ESAIM: M2AN 54 (2020) 591–618
https://doi.org/10.1051/m2an/2019075

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

ENHANCED POSITIVE VERTEX-CENTERED FINITE VOLUME SCHEME FOR ANISOTROPIC CONVECTION-DIFFUSION EQUATIONS

El Houssaine Quenjel*

Abstract. This article is about the development and the analysis of an enhanced positive control volume finite element scheme for degenerate convection-diffusion type problems. The proposed scheme involves only vertex unknowns and features anisotropic fields. The novelty of the approach is to devise a reliable upwind approximation with respect to flux-like functions for the elliptic term. Then, it is shown that the discrete solution remains nonnegative. Under general assumptions on the data and the mesh, the convergence of the numerical scheme is established owing to a recent compactness argument. The efficiency and stability of the methodology are numerically illustrated for different anisotropic ratios and nonlinearities.

Mathematics Subject Classification. 35K65, 65M08, 65M12.

Received May 17, 2019. Accepted October 17, 2019.

1. Introduction

Convection-diffusion equations arise from uncountable physical processes such as modeling complex flows in porous media [14, 15], environmental sciences [31], chemistry and biology, etc. [33]. Performing numerical methods for the approximation of such models is of significant importance in understanding the behavior of their solutions. Despite the large number of works on the topic, devising stable and convergent discretizations is still a challenging task.

In the present paper we focus our attention on the design and analysis of a positive finite volume scheme for a particular class of nonlinear degenerate convection-diffusion equations recast in the following formulation

$$\partial_t u - \operatorname{div}\left(\varphi(u)\Lambda\nabla u\right) + \operatorname{div}\left(f(u)\mathbf{V}\right) = 0 \quad \text{in } Q_{\mathfrak{T}} := \Omega \times (0, \mathfrak{T}). \tag{1.1}$$

The domain Ω is a polygonal connected bounded open of $\mathbb{R}^d$ $(d \geq 2)$ and $\mathfrak{T}$ represents the physical time. The velocity of the convection process is provided by $\mathbf{V}$. In addition, φ is a nonlinear nonnegative function that accounts for the behavior of the diffusion within Ω. The matrix Λ gives information on the anisotropy of the considered domain. The primary unknown of the problem is u, which may stand for the concentration, saturation, density, etc. To close the model problem, a boundary condition of Dirichlet–Neumann type is prescribed:

$$u = 0, \quad \text{on} \quad \Gamma_1 \times (0, \mathfrak{T}), \quad \Big(\varphi(u)\Lambda\nabla u - f(u)\mathbf{V}\Big)\cdot\mathbf{n} = 0, \quad \text{on} \quad \Gamma_2 \times (0, \mathfrak{T}), \tag{1.2}$$

Keywords and phrases. Finite volume, positive, convergence, convection, diffusion.

University of Nice Sophia-Antipolis, LJAD, CNRS UMR 7351, and COFFEE Team, INRIA Sophia Antipolis Méditerran'ee, Parc Valrose, 06108 Nice Cedex 02, France.
*Corresponding author: quenjel@unice.fr

Article published by EDP Sciences

where $\{\Gamma_1, \Gamma_2\}$ is a disjoint partition of the boundary $\partial\Omega$ such that $|\Gamma_1| > 0$ and $\mathbf{n}$ the outward unit normal vector to Γ_2. The initial solution is known at $t = 0$:

$$u(\cdot, 0) = u^0, \quad \text{in} \quad \Omega, \tag{1.3}$$

where u^0 is given. It is worth mentioning that (1.1) includes several major difficulties encountered in the development and the convergence analysis of modern finite volume schemes addressed to such problems. This is mainly due to the degeneracy of the function φ together with the anisotropy of the tensor Λ and the geometric configuration of the mesh.

Before defining the nature of the approximate solution, the following hypotheses are required on the data. They will be assumed to hold throughout the rest of this paper.

($\mathbf{A}_1$) u^0 is an $L^2(\Omega)$ function with $u^0(x) \geq 0$ *a.e.* $x \in \Omega$.

($\mathbf{A}_2$) $\Lambda : \Omega \longrightarrow \mathbb{R}^{d\times d}$ is a symmetric uniformly coercive matrix *i.e.* there exist positive constants $\underline{\Lambda}$ and $\overline{\Lambda}$ such that

$$\underline{\Lambda}|\zeta|^2 \leq \Lambda(x)\zeta \cdot \zeta \leq \overline{\Lambda}|\zeta|^2, \text{ for all } \zeta \in \mathbb{R}^d \quad \text{and} \quad a.e.\ x \in \Omega.$$

($\mathbf{A}_3$) $\mathbf{V}$ belongs to $C^0(\overline{\Omega} \times [0, \mathfrak{T}])$ and satisfies $\operatorname{div} \mathbf{V} = 0$.

($\mathbf{A}_4$) φ is a continuous nondecreasing function from $\mathbb{R}^+$ into $\mathbb{R}^+$ such that $\varphi(0) = 0$ and $\varphi(s) > 0$ for all $s > 0$. It is further extended by 0 on $(-\infty, 0)$.

($\mathbf{A}_5$) Kirchhoff's function ξ is defined by

$$\xi(v) = \int_0^v \sqrt{\varphi(s)}\, \mathrm{d}s, \quad \forall v \in \mathbb{R}^+. \tag{1.4}$$

The function ξ^{-1} exists, is continuous and increasing. We also need to control the behavior of φ for large values by supposing the growth assumption

$$\exists \varepsilon > 0 \quad \forall s \in \mathbb{R}^+ \quad \varphi(s) \leq C(1 + \xi(s)^{2-\varepsilon}). \tag{1.5}$$

Moreover, we assume that

$$\forall s \in \mathbb{R}^+ \quad s \leq C(1 + \xi(s)). \tag{1.6}$$

($\mathbf{A}_6$) The mobility-like function f is increasing, assumed to be in $\mathcal{C}^1(\mathbb{R}^+, \mathbb{R}^+)$ and fulfills

$$f(0) = 0, \quad \|f'\|_\infty < \infty.$$

It is also extended by 0 on $(-\infty, 0)$.

Remark 1.1. Assumption ($\mathbf{A}_4$) especially (1.5) accounts for the weakest possible requirement on the nonlinear diffusion function. For instance, it allows any polynomial growth of φ and it is satisfied by a wide range of nonlinearities. Notice that ε should take very small positive values. For a convenient fixed ε, there holds

$$\frac{\xi(s)^2}{\xi(s)^{2-\varepsilon}} \longrightarrow +\infty \quad \text{as} \quad s \longrightarrow +\infty.$$

Consequently, there exists $\delta_\varepsilon > 0$ and $C_{\delta_\varepsilon} > 0$ fulfilling

$$\forall s \in \mathbb{R}^+ \quad \xi(s)^{2-\varepsilon} \leq C_{\delta_\varepsilon} + \delta_\varepsilon \xi(s)^2. \tag{1.7}$$

The hypothesis (1.6) incorporates a class of functions obeying the following limit on ξ

$$\frac{\xi(s)}{s} \longrightarrow l \in (0, +\infty] \quad \text{as} \quad s \longrightarrow +\infty.$$

Remark 1.2. (i) Under Assumption ($\mathbf{A}_6$), especially $f(0)=0$, $\|f'\|_\infty < \infty$, f is a globally Lipschitz continuous function. Then

$$\forall s \in \mathbb{R}^+ \quad |f(s)| \le \|f'\|_\infty \, s \le C\Big(1+\xi(s)\Big). \tag{1.8}$$

(ii) The monotonicity condition on f in Assumption ($\mathbf{A}_6$) might seem restrictive. It is only taken to simplify the presentation of the upwind discretization of the hyperbolic term. The general case can be tackled by considering a classical splitting of type $f = f_\uparrow + f_\downarrow$ where $f_\uparrow$ is the increasing part while $f_\downarrow$ is the decreasing part. Then, this convective term rewrites

$$\operatorname{div}\Big(f(u)\mathbf{V}\Big) = \operatorname{div}\Big(f_\uparrow \mathbf{V}\Big) + \operatorname{div}\Big((-f_\downarrow)(-\mathbf{V})\Big).$$

Therefore, the approximation of the right hand side of this equality using a first order upstream scheme yields the Engquist–Osher numerical flux. Hence, the analysis carried out through the paper remains valid.

We recall the classical Sobolev space

$$H^1_{\Gamma_1}(\Omega) = \{u \in H^1(\Omega) \ / \ u = 0 \text{ on } \Gamma_1\},$$

equipped by the standard norm

$$||u||_{H^1_{\Gamma_1}} = ||\nabla u||_{L^2(\Omega)^d}.$$

We next define the notion of a weak solution to the continuous model.

Definition 1.3. Under Assumptions ($\mathbf{A}_1$)–($\mathbf{A}_6$,), we say that u is a weak solution to the problem (1.1)–(1.3) if:

$$u, \varphi(u) \in L^1(Q_\mathfrak{T}),\ u \ge 0 \ a.e.\ in\ Q_\mathfrak{T},\ \xi(u) \in L^2(0,\mathfrak{T}; H^1_{\Gamma_1}(\Omega)),$$

and for all $\forall \psi \in \mathcal{C}^\infty_c(\overline{\Omega} \times [0,\mathfrak{T}))$, such that $\psi = 0$ on $\Gamma_1 \times [0,\mathfrak{T})$, one has

$$\begin{aligned} &-\int_{Q_\mathfrak{T}} u\partial_t\psi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega u^0\psi(x,0)\,\mathrm{d}x + \int_{Q_\mathfrak{T}} \sqrt{\varphi}(u)\Lambda\nabla\xi(u)\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t \\ &-\int_{Q_\mathfrak{T}} f(u)\mathbf{V}\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t = 0. \end{aligned} \tag{1.9}$$

Existence of a weak solution to (1.1)–(1.3) has been proved in [14]. Under smoothness assumptions on the data, especially on f and $f \circ \xi^{-1}$, one can retain the uniqueness of the solution. See for instance [28] for more information.

During the past few decades, enormous amount of numerical methods were developed for approaching the solutions to problems having the form (1.1). In particular, finite element schemes were proposed and studied in [6,14]. Approximations based on finite elements and finite volumes were discussed and analyzed in [3,25]. Owing to their attractive properties and cheaper computational cost, finite volume schemes have received an increasing interest on both: the application and the theoretical sides. They have been extensively applied to and preferred for systems resulting from conservation laws [35] written like (1.1). In fact, the main idea of such an approach is based on a balance equation where the approximation of the fluxes through the interfaces defines the method in question. A simple and very practical finite volume scheme is the pioneer two-point flux approximation (TPFA) method, which has been used in Computational Fluid Dynamics in a large context. In the framework of porous media flows, the convergence study for such a scheme has been carried out for instance in [9,23,24,39]. Contrary to the TPFA approach, the control volume finite element methodology (CVFE) [10,22,34] features cell grids centered on the vertices. Its structure resembles to that of the TPFA scheme. However, schemes written in the two-point formulation may lose consistency in the presence of anisotropy or lack robustness on distorted meshes. This has led to the development of alternatives utilizing more than two points. Depending on the stencil, many propositions exist in the literature. We may cite multi-point flux approximation schemes (MPFA) [1, 2], discrete duality finite volume methods (DDFV) [4, 17], vertex approximation gradient approach (VAG) [8, 27]. We also mention the mixed finite volume scheme (MFV) [19] as well as the hybrid strategy (HFV) [26]. Except TPFA-based methods, the aforementioned discretizations belong to the gradient discretization formalism [20].

The primordial advantage of TPFA-like schemes [23] resides in theirs native unconditional monotony. Yet, this comes with the price of a geometric shape condition on the grids of the mesh. Such a constraint is too restrictive in practice. Without more assumptions on the anisotropy of the tensor Λ or on the mesh, the monotonicity property is not an easy outcome for schemes based on the gradient discretization methods as pointed out in [18, 32]. The latter are somehow based on central approximations of the fluxes across edges which may offer a high resolution even in the presence of heterogeneities, especially for diffusion processes. On the hand, they may induce undershoots and/or overshoots on coarse meshes in case of strong anisotropic fields and advected-dominated regimes. To eliminate such oscillations, many works have been devoted to this subject. Among them, only a few papers were focused on the convergence of the numerical scheme [11,12,30,38,40]. The authors in [12] suggested a positive finite volume scheme that is applicable to complex situations. Nevertheless, it suffers from an excessive numerical viscosity and thus the convergence rate is completely deteriorated in case of zero-flux boundary conditions. To alleviate this issue, we have proposed a correction in [29] that shows a considerable improvement, especially for isotropic media. The point is that the correction occurs when it is only necessary. Practically, the idea still however necessitates an enhancement targeting big anisotropic ratios. As a consequence, designing novel positive schemes which ameliorate the numerical convergence is of chief interest.

In the current contribution we develop a new positive vertex-centered finite volume discretization for the model example (1.1). The scheme is derived thanks to a direct approximation of the fluxes across the interfaces of the control volumes. This is achieved by replacing the continuous gradient by its discrete counterpart. To avoid the use of Kirchhoff's transform, a special remedy is proposed. The originality of our approach consists in eliminating the "bad" terms by using an adequate upwind scheme with respect to nonlinear flux-like functions. This technique allows to prevent the scheme from producing too artificial diffusion in case of strong anisotropic rates. The numerical analysis of the method stems some inspiration from [13]. However, our approach uses only vertex unknowns without processing centered ones. Numerical tests confirm our prediction where optimal convergence rate are achieved as known for popular approximations to (1.1). We emphasize that our technique can be extended to more complex problems involving (1.1) without any major issue.

The remainder of this article is organized as follows. Section 2 exposes the discrete setting of the considered vertex-centered discretization. Section 3 describes the positive numerical scheme as well as some preliminary results. Section 4 is concerned with the proof of a discrete maximum principle on the solution and the *a priori* estimates on the gradient of the Kirchhoff function. In Section 5 we study the convergence of the finite volume scheme by means of a recent compactness argument. Finally, Section 6 reports a number of academic tests in order to illustrate the efficiency of our methodology and its ability to deal with sever anisotropic situations.

