Convergence of nonlinear finite volume schemes for heterogeneous anisotropic diffusion on general meshes

Martin Schneider, Léo Agélas, Guillaume Enchéry, Bernd Flemisch

Convergence of nonlinear finite volume schemes for heterogeneous anisotropic diffusion on general meshes[☆]

Léo Agélas[a], Guillaume Enchery[a], Bernd Flemisch[b], Martin Schneider[b,*]

[a] *IFP Energies nouvelles, 1 & 4 avenue du Bois-Préau, 92852 Rueil-Malmaison Cedex, France*
[b] *Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Pfaffenwaldring 61, 70569 Stuttgart, Germany*

Abstract

In the present work, we deal with the convergence of cell-centered nonlinear finite volume schemes for anisotropic and heterogeneous diffusion operators. A general framework for the convergence study of finite volume methods is provided and used to establish the convergence of the new methods. Thorough assessment on a set of anisotropic heterogeneous problems as well as a comparison with linear finite volume schemes is provided.

Keywords: monotone, finite volume methods, heterogeneous anisotropic diffusion, multi-point flux approximation, convergence analysis

1. Introduction

In a variety of physical problems, as for example multi-phase flow in porous media, efficient and robust schemes are required for the discretization of Darcy-type equations. One of the key ingredients for the numerical solution of this type of equations is the discretization of anisotropic heterogeneous elliptic terms [1] on highly complex unstructured grids. In order to maintain mass conservation, the most commonly used schemes applied to Darcy-type equations are either cell-centered finite volume methods, such as multi-point flux approximation methods (MPFA) [2, 3, 4, 5, 6, 7], or mixed and hybrid schemes, such as the mixed finite element (MFE) [8, 9], the mimetic finite difference (MFD) [10, 11] or the hybrid finite volume schemes (HFV) [12, 13]. These mixed or hybrid methods introduce additional face unknowns, whereas MPFA schemes use interpolation rules to eliminate these additional degrees of freedom.

None of these schemes are unconditionally monotone for general heterogeneous and anisotropic elliptic terms and grids. For example, it is proven in [14] that there exist no linear higher-order unconditionally monotone control-volume schemes. Monotone schemes are not only desirable in

[☆] Authors listed alphabetically
[*] Corresponding author
Email addresses: leo.agelas@ifp.fr (Léo Agélas), guillaume.enchery@ifp.fr (Guillaume Enchery), bernd.flemisch@iws.uni-stuttgart.de (Bernd Flemisch), martin.schneider@iws.uni-stuttgart.de (Martin Schneider)

terms of reliability, but also because of the improved robustness. Thinking of highly nonlinear coupled partial differential equations, where secondary variables are calculated using physical laws and relationships that non-linearly depend on primary variables, unphysical solutions can cause convergence problems of linear and nonlinear solvers during the simulation run. Relaxation of the linearity requirement of the schemes allows the construction of nonlinear monotone finite volume schemes. The first concepts of positivity-preserving or discrete extremum-principles-preserving schemes have been presented in [15, 16, 17, 18].

In this article, the proof of convergence of a family of numerical methods is given. The proof relies on concepts that have been developed in [4]. It generalizes the one given in [19] and allows to prove the convergence for the nonlinear finite volume schemes introduced in [15, 16, 17, 18, 20, 21] for which no proof yet existed, as mentioned in [22].

This work is organized as follows: In Section 2, a generic finite volume framework is given, including the proof of convergence under some hypotheses. In Section 3, this framework is used to prove the convergence for a specific family of discretizations. The idea of schemes belonging to this family is the construction of face flux approximations as a convex combination of consistent linear approximations. In Section 4, two representatives of this family, a nonlinear two-point flux approximation (NLTPFA) and a nonlinear multi-point flux approximation (NLMPFA), are derived. These approximations are constructed such that the NLTPFA scheme is monotone and the NLMPFA satisfies discrete extremum principles. Furthermore, sufficient conditions are derived to guarantee the strong consistency of the fluxes. Additionally, possible face interpolators, for which the convergence theory holds, are presented. These schemes are compared to linear ones in Section 5. In the first part 5.1, the convergence of the schemes is analyzed for a mildly and highly anisotropic test case on unstructured grids. In the second part 5.2, the schemes are tested for the extremum-principle-preservation property and it is demonstrated that linear schemes produce negative solution values, in contrast to nonlinear ones. In the last part of Section 5, the linearity-preservation property is investigated and the Northeast German Basin serves as a benchmark problem.

2. Abstract framework

In this section, we present a generic finite volume framework, following ideas that have been introduced in [4]. In Section 2.1, we define the model problem together with generic finite volume discretization schemes. In Section 2.2, proof of convergence of these schemes is given. Furthermore, the existence of discrete solutions is discussed in Section 2.3.

2.1. Model problem and finite volume discretization

Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$, be an open bounded connected polygonal domain with boundary $\partial\Omega$. Let Λ be a symmetric tensor-valued function such that (s.t.) there exist $0 < \alpha_0 < \beta_0 < +\infty$ so that, for almost every (a.e.) $x \in \overline{\Omega}$, the spectrum of $\Lambda(x)$ is contained in $[\alpha_0, \beta_0]$. In the following, the problem

$$\begin{cases} \nabla\cdot(-\Lambda\ \nabla\overline{u}) = f & \text{in } \Omega, \\ \overline{u} = 0 & \text{on } \partial\Omega, \end{cases} \tag{1}$$

is considered, where $f \in L^r(\Omega)$ with $r > 1$ if $d = 2$ and $r = \frac{2d}{d+2}$ if $d > 2$. The existence and uniqueness of a weak solution $\overline{u} \in H_0^1(\Omega)$ of problem (1) is a classical result.

Remark 1. Other standard types of boundary conditions can be considered. However, for ease of presentation, homogeneous Dirichlet conditions are considered within this section.

In what follows, the definition of finite volume discretizations for problem (1) and a generic framework covering fairly general (possibly non-conforming) polygonal meshes is provided.

Definition 1 (Admissible family of discretizations)**.** *An admissible family of finite volume discretizations $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ is a triplet $\mathcal{D}_n = (\mathcal{T}_n, \mathcal{E}_n, \mathcal{P}_n)$, where*

(i) *$\mathcal{T}_n$ is a finite family of non-empty connected open disjoint subsets of Ω (the* cells *or* control volumes*) s.t. $\overline{\Omega} = \cup_{K\in\mathcal{T}_n}\overline{K}$. For all $K \in \mathcal{T}_n$, we denote by $\mathrm{m}_K > 0$ its d-dimensional measure (the* volume*) and let $\partial K \overset{\text{def}}{=} \overline{K} \setminus K$;*

(ii) *$\mathcal{E}_n$ is a finite family of subsets of $\overline{\Omega}$ (the* faces*) s.t., for all $\sigma \in \mathcal{E}_n$, σ is a non-empty closed subset of a hyperplane of $\mathbb{R}^d$ with $(d-1)$-dimensional measure $\mathrm{m}_\sigma > 0$ (the* area*), and s.t. the intersection of two different faces has zero $(d-1)$-dimensional measure. We assume that, for all $K \in \mathcal{T}_n$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}_n$ such that $\partial K = \cup_{\sigma\in\mathcal{E}_K}\sigma$. For a given $\sigma \in \mathcal{E}_n$, either $\mathcal{T}_\sigma \overset{\text{def}}{=} \{K \in \mathcal{T}_n \,|\, \sigma \in \mathcal{E}_K\}$ has exactly one element and then $\sigma \subset \partial\Omega$ (*boundary face*) or $\mathcal{T}_\sigma$ has exactly two elements (*inner face*); the sets of inner and boundary faces are denoted by $\mathcal{E}_{n,\mathrm{int}}$ and $\mathcal{E}_{n,\mathrm{ext}}$ respectively;*

(iii) *$\mathcal{P}_n = \{x_K\}_{K\in\mathcal{T}_n}$ is a family of points of Ω indexed by $\mathcal{T}_n$ (the* cell centers*, not required to be the barycenters) s.t. $x_K \in K$ and K is star-shaped with respect to x_K. For all $K \in \mathcal{T}_n$ and for all $\sigma \in \mathcal{E}_K$ we denote by $d_{K,\sigma}$ the Euclidean distance between x_K and the hyperplane supporting σ. We suppose that there exist $0 < \varrho_1,\ \varrho_2,\ \varrho_3 < +\infty$ independent of n s.t.*

$$\min_{K\in\mathcal{T}_n,\,\sigma\in\mathcal{E}_K} \frac{d_{K,\sigma}}{\operatorname{diam}(K)} \geq \varrho_1, \quad \min_{\sigma\in\mathcal{E}_{n,\mathrm{int}},\,\mathcal{T}_\sigma=\{K,L\}} \frac{\min(d_{K,\sigma}, d_{L,\sigma})}{\max(d_{K,\sigma}, d_{L,\sigma})} \geq \varrho_2, \quad \min_{K\in\mathcal{T}_n} \frac{\operatorname{diam}(K)}{h_{\mathcal{D}_n}} \geq \varrho_3, \tag{2}$$

where $h_{\mathcal{D}_n}$ denotes the size of the discretization defined by $h_{\mathcal{D}_n} \overset{\text{def}}{=} \sup_{K\in\mathcal{T}_n} \operatorname{diam}(K)$.

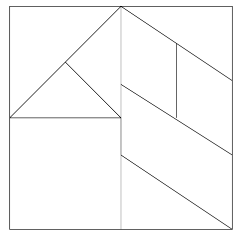


Figure 1: An example of admissible mesh for $d = 2$.

Figure 1 presents an example of an admissible mesh in two space dimensions. With items (ii) and (iii), and since $\frac{\mathrm{m}_\sigma\, d_{K,\sigma}}{d}$ is the measure of the convex hull $\triangle_{K,\sigma}$ of x_K and σ, it is inferred that

$$\forall K \in \mathcal{T}_n, \quad \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma\, d_{K,\sigma} = d\,\mathrm{m}_K\,. \tag{3}$$

For all $K \in \mathcal{T}_n$ and $\sigma \in \mathcal{E}_K$, we denote the unit vector that is normal to σ and outward to K with the term $\mathbf{n}_{K,\sigma}$. For all $K \in \mathcal{T}$ and for all $\Phi \in L^1(K)$, we set $\langle\Phi\rangle_K \stackrel{\text{def}}{=} \mathrm{m}_K^{-1}\int_K \Phi\,\mathrm{d}x$. For vectorial functions, this notation is meant component-wise. For all vectors $x \in \mathbb{R}^n$, $n \in \mathbb{N}^*$, the Euclidean norm will be denoted by $|x| \stackrel{\text{def}}{=} \sqrt{x{\cdot}x}$; for all matrices $A \in \mathbb{R}^n \times \mathbb{R}^n$, $n \in \mathbb{N}^*$, we shall denote by $|A|$ the norm induced by the scalar product of $\mathbb{R}^n$, i.e., $|A| \stackrel{\text{def}}{=} \sup_{x\in\mathbb{R}^d}\dfrac{|Ax|}{|x|}$. The vector space of bounded linear operators from E to F will be denoted by $\mathcal{L}(E;F)$.

In what follows, when referring to a generic element $\mathcal{D}_n$ of an admissible family of discretizations $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$, the subscript n will be dropped for the ease of reading in the case that no ambiguity arises. The space of piecewise constant functions on $\mathcal{T}$ is defined as

$$H_{\mathcal{T}}(\Omega) \stackrel{\text{def}}{=} \{v \in L^2(\Omega)\,|\, v_{|K} \in \mathbb{P}^0(K),\ \forall K \in \mathcal{T}\}.$$

For all $v \in H_{\mathcal{T}}$ and for all $K \in \mathcal{T}$, v_K will denote the (constant) value of v on K, i.e., $v_{|K}(x) = v_K$ for all $x \in K$. In order to endow $H_{\mathcal{T}}$ with a discrete H^1 norm, it is equipped with the following norm

$$||v||_{\mathcal{T}} \stackrel{\text{def}}{=} \left(\sum_{K\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_K}\frac{\mathrm{m}_\sigma}{d_{K,\sigma}}|\gamma_\sigma v - v_K|^2\right)^{1/2},$$

where $\gamma_\sigma \in \mathcal{L}(H_{\mathcal{T}}(\Omega);\mathbb{P}^0(\sigma))$ is defined as

$$\forall v \in H_{\mathcal{T}}(\Omega), \qquad \begin{cases} \gamma_\sigma v = \dfrac{d_{L,\sigma}v_K + d_{K,\sigma}v_L}{d_{K,\sigma}+d_{L,\sigma}} & \text{if } \sigma\in\mathcal{E}_{\text{int}} \text{ with } \mathcal{T}_\sigma = \{K, L\},\\[2ex] \gamma_\sigma v = 0 & \text{if } \sigma \in \mathcal{E}_{\text{ext}}. \end{cases}$$

Let $a_{\mathcal{T}}(u, v, w)$ be a form defined for all $(u, v, w) \in [H_{\mathcal{T}}(\Omega)]^3$. In what follows, discretizations for (1) of the form

$$\text{Find } u \in H_{\mathcal{T}}(\Omega) \text{ s.t. } a_{\mathcal{T}}(u, u, v) = \int_\Omega f v\,\mathrm{d}x \quad \text{ for all } v \in H_{\mathcal{T}}(\Omega) \tag{4}$$

are considered.

Remark 2. Any conservative finite volume scheme is equivalent to a discrete problem of type (4). For all $K \in \mathcal{T}$, and for all $\sigma \in \mathcal{E}_K$, let $F_{K,\sigma} : H_{\mathcal{T}}(\Omega) \times H_{\mathcal{T}}(\Omega) \mapsto \mathbb{P}^0(\sigma)$ be a numerical flux function meant to approximate the diffusive flux flowing out of K through σ such that the finite volume scheme reads: For all $K \in \mathcal{T}$,

$$-\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}(u, u) = \int_K f \,\mathrm{d}x, \tag{5}$$

with locally conservative fluxes: for all $(u, v) \in H_{\mathcal{T}}(\Omega) \times H_{\mathcal{T}}(\Omega)$, $\sigma \in \mathcal{E}_{\text{int}}$ and $\mathcal{T}_\sigma = \{K, L\}$,

$$F_{K,\sigma}(u, v) + F_{L,\sigma}(u, v) = 0. \tag{6}$$

Then, for all $v \in H_{\mathcal{T}}(\Omega)$, by multiplying equation (5) with v_K, $K \in \mathcal{T}$, summing up the resulting equation over $K \in \mathcal{T}$, we obtain for any $v \in H_{\mathcal{T}}(\Omega)$,

$$-\sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}(u, u) v_K = \int_\Omega f v \,\mathrm{d}x. \tag{7}$$

Thus, for all $(u, v, w) \in [H_{\mathcal{T}}(\Omega)]^3$, we define the form

$$a_{\mathcal{T}}(u, v, w) \stackrel{\text{def}}{=} -\sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}(u, v) w_K. \tag{8}$$

Then, thanks to (7) and (8), we obtain a discrete problem of type (4) with $a_{\mathcal{T}}$ defined by (8). Furthermore, starting from the discrete problem (4) with $a_{\mathcal{T}}$ defined by (8), equation (5) is obtained by taking for each $K \in \mathcal{T}$, $v_K = 1$ and $v_{K'} = 0$ for all $K' \in \mathcal{T}$ s.t. $K' \neq K$.

Remark 3. One can also easily verify that the discrete problem of type (4) is equivalent to the problem: Find $u \in H_{\mathcal{T}}(\Omega)$ such that for all $K \in \mathcal{T}$

$$\mathbb{A}_{\mathcal{T}}(u) = \int_K f \,\mathrm{d}x,$$

where the function $\mathbb{A}_{\mathcal{T}} : v \mapsto \mathbb{A}_{\mathcal{T}}(v)$, a mapping from $H_{\mathcal{T}}(\Omega)$ to $H_{\mathcal{T}}(\Omega)$, is defined as

$$(\mathbb{A}_{\mathcal{T}}(v))_K \stackrel{\text{def}}{=} a_{\mathcal{T}}(v, v, \mathbf{1}_K), \tag{9}$$

for each $K \in \mathcal{T}$, where $\mathbf{1}_K$ is the element of $H_{\mathcal{T}}(\Omega)$ equal to one on K and zero elsewhere.

