Stability and Error Analysis of Mixed Finite Volume Methods for Advection Dominated Problems

F. Brezzi[1,2] , L. D. Marini[1,2], S. Micheletti[3], P. Pietra[2], R. Sacco[3]

[1] Dipartimento di Matematica "F. Casorati", Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy
[2] Istituto di Matematica Applicata e Tecnologie Informatiche-CNR, Via Ferrata 1,27100 Pavia,Italy
[3] MOX, Dipartimento di Matematica "F. Brioschi", Politecnico di Milano,
Via Bonardi 9, 20133 Milano, Italy

Abstract

We consider a convection–diffusion–reaction problem, and we analyze a stabilized mixed finite volume scheme introduced in [23]. The scheme is presented in the format of Discontinuous Galerkin methods, and error bounds are given, proving $\mathcal{O}(h^{1/2})$ convergence in the L^2-norm for the scalar variable, which is approximated with piecewise constant elements.

1 Introduction

Advection-diffusion-reaction equations constitute a well-established model to describe a wide variety of problems in real-life applications. Transport and diffusion of heat in a body or of a pollutant substance flowing into water, oxigen exchange across an arterial wall in haemodynamics, electron and hole current flow in a semiconductor device are just a few but relevant examples of the use of advective-diffusive models in applied sciences.

Here, we consider the stationary convection-diffusion-reaction model problem

$$\begin{cases} -\mathrm{div}\,(\varepsilon\nabla u) + \mathrm{div}\,(\boldsymbol{\beta} u) + \gamma u = f \ \text{ in } \ \Omega, & \\ u = g & \text{on } \ \Gamma_D, \\ (\varepsilon\nabla u - \boldsymbol{\beta} u)\cdot \mathbf{n} = 0 & \text{on } \ \Gamma_N, \end{cases} \tag{1.1}$$

where Ω is a convex polygonal domain in $\mathbb{R}^2$ with boundary $\partial\Omega \equiv \Gamma = \Gamma_D \cup \Gamma_N$, $\mathbf{n}$ is the unit outward normal vector, and $f,\ g$ are given functions, with $f \in L^2(\Omega)$, and $g \in H^{1/2}(\Gamma_D)$. Moreover, $\varepsilon = \varepsilon(x)$, $\boldsymbol{\beta} = \boldsymbol{\beta}(x)$, and $\gamma = \gamma(x)$ are given regular

functions on $\overline{\Omega}$ such that

$$\exists\, \varepsilon_0,\, \varepsilon_M \quad \text{such that} \qquad \varepsilon_M \geq \varepsilon(x) \geq \varepsilon_0 > 0, \tag{1.2}$$

$$\exists\, \gamma_0,\, \gamma_M \quad \text{such that} \qquad \gamma_M \geq \gamma(x) \geq \gamma_0 \geq 0, \tag{1.3}$$

$$\exists\, b_0 \quad \text{such that} \qquad \gamma + \frac{1}{2}\text{div}\, \boldsymbol{\beta} \geq b_0 > 0. \tag{1.4}$$

Existence and uniqueness of the solution of (1.1) then follows by the maximum principle. Moreover, under the additional assumption

$$\boldsymbol{\beta} \cdot \mathbf{n} \leq 0 \quad \text{on } \Gamma_N, \tag{1.5}$$

the usual *coercivity* bound holds

$$\int_\Omega (\varepsilon_0 |\nabla u|^2 + b_0 u^2)\, dx - \frac{1}{2}\int_{\Gamma_N} \boldsymbol{\beta}\cdot\mathbf{n} u^2\, ds \leq \int_\Omega f u\, dx + \int_{\Gamma_D} g\varepsilon \nabla u \cdot \mathbf{n}\, ds - \frac{1}{2}\int_{\Gamma_D} g^2 \boldsymbol{\beta}\cdot\mathbf{n}\, ds. \tag{1.6}$$

In the present paper we shall analyze a discretization of (1.1) based on a Mixed Finite Volume approach described in [23, 10]. Essentially, this approach consists in writing (1.1) in the mixed form, and to discretize the flux variable by the lowest order Raviart-Thomas element, and the scalar variable by piecewise constants. The use of a suitable quadrature formula (see [15, 4, 5]), which diagonalizes the "mass" matrix applied to the flux vector variable, allows then to eliminate the flux variable from the mixed system. In such a way the final scheme, acting on the scalar variable only, can be regarded as a Mixed Finite Volume (MFV) cell-centered approximation of problem (1.1). Other approaches for the"mass" matrix diagonalization in the case of rectangular and triangular grids have been proposed and analyzed in [1, 11, 14, 19]. In the present paper particular attention is given to the case of advection dominated problems, for which it is well known that a stabilization procedure is necessary. This is done (see [23, 10]) by introducing a suitable artificial diffusion term at each edge of the computational grid. For an application to semiconductor device simulation see [26].

The paper is organized as follows. In Sect. 2 we recall the mixed formulation of (1.1), and the discretization via the lowest-order Raviart-Thomas element. In Sect. 3 we introduce the quadrature formula used to diagonalize the "mass" matrix, and we recast the MFV scheme within the more general format of Discontinuous Galerkin methods. This allows us to write the MFV approach as a generalized Galerkin method using piecewise constant finite elements for the scalar variable. The stabilization of the MFV procedure is described in Sect. 4. Then, in Sect. 5 the error analysis of the stabilized MFV scheme is carried out, proving $\mathcal{O}(h^{1/2})$ convergence in the L^2-norm for the approximate scalar variable. This error estimate can be regarded as optimal, since the loss of half a power of h is sort of physiological in advection dominated problems. Moreover, it is independent of the size of the diffusion coefficient, so that it does not blow up in the limit of vanishing viscosity.

2 Mixed finite element discretization

In order to write problem (1.1) in mixed form, we introduce the *flux* $\boldsymbol{\sigma} = \varepsilon\nabla u - \boldsymbol{\beta} u$ as an independent variable, so that (1.1) becomes

$$\begin{array}{llll} \boldsymbol{\sigma} = \varepsilon\nabla u - \boldsymbol{\beta} u, & \text{and} & -\text{div}\,\boldsymbol{\sigma} + \gamma u = f & \text{in } \Omega, \\ u = g \quad \text{on} \quad \Gamma_D, & & \boldsymbol{\sigma}\cdot\mathbf{n} = 0 \quad \text{on} \quad \Gamma_N. & \end{array} \tag{2.1}$$

Defining the spaces

$$\Sigma = \left\{\boldsymbol{\tau} \in (L^2(\Omega))^2 \,|\, \text{div}\,\boldsymbol{\tau} \in L^2(\Omega),\, \boldsymbol{\tau}\cdot\mathbf{n} = 0 \text{ on}\, \Gamma_N\right\} \subset H(\text{div}\,;\Omega), \tag{2.2}$$

$$V = L^2(\Omega), \tag{2.3}$$

with norms

$$\|v\|_V := \|v\|_{L^2(\Omega)}, \tag{2.4}$$

$$\|\boldsymbol{\tau}\|^2_\Sigma := \|\boldsymbol{\tau}\|^2_{H(\text{div}\,;\Omega)} = \|\boldsymbol{\tau}\|^2_{L^2(\Omega)} + \|\text{div}\,\boldsymbol{\tau}\|^2_{L^2(\Omega)}, \tag{2.5}$$

the mixed variational formulation of problem (1.1) can be written as

$$\begin{cases} \text{Find } (\boldsymbol{\sigma}, u) \in \Sigma \times V \ \text{ such that} \\ a(\boldsymbol{\sigma}, \boldsymbol{\tau}) + b_1(u, \boldsymbol{\tau}) = < g, \boldsymbol{\tau}\cdot\mathbf{n} > \ \forall\, \boldsymbol{\tau} \in \Sigma, \\ b_2(v, \boldsymbol{\sigma}) - c(u, v) = -(f, v) \qquad \forall\, v \in V, \end{cases} \tag{2.6}$$

where, with $\alpha := \varepsilon^{-1}$, we set

$$\begin{array}{lll} a(\boldsymbol{\sigma}, \boldsymbol{\tau}) = \displaystyle\int_\Omega \alpha \boldsymbol{\sigma}\cdot\boldsymbol{\tau}\, dx, & & \boldsymbol{\sigma}, \boldsymbol{\tau} \in \Sigma, \\ b_1(v, \boldsymbol{\tau}) = \displaystyle\int_\Omega v\, \text{div}\,\boldsymbol{\tau}\, dx + \int_\Omega \alpha\, v\, \boldsymbol{\beta}\cdot\boldsymbol{\tau}\, dx, & & v \in V, \boldsymbol{\tau} \in \Sigma, \\ b_2(v, \boldsymbol{\tau}) = \displaystyle\int_\Omega v\, \text{div}\,\boldsymbol{\tau}\, dx, & & v \in V, \boldsymbol{\tau} \in \Sigma, \\ c(u, v) = \displaystyle\int_\Omega \gamma\, u\, v\, dx, & & u, v \in V. \end{array} \tag{2.7}$$

In (2.6) the brackets $< .,. >$ denote the duality between $H^{1/2}(\Gamma)$ and its dual space $H^{-1/2}(\Gamma)$, and $(\cdot,\cdot)$ denotes the L^2-scalar product. A way to prove existence and uniqueness of the solution of problem (2.6) is to check that a solution of (2.1) (in the distributional sense) is a solution of (2.6) and vice-versa, and use the obvious equivalence of (2.1) and (1.1).