From now non, we focus on the two-dimensional setting to ease the readability of our approach. The extension to three dimension follows in the same fashion.

2. Discrete framework

This section aims to specify and fix some terminologies that will be used in the sequel. For expositional convenience, we would like to keep most of our notations as in the same spirit of the standard finite volumes so that our scheme can be understandable by the reader.

The finite volume or the dual mesh $\mathcal{M}$ is defined around vertices of a given primal partition $\mathcal{T}$ of Ω. In the two dimensional case, the considered primal mesh is nothing more than a conforming triangulation [16]. $\mathcal{T}$ is about a collection of open subsets (triangles in 2D or simplices in 3D setting) covering Ω. In 2D the intersection of triangles is either an edge, a point or the empty set. In 3D the simplices may share a face, a straight segment, a vertex or the empty set. We refer to $\mathcal{V}$ as the set of the mesh nodes. We consider $\mathcal{V}_T$ the set of the vertices of

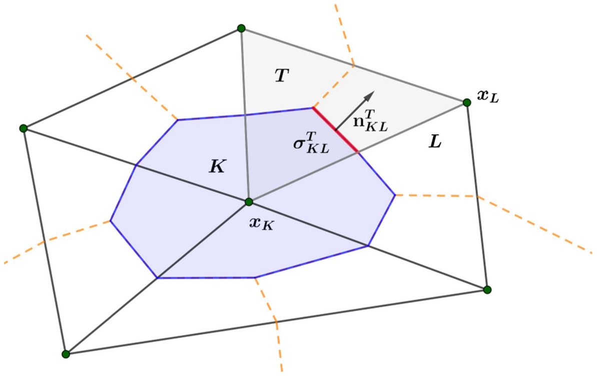


FIGURE 1. Schematic illustration of the control volume and some related notations in the two dimensional case.

the element T. We denote $|T|$ (resp. x_T, h_T, ϱ_T) the Lebesgue measure (resp. barycenter, diameter, the diameter of the biggest inscribed ball) of the simplex $T \in \mathcal{T}$. We consider $h_{\mathcal{T}} = \max\{h_T,\ T \in \mathcal{T}\}$.

The control volumes or the dual cells are then constructed around the vertices of $\mathcal{T}$. This is performed locally in each element. For instance, in the 2D case, it suffices to connect the midpoints of the edges of a triangle $T \in \mathcal{T}$ with the barycenter of the same triangle as depicted in Figure 1. Following the finite volumes philosophy we define $\mathcal{E}$ the set of dual interfaces of $\mathcal{M}$. We also consider $\mathcal{E}_T = \{\sigma \in \mathcal{E}/\sigma \subset T\}$, for all $T \in T$. This gives $\mathcal{E} = \cup_{T\in T}\mathcal{E}_T$. In the two dimensional setting each dual interface inside the simplex T is obtained through two vertices x_K and x_L: $\sigma_{KL}^T := \overline{K} \cap \overline{L} \cap \overline{T}$. Moreover, the vector $\mathbf{n}_{KL}^T$ stands for to the unit normal to σ_{KL}^T pointed from K to L. We designate by $\left|\sigma_{KL}^T\right|$ the $(d-1)$-dimensional measure of this face.

We hereafter would like to emphasize that the 2D aspect is mostly employed through this paper while the 3D version is fully analogous.

Let $\{\mathcal{T}_h\}_h$ be a sequence of refined meshes of $\mathcal{T}$. We assume that $\{\mathcal{T}_h\}_h$ is regular in the sense that the following shape condition holds:

$$\exists \Theta > 0 : \max_{T\in\mathcal{T}_h} \frac{h_T}{\varrho_T} \leq \Theta, \quad \forall h_{\mathcal{T}_h} \in (0, h_0). \tag{2.1}$$

We also suppose that the articulation points of Γ_1 and Γ_2 are the midpoint of some boundary edges of $\mathcal{T}$.

Next, a one-step approximation in time is taken into account. To this end, the time interval $(0, \mathfrak{T})$ is broken into a collection of sub-intervals whose boundaries can be defined by an increasing sequence $(t^n)_{n\in[\![0,N]\!]}$. We moreover take a uniform time step δt for sake of simplicity. One can get $t^n = n\delta t$ for all $n \in [\![0, N]\!]$. Then, we deduce that $\delta t = \mathfrak{T}/N$

Define

$$V_h = \{u_h \in \mathcal{C}^0(\overline{\Omega}),\ u_{h|T} \in \mathbb{P}_1(T),\ \forall T \in \mathcal{T}\},$$

where $\mathbb{P}_1(T)$ is the set of polynomial functions of degree at most 1 on T. The Lagrange finite element basis of V_h is denoted by $(\phi_K)_{K\in\mathcal{M}}$ where $\phi_K(x_L) = 1$ for $K = L$ and $\phi_K(x_L) = 0$ otherwise. Then, every function of V_h writes

$$u_h = \sum_{K\in\mathcal{M}} u_K \phi_K.$$

We hereafter set $\mathcal{M}_D = \{K \in \mathcal{M}/x_K \in \Gamma_1\}$ and $\mathcal{M}_D^c = \mathcal{M}\backslash\mathcal{M}_D$. We consider the functional space of V_h defined by

$$V_h^0 = \{u_h \in V_h :\ u_h(x_K) = 0,\ \forall K \in \mathcal{M}_D\} \subset H^1_{\Gamma_1}(\Omega).$$

The standard metric on V_h^0 comes from the energy norm

$$\|u_h\|_{V_h^0} = \int_\Omega |\nabla u_h|^2 \,\mathrm{d}x.$$

To perform the analysis of the numerical scheme we need to specify the trial space. We talk about the finite volume space W_h consisting of piecewise constant functions on the dual cells. Each element of W_h is uniquely represented by

$$\widetilde{u}_h = \sum_{K\in\mathcal{M}} u_K \mathbf{1}_{\mathring{K}}, \quad (u_K \in \mathbb{R},\ \forall K \in \mathcal{M}),$$

where $\mathbf{1}_{\mathring{K}}$ designates the characteristic of the topological interior of A_K.

Finally, we associate $z_h^{n+1} \in \{u_h^{n+1}, \widetilde{u}_h^{n+1}\}$, for $n \in [\![0, N-1]\!]$, with the discrete function $z_{h,\delta t}$ which is defined in each sub-interval $(t^n, t^{n+1}]$ by z_h^{n+1}. We then define

$$W_{h,\delta t}^0 = \{\widetilde{u}_{h,\delta t} \ / \ \widetilde{u}_{h,\delta t}(x_K, \cdot) = 0,\ \forall K \in \mathcal{M}_D\}.$$

In case of a nonlinear function G, the reconstruction of the composition $G \circ u$ is provided by the interpolation. This means that

$$G(u_{h,\delta t}) = (G\circ u)_{h,\delta t} \text{ and } \widetilde{G}(u_{h,\delta t}) = \widetilde{(G\circ u)}_{h,\delta t}.$$

We recall the discrete integration by parts formula

$$\int_T \Lambda\nabla u_h \cdot \nabla v_h \,\mathrm{d}x = \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \Lambda_{KL}^T (u_K - u_L)(v_K - v_L), \quad \forall T \in \mathcal{T}, \tag{2.2}$$

where Λ_{KL}^T is termed the transmissibility coefficient:

$$\Lambda_{KL}^T = -\int_T \Lambda\nabla\phi_K \cdot \nabla\phi_L \,\mathrm{d}x.$$

Remark 2.1. (i) We would like to point out that the tensor Λ can be assumed to be constant on the elements of $\mathcal{T}$. In the general case, one can retain this assumption by taking the average of Λ over T:

$$(\Lambda^h)_{|T} = \frac{1}{|T|}\int_T \Lambda(x)\,\mathrm{d}x, \quad \forall T \in \mathcal{T}.$$

(ii) If the triangulation $\mathcal{T}$ satisfies Delaunay's condition and the tensor Λ reduces to a positive scalar function, then all Λ_{KL}^T are nonnegative [21].

Let us consider $T \in \mathcal{T}$. Choosing for instance the positive rotation sense, we define the permutation τ^T which allows to connect two vertices of T. To fix the ideas, we assume that $\mathcal{V}_T = \{x_K, x_L, x_S\}$. For each vertex $i \in \mathcal{V}_T$ we associate with a neighbor vertex $\tau^T(i)$ of the same triangle. Then, one gets $\tau^T(\mathcal{V}_T) = \{x_L, x_S, x_K\}$. As a consequence, one can write $\mathcal{E}_T$ such that $\mathcal{E}_T = \{\sigma_{i\tau^T(i)}^T\}_{i\in\mathcal{V}_T}$. By abuse of notation we may keep τ instead of τ^T. Now, using the fact that

$$|T|\,\nabla\phi_i = -\Big(\left|\sigma_{i\tau(i)}^T\right| \mathbf{n}_{i\tau(i)}^T + \left|\sigma_{i\tau\circ\tau(i)}^T\right| \mathbf{n}_{i\tau\circ\tau(i)}^T\Big),$$

one can see in a straightforward way that the discrete gradient can be reformulated as follows:

$$\nabla u_{h|T} = -\frac{1}{|T|}\sum_{i\in\mathcal{V}_T} \left|\sigma_{i\tau(i)}^T\right| (u_i - u_{\tau(i)})\mathbf{n}_{i\tau(i)}^T. \tag{2.3}$$

We next introduce a local matrix $\mathbb{M}^T = (\beta_{ij})_{1\le i,j\le \#\mathcal{V}_T}$ whose entries read

$$\beta_{ij} = \frac{1}{|T|}\left|\sigma^T_{i\tau(i)}\right|\left|\sigma^T_{j\tau(j)}\right| \Lambda \mathbf{n}^T_{i\tau(i)} \cdot \mathbf{n}^T_{j\tau(j)}. \tag{2.4}$$

Combining (2.3) and (2.4), one obtains the two-point like relation

$$-\nabla u_h \cdot \left|\sigma^T_{j\tau(j)}\right| \mathbf{n}^T_{j\tau(j)} = \sum_{i\in\mathcal{V}_T} \beta_{ij}(u_i - u_{\tau(i)}). \tag{2.5}$$

Therefore, the integral (2.2) is expressed by

$$\begin{aligned}\int_T \Lambda\nabla u_h\cdot\nabla v_h\,\mathrm{d}x &= \sum_{i\in\mathcal{V}_T}\sum_{j\in\mathcal{V}_T} \beta_{ij}(u_i - u_{\tau(i)})(v_j - v_{\tau(j)})\\ &= \delta_T u\cdot \mathbb{M}^T\delta_T v, \quad \forall T\in\mathcal{T},\end{aligned} \tag{2.6}$$

where we have set

$$\delta_T u = \begin{pmatrix} u_i - u_{\tau(i)}\\ u_{\tau(i)} - u_{\tau\circ\tau(i)}\\ u_{\tau\circ\tau(i)} - u_i\end{pmatrix}.$$

Whence, one gets the following crucial relationship that is drawn from (2.2) and (2.6).

Lemma 2.2. *For every $T\in\mathcal{T}$, one has*

$$\int_T \Lambda\nabla u_h\cdot\nabla v_h\,\mathrm{d}x = \delta_T u\cdot\mathbb{M}^T\delta_T v = \sum_{\sigma^T_{KL}\in\mathcal{E}_T} \Lambda^T_{KL}(u_K - u_L)(v_K - v_L), \quad \forall u,v\in\mathbb{R}^{\#\mathcal{V}_T}.$$

As a direct consequence of this result, one claims that the matrix $\mathbb{M}^T$ is definite-positive since the tensor Λ is so. On the other hand, we observe that : if all $\Lambda^T_{KL}\ge 0$, the transformation (2.6) yields a stiffness matrix which preserves the M-matrix structure [41].

The following result confirms the equivalence of some specific norms on $\mathbb{R}^{\#\mathcal{V}_T}$ that will be useful later on. It shows on a second place that the condition number of $\mathbb{M}^T$ is uniformly bounded independently of the mesh steps.

Lemma 2.3. *There exist positive constants C_1, C_2 depending only on $\Theta, \underline{\Lambda}$ and $\overline{\Lambda}$ such that for ever every $T\in\mathcal{T}$ there holds*

$$C_1\delta_T u\cdot\delta_T u \le \delta_T u\cdot\mathbb{M}^T\delta_T u \le C_2\delta_T u\cdot\delta_T u, \quad \forall u\in\mathbb{R}^{\#\mathcal{V}_T}. \tag{2.7}$$

Proof. The proof is given in the two dimensional case. Using Lemma 2.2 yields

$$\delta_T u\cdot\mathbb{M}^T\delta_T u \le \sum_{\sigma^T_{KL}\in\mathcal{E}_T} \left|\Lambda^T_{KL}\right|(u_K - u_L)^2.$$

Now, the regularity of the mesh (2.1) guarantees the existence of a $\overline{C}>0$ depending only on Θ and $\overline{\Lambda}$

$$\left|\Lambda^T_{KL}\right| \le \overline{C}, \quad \forall\sigma^T_{KL}\in\mathcal{E}_T.$$

Let $\mu^T_{\max}$ (resp. $\mu^T_{\min}$) stand for the biggest (reps. smallest) eigenvalue of the local matrix $\mathbb{M}^T$. We denote by $v_{\mu^T_{\max}}$, $v_{\mu^T_{\min}}$ some eigenvectors associated to $\mu^T_{\max}$, $\mu^T_{\min}$ respectively. It is easy to check that the operator $\delta_T : w\mapsto\delta_T w$ defines a surjection from $\mathbb{R}^{\#\mathcal{V}_T}$ into $\mathbb{R}^{\#\mathcal{V}_T}$. Therefore, we make use of the latter fact to deduce

$$\mu^T_{\max} \le \overline{C}.$$

By taking $C_2 = \overline{C}$, we show the second inequality of the required result (2.7). Following [8, 12] and bearing in mind the coercivity of Λ, one can find a $C_1' > 0$:

$$\delta_T u \cdot \delta_T u = \sum_{\sigma_{KL}^T \in \mathcal{E}_T} (u_K - u_L)^2 \le C_1' \left\| \sqrt{\Lambda} \nabla u_h \right\|_{L^2(T)^2}^2 .$$

Combine (2.2) and once more Lemma 2.2 to obtain

$$\frac{1}{C_1'} \le \mu_{\min}^T .$$

This conclude the proof by setting $C_1 = 1/C_1'$. □

We now show that the difference between the finite volume and finite element reconstruction is dominated by the norm of the discrete gradient up to the size of $\mathcal{T}$.

