A monotone numerical flux for quasilinear convection diffusion equation

Claire Chainais-Hillairet, Robert Eymard, Jürgen Fuhrmann

A monotone numerical flux for quasilinear convection diffusion equation

C. Chainais-Hillairet, R. Eymard and J. Fuhrmann

Abstract

We propose a new numerical 2-point flux for a quasilinear convection–diffusion equation. This numerical flux is shown to be an approximation of the numerical flux derived from the solution of a two-point Dirichlet boundary value problem for the projection of the continuous flux onto the line connecting neighboring collocation points. The later approach generalizes an idea first proposed by Scharfetter and Gummel for linear drift-diffusion equations. We establish first that the new flux satisfies sufficient properties ensuring the convergence of the associate finite volume scheme, while respecting the maximum principle. Then, we pay attention to the long time behavior of the scheme: we show relative entropy decay properties satisfied by the new numerical flux as well as by the generalized Scharfetter-Gummel flux. The proof of these properties uses a generalization of some discrete (and continuous) log-Sobolev inequalities. The corresponding decay of the relative entropy of the continuous solution is proved in the appendix. Some 1D numerical experiments confirm the theoretical results.

Keywords: Quasilinear convection–diffusion equation, Scharfetter–Gummel flux, long time behavior, log-Sobolev inequalities.

2021 AMS Classification: 65N08, 65N12, 35A23, 35B40

1 Introduction

Many problems arising in physics, biology or engineering involve convection-diffusion equations. They consist of conservation laws of the form

$$u_t + \mathrm{div}\boldsymbol{J} = 0, \text{ with } \boldsymbol{J} = -\nabla\zeta(u) + \eta(u)\boldsymbol{q}. \tag{1}$$

Under this general form, the diffusion as well as the convection are assumed to be nonlinear functions, respectively ζ and η, of the unknown u, which may represent for instance the density of a given species. In general, the equation (1) is set on a bounded domain $\Omega \subset \mathbb{R}^d$ ($d \geq 1$) and supplemented with initial and boundary conditions. We will assume that the boundary $\partial\Omega$ of the domain splits into two parts Γ^D and Γ^N and consider Dirichlet boundary conditions on Γ^D and no-flux boundary conditions on Γ^N.

Let us give some examples of applications leading to (1). The porous media equation, which describes the flow of a gas through a porous interface, may be rewritten with a time dependent-scaling into the nonlinear form (1) with $\zeta(u) = u^m$, $m > 1$ and $\boldsymbol{q} = -\boldsymbol{x}$, see [7]. In order to describe the dynamics of bosons and fermions, Kaniadakis in [18] considers linear diffusion $\zeta(u) = u$ with nonlinear convection $\eta(u) = u(1 + ku)$ ($k = 1$ for the bosons, $k = -1$ for the fermions) with $\boldsymbol{q} = -\boldsymbol{x}$. This equation has been studied later by Carrillo *et al.* in [6, 5]. In order

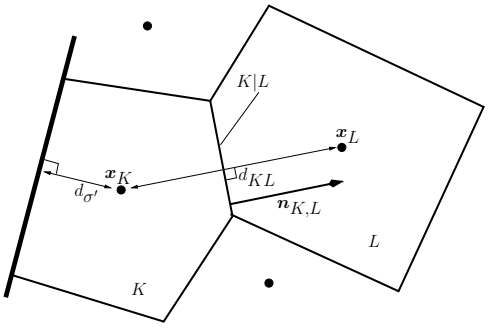

Fig. 1: Two control volumes in a finite volume scheme.

to describe electrochemical processes, Onsager in [20, 21], describes the flux $\boldsymbol{J}$ as proportional to the gradient of an electrochemical potential. It means that

$$\boldsymbol{J} = -\eta(u)\nabla\left(\mu(u) + z_u V(\boldsymbol{x})\right), \tag{2}$$

where $\mu(u)$ defines the chemical potential of a given charged species, z_u its charge, $\eta(u)$ its mobility and $V(\boldsymbol{x})$ an external electrical potential. The same definition of the flux is used in the modeling of semiconductor devices. In this framework, the mobility is assumed to be linear $\eta(u) = u$ and the quantity $\mu(u) + z_u V(\boldsymbol{x})$ is referred as the quasi-Fermi potential of the charged species (electrons or holes), see for instance the seminal papers by Gajewski and Gröger [14, 15] . Moreover, in the modeling of semiconductor devices, the electrical potential V is defined through a Poisson equation, so that (1) is just a part of a coupled system of equations.
In this paper, we are interested in the numerical approximation of (1) by finite volume schemes based on a two-point flux approximation (the notations will be introduced in Section 3.2 and are already shown on Figure 1). The discretization in time is a backward Euler discretization, with a time step Δt. Such a scheme has the generic form

$$\mathrm{m}_K \frac{u_K^{n+1} - u_K^n}{\Delta t} + \sum_{\sigma \in \mathcal{E}_{K,\mathrm{int}} \cup \mathcal{E}_K^D} \mathrm{m}_\sigma \mathcal{F}_{K\sigma}^{n+1} = 0. \tag{3}$$

In (3), the quantity $\mathcal{F}_{K\sigma}^{n+1}$ represents the numerical flux through the edge σ of the control volume K, outward K. It should be a consistent and conservative approximation of

$$\frac{1}{\mathrm{m}_\sigma} \int_\sigma \boldsymbol{J} \cdot n_{K,\sigma}.$$

When we consider two-point flux approximations, $\mathcal{F}_{K\sigma}^{n+1}$ is defined as a function of u_K^{n+1} and u_L^{n+1} if σ is the common edge to the control volumes K and L (we will denote $\sigma = K|L$, and $\sigma \in \mathcal{E}_{K,\mathrm{int}}$). It is a function of u_K^{n+1} and u_σ^D if σ is an edge of K included in the Dirichlet boundary Γ^D (we will denote $\sigma \in \mathcal{E}_K^D$ in this case and define by u_σ^D an approximation of the Dirichlet data u^D on σ).
In the linear case ($\eta(u) = u$, $\zeta(u) = u$), different choices for the numerical fluxes have already been proposed and studied. As shown by Chainais-Hillairet and Droniou in [8], some of them

may be written under a generic form:

$$\mathcal{F}_{K\sigma}^{n+1} = \mathcal{F}(u_K^{n+1}, u_{K\sigma}^{n+1}, q_{K,\sigma}, d_\sigma), \tag{4}$$

$$\text{with } \mathcal{F} : \mathbb{R}^3 \times \mathbb{R}_+ \to \mathbb{R} \text{ such that } \mathcal{F}(a,b,q,h) = \frac{aB(-qh) - bB(qh)}{h} \tag{5}$$

$$\text{letting } u_{K\sigma}^{n+1} = u_L^{n+1} \text{ if } \sigma = K|L \text{ and } u_{K\sigma}^{n+1} = u_\sigma^D \text{ if } \sigma \in \mathcal{E}_K^D, \tag{6}$$

and where $q_{K,\sigma}$ is defined by

$$q_{K,\sigma} = \frac{1}{\mathrm{m}_\sigma} \int_\sigma \boldsymbol{q} \cdot \boldsymbol{n}_{K\sigma} \mathrm{d}s \quad \forall \sigma \in \mathcal{E}_K$$

and d_σ is introduced on Figure 1. With a classical two-point flux approximation of the linear diffusion term, an upwinding of the convection term corresponds to the case $B(s) = 1 + (-s)\top 0$, while a centered approximation corresponds to the case $B(s) = 1 - s/2$. The choice of the Bernoulli function, $B(s) = s/(e^s - 1)$, leads to the Scharfetter-Gummel numerical fluxes introduced in [22].

The main advantage of the Scharfetter-Gummel numerical fluxes is that they preserve the thermal equilibria. Indeed, when $\boldsymbol{q} = -\nabla V$, if u is proportional to e^{-V}, the flux vanishes ($\boldsymbol{J} = 0$) and this property is preserved at the discrete level: if B is the Bernoulli function, $q_{K,\sigma}$ is defined by $q_{K,\sigma} = (V_L - V_K)/d_\sigma$, $u_K^{n+1} = e^{-V_K}$ and $u_L^{n+1} = e^{-V_L}$, then $\mathcal{F}_{K\sigma}^{n+1} = 0$ for $\sigma = K|L$. Recently, Heida, Kantner and Stephan in [16] have proposed a family of fluxes which are designed for the preservation of thermal equilibria. These fluxes have the generic form (4), (5), (6) with B defined by $B(x) = S_{\alpha,\beta}(1, e^{-x})$, where $S_{\alpha,\beta}$ is a Stolarsky mean defined by

$$S_{\alpha,\beta}(x,y) = \left(\frac{\beta}{\alpha} \frac{x^\alpha - y^\alpha}{x^\beta - y^\beta} \right)^{\frac{1}{\alpha-\beta}}, \text{ for } \alpha \neq 0, \beta \neq 0, \alpha \neq \beta, x \neq y,$$

and extended in a continuous way in the critical points. With $\alpha = 0$ and $\beta = -1$, one recovers the Scharfetter-Gummel fluxes.

As shown by Lazarov, Mishev and Vassilevski in [19], the Scharfetter-Gummel scheme has an order 2 in space, as the centered scheme but without any stability assumption on the Péclet number. Therefore, the question of its extension to nonlinear equations seems of great interest and has already given rise to numerous studies. Let us mention the works by Jüngel and Pietra [17] and by Bessemoulin-Chatard [1] where only the diffusion is assumed to be nonlinear, by Eymard, Fuhrmann and Gärtner [10], and by Bessemoulin-Chatard and Filbet [2] which deal with nonlinear diffusion and convection. Except in [17], all the schemes preserve thermodynamical equilibria. The only scheme which is second order in space even when the diffusion degenerates in the one proposed in [2].

As we will recall in Section 2.2, the construction of the scheme proposed by Eymard, Fuhrmann and Gärtner in [10] (we will denote it in what follows SGnl-scheme) follows the main lines of the original idea by Scharfetter and Gummel. It leads to numerical fluxes defined as the solution of a nonlinear boundary-value problem for each interface of the mesh. They are therefore defined in an implicit way, which is the main drawback of the SGnl-scheme and which limits its use.

Therefore, the aim of this paper is to introduce and analyze a numerical scheme for (1) based on the SGnl-scheme but which offers an easy computation of the numerical fluxes. The new numerical fluxes will consist in an approximation of the numerical fluxes defined in the SGnl-scheme, up to a precision δ not depending on the size of the mesh.

The organization of the paper is the following. In Section 2, we analyze both the SGnl-flux and the δ-approximation of this flux. We prove that the δ-approximation satisfies a series of properties gathered in Definition 2.1, which are sufficient for yielding the convergence of a finite volume scheme based on this flux to the transient solution of a convection-diffusion problem. This first result is stated in Proposition 2.9. We then turn to numerical analysis of the scheme in Section 3 in the special case of fluxes defined by (2). After the proof of standard properties (including the existence and uniqueness of a discrete solution to the scheme), we turn to the study of the long time behavior of the scheme in Section 3.3. This study extends the one which is done in [3] to the more general choices for η and μ than $\eta(s) = s$ and $\mu(s) = \log(s)$. New difficulties must then be handled, like the proof of the existence of a strictly positive lower bound of the discrete solution, or the adaptation of the mean discrete Poincaré inequality to a nonlinear setting (log-Sobolev inequalities no longer hold in this case). We establish in Theorem 3.11 the exponential decay of the discrete solution to the scheme towards a discrete thermal equilibrium, up to the δ-approximation of the SGnl-flux. Finally, a series of numerical results in Section 4 show that the theoretical result of approximate exponential decay is observed in practice.

In an appendix, we briefly state the exponential decay properties which hold for the long term behavior of the continuous problem involving the same general framework for η and μ as the one which is studied in Section 3.

Notation for the whole paper: for $a, b \in \mathbb{R}$, we denote by $a \top b$ (resp. $a \perp b$) the maximum (resp. the minimum) between a and b. We also denote by $a^+ = \max(a, 0)$ the positive part of a and $a^- = \max(-a, 0)$ its negative part.

2 Numerical fluxes

2.1 General framework

In the spirit of (5), we investigate the definition and the needed properties for a function $\mathcal{F} : \mathbb{R}^3 \times \mathbb{R}_+ \to \mathbb{R}$ used in the definition of numerical fluxes (4), for nonlinear functions η and ζ. Let us first specify the hypotheses we consider on the nonlinearities η and ζ involved in (1).

Assumption 2.1:

i) $\eta \in \mathcal{C}(\mathbb{R}, \mathbb{R})$ is a Lipschitz continuous function and we denote by L_η its Lipschitz constant.

ii) $\zeta \in \mathcal{C}^1(\mathbb{R}, \mathbb{R})$ is a Lipschitz continuous function with a Lipschitz constant denoted by L_ζ. We assume moreover:
$$\exists r > 0 \text{ such that } \zeta'(s) \geq r \quad \forall s \in \mathbb{R}.$$
We also define $\xi : \mathbb{R} \to \mathbb{R}$ by $\xi(s) = \displaystyle\int_0^s \sqrt{\zeta'(t)}\mathrm{d}t$.

The convergence analysis of the finite volume scheme proposed in [10] relies on some sufficient properties satisfied by the numerical fluxes. We recall these properties in Definition 2.1.

