Finite volume schemes for non-coercive elliptic problems with Neumann boundary conditions

Claire Chainais-Hillairet [1], Jérôme Droniou [2].
06/01/2009

Abstract

We consider a convective-diffusive elliptic problem with Neumann boundary conditions: the presence of the convective term entails the non-coercivity of the continuous equation and, because of the boundary conditions, the equation has a kernel. We discretize this equation with finite volume techniques and in a general framework which allows to consider several treatments of the convective term: either via a centered scheme, an upwind scheme (widely used in fluid mechanics problems) or a Scharfetter-Gummel scheme (common to semiconductor literature). We prove that these schemes satisfy the same properties as the continuous problem (one-dimensional kernel spanned by a positive function for instance) and that their kernel and solution converge to the kernel and solution of the PDE. We also present several numerical implementations, studying the effects of the choice of one scheme or the other in the approximation of the solution or the kernel.

Keywords: convection-diffusion equations, Neumann boundary conditions, finite volume schemes, numerical analysis.

1 Introduction

We are interested in the finite volume approximation of the following convection-diffusion equation with Neumann boundary conditions:

$$\begin{cases} -\Delta \bar{u} + \operatorname{div}(\mathbf{V}\bar{u}) = g & \text{in } \Omega \\ \nabla \bar{u} \cdot \mathbf{n} - (\mathbf{V} \cdot \mathbf{n})\bar{u} = 0 & \text{on } \partial\Omega \end{cases} \tag{1.1}$$

where

$$\begin{gathered} \Omega \text{ is a bounded polygonal connected domain of } \mathbb{R}^d \ (d \geq 2), \\ g \in L^2(\Omega)\,, \\ V \in L^p(\Omega)^d \text{ with } 2 < p < +\infty \text{ if } d = 2 \text{ and } p = d \text{ if } d \geq 3. \end{gathered} \tag{1.2}$$

The solution to (1.1) is understood in the usual weak sense

$$\begin{cases} \bar{u} \in H^1(\Omega)\,, \\ \forall \varphi \in H^1(\Omega)\,, \ \displaystyle\int_\Omega \nabla \bar{u} \cdot \nabla \varphi - \int_\Omega \bar{u}\mathbf{V} \cdot \nabla \varphi = \int_\Omega g\varphi. \end{cases} \tag{1.3}$$

[1]Laboratoire de Mathématiques, UMR CNRS 6620, Université Blaise Pascal, 63177 Aubière Cedex, France. email: Claire.Chainais@math.univ-bpclermont.fr

[2]Département de Mathématiques, UMR CNRS 5149, CC 051, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier cedex 5, France. email: droniou@math.univ-montp2.fr

The study of such convection-diffusion problems is not difficult if we assume that $\mathrm{div}(\mathbf{V}) \geq 0$ in Ω and $\mathbf{V} \cdot \mathbf{n} \leq 0$ on $\partial\Omega$ (in weak senses, if $\mathbf{V}$ is not regular). Indeed, in this case, the bilinear form in (1.3) is clearly coercive and the Lax-Milgram theorem can be applied. These conditions can in fact be relaxed a little bit because the bilinear form remains coercive if $\mathrm{div}(\mathbf{V})$ and $-\mathbf{V} \cdot \mathbf{n}$ are not too negative. But, the problem becomes more complex if we impose no assumption on these quantities: the bilinear form in (1.3) is then not coercive in general and the existence of a solution to this equation is not obvious. This situation can appear if we consider a passive scalar quantity which simultaneously diffuses into a compressible flow and is convected by it: in this case, the scalar satisfies the transient form of (1.1) and, if we look for a stationary regime, we arrive at (1.1) without any specific assumption on $\mathrm{div}(\mathbf{V})$ (because of the compressibility of the flow).
Despite the lack of assumption on $\mathbf{V}$, existence and uniqueness of the weak solution to (1.1) has been proved by J. Droniou and J.-L. Vázquez in [7]. The same result had been previously obtained by J. Droniou in [4] in the case of Dirichlet, Fourier and mixed boundary conditions. The main difference between the two families of boundary conditions is that in one case (Dirichlet/Fourier/mixed) there always exists a unique solution, whereas in the other (Neumann), the right hand side of (1.1) must satisfy

$$\int_\Omega g = 0$$

in order that there exists a weak solution, and this solution is never unique (the operator has a kernel). Moreover, for Dirichlet, Fourier or mixed boundary conditions, the existence of a solution is obtained via direct explicit estimates (see [4]), while this is not the case for Neumann boundary conditions ([7] relies on abstract functional analysis — the Fredholm theory, mainly). Indeed, if we consider (1.1) with a lower order term:

$$\left\{ \begin{array}{ll} -\Delta\bar{u} + \mathrm{div}(\mathbf{V}\bar{u}) + \gamma\bar{u} = g & \text{in } \Omega \\ \nabla\bar{u} \cdot \mathbf{n} - (\mathbf{V} \cdot \mathbf{n})\bar{u} = 0 & \text{on } \partial\Omega \end{array} \right. \tag{1.4}$$

with $\gamma > 0$, it becomes possible to make direct estimates to prove the existence and uniqueness of a solution to the weak formulation of (1.4) (this has been done in [4], and the technique can be adapted to Neumann boundary conditions). In fact, the study of (1.1) requires to make estimates on (1.4) at least for a large γ.

As shown in the book by R. Eymard, T. Gallouët and R. Herbin [10], finite volume schemes are based on the conservation of physical quantities (conservation and balance of the fluxes, for example), and are therefore well suited to discretize equations coming in particular from fluid mechanics (the finite volume techniques are particularly popular in the oil engineering field). In this paper, we intend to propose some finite volume schemes for the numerical approximation of (1.1) and to establish their convergence. As a by-product, we also get the convergence of some schemes for (1.4).
In finite volume methods, the proof of convergence of a subsequence of approximate solutions toward the solution $\bar{u}$ of the PDE usually does not require theoretical results on the PDE itself: the classical finite volume techniques rely on *a priori* estimates and

compactness properties which are proved on the approximate solution by simply adapting some continuous functional analysis to the discrete setting. Thus, the study of these schemes gives, as a by-product, the existence of a solution to the PDE (see [10]). For example, the technique of [4] (for non-coercive elliptic equations with Dirichlet boundary conditions) was adapted to the discrete setting in [6] by J. Droniou and T. Gallouët. The situation for (1.1) is quite different: because the technique of estimate in [7] is quite abstract, we cannot adapt it to find direct estimates in the discrete setting. During the study of the scheme for (1.1), we will therefore much more rely, even only to prove *a priori* estimate on the solutions to the scheme, on the known results of [7] on (1.1). The proof of convergence we propose in this paper is therefore not classical for finite volume methods.

There exists several studies of numerical methods to approximate non-coercive problems, either in a general setting [17] or more specifically for convection-diffusion equations [14] or the Helmholtz equation [15, 11]; to our best knowledge, these studies concern finite element or discontinuous Galerkin methods, which are either conformal (i.e. the approximate solutions belong to the same space as the solution to the continuous problem) or close to conformal approximations. Moreover, they all require that the continuous variational problem (and its adjoint) has a unique solution, and they only ensure that the linear system corresponding to the scheme is solvable for a mesh size small enough (an estimate of this smallness can be made, but it does not seem very explicit or easy to use in practice, see [1]). In the present situation, [7] shows that (1.1) has a unique solution in the space of functions in $H^1(\Omega)$ with mean value zero, and the results in the literature thus ensure that, using for example a conformal finite element approximation of (1.1) in this space, we would obtain a scheme which converges to the unique solution of (1.1) with mean value zero. However, the existence of the approximate solution would only be ensured for a mesh size small enough, and this method would not provide a practical approximation of the kernel of the PDE.
The finite volume approximation we suggest here does not require to eliminate from the start the kernel of (1.1) and, as a consequence, makes it possible to approximate not only the solution (with mean value zero) to this problem, but also its kernel. Moreover, thanks to the maximum principle satisfied by the finite volume method, no restriction on the mesh size is required to establish the solvability of the finite dimensional linear system giving the approximate solution.

The paper is organized as follows. In the next chapter, we introduce the finite volume schemes for (1.1) and (1.4); we use a general formulation which allows for several different discretizations of the convection term, such as the centered, upwind or Scharfetter-Gummel discretization (we give the expressions of each of these specific schemes). We conclude this next section by stating the main theoretical results of the paper: existence, uniqueness and convergence of the approximate solutions. As said above, Problem (1.1) has a kernel, and we also include in the study of the solutions to the scheme the fact that the kernel of the discretization converge to the kernel of (1.1). In Section 3, we give some properties of the matrices associated with the schemes, which allow us to study the existence and uniqueness of the solutions and the kernel of the schemes. Section 4 is

devoted to obtaining estimates on these solutions, first for large γ and then for any $\gamma \geq 0$, and to prove the convergence of the scheme (along with its kernel) toward the continuous problem. We provide in Section 5 numerical results obtained with the scheme; we illustrate both the convergence of the solution and of the kernel, giving numerical orders of convergence for various choices of schemes (centered, upwind or Scharfetter-Gummel). Finally, an appendix (Section 6) gives the proof of a discrete Sobolev inequality needed during the theoretical study of the schemes.

2 The finite volume schemes

We first define the notion of admissible discretization of Ω, following Definition 5.1 in [10].

Definition 2.1 (Admissible mesh) *An admissible mesh $\mathcal{M}$ of Ω is given by a finite family $\mathcal{T}$ of disjoint open convex polygonal subsets of Ω (the control volumes), a finite family $\mathcal{E}$ of disjoint subsets of $\overline{\Omega}$ (the edges) consisting in non-empty open convex subsets of affine hyperplanes and a family $\mathcal{P} = (x_K)_{K\in\mathcal{T}}$ of points in Ω such that:*

- $\overline{\Omega} = \cup_{K\in\mathcal{T}} \overline{K}$,
- *each $\sigma \in \mathcal{E}$, is contained in ∂K for some $K \in \mathcal{T}$,*
- *for all $K \in \mathcal{T}$, denoting $\mathcal{E}_K = \{\sigma \in \mathcal{E}\,,\ \sigma \subset \partial K\}$, $\partial K = \cup_{\sigma\in\mathcal{E}_K} \sigma$,*
- *for all $K \neq L$ in $\mathcal{T}$, either the $(d-1)$-dimensional measure of $\overline{K} \cap \overline{L}$ is zero or $\overline{K} \cap \overline{L} = \overline{\sigma}$ for some $\sigma \in \mathcal{E}$, which is then denoted $\sigma = K|L$,*
- *for all $K \in \mathcal{T}$, $x_K \in K$,*
- *for all $\sigma = K|L \in \mathcal{E}$, the straight line (x_K, x_L) is orthogonal to σ,*
- *for all $\sigma \in \mathcal{E}$ such that $\sigma \subset \partial\Omega \cap \partial K$, the line which is orthogonal to σ and goes through x_K intersects σ.*

It will be useful to introduce a few more notations associated with an admissible mesh. In the set of edges $\mathcal{E}$, we distinguish the set of interior edges $\mathcal{E}_{\rm int}$ (the edges contained in Ω) and the set of boundary edges $\mathcal{E}_{\rm ext}$ (the edges contained in $\partial\Omega$). The d-dimensional measure of a control volume K is $\mathrm{m}(K)$; similarly, the $(d-1)$-dimensional measure of an edge σ is $\mathrm{m}(\sigma)$. For a control volume $K \in \mathcal{T}$, we denote by $\mathcal{E}_{K,\rm int} = \mathcal{E}_{\rm int} \cap \mathcal{E}_K$ the set of its interior edges and by $N(K)$ the set of its neighbouring control volumes (i.e. the control volumes L such that $\overline{K} \cap \overline{L}$ is an edge of the discretization). For $\sigma \in \mathcal{E}_K$, $\mathbf{n}_{K,\sigma}$ is the unit normal to σ outwards K. If $\sigma \in \mathcal{E}_{K,\rm int}$, it means that $\sigma = K|L$ and we then denote d_σ the distance $\mathrm{d}(x_K, x_L)$. The size of the mesh is $\mathrm{size}(\mathcal{M}) = \sup_{K\in\mathcal{T}} \mathrm{diam}(K)$. See Figure 1 for the illustration of some of these assumptions and notations.