Lemma 2.4. *Let w_h be an element of V_h. Define two piecewise functions $\overline{w}_h, \underline{w}_h$ such that*

$$\overline{w}_{h|T} = \max_{x \in T} w_h(x), \quad \underline{w}_{h|T} = \min_{x \in T} w_h(x), \quad \forall T \in \mathcal{T}.$$

Then

$$\| \overline{w}_h - \underline{w}_h \|_{L^2(\Omega)} \le \# \mathcal{V}_T h_{\mathcal{T}} \| \nabla w_h \|_{L^2(\Omega)^d} , \quad \| \widetilde{w}_h - w_h \|_{L^2(\Omega)} \le \# \mathcal{V}_T h_{\mathcal{T}} \| \nabla w_h \|_{L^2(\Omega)^d} .$$

Proof. Let us select a T in $\mathcal{T}$. By the definition of the function w_h we get

$$\left| \overline{w}_{h|T} - \underline{w}_{h|T} \right| \le \# \mathcal{V}_T \max_{i \in \mathcal{V}_T} \left| w_i - w_{\tau(i)} \right| \le \# \mathcal{V}_T h_{\mathcal{T}} \left| \nabla w_{h|T} \right| .$$

Therefore

$$\| \overline{w}_h - \underline{w}_h \|_{L^2(\Omega)}^2 = \sum_{T \in \mathcal{T}} |T| \left| \overline{w}_{h|T} - \underline{w}_{h|T} \right|^2 \le (\# \mathcal{V}_T)^2 h_{\mathcal{T}}^2 \sum_{T \in \mathcal{T}} |T| \left| \nabla w_{h|T} \right|^2 .$$

This proves the first inequality. Concerning the second one, it suffices to observe that

$$\left| \widetilde{w}_{h|T} - w_{h|T} \right| \le \left| \overline{w}_{h|T} - \underline{w}_{h|T} \right| .$$

Hence, one uses the previous inequality to conclude. The proof is then finished. □

Now, we are in a position to introduce the positive finite volume scheme.

3. Finite volume scheme

In view of stability reasons, we are led to perform a fully Euler implicit scheme in time. As usual, the finite volume discretization is derived by integrating model's equation on the dual cells and applying the divergence theorem to get balance relationships across dual edges. We point out that the hyperbolic fluxes are approximated by a classical upstream scheme. Here our contribution targets the diffusive terms. More precisely, through the approximation process the discrete diffusive flux

$$-\nabla \xi(u_h) \cdot \left| \sigma_{j\tau(j)}^T \right| \mathbf{n}_{j\tau(j)}^T = \sum_{i \in \mathcal{V}_T} \beta_{ij} \Big(\xi(u_i) - \xi(u_{\tau(i)}) \Big),$$

is replaced by a nonlinear function $\mathcal{F}_{KL,T}^{n+1}(u)$ to avoid the introduction of Kirchhoff's transformation. At this stage, we decide to upwind with respect to the underlined function so that the positivity can be reinforced.

Then, the positive scheme we propose consists in finding a finite family $(u_K^{n+1})_{K\in\mathcal{M},n\in[\![0,\mathbb{N}-1]\!]}$ fulfilling the following algebraic system made from:

$$u_K^0 = \begin{cases} \dfrac{1}{|K|}\displaystyle\int_K u^0(x)\,\mathrm{d}x, & \text{for} \quad K\in\mathcal{M}_D^c \\ 0 & \text{for} \quad K\in\mathcal{M}_D \end{cases}, \tag{3.1}$$

and the following set of equations at each time level $n\in[\![0,\mathbb{N}-1]\!]$

$$\begin{cases} \dfrac{|K|}{\delta t}(u_K^{n+1}-u_K^n) + \displaystyle\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}} \sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}_{KL,T}^{n+1}(u) & \\ + \displaystyle\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}} f(u_{KL}^{up,n+1})\mathbf{V}_{KL,T}^{n+1} = 0, & \text{for} \quad K\in\mathcal{M}_D^c \\ u_K^{n+1}=0, & \text{for} \quad K\in\mathcal{M}_D \end{cases} \cdot \tag{3.2}$$

The convective flux is written in the upstream fashion where

$$u_{KL}^{up,n+1} = \begin{cases} u_K^{n+1} & \text{if} \quad \mathbf{V}_{KL,T}^{n+1}\geq 0 \\ u_L^{n+1} & \text{otherwise} \end{cases}, \tag{3.3}$$

and the velocity at the interface σ_{KL}^T is computed by

$$\mathbf{V}_{KL,T}^{n+1} = \int_{\sigma_{KL}^T} \mathbf{V}^{n+1}\cdot\mathbf{n}_{KL}^T\,\mathrm{d}\sigma \ \text{ and } \mathbf{V}^{n+1} = \mathbf{V}(\cdot,t^{n+1}). \tag{3.4}$$

$\mathrm{d}\sigma$ is the $(d-1)$-dimensional measure. As we mentioned before, we aim to maintain the nonnegativity of the discrete solution. We are aware that possible undershoots may arise from the diffusion. To cope with this issue, we treat the nonlinear elliptic term is the same spirit of upstream schemes as proposed in [12] where the authors upwinded with respect to the transmissibility coefficients. The latter approach suffers from an excessive artificial diffusion. Here our idea draws some inspiration from [12], but the construction of the scheme is different. As a matter of fact, it consists in upwinding with regards to the discrete flux-like function defined by

$$\mathcal{F}_{KL,T}^{n+1}(u) = \frac{1}{|T|}\left(\sum_{i\in\mathcal{V}_T}\sqrt{\varphi_{i\tau(i)}^{n+1}}(u_i^{n+1}-u_{\tau(i)}^{n+1})\left|\sigma_{i\tau(i)}^T\right|\Lambda\mathbf{n}_{i\tau(i)}^T\right)\cdot\left|\sigma_{KL}^T\right|\mathbf{n}_{KL}^T, \tag{3.5}$$

which plays the role of the velocity in the context of the hyperbolic framework. We denote $a_{i,KL}^T = \dfrac{1}{|T|}\left|\sigma_{i\tau(i)}^T\right|\Lambda\mathbf{n}_{i\tau(i)}^T\cdot\left|\sigma_{KL}^T\right|\mathbf{n}_{KL}^T$. Therefore one can rewrite $\mathcal{F}_{KL,T}^{n+1}(u)$ as follows

$$\mathcal{F}_{KL,T}^{n+1}(u) = \sum_{i\in\mathcal{V}_T} a_{i,KL}^T\sqrt{\varphi_{i\tau(i)}^{n+1}}(u_i^{n+1}-u_{\tau(i)}^{n+1}).$$

As we are avoiding Kirchhoff's transformation, the factor $\varphi_{i\tau(i)}^{n+1}$ may be expressed as

$$\varphi_{i\tau(i)}^{n+1} = \frac{\varphi(u_i^{n+1})+\varphi(u_{\tau(i)}^{n+1})}{2}, \quad \forall i\in\mathcal{V}_T.$$

Like $u_{KL}^{up,n+1}$, we choose $u_{KL}^{\mathrm{diff},n+1}$ in the upstream sense with regards to the sign of $\mathcal{F}_{KL,T}^{n+1}(u)$:

$$u_{KL}^{\mathrm{diff},n+1} = \begin{cases} u_K^{n+1} & \text{if} \quad \mathcal{F}_{KL,T}^{n+1}(u)\geq 0 \\ u_L^{n+1} & \text{else} \end{cases}. \tag{3.6}$$

In order to make an easily readable analysis of the scheme, we can introduce two practical matrices as done for instance in [13]. We begin with defining $\mathbb{D}^T(u^{n+1})$ by

$$\forall i, j \in \mathcal{V}_T, \quad \mathbb{D}^T(u^{n+1})_{ij} = \begin{cases} \sqrt{\varphi^{n+1}_{i\tau(i)}} & \text{if} \quad j = \tau(i) \\ 0 & \text{if} \quad j \neq \tau(i) \end{cases}. \tag{3.7}$$

We next set

$$\mathbb{A}^T(u^{n+1}) = \mathbb{D}^T(u^{n+1})\mathbb{M}^T\mathbb{D}^T(u^{n+1}). \tag{3.8}$$

It follows that $\mathbb{A}^T := \mathbb{A}^T(u^{n+1})$ is symmetric and semi-definite since φ may degenerate at 0. The point of introducing this matrix lies in the crucial relationship

$$\sum_{\sigma^T_{KL} \in \mathcal{E}_T} \sqrt{\frac{\varphi(u^{n+1}_K) + \varphi(u^{n+1}_L)}{2}} \mathcal{F}^{n+1}_{KL,T}(u)(u^{n+1}_K - u^{n+1}_L) = \delta^{n+1}_T u \cdot \mathbb{A}^T \delta^{n+1}_T u,$$

where $(\delta^{n+1}_T u)_i = u^{n+1}_i - u^{n+1}_{\tau(i)}$ for all $i \in \mathcal{V}_T$. More importantly, this structure and the prominent properties of $\mathbb{M}^T$ will provide the key path to the coercivity of the scheme as it will be developed and detailed in Section 4.

Remark 3.1. (i) Due to the homogeneous Neumann condition (1.2) prescribed on Γ_2 and the assumption on $\overline{\Gamma}_1 \cap \overline{\Gamma}_2$, we indicate that the fluxes across the dual edges located on this part of the boundary do not contribute to (3.2).

(ii) According to Assumption $(\mathbf{A}_3)$ together with the choice (3.4), one gets the following discrete divergence-free equality $\sum\limits_{T \cap K \neq \emptyset} \sum\limits_{\sigma^T_{KL} \in \mathcal{E}_{K \cap T}} \mathbf{V}^{n+1}_{KL,T} = 0$ for all $K \in \mathcal{M}$. Thereby,

$$\begin{aligned} &\sum_{T \cap K \neq \emptyset} \sum_{\sigma^T_{KL} \in \mathcal{E}_{K \cap T}} f(u^{up,n+1}_{KL}) \mathbf{V}^{n+1}_{KL,T} \\ &\qquad = \sum_{T \cap K \neq \emptyset} \sum_{\sigma^T_{KL} \in \mathcal{E}_{K \cap T}} \Big(f(u^{n+1}_K) - f(u^{n+1}_L) \Big) \Big(- \mathbf{V}^{n+1}_{KL,T} \Big)^+, \end{aligned} \tag{3.9}$$

where we hereafter adopt the notation $x = x^+ + x^-$ where $x^+ = \max(x, 0)$ and $x^- = \min(x, 0)$ for all $x \in \mathbb{R}$.

(iii) We can switch the role of f by a nonlinear increasing function $F : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$. As a consequence of a discrete integration by parts computation and the fact that $\mathbf{V}^{n+1}_{KL,T} = -\mathbf{V}^{n+1}_{LK,T}$, we infer

$$\sum_{K \in \mathcal{M}} \sum_{T \cap K \neq \emptyset} \sum_{\sigma^T_{KL} \in \mathcal{E}_{K \cap T}} \Big(F(u^{n+1}_K) - F(u^{n+1}_L) \Big) \Big(- \mathbf{V}^{n+1}_{KL,T} \Big)^+ = 0. \tag{3.10}$$

4. *A priori* analysis

This section claims stability results that serve to prove the existence for the numerical scheme and study its convergence. We first begin by establishing a lower bound on any discrete solution. It is about processing the sign of scheme terms appropriately.

Lemma 4.1. *For every $n \in \llbracket 0, \mathbb{N} \rrbracket$, the numerical scheme* (3.1) *and* (3.2) *is positivity-preserving i.e.*

$$u^n_K \geq 0 \quad \forall K \in \mathcal{M}.$$

Proof. The proof is carried out by an induction argument. It is clear that the statement holds for $n = 0$. Select $K \in \mathcal{M}$ such that $u_K^{n+1} = \min\limits_{L\in\mathcal{M}} u_L^{n+1}$ that we assume negative. Multiplying the equation associated to K by $\left(u_K^{n+1}\right)^- = \min(u_K^{n+1}, 0) < 0$ gives

$$\frac{|K|}{\delta t}\left|(u_K^{n+1})^-\right|^2 - u_K^n\left(u_K^{n+1}\right)^- + \sum_{T\cap K\neq\emptyset}\ \sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}} \sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}_{KL,T}^{n+1}(u)\left(u_K^{n+1}\right)^-$$
$$+ \sum_{T\cap K\neq\emptyset}\ \sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}} f(u_{KL}^{up,n+1})\mathbf{V}_{KL,T}^{n+1}\left(u_K^{n+1}\right)^- = 0.$$

Assumption $(\mathbf{A}_4)$ and the definition of the upstream value (3.6) for $u_{KL}^{\mathrm{diff},n+1}$ ensures that

$$\sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}_{KL,T}^{n+1}(u)\left(u_K^{n+1}\right)^- \geq \sqrt{\varphi}(u_K^{n+1})\mathcal{F}_{KL,T}^{n+1}(u)\left(u_K^{n+1}\right)^- = 0,$$
$$\forall\sigma_{KL}^T \in \mathcal{E}_{K\cap T},\ \forall T : T\cap K \neq \emptyset.$$

Now, taking advantage of Assumption $(\mathbf{A}_6)$, we check in a similar manner that

$$f(u_{KL}^{up,n+1})\mathbf{V}_{KL,T}^{n+1}\left(u_K^{n+1}\right)^- \geq f(u_K^{n+1})\mathbf{V}_{KL,T}^{n+1}\left(u_K^{n+1}\right)^- = 0,$$
$$\forall\sigma_{KL}^T \in \mathcal{E}_{K\cap T},\ \forall T : T\cap K \neq \emptyset.$$

As a result $\left|(u_K^{n+1})^-\right|^2 - u_K^n\left(u_K^{n+1}\right)^- \leq 0$. Introduce the induction hypothesis to conclude that $\left(u_K^{n+1}\right)^- = 0$, which yields a contradiction. Hence, the proof is complete. □

We next state and prove energy estimates.

