Positive Scharfetter-Gummel finite volume method for convection-diffusion equations on polygonal meshes

El-Houssaine Quenjel

Positive Scharfetter-Gummel finite volume method for convection-diffusion equations on polygonal meshes

El Houssaine QUENJEL

Chair of Biotechnology, LGPM, CentraleSupélec, CEBB, 3 rue des Rouges Terres, 51110 Pomacle, France
el-houssaine.quenjel@centralesupelec.fr

June 14, 2021

Abstract

In this paper, we develop and study a fully implicit positive finite volume scheme allowing to accurately approximate the nonlinear highly anisotropic convection-diffusion equations on almost arbitrary girds. The key idea is to relate the unstable fluxes, including the convective ones, to the normal monotonic diffusive flux thanks to a technique used in the Scharfetter-Gummel discretizations. Then, we obtain a nonlinear two-point-like scheme with positive coefficients on primal and dual meshes. We check that the underlined structure naturally ensures the nonnegativity of the approximate solutions. We also establish energy estimate, which enables the existence of the numerical solutions. This study is accompanied with a series of numerical results and simulations. They highlight the satisfied discrete maximum principle, the optimal accuracy, and the robustness with respect to the mesh and to important anisotropic ratios.

1 Introduction

The convection-diffusion equation is fundamental in modeling and numerically solving a broad range of physical, biological and engineering problems. Typical ones are fluid flows in complex porous media, chemotaxis and semiconductor device simulation. In the current work, the primary unknown describes the diffusion of a quantity e.g. a tracer, mass density, fluid saturation, or temperature while its displacement is governed by a given velocity field.

We next state the mathematical model we are interested in, we survey the relation with the existing literature works and we highlight the originality as well as the novelty of this work.

1.1 Model problem

Let Ω be a bounded connected polygonal open domain of $\mathbb{R}^d$ with $d \geq 1$. We denote by $\partial\Omega = \Gamma^D \cup \Gamma^N$ its boundary. Let $t_f > 0$ denote time. We consider the nonlinear unsteady

convection-diffusion problem written as

$$u_t - \nabla \cdot \Big(\kappa(u)\Lambda\nabla\gamma(u) - \gamma(u)\mathbf{V}\Big) = f \quad \text{in} \quad \Omega \times (0,t_f), \tag{1.1}$$

$$\Big(\kappa(u)\Lambda\nabla\gamma(u) - \gamma(u)\mathbf{V}\Big)\cdot \mathbf{n} = 0 \quad \text{on} \quad \Gamma^N \times (0,t_f), \tag{1.2}$$

$$u = 0 \quad \text{on} \quad \Gamma^D \times (0,t_f), \tag{1.3}$$

$$u(\cdot,0) = u^0 \quad \text{in} \quad \Omega. \tag{1.4}$$

The function $u(x,t)$ represents the primary unknown. The diffusivity function is κ. The potential and the flux transport function γ are considered identical. The space-dependent tensor Λ refers to the diffusion matrix. The vector $\mathbf{V}$ is a given velocity field. The right-hand side f accounts for source or sink contributions. The vector $\mathbf{n}$ is the unit outer normal to Γ^N. The numerical analysis of the scheme derived from the model (1.1)-(1.4) requires assumptions on the data. We set $Q_{t_f} = \Omega \times (0,t_f)$.

- **($\mathbf{A}_0$)** The initial condition u^0 is nonnegative i.e. $u^0 \geq 0$.
- **($\mathbf{A}_1$)** The diffusion function κ is nondecreasing continuous from $\mathbb{R}^+$ into $\mathbb{R}^+$. Assume that κ does not degenerate on a whole interval of $\mathbb{R}^+$. Consider that γ is increasing and continuous from $\mathbb{R}^+$ into $\mathbb{R}^+$ with $\gamma(0) = 0$. It is extended by 0 on $(-\infty, 0)$. Assume that $J(u^0)$ is L^1-integrable on Ω, where $J(s) = \int_0^s \gamma(v)\,\mathrm{d}v$. Suppose that $\sqrt{\kappa}\gamma'$ is also $\mathrm{L}^1_{\mathrm{loc}}$-integrable on $\mathbb{R}^+$. Then, define the Kirchhoff transform $\mu(s) = \int_0^s \sqrt{\kappa(v)}\gamma'(v)\,\mathrm{d}v$. Assume that there exists $C_\gamma > 0$ such that

$$\gamma(s) \leq C_\gamma + \mu(s), \quad \forall s \geq 0. \tag{1.5}$$

- **($\mathbf{A}_2$)** The symmetric matrix Λ is a measurable function from Ω to $L^\infty(\Omega)^{d\times d}$. There exist $\underline{\lambda}, \overline{\lambda} > 0$ such that

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

- **($\mathbf{A}_3$)** The velocity $\mathbf{V}$ belongs to $\mathcal{C}^1(\overline{\Omega})$ with $\nabla\cdot\mathbf{V} \geq 0$.
- **($\mathbf{A}_4$)** The function $f \geq 0$ is in $L^2(Q_{t_f})$.

1.2 Literature works

Several discretizations methods are available in the literature for approximating the convection-diffusion equation (1.1) with various focuses. For instance, in the context of porous media flows, we find the famous (Two-Point Flux Approximation) TPFA method [16, 17] thanks to its practical implementation, computational speed and robustness. As its names states, the two-point scheme approaches the flux using only two degrees of freedom per every interface shared by two cells and neglects other contributions by imposing an orthogonality condition on the mesh. Out of the orthogonal setting, multi-point approximations of the flux were developed to take into account general meshes and fully anisotropic tensors, e.g. see [13, 14] and the references are therein. Also, combined finite volume/finite element schemes were proposed in [1, 3, 18, 24]. In this work we will be concerned with the Discrete Duality Finite Volume (DDFV) method due to special approximation of the diffusive flux [2, 6, 12, 21]. We also conceive a positive nonlinear finite volume Scharfetter-Gummel (S. G.) strategy based on the DDFV fluxes (see the next subsection).

The standard S. G. scheme was firstly introduced in [29] for the 1D drift-diffusion system modeling charge carrier transport in semiconductors. It has been extended to the context of finite volume methods in [5, 10, 23]. *Brezzi et al.* proposed a mixed finite element S. G. scheme in the linear case based on the non-logarithmic Landau formulation [8]. *Da Veig et al.* suggested a S. G. approach to handle convection terms in finite volumes and mimetic discretization methods for elliptic problems [11]. However, the scheme of [11] does not satisfy the discrete maximum principle. In terms of accuracy, the S. G. methodology provides a better alternative to the upwind scheme for the transport term, especially in the diffusive regime.

On the other hand, we recently developed some recent positive nonlinear schemes with limitations. The work based on suitable upwind flux approximations was carried out in [26]. It is only first order accurate in space which is due to the additional artificial viscosity spanned by upwinding. A similar idea has been generalized to DDFV framework in [28]. It was observed that the accuracy is reduced when the solution is only continuous. Their extensions to complex flows in porous media have been investigated in [7, 19]. Contrary to upwind based-discretizations, central positive nonlinear schemes have been proposed in [9, 25, 27]. These strategies make use of some singular potential functionals to reinforce the solution positivity in the case of nonlinear diffusion or linear diffusion with a drift. They are merely problem-dependent. Generally, it is not known how to intrinsically extend them to the context of the porous medium equation when the convective effects are present.

1.3 Originality and contribution

The present work proposes a new finite volume scheme for (1.1)-(1.4), which is stable, accurate and robust. Stability allows to ensures the production of physical solutions and enables the coercivity of the scheme. Accuracy is referred to as the quadratic convergence rate when the solution is regular. By robustness, we mean that the stability and accuracy are independent of the mesh type and anisotropic tensors.

The originality of this work lies in the flux approximation based on the multi-point DDFV method. Precisely, at each interface between two control volumes, the DDFV strategy particularly provides normal and tangential fluxes for the diffusion. The first one is stable whereas the second one is unstable on generic meshes. Being inspired by the Scharfetter-Gummel technique, widely used in the TPFA context for diffusion problems with drift, a key element in our contribution is to connect this oscillatory term as well as the convection to the stable monotonic normal diffusive flux in a consistent way. To our knowledge, this technique has not been employed yet for the diffusion in the multi-point setting. This leads to a scheme having a two-point structure with positive coefficients. Such a formulation naturally entails the nonnegativity of the solution. Additionally, the scheme is still coercive. The coercivity result particularly serves to prove the existence of approximate solutions.

The numerical section brings out the major strengths of our novel scheme. Indeed, we test the accuracy and the robustness by varying the mesh and anisotropy. The precision turns out to be of second order for sufficiently smooth analytical solutions and optimal otherwise. This is the case of the convective porous medium equation with a low regularity solution. Beside to the positivity feature, our new scheme provides convergence rates similar to the ones produced by unstable quasi-linear schemes of the literature.

To sum up, the significant assets of the current scheme are provided in the following list.

- Include nonlinear diffusivity and transport functions (κ, γ).
- Handle meshes with quite generic shapes (e.g. distorted, non-conforming, etc).

- Incorporate highly anisotropic and heterogeneous diffusion tensors.
- Honor the physical range of the solution.
- Existence of approximate solutions.
- Second order numerical accuracy.
- Robustness with respect to the mesh and the anisotropic ratio.

1.4 Outline of the paper

In Section 2, we present and analyze the positive finite volume scheme discretizing the convection-diffusion model (1.1)-(1.4). More specifically, we first describe the discrete setting consisting of the spatial mesh, the discrete gradient and the discretization of the time interval. We derive the proposed flux approximation by means of the Scharfetter-Gummel technique. The final scheme is obtained by considering a fully implicit Euler approximation. Next, we study its crucial mathematical properties. Precisely, we check that the discrete maximum principle holds true. We demonstrate the coercivity of the scheme, and prove the existence of the numerical solutions. Section 2 is devoted to the numerical results where a particular emphasis is placed on the accuracy assessment, the validation of the proved maximum principle and the simulation of the quarter five spot problem with barriers. Finally, in Section 4, we cloture with a conclusion.

2 The finite volume scheme and its analysis

In this second section, we will elaborate the proposed nonlinear finite volume scheme. We are also going to study some of its substantial mathematical properties. Before that, we need to specify the meshed domain and some relevant notations.

2.1 Domain discretization and discrete gradient

The cornerstone point of the discrete duality finite volume methodology resides in the construction of a whole gradient in 2D using only geometrical objects of the mesh. As a result, two different directions are locally required to approach the gradient in a consistent and a stable way. To this purpose, the normal component is approximated on the primal mesh. The tangential component is conceived on the dual mesh.

The primal mesh $\mathcal{P}$ is a partition of Ω consisting of nonoverlapping open subset, usually called control volumes (A), such that $\bigcup_{A\in\mathcal{P}} \overline{A} = \overline{\Omega}$. Let $\mathbf{x}_A$ be the centeroid of A. The notation ∂A refers to the boundary of the cell A, and $|A|$ the area of A. The family $\mathcal{E}_A$ denotes the set of edges of A. We assume that each inner interface σ is only shared by two adjacent cells. So, $\sigma \in \mathcal{E}_A$ yields the existence of a unique $B \in \mathcal{P}$ such that $\overline{\sigma} = \overline{A} \cap \overline{B}$. We sometimes write $\sigma = A|B$ or σ_{AB} to indicate the cell neighbors of σ which are A and B. By convention, this notation can also cover the exterior edges. Indeed, if A is a boundary cell we still denote by B the boundary edge, seen as a degenerate cell. The boundary cells are all gathered in the set $\partial\mathcal{P}$. The symbol $|\sigma|$ defines the length of the interface σ. The unit normal to σ pointing from A to B is labeled by $\mathbf{n}_{AB}$. Keeping the underlined orientation, the unit tangent is given by $\mathbf{t}_{AB}$.

The dual mesh $\overline{\mathcal{P}^\star}$ is constructed and centered on the primal vertices. Let $A^\star$ denote the dual control volume. Its center coincides with the vertex $\mathbf{x}_{A^\star}$. In fact, $A^\star$ is a subdomain

obtained by connecting the centeroids of the primal cells sharing the vertex $\mathbf{x}_{A^\star}$. The dual edges of $A^\star$ are then collected in the set $\mathcal{E}_{A^\star}$. The length of $\sigma^\star \in \mathcal{E}_{A^\star}$ is designated by $|\sigma^\star|$. Analogously to the primal case, the vector $\mathbf{n}_{A^\star B^\star}$ (resp. $\mathbf{t}_{A^\star B^\star}$) denotes the unit dual normal (resp. tangent) to $\sigma^\star = A^\star | B^\star$.