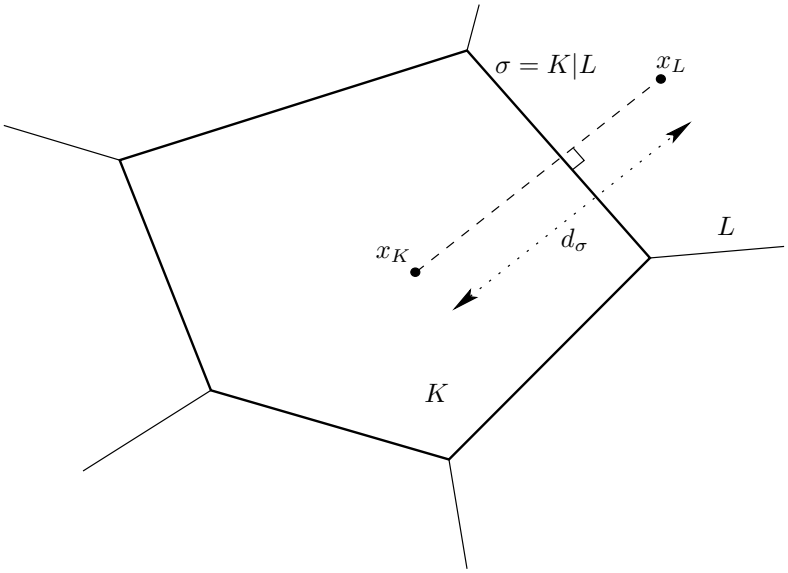


Figure 1: Illustration of the assumptions and notations on the discretization.

We will need discrete Sobolev inequalities (see Lemma 6.1), which depend on the constant ζ appearing in the following assumption.

$$\exists \zeta > 0 \text{ such that } \forall K \in \mathcal{T}\ ,\ \forall \sigma \in \mathcal{E}_K\ ,\ \mathrm{d}(x_K, \sigma) \geq \zeta d_\sigma. \tag{2.5}$$

The principle of finite volume schemes for convection-diffusion problems is to write a flux balance, using quantities $\mathcal{F}_{K,\sigma}$ which approximate $\int_\sigma (-\nabla \bar{u} \cdot \mathbf{n}_{K,\sigma} + (\mathbf{V} \cdot \mathbf{n}_{K,\sigma})\bar{u})$. In order to do so, we will need discretizations of the fluxes of $\mathbf{V}$ through the edges of the mesh:

$$v_{K,\sigma} = \frac{1}{\mathrm{m}(\mathcal{D}_\sigma)} \int_{\mathcal{D}_\sigma} \mathbf{V} \cdot \mathbf{n}_{K,\sigma}\, dx, \tag{2.6}$$

where $\mathcal{D}_\sigma$ is the diamond around σ, i.e. the convex hull of σ and $\{x_K, x_L\}$ if $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$ or the convex hull of σ and x_K if $\sigma \in \mathcal{E}_{K,\mathrm{int}}$.

Remark 2.2 *Other definitions of $v_{K,\sigma}$ are possible. In the case where $\mathbf{V}$ is continuous, a classical choice is $v_{K,\sigma} = \frac{1}{\mathrm{m}(\sigma)} \int_\sigma \mathbf{V} \cdot \mathbf{n}_{K,\sigma}$ or $v_{K,\sigma} = \mathbf{V}(x_\sigma) \cdot \mathbf{n}_{K,\sigma}$ (with $x_\sigma \in \sigma$). If $\mathbf{V}$ comes from a potential Φ ($\mathbf{V} = \nabla \Phi$), we can also choose $v_{K,\sigma} = \frac{\Phi(x_L) - \Phi(x_K)}{\mathrm{d}_\sigma}$ (see the interest of this choice in Remark 3.3). As one can convince himself by reading the proofs below, all these different choices (provided that $\mathbf{V}$ is regular enough so that they make sense) entail very little changes in the study of the scheme.*

We need now to explain how to construct the approximation of the fluxes $\mathcal{F}_{K,\sigma}$, using only approximate values $(u_K)_{K \in \mathcal{T}}$ of the solution inside the control volumes (these will be the unknowns of the system describing the scheme). Let us first consider three particular and well-known cases.

2.1 Centered fluxes

A simple definition of the numerical flux can be

$$\mathcal{F}_{K,\sigma} = \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}(u_K - u_L) + \mathrm{m}(\sigma) v_{K,\sigma} \frac{u_K + u_L}{2}, \quad \forall \sigma = K|L \in \mathcal{E}_{\mathrm{int}}.$$

Assuming that u_K (*resp.* u_L) is an approximation of $\bar{u}$ at x_K (*resp.* x_L), the orthogonality between (x_K, x_L) and σ ensures that $\frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}(u_K - u_L)$ is a consistent approximation of $\int_\sigma -\nabla \bar{u} \cdot \mathbf{n}_{K,\sigma}$. The quantity $\mathrm{m}(\sigma) v_{K,\sigma} \frac{u_K + u_L}{2}$ is a simple approximation of the convective flux $\int_\sigma \bar{u} \mathbf{V} \cdot \mathbf{n}_{K,\sigma}$, taking as value of $\bar{u}$ on σ the average of the values on both sides of the edge.
Note that by defining $B_{\mathrm{ce}}(s) = 1 - \dfrac{s}{2}$, these fluxes can be written

$$\mathcal{F}_{K,\sigma} = \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma} \Big(B_{\mathrm{ce}}(-v_{K,\sigma}\mathrm{d}_\sigma) u_K - B_{\mathrm{ce}}(v_{K,\sigma}\mathrm{d}_\sigma) u_L \Big).$$

The centered scheme introduces very little numerical diffusion in the discretization of the convective term, but is therefore also not very stable if the convection is much stronger than the natural diffusion (a Peclet condition must be imposed to prove the stability of the approximate solution). Other fluxes are therefore usually considered.

2.2 Upwind fluxes

Another definition of the numerical flux can be

$$\mathcal{F}_{K,\sigma} = \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}(u_K - u_L) + \mathrm{m}(\sigma)(v^+_{K,\sigma} u_K - v^-_{K,\sigma} u_L), \quad \forall \sigma = K|L \in \mathcal{E}_{\mathrm{int}},$$

where $s^+ = \max(s, 0)$ and $s^- = \max(-s, 0)$ are the positive and negative parts of a real number s. The discretization of the diffusive part is the same as before, but an upwind dicretization is used for the convective part, which stabilizes the scheme (at the cost of the introduction of an additional numerical diffusion).
Defining $B_{\mathrm{up}}(s) = 1 + (-s)^+ = 1 + s^-$, the fluxes of the upwind scheme can be written

$$\mathcal{F}_{K,\sigma} = \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma} \Big(B_{\mathrm{up}}(-v_{K,\sigma}\mathrm{d}_\sigma) u_K - B_{\mathrm{up}}(v_{K,\sigma}\mathrm{d}_\sigma) u_L \Big).$$

2.3 The Scharfetter-Gummel fluxes

The fluxes we now introduce are well-known in the semiconductor framework. They have been proposed by D.L. Scharfetter and H.K. Gummel in [16] for the numerical approximation of the 1D drift-diffusion model and numerical simulation of a silicon Read diode (see [13] for a detailed presentation). We also refer to the work by A.M. Il'in [12] where the same kind of fluxes is introduced for 1D finite difference schemes.

In a multidimensional context, the numerical fluxes of Scharfetter and Gummel are obtained by solving a one-dimensional ODE on the staight line $[x_K, x_L]$. On this line, we set $u(x) = u(x_K + \theta(x_L - x_K)) = \tilde{u}(\theta)$ with $\theta \in [0,1]$. Therefore $\tilde{u}'(\theta) = \mathrm{d}_\sigma \nabla u(x_K + \theta(x_L - x_K)) \cdot \mathbf{n}_{K,\sigma}$. Solving

$$\begin{cases} -\dfrac{1}{\mathrm{d}_\sigma}\tilde{u}'(\theta) + v_{K,\sigma}\tilde{u}(\theta) = \dfrac{\mathcal{F}_{K,\sigma}}{\mathrm{m}(\sigma)}\,, \qquad \theta \in [0,1], \\ \tilde{u}(0) = u_K. \end{cases}$$

we get

$$\tilde{u}(\theta) = \frac{\mathcal{F}_{K,\sigma}}{\mathrm{m}(\sigma)v_{K,\sigma}} + \left(u_K - \frac{\mathcal{F}_{K,\sigma}}{\mathrm{m}(\sigma)v_{K,\sigma}}\right) e^{v_{K,\sigma}\mathrm{d}_\sigma\theta}\,, \quad \forall \theta \in [0,1],$$

and the flux $\mathcal{F}_{K,\sigma}$ is then defined by imposing $\tilde{u}(1) = u_L$, which leads to

$$\mathcal{F}_{K,\sigma} = \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}\Big(B_{\mathrm{sg}}(-v_{K,\sigma}\mathrm{d}_\sigma)u_K - B_{\mathrm{sg}}(v_{K,\sigma}\mathrm{d}_\sigma)u_L\Big)\,, \quad \forall \sigma = K|L, \tag{2.7}$$

where $B_{\mathrm{sg}} = \dfrac{s}{e^s - 1}$ is the Bernoulli function.
An extension of the Scharfetter-Gummel scheme has been studied by R. Eymard, J. Fuhrmann and K. Gärtner [8] in the case where the convection and diffusion terms are nonlinear.

2.4 Definition of the generic scheme

We notice that the functions B_{up} and B_{sg} satisfy

$$B \text{ is Lipschitz-continuous on } \mathbb{R}, \tag{2.8}$$
$$B(0) = 1 \text{ and } B(s) > 0\,, \quad \forall s \in \mathbb{R}, \tag{2.9}$$
$$B(s) - B(-s) = -s\,, \quad \forall s \in \mathbb{R}. \tag{2.10}$$

The function B_{ce} also satisfies (2.8) and (2.10), but (2.9) only if $s < 2$ (since B_{ce} is applied to $v_{K,\sigma}\mathrm{d}_\sigma$, this translates into the Peclet condition: the convection must not be too large with respect to the diffusion).
The generic form of finite volume schemes for (1.1) therefore consists in defining

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

and in writing the following system on $(u_K)_{K\in\mathcal{T}}$:

$$\sum_{\sigma\in\mathcal{E}_{K,\mathrm{int}}} \mathcal{F}_{K,\sigma} = \mathrm{m}(K)g_K\,, \quad \forall K \in \mathcal{T}, \tag{2.11}$$

$$\mathcal{F}_{K,\sigma} = \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}\Big(B(-v_{K,\sigma}\mathrm{d}_\sigma)u_K - B(v_{K,\sigma}\mathrm{d}_\sigma)u_L\Big)\,, \quad \forall \sigma = K|L, \tag{2.12}$$

where B is a function satisfying (2.8)–(2.10). The first equation (2.11) comes from the physical balance of fluxes; it can also be obtained, if we recall that $\mathcal{F}_{K,\sigma}$ approximates $\int_\sigma(-\nabla\bar{u}\cdot\mathbf{n}_{K,\sigma}+(\mathbf{V}\cdot\mathbf{n}_{K,\sigma})\bar{u})$, by integrating the PDE in (1.1) and using Stokes' formula and the boundary conditions. The second equation (2.12) defines the approximate fluxes. Note that the homogeneous Neumann boundary condition in (1.1) is taken into account in the fact that the sum in (2.11) is restricted to the interior edges of each control volume: in fact, the boundary condition consists in imposing $\mathcal{F}_{K,\sigma}=0$ for $\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}$.