Finally, we introduce the discrete gradient $\widetilde{\nabla}_{\mathcal{D}} \in \mathcal{L}(H_{\mathcal{T}}(\Omega); [H_{\mathcal{T}}(\Omega)]^d)$ which is defined such that for all $K \in \mathcal{T}$ and all $v \in H_{\mathcal{T}}(\Omega)$,

$$\widetilde{\nabla}_{\mathcal{D}} v_{|K} = \frac{1}{\mathrm{m}_K} \sum_{\sigma \in \mathcal{E}_K} \mathrm{m}_\sigma (\gamma_\sigma v - v_K) \mathbf{n}_{K,\sigma}. \tag{10}$$

For all $v \in H_{\mathcal{T}}$ and for all $K \in \mathcal{T}$, $(\widetilde{\nabla}_{\mathcal{D}} v)_K$ will denote the (constant) value of $\widetilde{\nabla}_{\mathcal{D}} v$ on K, i.e., $\widetilde{\nabla}_{\mathcal{D}} v_{|K}(x) = (\widetilde{\nabla}_{\mathcal{D}} v)_K$ for all $x \in K$. Let us notice that Equation (3) together with the Cauchy-Schwarz inequality yield

$$\|\widetilde{\nabla}_{\mathcal{D}} v\|_{[L^2(\Omega)]^d} \leq \sqrt{d} \|v\|_{\mathcal{T}} \quad \forall v \in H_{\mathcal{T}}(\Omega). \tag{11}$$

2.2. Convergence analysis

The aim of this section is to carry out a convergence analysis for finite volume schemes of type (4) by assuming the following properties of the form $a_{\mathcal{T}}(u,v,w)$.

Hypotheses 1. *Let $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a family of discretizations matching Definition 1 s.t. $h_{\mathcal{D}_n} \to 0$ as $n \to \infty$. Let $\mathfrak{D}$ be a dense subspace of $H_0^1(\Omega)$ s.t. $\mathfrak{D} \subset C_0(\overline{\Omega})$, where $C_0(\overline{\Omega})$ denotes the space of continuous functions which vanish on $\partial\Omega$. For all $\varphi \in \mathfrak{D}$, we denote by $\varphi_{\mathcal{T}_n}$ the element of $H_{\mathcal{T}_n}(\Omega)$ s.t., for all $K \in \mathcal{T}_n$, $\varphi_{\mathcal{T}_n|K} = \varphi(\mathbf{x}_K)$. We suppose that:*

(P1) for any $v \in H_{\mathcal{T}_n}(\Omega)$, $v \mapsto a_{\mathcal{T}_n}(v,\cdot,\cdot)$ is a bilinear form;

(P2) $a_{\mathcal{T}_n}$ is uniformly coercive, i.e., there is $0 < \gamma_1 < +\infty$ independent of n s.t.

$$\forall (u,v) \in H_{\mathcal{T}_n}(\Omega) \times H_{\mathcal{T}_n}(\Omega), \quad a_{\mathcal{T}_n}(u,v,v) \geq \gamma_1 \|v\|_{\mathcal{T}_n}^2;$$

(P3) $a_{\mathcal{T}_n}$ is weakly consistent on $\mathfrak{D}$, i.e., for all $\varphi \in \mathfrak{D}$,

$$\epsilon_{\mathcal{D}_n}(\varphi) \stackrel{\text{def}}{=} \max_{(u,v)\in[H_{\mathcal{T}_n}(\Omega)]^2, v\neq 0} \frac{1}{\|v\|_{\mathcal{T}_n}} \left| a_{\mathcal{T}_n}(u,\varphi,v) - \int_\Omega \Lambda\nabla\varphi\cdot\widetilde{\nabla}_{\mathcal{D}_n} v \ \mathrm{d}x \right| \to 0 \ \text{as } n \to \infty. \tag{12}$$

Remark 4. Owing to (3), for a form $a_{\mathcal{T}_n}$ such as (8) derived from a conservative finite volume method, Property (P3) holds for strongly consistent numerical fluxes, i.e. fluxes, for which there is $0 < C_1 < +\infty$ independent of n, s.t. for all $\varphi \in \mathfrak{D}$,

$$\forall K \in \mathcal{T}_n, \ \forall \sigma \in \mathcal{E}_K, \quad \max_{u\in H_{\mathcal{T}_n}(\Omega)} |F_{K,\sigma}(u,\varphi_{\mathcal{T}_n}) - \mathrm{m}_\sigma \langle \Lambda\nabla\varphi\rangle_K \cdot \mathbf{n}_{K,\sigma}| \leq C_1\, \mathrm{m}_\sigma\, h_{\mathcal{D}_n}. \tag{13}$$

Indeed, thanks to the conservation of the fluxes (6), after inserting for each $\sigma \in \mathcal{E}_{\text{int}}$, $\gamma_\sigma v$ in the expression of $a_{\mathcal{T}_n}(u,\varphi,v)$ given by (8), we get

$$a_{\mathcal{T}_n}(u,\varphi,v) = \sum_{K\in\mathcal{T}_n} \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}(u,\varphi_{\mathcal{T}_n})(\gamma_\sigma v - v_K). \tag{14}$$

Furthermore, using (10), we have

$$\int_\Omega \Lambda\nabla\varphi\cdot\widetilde{\nabla}_{\mathcal{D}_n} v \,\mathrm{d}x = \sum_{K\in\mathcal{T}_n} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma \langle \Lambda\nabla\varphi\rangle_K \cdot \mathbf{n}_{K,\sigma}(\gamma_\sigma v - v_K). \tag{15}$$

Hence, by taking the difference between (14) and (15), using (13) and Cauchy-Schwarz inequality along with (3), we deduce that $\epsilon_{\mathcal{D}_n}(\varphi) \leq C_1\sqrt{d\,\mathrm{m}_\Omega}\, h_{\mathcal{D}_n}$, leading to (P3).

The main result of this section is stated in the theorem below.

Theorem 1 (Convergence)**.** *Let us assume that Hypotheses 1 hold and that for each $n \in \mathbb{N}$, there exists at least one solution $u_n \in H_{\mathcal{D}_n}(\Omega)$ to the problem (4). Then, as $n \to \infty$, the sequence of discrete solutions of problem (4), denoted as $\{u_n\}_{n\in\mathbb{N}}$, converges to the solution $\overline{u}$ of (1) in $L^q(\Omega)$ for all $q \in [1, 2d/(d-2))$ (and weakly in $L^{2d/(d-2)}(\Omega)$ if $d > 2$).*

Proof. The proof is based on a few technical propositions which are reminded in Section 7. Owing to the stability estimate (68) together with Theorem 2, there is $\widetilde{u} \in H_0^1(\Omega)$ s.t., up to a subsequence, (i) $\{u_n\}_{n\in\mathbb{N}}$ converges to $\widetilde{u}$ in $L^q(\Omega)$ for all $q \in [1, 2d/(d-2))$ (and weakly in $L^{2d/(d-2)}(\Omega)$ if $d > 2$) and (ii) $\{\widetilde{\nabla}_{\mathcal{D}_n} u_n\}_{n\in\mathbb{N}}$ weakly converges to $\nabla\widetilde{u}$ in $[L^2(\Omega)]^d$. It only remains to prove that $\widetilde{u} = \overline{u}$. Let $\varphi \in \mathfrak{D}$. Owing to (11) together with (P2) and (P1), we infer

$$\|\widetilde{\nabla}_{\mathcal{D}_n}(u_n - \varphi_{\mathcal{T}_n})\|^2_{[L^2(\Omega)]^d} \leq d\|u_n - \varphi_{\mathcal{T}_n}\|^2_{\mathcal{T}_n} \leq \frac{d}{\gamma_1} a_{\mathcal{T}_n}(u_n, u_n - \varphi_{\mathcal{T}_n}, u_n - \varphi_{\mathcal{T}_n}) = \frac{d}{\gamma_1}(T_1 + T_2), \quad (16)$$

where $T_1 \stackrel{\text{def}}{=} \int_\Omega f(u_n - \varphi_{\mathcal{T}_n})\,\mathrm{d}x$ and $T_2 \stackrel{\text{def}}{=} a_{\mathcal{T}_n}(u_n, \varphi_{\mathcal{T}_n}, \varphi_{\mathcal{T}_n} - u_n)$. Since $f \in L^r(\Omega)$ and $\{u_n\}_{n\in\mathbb{N}}$ weakly converges towards $\widetilde{u}$ in $L^q(\Omega)$ for all $q < +\infty$ if $d = 2$ and for all $q = \frac{2d}{d-2}$ if $d > 2$, we have

$$T_1 \to \int_\Omega f(\widetilde{u} - \varphi)\,\mathrm{d}x \text{ as } n \to \infty. \quad (17)$$

Furthermore, we have

$$\begin{aligned} a_{\mathcal{T}_n}(u_n, \varphi_{\mathcal{T}_n}, u_n) = &\left(a_{\mathcal{T}_n}(u_n, \varphi_{\mathcal{T}_n}, u_n) - \int_\Omega \Lambda\nabla\varphi\cdot\widetilde{\nabla}_{\mathcal{D}_n} u_n\,\mathrm{d}x\right) \\ &+ \int_\Omega \Lambda\nabla\varphi\cdot\widetilde{\nabla}_{\mathcal{D}_n} u_n\,\mathrm{d}x \stackrel{\text{def}}{=} T_{2,1} + T_{2,2}. \end{aligned}$$

We observe that $T_{2,1} \leq \epsilon_{\mathcal{D}_n}(\varphi)\|u_n\|_{\mathcal{T}_n}$. Thanks to Proposition 6, $\|u_n\|_{\mathcal{T}_n}$ is uniformly bounded with respect to n. Thus, according to property (P3), $T_{2,1} \to 0$ as $n \to \infty$. The weak convergence of $\{\widetilde{\nabla}_{\mathcal{D}_n} u_n\}_{n\in\mathbb{N}}$ also leads to $T_{2,2} \to \int_\Omega \Lambda\nabla\varphi\cdot\nabla\widetilde{u}\,\mathrm{d}x$ as $n \to \infty$.

Let us now consider T_2. By Proposition 5, $\|\varphi_{\mathcal{T}_n}\|_{\mathcal{T}_n}$ is uniformly bounded with respect to n; since $\varphi_{\mathcal{T}_n}$ obviously converges to φ, it is then easy, using Theorem 2, to see that $\widetilde{\nabla}_{\mathcal{D}_n}\varphi_{\mathcal{T}_n}$ weakly converges to $\nabla\varphi$. Proceeding in a similar way as for $a_{\mathcal{T}_n}(u_n, \varphi_{\mathcal{T}_n}, u_n)$, we can thus prove that $a_{\mathcal{T}_n}(u_n, \varphi_{\mathcal{T}_n}, \varphi_{\mathcal{T}_n}) \to \int_\Omega \Lambda\nabla\varphi\cdot\nabla\varphi\,\mathrm{d}x$ as $n \to \infty$. Therefore,

$$T_2 \to \int_\Omega \Lambda\nabla\varphi\cdot\nabla(\varphi - \widetilde{u})\,\mathrm{d}x \text{ as } n \to \infty. \quad (18)$$

Using the weak convergence of $\widetilde{\nabla}_{\mathcal{D}_n}(u_n - \varphi_{\mathcal{T}_n})$ in $[L^2(\Omega)]^d$, we get that $\liminf_{n\to\infty} \|\widetilde{\nabla}_{\mathcal{D}_n}(u_n - \varphi_{\mathcal{T}_n})\|_{[L^2(\Omega)]^d} \geq \|\nabla(\widetilde{u} - \varphi)\|_{[L^2(\Omega)]^d}$.

Plugging (17) and (18) into the right hand side of (16), we conclude that, for all $\varphi \in \mathfrak{D}$,

$$\|\nabla(\widetilde{u} - \varphi)\|^2_{[L^2(\Omega)]^d} \leq \frac{d}{\gamma_1}\left(\int_\Omega f(\widetilde{u} - \varphi)\,\mathrm{d}x + \int_\Omega \Lambda\nabla\varphi\cdot\nabla(\varphi - \widetilde{u})\,\mathrm{d}x\right).$$

Thanks to the definition of the test space, we can apply this inequality to a sequence $\{\varphi_m\}_{m\in\mathbb{N}} \in \mathfrak{D}$ which tends to $\overline{u}$ in $H_0^1(\Omega)$ and let $m \to \infty$; since $\overline{u}$ solves problem (1), we obtain

$$\|\nabla(\widetilde{u} - \overline{u})\|^2_{[L^2(\Omega)]^d} \leq \frac{d}{\gamma_1}\left[\int_\Omega f(\widetilde{u} - \overline{u})\,\mathrm{d}x - \int_\Omega \Lambda\nabla\overline{u}\cdot\nabla(\widetilde{u} - \overline{u})\,\mathrm{d}x\right] = 0,$$

i.e., $\widetilde{u} = \overline{u}$. Due to the uniqueness of the solution of (1), we deduce that the entire sequence $\{u_n\}_{n\in\mathbb{N}}$ converges to $\overline{u}$ in $L^q(\Omega)$ for all $q \in [1, 2d/(d-2))$ (and weakly in $L^{2d/(d-2)}(\Omega)$ if $d > 2$). Note that the order in which the limits for $n \to \infty$ and $m \to \infty$ are taken cannot be exchanged, since the sequence $\{\|(\varphi_m)_{\mathcal{T}_n}\|_{\mathcal{T}_n,I}\}_{m\in\mathbb{N}}$ is possibly unbounded. This concludes the proof. $\square$

2.3. Existence of a discrete solution

In this section, we briefly discuss the existence of discrete solutions for problem (4). Thanks to Proposition 6, Remark 3 and the application of Brouwer's topological degree leads to the proposition below whose proof is omitted here (see Proposition 3.4 in [19, 23] for more details).

Proposition 1 (Existence of a discrete solution)**.** *Assume that property (P2) of Hypotheses 1 holds and that for each $n \in \mathbb{N}$, $\mathbb{A}_{\mathcal{T}_n}$ is continuous on $H_{\mathcal{T}_n}(\Omega)$. Then, problem* (4) *admits at least one solution $u_n \in H_{\mathcal{T}_n}(\Omega)$ for each $n \in \mathbb{N}$.*

3. Application to some nonlinear finite volume schemes

An established idea to obtain monotone or extremum-principles-preserving schemes, as those developed in [15, 16, 17, 18, 20, 21, 24, 19], is to compute for each interior edge $\sigma \in \mathcal{E}_{\text{int}}$, with $\mathcal{T}_\sigma = \{K, L\}$, two consistent linear flux approximations $\tilde{F}_{K,\sigma}(u)$ and $\tilde{F}_{L,\sigma}(u)$ depending on the unknown $u \in H_{\mathcal{T}}(\Omega)$, and to define the final flux $F_{K,\sigma}(u,u)$ as a convex combination of these fluxes with coefficients also depending on u:

$$\begin{gathered} F_{K,\sigma}(u,u) = \mu_{K,\sigma}(u)\tilde{F}_{K,\sigma}(u) - \mu_{L,\sigma}(u)\tilde{F}_{L,\sigma}(u), \\ \text{with } \mu_{K,\sigma}(u) \geq 0, \mu_{L,\sigma}(u) \geq 0 \text{ and } \mu_{K,\sigma}(u) + \mu_{L,\sigma}(u) = 1. \end{gathered} \tag{19}$$

For any $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\text{int}}$, the linear flux $\tilde{F}_{K,\sigma}(u)$ is built in order to ensure the strong consistency, i.e, there exist $\mathfrak{D} \subset C_0(\overline{\Omega})$, a dense subspace of $H_0^1(\Omega)$, and $0 < C_1 < +\infty$ depending only on the mesh regularity (2), s.t. for all $\varphi \in \mathfrak{D}$,

$$\forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_K, \quad \left| \tilde{F}_{K,\sigma}(\varphi_{\mathcal{T}}) - \mathrm{m}_\sigma \langle \Lambda \nabla \varphi \rangle_K \cdot \mathbf{n}_{K,\sigma} \right| \leq C_1 \, \mathrm{m}_\sigma \, h_{\mathcal{D}}. \tag{20}$$

In (41) and (42) of Section 4, we specify the choice of the space $\mathfrak{D}$ related to the strong consistency property (20).