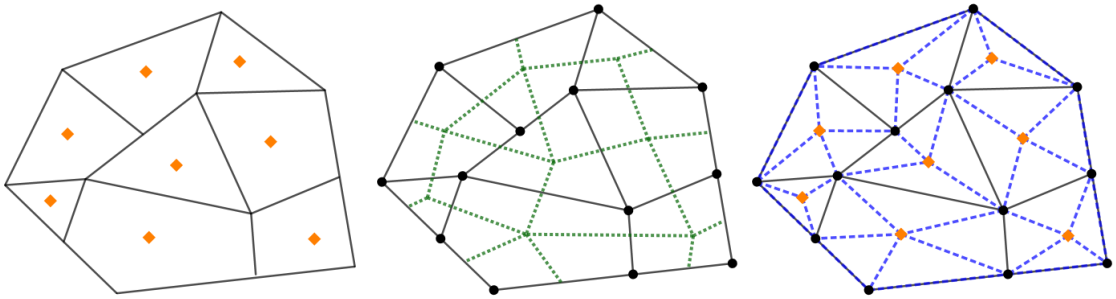

Figure 1: Schematic draw showing the primal mesh (left with solid lines), its associated dual (medium with dotted lines) and diamond meshes (right with dashed lines) as well as the location of degrees of freedom.

The diamond mesh $\mathfrak{D}$ consists of diamond cells. It is formed around the primal and dual interfaces. Given $\sigma = A|B$ with vertices $\mathbf{x}_{A^\star}, \mathbf{x}_{B^\star}$, we associate a unique diamond $\mathcal{D} \in \mathfrak{D}$. Practically, the cell $\mathcal{D}$ is the polygon whose vertices are $\{\mathbf{x}_A, \mathbf{x}_{A^\star}, \mathbf{x}_B, \mathbf{x}_{B^\star}\}$. It can be checked that $\bigcup_{\mathcal{D}\in\mathfrak{D}} \overline{\mathcal{D}} = \overline{\Omega}$. Let $\theta_\mathcal{D} \in]0, \pi/2]$ be the angel defined by $\theta_\mathcal{D} = \arcsin(\mathbf{n}_{AB}, \mathbf{t}_{AB})$. The diamond area $|\mathcal{D}|$ is measured by $|\mathcal{D}| = \sin(\theta_\mathcal{D})\,|\sigma|\,|\sigma^\star|\,/2$. The primal, dual and diamond meshes are depicted in Figure 1. The aforementioned geometrical objects per each diamond are illustrated in Figure 2.

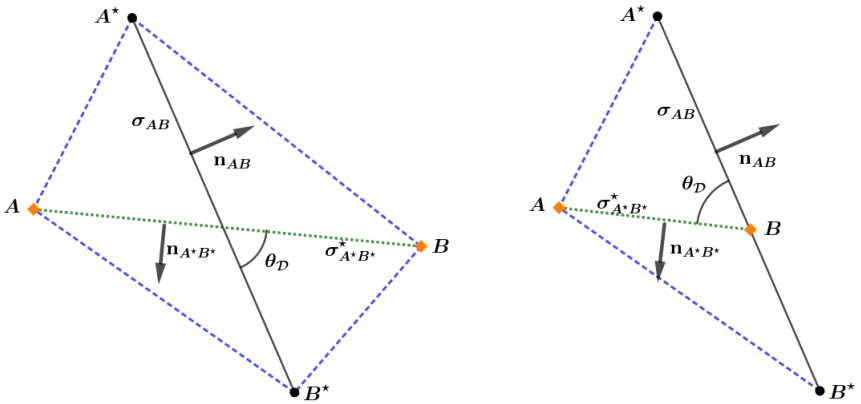

Figure 2: Left : example of a whole diamond cell constructed around the interior primal edge σ_{AB} together with its corresponding geometrical entities. Right : half of diamond where σ_{AB} is a boundary edge.

We consider that the borders Γ^D and Γ^N are articulated on some vertices of the primal mesh. We specify the sets of the different boundary cells

$$\partial\mathcal{P}^D = \left\{B \in \partial\mathcal{P} / \mathbf{x}_B \in \Gamma^D\right\}, \quad \partial\mathcal{P}^N = \partial\mathcal{P} \setminus \partial\mathcal{P}^D,$$
$$\partial\mathcal{P}^{\star,D} = \left\{A^\star \in \overline{\mathcal{P}^\star} / \mathbf{x}_{A^\star} \in \Gamma^D\right\}, \quad \mathcal{P}^\star = \overline{\mathcal{P}^\star} \setminus \partial\mathcal{P}^{\star,D}.$$

Let us set $\mathcal{T} = \mathcal{P} \cup \overline{\mathcal{P}^\star}$ and denote by h_W the diameter of the subset $W \in \mathfrak{D} \cup \mathcal{T}$. Define

$$\xi_{\mathcal{T}} = \max\left(\max_{\mathcal{D}\in\mathfrak{D}} \frac{1}{\sin(\theta_{\mathcal{D}})}, \max_{\mathcal{D}\in\mathfrak{D}} \frac{h_{\mathcal{D}}}{\sqrt{|\mathcal{D}|}}, \max_{K\in\mathcal{T}} \frac{h_K}{\sqrt{|K|}}\right).$$

In particular, this indicator gives information on how flat diamonds are. To control their shapes from degeneracy and too flatting we assume a regularity condition on the mesh. It claims that

$$\xi_{\mathcal{T}} \leq \xi,$$

for some positive constant ξ.

The degrees of freedom are placed at the cell centers and at the vertices. We denote by $\mathbb{R}^{\mathcal{T}}$ the space of discrete unknowns written as

$$\mathbf{s}_{\mathcal{T}} = \big((\mathbf{s}_A)_{A\in\mathcal{T}}, (\mathbf{s}_{A^\star})_{A^\star\in\mathcal{P}^\star}\big).$$

The function reconstruction on Ω is prescribed by the discrete operator $\mathcal{I}_{\mathcal{T}}$ defined from $\mathbb{R}^{\mathcal{T}}$ into $L^2(\Omega)$ as follows

$$\mathcal{I}_{\mathcal{T}}\mathbf{s}_{\mathcal{T}} = \frac{1}{2}\sum_{K\in\mathcal{T}} \mathbf{s}_K \mathbf{1}_K,$$

where $\mathbf{1}_K$ is the characteristic function of the subset K.

As mentioned in the beginning of the present subsection, the definition of the discrete gradient formula is a key element in the flux approximation procedure when the mesh is quite general. In the DDFV case, it is locally expressed per each diamond cell and uniquely defined in the corresponding basis $(\mathbf{n}_{AB}, \mathbf{n}_{A^\star B^\star})$. In fact, the reconstructed gradient $\nabla^{\mathfrak{D}}$ is a linear operator defined from $\mathbb{R}^{\mathcal{T}}$ into $L^2(\Omega)^2$ by

$$\nabla^{\mathfrak{D}}_{|\mathcal{D}}\mathbf{s}_{\mathcal{T}} := \nabla^{\mathcal{D}}\mathbf{s}_{\mathcal{T}} = \frac{1}{\sin(\theta_{\mathcal{D}})}\left(\frac{\mathbf{s}_B - \mathbf{s}_A}{|\sigma^\star|}\mathbf{n}_{AB} + \frac{\mathbf{s}_{B^\star} - \mathbf{s}_{A^\star}}{|\sigma\star|}\mathbf{n}_{A^\star B^\star}\right), \quad \forall \mathcal{D} \in \mathfrak{D}. \tag{2.1}$$

The function reconstruction and the approximate gradient are related through the discrete Poincaré inequality. Its proof can be found in [2, Lemma 3.3]. It is recalled hereafter.

Lemma 2.1. *There exists a constant C_p depending only on the diameter of Ω and the mesh regularity ξ such that for all $\boldsymbol{s}_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ one has*

$$\|\mathcal{I}_{\mathcal{T}}\boldsymbol{s}_{\mathcal{T}}\|_{L^2(\Omega)} \leq C_p \|\nabla^{\mathfrak{D}}\boldsymbol{s}_{\mathcal{T}}\|_{L^2(\Omega)^2}.$$

Since $|\mathcal{D}| = \sin(\theta_{\mathcal{D}})\,|\sigma|\,|\sigma^\star|\,/2$ the expression of the gradient is equivalent to

$$\nabla^{\mathcal{D}}\mathbf{s}_{\mathcal{T}} = \frac{1}{2\,|\mathcal{D}|}\Big(\sigma(\mathbf{s}_B - \mathbf{s}_A)\mathbf{n}_{AB} + \sigma^\star\,(\mathbf{s}_{B^\star} - \mathbf{s}_{A^\star})\,\mathbf{n}_{A^\star B^\star}\Big), \quad \forall \mathcal{D} \in \mathfrak{D}.$$

The anisotropic tensor $\Lambda(\cdot)$ is attached to the diffusion and therefore to the gradient. It is also approximated per diamonds using the integral average

$$\Lambda^{\mathcal{D}} = \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \Lambda(x)\,\mathrm{d}x, \quad \forall \mathcal{D} \in \mathfrak{D}.$$

In the rest of the paper, we adopt the notations

$$\tau_{AB} = \frac{|\sigma|^2\,\Lambda^{\mathcal{D}}\mathbf{n}_{AB}\cdot\mathbf{n}_{AB}}{2\,|\mathcal{D}|}, \quad \tau_{A^\star B^\star} = \frac{|\sigma^\star|^2\,\Lambda^{\mathcal{D}}\mathbf{n}_{A^\star B^\star}\cdot\mathbf{n}_{A^\star B^\star}}{2\,|\mathcal{D}|}, \quad \eta_{\mathcal{D}} = \frac{|\sigma|\,|\sigma^\star|\,\Lambda^{\mathcal{D}}\mathbf{n}_{AB}\cdot\mathbf{n}_{A^\star B^\star}}{2\,|\mathcal{D}|}.$$

Define the 2×2 symmetric matrix $\mathcal{R}_{\mathcal{D}}$ and the vector $\delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}}$ by

$$\mathcal{R}_{\mathcal{D}} = \begin{pmatrix} \tau_{AB} & \eta_{\mathcal{D}} \\ \eta_{\mathcal{D}} & \tau_{A^\star B^\star}. \end{pmatrix}, \qquad \delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}} = \begin{pmatrix} s_A - s_B \\ s_{A^\star} - s_{B^\star} \end{pmatrix}. \tag{2.2}$$

It can be easily checked that

$$|\mathcal{D}|\, \Lambda^{\mathcal{D}} \nabla^{\mathcal{D}}\mathbf{s}_{\mathcal{T}} \cdot \nabla^{\mathcal{D}}\mathbf{s}_{\mathcal{T}} = \frac{1}{2}\mathcal{R}_{\mathcal{D}}\delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}} \cdot \delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}}. \tag{2.3}$$

To establish the coercivity of the scheme we need to make sure that the left hand side of (2.3) is only controlled by the usual $\mathbb{R}^2$-norm of $\delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}}$ independently of the mesh. This is addressed in the following result.

Lemma 2.2. *There exist two positive constants m_0 and m_1 depending only on $\underline{\lambda}$, $\overline{\lambda}$ and the mesh regularity ξ such that*

$$m_0 \left|\delta^{\mathcal{D}}\boldsymbol{s}_{\mathcal{T}}\right|^2 \leq \mathcal{R}_{\mathcal{D}}\delta^{\mathcal{D}}\boldsymbol{s}_{\mathcal{T}} \cdot \delta^{\mathcal{D}}\boldsymbol{s}_{\mathcal{T}} \leq m_1 \left|\delta^{\mathcal{D}}\boldsymbol{s}_{\mathcal{T}}\right|^2, \quad \forall \boldsymbol{s}_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}, \quad \forall \mathcal{D} \in \mathfrak{D}.$$

Proof. Let $\mathbf{s}_{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ and $\mathcal{D} \in \mathfrak{D}$. Thanks to the ellipticity of Λ, given in $(\mathbf{A}_2)$, the equality (2.3) entails

$$2\underline{\lambda}\,|\mathcal{D}|\left|\nabla^{\mathcal{D}}\mathbf{s}_{\mathcal{T}}\right|^2 \leq \mathcal{R}_{\mathcal{D}}\delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}} \cdot \delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}} \leq 2\overline{\lambda}\,|\mathcal{D}|\left|\nabla^{\mathcal{D}}\mathbf{s}_{\mathcal{T}}\right|^2.$$

Following the result [27, Lemma 2.1.], there exist $C_0 > 0$ and $C_1 > 0$ depending only on the mesh regularity such that

$$C_0 \left|\delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}}\right|^2 \leq |\mathcal{D}|\left|\nabla^{\mathcal{D}}\mathbf{s}_{\mathcal{T}}\right|^2 \leq C_1 \left|\delta^{\mathcal{D}}\mathbf{s}_{\mathcal{T}}\right|^2.$$

As a consequence, the required inequality is obtained by setting

$$m_0 = 2\underline{\lambda}C_0, \quad m_1 = 2\overline{\lambda}C_1,$$

which concludes the proof. □

Let $N_{t_f} \in \mathbb{N} \setminus \{0\}$. We denote $[\![0, N_{t_f}]\!] = [0, N_{t_f}] \cap \mathbb{N}$. Without loss of generality, the time interval $(0, t_f)$ is partitioned into N_{t_f} equidistant sub-intervals whose size is denoted by $\triangle t$. Their end-points are given by an increasing sequence $(t^n)_{n\in[\![0,N_{t_f}]\!]}$ where $t^0 = 0$ and $t^{N_{t_f}} = t_f$. Any time-dependent variable denoted by $\mathbf{s}_{\mathcal{T}}^{n+1}$ means that it is computed at the $(n+1)$th level.