Remark 2.3 *We could easily (both in (1.1) and in its discretization (2.11)–(2.12)) consider non-homogeneous Neumann boundary conditions $\nabla\bar{u}\cdot\mathbf{n}-(\mathbf{V}\cdot\mathbf{n})\bar{u}=h$. It is well known that it simply adds a supplementary term in the right-hand side of the weak formulation (1.3) and in the balance of fluxes (2.11), which becomes*

$$\sum_{\sigma\in\mathcal{E}_{K,\rm int}}\mathcal{F}_{K,\sigma}=\mathrm{m}(K)g_K+\sum_{\sigma\in\mathcal{E}_{K,\rm ext}}\int_\sigma h\,dx.$$

It will also be useful to notice that, since $v_{K,\sigma}=-v_{L,\sigma}$ whenever $\sigma=K|L$, the fluxes defined by (2.12) are conservative:

$$\mathcal{F}_{K,\sigma}=-\mathcal{F}_{L,\sigma}\,,\quad\forall\sigma=K|L.$$

Obviously, the scheme for (1.4) consists in replacing (2.11) by

$$\sum_{\sigma\in\mathcal{E}_{K,\rm int}}\mathcal{F}_{K,\sigma}+\gamma\mathrm{m}(K)u_K=\mathrm{m}(K)g_K\,,\quad\forall K\in\mathcal{T}. \tag{2.13}$$

Remark 2.4 *It is also possible to consider an heterogeneous and/or anisotropic version of (1.1), in which $-\Delta u$ is replaced by $-\mathrm{div}(A\nabla u)$ with $A:\Omega\to M_d(\mathbb{R})$ a bounded measurable uniformly elliptic matrix. If $A(x)=a(x)\mathrm{I}$ (isotropic case), the modification of the scheme is minimal; if A is anisotropic, then one must change the notion of admissible discretization of Ω (the orthogonality between (x_K,x_L) and $\sigma=K|L$ should then be understood with respect to a scalar product defined using A, see [10]).*

2.5 Main results

In the following, we freely identify a vector $u=(u_K)_{K\in\mathcal{T}}$ with the piecewise-constant function u equal to u_K in $K\in\mathcal{T}$.
In the continuous case, it is known (see [7]) that (1.4) has a unique solution for all $\gamma>0$ and all $g\in L^2(\Omega)$. On the other hand, (1.1) has a kernel (i.e. a set of solutions with $g=0$) of dimension 1, spanned by an everywhere positive function, and has a solution only if $\int_\Omega g=0$; this solution is unique if its mean value is fixed (for example to zero). Theorems 2.5 and 2.6 state that the discretizations of these continuous problems enjoy the same properties. Theorems 2.7 and 2.8 state convergence results.

Theorem 2.5 (Existence of a solution for the scheme on (1.1)) *Assume (1.2), and let $\mathcal{M}$ be an admissible mesh of Ω.*

1. *The kernel of the scheme (2.8)–(2.12) has dimension 1 and is spanned by a function $\widehat{u} = (\widehat{u}_K)_{K\in\mathcal{T}}$ which is everywhere strictly positive.*
2. *If $\int_\Omega g = 0$, then there exists a unique solution $u = (u_K)_{K\in\mathcal{T}}$ of (2.8)–(2.12) such that $\int_\Omega u = 0$.*

Theorem 2.6 (Existence of a solution for the scheme on (1.4)) *Assume (1.2) and let $\mathcal{M}$ be an admissible mesh of Ω. Then, for any $\gamma > 0$, there exists a unique solution $u = (u_K)_{K\in\mathcal{T}}$ to ((2.8)–(2.10),(2.12),(2.13)) .*

Theorem 2.7 (Convergence of the kernel and the solution for the scheme on (1.1)) *Assume (1.2) and let $(\mathcal{M}_n)_{n\geq 1}$ be a sequence of admissible meshes which satisfy (2.5) with ζ not depending on n, and such that $\mathrm{size}(\mathcal{M}_n) \to 0$ as $n \to \infty$.*

1. *Let $\widehat{u}_n$ be the unique positive element, with norm in $L^2(\Omega)$ equal to 1, in the kernel of (2.8)–(2.12) for $\mathcal{M} = \mathcal{M}_n$. Then, as $n \to \infty$, $\widehat{u}_n \to \widehat{\bar{u}}$ in $L^2(\Omega)$, where $\widehat{\bar{u}} \in H^1(\Omega)$ is the unique positive element, with norm in $L^2(\Omega)$ equal to 1, in the kernel of the operator in (1.1) (i.e. a weak solution to this problem with $g = 0$).*
2. *Assume that $\int_\Omega g = 0$ and let u_n be the unique solution to (2.8)–(2.12) with zero mean value. Then, as $n \to \infty$, $u_n \to \bar{u}$ in $L^2(\Omega)$, where $\bar{u} \in H^1(\Omega)$ is the unique weak solution to (1.1) with zero mean value.*

Theorem 2.8 (Convergence of the solution for the scheme on (1.4)) *Assume (1.2) and let $(\mathcal{M}_n)_{n\geq 1}$ be a sequence of admissible meshes which satisfy (2.5) with ζ not depending on n, and such that $\mathrm{size}(\mathcal{M}_n) \to 0$ as $n \to \infty$. Let $\gamma > 0$ and u_n be the unique solution to ((2.8)–(2.10),(2.12),(2.13)) . Then, as $n \to \infty$, $u_n \to \bar{u}$ in $L^2(\Omega)$, where $\bar{u} \in H^1(\Omega)$ is the unique weak solution to (1.4).*

3 Existence and uniqueness of the solutions to the schemes

We intend here to prove Theorems 2.5 and 2.6 by means of linear algebra tools. The scheme (2.8)–(2.12) leads to a linear system of equations, which can be written

$$\mathbb{A}U = G$$

where $U = (u_K)_{K\in\mathcal{T}}$, $G = (\mathrm{m}(K)g_K)_{K\in\mathcal{T}}$ and $\mathbb{A}$ is the square matrix of size $\mathrm{Card}(\mathcal{T}) \times \mathrm{Card}(\mathcal{T})$ with entries

$$\begin{aligned}
&\mathbb{A}_{K,K} = \sum_{\sigma\in\mathcal{E}_{K,\mathrm{int}}} \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma} B(-v_{K,\sigma}\mathrm{d}_\sigma)\,, \quad \forall K \in \mathcal{T},\\
&\mathbb{A}_{K,L} = -\frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma} B(v_{K,\sigma}\mathrm{d}_\sigma)\,, \qquad \forall K \in \mathcal{T}\,, \forall L \in N(K), \text{ with } \sigma = K|L,\\
&\mathbb{A}_{K,L} = 0\,, \qquad \forall K \in \mathcal{T}\,, \forall L \notin N(K).
\end{aligned} \tag{3.14}$$

The scheme ((2.8)–(2.10),(2.12),(2.13)) for (1.4) also leads to a linear system of equations $\mathbb{A}_\gamma U = G$ where, for all $\gamma > 0$, the matrix $\mathbb{A}_\gamma$ is defined by $\mathbb{A}_\gamma = \mathbb{A} + \gamma\mathbb{D}$ with $\mathbb{D}$ the diagonal matrix whose diagonal entries are $\mathbb{D}_{K,K} = \mathrm{m}(K)$. Theorems 2.5 and 2.6 are thus easy consequences of the following proposition.

Proposition 3.1 *The matrix $\mathbb{A}$ of the scheme (2.8)–(2.12) satisfies:*

1. $\mathrm{Ker}(\mathbb{A})$ *has dimension 1 and, if* $U = (u_K)_{K\in\mathcal{T}} \in \mathrm{Ker}(\mathbb{A})\backslash\{0\}$, *then either* $u_K > 0$ *for all* $K \in \mathcal{T}$ *or* $u_K < 0$ *for all* $K \in \mathcal{T}$.

2. $\mathrm{Ker}(\mathbb{A}^*) = \mathbb{R}(1, 1, \cdots, 1)^*$ *(* denotes the transpose) and thus*

$$\mathrm{Im}(\mathbb{A}) = \left\{ (G_K)_{K\in\mathcal{T}} \ ; \ \sum_{K\in\mathcal{T}} G_K = 0 \right\}.$$

3. *For all $\gamma > 0$, the diagonal coefficients of $\mathbb{A}_\gamma$ are positive, the extra-diagonal coefficients of $\mathbb{A}_\gamma$ are non-positive, and the sum of the coefficients in each column of $\mathbb{A}_\gamma$ is positive. Therefore, $\mathbb{A}_\gamma$ is an M-matrix and is invertible.*

Proof of Proposition 3.1
Thanks to (3.14) and (2.9), we notice that all the diagonal entries of $\mathbb{A}$ are strictly positive, whereas the extra-diagonal coefficients are non-positive; this is thus also the case for $\mathbb{A}_\gamma$ for all $\gamma > 0$. Moreover, since $v_{L,\sigma} = -v_{K,\sigma}$ whenever $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$, we have

$$\mathbb{A}_{K,K} = -\sum_{L\in N(K)} \mathbb{A}_{L,K}\,, \quad \forall K \in \mathcal{T}. \tag{3.15}$$

In other words, in each column, the diagonal term is the opposite of the sum of the extra-diagonal terms. This has two consequences:

- $(1, 1, \cdots, 1)^* \in \mathrm{Ker}(\mathbb{A}^*)$, and thus $\mathrm{Ker}(\mathbb{A}^*)$ and $\mathrm{Ker}(\mathbb{A})$ have at least dimension 1.
- the sum of the coefficients in the column K of $\mathbb{A}_\gamma$ is equal to $\gamma\mathrm{m}(K)$, and Item 3 of the proposition is satisfied.

It remains to prove that $\mathrm{Ker}(\mathbb{A})$ has dimension 1 and that any non-zero vector inside this kernel is either strictly positive or strictly negative. This could be done by adapting the technique used in [7] for the continuous equation, but since there is no issue of regularity of solutions in the discrete setting we prefer to use a different (and probably more straightforward) technique; notice that, in the case of regular solutions to (1.1), the following technique could be used in the continuous case.
Let us take $U = (U_K)_{K\in\mathcal{T}} \in \mathrm{Ker}(\mathbb{A})$ which is not zero; upon changing U in $-U$, we can assume that one of the coefficients of U is strictly positive. Denote then

$$\mathcal{T}_U^+ = \{K \in \mathcal{T} \ ; \ U_K > 0\} \neq \emptyset,$$
$$\mathcal{T}_U^- = \{K \in \mathcal{T} \ ; \ U_K \leq 0\}.$$

Since $U \in \mathrm{Ker}(\mathbb{A})$, by (3.15) we have, for all $K \in \mathcal{T}$,

$$\mathbb{A}_{K,K}U_K + \sum_{L\in N(K)} \mathbb{A}_{K,L}U_L = -\sum_{L\in N(K)} \mathbb{A}_{L,K}U_K + \sum_{L\in N(K)} \mathbb{A}_{K,L}U_L = 0.$$

Let us sum these equations over $K \in \mathcal{T}_U^-$:

$$-\sum_{K\in\mathcal{T}_U^-}\sum_{L\in N(K)} \mathbb{A}_{L,K}U_K + \sum_{K\in\mathcal{T}_U^-}\sum_{L\in N(K)} \mathbb{A}_{K,L}U_L = 0.$$

We now gather this sum by edges; an edge $\sigma = K|L$ between two control volumes in $\mathcal{T}_U^-$ brings two contributions to this sum, namely $-\mathbb{A}_{L,K}U_K+\mathbb{A}_{K,L}U_L$ and $-\mathbb{A}_{K,L}U_L+\mathbb{A}_{L,K}U_K$, which cancel out each other. Hence, in this sum, the only remaining contributions are those of edges between control volumes in $\mathcal{T}_U^-$ and $\mathcal{T}_U^+$; denoting $\mathcal{T}_U^{-+} = \{\sigma = K|L \in \mathcal{E}_{\mathrm{int}} \, , K \in \mathcal{T}_U^- \ , \ L \in \mathcal{T}_U^+\}$, this leads to

$$-\sum_{\sigma=K|L\in\mathcal{T}_U^{-+}} \mathbb{A}_{L,K}U_K + \sum_{\sigma=K|L\in\mathcal{T}_U^{-+}} \mathbb{A}_{K,L}U_L = 0$$

(where, in concordance with the preceding notations, for each edge $\sigma = K|L \in \mathcal{T}_U^{-+}$, K is the control volume in $\mathcal{T}_U^-$ and L is the control volume in $\mathcal{T}_U^+$).
But, if $\mathcal{T}_U^{-+}$ is not empty then, by (2.9) and (3.14), $\sum_{\sigma=K|L\in\mathcal{T}_U^{-+}} \mathbb{A}_{L,K}U_K \geq 0$ and $\sum_{\sigma=K|L\in\mathcal{T}_U^{-+}} \mathbb{A}_{K,L}U_L < 0$, which lead to a contradiction. Hence, $\mathcal{T}_U^{-+} = \emptyset$ and, since Ω is connected and $\mathcal{T}_U^+$ is not empty, this implies that $\mathcal{T}_U^-$ is empty; all the coefficients of U are therefore strictly positive.
It follows that $\mathrm{Ker}(\mathbb{A})$ is of dimension at most 1: indeed, if $U \neq 0$ and V belong to $\mathrm{Ker}(\mathbb{A})$, then so does $V + \lambda U$ for all $\lambda \in \mathbb{R}$; all the coefficients of $V + \lambda U$ must therefore be either strictly positive, or strictly negative, or zero. Since it is always possible to choose λ such that one coefficient at leat of $V + \lambda U$ is zero, such a choice gives $V + \lambda U = 0$, that is to say the colinearity of U and V. The proof is then concluded. ■