The coefficients $\mu_{K,\sigma}(u)$ and $\mu_{L,\sigma}(u)$ are chosen to eliminate the "bad" parts of $\tilde{F}_{K,\sigma}(u)$ and $\tilde{F}_{L,\sigma}(u)$, that are responsible for the possible loss of monotonicity. For any $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\text{int}}$ and $L \in \mathcal{T}_K$ such that $\mathcal{T}_\sigma = \{K, L\}$, we thus get from (19) the function $F_{K,\sigma}(\cdot,\cdot)$, defined for all $(u,v) \in [H_{\mathcal{T}}(\Omega)]^2$, as

$$F_{K,\sigma}(u,v) = \mu_{K,\sigma}(u)\tilde{F}_{K,\sigma}(v) - \mu_{L,\sigma}(u)\tilde{F}_{L,\sigma}(v). \tag{21}$$

It is observed that for any $\sigma \in \mathcal{E}_{\text{int}}$ with $\mathcal{T}_\sigma = \{K, L\}$, the fluxes are conservative, i.e, $F_{K,\sigma}(u,v) + F_{L,\sigma}(u,v) = 0$. Thus, from Section 2, the finite volume scheme (5) defined from the fluxes (19) is equivalent to problem (4) with the form $a_{\mathcal{T}}$ (8), which is defined from the fluxes (21). Therefore, the following corollary can be deduced.

Corollary 1. *Let $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be an admissible family of discretizations matching Definition 1 s.t. $h_{\mathcal{D}_n} \to 0$ as $n \to \infty$. We have the following results:*

- *if, for all $n \in \mathbb{N}$, $K \in \mathcal{T}_n$ and $\sigma \in \mathcal{E}_K$, the functions $v \mapsto F_{K,\sigma}(v,v)$, defined by (19), are continuous on $H_{\mathcal{T}_n}(\Omega)$ and if the uniform coercivity property (P2) holds, then there exists at least one solution $u_n \in H_{\mathcal{T}_n}(\Omega)$ of problem (5);*

- *if, in addition, the strong consistency property (20) is satisfied, then the sequence $\{u_n\}_{n\in\mathbb{N}}$ of discrete solutions of problem (5), with numerical fluxes defined by (19), converges to the solution $\overline{u}$ of the continuous problem (1) in $L^q(\Omega)$ for all $q \in [1, 2d/(d-2))$ (and weakly in $L^{2d/(d-2)}(\Omega)$ if $d > 2$) as $n \to \infty$.*

Proof. To prove this result, we use the equivalence between the problem (5) and (4) with $a_{\mathcal{T}_n}$ defined by (8). By assumption, we get that for any $n \in \mathbb{N}$ and for all $K \in \mathcal{T}_n$ and $\sigma \in \mathcal{E}_K$, the function $v \mapsto F_{K,\sigma}(v,v)$ is continuous. From (8) and (9), we notice that the function $\mathbb{A}_{\mathcal{T}_n}$, defined here by $(\mathbb{A}_{\mathcal{T}_n}(v))_K = -\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}(v,v)$ for all $K \in \mathcal{T}_n$, is continuous on $H_{\mathcal{T}_n}(\Omega)$. Therefore, thanks to Proposition 1, we infer that for each $n \in \mathbb{N}$, there exists at least one solution $u_n \in H_{\mathcal{T}_n}(\Omega)$ to the problem (5), which gives the first result. The second one is a consequence of Theorem 1 since

- the fluxes $\{\tilde{F}_{K,\sigma}(\cdot)\}_{K\in\mathcal{T}_n,\sigma\in\mathcal{E}_K}$ are linear on $H_{\mathcal{T}_n}(\Omega)$, which gives (P1),

- the consistency of the fluxes (P3) can be obtained by proving the strong consistency of the fluxes $F_{K,\sigma}$ given by (21) (see Remark 4 which holds by assumption (20).

□

4. Construction of nonlinear finite volume schemes

In the previous section, the proof of the convergence of nonlinear finite volume schemes of type (19) has been given. In this section, we describe two schemes existing in the literature with some improvements, where the first scheme is monotone (see [15, 16, 17, 18, 21]) and the second one satisfies discrete extremum principles (see [19, 24, 25, 20]). Please note that for nonlinear schemes monotonicity only guarantees that the scheme is positivity-preserving. The presented schemes differ in the choice of the weights $\mu_{K,\sigma}, \mu_{L,\sigma}$ (19).

4.1. Consistent flux approximations

In the following, the fluxes $\tilde{F}_{K,\sigma}(u), \tilde{F}_{L,\sigma}(u)$ are constructed such that (20) holds. The decomposition of the conormal, defined as $\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma}$, in a basis $(\mathbf{x}_{\sigma'} - \mathbf{x}_K)_{\{\sigma'\in\mathcal{S}_{K,\sigma}\}}$ with coordinates $(\alpha_{K,\sigma\sigma'})_{\{\sigma'\in\mathcal{S}_{K,\sigma}\}}$ with $\mathcal{S}_{K,\sigma} \subset \mathcal{E}_K$ is calculated by solving the following optimization problem

$$\min_{\gamma\geq 0,\ \tilde{\alpha}\in\mathbb{R}^{|\mathcal{E}_K|}} \kappa\gamma + \sum_{\sigma'\in\mathcal{E}_K} \tilde{\alpha}_{\sigma'} \qquad \text{subject to} \quad \frac{\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma}}{|\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma}|} = \sum_{\sigma'\in\mathcal{E}_K} \tilde{\alpha}_{\sigma'} \frac{\mathbf{x}_{\sigma'} - \mathbf{x}_K}{|\mathbf{x}_{\sigma'} - \mathbf{x}_K|}$$
$$\sum_{\sigma'\in\mathcal{E}_K} \tilde{\alpha}_{\sigma'} \frac{|\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma}|}{|\mathbf{x}_{\sigma'} - \mathbf{x}_K|} \geq \delta, \quad -C_\alpha \leq -\gamma \leq \tilde{\alpha}_{\sigma'} \leq C_\alpha, \tag{22}$$

for given strictly positive parameters δ and C_α. Specifying the final coefficients as

$$\alpha_{K,\sigma\sigma'} \stackrel{\text{def}}{=} \tilde{\alpha}_{\sigma'} \frac{|\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma}|}{|\mathbf{x}_{\sigma'} - \mathbf{x}_K|}, \tag{23}$$

results in the following conormal decomposition

$$\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma} = \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} \alpha_{K,\sigma\sigma'} (\mathbf{x}_{\sigma'} - \mathbf{x}_K), \tag{24}$$

where the face stencil is defined as

$$\mathcal{S}_{K,\sigma} \stackrel{\text{def}}{=} \{\sigma' \in \mathcal{E}_K \mid \alpha_{K,\sigma\sigma'} \neq 0\}. \tag{25}$$

This decomposition is used to define consistent flux approximations $\tilde{F}_{K,\sigma}(u), \tilde{F}_{L,\sigma}(u)$. The idea of formulating the conormal decomposition as an optimization problem has been recently introduced in [26].

Proposition 2. *Let $\mathcal{D}$ be an element of a family of discretizations matching Definition 1 and let $\alpha_{K,\sigma\sigma'}$ be calculated from (22)-(23). Then, for any $\varphi \in C^2(\mathcal{T}) \cap C_0(\overline{\Omega})$ and $K \in \mathcal{T}$, we have the following estimate:*

$$\left| \mathrm{m}_\sigma \langle\Lambda\nabla\varphi\rangle_K \cdot \mathbf{n}_{K,\sigma} - \mathrm{m}_\sigma \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} \alpha_{K,\sigma\sigma'} (\varphi(\mathbf{x}_{\sigma'}) - \varphi(\mathbf{x}_K)) \right| \leq C\, \mathrm{m}_\sigma \operatorname{diam}(K). \tag{26}$$

Proof. We observe that for any $\varphi \in C^2(\mathcal{T}) \cap C_0(\overline{\Omega})$ and $K \in \mathcal{T}$,

$$\begin{aligned} \mathrm{m}_\sigma \langle\Lambda\nabla\varphi\rangle_K \cdot \mathbf{n}_{K,\sigma} &= \frac{\mathrm{m}_\sigma}{\mathrm{m}_K} \int_K \Lambda\nabla\varphi \cdot \mathbf{n}_{K,\sigma}\, \mathrm{d}x \\ &= \frac{\mathrm{m}_\sigma}{\mathrm{m}_K} \int_K \Lambda(x)(\nabla\varphi(x) - \nabla\varphi(\mathbf{x}_K)) \cdot \mathbf{n}_{K,\sigma}\, \mathrm{d}x + \mathrm{m}_\sigma \langle\Lambda\rangle_K \nabla\varphi(\mathbf{x}_K) \cdot \mathbf{n}_{K,\sigma}. \end{aligned} \tag{27}$$

Since K is star-shaped with respect to $\mathbf{x}_K$ Taylor's Theorem can be used to infer

$$\left| \frac{\mathrm{m}_\sigma}{\mathrm{m}_K} \int_K \Lambda(x)(\nabla\varphi(x) - \nabla\varphi(\mathbf{x}_K)) \cdot \mathbf{n}_{K,\sigma}\, \mathrm{d}x \right| \leq C_\varphi \beta_0\, \mathrm{m}_\sigma \operatorname{diam}(K), \tag{28}$$

where $C_\varphi = \mathcal{O}(\|\varphi\|_{C^2(K)})$.

Let us now estimate the second term in the right hand side of equation (27). Inserting the conormal decomposition (24) yields

$$\mathrm{m}_\sigma\, \nabla\varphi(\mathbf{x}_K) \cdot \langle\Lambda\rangle_K \mathbf{n}_{K,\sigma} = \mathrm{m}_\sigma \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} \alpha_{K,\sigma\sigma'} \nabla\varphi(\mathbf{x}_K) \cdot (\mathbf{x}_{\sigma'} - \mathbf{x}_K). \tag{29}$$

Since K is star-shaped with respect to $\mathbf{x}_K$, Taylor's Theorem can again be used to deduce that for all $\sigma' \in \mathcal{S}_{K,\sigma}$,

$$|\varphi(\mathbf{x}_{\sigma'}) - \varphi(\mathbf{x}_K) - \nabla\varphi(\mathbf{x}_K)\cdot(\mathbf{x}_{\sigma'} - \mathbf{x}_K)| \leq C_\varphi \mathrm{diam}(K)^2. \tag{30}$$

Owing to (29) and (30), we get

$$\left| \mathrm{m}_\sigma \, \nabla\varphi(\mathbf{x}_K)\cdot\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma} - \mathrm{m}_\sigma \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} \alpha_{K,\sigma\sigma'}(\varphi(\mathbf{x}_{\sigma'}) - \varphi(\mathbf{x}_K)) \right| \leq \mathrm{m}_\sigma \, C_\varphi \mathrm{diam}(K)^2 \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} |\alpha_{K,\sigma\sigma'}|. \tag{31}$$

Due to the constraints of the optimization problem (22), we observe that for all $\sigma' \in \mathcal{S}_{K,\sigma}$,

$$|\alpha_{K,\sigma\sigma'}| \leq C_\alpha \frac{|\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma}|}{|\mathbf{x}_{\sigma'} - \mathbf{x}_K|}. \tag{32}$$

We thus deduce from (2) that for all $\sigma' \in \mathcal{S}_{K,\sigma}$,

$$|\alpha_{K,\sigma\sigma'}| \leq \frac{C_\alpha \beta_0}{\varrho_1 \mathrm{diam}(K)}. \tag{33}$$

Using (31) and (33), it follows that

$$\left| \mathrm{m}_\sigma \, \nabla\varphi(\mathbf{x}_K)\cdot\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma} - \mathrm{m}_\sigma \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} \alpha_{K,\sigma\sigma'}(\varphi(\mathbf{x}_{\sigma'}) - \varphi(\mathbf{x}_K)) \right| \leq |\mathcal{E}_K| \, \mathrm{m}_\sigma \, C_\varphi C_\alpha \frac{\beta_0}{\varrho_1} \, \mathrm{diam}(K). \tag{34}$$

Then, including (28) and (34) the following desired estimate is obtained from (27)

$$\left| \mathrm{m}_\sigma \langle\Lambda\nabla\varphi\rangle_K\cdot\mathbf{n}_{K,\sigma} - \mathrm{m}_\sigma \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} \alpha_{K,\sigma\sigma'}(\varphi(\mathbf{x}_{\sigma'}) - \varphi(\mathbf{x}_K)) \right| \leq C_\varphi \beta_0 \left(1 + \frac{C_\alpha |\mathcal{E}_K|}{\varrho_1}\right) \mathrm{m}_\sigma \, \mathrm{diam}(K), \tag{35}$$

which completes the proof. □

Corollary 2 (Strong consistency)**.** *Let $\mathcal{D}$ be an element of a family of discretizations matching Definition 1. Let $\alpha_{K,\sigma\sigma'}$ be calculated from* (22)*-*(23)*. Let $\mathfrak{D}$ be a dense subspace of $H_0^1(\Omega)$ s.t. $\mathfrak{D} \subset C^2(\mathcal{T}) \cap C_0(\overline{\Omega})$. For $\sigma \in \mathcal{E}$, let $I_\sigma \in \mathcal{L}(H_\mathcal{T}(\Omega); \mathbb{P}^0(\sigma))$, be a trace reconstruction operator such that for all $\varphi \in \mathfrak{D}$*

$$|I_\sigma \varphi_\mathcal{T} - \varphi(\mathbf{x}_\sigma)| \leq \varrho h_\mathcal{D}^2, \tag{36}$$

where $\varrho > 0$ only depends on the mesh regularities (2)*. Then, the linear fluxes defined as*

$$\tilde{F}_{K,\sigma}(v) \stackrel{\text{def}}{=} \mathrm{m}_\sigma \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} \alpha_{K,\sigma\sigma'}(I_{\sigma'} v - v_K), \quad \forall\, v \in H_\mathcal{T}(\Omega), K \in \mathcal{T}, \sigma \in \mathcal{E}_K, \tag{37}$$

satisfy the strong consistency assumption (20)*.*

Proof. Thanks to Proposition 2, we obtain that for all $\varphi \in \mathfrak{D}$

$$\begin{aligned} \left| \mathrm{m}_\sigma \langle\Lambda\nabla\varphi\rangle_K\cdot\mathbf{n}_{K,\sigma} - \tilde{F}_{K,\sigma}(\varphi_\mathcal{T}) \right| \quad &\leq C \, \mathrm{m}_\sigma \, \mathrm{diam}(K) \\ &+ \mathrm{m}_\sigma \max_{\sigma'\in\mathcal{S}_{K,\sigma}} |I_{\sigma'}\varphi_\mathcal{T} - \varphi(\mathbf{x}_{\sigma'})| \sum_{\sigma'\in\mathcal{S}_{K,\sigma}} |\alpha_{K,\sigma\sigma'}|. \end{aligned} \tag{38}$$

However, thanks to (33), we have $\displaystyle\sum_{\sigma'\in\mathcal{S}_{K,\sigma}} |\alpha_{K,\sigma\sigma'}| \leq \frac{C_\alpha|\mathcal{E}_K|\beta_0}{\varrho_1 \mathrm{diam}(K)}$ and then from (38) we deduce that

$$\begin{aligned}\left|\mathrm{m}_\sigma\langle\Lambda\nabla\varphi\rangle_K\cdot\mathbf{n}_{K,\sigma} - \tilde{F}_{K,\sigma}(\varphi_{\mathcal{T}})\right| &\leq C\,\mathrm{m}_\sigma\,\mathrm{diam}(K)\\ &+\frac{C_\alpha|\mathcal{E}_K|\beta_0}{\varrho_1}\frac{\mathrm{m}_\sigma}{\mathrm{diam}(K)}\max_{\sigma'\in\mathcal{S}_{K,\sigma}}|I_{\sigma'}\varphi_{\mathcal{T}}-\varphi(\mathbf{x}_{\sigma'})|.\end{aligned} \tag{39}$$

On one hand, for any $K\in\mathcal{T}$, we have $\mathrm{diam}(K)\leq h_{\mathcal{D}}$. On the other hand, thanks to (2), for any $K\in\mathcal{T}$, we get $\dfrac{1}{\mathrm{diam}(K)}\leq\dfrac{1}{\varrho_3 h_{\mathcal{D}}}$. Therefore from (39), we infer

$$\begin{aligned}\left|\mathrm{m}_\sigma\langle\Lambda\nabla\varphi\rangle_K\cdot\mathbf{n}_{K,\sigma} - \tilde{F}_{K,\sigma}(\varphi_{\mathcal{T}})\right| &\leq C\,\mathrm{m}_\sigma\,h_{\mathcal{D}}\\ &+\frac{C_\alpha|\mathcal{E}_K|\beta_0}{\varrho_1\varrho_3}\frac{\mathrm{m}_\sigma}{h_{\mathcal{D}}}\max_{\sigma'\in\mathcal{S}_{K,\sigma}}|I_{\sigma'}\varphi_{\mathcal{T}}-\varphi(\mathbf{x}_{\sigma'})|.\end{aligned} \tag{40}$$

The strong consistency of the fluxes follows due to assumption (36). □

4.2. Choice of trace reconstruction operators

With the result obtained in the last section, we now propose choices for the space $\mathfrak{D}$ and the trace reconstruction operators $I_\sigma\in\mathcal{L}(H_{\mathcal{T}}(\Omega);\mathbb{P}^0(\sigma))$. The first choice consists in taking for all $u\in H_{\mathcal{T}}(\Omega),\sigma\in\mathcal{E}_{\mathrm{int}}$:

$$\begin{aligned}&\mathfrak{D}=C_c^\infty(\Omega),\\ &I_\sigma u=\sum_{K\in\mathcal{B}_\sigma}\beta_{K,\sigma}u_K,\end{aligned} \tag{41}$$

where $\mathcal{B}_\sigma$ is a subset of $\mathcal{T}$ with $\mathrm{card}(\mathcal{B}_\sigma)\geq d$, and $(\beta_{K,\sigma})_{K\in\mathcal{B}_\sigma}$ is a family of nonnegative real numbers such that $\displaystyle\sum_{K\in\mathcal{B}_\sigma}\beta_{K,\sigma}=1$ and $\mathbf{x}_\sigma=\displaystyle\sum_{K\in\mathcal{B}_\sigma}\beta_{K,\sigma}\mathbf{x}_K$. Both choices in (41) ensure that Corollary 1 is satisfied for the nonlinear finite volume schemes considered in Section 3 with the assumption that the permeability Λ belongs to $L^\infty(\Omega)$. Our result is thus an improvement of the convergence result obtained in [19] which requires Λ to be piecewise Lipschitz-continuous on Ω.