The temporal discretization is of chief important regarding stability. In this paper, we will take into account a fully implicit scheme to ensure the unconditional stability in time. In other words, no CFL condition is required for the numerical analysis of the scheme.

2.2 Generalized Scharfetter-Gummel finite volume scheme

In this part, we elaborate the proposed nonlinear finite volume scheme. We forget the time for the moment. Our key contribution particularly targets the space discretization of the model. First, we derive a consistent and coercive approximation of the fluxes based on the DDFV method. We next show how to extend the Scharfetter-Gummel strategy to handle both : the oscillatory diffusive fluxes and the convective term. Then, we provide an optimal positive structure to the whole approximated flux. We only focus on the scheme construction in the context of the primal mesh. Same ideas are naturally carried out in the case of the dual mesh.

Let $A \in \mathcal{P}$ be a control volume. Integrating the divergence term of (1.1) on A and applying Green's formula yield

$$X_A = -\int_A \nabla \cdot \Big(\kappa(u)\Lambda\nabla\gamma(u) - \gamma(u)\mathbf{V}\Big)\,\mathrm{d}x \\ = -\sum_{\sigma=A|B\in\mathcal{E}_A}\int_\sigma \Big(\kappa(u)\Lambda\nabla\gamma(u) - \gamma(u)\mathbf{V}\Big)\cdot\mathbf{n}_{AB}\,\mathrm{d}\sigma(x).$$

Approaching the continuous gradient and the tensor Λ by their discrete counterparts permits us to approximate X_A as follows

$$X_A \approx -\sum_{\sigma=A|B\in\mathcal{E}_A}\int_\sigma \Big(\kappa(u)\Lambda^{\mathcal{D}}\nabla^{\mathcal{D}}\gamma(u_{\mathcal{T}}) - \gamma(u)\mathbf{V}\Big)\cdot\mathbf{n}_{AB}\,\mathrm{d}\sigma(x). \tag{2.4}$$

Developing the right hand side (RHS) of (2.4) one obtains at each interface σ

$$-\int_\sigma \Big(\kappa(u)\Lambda^{\mathcal{D}}\nabla^{\mathcal{D}}\gamma(u_{\mathcal{T}}) - \gamma(u)\mathbf{V}\Big)\cdot\mathbf{n}_{AB}\,\mathrm{d}\sigma(x) \\ = \kappa^{\mathcal{D}}(u_{\mathcal{T}})\Big(\tau_{AB}\big(\gamma(u_A)-\gamma(u_B)\big) + \eta_{\mathcal{D}}\big(\gamma(u_{A^\star})-\gamma(u_{B^\star})\big)\Big) + \int_{\sigma_{AB}}\gamma(u)\mathbf{V}\cdot\mathbf{n}_{AB}\,\mathrm{d}\sigma(x),$$

where the diffusivity function is approximated thanks to a centered scheme on the diamond $\mathcal{D}$, that is

$$\kappa^{\mathcal{D}}(u_{\mathcal{T}}) = \frac{1}{4}\Big(\kappa(u_A)+\kappa(u_B)+\kappa(u_{A^\star})+\kappa(u_{B^\star})\Big). \tag{2.5}$$

A well-known fashion to discretize the convective term in the RHS of (2.4) consists in the standard upwind scheme. In contrast to the centered scheme, it takes into account the sign of the projected velocity on $\mathbf{n}_{AB}$ based on physical principles. This approach only offers an accuracy of first order on the solution due to the additional numerical viscosity. To improve the accuracy, at least where the diffusion is important, one can make use of the Scharfetter-Gummel scheme. The later scheme has been intensively studied in the case of TPFA methods. Its generalization to multi-point finite volumes schemes has been addressed in [11] for the linear unsteady convection-diffusion problem by separating the discretization of the connective term from the diffusion part. In the current work, we develop a novel idea enabling to connect the unsigned tangential diffusive flux together with the convective flux to the normal monotonic diffusive flux. To translate that into formulas, let us first set

$$\varphi(s) = s\coth(s/2)/2 \;\; (s\neq 0), \quad \gamma_{AB} = \frac{1}{2}\big(\gamma\left(u_A\right)+\gamma\left(u_B\right)\big), \quad \mathbf{V}_{AB} = \int_{\sigma_{AB}}\mathbf{V}\cdot\mathbf{n}_{AB}\,\mathrm{d}\sigma(x).$$

Note that φ is extended by continuity at 0 i.e. $\varphi(0) = 1$. The proposed flux, which is subsequently denoted by $\mathcal{F}_{AB}(u_{\mathcal{T}})$, centers the convection and weightens the monotonic normal flux i.e.

$$-\int_\sigma \Big(\kappa(u)\Lambda^{\mathcal{D}}\nabla^{\mathcal{D}}\gamma(u_{\mathcal{T}}) - \gamma(u)\mathbf{V}\Big)\cdot\mathbf{n}_{AB}\,\mathrm{d}\sigma(x) \\ \approx \mathcal{F}_{AB}(u_{\mathcal{T}}) := \kappa^{\mathcal{D}}(u_{\mathcal{T}})\tau_{AB}\varphi\big(g_{AB}(u_{\mathcal{T}})\big)\Big(\gamma(u_A)-\gamma(u_B)\Big) \\ + \kappa^{\mathcal{D}}(u_{\mathcal{T}})\gamma_{AB}\frac{\eta_{\mathcal{D}}\big(\gamma(u_{A^\star})-\gamma(u_{B^\star})\big)}{\big(\gamma_{AB}+\varepsilon\big)} + \gamma_{AB}\mathbf{V}_{AB}, \tag{2.6}$$

where $g_{AB}(u_{\mathcal{T}})$ refers to terms with uncontrolled signs that are gathered under the following form

$$g_{AB}(u_{\mathcal{T}}) = \begin{cases} \dfrac{\eta_{\mathcal{D}}\big(\gamma(u_{A^\star}) - \gamma(u_{B^\star})\big)}{\tau_{AB}\big(\gamma_{AB} + \varepsilon\big)} + \dfrac{\mathbf{V}_{AB}}{\tau_{AB}\kappa^{\mathcal{D}}(u_{\mathcal{T}})} & \text{if} \quad \kappa^{\mathcal{D}}(u_{\mathcal{T}}) \neq 0 \\ \\ 0 & \text{if} \quad \kappa^{\mathcal{D}}(u_{\mathcal{T}}) = 0 \end{cases}. \tag{2.7}$$

The role of the parameter ε is to ensure the continuity of the function g_{AB} with respect to $u_{\mathcal{T}}$. To avoid it influence from a numerical perspective, it can be fixed to $\varepsilon = 10^{-14}$. The case corresponding to $\kappa^{\mathcal{D}}(u_{\mathcal{T}}) = 0$ has no impact since its associated flux is vanished.

As a consequence, the final proposed approximation of X_A is derived thanks to a generalization of the Scharfetter-Gummel scheme. It is written in the weighted two-point formulation

$$X_A \approx \sum_{\sigma = A|B \in \mathcal{E}_A} \kappa^{\mathcal{D}}(u_{\mathcal{T}})\tau_{AB}\Big(\boldsymbol{\alpha_{AB}}\gamma(u_A) - \boldsymbol{\alpha_{BA}}\gamma(u_B)\Big). \tag{2.8}$$

For readability, the remained coefficients $\boldsymbol{\alpha_{AB}}, \boldsymbol{\alpha_{BA}}$ are nonlinearly expressed as

$$\boldsymbol{\alpha_{AB}} = \boldsymbol{\alpha}\big(-g_{AB}(u_{\mathcal{T}})\big), \quad \boldsymbol{\alpha_{BA}} = \boldsymbol{\alpha}\big(g_{AB}(u_{\mathcal{T}})\big),$$

where $\boldsymbol{\alpha}$ is the Lipschitz-continuous function of Bernoulli defined by

$$\boldsymbol{\alpha}(r) = \frac{r}{e^r - 1} = \frac{r}{2}\Big(\coth\Big(\frac{r}{2}\Big) - 1\Big), \quad \forall r \in \mathbb{R} \setminus \{0\}, \quad \boldsymbol{\alpha}(0) = 1.$$

It further satisfies

$$\boldsymbol{\alpha}(-r) - \boldsymbol{\alpha}(r) = r, \quad \forall r \in \mathbb{R} \setminus \{0\}. \tag{2.9}$$

Remark 2.1.

(i) Observe that $\boldsymbol{\alpha}(r) \underset{\infty}{\sim} \max(-r, 0)$. *This means that the Scharfetter-Gummel scheme reduces to the uwpwind scheme in regions where convection is highly dominated.*

(ii) Even if the compact approximation (2.8) *seems simple, the formula* (2.6) *is quite practical for the derivation of the energy estimate, since it segregates the unstable fluxes from the monotonic flux.*

Proceeding similarly on the dual cells we find

$$\begin{aligned} &-\int_{A^\star} \nabla \cdot \Big(\kappa(u)\Lambda\nabla\gamma(u) - \gamma(u)\mathbf{V}\Big)\,\mathrm{d}x \\ &\approx \mathcal{F}_{A^\star B^\star}(u_{\mathcal{T}}) := \sum_{\sigma = A^\star|B^\star \in \mathcal{E}_{A^\star}} \kappa^{\mathcal{D}}(u_{\mathcal{T}})\tau_{A^\star B^\star}\Big(\boldsymbol{\alpha_{A^\star B^\star}}\gamma(u_{A^\star}) - \boldsymbol{\alpha_{B^\star A^\star}}\gamma(u_{B^\star})\Big). \end{aligned}$$

One can check that $\mathcal{F}_{AB}(u_{\mathcal{T}})$ and $\mathcal{F}_{A^\star B^\star}(u_{\mathcal{T}})$ are conservative using (2.9).

Concerning the temporal discretization we consider a fully implicit Euler scheme in order to reinforce the unconditional stability of the scheme. This allows to avoid small time steps in highly heterogeneous permeability regions.

The initial datum and the source term are approximated thanks to the mean integral i.e.

$$u_K^0 = \frac{1}{|K|}\int_K u(x)\,\mathrm{d}x, \qquad f_K^{n+1} = \frac{1}{\triangle t\,|K|}\int_{t^n}^{t^{n+1}}\int_K f(x,t)\,\mathrm{d}x\,\mathrm{d}t, \quad K = A, A^\star;\ n \geq 0.$$

To sum up, the discrete convection-diffusion process is governed by the nonlinear finite volume scheme consisting of the algebraic equations et each time level n:

$$\frac{|A|}{\triangle t}\big(u_A^{n+1}-u_A^n\big)+\sum_{\sigma_{AB}\in\mathcal{E}_A}\mathcal{F}_{AB}(u_{\mathcal{T}}^{n+1})=|A|\,f_A^{n+1},\quad \forall A\in\mathcal{P}, \tag{2.10}$$

$$\frac{|A|}{\triangle t}\big(u_{A^\star}^{n+1}-u_{A^\star}^n\big)+\sum_{\sigma^\star_{A^\star B^\star}\in\mathcal{E}_{A^\star}}\mathcal{F}_{A^\star B^\star}(u_{\mathcal{T}}^{n+1})=|A^\star|\,f_{A^\star}^{n+1},\quad \forall A^\star\in\mathcal{P}^\star, \tag{2.11}$$

$$u_B^{n+1}=0,\quad \forall B\in\partial\mathcal{P}^D,\quad u_{A^\star}^{n+1}=0,\quad \forall A^\star\in\partial\mathcal{P}^{\star,D}, \tag{2.12}$$

$$\mathcal{F}_{AB}(u_{\mathcal{T}}^{n+1})=0,\quad \forall B\in\partial\mathcal{P}^N. \tag{2.13}$$

The first line represents the balance equation on each primal cell. The second one accounts for the balance equation on the dual control volumes. The latter ones are necessary to equalize the number of unknowns with the system size and to ensure the scheme stability as we are going to see below. The equation (2.12) is the discrete counterpart of the imposed Dirichlet boundary condition. The last raw is nothing more than the homogeneous Neumann boundary condition, which allows to determine u_B^{n+1} whenever B is in $\partial\mathcal{P}^N$.