In order to discretize problem (2.6), let $\{\mathcal{T}_h\}_h$ be a family of regular decompositions of $\overline{\Omega}$ into triangles T [12], such that there is always a vertex of $\mathcal{T}_h$ on the interface between Γ_D and Γ_N.
We shall approximate the scalar variable u with piecewise constant functions on $\mathcal{T}_h$, and the vector variable $\boldsymbol{\sigma}$ with the lowest-order Raviart-Thomas element (see [24] and [6]) defined, on each $T \in \mathcal{T}_h$, by

$$\mathbb{RT}_0(T) = \text{span}\left\{(1,0),\ (0,1),\ (x,y)\right\}. \tag{2.8}$$

Next, we form the finite element spaces as

$$\Sigma_h = \left\{ \boldsymbol{\tau}_h \in \Sigma \mid \boldsymbol{\tau}_{h|T} \in \mathbb{RT}_0(T)\,, \forall\, T \in \mathcal{T}_h \right\}, \tag{2.9}$$

$$V_h = \left\{ v_h \in V \mid v_{h|T} \in \mathbb{P}_0(T)\,, \forall\, T \in \mathcal{T}_h \right\}. \tag{2.10}$$

Then, the discrete formulation of (2.6) is:

$$\begin{cases} \text{Find } (\boldsymbol{\sigma}_h, u_h) \in \Sigma_h \times V_h \ \text{ such that} \\ a(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + b_1(u_h, \boldsymbol{\tau}_h) = < g, \boldsymbol{\tau}_h \cdot \mathbf{n} > \ \forall\, \boldsymbol{\tau}_h \in \Sigma_h, \\ b_2(v_h, \boldsymbol{\sigma}_h) - c(u_h, v_h) = -(f, v_h) \qquad \forall\, v_h \in V_h. \end{cases} \tag{2.11}$$

For future purposes it is convenient to assume that the convective field $\boldsymbol{\beta}$ in (2.11) has continuous normal component across each edge of the triangulation. We therefore assume that $\boldsymbol{\beta}$ is itself a Raviart–Thomas element vector field. The algebraic form of (2.11) reads

$$\begin{pmatrix} A & B_1 \\ B_2 & C \end{pmatrix} \begin{pmatrix} \Phi_h \\ U_h \end{pmatrix} = \begin{pmatrix} G_h \\ F_h \end{pmatrix} \tag{2.12}$$

where Φ_h is the vector of the unknown fluxes of $\boldsymbol{\sigma}_h$ across each edge of $\mathcal{T}_h$, and U_h is the vector of the unknown values of u_h in each $T \in \mathcal{T}_h$. Eliminating Φ_h leads to the following scheme for U_h

$$(C - B_2 A^{-1} B_1) U_h = F_h - B_2 A^{-1} G_h.$$

The matrix $M \equiv C - B_2 A^{-1} B_1$ is full and, in general, neither symmetric nor positive definite, so that solving this system can be quite expensive. It is also well known that M is not an M-matrix for any value of γ, as pointed out in [6, 20, 21] in the case of reaction–diffusion problems. Moreover, for advection dominated problems the scheme is not stable.
The reduced integration for the "mass" matrix and the connected stabilization procedure developed in the forthcoming sections will allow us to circumvent the drawbacks of the $\mathbb{RT}_0$ approximation, leading to stable cell-centered finite volume methods that preserve the good approximation properties provided by the mixed approach, though at a reduced computational cost.

3 The mixed finite volume formulation

In this section we introduce, starting from formulation (2.11), the mixed finite volume (MFV) discretization of problem (1.1). As a first step, however, we need to introduce convenient notation.

3.1 Notation

For a given regular triangulation $\mathcal{T}_h$ [12], we denote by N_E and N_T the total number of edges and triangles of $\mathcal{T}_h$, respectively. For every triangle $T_k \in \mathcal{T}_h$, let h_T denote the diameter of T_k, and $h = \max_{T_k \in \mathcal{T}_h} h_T$. In what follows, we then agree that

- superscripts will be used for edges (as $e^r,\ 1 \le r \le N_E$),
- subscripts will be used for triangles (as $T_k,\ 1 \le k \le N_T$),

and we introduce the following notation.

- $\mathcal{T}_h$ denotes as well the set of triangles of the triangulation $\mathcal{T}_h$.
- $\mathcal{E}_h$ denotes the set of edges in $\mathcal{T}_h$, and $\mathcal{E}_h^0$ the subset of those that do not belong to Γ_N.
- For $r = 1, ..., N_E$ the set $T(r)$ contains the indices of the triangles having e^r as an edge.
- For $k = 1, ..., N_T$ the set $E(k)$ contains the indices of the edges of T_k.
- For $k = 1, ..., N_T$ and $r \in E(k)$ we denote by $\mathbf{n}_k^r$ the unit vector normal to e^r and pointing out of T_k.
- For $k = 1, ..., N_T$, with $E(k) = (\ell, r, s)$, we also define the *vectors* $\underline{e}_k^\ell, \underline{e}_k^r, \underline{e}_k^s$ obtained by orienting the boundary of T_k counterclockwise. Observe that $\underline{e}_j^r = -\underline{e}_k^r$ for $j, k \in T(r)$.

Hence $r \in E(k)$, or, equivalently, $k \in T(r)$, means that e^r is an edge of the triangle T_k. Clearly, $E(k)$ will always contain three indices, while $T(r)$ might contain one index or two, according to whether or not the edge e^r is a boundary edge. For future purposes, it will also be useful to recall some notation typically used in the treatment of Discontinuous Galerkin methods. Assume that φ is a piecewise smooth scalar function and $\mathbf{q}$ a piecewise smooth vector valued function on $\mathcal{T}_h$.

- For each internal edge e^r, with $T(r) = \{j, k\}$ we define averages and jumps as follows.

$$\begin{aligned} \{\varphi\}^r &:= \frac{\varphi_j + \varphi_k}{2}, & \{\mathbf{q}\}^r &:= \frac{\mathbf{q}_j + \mathbf{q}_k}{2}, \\ [\varphi]^r &:= \varphi_j \mathbf{n}_j^r + \varphi_k \mathbf{n}_k^r, & [\mathbf{q}]^r &:= \mathbf{q}_j \cdot \mathbf{n}_j^r + \mathbf{q}_k \cdot \mathbf{n}_k^r. \end{aligned} \tag{3.1}$$

- On a boundary edge e^r with $T(r) = \{k\}$ we set instead

$$\{\varphi\}^r := \frac{\varphi_k}{2}, \qquad \{\mathbf{q}\}^r := \frac{\mathbf{q}_k}{2}, \qquad [\varphi]^r := \varphi_k \mathbf{n}_k^r, \qquad [\mathbf{q}]^r := \mathbf{q}_k \cdot \mathbf{n}_k^r. \tag{3.2}$$

The superscript r will sometimes be omitted, when no confusion can occur. We point out that the jump of a scalar is a vector (normal to the edge) and the jump of a vector is a scalar (that, in particular, only depends on the normal component). We recall immediately the following *basic identity*

$$\sum_{T_k \in \mathcal{T}_h} \int_{\partial T_k} \varphi_k \mathbf{q}_k \cdot \mathbf{n}_k \, ds = \sum_{e^r \in \mathcal{E}_h} \int_{e^r} [\mathbf{q}]^r \{\varphi\}^r \, ds + \sum_{e^r \in \mathcal{E}_h} \int_{e^r} \{\mathbf{q}\}^r \cdot [\varphi]^r \, ds, \tag{3.3}$$

that can be easily deduced by rearranging terms (see e.g. [7] or [3])).