Lemma 4.2. *Let* $(u_K^{n+1})_{K\in\mathcal{M},n\in[\![0,\mathbb{N}-1]\!]}$ *be a family of reals that defines the numerical scheme* (3.1) *and* (3.2). *Then, there exists* C *such that*

$$\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{i\in\mathcal{V}_T}\varphi_{i\tau(i)}^{n+1}(u_i^{n+1} - u_{\tau(i)}^{n+1})^2 \leq C. \tag{4.1}$$

$$\sum_{n=0}^{N-1}\delta t\left\|\nabla\xi(u_h^{n+1})\right\|_{L^2(\Omega)^d}^2 \leq C. \tag{4.2}$$

$$\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}}\ \sum_{T\cap K\neq\emptyset}\ \sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}\left(f(u_K^{n+1}) - f(u_L^{n+1})\right)^2\left(-\boldsymbol{V}_{KL,T}^{n+1}\right)^+ \leq C. \tag{4.3}$$

Proof. For every $K \in \mathcal{M}$, we multiply the line corresponding to K in the system (3.2) by $\delta t u_K^{n+1}$. We sum on all $K \in \mathcal{M}$ and $n \in [\![0, \mathbb{N} - 1]\!]$:

$$\mathcal{A}_1 + \mathcal{A}_2 + \mathcal{A}_3 = 0, \tag{4.4}$$

where each term is given by

$$\mathcal{A}_1 = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{M}}|K|\,(u_K^{n+1} - u_K^n)u_K^{n+1},$$

$$\mathcal{A}_2 = \sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}}\ \sum_{T\cap K\neq\emptyset}\ \sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}\sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}_{KL,T}^{n+1}(u)u_K^{n+1},$$

$$\mathcal{A}_3=\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}}\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}} f(u_{KL}^{up,n+1})\mathbf{V}_{KL,T}^{n+1}u_K^{n+1}.$$

Thanks to the elementary inequality $\dfrac{a^2-b^2}{2}\leq(a-b)a$ we directly show that

$$\mathcal{A}_1\geq\sum_{K\in\mathcal{M}}|K|\Big((u_K^N)^2-(u_K^0)^2\Big).\tag{4.5}$$

Being rearranged by dual edges, the diffusion term becomes

$$\mathcal{A}_2=\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}_{KL,T}^{n+1}(u)(u_K^{n+1}-u_L^{n+1}).$$

The crucial choice $u_{KL}^{\mathrm{diff},n+1}$ given in (3.6) entails

$$\sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}_{KL,T}^{n+1}(u)(u_K^{n+1}-u_L^{n+1})\geq\sqrt{\frac{\varphi(u_K^{n+1})+\varphi(u_L^{n+1})}{2}}\mathcal{F}_{KL,T}^{n+1}(u)(u_K^{n+1}-u_L^{n+1}),$$

regardless the sign of $\mathcal{F}_{KL,T}^{n+1}(u)(u_K^{n+1}-u_L^{n+1})$. Indeed, we should distinguish two cases:

(i) If $\mathcal{F}_{KL,T}^{n+1}(u)(u_K^{n+1}-u_L^{n+1})\geq 0$, then the definition of $u_{KL}^{\mathrm{diff},n+1}$ ensures that $u_{KL}^{\mathrm{diff},n+1}=\max(u_K^{n+1},u_L^{n+1})$. Now, using the fact that $\sqrt{\varphi}$ is a nondecreasing function, we automatically obtain

$$\sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\geq\sqrt{\frac{\varphi(u_K^{n+1})+\varphi(u_L^{n+1})}{2}}.$$

(ii) Otherwise, if $\mathcal{F}_{KL,T}^{n+1}(u)(u_K^{n+1}-u_L^{n+1})<0$, the relationship (3.6) amounts to writing $u_{KL}^{\mathrm{diff},n+1}=\min(u_K^{n+1},u_L^{n+1})$ and therefore one gets

$$\sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\leq\sqrt{\frac{\varphi(u_K^{n+1})+\varphi(u_L^{n+1})}{2}},$$

where we also used the monotonicity of $\sqrt{\varphi}$.

We recall the crucial relationship

$$\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\sqrt{\frac{\varphi(u_K^{n+1})+\varphi(u_L^{n+1})}{2}}\mathcal{F}_{KL,T}^{n+1}(u)(u_K^{n+1}-u_L^{n+1})=\delta_T^{n+1}u\cdot\mathbb{A}^T\delta_T^{n+1}u.$$

where the local matrix $\mathbb{A}^T$ is expressed by (3.8). Then, Lemma 2.3 implies

$$\begin{aligned}\mathcal{A}_2&\geq\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\mathbb{D}^T(u^{n+1})\delta_T^{n+1}u\cdot\mathbb{M}^T\mathbb{D}^T(u^{n+1})\delta_T^{n+1}u\\&\geq C_1\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\mathbb{D}^T(u^{n+1})\delta_T^{n+1}u\cdot\mathbb{D}^T(u^{n+1})\delta_T^{n+1}u\\&=C_1\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{i\in\mathcal{V}_T}\varphi_{i\tau(i)}^{n+1}(u_i^{n+1}-u_{\tau(i)}^{n+1})^2.\end{aligned}$$

Utilizing the fact that φ is a nondecreasing function and once more Lemma 2.3 entails

$$\begin{aligned}\sum_{i\in\mathcal{V}_T}\varphi_{i\tau(i)}^{n+1}(u_i^{n+1}-u_{\tau(i)}^{n+1})^2 &\geq \frac{1}{2}\sum_{i\in\mathcal{V}_T}\Big(\max\Big\{\varphi(u_i^{n+1}),\varphi(u_{\tau(i)}^{n+1})\Big\}\Big)\Big(u_i^{n+1}-u_{\tau(i)}^{n+1}\Big)^2\\ &\geq \frac{1}{2}\delta_T^{n+1}\xi(u)\cdot\delta_T^{n+1}\xi(u)\\ &\geq \frac{1}{2C_2}\delta_T^{n+1}\xi(u)\cdot\mathbb{M}^T\delta_T^{n+1}\xi(u).\end{aligned}$$

Hence, we make use of Lemma 2.2 to deduce

$$\frac{C_1}{2C_2}\sum_{n=0}^{N-1}\delta t\left\|\nabla\xi(u_h^{n+1})\right\|_{L^2(\Omega)^d}^2 \leq \sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{i\in\mathcal{V}_T}\varphi_{i\tau(i)}^{n+1}(u_i^{n+1}-u_{\tau(i)}^{n+1})^2 \leq \mathcal{A}_2. \tag{4.6}$$

Following [3,23], we intend to establish a sort of a weak BV estimate on the hyperbolic term. By the equality (3.9) we rewrite $\mathcal{A}_3$ such that

$$\mathcal{A}_3=\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}}\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}\Big(f(u_K^{n+1})-f(u_L^{n+1})\Big)\Big(-\mathbf{V}_{KL,T}^{n+1}\Big)^+u_K^{n+1}.$$

First, we define $G(z)=\displaystyle\int_0^s zf'(z)\,\mathrm{d}z$. In the same spirit of [3] we find

$$\Big(f(u_K^{n+1})-f(u_L^{n+1})\Big)^2\leq 2\Big(G(u_L^{n+1})-G(u_K^{n+1})+(f(u_K^{n+1})-f(u_L^{n+1}))u_K^{n+1}\Big)\left\|f'\right\|_\infty.$$

Next, by virtue of Assumption $(\mathbf{A}_6)$ and the point (iii) of Remark 3.1 we get

$$\sum_{K\in\mathcal{M}}\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}\Big(G(u_L^{n+1})-G(u_K^{n+1})\Big)\Big(-\mathbf{V}_{KL,T}^{n+1}\Big)^+=0.$$

Thus, one concludes the estimation on $\mathcal{A}_3$ by writing

$$\frac{1}{2\left\|f'\right\|_\infty}\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}}\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}\Big(f(u_K^{n+1})-f(u_L^{n+1})\Big)^2\Big(-\mathbf{V}_{KL,T}^{n+1}\Big)^+\leq\mathcal{A}_3. \tag{4.7}$$

Finally, we gather (4.4)–(4.7) for the conclusion. □

Following we show that our numerical scheme is well-defined.

Lemma 4.3. *For every* $[\![0,\mathrm{N}-1]\!]$, *the algebraic system resulting from the numerical scheme* (3.1) *and* (3.2) *has at least one solution* $(u_K^{n+1})_{K\in\mathcal{M}}$.

Proof. We use the induction argument to prove the existence result. We denote $U^{n+1}=(u_K^{n+1})_{K\in\mathcal{M}}$. Let us consider $\mathbb{R}^{\#\mathcal{M}}$ equipped with the usual inner product. We next define from $\mathbb{R}^{\#\mathcal{M}}$ into itself the continuous operator Υ whose components are

$$\begin{aligned}\Upsilon(U^{n+1})_{|K}&=\frac{|K|}{\delta t}(u_K^{n+1}-u_K^n)+\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}\sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}_{KL,T}^{n+1}(u)\\ &\quad+\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}f(u_{KL}^{up,n+1})\mathbf{V}_{KL,T}^{n+1},\quad\forall K\in\mathcal{M}_D^c,\\ \Upsilon(U^{n+1})_{|K}&=0,\quad\forall K\in\mathcal{M}_D.\end{aligned}$$

With this notation, the numerical scheme (3.1)-(3.2) takes the vector form:

$$\text{find} \quad U^{n+1} \in \mathbb{R}^{\#\mathcal{M}} : \quad \Upsilon(U^{n+1}) = 0.$$

It follows from the proof of Lemma 4.2 that

$$\Upsilon(U^{n+1}) \cdot U^{n+1} \geq C_{h,\delta t}\left|U^{n+1}\right|^2 - C'_{h,\delta t}|U^n|^2.$$

Note that the constants $C_{h,\delta t}, C'_{h,\delta t} > 0$ are depending this time on the discretization parameters. They are obtained by the fact that all norms are equivalent on $\mathbb{R}^{\#\mathcal{M}}$. For a sufficiently large $\left|U^{n+1}\right|$ we get

$$\Upsilon(U^{n+1}) \cdot U^{n+1} > 0.$$

Applying a fixed point criterion [36] ensures the existence of $U^{n+1} \in \mathbb{R}^{\#\mathcal{M}}$ to the nonlinear system

$$\Upsilon(U^{n+1}) = 0.$$

Hence, the numerical scheme admits a solution as required. □

5. Convergence

This section is devoted to the convergence proof of the finite volume scheme. To this end, we first specify without proof compactness estimates on the space translates of the discrete functions.

Lemma 5.1. *There exists a constant C depending only on the data and on the regularity of the mesh Θ such that*

$$\int_{\Omega_y \times (0,\mathfrak{T})} |\tilde{v}_{h,\delta t}(x+y,t) - \tilde{v}_{h,\delta t}(x,t)| \,\mathrm{d}x\,\mathrm{d}t \leq C\,|y| \left(\sum_{n=0}^{N-1} \delta t \left\| \nabla v_h^{n+1} \right\|^2_{L^2(\Omega)^d} \right)^{\frac{1}{2}}, \tag{5.1}$$

$\forall y \in \mathbb{R}^d$, $\forall \tilde{v}_{h,\delta t} \in W^0_{h,\delta t}$, where $\Omega_y = \{x \in \Omega : \ [x,y] \subset \Omega\}$.

Proof. The proof mimics classical ideas as done for instance in [23,30]. □

In order to make use of the compactness criterion given in [5] we require the following result.

Lemma 5.2. *Let us select $\psi \in \mathcal{C}_c^\infty(\overline{\Omega} \times [0,\mathfrak{T}))$ with $\psi = 0$ on $\Gamma_1 \times [0,\mathfrak{T})$. We set $\psi_K^{n+1} = \psi(x_K, t^{n+1}), \forall K \in \mathcal{M}, n \in [\![0, N-1]\!]$. Then, there exists a constant depending only on the physical data, the mesh regularity Θ and ε such that*

$$\sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}_h} |K| \,(u_K^{n+1} - u_K^n)\psi_K^{n+1} \leq C \left\|\nabla\psi\right\|_\infty. \tag{5.2}$$

Proof. Multiplying (3.2) by $\delta t \psi_K^{n+1}$ summing over both $K \in \mathcal{M}$ and $n \in [\![0, N-1]\!]$ together with the integration by parts procedure yields

$$D_1 = D_2 + D_3, \tag{5.3}$$

where

$$D_1 = \sum_{n=0}^{N-1} \sum_{K \in \mathcal{M}} |K| \,(u_K^{n+1} - u_K^n)\psi_K^{n+1},$$

$$D_2 = -\sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1}) \mathcal{F}_{KL,T}^{n+1}(u)(\psi_K^{n+1} - \psi_L^{n+1}),$$

$$D_3 = -\sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} f(u_{KL}^{up,n+1}) \mathbf{V}_{KL,T}^{n+1}(\psi_K^{n+1} - \psi_L^{n+1}).$$

It follows from the smoothness of the test function and the regularity of the mesh that

$$\left|\psi_K^{n+1} - \psi_L^{n+1}\right| \le C\sqrt{|T_{KL}|}\, \|\nabla\psi\|_\infty ,$$

for some $C > 0$, where T_{KL} is the sub-triangle made by K, L and the barycenter of T. In addition, we have

$$\left|\mathcal{F}_{KL,T}^{n+1}(u)\right| \le C \sum_{i\in\mathcal{V}_T} \varphi_{i\tau(i)}^{n+1}(u_i^{n+1} - u_{\tau(i)}^{n+1})^2.$$

According to these materials, the Cauchy–Schwarz inequality, the estimate (4.1) and the inequality (1.7) we find

$$\begin{aligned}
|D_2| &\le C\,\|\nabla\psi\|_\infty \left(\sum_{n=0}^{N-1}\delta t \sum_{T\in\mathcal{T}}\sum_{i\in\mathcal{V}_T} \varphi_{i\tau(i)}^{n+1}(u_i^{n+1} - u_{\tau(i)}^{n+1})^2\right)^{\frac{1}{2}} \\
&\quad\times \left(\sum_{n=0}^{N-1}\delta t \sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T} |T_{KL}|\,\varphi(u_{KL}^{\mathrm{diff},n+1})\right)^{\frac{1}{2}} \\
&\le C_\varepsilon\,\|\nabla\psi\|_\infty \left(\sum_{n=0}^{N-1}\delta t \sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T} |T_{KL}|\left(\xi(u_{KL}^{\mathrm{diff},n+1})\right)^2\right)^{\frac{1}{2}}.
\end{aligned}$$

By (4.2) and Poincaré's inequality we prove the existence of $C' > 0$ so that we estimate

$$\|\xi(u_{h,\delta t})\|_{L^2(Q_{\mathfrak{T}})} \le C'.$$

As a consequence of Lemma 2.4 and again (4.2) we deduce

$$|D_2| \le C'_\varepsilon\,\|\nabla\psi\|_\infty \left(\sum_{n=0}^{N-1}\delta t \left\|\xi(\overline{u_{h,\delta t}}) - \xi(\underline{u_{h,\delta t}})\right\|_{L^2(Q_{\mathfrak{T}})}^2 + \|\xi(u_{h,\delta t})\|_{L^2(Q_{\mathfrak{T}})}^2\right) \le C''_\varepsilon\,\|\nabla\psi\|_\infty .$$

Similar guidelines can be adapted to establish $|D_3| \le C\,\|\nabla\psi\|_\infty$. The cornerstone element of the latter proof relies heavily on Assumption $(\mathbf{A}_6)$ and particularly on inequality (1.8). □

We now state and demonstrate the main result of this paper.