Remark 2.2.

- *When κ is constant and γ is linear, note that the flux $\mathcal{F}_{AB}(u_{\mathcal{T}}^{n+1})$ always remains nonlinear contrary to the traditional DDFV flux which is linear.*
- *Assume that the mesh is orthogonal in the sense of* Eymard et al. *[16] and the tensor Λ is a scalar function. Such a situation may happen on Cartesian meshes or on triangulations which the angle condition. In particular, this yields $\eta_{\mathcal{D}}=0$ for all $\mathcal{D}\in\mathfrak{D}$. As a result, the system (2.10)-(2.13) reduces to the nonlinear conventional Scharfetter-Gummel finite volume scheme [5, 22, 29] on the primal and on the dual meshes separately.*
- *We emphasize the fact that the boundary conditions are strongly simplified in order to highlight the key points of our contribution. The results are still valid if the Dirichlet constraint is prescribed by a quite general function u^{Dir}. We refer to [2] for the derivation and the analysis of the discrete Dirichlet condition using center or averaged values of u^{Dir}.*

2.3 Stability analysis and existence result

In this subsection, we prove that the new nonlinear finite volume scheme is stable in the following sense. First, the numerical solution is nonnegative as in the continuous setting. Also, the discrete gradient of the Kirchhoff function is uniformly bounded with regard to the energy norm. Both properties are independent of the chosen mesh. The coercivity property allows to prove the existence result.

To start off, let us look at the central statement about the nonnegativity of the solution.

Proposition 2.1. *Any solution $(u_{\mathcal{T}}^{n+1})_{n\in[\![0,N_{t_f}-1]\!]}$ to the nonlinear system (2.10)-(2.13) respects its physical range, that is*

$$u_K^{n+1}\geq 0,\quad \forall A\in\mathcal{T},\forall n\in[\![0,N_{t_f}-1]\!].$$

Proof. It is suffices to establish

$$u_K^{n+1}\geq 0,\quad \forall K\in\mathcal{P}, \tag{2.14}$$

by induction on n. The property on the dual mesh can be drawn in the same manner. Choose $A \in \mathcal{P}$ so that $u_A^{n+1} = \min_{K\in\mathcal{P}} u_K^{n+1}$. There is no thing to prove if $u_{\mathcal{T}}^{n+1}$ is constant. Avoiding the trivial case, the claim (2.14) is equivalent to showing $u_A^{n+1} \geq 0$. Let us prove the latter inequality by contradiction. Assume that $u_A^{n+1} < 0$. Multiplying the line of (2.10) corresponding to A by u_A^{n+1} yields

$$\frac{|A|}{\triangle t}(u_A^{n+1} - u_A^n)u_A^{n+1} + \sum_{\sigma_{AB}\in\mathcal{E}_A} \mathcal{F}_{AB}(u_{\mathcal{T}}^{n+1})u_A^{n+1} = |A|\, f_A^{n+1}u_A^{n+1} \leq 0. \tag{2.15}$$

Let us study the left hand side of (2.15). First, it obvious that

$$\frac{|A|}{\triangle t}(u_A^{n+1} - u_A^n) < 0.$$

The assumption $u_A^{n+1} = \min_{K\in\mathcal{P}} u_K^{n+1} < 0$ gives $\gamma(u_A^{n+1}) = 0$ since γ is extended by 0 on $(-\infty, 0)$. The function γ is being increasing and the coefficients $\boldsymbol{\alpha}_{\boldsymbol{BA}}^{n+1}$ are positive. Therefore, we obtain

$$\begin{aligned}
\sum_{\sigma_{AB}\in\mathcal{E}_A} \mathcal{F}_{AB}(u_{\mathcal{T}}^{n+1}) &= \sum_{\sigma_{AB}\in\mathcal{E}_A} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\big(\boldsymbol{\alpha}_{\boldsymbol{AB}}^{n+1}\gamma(u_A^{n+1}) - \boldsymbol{\alpha}_{\boldsymbol{BA}}^{n+1}\gamma(u_B^{n+1})\big) \\
&= \sum_{\sigma_{AB}\in\mathcal{E}_A} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\boldsymbol{\alpha}_{\boldsymbol{BA}}^{n+1}\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big) \\
&\qquad + \gamma(u_A^{n+1}) \sum_{\sigma_{AB}\in\mathcal{E}_A} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\big(\boldsymbol{\alpha}_{\boldsymbol{AB}}^{n+1} - \boldsymbol{\alpha}_{\boldsymbol{BA}}^{n+1}\big) \\
&= \sum_{\sigma_{AB}\in\mathcal{E}_A} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\boldsymbol{\alpha}_{\boldsymbol{BA}}^{n+1}\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big) \leq 0.
\end{aligned}$$

Accordingly, the left hand side of (2.15) is negative whereas the source term is nonpositive. The relationship is not true except if $u_A^{n+1} = 0$. This is a contradiction with the hypothesis $u_A^{n+1} < 0$. To conclude, the nonlinear scheme prohibits the arsing of undershootss by construction. □

Remark 2.3. *If γ is linear, i.e. $\gamma(u) = u$, the whole numerical scheme (2.10)-(2.13) can be recast under the symmetric matrix form $R(U_{\mathcal{T}}^{n+1})U_{\mathcal{T}}^{n+1} = q(U_{\mathcal{T}}^n)$, at each time level n. Hence, the main entries of R are*

$$\begin{aligned}
R(U_{\mathcal{T}}^{n+1})_{AA} &= \frac{|A|}{\triangle t} + \sum_{\sigma_{AB}\in\mathcal{E}_A} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\boldsymbol{\alpha}_{\boldsymbol{AB}}^{n+1} > 0, \\
R(U_{\mathcal{T}}^{n+1})_{AB} &= -\kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\boldsymbol{\alpha}_{\boldsymbol{BA}}^{n+1} \leq 0.
\end{aligned}$$

The coefficients $R(U_{\mathcal{T}}^{n+1})_{A^\star A^\star}$ and $R(U_{\mathcal{T}}^{n+1})_{A^\star B^\star}$ are defined in the same way. Since homogeneous boundary conditions are prescribed, the right hand side $q(U_{\mathcal{T}}^n)$ includes the old level solution and the source term. Furthermore, $R(U_{\mathcal{T}}^{n+1})$ is strictly diagonally dominant and its graph is connected. As a result, $R(U_{\mathcal{T}}^{n+1})$ has the M-Matrix structure i.e. it is invertible and all the entries of its inverse are nonnegative [4]. Consequently, any solution to the proposed finite volume scheme is nonnegative. This gives an alternative proof to Proposition 2.1.

In the following result, we prove the scheme coercivity.

Proposition 2.2. *The energy norm of the approximate gradient of the Kirchhoff functional is controlled by a constant C depending only on the geometrical regularity of the mesh ξ, $\underline{\lambda}, \overline{\lambda}$ and the source function f. In other words, there holds*

$$\sum_{n=0}^{N_{t_f}-1} \triangle t \|\nabla^{\mathfrak{D}} \mu(u_{\mathcal{T}}^{n+1})\|_{L^2(\Omega)^2}^2 \leq C.$$

Proof. Multiply the balance equation (2.10) by $\triangle t u_A^{n+1}$ and sum over the primal cells. Similarly, multiply (2.11) by $\triangle t u_{A^\star}^{n+1}$ and sum on the dual control volumes of $\mathcal{P}^\star$. Add the resulting equations on both meshes, perform a discrete integration-by-parts and use the identity (2.6) to obtain

$$\mathcal{Z}_0 + \mathcal{Z}_1 + \mathcal{Z}_2 + \mathcal{Z}_3 + \mathcal{Z}_4 + \mathcal{Z}_0^\star + \mathcal{Z}_1^\star + \mathcal{Z}_2^\star + \mathcal{Z}_3^\star + \mathcal{Z}_4^\star = 0,$$

where $\mathcal{Z}_0, \cdots, \mathcal{Z}_4$ write

$$\begin{aligned}
\mathcal{Z}_0 &= \sum_{n=0}^{N_{t_f}-1} \sum_{A\in\mathcal{P}} |A| \left(u_A^{n+1} - u_A^n\right)\gamma(u_A^{n+1}),\\
\mathcal{Z}_1 &= \sum_{n=0}^{N_{t_f}-1} \triangle t \sum_{\mathcal{D}\in\mathfrak{D}} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\varphi\big(g_{AB}(u_{\mathcal{T}}^{n+1})\big)\Big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\Big)\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big),\\
\mathcal{Z}_2 &= \sum_{n=0}^{N_{t_f}-1} \triangle t \sum_{\mathcal{D}\in\mathfrak{D}} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\gamma_{AB}^{n+1} \frac{\eta_{\mathcal{D}}\big(\gamma(u_{A^\star}^{n+1}) - \gamma(u_{B^\star}^{n+1})\big)}{\big(\gamma_{AB}^{n+1} + \varepsilon\big)}\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big),\\
\mathcal{Z}_3 &= \sum_{n=0}^{N_{t_f}-1} \triangle t \sum_{A\in\mathcal{P}} \sum_{\sigma_{AB}\in\mathcal{E}_A} \gamma_{AB}^{n+1}\mathbf{V}_{AB}, \quad \mathcal{Z}_4 = \sum_{n=0}^{N_{t_f}-1} \triangle t \sum_{A\in\mathcal{P}} |A|\, f_A^{n+1}\gamma(u_A^{n+1}).
\end{aligned}$$

The dual terms $\mathcal{Z}_0, \cdots, \mathcal{Z}_4$ are written in the same fashion. They are omitted for legibility. The accumulation term is commonly estimated with a telescopic series leading to

$$\mathcal{Z}_0 \geq \sum_{A\in\mathcal{P}} |A| \left(J(u_A^N) - J(u_A^0)\right) \geq -\left\|J(u^0)\right\|_{L^1(\Omega)},$$

where the convexity of J and Jensen's inequality are applied. Same inequality holds for $\mathcal{Z}_0$. Next, the function φ satisfies $\varphi(s) \geq 1$ for all s. As a consequence

$$\begin{aligned}
\mathcal{Z}_1 &= \sum_{n=0}^{N_{t_f}-1} \triangle t \sum_{\mathcal{D}\in\mathfrak{D}} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\varphi\big(g_{AB}(u_{\mathcal{T}}^{n+1})\big)\Big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\Big)\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big)\\
&\geq \sum_{n=0}^{N_{t_f}-1} \triangle t \sum_{\mathcal{D}\in\mathfrak{D}} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\tau_{AB}\Big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\Big)\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big).
\end{aligned}$$

$\mathcal{Z}_1^\star$ is analogously estimated. Define

$$\mathfrak{D}^{\pm} = \Big\{\mathcal{D}\in\mathfrak{D},\ \mathbf{sgn}\ \Big(\eta_{\mathcal{D}}\big(\gamma(u_{A^\star}^{n+1}) - \gamma(u_{B^\star}^{n+1})\big)\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big)\Big) = \pm 1\Big\},$$

sgn is the sign function i.e. **sgn** $(x) = x/|x|$ for all $x \neq 0$. Using this notation, one finds that

$$\mathcal{Z}_2 \geq \sum_{\mathcal{D}\in\mathfrak{D}^-} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\Big(\eta_{\mathcal{D}}\big(\gamma(u_{A^\star}^{n+1}) - \gamma(u_{B^\star}^{n+1})\big)\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big)\Big).$$

Similar inequality is valid for $\mathcal{Z}_2^\star$. Let $\widetilde{\mathcal{R}}_{\mathcal{D}} = \mathbf{diag}\big(\mathcal{R}_{\mathcal{D}}\big)$ be a diagonal matrix defined by extracting the diagonal of $\mathcal{R}_{\mathcal{D}}$ introduced in (2.2). It is now possible to gather the previous terms to compute

$$\begin{aligned}\overline{Z} := \mathcal{Z}_1 + \mathcal{Z}_2 + \mathcal{Z}_1^\star + \mathcal{Z}_2^\star \geq \sum_{n=0}^{N_{t_f}-1} \triangle t \Bigg(&\sum_{\mathcal{D}\in\mathfrak{D}^+} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\widetilde{\mathcal{R}}_{\mathcal{D}}\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1})\cdot\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1}) \\ &+ \sum_{\mathcal{D}\in\mathfrak{D}^-} \kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\mathcal{R}_{\mathcal{D}}\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1})\cdot\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1})\Bigg).\end{aligned}$$

Using the Cauchy-Schwarz inequality on the term

$$\frac{1}{2\,|\mathcal{D}|}\Big(\big(\gamma(u_A^{n+1}) - \gamma(u_B^{n+1})\big)\,|\sigma|\,\sqrt{\Lambda^{\mathcal{D}}}\mathbf{n}_{AB}\Big)\cdot\Big(\big(\gamma(u_{A^\star}^{n+1}) - \gamma(u_{B^\star}^{n+1})\big)\,|\sigma^\star|\,\sqrt{\Lambda^{\mathcal{D}}}\mathbf{n}_{A^\star B^\star}\Big),$$

we deduce that

$$\widetilde{\mathcal{R}}_{\mathcal{D}}\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1})\cdot\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1}) \geq \frac{1}{2}\mathcal{R}_{\mathcal{D}}\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1})\cdot\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1}).$$