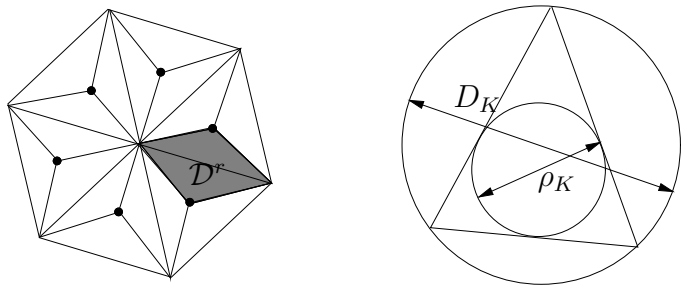

FIGURE 1
Primal triangulation $\mathcal{T}_h$ with the corresponding lumping regions $\mathcal{D}^r$ (left), mesh parameters (right).

Our next step will be to define the so-called *lumping regions*. For this, we need further assumptions on the triangulation. Namely, we assume that $\mathcal{T}_h$ is a Delaunay triangulation (see [13]). We then consider the dual tessellation $\mathcal{D}_h$ of $\mathcal{T}_h$, which is constructed in the following way.

- For every edge e^r and for every index $k \in T(r)$ we denote by C_k the circumcenter of T_k.
- For every edge e^r and for every index $k \in T(r)$ we denote by T_k^r the sub-triangle of T_k having e^r as an edge and C_k as opposite vertex. If C_k belongs to e^r (that means that the angle of T_k opposite to e^r is $\pi/2$) then sub-triangle T_k^r degenerates and we consider it to be empty.
- For every edge e^r the corresponding lumping region $\mathcal{D}^r$ is then given as (see Figure 2)
$$\mathcal{D}^r := \bigcup_{k \in T(r)} T_k^r. \tag{3.4}$$

We define now some additional averages of functions and vectors on the mesh $\mathcal{T}_h$ or on its dual tessellation $\mathcal{D}_h$.

- For any $T_k \in \mathcal{T}_h$ and for any integrable function φ, we define its mean value as
$$\overline{\varphi}_k = \frac{1}{|T_k|} \int_{T_k} \varphi \, dx \tag{3.5}$$
where $|T_k|$ is the area of T_k, and we denote by $\overline{\varphi}$ the corresponding piecewise constant function assuming the value $\overline{\varphi}_k$ in T_k for every k.
- For any $e^r \in \mathcal{E}_h$ and for any integrable function φ, we define its mean value on $\mathcal{D}^r$ as
$$\widehat{\varphi}^r = \frac{1}{|\mathcal{D}^r|} \int_{\mathcal{D}^r} \varphi \, dx \tag{3.6}$$
where $|\mathcal{D}^r|$ is the area of $\mathcal{D}^r$, and we denote by $\widehat{\varphi}$ the corresponding piecewise constant function assuming the value $\widehat{\varphi}^r$ in $\mathcal{D}^r$ for every r.
- Finally, for every piecewise smooth vector valued function $\mathbf{q}$ having continuous normal component on the edges in $\mathcal{E}_h$, and for every $e^r \in \mathcal{E}_h$ we define its *normal flux vector* $\widehat{\mathbf{q}}^r$ by
$$\widehat{\mathbf{q}}^r := \frac{1}{|e^r|} \left(\int_{e^r} \mathbf{q} \cdot \mathbf{n}^r \, ds \right) \mathbf{n}^r, \tag{3.7}$$

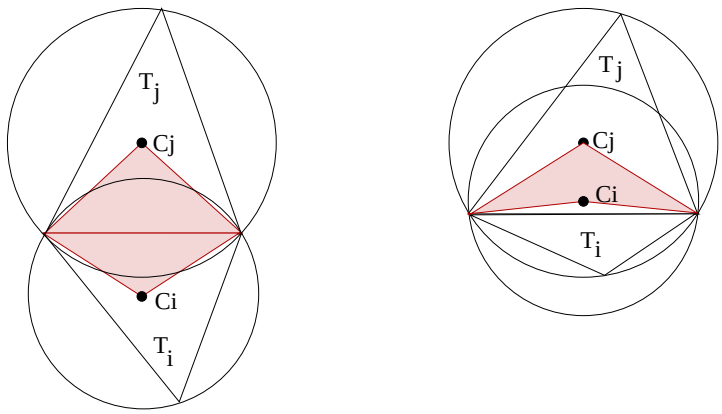

FIGURE 2
Examples of lumping regions for acute (left) and obtuse (right) triangles.

where $\mathbf{n}^r$ is (any) unit vector normal to e^r. We note that, for the particular case of $\mathbf{q} \in \Sigma_h$, we have that $\widehat{\mathbf{q}}^r$ corresponds to the *normal part* of $\mathbf{q}$ (the one that is continuous).

- In general, a function denoted with an over-bar will always be assumed to be piecewise constant on the triangulation, while a function denoted with a hat will be assumed to be constant in each lumping region.

We are now ready to introduce the mixed finite volume discretization of (2.1), setting, without loss of generality, $g \equiv 0$ in order to simplify the exposition. Our main step will be the use of a suitable numerical integration to approximate the bilinear forms a, b_1, b_2 and c appearing in (2.11).

3.2 The integration formula and the scheme

To simplify the notation, throughout this section we shall drop the subscript h from our discrete unknowns and test functions. We shall get back to the proper notation in the last section, where, for obtaining error estimates, it will be necessary to distinguish $\boldsymbol{\sigma}_h$ from $\boldsymbol{\sigma}$ and u_h from u.
To approximate some of the integrals in our mixed formulation we shall use a quadrature formula based on that proposed and analyzed in [15, 4, 5], that we recall here briefly. Let $T_k \in \mathcal{T}_h$, let $\mathbf{q}$ and $\mathbf{p}$ be smooth vector valued functions on T_k and let μ be a smooth scalar function on Ω. We take

$$\int_{T_k} \mu \mathbf{q} \cdot \mathbf{p} \, dx \simeq \sum_{r \in E(k)} \widehat{\mu}^r \, \widehat{\mathbf{q}}^r \cdot \widehat{\mathbf{p}}^r \, |e^r|^2 \, \omega_k^r. \tag{3.8}$$

Notice that formula (3.8) amounts to a diagonalization of the "mass" matrix when $\mathbf{p}, \mathbf{q}$ are $\mathbb{RT}_0$ vectors with degrees of freedom chosen as the edge fluxes. Moreover, it can be proved that the formula (3.8) is exact for constant $\mathbf{q}$, $\mathbf{p}$, and μ, if and only if the weights ω_k^r are given by the formula:

$$\omega_k^r = -\frac{\underline{e}_k^i \cdot \underline{e}_k^j}{4|T_k|} \qquad i, j, r \in E(k)\,,\ i \neq r, j \neq r, i \neq j. \tag{3.9}$$

We point out that the quantities ω_k^r can also be computed using the formula

$$\omega_k^r = \frac{d_k^r}{|e^r|}, \tag{3.10}$$

where d_k^r is the distance between the circumcenter C_k and the edge e^r.

Remark 3.1 *Actually, formula* (3.10) *could as well be used if T_k has an obtuse angle, although ω_k^r (when e^r is opposite to the obtuse angle) becomes negative. In this case formula (3.10) will also hold, but taking d_k^r to be minus the distance between the circumcenter C_k (that now is external to T_k) and the edge e^r (see [10] for a detailed discussion). Expression* (3.10) *is very important in view of the finite volume interpretation of the numerical method obtained with the quadrature formula* (3.8)*. However, we point out that expression* (3.9) *is easier to compute, and is actually used in the implementation of the method.*

The analysis and examples of application of (3.8)-(3.10) can be found in [10, 23, 26, 27, 22].
Applying the quadrature formula (3.8) to the bilinear form a appearing in (2.7) we get

$$a(\boldsymbol{\sigma}, \boldsymbol{\tau}) \equiv \int_\Omega \alpha \boldsymbol{\sigma} \cdot \boldsymbol{\tau} \, dx \simeq \sum_{T_k \in \mathcal{T}_h} \sum_{r \in E(k)} \widehat{\alpha}^r \widehat{\boldsymbol{\sigma}}^r \cdot \widehat{\boldsymbol{\tau}}^r \, |e^r|^2 \omega_k^r. \tag{3.11}$$