Proposition 5.3. *Assume that hypotheses $(\boldsymbol{A}_1)$–$(\boldsymbol{A}_6)$ hold. Let $(\mathcal{T}_h)_h$ be a family of refined meshes to Ω such that the geometric condition (2.1) is satisfied. Let $(u_{h,\delta t})$ be a sequence of discrete solutions to the finite volume scheme (3.1) and (3.2). When $h, \delta t$ tend to 0, one gets up to a subsequence:*

$$\tilde{u}_{h,\delta t}, u_{h,\delta t} \longrightarrow u \qquad a.e. \;\text{in}\; Q_{\mathfrak{T}} \text{ and strongly in } L^1(Q_{\mathfrak{T}}), \tag{5.4}$$

$$\varphi(\tilde{u}_{h,\delta t}), \varphi(u_{h,\delta t}) \longrightarrow \varphi(u) \qquad a.e. \;\text{in}\; Q_{\mathfrak{T}} \text{ and strongly in } L^1(Q_{\mathfrak{T}}), \tag{5.5}$$

$$\nabla\xi(u_{h,\delta t}) \longrightarrow \nabla\xi(u) \qquad \text{weakly in } L^2(Q_{\mathfrak{T}})^d. \tag{5.6}$$

Moreover, one has $u \ge 0$ a.e. in $Q_{\mathfrak{T}}$ and $\xi(u) \in L^2(0,\mathfrak{T}; H^1_{\Gamma_1}(\Omega))$. Finally, the limit function u is then a weak solution to the continuous equations (1.1)–(1.3) as specified in Definition 1.3.

Proof. First, Poincaré's inequality ensures a uniform $L^2(Q_{\mathfrak{T}})$ bound on $\xi(u_{h,\delta t})$ *i.e.* there exists a positive constant $C > 0$ such that

$$\|\xi(u_{h,\delta t})\|_{L^2(Q_{\mathfrak{T}})} \le C. \tag{5.7}$$

Owing once more to Lemma 2.4 and the inequality (4.2), we check that

$$\|\xi(\tilde{u}_{h,\delta t})\|_{L^2(Q_{\mathfrak{T}})} \leq C, \tag{5.8}$$

for some $C > 0$. Now, by the estimations of Lemma 5.1 and making use of the compactness criterion ([5], Thm. 3.9) we get

$$\tilde{u}_{h,\delta t} \longrightarrow u \quad a.e. \quad \text{in} \quad Q_{\mathfrak{T}}. \tag{5.9}$$

Applying Lemma 2.4 gives

$$\xi(\tilde{u}_{h,\delta t}) - \xi(u_{h,\delta t}) \longrightarrow 0 \quad a.e. \quad \text{in} \quad Q_{\mathfrak{T}}, \tag{5.10}$$

up to the extraction of another subsequence. Combining (5.9) and the continuity of ξ^{-1} we claim

$$u_{h,\delta t} \longrightarrow u \quad a.e. \quad \text{in} \quad Q_{\mathfrak{T}}. \tag{5.11}$$

By virtue of Lemma 4.1 we check that $u \geq 0 \quad a.e. \quad \text{in} \quad Q_{\mathfrak{T}}$. Thanks to the latter and (5.7), (5.8) we get the equi-integrability of $\xi(u_{h,\delta t})$ and $\xi(\tilde{u}_{h,\delta t})$ in $L^{2-\delta}(Q_{\mathfrak{T}})$ for every small $\delta > 0$ thanks to De La Vallée Poussin equi-integrability criterion ([7], Thm. 4.5.9). Now, we employ Vitali's convergence argument to obtain

$$\xi(\tilde{u}_{h,\delta t}),\ \xi(u_{h,\delta t}) \longrightarrow \xi(u) \text{ strongly in } L^{2-\delta}(Q_{\mathfrak{T}}), \tag{5.12}$$

for all small $\delta > 0$. According to (1.6) we deduce the $L^1(Q_{\mathfrak{T}})$-equi-integrability of the sequences $u_{h,\delta t}, \tilde{u}_{h,\delta t}$. As a consequence of Vitali's convergence theorem we affirm that

$$\tilde{u}_{h,\delta t}, u_{h,\delta t} \longrightarrow u \quad \text{strongly in } L^1(Q_{\mathfrak{T}}), \tag{5.13}$$

completing the proof of (5.4). Next, the continuity of φ and the *a.e.* convergence (5.4) lead to

$$\sqrt{\varphi}(\tilde{u}_{h,\delta t}),\ \sqrt{\varphi}(u_{h,\delta t}) \longrightarrow \sqrt{\varphi}(u) \quad a.e. \quad \text{in} \quad Q_{\mathfrak{T}}. \tag{5.14}$$

We fix $0 < \delta < \varepsilon$ where ε is mentioned in (1.5). Recall that $\xi(u_{h,\delta t})$ and $\xi(\tilde{u}_{h,\delta t})$ are equi-integrable in $L^{2-\delta}(Q_{\mathfrak{T}})$. By virtue of the growth assumption (1.5) we check once again that

$$\|\sqrt{\varphi}(u_{h,\delta t})\|_{L^{2+\alpha}(Q_{\mathfrak{T}})} + \|\sqrt{\varphi}(\tilde{u}_{h,\delta t})\|_{L^{2+\alpha}(Q_{\mathfrak{T}})} \leq C, \tag{5.15}$$

for every $0 < \alpha < 2(\varepsilon - \delta)/(2 - \varepsilon)$. Hence, the equi-integrability of the both sequences $\sqrt{\varphi}(u_{h,\delta t})$ and $\sqrt{\varphi}(\tilde{u}_{h,\delta t})$ holds in $L^2(Q_{\mathfrak{T}})$ by another application of De La Vallée Poussin equi-integrability criterion. As a result, Vitali's convergence theorem guarantees that the limit (5.14) is enhanced and it holds now strongly in $L^2(Q_{\mathfrak{T}})$ which concludes the proof of (5.5). On the other hand, Lemma 4.2 guarantees the existence of $G \in L^2(Q_{\mathfrak{T}})^d$ such that

$$\nabla\xi(u_{h,\delta t}) \longrightarrow G \quad \text{weakly in } L^2(Q_{\mathfrak{T}})^d.$$

We finally apply (5.12) and the identification of the limit process to get $G = \nabla\xi(u)$. This proves by the way that $\xi(u) \in L^2(0, \mathfrak{T}; H^1_{\Gamma_1}(\Omega))$.

It remains to establish that u is a weak solution to (1.1)–(1.3) in the sense of Definition 1.3. To this purpose let us pick a test function $\psi \in \mathcal{C}^\infty_c(\overline{\Omega} \times [0, \mathfrak{T}))$, such that $\psi = 0$ on $\Gamma_1 \times [0, \mathfrak{T})$. We denote $\psi^{n+1}_K = \psi(x_K, t^{n+1}), \forall K \in \mathcal{M}, n \in [\![0, N-1]\!]$. We multiply (3.2) by $\delta t \psi^{n+1}_K$ and sum on both $K \in \mathcal{M}$ and $n \in [\![0, N-1]\!]$. This gives

$$\mathcal{G}^1_{h,\delta t} + \mathcal{G}^2_{h,\delta t} + \mathcal{G}^3_{h,\delta t} = 0, \tag{5.16}$$

where

$$\mathcal{G}^1_{h,\delta t} = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{M}_h} |K|\,(u_K^{n+1} - u_K^n)\psi_K^{n+1},$$

$$\mathcal{G}^2_{h,\delta t} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}_h} \sum_{T\cap K\neq\emptyset} \sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}} \sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}^{n+1}_{KL,T}(u)\psi_K^{n+1},$$

$$\mathcal{G}^3_{h,\delta t} = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{M}_h} \sum_{T\cap K\neq\emptyset} \sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}} f(u_{KL}^{up,n+1})\mathbf{V}^{n+1}_{KL,T}\psi_K^{n+1}.$$

The discrete integration by parts in time with the fact that $\psi_K^N = 0$ yields

$$\begin{aligned}\mathcal{G}^1_{h,\delta t} &= -\sum_{n=0}^{N-1} \sum_{K\in\mathcal{M}_h} |K|\,u_K^n(\psi_K^{n+1} - \psi_K^n) - \sum_{K\in\mathcal{M}_h} |K|\,u_K^0\psi_K^0 \\ &= -\sum_{n=0}^{N-1} \int_{t^n}^{t^{n+1}} \int_K \widetilde{u}_h^n \partial_t\psi(x_K,t)\,\mathrm{d}x\,\mathrm{d}t - \int_{Q_{\mathfrak{T}}} u^0\widetilde{\psi}_h(x,0)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

In light of the regularity of ψ and the strong convergence of $\tilde{u}_{h,\delta t}$ towards u we infer

$$\mathcal{G}^1_{h,\delta t} \longrightarrow -\int_{Q_{\mathfrak{T}}} u\partial_t\psi\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega u^0\psi(x,0)\,\mathrm{d}x.$$

Next, we focus on the convergence of the diffusive term. We first reorder $\mathcal{G}^2_{h,\delta t}$ by dual edges. Then

$$\mathcal{G}^2_{h,\delta t} = \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \sum_{\sigma^T_{KL}\in\mathcal{E}_T} \sqrt{\varphi}(u_{KL}^{\mathrm{diff},n+1})\mathcal{F}^{n+1}_{KL,T}(u)(\psi_K^{n+1} - \psi_L^{n+1}).$$

To study the convergence of $\mathcal{G}^2_{h,\delta t}$, we need to define two discrete functions $\overline{u_{h,\delta t}}$ and $\underline{u_{h,\delta t}}$ so that their restrictions respectively to $T\times(t^n,t^{n+1}]$ are set to

$$\overline{u}_T^{n+1} = \max_{i\in\mathcal{V}_T} u_i^{n+1}, \quad \underline{u}_T^{n+1} = \min_{i\in\mathcal{V}_T} u_i^{n+1}. \tag{5.17}$$

We observe that

$$\begin{aligned}\|\xi(\overline{u_{h,\delta t}}) - \xi(u)\|_{L^{2-\delta}(Q_{\mathfrak{T}})} &\le C\Big(\|\xi(\tilde{u}_{h,\delta t}) - \xi(u)\|_{L^{2-\delta}(Q_{\mathfrak{T}})} + \|\xi(\overline{u_{h,\delta t}}) - \xi(\tilde{u}_{h,\delta t})\|_{L^2(Q_{\mathfrak{T}})}\Big) \\ &\le C'\Big(\|\xi(\tilde{u}_{h,\delta t}) - \xi(u)\|_{L^{2-\delta}(Q_{\mathfrak{T}})} + \Big\|\xi(\overline{u_{h,\delta t}}) - \xi(\underline{u_{h,\delta t}})\Big\|_{L^2(Q_{\mathfrak{T}})}\Big).\end{aligned}$$

So, Lemma 2.4, the energy estimate (4.2) and the strong limit (5.12) allow us to establish that

$$\begin{aligned}\Big\|\xi(\overline{u_{h,\delta t}}) - \xi(\underline{u_{h,\delta t}})\Big\|^2_{L^2(Q_{\mathfrak{T}})} &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} |T|\,\big|\xi(\overline{u}_T^{n+1}) - \xi(\underline{u}_T^{n+1})\big|^2 \\ &\le Ch_{\mathcal{T}_h} \sum_{n=0}^{N-1} \delta t\,\big\|\nabla\xi(u_h^{n+1})\big\|^2_{L^2(\Omega)^d} \le Ch_{\mathcal{T}_h} \longrightarrow 0.\end{aligned}$$

Consequently, $\xi(\overline{u_{h,\delta t}}),\ \xi(\underline{u_{h,\delta t}}) \longrightarrow \xi(u)$ strongly in $L^{2-\delta}(Q_{\mathfrak{T}})$. Then, using the continuity of ξ^{-1}, there exists a new subsequence such that

$$\overline{u_{h,\delta t}},\ \underline{u_{h,\delta t}} \longrightarrow u \quad a.e. \quad \text{in} \quad Q_{\mathfrak{T}}. \tag{5.18}$$