Definition 2.1 (Admissible numerical flux): Under Assumption 2.1, the function $\mathcal{F} : (a, b, q, h) \in \mathbb{R}^3 \times \mathbb{R}_+ \mapsto \mathcal{F}(a, b, q, h) \in \mathbb{R}$ defines an admissible numerical flux if it satisfies the following properties:

i) $\mathcal{F}$ is Lipschitz-continuous with respect to a and b.

ii) $\mathcal{F}$ is increasing with respect to a, decreasing with respect to b.

iii) $\mathcal{F}(a,b,q,h) + \mathcal{F}(b,a,-q,h) = 0$ for all $(a,b,q,h) \in \mathbb{R}^3 \times \mathbb{R}_+$.

iv) There exists $c \in [a\perp b, a\top b]$ such that $\mathcal{F}(a,b,q,h) = q\eta(c) - \dfrac{\zeta(b)-\zeta(a)}{h}$.

v) There holds $(a-b)\mathcal{F}(a,b,q,h) \geq -\displaystyle\int_a^b q\eta(s)\mathrm{d}s + \frac{(\xi(b)-\xi(a))^2}{h}$.

Remark 2.2: In the B-schemes for the case $\eta(s) = s$ and $\zeta(s) = \xi(s) = s$, the function $\mathcal{F}$ is defined by (5), for a given function B. The following assumptions are sufficient conditions to ensure that the associated flux $\mathcal{F}$ defined by (5) is admissible in the sense of Definition 2.1:

$$\begin{aligned} &B \text{ is a nonnegative and nonincreasing Lipschitz continuous function,} \\ &B(0) = 1, \\ &B(s) - B(-s) = -s \text{ for all } s \in \mathbb{R}, \\ &B(s) \geq 1 - s/2 \text{ for all } s \in \mathbb{R}. \end{aligned} \tag{7}$$

Let us mention that the functions defined by $B(s) = 1 + (-s)\top 0$ (leading to the upwind fluxes) and $B(s) = s/(e^s - 1)$ (leading to the Scharfetter-Gummel fluxes) verify (7). The centered fluxes, for which $B(s) = 1 - s/2$, satisfy the positivity assumption only for $s \leq 2$, so that the convergence analysis needs an additional condition on the meshsize h related to the velocity field (Péclet condition).
Let us also recall that, in the case of the Scharfetter-Gummel fluxes (which means $B(s) = s/(e^s - 1)$), then $\mathcal{F}(a,b,q,h)$ defined by (5) is equal to the constant value of $qy(s) - y'(s)$ for $s \in [0,h]$, when $y : [0,h] \to [a\perp b, a\top b]$ is the solution to the boundary-value problem

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}s}\Big(q\ y(s) - y'(s)\Big) = 0, \\ y(0) = a, \\ y(h) = b, \end{cases} \tag{8}$$

as it is initially noticed in [22].

2.2 The nonlinear Scharfetter-Gummel numerical fluxes

The numerical fluxes, introduced by Eymard, Fuhrmann and Gärtner in [10], are an adaptation of the initial construction of the Scharfetter-Gummel fluxes, based on the solution to (8), to the nonlinear case ; they will be called "SGnl-fluxes" in this paper. They are obtained by replacing Problem (8) by the following one: search for the solution $y : [0,h] \to [a\perp b, a\top b]$ to the following boundary-value problem

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}s}\Big(q\ \eta(y(s)) - (\zeta(y))'(s)\Big) = 0, \\ y(0) = a, \\ y(h) = b. \end{cases} \tag{9}$$

Then $\mathcal{F}(a,b,q,h)$ is defined as the constant value of the expression $q\ \eta(y(s)) - (\zeta(y))'(s)$. Let us start with recalling how the function $\mathcal{F}$ can be obtained.
We first remark that when $a = b$, $y(s) = a$ is the solution to (9), so that

$$\mathcal{F}(a,a,q,h) = q\eta(a). \tag{10}$$

When $a \neq b$, if y is a solution to (9), then $z(t) = y(h-t)$ verifies

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}\Big(- q\ \eta(z(t)) - (\zeta(z))'(t)\Big) = 0, \\ z(0) = b, \\ z(h) = a. \end{cases}$$

This implies the conservativity of the numerical fluxes given in Definition 2.1 iii). Then, it is sufficient to define the numerical fluxes for $a < b$ following [10]. Therefore, let us recall the definition of the Godunov flux $\mathcal{F}^{(q)}_{\mathrm{god}}(a,b)$ for the approximation of the convective term:

$$\mathcal{F}^{(q)}_{\mathrm{god}}(a,b) = \begin{cases} \min\limits_{s\in[a,b]} \ q\eta(s) \text{ if } a \leq b, \\ \max\limits_{s\in[b,a]} \ q\eta(s) \text{ if } b \leq a. \end{cases} \tag{11}$$

In the case $a < b$, the solution y to (9) is such that $y'(s) > 0$ (see Lemma 3.1 in [10]), which implies

$$\mathcal{F}(a,b,q,h) < q\eta(y(s)) \text{ for all } s \in [0,h]$$

and

$$\mathcal{F}(a,b,q,h) < \mathcal{F}^{(q)}_{\mathrm{god}}(a,b).$$

It is then proved that $\mathcal{F}(a,b,q,h)$ is such that there holds

$$\int_0^h \frac{\zeta'(y(s))y'(s)}{q\eta(y(s)) - \mathcal{F}(a,b,q,h)}\mathrm{d}s = h.$$

Applying the change of variable $s \to y(s)$, we define for any $x \in (-\infty, \mathcal{F}^{(q)}_{\mathrm{god}}(a,b))$

$$H(x) = \int_a^b \frac{\zeta'(s)}{q\eta(s) - x}\mathrm{d}s, \tag{12}$$

and $\mathcal{F}(a,b,q,h)$ is defined as the unique solution to the nonlinear equation:

$$H(\mathcal{F}(a,b,q,h)) = h. \tag{13}$$

The following result is proved in [10].

Lemma 2.3: Under Assumption 2.1, the flux $\mathcal{F}(a,b,q,h)$ is uniquely defined as the solution to (13) if $a < b$, by (10) if $a = b$ and by the conservativity relation iii) if $a > b$. Moreover it is an admissible flux in the sense of Definition 2.1.

2.3 Approximation of $\mathcal{F}(a,b,q,h)$: the δ-fluxes $\mathcal{F}_\delta(a,b,q,h)$

As stated in the previous section, there is generally no clear way for getting an explicit expression of $\mathcal{F}(a,b,q,h)$ solution to (13). Therefore we consider in this paper a new method for defining a numerical flux, which can approximate as closely as desired the expression given by (13), and which is easy to compute.

Let $(\overline{y}_i)_{i\in\mathbb{Z}} \subset \mathbb{R}$ be a given sequence, independent of a, b, q, h, such that:

1. the sequence $(\overline{y}_i)_{i\in\mathbb{Z}}$ is strictly increasing,
2. $\overline{y}_i$ tends to $\pm\infty$ as $i \to \pm\infty$,
3. $\sup_{i\in\mathbb{Z}}(\overline{y}_{i+1} - \overline{y}_i) = \delta \in (0, +\infty)$.

We then use δ in an abuse of notation for indexing the new flux, denoted by $\mathcal{F}_\delta(a,b,q,h)$, for $q \in \mathbb{R}$ and $h > 0$ (it should be indexed by the whole sequence $(\overline{y}_i)_{i\in\mathbb{Z}}$).
In the case $a = b$, we let

$$\mathcal{F}_\delta(a,a,q,h) = q\eta(a). \tag{14}$$

Moreover, we impose the conservativity relation iii) for $\mathcal{F}_\delta$, so that it is sufficient to consider now that $a < b$.
Let $i_a \in \mathbb{Z}$ be such that $a \in [\overline{y}_{i_a}, \overline{y}_{i_a+1})$ and let $i_b \in \mathbb{Z}$ be such that $b \in (\overline{y}_{i_b}, \overline{y}_{i_b+1}]$. If $i_a = i_b$, we let $N = 1$, and we define $y_0 = a$, $y_1 = b$. Otherwise, we have $i_a < i_b$, and we define $N = 1 + i_b - i_a$, $y_0 = a, y_1 = \overline{y}_{i_a+1}, \dots, y_{N-1} = \overline{y}_{i_b}, y_N = b$.
Let us then define the function $H_\delta(x)$, for any $x \in (-\infty, \mathcal{F}^{(q)}_{\rm god}(a,b))$, by

$$H_\delta(x) = \sum_{i=0}^{N-1} \frac{\zeta(y_{i+1}) - \zeta(y_i)}{\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - x}, \tag{15}$$

where we recall that $\mathcal{F}^{(q)}_{\rm god}$ is defined by (11), and therefore satisfies $\mathcal{F}^{(q)}_{\rm god}(a,b) \le \mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1})$ for all $i = 0, \dots, N-1$. Then $\mathcal{F}_\delta(a,b,q,h)$ is defined as the solution to the nonlinear equation:

$$h = H_\delta(\mathcal{F}_\delta(a,b,q,h)). \tag{16}$$

The aim of this section is now to establish that the δ-fluxes are well defined and that they are admissible in the sense of Definition 2.1, see Proposition 2.9. We start with Lemma 2.4 which states the existence and uniqueness of the δ-flux $\mathcal{F}_\delta(a,b,q,h)$.

Lemma 2.4: Let assume Assumption 2.1 and $a < b$. Then the function H_δ defined by (15) is strictly increasing and strictly convex, and there holds

$$\lim_{x\to-\infty} H_\delta(x) = 0 \text{ and } \lim_{x \overset{>}{\to} \mathcal{F}^{(q)}_{\rm god}(a,b)} H_\delta(x) = +\infty. \tag{17}$$

As a consequence, $\mathcal{F}_\delta(a,b,q,h)$ is well defined as the unique solution to the equation $H_\delta(x) = h$.

Proof. From the definition of H_δ, we compute the first and second order derivatives of H_δ:

$$H'_\delta(x) = \sum_{i=0}^{N-1} \frac{\zeta(y_{i+1}) - \zeta(y_i)}{(\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - x)^2} \quad \text{and} \quad H''_\delta(x) = 2\sum_{i=0}^{N-1} \frac{\zeta(y_{i+1}) - \zeta(y_i)}{(\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - x)^3}.$$

As ζ is a strictly increasing function,we deduce that H_δ is strictly increasing and strictly convex on $(-\infty, \mathcal{F}^{(q)}_{\rm god}(a,b))$. The limit of H_δ in $-\infty$ is straightforward. We now remark that there exists $i_0 \in \{0,\dots,N-1\}$ such that

$$\mathcal{F}^{(q)}_{\rm god}(y_{i_0}, y_{i_0+1}) = \mathcal{F}^{(q)}_{\rm god}(a,b),$$

since $\bigcup_{i=0}^{N-1} [y_i, y_{i+1}] = [a,b]$. We therefore have

$$H_\delta(x) \geq \frac{\zeta(y_{i_0+1}) - \zeta(y_{i_0})}{\mathcal{F}^{(q)}_{\rm god}(a,b) - x},$$

which provides the limit of H_δ in $\mathcal{F}^{(q)}_{\rm god}(a,b)$. ∎

Letting $\delta \to +\infty$ in the definition of the function $\mathcal{H}_\delta$ yields

$$H_\infty(x) = \frac{\zeta(b) - \zeta(a)}{\mathcal{F}^{(q)}_{\rm god}(a,b) - x},$$

so that we may define the associate flux $\mathcal{F}_\infty(a,b,q,h)$ by

$$\mathcal{F}_\infty(a,b,q,h) = \mathcal{F}^{(q)}_{\rm god}(a,b) - \frac{\zeta(b) - \zeta(a)}{h}. \tag{18}$$

Lemma 2.5 brings now a comparison between the δ-flux and its limits $\delta \to \infty$, $\mathcal{F}_\infty(a,b,q,h)$, and $\delta \to 0$, $\mathcal{F}(a,b,q,h)$.

Lemma 2.5: Letting $Q \geq |q|$, and $a \leq b$, there holds

$$\mathcal{F}_\infty(a,b,q,h) \leq \mathcal{F}_\delta(a,b,q,h) \leq \mathcal{F}(a,b,q,h) \leq \mathcal{F}^{(q)}_{\rm god}(a,b) - \delta_F,$$

with

$$\delta_F = \frac{QL_\eta(b-a)}{2(\exp(\frac{QL_\eta h}{r}) - 1)}.$$

Proof. Thanks to (11) and (18), we first notice that

$$\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - \mathcal{F}_\infty(a,b,q,h) \geq \frac{\zeta(b) - \zeta(a)}{h}, \quad \forall i \in \mathbb{Z},$$

so that the definition (15) of H_δ, together with the definition (18) of $\mathcal{F}_\infty$, implies

$$H_\delta(\mathcal{F}_\infty(a,b,q,h)) \leq h.$$

We then deduce the first inequality $\mathcal{F}_\infty(a,b,q,h)) \leq \mathcal{F}_\delta(a,b,q,h)$ from the monotonicity of H_δ. Using $\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) \leq q\eta(s)$ for all $s \in [y_i, y_{i+1}]$, we have

$$\sum_{i=0}^{N-1} \int_{y_i}^{y_{i+1}} \frac{\zeta'(s)}{\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - \mathcal{F}_\delta(a,b,q,h)} \mathrm{d}s \geq \sum_{i=0}^{N-1} \int_{y_i}^{y_{i+1}} \frac{\zeta'(s)}{q\eta(s) - \mathcal{F}_\delta(a,b,q,h)} \mathrm{d}s,$$

which can be rewritten $H_\delta(\mathcal{F}_\delta(a,b,q,h)) \geq H(\mathcal{F}_\delta(a,b,q,h))$. But $H_\delta(\mathcal{F}_\delta(a,b,q,h)) = h = H(\mathcal{F}(a,b,q,h))$, so that the monotonicity of H provides the second inequality $\mathcal{F}_\delta(a,b,q,h) \leq \mathcal{F}(a,b,q,h)$.
It remains to prove the last inequality. Let us set $e = \mathcal{F}^{(q)}_{\rm god}(a,b) - \mathcal{F}(a,b,q,h)$. We already know that $e \geq 0$ but want to establish that $e \geq \delta_F$. Let consider $s_0 \in [a,b]$ such that $\mathcal{F}^{(q)}_{\rm god}(a,b) = q\eta(s_0)$, we have that, for all $s \in [a,b]$,

$$q\eta(s) - \mathcal{F}(a,b,q,h) = q(\eta(s) - \eta(s_0)) + e$$

and

$$q\eta(s) - \mathcal{F}(a,b,q,h) \leq QL_\eta|s - s_0| + e.$$

As $h = H(\mathcal{F}(a,b,q,h))$ and $\zeta'(s) \geq r$ for all $s \in \mathbb{R}$, this yields

$$h \geq \int_a^b \frac{r}{QL_\eta|s - s_0| + e}\mathrm{d}s,$$

and

$$h \geq r\int_0^{\frac{b-a}{2}} \frac{1}{QL_\eta x + e}\mathrm{d}x = \frac{r}{QL_\eta}\log\frac{QL_\eta\frac{b-a}{2} + e}{e}.$$

We then deduce the expected inequality:

$$e \geq \frac{QL_\eta(b-a)}{2(\exp(\frac{QL_\eta h}{r}) - 1)}.$$

■

We give with Lemma 2.6 a bound of the distance between $\mathcal{F}(a,b,q,h)$ and $\mathcal{F}_\delta(a,b,q,h)$ in term of δ.

