\boldsymbol{a}_d} - u_{\boldsymbol{a}_0}\end{pmatrix}. \tag{2.10}$$

Thanks to classical arguments from the finite element theory, we can show that there exists C_1 depending only on the regularity on the mesh $\theta_{\mathcal{T}}$ and on the anisotropy ratio of $\mathbf{\Lambda}$ such that

$$\mathrm{cond}_2(\mathbf{A}_\mathrm{T}) \leq \mathrm{C}_1, \qquad \forall \mathrm{T} \in \mathcal{T}. \tag{2.11}$$

A similar inequality was derived in the contexts of the VAG scheme in [15] and of the DDFV scheme [13].

Following [15, Lemma A.2], we deduce from inequality (2.11) that there exists C_2 depending only on $\mathbf{\Lambda}, \theta_{\mathcal{T}}$, and d such that, for any $T \in \mathcal{T}$ and for any $\boldsymbol{v} = (v_{\boldsymbol{a}_i})_{i=1,\cdots,d} \in \mathbb{R}^d$

$$\sum_{i=1}^d \left(\sum_{j=1}^d |\alpha_{i,j}^T|\right)(v_{\boldsymbol{a}_i})^2 \leq C_2 \boldsymbol{v}\cdot\mathbf{A}_T\boldsymbol{v}. \tag{2.12}$$

2.2. Time discretization. We are concerned with the discretization of the evolutionary problem (1.1). It will be performed thanks to the backward Euler scheme. To this end, we consider a partition $(t_n)_{0\leq n\leq N}$ of the interval $[0, t_\mathrm{f}]$, i.e.,

$$0 = t_0 < t_1 < \cdots < t_N = t_\mathrm{f}.$$

We denote by $\tau_n := t_n - t_{n-1}$, $1 \leq n \leq N$, and by

$$\tau := \max_{1\leq n\leq N} \tau_n.$$

We define the time-and-space discretization spaces $V_{h\tau}$, $X_{h\tau}$, and $\widetilde{X}_{h\tau}$ by

$$V_{h\tau} := \Big\{ v_{h\tau} \in D((-\infty, t_{\rm f}]; V_h) \ : \ v_{h\tau}(t,\cdot) = v_{h\tau}(t_n,\cdot), \qquad \forall t \in (t_{n-1}, t_n],\ 0 \le n \le N \Big\}, \tag{2.13a}$$

$$X_{h\tau} := \Big\{ v_{h\tau} \in D((-\infty, t_{\rm f}]; X_h) \ : \ v_{h\tau}(t,\cdot) = v_{h\tau}(t_n,\cdot), \qquad \forall t \in (t_{n-1}, t_n],\ 0 \le n \le N \Big\}, \tag{2.13b}$$

$$\widetilde{X}_{h\tau} := \Big\{ v_{h\tau} \in D((-\infty, t_{\rm f}]; \widetilde{X}_h) \ : \ v_{h\tau}(t,\cdot) = v_{h\tau}(t_n,\cdot), \qquad \forall t \in (t_{n-1}, t_n],\ 0 \le n \le N \Big\}, \tag{2.13c}$$

where $D(I,E)$ denotes the set of the left-continuous with right-limit (càglàd) functions from the interval I to the space E. In (2.13), we have set $t_{-1} = -\infty$, so that functions of $V_{h\tau}$, $X_{h\tau}$, and $\widetilde{X}_{h\tau}$ are constants on $\{t \le 0\}$.

The mapping $\pi_0 : C(\overline{\Omega}) \to X_h$ (resp. $\pi_1 : C(\overline{\Omega}) \to V_h$) is naturally extended to the time-and-space framework into an operator still denoted by $\pi_0 : C([0,t_{\rm f}] \times \overline{\Omega}) \to X_{h\tau}$ (resp. $\pi_1 : C([0,t_{\rm f}] \times \overline{\Omega}) \to V_{h\tau}$).

In what follows, we will also need the space $\hat{V}_{h\tau}$ of piecewise linear reconstructions in both space and time, and the space $\check{X}_{h\tau}$ of piecewise linear in time and piecewise constant in space reconstructions, i.e.,

$$\hat{V}_{h\tau} = \Big\{ \hat{v}_{h\tau} \in C([0,t_{\rm f}]; V_h) \ : \hat{v}_{h\tau}(t,\cdot) = \frac{t_n - t}{\tau_n} \hat{v}_{h\tau}(t_{n-1},\cdot) + \frac{t - t_{n-1}}{\tau_n} \hat{v}_{h\tau}(t_n,\cdot), \qquad \forall t \in [t_{n-1}, t_n],\ 1 \le n \le N \Big\},$$

$$\check{X}_{h\tau} = \Big\{ \check{v}_{h\tau} \in C([0,t_{\rm f}]; X_h) \ : \check{v}_{h\tau}(t,\cdot) = \frac{t_n - t}{\tau_n} \check{v}_{h\tau}(t_{n-1},\cdot) + \frac{t - t_{n-1}}{\tau_n} \check{v}_{h\tau}(t_n,\cdot), \qquad \forall t \in [t_{n-1}, t_n],\ 1 \le n \le N \Big\}.$$

We define the linear subspace $V_{h\tau}^0$ and the subset $V_{h\tau}^D$ of $V_{h\tau}$ by

$$V_{h\tau}^0 = \left\{ v_{h\tau} \in V_{h\tau} : v_h^n \in V_h^0, 0 \le n \le N \right\},$$
$$V_{h\tau}^D = \Big\{ v_{h\tau} \in V_{h\tau} : v_h^n \in V_h^{D,n}, 0 \le n \le N \Big\}.$$

Notation: In order to lighten the notations, when applied to a function v_h of V_h or to a function $v_{h\tau}$ of $V_{h\tau}$, the operator π_0 is replaced by $\bar{\cdot}$, i.e.,

$$\pi_0 v_h = \overline{v}_h, \quad \forall v_h \in V_h, \quad \text{and} \quad \pi_0 v_{h\tau} = \overline{v}_{h\tau}, \quad \forall v_{h\tau} \in V_{h\tau}.$$

Moreover, for $v_{h\tau} \in V_{h\tau}$, we denote by $\hat{v}_{h\tau}$ (resp. $\check{v}_{h\tau}$) the unique element of $\hat{V}_{h\tau}$ (resp. $\check{X}_{h\tau}$) such that

$$v_{h\tau}(t_n,\cdot) = \hat{v}_{h\tau}(t_n,\cdot), \quad \overline{v}_{h\tau}(t_n,\cdot) = \check{v}_{h\tau}(t_n,\cdot) \qquad \forall n \ge 0.$$

REMARK 2.1 (Mesh adaptivity). *We consider in our presentation that the mesh $\mathcal{T}$ is fixed along time. Based on the a posteriori analysis carried out in §5 for proving Theorem 2.5, we could have considered dynamic mesh adaptation. This is possible in our setting but we avoid it for the sake of simplicity.*

2.3. Presentation of the scheme. Let us start by discretizing the initial data by setting

$$u_{\boldsymbol{a}}^0 := \begin{cases} \dfrac{1}{|s_{\boldsymbol{a}}|}\displaystyle\int_{s_{\boldsymbol{a}}} u_0, & \forall \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}} \setminus \mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}, \\ u_{\mathrm{D}}(0,\boldsymbol{a}), & \forall \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}, \end{cases} \tag{2.14}$$

and $u_h^0 := \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} u_{\boldsymbol{a}}^0 \phi_{\boldsymbol{a}} \in V_h^{\mathrm{D},0}$. As a consequence of Jensen's inequality and since the energy density function E defined by (1.4) is non-negative, one has

$$\sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}\setminus\mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}} \int_{s_{\boldsymbol{a}}} \pi_0 E(u_h^0) \le \int_\Omega E(u_0).$$

Concerning the vertices on the Dirichlet part of the boundary we have

$$\sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}} \int_{s_{\boldsymbol{a}}} \pi_0 E(u_h^0) = \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}} \int_{s_{\boldsymbol{a}}} E(u_{\mathrm{D}}(0,\boldsymbol{a})) \le C h_{\mathcal{T}} \left\| E(u_{\mathrm{D}}) \right\|_{L^\infty(Q_{t_{\mathrm{f}}})}.$$

Thus, thanks to (2.6), (1.9), and (1.11), we obtain

$$\int_\Omega E(\pi_0 u_h^0) = \int_\Omega \pi_0 E(u_h^0) \le \int_\Omega E(u_0) + C h_{\mathcal{T}} \left\| E(u_{\mathrm{D}}) \right\|_{L^\infty(Q_{t_{\mathrm{f}}})} < \infty. \tag{2.15}$$

Concerning the source term, we set, with $u_{\boldsymbol{a}}^n$ defined below, (see (2.21))

$$f_{\boldsymbol{a}}^n := f_{\mathrm{inj},\boldsymbol{a}}^n - \eta((u_{\boldsymbol{a}}^n)^+) f_{\mathrm{out},\boldsymbol{a}}^n, \quad \forall \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}},\ \forall n \in \{0,\cdots,N\}, \tag{2.16}$$

where for any $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}$ and $n \in \{0,\cdots,N\}$ we have set

$$f_{*,\boldsymbol{a}}^n := \frac{1}{|s_{\boldsymbol{a}}|\tau_n} \int_{\omega_{\boldsymbol{a}}} \int_{t_{n-1}}^{t_n} f_*(t,\boldsymbol{x})\phi_{\boldsymbol{a}}, \text{ with } * = \{\mathrm{inj},\mathrm{out}\}. \tag{2.17}$$

With this we can define the piecewise affine function

$$f_h^n := \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} f_{\boldsymbol{a}}^n \phi_{\boldsymbol{a}} \in V_h, \tag{2.18}$$

the corresponding piecewise constant function $\overline{f}_h^n \in X_h$, as well as the space-time reconstructions $f_{h\tau} \in V_{h\tau}$ and $\overline{f}_{h\tau} \in X_{h\tau}$.

Now, we have at hand all the necessary material to define the numerical scheme. Let $n \ge 1$, and assume that $u_h^{n-1} \in V_h^{\mathrm{D},n-1}$ is known and fulfills

$$\int_\Omega E(\overline{u}_h^{n-1}) = \int_\Omega E(\pi_0 u_h^{n-1}) < \infty.$$

Denoting

$$\Psi_h := \pi_1 \Psi, \quad \eta_h^n := \pi_1 \eta(u_h^n), \quad p_h^n := \pi_1 p(u_h^n) \tag{2.19}$$

and using the notation

$$\overline{w}_h = \pi_0 w_h, \qquad w_h \in V_h, \tag{2.20}$$

we look for

$$u_h^n := \sum_{\boldsymbol{a} \in \mathcal{V}_\mathcal{T}} u_{\boldsymbol{a}}^n \phi_{\boldsymbol{a}} \in V_h^{\mathrm{D},n} \tag{2.21}$$

such that, $\forall\, v_h \in V_h^0$,

$$\int_\Omega \frac{\overline{u}_h^n - \overline{u}_h^{n-1}}{\tau_n} \overline{v}_h + \int_\Omega \eta_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h) \cdot \boldsymbol{\nabla} v_h = \int_\Omega \overline{f}_h^n \overline{v}_h. \tag{2.22}$$

For the analysis below, it is useful to recall that the Lagrange vertex-quadrature formula resulting from (2.5) is exact on $\mathbb{P}_1$:

Remark 2.2 (Lagrange vertex-quadrature formula). *For any $v_h, w_h \in V_h$ and any $z \in C(\overline{\Omega})$, one has*

$$\int_\Omega \pi_1 z \boldsymbol{\Lambda}_h \boldsymbol{\nabla} v_h \cdot \boldsymbol{\nabla} w_h = \int_\Omega \widetilde{\pi}_0 z \boldsymbol{\Lambda}_h \boldsymbol{\nabla} v_h \cdot \boldsymbol{\nabla} w_h.$$

Indeed, thanks to definition (2.8) of $\boldsymbol{\Lambda}_h$ and since $v_h, w_h \in V_h$, we have

$$\int_\Omega \pi_1 z \boldsymbol{\Lambda}_h \boldsymbol{\nabla} v_h \cdot \boldsymbol{\nabla} w_h = \sum_{T \in \mathcal{T}} \boldsymbol{\Lambda}_T \boldsymbol{\nabla} v_h|_T \cdot \boldsymbol{\nabla} w_h|_T \int_T \pi_1 z,$$

so that relation (2.7) yields the assertion.

This remark allows to rewrite the scheme under an equivalent form to be used later on in the convergence proof, that is

$$\int_\Omega \frac{\overline{u}_h^n - \overline{u}_h^{n-1}}{\tau_n} \overline{v}_h + \int_\Omega \widetilde{\eta}_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h) \cdot \boldsymbol{\nabla} v_h = \int_\Omega \overline{f}_h^n \overline{v}_h, \quad \forall v_h \in V_h^0, \tag{2.23}$$

where we denote

$$\widetilde{\eta}_h^n = \widetilde{\pi}_0 \eta(u_h^n). \tag{2.24}$$

At each time step $n \in \{1, \dots, N\}$, the scheme gives rise to a nonlinear system of $\#\mathcal{V}_\mathcal{T} - \#\mathcal{V}_\mathcal{T}^{\mathrm{ext,D}}$ algebraic equations

$$\boldsymbol{\mathcal{F}}^n \left((u_{\boldsymbol{a}}^n)_{\boldsymbol{a} \in \mathcal{V}_\mathcal{T} \setminus \mathcal{V}_\mathcal{T}^{\mathrm{ext,D}}} \right) = \boldsymbol{0}, \tag{2.25}$$

with each line obtained by choosing $v_h = \phi_{\boldsymbol{a}'}$, $\boldsymbol{a}' \in \mathcal{V}_\mathcal{T} \setminus \mathcal{V}_\mathcal{T}^{\mathrm{ext,D}}$, in (2.22).

As it will be proved later on, see Theorem 2.3, the scheme (2.22) above gives rise to a nonlinear system of algebraic equations which admits at least one solution $u_h^n \in V_h^{D,n}$, $n \geq 1$. This allows to construct recursively a solution $u_{h\tau} \in V_{h\tau}^D$ which satisfies: $\forall v_{h\tau} \in V_{h\tau}^0$,

$$\iint_{Q_{t_\mathrm{f}}} \partial_t \check{u}_{h\tau} \overline{v}_{h\tau} + \iint_{Q_{t_\mathrm{f}}} \eta_{h\tau} \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_{h\tau} + \Psi_h) \cdot \boldsymbol{\nabla} v_{h\tau} = \iint_{Q_{t_\mathrm{f}}} \overline{f}_{h\tau} \overline{v}_{h\tau}. \tag{2.26}$$

In the above statement, $\eta_{h\tau}$ and $p_{h\tau}$ denote the unique elements of $V_{h\tau}$ such that $\eta_{h\tau}(t_n, \cdot) = \eta_h^n$ and $p_{h\tau}(t_n, \cdot) = p_h^n$.

2.4. Main results. As highlighted in (2.25), the scheme yields a nonlinear system to be solved at each time step. The existence of a solution $(u_{\boldsymbol{a}}^n)_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}\setminus\mathcal{V}_\mathcal{T}^{\text{ext,D}}}$ to theses systems, and thus of an approximate solution $u_{h\tau}$, is far from being obvious. One must in particular check that $u_{\boldsymbol{a}}^n \in I_p$ for all $\boldsymbol{a}\in\mathcal{V}_\mathcal{T}$ so that $(u_{\boldsymbol{a}}^n)$ is in the domain of $\boldsymbol{\mathcal{F}}^n$. This can be deduced from the energy stability of the scheme. More precisely, define the discrete counterpart

$$I_n(\overline{u}_h^n) := \overline{u}_h^n\left(p(1) - \overline{p}_{\mathrm{D},h}^n\right) \in X_h \tag{2.27}$$

of $I(u)$ defined in (1.12), where $p_{\mathrm{D},h}^n = \pi_1 p_\mathrm{D}(t^n,.)$ and $\overline{p}_{\mathrm{D},h}^n = \pi_0 p_\mathrm{D}(t^n,.) = \pi_0 p_{\mathrm{D},h}^n$. Then our first main result is devoted to the analysis of the scheme at fixed grid and is the purpose of the following statement.

Theorem 2.3 (Discrete solution and local conservativity). *For any given $u_h^{n-1}\in V_h^{\mathrm{D},n-1}$, there exists at least one solution $u_h^n \in V_h^{\mathrm{D},n}$ to the scheme (2.22) satisfying $u_h^n(\boldsymbol{a})\in I_p$ for all $\boldsymbol{a}\in\mathcal{V}_\mathcal{T}$ and*

$$\int_\Omega \left(E(\overline{u}_h^n) + I_n(\overline{u}_h^n)\right) \le \left(1+\frac{\tau_n}{t_\mathrm{f}}\right)\int_\Omega \left(E(\overline{u}_h^{n-1}) + I_{n-1}(\overline{u}_h^{n-1})\right) + C_3\tau_n, \tag{2.28}$$

for some C_3 depending on the data of the continuous problem but neither on τ_n nor on $\mathcal{T}$. This yields the existence of a discrete solution $u_{h\tau}\in V_{h\tau}^D$ with u_h^0 given by (2.14) fulfilling (2.26). Moreover, the scheme is locally conservative in the sense that there exists $\boldsymbol{\sigma}_{h\tau}\in L^2((0,t_\mathrm{f});\mathbf{H}(\mathrm{div},\Omega))$, a local postprocessing of $u_{h\tau}$ which is piecewise constant in time with values in $\mathbf{RTN}_1$ given by (5.1) below, satisfying

$$\partial_t \hat{u}_{h\tau} + \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{h\tau} = f_{h\tau} \text{ in } \Omega \quad \text{and} \quad \boldsymbol{\sigma}_{h\tau}\cdot\boldsymbol{n} = 0 \text{ on } (0,t_\mathrm{f})\times\Sigma_\mathrm{N},$$

where $\hat{u}_{h\tau}\in\hat{V}_{h\tau}$ is the piecewise affine in space and in time approximation built from $u_{h\tau}$ by

$$u_{h\tau}(t_n,\cdot) = \hat{u}_{h\tau}(t_n,\cdot) \qquad \forall n\ge 0. \tag{2.29}$$

Our second main result concerns the convergence of the scheme when the discretization parameters $h_\mathcal{T}$ and τ tend to 0. Let $(\mathcal{T}_m)_{m\ge1}$ be a sequence of simplicial meshes with bounded regularity and size tending to 0, i.e.,

$$\sup_{m\ge1}\theta_{\mathcal{T}_m} \le \theta^\star < \infty, \qquad \lim_{m\to\infty} h_{\mathcal{T}_m} = 0,$$

and let $\left(\left(t_n^{(m)}\right)_{0\le n\le N_m}\right)_{m\ge1}$ be a sequence of time discretizations of $[0,t_\mathrm{f}]$. We denote by $\tau_n^{(m)} := t_n^{(m)} - t_{n-1}^{(m)}$ and by

$$\tau^{(m)} = \max_{n\ge1}\tau_n^{(m)} \underset{m\to\infty}{\longrightarrow} 0.$$

We denote by $\left(\overline{u}_{h\tau}^{(m)}\right)_{m\ge1}$ a sequence of piecewise space-time constant approximate solutions provided by the scheme (2.26) corresponding to the simplicial meshes $\mathcal{T}_m$ and the time discretizations $\left((t_n^{(m)})_n\right)_m$. Then the following convergence result holds.

Theorem 2.4 (Convergence of the scheme). *There exists a weak solution u to the problem (1.1) in the sense of Definition 1 such that, up to the extraction of an unlabeled subsequence, there holds*

$$\overline{u}_{h\tau}^{(m)} \xrightarrow[m\to\infty]{} u \quad \textit{in } L^1(Q_{t_{\mathrm{f}}}).$$

The proof of Theorem 2.4 relies on compactness arguments, hence it provides no information on the speed of convergence of the scheme. We are not able to provide an error estimate for some norm of the error $\|u - u_{h\tau}\|$, but we manage to provide an a posteriori error estimate expressed in the dual norm of the residual.

Theorem 2.5 (Guaranteed upper bound). *Let $u \in X$ be a weak solution to problem (1.19) and let $\hat{u}_{h\tau}$ be the piecewise linear reconstruction in $\hat{V}_{h\tau}$ of the approximate solution $u_{h\tau}$ given by (2.29). Let $\mathcal{J}(\hat{u}_{h\tau})$ be the dual norm of the residual be given by (5.6) below. Then*

$$\mathcal{J}(\hat{u}_{h\tau}) \le \eta_{\mathrm{F}} + \eta_{\mathrm{IC}} + \eta_{\mathrm{f_{inj}}} + \eta_{\mathrm{qd}},$$

where the flux estimator, the data (source term) oscillation estimator, and the initial condition estimator are respectively defined by

$$\begin{aligned}
\eta_{\mathrm{F}} &:= \int_0^{t_{\mathrm{f}}} \|\boldsymbol{\Lambda}(\boldsymbol{\nabla}\gamma(\hat{u}_{h\tau}) + \eta(\hat{u}_{h\tau})\boldsymbol{\nabla}\Psi) + \boldsymbol{\sigma}_{h\tau}\|_{L^1(\Omega)} \\
\eta_{\mathrm{f_{inj}}} &:= \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} h_{\omega_{\boldsymbol{a}}} \int_{t^{n-1}}^{t^n} \left\| f_{\mathrm{inj}} - f_{\mathrm{inj},\boldsymbol{a}}^{n} \right\|_{L^1(\omega_{\boldsymbol{a}})} \\
&\quad + \max_{n=1,\dots,N} \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} \int_{t^{n-1}}^{t^n} \left\| f_{\mathrm{inj}} - f_{\mathrm{inj},\boldsymbol{a}}^{n} \right\|_{L^1(\omega_{\boldsymbol{a}})} \\
\eta_{\mathrm{IC}} &:= \|\hat{u}_{h\tau}(0,\cdot) - u_0\|_{L^1(\Omega)},
\end{aligned} \tag{2.30a}$$

and the quadrature error for the source term is defined by

$$\begin{aligned}
\eta_{\mathrm{qd}} &:= \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} \int_{t^{n-1}}^{t^n} \left\| \left(\eta((\hat{u}_{h\tau})^+) - \eta((u_{\boldsymbol{a}}^n)^+)\right) f_{\mathrm{out}} \right\|_{L^1(\omega_{\boldsymbol{a}})} \\
&\quad + \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} \eta((u_{\boldsymbol{a}}^n)^+) h_{\omega_{\boldsymbol{a}}} \int_{t^{n-1}}^{t^n} \left\| f_{\mathrm{out}} - f_{\mathrm{out},\boldsymbol{a}}^{n} \right\|_{L^1(\omega_{\boldsymbol{a}})} \\
&\quad + \max_{n=1,\dots,N} \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} \eta((u_{\boldsymbol{a}}^n)^+) \int_{t^{n-1}}^{t^n} \left\| f_{\mathrm{out}} - f_{\mathrm{out},\boldsymbol{a}}^{n} \right\|_{L^1(\omega_{\boldsymbol{a}})},
\end{aligned} \tag{2.30b}$$

where we note that the space-time function $\left(\eta((\hat{u}_{h\tau})^+) - \eta((u_{\boldsymbol{a}}^n)^+)\right)$ is zero at the points $(t_n, \boldsymbol{a})$.

Remark 2.6. *The method we propose relies on lowest-order conforming finite elements. The extension to a quite general family of low-order methods is natural using the framework presented in [11, Section 3], at least concerning the energy stability stated in Theorem 2.3 and the convergence result stated in Theorem 2.4. Extensions to higher-order methods were not studied here as they appear more subtle. In particular, in order to build some higher-order extension, attention needs to be paid to construct a scheme in a way such that for smooth functions, the discretization of $\eta(u)\boldsymbol{\nabla}p(u)$ approximates $\boldsymbol{\nabla}\gamma(u)$ at the right order. Moreover, solutions to degenerate parabolic* *problems of the form* (1.1) *do not have smooth solutions in general, so that higher-order methods could potentially only be beneficial on adaptively-refined meshes, but not on uniformly-refined meshes.*

3. Energy stability and existence of a discrete solution.

Here we focus, for a given mesh, on the existence of a discrete solution of the numerical scheme (2.22), that is we prove the first part of Theorem 2.3. We do so by means of some a priori estimates that will also be useful to perform the convergence analysis of Theorem 2.4.