We mimic an analogous proof as we have conducted for (5.5) to justify

$$\sqrt{\varphi}(\overline{u_{h,\delta t}}), \sqrt{\varphi}(\underline{u_{h,\delta t}}) \longrightarrow \sqrt{\varphi}(u) \quad \text{strongly in } L^2(Q_{\mathfrak{T}}). \tag{5.19}$$

Now, we decompose $\mathcal{G}^2_{h,\delta t}$ into three parts as follows:

$$\mathcal{G}^2_{h,\delta t} = \mathcal{G}^{2,1}_{h,\delta t} + \mathcal{G}^{2,2}_{h,\delta t} + \mathcal{G}^{2,3}_{h,\delta t},$$

such that

$$\begin{aligned}
\mathcal{G}^{2,1}_{h,\delta t} &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \sqrt{\varphi}(\underline{u}_T^{n+1}) \sum_{\sigma^T_{KL}\in\mathcal{E}_T} \sum_{i\in\mathcal{V}_T} a^T_{i,KL}\Big(\xi(u_i^{n+1}) - \xi(u^{n+1}_{\tau(i)})\Big)(\psi_K^{n+1} - \psi_L^{n+1}),\\
\mathcal{G}^{2,2}_{h,\delta t} &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \sqrt{\varphi}(\underline{u}_T^{n+1}) \sum_{\sigma^T_{KL}\in\mathcal{E}_T} \sum_{i\in\mathcal{V}_T} a^T_{i,KL}\left(\sqrt{\varphi^{n+1}_{i\tau(i)}}(u_i^{n+1} - u^{n+1}_{\tau(i)}) - \Big(\xi(u_i^{n+1}) - \xi(u^{n+1}_{\tau(i)})\Big)\right)\\
&\quad \times (\psi_K^{n+1} - \psi_L^{n+1}),\\
\mathcal{G}^{2,3}_{h,\delta t} &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \sum_{\sigma^T_{KL}\in\mathcal{E}_T} \Big(\sqrt{\varphi}(u^{\mathrm{diff},n+1}_{KL}) - \sqrt{\varphi}(\underline{u}_T^{n+1})\Big)\mathcal{F}^{n+1}_{KL,T}(u)(\psi_K^{n+1} - \psi_L^{n+1}).
\end{aligned}$$

Thanks to Lemma 2.6 we can express $\mathcal{G}^{2,1}_{h,\delta t}$ in the following integral form

$$\begin{aligned}
\mathcal{G}^{2,1}_{h,\delta t} &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \sqrt{\varphi}(\underline{u_T}^{n+1})\delta_T^{n+1}\xi(u)\cdot \mathbb{M}^T\delta_T^{n+1}\psi\\
&= \int_{Q_{\mathfrak{T}}} \sqrt{\varphi}(\underline{u_{h,\delta t}})\Lambda\nabla\xi(u_{h,\delta t})\cdot\nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t.
\end{aligned}$$

Thereby, using the weak convergence of $\nabla\xi(u_{h,\delta t})$ and the strong convergence of $\nabla\psi_{h,\delta t}$ together with (5.19), we can pass to the limit in $\mathcal{G}^{2,1}_{h,\delta t}$:

$$\mathcal{G}^{2,1}_{h,\delta t} \longrightarrow \int_{Q_{\mathfrak{T}}} \sqrt{\varphi}(u)\Lambda\nabla\xi(u)\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t.$$

Let us now move on to study the limit of $\mathcal{G}^{2,1}_{h,\delta t}$. The function φ being nondecreasing, we notice that

$$\sqrt{\varphi}(\underline{u}_T^{n+1}) \le \sqrt{\frac{\varphi(u_i^{n+1}) + \varphi(u^{n+1}_{\tau(i)})}{2}} = \sqrt{\varphi^{n+1}_{i\tau(i)}}, \quad \forall i\in\mathcal{V}_T. \tag{5.20}$$

We define $u^{*,n+1}_{i\tau(i)} \in [\min(u_i^{n+1}, u^{n+1}_{\tau(i)}), \max(u_i^{n+1}, u^{n+1}_{\tau(i)})]$ such that

$$\forall i\in\mathcal{V}_T, \quad \sqrt{\varphi}(u^{*,n+1}_{i\tau(i)}) = \begin{cases} \dfrac{\xi(u_i^{n+1}) - \xi(u^{n+1}_{\tau(i)})}{u_i^{n+1} - u^{n+1}_{\tau(i)}} & \text{if} \quad u_i^{n+1} \neq u^{n+1}_{\tau(i)}\\ \sqrt{\varphi}(u_i^{n+1}) & \text{if} \quad u_i^{n+1} = u^{n+1}_{\tau(i)} \end{cases}. \tag{5.21}$$

Observe that such a real exists *via* the mean value theorem. Gathering (5.20) and (5.21) leads to

$$
\begin{aligned}
\left|\mathcal{G}_{h,\delta t}^{2,2}\right| \leq{} & \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \sqrt{\varphi}(\underline{u}_T^{n+1}) \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \left(\sum_{i\in\mathcal{V}_T} \left|a_{i,KL}^T\right| \left(\sqrt{\varphi_{i\tau(i)}^{n+1}} - \sqrt{\varphi}(u_{i\tau(i)}^{*,n+1})\right) (u_i^{n+1} - u_{\tau(i)}^{n+1}) \right) \\
& \times (\psi_K^{n+1} - \psi_L^{n+1}) \\
\leq{} & 2 \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \left(\sum_{i\in\mathcal{V}_T} \left|a_{i,KL}^T\right| \sqrt{\varphi_{i\tau(i)}^{n+1}} \left(\sqrt{\varphi_{i\tau(i)}^{n+1}} - \sqrt{\varphi}(u_{i\tau(i)}^{*,n+1})\right) (u_i^{n+1} - u_{\tau(i)}^{n+1}) \right) \\
& \times (\psi_K^{n+1} - \psi_L^{n+1}) \\
\leq{} & 2 \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \left|\sqrt{\varphi}(\overline{u}_T^{n+1}) - \sqrt{\varphi}(\underline{u}_T^{n+1})\right| \\
& \times \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \sum_{i\in\mathcal{V}_T} \left|a_{i,KL}^T\right| \sqrt{\varphi_{i\tau(i)}^{n+1}} \left(u_i^{n+1} - u_{\tau(i)}^{n+1}\right) \left(\psi_K^{n+1} - \psi_L^{n+1}\right).
\end{aligned}
$$

We introduce the regularity of the mesh to see that $\left|a_{i,KL}^T\right| \leq C$ regardless σ_{KL}^T and $T \in \mathcal{T}$. The smoothness of ψ implies that $\left|\psi_K^{n+1} - \psi_L^{n+1}\right| \leq \sqrt{|T|} C_\psi$. As a result of the latter and the Cauchy–Schwarz inequality we explore

$$
\begin{aligned}
\left|\mathcal{G}_{h,\delta t}^{2,2}\right| \leq C_\psi & \left(\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} |T| \left|\sqrt{\varphi}(\overline{u}_T^{n+1}) - \sqrt{\varphi}(\underline{u}_T^{n+1})\right|^2 \right)^{\frac{1}{2}} \\
& \times \underbrace{\left(\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}_h} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \sum_{i\in\mathcal{V}_T} \varphi_{i\tau(i)}^{n+1} \left(u_i^{n+1} - u_{\tau(i)}^{n+1}\right)^2 \right)^{\frac{1}{2}}}_{\mathfrak{X}_{h,\delta t}}.
\end{aligned}
$$

Owing to energy estimate (4.1) we have $\mathfrak{X}_{h,\delta t} \leq C$. As a consequence of the strong convergence (5.19) we conclude that

$$
\mathcal{G}_{h,\delta t}^{2,2} \longrightarrow 0.
$$

One can draw the same conclusion for the term $\mathcal{G}_{h,\delta t}^{2,3}$.

Finally, it is left to identify the limit of the hyperbolic term. To this purpose, we rewrite its otherwise keeping in mind that $\sum\limits_{T\cap K\neq\emptyset} \sum\limits_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}} \mathbf{V}_{KL,T}^{n+1} = 0$. First, we set

$$
\psi_{\sigma_{KL}^T}^{n+1} = \frac{1}{\delta t \left|\sigma_{KL}^T\right|} \int_{t^n}^{t^{n+1}} \int_{\sigma_{KL}^T} \psi \,\mathrm{d}\sigma \,\mathrm{d}t, \qquad \text{and} \qquad \widehat{\mathbf{V}_{\sigma_{KL}^T}^{n+1}} = \frac{1}{\left|\sigma_{KL}^T\right|} \int_{\sigma_{KL}^T} \mathbf{V}^{n+1} \,\mathrm{d}\sigma.
$$

Then

$$\mathcal{G}^3_{h,\delta t} = \underbrace{-\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h}\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}} f(u^{n+1}_K)\psi^{n+1}_{\sigma^T_{KL}}\mathbf{V}^{n+1}_{KL,T}}_{\mathcal{G}^{3,1}_{h,\delta t}}$$
$$+\underbrace{\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h}\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\Big(f(u^{up,n+1}_{KL})-f(u^{n+1}_K)\Big)\mathbf{V}^{n+1}_{KL,T}\Big(\psi^{n+1}_K-\psi^{n+1}_{\sigma^T_{KL}}\Big)}_{\mathcal{G}^{3,2}_{h,\delta t}}.$$

Let us first prove that $\mathcal{G}^{3,2}_{h,\delta t}\longrightarrow 0$. It follows from the definition of $u^{up,n+1}_{KL}$ stated in (3.3) that

$$\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\Big|f(u^{up,n+1}_{KL})-f(u^{n+1}_K)\Big|\Big|\mathbf{V}^{n+1}_{KL,T}\Big| = \sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\big|f(u^{n+1}_L)-f(u^{n+1}_K)\big|\Big(-\mathbf{V}^{n+1}_{KL,T}\Big)^+.$$

This identity together with the Cauchy–Schwarz inequality gives

$$\Big|\mathcal{G}^{3,2}_{h,\delta t}\Big| \le \left(\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h}\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\big|f(u^{n+1}_K)-f(u^{n+1}_L)\big|^2\Big(-\mathbf{V}^{n+1}_{KL,T}\Big)^+\right)^{\frac{1}{2}}$$
$$\times\left(\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h}\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\|\mathbf{V}\|_\infty\,\big|\sigma^T_{KL}\big|\Big(\psi^{n+1}_K-\psi^{n+1}_{\sigma^T_{KL}}\Big)^2\right)^{\frac{1}{2}}.$$

Thanks to the estimation (4.3), the smoothness of the test function ψ and the regularity of the mesh (2.1) we get

$$\Big|\mathcal{G}^{3,2}_{h,\delta t}\Big| \le C_\psi h^{\frac{1}{2}} \longrightarrow 0.$$

On the other hand, we split $\mathcal{G}^{3,1}_{h,\delta t}$ in its turn into two parts

$$\mathcal{G}^{3,1}_{h,\delta t} = \mathfrak{R}^1_{h,\delta t}+\mathfrak{R}^2_{h,\delta t}$$

where each one writes

$$\mathfrak{R}^1_{h,\delta t} = -\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h} f(u^{n+1}_K)\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\int_{\sigma^T_{KL}}\psi\mathbf{V}^{n+1}\cdot\mathbf{n}^T_{KL}\,\mathrm{d}\sigma,$$
$$\mathfrak{R}^2_{h,\delta t} = -\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h} f(u^{n+1}_K)\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\left(\int_{\sigma^T_{KL}}(\psi^{n+1}_{\sigma^T_{KL}}-\psi)\mathbf{V}^{n+1}\cdot\mathbf{n}^T_{KL}\,\mathrm{d}\sigma\right).$$

We utilize Green's formula and the condition div $\mathbf{V}=0$ to obtain

$$\mathfrak{R}^1_{h,\delta t} = -\int_{Q_{\mathfrak{T}}} f(\tilde{u}_{h,\delta t})\nabla\psi\cdot\mathbf{V}_{\delta t}\,\mathrm{d}x\,\mathrm{d}t \quad \text{with} \quad \mathbf{V}_{\delta t_{|(t^n,t^{n+1}]}} = \mathbf{V}^{n+1}.$$

By Assumption $(\mathbf{A}_6)$, it can be easily seen that

$$\|f(\tilde{u}_{h,\delta t})-f(u)\|_{L^1(Q_{\mathfrak{T}})} \le \|f'\|_\infty\,\|\tilde{u}_{h,\delta t}-u\|_{L^1(Q_{\mathfrak{T}})} \longrightarrow 0.$$

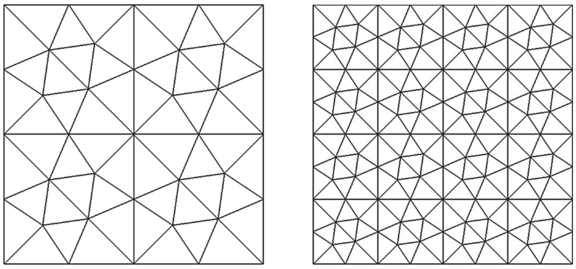


FIGURE 2. First (left) and second (right) triangular meshes used for the tests.