However, the choice of a convex combination made in (41), and in [19] as well, does not allow us to retrieve exactly piecewise linear solutions of problem (1) for heterogeneous permeabilities Λ which are cell-wise C^2 on Ω. This choice may lead to non-physical solutions of problem (1) for this kind of permeability functions. To handle these cases, we propose a second choice for $\mathfrak{D}$ and the trace reconstruction operators. To that purpose, we make the following hypotheses.

Hypotheses 2. *(Q1)* $P_\Omega\overset{\mathrm{def}}{=}\{\Omega_i\}_{i=1\ldots N_\Omega}$ *is a finite partition of* Ω *into open connected disjoint polygonal subsets,*

(Q2) Λ *is a symmetric tensor-valued function such that* $\Lambda_{|\Omega_i}\in[C^2(\overline{\Omega_i})]^{d\times d}$ *for all* $i=1\ldots N_\Omega$*,*

(Q3) $\mathcal{T}$ *is compatible with* P_Ω *(each cell is contained in one element of the partition* P_Ω*).*

We then suggest, with these additional assumptions, to take, for all $u \in H_{\mathcal{T}}(\Omega), \sigma \in \mathcal{E}_{\text{int}}$:

$$\begin{aligned} &\mathfrak{D} = \mathcal{Q}, \\ &I_\sigma u = \omega_K u_K + \omega_L u_L, \end{aligned} \tag{42}$$

where $\mathcal{Q}$ is defined and proved to be dense in $H_0^1(\Omega)$, as described in Proposition 3, and ω_K and ω_L given below, are the coefficients defining the harmonic averaging interpolator introduced in [27]:

$$\begin{aligned} &\omega_K = \frac{d_{L,\sigma}\tau_{K,\sigma}}{d_{L,\sigma}\tau_{K,\sigma} + d_{K,\sigma}\tau_{L,\sigma}}, \quad \omega_L = \frac{d_{K,\sigma}\tau_{L,\sigma}}{d_{L,\sigma}\tau_{K,\sigma} + d_{K,\sigma}\tau_{L,\sigma}}, \\ &\tau_{K,\sigma} = \mathbf{n}_{K,\sigma}\langle\Lambda\rangle_K \mathbf{n}_{K,\sigma}, \quad \tau_{L,\sigma} = \mathbf{n}_{L,\sigma}\langle\Lambda\rangle_L \mathbf{n}_{L,\sigma}, \\ &\mathbf{x}_\sigma = \omega_K \mathbf{x}_K + \omega_L \mathbf{x}_L + \frac{d_{K,\sigma}d_{L,\sigma}}{d_{L,\sigma}\tau_{K,\sigma} + d_{K,\sigma}\tau_{L,\sigma}}(\langle\Lambda\rangle_K - \langle\Lambda\rangle_L)\mathbf{n}_{K,\sigma}. \end{aligned}$$

With the same ideas as the ones used for the proof of Lemma 7 in [4] and the additional Hypotheses 2, the property (36) is satisfied with the choices (42).

The previous strategies can be generalized with the following reconstruction operator

$$I_\sigma u = \sum_{M \in \mathcal{I}_\sigma} \omega_{M,\sigma} u_M, \qquad \sum_{M \in \mathcal{I}_\sigma} \omega_{M,\sigma} = 1, \quad \omega_{M,\sigma} \geq 0, \tag{43}$$

with interpolation index set $\mathcal{I}_\sigma$. It is assumed that $\omega_{M,\sigma} = 0$ if $M \notin \mathcal{I}_\sigma$. In the next sections, two nonlinear schemes are derived by using the consistent flux approximations (37) with trace reconstruction operators (43).

4.3. Nonlinear Two-Point Flux Approximation

In this section, a nonlinear two-point flux approximation (NLTPFA) is derived, using concepts presented in [16, 17, 18, 21]. Inserting (37) into (21), using the reconstruction operator (43), reordering the terms and using the fact that $\sum\limits_{M \in \mathcal{I}_\sigma} \omega_M = 1$ yield:

$$F_{K,\sigma}(u,v) = t_{L,\sigma}(u)v_L - t_{K,\sigma}(u)v_K - \underbrace{(\mu_{L,\sigma}(u)\lambda_{L,\sigma}(v) - \mu_{K,\sigma}(u)\lambda_{K,\sigma}(v))}_{\stackrel{\text{def}}{=} R_{K,\sigma}(u,v)}, \tag{44}$$

with the transmissibilities

$$\begin{aligned} t_{K,\sigma}(u) &= \mathrm{m}_\sigma \left(\mu_{K,\sigma}(u) \sum_{\sigma' \in \mathcal{S}_{K,\sigma}} \sum_{M \in \{\mathcal{I}_{\sigma'} \setminus \{K\}\}} \alpha_{K,\sigma\sigma'}\omega_{M,\sigma'} + \mu_{L,\sigma}(u) \sum_{\sigma' \in \mathcal{S}_{L,\sigma}} \sum_{M \in \{\mathcal{I}_{\sigma'} \cap \{K\}\}} \alpha_{L,\sigma\sigma'}\omega_{M,\sigma'} \right), \\ t_{L,\sigma}(u) &= \mathrm{m}_\sigma \left(\mu_{L,\sigma}(u) \sum_{\sigma' \in \mathcal{S}_{L,\sigma}} \sum_{M \in \{\mathcal{I}_{\sigma'} \setminus \{L\}\}} \alpha_{L,\sigma\sigma'}\omega_{M,\sigma'} + \mu_{K,\sigma}(u) \sum_{\sigma' \in \mathcal{S}_{K,\sigma}} \sum_{M \in \{\mathcal{I}_{\sigma'} \cap \{L\}\}} \alpha_{K,\sigma\sigma'}\omega_{M,\sigma'} \right), \end{aligned} \tag{45}$$

and

$$\begin{aligned} \lambda_{K,\sigma}(v) &\stackrel{\text{def}}{=} \mathrm{m}_\sigma \sum_{\sigma' \in \mathcal{S}_{K,\sigma}} \sum_{M \in \{\mathcal{I}_{\sigma'} \setminus \{K,L\}\}} \alpha_{K,\sigma\sigma'}\omega_{M,\sigma'} v_M, \\ \lambda_{L,\sigma}(v) &\stackrel{\text{def}}{=} \mathrm{m}_\sigma \sum_{\sigma' \in \mathcal{S}_{L,\sigma}} \sum_{M \in \{\mathcal{I}_{\sigma'} \setminus \{K,L\}\}} \alpha_{L,\sigma\sigma'}\omega_{M,\sigma'} v_M. \end{aligned} \tag{46}$$

In order to obtain a nonlinear two-point flux approximation, the following weights are considered:

$$\begin{aligned}
&\mu_{K,\sigma}(u) = 0.5, \quad \mu_{L,\sigma}(u) = 0.5, \qquad \text{if } \lambda_{L,\sigma}(u) = \lambda_{K,\sigma}(u) = 0, \\
&\mu_{K,\sigma}(u) = \frac{|\lambda_{L,\sigma}(u)|}{|\lambda_{K,\sigma}(u)| + |\lambda_{L,\sigma}(u)|},\ \mu_{L,\sigma}(u) = \frac{|\lambda_{K,\sigma}(u)|}{|\lambda_{K,\sigma}(u)| + |\lambda_{L,\sigma}(u)|}, \qquad \text{otherwise.}
\end{aligned} \tag{47}$$

Therefore, from (44), the flux $F_{K,\sigma}(u,u)$ reads:

$$F_{K,\sigma}(u,u) = t_{L,\sigma}(u)u_L - t_{K,\sigma}(u)u_K - R_{K,\sigma}(u,u). \tag{48}$$

Under the assumption that $\lambda_{L,\sigma}(u)\lambda_{K,\sigma}(u) \geq 0$, it is inferred from (48) that:

$$F_{K,\sigma}(u,u) = t_{L,\sigma}(u)u_L - t_{K,\sigma}(u)u_K. \tag{49}$$

By virtue of (49), we thus get a nonlinear two-point flux approximation. However, to get the convergence of the finite volume scheme defined by the fluxes (48) using Corollary 1, the function $u \mapsto F_{K,\sigma}(u,u)$ must be continuous, which is not a priori the case. The problem comes from the definition (47) of the function $u \mapsto \mu_{K,\sigma}(u)$ for which discontinuities can appear. Thus, in order to guarantee the continuity of the function $u \mapsto F_{K,\sigma}(u,u)$, we finally choose the weights as:

$$\mu_{K,\sigma}(u) = \frac{|\lambda_{L,\sigma}(u)| + \epsilon}{|\lambda_{K,\sigma}(u)| + |\lambda_{L,\sigma}(u)| + 2\epsilon}, \quad \mu_{L,\sigma}(u) = \frac{|\lambda_{K,\sigma}(u)| + \epsilon}{|\lambda_{K,\sigma}(u)| + |\lambda_{L,\sigma}(u)| + 2\epsilon}, \tag{50}$$

with $\epsilon > 0$ such that $0 < \epsilon \leq h_{\mathcal{D}} \min_{\sigma\in\mathcal{E}} \mathrm{m}_\sigma$. Thus, the convergence of the finite volume scheme defined by the fluxes (48) with weights (50) is obtained thanks to Corollary 1.

Let us now discuss the monotonicity of the finite volume scheme defined by the fluxes (48). First, we observe that, under some conditions, we can rewrite the flux $F_{K,\sigma}(u,u)$ given by the expression (48) to obtain a nonlinear two-point flux approximation. Indeed,

- if we have
$$R_{K,\sigma}(u,u) = 0, \tag{51}$$
then the flux $F_{K,\sigma}(u,u)$ given by (48) becomes:
$$F_{K,\sigma}(u,u) = t_{L,\sigma}(u)u_L - t_{K,\sigma}(u)u_K;$$
- if we have
$$R_{K,\sigma}(u,u) > 0 \text{ and } u_K \neq 0, \tag{52}$$
then the flux $F_{K,\sigma}(u,u)$ given by (48) can be rewritten as:
$$F_{K,\sigma}(u,u) = t_{L,\sigma}(u)u_L - \left(t_{K,\sigma}(u) + \frac{R_{K,\sigma}(u,u)}{u_K}\right)u_K.$$
- if we have
$$R_{K,\sigma}(u,u) < 0 \text{ and } u_L \neq 0, \tag{53}$$

then the flux $F_{K,\sigma}(u,u)$ given by (48) can be rewritten as:

$$F_{K,\sigma}(u,u) = \left(t_{L,\sigma}(u) - \frac{R_{K,\sigma}(u,u)}{u_L} \right) u_L - t_{K,\sigma}(u) u_K.$$

Furthermore, under the assumption that

$$\lambda_{L,\sigma}(u)\lambda_{K,\sigma}(u) \geq 0, \tag{54}$$

the flux $F_{K,\sigma}(u,u)$ defined by (48) with weights (50) can be rewritten as:

$$F_{K,\sigma}(u,u) = t_{L,\sigma}(u)u_L - t_{K,\sigma}(u)u_K - \epsilon \underbrace{\frac{\lambda_{L,\sigma}(u) - \lambda_{K,\sigma}(u)}{|\lambda_{K,\sigma}(u)| + |\lambda_{L,\sigma}(u)| + 2\epsilon}}_{\stackrel{\text{def}}{=} \mathfrak{E}_{K,\sigma}(u)}, \tag{55}$$

where we observe that

$$|\mathfrak{E}_{K,\sigma}(u)| \leq \epsilon. \tag{56}$$

Thus, thanks to Equation (55) and inequality (56), it is inferred that under the assumption that $\lambda_{L,\sigma}(u)\lambda_{K,\sigma}(u) \geq 0$, the flux $F_{K,\sigma}(u,u)$ defined by (48) with weights (50) is close to a nonlinear two-point flux approximation provided that ϵ is sufficiently small.

Thus, for the monotonicity property of the scheme, we get the following result: Provided that for all $\sigma \in \mathcal{E}_{\text{int}}$ with $\mathcal{T}_\sigma = \{K, L\}$, one of these four conditions (51),(52),(53) or (54) holds, and the values u_K as well as the α_K and ω_K coefficients are nonnegative, then the resulting discretization matrix is an M-matrix (for sufficiently small ϵ for the case that the condition (54) is used).

If in addition to that, the source term f is nonnegative, the positivity-preservation of the scheme using a Picard method can be proven (see [17]).