By virtue of this inequality, applying a couple of times Lemma 2.2, bearing in mind the fact that κ is nondecreasing and using the dentition of the Kirchhoff functional μ we infer

$$\begin{aligned}\overline{Z} &\geq \frac{m_0}{2}\sum_{n=0}^{N_{t_f}-1}\triangle t\sum_{\mathcal{D}\in\mathfrak{D}}\kappa^{\mathcal{D}}(u_{\mathcal{T}}^{n+1})\left|\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1})\right|^2 \\ &\geq \frac{m_0}{8}\sum_{n=0}^{N_{t_f}-1}\triangle t\sum_{\mathcal{D}\in\mathfrak{D}}\max_{s\in\{u_A^{n+1},u_B^{n+1},u_{A^\star}^{n+1},u_{B^\star}^{n+1}\}}\kappa(s)\left|\delta^{\mathcal{D}}\gamma(u_{\mathcal{T}}^{n+1})\right|^2 \\ &\geq \frac{m_0}{8}\sum_{n=0}^{N_{t_f}-1}\triangle t\sum_{\mathcal{D}\in\mathfrak{D}}\left|\delta^{\mathcal{D}}\mu(u_{\mathcal{T}}^{n+1})\right|^2 \\ &\geq \frac{m_0}{8m_1}\sum_{n=0}^{N_{t_f}-1}\triangle t\sum_{\mathcal{D}\in\mathfrak{D}}\mathcal{R}_{\mathcal{D}}\delta^{\mathcal{D}}\mu(u_{\mathcal{T}}^{n+1})\cdot\delta^{\mathcal{D}}\mu(u_{\mathcal{T}}^{n+1}) \\ &= m'\sum_{n=0}^{N_{t_f}-1}\triangle t\|\nabla^{\mathfrak{D}}\mu(u_{\mathcal{T}}^{n+1})\|_{L^2(\Omega)^2}^2, \qquad m' = \frac{m_0}{4m_1}.\end{aligned}$$

Rearranging and introducing the positivity assumption $(\mathbf{A}_2)$ on the divergence of the velocity field we deduce

$$\begin{aligned}\mathcal{Z}_3 &= \frac{1}{2}\sum_{n=0}^{N_{t_f}-1}\triangle t\sum_{A\in\mathcal{P}}\big(\gamma(u_A^{n+1})\big)^2\int_{\partial A}\mathbf{V}\cdot\mathbf{n}\,\mathrm{d}\sigma(x) \\ &= \frac{1}{2}\sum_{n=0}^{N_{t_f}-1}\triangle t\sum_{A\in\mathcal{P}}\big(\gamma(u_A^{n+1})\big)^2\int_A\nabla\cdot\mathbf{V}\,\mathrm{d}x \geq 0.\end{aligned}$$

It can be checked that $\mathcal{Z}_3^\star \geq 0$. Finally, thanks to the growth condition on γ mentioned in (1.5), the Cauchy-Schwarz, Young and discrete Poincaré inequalities there holds

$$|\mathcal{Z}_4| \leq C_\gamma \|f\|_{L^1(Q_{t_f})} + \frac{1}{m'}\|f\|^2_{L^2(Q_{t_f})} + \frac{m'}{4}\sum_{n=0}^{N_{t_f}-1} \triangle t \|\nabla^{\mathfrak{D}} \mu(u_{\mathcal{T}}^{n+1})\|^2_{L^2(\Omega)^2}$$

The estimation of $|\mathcal{Z}_4^\star|$ is similarly obtained. Thereby, combining all the contributions, one ends up with

$$\sum_{n=0}^{N_{t_f}-1} \triangle t \|\nabla^{\mathfrak{D}} \mu(u_{\mathcal{T}}^{n+1})\|^2_{L^2(\Omega)^2} \leq C := \frac{4}{m'} C_\gamma \|f\|_{L^1(Q_{t_f})} + \frac{4}{(m')^2}\|f\|^2_{L^2(Q_{t_f})} + \frac{4}{m'}\left\|J(u^0)\right\|_{L^1(\Omega)},$$

proving the coercivity as required. □

We next wonder if the scheme has a solution. The answer is positive. It is stated and proved in the following result.

Proposition 2.3. *There exits at least one nonnegative solution* $(u_{\mathcal{T}}^{n+1})_{n\in[\![0,N_{t_f}-1]\!]}$ *to the nonlinear finite volume scheme* (2.10)-(2.13).

Proof. An inductive argument on n is used for the existence proof. The initial solution trivially exists. Assume that $u_{\mathcal{T}}^n$ is known and let us prove the existence of $u_{\mathcal{T}}^{n+1}$. Set $m = \#(\mathcal{P} \cup \mathcal{P}^\star)$. The space $\mathbb{R}^m$ is equipped with the usual norm denoted $|\cdot|_{\mathbb{R}^m}$. Let $\mathbb{T} : \mathbb{R}^m \to \mathbb{R}^m$ be the continuous functional defined by

$$\begin{aligned}
\mathbb{T}(v_{\mathcal{T}})_{|A} &= |A|\left(v_A - u_A^n\right) + \triangle t \sum_{\sigma_{AB}\in\mathcal{E}_A} \mathcal{F}_{AB}(v_{\mathcal{T}}) - \triangle t\,|A|\, f_A^{n+1}, \quad \forall A \in \mathcal{P},\\
\mathbb{T}(v_{\mathcal{T}})_{|A^\star} &= |A^\star|\left(v_{A^\star} - u_{A^\star}^n\right) + \triangle t \sum_{\sigma_{A^\star B^\star}\in\mathcal{E}_{A^\star}} \mathcal{F}_{A^\star B^\star}(v_{\mathcal{T}}) - \triangle t\,|A^\star|\, f_{A^\star}^{n+1}, \quad \forall A^\star \in \mathcal{P}^\star,\\
\mathcal{F}_{AB}(v_{\mathcal{T}}) &= 0, \quad \forall B \in \partial\mathcal{P}^N.
\end{aligned}$$

If σ_{AB} is an edge belonging to the Dirichlet boundary, we maintain the condition $v_B = v_{A^\star} = v_{B^\star} = 0$. Then, $u_{\mathcal{T}}^{n+1}$ is solution to the finite volume scheme (2.10)-(2.13) if $u_{\mathcal{T}}^{n+1}$ solves $\mathbb{T}(u_{\mathcal{T}}^{n+1}) = 0$. Define the homeomorphism $\mathfrak{G} : \mathbb{R}^m \to \mathbb{R}^m$ such that $\mathfrak{G}(v_{\mathcal{T}}) = \gamma^{-1}(v_{\mathcal{T}})$. Set $\gamma_{\mathcal{T}} = \gamma(v_{\mathcal{T}})$ and $\Upsilon = \mathbb{T} \circ \mathfrak{G}$. It follows that $\Upsilon(\gamma_{\mathcal{T}}) = \mathbb{T}(v_{\mathcal{T}})$. In particular, a solution to $\Upsilon(\gamma_{\mathcal{T}}) = 0$ is automatically a solution to $\mathbb{T}(v_{\mathcal{T}}) = 0$ and vice-versa.

Reproducing the guidelines of the previous proof, making use of the discrete Poincaré inequality and retain the inequality (1.5) of Assumption $(\mathbf{A}_1)$ we deduce

$$\begin{aligned}
\Upsilon(\gamma_{\mathcal{T}})\cdot\gamma_{\mathcal{T}} &= \sum_{A\in\mathcal{P}} |A|\left(v_A - u_A^n\right)\gamma(v_A) + \triangle t \sum_{\sigma_{AB}\in\mathcal{E}_A} \mathcal{F}_{AB}(v_{\mathcal{T}})\big(\gamma(v_A) - \gamma(v_B)\big)\\
&\quad - \triangle t \sum_{A\in\mathcal{P}} |A|\, f_A^{n+1}\gamma(v_A) + \sum_{A^\star\in\mathcal{P}^\star} |A^\star|\left(v_{A^\star} - u_{A^\star}^n\right)\gamma(v_{A^\star})\\
&\quad + \triangle t \sum_{\sigma_{A^\star B^\star}\in\mathcal{E}^\star} \mathcal{F}_{A^\star B^\star}(v_{\mathcal{T}})\big(\gamma(v_{A^\star}) - \gamma(v_{B^\star})\big) - \triangle t \sum_{A^\star\in\mathcal{P}^\star} |A^\star|\, f_{A^\star}^{n+1}\gamma(v_{A^\star})\\
&\geq C_{\triangle t}\|\nabla^{\mathfrak{D}}\mu(v_{\mathcal{T}})\|^2_{L^2(\Omega)^2} - C \geq C'_{\triangle t}\|\mathcal{I}_{\mathcal{T}}\gamma_{\mathcal{T}}\|^2_{L^2(\Omega)} - C',
\end{aligned}$$

for some positive constants $C_{\triangle t}, C_{\triangle t}, C$ and C'. The equivalence of norms on the finite dimensional space $\mathbb{R}^m$ ensures the existence of $C_{\triangle t,\mathcal{T}} > 0$ depending on the mesh parameters such that

$$\Upsilon(\gamma_{\mathcal{T}}) \cdot \gamma_{\mathcal{T}} \geq C_{\triangle t,\mathcal{T}} \, |\gamma_{\mathcal{T}}|^2_{\mathbb{R}^m} - C'.$$

For all $\gamma_{\mathcal{T}}$ such that $|\gamma_{\mathcal{T}}|^2_{\mathbb{R}^m} \leq 1 + C_{\triangle t,\mathcal{T}}/C$, one infers that $\Upsilon(\gamma_{\mathcal{T}}) \cdot \gamma_{\mathcal{T}} > 0$. As a consequence of Brouwer's fixed point criterion [15, Lemma of page 493], the vector field $\mathbb{T}$ admits at least one zero. Hence, there exists a solution $u_{\mathcal{T}}^{n+1}$ to the proposed finite volume scheme for all n. Thanks to Proposition 2.1, this numerical solution is necessarily nonnegative and the proof is concluded. □

3 Numerical results

This numerical section puts into practice the extended positive Scharfetter-Gummel discretization allowing the resolution of the convection-diffusion problem (1.1)-(1.4). The main aim is to highlight the capability of the developed finite volume scheme to respect the physical range of the solution while the optimal accuracy is maintained.

In all the test, the computational domain is the unit square $\overline{\Omega} = [0,1] \times [0,1]$. It is discretized using five sequences of various meshes taken from the famous FVCA5 benchmark for diffusion problems [20]. Theses partitions are made of the triangular, randomly disturbed, locally refined, and Kershaw meshes. We label them with **Tri**, **LocRef**, **Rand** and **Kersh** respectively. Notice that **LocRef** is an example of nonconforming mesh and **Kersh** mesh shows the distortion along the x-direction.

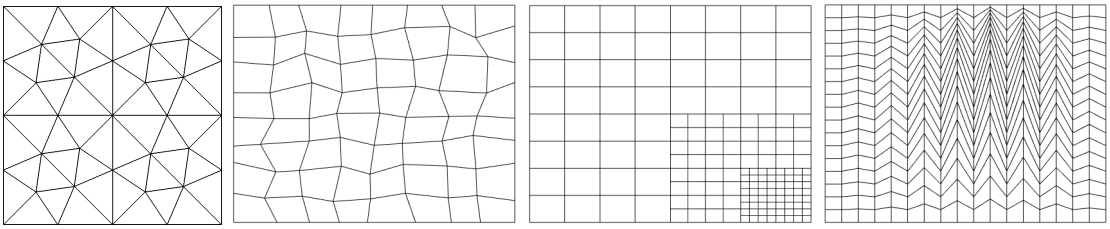

Figure 3: From left to right: triangular, random, locally refined, and Kershaw meshes.

The scheme (2.10)-(2.13) gives rise to a nonlinear algebraic system solved thanks to Newton's method. Its convergence is achieved if the difference between the successive iterates in the ℓ^2-norm is smaller than 10^{-10}.