As a consequence of Assumption ($\mathbf{A}_3$)

$$\mathfrak{R}^1_{h,\delta t} \longrightarrow -\int_{Q_{\mathfrak{T}}} f(u)\nabla\psi\cdot\mathbf{V}\,\mathrm{d}x\,\mathrm{d}t.$$

In addition, the residual term $\mathfrak{R}^2_{h,\delta t}$ can be rewritten as

$$\begin{aligned}\mathfrak{R}^2_{h,\delta t} = &-\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h} f(u_K^{n+1})\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\left(\int_{\sigma^T_{KL}}(\psi^{n+1}_{\sigma^T_{KL}}-\psi)(\mathbf{V}^{n+1}-\widehat{\mathbf{V}^{n+1}_{\sigma^T_{KL}}})\cdot\mathbf{n}^T_{KL}\,\mathrm{d}\sigma\right)\\ &\underbrace{-\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h} f(u_K^{n+1})\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\widehat{\mathbf{V}^{n+1}_{\sigma^T_{KL}}}\cdot\mathbf{n}^T_{KL}\left(\int_{\sigma^T_{KL}}(\psi^{n+1}_{\sigma^T_{KL}}-\psi)\,\mathrm{d}\sigma\right)}_{=0}.\end{aligned}$$

We ultimately finish the proof by seeing that

$$\begin{aligned}\left|\mathfrak{R}^2_{h,\delta t}\right| &\leq C_\psi h\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{M}_h} f(u_K^{n+1})\sum_{T\cap K\neq\emptyset}\sum_{\sigma^T_{KL}\in\mathcal{E}_{K\cap T}}\left|\sigma^T_{KL}\right|\left\|\mathbf{V}^{n+1}-\widehat{\mathbf{V}^{n+1}_{\sigma^T_{KL}}}\right\|_\infty\\ &\leq C_\psi\left\|f(\tilde{u}_{h,\delta t})\right\|_{L^1(Q_{\mathfrak{T}})}\sup_{\sigma^T_{KL}\in\mathcal{E}}\left\|\mathbf{V}^{n+1}-\widehat{\mathbf{V}^{n+1}_{\sigma^T_{KL}}}\right\|_\infty \longrightarrow 0.\end{aligned}$$

This last convergence holds since $\|f(\tilde{u}_{h,\delta t})\|_{L^1(Q_{\mathfrak{T}})}$ is bounded and thanks to the continuity assumption on the field $\mathbf{V}$. Hence, the proof is complete. □

6. NUMERICAL RESULTS

In this final section, we perform several numerical experiments in two dimensions in space to test the efficiency of our methodology and its ability to respect the lower bound on the computed solution, especially when the medium of interest is strongly anisotropic.

Here we take the computational domain as the unit square $\Omega = (0,1)^2$. This square is meshed thanks to a refined series of triangular meshes used for benchmarking problems [32]. An illustration of the latter is depicted in Figure 2.

Following, the tensor Λ is chosen to be diagonal so that we can determine analytical solutions

$$\Lambda = \begin{pmatrix}\lambda_x & 0\\ 0 & \lambda_y\end{pmatrix}.$$

Table 1. Test 1: Linear heat equation with scheme (6.1).

$h_{\mathcal{T}}$	$\|\tilde{u}_{h,\delta t} - u\|_2$	Rate	$\|\tilde{u}_{h,\delta t} - u\|_\infty$	Rate	$u_{\min}$
0.250	0.117 E-01	–	0.691 E-01	–	−0.0392
0.125	0.283 E-02	2.044	0.204 E-01	1.756	−0.0123
0.063	0.700 E-03	2.041	0.534 E-02	1.958	−0.0032
0.031	0.175 E-03	1.956	0.136 E-02	1.931	−0.0008
0.016	0.436 E-04	2.102	0.342 E-03	2.084	−0.0002

We have implemented a Newton-Raphson algorithm for the resolution of the nonlinear system (3.1) and (3.2) at each time iteration. Its stopping criterion is fixed to 10^{-10}. To assess the difference between the approximate solution and the exact one we compute the errors

$$\|\tilde{u}_{h,\delta t} - u\|_p = \|\tilde{u}_{h,\delta t} - u\|_{L^p(Q_{\mathfrak{T}})}, \quad p = 2, \infty.$$

6.1. Test 1: Pure diffusion

In this first example we compare the efficiency of the linear control volume finite element scheme and the improvement provided by the nonlinear versions. To this end, we focus on the linear pure diffusion where we neglect the convection effects *i.e.* $\mathbf{V} \equiv 0$. Let us consider the heat equation

$$\partial_t u - \operatorname{div}(\Lambda \nabla u) = 0,$$

which is supplemented with a zero-flux condition on $\partial\Omega \times (0, \mathfrak{T})$. We take the one-dimensional classical solution given by the formula

$$u(x, y, t) = \frac{1}{2}\Big(\cos(\pi x)e^{-\pi^2 \lambda_x t} + 1\Big), \quad \forall (x, y, t) \in \Omega \times (0, \mathfrak{T}),$$

where the final time is set to $\mathfrak{T} = 0.15$ and the tensor is chosen such as $\lambda_x = 1$ and $\lambda_y = 1000$. First, the linear control volume finite element discretization for the heat equation reads

$$\frac{|K|}{\delta t}(u_K^{n+1} - u_K^n) + \sum_{T \cap K \neq \emptyset} \sum_{\sigma_{KL}^T \in \mathcal{E}_{K \cap T}} \Lambda_{KL}^T \Big(u_K^{n+1} - u_L^{n+1}\Big). \tag{6.1}$$

In Table 1 we present the numerical errors, their corresponding convergence rates as well as the minimum value of the discrete solution computed by solving the algebraic system (6.1). From the same table, it is clear that the linear scheme is privileged by the superconvergence feature. However, it gives rise to undesirable undershoots linked to the anisotropy of Λ. To cope with this issue, we propose two nonlinear schemes. Both of them require a nonlinear reformulation of the heat equation as follows

$$\partial_t u - \operatorname{div}(u \Lambda \nabla \log(u)) = 0.$$

The first nonlinear approach is taken from [12]. The key idea consists in treating the diffusion fluxes as if they were of hyperbolic kind. Then, an upstream technique with respect to the transmissibility coefficients is introduced to reinforce the positivity of the solution. So, the approximation proposed in [12] for the underlined problem is:

$$\frac{|K|}{\delta t}(u_K^{n+1} - u_K^n) + \sum_{T \cap K \neq \emptyset} \sum_{\sigma_{KL}^T \in \mathcal{E}_{K \cap T}} \Lambda_{KL}^T u_{KL}^{n+1} \Big(\log(u_K^{n+1}) - \log(u_L^{n+1})\Big), \tag{6.2}$$

TABLE 2. Test 1: Nonlinear heat equation with scheme (6.2) and (6.3).

$h_{\mathcal{T}}$	$\Vert\tilde{u}_{h,\delta t}-u\Vert_2$	Rate	$\Vert\tilde{u}_{h,\delta t}-u\Vert_\infty$	Rate	$u_{\min}$
0.250	0.701 E-01	–	0.381 E-00	–	0
0.125	0.686 E-01	0.030	0.370 E-00	−0.001	0
0.063	0.644 E-01	0.091	0.337 E-00	0.072	0
0.031	0.579 E-01	0.150	0.292 E-00	0.135	0
0.016	0.492 E-01	0.245	0.239 E-00	0.231	0

TABLE 3. Test 1: Nonlinear heat equation with scheme (3.1) and (3.2).

$h_{\mathcal{T}}$	$\Vert\tilde{u}_{h,\delta t}-u\Vert_2$	Rate	$\Vert\tilde{u}_{h,\delta t}-u\Vert_\infty$	Rate	$u_{\min}$
0.250	0.142 E-01	–	0.778 E-01	–	0
0.125	0.430 E-02	1.724	0.253 E-01	1.618	0
0.063	0.148 E-02	1.549	0.101 E-01	1.332	0
0.031	0.616 E-03	1.241	0.401 E-02	1.313	0
0.016	0.282 E-03	1.178	0.168 E-02	1.312	0

such that u_{KL}^{n+1} is determined by the upwind relationship

$$u_{KL}^{n+1} = \begin{cases} \max(u_K^{n+1}, u_L^{n+1}) & \text{if} \quad \Lambda_{KL}^T \geq 0 \\ \min(u_K^{n+1}, u_L^{n+1}) & \text{else} \end{cases}. \tag{6.3}$$

Table 2 gives the obtained numerical results by running the nonlinear scheme (6.2) and (6.3). We observe that this technique allows to suppress the nonphysical oscillations. However, it comes with the price of adding too artificial diffusion which slows dramatically the convergence speed and deteriorates the accuracy. This is indeed confirmed since the rate is about 0.245 even on refined meshes. To cope with this deficiency of the above nonlinear methodology, we implement the scheme (3.1) and (3.2) and the results are reported in Table 3. It is seen that the convergence rate is significantly enhanced. This is mainly due to the crucial choice (3.6). The accuracy is of order 1 because of the unwinding which seems natural. On the other hand, the physical range of the computed solution is preserved. We deduce from the last table that there is a good agreement with the expected errors, which are known for upstream methods, together with the obtained results.

6.2. Test 2: Nonlinear diffusion with drift

As in the first example, this test is concerned with the porous medium equation with drift:

$$\partial_t u - \operatorname{div}(2u\Lambda\nabla u) + \operatorname{div} u\mathbf{V} = 0.$$

We set the velocity as $V = -\Lambda\nabla x = -(\lambda_x, 0)^t$. An analytical solution to the above equality is selected under the form

$$u(x,y,t) = \max(\alpha t - x, 0), \quad \forall (x,y,t) \in \Omega \times (0, \mathfrak{T}), \tag{6.4}$$

where $\alpha = 3\lambda_x$. In accordance with (6.4), a Dirichlet boundary condition is taken into account on $\partial\Omega \times (0,\mathfrak{T})$. Here, the final time is fixed to $\mathfrak{T} = 0.1$. The information on the anisotropy reads: $\lambda_x = 1$ and $\lambda_y = 100$. The behavior of our method is compared to the quasilinear discretization and the nonlinear approach suggested by Oulhaj *et al.* [37]. Let us first take a look at the quasilinear scheme. Here the fluxes are approached by a centered

TABLE 4. Test 2: Porous medium equation with quasilinear scheme (6.5).

$h_{\mathcal{T}}$	$\lVert\tilde{u}_{h,\delta t}-u\rVert_2$	Rate	$\lVert\tilde{u}_{h,\delta t}-u\rVert_\infty$	Rate	$u_{\min}$
0.250	0.139 E-01	–	0.110 E-00	–	−0.0679
0.125	0.622 E-02	1.158	0.658 E-01	0.746	−0.0286
0.063	0.261 E-02	1.267	0.420 E-01	0.656	−0.0117
0.031	0.112 E-02	1.187	0.220 E-01	0.909	−0.0059
0.016	0.488 E-03	1.264	0.118 E-01	0.937	−0.0029

approximation. Then, this finite volume scheme is given by

$$\frac{|K|}{\delta t}(u_K^{n+1}-u_K^n)+\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}\Lambda_{KL}^T\left((u_K^{n+1})^2-(u_L^{n+1})^2+\frac{u_K^{n+1}+u_L^{n+1}}{2}(x_K-x_L)\right)=0. \tag{6.5}$$

Note that x_K stands for the x-coordinate of the center of the control volume K. The errors of the discretization (6.5) as well as their corresponding convergence rates are displayed in Table 4. It is shown that the quasilinear scheme (6.5) converges with an order strictly less than 2. This is natural since the exact solution lacks regularity which is due to the degeneracy of the problem. We also notice that the method violates the discrete maximum principle. To circumvent the latter point, we propose two nonlinear corrections. The first one is defined in the same fashion of [12] by making a slight change of variables as investigated in [37]. Therefore, the scheme takes the form

$$\frac{|K|}{\delta t}(u_K^{n+1}-u_K^n)+\sum_{T\cap K\neq\emptyset}\sum_{\sigma_{KL}^T\in\mathcal{E}_{K\cap T}}u_{KL}^{n+1}\Lambda_{KL}^T\left(U_K^{n+1}-U_L^{n+1}\right)=0. \tag{6.6}$$

The new potential quantity includes the drift effects

$$U_K^{n+1}=2u_K^{n+1}+x_K. \tag{6.7}$$

The upstream value is now given by

$$u_{KL}^{n+1}=\begin{cases}u_K^{n+1} & \text{if}\quad \Lambda_{KL}^T(U_K^{n+1}-U_K^{n+1})\geq 0\\ u_L^{n+1} & \text{otherwise}\end{cases}. \tag{6.8}$$

The resolution of (6.6)–(6.8) gives the results exhibited in Table 5. We clearly observe that the method is positive, but the convergence speed is very slow. This drawback of the scheme goes back to the formulation (6.6) all together with the choice (6.8). In order to ameliorate the accuracy we have run our positive version defined by the system (3.1) and (3.2) and the results are presented in Table 6. It is visibly shown that we reach a better accuracy and stability compared to the previous nonlinear approach. This ensures that our method is efficient and can practically be adapted to problems whose solutions are not smooth.

6.3. Test 3: Nonlinear convection-diffusion

In this test we are interested in evaluating the behavior of our method in the presence of the nonlinear convection and diffusion together with the anisotropy of the medium.

$$\partial_t u-\operatorname{div}(\varphi(u)\Lambda\nabla u)+\operatorname{div}(f(u)\mathbf{V})=q. \tag{6.9}$$

The nonlinearities of this equation are:

$$\varphi(u)=\frac{u^2}{1+u^2},\quad f(u)=\frac{u}{1.5-u}\cdot$$

Table 5. Test 2: Porous medium equation with scheme (6.6)–(6.8).

$h_{\mathcal{T}}$	$\|\tilde{u}_{h,\delta t} - u\|_2$	Rate	$\|\tilde{u}_{h,\delta t} - u\|_\infty$	Rate	$u_{\min}$
0.250	0.193 E-01	–	0.136 E-00	–	0
0.125	0.160 E-01	0.260	0.118 E-00	0.200	0
0.063	0.127 E-01	0.331	0.980 E-01	0.277	0
0.031	0.986 E-02	0.365	0.801 E-01	0.282	0
0.016	0.744 E-02	0.425	0.655 E-01	0.305	0

Table 6. Test 2: Porous medium equation with scheme (3.1) and (3.2).

$h_{\mathcal{T}}$	$\|\tilde{u}_{h,\delta t} - u\|_2$	Rate	$\|\tilde{u}_{h,\delta t} - u\|_\infty$	Rate	$u_{\min}$
0.250	0.770 E-02	–	0.827 E-01	–	0
0.125	0.409 E-02	0.912	0.491 E-01	0.751	0
0.063	0.204 E-02	1.016	0.326 E-01	0.599	0
0.031	0.959 E-03	1.064	0.184 E-01	0.803	0
0.016	0.451 E-03	1.138	0.107 E-01	0.815	0

Table 7. Test 3: Nonlinear diffusion-convection with (3.1) and (3.2).