Then we define our approximate bilinear form a_h as:

$$a_h(\boldsymbol{\sigma}, \boldsymbol{\tau}) := \sum_{T_k \in \mathcal{T}_h} \sum_{r \in E(k)} \widehat{\alpha}^r \widehat{\boldsymbol{\sigma}}^r \cdot \widehat{\boldsymbol{\tau}}^r \, |e^r|^2 \omega_k^r. \tag{3.12}$$

Setting also

$$d^r := \sum_{j \in T(r)} d_j^r \quad \text{and } \omega^r := \sum_{j \in T(r)} \omega_j^r \equiv \frac{d^r}{|e^r|}, \tag{3.13}$$

we can write our bilinear form as

$$a_h(\boldsymbol{\sigma}, \boldsymbol{\tau}) := \sum_{e^r \in \mathcal{E}_h} \widehat{\alpha}^r \widehat{\boldsymbol{\sigma}}^r \cdot \widehat{\boldsymbol{\tau}}^r \, |e^r|^2 \omega^r \equiv \sum_{e^r \in \mathcal{E}_h} \widehat{\alpha}^r \widehat{\boldsymbol{\sigma}}^r \cdot \widehat{\boldsymbol{\tau}}^r \, |e^r| d^r \equiv 2 \sum_{e^r \in \mathcal{E}_h} \widehat{\alpha}^r \widehat{\boldsymbol{\sigma}}^r \cdot \widehat{\boldsymbol{\tau}}^r \, |\mathcal{D}^r|. \tag{3.14}$$

Remark 3.2 *We observe that, for each edge $e^r \in \mathcal{E}_h$, other choices of $\widehat{\alpha}^r$ are possible: here we have taken the average of α over the lumping region $\mathcal{D}^r$, but this is not mandatory. It suffices that $\widehat{\alpha}^r$ is constant over $\mathcal{D}^r$ (see [23] for alternative choices).*

We consider now the bilinear form b_1 appearing in (2.7). The first term does not require any special adjustment. Indeed, using our basic formula (3.3) and taking again into account the continuity of the normal component of the elements in Σ_h we have

$$\int_\Omega u \, \mathrm{div} \boldsymbol{\tau} \, dx = \sum_{T_k \in \mathcal{T}_h} \int_{\partial T_k} u_k \, \boldsymbol{\tau}_k \cdot \mathbf{n}_k \, ds = \sum_{e^r \in \mathcal{E}_h} \int_{e^r} [u]^r \cdot \{\boldsymbol{\tau}\}^r \, ds = \sum_{e^r \in \mathcal{E}_h} [u]^r \cdot \widehat{\boldsymbol{\tau}}^r \, |e^r|. \tag{3.15}$$

This gives us at once a new way of writing the bilinear form b_2 appearing in (2.7). Indeed, we set

$$b_{2,h}(v,\boldsymbol{\sigma}) := \sum_{e^r\in\mathcal{E}_h} [v]^r \cdot \widehat{\boldsymbol{\sigma}}^r\, |e^r| = \sum_{e^r\in\mathcal{E}_h} \int_{e^r} [v]^r \cdot \widehat{\boldsymbol{\sigma}}^r\, ds (\equiv b_2(v,\boldsymbol{\sigma})). \tag{3.16}$$

In order to apply the quadrature formula to the second integral appearing in the definition of $b_1(\cdot,\cdot)$ (see (2.7)), a *unique* value for u needs to be defined at each edge. It seems natural, at first, to take the average of u on e^r, as defined in (3.1) and (3.2). Then, applying quadrature formula (3.8) and arguing as in (3.14), we have

$$\int_\Omega \alpha\, u\, \boldsymbol{\beta}\cdot\boldsymbol{\tau}\, dx \simeq \sum_{T_k\in\mathcal{T}_h} \sum_{r\in E(k)} \widehat{\alpha}^r\, \{u\}^r\, \widehat{\boldsymbol{\beta}}^r \cdot \widehat{\boldsymbol{\tau}}^r\, |e^r|^2 \omega_k^r \equiv 2 \sum_{e^r\in\mathcal{E}_h} \widehat{\alpha}^r\, \{u\}^r\, \widehat{\boldsymbol{\beta}}^r \cdot \widehat{\boldsymbol{\tau}}^r\, |\mathcal{D}^r|. \tag{3.17}$$

Collecting (3.15) and (3.17) we can finally write

$$b_{1,h}(u,\boldsymbol{\tau}) := \sum_{e^r\in\mathcal{E}_h} [u]^r \cdot \widehat{\boldsymbol{\tau}}^r\, |e^r| + 2 \sum_{e^r\in\mathcal{E}_h} \widehat{\alpha}^r\, \{u\}^r\, \widehat{\boldsymbol{\beta}}^r \cdot \widehat{\boldsymbol{\tau}}^r\, |\mathcal{D}^r|. \tag{3.18}$$

To conclude, we take $c_h(u,v) \equiv c(u,v)$, as defined in (2.7), and we note that

$$c_h(u,v) \equiv c(u,v) = \sum_{T_k\in\mathcal{T}_h} \overline{\gamma}\, u_k\, v_k |T_k|. \tag{3.19}$$

Having defined the approximate bilinear forms a_h, $b_{1,h}$, $b_{2,h}$, and c_h, we can now write the following final form of our scheme:

$$\begin{cases} \text{Find } (\boldsymbol{\sigma},u)\in\Sigma_h\times V_h \ \text{ such that} & \\ a_h(\boldsymbol{\sigma},\boldsymbol{\tau}) + b_{1,h}(u,\boldsymbol{\tau}) = 0 & \forall\, \boldsymbol{\tau}\in\Sigma_h, \\ b_{2,h}(v,\boldsymbol{\sigma}) - c(u,v) = -(f,v) & \forall\, v\in V_h. \end{cases} \tag{3.20}$$

Now we would like to take advantage of the fact that our bilinear form a_h is diagonal, in order to eliminate $\boldsymbol{\sigma}$ from the first equation of (3.20) and insert it into the second, so that the final scheme could be written in terms of u only.

With this aim we recall from (3.14) and (3.18) that the first equation of (3.20) can be written as

$$\sum_{e^r\in\mathcal{E}_h} \Big(2\,\widehat{\alpha}^r \widehat{\boldsymbol{\sigma}}^r \cdot \widehat{\boldsymbol{\tau}}^r\, |\mathcal{D}^r| + [u]^r \cdot \widehat{\boldsymbol{\tau}}^r\, |e^r| + 2\,\widehat{\alpha}^r\, \{u\}^r\, \widehat{\boldsymbol{\beta}}^r \cdot \widehat{\boldsymbol{\tau}}^r\, |\mathcal{D}^r| \Big) = 0, \tag{3.21}$$

which gives immediately, for the edges $e^r \notin \Gamma_N$,

$$\widehat{\boldsymbol{\sigma}}^r = -\frac{[u]^r\, |e^r|}{2\widehat{\alpha}^r\, |\mathcal{D}^r|} - \{u\}^r\, \widehat{\boldsymbol{\beta}}^r = -\frac{[u]^r}{\widehat{\alpha}^r\, d^r} - \{u\}^r\, \widehat{\boldsymbol{\beta}}^r \qquad \forall e^r\in\mathcal{E}_h^0. \tag{3.22}$$

Substituting into the second equation of (3.20) and using (3.16) and (3.19) we have immediately

$$\sum_{e^r\in\mathcal{E}_h^0} \left(\frac{[u]^r\cdot[v]^r\, |e^r|}{\widehat{\alpha}^r\, d^r} + \{u\}^r\, \widehat{\boldsymbol{\beta}}^r \cdot [v]^r\, |e^r| \right) + \int_\Omega \gamma\, u\, v\, dx = \int_\Omega f\, v\, dx \quad \forall v\in V_h. \tag{3.23}$$

Setting now

$$\widehat{\varepsilon}^r := (\widehat{\alpha}^r)^{-1}, \tag{3.24}$$

recalling that

$$\frac{|e^r|}{d^r} = 2\frac{|\mathcal{D}^r|}{(d^r)^2}, \tag{3.25}$$

and finally recalling definition (3.7), relation (3.23) can also be written as

$$\sum_{e^r\in\mathcal{E}_h^0} 2\int_{\mathcal{D}^r} \widehat{\varepsilon}^r \frac{[u]^r}{d^r}\cdot\frac{[v]^r}{d^r}\,dx + \sum_{e^r\in\mathcal{E}_h^0}\int_{e^r} \{u\}^r\,\boldsymbol{\beta}\cdot[v]^r\,ds + \int_\Omega \gamma\,u\,v\,dx = \int_\Omega f\,v\,dx \quad \forall v\in V_h. \tag{3.26}$$