4.4. Nonlinear Multi-Point Flux Approximation

In this section, we mainly follow ideas presented in [19, 24, 25]. For the derivation of a nonlinear multi-point flux approximation (NLMPFA), the fluxes (37) are split as follows

$$\begin{aligned} \tilde{F}_{K,\sigma}(v) &\stackrel{\text{def}}{=} \tilde{F}^{(1)}_{K,\sigma}(v) + \tilde{F}^{(2)}_{K,\sigma}(v), \\ \tilde{F}_{L,\sigma}(v) &\stackrel{\text{def}}{=} \tilde{F}^{(1)}_{L,\sigma}(v) + \tilde{F}^{(2)}_{L,\sigma}(v), \end{aligned} \tag{57}$$

with

$$\begin{aligned} \tilde{F}^{(1)}_{K,\sigma}(v) &= \mathrm{m}_\sigma\, \alpha_{K,\sigma\sigma} \omega_{L,\sigma}(v_L - v_K), \\ \tilde{F}^{(1)}_{L,\sigma}(v) &= \mathrm{m}_\sigma\, \alpha_{L,\sigma\sigma} \omega_{K,\sigma}(v_K - v_L), \\ \tilde{F}^{(2)}_{K,\sigma}(v) &= \mathrm{m}_\sigma\, \alpha_{K,\sigma\sigma} \sum_{M \in \{\mathcal{I}_\sigma \setminus \{L\}\}} \omega_{M,\sigma}(v_M - v_K) + \sum_{\sigma' \in \{\mathcal{S}_{K,\sigma} \setminus \{\sigma\}\}} \mathrm{m}_\sigma\, \alpha_{K,\sigma\sigma'}(I_{\sigma'} v - v_K), \\ \tilde{F}^{(2)}_{L,\sigma}(v) &= \mathrm{m}_\sigma\, \alpha_{L,\sigma\sigma} \sum_{M \in \{\mathcal{I}_\sigma \setminus \{K\}\}} \omega_{M,\sigma}(v_M - v_L) + \sum_{\sigma' \in \{\mathcal{S}_{L,\sigma} \setminus \{\sigma\}\}} \mathrm{m}_\sigma\, \alpha_{L,\sigma\sigma'}(I_{\sigma'} v - v_L). \end{aligned}$$

The weights are chosen as

$$
\begin{aligned}
\mu_{K,\sigma} = \mu_{L,\sigma} = 0.5, \qquad \text{if } \tilde{F}^{(2)}_{K,\sigma} = \tilde{F}^{(2)}_{L,\sigma} = 0,\\
\mu_{K,\sigma} = \frac{|\tilde{F}^{(2)}_{L,\sigma}|}{|\tilde{F}^{(2)}_{K,\sigma}| + |\tilde{F}^{(2)}_{L,\sigma}|}, \quad \mu_{L,\sigma} = \frac{|\tilde{F}^{(2)}_{K,\sigma}|}{|\tilde{F}^{(2)}_{K,\sigma}| + |\tilde{F}^{(2)}_{L,\sigma}|}, \qquad \text{otherwise.}
\end{aligned} \tag{58}
$$

This choice results in the final flux approximations

$$
\begin{aligned}
F_{K,\sigma}(u,u) &= \mu_{K,\sigma}(u)\tilde{F}^{(1)}_{K,\sigma}(u) - \mu_{L,\sigma}(u)\tilde{F}^{(1)}_{L,\sigma}(u) + \mu_{K,\sigma}(u)\left(1 - \operatorname{sign}\left(\tilde{F}^{(2)}_{K,\sigma}(u)\tilde{F}^{(2)}_{L,\sigma}(u)\right)\right)\tilde{F}^{(2)}_{K,\sigma}(u),\\
F_{L,\sigma}(u,u) &= \mu_{L,\sigma}(u)\tilde{F}^{(1)}_{L,\sigma}(u) - \mu_{K,\sigma}(u)\tilde{F}^{(1)}_{K,\sigma}(u) + \mu_{L,\sigma}(u)\left(1 - \operatorname{sign}\left(\tilde{F}^{(2)}_{K,\sigma}(u)\tilde{F}^{(2)}_{L,\sigma}(u)\right)\right)\tilde{F}^{(2)}_{L,\sigma}(u),
\end{aligned} \tag{59}
$$

where the flux conservation $F_{K,\sigma}(u,u) + F_{L,\sigma}(u,u) = 0$ is obtained. Under the assumption of nonnegative coefficients $\omega_{M,\sigma}, \alpha_{K,\sigma\sigma'}$, discrete extremum principles can be proven for this scheme (see for instance [19, 20]).

Again, the function $u \mapsto F_{K,\sigma}(u,u)$ defined by (59) is not a priori continuous when $\tilde{F}^{(2)}_{K,\sigma}(u) = \tilde{F}^{(2)}_{L,\sigma}(u) = 0$. To guarantee the continuity, a splitting of the factors $\alpha_{K,\sigma\sigma}\omega_{L,\sigma}$ and $\alpha_{L,\sigma\sigma}\omega_{K,\sigma}$ is carried out in the following way

$$
\begin{aligned}
\alpha_{K,\sigma\sigma}\omega_{L,\sigma} &= \beta_\sigma + (\alpha_{K,\sigma\sigma}\omega_{L,\sigma} - \beta_\sigma),\\
\alpha_{L,\sigma\sigma}\omega_{K,\sigma} &= \beta_\sigma + (\alpha_{L,\sigma\sigma}\omega_{K,\sigma} - \beta_\sigma),
\end{aligned}
$$

with $\beta_\sigma = \min(\alpha_{K,\sigma\sigma}\omega_{L,\sigma}, \alpha_{L,\sigma\sigma}\omega_{K,\sigma})$. Thus, the fluxes $\tilde{F}_{K,\sigma}(u)$, $\tilde{F}_{L,\sigma}(u)$ from (57) are rewritten as follows

$$
\begin{aligned}
\tilde{F}_{K,\sigma}(v) &\stackrel{\text{def}}{=} \tilde{F}^{(1)}_{K,\sigma}(v) + \tilde{F}^{(2)}_{K,\sigma}(v),\\
\tilde{F}_{L,\sigma}(v) &\stackrel{\text{def}}{=} \tilde{F}^{(1)}_{L,\sigma}(v) + \tilde{F}^{(2)}_{L,\sigma}(v),
\end{aligned} \tag{60}
$$

with

$$
\begin{aligned}
\tilde{F}^{(1)}_{K,\sigma}(v) &= \mathrm{m}_\sigma\, \beta_\sigma (v_L - v_K),\\
\tilde{F}^{(1)}_{L,\sigma}(v) &= -\tilde{F}^{(1)}_{K,\sigma}(v),\\
\tilde{F}^{(2)}_{K,\sigma}(v) &= \mathrm{m}_\sigma(\alpha_{K,\sigma\sigma}\omega_{L,\sigma} - \beta_\sigma)(v_L - v_K) + \mathrm{m}_\sigma\, \alpha_{K,\sigma\sigma} \sum_{M \in \{\mathcal{I}_\sigma \setminus \{L\}\}} \omega_{M,\sigma}(v_M - v_K)\\
&\quad + \sum_{\sigma' \in \{\mathcal{S}_{K,\sigma} \setminus \{\sigma\}\}} \mathrm{m}_\sigma\, \alpha_{K,\sigma\sigma'}(I_{\sigma'} v - v_K),\\
\tilde{F}^{(2)}_{L,\sigma}(v) &= \mathrm{m}_\sigma(\alpha_{L,\sigma\sigma}\omega_{K,\sigma} - \beta_\sigma)(v_K - v_L) + \mathrm{m}_\sigma\, \alpha_{L,\sigma\sigma} \sum_{M \in \{\mathcal{I}_\sigma \setminus \{K\}\}} \omega_{M,\sigma}(v_M - v_L)\\
&\quad + \sum_{\sigma' \in \{\mathcal{S}_{L,\sigma} \setminus \{\sigma\}\}} \mathrm{m}_\sigma\, \alpha_{L,\sigma\sigma'}(I_{\sigma'} v - v_L).
\end{aligned}
$$

The weights, $\mu_{K,\sigma}$ and $\mu_{L,\sigma}$, and the fluxes, $F_{K,\sigma}(u,u)$ and $F_{L,\sigma}(u,u)$, are still defined by (58) and (59), respectively. Now, let us consider the case where $\tilde{F}^{(2)}_{K,\sigma}(u) = \tilde{F}^{(2)}_{L,\sigma}(u) = 0$ for which the functions $\mu_{K,\sigma}$ and $\mu_{L,\sigma}$ are not continuous. However, since $\tilde{F}^{(1)}_{L,\sigma}(v) = -\tilde{F}^{(1)}_{K,\sigma}(v)$, the final flux does not depend on these functions. In fact,

$$\begin{aligned} F_{K,\sigma}(u,u) &= \mu_{K,\sigma}(u)\tilde{F}^{(1)}_{K,\sigma}(u) - \mu_{L,\sigma}(u)\tilde{F}^{(1)}_{L,\sigma}(u) \\ &= (\mu_{K,\sigma}(u) + \mu_{L,\sigma}(u))\tilde{F}^{(1)}_{K,\sigma}(u) \\ &= \tilde{F}^{(1)}_{K,\sigma}(u), \end{aligned}$$

which means that for all $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_K$, the function $u \mapsto F_{K,\sigma}(u,u)$ is continuous on $H_{\mathcal{T}}(\Omega)$. The above flux splitting only makes sense if the coefficients $\alpha_{K,\sigma\sigma}, \alpha_{L,\sigma\sigma}$ are positive. This is done by adding the constraints

$$\alpha_{K,\sigma\sigma} \geq \delta_\alpha, \qquad \alpha_{L,\sigma\sigma} \geq \delta_\alpha, \tag{61}$$

to the optimization problem (22). Thus, the convergence of this scheme is obtained thanks to Corollary 1.

5. Numerical results

In this section, the behavior of the above mentioned nonlinear finite volume schemes is investigated and compared to linear schemes. The NLTPFA scheme is given by equation (48) with weights (50), the NLMPFA scheme by equation (59), (60), the weights (58) and the additional constraints (61) for the conormal decomposition. The scheme with fluxes (37) and constant weights $\mu_{K,\sigma} = \mu_{L,\sigma} = 0.5$, which results in a linear scheme, is denoted as AvgMPFA . In Section 5.1, the convergence behavior of these schemes is analyzed for a mildly and highly anisotropic test case. In Sections 5.2 and 5.3, we compare these schemes to the Box method [28, 29] that uses finite-element basis functions on each cell to calculate fluxes over sub-volume faces. Further, in Section 5.2 discrete extremum principles are investigated and in Section 5.3 benchmark test cases are considered. So far, the reconstruction operator I_σ has not been specified. From now on, the harmonic averaging interpolator (42) is used.

For measuring the coercivity of the scheme, the following estimate is defined

$$e_{\mathcal{T}}(u,v) \stackrel{\text{def}}{=} \frac{a_{\mathcal{T}}(u,v,v)}{\|v\|_{\mathcal{T}}}. \tag{62}$$

The impact of the term $R_{K,\sigma}(u,v)$ in the NLTPFA expression is quantified with

$$e_R(u,v) \stackrel{\text{def}}{=} \max_{K\in\mathcal{T},\sigma\in\mathcal{E}_K} |R_{K,\sigma}(u,v)|. \tag{63}$$

For simplicity, we define $e_{\mathcal{T},n} \stackrel{\text{def}}{=} e_{\mathcal{T}}(u_n,u_n)$, $\overline{e}_{\mathcal{T},n} \stackrel{\text{def}}{=} e_{\mathcal{T}}(u_n,u_n-\overline{u})$, and analogously $e_{R,n}, \overline{e}_{R,n}$.

All simulations are performed using the open-source simulator DuMu$^\text{x}$ [30], which comes in the form of an additional *DUNE* module [31]. Newton's method is used for solving the occuring nonlinear systems of equation. The nonlinear iteration loop is stopped if the absolute residual is below 10^{-5}. The optimization problem (22) is solved using a *Primal-Dual Simplex Method* provided by the open-source library *GNU Linear Programming Kit*[1] (GLPK).

5.1. Convergence rates

Within this section, the computational domain is chosen as $\Omega = [0,1]^2$. Furthermore, Dirichlet conditions are set on the whole boundary consistent with the exact solution. The grids that are used to analyze the convergence behavior of the schemes are shown in Figure 2. These meshes are refined such that the pattern remains unaffected.

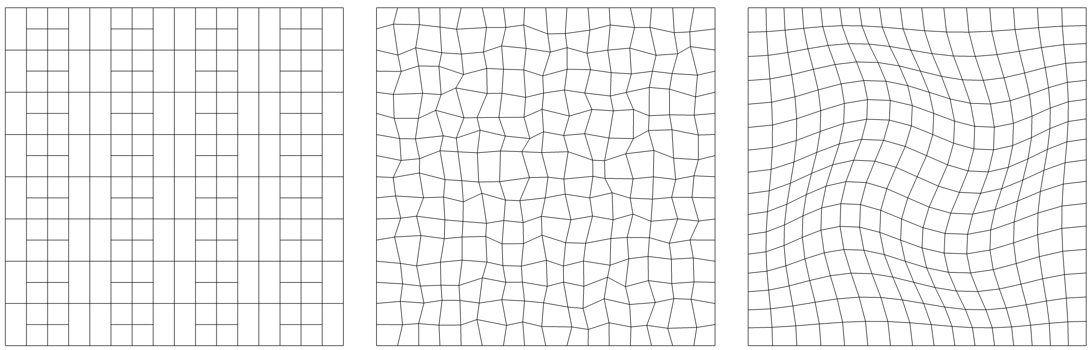


Figure 2: Grids used for the convergence tests. From left to right: non-matching, randomly distorted and twisted grid.

The first test case analyzes the convergence rates for a homogeneous mildly anisotropic tensor

$$\Lambda = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}, \tag{64}$$

with the exact solution $\overline{u}(x,y) = 1 + \sin(\pi x)\sin(\pi y)$ and the corresponding source term as $f = -\nabla\cdot(\Lambda\ \nabla\overline{u})$.

Table 1–3 list the error norms for the NLTPFA, NLMPFA and AvgMPFA schemes. It is observed that all schemes converge approximately with second order in the L^2-norm and at least first order in the H^1-norm. Furthermore, the coercivity estimates $e_{\mathcal{T},n}$, $\overline{e}_{\mathcal{T},n}$ seem to be bounded. The number of Newton iterations are quite small for the NLTPFA scheme. The Newton method converges within three iterations, whereas the NLMPFA method needs approximately $3-6$ iterations.

In the next example, the tensor is changed to investigate the behavior for high anisotropy ratios.

$$\Lambda(x,y) = \frac{1}{x^2+y^2}\begin{pmatrix} \beta x^2 + y^2 & (\beta-1)xy \\ (\beta-1)xy & x^2+\beta y^2 \end{pmatrix}, \tag{65}$$

[1] http://www.gnu.org/software/glpk/glpk.html

Table 1: Discrete error norms, convergence rates (cr) and number of nonlinear iterations (nIt) for the mild anisotropic test case on non-matching grids.

scheme	n	$\|u_n - \overline{u}\|_{L^2}$	cr	$\|u_n - \overline{u}\|_{\mathcal{T}}$	cr	$e_{\mathcal{T},n}$	$\overline{e}_{\mathcal{T},n}$	nIt	$h_{\mathcal{D}}$
	1	1.90e-02	0.00	1.45e-01	0.00	2.44	1.09	3	5.59e-01
	2	6.50e-03	1.54	8.42e-02	0.78	2.46	1.06	3	2.80e-01
	3	1.78e-03	1.87	4.28e-02	0.97	2.46	1.03	3	1.40e-01
NLTPFA	4	4.55e-04	1.97	2.13e-02	1.01	2.47	1.01	3	6.99e-02
	5	1.14e-04	2.00	1.05e-02	1.02	2.47	1.00	2	3.49e-02
	6	2.84e-05	2.00	5.18e-03	1.02	2.47	1.00	2	1.75e-02
	7	7.08e-06	2.00	2.57e-03	1.01	2.47	1.00	2	8.73e-03
	1	2.53e-02	0.00	2.06e-01	0.00	2.42	1.11	4	5.59e-01
	2	8.57e-03	1.56	1.26e-01	0.71	2.47	1.12	5	2.80e-01
	3	2.09e-03	2.04	5.55e-02	1.18	2.47	1.07	5	1.40e-01
NLMPFA	4	4.95e-04	2.08	2.47e-02	1.17	2.47	1.04	5	6.99e-02
	5	1.19e-04	2.06	1.14e-02	1.12	2.47	1.02	4	3.49e-02
	6	2.90e-05	2.03	5.41e-03	1.07	2.47	1.01	5	1.75e-02
	7	7.16e-06	2.02	2.63e-03	1.04	2.47	1.01	4	8.73e-03
	1	1.80e-02	0.00	1.37e-01	0.00	2.45	1.08	1	5.59e-01
	2	6.43e-03	1.49	8.19e-02	0.74	2.46	1.06	1	2.80e-01
	3	1.76e-03	1.87	4.14e-02	0.99	2.47	1.02	1	1.40e-01
AvgMPFA	4	4.50e-04	1.96	2.07e-02	1.00	2.47	1.01	1	6.99e-02
	5	1.13e-04	1.99	1.03e-02	1.01	2.47	1.00	1	3.49e-02
	6	2.83e-05	2.00	5.13e-03	1.00	2.47	1.00	1	1.75e-02
	7	7.07e-06	2.00	2.56e-03	1.00	2.47	1.00	1	8.73e-03

with $\beta = 10^{-3}$. The exact solution is the same than in the previous example. The anisotropy ratio is given as $\frac{1}{\beta}$. The integrated source term and the averaged tensor $\langle \Lambda \rangle_K$ are calculated using a fifth-order quadrature rule. For this test case, faces exist where the conormal cannot be decomposed with only positive coefficients. Negative coefficients especially occur on the randomly distorted grid. Therefore, the calculation of $e_{R,n}$, $\overline{e}_{R,n}$ is included. Please note that these values are rounded to the eighth decimal place.