$h_{\mathcal{T}}$	$\|\tilde{u}_{h,\delta t} - u\|_2$	Rate	$\|\tilde{u}_{h,\delta t} - u\|_\infty$	Rate	$u_{\min}$
0.250	0.352 E-04	–	0.411 E-03	–	0
0.125	0.252 E-04	0.480	0.376 E-03	0.128	0
0.063	0.141 E-04	0.842	0.225 E-03	0.749	0
0.031	0.736 E-05	0.923	0.126 E-03	0.815	0
0.016	0.363 E-05	1.066	0.660 E-04	0.982	0

The transport velocity is chosen as

$$\mathbf{V} = (t+0.2)\sin\left(\frac{\pi}{4}(x+y)\right)\left(1.5-(t+0.2)\sin\left(\frac{\pi}{4}(x+y)\right)\right)\begin{pmatrix}1\\-1\end{pmatrix}.$$

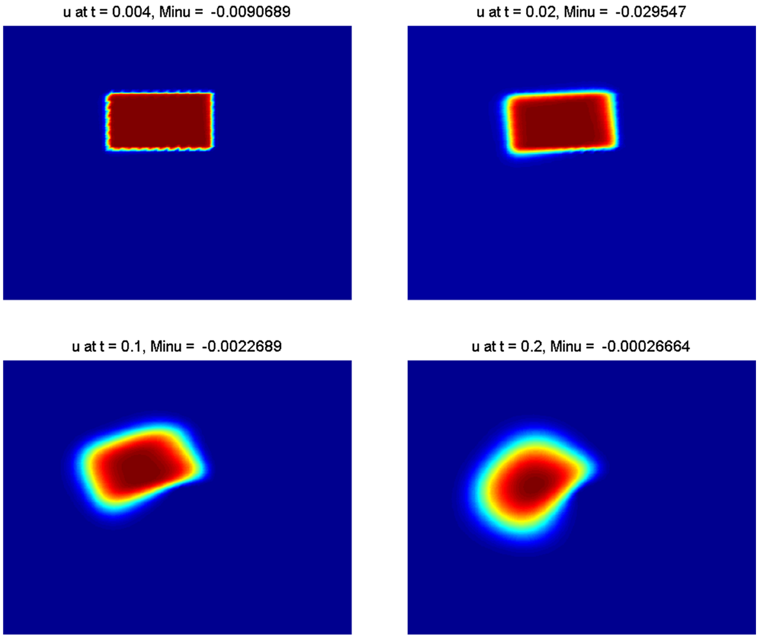
We consider the analytical solution

$$u(x,y,t) = (t+0.2)\sin\left(\frac{\pi}{4}(x+y)\right).$$

The final time is set to $\mathfrak{T} = 0.1$ and the eigenvalues of the tensor Λ are assigned to: $\lambda_x = 1$ and $\lambda_y = 0.001$. By substituting this expression in (6.9) we get a nonnegative source term $q \geq 0$. The accuracy results for this test are displayed in Table 7. As in the previous example, the latter reveals that the convergence rate achieves the order one, which is conforming to our predictions. We also notice that the minimum of the calculated solution stays nonnegative through the refinement procedure.

6.4. Test 4

This last test consists of illustrating the behavior of our approach to approximate a nonlinear convection-diffusion equation of type (1.1) with anisotropy in the case where the analytical solution is unknown. For example, this kind of models describes the transport of a contaminant that diffuses within a medium filled of



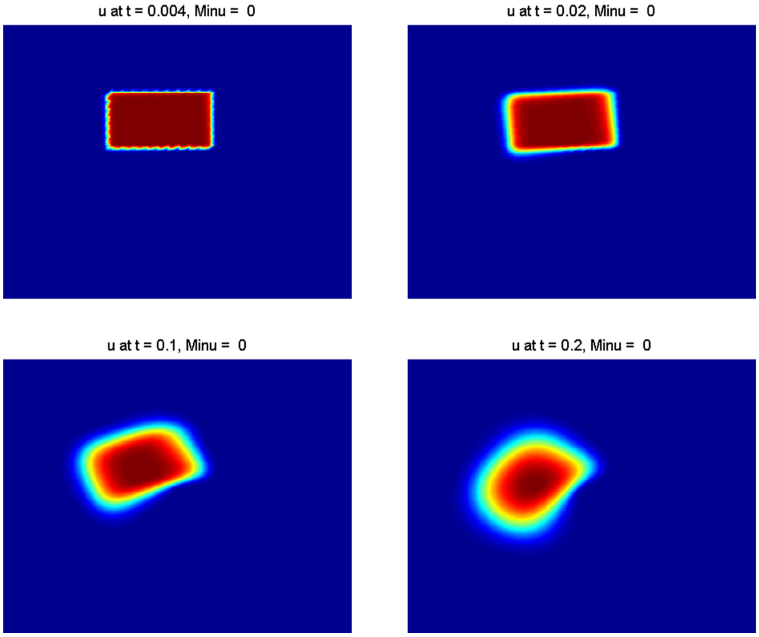
u at t = 0.004, Minu = -0.0090689

u at t = 0.02, Minu = -0.029547

u at t = 0.1, Minu = -0.0022689

u at t = 0.2, Minu = -0.00026664

FIGURE 3. Test 4: The approximate density for simulation times $t_1 = 0.004$, $t_2 = 0.02s$, $t_3 = 0.1$ and $t_4 = 0.2$ using a centered scheme without correction.



u at t = 0.004, Minu = 0

u at t = 0.02, Minu = 0

u at t = 0.1, Minu = 0

u at t = 0.2, Minu = 0

FIGURE 4. Test 4: The approximate density for simulation times $t_1 = 0.004$, $t_2 = 0.02s$, $t_3 = 0.1$ and $t_2 = 0.2$. using the positive scheme (3.1) and (3.2).

water. The purpose here is to stress the capability of our methodology to preserve the nonnegativity of the computed density. To confirm this, let us first consider the following data $\varphi(u) = 0.01u^2$, $f(u) = u$ and

$$\mathbf{V} = 10 \begin{pmatrix} (x - x^2)(1 - 2x) \\ -(y - y^2)(1 - 2y) \end{pmatrix}.$$

The diffusion tensor Λ is described by $\lambda_x = 1$ and $\lambda_y = 0.001$. The final simulation time is $\mathfrak{T} = 0.2$. This example is about a nonlinear diffusion of the quantity u within a fully saturated medium Ω. The underlined process occurs simultaneously with a linear advection whose velocity is expressed by a rotating vector field $\mathbf{V}$. The time step reads $\delta t = 0.002$. We here consider no source term. The boundary condition is prescribed by $u_{|\partial\Omega} = 0$. The initial density is given by

$$u(x, y, 0) = \begin{cases} 1 & \text{in} \quad \Omega' := [0.3, 0.6] \times [0.55, 0.75] \\ 0 & \text{in} \quad \Omega \backslash \Omega' \end{cases}. \tag{6.10}$$

We run our code on the fifth triangular mesh of the family illustrated in Figure 2 which is made of 14 336 triangles. We next compare our strategy to the centered scheme obtained by using Kirchhoff's transformation on the diffusion term and upstreaming the convective one.

Running the centered scheme, the plot given in Figure 3 depicts the simulation results for different instants. The convection effects are dominated with respect to the diffusion. Traditionally, the use of first order upwind approximation on the hyperbolic term permits to produce more artificial viscosity. This regularizes rapidly the solution as seen on the figure in question. In addition, the solution diffuses horizontally which goes back to the strong presence of anisotropy in the x-direction. Furthermore, noticeable undershoots are recorded at each time step of the solver.

On the contrary, we observe that these oscillations disappear on Figure 4 where the same test is carried out by implementing our positive alternative. Analogous remarks can be made as in the previous test. We notice that both solutions posses similar patterns. However, the solution of the positive scheme is expected to diffuse a little bit more than the solution of the centered scheme since our correction is based on the upstreaming technique. This can not be directly viewed on the figure since the convection is too dominated. To see this it suffices to consider the case of the pure diffusion as done in the first example by neglecting the velocity $\mathbf{V}$. We conclude that our approach is more stable and provides satisfactory results.

Acknowledgements. The author would like to thank the referees for theirs fruitful comments and remarks that helped improve the quality and the scope of this work.

References

[1] I. Aavatsmark, T. Barkve, Ø. Bøe and T. Mannseth, Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media. *J. Comput. Phys.* **127** (1996) 2–14.

[2] I. Aavatsmark, G. Eigestad, R. Klausen, M. Wheeler and I. Yotov, Convergence of a symmetric MPFA method on quadrilateral grids. *Comput. Geosci.* **11** (2007) 333–345.

[3] M. Afif and B. Amaziane, Convergence of finite volume schemes for a degenerate convection–diffusion equation arising in flow in porous media. *Comput. Methods Appl. Mech. Eng.* **191** (2002) 5265–5286.

[4] B. Andreianov, F. Boyer and F. Hubert, Discrete duality finite volume schemes for Leray- Lions- type elliptic problems on general 2D meshes. *Numer. Methods Partial Differ. Equ.* **23** (2007) 145–195.

[5] 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) 3633–3670.

[6] T. Arbogast and M.F. Wheeler, A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM J. Numer. Anal.* **33** (1996) 1669–1687.

[7] V.I. Bogachev, Measure Theory. Springer Science & Business Media **1** (2007).

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

[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) 573–597.

[10] Z. Cai, On the finite volume element method. *Numer. Math.* **58** (1990) 713–735.

[11] C. Cancès, M. Cathala and C. Le Potier, Monotone corrections for generic cell-centered finite volume approximations of anisotropic diffusion equations. *Numer. Math.* **125** (2013) 387–417.

[12] C. Cancès and C. Guichard, Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. *Math. Comput.* **85** (2016) 549–580.

[13] 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) 1525–1584.

[14] G. Chavent and J. Jaffré, Mathematical models and finite elements for reservoir simulation: single phase, multiphase and multicomponent flows through porous media. In: Vol. 17 of *Stud. Math. Appl.* North-Holland, Amsterdam (1986).

[15] Z. Chen, G. Huan and Y. Ma, Computational Methods for Multiphase Flows in Porous Media. SIAM **2** (2006).

[16] P. Ciarlet, The Finite Element Method for Elhptic Problems. North-Holland, Amsterdam (1978).

[17] K. Domelevo and P. Omnes, A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *ESAIM: M2AN* **39** (2005) 1203–1249.

[18] J. Droniou, Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Math. Models Methods Appl. Sci.* **24** (2014) 1575–1619.

[19] J. Droniou and R. Eymard, A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numer. Math.* **105** (2006) 35–71.

[20] J. Droniou, R. Eymard, T. Gallouët, C. Guichard and R. Herbin, The Gradient Discretisation Method. Springer **82** (2018).

[21] A. Ern and J.-L. Guermond, Theory and Practice of Finite Elements. Springer Science & Business Media **159** (2013).

[22] R.E. Ewing, T. Lin and Y. Lin, On the accuracy of the finite volume element method based on piecewise linear polynomials. *SIAM J. Numer. Anal.* **39** (2002) 1865–1888.

[23] R. Eymard, T. Gallouët and R. Herbin, Finite volume methods. In: Vol. 7 of *Handbook of Numerical Analysis.* Elsevier (2000) 713–1018.

[24] R. Eymard, T. Gallouït, R. Herbin and A. Michel, Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. *Numer. Math.* **92** (2002) 41–82.

[25] 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) 73–131.

[26] R. Eymard, T. Gallouët and R. Herbin, Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.* **30** (2009) 1009–1043.

[27] R. Eymard, C. Guichard and R. Herbin, Small-stencil 3D schemes for diffusive flows in porous media. *ESAIM: M2AN* **46** (2012) 265–290.

[28] G. Gagneux and M. Madaune-Tort, Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière. Springer Science & Business Media **22** (1995).

[29] M. Ghilani, E.H. Quenjel and M. Saad, Convergence of a positivity-preserving finite volume scheme for compressible two-phase flows in anisotropic porous media: the densities are depending on the physical pressures. submitted (2019).

[30] M. Ghilani, E.H. Quenjel and M. Saad. Positive control volume finite element scheme for a degenerate compressible two-phase flow in anisotropic porous media. *Comput. Geosci.* **23** (2019) 55–79.

[31] R. Helmig, Multiphase Flow and Transport Processes in the Subsurface: A Contribution to the Modeling of Hydrosystems. Springer-Verlag (1997).

[32] R. Herbin and F. Hubert, Benchmark on discretization schemes for anisotropic diffusion problems on general grids, edited by R. Eymard and J.-M. Herard. In: *Finite Volumes for Complex Applications V.* Wiley (2008) 659–692.

[33] W. Hundsdorfer and J.G. Verwer, Numerical Solution of Time-dependent Advection-diffusion-reaction Equations. Springer Science & Business Media **33** (2013).

[34] M. Ibrahim and M. Saad, On the efficacy of a control volume finite element method for the capture of patterns for a volume-filling chemotaxis model. *Comput. Math. App.* **68** (2014) 1032–1051.

[35] R.J. LeVeque, Finite Volume Methods for Hyperbolic Problems. Cambridge University Press **31** (2002).

[36] J.L. Lions, Quelques méthodes de résolution des problèmes aux limites non linéaires. Dunod (1969).

[37] A.A.H. 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: M2AN* **52** (2018) 1533–1567.

[38] E.H. Quenjel, M. Saad, M. Ghilani and M. Bessemoulin-Chatard, On the positivity of a discrete duality finite volume scheme for degenerate nonlinear diffusion equations. submitted (2018).

[39] B. Saad and M. Saad, Study of full implicit petroleum engineering finite-volume scheme for compressible two-phase flow in porous media. *SIAM J. Numer. Anal.* **51** (2013) 716–741.

[40] M. Schneider, L. Agélas, G. Enchéry and B. Flemisch, Convergence of nonlinear finite volume schemes for heterogeneous anisotropic diffusion on general meshes. *J. Comput. Phys.* **351** (2017) 80–107.

[41] R.S. Varga, Matrix Iterative Analysis. Springer Science & Business Media **27** (2009).