This allows us to write the final formulation of our MFV scheme in terms of the scalars u and v only. Indeed, we can set

$$\mathcal{L}(u,v) := \sum_{e^r\in\mathcal{E}_h^0} 2\int_{\mathcal{D}^r} \widehat{\varepsilon}^r \frac{[u]^r}{d^r}\cdot\frac{[v]^r}{d^r}\,dx + \sum_{e^r\in\mathcal{E}_h^0}\int_{e^r} \{u\}^r\,\boldsymbol{\beta}\cdot[v]^r\,ds + \int_\Omega \gamma\,u\,v\,dx, \tag{3.27}$$

and write our discrete problem as

$$\begin{cases} \text{Find } u\in V_h \ \text{ such that} \\ \mathcal{L}(u,v) = (f,v) \qquad\qquad \forall\, v\in V_h. \end{cases} \tag{3.28}$$

We notice that, as far as the diffusive and reactive parts of the bilinear form $\mathcal{L}(u,v)$ are concerned, i.e., the first and third terms in (3.27), respectively, it can be proved that they give rise to an M-matrix (see e.g., [10]). In particular, the third term yields a positive diagonal matrix, while the first term provides an M-matrix, provided that the terms d^r appearing in (3.27) and defined in (3.13) are positive. This is guaranteed if $\mathcal{T}_h$ is a Delaunay triangulation. Actually, thanks to this property, though one of the terms d_j^r in (3.13) may be negative (when the angle opposite to edge e^r in triangle K_j is obtuse), the term d^r is always positive. However, the M-matrix property is lost when in (3.27) advection dominates. In the next section a stabilization of the MFV scheme (3.28) is introduced, with the effect that it always yields an M-matrix, independently of the strength of the advective field $\boldsymbol{\beta}$.

4 Stabilization of the mixed finite volume scheme

We start by noticing that, taking $u = v$ in (3.27), we have

$$\mathcal{L}(v,v) := \sum_{e^r\in\mathcal{E}_h^0} 2\int_{\mathcal{D}^r} \widehat{\varepsilon}^r \left|\frac{[v]^r}{d^r}\right|^2 dx + \sum_{e^r\in\mathcal{E}_h^0}\int_{e^r} \{v\}^r\,\boldsymbol{\beta}\cdot[v]^r\,ds + \int_\Omega \gamma\,v^2\,dx. \tag{4.1}$$

We also note that

$$2\,\{v\}^r\,[v]^r = [v^2]^r, \tag{4.2}$$

so that using our basic equation (3.3) with $\varphi = v^2$ and $\mathbf{q} = \boldsymbol{\beta}$, and recalling that $[\boldsymbol{\beta}] = 0$, we get

$$2\sum_{e^r\in\mathcal{E}_h^0}\int_{e^r} \{v\}^r\,\boldsymbol{\beta}\cdot[v]^r\,ds = \sum_{T_k\in\mathcal{T}_h}\int_{\partial T_k\setminus\Gamma_N} \boldsymbol{\beta}\cdot\mathbf{n}_k\,v_k^2\,ds = \int_\Omega \operatorname{div}\boldsymbol{\beta}\,v^2\,dx - \int_{\Gamma_N} \boldsymbol{\beta}\cdot\mathbf{n}\,v^2\,ds. \tag{4.3}$$

Combining (4.1), (4.3), and assumption (1.4), we finally have

$$\begin{aligned}\mathcal{L}(v,v) &= 2\sum_{e^r\in\mathcal{E}_h^0}\int_{\mathcal{D}^r}\widehat{\varepsilon}^r\left|\frac{[v]^r}{d^r}\right|^2 dx + \int_\Omega\left(\frac{1}{2}\mathrm{div}\,\boldsymbol{\beta}+\gamma\right)v^2\,dx - \frac{1}{2}\int_{\Gamma_N}\boldsymbol{\beta}\cdot\mathbf{n}\,v^2\,ds\\ &\geq 2\sum_{e^r\in\mathcal{E}_h^0}\int_{\mathcal{D}^r}\widehat{\varepsilon}^r\left|\frac{[v]^r}{d^r}\right|^2 dx + b_0\,||v||_0^2 \underbrace{-\frac{1}{2}\int_{\Gamma_N}\boldsymbol{\beta}\cdot\mathbf{n}\,v^2\,ds}_{\geq 0},\end{aligned} \tag{4.4}$$

where the last term in (4.4) is nonnegative, due to (1.5). However, as ε can be very small, the coercivity bound provided by (4.4) could be very poor, and insufficient to prove error bounds with constants independent of ε. We are going to add, therefore, some sort of additional diffusion. Actually, for every $e^r\in\mathcal{E}_h^0$ we define a real number θ^r with the assumption that

$$1/2 \geq \theta^r \geq \theta_0 > 0, \tag{4.5}$$

where θ_0 is a constant independent of the decomposition. Then we set, always for every $e^r\in\mathcal{E}_h^0$,

$$\widehat{\rho}^r := \theta^r\,d^r\,|\widehat{\boldsymbol{\beta}}^r|. \tag{4.6}$$

Then we consider the *stabilized* bilinear form $\mathcal{L}_s(u,v)$ defined as

$$\mathcal{L}_s(u,v) := 2\sum_{e^r\in\mathcal{E}_h^0}\int_{\mathcal{D}^r}(\widehat{\varepsilon}^r+\widehat{\rho}^r)\frac{[u]^r}{d^r}\cdot\frac{[v]^r}{d^r}\,dx + \sum_{e^r\in\mathcal{E}_h^0}\int_{e^r}\{u\}^r\,\boldsymbol{\beta}\cdot[v]^r\,ds + \int_\Omega\gamma\,u\,v\,dx. \tag{4.7}$$

It is clear that, instead of (4.4), we have now

$$\mathcal{L}_s(v,v) \geq 2\sum_{e^r\in\mathcal{E}_h^0}\int_{\mathcal{D}^r}(\widehat{\varepsilon}^r+\widehat{\rho}^r)\left|\frac{[v]^r}{d^r}\right|^2 dx + b_0\,||v||_0^2 - \frac{1}{2}\int_{\Gamma_N}\boldsymbol{\beta}\cdot\mathbf{n}\,v^2\,ds. \tag{4.8}$$

We shall show that different choices of θ^r in (4.6) correspond to modify definition (3.18) of the bilinear form $b_{1,h}(u,\boldsymbol{\tau})$, by taking proper values of u on the edge e^r instead of the average $\{u\}^r$. In particular, we shall consider two choices of θ^r that lead to two well known stabilization methods, namely, the upwind scheme and the Scharfetter-Gummel (SG) scheme. The SG stabilization amounts to introducing exponential fitting into the MFV formulation and is the most widely used technique in the numerical simulation of semiconductor devices using drift-diffusion and energy-transport models [28].
By defining the upwind value of u on the edge e^r

$$u^r_{\mathrm{upw}} = \begin{cases}\dfrac{1}{2|\widehat{\boldsymbol{\beta}}^r\cdot\mathbf{n}^r|}\displaystyle\sum_{j\in T(r)}u_j\left(\widehat{\boldsymbol{\beta}}^r\cdot\mathbf{n}^r_j+|\widehat{\boldsymbol{\beta}}^r\cdot\mathbf{n}^r_j|\right), & \widehat{\boldsymbol{\beta}}^r\cdot\mathbf{n}^r\neq 0\\ \{u\}^r, & \widehat{\boldsymbol{\beta}}^r\cdot\mathbf{n}^r = 0,\end{cases}$$

it can be seen that taking $\theta^r = 1/2$ in (4.6) corresponds to using the upwind value u^r_{upw} of u instead of the average $\{u\}^r$ in definition (3.18) of $b_{1,h}$. Indeed, taking

$$\sum_{e^r\in\mathcal{E}_h}\widehat{\alpha}^r\,u^r_{\mathrm{upw}}\,\widehat{\boldsymbol{\beta}}^r\cdot\widehat{\boldsymbol{\tau}}^r|\mathcal{D}^r|, \quad \text{instead of} \quad \sum_{e^r\in\mathcal{E}_h}\widehat{\alpha}^r\,\{u\}^r\,\widehat{\boldsymbol{\beta}}^r\cdot\widehat{\boldsymbol{\tau}}^r\,|\mathcal{D}^r| \tag{4.9}$$