Table 4–6 list the error norms of the NLTPFA, NLMPFA and AvgMPFA schemes for the high anisotropy test case. It is observed that all schemes converge approximately with order $1.5-2.0$ in the L^2-norm and order $0.7-2.0$ in the H^1-norm. Furthermore, the coercivity estimates $e_{\mathcal{T},n}$, $\overline{e}_{\mathcal{T},n}$ seem to be bounded. However, the behavior of $\overline{e}_{\mathcal{T},n}$ is unclear for the non-matching grid. The number of Newton iterations are again quite small for the NLTPFA scheme. The Newton method

Table 2: Discrete error norms, convergence rates (cr) and number of nonlinear iterations (nIt) for the mild anisotropic test case on randomly distorted grids.

scheme	n	$\|u_n - \overline{u}\|_{L^2}$	cr	$\|u_n - \overline{u}\|_{\mathcal{T}}$	cr	$e_{\mathcal{T},n}$	$\overline{e}_{\mathcal{T},n}$	nIt	$h_{\mathcal{D}}$
	1	2.26e-02	0.00	1.71e-01	0.00	2.77	0.97	3	4.18e-01
	2	7.27e-03	2.02	8.88e-02	1.17	2.52	1.08	3	2.38e-01
	3	2.10e-03	1.77	3.61e-02	1.28	2.54	1.06	2	1.18e-01
NLTPFA	4	6.12e-04	2.04	1.77e-02	1.17	2.51	1.07	2	6.46e-02
	5	1.59e-04	1.96	9.10e-03	0.97	2.51	1.05	2	3.25e-02
	6	4.05e-05	1.97	4.52e-03	1.01	2.50	1.07	2	1.63e-02
	7	1.08e-05	1.93	2.27e-03	1.01	2.50	1.07	2	8.18e-03
	1	3.14e-02	0.00	2.53e-01	0.00	2.74	0.98	4	4.18e-01
	2	8.05e-03	2.42	1.03e-01	1.60	2.49	1.01	5	2.38e-01
	3	2.21e-03	1.84	4.53e-02	1.17	2.52	1.00	6	1.18e-01
NLMPFA	4	1.10e-03	1.15	2.16e-02	1.23	2.50	1.01	6	6.46e-02
	5	2.95e-04	1.92	1.01e-02	1.10	2.50	1.03	6	3.25e-02
	6	8.36e-05	1.82	4.80e-03	1.08	2.50	1.05	6	1.63e-02
	7	2.23e-05	1.92	2.32e-03	1.06	2.50	1.06	6	8.18e-03
	1	2.69e-02	0.00	1.93e-01	0.00	2.82	0.95	1	4.18e-01
	2	8.84e-03	1.98	9.16e-02	1.32	2.54	1.03	1	2.38e-01
	3	2.45e-03	1.83	3.58e-02	1.34	2.54	1.03	1	1.18e-01
AvgMPFA	4	6.92e-04	2.10	1.74e-02	1.19	2.51	1.06	1	6.46e-02
	5	1.73e-04	2.01	8.86e-03	0.98	2.51	1.06	1	3.25e-02
	6	4.41e-05	1.98	4.40e-03	1.01	2.50	1.07	1	1.63e-02
	7	1.17e-05	1.93	2.21e-03	1.01	2.50	1.07	1	8.18e-03

converges within three iterations, whereas, the NLMPFA needs more iterations. In particular for the randomly distorted grid, the number of Newton iterations increases with grid refinement for the NLMPFA scheme. Moreover, the estimates $e_{R,n}$, $\overline{e}_{R,n}$ are quite small and bounded, such that this term is in $\mathcal{O}(1)$.

In the last examples, it has been observed that the convergence behavior of the NLTPFA, NLMPFA and AvgMPFA schemes is quite similar. Furthermore, the schemes seem to be coercive for these test cases. The main drawback of the NLMPFA scheme is the fact that it requires more Newton iterations, and that the number of iterations partly depends on the discretization length $h_{\mathcal{D}}$.

Table 3: Discrete error norms, convergence rates (cr) and number of nonlinear iterations (nIt) for the mild anisotropic test case on twisted grids.

scheme	n	$\Vert u_n - \overline{u}\Vert_{L^2}$	cr	$\Vert u_n - \overline{u}\Vert_{\mathcal{T}}$	cr	$e_{\mathcal{T},n}$	$\overline{e}_{\mathcal{T},n}$	nIt	$h_{\mathcal{D}}$
NLTPFA	1	1.70e-02	0.00	1.32e-01	0.00	2.74	0.95	3	4.26e-01
	2	8.21e-03	1.24	8.14e-02	0.82	2.57	0.97	3	2.37e-01
	3	3.03e-03	1.46	3.11e-02	1.41	2.49	0.76	2	1.20e-01
	4	8.95e-04	1.79	9.51e-03	1.73	2.46	0.64	2	6.06e-02
	5	2.38e-04	1.92	2.57e-03	1.89	2.45	0.59	2	3.04e-02
	6	6.10e-05	1.97	6.57e-04	1.97	2.45	0.57	2	1.52e-02
	7	1.54e-05	1.99	1.65e-04	1.99	2.45	0.57	2	7.60e-03
NLMPFA	1	2.31e-02	0.00	1.93e-01	0.00	2.66	1.02	3	4.26e-01
	2	6.88e-03	2.07	8.48e-02	1.40	2.52	1.00	5	2.37e-01
	3	4.83e-03	0.52	5.88e-02	0.54	2.50	0.70	5	1.20e-01
	4	1.63e-03	1.59	2.74e-02	1.12	2.47	0.57	5	6.06e-02
	5	4.35e-04	1.91	9.87e-03	1.48	2.45	0.55	5	3.04e-02
	6	1.12e-04	1.96	3.34e-03	1.56	2.45	0.60	5	1.52e-02
	7	2.97e-05	1.92	1.12e-03	1.58	2.45	0.66	5	7.60e-03
AvgMPFA	1	2.05e-02	0.00	1.43e-01	0.00	2.78	0.91	1	4.26e-01
	2	9.94e-03	1.23	8.66e-02	0.86	2.59	0.91	1	2.37e-01
	3	3.78e-03	1.42	3.56e-02	1.31	2.50	0.69	1	1.20e-01
	4	1.13e-03	1.77	1.15e-02	1.66	2.46	0.56	1	6.06e-02
	5	3.00e-04	1.91	3.17e-03	1.86	2.45	0.51	1	3.04e-02
	6	7.65e-05	1.97	8.15e-04	1.96	2.45	0.49	1	1.52e-02
	7	1.93e-05	1.99	2.05e-04	1.99	2.45	0.49	1	7.60e-03

5.2. Discrete extremum principles

The following two examples investigate whether the schemes satisfy discrete extremum principles. In the first example, the tensor (65) is again considered. The boundary conditions are $u = 0$ on $\partial\Omega$ and $\Omega = [0, 1]^2$ is discretized with a regular cartesian grid. The source term is $f = 10$ in $(0.5, 1)^2$ and $f = 0$ elsewhere. The weak solution of this test problem is positive within the domain, because of the non-negativity of the source term and the chosen boundary conditions. Figure 3 shows the numerical results of the Box, AvgMPFA, NLTPFA and NLMPFA schemes. It can be seen that the linear schemes produce unphysical negative solution values, whereas the undershoots produced by the nonlinear schemes are in the range of the solver tolerance.

The next example investigates another test case without a source term. The domain and the grid are shown in Figure 4, with an inner and an outer boundary. The Dirichlet values $u = 10^5$ and

Table 4: Discrete error norms, convergence rates (cr) and number of nonlinear iterations (nIt) for the high anisotropy test case on non-matching grids.

scheme	n	$\|u_n - \overline{u}\|_{L^2}$	cr	$\|u_n - \overline{u}\|_{\mathcal{T}}$	cr	$e_{\mathcal{T},n}$	$\overline{e}_{\mathcal{T},n}$	$e_{R,n}$	$\overline{e}_{R,n}$	nIt
	1	5.99e-02	0.00	5.20e-01	0.00	1.01	0.51	0	2.34e-02	3
	2	1.76e-02	1.76	2.62e-01	0.99	0.70	0.40	0	1.23e-02	3
	3	6.45e-03	1.45	1.61e-01	0.70	0.40	0.28	0	4.23e-03	3
NLTPFA	4	2.35e-03	1.46	1.10e-01	0.55	0.28	0.18	0	1.44e-03	3
	5	8.00e-04	1.55	7.35e-02	0.59	0.24	0.11	0	4.45e-04	3
	6	2.56e-04	1.64	4.68e-02	0.65	0.23	0.08	0	1.26e-04	3
	7	7.81e-05	1.71	2.84e-02	0.72	0.23	0.05	0	3.30e-05	3
	1	6.81e-02	0.00	5.98e-01	0.00	1.01	0.49	0	0	6
	2	2.05e-02	1.74	3.24e-01	0.88	0.71	0.41	0	0	6
	3	6.78e-03	1.59	1.80e-01	0.85	0.41	0.29	0	0	10
NLMPFA	4	2.41e-03	1.49	1.17e-01	0.63	0.28	0.18	0	0	12
	5	8.20e-04	1.56	7.65e-02	0.61	0.24	0.11	0	0	9
	6	2.62e-04	1.64	4.82e-02	0.66	0.23	0.08	0	0	13
	7	7.95e-05	1.72	2.91e-02	0.73	0.23	0.05	0	0	18
	1	5.70e-02	0.00	4.94e-01	0.00	1.00	0.51	0	0	1
	2	1.67e-02	1.77	2.52e-01	0.97	0.70	0.42	0	0	1
	3	6.23e-03	1.42	1.59e-01	0.67	0.40	0.28	0	0	1
AvgMPFA	4	2.34e-03	1.41	1.11e-01	0.51	0.28	0.17	0	0	1
	5	8.10e-04	1.53	7.51e-02	0.57	0.24	0.11	0	0	1
	6	2.61e-04	1.63	4.78e-02	0.65	0.23	0.07	0	0	1
	7	7.93e-05	1.72	2.90e-02	0.72	0.23	0.05	0	0	1

$u = 0$ are set at the inner and outer boundaries, respectively. Therefore, the solution is expected to be within these bounds.

Figure 5 shows the numerical solutions of the Box, AvgMPFA, NLTPFA and NLMPFA schemes on a three times refined grid. All schemes fulfill the maximum principle, whereas the minimum principle is violated by the linear schemes. The undershoots of the AvgMPFA scheme are above 4% and those of the Box scheme above 2%.

The small negative undershoots of the nonlinear schemes are caused by Newton's method. These undershoots can be prevented by using other nonlinear solvers such as Picard's method or enhanced solvers [32].

The above test cases exhibit how nonlinear schemes are capable to reproduce physical solutions, whereas linear schemes can produce negative values. When solving highly complex partial

Table 5: Discrete error norms, convergence rates (cr) and number of nonlinear iterations (nIt) for the high anisotropy test case on randomly distorted grids.

scheme	n	$\|u_n - \overline{u}\|_{L^2}$	cr	$\|u_n - \overline{u}\|_{\mathcal{T}}$	cr	$e_{\mathcal{T},n}$	$\overline{e}_{\mathcal{T},n}$	$e_{R,n}$	$\overline{e}_{R,n}$	nIt
NLTPFA	1	7.26e-02	0.00	5.88e-01	0.00	1.36	0.52	0	5.61e-02	3
	2	2.97e-02	1.59	4.19e-01	0.60	1.37	0.37	0.45	4.90e-02	3
	3	8.66e-03	1.76	2.03e-01	1.03	1.46	0.49	0	8.06e-03	2
	4	9.37e-03	-0.13	3.55e-01	-0.93	1.40	0.27	0.79	1.77e-02	2
	5	3.63e-03	1.38	2.59e-01	0.46	1.42	0.22	1.15	1.39e-02	2
	6	1.12e-03	1.69	1.34e-01	0.95	1.44	0.30	0.84	5.23e-03	2
	7	2.83e-04	2.01	6.81e-02	0.99	1.44	0.30	1.25	1.71e-03	2
NLMPFA	1	9.87e-02	0.00	7.63e-01	0.00	1.29	0.47	0	0	5
	2	6.07e-02	0.86	7.61e-01	0.00	1.19	0.21	0	0	7
	3	1.62e-02	1.88	2.49e-01	1.59	1.40	0.37	0	0	9
	4	2.84e-02	-0.93	6.24e-01	-1.52	1.27	0.22	0	0	16
	5	1.09e-02	1.40	3.83e-01	0.71	1.37	0.18	0	0	18
	6	3.83e-03	1.50	1.80e-01	1.09	1.42	0.24	0	0	24
	7	1.30e-03	1.58	8.30e-02	1.13	1.43	0.26	0	0	54
AvgMPFA	1	6.60e-02	0.00	5.22e-01	0.00	1.58	0.60	0	0	1
	2	3.13e-02	1.33	3.97e-01	0.49	1.42	0.55	0	0	1
	3	1.38e-02	1.16	2.57e-01	0.62	1.49	0.91	0	0	1
	4	9.14e-03	0.69	3.35e-01	-0.44	1.43	1.24	0	0	1
	5	3.32e-03	1.47	2.31e-01	0.54	1.44	0.49	0	0	1
	6	1.36e-03	1.29	1.42e-01	0.70	1.44	0.77	0	0	1
	7	3.93e-04	1.81	7.17e-02	1.00	1.44	0.38	0	0	1

differential equations, where secondary variables non-linearly depend on primary variables, such negative values can strongly influence the efficiency of the scheme, in terms of linear and nonlinear solver convergence.

5.3. Benchmark examples

In this last section, three-dimensional benchmark test cases are considered. The first example investigates the linearity-preservation property of the schemes. The considered domain and the grid are shown in Figure 6 (right). The domain consists of two sub-domains Ω_1 and Ω_2. The

Table 6: Discrete error norms, convergence rates (cr) and number of nonlinear iterations (nIt) for the high anisotropy test case on twisted grids.

scheme	n	$\|u_n - \overline{u}\|_{L^2}$	cr	$\|u_n - \overline{u}\|_{\mathcal{T}}$	cr	$e_{\mathcal{T},n}$	$\overline{e}_{\mathcal{T},n}$	$e_{R,n}$	$\overline{e}_{R,n}$	nIt
	1	5.15e-02	0.00	4.14e-01	0.00	1.59	0.52	0	4.66e-02	3
	2	1.87e-02	1.73	2.29e-01	1.01	1.44	0.43	0.23	2.67e-02	3
	3	1.42e-02	0.41	2.30e-01	-0.01	1.40	0.32	0	1.30e-02	3
NLTPFA	4	6.65e-03	1.11	1.20e-01	0.95	1.41	0.33	0	3.23e-03	2
	5	2.20e-03	1.60	4.25e-02	1.50	1.41	0.33	0	6.69e-04	2
	6	6.08e-04	1.85	1.21e-02	1.81	1.41	0.33	0	1.20e-04	2
	7	1.57e-04	1.95	3.21e-03	1.92	1.41	0.32	0	1.67e-05	2
	1	7.36e-02	0.00	5.44e-01	0.00	1.59	0.48	0	0	6
	2	3.27e-02	1.38	4.09e-01	0.48	1.39	0.31	0	0	6
	3	2.69e-02	0.29	4.19e-01	-0.04	1.36	0.24	0	0	7
NLMPFA	4	1.49e-02	0.87	2.55e-01	0.73	1.38	0.23	0	0	14
	5	6.04e-03	1.31	1.08e-01	1.25	1.40	0.23	0	0	14
	6	2.25e-03	1.43	4.49e-02	1.26	1.41	0.19	0	0	9
	7	8.20e-04	1.46	2.04e-02	1.14	1.41	0.14	0	0	13
	1	5.79e-02	0.00	4.53e-01	0.00	1.64	0.55	0	0	1
	2	1.71e-02	2.09	2.22e-01	1.21	1.44	0.42	0	0	1
	3	1.04e-02	0.72	2.02e-01	0.14	1.43	0.56	0	0	1
AvgMPFA	4	4.37e-03	1.28	1.01e-01	1.02	1.41	0.46	0	0	1
	5	1.51e-03	1.54	3.49e-02	1.54	1.41	0.38	0	0	1
	6	4.36e-04	1.80	1.03e-02	1.77	1.41	0.36	0	0	1
	7	1.14e-04	1.93	2.78e-03	1.89	1.41	0.33	0	0	1