3.1. Energy estimates.

The energy estimate is one of the key point for the analysis to follow. Let us extend to the discrete level the calculations provided in Section 1.2. Thanks to definition (2.27) of I_n and estimate (1.8) (with $\widetilde{\varepsilon} = \varepsilon / \|p(1) - p_D\|_{L^\infty(Q_{t_\mathrm{f}})}$), one has that for any $\epsilon > 0$, there exists $\widetilde{C}_\epsilon > 0$ depending on p, p_D, and η such that for any $u_h \in V_h$,

$$|I_n(\overline{u}_h)| \le \|p(1) - p_D\|_{L^\infty(Q_{t_\mathrm{f}})} |\overline{u}_h| \le \epsilon E(\overline{u}_h) + \widetilde{C}_\epsilon. \tag{3.1}$$

Now we state the one-step energy estimate which is the cornerstone to prove the stability of the scheme, the existence of a solution, and the convergence of the scheme.

Proposition 3.1. *Let* $(u_h^n)_{1\le n\le N}$*, with* $u_h^n \in V_h^{\mathrm{D},n}$ *for any* n*, be a solution to the problem* (2.22) *associated with the initial data* u_0*. Then there exist constants* $C_3, C_4 > 0$ *depending only on* $\Omega, t_\mathrm{f}, \mathbf{\Lambda}, \Psi, f_\mathrm{inj}, f_\mathrm{out}, p_\mathrm{D}, p$*, and* η *such that*

$$\begin{aligned}\int_\Omega (E(\overline{u}_h^n) + I_n(\overline{u}_h^n)) + \frac{\tau_n}{2}\int_\Omega \eta_h^n \mathbf{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(p_h^n + \Psi_h) \\ \le \left(1 + \frac{\tau_n}{t_\mathrm{f}}\right)\int_\Omega \left(E(\overline{u}_h^{n-1}) + I_{n-1}(\overline{u}_h^{n-1})\right) + C_3\tau_n.\end{aligned} \tag{3.2}$$

Furthermore, we have the following uniform energy stability property:

$$\int_\Omega E(\overline{u}_h^n) \le C_4\left(\int_\Omega E(u_0) + 1\right), \ \forall n \in \{0, \cdots, N\}. \tag{3.3}$$

Proof. We begin by proving estimate (3.2). Choosing $v_h = p_h^n - p_{\mathrm{D},h}^n \in V_h^0$ as test function in equation (2.22) yields

$$\begin{aligned}&\int_\Omega \left(\overline{u}_h^n - \overline{u}_h^{n-1}\right)(\overline{p}_h^n - p(1)) + \int_\Omega \left(\overline{u}_h^n - \overline{u}_h^{n-1}\right)\left(p(1) - \overline{p}_{\mathrm{D},h}^n\right) \\ &+ \tau_n \int_\Omega \eta_h^n \mathbf{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(p_h^n + \Psi_h) \\ &= \tau_n \int_\Omega \eta_h^n \mathbf{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(\Psi_h + p_{\mathrm{D},h}^n) + \tau_n \int_\Omega \overline{f}_h^n(\overline{p}_h^n - \overline{p}_{\mathrm{D},h}^n).\end{aligned} \tag{3.4}$$

Thanks to the convexity of E, the first term on the left-hand side satisfies

$$\int_\Omega \left(\overline{u}_h^n - \overline{u}_h^{n-1}\right)(\overline{p}_h^n - p(1)) \ge \int_\Omega \left(E(\overline{u}_h^n) - E(\overline{u}_h^{n-1})\right). \tag{3.5}$$

Then, for the second term, owing to definition (2.27) of I_n and since $p_\mathrm{D} \in W^{1,\infty}(Q_{t_\mathrm{f}})$, one has

$$\int_\Omega \left(\overline{u}_h^n - \overline{u}_h^{n-1}\right)\left(p(1) - \overline{p}_{\mathrm{D},h}^n\right) = \int_\Omega \left(I_n(\overline{u}_h^n) - I_{n-1}(\overline{u}_h^{n-1}) + \tau_n \overline{u}_h^{n-1} \frac{\overline{p}_{\mathrm{D},h}^n - \overline{p}_{\mathrm{D},h}^{n-1}}{\tau_n} \right)$$
$$\geq \int_\Omega I_n(\overline{u}_h^n) - \int_\Omega I_{n-1}(\overline{u}_h^{n-1}) - \tau_n \left\| \partial_t p_{\mathrm{D}} \right\|_{L^\infty(Q_{t_{\mathrm{f}}})} \int_\Omega |\overline{u}_h^{n-1}|.$$

Then estimate (1.8) yields

$$\begin{aligned}\int_\Omega \left(\overline{u}_h^n - \overline{u}_h^{n-1}\right)\left(p(1) - \overline{p}_{\mathrm{D},h}^n\right) \geq& \int_\Omega I_n(\overline{u}_h^n) - \int_\Omega I_{n-1}(\overline{u}_h^{n-1}) \\ &- \tau_n \left\| \partial_t p_{\mathrm{D}} \right\|_{L^\infty(Q_{t_{\mathrm{f}}})} \left(\varepsilon \int_\Omega E(\overline{u}_h^{n-1}) + C_\varepsilon |\Omega| \right).\end{aligned} \tag{3.6}$$

Thanks to the Young inequality, assumption (1.2) on $\mathbf{\Lambda}$, and the L^∞ stability of the Lagrange interpolate π_1 on $W^{1,\infty}(\Omega)$ functions (cf. [25] or the proof of similar property (A.3) in Appendix A below), the first term in the right-hand side of (3.4) satisfies

$$\int_\Omega \eta_h^n \mathbf{\Lambda}_h \mathbf{\nabla}(p_h^n + \Psi_h)\cdot \mathbf{\nabla}(\Psi_h + p_{\mathrm{D},h}^n) \leq \frac{1}{2} \int_\Omega \eta_h^n \mathbf{\Lambda}_h \mathbf{\nabla}(p_h^n + \Psi_h)\cdot \mathbf{\nabla}(p_h^n + \Psi_h)$$
$$+ \frac{\lambda^\star}{2} \left\| \mathbf{\nabla}(\Psi + p_{\mathrm{D}}) \right\|^2_{L^\infty(Q_{t_{\mathrm{f}}})} \int_\Omega \eta_h^n.$$

Recall notations (2.19)–(2.20) and property (2.7). Then one has

$$\begin{aligned}\int_\Omega \eta_h^n =& \int_\Omega \pi_1 \eta(u_h^n) = \int_\Omega \widetilde{\pi}_0 \eta(u_h^n) = \int_\Omega \pi_0 \eta(u_h^n) = \sum_{T\in\mathcal{T}} \int_T \pi_0 \eta(u_h^n) \\ =& \sum_{T\in\mathcal{T}} \sum_{\boldsymbol{a}\in\mathcal{V}_T} \frac{|T|}{d+1} (\eta(u_h^n))(\boldsymbol{a}) = \sum_{T\in\mathcal{T}} \sum_{\boldsymbol{a}\in\mathcal{V}_T} \frac{|T|}{d+1} (\eta(\overline{u}_h^n))(\boldsymbol{a}) = \int_\Omega \eta(\overline{u}_h^n),\end{aligned} \tag{3.7}$$

so that using also estimate (1.8), we infer

$$\begin{aligned}\int_\Omega \eta_h^n \mathbf{\Lambda}_h \mathbf{\nabla}(p_h^n + \Psi_h)\cdot \mathbf{\nabla}(\Psi_h + p_{\mathrm{D},h}^n) \leq& \frac{1}{2} \int_\Omega \eta_h^n \mathbf{\Lambda}_h \mathbf{\nabla}(p_h^n + \Psi_h)\cdot \mathbf{\nabla}(p_h^n + \Psi_h) \\ &+ \frac{\lambda^\star}{2} \left\| \mathbf{\nabla}(\Psi + p_{\mathrm{D}}) \right\|^2_{L^\infty(Q_{t_{\mathrm{f}}})} \left(\varepsilon \int_\Omega E(\overline{u}_h^n) + C_\varepsilon |\Omega| \right).\end{aligned} \tag{3.8}$$

Now we have to deal with the last term in the right-hand side of (3.4). Thanks to definition (1.10) of f it can be decomposed in four terms. First, one has

$$\int_\Omega \overline{f}_{\mathrm{inj},h}^n \overline{p}_{\mathrm{D},h}^n \leq |\Omega| \left\| f_{\mathrm{inj}} \right\|_{L^\infty(Q_{t_{\mathrm{f}}})} \left\| p_{\mathrm{D}} \right\|_{L^\infty(Q_{t_{\mathrm{f}}})}.$$

Then, thanks to estimate (1.8), by denoting $p^+(u) = \max(0, p(u))$ we obtain,

$$\int_\Omega \overline{f}_{\mathrm{inj},h}^n \overline{p}_h^n \leq \left\| f_{\mathrm{inj}} \right\|_{L^\infty(Q_{t_{\mathrm{f}}})} \int_\Omega p^+(\overline{u}_h^n) \leq \left\| f_{\mathrm{inj}} \right\|_{L^\infty(Q_{t_{\mathrm{f}}})} \left(\varepsilon \int_\Omega E(\overline{u}_h^n) + C_\varepsilon |\Omega| \right),$$

and

$$\int_\Omega \eta((\overline{u}_h^n)^+)\,\overline{f}_{\mathrm{out},h}^n\,\overline{p}_{\mathrm{D},h}^n \le \|f_{\mathrm{out}}\|_{L^\infty(Q_{t_\mathrm{f}})}\,\|p_\mathrm{D}\|_{L^\infty(Q_{t_\mathrm{f}})}\left(\varepsilon\int_\Omega E(\overline{u}_h^n) + C_\varepsilon|\Omega|\right).$$

Concerning the last term, noting $p^-(u) = \max(0, -p(u))$ we have

$$-\int_\Omega \eta((\overline{u}_h^n)^+)\overline{f}_{\mathrm{out},h}^n\overline{p}_h^n \le \int_\Omega \eta((\overline{u}_h^n)^+)\overline{f}_{\mathrm{out},h}^n p^-(\overline{u}_h^n),$$

and we consider two cases.

- If $p(0) \ge 0$, since p is increasing we have $\eta((\overline{u}_h^n)^+)p^-(\overline{u}_h^n) = 0$ and
$$-\int_\Omega \eta((\overline{u}_h^n)^+)\overline{f}_{\mathrm{out},h}^n\overline{p}_h^n \le 0.$$
- If $p(0) < 0$, there exists $u_\star > 0$ such that $p(u_\star) = 0$. Since $p^-(u)$ (resp. $\eta(u^+)$) is a continuous and decreasing (resp. increasing) non-negative function which vanishes at $u_\star$ (resp. 0), the function $\eta(u^+)p^-(u)$ is continuous and vanishes outside of $]0, u_\star[$. Thus it is bounded, so there exists $C_{\eta,p}$ such that
$$-\int_\Omega \eta((\overline{u}_h^n)^+)\overline{f}_{\mathrm{out},h}^n\overline{p}_h^n \le |\Omega| C_{\eta,p}\,\|f_{\mathrm{out}}\|_{L^\infty(Q_{t_\mathrm{f}})}.$$

Combining all these inequalities there exists $C > 0$ depending on $\Omega, f_{\mathrm{inj}}, f_{\mathrm{out}}, p_\mathrm{D}, p$ and η such that, for all $\varepsilon > 0$, there holds

$$\int_\Omega \overline{f}_h^n(\overline{p}_h^n - \overline{p}_{\mathrm{D},h}^n) \le C\left(\varepsilon\int_\Omega E(\overline{u}_h^n) + C_\varepsilon\right). \tag{3.9}$$

Gathering estimates (3.5)–(3.9) in (3.4) leads to

$$\begin{aligned}&\int_\Omega (E(\overline{u}_h^n) + I_n(\overline{u}_h^n)) + \frac{\tau_n}{2}\int_\Omega \eta_h^n \mathbf{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(p_h^n + \Psi_h)\\ &\quad\le \int_\Omega \left(E(\overline{u}_h^{n-1}) + I_{n-1}(\overline{u}_h^{n-1})\right) + C\tau_n\varepsilon\int_\Omega E(\overline{u}_h^n) + C\tau_n\varepsilon\int_\Omega E(\overline{u}_h^{n-1}) + C'_\varepsilon\tau_n,\end{aligned}$$

and, consequently,

$$\begin{aligned}&(1 - C\tau_n\varepsilon)\int_\Omega (E(\overline{u}_h^n) + I_n(\overline{u}_h^n)) + C\tau_n\varepsilon\int_\Omega I_n(\overline{u}_h^n)\\ &\quad + \frac{\tau_n}{2}\int_\Omega \eta_h^n \mathbf{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(p_h^n + \Psi_h)\\ &\le (1 + C\tau_n\varepsilon)\int_\Omega \left(E(\overline{u}_h^{n-1}) + I_{n-1}(\overline{u}_h^{n-1})\right) - C\tau_n\varepsilon\int_\Omega I_{n-1}(\overline{u}_h^{n-1}) + C'_\varepsilon\tau_n.\end{aligned} \tag{3.10}$$

Using (3.1), one gets that

$$\int_\Omega I_n(\overline{u}_h^n) \ge -\int_\Omega (E(\overline{u}_h^n) + I_n(\overline{u}_h^n)) - C$$

and similarly for $\overline{u}_h^{n-1}$. Using this in (3.10) yields

$$\begin{aligned}&\int_\Omega (E(\overline{u}_h^n) + I_n(\overline{u}_h^n)) + \frac{\tau_n}{2(1 - 2C\varepsilon\tau_n)}\int_\Omega \eta_h^n \mathbf{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(p_h^n + \Psi_h)\\ &\le \frac{1 + 2C\varepsilon\tau_n}{1 - 2C\varepsilon\tau_n}\int_\Omega \left(E(\overline{u}_h^{n-1}) + I_{n-1}(\overline{u}_h^{n-1})\right) + \frac{C''_\varepsilon\tau_n}{1 - 2C\varepsilon\tau_n}.\end{aligned}$$

By noticing that $\tau_n \le t_{\rm f}$ and that if $x \le \dfrac{1}{2}$ then $\dfrac{1+x}{1-x} \le 1+4x$, the claim (3.2) follows by choosing $\varepsilon = \dfrac{1}{8Ct_{\rm f}}$, since then $\dfrac{1+2C\varepsilon\tau_n}{1-2C\varepsilon\tau_n} \le 1 + \dfrac{\tau_n}{t_{\rm f}}$ and $1 \le \dfrac{1}{1-2C\varepsilon\tau_n} \le \dfrac{4}{3}$. Combining this with the discrete Gronwall lemma, one has

$$\begin{aligned}\int_\Omega \left(E(\overline{u}_h^n) + I_n(\overline{u}_h^n)\right) &\le e^{\frac{t_n-t_0}{t_{\rm f}}} \int_\Omega \left(E(\overline{u}_h^0) + I_n(\overline{u}_h^0)\right) + \frac{4}{3}C''_\varepsilon \sum_{i=0}^{n-1} \tau_n e^{\frac{t_n-t_{i+1}}{t_{\rm f}}} \\ &\le e^1 \int_\Omega \left(E(\overline{u}_h^0) + I_0(\overline{u}_h^0)\right) + \frac{4}{3}C''_\varepsilon t_{\rm f} e^1.\end{aligned}$$

Thanks to estimate (3.1) (with $\varepsilon = 1$) and estimate (2.15) we obtain

$$\int_\Omega \left(E(\overline{u}_h^n) + I_n(\overline{u}_h^n)\right) \le e^1 \left(2\int_\Omega E(u_0) + 2Ch_{\mathcal{T}} \left\|E(u_{\rm D})\right\|_{L^\infty(Q_{t_{\rm f}})} + \widetilde{C}_1|\Omega| + \frac{4}{3}C''_\varepsilon t_{\rm f}\right).$$

Using estimate (3.1) with $\varepsilon = \dfrac{1}{2}$, we recover estimate (3.3). □

3.2. A pressure estimate. Estimate (3.2) provides a control on the energy dissipation. This information can be used to derive some weighted estimated on the variations of the pressure. Such an estimate is the purpose of the following lemma.

LEMMA 3.2. *Let $u_h^{n-1} \in V_h^{{\rm D},n-1}$ be given. Let $u_h^n = \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} u_{\boldsymbol{a}}^n \phi_{\boldsymbol{a}} \in V_h^{{\rm D},n}$ be a solution to problem (2.22). Then there exists a constant $C_5 > 0$ depending only on $\Omega, \boldsymbol{\Lambda}, \Psi, p, \eta, \tau_n, t_{\rm f}, \theta_{\mathcal{T}}$, and d such that*

$$\sum_{T\in\mathcal{T}} \widetilde{\eta}_T^n \sum_{i=1}^d \left(\sum_{j=1}^d |\alpha_{i,j}^T|\right) \left(p(u_{\boldsymbol{a}_i}^n) - p(u_{\boldsymbol{a}_0}^n)\right)^2 \le C_5 \left(1 + \int_\Omega E(\overline{u}_h^{n-1})\right).$$

Proof. Let $T \in \mathcal{T}$. Thanks to property (2.10) and to inequality (2.12), one has

$$\begin{aligned}\sum_{i=1}^d \left(\sum_{j=1}^d |\alpha_{i,j}^T|\right) \left(p(u_{\boldsymbol{a}_i}^n) - p(u_{\boldsymbol{a}_0}^n)\right)^2 &\le C_2 \int_T \boldsymbol{\Lambda}_T \boldsymbol{\nabla} p_h^n \cdot \boldsymbol{\nabla} p_h^n \\ &\le 2C_2 \int_T \boldsymbol{\Lambda}_T \boldsymbol{\nabla}(p_h^n + \Psi_h) \cdot \boldsymbol{\nabla}(p_h^n + \Psi_h) + 2C_2 \int_T \boldsymbol{\Lambda}_T \boldsymbol{\nabla}\Psi_h \cdot \boldsymbol{\nabla}\Psi_h.\end{aligned}$$

We now bound the last term in the right-hand side of this inequality. Assumption (1.2) and the L^∞ stability of the Lagrange interpolate π_1 on $W^{1,\infty}(\Omega)$ functions give

$$\int_T \boldsymbol{\Lambda}_T \boldsymbol{\nabla}\Psi_h \cdot \boldsymbol{\nabla}\Psi_h \le \lambda^\star |T| \left\|\boldsymbol{\nabla}\Psi\right\|_{L^\infty(\Omega)}^2.$$

Thus, owing to notation (2.24) and equality (3.7) and relation (1.8), one has

$$\begin{aligned}\sum_{T\in\mathcal{T}} \widetilde{\eta}_T \int_T \boldsymbol{\Lambda}_T \boldsymbol{\nabla}\Psi_h \cdot \boldsymbol{\nabla}\Psi_h &\le \lambda^\star \left\|\boldsymbol{\nabla}\Psi\right\|_{L^\infty(\Omega)}^2 \int_\Omega \eta(\overline{u}_h^n) \\ &\le \lambda^\star \left\|\boldsymbol{\nabla}\Psi\right\|_{L^\infty(\Omega)}^2 \left(\varepsilon \int_\Omega E(\overline{u}_h^n) + |\Omega| C_\varepsilon\right).\end{aligned}$$

Choosing ε in an appropriate way, one gets that

$$\sum_{T\in\mathcal{T}} \widetilde{\eta}_T \sum_{i=1}^{d} \left(\sum_{j=1}^{d} |\alpha_{i,j}^T| \right) \left(p(u_{\boldsymbol{a}_i}^n) - p(u_{\boldsymbol{a}_0}^n)\right)^2 \\ \leq C_6 \left(1 + \int_\Omega E(\overline{u}_h^n) + \int_\Omega \widetilde{\eta}_h \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(p_h^n + \Psi_h) \right) \tag{3.11}$$

for some C_6 depending only on $\Omega, \boldsymbol{\Lambda}, \Psi, p, \eta, \theta_{\mathcal{T}}$, and d. Since $\tau_n \leq t_{\mathrm{f}}$, this leads to

$$\sum_{T\in\mathcal{T}} \widetilde{\eta}_T^n \sum_{i=1}^{d} \left(\sum_{j=1}^{d} |\alpha_{i,j}^T| \right) \left(p(u_{\boldsymbol{a}_i}^n) - p(u_{\boldsymbol{a}_0}^n)\right)^2 \\ \leq \frac{C_7}{\tau_n} \left(1 + \int_\Omega E(\overline{u}_h^n) + \tau_n \int_\Omega \widetilde{\eta}_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(p_h^n + \Psi_h) \right) \tag{3.12}$$

for C_7 that additionally depends on t_{f}.

Moreover, thanks to relations (3.2) and (3.1) (with $\varepsilon = 1/2$) we also have,

$$\int_\Omega E(\overline{u}_h^n) + \tau_n \int_\Omega \widetilde{\eta}_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}(p_h^n + \Psi_h) \\ \leq 2\left(1 + \frac{\tau_n}{t_{\mathrm{f}}}\right) \int_\Omega \left(E(\overline{u}_h^{n-1}) + I_{n-1}(\overline{u}_h^{n-1})\right) + 2C_3\tau_n + 2\widetilde{C}_{\frac{1}{2}}. \tag{3.13}$$

The claim follows from combining (3.12) and (3.13) together with (3.1). □

In the case where $p(0) = -\infty$, the functional $\boldsymbol{\mathcal{F}}^n$ of (2.25) is continuous on $(0,\infty)^{\#\mathcal{V}_{\mathcal{T}}\setminus\mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}}$ but blows up when one $u_{\boldsymbol{a}}^n$ goes to 0. Fortunately, this situation is prevented for the solution of scheme (2.22) thanks to the control on the discrete pressure proven in Lemma 3.2, as shown in the next lemma.

Lemma 3.3. *Let $u_h^n \in V_h^{\mathrm{D},n}$ be a solution to scheme (2.22). Assume that $p(0) = -\infty$. Then there exists $\varepsilon_{\mathcal{T},\tau_n} > 0$ depending only on $\Omega, \boldsymbol{\Lambda}, \Psi, p, \eta, \tau_n, t_{\mathrm{f}}, \theta_{\mathcal{T}}$, and d such that*

$$u_{\boldsymbol{a}}^n \geq \varepsilon_{\mathcal{T},\tau_n} \qquad \forall \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}. \tag{3.14}$$

Proof. Since p is increasing with $p(0) = -\infty$ and $\lim_{u\to+\infty} p(u) = +\infty$ one has $u_{\mathrm{D}}= p^{-1}(p_{\mathrm{D}}) > 0$. Moreover there exists at least one vertex belonging to the Dirichlet boundary Σ_{D}, that is there exists $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}$ such that $u_{\boldsymbol{a}}^n = u_{\mathrm{D}}(t_n, \boldsymbol{a}) > 0$. We can now follow the reasoning given in [15, Lemma 3.7] to conclude, thus we do not give the details here. □

3.3. Existence of the discrete solution. We have now all the necessary tools at hand to prove the existence of a solution to the nonlinear system (2.25) and thus to the scheme (2.22). Since the proof is very similar to the one of [15, Proposition 3.8], we do not give the details here.

Proposition 3.4. *For all $n \geq 1$, there exists at least one solution $u_h^n \in V_h^{\mathrm{D},n}$ to the scheme (2.22).*

We have thus proven Theorem 2.3 up to the local conservativity statement. This will be done later on in Section 5.