Lemma 2.6: There exists $C_F > 0$, only depending on Q, M, L_η, L_ζ and r where $M \geq h$ and $Q > |q|$, such that, for $a \leq b$,

$$0 \leq \mathcal{F}(a,b,q,h) - \mathcal{F}_\delta(a,b,q,h) \leq C_F\delta, \tag{19}$$

which therefore implies

$$\forall a,b \in \mathbb{R},\ (a-b)(\mathcal{F}_\delta(a,b,q,h) - \mathcal{F}(a,b,q,h)) \geq 0. \tag{20}$$

Proof. The lower bound in (19) has already been established in Lemma 2.5 and implies (20) due to the conservativity of the fluxes. It remains to prove the upper bound in (19).
We have $H(\mathcal{F}(a,b,q,h)) - H_\delta(\mathcal{F}_\delta(a,b,q,h)) = 0$, which can be rewritten as

$$\sum_{i=0}^{N-1}\int_{y_i}^{y_{i+1}} \zeta'(s)\left(\frac{\left(\mathcal{F}^{(q)}_{\rm god}(y_i,y_{i+1}) - q\eta(s)\right) + \left(\mathcal{F}(a,b,q,h) - \mathcal{F}_\delta(a,b,q,h)\right)}{\left(q\eta(s) - \mathcal{F}(a,b,q,h)\right)\left(\mathcal{F}^{(q)}_{\rm god}(y_i,y_{i+1}) - \mathcal{F}_\delta(a,b,q,h)\right)}\right)\mathrm{d}s = 0.$$

Remarking that $|q\eta(s) - \mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1})| \leq QL_\eta\delta$ for all $s \in [y_i, y_{i+1}]$ and introducing

$$A = \sum_{i=0}^{N-1} \int_{y_i}^{y_{i+1}} \left(\frac{\zeta'(s)}{\Big(q\eta(s) - \mathcal{F}(a,b,q,h)\Big)\Big(\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - \mathcal{F}_\delta(a,b,q,h)\Big)} \right) \mathrm{d}s,$$
$$B = \sum_{i=0}^{N-1} \int_{y_i}^{y_{i+1}} \left(\frac{\zeta'(s)QL_\eta}{\Big(q\eta(s) - \mathcal{F}(a,b,q,h)\Big)\Big(\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - \mathcal{F}_\delta(a,b,q,h)\Big)} \right) \mathrm{d}s,$$

we obtain that $A|\mathcal{F}_\delta(a,b,q,h) - \mathcal{F}(a,b,q,h)| \leq \delta B$. But, from Lemma 2.5, we have

$$\begin{aligned} &\mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - \mathcal{F}_\delta(a,b,q,h) \geq \mathcal{F}^{(q)}_{\rm god}(a,b) - \mathcal{F}_\delta(a,b,q,h) \geq \delta_F \\ \text{and } &q\eta(s) - \mathcal{F}(a,b,q,h) \geq \mathcal{F}^{(q)}_{\rm god}(a,b) - \mathcal{F}(a,b,q,h) \geq \delta_F \quad \forall s \in [y_i, y_{i+1}], \end{aligned} \tag{21}$$

so that

$$B \leq \frac{QL_\eta}{\delta_F^2}(\zeta(b) - \zeta(a)). \tag{22}$$

We also have that

$$\begin{aligned} &0 \leq \mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - \mathcal{F}_\delta(a,b,q,h) \leq \mathcal{F}^{(q)}_{\rm god}(y_i, y_{i+1}) - \mathcal{F}_\infty(a,b,q,h) \leq QL_\eta|b-a| + \frac{\zeta(b)-\zeta(a)}{h}, \\ \text{and } &0 \leq q\eta(s) - \mathcal{F}(a,b,q,h) \leq QL_\eta|b-a| + \frac{\zeta(b)-\zeta(a)}{h} \quad \forall s \in [a,b], \end{aligned} \tag{23}$$

yielding

$$A \geq \frac{\zeta(b) - \zeta(a)}{\Big(\frac{\zeta(b)-\zeta(a)}{h} + QL_\eta(b-a)\Big)^2}. \tag{24}$$

From (22) and (24), we deduce that

$$\begin{aligned} |\mathcal{F}_\delta(a,b,q,h) - \mathcal{F}(a,b,q,h)| &\leq \delta\frac{QL_\eta}{\delta_F^2}\left(\frac{\zeta(b)-\zeta(a)}{h} + QL_\eta(b-a)\right)^2 \\ &\leq \delta QL_\eta\Big(\frac{\zeta(b)-\zeta(a)}{h} + QL_\eta(b-a)\Big)^2 \frac{4\Big(\exp(\frac{QL_\eta h}{r}) - 1\Big)^2}{\Big(QL_\eta(b-a)\Big)^2}, \\ &\leq \delta\frac{4}{QL_\eta}\Big(\exp(\frac{QL_\eta h}{r}) - 1\Big)^2\left(\frac{1}{h}\frac{\zeta(b)-\zeta(a)}{b-a} + QL_\eta\right)^2. \end{aligned}$$

This leads finally to

$$|\mathcal{F}_\delta(a,b,q,h) - \mathcal{F}(a,b,q,h)| \leq 4\delta\frac{\Big(L_\zeta + QL_\eta h\Big)^2}{QL_\eta}\Big(\frac{\exp(\frac{QL_\eta h}{r}) - 1}{h}\Big)^2.$$

We conclude the proof of the lemma by setting

$$C_F = 4\max_{x\in[0,M]} \frac{\Big(L_\zeta + QL_\eta hx\Big)^2}{QL_\eta}\Big(\frac{\exp(\frac{QL_\eta x}{r}) - 1}{x}\Big)^2$$

■

Let us now establish the regularity and the monotonicity of the fluxes.

Lemma 2.7: The function $\mathcal{F}_\delta(a,b,q,h)$ is Lipschitz-continuous with respect to a and b, increasing with respect to a, decreasing with respect to b.

Proof. Since $\mathcal{F}_\delta(a,b,q,h)$ is piecewise Lipschitz continuous, for given $a<b$, we have by differentiation of (16): $A_1\partial_1\mathcal{F}_\delta(a,b,q,h) = A_2\zeta'(a) + A_3\partial_1\mathcal{F}^{(q)}_{\rm god}(a,y_1)$, with

$$A_1 = \sum_{i=0}^{N-1} \frac{\zeta(y_{i+1})-\zeta(y_i)}{(\mathcal{F}^{(q)}_{\rm god}(y_i,y_{i+1}) - \mathcal{F}_\delta(a,b,q,h))^2},$$

$$A_2 = \frac{1}{\mathcal{F}^{(q)}_{\rm god}(a,y_1) - \mathcal{F}_\delta(a,b,q,h)},$$

and

$$A_3 = \frac{\zeta(y_1)-\zeta(a)}{(\mathcal{F}^{(q)}_{\rm god}(a,y_1) - \mathcal{F}_\delta(a,b,q,h))^2}.$$

As A_1, A_2 and A_3 are nonnegative, we first notice that $\partial_1\mathcal{F}_\delta(a,b,q,h)\geq 0$. Using the bound (23) and (16), we obtain that

$$A_1 \geq \frac{h}{\frac{\zeta(b)-\zeta(a)}{h} + QL_\eta(b-a)} \geq \frac{h}{(b-a)(\frac{L_\zeta}{h}+QL_\eta)}.$$

We also have that $A_3\leq A_1$ and $\partial_1\mathcal{F}^{(q)}_{\rm god}(a,y_1)\leq QL_\eta$. Moreover, due to (21), $A_2\leq 1/\delta_F$ and we get that

$$0\leq \partial_1\mathcal{F}_\delta(a,b,q,h) \leq L_\zeta(\frac{L_\zeta}{h}+QL_\eta)\frac{2(\exp(\frac{QL_\eta h}{r})-1)}{hQL_\eta} + QL_\eta.$$

Similar reasoning on b proves the result. ■

In order to conclude the proof of admissibility of the δ-fluxes in the sense of Definition 2.1, it remains to establish that they satisfy the properties iv) and v). This is the aim of Lemma 2.8

Lemma 2.8: The δ-fluxes satisfy the consistency and dissipativity properties: for all $a,b\in\mathbb{R}$,

$$\exists c\in[a\perp b, a\top b] \text{ such that } \mathcal{F}_\delta(a,b,q,h) = q\eta(c) - \frac{\zeta(b)-\zeta(a)}{h},$$
$$\text{and } (a-b)\mathcal{F}_\delta(a,b,q,h) \geq \frac{(\xi(b)-\xi(a))^2}{h} - q\int_a^b \eta(s)\mathrm{d}s.$$

Proof. Due to the conservativity of the fluxes, we can restrict the proof to the case $a<b$. By definition of the Godunov fluxes,

$$\min_{s\in[a,b]} q\eta(s) \leq \mathcal{F}^{(q)}_{\rm god}(y_i,y_{i+1}) \leq \max_{s\in[a,b]} q\eta(s) \quad \forall i\in\mathbb{Z}.$$

Using (16), this implies

$$\min_{s\in[a,b]} q\eta(s) \leq \mathcal{F}_\delta(a,b,q,h) + \frac{\zeta(b)-\zeta(a)}{h} \leq \max_{s\in[a,b]} q\eta(s),$$

so that there exists $c \in [a\perp b, a\top b]$ such that

$$\mathcal{F}_\delta(a,b,q,h) + \frac{\zeta(b)-\zeta(a)}{h} = q\eta(c).$$

Since we have established in Lemma 2.5 that $\mathcal{F}_\delta(a,b,q,h) \le \mathcal{F}(a,b,q,h)$ when $a < b$, we obtain

$$(a-b)\mathcal{F}_\delta(a,b,q,h) \ge (a-b)\mathcal{F}(a,b,q,h)$$

and the dissipativity property of the δ-fluxes is a direct consequence of the dissipativity property of the SGnl-fluxes established in [10, Lemma 2.5]. ∎

We are now able to state that the δ-fluxes are admissible in the sense of Definition 2.1. This is a consequence of Lemmas 2.7 and 2.8.

Proposition 2.9: The δ-fluxes defined by (14)-(15)-(16) for $a \le b$ and the conservativity relation for $a > b$ are admissible in the sense of Definition 2.1.

Since the convergence analysis of the finite volume scheme in [10] relies on the admissibility of the fluxes, it still holds with $\mathcal{F}_\delta(a,b,q,h)$ instead of $\mathcal{F}(a,b,q,h)$.

3 Numerical analysis of the schemes

In this section, we will let $d = 2$ or $d = 3$.

3.1 Specification on the continuous model

The convergence analysis of the SGnl-scheme has been achieved in [10] for the convection-diffusion equation (1) under Assumption 2.1, with $\mathrm{div}\boldsymbol{q} = 0$, Dirichlet boundary conditions and an initial condition $u_0 \in L^2(\Omega)$. As already mentioned, this analysis extends to the δ-schemes under the same hypotheses as the keypoint is the admissibility of the fluxes established in Proposition 2.9.
In what follows, we want to focus on the long-time behavior of the SGnl- and δ-schemes. Therefore, we consider additional hypotheses on the continuous problem. Indeed, our aim is to deal with generic fluxes *à la Onsager* (2). Let us specify the new framework of this Section.

Assumption 3.1:

i) The mobility function satisfies $\eta(0) = 0$ and $\eta(s) > 0$ for all $s > 0$.

ii) The convection field derives from a potential $V \in L^\infty(\Omega) \cap H^1(\Omega)$ such that $\displaystyle\int_\Omega V = 0$ and $g := -\Delta V \in L^2(\Omega)$. We set $\boldsymbol{q} = -\nabla V$, and we assume that

$$\boldsymbol{q}\cdot\boldsymbol{n} = 0 \text{ on } \Gamma = \partial\Omega.$$

iii) The boundary conditions are no-flux boundary conditions: $\boldsymbol{J}\cdot\boldsymbol{n} = 0$ on Γ, with $\boldsymbol{J} = -\nabla\zeta(u) + \eta(u)\boldsymbol{q}$.

iv) $u_0 \in L^\infty(\Omega)$ with $\underline{u}_0 := \operatorname{essinf} u_0 > 0$ and $\overline{u}_0 := \operatorname{esssup} u_0$.