To evaluate the discrete errors we compute

$$E_{\text{sol}} = \max_{n \in [\![0, N_{t_f}]\!]} \|u_{\text{ex}}(.,t^n) - \mathcal{I}_{\mathcal{T}} u_{\mathcal{T}}^n\|_{L^2(\Omega)}, \quad E_{\text{grad}} = \left(\sum_{n=0}^{N_{t_f}} \triangle t \|\nabla u_{\text{ex}} - \nabla^{\mathfrak{D}} u_{\mathcal{T}}\|^2_{L^2(\Omega)^2} \right)^{1/2}.$$

3.1 Example 1 : linear Fokker-Planck equation

In this example, we are interested in the heat equation with an additional linear convection. This model is usually called linear Fokker-Planck equation (FPE). Our goal is to study the numerical convergence of the proposed "Positive Scharfetter-Gummel Finite Volume" (PSGFV)

approach with regard to different meshes depicted in Figure 3. Also, the scheme precision is assessed in the presence or the absence of anisotropy effects in the y-direction. Concerning te used data-set, we first consider a diagonal tensor Λ as

$$\Lambda = \begin{pmatrix} a_x & 0 \\ 0 & a_y \end{pmatrix}.$$

We take $\kappa(u) = 1$ and $\gamma(u) = u$. Setting the right-hand side $f = 0$ and the mono-directional velocity $\mathbf{V} = a_y \vec{\mathbf{e}}_y = (0, a_y)$, it is possible to manufacture a two-dimensional solution to the model (1.1)-(1.4) as follows

$$u_{\text{ex}}(x, y, t) = \frac{1}{2}\big(\cos(\pi x)\exp(-\pi^2 a_x t) + 1\big)\exp(y), \quad (x, y) \in \Omega, \quad t \in [0, tf].$$

The problem is subject to no-flux Neumann boundary condition on the whole boundary. The initial state stems from this analytical solution. For all the tests of the current example, the final time is set to $t_f = 0.15$ and the time step is proportional to the squared mesh size, i.e. $\triangle t = 0.1 h_{\mathcal{T}}^2$. This $\triangle t$ allows to evaluate the spatial accuracy on the refined meshes.

The numerical results are graphically shown in Figures 4-7 in the log-scale using an increasing anisotropic ratio varying from 0.1 to 100. Regardless anisotropy, a quadratic convergence is obtained in the L^2 norm and rates greater than 1 are noticed for the gradients.

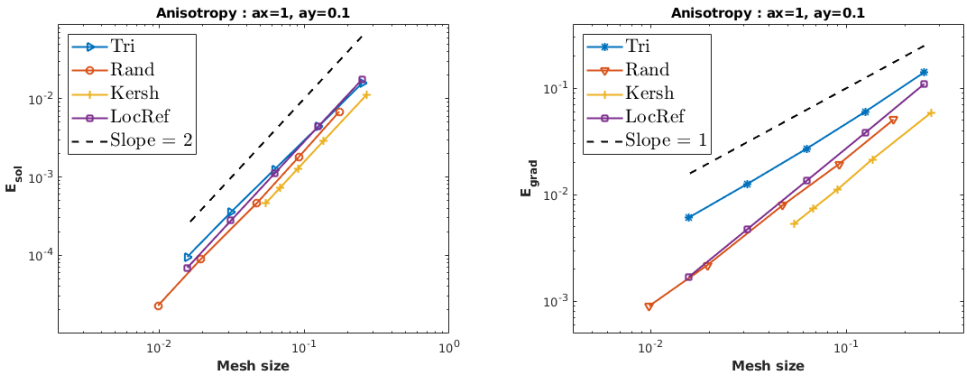

Figure 4: Accuracy results for FPE with $a_y = 0.1$

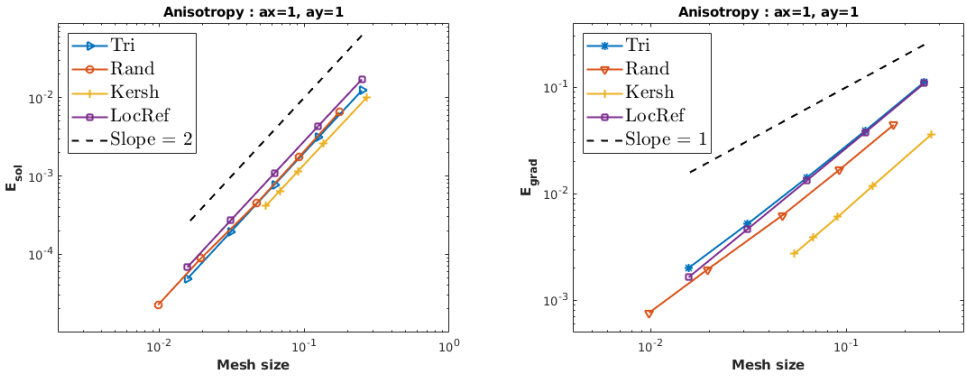

Figure 5: Accuracy results for FPE with $a_y = 1$

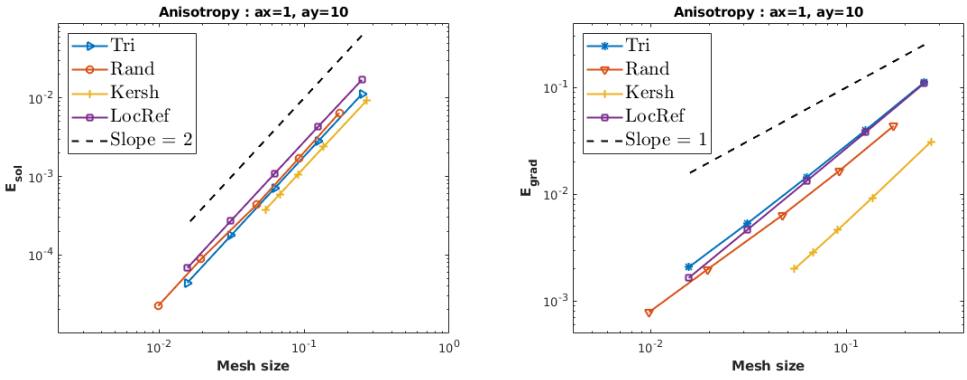

Figure 6: Accuracy results for FPE with $a_y = 10$

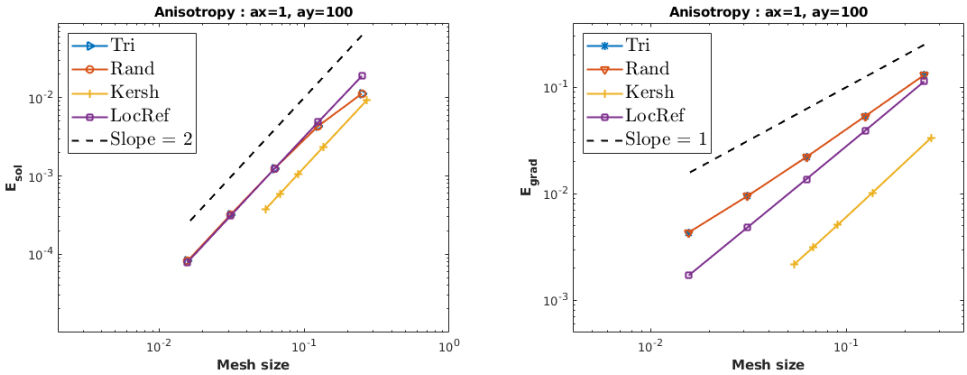

Figure 7: Accuracy results for FPE with $a_y = 100$

In Table 1 we report the minimum of the computed solution after the first time iteration together with the required Newton's iterations for the all considered situations. It is observed that our approach reinforces the positivity of the solution independently of the mesh and anisotropy. The computational cost, leaded by the Newton's solver increases on the coarse meshes as a_y becomes important.

Triangular meshes							
$a_y = 0.1$		$a_y = 1$		$a_y = 10$		$a_y = 100$	
2.8E-2	74	2.7E-2	73	2.8E-2	97	3.6E-2	155
7.4E-3	199	7.2E-3	282	7.4E-3	307	9.9E-3	460
1.8E-3	772	1.8E-3	772	1.8E-3	919	2.5E-3	1247
4.6E-4	3075	4.5E-4	3074	4.6E-4	3090	6.3E-4	3126
1.1E-4	12290	1.1E-4	12290	1.1E-4	12291	1.5E-4	12297
Random meshes							
$a_y = 0.1$		$a_y = 1$		$a_y = 10$		$a_y = 100$	
1.4E-2	134	1.4E-2	151	1.4E-2	193	3.6E-2	155
3.9E-3	363	3.9E-3	367	3.9E-3	535	9.9E-3	460
1.1E-3	1351	1.1E-3	1351	1.1E-3	1407	2.5E-3	1247
3.9E-9	7928	3.8E-9	7928	1.8E-4	7930	6.3E-4	3126
1.4E-9	31120	1.3E-9	31120	1.3E-9	31120	1.5E-4	12297
Locally refined meshes							
$a_y = 0.1$		$a_y = 1$		$a_y = 10$		$a_y = 100$	
2.8E-2	65	2.8E-2	65	2.8E-2	68	2.6E-2	86
7.5E-3	201	7.5E-3	201	7.5E-3	203	7.4E-3	209
1.9E-3	772	1.9E-3	772	1.9E-3	772	1.9E-3	777
4.8E-4	3075	4.8E-4	3075	4.7E-4	3075	4.7E-4	3077
1.8E-9	12289	1.8E-9	12289	1.1E-4	12289	1.2E-4	12293
Kershaw meshes							
$a_y = 0.1$		$a_y = 1$		$a_y = 10$		$a_y = 100$	
3.4E-2	76	3.4E-2	68	3.4E-2	99	3.4E-2	165
8.9E-3	171	8.9E-3	228	8.9E-3	267	8.8E-3	413
4.0E-3	371	3.9E-3	371	3.9E-3	480	3.9E-3	643
2.2E-3	657	2.2E-3	657	2.2E-3	657	2.2E-3	882
1.4E-3	1025	1.4E-3	1025	1.4E-3	1025	1.4E-3	1060

Table 1: FPE : lower bound of the numerical solution and total number of Newton's iterations with respect to anisotropy on the four used meshes.

3.2 Example 2 : heterogeneous rotating anisotropy

In this test we compare the numerical solution produced by the linear DDFV scheme and the one obtained by our nonlinear finite volume method when applying a heterogeneous rotating anisotropic tensor. Precisely, we are interested in assessing the positivity preservation for both schemes. The continuous model consists of the linear heat equation together with fully homogeneous Dirichlet boundary conditions. The convection is neglected. The diffusivity function κ is constant and the potential γ is linear i.e. $\kappa(u) = 1$ and $\gamma(u) = u$. The anisotropy matrix is given by

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

where δ accounts for the anisotropy ratio that is fixed to 10^{-3}. The initial condition is chosen as $u(x, y, 0) = \sin(\pi x)\sin(\pi y)$ and there is no contribution of the source term, i.e. $f = 0$.

Recall that the DDFV scheme is characterized by the linear fluxes

$$\mathcal{F}_{AB}(u_{\mathcal{T}}^{n+1}) = -|\sigma|\,\Lambda^{\mathcal{D}}\nabla^{\mathcal{D}}u_{\mathcal{T}}^{n+1}\cdot\mathbf{n}_{AB}, \quad \mathcal{F}_{A^\star B^\star}(u_{\mathcal{T}}^{n+1}) = -|\sigma^\star|\,\Lambda^{\mathcal{D}}\nabla^{\mathcal{D}}u_{\mathcal{T}}^{n+1}\cdot\mathbf{n}_{A^\star B^\star}.$$

The taken time step is $\triangle t = 2.3\cdot 10^{-4}$. The final time is $t_f = 0.1$. We use the third randomly disturbed mesh for the simulation.

The outputs of the test are shown on Figure 8 for three instances $t = 0.01, 0.05, 0.1$. The first row corresponds to the results of the standard DDFV scheme while the second row indicates the results of the proposed PSGFV scheme. As shown in the upper figures, a violation of the discrete maximum principle is recorded. Indeed, the red dots illustrate the positions where the DDFV solution presents negative and nonphysical values. These undershoots are accumulating in time. As in Remark 2.3, the DDFV scheme can be written in a linear matrix formulation $\mathcal{H}U_{\mathcal{T}}^{n+1} = q(U_{\mathcal{T}}^{n})$, where $\mathcal{H}$ does not have the M-matrix structure when Λ is anisotropic or the mesh is non-orthogonal. On the other hand, no oscillations are noticed in case of the new scheme. In terms of accuracy, one can clearly see that both solutions are quite similar.