4. Convergence analysis. In Section 3, for a given mesh $\mathcal{T}$ and a given time step τ_n, we proved the existence of a discrete solution u_h^n for any $n \in \{0, \cdots, N\}$. Recalling definition (2.13), we can reconstruct an approximate solution $u_{h\tau} \in V_{h\tau}$ to problem (2.22) associated with the initial data u_0, the mesh $\mathcal{T}$, and the time steps τ_n as a function piecewise constant in time such that $u_{h\tau}(t, \cdot) = u_h^n$, $t \in (t_{n-1}, t_n]$, $n \in \{1, \cdots, N\}$. In the sequel we denote

$$p_{h\tau} = \pi_1 p(u_{h\tau}),\ \ \xi_{h\tau} = \pi_1 \xi(u_{h\tau}),\ \ \eta_{h\tau} = \pi_1 \eta(u_{h\tau}),\ \ \widetilde{\eta}_{h\tau} = \widetilde{\pi}_0 \eta(u_{h\tau}),\ \ \overline{u}_{h\tau} = \pi_0 u_{h\tau}.$$

The goal of the current section is to prove Theorem 2.4. The proof is based on compactness arguments.

Rather than considering a single mesh $\mathcal{T}$ and time discretization $(t_n)_{0\le n\le N}$ of $[0, t_\mathrm{f}]$, we focus now on the situation where we have a sequence of meshes $(\mathcal{T}_m)_{m\ge1}$ and a sequence $\left((t_n^{(m)})_{0\le n\le N_m}\right)_{m\ge1}$ of time discretizations, with

$$h_{\mathcal{T}_m} \xrightarrow[m\to\infty]{} 0, \qquad \theta_{\mathcal{T}_m} \le \theta^\star, \qquad \tau^{(m)} \xrightarrow[m\to\infty]{} 0$$

yielding a sequence of piecewise constant in time and piecewise affine in space approximate solutions $u_{h\tau}^{(m)}$. Our goal is to show thanks to compactness arguments that the sequence of piecewise constant reconstructions $\left(\overline{u}_{h\tau}^{(m)}\right)_{m\ge1}$ converges towards a weak solution u in some appropriate sense. The proof will be made in three steps. We first state some stability estimates in Section 4.1 to be used then in Section 4.2 to infer some compactness properties on the sequence of approximate solutions. Finally we identify any limit point as a weak solution in Section 4.3.

In order to lighten the notations, we will get rid of the index m in the presentation below. Hence we consider the limit $h_\mathcal{T}, \tau \to 0$ instead of the limit $m \to \infty$.

4.1. Stability estimates. In order to obtain compactness results, we need to obtain further estimates.

LEMMA 4.1. *There exist constants $C_8, C_9 > 0$ depending on $\Omega, t_\mathrm{f}, \mathbf{\Lambda}, \Psi, f_\mathrm{inj}, f_\mathrm{out}, p_\mathrm{D}, p$, and η such that*

$$\iint_{Q_{t_\mathrm{f}}} \widetilde{\eta}_{h\tau} \mathbf{\Lambda}_h \boldsymbol{\nabla}(p_{h\tau} + \Psi_h) \cdot \boldsymbol{\nabla}(p_{h\tau} + \Psi_h) \le C_8 \left(\int_\Omega E(u_0) + 1 \right), \tag{4.1}$$

$$\iint_{Q_{t_\mathrm{f}}} \widetilde{\eta}_{h\tau} \mathbf{\Lambda}_h \boldsymbol{\nabla} p_{h\tau} \cdot \boldsymbol{\nabla} p_{h\tau} \le C_9 \left(\int_\Omega E(u_0) + 1 \right). \tag{4.2}$$

Moreover, there exists a constant $C_{10} > 0$ depending on $u_0, \Omega, t_\mathrm{f}, \theta_\mathcal{T}, \mathbf{\Lambda}, \Psi, f_\mathrm{inj}, f_\mathrm{out}, p_\mathrm{D}, p$, and η such that

$$\|E(\overline{u}_{h\tau})\|_{L^\infty((0,t_\mathrm{f});L^1(\Omega))} \le C_{10}, \tag{4.3}$$

$$\iint_{Q_{t_\mathrm{f}}} \mathbf{\Lambda}_h \boldsymbol{\nabla} \xi_{h\tau} \cdot \boldsymbol{\nabla} \xi_{h\tau} \le C_{10}, \tag{4.4}$$

$$\sum_{n=1}^N \tau_n \sum_{T\in\mathcal{T}} \widetilde{\eta}_T^n \sum_{i=1}^d \left(\sum_{j=1}^d |\alpha_{i,j}^T| \right) \left(p(u_{\boldsymbol{a}_i}^n) - p(u_{\boldsymbol{a}_0}^n) \right)^2 \le C_{10}, \tag{4.5}$$

$$\|\xi(\overline{u}_{h\tau})\|_{L^2((0,t_\mathrm{f});L^6(\Omega))} \le C_{10}, \tag{4.6}$$

$$\|\overline{u}_{h\tau}\|_{L^1((0,t_\mathrm{f});L^3(\Omega))} \le C_{10}. \tag{4.7}$$

Proof. Estimate (4.3) has already been established in (3.3), and (4.5) follows immediately from Lemma 3.2 and (3.3). To obtain relation (4.1), we sum (3.2) over $n \in \{1, \cdots, N\}$ and we conclude by using (3.1) and (3.3). Mimicking the proof of Lemma 3.2 and using the multistep a priori estimates given in estimates (3.3) and (4.1), we obtain the multistep a priori estimate (4.2).

Let us focus on the proof of estimate (4.4). We note $\mathcal{S}^n$ the simplex whose vertices are the values $(u^n_{\boldsymbol{a}_i})_{i=0,\cdots,d}$. Then, since the function η is non-decreasing, one has

$$\max_{u\in\mathcal{S}^n} \eta(u) = \max_{i=0,\cdots,d} \eta(u^n_{\boldsymbol{a}_i}).$$

This leads to, recalling (2.24) and (2.5),

$$\widetilde{\eta}^n_T = \frac{1}{d+1}\sum_{i=0}^{d} \eta(u^n_{\boldsymbol{a}_i}) \geq \frac{1}{d+1}\max_{u\in\mathcal{S}^n} \eta(u).$$

Moreover thanks to the definition (1.15) of the semi-Kirchhoff transform ξ, for any $T \in \mathcal{T}$, and for vertices $\boldsymbol{a}_0, \boldsymbol{a}_i \in \mathcal{V}_T$ we have

$$\begin{aligned}\left(\xi(u^n_{\boldsymbol{a}_i}) - \xi(u^n_{\boldsymbol{a}_0})\right)^2 &\leq \left(\max_{u\in\mathcal{S}^n}\eta(u)\right)\left(p(u^n_{\boldsymbol{a}_i}) - p(u^n_{\boldsymbol{a}_0})\right)^2 \\ &\leq (d+1)\widetilde{\eta}^n_T \left(p(u^n_{\boldsymbol{a}_i}) - p(u^n_{\boldsymbol{a}_0})\right)^2.\end{aligned}$$

Thus, one has

$$\sum_{i=1}^{d} \widetilde{\eta}^n_T \left(p(u^n_{\boldsymbol{a}_i}) - p(u^n_{\boldsymbol{a}_0})\right)^2 \geq \frac{1}{d+1}\sum_{i=1}^{d}\left(\xi(u^n_{\boldsymbol{a}_i}) - \xi(u^n_{\boldsymbol{a}_0})\right)^2.$$

Noting that $\boldsymbol{v}\cdot\mathbf{A}_T\boldsymbol{v} \geq \boldsymbol{w}\cdot\mathbf{A}_T\boldsymbol{w}$ for any $\boldsymbol{v}, \boldsymbol{w}$ such that $|\boldsymbol{v}|^2 \geq \mathrm{cond}_2(\mathbf{A}_\mathrm{T})|\boldsymbol{w}|^2$ and using (2.10), we finally obtain

$$\int_T \widetilde{\eta}^n_T \boldsymbol{\Lambda}_T \boldsymbol{\nabla} p^n_h \cdot \boldsymbol{\nabla} p^n_h \geq \frac{1}{(d+1)\mathrm{cond}_2(\mathbf{A}_\mathrm{T})}\int_T \boldsymbol{\Lambda}_T \boldsymbol{\nabla}\xi^n_h \cdot \boldsymbol{\nabla}\xi^n_h.$$

Using relation (2.11), multiplying the resulting estimate by τ_n and then summing up over $T \in \mathcal{T}$ and $n \in \{0, \ldots, N\}$, Lemma 4.1 gives the expected bound.

Thanks to estimate (4.4) and to the control on $\xi_{h\tau}$ on the boundary Σ_D, we get from the Poincaré inequality that $\xi_{h\tau}$ is bounded in $L^2(Q_{t_\mathrm{f}})$, and then from Sobolev inequality that

$$\|\xi_{h\tau}\|_{L^2((0,t_\mathrm{f});L^6(\Omega))} \leq C.$$

Then we deduce (4.6) from the previous inequality together with [15, Lemma 6.6]. Finally, estimate (4.7) follows from the combination of (4.6) with (1.17). □

4.2. Some compactness properties on the approximate solutions. As a first step, we prove that the piecewise constant reconstructions $(\overline{u}_{h\tau})_{h,\tau}$ is sequentially relatively compact in $L^1(Q_{t_\mathrm{f}})$.

Proposition 4.2. *There exists a measurable function $u : Q_{t_\mathrm{f}} \to \mathbb{R}$ such that, up to the extraction of a subsequence, there holds*

$$\begin{aligned}&\overline{u}_{h\tau} \xrightarrow[h_\mathcal{T},\tau\to 0]{} u \ \textit{a.e. in } Q_{t_\mathrm{f}} \textit{ and strongly in } L^1(Q_{t_\mathrm{f}}),\\ &\xi_{h\tau} \xrightarrow[h_\mathcal{T},\tau\to 0]{} \xi(u) \ \textit{weakly in } L^2((0,t_\mathrm{f}); H^1(\Omega)).\end{aligned}$$

Proof. The proof relies on the time-compactness result proposed in [3] and recast to our framework in Appendix A. In order to apply Theorem A.1, we first note that the functions $\xi_{h\tau}$ are uniformly bounded in $L^2((0,t_{\rm f});H^1(\Omega))$ via (4.4) and the control prescribed on the boundary $\Sigma_{\rm D}$. Next, it results from (4.3), (1.8), and (4.7) that

$$\|\overline{u}_{h\tau}\|_{L^\infty((0,t_{\rm f});L^1(\Omega))} \le C, \qquad \|\overline{u}_{h\tau}\|_{L^1((0,t_{\rm f});L^3(\Omega))} \le C.$$

Thus,

$$\|\overline{u}_{h\tau}\|_{L^2((0,t_{\rm f});L^{4/3}(\Omega))} \le C$$

is a consequence of the Riesz–Thorin interpolation theorem. Therefore, it only remains to check that (A.1) holds, i.e.,

$$\iint_{Q_{t_{\rm f}}} \partial_t \check{u}_{h\tau}\overline{v}_{h\tau} \le C\|\boldsymbol{\nabla} v_{h\tau}\|_{L^\infty(Q_{t_{\rm f}})}, \qquad \forall v_{h\tau} \in V_{h\tau}^0, \tag{4.8}$$

for some C depending neither on τ nor on $h_{\mathcal{T}}$ (but on the data of the continuous problem and on the mesh regularity $\theta_{\mathcal{T}}$).

Thanks to (2.26), there holds

$$\iint_{Q_{t_{\rm f}}} \partial_t \check{u}_{h\tau}\overline{v}_{h\tau} = -\iint_{Q_{t_{\rm f}}} \eta_{h\tau}\boldsymbol{\Lambda}_h\boldsymbol{\nabla}(p_{h\tau}+\Psi_h)\cdot\boldsymbol{\nabla} v_{h\tau} + \iint_{Q_{t_{\rm f}}} \overline{f}_{h\tau}\overline{v}_{h\tau}. \tag{4.9}$$

It follows from the Cauchy–Schwarz inequality that

$$-\iint_{Q_{t_{\rm f}}} \eta_{h\tau}\boldsymbol{\Lambda}_h\boldsymbol{\nabla}(p_{h\tau}+\Psi_h)\cdot\boldsymbol{\nabla} v_{h\tau} \le$$
$$\left(\iint_{Q_{t_{\rm f}}} \eta_{h\tau}\boldsymbol{\Lambda}_h\boldsymbol{\nabla}(p_{h\tau}+\Psi_h)\cdot\boldsymbol{\nabla}(p_{h\tau}+\Psi_h)\right)^{1/2}\left(\iint_{Q_{t_{\rm f}}} \eta_{h\tau}\boldsymbol{\Lambda}_h\boldsymbol{\nabla} v_{h\tau}\cdot\boldsymbol{\nabla} v_{h\tau}\right)^{1/2}.$$

Owing to Remark 2.2 and (4.1), the first term of the right-hand side is uniformly bounded. The second term of the right-hand side can be estimated by

$$\iint_{Q_{t_{\rm f}}} \eta_{h\tau}\boldsymbol{\Lambda}_h\boldsymbol{\nabla} v_{h\tau}\cdot\boldsymbol{\nabla} v_{h\tau} \le \lambda^\star\|\boldsymbol{\nabla} v_{h\tau}\|^2_{L^\infty(Q_{t_{\rm f}})}\|\eta_{h\tau}\|_{L^1(Q_{t_{\rm f}})}.$$

We deduce from (3.7), (1.8), and (3.3) that $\|\eta_{h\tau}\|_{L^1(Q_{t_{\rm f}})}$ is uniformly bounded, hence

$$-\iint_{Q_{t_{\rm f}}} \eta_{h\tau}\boldsymbol{\Lambda}_h\boldsymbol{\nabla}(p_{h\tau}+\Psi_h)\cdot\boldsymbol{\nabla} v_{h\tau} \le C\|\boldsymbol{\nabla} v_{h\tau}\|_{L^\infty(Q_{t_{\rm f}})}.$$

The second term of the right-hand side of (4.9) can be estimated

$$\begin{aligned}\iint_{Q_{t_{\rm f}}} \overline{f}_{h\tau}\overline{v}_{h\tau} &\le \|\overline{f}_{h\tau}\|_{L^1(Q_{t_{\rm f}})}\|\overline{v}_{h\tau}\|_{L^\infty(Q_{t_{\rm f}})} \\ &\le \Big(|\Omega|t_{\rm f}\|f_{\rm inj}\|_{L^\infty(Q_{t_{\rm f}})} + \|f_{\rm out}\|_{L^\infty(Q_{t_{\rm f}})}\|\eta_{h\tau}\|_{L^1(Q_{t_{\rm f}})}\Big)\|v_{h\tau}\|_{L^\infty(Q_{t_{\rm f}})}.\end{aligned}$$

We use once again the fact that $\|\eta_{h\tau}\|_{L^1(Q_{t_{\rm f}})}$ is uniformly bounded, while the Poincaré inequality ensures that $\|v_{h\tau}\|_{L^\infty(Q_{t_{\rm f}})} \le C\|\boldsymbol{\nabla} v_{h\tau}\|_{L^\infty(Q_{t_{\rm f}})}$, so that (4.8) holds. Then Theorem A.1 provides that

$$\overline{u}_{h\tau} \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} u \text{ a.e. in } Q_{t_{\rm f}}$$

and

$$\xi_{h\tau} \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} \xi(u) \text{ weakly in } L^2((0,t_{\rm f});H^1(\Omega)).$$

Finally, the strong $L^1(Q_{t_{\rm f}})$ convergence follows from Vitali's convergence theorem (see for instance [51, Proposition 3.11]). □

The following lemma will be useful to identify the limits in what follows.

Lemma 4.3. *We define $w_{h\tau} \in \widetilde{X}_{h\tau}$ by*

$$w_{h\tau}|_{T\times(t_{n-1},t_n]} = w_T^n := \max_{\boldsymbol{a},\boldsymbol{a}'\in\mathcal{V}_T} |\xi(u_{\boldsymbol{a}}^n) - \xi(u_{\boldsymbol{a}'}^n)|\,,\ \forall T\in\mathcal{T}, \forall n\in\{0,\cdots,N\}. \tag{4.10}$$

Then

$$w_{h\tau} \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} 0 \text{ strongly in } L^2(Q_{t_{\rm f}}).$$

Proof. Let $T\in\mathcal{T}$, $\boldsymbol{a},\boldsymbol{a}'\in\mathcal{V}_T$, and $n\in\{0,\cdots,N\}$. Let also $\xi_h^n := \pi_1\xi(u_h^n)$. Then

$$\sqrt{|T|}\,|\xi(u_{\boldsymbol{a}}^n)-\xi(u_{\boldsymbol{a}'}^n)| = \sqrt{|T|}\,|\boldsymbol{\nabla}\xi_h^n|_T\cdot(\boldsymbol{a}-\boldsymbol{a}')| \le \sqrt{|T|}\,|\boldsymbol{\nabla}\xi_h^n|_T|\,h_T = h_T\,\|\boldsymbol{\nabla}\xi_h^n\|_{L^2(T)}.$$

Thus

$$\sqrt{|T|}w_T^n \le h_T\,\|\boldsymbol{\nabla}\xi_h^n\|_{L^2(T)}\,.$$

Summing up over $T\in\mathcal{T}$, we finally obtain

$$\left\|w_{h\tau}|_{(t_{n-1},t_n]}\right\|_{L^2(\Omega)} \le h_{\mathcal{T}}\,\|\boldsymbol{\nabla}\xi_h^n\|_{L^2(\Omega)}\,.$$

Thus, owing to estimate (4.4) and assumption (1.2), we conclude the proof. □

We identified in Proposition 4.2 a limit u of $u_{h\tau}$. The convergence being strong, it is enough to pass in the nonlinearities provided we do not introduce too much error in the different reconstructions of the functions. This is what we establish now.

Lemma 4.4. *Let u be a limit of $u_{h\tau}$ as in Proposition 4.2, then*

$$\overline{\eta}_{h\tau} = \pi_0\eta(u_{h\tau}) \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} \eta(u),\ \text{strongly in } L^1(Q_{t_{\rm f}}), \tag{4.11a}$$

$$\widetilde{\eta}_{h\tau} := \widetilde{\pi}_0\eta(u_{h\tau}) \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} \eta(u) \text{ strongly in } L^1(Q_{t_{\rm f}}). \tag{4.11b}$$

Proof. To obtain the convergences (4.11), we want to apply the Vitali's convergence theorem. So, first, we have to check that the sequences $(\overline{\eta}_{h\tau})_{h,\tau}$ and $(\widetilde{\eta}_{h\tau})_{h,\tau}$ are uniformly equi-integrable, i.e., for all $\varepsilon>0$, there exists $\alpha>0$ such that

$$|U|\le\alpha^2 \quad\implies\quad \iint_U \overline{\eta}_{h\tau} \le \varepsilon. \tag{4.12}$$

Let $\varepsilon>0$, and let $\alpha>0$ to be fixed later on. Let $U\subset Q_{t_{\rm f}}$ be such that $|U|\le\alpha^2$. First, we write $U = \bigcup_{t\in[0,t_{\rm f}]}\{t\}\times U_t$, and denote by $J(\alpha)=\{t\,|\,|U_t|\ge\alpha\}\subset[0,t_{\rm f}]$. Then Markov's inequality ensures that $|J(\alpha)|\le\alpha$. Now, we decompose

$$\iint_U \overline{\eta}_{h\tau} = \int_{J(\alpha)}\int_{U_t}\overline{\eta}_{h\tau} + \int_{J(\alpha)^c}\int_{U_t}\overline{\eta}_{h\tau}. \tag{4.13}$$

As a consequence of (1.8) and (4.3), $\overline{\eta}_{h\tau}$ is uniformly bounded in $L^\infty((0,t_{\rm f});L^1(\Omega))$, i.e.,

$$\|\overline{\eta}_{h\tau}\|_{L^\infty((0,t_{\rm f});L^1(\Omega))} \le C_{11}.$$

Therefore

$$\int_{J(\alpha)}\int_{U_t} \overline{\eta}_{h\tau} \le |J(\alpha)| \|\overline{\eta}_{h\tau}\|_{L^\infty((0,t_\mathrm{f});L^1(\Omega))} \le C_{11}\alpha. \tag{4.14}$$

Let us now focus on the second term in the right-hand side of (4.13). Thanks to (1.8), there exists $C_\varepsilon > 0$ such that

$$\eta(u) \le \frac{\varepsilon}{3t_\mathrm{f}C_{10}} E(u) + C_\varepsilon, \qquad \forall u \in \overline{I}_p.$$

Let $t \in J(\alpha)^c$, so that $|U_t| \le \alpha$, and let $n \in \{0, \ldots, N\}$ be such that $t \in (t_{n-1}, t_n]$. Then there holds

$$\int_{U_t} \eta(\overline{u}_h^n) \le \frac{\varepsilon}{3t_\mathrm{f}C_{10}} \int_{U_t} E(\overline{u}_h^n) + C_\varepsilon |U_t| \le \frac{\varepsilon}{3t_\mathrm{f}} + \alpha C_\varepsilon.$$

Therefore, we obtain that

$$\int_{J(\alpha)^c}\int_{U_t} \overline{\eta}_{h\tau} \le \frac{\varepsilon}{3} + \alpha t_\mathrm{f} C_\varepsilon. \tag{4.15}$$

Choosing $\alpha = \min\left(\varepsilon\dfrac{C_{11}}{3}, \dfrac{\varepsilon}{3t_\mathrm{f}C_\varepsilon}\right)$ in (4.14) and (4.15) and combining the results in (4.13) provides (4.12). Thereby the sequence $(\overline{\eta}_{h\tau})_{h,\tau}$ is uniformly equi-integrable and thanks to Proposition 4.2 and the continuity of η we obtain

$$\overline{\eta}_{h\tau} = \pi_0 \eta(u_{h\tau}) \xrightarrow[h_\mathcal{T},\tau\to 0]{} \eta(u) \text{ a.e. in } Q_{t_\mathrm{f}}.$$

Applying the Vitali theorem we get (4.11a).

Proving that $\widetilde{\eta}_{h\tau}$ is uniformly equi-integrable is similar and additionally uses that

$$\int_{U_t} E(\widetilde{\pi}_0 u_h^n) \le \int_{U_t} E(\overline{u}_h^n), \qquad \forall t \in (t_{n-1}, t_n],$$

which is a consequence of Jensen's inequality (recall that E is convex).

To conclude the proof, it remains to prove that $\eta(\overline{u}_{h\tau})$ and $\widetilde{\eta}_{h\tau}$ have the same limit. Let $n \in \{0, \cdots, N\}$ and $\boldsymbol{x} \in T \cap s_{\boldsymbol{a}_i}$, $T \in \mathcal{T}$, $\boldsymbol{a}_i \in \mathcal{V}_T$. Then

$$\sqrt{\overline{\eta}_h^n(\boldsymbol{x})} - \sqrt{\widetilde{\eta}_h^n(\boldsymbol{x})} = \sqrt{\eta(u_{\boldsymbol{a}_i}^n)} - \frac{1}{\sqrt{d+1}}\sqrt{\sum_{\boldsymbol{a}\in\mathcal{V}_T} \eta(u_{\boldsymbol{a}}^n)}.$$

Noting that for any $T \in \mathcal{T}$, one has

$$\min_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u_{\boldsymbol{a}}^n)} \le \frac{1}{\sqrt{d+1}}\sqrt{\sum_{\boldsymbol{a}\in\mathcal{V}_T} \eta(u_{\boldsymbol{a}}^n)} \le \max_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u_{\boldsymbol{a}}^n)}$$

this yields

$$\begin{aligned}\left|\sqrt{\overline{\eta}_h^n(\boldsymbol{x})} - \sqrt{\widetilde{\eta}_h^n(\boldsymbol{x})}\right| &\le \max_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u_{\boldsymbol{a}}^n)} - \min_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u_{\boldsymbol{a}}^n)} \\ &= \max_{\boldsymbol{a},\boldsymbol{a}'\in\mathcal{V}_T} \left(\sqrt{\eta(u_{\boldsymbol{a}}^n)} - \sqrt{\eta(u_{\boldsymbol{a}'}^n)}\right).\end{aligned}$$

Let $\boldsymbol{a}, \boldsymbol{a}' \in \mathcal{V}_T$. Since the function $\sqrt{\eta \circ \xi^{-1}}$ is absolutely continuous, one has

$$\begin{aligned} \left|\sqrt{\eta(u^n_{\boldsymbol{a}})} - \sqrt{\eta(u^n_{\boldsymbol{a}'})}\right| &= \left|\sqrt{\eta \circ \xi^{-1}(\xi(u^n_{\boldsymbol{a}}))} - \sqrt{\eta \circ \xi^{-1}(\xi(u^n_{\boldsymbol{a}'}))}\right| \\ &\leq \varpi\left(|\xi(u^n_{\boldsymbol{a}}) - \xi(u^n_{\boldsymbol{a}'})|\right) \leq \varpi(w^n_T), \end{aligned} \tag{4.16}$$

where w^n_T is defined by (4.10), and where ϖ is the modulus of continuity of $\sqrt{\eta \circ \xi^{-1}}$. Then, thanks to Lemma 4.3 we conclude the proof. □

The last lemma of this section is a technical lemma to be used to identify the limit u as a weak solution later on.