Remark 3.2 (Extension of the results of Proposition 3.1 to the centered scheme) *The results of Proposition 3.1 remain true for the centered scheme if* $B_{\rm ce}(-v_{K,\sigma}{\rm d}_\sigma) > 0$ *and* $B_{\rm ce}(v_{K,\sigma}{\rm d}_\sigma) > 0$ *for every interior edge* σ *of the mesh. Using the definition of* $B_{\rm ce}$ *for the centered scheme, it rewrites as a Péclet condition:*

$$|v_{K,\sigma}{\rm d}_\sigma| < 2, \quad \forall K \in \mathcal{T},\ \forall \sigma \in \mathcal{E}_{K,{\rm int}}. \tag{3.16}$$

Remark 3.3 (The kernel in the case of the Scharfetter-Gummel scheme with a gradient velocity) *Assume that* $\mathbf{V} = \nabla\Phi$ *with* $\Phi \in C^1(\bar\Omega)$*. Then it is easy to check that the kernel of the continuous problem (1.1) is* $\mathbb{R}e^\Phi$*; for the Scharfetter-Gummel scheme, this property is also valid at the discrete level. Indeed, if we associate with* Φ *the vector, also denoted by* Φ*, defined by* $\Phi = (\Phi_K = \Phi(x_K))_{K\in\mathcal{T}}$ *and if we take*

$$v_{K,\sigma} = \frac{\Phi_L - \Phi_K}{{\rm d}_\sigma}, \quad \forall \sigma = K|L,$$

then, plugging $u = e^\Phi$ *in (2.7) we obtain* $\mathcal{F}_{K,\sigma} = 0$ *for all* $\sigma = K|L$*, and thus* ${\rm Ker}(\mathbb{A}) = \mathbb{R}e^\Phi$*.*

Proof of Theorems 2.5 and 2.6
Theorem 2.6 is an immediate consequence of Item 3 in Proposition 3.1.
The first item of Theorem 2.5 follows from the first item in Proposition 3.1. By Item 2 in this proposition, we know that, as soon as $\int_\Omega g = 0 = \sum_{K\in\mathcal{T}} {\rm m}(K) g_K$, $G = ({\rm m}(K)g_K)_{K\in\mathcal{T}}$ belongs to ${\rm Im}(\mathbb{A})$. Therefore, there exists at least one solution to the scheme, denoted by u_0. By Item 1 in Theorem 2.5, all the solutions to (2.8)–(2.12) can be written $u = u_0 + \lambda\widehat{u}$ with $\lambda \in \mathbb{R}$, and the fact that $\int_\Omega \widehat{u} \neq 0$ shows that there is only one solution which satisfies $\int_\Omega u = 0$: it corresponds to the choice $\lambda = -\int_\Omega u_0/(\int_\Omega \widehat{u})$. ■

4 Convergence of the schemes

4.1 Estimates for large γ

The estimates on finite volume schemes are usually performed using an adequate H^1-semi-norm associated with the scheme. The H^1-semi-norm of $u = (u_K)_{K\in\mathcal{T}}$ is here defined by

$$|u|_{1,\mathcal{M}} = \left(\sum_{\sigma = K|L \in \mathcal{E}_{\rm int}} \frac{{\rm m}(\sigma)}{{\rm d}_\sigma} (u_K - u_L)^2 \right)^{1/2}. \tag{4.17}$$

The following proposition and corollary check that our schemes enjoy properties which are well-known for the continuous equations, namely: if γ is large enough, it is easy to obtain *a priori* estimates on the solution to (1.4) and, for any γ, the L^2 norm of the gradient of the solution to (1.4) is always controlled by the L^2 norm of the solution. Of course, the main difference with respect to the continuous case is that, at the discrete level, we have to make sure that these estimates do not depend on the size of the mesh.

Proposition 4.1 *Assume that (1.2) holds and that $\mathcal{M}$ satisfies (2.5). There exists $\gamma_0 = \gamma_0(\Omega, \mathbf{V}, B, \zeta) > 0$ and $C_1 = C_1(\Omega, \mathbf{V}, B, \zeta) > 0$ such that, for all $\gamma \geq \gamma_0$, if u is a solution to the scheme ((2.8)–(2.10),(2.12),(2.13)) for (1.4), then*

$$|u|_{1,\mathcal{M}}^2 + ||u||_{L^2(\Omega)}^2 \leq \frac{C_1}{\gamma}||g||_{L^2(\Omega)}^2.$$

Remark 4.2 *In the case of an upwind flux $B = B_{\rm up}$, then a simple adaptation of the technique in [6] allows to see that one can in fact take $\gamma_0 = 0$.*

Proof of Proposition 4.1
All the constants C_i appearing in this proof only depend on Ω, $\mathbf{V}$, B and ζ. We multiply (2.13) by u_K and sum over $K \in \mathcal{T}$. Thanks to the conservativity of the fluxes and to (2.12), gathering by edges we find

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \frac{{\rm m}(\sigma)}{{\rm d}_\sigma}\Big[B(-v_{K,\sigma}{\rm d}_\sigma)u_K - B(v_{K,\sigma}{\rm d}_\sigma)u_L\Big](u_K - u_L) + \gamma||u||_{L^2(\Omega)}^2 \leq ||g||_{L^2(\Omega)}||u||_{L^2(\Omega)}.$$

Subtracting and adding $B(0) = 1$ to $B(-v_{K,\sigma}{\rm d}_\sigma)$ and $B(v_{K,\sigma}{\rm d}_\sigma)$ and using the Lipschitz-continuity of B we obtain, thanks to the Cauchy-Schwarz and Young's inequalities,

$$\begin{aligned}
&\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \frac{{\rm m}(\sigma)}{{\rm d}_\sigma}(u_K - u_L)^2 + \frac{\gamma}{2}||u||_{L^2(\Omega)}^2 \\
&\qquad \leq \frac{1}{2\gamma}||g||_{L^2(\Omega)}^2 + C_2 \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} {\rm m}(\sigma)|v_{K,\sigma}|\left(|u_K| + |u_L|\right)|u_K - u_L| \\
&\qquad \leq \frac{1}{2\gamma}||g||_{L^2(\Omega)}^2 + C_2 \left(\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \frac{{\rm m}(\sigma)}{{\rm d}_\sigma}(u_K - u_L)^2\right)^{1/2} \\
&\qquad\qquad \times \left(\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} {\rm m}(\sigma){\rm d}_\sigma v_{K,\sigma}^2\left(|u_K| + |u_L|\right)^2\right)^{1/2}.
\end{aligned}$$

We deduce, using Young's inequality,

$$|u|_{1,\mathcal{M}}^2 + \gamma||u||_{L^2(\Omega)}^2 \leq \frac{C_3}{\gamma}||g||_{L^2(\Omega)}^2 + C_3 \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} {\rm m}(\sigma){\rm d}_\sigma v_{K,\sigma}^2\left(|u_K| + |u_L|\right)^2. \qquad (4.18)$$

We now take $M > 0$ and define $v_{K,\sigma}^M = \frac{1}{{\rm m}(\mathcal{D}_\sigma)}\int_{\mathcal{D}_\sigma} \mathbf{1}_{\{|\mathbf{V}|\geq M\}}\mathbf{V}\cdot\mathbf{n}_{K,\sigma}$; we notice that $|v_{K,\sigma} - v_{K,\sigma}^M| \leq M$ and Estimate (4.18) therefore gives

$$\begin{aligned}
|u|_{1,\mathcal{M}}^2 + \gamma||u||_{L^2(\Omega)}^2 \;\leq\; &\frac{C_3}{\gamma}||g||_{L^2(\Omega)}^2 + C_4M^2 \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} {\rm m}(\sigma){\rm d}_\sigma(|u_K|^2 + |u_L|^2) \\
&+C_5 \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} {\rm m}(\sigma){\rm d}_\sigma(v_{K,\sigma}^M)^2(|u_K| + |u_L|)^2.
\end{aligned}$$

But, from (2.5), we have $\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)\mathrm{d}_\sigma \le \frac{1}{\zeta}\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)\mathrm{d}(x_K,\sigma) = \frac{d}{\zeta}\mathrm{m}(K)$ for all $K\in\mathcal{T}$ and therefore, gathering by control volumes,

$$\begin{aligned} |u|_{1,\mathcal{M}}^2 + \gamma||u||_{L^2(\Omega)}^2 &\le \frac{C_3}{\gamma}||g||_{L^2(\Omega)}^2 + C_6M^2||u||_{L^2(\Omega)}^2 \\ &\quad + C_6\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)\mathrm{d}_\sigma(v_{K,\sigma}^M)^2(|u_K|+|u_L|)^2. \end{aligned} \tag{4.19}$$

Let $p>2$ be the exponent given by (1.2). Using Hölder's inequality with exponents $\frac{p}{2}$ and $\frac{p}{p-2}$, we find

$$\begin{aligned} \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)\mathrm{d}_\sigma(v_{K,\sigma}^M)^2(|u_K|+|u_L|)^2 &\le \left(\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)\mathrm{d}_\sigma|v_{K,\sigma}^M|^p\right)^{2/p} \\ &\quad \left(\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)\mathrm{d}_\sigma(|u_K|+|u_L|)^{\frac{2p}{p-2}}\right)^{(p-2)/p}. \end{aligned}$$

We use once more $\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)\mathrm{d}_\sigma \le \frac{d}{\zeta}\mathrm{m}(K)$, apply Jensen's inequality to $|v_{K,\sigma}^M|^p = |\frac{1}{\mathrm{m}(\mathcal{D}_\sigma)}\int_{\mathcal{D}_\sigma} \mathbf{1}_{\{|\mathbf{V}|\ge M\}}\mathbf{V}\cdot\mathbf{n}_{K,\sigma}|^p$ and use the fact that $\mathrm{m}(\mathcal{D}_\sigma) = \frac{\mathrm{m}(\sigma)\mathrm{d}_\sigma}{d}$ to deduce

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)\mathrm{d}_\sigma(v_{K,\sigma}^M)^2(|u_K|+|u_L|)^2 \le C_7||\mathbf{1}_{\{|\mathbf{V}|\ge M\}}\mathbf{V}||_{L^p(\Omega)^d}^2||u||_{L^{\frac{2p}{p-2}}(\Omega)}^2.$$

By assumption on p, we can apply Lemma 6.1 (in the appendix) with $q=\frac{2p}{p-2}$ to find

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)\mathrm{d}_\sigma(v_{K,\sigma}^M)^2(|u_K|+|u_L|)^2 \le C_8||\mathbf{1}_{\{|\mathbf{V}|\ge M\}}\mathbf{V}||_{L^p(\Omega)^d}^2(|u|_{1,\mathcal{M}}^2 + ||u||_{L^2(\Omega)}^2)$$

and, coming back to (4.19),

$$\begin{aligned} |u|_{1,\mathcal{M}}^2 + \gamma||u||_{L^2(\Omega)}^2 &\le \frac{C_3}{\gamma}||g||_{L^2(\Omega)}^2 + C_6M^2||u||_{L^2(\Omega)}^2 \\ &\quad + C_9||\mathbf{1}_{\{|\mathbf{V}|\ge M\}}\mathbf{V}||_{L^p(\Omega)^d}^2(|u|_{1,\mathcal{M}}^2 + ||u||_{L^2(\Omega)}^2). \end{aligned}$$

We notice that, as $M\to\infty$, $||\mathbf{1}_{\{|\mathbf{V}|\ge M\}}\mathbf{V}||_{L^p(\Omega)^d}\to 0$; we can thus fix $M=M(\Omega,\mathbf{V},B,\zeta)$ such that $C_9||\mathbf{1}_{\{|\mathbf{V}|\ge M\}}\mathbf{V}||_{L^p(\Omega)^d}^2 \le 1/2$ and the proof is complete if we take $\gamma_0 = C_6M^2+1$ and $C_1 = 2C_3$. ■

Corollary 4.3 *Assume that (1.2) holds and that $\mathcal{M}$ satisfies (2.5). Then there exists $C_{10} = C_{10}(\Omega,\mathbf{V},B,\zeta)$ such that, for all $\gamma\ge 0$, if u is a solution to the scheme ((2.8)–(2.10),(2.12),(2.13)) for (1.4) then $|u|_{1,\mathcal{M}} \le C_{10}(||g||_{L^2(\Omega)} + ||u||_{L^2(\Omega)})$.*

Proof of Corollary 4.3
Take γ_0 given by Proposition 4.1 and remark that if u is the solution to the scheme ((2.8)–(2.10),(2.12),(2.13)) for some γ, it is also solution to the same scheme with γ replaced by $\gamma+\gamma_0\ge\gamma_0$ and g replaced by $g+\gamma_0 u$. Proposition 4.1 gives thus C_{11} such that $|u|_{1,\mathcal{M}} \le C_{11}||g+\gamma_0 u||_{L^2(\Omega)}$ and the corollary is proved. ■

4.2 Estimates for any γ

Let us first begin with a lemma, which basically states that if a sequence of approximate solutions given by the schemes converges, then its limit is a solution to the initial problem (1.3). This lemma will be useful both in obtaining estimates on the approximate solutions and, of course, in proving their convergence toward the solution of the continuous problem.