can be easily tracked to produce $u^r_{\rm upw}\,\widehat{\boldsymbol{\beta}}^r$ instead of $\{u\}^r\,\widehat{\boldsymbol{\beta}}^r$ in (3.22), ending up with

$$\sum_{e^r\in\mathcal{E}_h^0}\int_{e^r} u^r_{\rm upw}\,\boldsymbol{\beta}\cdot[v]^r\,ds \quad \text{instead of} \quad \sum_{e^r\in\mathcal{E}_h^0}\int_{e^r} \{u\}^r\,\boldsymbol{\beta}\cdot[v]^r\,ds \tag{4.10}$$

in the final definition (3.27) of $\mathcal{L}$. It is easy to check that, if we take $\mathbf{n}^r_\beta$ to be such that $\boldsymbol{\beta}\cdot\mathbf{n}^r_\beta\geq 0$, then

$$u^r_{\rm upw}-\{u\}^r=\frac{1}{2}\mathbf{n}^r_\beta\cdot[u]^r \tag{4.11}$$

so that

$$u^r_{\rm upw}\,\boldsymbol{\beta}\cdot[v]^r-\{u\}^r\,\boldsymbol{\beta}\cdot[v]^r=\theta^r_{\rm upw}|\widehat{\boldsymbol{\beta}}^r|[u]^r\cdot[v]^r, \tag{4.12}$$

with $\theta^r_{\rm upw}=1/2$ and where we took the absolute value to recall that we always have a nonnegative term. Taking the integral of (4.12) over e^r gives

$$\int_{e^r}\theta^r_{\rm upw}|\widehat{\boldsymbol{\beta}}^r|[u]^r\cdot[v]^r\,ds=\theta^r_{\rm upw}|e^r|\,(d^r)^2\,|\widehat{\boldsymbol{\beta}}^r|\frac{[u]^r}{d^r}\cdot\frac{[v]^r}{d^r}\;=2\int_{\mathcal{D}^r}\theta^r_{\rm upw}d^r\,|\widehat{\boldsymbol{\beta}}^r|\frac{[u]^r}{d^r}\cdot\frac{[v]^r}{d^r}\,dx, \tag{4.13}$$

where $\theta^r_{\rm upw}d^r\,|\widehat{\boldsymbol{\beta}}^r|$ is precisely $\widehat{\rho}^r$ with $\theta^r=\theta^r_{\rm upw}$.

To show that we also recover the SG scheme, let us first define the "edge" value of the scalar u

$$u^r_{\rm SG}=\sum_{j\in T(r)}u_j\left(\frac{\mathcal{B}(-2{\rm Pe}^r_j)-1}{2{\rm Pe}^r_j}\right), \tag{4.14}$$

where

$$\mathcal{B}(t)=\begin{cases}\dfrac{t}{\exp(t)-1}, & t\neq 0,\\ 1, & t=0,\end{cases}$$

is the Bernoulli function, and

$${\rm Pe}^r_j=\frac{\widehat{\boldsymbol{\beta}}^r\cdot\mathbf{n}^r_j}{2\widehat{\alpha}^r d^r}$$

is the local Péclet number. Notice that $0<(\mathcal{B}(-t)-1)/t<1$, for $t\neq 0$, and it is understood that $(\mathcal{B}(-t)-1)/t=1/2$ at $t=0$. As before, it follows that

$$u^r_{\rm SG}\,\boldsymbol{\beta}\cdot[v]^r-\{u\}^r\,\boldsymbol{\beta}\cdot[v]^r=\theta^r_{\rm SG}|\widehat{\boldsymbol{\beta}}^r|[u]^r\cdot[v]^r, \tag{4.15}$$

where

$$\theta^r_{\rm SG}=\frac{\mathcal{B}(-2{\rm Pe}^r_\beta)-1}{2{\rm Pe}^r_\beta}-\frac{1}{2},$$

in which

$${\rm Pe}^r_\beta=\frac{\widehat{\boldsymbol{\beta}}^r\cdot\mathbf{n}^r_\beta}{2\widehat{\alpha}^r d^r}>0.$$

Notice that $0<\theta^r_{\rm SG}<1/2$ and that the upwind value of $\theta^r_{\rm upw}$ is recovered from $\theta^r_{\rm SG}$ for infinite local Péclet number.

The analogous result for the SG case, obtained from (4.15), holds with $\theta^r_{\rm SG}$. For more details see [10] and [9].

5 Error estimates

In order to prove error bounds for the stabilized mixed finite volume scheme corresponding to using (4.8), we need some stricter assumptions on the decomposition $\mathcal{T}_h$. In particular, we need that the coefficients d^r appearing in (3.13) and used in the numerical integration formula (3.8)-(3.10), are uniformly bounded from below as

$$d_1|e^r| \geq d^r \geq d_0|e^r| \tag{5.1}$$

where d_1 and d_0 are some given constants independent of r and h. We also assume, for simplicity, that the sequence of triangulations $\{\mathcal{T}_h\}_{h>0}$ is quasi-uniform, in the sense that there is a constant C^*, independent of the triangulation, such that

$$h_T \geq C^* h \quad \forall T \in \mathcal{T}_h. \tag{5.2}$$

As previously announced, in this section we go back to the original (and more precise) notation of Section 2, reintroducing the index h for discrete solutions. In particular, we shall indicate by u_h the solution of the *discretized stabilized problem*

$$\begin{cases} \text{Find } u_h \in V_h \ \text{ such that} \\ \mathcal{L}_s(u_h, v) = (f, v) \qquad\qquad \forall\, v \in V_h, \end{cases} \tag{5.3}$$

where $\mathcal{L}_s$ is the stabilized bilinear form defined in (4.7). The ellipticity property (4.8) easily implies existence and uniqueness of the solution of (5.3).

We recall that error estimates for the simpler case in which $\boldsymbol{\beta} = 0$ and $\gamma \geq 0$ have already been derived in [10]. In order to use these estimates, we set

$$\widetilde{f} := -\mathrm{div}\,(\varepsilon \nabla u) \text{ in } \Omega \qquad \widetilde{g}_N := \varepsilon \nabla u \cdot \mathbf{n} \equiv \boldsymbol{\beta} \cdot \mathbf{n}\, u \text{ on } \Gamma_N, \tag{5.4}$$

and we consider the auxiliary problem

$$\begin{cases} \text{Find } w \in \mathrm{H}^1(\Omega) \ \text{ such that} \\ -\mathrm{div}\,(\varepsilon \nabla w) = \widetilde{f}, \\ w = 0 \text{ on } \Gamma_D, \qquad \varepsilon \nabla w \cdot \mathbf{n} = \widetilde{g}_N \text{ on } \Gamma_N, \end{cases} \tag{5.5}$$

whose solution is obviously $w \equiv u$. We then consider the discrete solution $w_h \in V_h$ of (5.5) by means of the MFV scheme (3.28), that, in this case, becomes

$$\mathcal{L}_{\mathrm{diff}}(w_h, v) := 2 \sum_{e^r \in \mathcal{E}_h^0} \int_{\mathcal{D}^r} \widehat{\varepsilon}^r \frac{[w_h]^r}{d^r} \cdot \frac{[v]^r}{d^r}\, dx = \int_\Omega \widetilde{f} v\, dx + \int_{\Gamma_N} \widetilde{g}_N v\, ds \qquad \forall v \in V_h, \tag{5.6}$$

and for which, under suitable hypotheses, we have the error estimate [5, 10]:

$$\|u - w_h\|_{0,\Omega} \leq C\, h\, \|u\|_{2,\Omega}, \tag{5.7}$$

where the constant C only depends on the geometric constants of the triangulation $\mathcal{T}_h$, and on the maximum norm of γ, $\boldsymbol{\beta}$, and of the derivatives of ε. Therefore, in

order to get error estimates for (5.3), we can as well compare u_h with w_h. Setting, for $v \in V_h$,

$$|\!|\!| v |\!|\!|^2 := 2 \sum_{e^r \in \mathcal{E}_h^0} \int_{\mathcal{D}^r} (\widehat{\varepsilon}^r + \widehat{\rho}^r) \left| \frac{[v]^r}{d^r} \right|^2 dx + b_0 \, \|v\|_0^2 - \frac{1}{2} \int_{\Gamma_N} \boldsymbol{\beta} \cdot \mathbf{n} \, v^2 \, ds, \tag{5.8}$$

and setting $\delta := u_h - w_h$ we have from (4.8), (5.3) and the definitions (3.27) and (4.7) of $\mathcal{L}$ and $\mathcal{L}_s$, respectively

$$|\!|\!| \delta |\!|\!|^2 \leq \mathcal{L}_s(\delta, \delta) = \int_\Omega f \, \delta \, dx - \mathcal{L}(w_h, \delta) - 2 \sum_{e^r \in \mathcal{E}_h^0} \int_{\mathcal{D}^r} \widehat{\rho}^r \frac{[w_h]^r}{d^r} \cdot \frac{[\delta]^r}{d^r} \, dx. \tag{5.9}$$