Lemma 4.5. *We define $\mu_{h\tau} \in \widetilde{X}_{h\tau}$ such that for any $T \in \mathcal{T}$ and $n \in \{0, \cdots, N\}$,*

$$\mu^n_T := \begin{cases} \max\limits_{\boldsymbol{a},\boldsymbol{a}' \in \mathcal{V}_T} \left(\sqrt{\eta(u^n_{\boldsymbol{a}})} - \sqrt{\eta(u^n_{\boldsymbol{a}'})}\right) & \text{if } u_{\boldsymbol{a}} u_{\boldsymbol{a}'} \geq 0, \ \forall \boldsymbol{a}, \boldsymbol{a}' \in \mathcal{V}_T, \\ \max\limits_{\boldsymbol{a} \in \mathcal{V}_T} \sqrt{\eta(u^n_{\boldsymbol{a}})} & \text{otherwise.} \end{cases} \tag{4.17}$$

Then

$$\mu_{h\tau} \xrightarrow[h_{\mathcal{T}},\tau \to 0]{} 0 \text{ strongly in } L^2(Q_{t_f}).$$

Proof. Thanks to definition (4.17) of $\mu_{h\tau}$ we have, for any $T \in \mathcal{T}$ and $n \in \{0, \cdots, N\}$,

$$(\mu^n_T)^2 \leq \max_{\boldsymbol{a} \in \mathcal{V}_T} \eta(u^n_{\boldsymbol{a}}) \leq \sum_{\boldsymbol{a} \in \mathcal{V}_T} \eta(u^n_{\boldsymbol{a}}) \leq (d+1)\widetilde{\eta}^n_T.$$

Since the sequence $(\widetilde{\eta}_{h\tau})_{h,\tau}$ is uniformly equi-integrable (see Lemma 4.4), the sequence $(\mu_{h\tau})_{h,\tau}$ is uniformly L^2-equi-integrable.

Let again $T \in \mathcal{T}$ and $n \in \{0, \cdots, N\}$. The case where all the products $u^n_{\boldsymbol{a}_i} u^n_{\boldsymbol{a}_j} \geq 0$ (for $i, j \in \{0, \cdots, d\}$) is exactly the same as (4.16). Let us consider the second case. Let $\boldsymbol{a} \in \mathcal{V}_T$ be such that $\eta(u^n_{\boldsymbol{a}}) = \max\limits_{\boldsymbol{a}' \in \mathcal{V}_T} \eta(u^n_{\boldsymbol{a}'})$. Recalling that the semi-Kirchhoff transform ξ defined by (1.15) satisfies $\xi(0) = 0$, one has

$$\sqrt{\eta(u^n_{\boldsymbol{a}})} = \sqrt{\eta \circ \xi^{-1}(\xi(u^n_{\boldsymbol{a}}))} - \sqrt{\eta \circ \xi^{-1}(\xi(0))} \leq \varpi\left(|\xi(u^n_{\boldsymbol{a}})|\right).$$

Let $\boldsymbol{a}' \in \mathcal{V}_T$, $\boldsymbol{a}' \neq \boldsymbol{a}$, be such that $u^n_{\boldsymbol{a}} u^n_{\boldsymbol{a}'} < 0$. Then, since also $\xi(x) \geq 0$ for $x \geq 0$, and $\xi(x) \leq 0$ for $x \leq 0$,

$$|\xi(u^n_{\boldsymbol{a}})| \leq |\xi(u^n_{\boldsymbol{a}}) - \xi(u^n_{\boldsymbol{a}'})|.$$

Thus, since the function ϖ is increasing, we obtain

$$\max_{\boldsymbol{a}' \in \mathcal{V}_T} \sqrt{\eta(u^n_{\boldsymbol{a}'})} = \sqrt{\eta(u^n_{\boldsymbol{a}})} \leq \varpi\left(|\xi(u^n_{\boldsymbol{a}}) - \xi(u^n_{\boldsymbol{a}'})|\right) \leq \varpi(w^n_T),$$

and Lemma 4.3 gives the claim. □

4.3. Identification of the limit as a weak solution. In order to end the proof of Theorem 2.4, it remains to prove that the limit u of u_h exhibited in the previous section is a weak solution in the sense of Definition 1. This is the purpose of the following proposition.

Proposition 4.6. *Let u be the function from Proposition 4.2. Then u is a weak solution of Problem 1.1 in the sense of Definition 1.*

Proof. Let $\varphi \in \mathcal{C}_c^\infty([0,t_f),\overline{\Omega})$ such that $\varphi = 0$ on $[0,t_f)\times\Sigma_D$. We note $\varphi_h^n := \pi_1\varphi(t^n,\cdot)$.

Choosing $v_h = \varphi_h^{n-1}$ in equation (2.23), multiplying by τ_n, summing over all time levels, and integrating discretely by parts the first term, one has

$$\begin{aligned} -\int_\Omega \overline{u}_h^0\overline{\varphi}_h^0 - \sum_{n=1}^N \tau_n \int_\Omega \left(\frac{\overline{\varphi}_h^n - \overline{\varphi}_h^{n-1}}{\tau_n}\right)\overline{u}_h^n \\ + \sum_{n=1}^N \tau_n \int_\Omega \widetilde{\eta}_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\cdot\boldsymbol{\nabla}\varphi_h^{n-1} = \sum_{n=1}^N \tau_n \int_\Omega \overline{f}_h^n \overline{\varphi}_h^{n-1}. \end{aligned} \tag{4.18}$$

First, since $\overline{u}_h^0$ strongly converges towards u_0 in $L^1(\Omega)$ and $\overline{\varphi}_h^0$ converges uniformly towards $\varphi(0,\cdot)$, the first term in (4.18) satisfies

$$-\int_\Omega \overline{u}_h^0\overline{\varphi}_h^0 \xrightarrow[h_{\mathcal{T}}\to 0]{} -\int_\Omega u_0\varphi(0,\cdot). \tag{4.19}$$

Moreover, thanks to the regularity of φ, the piecewise constant function

$$\sum_{n\geq 1} \frac{\overline{\varphi}_h^n - \overline{\varphi}_h^{n-1}}{\tau_n}\mathbf{1}_{(\tau_{n-1},\tau_n]}$$

converges uniformly towards $\partial_t\varphi$. Combining this property with the strong $L^1(Q_{t_f})$ convergence of $u_{h\tau}$ towards u (cf. Proposition 4.2), we obtain that

$$-\sum_{n=1}^N \tau_n \int_\Omega \left(\frac{\overline{\varphi}_h^n - \overline{\varphi}_h^{n-1}}{\tau_n}\right)\overline{u}_h^n \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} -\iint_{Q_{t_f}} u\partial_t\varphi.$$

Let us now consider the third term in the left-hand side of (4.18). We define $\breve{\varphi}_{h\tau} \in V_{h\tau}$ by $\breve{\varphi}_h^0 = \varphi_h^0$ and $\breve{\varphi}_h^n = \varphi_h^{n-1}$. Then one has

$$\begin{aligned} \sum_{n=1}^N \tau_n \int_\Omega \widetilde{\eta}_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla} p_h^n \cdot \boldsymbol{\nabla}\varphi_h^{n-1} = \int_0^{t_f}\int_\Omega \sqrt{\widetilde{\eta}_{h\tau}}\boldsymbol{\Lambda}_h\boldsymbol{\nabla}\xi_{h\tau}\cdot\boldsymbol{\nabla}\breve{\varphi}_{h\tau} \\ + \sum_{n=1}^N \tau_n \int_\Omega \sqrt{\widetilde{\eta}_h^n}\boldsymbol{\Lambda}_h\left(\sqrt{\widetilde{\eta}_h^n}\boldsymbol{\nabla} p_h^n - \boldsymbol{\nabla}\xi_h^n\right)\cdot\boldsymbol{\nabla}\breve{\varphi}_h^n. \end{aligned} \tag{4.20}$$

Owing to Lemma 4.4, one has

$$\sqrt{\widetilde{\eta}_{h\tau}} \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} \sqrt{\eta(u)} \text{ strongly in } L^2(Q_{t_f}),$$
$$\text{and } \boldsymbol{\nabla}\xi_{h\tau} \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} \boldsymbol{\nabla}\xi(u) \text{ weakly in } L^2(Q_{t_f}).$$

Thus, since $\boldsymbol{\nabla}\breve{\varphi}_{h\tau}$ uniformly converges towards $\boldsymbol{\nabla}\varphi$ and $\boldsymbol{\Lambda}_h$ converges almost everywhere towards $\boldsymbol{\Lambda}$, we obtain

$$\int_0^{t_f}\int_\Omega \sqrt{\widetilde{\eta}_{h\tau}}\boldsymbol{\Lambda}_h\boldsymbol{\nabla}\xi_{h\tau}\cdot\boldsymbol{\nabla}\breve{\varphi}_{h\tau} \xrightarrow[h_{\mathcal{T}},\tau\to 0]{} \int_0^{t_f}\int_\Omega \sqrt{\eta(u)}\boldsymbol{\Lambda}\boldsymbol{\nabla}\xi(u)\cdot\boldsymbol{\nabla}\varphi. \tag{4.21}$$

Employing (2.10) and thanks to the definition (1.15) of ξ, for any $T \in \mathcal{T}$ and for any $\boldsymbol{a}_i \in \mathcal{V}_T$, there exists $u_i^n \in [\min(u^n_{\boldsymbol{a}_0}, u^n_{\boldsymbol{a}_i}), \max(u^n_{\boldsymbol{a}_0}, u^n_{\boldsymbol{a}_i})]$ such that

$$\begin{aligned}
\int_\Omega \sqrt{\widetilde{\eta}_h^n}\boldsymbol{\Lambda}_h \left(\sqrt{\widetilde{\eta}_h^n}\boldsymbol{\nabla} p_h^n - \boldsymbol{\nabla}\xi_h^n\right)\cdot \boldsymbol{\nabla}\breve{\varphi}_h^n& \\
&= \sum_{T\in\mathcal{T}} \sqrt{\widetilde{\eta}_T^n} \sum_{i=1}^d \left(\sqrt{\widetilde{\eta}_T^n}(p^n_{\boldsymbol{a}_0} - p^n_{\boldsymbol{a}_i}) - (\xi^n_{\boldsymbol{a}_0} - \xi^n_{\boldsymbol{a}_i})\right) \sum_{j=1}^d \alpha^T_{i,j}(\breve{\varphi}^n_{\boldsymbol{a}_0} - \breve{\varphi}^n_{\boldsymbol{a}_j}) \\
&= \sum_{T\in\mathcal{T}} \sqrt{\widetilde{\eta}_T^n} \sum_{i=1}^d (p^n_{\boldsymbol{a}_0} - p^n_{\boldsymbol{a}_i}) \left(\sqrt{\widetilde{\eta}_T^n} - \sqrt{\eta(u_i^n)}\right) \sum_{j=1}^d \alpha^T_{i,j}(\breve{\varphi}^n_{\boldsymbol{a}_0} - \breve{\varphi}^n_{\boldsymbol{a}_j}).
\end{aligned}$$

We claim that for any $T \in \mathcal{T}$ and for any $\boldsymbol{a}, \boldsymbol{a}' \in \mathcal{V}_T$, one has

$$\left|\sqrt{\widetilde{\eta}_T^n} - \sqrt{\eta(u_i^n)}\right| \le \mu_T^n,$$

where, recall, μ_T^n is given by (4.17). We have to consider two cases.

1. If for any $i, j \in \{0, \cdots, d\}$, $u^n_{\boldsymbol{a}_i} u^n_{\boldsymbol{a}_j} \ge 0$, then

$$\min\left(\sqrt{\eta(u^n_{\boldsymbol{a}_0})}, \sqrt{\eta(u^n_{\boldsymbol{a}_i})}\right) \le \sqrt{\eta(u_i^n)} \le \max\left(\sqrt{\eta(u^n_{\boldsymbol{a}_0})}, \sqrt{\eta(u^n_{\boldsymbol{a}_i})}\right)$$

and definition (4.17) of $\mu_{h\tau}$ gives

$$\left|\sqrt{\widetilde{\eta}_T^n} - \sqrt{\eta(u_i^n)}\right| \le \max_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u^n_{\boldsymbol{a}})} - \min_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u^n_{\boldsymbol{a}})} \le \mu_T^n.$$

2. If there exists $i, j \in \{0, \cdots, d\}$ such that $u^n_{\boldsymbol{a}_i} u^n_{\boldsymbol{a}_j} \le 0$, then one has

$$0 \le \sqrt{\widetilde{\eta}_T^n} \le \max_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u^n_{\boldsymbol{a}})},$$

and

$$0 \le \sqrt{\eta(u_i^n)} \le \max\left(\sqrt{\eta(u^n_{\boldsymbol{a}_0})}, \sqrt{\eta(u^n_{\boldsymbol{a}_i})}\right) \le \max_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u^n_{\boldsymbol{a}})}.$$

Thus we obtain,

$$\left|\sqrt{\widetilde{\eta}_T^n} - \sqrt{\eta(u_i^n)}\right| \le \max_{\boldsymbol{a}\in\mathcal{V}_T} \sqrt{\eta(u^n_{\boldsymbol{a}})} = \mu_T^n.$$

Hence, the Cauchy–Schwarz inequality gives

$$\begin{aligned}
\sum_{n=1}^N \tau_n \int_\Omega \sqrt{\widetilde{\eta}_h^n}\boldsymbol{\Lambda}_h \left(\sqrt{\widetilde{\eta}_h^n}\boldsymbol{\nabla} p_h^n - \boldsymbol{\nabla}\xi_h^n\right)\cdot \boldsymbol{\nabla}\breve{\varphi}_h^n& \\
&\le \left(\sum_{n=1}^N \tau_n \sum_{T\in\mathcal{T}} \widetilde{\eta}_T^n \sum_{i=1}^d \left(\sum_{j=1}^d |\alpha^T_{i,j}|\right)(p^n_{\boldsymbol{a}_0} - p^n_{\boldsymbol{a}_i})^2\right)^{\frac{1}{2}} \\
&\quad \left(\sum_{n=1}^N \tau_n \sum_{T\in\mathcal{T}} (\mu_T^n)^2 \sum_{j=1}^d \left(\sum_{i=1}^d |\alpha^T_{i,j}|\right)(\breve{\varphi}^n_{\boldsymbol{a}_0} - \breve{\varphi}^n_{\boldsymbol{a}_j})^2\right)^{\frac{1}{2}}.
\end{aligned}$$

Thanks to (4.5), (2.12), (2.10), and the regularity of φ, one has

$$\begin{aligned}\sum_{n=1}^{N} \tau_n \int_\Omega \sqrt{\widetilde{\eta}_h^n} \mathbf{\Lambda}_h \left(\sqrt{\widetilde{\eta}_h^n} \boldsymbol{\nabla} p_h^n - \boldsymbol{\nabla} \xi_h^n\right) \cdot \boldsymbol{\nabla} \breve{\varphi}_h^n \\ &\leq \sqrt{C_{10} C_2} \left(\sum_{n=1}^{N} \tau_n \int_\Omega (\mu_h^n)^2 \mathbf{\Lambda}_h \boldsymbol{\nabla} \breve{\varphi}_h^n \cdot \boldsymbol{\nabla} \breve{\varphi}_h^n \right)^{\frac{1}{2}} \\ &\leq C' \left\| \mu_{h\tau} \right\|_{L^2(Q_{t_\mathrm{f}})} .\end{aligned}$$

Hence, owing to Lemma 4.5, we obtain

$$\sum_{n=1}^{N} \tau_n \int_\Omega \sqrt{\widetilde{\eta}_h^n} \mathbf{\Lambda}_h \left(\sqrt{\widetilde{\eta}_h^n} \boldsymbol{\nabla} p_h^n - \boldsymbol{\nabla} \xi_h^n\right) \cdot \boldsymbol{\nabla} \breve{\varphi}_h^n \xrightarrow[h_\mathcal{T}, \tau \to 0]{} 0,$$

and, since (4.21) holds, we obtain from (4.20)

$$\sum_{n=1}^{N} \tau_n \int_\Omega \widetilde{\eta}_h^n \mathbf{\Lambda}_h \boldsymbol{\nabla} p_h^n \cdot \boldsymbol{\nabla} \varphi_h^{n-1} \xrightarrow[h_\mathcal{T}, \tau \to 0]{} \int_0^{t_\mathrm{f}} \int_\Omega \sqrt{\eta(u)} \mathbf{\Lambda} \boldsymbol{\nabla} \xi(u) \cdot \boldsymbol{\nabla} \varphi. \tag{4.22}$$

Since $\widetilde{\eta}_{h\tau}$ strongly converges towards $\eta(u)$ in $L^1(Q_{t_\mathrm{f}})$ (see Lemma 4.4) and $\boldsymbol{\nabla} \Psi_h$ (resp. $\boldsymbol{\nabla} \breve{\varphi}_h^n$) uniformly converges towards $\boldsymbol{\nabla} \Psi$ (resp. $\boldsymbol{\nabla} \varphi$), we also have

$$\sum_{n=1}^{N} \tau_n \int_\Omega \widetilde{\eta}_h^n \mathbf{\Lambda}_h \boldsymbol{\nabla} \Psi_h \cdot \boldsymbol{\nabla} \varphi_h^{n-1} \xrightarrow[h_\mathcal{T}, \tau \to 0]{} \int_0^{t_\mathrm{f}} \int_\Omega \eta(u) \mathbf{\Lambda} \boldsymbol{\nabla} \Psi \cdot \boldsymbol{\nabla} \varphi. \tag{4.23}$$

Finally, we have to deal with the right-hand side of (4.18). We note from (1.10) and (2.16) that

$$\begin{aligned}\int_0^{t_\mathrm{f}} \int_\Omega \left| f(u) - \overline{f}_{h\tau} \right| \leq & \sum_{n=1}^{N} \sum_{\boldsymbol{a} \in \mathcal{V}_\mathcal{T}} \int_{t^{n-1}}^{t^n} \int_{s_{\boldsymbol{a}}} \left| f_\mathrm{inj} - f_{\mathrm{inj},\boldsymbol{a}}^n \right| \\ & + \sum_{n=1}^{N} \sum_{\boldsymbol{a} \in \mathcal{V}_\mathcal{T}} \int_{t^{n-1}}^{t^n} \int_{s_{\boldsymbol{a}}} \left| \eta(u^+) - \eta\left((u_{\boldsymbol{a}}^n)^+\right) \right| f_\mathrm{out} \\ & + \sum_{n=1}^{N} \sum_{\boldsymbol{a} \in \mathcal{V}_\mathcal{T}} \int_{t^{n-1}}^{t^n} \int_{s_{\boldsymbol{a}}} \eta\left((u_{\boldsymbol{a}}^n)^+\right) \left| f_\mathrm{out}(t, \boldsymbol{x}) - f_{\mathrm{out},\boldsymbol{a}}^n \right| .\end{aligned}$$

The function $\overline{f}_{\mathrm{inj},h\tau}$ converges strongly towards f_inj in $L^1(Q_{t_\mathrm{f}})$ (see for instance [17, Appendix B]). Furthermore, thanks to a similar reasoning to that given in Lemma 4.4, we can prove that $\eta((\overline{u}_{h\tau})^+)$ strongly converges in $L^1(Q_{t_\mathrm{f}})$ towards $\eta(u^+)$. Thus, since f_out is bounded in $L^\infty(Q_{t_\mathrm{f}})$, the second term in the right-hand side of this inequality tends to 0. Finally, since $\overline{f}_{\mathrm{out},h\tau}$ converges almost everywhere towards f_out, the quantity $\overline{f}_{\mathrm{out},h\tau} - f_\mathrm{out}$ is bounded in $L^\infty(Q_{t_\mathrm{f}})$ and $\eta((\overline{u}_{h\tau})^+)$ is uniformly equi-integrable. We conclude that the last term in the right-hand side tends to 0, and this finishes the proof. □

5. Flux reconstruction and a posteriori error indicator. We derive now an a posteriori error estimate for the discretization of problem (1.1) by the $\mathbb{P}_1$ finite elements with mass lumping (2.22).

5.1. Equilibrated flux reconstruction. We first devise an equilibrated flux reconstruction in the sense of Theorem 2.3. We will use for this purpose the space

$$\mathbf{RTN}_1 := \{\boldsymbol{v}_h \in \mathbf{H}(\mathrm{div},\Omega) : \ \boldsymbol{v}_h|_T \in \mathbf{RTN}_1(T), \ \forall T \in \mathcal{T}\}, \tag{5.1}$$

where $\mathbf{RTN}_1(T) := [\mathbb{P}_1(T)]^d + \boldsymbol{x}\mathbb{P}_1(T)$ is the Raviart–Thomas–Nédélec finite element space of order 1. We extend to the present setting the procedure from [22, 8, 28].

Proposition 5.1 (Space-time equilibrated flux reconstruction). *There exists a locally defined flux reconstruction $\boldsymbol{\sigma}_{h\tau} \in L^2((0,t_{\mathrm{f}});\mathbf{H}(\mathrm{div},\Omega))$, piecewise constant in time with values in $\mathbf{RTN}_1$, satisfying*

$$\partial_t \hat{u}_{h\tau} + \boldsymbol{\nabla}{\cdot}\boldsymbol{\sigma}_{h\tau} = f_{h\tau} \ \text{in } \Omega \quad \text{and} \quad \boldsymbol{\sigma}_{h\tau}{\cdot}\boldsymbol{n} = 0 \ \text{on } (0,t_{\mathrm{f}}) \times \Sigma_{\mathrm{N}}. \tag{5.2}$$

Remark 5.2. *We believe that the equilibration property (5.2) is remarkable: the scheme (2.22) uses mass lumping for both the time and source terms, and yet we recover that the divergence of the flux $\boldsymbol{\sigma}_{h\tau}$ is equal to the piecewise affine-in-space $f_{h\tau} - \partial_t\hat{u}_{h\tau}$.*

Proof. We consider the hat function $\phi_{\boldsymbol{a}}$ associated to the vertex $\boldsymbol{a}$, defined in (2.2), as test function in equation (2.22). Thanks to (2.3), we obtain the following hat-function orthogonality: for any $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}} \backslash \mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}$,

$$\int_{\omega_{\boldsymbol{a}}} \frac{u_{\boldsymbol{a}}^n - u_{\boldsymbol{a}}^{n-1}}{\tau_n}\phi_{\boldsymbol{a}} + \int_{\omega_{\boldsymbol{a}}} \eta_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h){\cdot}\boldsymbol{\nabla}\phi_{\boldsymbol{a}} = \int_{\omega_{\boldsymbol{a}}} f_{\boldsymbol{a}}^n \phi_{\boldsymbol{a}}. \tag{5.3}$$

Denote by $Q_h := \mathbb{P}_1(\mathcal{T})$ the broken polynomial space spanned by the functions $v_h \in L^1(\Omega)$ such that for any $T \in \mathcal{T}$, $v_h|_T \in \mathbb{P}_1(T)$. We will also use the shorthand notation $\mathbf{W}_h := \mathbf{RTN}_1$ and let $\mathbf{W}_h|_{\omega_{\boldsymbol{a}}}$ (resp. $Q_h|_{\omega_{\boldsymbol{a}}}$) be the restriction of $\mathbf{W}_h$ (resp. Q_h) to the patch $\omega_{\boldsymbol{a}}$, $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}$.