Under Assumptions 2.1 and 3.1, we may define $\mu : (0,+\infty) \to \mathbb{R}$ by

$$\forall t \in (0,+\infty),\ \mu(t) = \int_1^t \frac{\zeta'(s)}{\eta(s)} \mathrm{d}s. \tag{25}$$

The function μ is obviously strictly increasing. Moreover, as $0 \leq \eta(s) \leq L_\eta s$, for all $s \in [0,+\infty)$, we have

$$\mu(1) - \mu(s) \geq -\frac{r}{L_\eta} \log s \quad \forall s \in (0,1),$$
$$\text{and } \mu(s) - \mu(1) \geq \frac{r}{L_\eta} \log s \quad \forall s \in (1,+\infty),$$

yielding

$$\lim_{s\to 0} \mu(s) = -\infty \text{ and } \lim_{s\to+\infty} \mu(s) = +\infty.$$

The new assumptions permit to rewrite the convection-diffusion flux $\boldsymbol{J}$ in (1) as

$$\boldsymbol{J} = -\Big(\nabla\zeta(u) - \eta(u)\boldsymbol{q}(\boldsymbol{x}\Big) = -\eta(u)\nabla\Big(\mu(u) + V(\boldsymbol{x})\Big), \tag{26}$$

which corresponds to (2) with $z_u = 1$. Some results – existence and uniqueness of a solution to the model (1) with (26), existence of a thermal equilibrium and exponential decay in time towards this equilibrium- are presented in Appendix A.
Lemma 3.1 shows that the SGnl-fluxes defined by (12)-(13) are consistent with this formulation of the continuous fluxes.

Lemma 3.1: Under Assumptions 2.1 and 3.1, for all $a, b \in (0,+\infty)$, there exists $c \in [a \perp b, a \top b]$, denoted in the sequel $c = \chi(a,b)$, such that

$$\mathcal{F}(a,b,q,h) = -\eta(c)\left(\frac{\mu(b)-\mu(a)}{h} - q\right). \tag{27}$$

Proof. If $a = b$, we have $\mathcal{F}(a,b,q,h) = q\eta(a)$ and (27) holds with $c = a$. Let us now assume $a < b$. We rewrite the function H defined by (12) as

$$H(x) = \int_a^b \frac{\eta(s)\mu'(s)}{q\eta(s) - x} ds.$$

The function $t \mapsto \frac{t}{qt-x}$ is nondecreasing on the interval $[\min_{[a,b]}\eta, \max_{[a,b]}\eta]$ if $x \leq 0$ while nonincreasing if $x \geq 0$ on the same interval. Assume first that $\mathcal{F}(a,b,q,h)$ defined by (12)-(13) is nonpositive, then we have

$$\frac{(\mu(b)-\mu(a))\min_{[a,b]}\eta}{q\min_{[a,b]}\eta - \mathcal{F}(a,b,q,h)} \leq h \leq \frac{(\mu(b)-\mu(a))\max_{[a,b]}\eta}{q\max_{[a,b]}\eta - \mathcal{F}(a,b,q,h)},$$

which implies

$$-\max_{[a,b]}\eta\left(\frac{\mu(b)-\mu(a)}{h} - q\right) \leq \mathcal{F}(a,b,q,h) \leq -\min_{[a,b]}\eta\left(\frac{\mu(b)-\mu(a)}{h} - q\right).$$

A similar inequality holds, up to an exchange of the min and the max, when $\mathcal{F}(a,b,q,h)$ is nonnegative. And, in any case, we can deduce (27). ∎

3.2 Numerical schemes and existence results

Let us first introduce the notations describing the mesh. The mesh $\mathcal{M} = (\mathcal{T}, \mathcal{E}, \mathcal{P})$ of the domain Ω is given by a family $\mathcal{T}$ of open polygonal or polyhedral control volumes, a family $\mathcal{E}$ of edges (or faces), and a family $\mathcal{P} = (\boldsymbol{x}_K)_{K\in\mathcal{T}}$ of points such that $\boldsymbol{x}_K \in K$ for all $K \in \mathcal{T}$. We assume that there exists a path between any pair of control volumes. As it is classical for TPFA finite volume discretizations including diffusive terms, we also assume that the mesh is admissible in the sense of [11, Definition 9.1]. It implies that the straight line between two neighboring centers of cells $(\boldsymbol{x}_K, \boldsymbol{x}_L)$ is orthogonal to the face $\sigma = K|L$.
In the set of edges $\mathcal{E}$, we distinguish the interior edges $\mathcal{E}_{\rm int}$ and the boundary edges $\mathcal{E}_{\rm ext}$. For a control volume $K \in \mathcal{T}$, we define the set of its edges $\mathcal{E}_K$, which is also partitioned into $\mathcal{E}_K = \mathcal{E}_{K,\rm int} \cup \mathcal{E}_{K,\rm ext}$. We assume that, for each edge $\sigma \in \mathcal{E}_{K,\rm int}$, there exist exactly two cells $K, L \in \mathcal{T}$ such that $\sigma \in \mathcal{E}_K \cap \mathcal{E}_L$, and we denote $\sigma = K|L$. We define $\boldsymbol{x}_\sigma$ as the center of gravity of σ for all $\sigma \in \mathcal{E}$.
For all control volume $K \in \mathcal{T}$ (*resp.* $\sigma \in \mathcal{E}$), we denote by m_K (*resp.* m_σ) its d-dimensional measure (*resp.* $d-1$-dimensional measure). For all $\sigma \in \mathcal{E}_{\rm int}$, $\sigma = K|L$, we define $\mathrm{d}_\sigma = d(\boldsymbol{x}_K, \boldsymbol{x}_L)$. Then the transmissibility coefficient is defined by $\tau_\sigma = \mathrm{m}_\sigma/\mathrm{d}_\sigma$, for all $\sigma \in \mathcal{E}_{\rm int}$.
Let us recall that, from [11, Lemma 10.2], there exists C_P, only depending on Ω, such that the following discrete Poincaré inequality holds (in the case $d = 2$ or $d = 3$).

$$\forall (u_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}},\ \sum_{K\in\mathcal{T}} \mathrm{m}_K (u_K - \frac{1}{\mathrm{m}_\Omega} \sum_{L\in\mathcal{T}} \mathrm{m}_L u_L)^2 \le C_P |u|^2_{1,\mathcal{M}}, \tag{28}$$

with

$$|u|^2_{1,\mathcal{M}} = \sum_{\sigma\in\mathcal{E}_{\rm int}, \sigma = K|L} \tau_\sigma (u_K - u_L)^2. \tag{29}$$

For any $K \in \mathcal{T}$, we use the simplified notation

$$\sum_{\sigma = K|L} \mathrm{term}(\sigma, K, L) \text{ instead of } \sum_{\sigma\in\mathcal{E}_{K,\rm int}, \sigma = K|L} \mathrm{term}(\sigma, K, L).$$

We first define the approximation of the potential V and the associate approximation of the convective field $\boldsymbol{q}$. The scheme is the usual TPFA-finite volume scheme for the Poisson equation with no-flux boundary conditions. Defining

$$g_K = \frac{1}{\mathrm{m}_K} \int_K g(x) dx,$$

the family $(V_K)_{K\in\mathcal{T}}$ is uniquely defined by

$$\begin{aligned} &- \sum_{\sigma = K|L} \tau_\sigma (V_L - V_K) = \mathrm{m}_K g_K, \quad \forall K \in \mathcal{T}, \\ &\sum_{K\in\mathcal{T}} \mathrm{m}_K V_K = 0. \end{aligned} \tag{30}$$

Let us now define

$$q_{K,L} = -\frac{V_L - V_K}{\mathrm{d}_\sigma} \quad \forall \sigma = K|L \in \mathcal{E}_{\rm int}. \tag{31}$$

Let $\Delta t > 0$ be the time step, we consider a Euler implicit in time scheme combined with a TPFA-finite volume scheme in space, based on the δ-numerical fluxes. For a given $\delta \geq 0$ ($\delta = 0$ corresponding to the SGnl-fluxes), we define the δ- scheme by

$$u_K^0 = \frac{1}{\mathrm{m}_K} \int_K u_0(x)\mathrm{d}x, \quad \forall K \in \mathcal{T}, \tag{32}$$

$$\mathrm{m}_K \frac{u_K^{n+1} - u_K^n}{\Delta t} + \sum_{\sigma=K|L} \mathrm{m}_\sigma \, \mathcal{F}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma) = 0, \quad \forall K \in \mathcal{T}, \forall n \geq 0. \tag{33}$$

The conservativity of the numerical fluxes ensures the preservation of the initial mass along time, as stated in Lemma 3.2.

Lemma 3.2 (Preservation of mass): Let $(u_K^n)_{K\in\mathcal{T},n\in\mathbb{N}}$ be such that (33) holds. Then we have

$$\sum_{K\in\mathcal{T}} \mathrm{m}_K u_K^n = \sum_{K\in\mathcal{T}} \mathrm{m}_K u_K^0, \quad \forall n \geq 0. \tag{34}$$

Lemma 3.3 (Monotony of the δ-scheme): For given $n \in \mathbb{N}$, let $(u_K^n)_{K\in\mathcal{T}}$ and $(v_K^n)_{K\in\mathcal{T}}$ be given, with $u_K^n \leq v_K^n$ (respectively $u_K^n \geq v_K^n$) for all $K \in \mathcal{T}$. Let $(u_K^{n+1})_{K\in\mathcal{T}}$ and $(v_K^{n+1})_{K\in\mathcal{T}}$ be the corresponding solutions to the δ-scheme (33) with $\delta \geq 0$. Then these values satisfy $u_K^{n+1} \leq v_K^{n+1}$ (respectively $u_K^{n+1} \geq v_K^{n+1}$) for all $K \in \mathcal{T}$.

Proof. Let $(u_K^n)_{K\in\mathcal{T}}$ and $(v_K^n)_{K\in\mathcal{T}}$ be given. Starting from (33), the monotonicity properties of $\mathcal{F}_\delta$ ensure that, for all $K \in \mathcal{T}$,

$$\begin{aligned}
\mathrm{m}_K u_K^{n+1} &\geq \mathrm{m}_K u_K^n \perp v_K^n - \Delta t \sum_{\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}_\delta(u_K^{n+1}, u_L^{n+1} \perp v_L^{n+1}, q_{K,L}, d_\sigma) \\
\text{and } \mathrm{m}_K v_K^{n+1} &\geq \mathrm{m}_K v_K^n \perp u_K^n - \Delta t \sum_{\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}_\delta(v_K^{n+1}, v_L^{n+1} \perp u_L^{n+1}, q_{K,L}, d_\sigma).
\end{aligned}$$

Hence, since the minimum value between u_K^{n+1} and v_K^{n+1} is one of them, we get

$$\mathrm{m}_K u_K^{n+1} \perp v_K^{n+1} \geq \mathrm{m}_K u_K^n \perp v_K^n - \Delta t \sum_{\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}_\delta(u_K^{n+1} \perp v_K^{n+1}, u_L^{n+1} \perp v_L^{n+1}, q_{K,\sigma}, d_\sigma).$$

Subtracting the above inequality to (33) written for u, we obtain, for all $K \in \mathcal{T}$,

$$\begin{aligned}
\mathrm{m}_K \max(u_K^{n+1} - v_K^{n+1}, 0) &\leq \mathrm{m}_K \max(u_K^n - v_K^n, 0) \\
&\quad - \Delta t \sum_{\sigma=K|L} \mathrm{m}_\sigma \Big(\mathcal{F}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,\sigma}, d_\sigma) - \mathcal{F}_\delta(u_K^{n+1} \perp v_K^{n+1}, u_L^{n+1} \perp v_L^{n+1}, q_{K,\sigma}, d_\sigma) \Big).
\end{aligned}$$

Summing over $K \in \mathcal{T}$, we get

$$\sum_{K\in\mathcal{T}} \mathrm{m}_K \max(u_K^{n+1} - v_K^{n+1}, 0) \leq \sum_{K\in\mathcal{T}} \mathrm{m}_K \max(u_K^n - v_K^n, 0).$$

If $u_K^n \leq v_K^n$ for all $K \in \mathcal{T}$, the left-hand-side vanishes, implying that $u_K^{n+1} \leq v_K^{n+1}$ for all $K \in \mathcal{T}$. If $u_K^n \geq v_K^n$ for all $K \in \mathcal{T}$, we obtain that $u_K^{n+1} \geq v_K^{n+1}$ for all $K \in \mathcal{T}$ by exchanging the role of u and v. ∎

We establish now the *a priori*-positivity of a solution $(u_K^n)_{K\in\mathcal{T},n\in\mathbb{N}}$ to (33). Combined with Lemma 3.2, Lemma 3.4 brings the existence of a solution to the scheme stated in Proposition 3.5.