On the other hand, using (3.27), then (5.6), and finally (5.4), we easily have

$$\begin{aligned}
\mathcal{L}(w_h, \delta) &= \mathcal{L}_{\text{diff}}(w_h, \delta) + \sum_{e^r \in \mathcal{E}_h^0} \int_{e^r} \{w_h\}^r \, \boldsymbol{\beta} \cdot [\delta]^r \, ds + \int_\Omega \gamma \, w_h \, \delta \, dx \\
&= \int_\Omega \widetilde{f} \, \delta \, dx + \int_{\Gamma_N} \widetilde{g}_N \delta \, ds + \sum_{e^r \in \mathcal{E}_h^0} \int_{e^r} \{w_h\}^r \, \boldsymbol{\beta} \cdot [\delta]^r \, ds + \int_\Omega \gamma \, w_h \, \delta \, dx \\
&= \int_\Omega -\text{div}\,(\varepsilon \nabla u) \, \delta \, dx + \int_{\Gamma_N} \boldsymbol{\beta} \cdot \mathbf{n} \, u \delta \, ds + \sum_{e^r \in \mathcal{E}_h^0} \int_{e^r} \{w_h\}^r \, \boldsymbol{\beta} \cdot [\delta]^r \, ds + \int_\Omega \gamma \, w_h \, \delta \, dx,
\end{aligned} \tag{5.10}$$

that using (1.1) becomes

$$\begin{aligned}
\mathcal{L}(w_h, \delta) = &\int_\Omega (f - \text{div}\,(\boldsymbol{\beta} u) - \gamma u) \, \delta \, dx + \int_{\Gamma_N} (\boldsymbol{\beta} \cdot \mathbf{n}) \, u \delta \, ds \\
&+ \sum_{e^r \in \mathcal{E}_h^0} \int_{e^r} \{w_h\}^r \, \boldsymbol{\beta} \cdot [\delta]^r \, ds + \int_\Omega \gamma \, w_h \, \delta \, dx.
\end{aligned} \tag{5.11}$$

This easily gives, integrating by parts the term with the divergence and using the basic property (3.3),

$$\int_\Omega f \, \delta \, dx - \mathcal{L}(w_h, \delta) = \sum_{e^r \in \mathcal{E}_h^0} \int_{e^r} \{u - w_h\}^r \, \boldsymbol{\beta} \cdot [\delta]^r \, ds + \int_\Omega \gamma (u - w_h) \, \delta \, dx. \tag{5.12}$$

Combining (5.9)-(5.12) we get

$$|\!|\!| \delta |\!|\!|^2 \leq \sum_{e^r \in \mathcal{E}_h^0} \int_{e^r} \{u - w_h\}^r \, \boldsymbol{\beta} \cdot [\delta]^r \, ds + \int_\Omega \gamma (u - w_h) \, \delta \, dx - 2 \sum_{e^r \in \mathcal{E}_h^0} \int_{\mathcal{D}^r} \widehat{\rho}^r \frac{[w_h]^r}{d^r} \cdot \frac{[\delta]^r}{d^r} \, dx. \tag{5.13}$$

We shall bound the three terms in the right-hand side of (5.13) separately. For the first term, we easily get

$$\int_{e^r} \{u - w_h\}^r \, \boldsymbol{\beta} \cdot [\delta]^r \, ds \leq \|u - \{w_h\}^r\|_{0,e^r} \, \|\boldsymbol{\beta} \cdot [\delta]^r\|_{0,e^r} \leq \|u - \{w_h\}^r\|_{0,e^r} \, |e^r|^{1/2} \, |\boldsymbol{\beta} \cdot [\delta]^r|. \tag{5.14}$$

We also recall the following trace inequalities, that could be easily deduced from the so-called Agmon inequality (see e.g. [2]) and our assumption (5.1): for all function $\varphi \in \mathrm{H}^1(\Omega)$, for every $v \in V_h$, and for every edge $e^r \in \mathcal{E}_h$

$$\begin{aligned} \|\varphi - \{v\}^r\|_{0,e^r}^2 &\le C\,(|e^r|^{-1}\|\varphi - v\|_{0,\mathcal{D}^r}^2 + |e^r||\varphi|_{1,\mathcal{D}^r}^2), \\ \|\varphi - [v]^r\|_{0,e^r}^2 &\le C\,(|e^r|^{-1}\|\varphi - v\|_{0,\mathcal{D}^r}^2 + |e^r||\varphi|_{1,\mathcal{D}^r}^2). \end{aligned} \tag{5.15}$$

Using $(5.15)_1$, we easily have

$$\|u - \{w_h\}^r\|_{0,e^r}^2 \le C\,(|e^r|^{-1}\|u - w_h\|_{0,\mathcal{D}^r}^2 + |e^r||u|_{1,\mathcal{D}^r}^2), \tag{5.16}$$

while recalling the definition (4.6) of $\widehat{\rho}^r$, and the boundedness of $\boldsymbol{\beta}$ we have

$$|e^r|^{1/2}|\boldsymbol{\beta}\cdot[\delta]^r| \le C\,(|\widehat{\boldsymbol{\beta}}^r|d^r)^{1/2}\frac{|[\delta]^r|}{d^r}(d^r|e^r|)^{1/2} \le C(\widehat{\rho}^r)^{1/2}\frac{|[\delta]^r|}{d^r}|\mathcal{D}^r|^{1/2}. \tag{5.17}$$

Combining (5.14)-(5.17), using (5.2) and (5.7), and recalling the definition (5.8) of the triple-bar norm, we then have

$$\sum_{e^r\in\mathcal{E}_h^0}\int_{e^r}\{u - w_h\}^r\,\boldsymbol{\beta}\cdot[\delta]^r\,ds \le C\,h^{1/2}\,\|u\|_{2,\Omega}\,|\!|\!|\delta|\!|\!|, \tag{5.18}$$

that bounds the first term in the right-hand side of (5.13). The second term is easy. We immediately get

$$\int_\Omega \gamma(u - w_h)\,\delta\,dx \le C\,h\,\|\gamma\|_{L^\infty(\Omega)}\,\|u\|_{2,\Omega}\|\delta\|_{0,\Omega}. \tag{5.19}$$

We are left with the last term. For this we first have easily

$$-2\sum_{e^r\in\mathcal{E}_h^0}\int_{\mathcal{D}^r}\widehat{\rho}^r\frac{[w_h]^r}{d^r}\cdot\frac{[\delta]^r}{d^r}\,dx \le 2\Big(\sum_{e^r\in\mathcal{E}_h^0}\int_{\mathcal{D}^r}\widehat{\rho}^r\left|\frac{[w_h]^r}{d^r}\right|^2dx\Big)^{1/2}|\!|\!|\delta|\!|\!|. \tag{5.20}$$

Then we estimate the term containing w_h. Recalling again that $2|\mathcal{D}^r| = d^r|e^r|$, and the definition (4.6) of $\widehat{\rho}^r$, we have first

$$2\sum_{e^r\in\mathcal{E}_h^0}\int_{\mathcal{D}^r}\widehat{\rho}^r\left|\frac{[w_h]^r}{d^r}\right|^2dx = \sum_{e^r\in\mathcal{E}_h^0}|e^r|d^r\widehat{\theta}^r|\widehat{\boldsymbol{\beta}}^r|d^r\left|\frac{[w_h]^r}{d^r}\right|^2 = \sum_{e^r\in\mathcal{E}_h^0}\widehat{\theta}^r|\widehat{\boldsymbol{\beta}}^r|\,\|[w_h]^r\|_{0,e^r}^2. \tag{5.21}$$

As $\widehat{\theta}^r|\widehat{\boldsymbol{\beta}}^r|$ is easily bounded from above, we just deal with the L^2 norm of the jumps of w_h, that actually coincide with the jumps of $u - w_h$, since u is continuous. By $(5.15)_2$, we have

$$\|[w_h]^r\|_{0,e^r}^2 \equiv \|[u - w_h]^r\|_{0,e^r}^2 \le C\,(|e^r|^{-1}\|u - w_h\|_{0,\mathcal{D}^r}^2 + |e^r|\,|u|_{1,\mathcal{D}^r}^2). \tag{5.22}$$