For any $n \in \{1, \cdots, N\}$, we define the flux reconstruction $\boldsymbol{\sigma}_h^n$ as follows

$$\boldsymbol{\sigma}_h^n := \sum_{\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}} \boldsymbol{\sigma}_{h,\boldsymbol{a}}^n; \tag{5.4}$$

then $\boldsymbol{\sigma}_{h\tau}$ is piecewise constant in time, given by $\boldsymbol{\sigma}_h^n$ on any $(t_{n-1}, t_n]$. For any $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}$, the patchwise contributions are given by the solution of the following mixed finite element Laplace problems with homogeneous Neumann boundary condition (except for Dirichlet boundary vertices): find $\boldsymbol{\sigma}_{h,\boldsymbol{a}}^n \in \mathbf{W}_h^a$ and $r_h^{\boldsymbol{a}} \in Q_h^a$ such that

$$\int_{\omega_{\boldsymbol{a}}} \boldsymbol{\sigma}_{h,\boldsymbol{a}}^n \boldsymbol{v}_h - \int_{\omega_{\boldsymbol{a}}} \boldsymbol{\nabla}{\cdot}\boldsymbol{v}_h r_h^{\boldsymbol{a}} = -\int_{\omega_{\boldsymbol{a}}} \phi_{\boldsymbol{a}} \eta_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h)\boldsymbol{v}_h, \tag{5.5a}$$

$$\int_{\omega_{\boldsymbol{a}}} \boldsymbol{\nabla}{\cdot}\boldsymbol{\sigma}_{h,\boldsymbol{a}}^n q_h = \int_{\omega_{\boldsymbol{a}}} \left[\left(f_{\boldsymbol{a}}^n - \frac{u_{\boldsymbol{a}}^n - u_{\boldsymbol{a}}^{n-1}}{\tau_n}\right)\phi_{\boldsymbol{a}} - \eta_h^n \boldsymbol{\Lambda}_h \boldsymbol{\nabla}(p_h^n + \Psi_h){\cdot}\boldsymbol{\nabla}\phi_{\boldsymbol{a}}\right] q_h, \tag{5.5b}$$

for all $\boldsymbol{v}_h \in \mathbf{W}_h^a$ and $q_h \in Q_h^a$, where

$$\mathbf{W}_h^a := \begin{cases} \{\boldsymbol{v}_h \in \mathbf{W}_h|_{\omega_{\boldsymbol{a}}} : \boldsymbol{v}_h{\cdot}\boldsymbol{n}_{\omega_{\boldsymbol{a}}} = 0 \text{ on } \partial\omega_{\boldsymbol{a}}\}, & \text{if } \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}^{\mathrm{int}}, \\ \{\boldsymbol{v}_h \in \mathbf{W}_h|_{\omega_{\boldsymbol{a}}} : \boldsymbol{v}_h{\cdot}\boldsymbol{n}_{\omega_{\boldsymbol{a}}} = 0 \text{ on } \partial\omega_{\boldsymbol{a}} \backslash \Sigma_{\mathrm{D}}\}, & \text{if } \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext}}, \end{cases}$$

$$Q_h^a := \begin{cases} \left\{q_h \in Q_h|_{\omega_{\boldsymbol{a}}} \text{ s.t. } \displaystyle\int_{\omega_{\boldsymbol{a}}} q_h = 0\right\}, & \text{if } \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}} \backslash \mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}, \\ \qquad\qquad\qquad\qquad Q_h|_{\omega_{\boldsymbol{a}}}, & \text{if } \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}. \end{cases}$$

If the vertex $\boldsymbol{a}$ lies inside Ω or inside the Neumann boundary Σ_{N}, $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}} \backslash \mathcal{V}_{\mathcal{T}}^{\mathrm{ext,D}}$, then, since $\boldsymbol{\sigma}_{h,\boldsymbol{a}}^n \in \mathbf{W}_h^a$, we have $\boldsymbol{\sigma}_{h,\boldsymbol{a}}^n \cdot \boldsymbol{n}_{\omega_{\boldsymbol{a}}} = 0$ on $\partial\omega_{\boldsymbol{a}}$. Thus, using the hat-function orthogonality (5.3) and the Green theorem, we remark that equation (5.5b) also holds for constants on the patch $\omega_{\boldsymbol{a}}$, and consequently for all functions in $Q_h|_{\omega_{\boldsymbol{a}}}$ (and not only those with mean value zero).

Thanks to the definition of the flux reconstruction, it is clear that $\boldsymbol{\sigma}_h^n \in \mathbf{H}(\mathrm{div}, \Omega)$. Furthermore, let $T \in \mathcal{T}$ and $q_h \in Q_h(T)$; remark that the space Q_h is discontinuous, so that choosing q_h only supported on one element T is possible. Then, since $\sum_{\boldsymbol{a} \in \mathcal{V}_T} \phi_{\boldsymbol{a}} = 1$, we obtain from (5.4) and (5.5b)

$$\int_T \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h^n q_h = \sum_{\boldsymbol{a}\in\mathcal{V}_T} \int_T \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{h,\boldsymbol{a}}^n q_h = \sum_{\boldsymbol{a}\in\mathcal{V}_T} \int_T \left(f_{\boldsymbol{a}}^n - \frac{u_{\boldsymbol{a}}^n - u_{\boldsymbol{a}}^{n-1}}{\tau_n} \right) \phi_{\boldsymbol{a}} q_h.$$

Since $u_h^n, u_h^{n-1}, f_h^n \in V_h$ are respectively prescribed by (2.21) and (2.18), we infer

$$\int_T \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h^n q_h = \int_T \left(f_h^n - \frac{u_h^n - u_h^{n-1}}{\tau_n} \right) q_h,$$

and the claim follows. □

5.2. Guaranteed a posteriori error estimate. We are now in position to obtain the error upper bound on the residual. We consider the space X defined in (1.20), associated with the norm

$$\|\varphi\|_X := \|\boldsymbol{\nabla}\varphi\|_{L^\infty(Q_{t_{\mathrm{f}}})} + \int_0^{t_{\mathrm{f}}} \|\partial_t \varphi\|_{L^\infty(\Omega)}, \qquad \varphi \in X.$$

Let v and $\eta(v)$ belong to $L^\infty((0,t_{\mathrm{f}}); L^1(\Omega))$ and $\xi(v)$ belong to $L^2((0,t_{\mathrm{f}}); H^1(\Omega))$ with $\xi(v) = \xi(u_{\mathrm{D}})$ a.e. on $(0,t_{\mathrm{f}}) \times \Sigma_{\mathrm{D}}$. We define the residual $R(v) \in X'$ such that for any $\varphi \in X$,

$$\begin{aligned}\langle R(v), \varphi \rangle_{X',X} := \iint_{Q_{t_{\mathrm{f}}}} v \partial_t \varphi + \int_\Omega u_0 \varphi(0,\cdot) \\ - \iint_{Q_{t_{\mathrm{f}}}} (\boldsymbol{\nabla}\gamma(v) + \eta(v)\boldsymbol{\nabla}\Psi) \cdot \boldsymbol{\Lambda}\boldsymbol{\nabla}\varphi + \iint_{Q_{t_{\mathrm{f}}}} f(v)\varphi.\end{aligned}$$

We note that the residual vanishes if and only if v is solution to the weak formulation (1.19). Then the error measure $\mathcal{J}(\hat{u}_{h\tau})$ is the dual norm of the residual defined by

$$\mathcal{J}(\hat{u}_{h\tau}) := \sup_{\varphi \in X, \|\varphi\|_X = 1} \langle R(\hat{u}_{h\tau}), \varphi \rangle_{X',X}. \tag{5.6}$$

Note that for linear problems, one can typically identify a suitable setting such that the dual norm of the residual is the difference $u - \hat{u}_{h\tau}$ measured in a norm, see, e.g., [26, Theorem 2.1 and relation (2.7)] and the references therein.

Proof (of Theorem 2.5). Let $\varphi \in X$ be such that $\|\varphi\|_X = 1$. Since $\varphi = 0$ on $(0,t_{\mathrm{f}}) \times \Sigma_{\mathrm{D}}$ and $\boldsymbol{\sigma}_{h\tau}\cdot\boldsymbol{n} = 0$ on $(0,t_{\mathrm{f}}) \times \Sigma_{\mathrm{N}}$, the Green formula gives

$$\iint_{Q_{t_{\mathrm{f}}}} \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{h\tau}\varphi + \iint_{Q_{t_{\mathrm{f}}}} \boldsymbol{\sigma}_{h\tau}\cdot\boldsymbol{\nabla}\varphi = 0.$$

By integration by parts

$$\iint_{Q_{t_{\mathrm{f}}}} \partial_t \hat{u}_{h\tau}\varphi + \iint_{Q_{t_{\mathrm{f}}}} \hat{u}_{h\tau}\partial_t\varphi = -\int_\Omega \hat{u}_{h\tau}(0,\cdot)\varphi(0,\cdot).$$

Moreover, since $\varphi(0,\cdot) = -\displaystyle\int_0^{t_{\mathrm{f}}} \partial_t\varphi$, the residual can be written as

$$\begin{aligned}\langle R(\hat{u}_{h\tau}),\varphi\rangle_{X',X} = &\iint_{Q_{t_{\mathrm{f}}}} \left(f_{h\tau} - \partial_t\hat{u}_{h\tau} - \boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_{h\tau}\right)\varphi \\ &- \iint_{Q_{t_{\mathrm{f}}}} \left(\boldsymbol{\Lambda}(\boldsymbol{\nabla}\gamma(\hat{u}_{h\tau}) + \eta(\hat{u}_{h\tau})\boldsymbol{\nabla}\Psi) + \boldsymbol{\sigma}_{h\tau}\right)\cdot\boldsymbol{\nabla}\varphi \\ &+ \iint_{Q_{t_{\mathrm{f}}}} \left(\hat{u}_{h\tau}(0,\cdot) - u_0\right)\partial_t\varphi + \iint_{Q_{t_{\mathrm{f}}}} \left(f(\hat{u}_{h\tau}) - f_{h\tau}\right)\varphi.\end{aligned}$$

Thanks to Proposition 5.1, the first term vanishes. Using that $\|\varphi\|_X = 1$, the two next terms satisfy

$$\begin{aligned}-&\iint_{Q_{t_{\mathrm{f}}}} \left(\boldsymbol{\Lambda}(\boldsymbol{\nabla}\gamma(\hat{u}_{h\tau}) + \eta(\hat{u}_{h\tau})\boldsymbol{\nabla}\Psi) + \boldsymbol{\sigma}_{h\tau}\right)\cdot\boldsymbol{\nabla}\varphi \\ &\qquad \le \|\boldsymbol{\nabla}\varphi\|_{L^\infty(Q_{t_{\mathrm{f}}})} \iint_{Q_{t_{\mathrm{f}}}} \left|\boldsymbol{\Lambda}(\boldsymbol{\nabla}\gamma(\hat{u}_{h\tau}) + \eta(\hat{u}_{h\tau})\boldsymbol{\nabla}\Psi) + \boldsymbol{\sigma}_{h\tau}\right| \le \eta_{\mathrm{F}},\end{aligned}$$

and

$$\iint_{Q_{t_{\mathrm{f}}}} \left(\hat{u}_{h\tau}(0,\cdot) - u_0\right)\partial_t\varphi \le \int_0^{t_{\mathrm{f}}} \|\partial_t\varphi\|_{L^\infty(\Omega)} \int_\Omega \left|u_{h\tau}(0,\cdot) - u_0\right| \le \eta_{\mathrm{IC}}.$$

To finish, we have to deal with the last term due to the right-hand side f. This can be decomposed, using (1.10) and (2.16)–(2.18) as

$$\begin{aligned}\iint_{Q_{t_{\mathrm{f}}}} \left(f(\hat{u}_{h\tau}) - f_{h\tau}\right)\varphi = &\iint_{Q_{t_{\mathrm{f}}}} \left(f_{\mathrm{inj}} - f_{\mathrm{inj},h\tau}\right)\varphi \\ &- \sum_{n=1}^N \int_{t^{n-1}}^{t^n}\int_\Omega \left(\eta((\hat{u}_{h\tau})^+)f_{\mathrm{out}} - \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \eta((u_{\boldsymbol{a}}^n)^+) f_{\mathrm{out},\boldsymbol{a}}^n\phi_{\boldsymbol{a}}\right)\varphi.\end{aligned}$$

Concerning the term related to f_{inj}, using that $\displaystyle\int_{t^{n-1}}^{t^n}\int_{\omega_{\boldsymbol{a}}} (f_{\mathrm{inj}} - f_{\mathrm{inj},\boldsymbol{a}}^n)\phi_{\boldsymbol{a}} = 0$ (which follows from (2.17) and (2.3)) and $\sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \phi_{\boldsymbol{a}} = 1$, one has

$$
\begin{aligned}
\iint_{Q_{t_\mathrm{f}}} (f_\mathrm{inj} - f_{\mathrm{inj},h\tau})\,\varphi &= \sum_{n=1}^{N} \int_{t^{n-1}}^{t^n} \int_\Omega \left(f_\mathrm{inj} - \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} f^n_{\mathrm{inj},\boldsymbol{a}} \phi_{\boldsymbol{a}} \right) \varphi \\
&= \sum_{n=1}^{N} \int_{t^{n-1}}^{t^n} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \int_{\omega_{\boldsymbol{a}}} \left(f_\mathrm{inj} - f^n_{\mathrm{inj},\boldsymbol{a}} \right) \phi_{\boldsymbol{a}} \varphi \\
&= \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \int_{t^{n-1}}^{t^n} \int_{\omega_{\boldsymbol{a}}} \left(f_\mathrm{inj} - f^n_{\mathrm{inj},\boldsymbol{a}} \right) \phi_{\boldsymbol{a}} \left(\varphi - \varphi_{\omega_{\boldsymbol{a}}}(t_n) \right) \\
&= \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \int_{t^{n-1}}^{t^n} \int_{\omega_{\boldsymbol{a}}} \left(f_\mathrm{inj} - f^n_{\mathrm{inj},\boldsymbol{a}} \right) \phi_{\boldsymbol{a}} \left(\varphi - \varphi(t_n,\cdot) + \varphi(t_n,\cdot) - \varphi_{\omega_{\boldsymbol{a}}}(t_n) \right) \\
&\leq \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \|\phi_{\boldsymbol{a}}\|_{L^\infty(\omega_{\boldsymbol{a}})} \int_{t^{n-1}}^{t^n} \int_{\omega_{\boldsymbol{a}}} \left\{ \left| f_\mathrm{inj} - f^n_{\mathrm{inj},\boldsymbol{a}} \right| \left| \int_{t_n}^{t} \partial_t \varphi \right| \right\} \\
&+ \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} h_{\omega_{\boldsymbol{a}}} \|\phi_{\boldsymbol{a}}\|_{L^\infty(\omega_{\boldsymbol{a}})} \|\boldsymbol{\nabla}\varphi(t_n,\cdot)\|_{L^\infty(\omega_{\boldsymbol{a}})} \int_{t^{n-1}}^{t^n} \int_{\omega_{\boldsymbol{a}}} \left| f_\mathrm{inj} - f^n_{\mathrm{inj},\boldsymbol{a}} \right| \leq \eta_{f_\mathrm{inj}},
\end{aligned}
$$

where $\varphi_{\omega_{\boldsymbol{a}}}(t_n) := \frac{1}{|\omega_{\boldsymbol{a}}|} \int_{\omega_{\boldsymbol{a}}} \varphi(t_n,\cdot)$. Finally,

$$
\begin{aligned}
&\sum_{n=1}^{N} \int_{t^{n-1}}^{t^n} \int_\Omega \left(\eta((u_{h\tau})^+) f_\mathrm{out} - \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \eta((u^n_{\boldsymbol{a}})^+) f^n_{\mathrm{out},\boldsymbol{a}} \phi_{\boldsymbol{a}} \right) \varphi \\
&\qquad = \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \int_{t^{n-1}}^{t^n} \int_{\omega_{\boldsymbol{a}}} \left(\eta((u_{h\tau})^+) - \eta((u^n_{\boldsymbol{a}})^+) \right) f_\mathrm{out} \phi_{\boldsymbol{a}} \varphi \\
&\qquad\qquad + \sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \eta((u^n_{\boldsymbol{a}})^+) \int_{t^{n-1}}^{t^n} \int_{\omega_{\boldsymbol{a}}} \left(f_\mathrm{out} - f^n_{\mathrm{out},\boldsymbol{a}} \right) \phi_{\boldsymbol{a}} \varphi.
\end{aligned}
$$

The second term is treated exactly as the term with f_inj above and leads to the last two terms in (2.30b). To treat the first term, we note that

$$
\begin{aligned}
&\sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \int_{t^{n-1}}^{t^n} \int_{\omega_{\boldsymbol{a}}} \left(\eta((u_{h\tau})^+) - \eta((u^n_{\boldsymbol{a}})^+) \right) f_\mathrm{out} \phi_{\boldsymbol{a}} \varphi \\
&= -\sum_{n=1}^{N} \sum_{\boldsymbol{a}\in\mathcal{V}_\mathcal{T}} \int_{t^{n-1}}^{t^n} \int_{\omega_{\boldsymbol{a}}} \left\{ \left(\eta((u_{h\tau})^+) - \eta((u^n_{\boldsymbol{a}})^+) \right) f_\mathrm{out} \phi_{\boldsymbol{a}} \int_t^{t_\mathrm{f}} \partial_t \varphi \right\}.
\end{aligned}
$$

Using that $\int_0^{t_\mathrm{f}} \|\partial_t \varphi\|_{L^\infty(\Omega)}\,\mathrm{d}t \leq 1$, we obtain the first term in (2.30b), and the proof is finished. □

Remark 5.3. *Note that in the error estimators (2.30), one could further distinguish the different error components (spatial, temporal, numerical quadrature and possibly also linearization and algebraic). This is possible following [27, 16, 23] and the references therein.*

6. Numerical results.

We present here the results of several numerical experiments using scheme (2.22) in the 2-dimensional case. We use the FreeFem++ software (see [39]). For linearization, we employ the Newton method, and GMRES or UMFPACK is the employed algebraic solver. We give now some precisions on the Newton algorithm. First, we recall that for a given $u_h^{n-1} \in V_h^{\mathrm{D},n-1}$, we construct with the Newton method a sequence $(u_h^{n,\ell})_{\ell \geq 0}$ which should converge towards u_h^n. In this paper, we first choose a rather classical stopping criterion on the L^∞-norm between two successive iterations, i.e., the algorithm said to numerically converge and is stopped if

$$\|u_h^{n,\ell+1} - u_h^{n,\ell}\|_\infty \leq \varepsilon; \tag{6.1}$$

in the computations, we use $\varepsilon = 10^{-8}$. Later in Section 6.3, we then show how a posteriori error estimates can be used to design adaptive stopping criteria for the Newton iterative linearization, in place of (6.1). Thanks to Lemma 3.3, we know that when p is singular, for any vertex $\boldsymbol{a} \in \mathcal{V}_\mathcal{T}$ and any time step $n \in \{0, \cdots, N\}$, the discrete nodal values satisfy $u_{\boldsymbol{a}}^n > 0$; thus in this case we initialize the Newton method as follows: $u_{\boldsymbol{a}}^{n,0} = \max(u_{\boldsymbol{a}}^{n-1}, 10^{-12})$. Otherwise we choose $u_h^{n,0} = u_h^{n-1}$.

6.1. Convergence orders and a posteriori error estimates for known solutions.

We begin with several cases where we know the exact solution and for which the source term vanishes, that is $f_{\mathrm{inj}} = f_{\mathrm{out}} = 0$. In each case, we (approximately) compute the error between the exact and the approximate solution for the following norms: $L^1((0,t_{\mathrm{f}}) \times \Omega)$, $L^2((0,t_{\mathrm{f}}) \times \Omega)$, and $L^\infty((0,t_{\mathrm{f}}) \times \Omega)$, as well as the corresponding rates of convergence. The quantity N_{v} is the number of vertices in the mesh.

First, we consider the unit square $\Omega =]0,1[^2$ whose computational mesh $\mathcal{T}$ is constituted by triangles. We choose the functions $\eta(u) = u$, $p(u) = \log(u)$, and $\Psi(x,y) = -gx$, that is we consider the following linear Fokker–Planck equation (but discretized in a nonlinear form):

$$\partial_t u - \boldsymbol{\nabla}\cdot(\boldsymbol{\Lambda}(\boldsymbol{\nabla} u - \boldsymbol{g}_x u)) = 0 \quad \text{where } \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_x & 0 \\ 0 & \lambda_y \end{pmatrix} \text{ and } \boldsymbol{g}_x = \begin{pmatrix} g \\ 0 \end{pmatrix}.$$

The exact solution of this problem with no-flux boundary condition is given by the 1-dimensional function

$$u(t,(x,y)) = e^{-\beta t + \frac{g}{2}x}\left(\pi\cos(\pi x) + \tfrac{g}{2}\sin(\pi x)\right) + \pi e^{g(x-\frac{1}{2})},$$

with $\beta = \lambda_x(\pi^2 + \frac{g}{4})$, $g = 1$, $\lambda_x = 1$, and the final time $t_{\mathrm{f}} = 0.25$. We test the numerical scheme for several values of the coefficient λ_y.