Lemma 3.4 (Values of u_K^n are nonnegative for nonnegative initial values): Let $(u_K^n)_{K\in\mathcal{T},n\in\mathbb{N}}$ be a solution to (33), with $u_K^0 \geq 0$ for all $K \in \mathcal{T}$. Then, for all $n \in \mathbb{N}$, there holds

$$\frac{\sum_{K\in\mathcal{T}} \mathrm{m}_K u_K^0}{\min_K \mathrm{m}_K} \geq u_K^n \geq 0 \quad \forall K \in \mathcal{T}. \tag{35}$$

Proof. We observe that the family $(0)_{K\in\mathcal{T}}$ is solution to (33) since $\mathcal{F}_\delta(0,0,q_{K,\sigma},d_\sigma) = 0$ (recall that $\eta(0) = 0$). Applying Lemma 3.3 provides $u_K^n \geq 0$ forall $K \in \mathcal{T}$ and $n \in \mathbb{N}$. Then the relation (34) gives the left inequality of (35). ∎

Proposition 3.5 (Existence and uniqueness of a solution to the scheme): Let $(u_K^0)_{K\in\mathcal{T}}$ be given, such that $u_K^0 \geq 0$ for all $K \in \mathcal{T}$. Then there exists a unique solution $(u_K^n)_{K\in\mathcal{T},n\in\mathbb{N}}$ to the scheme (33), which satisfies $u_K^n \geq 0$ for all $K \in \mathcal{T}$ and $n \geq 0$.

Proof. Reasoning by induction, we consider the mapping $F : \mathbb{R}^{\mathcal{T}} \times [0,1] \to \mathbb{R}^{\mathcal{T}}$ defined, for $\theta \in [0,1]$ and $K \in \mathcal{T}$, by

$$F_K(v,\theta) = \mathrm{m}_K \frac{v_K - u_K^n}{\Delta t} + \theta \sum_{\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}_\delta(v_K, v_L, q_{K,L}, d_\sigma).$$

Since $F(v,\theta) = 0$ implies $v_K \in [0, (\sum_{K\in\mathcal{T}} \mathrm{m}_K u_K^0)/\min_K \mathrm{m}_K]$ by Lemma 3.4, and since $F(v,0) = 0$ has a solution with the same bounds, the invariance of the topological degree by homotopy enables to conclude that $F(v,1) = 0$ has at least one solution. The uniqueness of the solution is a direct consequence of the monotony of the scheme stated in Lemma 3.3. ∎

The analysis of the convergence of the scheme to a solution of (1) on $\Omega \times (0,T)$ for a given $T > 0$ can then be completed, by a simple adaptation of the proof done in [10] and is not detailed here.

3.3 Long time behavior of the SGnl-scheme and of the δ-scheme

A particularity of the Scharfetter-Gummel numerical fluxes designed for linear convection-diffusion equations is that they preserve any thermal equilibrium. Indeed, for all $\lambda \geq 0$, $u_{\mathrm{th}} = \lambda e^{-V}$ is a thermal equilibrium (depending on the boundary conditions) as the continuous flux verifies $\boldsymbol{J} = -\nabla u_{\mathrm{th}} - u_{\mathrm{th}} \nabla V = 0$. With the Scharfetter-Gummel numerical fluxes, the numerical fluxes similarly vanish for any discrete thermal equilibrium :

$$u_{\mathrm{th},K} = \lambda e^{-V_K} \implies \mathcal{F}_{K,\sigma} = 0, \quad \forall K \in \mathcal{T}, \forall \sigma \in \mathcal{E}_K.$$

Under Assumptions 2.1 and 3.1, a thermal equilibrium satisfies $\mu(u) + V = \lambda$, with $\lambda \in \mathbb{R}$. In practice, when the boundary conditions are no-flux boundary conditions, the value of λ is prescribed by the initial mass, which is preserved along time. Lemma 3.6 shows that the SG-nl scheme preserves the thermal equilibria.

Lemma 3.6 (Thermal equilibrium): For any $M^0 > 0$, there exists one and only one $(u_K)_{K\in\mathcal{T}}$ with $u_K \geq 0$ such that

$$\sum_{K\in\mathcal{T}} \mathrm{m}_K u_K = M^0 > 0, \tag{36}$$

and

$$\forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_{K,\mathrm{int}}, \sigma = K|L,\ \mathcal{F}(u_K, u_L, q_{K,L}, d_\sigma) = 0. \tag{37}$$

Moreover, we have $u_K > 0$ for all $K \in \mathcal{T}$, and there exists one and only one $\lambda \in \mathbb{R}$ such that

$$\forall K \in \mathcal{T},\ \mu(u_K) + V_K = \lambda.$$

Proof. Let us consider $q \in \mathbb{R}$, $h > 0$ and $a, b \in [0, +\infty)$, with $a \leq b$. If $\mathcal{F}(a, b, q, h) = 0$, then a consequence of Lemma 3.1 is that: either $\mu(b) - \mu(a) = qh$ or $\eta(\chi(a, b)) = 0$. The first condition can be satisfied only if $a > 0$ and $b > 0$, as $\lim_0 \mu = -\infty$. The second one implies $\chi(a, b) = 0$ and then $a = 0$ and $\min_{[a,b]} q\eta \leq 0$. In this case, $b > 0$ is impossible, since for $b > a = 0$, the numerical flux is the unique solution x to the equation $H(x) = h$ and must satisfy $x < \min_{[a,b]} q\eta$. This means that $b = a = 0$. We deduce the following alternative to get $\mathcal{F}(a, b, q, h) = 0$:

- either $a > 0$, $b > 0$ and $\mu(b) - \mu(a) = qh$,
- or $a = b = 0$.

We obtain the same result if $a \geq b$ by exchanging the role of a and b and using the conservativity relation.
Let $(u_K)_{K\in\mathcal{T}}$ satisfying (36)-(37), with $u_K \geq 0$ for all $K \in \mathcal{T}$. If there exists $K \in \mathcal{T}$ such that $u_K = 0$, then all neighboring values are null as well. This implies, since we assume that all the control volumes are connected, that all the values $(u_K)_{K\in\mathcal{T}}$ are null. This is in contradiction with $M^0 > 0$.
Hence we get that necessarily $u_K > 0$ for all $K \in \mathcal{T}$ and that

$$\forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_{K,\mathrm{int}},\ \sigma = K|L,\ \mu(u_L) - \mu(u_K) = V_K - V_L.$$

This implies, using again that all control volumes are connected, that there exists $\lambda \in \mathbb{R}$ such that

$$\forall K \in \mathcal{T},\ \mu(u_K) = -V_K + \lambda. \tag{38}$$

Using (36), we obtain that $\lambda \in \mathbb{R}$ is the unique solution of the equation

$$f(\lambda) := \sum_{K\in\mathcal{T}} \mathrm{m}_K \mu^{-1}(-V_K + \lambda) = M^0,$$

owing to the fact that f is a strictly increasing function satisfying $f(s) \to 0$ as $\lambda \to -\infty$ and $f(s) \to +\infty$ as $\lambda \to +\infty$. ∎

Let us introduce $\Phi : s \in (0, +\infty) \mapsto \int_1^s \mu(t)dt$. Φ is obviously a convex function, as the following one:

$$s \mapsto \Phi(s) - \Phi(s^\infty) - \mu(s^\infty)(s - s^\infty) \text{ for any } s^\infty \in (0, \infty).$$

We can now prove that under Assumptions 2.1 and 3.1, the δ-scheme satisfies a discrete entropy/entropy dissipation property, as stated in Lemma 3.7.

Lemma 3.7 (Decay of the relative entropy): Let $(u_K^0)_{K\in\mathcal{T}}$ with $u_K^0 > 0$ be given. Let $(u_K^\infty)_{K\in\mathcal{T}}$ be the thermal equilibrium given by Lemma 3.6 for $M^0 = \sum_{K\in\mathcal{T}} \mathrm{m}_K u_K^0$, and let $(u_K^n)_{K\in\mathcal{T},n\geq 0}$ be the solution to the δ-scheme ($\delta \geq 0$), defined by (30)–(33). For any $n \in \mathbb{N}$, we define the discrete entropy E^n and the associated discrete dissipation D^{n+1} by:

$$E^n = \sum_{K\in\mathcal{T}} \mathrm{m}_K \Big(\Phi(u_K^n) - \Phi(u_K^\infty) - \mu(u_K^\infty)(u_K^n - u_K^\infty)\Big) \tag{39}$$

$$D^{n+1} = \sum_{\sigma\in\mathcal{E}_{\rm int},\sigma=K|L} \tau_\sigma \eta(\overline{u}_\sigma^{n+1})\Big(\mu(u_K^{n+1}) - \mu(u_L^{n+1}) - \mu(u_K^\infty) + \mu(u_L^\infty)\Big)^2, \tag{40}$$

with $\overline{u}_\sigma^{n+1} = \chi(u_K^{n+1}, u_L^{n+1})$ defined by (27). Then there exists $\beta \geq 0$, only depending on Ω, $\|g\|_{L^2(\Omega)}$, L_η, L_ζ, such that

$$\frac{E^{n+1} - E^n}{\Delta t} + D^{n+1} \leq \beta\delta. \tag{41}$$

Proof. Let us first notice that we can rewrite the dissipation term (40), thanks to (27) and (38), as

$$D^{n+1} = \sum_{\sigma\in\mathcal{E}_{\rm int},\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma)\Big(\mu(u_K^{n+1}) - \mu(u_L^{n+1}) - \mu(u_K^\infty) + \mu(u_L^\infty)\Big)$$

Due to the convexity of the function Φ, we have

$$E^{n+1} - E^n \leq \sum_{K\in\mathcal{T}} \mathrm{m}_K (u_K^{n+1} - u_K^n)(\mu(u_K^{n+1}) - \mu(u_K^\infty)).$$

Then, multipling (33) by $\mu(u_K^{n+1}) - \mu(u_K^\infty)$ and summing over $K \in \mathcal{T}$, we get

$$\begin{aligned}
E^{n+1} - E^n &\leq -\Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma)(\mu(u_K^{n+1}) - \mu(u_K^\infty)) \\
&\leq -\Delta t \sum_{\sigma\in\mathcal{E}_{\rm int},\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma)\Big(\mu(u_K^{n+1}) - \mu(u_L^{n+1}) - \mu(u_K^\infty) + \mu(u_L^\infty)\Big), \\
&\leq -\Delta t (D^{n+1} + R^{n+1}),
\end{aligned} \tag{42}$$

with, setting $\mathcal{G}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma) = \mathcal{F}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma) - \mathcal{F}(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma)$,

$$R^{n+1} = \sum_{\sigma\in\mathcal{E}_{\rm int},\sigma=K|L} \mathrm{m}_\sigma \mathcal{G}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma)\Big(\mu(u_K^{n+1}) - \mu(u_L^{n+1}) - \mu(u_K^\infty) + \mu(u_L^\infty)\Big).$$

But, using (20) and the monotonicity of μ, we have that

$$\mathcal{G}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma)\Big(\mu(u_K^{n+1}) - \mu(u_L^{n+1})\Big) \geq 0 \quad \forall \sigma = K|L$$

and therefore, using (38) again and applying the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
R^{n+1} &\geq \sum_{\sigma\in\mathcal{E}_{\rm int},\sigma=K|L} \mathrm{m}_\sigma \mathcal{G}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma)(V_K - V_L), \\
&\geq -\left(\sum_{\sigma\in\mathcal{E}_{\rm int},\sigma=K|L} \mathrm{m}_\sigma \mathrm{d}_\sigma (\mathcal{G}_\delta(u_K^{n+1}, u_L^{n+1}, q_{K,L}, d_\sigma))^2\right)^{1/2} |V|_{1,\mathcal{M}},
\end{aligned}$$

where we use the discrete H^1-seminorm defined by (29). Owing to Lemma 2.6, we finally obtain

$$R^{n+1} \geq C_F \delta (d\mathrm{m}_\Omega)^{1/2} |V|_{1,\mathcal{M}}.$$

Note that, using (30) and the discrete Poincaré inequality with null average (28), we have

$$|V|_{1,\mathcal{M}} \leq C_P \|g\|_2,$$

where C_P only depends on Ω. This, together with (42), yields (41). This concludes the proof of Lemma 3.7. ∎

Lemma 3.8 (Existence of strictly positive stationary states, with imposed lower or upper bounds): For any $A > 0$, there exists $B > 0$ only depending on A, $\max_K V_K$, $\min_K V_K$ and μ, and there exists $\delta_0 > 0$, only depending on A, $\max_K V_K$, $\min_K V_K$, μ and on $\mathcal{M}$, such that, for any $\delta \in [0, \delta_0)$, one can find $(u_K^{(\delta)})_{K\in\mathcal{T}}$ with $A \geq u_K^{(\delta)} \geq B$ (respectively $A \leq u_K^{(\delta)} \leq B$) for all $K \in \mathcal{T}$ and

$$\forall K \in \mathcal{T}, \ \sum_{\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}_\delta(u_K^{(\delta)}, u_L^{(\delta)}, q_{K,L}, d_\sigma) = 0. \tag{43}$$

Proof. The starting point of the proof is similar to that of Lemma A.3. Let $\lambda \in \mathbb{R}$ be defined by $\mu^{-1}(\lambda - \min_K V_K) = A/2$. We then consider the thermal equilibrium $(u_K^\infty)_{K\in\mathcal{T}}$ with total mass M^∞ defined by

$$\forall K \in \mathcal{T}, \ u_K^\infty := \mu^{-1}(\lambda - V_K) \text{ and } M^\infty := \sum_{K\in\mathcal{T}} \mathrm{m}_K u_K^\infty.$$