Lemma 4.4 *Let $\mathcal{M}_n$ be a sequence of discretizations such that* $\mathrm{size}(\mathcal{M}_n) \to 0$ *as $n \to \infty$ and which satisfy (2.5) with ζ not depending on n. Let $u_n = (u_K^n)_{K\in\mathcal{T}}$ be such that $(|u_n|_{1,\mathcal{M}_n} + ||u_n||_{L^2(\Omega)})_{n\geq 1}$ is bounded and $u_n \to \bar{u}$ in $L^2(\Omega)$ as $n \to \infty$, with $\bar{u} \in H^1(\Omega)$. Then, for all $\varphi \in C^\infty(\overline{\Omega})$,*

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n} \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}\Big(B(-v_{K,\sigma}\mathrm{d}_\sigma)u_K^n - B(v_{K,\sigma}\mathrm{d}_\sigma)u_L^n\Big)(\varphi(x_K)-\varphi(x_L)) \longrightarrow \int_\Omega \nabla\bar{u}\cdot\nabla\varphi - \int_\Omega \bar{u}\mathbf{V}\cdot\nabla\varphi \quad \text{as } n\to\infty \tag{4.20}$$

(in this equation, $\mathcal{E}_{\rm int}^n$ denotes the interior edges of $\mathcal{M}_n$).

Proof of Lemma 4.4
We have, by (2.10),

$$B(-v_{K,\sigma}\mathrm{d}_\sigma)u_K^n - B(v_{K,\sigma}\mathrm{d}_\sigma)u_L^n = B(-v_{K,\sigma}\mathrm{d}_\sigma)(u_K^n-u_L^n) + v_{K,\sigma}\mathrm{d}_\sigma u_L^n$$

and

$$B(-v_{K,\sigma}\mathrm{d}_\sigma)u_K^n - B(v_{K,\sigma}\mathrm{d}_\sigma)u_L^n = v_{K,\sigma}\mathrm{d}_\sigma u_K^n + B(v_{K,\sigma}\mathrm{d}_\sigma)(u_K^n-u_L^n),$$

so that, averaging these two quantities,

$$\begin{aligned} B(-v_{K,\sigma}\mathrm{d}_\sigma)u_K^n - & B(v_{K,\sigma}\mathrm{d}_\sigma)u_L^n \\ &= \frac{B(-v_{K,\sigma}\mathrm{d}_\sigma)+B(v_{K,\sigma}\mathrm{d}_\sigma)}{2}(u_K^n-u_L^n) + v_{K,\sigma}\mathrm{d}_\sigma\frac{u_K^n+u_L^n}{2} \\ &= (u_K^n-u_L^n) + v_{K,\sigma}\mathrm{d}_\sigma\frac{u_K^n+u_L^n}{2} + R_{K,\sigma}^n \end{aligned}$$

where, since $B(0)=1$ and B is Lipschitz-continuous, $|R_{K,\sigma}^n| \leq \mathrm{Lip}(B)\mathrm{d}_\sigma|v_{K,\sigma}|\,|u_K^n-u_L^n|$. Therefore, we can write

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n} \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}\Big(B(-v_{K,\sigma}\mathrm{d}_\sigma)u_K^n - B(v_{K,\sigma}\mathrm{d}_\sigma)u_L^n\Big)(\varphi(x_K)-\varphi(x_L)) = T_1^n + T_2^n + \mathrm{Lip}(B)T_3^n.$$

with

$$\begin{aligned} T_1^n &= \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n} \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}(u_K^n-u_L^n)(\varphi(x_K)-\varphi(x_L)), \\ T_2^n &= \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n} \mathrm{m}(\sigma)v_{K,\sigma}\frac{u_K^n+u_L^n}{2}(\varphi(x_K)-\varphi(x_L)), \\ |T_3^n| &\leq \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n} \mathrm{m}(\sigma)|v_{K,\sigma}|\,|u_K^n-u_L^n|\,|\varphi(x_K)-\varphi(x_L)|. \end{aligned}$$

By regularity of φ, there exists C_{12} only depending on φ such that

$$\begin{aligned}|T_3^n| &\leq C_{12}\mathrm{size}(\mathcal{M}_n)\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n}\mathrm{m}(\sigma)|v_{K,\sigma}|\,|u_K^n-u_L^n|\\ &\leq C_{12}\mathrm{size}(\mathcal{M}_n)\left(\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n}\frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}(u_K^n-u_L^n)^2\right)^{1/2}\left(\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n}\mathrm{m}(\sigma)\mathrm{d}_\sigma v_{K,\sigma}^2\right)^{1/2}\end{aligned}$$

and the bound on $|u_n|_{1,\mathcal{M}_n}$ and the fact that $\mathbf{V}\in L^2(\Omega)^d$ imply that T_3^n tends to 0 as $n\to\infty$.
The convergence of T_1^n and T_2^n respectively toward $\int_\Omega\nabla\bar{u}\cdot\nabla\varphi$ and $-\int_\Omega\bar{u}\mathbf{V}\cdot\nabla\varphi$ is quite classical. It can be proved, for example, using the approximate gradient introduced equivalently in [2, Lemma 4.4], [5, Lemma 6.5] or [9, Definition 2] : the approximate gradient $\nabla_{\mathcal{M}_n}v$ of a function $v=(v_K)_{K\in\mathcal{T}_n}$ is defined as a piecewise constant function, equal to $\frac{\mathrm{m}(\sigma)}{\mathrm{m}(\mathcal{D}_\sigma)}(v_L-v_K)\mathbf{n}_{K,\sigma}$ on each diamond $\mathcal{D}_\sigma$ around the interior edges $\sigma=K|L$ (and, for example, to 0 on the diamonds corresponding to boundary edges). Thanks to the bound on $|u_n|_{1,\mathcal{M}_n}$ and the regularity of φ, one can see (as in [2] and [5] ([3])) that $\nabla_{\mathcal{M}_n}u_n\to\nabla\bar{u}$ weakly in $L^2(\Omega)^d$ and $\nabla_{\mathcal{M}_n}\varphi\to\nabla\varphi$ weakly-$*$ in $L^\infty(\Omega)^d$ as $n\to\infty$ ($\nabla_{\mathcal{M}_n}\varphi$ being constructed from the values $(\varphi(x_K))_{K\in\mathcal{T}_n}$). Up to errors which tend to 0 as $n\to\infty$, we have $T_1^n\approx\int_\Omega\nabla_{\mathcal{M}_n}u_n\cdot\nabla\varphi$ and $T_2^n\approx-\int_\Omega u_n\mathbf{V}\cdot\nabla_{\mathcal{M}_n}\varphi$ and the passage to the limit $n\to\infty$ is then straightforward, since the convergence of u_n in $L^2(\Omega)$ and the fact that $\mathbf{V}$ belongs to $L^2(\Omega)^d$ ensure that $u_n\mathbf{V}\to\bar{u}\mathbf{V}$ in $L^1(\Omega)^d$. ■

We can now prove, by way of contradiction, *a priori* estimates on the solutions to the schemes.

Proposition 4.5 *Let $(\mathcal{M}_n)_{n\geq1}$ be a sequence of discretizations such that* $\mathrm{size}(\mathcal{M}_n)\to0$ *as $n\to\infty$, and such that $\mathcal{M}_n$ satisfies (2.5) with ζ not depending on n. Let $\gamma\geq0$ and u_n be the solution of ((2.8)–(2.10),(2.12),(2.13)) for $\mathcal{M}=\mathcal{M}_n$. If $\gamma=0$, we moreover assume that $\int_\Omega u_n=0$. Then $(|u_n|_{1,\mathcal{M}_n}+||u_n||_{L^2(\Omega)})_{n\geq1}$ is bounded.*

Proof of Proposition 4.5
By Corollary 4.3, we only need to prove that $(u_n)_{n\geq1}$ remains bounded in $L^2(\Omega)$. If this is not the case then, up to a subsequence, we can assume that $||u_n||_{L^2(\Omega)}\to\infty$ as $n\to\infty$. We notice that $w_n=\frac{u_n}{||u_n||_{L^2(\Omega)}}$ is the solution to ((2.8)–(2.10),(2.12),(2.13)) with g replaced by $\frac{g}{||u_n||_{L^2(\Omega)}}$ and, since $||w_n||_{L^2(\Omega)}=1$, Corollary 4.3 then shows that $(|w_n|_{1,\mathcal{M}_n}+||w_n||_{L^2(\Omega)})_{n\geq1}$ is bounded. As in Step 1 in the proof of [10, Theorem 10.3], we infer that, upon extracting a subsequence, $w_n\to\bar{w}$ in $L^2(\Omega)$, with $\bar{w}\in H^1(\Omega)$ and $||\bar{w}||_{L^2(\Omega)}=1$. If $\gamma=0$, each w_n having a zero mean value, this is also the case for $\bar{w}$.

[3] The different boundary conditions we consider here, with respect to these references, modify nothing to the proof.

Let $\varphi \in C^\infty(\overline{\Omega})$, multiply (2.13) for w_n by $\varphi(x_K)$, sum over $K \in \mathcal{T}$ and gather by edges using the conservativity of the fluxes. This gives

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n} \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}(B(-v_{K,\sigma}\mathrm{d}_\sigma)w_K^n - B(v_{K,\sigma}\mathrm{d}_\sigma)w_L^n)(\varphi(x_K)-\varphi(x_L))$$
$$+\gamma\sum_{K\in\mathcal{T}}\mathrm{m}(K)w_K^n\varphi(x_K) = \frac{1}{||u_n||_{L^2(\Omega)}}\sum_{K\in\mathcal{T}}\mathrm{m}(K)g_K\varphi(x_K).$$

Passing to the limit $n\to\infty$ thanks to Lemma 4.4, to the convergence in $L^2(\Omega)$ of $(w_n)_{n\geq 1}$ and to the fact that $||u_n||_{L^2(\Omega)}\to\infty$, we obtain

$$\int_\Omega \nabla\bar{w}\cdot\nabla\varphi - \int_\Omega \bar{w}\mathbf{V}\cdot\nabla\varphi + \gamma\int_\Omega\bar{w}\varphi = 0,$$

that is to say $\bar{w}$ is a weak solution to (1.4) with $g=0$. But, if $\gamma>0$, Problem (1.4) has a trivial kernel (see [7, Proposition 5.1]) and, if $\gamma=0$, (1.4) is (1.1), whose only element with zero mean value in the kernel is the zero function (see [7, Proposition 2.2]). Hence, in either case we see that $\bar{w}=0$, which is a contradiction with $||\bar{w}||_{L^2(\Omega)}=1$. This proves that $(u_n)_{n\geq1}$ is bounded in $L^2(\Omega)$ and concludes the proof. ■

The proof of the convergence of the schemes is now easy.