τ	$h_\mathcal{T}$	N_{v}	$\min(u_{h\tau})$	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
$1 \cdot 10^{-2}$	0.354	25	0.434	$2.806 \cdot 10^{-2}$	–	$3.637 \cdot 10^{-2}$	–	0.108	–
$2.5 \cdot 10^{-3}$	0.177	81	0.128	$7.414 \cdot 10^{-3}$	2.265	$9.658 \cdot 10^{-3}$	2.256	$3.485 \cdot 10^{-2}$	1.932
$6.25 \cdot 10^{-4}$	$8.839 \cdot 10^{-2}$	289	$3.391 \cdot 10^{-2}$	$1.886 \cdot 10^{-3}$	2.152	$2.455 \cdot 10^{-3}$	2.154	$1.077 \cdot 10^{-2}$	1.846
$1.563 \cdot 10^{-4}$	$4.419 \cdot 10^{-2}$	1089	$8.651 \cdot 10^{-3}$	$4.743 \cdot 10^{-4}$	2.081	$6.167 \cdot 10^{-4}$	2.083	$3.27 \cdot 10^{-3}$	1.797
$3.906 \cdot 10^{-5}$	$2.21 \cdot 10^{-2}$	4225	$2.18 \cdot 10^{-3}$	$1.188 \cdot 10^{-4}$	2.042	$1.544 \cdot 10^{-4}$	2.043	$9.8 \cdot 10^{-4}$	1.778

Table 1: The linear Fokker–Planck equation with $\lambda_y = 0.1$

τ	$h_{\mathcal{T}}$	N_{v}	$\min(u_{h\tau})$	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
$1 \cdot 10^{-2}$	0.354	25	0.456	$2.26 \cdot 10^{-2}$	–	$2.773 \cdot 10^{-2}$	–	$5.335 \cdot 10^{-2}$	–
$2.5 \cdot 10^{-3}$	0.177	81	0.133	$5.772 \cdot 10^{-3}$	2.322	$7.017 \cdot 10^{-3}$	2.338	$1.354 \cdot 10^{-2}$	2.332
$6.25 \cdot 10^{-4}$	$8.839 \cdot 10^{-2}$	289	$3.493 \cdot 10^{-2}$	$1.454 \cdot 10^{-3}$	2.168	$1.758 \cdot 10^{-3}$	2.176	$3.516 \cdot 10^{-3}$	2.121
$1.563 \cdot 10^{-4}$	$4.419 \cdot 10^{-2}$	1089	$8.895 \cdot 10^{-3}$	$3.645 \cdot 10^{-4}$	2.086	$4.4 \cdot 10^{-4}$	2.088	$9.201 \cdot 10^{-4}$	2.021
$3.906 \cdot 10^{-5}$	$2.21 \cdot 10^{-2}$	4225	$2.24 \cdot 10^{-3}$	$9.12 \cdot 10^{-5}$	2.044	$1.101 \cdot 10^{-4}$	2.044	$2.414 \cdot 10^{-4}$	1.974

Table 2: The linear Fokker–Planck equation with $\lambda_y = 10$

τ	$h_{\mathcal{T}}$	N_{v}	$\min(u_{h\tau})$	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
$1 \cdot 10^{-2}$	0.354	25	0.467	$2.253 \cdot 10^{-2}$	–	$2.758 \cdot 10^{-2}$	–	$4.769 \cdot 10^{-2}$	–
$2.5 \cdot 10^{-3}$	0.177	81	0.135	$5.76 \cdot 10^{-3}$	2.321	$6.986 \cdot 10^{-3}$	2.336	$1.153 \cdot 10^{-2}$	2.415
$6.25 \cdot 10^{-4}$	$8.839 \cdot 10^{-2}$	289	$3.548 \cdot 10^{-2}$	$1.451 \cdot 10^{-3}$	2.167	$1.752 \cdot 10^{-3}$	2.175	$2.891 \cdot 10^{-3}$	2.176
$1.563 \cdot 10^{-4}$	$4.419 \cdot 10^{-2}$	1089	$9.028 \cdot 10^{-3}$	$3.639 \cdot 10^{-4}$	2.086	$4.385 \cdot 10^{-4}$	2.088	$7.323 \cdot 10^{-4}$	2.070
$3.906 \cdot 10^{-5}$	$2.21 \cdot 10^{-2}$	4225	$2.273 \cdot 10^{-3}$	$9.105 \cdot 10^{-5}$	2.044	$1.097 \cdot 10^{-4}$	2.044	$1.859 \cdot 10^{-4}$	2.023

Table 3: The linear Fokker–Planck equation with $\lambda_y = 100$

We observe in Tables 1, 2, and 3 second-order convergence in all norms, and this uniformly with respect to the anisotropy. Furthermore, since the pressure function p is singular, we can check (with the quantity $\min(u_{h\tau})$) that the solution remains non-negative.

Now, we turn to a posteriori error estimates and consider the following quantities, for each discrete time t^n, $n \in \{1, \cdots, N\}$, and each mesh element $T \in \mathcal{T}$:

- the actual error distribution

$$\int_{t^{n-1}}^{t^n} \|\boldsymbol{\Lambda}(\boldsymbol{\nabla}\gamma(\hat{u}_{h\tau}) + \eta(\hat{u}_{h\tau})\boldsymbol{\nabla}\Psi) - \boldsymbol{\Lambda}(\boldsymbol{\nabla}\gamma(u) + \eta(u)\boldsymbol{\nabla}\Psi)\|_{L^1(T)}\,; \tag{6.2}$$

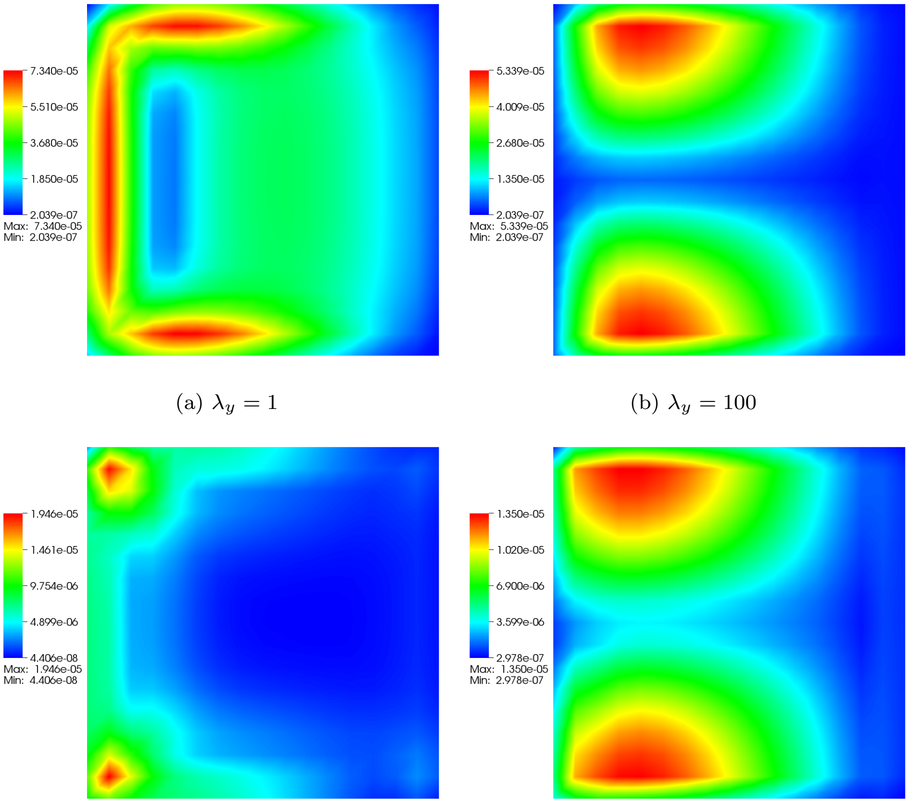
- the predicted error distribution (both in time and in space)

$$\eta^n_{\mathrm{F},T} := \int_{t^{n-1}}^{t^n} \|\boldsymbol{\Lambda}(\boldsymbol{\nabla}\gamma(\hat{u}_{h\tau}) + \eta(\hat{u}_{h\tau})\boldsymbol{\nabla}\Psi) + \boldsymbol{\sigma}_{h\tau}\|_{L^1(T)}\,; \tag{6.3}$$

- the predicted error distribution in space only

$$\int_{t^{n-1}}^{t^n} \|\boldsymbol{\Lambda}(\boldsymbol{\nabla}\gamma(u_{h\tau}) + \eta(u_{h\tau})\boldsymbol{\nabla}\Psi) + \boldsymbol{\sigma}_{h\tau}\|_{L^1(T)}\,. \tag{6.4}$$

For the third pair of meshes of this test case (that is $\tau = 0.000625$ and $h_{\mathcal{T}} = 0.08839$), using the trapezoidal rule, we approximately compute and plot in Fig 2 these errors after 10 time steps (that is at time $t^n = 0.00625$). We observe that for $\lambda_y = 100$ the actual error (Fig. 2b) is very close to the predicted error in time and in space (Fig. 2d). This is rather remarkable in view of the complexity of this unsteady nonlinear test problem. One may remark, though, that the estimators $\eta^n_{\mathrm{F},T}$ of (6.3) underestimate the error (6.2); we can actually only prove global space-time upper bound for the dual norm of the residual $\mathcal{J}(\hat{u}_{h\tau})$ given by (5.6) which lies below the space-time L^1 norm (6.2). The estimate of Fig. 2c for $\lambda_y = 1$ appears less sharp. Note that in Fig. 2 (and similarly in Fig. 3 below), for a fixed value of λ_y, the color scales are different. We have done this on purpose, since this allows to 1) best see where the (estimated) error is located; 2) appreciate that the predicted location of the error matches quite nicely with the exact one.



(a) $\lambda_y = 1$

(b) $\lambda_y = 100$

(c) $\lambda_y = 1$

(d) $\lambda_y = 100$

Fig. 2: Comparison of the actual error distribution (6.2) (first line) with the predicted error distribution in time and in space (6.3) (second line) after 10 time steps for two values of λ_y

Now, with the same domain Ω, we consider the functions $\eta(u) = |u|$, $p(u) = 2u$, and $\Psi(x, y) = -gx$ with $g = 1$. Thus we study the porous medium equation with drift,

$$\partial_t u - \boldsymbol{\nabla}\cdot(\boldsymbol{\Lambda}(2|u|\boldsymbol{\nabla} u - \boldsymbol{g}_x u)) = 0,$$

associated with Dirichlet boundary conditions, for which the exact solution is given by the 1-dimensional function

$$u(t, (x, y)) = \max\left(\beta t - x, 0\right), \tag{6.5}$$

with $\beta = \lambda_x(2 + g)$, $g = 1$, $\lambda_x = 1$, the final time $t_{\mathrm{f}} = 0.25$, and various values of λ_y.

Since the exact solution $u(t, \cdot)$ given by (6.5) is no longer in $H^2(\Omega)$ but only in $H^{3/2-\varepsilon}(\Omega)$ for $t > 0$, the expected order of convergence for the L^2 norm is slightly smaller than 3/2, as indeed observed in Tables 4,5,6 and 7. The approximate solution suffers of small undershoots, which is possible since p is not singular here, i.e., $I_p = \mathbb{R}$. But we see that even with an anisotropy ratio of 100, we recover the expected convergence rate of convergence. Here again, the error appears to be remarkably

τ	$h_{\mathcal{T}}$	N_{v}	$\min(u_{h\tau})$	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
$1.25 \cdot 10^{-2}$	0.177	81	$-8.536 \cdot 10^{-34}$	$1.173 \cdot 10^{-2}$	–	$1.462 \cdot 10^{-2}$	–	$2.723 \cdot 10^{-2}$	–
$3.125 \cdot 10^{-3}$	$8.839 \cdot 10^{-2}$	289	$-1.609 \cdot 10^{-35}$	$4.901 \cdot 10^{-3}$	1.372	$8.088 \cdot 10^{-3}$	0.930	$1.972 \cdot 10^{-2}$	0.507
$7.813 \cdot 10^{-4}$	$4.419 \cdot 10^{-2}$	1089	$-3.312 \cdot 10^{-36}$	$1.63 \cdot 10^{-3}$	1.659	$3.544 \cdot 10^{-3}$	1.244	$1.075 \cdot 10^{-2}$	0.915
$1.953 \cdot 10^{-4}$	$2.21 \cdot 10^{-2}$	4225	$-4.734 \cdot 10^{-37}$	$4.876 \cdot 10^{-4}$	1.781	$1.364 \cdot 10^{-3}$	1.408	$5.359 \cdot 10^{-3}$	1.026
$4.883 \cdot 10^{-5}$	$1.105 \cdot 10^{-2}$	16641	$-6.149 \cdot 10^{-38}$	$1.352 \cdot 10^{-4}$	1.871	$4.904 \cdot 10^{-4}$	1.492	$2.599 \cdot 10^{-3}$	1.056

Table 4: Porous medium equation with drift with $\lambda_y = 100$

τ	$h_{\mathcal{T}}$	N_{v}	$\min(u_{h\tau})$	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
$1.25 \cdot 10^{-2}$	0.177	81	$-1.396 \cdot 10^{-34}$	$3.312 \cdot 10^{-2}$	–	$3.552 \cdot 10^{-2}$	–	$5.514 \cdot 10^{-2}$	–
$3.125 \cdot 10^{-3}$	$8.839 \cdot 10^{-2}$	289	$-2.612 \cdot 10^{-35}$	$1.077 \cdot 10^{-2}$	1.766	$1.432 \cdot 10^{-2}$	1.428	$2.751 \cdot 10^{-2}$	1.094
$7.813 \cdot 10^{-4}$	$4.419 \cdot 10^{-2}$	1089	$-3.861 \cdot 10^{-36}$	$3.068 \cdot 10^{-3}$	1.894	$5.057 \cdot 10^{-3}$	1.569	$1.263 \cdot 10^{-2}$	1.174
$1.953 \cdot 10^{-4}$	$2.21 \cdot 10^{-2}$	4225	$-5.082 \cdot 10^{-37}$	$8.217 \cdot 10^{-4}$	1.943	$1.687 \cdot 10^{-3}$	1.620	$5.763 \cdot 10^{-3}$	1.157
$4.883 \cdot 10^{-5}$	$1.105 \cdot 10^{-2}$	16641	$-6.465 \cdot 10^{-38}$	$2.134 \cdot 10^{-4}$	1.967	$5.547 \cdot 10^{-4}$	1.622	$2.689 \cdot 10^{-3}$	1.112

Table 5: Porous medium equation with drift with $\lambda_y = 10$

τ	$h_{\mathcal{T}}$	N_{v}	$\min(u_{h\tau})$	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
$1.25 \cdot 10^{-2}$	0.177	81	$-2.141 \cdot 10^{-34}$	$5.432 \cdot 10^{-2}$	–	$5.24 \cdot 10^{-2}$	–	$7.19 \cdot 10^{-2}$	–
$3.125 \cdot 10^{-3}$	$8.839 \cdot 10^{-2}$	289	$-3.14 \cdot 10^{-35}$	$1.6 \cdot 10^{-2}$	1.922	$1.842 \cdot 10^{-2}$	1.644	$3.126 \cdot 10^{-2}$	1.310
$7.813 \cdot 10^{-4}$	$4.419 \cdot 10^{-2}$	1089	$-4.148 \cdot 10^{-36}$	$4.316 \cdot 10^{-3}$	1.975	$5.972 \cdot 10^{-3}$	1.698	$1.343 \cdot 10^{-2}$	1.274
$1.953 \cdot 10^{-4}$	$2.21 \cdot 10^{-2}$	4225	$-5.306 \cdot 10^{-37}$	$1.122 \cdot 10^{-3}$	1.987	$1.881 \cdot 10^{-3}$	1.704	$5.941 \cdot 10^{-3}$	1.203
$4.883 \cdot 10^{-5}$	$1.105 \cdot 10^{-2}$	16641	$-6.67 \cdot 10^{-38}$	$2.868 \cdot 10^{-4}$	1.991	$5.944 \cdot 10^{-4}$	1.681	$2.73 \cdot 10^{-3}$	1.134

Table 6: Porous medium equation with drift with $\lambda_y = 1$

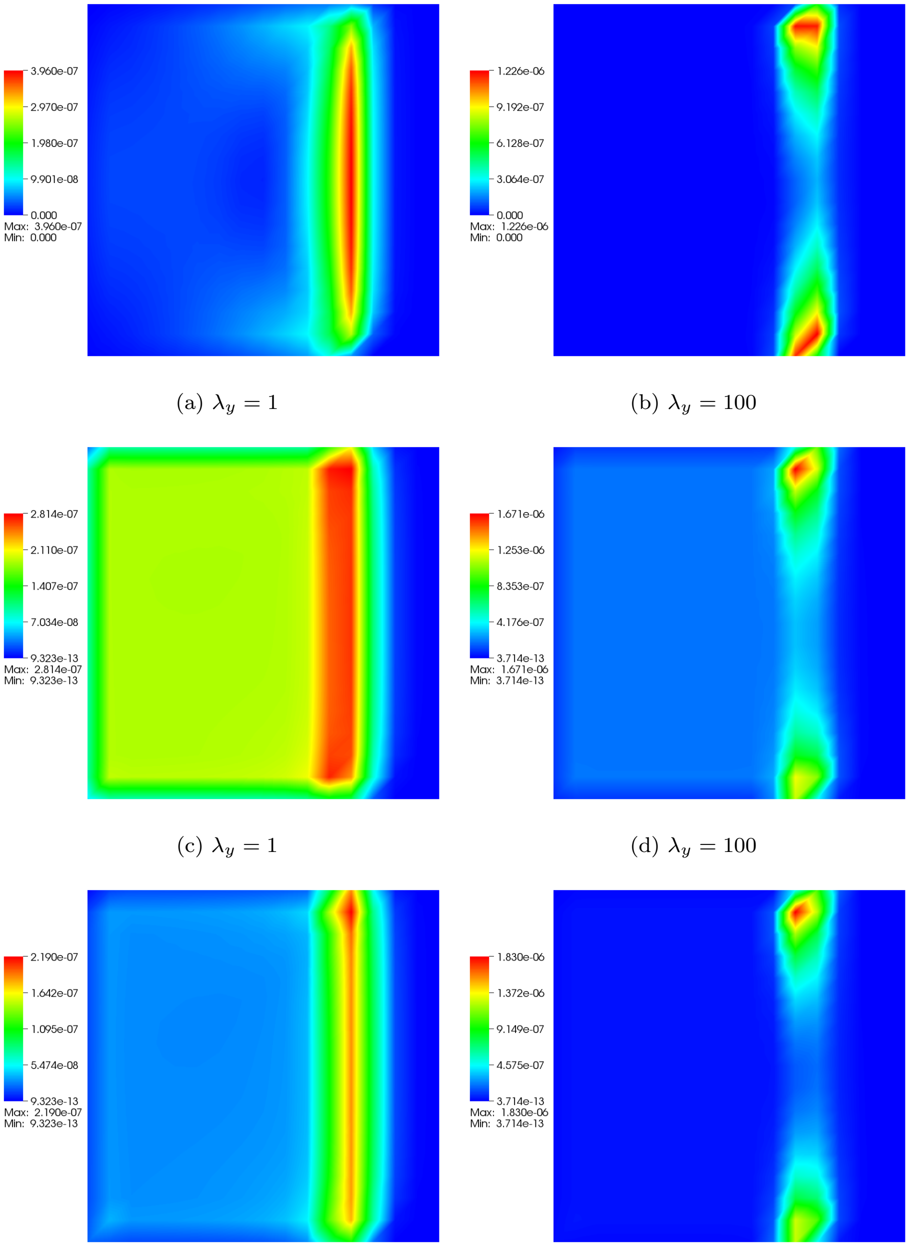
τ	$h_{\mathcal{T}}$	N_{v}	$\min(u_{h\tau})$	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
$1.25 \cdot 10^{-2}$	0.177	81	$-2.241 \cdot 10^{-34}$	$6.068 \cdot 10^{-2}$	–	$5.691 \cdot 10^{-2}$	–	$7.368 \cdot 10^{-2}$	–
$3.125 \cdot 10^{-3}$	$8.839 \cdot 10^{-2}$	289	$-3.368 \cdot 10^{-35}$	$1.77 \cdot 10^{-2}$	1.938	$1.955 \cdot 10^{-2}$	1.680	$3.147 \cdot 10^{-2}$	1.338
$7.813 \cdot 10^{-4}$	$4.419 \cdot 10^{-2}$	1089	$-4.296 \cdot 10^{-36}$	$4.75 \cdot 10^{-3}$	1.983	$6.244 \cdot 10^{-3}$	1.721	$1.346 \cdot 10^{-2}$	1.280
$1.953 \cdot 10^{-4}$	$2.21 \cdot 10^{-2}$	4225	$-5.452 \cdot 10^{-37}$	$1.231 \cdot 10^{-3}$	1.992	$1.943 \cdot 10^{-3}$	1.722	$5.948 \cdot 10^{-3}$	1.205
$4.883 \cdot 10^{-5}$	$1.105 \cdot 10^{-2}$	16641	$-6.792 \cdot 10^{-38}$	$3.139 \cdot 10^{-4}$	1.994	$6.075 \cdot 10^{-4}$	1.696	$2.732 \cdot 10^{-3}$	1.135

Table 7: Porous medium equation with drift with $\lambda_y = 0.1$

stable w.r.t. the anisotropy ratio, in opposition to the methods based on upwinding presented in [14, 1, 11].

For the second mesh of this test case (that is $\tau = 0.003125$ and $h_{\mathcal{T}} = 0.0839$), we plot in Fig 3 the actual error distribution (6.2), the predicted error distribution in time and in space (6.3), as well as the predicted error distribution in space only (6.4). The results are presented for the last time interval $[t^{N-1}, t^N]$ and for $\lambda_y = 1$ and 100. We observe that the actual error (Figs. 3a and 3b) and the predicted error in space only (Figs. 3e and 3f) are very similar. The predicted error in time and in space (Figs. 3c and 3d) is less satisfactory, even though the scale is the same. This is linked to a known deficiency of the present estimates which overestimate the error in time for simple time-behaviors like (6.5), see [24, 26] and the references therein (note that $\boldsymbol{\sigma}_{h\tau}$ is constant in time in (6.3) whereas $\hat{u}_{h\tau}$ is approximately affine in time just as the exact solution.)

Finally, we consider the porous medium equation, that is $\eta(u) = 2u$, $p(u) = u$,



(a) $\lambda_y = 1$

(b) $\lambda_y = 100$

(c) $\lambda_y = 1$

(d) $\lambda_y = 100$

(e) $\lambda_y = 1$

(f) $\lambda_y = 100$

Fig. 3: Comparison between the actual error distribution (6.2) (first line), the predicted error distribution in time and in space (6.3) (second line), and the predicted error distribution in space only (third line) at final time $t_{\mathrm{f}} = 0.25$ for two values of λ_y

and $\Psi = 0$, with Dirichlet boundary conditions, i.e.,

$$\partial_t u - \boldsymbol{\nabla}\cdot(2u\boldsymbol{\Lambda}\boldsymbol{\nabla} u) = 0,$$

with the initial time $t_0 = 0.005$, the final time $t_\mathrm{f} = 0.25$, and $\lambda_x = 1$. We consider two cases corresponding to two domains $\Omega = \{(x,y) \in \mathbb{R}^2 : x^2 + \frac{y^2}{\lambda_y} = R\}$ with $\lambda_y = 1$ and $R = 2.6$ for the disk and $\lambda_y = 0.1$ and $R = 3$ for the ellipse. The Barenblatt's solution to the porous medium equation is given by

$$u(t,(x,y)) = \frac{1}{2\sqrt{t}} \max\left(0, \frac{1}{\sqrt{2\pi}} - \frac{x^2 + \frac{y^2}{\lambda_y}}{8\sqrt{t}}\right).$$

τ	$h_\mathcal{T}$	N_v	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
0.123	0.643	54	0.284	–	0.322	–	0.761	–
$4.9 \cdot 10^{-2}$	0.478	117	0.172	1.306	0.197	1.272	0.446	1.382
$2.042 \cdot 10^{-2}$	0.342	243	0.103	1.394	0.128	1.181	0.324	0.873
$9.074 \cdot 10^{-3}$	0.221	536	$5.416 \cdot 10^{-2}$	1.629	$7.076 \cdot 10^{-2}$	1.500	0.156	1.854
$3.952 \cdot 10^{-3}$	0.156	1163	$2.775 \cdot 10^{-2}$	1.727	$3.756 \cdot 10^{-2}$	1.635	$8.247 \cdot 10^{-2}$	1.640
$1.738 \cdot 10^{-3}$	0.110	2603	$1.345 \cdot 10^{-2}$	1.799	$1.863 \cdot 10^{-2}$	1.740	$4.074 \cdot 10^{-2}$	1.751
$7.729 \cdot 10^{-4}$	$6.646 \cdot 10^{-2}$	5884	$6.27 \cdot 10^{-3}$	1.871	$8.963 \cdot 10^{-3}$	1.794	$2.949 \cdot 10^{-2}$	0.793
$3.427 \cdot 10^{-4}$	$4.961 \cdot 10^{-2}$	13030	$2.899 \cdot 10^{-3}$	1.941	$4.277 \cdot 10^{-3}$	1.861	$2.083 \cdot 10^{-2}$	0.874
$1.523 \cdot 10^{-4}$	$3.333 \cdot 10^{-2}$	29104	$1.308 \cdot 10^{-3}$	1.980	$2.054 \cdot 10^{-3}$	1.825	$1.499 \cdot 10^{-2}$	0.819

Table 8: Porous medium equation in the disk with $\lambda_y = 1$

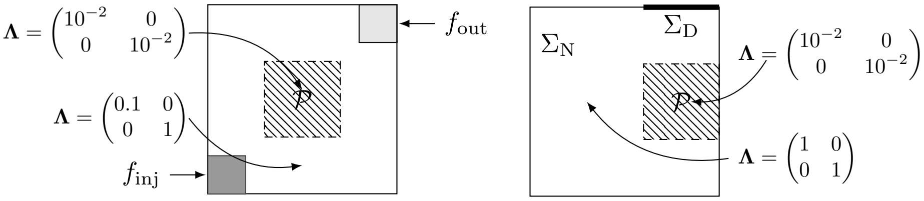
τ	$h_\mathcal{T}$	N_v	L^1-error	L^1 rate	L^2-error	L^2 rate	L^∞-error	L^∞ rate
0.123	0.687	32	1.626	–	1.513	–	1.486	–
$4.9 \cdot 10^{-2}$	0.447	59	1.000	1.589	1.000	1.353	1.000	1.295
$2.042 \cdot 10^{-2}$	0.308	128	0.274	3.345	0.295	3.151	0.459	2.013
$9.074 \cdot 10^{-3}$	0.220	264	0.172	1.291	0.182	1.340	0.260	1.566
$3.952 \cdot 10^{-3}$	0.145	565	$8.726 \cdot 10^{-2}$	1.778	0.100	1.558	0.204	0.634
$1.738 \cdot 10^{-3}$	$9.482 \cdot 10^{-2}$	1259	$3.178 \cdot 10^{-2}$	2.521	$3.708 \cdot 10^{-2}$	2.488	$6.361 \cdot 10^{-2}$	2.913
$7.729 \cdot 10^{-4}$	$6.359 \cdot 10^{-2}$	2832	$1.497 \cdot 10^{-2}$	1.857	$1.822 \cdot 10^{-2}$	1.752	$5.003 \cdot 10^{-2}$	0.593
$3.427 \cdot 10^{-4}$	$4.409 \cdot 10^{-2}$	6257	$7.333 \cdot 10^{-3}$	1.802	$9.723 \cdot 10^{-3}$	1.585	$3.388 \cdot 10^{-2}$	0.984
$1.523 \cdot 10^{-4}$	$3.006 \cdot 10^{-2}$	13920	$3.732 \cdot 10^{-3}$	1.689	$5.4 \cdot 10^{-3}$	1.471	$3.187 \cdot 10^{-2}$	0.152

Table 9: Porous medium equation in the ellipse with $\lambda_y = 0.1$

We again observe fast convergence, except for the L^∞ norm, in Tables 8–9.