Setting $B = \frac{1}{2}\mu^{-1}(\lambda - \max_K V_K)$, this thermal equilibrium satisfies $u_K^\infty \in [2B, A/2]$ for all $K \in \mathcal{T}$.
For any $\delta \geq 0$ and $k > 0$, we define $(u_K(k))_{K\in\mathcal{T}}$ as the unique solution to

$$\forall K \in \mathcal{T}, \ \mathrm{m}_K \frac{u_K(k) - u_K^\infty}{k} + \sum_{\sigma=K|L} \mathrm{m}_\sigma \mathcal{F}_\delta(u_K(k), u_L(k), q_{K,L}, d_\sigma) = 0.$$

It corresponds to the first iterate in time of the δ-scheme for an initial condition coinciding with the thermal equilibrium and with a time step k. Thanks to Lemma 3.4, which holds since $u_K^\infty > 0$ for all $K \in \mathcal{T}$, the solution satisfies $u_K(k) \in [0, U]$ for all $K \in \mathcal{T}$ with $U = \frac{M^\infty}{\min_K \mathrm{m}_K}$. Moreover, as the initial discrete entropy is equal to zero, Lemma 3.7 implies that

$$\frac{1}{k} E^1 + D^1 \leq \beta\delta,$$

with $E^1 \geq 0$ and

$$D^1 = \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}, \sigma=K|L} \tau_\sigma \eta(\overline{u}_\sigma(k)) \Big(\mu(u_K(k)) - \mu(u_L(k)) - \mu(u_K^\infty) + \mu(u_L^\infty)\Big)^2.$$

Notice that we have

$$\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}, \sigma=K|L} \mathrm{m}_\sigma \mathrm{d}_\sigma (\mathcal{F}(u_K(k), u_L(k), q_{K,L}, d_\sigma))^2 =$$

$$\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}, \sigma=K|L} \tau_\sigma \eta(\overline{u}_\sigma(k))^2 \Big(\mu(u_K(k)) - \mu(u_L(k)) - \mu(u_K^\infty) + \mu(u_L^\infty)\Big)^2 \leq \max_{[0,U]} \eta \ D^1.$$

We then consider a sequence $(k_m)_m$ which tends to $+\infty$. For any m, we then have

$$\sum_{\sigma\in\mathcal{E}_{\rm int},\sigma=K|L} {\rm m}_\sigma {\rm d}_\sigma(\mathcal{F}(u_K(k_m),u_L(k_m),q_{K,L},d_\sigma))^2 \le \max_{[0,U]}\eta\ \beta\delta,$$

and therefore, denoting by $C_2=\beta\frac{\max_{[0,U]}\eta}{\min_{\sigma\in\mathcal{E}_{\rm int}}({\rm m}_\sigma d_\sigma)}$,

$$\forall\sigma=K|L,\ \mathcal{F}(u_K(k_m),u_L(k_m),q_{K,L},d_\sigma)^2\le C_2\delta.$$

Up to the extraction of a subsequence, we obtain a converging sequence in $[0,U]^{\mathcal{T}}$, whose limit as $m\to\infty$ is denoted $(u_K^{(\delta)})_{K\in\mathcal{T}}$. It satisfies $u_K^{(\delta)}\in[0,U]$ for all $K\in\mathcal{T}$,

$$\sum_{K\in\mathcal{T}} {\rm m}_K u_K^{(\delta)} = M^\infty,$$

$$\forall K\in\mathcal{T},\ \sum_{\sigma=K|L} {\rm m}_\sigma\,\mathcal{F}_\delta(u_K^{(\delta)},u_L^{(\delta)},q_{K,\sigma},d_\sigma)=0,$$

and

$$\forall\sigma=K|L,\ \mathcal{F}(u_K^{(\delta)},u_L^{(\delta)},q_{K,L},d_\sigma)^2\le C_2\delta.$$

It remains to prove that there exists $\delta_0>0$ such that, for all $\delta<\delta_0$, $u_K^{(\delta)}\in[B,A]$ for all $K\in\mathcal{T}$. Therefore, we consider a sequence (δ_m) which tends to 0. We can extract a subsequence such that $(u_K^{(\delta_m)})_m$ converges. Then, its limit denoted by $(u_K)_{K\in\mathcal{T}}$ satisfies $u_K\in[0,U]$ for all $K\in\mathcal{T}$,

$$\sum_{K\in\mathcal{T}} {\rm m}_K u_K = M^\infty,$$

and, by continuity of $\mathcal{F}(a,b,q,h)$ with respect to a and b,

$$\forall\sigma=K|L,\ \mathcal{F}(u_K,u_L,q_{K,L},d_\sigma)=0.$$

This means that $(u_K)_{K\in\mathcal{T}}$ is a nonnegative thermal equilibrium with the mass M^∞, therefore equal to $(u_K^\infty)_{K\in\mathcal{T}}$ as shown by Lemma 3.6. By uniqueness of the limit $(u_K^\infty)_{K\in\mathcal{T}}$, we get that the whole sequence $(u_K^{(\delta)})_{K\in\mathcal{T}}$ converges to $(u_K^\infty)_{K\in\mathcal{T}}$ as $\delta\to0$. Hence, one can find δ_0 such that

$$\forall\delta<\delta_0,\ \forall K\in\mathcal{T},\ |u_K^{(\delta)}-u_K^\infty|\le\min(B,\frac{A}{2}),$$

which then implies that

$$\forall\delta<\delta_0,\ \forall K\in\mathcal{T},\quad A\ge u_K^{(\delta)}\ge B.$$

The second case where A is a fixed lower bound instead of the upper bound is similarly handled, defining $\lambda\in\mathbb{R}$ by $\mu^{-1}(\lambda-\max_K V_K)=2A$. ∎

Lemma 3.9 (Discrete solutions are uniformly bounded for δ small enough):
Let $(u_K^0)_{K\in\mathcal{T}}$ be given, with $u_K^0>0$ for all $K\in\mathcal{T}$. Then there exist $\underline{B}>0$ and $\overline{B}>0$, only depending on $\max_K u_K^0$, $\min_K u_K^0$, $\max_K V_K$, $\min_K V_K$ and μ, and $\delta_0>0$, only depending on $\max_K u_K^0$, $\min_K u_K^0$, $\max_K V_K$, $\min_K V_K$, μ and on $\mathcal{M}$, such that, for all $\delta<\delta_0$, any solution $(u_K^n)_{K\in\mathcal{T},n\ge0}$ of the δ- scheme (33) is such that

$$\forall n\in\mathbb{N},\ \forall K\in\mathcal{T},\ 0<\underline{B}\le u_K^n\le\overline{B}.$$

Proof. Let $A = \min_{K\in\mathcal{T}} u_K^0$, and let $\delta_0^{(1)} > 0$ and $\underline{B} := B > 0$ be given by the first statement of Lemma 3.8. For $\delta < \delta_0^{(1)}$, let $(u_K^{(\delta)})_{K\in\mathcal{T}}$ be a stationary solution given by Lemma 3.8 (which therefore satisfies the δ- scheme (33)). It satisfies

$$u_K^0 \geq u_K^{(\delta)} \geq \underline{B}, \quad \forall K \in \mathcal{T}.$$

Therefore, the monotonicity of the δ-scheme stated in Lemma 3.3 yields, by induction,

$$u_K^n \geq u_K^{(\delta)} \geq \underline{B} \quad \forall K \in \mathcal{T}, \forall n \geq 0.$$

Let $A = \max_{K\in\mathcal{T}} u_K^0$, and let $\delta_0^{(2)} > 0$ and $\overline{B} := B > 0$ be given by the second statement of Lemma 3.8. For $\delta < \delta_0^{(2)}$, let $(u_K^{(\delta)})_{K\in\mathcal{T}}$ be a stationary solution given by Lemma 3.8. It satisfies

$$u_K^0 \leq u_K^{(\delta)} \leq \overline{B}, \quad \forall K \in \mathcal{T}.$$

and we obtain, as previously,

$$u_K^n \leq u_K^{(\delta)} \leq \overline{B} \quad \forall K \in \mathcal{T}, \forall n \geq 0.$$

We conclude the proof by taking $\delta_0 = \min(\delta_0^{(1)}, \delta_0^{(2)})$.
The following lemma replaces the log-Sobolev inequalities used for the exponential decay of the relative entropy in the case $\eta(s) = s$ and $\mu(s) = \log(s)$ [3].

Lemma 3.10 (Discrete nonlinear mean Poincaré inequality): Let A, B be two reals such that $0 < B \leq A$. Then there exists C_P, only depending on Ω, such that, for any $(u_K)_{K\in\mathcal{T}}$ and $(v_K)_{K\in\mathcal{T}}$ belonging to $[B, A]^{\mathcal{T}}$ and satisfying

$$\sum_{K\in\mathcal{T}} \mathrm{m}_K u_K = \sum_{K\in\mathcal{T}} \mathrm{m}_K v_K,$$

we have

$$\sum_{K\in\mathcal{T}} \mathrm{m}_K (u_K - v_K)(\mu(u_K) - \mu(v_K)) \leq \frac{1}{\min_{[B,A]} \mu'} C_P |\mu(u_K) - \mu(v_K)|_{1,\mathcal{M}}^2.$$

Proof. This proof is similar to that of the Poincaré inequality proved in Lemma A.4. Let us introduce

$$C = \frac{1}{m_\Omega} \sum_{K\in\mathcal{T}} \mathrm{m}_K (\mu(u_K) - \mu(v_K)).$$

For any $K \in \mathcal{T}$, let us denote $w_K = \mu(u_K) - \mu(v_K) - C$. The mean discrete Poincaré inequality (28) can be applied to $(w_K)_{K\in\mathcal{T}}$, yielding

$$\sum_{K\in\mathcal{T}} \mathrm{m}_K w_K^2 \leq C_P |\mu(u_K) - \mu(v_K)|_{1,\mathcal{M}}^2, \tag{44}$$

where $C_P \geq 0$ is only depending on Ω. We then have

$$u_K = \mu^{-1}(\mu(v_K) + C + w_K) = v_K + (\mu^{-1})'(z_K)(C + w_K),$$

with $z_K \in [\mu(u_K) \perp \mu(v_K), \mu(u_K) \top \mu(v_K)] \subset [\mu(B), \mu(A)]$. Then we introduce

$$\alpha_K := (\mu^{-1})'(z_K) = \frac{1}{\mu'(\mu^{-1}(z_K))} \in [\underline{\alpha}, \overline{\alpha}],$$

with

$$\underline{\alpha} = \frac{1}{\max_{s\in[B,A]} \mu'(s)} \text{ and } \overline{\alpha} = \frac{1}{\min_{s\in[B,A]} \mu'(s)}.$$

We therefore have

$$\sum_{K\in\mathcal{T}} \mathrm{m}_K(v_K + \alpha_K(C + w_K)) = \sum_{K\in\mathcal{T}} \mathrm{m}_K v_K,$$

which implies

$$C = -\frac{\sum_{K\in\mathcal{T}} \mathrm{m}_K \alpha_K w_K}{\sum_{K\in\mathcal{T}} \mathrm{m}_K \alpha_K}.$$

We then get

$$\begin{aligned}\sum_{K\in\mathcal{T}} \mathrm{m}_K(u_K - v_K)(\mu(u_K) - \mu(v_K)) &= \sum_{K\in\mathcal{T}} \mathrm{m}_K \alpha_K (C + w_K)^2 \\ &= -\frac{\left(\sum_{K\in\mathcal{T}} \mathrm{m}_K \alpha_K w_K\right)^2}{\sum_{K\in\mathcal{T}} \mathrm{m}_K \alpha_K} + \sum_{K\in\mathcal{T}} \mathrm{m}_K \alpha_K w_K^2 \leq \overline{\alpha} \sum_{K\in\mathcal{T}} \mathrm{m}_K w_K^2\end{aligned}$$

Together with (44), this concludes the proof. ∎

Theorem 3.11 (Convergence to the discrete thermal equilibrium): Let (u_K^0) be given, with $u_K^0 > 0$ for all $K \in \mathcal{T}$ and M^0 the associate mass. Let $(u_K^n)_{K\in\mathcal{T}, n\geq 0}$ the solution to the scheme (33) and $(u_K^\infty)_{K\in\mathcal{T}}$ the thermal equilibrium defined by Lemma 3.6. Let $\delta_0 > 0$, $\underline{B} > 0$ and $\overline{B} > 0$ be given by Lemma 3.9, and $\beta > 0$ be given by Lemma 3.7. Then there exists $\alpha > 0$ only depending on μ, η, $\underline{B}$ and $\overline{B}$, such that, for any $\delta \in [0, \delta_0)$ and for any $n \in \mathbb{N}$, it holds

$$\frac{1}{2} \min_{[\underline{B},\overline{B}]} \mu' \sum_{K\in\mathcal{T}} \mathrm{m}_K (u_K^n - u_K^\infty)^2 \leq E^n \leq \beta\delta \frac{(1 - (1 + \alpha\Delta t)^{-n})}{\alpha} + E^0 (1 + \alpha\Delta t)^{-n}.$$

Note that, for $\Delta t \leq 1/\alpha$, it holds $(1 + \alpha\Delta t)^{-n} \leq \exp(-\frac{1}{2}\alpha n \Delta t)$, which shows in this case the exponential decay of E^n, up to δ.