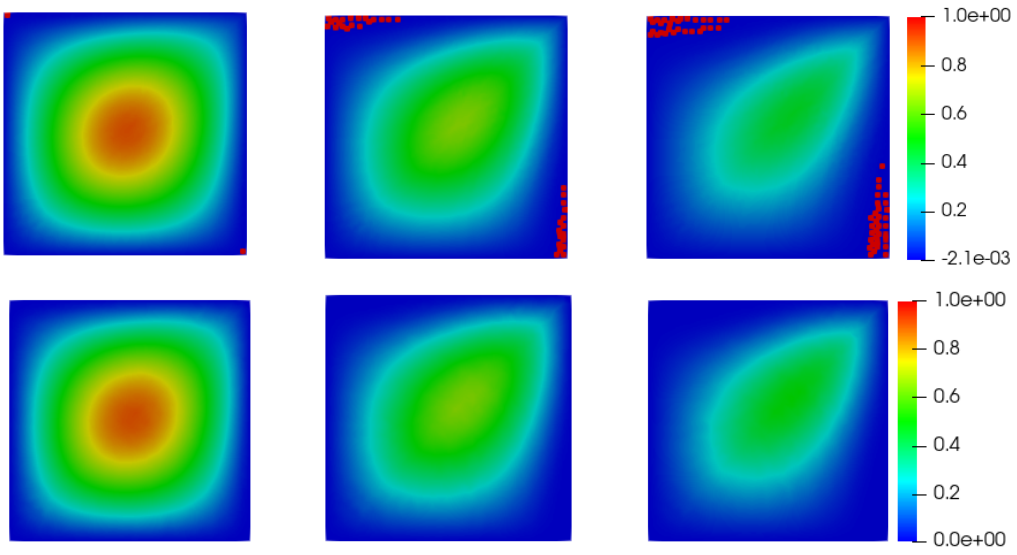

Figure 8: Solution of the heat equation computed by the linear DDFV scheme (top row) and the nonlinear PSGFV scheme (bottom row) for three times $t = 0.01, 0.05, 0.1$.

3.3 Example 3 : porous medium equation with drift

The intention of this experiment is to demonstrate how well the developed scheme can approximate the low regular solution of the degenerate porous medium equation with drift. This test-case is being inspired from [9].

For the moment, the tensor Λ is diagonal as in the test-case of Subsection 3.1. The nonlinearities are taken as

$$\kappa(u) = 2\,|u|\,, \quad \gamma(u) = u.$$

Then $\mu(u) = |u|\,u$. We put the velocity vector $\mathbf{V} = (a_x/2, 0)$. Under this setting, a one-dimensional continuous solution to the equation (1.1) is expressed by

$$u_{\text{ex}}(x, y, t) = \max(3a_x t - x, 0), \quad (x, y) \in \Omega, \quad t \in [0, tf].$$

We consider the final time $t_f = 0.2$. For accuracy measurement, the time step is computed by $\triangle t = 0.1h_{\mathcal{T}}^2$. To close the model, the boundary and initial conditions agree with the exact solution. Note that this solution is not smooth enough in space. Accordingly, the errors E_{sol} are expected to decrease as the mesh is refined with a rate which is strictly less than 2 (around 3/2). This fact is obviously concertized on Figures 9-12. Improved convergence orders are obtained for most cases, especially when using the Kershaw mesh.

In Table 2 we display the lower bound on the numerical solution as well as the behavior of the Newton solver in terms of its total iterations by mesh and anisotropic tensor. First, the algorithm does not come with an extra computational cost for low values of a_y. More iterations are required as strong anisotropy takes place. The solution remains nonnegative in all the cases, which is in excellent accordance with our expectation. This is not the case of the accurate VAG (Vertex Approximate Gradient) scheme presented in [9, Test 3] where undershoots were observed for a similar problem.

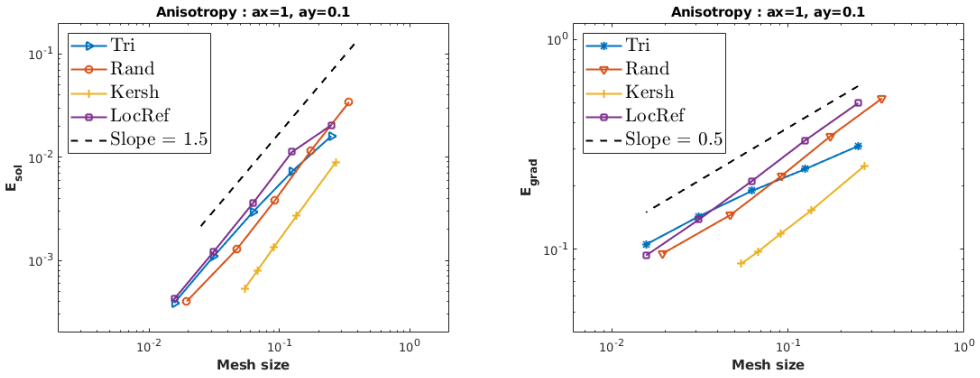

Figure 9: Accuracy results for PME with $a_y = 0.1$

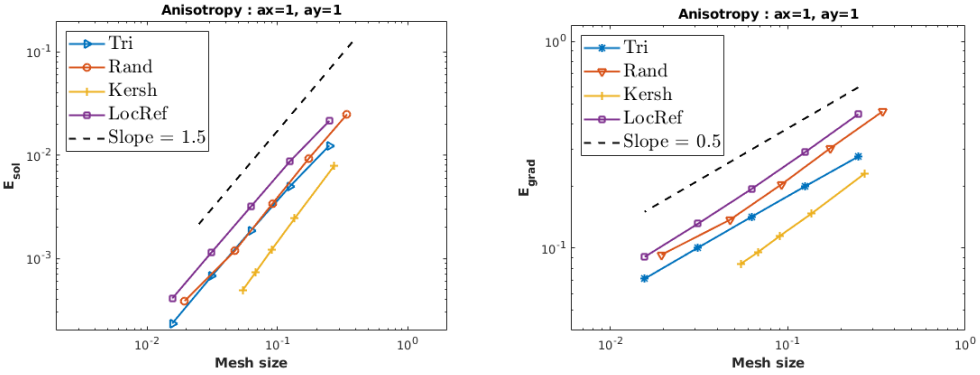

Figure 10: Accuracy results for PME with $a_y = 1$

Next, we compare the robustness of the well known DDFV scheme and the proposed PSGFV methodology. To this purpose, we keep the porous medium equation with the same

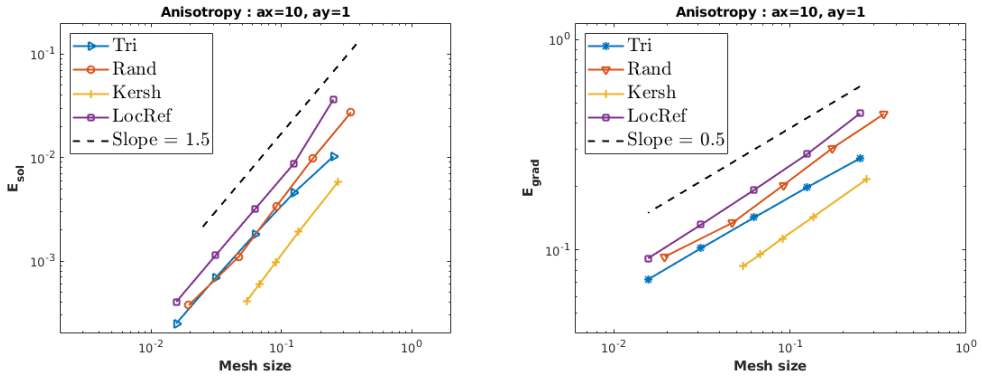

Figure 11: Accuracy results for PME with $a_y = 10$

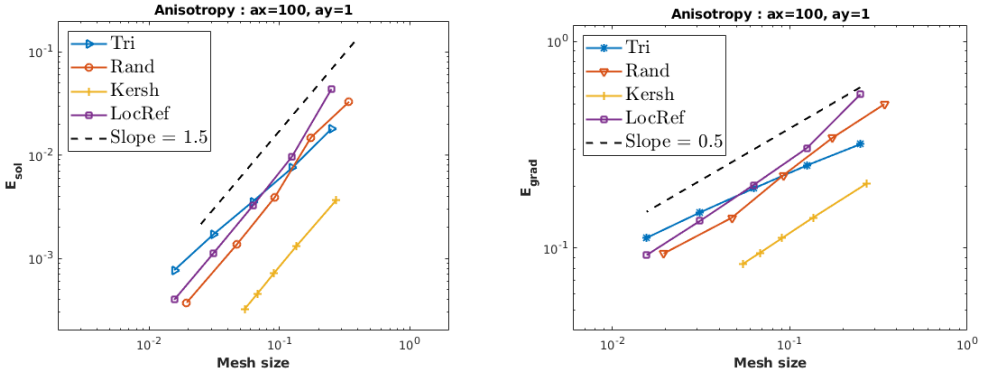

Figure 12: Accuracy results for PME with $a_y = 100$

Triangular meshes			
$a_y = 0.1$	$a_y = 1$	$a_y = 10$	$a_y = 100$
0 126	0 126	0 125	0 148
0 383	0 383	0 509	0 606
0 1535	0 1535	0 1539	0 1977
0 6144	0 6144	0 6144	0 6324
0 24577	0 24577	0 24577	0 24590

Random meshes			
$a_y = 0.1$	$a_y = 1$	$a_y = 10$	$a_y = 100$
0 66	0 66	0 66	0 74
0 197	0 97	0 261	0 262
0 710	0 710	0 764	0 930
0 2993	0 2693	0 2693	0 2843
0 15856	0 15856	0 15856	0 15856

Locally refined meshes			
$a_y = 0.1$	$a_y = 1$	$a_y = 10$	$a_y = 100$
0 96	0 100	0 126	0 122
0 383	0 383	0 403	0 491
0 1534	0 1534	0 1534	0 1642
0 6144	0 6144	0 6144	0 6144
0 24576	0 24576	0 24576	0 24576

Kershaw meshes			
$a_y = 0.1$	$a_y = 1$	$a_y = 10$	$a_y = 100$
0 112	0 112	0 117	0 137
0 435	0 435	0 435	0 439
0 827	0 838	0 978	0 978
0 1347	0 1354	0 1524	0 1742
0 2045	0 2045	0 2170	0 2524

Table 2: PME : lower bound of the numerical solution and total number of Newton's iterations with respect to anisotropy on and the level the four used meshes.

nonlinearities. The convection is neglected and the anisotropic tensor is now set to

$$\Lambda = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 20 \end{pmatrix} \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix},$$

where $\theta = \pi/8$. The initial state of the solution is considered to be $u^0 = 0$, and the imposed boundary Dirichlet conditions are chosen such that

$$u(x, y, t) = \max(2t - x, 0), \quad (x, y) \in \partial\Omega, \quad t \in [0, tf].$$

We underline that the DDFV scheme for the porous medium equation is characterized by the nonlinear fluxes

$$\mathcal{F}_{AB}(u_{\mathcal{T}}^{n+1}) = -|\sigma| \Lambda^{\mathcal{D}} \nabla^{\mathcal{D}} \mu(u_{\mathcal{T}}^{n+1}) \cdot \mathbf{n}_{AB}, \quad \mathcal{F}_{A^\star B^\star}(u_{\mathcal{T}}^{n+1}) = -|\sigma^\star| \Lambda^{\mathcal{D}} \nabla^{\mathcal{D}} \mu(u_{\mathcal{T}}^{n+1}) \cdot \mathbf{n}_{A^\star B^\star}.$$

We utilize the first element of Kershaw meshes. The simulation stops as it reaches $t_f = 0.198$ by the time step $\triangle t = 0.073$. The results are exhibited on Figure 13. Since the DDFV scheme is not positive, undershoots on the numerical solution are expected to occur as the left part of the same figure indicates. On the other hand, the PSGFV scheme is free of such instabilities. This is once again confirmed by the right part of Figure 13.

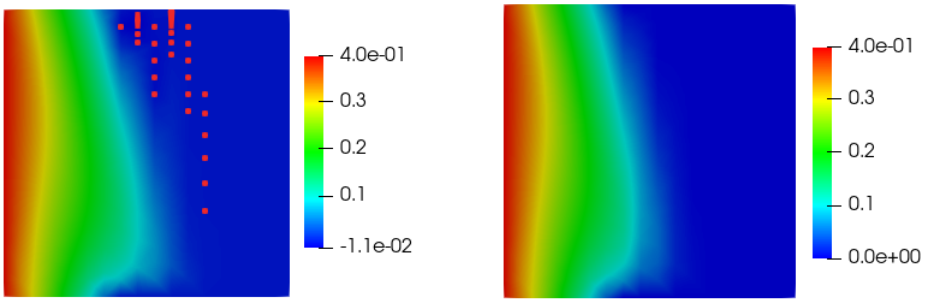

Figure 13: DDFV solution (left) and PSGFV (righ) for the anisotropic PME at $t_f = 0.198$. The red dots account for the locations of occurred undershoots.

3.4 Example 4 : quarter five spot problem

We now look at the quarter five spot problem widely known in the context of multi-phase flows in porous media. In this study, it serves to accurately capture, quantify and track the behavior of water diffusion with a heterogeneous porous medium such as wood, underground, etc, regardless the used mesh. Then, the solution may account for the water content or saturation.