6.2. Test cases with heterogeneous permeability tensors. We consider now two test-cases for which we do not know the exact solution. For both, the domain Ω is the unit square $]0,1[^2$ and we use a simple adaptive time-stepping strategy: the initial time is $t_0 = 0$, the initial time step is $\tau_0 = 10^{-6}$, and for $n \geq 1$, if the number of Newton iterations is less than 4 (and $\tau_{n-1} < 0.9$) then $\tau_n = 2\tau_{n-1}$, otherwise $\tau_n = \tau_{n-1}$. We perform 5000 iterations in time. Moreover, the value of the permeability tensor field $\mathbf{\Lambda}$ is not constant (see Fig. 4):

$$\mathbf{\Lambda}(\boldsymbol{x}) = \begin{cases} \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^{-2} \end{pmatrix} & \text{if } \boldsymbol{x} \in \mathcal{P}, \\ \begin{pmatrix} \beta & 0 \\ 0 & 1 \end{pmatrix} & \text{if } \boldsymbol{x} \notin \mathcal{P}. \end{cases}$$

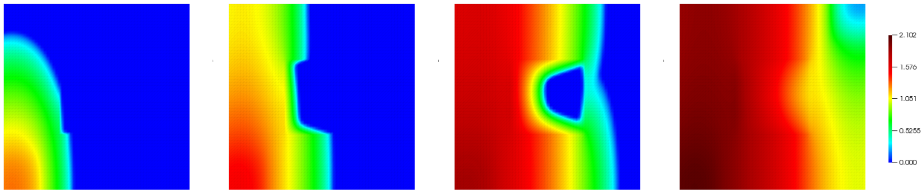

$\mathbf{\Lambda} = \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^{-2} \end{pmatrix}$
f_{out}
$\mathcal{P}$
$\mathbf{\Lambda} = \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}$
f_{inj}
Σ_{N}
Σ_{D}
$\mathbf{\Lambda} = \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^{-2} \end{pmatrix}$
$\mathcal{P}$
$\mathbf{\Lambda} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

(a) Quarter Five Spot

(b) Barrier

Fig. 4: Setting of the tests with heterogeneity

Quarter five spot test-case. For the first test-case (described in Fig. 4a), the equation is associated with homogeneous Neumann boundary conditions (that is $\Sigma_{\text{N}} = \partial\Omega$ and $\Sigma_{\text{D}} = \emptyset$), and we choose the functions $\eta(u) = u^2$, $p(u) = 2u$, $\Psi(x, y) = -gx$ with $g = 1$, $f_{\text{inj}} = 1_{[0,0.2]^2}$, and $f_{\text{out}} = 1_{[0.8,1]^2}$. The domain $\mathcal{P}$ is the square $[0.3, 0.7]^2$ and $\beta = 0.1$. We initialize the numerical scheme with $u_0 = 0$ and we plot the value of the approximate solution for 4 different times.



(a) $t = 0.252415$

(b) $t = 0.600575$

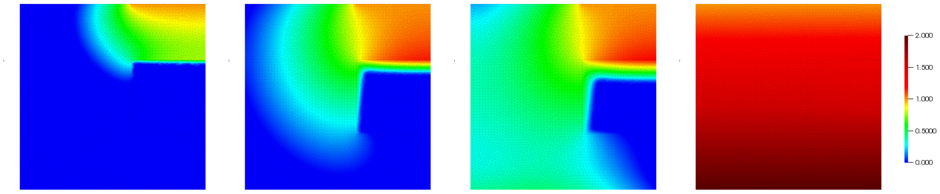
(c) $t = 1.28$

(d) $t = 2.24$

Fig. 5: Quarter five spot: approximate solution $u_{h\tau}$ in Ω for different times t

At the beginning (see Fig. 5a) the solution is in the complement of $\mathcal{P}$. Thus, thanks to the definition of the permeability tensor field $\mathbf{\Lambda}$, we observe anisotropy in the vertical direction. Then, since the area of $\mathcal{P}$ is less permeable than the complement of $\mathcal{P}$, we can see that the solution remains outside $\mathcal{P}$ (see Figs. 5b and 5c). Moreover, we can note all along the simulation the influence of the injection and extraction wells (due to the functions f_{inj} and f_{out} respectively).

Barrier test-case. For the second example (described in Fig. 4b), we impose the Dirichlet boundary condition $u = 1$ on $\Sigma_{\text{D}} =]0.6, 1[\times \{1\}$ and a homogeneous Neumann boundary condition on $\Sigma_{\text{N}} = \partial\Omega \backslash \Sigma_{\text{D}}$. We do not have sources, that is $f_{\text{inj}} = f_{\text{out}} = 0$, and we choose the functions $\eta(u) = |u|$, $p(u) = u$, and $\Psi(x, y) = gy$ with $g = 1$. The domain $\mathcal{P}$ is here the rectangle $[0.6, 1] \times [0.3, 0.7]$ and $\beta = 1$.



(a) $t = 0.067215$ (b) $t = 0.272015$ (c) $t = 0.551055$ (d) $t = 7.6996$

Fig. 6: Barrier: Approximate solution $u_{h\tau}$ in Ω for different times t

We can observe in Fig. 6 that, thanks to the definition of the external potential Ψ and since the area of $\mathcal{P}$ is less permeable than the complement of $\mathcal{P}$, that the solution moves in the longitudinal direction while avoiding the $\mathcal{P}$ area and satisfying the Dirichlet boundary condition on Σ_{D}.

6.3. Linearization adaptive stopping criteria. We finally show how the distinction of the error components mentioned in Remark 5.3 can be used to design adaptive stopping criteria for the Newton iterative linearization. At the n^{th} time step, we know u_h^{n-1} and we search for u_h^n, solution to the nonlinear system (2.22). The Newton method used to solve this scheme can be written as follows. At the ℓ^{th} iteration, we know $u_h^{n,\ell-1}$, and $u_h^{n,\ell}$ is given as the solution to the linear system

$$\int_{\omega_{\boldsymbol{a}}} \boldsymbol{\zeta}_h^{n,\ell-1}(u_h^{n,\ell})\cdot\boldsymbol{\nabla}\phi_{\boldsymbol{a}} = \int_{\omega_{\boldsymbol{a}}} \left(f_{\boldsymbol{a}}^n - \frac{u_{\boldsymbol{a}}^{n,\ell} - u_{\boldsymbol{a}}^{n-1}}{\tau_n} \right) \phi_{\boldsymbol{a}}, \quad \forall \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}} \setminus \mathcal{V}_{\mathcal{T}}^{\text{ext,D}},$$

where $\boldsymbol{\zeta}_h^{n,\ell-1}(u_h^{n,\ell})$ is defined by

$$\begin{aligned}\boldsymbol{\zeta}_h^{n,\ell-1}(u_h^{n,\ell}) = {}& \left(u_h^{n,\ell} - u_h^{n,\ell-1}\right)(\eta')_h^{n,\ell-1}\boldsymbol{\Lambda}_h\boldsymbol{\nabla}\left(p_h^{n,\ell-1} + \Psi_h\right) \\ & + \eta_h^{n,\ell-1}\left(u_h^{n,\ell} - u_h^{n,\ell-1}\right)\boldsymbol{\Lambda}_h\boldsymbol{\nabla}\left((p')_h^{n,\ell-1}\right) \\ & + \eta_h^{n,\ell-1}\boldsymbol{\Lambda}_h(p')_h^{n,\ell-1}\boldsymbol{\nabla}\left(u_h^{n,\ell} - u_h^{n,\ell-1}\right) + \eta_h^{n,\ell-1}\boldsymbol{\Lambda}_h\boldsymbol{\nabla}(p_h^{n,\ell-1} + \Psi_h).\end{aligned}$$

In the above expression, we used again the notation $g_h^{n,\ell-1} = \pi_1 g(u_h^{n,\ell-1})$ for $g : I_p \to \mathbb{R}$. We only give here the details relative to the stopping criterion for the nonlinear solver but all the details on the complete numerical method and the different error components can be found in [16, 23, 27]. At the n^{th} time step and the ℓ^{th} linearization step we can decompose the flux reconstruction $\boldsymbol{\sigma}_h^{n,\ell}$ defined in Proposition 5.1 as $\boldsymbol{\sigma}_h^{n,\ell} = \mathbf{d}_h^{n,\ell} + \mathbf{l}_h^{n,\ell}$ where

- $\mathbf{d}_h^{n,\ell} \in [L^1(\Omega)]^d$ is an approximation of the discretization flux
$$-\eta(u_h^{n,\ell})\boldsymbol{\Lambda}_h\boldsymbol{\nabla}\left(p(u_h^{n,\ell}) + \Psi_h\right);$$
- $\mathbf{l}_h^{n,\ell} \in [L^1(\Omega)]^d$ represents the linearization error and satisfies $\left\|\mathbf{l}_h^{n,\ell}\right\|_{L^1(\Omega)} \to 0$ when the nonlinear solver converges.

We are now able to give the adaptive stopping criterion that we use for the Newton method

$$\eta_{\text{lin}}^{n,\ell} \leq \gamma\, \eta_{\text{disc}}^{n,\ell}, \tag{6.6}$$

where

- $\eta_{\mathrm{lin}}^{n,\ell}$ is the linearization estimator
$$\eta_{\mathrm{lin}}^{n,\ell} = \tau_n \left\| \mathbf{l}_h^{n,\ell} \right\|_{L^1(\Omega)} \quad \text{with } \mathbf{l}_h^{n,\ell} = \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} \mathbf{l}_{\boldsymbol{a}}^{n,\ell}.$$
- $\eta_{\mathrm{disc}}^{n,\ell}$ is the spatial discretization estimator
$$\eta_{\mathrm{disc}}^{n,\ell} = \tau_n \left\| \eta_h^{n,\ell} \boldsymbol{\Lambda}_h \boldsymbol{\nabla} \left(p_h^{n,\ell} + \Psi_h \right) + \mathbf{d}_h^{n,\ell} \right\|_{L^1(\Omega)} \quad \text{with } \mathbf{d}_h^{n,\ell} = \sum_{\boldsymbol{a}\in\mathcal{V}_{\mathcal{T}}} \mathbf{d}_{\boldsymbol{a}}^{n,\ell}.$$
- γ is a positive parameter expressing the desired ratio of the linearization error to the discretization error (typically of order 0.1).

For $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}}$, the quantities $\mathbf{d}_{\boldsymbol{a}}^{n,\ell}$ and $\mathbf{l}_{\boldsymbol{a}}^{n,\ell}$ are computed by solving the following problems.

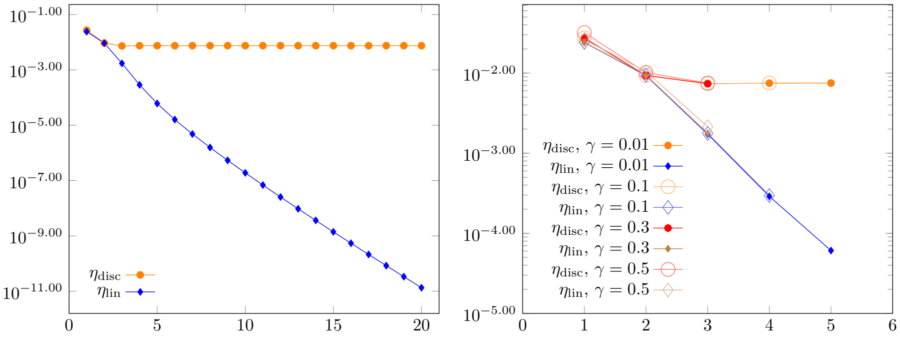
- Construction of $(\mathbf{d}_{\boldsymbol{a}}^{n,\ell} + \mathbf{l}_{\boldsymbol{a}}^{n,\ell})$: find $\mathbf{d}_{\boldsymbol{a}}^{n,\ell} + \mathbf{l}_{\boldsymbol{a}}^{n,\ell} \in \mathbf{W}_h^a$ and $r_h^{\boldsymbol{a}} \in Q_h^a$ such that
$$\int_{\omega_{\boldsymbol{a}}} \left(\mathbf{d}_{\boldsymbol{a}}^{n,\ell} + \mathbf{l}_{\boldsymbol{a}}^{n,\ell} \right) \boldsymbol{v}_h - \int_{\omega_{\boldsymbol{a}}} \boldsymbol{\nabla}\cdot \boldsymbol{v}_h r_h^{\boldsymbol{a}} = - \int_{\omega_{\boldsymbol{a}}} \phi_{\boldsymbol{a}} \boldsymbol{\zeta}_h^{n,\ell-1}(u_h^{n,\ell})\cdot \boldsymbol{v}_h,$$
$$\int_{\omega_{\boldsymbol{a}}} \boldsymbol{\nabla}\cdot \left(\mathbf{d}_{\boldsymbol{a}}^{n,\ell} + \mathbf{l}_{\boldsymbol{a}}^{n,\ell} \right) q_h = \int_{\omega_{\boldsymbol{a}}} \left[\left(f_{\boldsymbol{a}}^{n,\ell} - \frac{u_{\boldsymbol{a}}^{n,\ell} - u_{\boldsymbol{a}}^{n-1}}{\tau_n} \right) \phi_{\boldsymbol{a}} - \boldsymbol{\zeta}_h^{n,\ell-1}(u_h^{n,\ell})\cdot \boldsymbol{\nabla}\phi_{\boldsymbol{a}} \right] q_h,$$
for all $\boldsymbol{v}_h \in \mathbf{W}_h^a$ and $q_h \in Q_h^a$.
- Construction of $\mathbf{d}_{\boldsymbol{a}}^{n,\ell}$: find $\mathbf{d}_{\boldsymbol{a}}^{n,\ell} \in \mathbf{W}_h^a$ and $\overline{r}_h^{\boldsymbol{a}} \in Q_h^a$ such that
$$\int_{\omega_{\boldsymbol{a}}} \mathbf{d}_{\boldsymbol{a}}^{n,\ell} \boldsymbol{v}_h - \int_{\omega_{\boldsymbol{a}}} \boldsymbol{\nabla}\cdot \boldsymbol{v}_h \overline{r}_h^{\boldsymbol{a}} = - \int_{\omega_{\boldsymbol{a}}} \phi_{\boldsymbol{a}} \eta_h^{n,\ell} \boldsymbol{\Lambda}_h \boldsymbol{\nabla} \left(p_h^{n,\ell} + \Psi_h \right) \cdot \boldsymbol{v}_h,$$
$$\int_{\omega_{\boldsymbol{a}}} \boldsymbol{\nabla}\cdot \mathbf{d}_{\boldsymbol{a}}^{n,\ell} q_h = \int_{\omega_{\boldsymbol{a}}} \left[\left(f_{\boldsymbol{a}}^{n,\ell} - \frac{u_{\boldsymbol{a}}^{n,\ell} - u_{\boldsymbol{a}}^{n-1}}{\tau_n} \right) \phi_{\boldsymbol{a}} - \eta_h^{n,\ell} \boldsymbol{\Lambda}_h \boldsymbol{\nabla} \left(p_h^{n,\ell} + \Psi_h \right) \cdot \boldsymbol{\nabla}\phi_{\boldsymbol{a}} \right] q_h,$$
for all $\boldsymbol{v}_h \in \mathbf{W}_h^a$ and $q_h \in Q_h^a$.

Let the domain Ω be a unit disk with radius $R = 2.6$ and mesh size $h_{\mathcal{T}} \sim 0.16$. We consider the functions $\eta(u) = u$, $p(u) = \dfrac{m}{m-1} u^{m-1}$ with $m = 4$, $\Psi = f_{\mathrm{inj}} = f_{\mathrm{out}} = 0$, and the tensor field $\boldsymbol{\Lambda} = I_d$. Thus, the exact solution is the following Barenblatt's solution
$$u(t,(x,y)) = \left(\frac{1}{t+1} \left(\left[1 - \frac{m-1}{4m^2} \frac{x^2+y^2}{(t+1)^m} \right]^+ \right)^{\frac{m}{m-1}} \right)^{\frac{1}{m}},$$
to which we associate the corresponding Dirichlet boundary conditions. The initial time is $t_0 = 0$, the final time $t_{\mathrm{f}} = 0.1$, and the time step $\tau_n = 0.01$. In the computations which follow, we employ 10 time steps. We give in Table 10 the number of Newton iterations for an "exact" solver corresponding to the stopping criterion (6.1), and for the a posteriori strategy (6.6) for each iteration. In the last column, we report the number of cumulative Newton iterations for the whole simulation in each case.

We can observe in Table 10 that even with a small parameter $\gamma = 0.01$, the gain in the number of Newton iterations is significant. Indeed, with only 10 iterations in time, the number of cumulative Newton iterations is equal to 170 with the exact solver whereas it is equal to 44 with the a posteriori strategy with $\gamma = 0.01$. Moreover, as expected, we can see that for the adaptive stopping criteria, the greater is the

Time	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	Cumulated iterations
Exact solver	22	20	18	18	16	15	15	16	15	15	170
$\gamma = 0.01$	7	5	5	4	4	4	4	4	4	3	44
$\gamma = 0.1$	6	4	3	3	3	3	3	2	2	2	31
$\gamma = 0.3$	6	3	3	3	2	2	2	2	2	2	27
$\gamma = 0.5$	5	3	2	2	2	2	1	2	1	1	21

Table 10: Number of Newton iterations for each time step



(a) "Exact" solver (6.1)

(b) Adaptive stopping criterion (6.6)

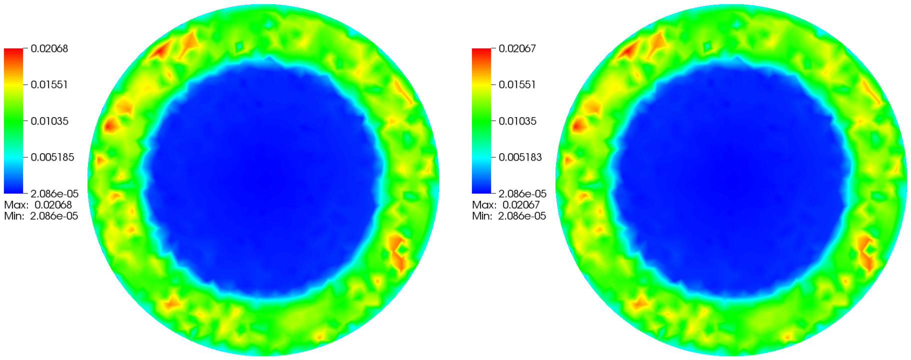
Fig. 7: Linearization and discretization estimators in functions of Newton iterations

parameter γ, the more the number of Newton iterations diminishes. To understand this phenomenon, we plot in Fig. 7 the estimators η_{lin} and η_{disc} in function of the Newton iterations for $t = 0.02$ (that is, on the second time step).

We can observe in Fig. 7a that as soon as we perform two Newton iterations, the linearization estimator η_{lin} is smaller than the discretization estimator η_{disc}. Moreover, at the end of the simulation, the difference between the two estimators is greater than eight orders of magnitude, and it is apparently not necessary to perform as many Newton iterations. Fig 7b confirms that even with a larger γ, for example $\gamma = 0.5$, the linearization estimator becomes quickly smaller than the discretization estimator, so that only a small number of Newton iterations is necessary.

To observe the distribution of the different local error components we plot in Fig. 8 and 9 the error distribution in three cases.

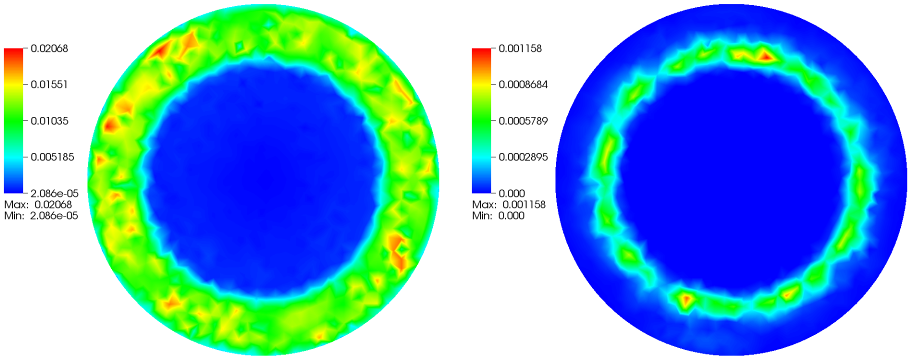
- The total error (Fig. 8) is the difference between the flux with the exact solution u and those obtained with the approximate solution at final time u_h^N with local contributions given by
$$\tau_N \left\| \mathbf{\Lambda}(\boldsymbol{\nabla}\gamma(u_h^N) + \eta(u_h^N)\boldsymbol{\nabla}\Psi) - \mathbf{\Lambda}(\boldsymbol{\nabla}\gamma(u(t_{\mathrm{f}},\cdot)) + \eta(u(t_{\mathrm{f}},\cdot))\boldsymbol{\nabla}\Psi) \right\|_{L^1(T)}.$$
- The discretization error (Fig. 9a) defined by
$$\tau_N \Big\| \mathbf{\Lambda}(\boldsymbol{\nabla}\gamma((u_h^N)^{\mathrm{ex}}) + \eta((u_h^N)^{\mathrm{ex}})\boldsymbol{\nabla}\Psi) \\ - \mathbf{\Lambda}(\boldsymbol{\nabla}\gamma(u(t_{\mathrm{f}},\cdot)) + \eta(u(t_{\mathrm{f}},\cdot))\boldsymbol{\nabla}\Psi) \Big\|_{L^1(T)},$$



(a) Error for the "exact" solve (6.1)

(b) Error for the adaptive linearization stopping criterion (6.6), $\gamma = 0.5$

Fig. 8: Comparison of the total errors for different resolution strategies at the final time



(a) Discretization error, $\gamma = 0.5$

(b) Linearization error, $\gamma = 0.5$

Fig. 9: Distributions of the error components, $\gamma = 0.5$ at the final time; note that the maximal linearization error is approximately 2 times smaller than the maximal discretization error

where $(u_h^N)^{\text{ex}}$ is computed by taking the stopping criterion (6.1) of the exact solver and for which we initialize the Newton method by $(u_h^{N,0})^{\text{ex}} = (u_h^{N-1})^{\text{adapt}}$ with $(u_h^{N-1})^{\text{adapt}}$ the approximate solution at time t^{N-1} obtained by taking the stopping criterion (6.6) of the adaptive strategy.