Proof.
The inequality

$$\forall a, b > 0, \Phi(b) - \Phi(a) - \mu(b)(b - a) = \int_a^b (\mu(s) - \mu(a)) \mathrm{d}s \leq (b - a)(\mu(b) - \mu(a))$$

yields, applying Lemma 3.10,

$$E^n \leq \frac{1}{\alpha} D^n \text{ with } \frac{1}{\alpha} := \frac{C_P}{\min_{[\underline{B},\overline{B}]} \mu'} \frac{1}{\min_{[\underline{B},\overline{B}]} \eta}.$$

We then get, by Lemma 3.7,

$$E^{n+1}(1+\alpha\Delta t) \leq E^n + \beta\delta\Delta t,$$

leading, by induction,

$$E^n \leq \beta\delta\frac{(1-(1+\alpha\Delta t)^{-n})}{\alpha} + E^0(1+\alpha\Delta t)^{-n} \quad \forall n \geq 0.$$

Moreover, the left hand side of the expected inequality is a straightforward consequence of the Taylor expansion up to order 2 of the function Φ in the definition (39) of E^n. ∎

4 Numerical experiments

4.1 Implementation of the scheme

In order to find an approximate value of the function $\mathcal{F}_\delta(a,b,q,h)$ defined by (16), we use the following algorithm.

1. find the smallest n such that $H_\delta(x_{\rm ini}^{(n)}) > h$ with $x_{\rm ini}^{(n)} = \mathcal{F}_{\rm god}^{(q)}(a,b) - \frac{\zeta(b)-\zeta(a)}{10^n h}$,
2. solve $H_\delta(x) = h$ by Newton's method, with $x_0 = x_{\rm ini}$:

The convergence of Newton's method is ensured, since H_δ is convex and the choice of the initial guess x_0 leads to $\mathcal{F}_\delta(a,b,q,h) \leq x_0$.
This algorithm is efficient and cheap, and takes a large benefit of parallel computing, since there is no sequential dependence between the values of the numerical fluxes. A byproduct of this algorithm is the easy computation of the partial derivatives of $\mathcal{F}_\delta(a,b,q,h)$, used in the implementation of the numerical scheme (33).
Note that the definition of $\mathcal{F}_\delta(a,b,q,h)$ implies that a sequence $(\overline{y}_i)_i$ be defined. This is completed in our implementation by fixing $\overline{y}_0 = 0$ and then by requiring

$$\forall i \in \mathbb{Z},\ \zeta(\overline{y}_{i+1}) - \zeta(\overline{y}_i) = \widetilde{\delta}, \tag{45}$$

for a given value of $\widetilde{\delta} > 0$. This enables to apply our scheme to cases where ζ is only nondecreasing. For the simplicity of the notation, in the remaining of this section we simply denote by δ the value $\widetilde{\delta}$.

4.2 Validation of the theoretical results

In this section, we consider $\Omega \subset \mathbb{R}$ and the data of (1) are given by

$$\zeta(u) = \alpha u + |u|^m \mathrm{sign}(u),\ \eta(u) = u \text{ and } \boldsymbol{q} = \rho_0 + \rho_1 x, \tag{46}$$

for different values of $m \geq 1$, $\alpha \geq 0$, $\rho_0, \rho_1 \in \mathbb{R}$ and for $x \in \Omega$, with various initial and boundary conditions. This example is inspired by [2].

Accuracy with respect to δ

We first investigate the influence of the value of δ used in (45) on the accuracy of the scheme. We consider the numerical solutions u_N corresponding to $\delta = 10^{-N}$, with $N = 0, 1, 2, 3, 4, 5$, together

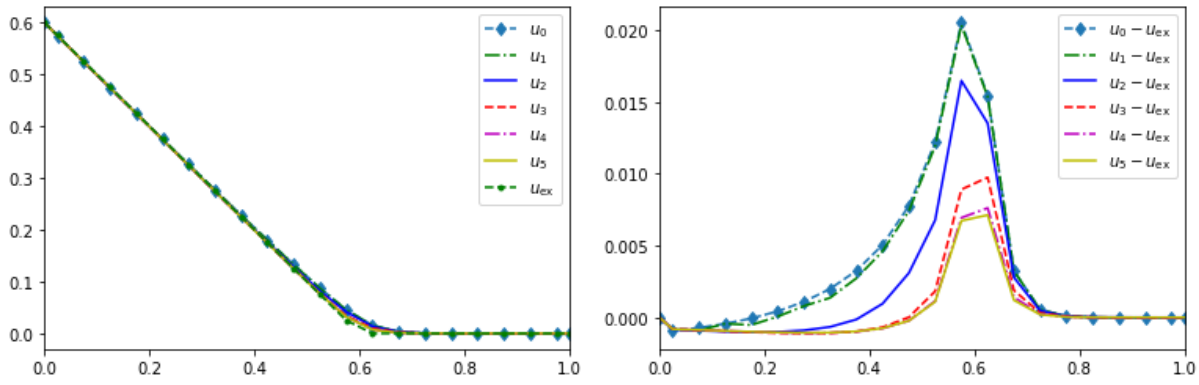

Fig. 2: Left: Numerical solution u_N at $T = 0.2$ with different values of $\delta = 10^{-N}$, with $N = 0, 1, 2, 3, 4, 5$. Right: difference $u_N - u_{\text{ex}}$.

with letting $m = 2$, $\alpha = 10^{-6}$, $\rho_0 = 1$, $\rho_1 = 0$ in (46). Furthermore, we define $\Omega = (0, 1)$, we consider the initial condition $u(x, 0) = 0$ and the nonhomogeneous Dirichlet boundary conditions $u(0, t) = 3t$ and $u(1, t) = \max(3t - 1, 0)$. Let us observe that, in this case, the numerical flux $\mathcal{F}_\delta(a, b, q, h)$ computed with $\delta = 1$ is identical to the numerical flux $\mathcal{F}_\infty(a, b, q, h)$ (this flux is defined in (18), and corresponds to the use of a simple Godunov scheme in the convection term combined with a 2-point diffusion flux).

We let the mesh step size constant equal to $1/20$ and the time step constant equal to $k = 0.001$. We obtain the results provided by Figure 2. In this figure, the function $u_{\text{ex}}(x, t) = \max(3t - x, 0)$ is the analytical solution of (1) in the case $\alpha = 0$ (then the difference between $u_{\text{ex}}(x, t)$ and the analytical solution of (1) with $\alpha = 10^{-6}$ is much smaller than the numerical error). We observe that $\delta = 10^{-N}$, with $N = 4, 5$ provides significantly more accurate results than the simple Godunov scheme (obtained with $N = 0$), and that u_5 is close to u_4. This suggests that it is not interesting to consider greater values for N.

Entropy decay in a degenerate parabolic case

Following [2, Example 8], we now let $\alpha = 0$, $\rho_0 = 0$, $\rho_1 = -1$ in (46), so that

$$\zeta(u) = |u|^m \text{sign}(u), \ \eta(u) = u, \ \boldsymbol{q} = \rho_0 + \rho_1 x. \tag{47}$$

We consider $\Omega = (-5.5, 5.5)$ and homogeneous Neumann boundary condition with the following initial value

$$u_0(x) = \begin{cases} 1, & 0.7 < |x| < 3.7 \\ 0, & \text{otherwise.} \end{cases}$$

In this case, the evolutive equation (1) has an equilibrium solution defined by

$$u^{\text{eq}} = \left(\bar{C} - \frac{(m-1)}{2m} |x|^2 \right)_+^{\frac{1}{m-1}}$$

where $\bar{C}$ is calculated from the mass conservation condition $\int_\Omega u_0(x)\mathrm{d}x = \int_\Omega u^{\text{eq}}(x)\mathrm{d}x$ by some nonlinear solution method. Here, we choose the default nonlinear solver from `NLSolve.jl`. Note

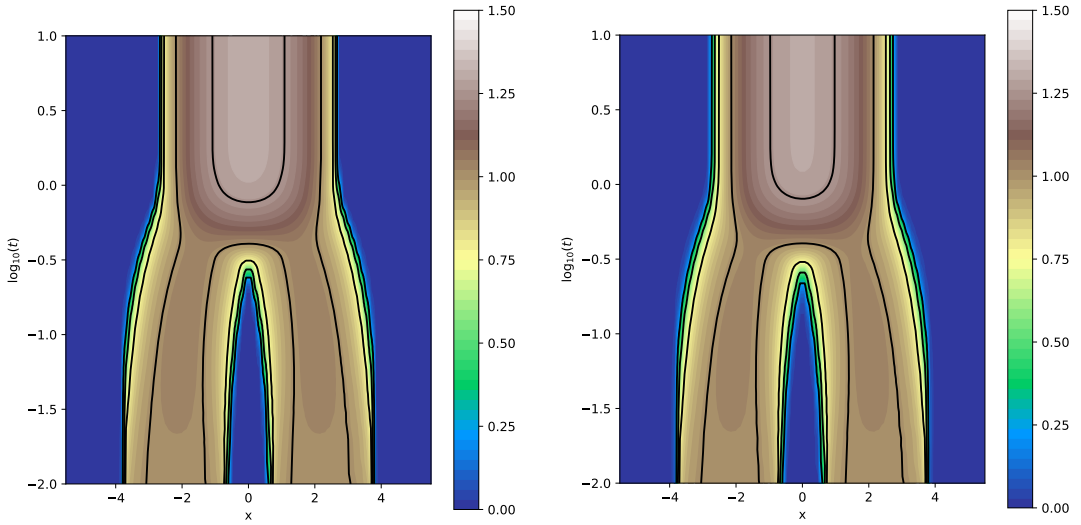

Fig. 3: Time evolution of the solution of (47) for $\delta = 0.01$. Left: $\alpha = 0$. Right: $\alpha = 0.1$.

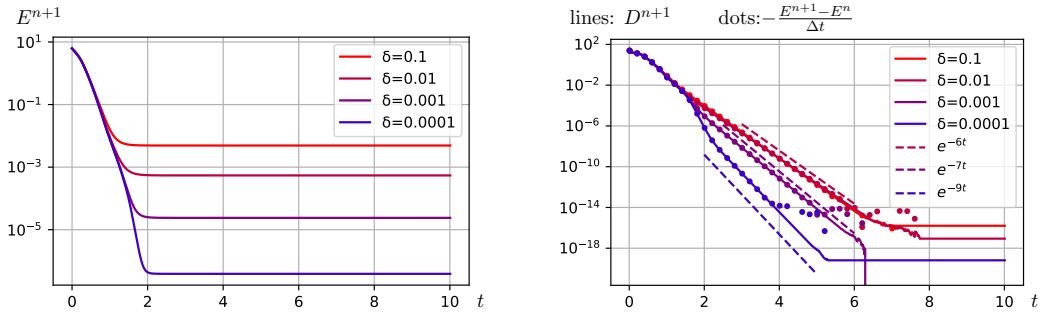

Fig. 4: Case $\alpha = 0$. Left: Evolution of the relative entropy E^{n+1} from (39) for different values of δ. Right: Evolution of the dissipation D^{n+1} from (40) (lines) and $-\frac{E^{n+1}-E^n}{\Delta t}$ (dots).

that this case does note enter into the mathematical framework of the paper, since 2.1 ii) does not hold. The solution for $t \in (0, 10)$ is shown in the left part of Fig. 3.
We observe in the left part of Fig. 4 that the asymptotic decay of relative entropy is prescribed by the value of δ. The right part of Fig. 4 shows that, in this degenerate case, the rate of the dissipation decay seems to depend on δ (such a result is not stated by Theorem 3.11, established in the case of nondegenerate problems). Moreover, the quantity $D^{n+1} + \frac{E^{n+1}-E^n}{\Delta t}$ seems to behave as the precision of the computer, which means that, in Lemma 3.7, there is a numerical evidence that $\beta = 0$ could be chosen. If we compare the results with those provided in [2, Example 8], we do not retrieve the rate -12, but we see that this rate decreases as δ tends to 0 (it is equal to -9 for the smallest value of δ). This can be intuitively expected, since $\mathcal{F}_\delta(a, b, q, h)$ converges to $\mathcal{F}(a, b, q, h)$ as $\delta \to 0$.

Entropy decay in a nondegenerate parabolic case

As noticed in the preceding example, the choice $\alpha = 0$ is not compatible with 2.1 ii). Therefore we investigate the case $\alpha = 0.1$ in (46) with the same initial value.
The equilibrium solution $u^{\text{eq}}(x)$ can be calculated from the condition

$$\alpha \log u + \frac{m}{m-1} u^{m-1} - (\bar{C} - \frac{x^2}{2}) = 0$$

using e.g. Newton's method. To achieve this, the logarithm is regularized such that for a given $\varepsilon > 0$ we replace, for any $x < \varepsilon$, $\log x$ by $\log \varepsilon + (x-\varepsilon)/\varepsilon$, and we can tolerate small negative values during the iteration. We note that in this case, the solution loses its finite support property, however for this particular example, the absolute values of $u^{eq}(x)$ and $u(x, 10)$ at the boundaries are less than 10^{-18}, and we get meaningful data for the comparison on a finite domain.

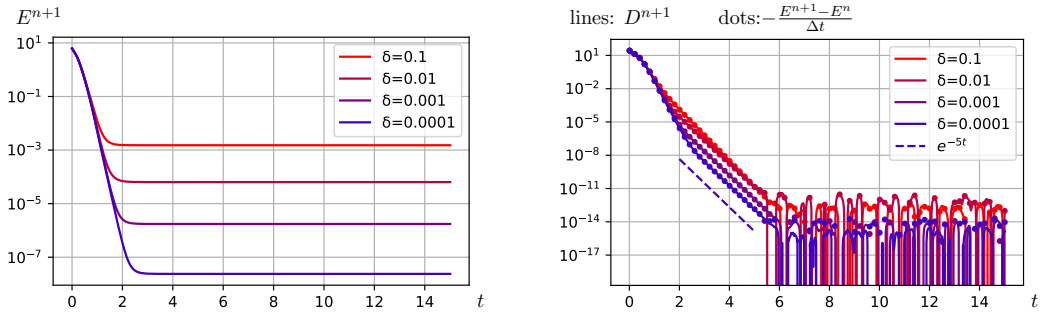

Fig. 5: Case $\alpha = 0.1$. Left: Evolution of the relative entropy E^{n+1} from (39) for different values of δ. Right: Evolution of the dissipation D^{n+1} from (40) (lines) and $-\frac{E^{n+1}-E^n}{\Delta t}$ (dots).