Inserting (5.22) into (5.21) and using (5.2) and (5.7), we then have

$$2\sum_{e^r\in\mathcal{E}_h^0}\int_{\mathcal{D}^r}\widehat{\rho}^r\left|\frac{[w_h]^r}{d^r}\right|^2dx \le C\,h\,\|u\|_{2,\Omega}^2, \tag{5.23}$$

so that in the end, inserting (5.23) into (5.20), the third term can be bounded as follows:

$$-2 \sum_{e^r \in \mathcal{E}_h^0} \int_{\mathcal{D}^r} \widehat{\rho}^r \frac{[w_h]^r}{d^r} \cdot \frac{[\delta]^r}{d^r} \, dx \leq C \, h^{1/2} \|u\|_{2,\Omega} |\!|\!| \delta |\!|\!|. \tag{5.24}$$

Collecting the three estimates (5.18), (5.19), and (5.24) and inserting them into (5.13), we have

$$|\!|\!| \delta |\!|\!|^2 \leq C \, h^{1/2} \|u\|_{2,\Omega} |\!|\!| \delta |\!|\!| + C \, h \|u\|_{2,\Omega} \|\delta\|_{0,\Omega}, \tag{5.25}$$

that gives easily

$$|\!|\!| \delta |\!|\!| \leq C \, h^{1/2} \|u\|_{2,\Omega}. \tag{5.26}$$

Using (5.26), (5.7), and the triangle inequality, we finally get the error estimate.

Theorem 5.1 *Let u be the solution of* (1.1)*, and let u_h be the solution of* (5.3)*. Assume moreover that $\{\mathcal{T}_h\}_h$ is a regular sequence of quasi-uniform Delaunay triangulations satisfying* (5.1)*. Then there exists a constant C (depending only on the geometric constants of the sequence $\{\mathcal{T}_h\}_h$, on the maximum norm of γ, $\boldsymbol{\beta}$, and of the derivatives of ε), such that*

$$\|u - u_h\|_{0,\Omega} \leq C \, h^{1/2} \|u\|_{2,\Omega}. \tag{5.27}$$

We notice that the above estimate could be considered as *optimal*, since we are using piecewise constant finite elements for u_h, and the loss of half a power of h is sort of physiological in these types of problems (see e.g. [18, 25, 8, 17, 16] and the references therein). It is not optimal, however, with respect to the norm of u used in the right-hand side of (5.27).

We believe that some improvement could be obtained by estimating *directly* the distance $u_h - u_I$ where u_I is the L^2-projection of u onto the space V_h of piecewise constants. Indeed the *trick* of comparing u_h with w_h avoids a lot of technicalities connected with the use of the numerical integration formula (3.8), but forces the use of the quasi-uniformity assumption that, very likely, is not strictly needed. This alone, however, cannot solve the problem of the use of the H^2–norm of u, and could at most trade it for some combination of the type $\varepsilon\|u\|_2 + \|u\|_1$, that would not improve much the quality of the estimate. The presence of the norm in H^2 seems indeed not avoidable in a scheme based on mixed methods, unless a totally different strategy of proof is employed.

References

[1] T. Arbogast, M.F. Wheeler, I. Yotov, Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences, SIAM J. Numer. Anal. **34**, 828–852 (1997).

[2] D.N. Arnold, An interior penalty finite element method with discontinuous elements, SIAM J. Numer. Anal. **19**, 742–760 (1982).

[3] D.N. Arnold, F. Brezzi, B. Cockburn, L.D. Marini, Unified Analysis of Discontinuous Galerkin Methods for Elliptic Problems, SIAM J. Numer. Anal. **39**, 1749–1779 (2002).

[4] J. Baranger, J.F. Maitre, F. Oudin, Application de la théorie des éléments finis mixtes à l'étude d'une classe de schémas aux volumes différences finis pour les problèmes elliptiques, C. R. Acad. Sci. Paris, **319**, série I 401–404 (1994).

[5] J. Baranger, J.F. Maitre, F. Oudin, Connection between finite volume and mixed finite element methods, M^2AN **30** (4), 445–465 (1996).

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, (1991).

[7] F. Brezzi, L.D. Marini, M. Manzini, P. Pietra, A. Russo, Discontinuous Galerkin approximations for elliptic problems, Numer. Meth. Part. Diff. Eq. **16**, 365–278 (2000).

[8] F. Brezzi, L.D. Marini, E. Süli, Residual-free bubbles for advection-diffusion problems: the general error analysis, Numer. Math. **85**, 31-47 (2000).

[9] F. Brezzi, L.D. Marini, E. Süli, Discontinuous Galerkin methods for first-order hyperbolic problems. NA-04/02, Oxford University Computing Laboratory, 2004.

[10] F. Brezzi, L.D. Marini, S. Micheletti, P. Pietra, R. Sacco, S. Wang, *Discretization of semiconductor device problems (I)*, in Handbook of Numerical Analysis, (W.H.A. Schilders, E.J.W. ter Maten, eds.), Numerical Methods for Electrodynamic Problems, North-Holland, Amsterdam (2003).

[11] Z. Cai, J.E. Jones, S.F. McCormick, T.F. Russell, Control-volume mixed finite element methods, Comput. Geosci. **1**, 289–315 (1997).

[12] Ph.G. Ciarlet, *The finite element method for elliptic problems*, North-Holland, Amsterdam (1978).

[13] B. Delaunay, Sur la sphère vide, Izv. Akad. Nauk. SSSR., Math. and Nat. Sci. Div. **6**, 793–800 (1934).

[14] R.E. Ewing, O. Saevareid, J. Shen, Discretization schemes on triangular grids, Comput. Meth. Appl. Mech. Engrng. **152**, 219–238 (1998).

[15] Y. Haugazeau, P. Lacoste, Condensation de la matrice masse pour les éléments finis mixtes de $H(rot)$, C. R. Acad. Sci. Paris, **316**, série I 509–512 (1993).

[16] P. Houston, Ch. Schwab, E. Süli, Discontinuous hp-finite element methods for advection-diffusion problems, SIAM J. Numer. Anal. (Submitted for publication)

[17] P. Houston, E. Süli, Stabilized hp-finite element approximation of partial differential equations with non-negative characteristic form, Computing **66**, 99-119 (2001).

[18] C. Johnson, U. Nävert, J. Pitkaranta, Finite element methods for linear hyperbolic problems, Comp. Meth. Appl. Mech. Engrg. **45**, 285–312 (1984).

[19] R.D. Lazarov, I.D. Michev, P.S. Vassilevski, Finite volume methods for convection-diffusion problems, SIAM J. Numer. Anal., **33** (1), 31–55 (1996).

[20] L.D. Marini, P. Pietra, An abstract theory for mixed approximations of second order elliptic problems, Mat. Aplic. Comp. **8**, 219–239 (1989).

[21] L.D. Marini, P. Pietra, New mixed finite element schemes for current continuity equations, Compel **9**, 257–268 (1990).

[22] S. Micheletti, R. Sacco, Stabilized mixed finite elements for fluid models in semiconductors, Comput. Visual. Sci. **2**, 139–147 (1999).

[23] S. Micheletti, R. Sacco, F. Saleri, On some mixed finite element methods with numerical integration, SIAM J. Sci. Comput. **23** (1), 245–270 (2001).

[24] P.A. Raviart, J.M. Thomas, A mixed finite element method for second order elliptic problems, in Mathematical Aspects of the Finite Element Method (I. Galligani, E. Magenes, eds.), Lecture Notes in Math., Springer-Verlag, New York, **606**, 292–315 (1977).

[25] H.G. Roos, M. Stynes, L. Tobiska, *Numerical methods for singularly perturbed differential equations*, Springer-Verlag, Berlin Heidelberg, (1996).

[26] R. Sacco, F. Saleri, Mixed finite volume methods for semiconductor device simulation, Numer. Meth. Part. Diff. Eq. **13**, 215–236 (1997).

[27] R. Sacco, F. Saleri, Stabilized mixed finite volume methods for convection-diffusion problems, East-West J. Numer. Math. **4** (5), 291–311 (1997).

[28] D.L. Scharfetter, H.K. Gummel, Large signal analysis of a silicon Read diode oscillator. IEEE Trans. Electron Devices **ED-16**, 64–77 (1969).