Proof of Theorems 2.7 and 2.8
Let us first consider the convergence of the kernel in Item 1 of Theorem 2.7. By choice, $(||\widehat{u}_n||_{L^2(\Omega)})_{n\geq1}$ is bounded and Corollary 4.3 thus shows that $(|\widehat{u}_n|_{1,\mathcal{M}_n})_{n\geq1}$ is also bounded ($\widehat{u}_n$ satisfies ((2.8)–(2.10),(2.12),(2.13)) with $\gamma=0$, $g=0$ and $\mathcal{M}=\mathcal{M}_n$ which satisfies (2.5) with ζ not depending on n). Hence, as in the proof of Proposition 4.5, Step 1 in [10, proof of Theorem 10.3] shows that, up to a subsequence, $\widehat{u}_n\to\widehat{\widehat{u}}$ in $L^2(\Omega)$, where $\widehat{\widehat{u}}\in H^1(\Omega)$. Then $\widehat{\widehat{u}}$ is non-negative (since $\widehat{u}_n$ is positive for all $n\geq1$) and has an L^2 norm equal to one (since this is the case for $\widehat{u}_n$ for all $n\geq1$). We now write (2.8)–(2.12) on $\mathcal{M}=\mathcal{M}_n$ with $g=0$ for $\widehat{u}_n$, multiply the flux balance (2.11) by $\varphi(x_K)$ for some $\varphi\in C^\infty(\overline{\Omega})$ and sum over the control volumes $K\in\mathcal{T}$. Gathering by edges, we obtain that the left-hand side of (4.20) with $u^n=\widehat{u}_n$ is equal to zero; we can pass to the limit thanks to Lemma 4.4 to see that, for all regular φ,

$$\int_\Omega\nabla\widehat{\widehat{u}}\cdot\nabla\varphi - \int_\Omega\widehat{\widehat{u}}\mathbf{V}\cdot\nabla\varphi = 0.$$

This means that $\widehat{\widehat{u}}$ belongs to the kernel of the operator in (1.1); since there is only one element in this kernel which is non-negative and has an L^2 norm equal to 1 (see [7, Proposition 2.2]; this element is in fact positive) and since the preceding reasoning can be done along any subsequence of $(\widehat{u}_n)_{n\geq1}$, this shows that the whole sequence $(\widehat{u}_n)_{n\geq1}$ converges to $\widehat{\widehat{u}}$ and concludes the proof of Item 1 in Theorem 2.7.
To prove Item 2 of the same theorem, we notice that Proposition 4.5 gives a bound on $(|u_n|_{1,\mathcal{M}_n}+||u_n||_{L^2(\Omega)})_{n\geq1}$ and thus, as before, we can assume that, up to a subsequence,

$u_n \to \bar{u}$ in $L^2(\Omega)$ with $\bar{u} \in H^1(\Omega)$ with zero mean value (since each u_n has a zero mean-value). Also following the preceding reasoning, we multiply the scheme satisfied by u_n by the values at the center of the control volumes of a regular function φ, sum over the control volumes and gather by edges, and we obtain

$$\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}^n} \frac{\mathrm{m}(\sigma)}{\mathrm{d}_\sigma}\Big(B(-v_{K,\sigma}\mathrm{d}_\sigma)u_K^n - B(v_{K,\sigma}\mathrm{d}_\sigma)u_L^n\Big)(\varphi(x_K)-\varphi(x_L)) = \sum_{K\in\mathcal{T}_n} \mathrm{m}(K) g_K \varphi(x_K).$$

Passing to the limit $n \to \infty$ in this equation (thanks to Lemma 4.4 for the left-hand side and to the regularity of φ for the right-hand side), we see that $\bar{u}$ is a weak solution to (1.1); since there is only one such weak solution with zero mean value, this proves that the whole sequence $(u_n)_{n\geq 1}$ converges to $\bar{u}$ and concludes the proof of Theorem 2.7.
The proof of Theorem 2.8 is completely similar and is therefore omitted. ■

Remark 4.6 *From the bound on the L^2 norm and on the discrete H^1 semi-norm of $(\widehat{u}_n)_{n\geq 1}$ and $(u_n)_{n\geq 1}$ and the discrete Sobolev inequalities, we see that the convergence of these functions holds not only in $L^2(\Omega)$ but also in $L^q(\Omega)$ for all $q < +\infty$ if $d = 2$ and all $q < 6$ if $d = 3$, and also weakly in $L^6(\Omega)$ if $d = 3$.*

5 Numerical results

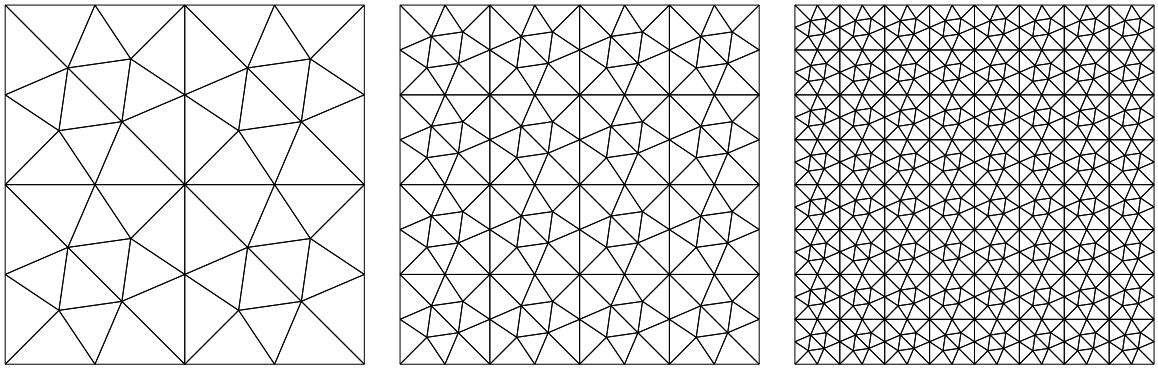
In this section, we present numerical results obtained with the scheme (2.8)–(2.12) for (1.1). The different choices of the function B correspond to the upwind, the centered and the Scharfetter-Gummel fluxes (we recall that in the case of the centered fluxes, B satisfies (2.9) only if $s < 2$). We are interested in the computation of a function spanning the kernel of (2.8)–(2.12) and also of solutions to this scheme for $g \neq 0$.
In all the test cases, the domain Ω is the square $[0,1]\times[0,1]$ and we use a sequence of triangular meshes numbered from 1 to 7. The initial mesh is made of 56 triangles and the other meshes come from successive refinements of Ω in squares, each one being partitioned in triangles using the initial mesh (Figure 2 shows a drawing of the first three grids; the size of mesh i is half the size of mesh $i-1$, and the size of mesh 1 is 0.25). All these meshes are admissible in the sense of Definition 2.1.
For the initial mesh, we have $\max_{\sigma\in\mathcal{E}} \mathrm{d}_\sigma \leq 0.15$. It follows that for the mesh i, we have $\max_{\sigma\in\mathcal{E}} \mathrm{d}_\sigma \leq \frac{0.15}{2^{i-1}}$. Therefore, when $\mathbf{V}$ is bounded, we set $\mathcal{V} = \sup_{x\in\Omega} \|\mathbf{V}(x)\|$ and the Péclet condition (3.16) is satisfied if

$$i \geq \frac{\ln(0.15\mathcal{V})}{\ln(2)}. \tag{5.21}$$

5.1 Computations of the kernel

In this section, we consider three test cases for which we compute a numerical approximation of the kernel by the different schemes. In the three cases, the convection field

Figure 2: Meshes 1, 2 and 3 used in the numerical tests.



$\mathbf{V}$ is continuous and we may choose $v_{K,\sigma} = \mathbf{V}(x_\sigma) \cdot \mathbf{n}_{K,\sigma}$, where x_σ is the center of the edge σ (choice 1). But when $\mathbf{V}$ comes from a potential Φ (Test cases 1 and 2), we may also choose for the Scharfetter-Gummel scheme $v_{K,\sigma} = \frac{\Phi(x_L)-\Phi(x_K)}{\mathrm{d}_\sigma}$ (choice 2).
The element of the kernel of (2.8)–(2.12) we consider is the one given by Item 1 in Theorem 2.5, normalized so that $||\widehat{u}||_{L^2} = 1$ (there is only one such element in the kernel). To compute it, we notice that, as it is usual in Neumann boundary problems, the sum of the lines of $\mathbb{A}$ vanishes (see Item 2 in Proposition 3.1); hence, solving all the equations but one of the system $\mathbb{A}U = 0$ is equivalent to solving the whole system; moreover, as the kernel of $\mathbb{A}$ has dimension 1 and is spanned by a positive vector, there exists a unique vector in this kernel such that $\sum_{K\in\mathcal{T}} \mathrm{m}(K)u_K = 1$. These constatations ensure that the system obtained by replacing any one of the lines of $\mathbb{A}U = 0$ by the line $\sum_{K\in\mathcal{T}} \mathrm{m}(K)u_K = 1$ is invertible and provides a vector in the kernel of $\mathbb{A}$. Then, we multiply the obtained vector by a positive number in order to normalize it in L^2.
For the upwind or Scharfetter-Gummel scheme, as we have proved in Theorem 2.5 and as we will see in Test 3, this gives the unique positive normalized element $\widehat{u}$ in the kernel of (2.8)–(2.12). The centered scheme does not necessarily have a positive solution in its kernel, but we use the same technique to compute a "would-be positive" solution.

Test case 1.

$$\mathbf{V}(x,y) = \begin{pmatrix} 10 \\ 0 \end{pmatrix} = \nabla\Phi \quad \text{with} \quad \Phi(x,y) = 10x.$$

The kernel is spanned by the function $\widehat{\widehat{u}}(x,y) = \exp(10x)$. As $\mathbf{V}$ is constant, the two different choices for $v_{K,\sigma}$ coincide. In this case, the condition (5.21) is satisfied by every mesh.
In Table 1, we present the error in L^2-norm between $\widehat{\widehat{u}}$ and $\widehat{u}$ (both functions being normalized to 1) for $B = B_{\mathrm{ce}}$, B_{up} or B_{sg}. It shows that the Scharfetter-Gummel scheme is exact in this case (as proved in Remark 3.3) and, recalling that the size of each grid is half the size of the preceding grid, that the centered scheme converges with an order

around 2 and the upwind scheme converges with an order around 1 (this is expected and well-known in other applications of finite volume methods).

Table 1: Error between the function $\widehat{\overline{u}}$ spanning the continuous kernel and its numerical approximation $\widehat{u}$ (both chosen positive and normalized to 1), for the **test case 1**.

Mesh	Number of triangles	$\|\widehat{u}-\widehat{\overline{u}}\|_{L^2(\Omega)}$ centered scheme	$\|\widehat{u}-\widehat{\overline{u}}\|_{L^2(\Omega)}$ upwind scheme	$\|\widehat{u}-\widehat{\overline{u}}\|_{L^2(\Omega)}$ SG scheme
1	56	4.48e-02	1.66e-01	5.73e-16
2	224	1.26e-02	1.05e-01	8.28e-16
3	896	3.14e-03	5.88e-02	8.48e-15
4	3584	7.51e-04	3.04e-02	6.84e-15
5	14336	1.84e-04	1.55e-02	2.35e-14
6	57344	4.85e-05	7.83e-03	6.26e-14
7	229376	1.14e-05	3.94e-03	6.78e-14

Table 2: Error between the function spanning the continuous kernel $\widehat{\overline{u}}$ and its numerical approximation $\widehat{u}$ (both chosen positive and normalized to 1), for the **test case 2**.

Mesh	$\|\widehat{u}-\widehat{\overline{u}}\|_{L^2(\Omega)}$ centered scheme	$\|\widehat{u}-\widehat{\overline{u}}\|_{L^2(\Omega)}$ upwind scheme	$\|\widehat{u}-\widehat{\overline{u}}\|_{L^2(\Omega)}$ SG scheme 1	$\|\widehat{u}-\widehat{\overline{u}}\|_{L^2(\Omega)}$ SG scheme 2
1	8.15e-03	3.29e-02	1.00e-02	1.29e-15
2	2.07e-03	1.50e-02	2.74e-03	5.60e-15
3	5.46e-04	7.22e-03	7.50e-04	1.30e-13
4	1.46e-04	3.56e-03	2.04e-04	2.72e-13
5	3.91e-05	1.77e-03	5.52e-05	5.80e-13
6	1.04e-05	8.81e-04	1.48e-05	5.73e-13
7	2.77e-06	4.40e-04	3.93e-06	1.39e-12

Test case 2.

$$\mathbf{V}(x,y) = \begin{pmatrix} \dfrac{1-2y}{x+y-2xy} \\ \dfrac{1-2x}{x+y-2xy} \end{pmatrix} = \nabla\Phi \quad \text{with} \quad \Phi(x,y) = \log(x+y-2xy).$$

The kernel is spanned by the function $\widehat{\overline{u}}(x,y) = x+y-2xy$. In this case, $\mathbf{V}$ really depends on x and y and we can test the two choices of $v_{K,\sigma}$ for the Scharfetter-Gummel scheme.