The computational domain is $\Omega = (0, 1) \times (0, 1)$. It is composed of three parts defined by

$$\Omega_1 = \Omega \setminus \left(\overline{\Omega}_1 \cup \overline{\Omega}_2\right), \quad \Omega_2 = (0.75, 1) \times (0.625, 0.75), \quad \Omega_2 = (0.25, 0.75) \times (0.25, 0.375).$$

The boundary $\partial\Omega$ is split into the outflow boundary Γ^{D_l}, the injection one Γ^{D_r} and the Neumann one $\Gamma^N = \partial\Omega \setminus \left(\Gamma^{D_l} \cup \Gamma^{D_l}\right)$ where

$$\Gamma^{D_l} = \{x = 0\} \times \{y \geq 0.9\} \bigcup \{x \leq 0.1\} \times \{y = 1\},$$
$$\Gamma^{D_r} = \{x \geq 0.9\} \times \{y = 0\} \bigcup \{x = 1\} \times \{y \leq 0.1\}$$

We refer to Figure 14 for an illustration of these subsets. In this test-case, we neglect convection

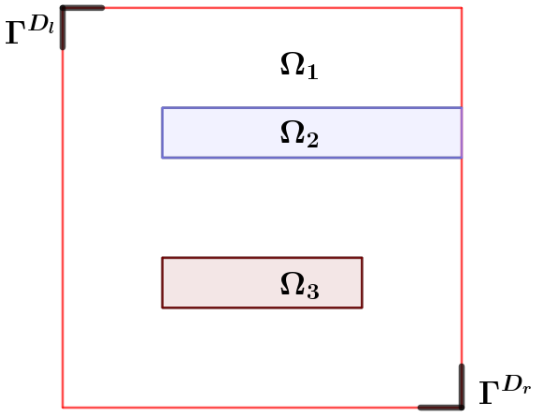

Figure 14: Schematic illustration of the porous medium with barriers Ω_2 and Ω_3 with low permeabilities. The injection zone is Γ^{D_r}. The production one is located at Γ^{D_l}. The rest of the boundary is impermeable.

effects and source term. Only diffusion is taken into account with a nonlinear diffusivity function

$$\kappa(u) = \frac{u^2}{u^2 + 8(1-u)^2},$$

and $\gamma(u) = u$. Highly heterogeneous permeabilities are considered in these subdomains as follows

$$\Lambda(x) = I_2, \quad x \in \Omega_1, \quad \Lambda(x) = \begin{pmatrix} 10^{-3} & 0 \\ 0 & 10^{-1} \end{pmatrix}, \quad x \in \Omega_2, \quad \Lambda(x) = 10^{-4} I_2, \quad x \in \Omega_3,$$

where I_2 is the identity matrix. The region of small permeability acts as a barrier.

Initially, there is no water in the medium. This amount to setting $u^0 = 0$. We put Ω in contact with water at Γ^{D_r} by imposing $u^{D_r} = 1$ in the course of time simulation. Letting an outflow condition on Γ^{D_l}, the fluid can exit the medium. Γ^{D_l} is refereed to as an exaction zone playing the role of well. The remained part Γ^N is impervious.

The domain Ω is discretized utilizing a triangular mesh made of 3584 elements and a locally refined mesh consisting of 2560 cells. The used time step is $\triangle t = 0.001$. In this experiment, we consider that the permeability is constant by cells. It is then discontinuous across the primal edges. To deal with that, we implemented the coefficient $\tau_{AB}, \tau_{A^\star B^\star}$ and $\eta_{\mathcal{D}}$ that are borrowed from the m-DDFV method. Such a strategy allows to maintain the quadratic convergence of our scheme, otherwise the optimal accuracy is only of first order. One can consult [6, 6. Examples] for more details on this method.

In Figure 15, snapshots of the saturation profile are exhibited for different times. The first row corresponds to the result on the triangular mesh whereas the second one refers to the results on the locally refined mesh. As expected, in the both cases, the fluid spreads from the injection zone towards the exit by avoiding the areas of lower permeabilities. It is clearly seen that Ω_3 takes more time to be occupied by water due its small permeability.

We here emphasize two key observations. On the one hand, no numerical instabilities such as undershoots are noticed during the simulation and the saturation honors its physical ranges, which is already justified by the scheme construction in Proposition 2.1. On the other hand, the Newton solver requires around 10 iterations to converge at the beginning while it needs 3 or 2 after that.

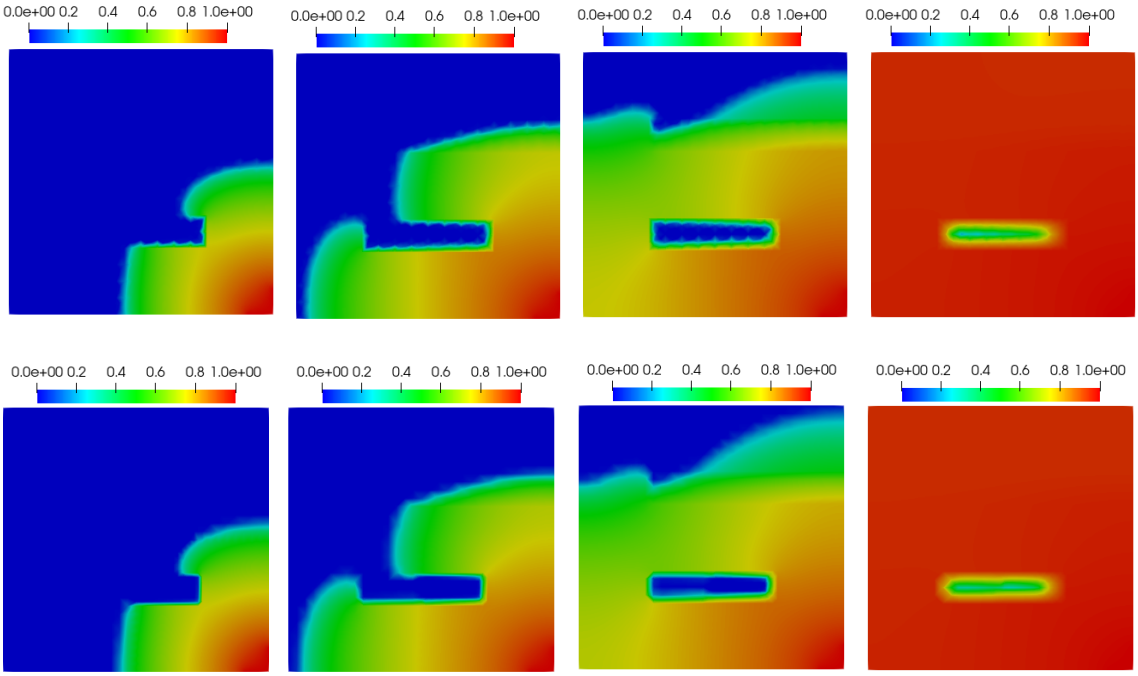

Figure 15: Saturation profile on the triangular mesh (top row) and on the locally refined mesh (bottom row) at $t = 0.5, t = 1.5$, $t = 3$ and $t = 8$.

4 Conclusion

In this work we presented an efficient and a robust positive nonlinear finite volume scheme for solving convection-diffusion equations on polygonal meshes with heterogeneous and anisotropic diffusivity tensors. The key idea consists in connecting nonlinearly the unstable tangential flux and the convection term thanks to a technique widely used in Scharfetter-Gummel approximations. As outcomes, the scheme is naturally free of numerical instabilities such as undershoots and the optimal accuracy is formally achieved in the diffusive regime independently of the mesh and anisotropy. These findings have been jointed with intensive numerical validations and illustrations focusing on the Fokker-Planck equation with a drift, the convective porous medium equation and the highly heterogeneous quarter five spot problem.

Acknowledgments: This study was carried out in the Centre Européen de Biotechnologie et de Bioéconomie (CEBB), supported by the Région Grand Est, Département de la Marne, Greater Reims and the European Union. In particular, the author would like to thank the Département de la Marne, Greater Reims, Région Grand Est and the European Union along with the European Regional Development Fund (ERDF Champagne Ardenne 2014-2020) for their financial support of the Chair of Biotechnology of CentraleSupélec.

References

[1] M. Afif and B. Amaziane. Convergence of finite volume schemes for a degenerate convection–diffusion equation arising in flow in porous media. *Computer Methods in Applied Mechanics and Engineering*, 191(46):5265–5286, 2002.

[2] B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray–Lions–type elliptic problems on general 2D meshes. *Numerical Methods for Partial Differential Equations*, 23(1):145–195, 2007.

[3] P. Angot, V. Dolejší, M. Feistauer, and J. Felcman. Analysis of a combined barycentric finite volume-nonconforming finite element method for nonlinear convection-diffusion problems. *Applications of Mathematics*, 43(4):263–310, 1998.

[4] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*. SIAM, Philadelphia, PA, USA, 1994.

[5] M. Bessemoulin-Chatard. A finite volume scheme for convection–diffusion equations with nonlinear diffusion derived from the Scharfetter–Gummel scheme. *Numerische Mathematik*, 121(4):637–670, 2012.

[6] F. Boyer and F. Hubert. Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM Journal on Numerical Analysis*, 46(6):3032–3070, 2008.

[7] K. Brenner, R. Masson, and E. H. Quenjel. Vertex Approximate Gradient Discretization preserving positivity for two-phase Darcy flows in heterogeneous porous media. *Journal of Computational Physics*, 409:109357, 2020.

[8] C. Buet and S. Dellacherie. On the Chang and Cooper scheme applied to a linear Fokker-Planck equation. *Communications in Mathematical Sciences*, 8(4):1079–1090, 2010.

[9] C. Cancès and C. Guichard. Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Foundations of Computational Mathematics*, 17(6):1525–1584, 2017.

[10] C. Chainais-Hillairet and J. Droniou. Finite-volume schemes for noncoercive elliptic problems with neumann boundary conditions. *IMA Journal of Numerical Analysis*, 31(1):61–85, 2011.

[11] B. Da Veiga, J. Droniou, and G. Manzini. A unified approach for handling convection terms in finite volumes and mimetic discretization methods for elliptic problems. *IMA Journal of Numerical Analysis*, 31(4):1357–1401, 2011.

[12] K. Domelevo and P. Omnès. A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *ESAIM: Mathematical Modelling and Numerical Analysis*, 39(6):1203–1249, 2005.

[13] J. Droniou. Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Mathematical Models and Methods in Applied Sciences*, 24(08):1575–1619, 2014.

[14] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. *The gradient discretisation method*, volume 82. Springer, 2018.

[15] L. C. Evans. *Partial differential equations*, volume 19. American Mathematical Society, 2010.

[16] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In *Handbook of Numerical Analysis*, volume 7, pages 713–1018. Elsevier, 2000.

[17] R. Eymard, T. Gallouït, R. Herbin, and A. Michel. Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. *Numerische Mathematik*, 92(1):41–82, 2002.

[18] R. Eymard, D. Hilhorst, and M. Vohralík. A combined finite volume–finite element scheme for the discretization of strongly nonlinear convection–diffusion–reaction problems on nonmatching grids. *Numerical Methods for Partial Differential Equations*, 26(3):612–646, 2010.

[19] M. Ghilani, E. H. Quenjel, and M. Saad. Positivity-preserving finite volume scheme for compressible two-phase flows in anisotropic porous media: The densities are depending on the physical pressures. *Journal of Computational Physics*, 407:109233, 2020.

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

[21] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *Journal of computational Physics*, 160(2):481–499, 2000.

[22] A. Jüngel. Numerical approximation of a drift-diffusion model for semiconductors with nonlinear diffusion. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 75(10):783–799, 1995.

[23] A. Jüngel and P. Pietra. A discretization scheme for a quasi-hydrodynamic semiconductor model. *Mathematical Models and Methods in Applied Sciences*, 7(07):935–955, 1997.

[24] I. S. Pop and W.-A. Yong. A numerical approach to degenerate parabolic equations. *Numerische Mathematik*, 92(2):357–381, 2002.

[25] E. H. Quenjel. Analysis of accurate and stable finite volume scheme for anisotropic diffusion equations with drift. *Preprint*, 2019.

[26] E. H. Quenjel. Enhanced positive vertex-centered finite volume scheme for anisotropic convection-diffusion equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 54(2):591–618, 2020.

[27] E. H. Quenjel. Nonlinear finite volume discretization for transient diffusion problems on general meshes. *Applied Numerical Mathematics*, 161:148–168, 2021.

[28] E. H. Quenjel, M. Saad, M. Ghilani, and M. Bessemoulin-Chatard. Convergence of a positive nonlinear DDFV scheme for degenerate parabolic equations. *Calcolo*, 57(19), 2020.

[29] D. L. Scharfetter and H. K. Gummel. Large-signal analysis of a silicon read diode oscillator. *IEEE Transactions on electron devices*, 16(1):64–77, 1969.