- The linearization error (Fig. 9b) is the difference between the flux with the previous solution $(u_h^N)^{\text{ex}}$ and the flux with the solution $(u_h^N)^{\text{adapt}}$ obtained upon the (6.6) stopping criterion for the Newton linearization, that is

$$\tau_N \Big\| \mathbf{\Lambda}(\mathbf{\nabla}\gamma((u_h^N)^{\mathrm{ex}}) + \eta((u_h^N)^{\mathrm{ex}})\mathbf{\nabla}\Psi) \\ - \mathbf{\Lambda}(\mathbf{\nabla}\gamma((u_h^N)^{\mathrm{adapt}}) + \eta((u_h^N)^{\mathrm{adapt}})\mathbf{\nabla}\Psi)\Big\|_{L^1(T)}.$$

We plot in Fig. 8a the total error with the exact solver and we compare the results with the error distribution in two components: the discretization error in Fig. 9a and the linearization error in Fig. 9b with the a posteriori strategy for $\gamma = 0.5$. As expected we observe that the total error is dominated by the discretization error, and that the linearization error is negligible.

Appendix A. A time compactness result.

The goal of this appendix is to briefly present the blackbox for proving the time-compactness of the sequence of approximate solutions in Proposition 4.2. This blackbox has been introduced in [3] and extends to the discrete setting some results of [44].

Let $(\mathcal{T}_m)_{m\geq 1}$ be a sequence of simplicial meshes with bounded regularity and size tending to 0, and let $\left(\left(t_n^{(m)}\right)_{0\leq n\leq N_m}\right)_{m\geq 1}$ be a sequence of time discretizations as in Theorem 2.4. In accordance with the notation of the core of the paper, we define for $m \geq 1$ the linear spaces

$$\begin{aligned} V_h^{(m)} &:= \left\{ f \in C(\overline{\Omega}) : f|_T \text{ is affine for all } T \in \mathcal{T}_m \right\}, \\ V_h^{0,(m)} &:= \left\{ f \in V_h^{(m)} : f \equiv 0 \text{ on } \Sigma_D \right\}, \\ X_h^{(m)} &:= \{ f \in L^\infty(\Omega) : f|_{s_{\boldsymbol{a}}} \text{ is constant for all } \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}_m} \} \end{aligned}$$

and

$$\begin{aligned} V_{h\tau}^{(m)} &:= \left\{ f \in D((-\infty, t_{\mathrm{f}}]; V_h^{(m)}) : f|_{(t_{n-1}^{(m)}, t_n^{(m)}]} \text{ is constant for all } n \in \{0, \dots, N_m\} \right\}, \\ V_{h\tau}^{0,(m)} &:= \left\{ f \in V_{h\tau}^{(m)} : f \equiv 0 \text{ on } [0, t_{\mathrm{f}}] \times \Sigma_D \right\}, \\ X_{h\tau}^{(m)} &:= \left\{ f \in D((-\infty, t_{\mathrm{f}}]; X_h^{(m)}) : f|_{(t_{n-1}^{(m)}, t_n^{(m)}]} \text{ is constant for all } n \in \{0, \dots, N_m\} \right\}, \\ \hat{V}_{h\tau}^{(m)} &:= \left\{ f \in C([0, t_{\mathrm{f}}]; V_h^{(m)}) : f|_{(t_{n-1}^{(m)}, t_n^{(m)}]} \text{ is affine for all } n \in \{0, \dots, N_m\} \right\}. \end{aligned}$$

Given an element $u_h^{(m)}$ of $V_h^{(m)}$, we denote by $\overline{u}_h^{(m)}$ the unique element of $X_h^{(m)}$ such that $u_h^{(m)}(\boldsymbol{a}) = \overline{u}_h^{(m)}(\boldsymbol{a})$ for all $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}_m}$. Similarly, given $u_{h\tau}^{(m)} \in V_{h\tau}^{(m)}$, we denote by $\overline{u}_{h\tau}^{(m)}$ and $\hat{u}_{h\tau}^{(m)}$ the unique element of $X_{h\tau}^{(m)}$ and $\hat{V}_{h\tau}^{(m)}$ respectively such that $u_{h\tau}^{(m)}(t_n, \boldsymbol{a}) = \overline{u}_{h\tau}^{(m)}(t_n, \boldsymbol{a}) = \hat{u}_{h\tau}^{(m)}(t_n, \boldsymbol{a})$ for all $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}_m}$ and all $n \in \{0, \dots, N_m\}$. The transposition to our setting of [3, Theorem 3.9] leads to the following statement.

THEOREM A.1. *Consider two sequences* $\left(u_{h\tau}^{(m)}\right)_{m\geq 1}$ *and* $\left(v_{h\tau}^{(m)}\right)_{m\geq 1}$ *such that* $u_{h\tau}^{(m)}, v_{h\tau}^{(m)} \in V_{h\tau}^{(m)}$ *for all* $m \geq 1$. *Assume that*

(i) the sequence $\left(\overline{u}_{h\tau}^{(m)}\right)_m$ *is bounded in* $L^2((0, t_{\mathrm{f}}); L^r(\Omega))$ *for some* $r > 6/5$;

(ii) the sequence $\left(v_{h\tau}^{(m)}\right)_m$ *is bounded in* $L^2((0, t_{\mathrm{f}}); H^1(\Omega))$;

(iii) there exists a non-decreasing continuous function ξ such that, for all $m \geq 1$, there holds

$$v_{h\tau}^{(m)}(t_n^{(m)}, \boldsymbol{a}) = \xi(u_{h\tau}^{(m)}(t_n^{(m)}, \boldsymbol{a})), \qquad \forall \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}_m},\ \forall n \in \{0, \dots, N_m\}.$$

This enforces in particular that $u_{h\tau}^{(m)}$ takes its values in Dom(ξ).

(iv) There exists $C > 0$ such that

$$\iint_{Q_{t_\mathrm{f}}} \partial_t \hat{u}_{h\tau}^{(m)} \overline{\varphi}_{h\tau} \leq C \|\boldsymbol{\nabla} \varphi_{h\tau}\|_{L^\infty(Q_{t_\mathrm{f}})}, \qquad \forall \varphi_{h\tau} \in V_{h\tau}^{0,(m)}. \tag{A.1}$$

Then there exists $u \in L^2((0, t_\mathrm{f}); L^r(\Omega))$ such that, up to the extraction of an unlabeled subsequence, there holds

$$\overline{u}_{h\tau}^{(m)} \underset{m\to\infty}{\longrightarrow} u \text{ a.e. in } Q_{t_\mathrm{f}}, \qquad \text{and} \qquad v_{h\tau}^{(m)} \underset{m\to\infty}{\longrightarrow} \xi(u) \text{ weakly in } L^2((0, t_\mathrm{f}); H^1(\Omega)).$$

In order to use [3, Theorem 3.9], we need to check a few assumptions referred as $(\mathbf{A_x}1)$, $(\mathbf{A_x}2)$, $(\mathbf{A_x}3)$, and $(\mathbf{A}_t)$ in [3]. Let us detail why these assumptions hold.

$(\mathbf{A}_t)$ This assumption is always fulfilled for one-step time-discretizations.

$(\mathbf{A_x}1)$ In our context, it amounts to check that for any sequence $(w_h^{(m)})$ of $V_h^{(m)}$ such that $\|w_h^{(m)}\|_{H^1(\Omega)} \leq C$, then $(w_h^{(m)})$ is relatively compact in $L^{\frac{r}{r-1}}(\Omega)$. This is a consequence of Sobolev's embedding (recall that $d \leq 3$).

$(\mathbf{A_x}2)$ This assumption is always fulfilled for piecewise constant reconstructions implemented in the space $X_h^{(m)}$.

$(\mathbf{A_x}3)$ Let $\varphi \in C_\mathrm{c}^\infty(\Omega)$, then define $\varphi_h^{(m)}$ as the unique function of $V_h^{(m)}$ such that

$$\varphi_h^{(m)}(\boldsymbol{a}) = \frac{1}{|s_{\boldsymbol{a}}|} \int_{s_{\boldsymbol{a}}} \varphi \qquad \text{for all } \boldsymbol{a} \in \mathcal{V}_{\mathcal{T}_m}. \tag{A.2}$$

In order to check this assumption, one has to verify that

$$\|\boldsymbol{\nabla} \varphi_h^{(m)}\|_\infty \leq C \|\boldsymbol{\nabla} \varphi\|_\infty \tag{A.3}$$

for some C depending only on the mesh regularity factor $\theta^\star$ and on the space dimension d (but not on m).

In order to establish (A.3), let us first remark that there exists a positive integer $M_{d,\theta^\star}$ depending only on $\theta^\star$ and d such that $\#\mathcal{T}_{\boldsymbol{a}} \leq M_{d,\theta^\star}$, where $\mathcal{T}_{\boldsymbol{a}}$ denotes the subset of $\mathcal{T}_m$ made of the simplices admitting $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}_m}$ as a vertex, and that

$$h_{\boldsymbol{a}} = \max_{T \in \mathcal{T}_{\boldsymbol{a}}} h_T \leq C_{d,\theta^\star} \min_{T \in \mathcal{T}_{\boldsymbol{a}}} h_T. \tag{A.4}$$

Mapping the simplex T on the reference simplex $\widehat{T}$ (see for instance [21] or [25]), we can establish that

$$h_T |\boldsymbol{\nabla} \phi_{\boldsymbol{a}}(\boldsymbol{x})| \leq C_{\theta^\star}, \qquad \forall \boldsymbol{x} \in T,\ \forall \boldsymbol{a} \in \mathcal{V}_T,\ \forall T \in \mathcal{T}_m. \tag{A.5}$$

Let $\varphi \in C_\mathrm{c}^\infty(\Omega)$ and let $\varphi_h^{(m)} \in V_h^{(m)}$ be defined by (A.2). Remark that for all $\boldsymbol{a} \in \mathcal{V}_{\mathcal{T}_m}$, there exists $\tilde{\boldsymbol{x}}_{\boldsymbol{a}} \in s_{\boldsymbol{a}}$ such that $\varphi_h^{(m)}(\boldsymbol{a}) = \varphi(\tilde{\boldsymbol{x}}_{\boldsymbol{a}})$. Fix now $T \in \mathcal{T}_m$.

Then, for all $\boldsymbol{x} \in T$, there holds

$$\begin{aligned}\left|\boldsymbol{\nabla}\varphi_h^{(m)}(\boldsymbol{x})\right| &= \left|\sum_{i=1}^{d}\left(\varphi_h^{(m)}(\boldsymbol{a}_i) - \varphi_h^{(m)}(\boldsymbol{a}_0)\right)\boldsymbol{\nabla}\phi_{\boldsymbol{a}_i}(\boldsymbol{x})\right| \\ &= \left|\sum_{i=1}^{d}\left(\varphi(\tilde{\boldsymbol{x}}_{\boldsymbol{a}_i}) - \varphi(\tilde{\boldsymbol{x}}_{\boldsymbol{a}_0})\right)\boldsymbol{\nabla}\phi_{\boldsymbol{a}_i}(\boldsymbol{x})\right| \\ &\leq \|\boldsymbol{\nabla}\varphi\|_\infty \sum_{i=1}^{d} |\tilde{\boldsymbol{x}}_{\boldsymbol{a}_i} - \tilde{\boldsymbol{x}}_{\boldsymbol{a}_0}|\,|\boldsymbol{\nabla}\phi_{\boldsymbol{a}_i}(\boldsymbol{x})|\,.\end{aligned}$$

Using the fact that $|\tilde{\boldsymbol{x}}_{\boldsymbol{a}_i} - \tilde{\boldsymbol{x}}_{\boldsymbol{a}_0}| \leq h_{\boldsymbol{a}}$ together with (A.4) and (A.5), we obtain (A.3).

REFERENCES

[1] A. Ait Hammou Oulhaj, C. Cancès, and C. Chainais-Hillairet, *Numerical analysis of a nonlinearly stable and positive Control Volume Finite Element scheme for Richards equation with anisotropy*, ESAIM Math. Model. Numer. Anal., 52 (2018), pp. 1532–1567.

[2] A. Ait Hammou Oulhaj, C. Cancès, C. Chainais-Hillairet, and P. Laurençot, *Large time behavior of a two phase extension of the porous medium equation*, Interfaces Free Bound., 21 (2019), pp. 199–229, doi:10.4171/IFB/421, https://doi.org/10.4171/IFB/421.

[3] B. Andreianov, C. Cancès, and A. Moussa, *A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic PDEs*, J. Funct. Anal., 273 (2017), pp. 3633–3670.

[4] L. A. Baughman and N. J. Walkington, *Co-volume methods for degenerate parabolic problems*, Numer. Math., 64 (1993), pp. 45–67.

[5] J. Bear and Y. Bachmat, *Introduction to modeling of transport phenomena in porous media*, vol. 4, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[6] M. Bessemoulin-Chatard and C. Chainais-Hillairet, *Exponential decay of a finite volume scheme to the thermal equilibrium for drift-diffusion systems*, J. Numer. Math., 25 (2017), pp. 147–168, doi:10.1515/jnma-2016-0007, https://doi.org/10.1515/jnma-2016-0007.

[7] M. Bessemoulin-Chatard, C. Chainais-Hillairet, and M.-H. Vignal, *Study of a finite volume scheme for the drift-diffusion system. Asymptotic behavior in the quasi-neutral limit*, SIAM J. Numer. Anal., 52 (2014), pp. 1666–1691, doi:10.1137/130913432, http://dx.doi.org/10.1137/130913432.

[8] D. Braess and J. Schöberl, *Equilibrated residual error estimator for edge elements*, Math. Comp., 77 (2008), pp. 651–672, doi:10.1090/S0025-5718-07-02080-7, http://dx.doi.org/10.1090/S0025-5718-07-02080-7.

[9] K. Brenner, C. Cancès, and D. Hilhorst, *Finite volume approximation for an immiscible two-phase flow in porous media with discontinuous capillary pressure*, Comput. Geosci., 17 (2013), pp. 573–597.

[10] K. Brenner and R. Masson, *Convergence of a vertex centered discretization of two-phase Darcy flows on general meshes*, Int. J. Finite Vol., 10 (2013), pp. 1–37.

[11] C. Cancès, *Energy stable numerical methods for porous media flow type problems*, Oil & Gas Science and Technology-Rev. IFPEN, 73 (2018), pp. 1–18.

[12] C. Cancès, C. Chainais-Hillairet, and S. Krell, *A nonlinear Discrete Duality Finite Volume Scheme for convection-diffusion equations*, in FVCA8 2017 - International Conference on Finite Volumes for Complex Applications VIII, C. C. . P. Omnes, ed., vol. 199 of Springer Proceedings in Mathematics & Statistics, Lille, France, 2017, Springer International Publishing, pp. 439–447.

[13] C. Cancès, C. Chainais-Hillairet, and S. Krell, *Numerical analysis of a nonlinear free-energy diminishing Discrete Duality Finite Volume scheme for convection diffusion equations*, Comput. Methods Appl. Math.„ 18 (2018), pp. 407–432, doi:10.1515/cmam-2017-0043, https://hal.archives-ouvertes.fr/hal-01529143.

[14] C. Cancès and C. Guichard, *Convergence of a nonlinear entropy diminishing Control Volume Finite Element scheme for solving anisotropic degenerate parabolic equations*, Math. Comp., 85 (2016), pp. 549–580.

[15] C. Cancès and C. Guichard, *Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure*, Found. Comput. Math., 17 (2017), pp. 1525–1584, doi:10.1007/s10208-016-9328-6, https://doi.org/10.1007/s10208-016-9328-6.

[16] C. Cancès, I. S. Pop, and M. Vohralík, *An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow*, Math. Comp., 83 (2014), pp. 153–188, doi:10.1090/S0025-5718-2013-02723-8, http://dx.doi.org/10.1090/S0025-5718-2013-02723-8.

[17] C. Chainais-Hillairet, *Schéma volumes finis pour des problèmes hyperboliques : convergence et estimations d'erreur*, PhD thesis, Université Paris 6, 1998.

[18] C. Chainais-Hillairet and F. Filbet, *Asymptotic behaviour of a finite-volume scheme for the transient drift-diffusion model*, IMA J. Numer. Anal., 27 (2007), pp. 689–716, doi:10.1093/imanum/drl045, http://dx.doi.org/10.1093/imanum/drl045.

[19] M. Chatard, *Asymptotic behavior of the Scharfetter-Gummel scheme for the drift-diffusion model*, in Finite volumes for complex applications VI. Problems & perspectives. Volume 1, 2, vol. 4 of Springer Proc. Math., Springer, Heidelberg, 2011, pp. 235–243, doi:10.1007/978-3-642-20671-9_25, https://doi.org/10.1007/978-3-642-20671-9_25.

[20] G. Chavent and J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation*, vol. 17, North-Holland, Amsterdam, stud. math. appl. ed., 1986.

[21] P. G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[22] P. Destuynder and B. Métivet, *Explicit error bounds in a conforming finite element method*, Math. Comp., 68 (1999), pp. 1379–1396, doi:10.1090/S0025-5718-99-01093-5, http://dx.doi.org/10.1090/S0025-5718-99-01093-5.

[23] D. A. Di Pietro, M. Vohralík, and S. Yousef, *Adaptive regularization, linearization, and discretization and a posteriori error control for the two-phase Stefan problem*, Math. Comp., 84 (2015), pp. 153–186, doi:10.1090/S0025-5718-2014-02854-8, http://dx.doi.org/10.1090/S0025-5718-2014-02854-8.

[24] V. Dolejší, A. Ern, and M. Vohralík, *A framework for robust a posteriori error control in unsteady nonlinear advection-diffusion problems*, SIAM J. Numer. Anal., 51 (2013), pp. 773–793, doi:10.1137/110859282, http://dx.doi.org/10.1137/110859282.

[25] A. Ern and J. L. Guermond, *Theory and Practice of Finite Elements*, vol. 159 of Applied Mathematical Series, Springer, New York, 2004.

[26] A. Ern, I. Smears, and M. Vohralík, *Guaranteed, locally space-time efficient, and polynomial-degree robust a posteriori error estimates for high-order discretizations of parabolic problems*, SIAM J. Numer. Anal., 55 (2017), pp. 2811–2834, doi:10.1137/16M1097626, https://doi.org/10.1137/16M1097626.

[27] A. Ern and M. Vohralík, *Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs*, SIAM J. Sci. Comput., 35 (2013), pp. A1761–A1791, doi:10.1137/120896918, http://dx.doi.org/10.1137/120896918.

[28] A. Ern and M. Vohralík, *Polynomial-degree-robust a posteriori estimates in a unified setting for conforming, nonconforming, discontinuous Galerkin, and mixed discretizations*, SIAM J. Numer. Anal., 53 (2015), pp. 1058–1081, doi:10.1137/130950100, http://dx.doi.org/10.1137/130950100.

[29] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods.* Ciarlet, P. G. (ed.) et al., in Handbook of numerical analysis. North-Holland, Amsterdam, pp. 713–1020, 2000.

[30] R. Eymard, C. Guichard, R. Herbin, and R. Masson, *Vertex-centred discretization of multiphase compositional Darcy flows on general meshes*, Comput. Geosci., 16 (2012), pp. 987–1005.

[31] R. Eymard, C. Guichard, R. Herbin, and R. Masson, *Gradient schemes for two-phase flow in heterogeneous porous media and Richards equation*, ZAMM - J. of App. Math. and Mech., 94 (2014), pp. 560–585.

[32] R. Eymard, M. Gutnic, and D. Hilhorst, *The finite volume method for Richards equation*, Comput. Geosci., 3 (1999), pp. 259–294, doi:10.1023/A:1011547513583.

[33] R. Eymard, R. Herbin, and A. Michel, *Mathematical study of a petroleum-engineering scheme*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 937–972.

[34] R. Eymard, D. Hilhorst, and M. Vohralík, *A combined finite volume–nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems*, Numer. Math., 105 (2006), pp. 73–131, doi:10.1007/s00211-006-0036-z.

[35] A. Fiebach, A. Glitzky, and A. Linke, *Uniform global bounds for solutions of an implicit Voronoi finite volume method for reaction–diffusion problems*, Numer. Math., 128 (2014),

p. 31–72, doi:10.1007/s00211-014-0604-6, https://doi.org/10.1007/s00211-014-0604-6.

[36] A. Fiebach, A. Glitzky, and A. Linke, *Convergence of an implicit voronoi finite volume method for reaction-diffusion problems*, Numer. Methods Partial Differential Equations, 32 (2016), p. 141–174, doi:10.1002/num.21990, https://doi.org/10.1002/num.21990.

[37] F. Filbet and M. Herda, *A finite volume scheme for boundary-driven convection-diffusion equations with relative entropy structure*, Numerische Mathematik, (2017), https://hal.archives-ouvertes.fr/hal-01326029.

[38] P. A. Forsyth, *A control volume finite element approach to NAPL groundwater contamination.*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1029–1057.

[39] F. Hecht, *New development in FreeFem++*, J. Numer. Math., 20 (2012), pp. 251–265.

[40] R. Herbin, *An error estimate for a finite volume scheme for a diffusion–convection problem on a triangular mesh*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 165–173, doi:10.1002/num.1690110205, https://doi.org/10.1002/num.1690110205.

[41] J. W. Jerome and M. E. Rose, *Error estimates for the multidimensional two-phase Stefan problem*, Math. Comp., 39 (1982), pp. 377–414, doi:10.2307/2007320, http://dx.doi.org/10.2307/2007320.

[42] A. M. Meirmanov, *The Stefan problem*, vol. 3 of de Gruyter Expositions in Mathematics, Walter de Gruyter & Co., Berlin, 1992. Translated from the Russian by Marek Niezgódka and Anna Crowley, With an appendix by the author and I. G. Götz.

[43] A. Michel, *A finite volume scheme for two-phase immiscible flow in porous media*, SIAM J. Numer. Anal., 41 (2003), pp. 1301–1317.

[44] A. Moussa, *Some variants of the classical Aubin-Lions Lemma*, J. Evol. Equ., 16 (2016), pp. 65–93.

[45] R. H. Nochetto, M. Paolini, and C. Verdi, *An adaptive finite element method for two-phase Stefan problems in two space dimensions. I. Stability and error estimates*, Math. Comp., 57 (1991), pp. 73–108, S1–S11, doi:10.2307/2938664, http://dx.doi.org/10.2307/2938664.

[46] R. H. Nochetto, A. Schmidt, and C. Verdi, *A posteriori error estimation and adaptivity for degenerate parabolic problems*, Math. Comp., 69 (2000), pp. 1–24.

[47] F. Otto, *L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations*, J. Differential Equations, 131 (1996), pp. 20–38.

[48] B. Perthame, *Parabolic equations in biology*, Lecture Notes on Mathematical Modelling in the Life Sciences, Springer, 2015.

[49] J. Rulla and N. J. Walkington, *Optimal rates of convergence for degenerate parabolic problems in two dimensions*, SIAM J. Numer. Anal., 33 (1996), pp. 56–67.

[50] J. L. Vázquez, *The porous medium equation*, Oxford Mathematical Monographs, The Clarendon Press Oxford University Press, Oxford, 2007. Mathematical theory.

[51] A. Visintin, *Models of phase transitions*, vol. 28 of Progress in nonlinear differential equations and their applications, Birkhäuser Boston, 1996.