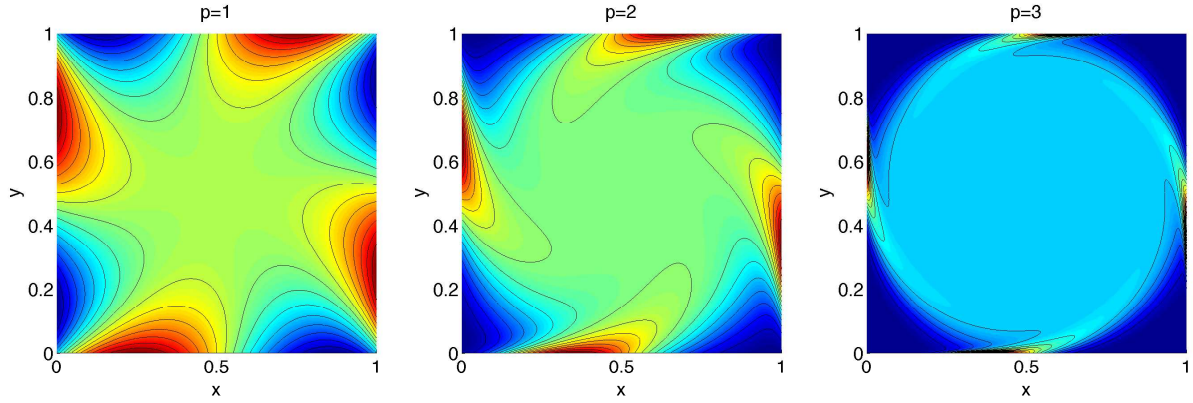
Table 2 confirms that the centered scheme converges with an order around 2, while the upwind scheme converges with an order 1. The Scharfetter-Gummel scheme is exact with the choice 2 of $v_{K,\sigma}$ while it converges with an order slightly less than 2 with the choice 1. In this case, the convective field $\mathbf{V}$ is not bounded on the domain and we cannot use the weak Péclet condition (5.21). Nevertheless, we can compute for each mesh the maximal value of $|v_{K,\sigma} \mathrm{d}_\sigma|$ on the interior edges. We obtain from mesh 1 to mesh 7 the following values : 0.6156, 0.6768, 0.7048, 0.7182, 0.7247, 0.7280, 0.7296. It shows that the Péclet condition (3.16) is satified on all the meshes.

Test case 3.

$$\mathbf{V}(x,y) = 10^p \begin{pmatrix} -(y-0.5) \\ \\ (x-0.5) \end{pmatrix}$$

We propose here a convection field which does not derive from a potential. Let us note that $\mathrm{div}(\mathbf{V}) = 0$ but that $\mathbf{V} \cdot \mathbf{n}$ does not remain nonpositive on the boundary of Ω (so that the operator associated with (1.3) is not necessarily coercive). The variation of the parameter p permits to study the effects of an increased convection strength. Figure 3 shows the positive normed generator of the kernel of (2.8)–(2.12) for $B = B_{\mathrm{sg}}$ and $p = 1$, 2 and 3.

Figure 3: A solution in the kernel for the **test case 3** ($p = 1, 2, 3$) computed with the Scharfetter-Gummel scheme on Mesh 7.



This test case also shows (see Table 3) that the positivity of an element spanning the kernel of the scheme, as stated in Theorem 2.5, is satisfied by the practical implementations of the upwind and Scharfetter-Gummel schemes, but not by the centered scheme unless a condition between the size of the mesh and the convection holds. In this case, we have $\mathcal{V} = \frac{\sqrt{2}}{2} 10^p$ and the weak Péclet condition (5.21) is satisfied by every mesh for $p = 1$, from Mesh 4 for $p = 2$ and only by Mesh 7 for $p = 3$. This result is illustrated in Table 3.

Table 3: Minimum and maximum values of $\widehat{u}$ obtained by the different schemes for the **test case 3** with $p = 3$.

Mesh	Centered scheme		Upwind scheme		SG scheme	
	min	max	min	max	min	max
1	-1.56e-02	1.55e+00	2.15e-01	1.35e+00	2.03e-01	1.34e+00
2	-7.86e-02	1.38e+00	4.41e-02	1.52e+00	3.47e-02	1.48e+00
3	-2.20e-01	2.08e+00	2.62e-03	1.87e+00	1.15e-03	1.80e+00
4	-7.70e-02	2.69e+00	4.67e-05	2.37e+00	5.09e-06	2.37e+00
5	-2.77e-03	3.24e+00	7.94e-07	2.85e+00	6.50e-09	2.96e+00
6	-1.07e-09	3.52e+00	1.82e-08	3.19e+00	1.24e-10	3.39e+00
7	1.00e-11	3.61e+00	9.44e-10	3.42e+00	2.34e-11	3.58e+00

5.2 Computations of solutions with non-vanishing right-hand sides

In this section, we compute approximate solutions for (1.1) using the scheme (2.8)–(2.12) with $B = B_{\rm ce}$, $B_{\rm up}$ and $B_{\rm sg}$ (with the two choices for $v_{K,\sigma}$ when $\mathbf{V}$ comes from a potential).

Test case 4.

$$\mathbf{V}(x,y) = \begin{pmatrix} 4(x-0.5)^2 \\ 0 \end{pmatrix} = \nabla\Phi \quad \text{with} \quad \Phi(x,y) = \frac{4}{3}(x-0.5)^3,$$
$$g(x,y) = \exp(x)(4(x-0.5)(x+1.5)-1).$$

The exact solution of (1.1) is $\bar{u}(x,y) = \exp(x)$; the mean value of this function whose mean value is not equal to 0, so we compute the solution u to the scheme (2.8)–(2.12) (with $g_K = g(x_K)$) which satisfies

$$\sum_{K\in\mathcal{T}} \mathrm{m}(K) u_K = \sum_{K\in\mathcal{T}} \mathrm{m}(K)\bar{u}_K \quad \text{(instead of 0)}.$$

Table 4 shows the errors for the schemes; the order of convergence is near 1 for the upwind scheme and near 2 for all other schemes.

Test case 5.

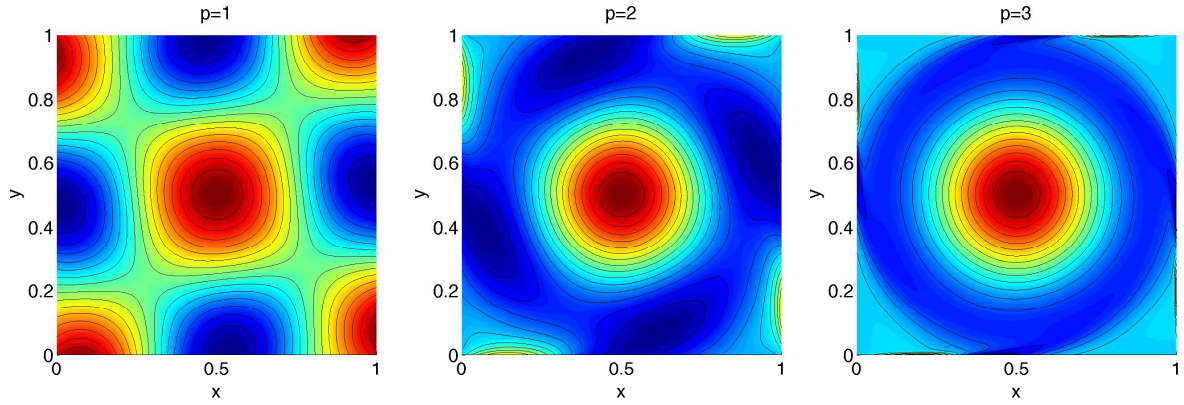
$$\mathbf{V}(x,y) = 10^p \begin{pmatrix} -(y-0.5) \\ \\ (x-0.5) \end{pmatrix}$$
$$g(x,y) = \cos(2\pi x)\cos(2\pi y)$$

The convection field is the same as for the test case 3. Figure 4 shows, for $p = 1$, 2 and 3, the solution to (2.8)–(2.12) (for $B = B_{\rm sg}$) which has a vanishing mean value.

Table 4: Error in L^2-norm between the exact solution $\bar{u}$ and its numerical approximation u, for the **test case 4**.

Mesh	$\|u-\bar{u}\|_{L^2(\Omega)}$ centered scheme	$\|u-\bar{u}\|_{L^2(\Omega)}$ upwind scheme	$\|u-\bar{u}\|_{L^2(\Omega)}$ SG scheme 1	$\|u-\bar{u}\|_{L^2(\Omega)}$ SG scheme 2
1	5.67e-03	4.04e-03	5.63e-03	8.81e-03
2	1.34e-03	2.56e-03	1.33e-03	1.90e-03
3	3.46e-04	1.37e-03	3.45e-04	4.63e-04
4	8.85e-05	7.05e-04	8.83e-05	1.16e-04
5	2.23e-05	3.59e-04	2.23e-05	2.91e-05
6	5.61e-06	1.81e-04	5.59e-06	7.29e-06
7	1.40e-06	9.10e-05	1.40e-06	1.82e-06

Figure 4: A solution of the **test case 5** ($p = 1, 2, 3$) computed with the Scharfetter-Gummel scheme on the mesh 7.



Test case 6 We finaly consider a case where $\mathbf{V}$ does not come from a potential and where $\mathrm{div}(\mathbf{V})$ is (strongly) negative in Ω:

$$\mathbf{V}(x,y) = -100 \begin{pmatrix} x+y \\ \\ y-x \end{pmatrix}$$

The right hand side g and the boundary conditions h (which are here non homogeneous, see Remark 2.3) have been chosen such that the exact solution is $\bar{u}(x,y) = 30x(1-x)y(1-y)$ ($\|\bar{u}\|_{L^2(\Omega)} = 1$) and Table 5 shows the approximation errors for the different schemes. Once again, despite the strong non-coercivity of the operator in (1.3) for this test case, the orders of convergence are 1 for the upwind scheme and 2 for the other schemes. Let us also note that, in this case $\mathcal{V} = 200$ and that the condition (5.21) is satisfied from Mesh 5.

Table 5: Error in L^2-norm between the exact solution $\bar{u}$ and its numerical approximation u, for the **test case 6**.

Mesh	$\|u-\bar{u}\|_{L^2(\Omega)}$ centered scheme	$\|u-\bar{u}\|_{L^2(\Omega)}$ upwind scheme	$\|u-\bar{u}\|_{L^2(\Omega)}$ SG scheme 1
1	3.01e-01	1.09e+00	2.15e+00
2	8.30e-02	3.93e-01	3.43e-01
3	2.17e-02	1.81e-01	9.64e-02
4	5.74e-03	9.14e-02	2.76e-02
5	1.49e-03	4.72e-02	7.40e-03
6	3.78e-04	2.42e-02	1.89e-03
7	9.50e-05	1.23e-02	4.76e-04

6 Appendix

Discrete Sobolev inequalities are classical tools for the study of finite volume discretizations of elliptic equations; they are especially useful when one has to handle a more difficult case that the simple Laplace equation with a L^2 right-hand side. In the framework of Dirichlet boundary conditions, discrete Sobolev inequalities are proved in [3]; the following lemma gives corresponding inequalities for functions which do not vanish on $\partial\Omega$.

Lemma 6.1 (Discrete Sobolev inequalities for non-Dirichlet boundary conditions) *Let Ω be a bounded polygonal open subset of $\mathbb{R}^d$ and let $\mathcal{M}$ be an admissible mesh (in the sense of Definition 2.1) which satisfies (2.5). Let $q < +\infty$ if $d = 2$ and $q = \frac{2d}{d-2}$ if $d \geq 3$. Then there exists $C = C(\Omega, \zeta, q)$ such that, for all $u = (u_K)_{K\in\mathcal{T}}$,*

$$||u||_{L^q(\Omega)} \leq C(|u|_{1,\mathcal{M}} + ||u||_{L^2(\Omega)}),$$

where $|\cdot|_{1,\mathcal{M}}$ is the discrete H^1 semi-norm defined by (4.17).