The solution for $t \in (0, 10)$ is shown in the right part of Fig. 3. As in the case $\alpha = 0$, we again observe the dependence of the asymptotic decay of relative entropy with respect to δ (left part of Fig. 5). But, contrary to the case $\alpha = 0$, the dissipation decay rate (see right part of Fig. 5) seems to no longer depend on δ. As stated by Theorem 3.11, the values of the dissipation are proportional to a linear function of δ for n large enough. Since the vertical axis is in logarithmic scale, this is approximately observed in the right part of Fig. 5 for sufficiently small values of δ and sufficiently large times. Here again, $D^{n+1} + \frac{E^{n+1}-E^n}{\Delta t}$ seem to behave as the precision of the computer, which again means that, in Lemma 3.7, there is a numerical evidence that $\beta = 0$ could be chosen. Note that the proof of this result is an open problem.

A Some properties of the continuous model

We have the following properties.

Lemma A.1 (Existence and uniqueness of a solution to the continuous problem):
Under Assumptions 2.1 and 3.1, there exists one and only one solution u such that, for all $T > 0$, $u \in L^2(0, T; H^1(\Omega))$ with $u_t \in L^2(0, T; H^1(\Omega)')$ to Problem (1) supplemented with Neumann

boundary conditions and an initial data u_0. Moreover, the solution u satisfies $u \geq 0$ a.e. and $\int_\Omega u(.,t) = \int_\Omega u_0$ for all $t \in \mathbb{R}^+$.

Proof. The hypothesis $\zeta' \geq r$ implies that the space operator is strongly elliptic. Standard results lead to the existence of a unique solution $u \in L^2(0,T;H^1(\Omega))$ with $u_t \in L^2(0,T;H^1(\Omega)')$. The integration of (1) on Ω provides the conclusion. ∎

Lemma A.2 (Thermal equilibrium): There exists one and only one function $u_\infty \in L^\infty(\Omega)$ with $\underline{u_\infty} := \operatorname{essinf} u_\infty > 0$ such that

$$\int_\Omega u_\infty = \int_\Omega u_0$$

and $\mu(u_\infty) + V$ is constant in Ω.

Proof. If such a function exists, it must verify that there exists $\lambda \in \mathbb{R}$ be such that $\mu(u_\infty)+V = \lambda$ a.e. in Ω. This value λ must therefore satisfy

$$\int_\Omega \mu^{-1}(\lambda - V) = \int_\Omega u_0. \tag{48}$$

By dominated convergence, we have

$$\lim_{\lambda \to -\infty} \int_\Omega \mu^{-1}(\lambda - V) = 0,$$
$$\text{and } \lim_{\lambda \to +\infty} \int_\Omega \mu^{-1}(\lambda - V) = +\infty.$$

Since the function $\lambda \mapsto \int_\Omega \mu^{-1}(\lambda - V)$ is continuous and strictly increasing, we obtain the existence and uniqueness of λ satisfying (48), which implies that the function u_∞ defined by

$$u_\infty := \mu^{-1}(\lambda - V)$$

is the only one which satisfies the conclusions of the lemma. ∎

Lemma A.3 (Strict positivity of the solution and bounds): There exist reals $\overline{B} \geq \underline{B} > 0$, only depending on $\operatorname{essinf} V$, $\operatorname{esssup} V$, $\underline{u}_0$, $\overline{u}_0$ and μ, such that $\overline{B} \geq u \geq \underline{B}$ a.e. in $\Omega \times \mathbb{R}^+$.

Proof. Let $\lambda \in \mathbb{R}$ such that

$$\mu^{-1}(-\operatorname{essinf} V + \lambda) \leq \underline{u}_0.$$

Then the function $w(x,t) := \mu^{-1}(-V(x)+\lambda)$ is solution to Problem (1) with the initial condition $w_0 := \mu^{-1}(-V+\lambda) \leq u_0$. By monotony, we deduce that $\mu^{-1}(-\operatorname{esssup} V + \lambda) \leq w \leq u$ a.e. in $\Omega \times \mathbb{R}^+$.
Similarly, let $\lambda \in \mathbb{R}$ such that

$$\mu^{-1}(-\operatorname{esssup} V + \lambda) \geq \overline{u}_0.$$

Then the function $w(x,t) := \mu^{-1}(-V(x)+\lambda)$ is solution to Problem (1) with the initial condition $w_0 := \mu^{-1}(-V+\lambda) \geq u_0$. By monotony, we deduce that $\mu^{-1}(-\operatorname{essinf} V + \lambda) \geq w \geq u$ a.e. in $\Omega \times \mathbb{R}^+$. ∎

Lemma A.4 (Long time behavior): Let $\Phi(s) := \int_1^s \mu(a)\mathrm{d}a$ and

$$E(t) = \int_\Omega \Big(\Phi(u(x,t)) - \Phi(u_\infty(x)) - \mu(u_\infty(x))(u(x,t) - u_\infty(x))\Big)\mathrm{d}x.$$

Then there exists $C > 0$, only depending on Ω, d, essinf V, esssup V, $\underline{u}_0$, $\overline{u}_0$, η and μ, such that $E(t) \leq E(0)\exp(Ct)$.

Proof. We have, multiplying Problem (1) by $\mu(u(x,t))-\mu(u_\infty(x))$ and integrating on $\Omega\times(0,T)$,

$$E(T) - E(0) + \int_0^T \int_\Omega \eta(u(x,t))|\nabla(\mu(u(x,t)) - \mu(u_\infty(x)))|^2\mathrm{d}x\mathrm{d}t = 0$$

Let

$$C(t) = \frac{1}{m_\Omega}\int_\Omega (\mu(u(x,t)) - \mu(u_\infty(x)))\mathrm{d}x.$$

From the Poincaré inequality on functions with null average, we know that there exists C_P such that

$$\int_\Omega w(x,t)^2\mathrm{d}x \leq C_P \int_\Omega |\nabla(\mu(u(x,t)) - \mu(u_\infty(x)))|^2\mathrm{d}x,$$

where we define $w(x,t) = \mu(u(x,t)) - \mu(u_\infty(x)) - C(t)$. Observing that, for all $t \in \mathbb{R}^+$, we have

$$\int_\Omega (u(x,t) - u_\infty(x))\mathrm{d}x = 0,$$

and that, applying the mean value theorem for the function μ^{-1} between $\mu(u(x,t))$ and $\mu(u_\infty(x))$

$$u(x,t) = u_\infty(x) + \alpha(x,t)(w(x,t) + C(t)),$$

where $\alpha(x,t) = (\mu^{-1})'(\beta(x,t))$, where $\beta(x,t)$ belongs to the interval with bounds $\mu(u(x,t))$ and $\mu(u_\infty(x))$. From Lemma A.3, this implies that

$$\alpha(x,t) \in [\frac{1}{\max_{\underline{B},\overline{B}}\mu'}, \frac{1}{\min_{\underline{B},\overline{B}}\mu'}]$$

This yields, by integration of the preceding equality on Ω,

$$\int_\Omega \alpha(x,t)(w(x,t) + C(t))\mathrm{d}x = 0,$$

and therefore

$$C(t) = -\frac{1}{\int_\Omega \alpha(x,t)\mathrm{d}x}\int_\Omega \alpha(x,t)w(x,t)\mathrm{d}x.$$

We therefore get that

$$\begin{aligned}\int_\Omega (u(x,t) - u_\infty(x))(\mu(u(x,t)) - \mu(u_\infty(x)))\mathrm{d}x &= \int_\Omega \alpha(x,t)(w(x,t) + C(t))^2\mathrm{d}x \\ &= C(t)^2\int_\Omega \alpha(x,t)\mathrm{d}x + 2C(t)\int_\Omega \alpha(x,t)w(x,t)\mathrm{d}x + \int_\Omega \alpha(x,t)w(x,t)^2\mathrm{d}x \\ &= -C(t)^2\int_\Omega \alpha(x,t)\mathrm{d}x + \int_\Omega \alpha(x,t)w(x,t)^2\mathrm{d}x \leq \frac{1}{\min_{\underline{B},\overline{B}}\mu'}\int_\Omega w(x,t)^2.\end{aligned}$$

Since we have

$$E(t) \leq \int_\Omega (u(x,t) - u_\infty(x))(\mu(u(x,t)) - \mu(u_\infty(x)))\mathrm{d}x,$$

we deduce that

$$E(t) \leq \operatorname{esssup} \alpha \; C_P \int_\Omega |\nabla(\mu(u(x,t)) - \mu(u_\infty(x)))|^2 \mathrm{d}x$$
$$\leq \frac{C_P}{\min_{\underline{B},\overline{B}} \mu' \min_{\underline{B},\overline{B}} \eta} \int_\Omega \eta(u(x,t)|\nabla(\mu(u(x,t)) - \mu(u_\infty(x)))|^2 \mathrm{d}x.$$

Gathering the previous inequalities, we get

$$E(T) - E(0) + \frac{\min_{\underline{B},\overline{B}} \mu' \min_{\underline{B},\overline{B}} \eta}{C_P} \int_0^T E(t)\mathrm{d}t \leq 0,$$

which provides the result using the Gronwall inequality. ∎

References

[1] Marianne 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, feb 2012.

[2] Marianne Bessemoulin-Chatard and Francis Filbet. A finite volume scheme for nonlinear degenerate parabolic equations. *SIAM Journal on Scientific Computing*, 34(5):B559–B583, jan 2012.

[3] Clément Cancès, Claire Chainais-Hillairet, Maxime Herda, and Stella Krell. Large time behavior of nonlinear finite volume schemes for convection-diffusion equations. *SIAM Journal on Numerical Analysis*, 58(5):2544–2571, jan 2020.

[4] José A. Carrillo, Ansgar Jüngel, Peter A. Markowich, Giuseppe Toscani, and Andreas Unterreiter. Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities. *Monatshefte für Mathematik*, 133(1):1–82, may 2001.

[5] José A. Carrillo, Philippe Laurençot, and Jesús Rosado. Fermi–Dirac–Fokker–Planck equation: Well-posedness & long-time asymptotics. *Journal of Differential Equations*, 247(8):2209–2234, oct 2009.

[6] José A. Carrillo, Jesús Rosado, and Francesco Salvarani. 1d nonlinear Fokker–Planck equations for fermions and bosons. *Applied Mathematics Letters*, 21(2):148–154, feb 2008.

[7] José A. Carrillo and Giuseppe Toscani. Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity. *Indiana University Mathematics Journal*, 49(1):0–0, 2000.

[8] Claire Chainais-Hillairet and Jérôme Droniou. Finite-volume schemes for noncoercive elliptic problems with neumann boundary conditions. *IMA Journal of Numerical Analysis*, 31(1):61–85, aug 2009.

[9] Claire Chainais-Hillairet and Maxime Herda. Large-time behaviour of a family of finite volume schemes for boundary-driven convection–diffusion equations. *IMA Journal of Numerical Analysis*, 40(4):2473–2504, nov 2019.

[10] Robert Eymard, Jürgen Fuhrmann, and Klaus Gärtner. A finite volume scheme for nonlinear parabolic equations derived from one-dimensional local Dirichlet problems. *Numer. Math.*, 102(3):463–495, 2006.

[11] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. In *Handbook of numerical analysis, Vol. VII*, Handb. Numer. Anal., VII, pages 713–1020. North-Holland, Amsterdam, 2000.

[12] Patricio Farrell, Thomas Koprucki, and Jürgen Fuhrmann. Computational and analytical comparison of flux discretizations for the semiconductor device equations beyond boltzmann statistics. *Journal of Computational Physics*, 346:497–513, oct 2017.

[13] Francis Filbet and Maxime Herda. A finite volume scheme for boundary-driven convection–diffusion equations with relative entropy structure. *Numerische Mathematik*, 137(3):535–577, apr 2017.

[14] Herbert Gajewski and Konrad Gröger. On the basic equations for carrier transport in semiconductors. *Journal of Mathematical Analysis and Applications*, 113(1):12–35, jan 1986.

[15] Herbert Gajewski and Konrad Gröger. Semiconductor equations for variable mobilities based on Boltzmann statistics or Fermi-Dirac statistics. *Mathematische Nachrichten*, 140(1):7–36, 1989.

[16] Martin Heida, Markus Kantner, and Artur Stephan. Consistency and convergence for a family of finite volume discretizations of the Fokker-Planck operator. *ESAIM: M2AN*, 55(6):3017–3042, 2021.

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

[18] Giorgio Kaniadakis. Generalized boltzmann equation describing the dynamics of bosons and fermions. *Physics Letters A*, 203(4):229–234, jul 1995.

[19] R. D. Lazarov, Ilya D. Mishev, and P. S. Vassilevski. Finite volume methods for convection-diffusion problems. *SIAM Journal on Numerical Analysis*, 33(1):31–55, feb 1996.

[20] Lars Onsager. Reciprocal relations in irreversible processes. i. *Physical Review*, 37(4):405–426, feb 1931.

[21] Lars Onsager. Reciprocal relations in irreversible processes. II. *Physical Review*, 38(12):2265–2279, dec 1931.

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