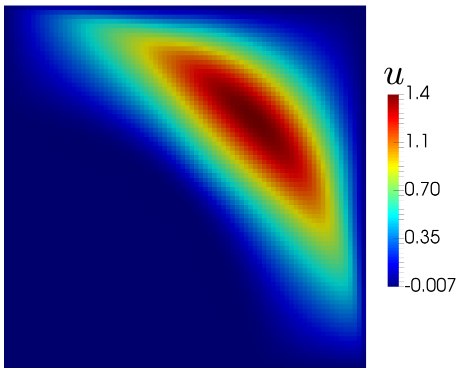
transition from Ω_1 to Ω_2 is located at $x = 0.6$, and the permeability tensors are chosen as

$$\Lambda_1 = \begin{pmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \Lambda_2 = \begin{pmatrix} 10 & 3 & 0 \\ 3 & 10 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{66}$$

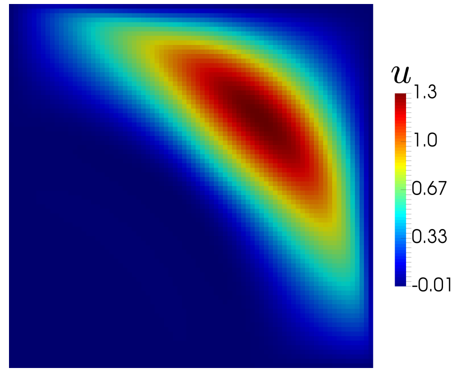
The exact solutions in the sub-domains are

$$\overline{u}_1 = 14x + y + z, \quad \overline{u}_2 = 4x + y + z + 6. \tag{67}$$

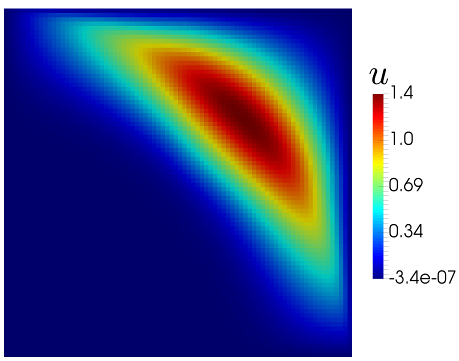
Figure 6 (left) depicts the exact solution. Please note that the exact solution and the corresponding flux function are globally continuous within the domain. It can also be seen that the grid is non-matching at the transition of the sub-domains. Such non-matching grids often occur in faulted geological environments. The grid in Figure 6 is defined by means of the standard corner-point



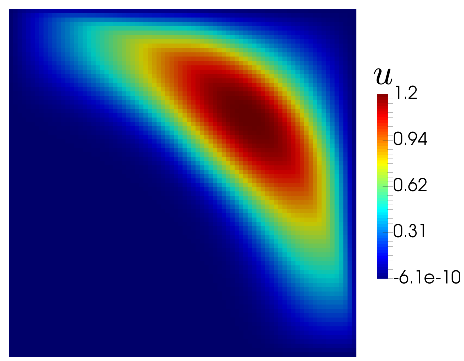
(a) Box scheme



(b) AvgMPFA scheme



(c) NLTPFA scheme



(d) NLMPFA scheme

Figure 3: Solution of Box, AvgMPFA, NLTPFA and NLMPFA schemes for the first extremum principle test case.

grid format and has been generated with the *Matlab Reservoir Simulation Toolbox* (MRST) [33]. To read in the grid, the *opm-grid* module from the *Open Porous Media (OPM) initiative*[2] has been used.

Table 7: Discrete error norms, number of non-zero entries in the Jacobian matrix (nnz) and the number of Newton iterations (nIt) needed for the linearity-preservation test case.

scheme	$\|u_n - \overline{u}\|_{L^2}$	$\|u_n - \overline{u}\|_{\mathcal{T}}$	nnz	nIt
NLTPFA	1.97e-08	8.11e-07	184111	4
NLMPFA	1.99e-08	8.31e-07	184202	7
AvgMPFA	1.99e-08	8.31e-07	184111	1
TPFA	9.11e-03	3.92e-01	107600	1

Table 7 lists the discrete error norms, the number of non-zero entries in the Jacobian matrix (nnz) and the number of Newton iterations (nIt) needed for the simulation run. It can be seen that the NLTPFA, the NLMPFA and the AvgMPFA all reproduce the exact solution, because the

[2] http://opm-project.org/

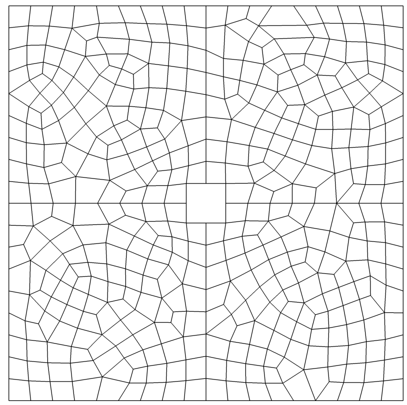


Figure 4: Unstructured grid used for the second discrete extremum principle test case.

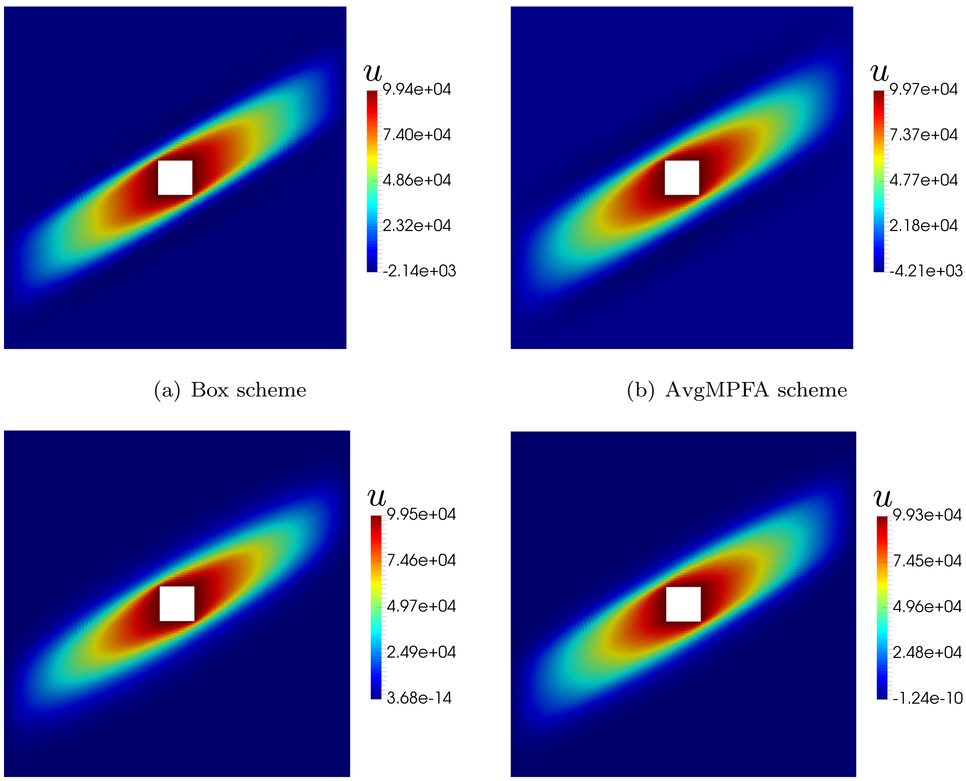
errors are within the range of the nonlinear and linear solver tolerance, whereas the errors of the linear TPFA scheme are approximately five orders of magnitude higher. It is well-known that the errors of the linear TPFA scheme are in $\mathcal{O}(1)$ for non-K-orthogonal grids. However, the improved accuracy of the other schemes comes with the cost of a larger face flux stencil, which is the reason why the corresponding Jacobian matrices are denser than the one of the TPFA scheme. When using Picard's method instead of Newton's method, the number of non-zero entries would be the same for the NLTPFA and TPFA scheme.

The next example is a synthetic model of sedimentary basin inspired by the 3D *Northeast German Basin* model presented in [34]. An approximate geometry of the basin was reconstructed using the software TemisFlow developed at IFPEN. For that case, the stationary heat equation is solved, where, here, Λ corresponds to the thermal conductivity $[\mathrm{W}/(\mathrm{m}\cdot\mathrm{K})]$ and u to the temperature [K]. The thermal conductivity has been computed using the following law

$$\Lambda = \left(\frac{\Lambda_w}{\Lambda_s}\right)^{\phi} \frac{\Lambda_s}{1+\alpha u},$$

where α is a coefficient used to express the thermal dependency, Λ_w and Λ_s denote the water and rock conductivities, and ϕ the porosity. A vertical geothermal gradient was assumed initially to evaluate the law. Salt diapirs within this model create high conductive regions, as shown in Figure 7, leading to thermal anomalies. A robust discretization with respect to the grid is required for this type of structure, in order to evaluate the temperature field and to perform thermohaline simulations. At the top and bottom boundaries, Dirichlet conditions are set to $281.15\,\mathrm{K}$ and $423.15\,\mathrm{K}$, respectively, whereas Neumann no-flow conditions are used elsewhere.

Figure 8 (a)-(c) show the numerical solutions of the TPFA, NLTPFA and the Box scheme. Additionally, the absolute difference between the TPFA and the NLTPFA is depicted in Figure 8 (d). It is observed that the TPFA scheme differs from the NLTPFA and Box scheme especially at



(a) Box scheme (b) AvgMPFA scheme

(c) NLTPFA scheme (d) NLMPFA scheme

Figure 5: Solution of Box, AvgMPFA, NLTPFA and NLMPFA schemes for the second discrete extremum principle test case.

the salt domes, where it seems that the TPFA scheme overestimates the temperature values.

Table 8 lists the discrete error norms $\|u_1 - u_2\|_{L^2}$ between the schemes. Please note that the total domain volume is approximately $|\Omega| \approx 1.75e14\,\mathrm{m}^3$, which explains why the errors are quite large. All schemes differ at most from the TPFA scheme, which shows a better accuracy of the schemes compared to a TPFA.

Table 8: Discrete error norms $\|u_1 - u_2\|_{L^2}$ between the different schemes.

scheme	NLTPFA	NLMPFA	AvgMPFA	TPFA	Box	nnz	nIt
NLTPFA	0	9.09e06	2.28e06	6.69e07	2.27e07	11967982	6
NLMPFA	9.09e06	0	8.98e06	6.57e07	2.26e07	11969149	9
AvgMPFA	2.28e06	8.98e06	0	6.69e07	2.26e07	11967982	1
TPFA	6.69e07	6.57e07	6.69e07	0	7.84e07	5974567	1
Box	2.27e07	2.26e07	2.26e07	7.84e07	0	23684992	1

Again, the number of non-zero entries of the NLTPFA, NLMPFA and AvgMPFA is approximately twice the number of the TPFA scheme. Moreover, the most dense matrix is the one of the

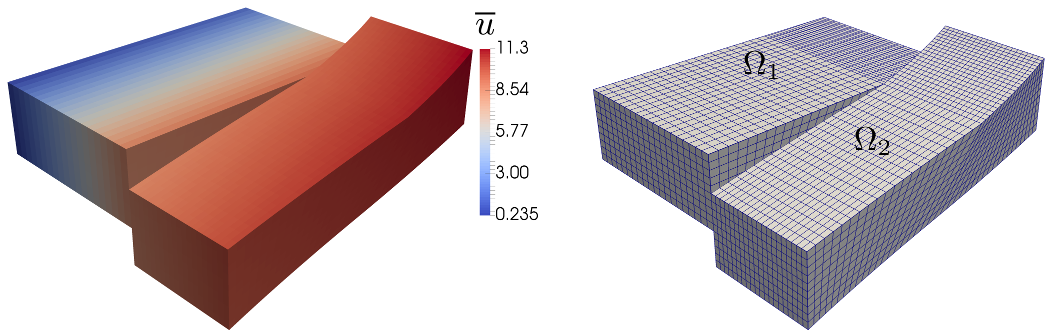


Figure 6: Exact solution for linearity-preservation test case (left); Grid used for the spatial discretization (right).

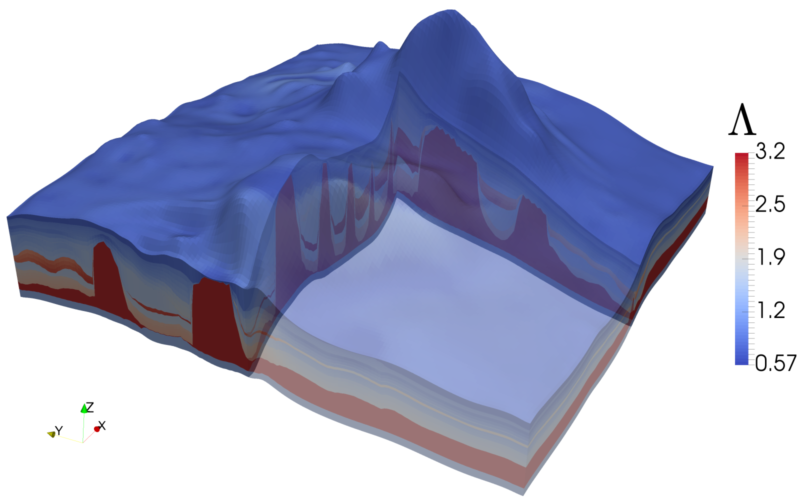


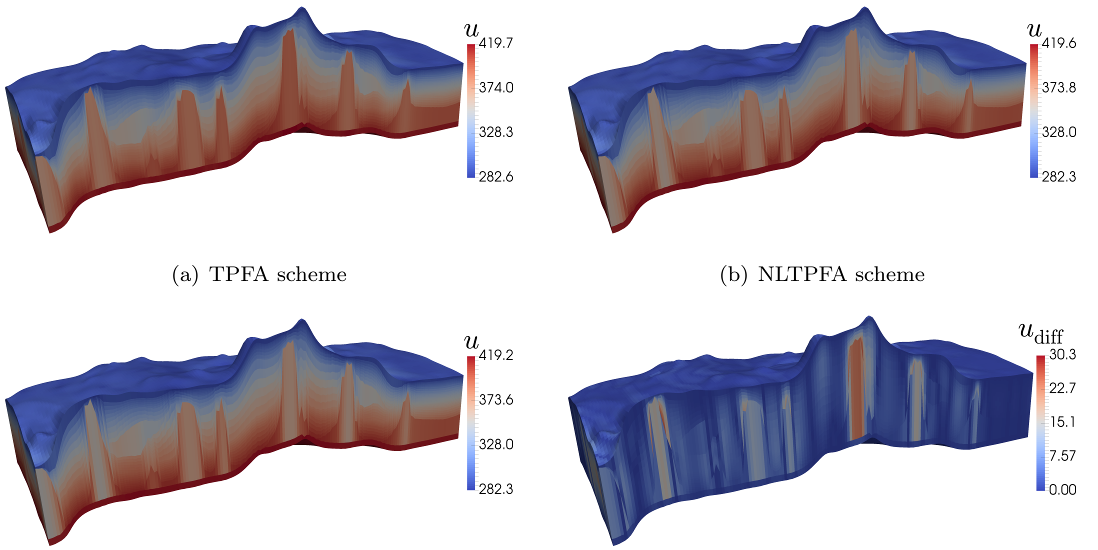
Figure 7: Thermal conductivity of the Northeast German Basin. The salt domes correspond to the high conductive regions. The domain lengths in coordinate directions are approximately 169 km (in the x-direction), 165 km (in the y-direction), and 17.57 km (in the z-direction).

Box scheme.

6. Conclusion

In this article, a family of cell-centered finite volume schemes has been introduced and analyzed. The construction of these schemes is based on a convex combination of two face flux approximations. These face flux approximations are designed to satisfy a strong consistency condition by choosing an appropriate face interpolator.