Proof of Lemma 6.1
Unless otherwise mentioned, all the constants C_i in this proof only depend on Ω, ζ and q. We define the discrete $W^{1,1}$ semi-norm of a function $v = (v_K)_{K\in\mathcal{T}}$ by

$$N_{\mathcal{M}}(v) = \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \mathrm{m}(\sigma)|v_K - v_L|.$$

Step 1: we first prove that, for any v,

$$||v||_{L^{\frac{d}{d-1}}(\Omega)} \leq C_{13}\left(N_{\mathcal{M}}(v) + ||v||_{L^1(\Omega)}\right). \tag{6.22}$$

The idea is first to obtain such an estimate with $||v||_{L^1(\Omega)}$ replaced by the L^1 norm of a trace of v on $\partial\Omega$ and then, using a trace result similar to [10, Lemma 10.5, p 807], to replace this boundary norm with an interior norm.

Let $(\mathbf{e}_i)_{i=1,\dots,d}$ be the cartesian basis of $\mathbb{R}^d$ and, for $\sigma \in \mathcal{E}$, $\chi^i_\sigma(x) = 1$ if $\sigma \cap (x + \mathbb{R}\mathbf{e}_i) \neq \emptyset$ and $\chi^i_\sigma(x) = 0$ otherwise. Summing all the jumps of u encountered from x to the exterior of Ω following the direction of $\mathbf{e}_i$, we have, for all $i = 1, \dots, d$,

$$|v(x)| \leq \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \chi^i_\sigma(x)|v_K - v_L| + \sum_{\sigma\in\mathcal{E}_{\rm ext}} \chi^i_\sigma(x)|v_{K(\sigma)}| := W^i(x)$$

where, for $\sigma \in \mathcal{E}_{\rm ext}$, $K(\sigma)$ is the unique control volume such that $\sigma \in \mathcal{E}_K$. We infer that $|v(x)|^{\frac{d}{d-1}} \leq \prod_{i=1}^d (W^i(x))^{\frac{1}{d-1}}$ and, noticing that $\chi^i_\sigma(x)$ does not depend on the i-th coordinate of x, we can apply the Gagliardo-Nirenberg inequality to see that

$$\int_\Omega |v|^{\frac{d}{d-1}} \leq \prod_{i=1}^d \left(\int_{\Omega^i} W^i\right)^{\frac{1}{d-1}} \tag{6.23}$$

where Ω^i is the projection of Ω on the hyperplane $\{x_i = 0\}$. Since χ^i_σ does not vanish only on a cylinder of base σ and direction $\mathbf{e}_i$, we have $\int_{\Omega^i} \chi^i_\sigma \leq {\rm m}(\sigma)$ and therefore $\int_{\Omega^i} W^i \leq N_\mathcal{M}(v) + ||\bar\gamma(v)||_{L^1(\partial\Omega)}$, where $\bar\gamma(v)$ is the trace of v defined by $\bar\gamma(v) = v_{K(\sigma)}$ on $\sigma \in \mathcal{E}_{\rm ext}$. Coming back to (6.23), this gives $||v||_{L^{\frac{d}{d-1}}(\Omega)} \leq N_\mathcal{M}(v) + ||\bar\gamma(v)||_{L^1(\partial\Omega)}$. Following the proof of [10, Lemma 10.5, p 807], it is quite easy to see that $||\bar\gamma(v)||_{L^1(\partial\Omega)} \leq C_{14}(N_\mathcal{M}(v) + ||v||_{L^1(\Omega)})$ (this is in fact more straightforward than the L^2 trace inequality in this reference), which concludes the proof of (6.22).

Step 2: we conclude from (6.22) by an induction process.
Let $u = (u_K)_{K\in\mathcal{T}}$, $s \geq \frac{3}{2}$ and define $v = |u|^s$. Since $||u_K|^s - |u_L|^s| \leq s(|u_K|^{s-1} + |u_L|^{s-1})|u_K - u_L|$, the Cauchy-Schwarz inequality and a gathering by control volumes show that

$$\begin{aligned}
N_\mathcal{M}(v) &\leq \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} {\rm m}(\sigma)s(|u_K|^{s-1} + |u_L|^{s-1})|u_K - u_L| \\
&\leq s\left(\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} {\rm m}(\sigma){\rm d}_\sigma(|u_K|^{s-1} + |u_L|^{s-1})^2\right)^{1/2} |u|_{1,\mathcal{M}} \\
&\leq s\left(\sum_{K\in\mathcal{T}} 2|u_K|^{2(s-1)} \sum_{\sigma\in\mathcal{E}_{K,\rm int}} {\rm m}(\sigma){\rm d}_\sigma\right)^{1/2} |u|_{1,\mathcal{M}}.
\end{aligned}$$

Using (2.5) and the fact that $\sum_{\sigma\in\mathcal{E}_{K,\rm int}} {\rm m}(\sigma){\rm d}(x_K, \sigma) \leq d{\rm m}(K)$, we infer

$$N_\mathcal{M}(v) \leq C_{15} s||u||^{s-1}_{L^{2(s-1)}(\Omega)}|u|_{1,\mathcal{M}}.$$

Hence, for any $s \geq \frac{3}{2}$ (which implies $2(s-1) \geq 1$), using $v = |u|^s$ in (6.22) and taking the power $1/s$ of the resulting inequality, we obtain

$$||u||_{L^{s\frac{d}{d-1}}(\Omega)} \leq C_{16}\left(||u||^{\frac{s-1}{s}}_{L^{2(s-1)}(\Omega)}|u|^{\frac{1}{s}}_{1,\mathcal{M}} + ||u||_{L^s(\Omega)}\right).$$

Assume now that $s \leq \frac{2(d-1)}{d-2}$; then $2(s-1) \leq s\frac{d}{d-1}$ and the $L^{2(s-1)}$ norm of u can be bounded by its $L^{s\frac{d}{d-1}}$ norm (with a multiplicative constant only depending on Ω and s). We infer from Young's inequality applied with exponents $\frac{s}{s-1}$ and s that, for all $\frac{3}{2} \leq s \leq \frac{2(d-1)}{d-2}$ (s finite), there exists $C_{17}(s) = C_{17}(s, \Omega, \zeta)$ such that

$$||u||_{L^{s\frac{d}{d-1}}(\Omega)} \leq C_{17}(s) \left(|u|_{1,\mathcal{M}} + ||u||_{L^s(\Omega)} \right). \tag{6.24}$$

The conclusion now follows from successive applications of this inequality. With $s = 2\frac{d}{d-1} \leq \frac{2(d-1)}{d-2}$, it gives

$$||u||_{L^{2(\frac{d}{d-1})^2}(\Omega)} \leq C_{18} \left(|u|_{1,\mathcal{M}} + ||u||_{L^{2\frac{d}{d-1}}(\Omega)} \right).$$

Using (6.24) with $s = 2$ we can bound the last term in this inequality and we obtain

$$||u||_{L^{2(\frac{d}{d-1})^2}(\Omega)} \leq C_{19} \left(|u|_{1,\mathcal{M}} + ||u||_{L^2(\Omega)} \right). \tag{6.25}$$

We can then apply (6.24) with $s = 2(\frac{d}{d-1})^2$ provided that $2(\frac{d}{d-1})^2 \leq \frac{2(d-1)}{d-2}$ (which is always true for $d \geq 2$); using (6.25) to bound the $L^{2(\frac{d}{d-1})^2}$ norm of u appearing in the left-hand side, this leads to

$$||u||_{L^{2(\frac{d}{d-1})^3}(\Omega)} \leq C_{20} \left(|u|_{1,\mathcal{M}} + ||u||_{L^2(\Omega)} \right).$$

A simple induction then shows that, as long as $2(\frac{d}{d-1})^r \leq \frac{2(d-1)}{d-2}$, we have

$$||u||_{L^{2(\frac{d}{d-1})^{r+1}}(\Omega)} \leq C_{21}(r) \left(|u|_{1,\mathcal{M}} + ||u||_{L^2(\Omega)} \right). \tag{6.26}$$

If $d = 2$, then any r satisfies $2(\frac{d}{d-1})^r \leq \frac{2(d-1)}{d-2} = +\infty$ and, since $(\frac{d}{d-1})^r \to \infty$ as $r \to \infty$, this gives the desired result (for any $q < \infty$, we can find r such that $2(\frac{d}{d-1})^{r+1} \geq q$ and the estimate (6.26) on the $L^{2(\frac{d}{d-1})^{r+1}}$ norm gives an estimate on the L^q norm). If $d \geq 3$, we take r_0 the greatest integer such that $2(\frac{d}{d-1})^{r_0} \leq \frac{2(d-1)}{d-2}$ (such a r_0 is finite), so that (6.26) holds with $r = r_0$; applying (6.24) with $s = \frac{2(d-1)}{d-2}$, we can write

$$||u||_{L^{\frac{2d}{d-2}}(\Omega)} \leq C_{22} \left(|u|_{1,\mathcal{M}} + ||u||_{L^{\frac{2(d-1)}{d-2}}(\Omega)} \right) \tag{6.27}$$

and, since $\frac{2(d-1)}{d-2} \leq 2(\frac{d}{d-1})^{r_0+1}$, (6.26) with $r = r_0$ allows to bound the right-hand side of (6.27) and to conclude the proof. ■

References

[1] Banjai L. and Sauter S., *A refined galerkin error and stability analysis for highly indefinite variational problems*, SIAM J. Numer. Anal. **45** (2007), no. 1, 37–53.

[2] Chainais-Hillairet C., Liu J.-G., Peng Y.-J., *Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis*, M2AN Math. Model. Numer. Anal. **37** (2003), no. 2, p.319-338.

[3] Coudière Y., Gallouët T., Herbin R., *Discrete Sobolev inequalities and L^p error estimates for finite volume solutions of convection diffusion equations*, M2AN Math. Model. Numer. Anal. **35** (2001), no. 4, 767–778.

[4] Droniou J., *Non-coercive Linear Elliptic Problems*, Potential Anal. **17** (2002), no. 2, 181-203.

[5] Droniou J., Eymard R., *Study of the mixed finite volume method for Stokes and Navier-Stokes equations*, Numerical Methods for Partial Differential Equations, **25** (2008), no. 1, 137-171.

[6] Droniou J., Gallouët T., *Finite volume methods for convection-diffusion equations with right-hand side in H^{-1}*, M2AN Math. Model. Numer. Anal. **36** (2002), no. 4, 705-724.

[7] Droniou J., Vázquez J. L., *Noncoercive convection-diffusion elliptic problems with Neumann boundary conditions*, to appear in Calculus of Variation and Partial Differential Equations.

[8] Eymard R., Fuhrmann J., Gärtner K., *A finite volume scheme for nonlinear parabolic equations derived from one-dimensional local Dirichlet problems*, Numer. Math. **102** (2006),no. 3,463–495.

[9] Eymard R., Gallouët T., *H-convergence and numerical schemes for elliptic problems*, SIAM J. Numer. Anal. **41** (2003), no. 2, 539–562.

[10] Eymard R., Gallouët T., Herbin R., Finite volume methods, *Handbook of numerical analysis, Vol. VII, North-Holland, Amsterdam* (2000), 713–1020.

[11] Houston P., Perugia I., Schneebeli A. and Schötzau D., *Mixed discontinuous Galerkin approximation of the Maxwell operator: The indefinite case*, M2AN Math. Model. Numer. Anal. **39** (2005, no. 4, 727–753.

[12] Il'in A. M., *A difference scheme for a differential equation with a small parameter multiplying the highest derivative*, Mat. Zametki **6** (1969), 237–248.

[13] Markowich P. A., The stationary semiconductor device equations, *Springer-Verlag, Computational Microelectronics, Vienna* (1986).

[14] Mishev I. D., *Finite volume element methods for non-definite problems*, Numer. Math. **83** (1999), 161–175.

[15] Monk P. and Süli E., *The Adaptive Computation of Far-Field Patterns by A Posteriori Error Estimation of Linear Functionals*, SIAM J. Numer. Anal. **36** (1998), no. 1, 251–274.

[16] Scharfetter D. L., Gummel H. K., *Large signal analysis of a silicon Read diode*, IEEE Trans. on Elec. Dev. **16** (1969), 64–77.

[17] Schatz A.H., *An Observation Concerning Ritz-Galerkin Methods with Indefinite Bilinear Forms*, Math. Comp. **28** (1974), 959–962.