In the first part of this work, a proof of the convergence of this family of schemes has been given. In Section 4, two representatives of this family have been constructed, namely the nonlinear two-point flux approximation (NLTPFA) and the nonlinear multi-point flux approximation (NLMPFA), such that the strong consistency assumption is fulfilled. To guarantee the existence of a discrete solution, the discrete flux approximations have been modified to be continuous in



(a) TPFA scheme

(b) NLTPFA scheme

(c) Box scheme

(d) Difference of TPFA and NLTPFA

Figure 8: Solution of TPFA, NLTPFA and Box scheme (a)-(c). Absolute difference of TPFA and NLTPFA scheme (d). The results are shown for a part of the domain.

$H_{\mathcal{T}_n}(\Omega)$. Moreover, the NLTPFA scheme has been extended to the case where negative coefficients arise in the conormal decomposition. This has been achieved by reformulating the residual term in the flux approximation.

Finally, in Section 5, the nonlinear schemes have been compared to linear ones. The convergence behavior has been analyzed for a mild and high anisotropy test case on non-matching, randomly distorted and twisted grids. It has been observed that there are almost no differences in the convergence rates between the linear AvgMPFA and the nonlinear schemes. In addition to that, estimates have shown the coercivity of the schemes for the considered test cases. The main difference between the NLTPFA and the NLMPFA is the number of Newton iterations needed for convergence. For all test cases, the NLTPFA requires less iterations than the NLMPFA scheme. The positivity-preserving property of the nonlinear schemes has been analyzed in Section 5.2, where it has been shown that linear schemes produce unphysical negative values, in contrast to the nonlinear ones. In Section 5.3, it has been demonstrated that the introduced schemes are linearly exact on non-matching grids. Furthermore, the schemes have been applied to a synthetic geological formation inspired by the Northeast German Basin, to solve the stationary heat equation with heterogeneous thermal conductivities. It has been shown that the standard linear TPFA scheme overestimates the temperature in salt domes, whereas the NLTPFA, NLMPFA, AvgMPFA and Box schemes all exhibit similar behavior.

Within this work, only linear elliptic problems have been considered. Therefore, using a nonlinear discretization method obviously deteriorates the efficiency of the computations compared to linear schemes. However, this drawback vanishes when solving highly nonlinear partial differential equations [21].

7. Appendix: Technical propositions

Proposition 3 (Density of a space of test-functions)**.** *Under Hypotheses 2, let $\mathcal{Q}$ be the space of functions $\varphi : \overline{\Omega} \to \mathbb{R}$ s.t.*

(i) (φ is continuous and piecewise regular*) $\varphi \in C_0(\overline{\Omega})$ and, for all $i = 1, \ldots, N_\Omega$, $\varphi \in C^2(\overline{\Omega_i})$,*

*(ii) (*the tangential derivatives of φ are continuous through the interfaces of P_Ω*) for all $i, j = 1, \ldots, N_\Omega$, for all vectors $\mathbf{t}$ parallel to $\partial\Omega_i \cap \partial\Omega_j$, $(\nabla\varphi)_{|\overline{\Omega_i}} \cdot \mathbf{t} = (\nabla\varphi)_{|\overline{\Omega_j}} \cdot \mathbf{t}$ on $\partial\Omega_i \cap \partial\Omega_j$, where $(\nabla\varphi)_{|\overline{\Omega_i}}$ refers to the value of $\nabla\varphi$ on $\partial\Omega_i$ computed from the values on $\overline{\Omega_i}$,*

*(iii) (*the flux of $\nabla\varphi$ directed by $\Lambda\mathbf{n}$ is continuous through the interfaces of P_Ω*) for all $i, j = 1, \ldots, N_\Omega$ s.t. $\partial\Omega_i \cap \partial\Omega_j$ has dimension $d-1$, $(\Lambda\nabla\varphi)_{|\overline{\Omega_i}} \cdot \mathbf{n}_i + (\Lambda\nabla\varphi)_{|\overline{\Omega_j}} \cdot \mathbf{n}_j = 0$ on $\partial\Omega_i \cap \partial\Omega_j$, where $\mathbf{n}_i$ is the outer normal to Ω_i.*

Then, $\mathcal{Q}$ is dense in $H_0^1(\Omega)$.

Proof. see [4]. □

Proposition 4 (Discrete Sobolev embeddings)**.** *Let $\mathcal{D}$ be an element of a family of discretizations matching Definition 1. Let $q \in [1, +\infty)$ if $d = 2$, and $q \in [1, 2d/(d-2)]$ if $d > 2$. Then, there exists a strictly positive parameter $C_2 > 0$ depending only on Ω, q, ϱ_1 and ϱ_2 s.t.*

$$\|u\|_{L^q(\Omega)} \leq C_2 \|u\|_{\mathcal{T}} \quad \forall u \in H_{\mathcal{T}}(\Omega).$$

Proof. This result can be proved following the guidelines of the proof in [12, §5.1.2], since all discrete norms considered in this work are equivalent under the mesh regularity assumptions of Definition 1. □

Theorem 2 (Discrete Rellich theorem)**.** *Let $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a sequence of admissible discretizations matching Definition 1 s.t. $h_{\mathcal{D}_n} \to 0$ as $n \to \infty$. Let $\{v_n\}_{n\in\mathbb{N}}$ be a sequence in $H_{\mathcal{T}_n}(\Omega)$ s.t. there exists $C > 0$ with $\|v_n\|_{\mathcal{T}_n} \leq C$ for all $n \in \mathbb{N}$. Then, there exist a subsequence of $\{v_n\}_{n\in\mathbb{N}}$ and a function $\widetilde{v} \in H_0^1(\Omega)$ s.t., as $n \to \infty$, (i) $v_n \to \widetilde{v}$ in $L^q(\Omega)$ for all $q \in [1, 2d/(d-2))$ (and weakly in $L^{2d/(d-2)}(\Omega)$ if $d > 2$); (ii) $\{\widetilde{\nabla}_{\mathcal{D}_n} v_n\}_{n\in\mathbb{N}}$ weakly converges to $\nabla\widetilde{v}$ in $[L^2(\Omega)]^d$.*

Proof. This theorem deduces from (11) using the same techniques as for [12, Lemmata 5.6–5.7]. □

Proposition 5 (Asymptotic stability of the interpolator)**.** *Under Hypotheses 1, we have*

$$\|\varphi_{\mathcal{T}}\|_{\mathcal{T}} \leq \frac{1}{\gamma_1} \left(\epsilon_{\mathcal{D}}(\varphi) + \beta_0 \sqrt{d} |\varphi|_{H^1(\Omega)} \right)$$

for all $\varphi \in \mathfrak{D}$.

Proof. Let $\varphi \in \mathfrak{D}$. Owing to (P2), we get

$$\begin{aligned}\gamma_1 \|\varphi_{\mathcal{T}}\|_{\mathcal{T}}^2 &\le a_{\mathcal{T}}(\varphi_{\mathcal{T}}, \varphi_{\mathcal{T}}, \varphi_{\mathcal{T}}) \\ &= \left(a_{\mathcal{T}}(\varphi_{\mathcal{T}}, \varphi_{\mathcal{T}}, \varphi_{\mathcal{T}}) - \int_\Omega \Lambda \nabla \varphi \cdot \widetilde{\nabla}_{\mathcal{D}} \varphi_{\mathcal{T}} \, \mathrm{d}x \right) + \int_\Omega \Lambda \nabla \varphi \cdot \widetilde{\nabla}_{\mathcal{D}} \varphi_{\mathcal{T}} \, \mathrm{d}x \\ &\le \epsilon_{\mathcal{D}}(\varphi) \|\varphi_{\mathcal{T}}\|_{\mathcal{T}} + \beta_0 |\varphi|_{H^1(\Omega)} \|\widetilde{\nabla}_{\mathcal{D}} \varphi_{\mathcal{T}}\|_{[L^2(\Omega)]^d} \le \left(\epsilon_{\mathcal{D}}(\varphi) + \beta_0 \sqrt{d} |\varphi|_{H^1(\Omega)} \right) \|\varphi_{\mathcal{T}}\|_{\mathcal{T}}. \quad \square\end{aligned}$$

Proposition 6 (Stability)**.** *Assume that Hypotheses 1 hold. Then, any solution $u_n \in H_{\mathcal{D}_n}(\Omega)$ of problem* (4) *for a given $n \in \mathbb{N}$ satisfies the stability estimate*

$$\|u_n\|_{\mathcal{T}_n} \le \frac{C_2}{\gamma_1} \|f\|_{L^r(\Omega)}. \tag{68}$$

Proof. Using the fact that $f \in L^r(\Omega)$ and thanks to (P2), Hölder's inequality and Proposition 4, we have

$$\gamma_1 \|u_n\|_{\mathcal{T}_n}^2 \le a_{\mathcal{T}_n}(u_n, u_n, u_n) = \int_\Omega f u_n \, \mathrm{d}x \le \|f\|_{L^r(\Omega)} \|u_n\|_{L^{r'}(\Omega)} \le C_2 \|f\|_{L^r(\Omega)} \|u_n\|_{\mathcal{T}_n},$$

with $r' \stackrel{\text{def}}{=} \frac{r}{r-1} = \frac{2d}{d-2}$. $\square$

Acknowledgements

The authors Bernd Flemisch and Martin Schneider would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/2) at the University of Stuttgart.

References

[1] L. Agélas, D. A. Di Pietro, R. Masson, A symmetric and coercive finite volume scheme for multiphase porous media flow with applications in the oil industry, in: R. Eymard, J.-M. Hérard (Eds.), Finite Volumes for Complex Applications V, John Wiley & Sons, 2008, pp. 35–52.

[2] I. Aavatsmark, T. Barkve, Ø. Bøe, T. Mannseth, Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media, Journal of Computational Physics 127 (1) (1996) 2–14.

[3] M. Edwards, C. Rogers, Finite volume discretization with imposed flux continuity for the general tensor pressure equation, Computational Geosciences 2 (4) (1998) 259–290.

[4] L. Agélas, D. Di Pietro, J. Droniou, The G method for heterogeneous anisotropic diffusion on general meshes, M2AN Math. Model. Numer. Anal. 44 (4) (2010) 597–625.

[5] M. Wolff, Y. Cao, B. Flemisch, R. Helmig, B. Wohlmuth, Multi-point flux approximation L-method in 3d: numerical convergence and application to two-phase flow through porous media, Radon Series on Computational and Applied Mathematics, De Gruyter 12 (2013) 39–80.

[6] L. Agélas, C. Guichard, R. Masson, Convergence of finite volume MPFA O type schemes for heterogeneous anisotropic diffusion problems on general meshes, Int. J. Finite Vol. 7 (2) (2010) 1–33.

[7] L. Agélas, R. Masson, Convergence of the finite volume MPFA O scheme for heterogeneous anisotropic diffusion problems on general meshes, C. R. Acad. Sci. Paris, Sér. I 346 (17-18) (2008) 1007–1012.

[8] D. Arnold, F. Brezzi, Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates, RAIRO-Modélisation mathématique et analyse numérique 19 (1) (1985) 7–32.

[9] P. Raviart, J. Thomas, A mixed finite element method for 2-nd order elliptic problems, in: Mathematical aspects of finite element methods, Springer, 1977, pp. 292–315.

[10] F. Brezzi, K. Lipnikov, V. Simoncini, A family of mimetic finite difference methods on polygonal and polyhedral meshes, Mathematical Models and Methods in Applied Sciences 15 (10) (2005) 1533–1551.

[11] F. Brezzi, K. Lipnikov, M. Shashkov, Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes, SIAM Journal on Numerical Analysis 43 (5) (2005) 1872–1896.

[12] R. Eymard, T. Gallouët, R. Herbin, Discretization of heterogeneous and anisotropic diffusion problems on general non-conforming meshes. SUSHI: a scheme using stabilization and hybrid interfaces, IMA J. Num. Anal. 30 (4) (2010) 1009–1043.

[13] R. Eymard, R. Herbin, C. Guichard, R. Masson, Vertex-centred discretization of multiphase compositional Darcy flows on general meshes, Comput. Geosci. 16 (4) (2012) 987–1005.

[14] J. Nordbotten, I. Aavatsmark, G. Eigestad, Monotonicity of control volume methods, Numerische Mathematik 106 (2) (2007) 255–288.

[15] C. L. Potier, Schéma volumes finis monotone pour des opérateurs de diffusion fortement anisotropes sur des maillages de triangles non structurés, C. R. Math. Acad. Sci. 341 (12) (2005) 787–792.

[16] G. Yuan, Z. Sheng, Monotone finite volume schemes for diffusion equations on polygonal meshes, J. Comput. Phys. 227 (12) (2008) 6288–6312.

[17] A. Danilov, Y. Vassilevski, A monotone nonlinear finite volume method for diffusion equations on conformal polyhedral meshes, Russ. J. Numer. Anal. Math. Modelling 24 (3) (2009) 207–227.

[18] K. Lipnikov, D. Svyatskiy, Y. Vassilevski, Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes, J. Comput. Phys. 228 (3) (2009) 703–716.

[19] J. Droniou, C. L. Potier, Construction and convergence study of schemes preserving the elliptic local maximum principle, SIAM J. Numer. Anal. 49 (2) (2011) 459–490.

[20] K. Lipnikov, D. Svyatskiy, Y. Vassilevski, Minimal stencil finite volume scheme with the discrete maximum principle, Russ. J. Numer. Anal. Math. Modelling 27 (4) (2012) 369–385.

[21] M. Schneider, B. Flemisch, R. Helmig, Monotone nonlinear finite-volume method for non-isothermal two-phase two-component flow in porous media, Int. J. Numer. Methods Fluids (2016) –.

[22] J. Droniou, Finite volume schemes for diffusion equations: introduction to and review of modern methods, Math. Mod. Meths. Appli. Sci. (M3AS) 24 (8) (2014) 1575–1619.

[23] M. C. C. Cancès, C. L. Potier, Monotone coercive cell-centered finite volume schemes for anisotropic diffusion equations, Numer. Math. 125 (3) (2013) 387–417.

[24] C. L. Potier, A nonlinear finite volume scheme satisfying maximum and minimum principles for diffusion operators, Int. J. Finite Vol. 6 (2).

[25] Z. Sheng, G. Yuan, The finite volume scheme preserving extremum principle for diffusion equations on polygonal meshes, Journal of Computational Physics 230 (7) (2011) 2588–2604.

[26] M. Schneider, B. Flemisch, R. Helmig, K. Terekhov, H. Tchelepi, Monotone nonlinear finite-volume method for challenging grids, SimTech preprint.
URL `http://www.simtech.uni-stuttgart.de/publikationen/prints.php?ID=1507`

[27] L. Agélas, R. Eymard, R. Herbin, A nine-point finite volume scheme for the simulation of diffusion in heterogeneous media, C. R. Math. Acad. Sci. 347 (11) (2009) 673–676.

[28] W. Hackbusch, On first and second order box schemes, Computing 41 (4) (1989) 277–296.

[29] R. Helmig, Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems., Springer-Verlag, 1997.

[30] J. Hommel, S. Ackermann, M. Beck, B. Becker, H. Class, T. Fetzer, B. Flemisch, D. Gläser, C. Grüninger, K. Heck, A. Kissinger, T. Koch, M. Schneider, G. Seitz, K. Weishaupt, DuMuX 2.10.0 (Sep. 2016). `doi:10.5281/zenodo.159007`.
URL `https://doi.org/10.5281/zenodo.159007`

[31] M. Blatt, A. Burchardt, A. Dedner, C. Engwer, J. Fahlke, B. Flemisch, C. Gersbacher, C. Gräser, F. Gruber, C. Grüninger, D. Kempf, R. Klöfkorn, T. Malkmus, S. Müthing, M. Nolte, M. Piatkowski, O. Sander, The distributed and unified numerics environment, version 2.4, Archive of Numerical Software 4 (100) (2016) 13–29.

[32] K. Terekhov, B. Mallison, H. Tchelepi, Cell-centered nonlinear finite-volume methods for the heterogeneous anisotropic diffusion problem, Journal of Computational Physics.

[33] S. Krogstad, K. Lie, O. Møyner, H. M. Nilsen, X. Raynaud, B. Skaflestad, MRST-AD–an open-source framework for rapid prototyping and evaluation of reservoir simulation problems, in: SPE reservoir simulation symposium, Society of Petroleum Engineers, 2015.

[34] M. Scheck, U. Bayer, Evolution of the northeast german basin - inferences from a 3d structural model and subsidence analysis, Tectonophysics 3 (3) (1999) 145–169.