Finite element and finite volume discretizations of Drift-Diffusion type fluid models for semiconductors

F. Brezzi[♯,§], L. D. Marini[♯,§], S. Micheletti[♭], P. Pietra[§], R. Sacco[♭], S. Wang[‡]

[♯] Dipartimento di Matematica "F. Casorati"
Università di Pavia - Via Ferrata 1
I-27100 Pavia, Italy
[§] Istituto di Matematica Applicata e
Tecnologie Informatiche - C.N.R.
Via Ferrata 1, I-27100 Pavia, Italy
[♭] MOX - Modeling and Scientific Computing
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
Via Bonardi 9, 20133 Milano, Italy
‡ University of Western Australia
Department of Mathematics and Statistics
Nedlands, Western Australia 6907, Australia

This paper is to appear as a Chapter of *Numerical Methods for Electrodynamic Problems*, W.H.A. Schilders, E.J.W. ter Maten, eds., in Handbook of Numerical Analysis, North-Holland - Elsevier Science Publishers B.V.

Acknowledgements. The first three authors and the fifth author acknowledge partial support from the italian MIUR Project PRIN 2001 "Advanced numerical methods for PDE's in applied sciences".

1 Fluid models for transport in semiconductors

In this section we recall the Drift-Diffusion model, already but quickly addressed in Chapter I, and we describe the initial/boundary conditions which complete the model. We also consider in more detail the Energy-Balance transport model. For a more complete discussion of these issues we suggest the books by Markowich [102], Markowich, Ringhofer and Schmeiser [105], Jerome [82], Jüngel [89] and Selberherr [150].

1.1 The Drift-Diffusion model

It is possible to view classical or semiclassical modelling of transport in semiconductors as a hierarchical structure, sweeping from the Boltzmann Transport Equation (BTE) down to the Drift-Diffusion (DD) model, passing through all the systems derived from the moments of the BTE with respect to increasing powers of the carrier group velocity (cf. Section 4.6 of Chapter I). The Hydrodynamic (HD) model and its inertial limit, that is a type of Energy–Transport (ET) model, are examples, among others, of these intermediate steps. The ET model will be briefly presented in Section 1.4. For the HD model we refer to Chapter V and to the references therein.

The DD model is by far the best understood of the above models. It was introduced by Van Roosbroeck back in 1950 [159] as a conservation system for electron and hole carriers, along with the electric field determined from the Poisson equation in a self-consistent fashion. Other derivations of the model are also possible, starting from the BTE and making use of the diffusion approximation, see e.g. [48, 139, 51]. For a mathematical theory on the subject we refer to [130, 132, 131, 68, 107, 103]. As a general comment, the DD model is a good approximation of the underlying physical phenomena when

- All of the scattering processes are elastic. The spatial variation of the relaxation times and of the band structure of the semiconductor are neglected, i.e., slow variations of the doping profile are assumed.
- Degeneration effects in the approximation of the collision integral are neglected.
- The electric field depends mildly on the spatial position and the magnetic induction vanishes.
- The carrier temperature coincides with the lattice temperature, which is constant. This prevents simulating hot carrier and velocity overshoot phenomena.
- The energy bands are parabolic.

From a mathematical viewpoint, when appropriate scalings are employed for the quantities appearing in the DD system, one finds that the conservation part of the system is convection dominated. In particular, there is a scaling which makes the order of

magnitude of the electron and hole concentrations equal to 1. The convection domination arises from the potential, solving Poisson equation, which can be singularly perturbed. This issue will be addressed in Section 1.3. Major breakthroughs in the solution of the model occurred during the 1960's thanks to the pioneering work by H. K. Gummel and D. L. Scharfetter. Gummel introduced in [72] the system decoupling through a nonlinear Gauss-Seidel iteration, and Scharfetter and Gummel provided an exponential fitting approximation for the continuity equations in 1-d [146] (this scheme had been previously introduced in fluidynamics applications by Allen-Southwell [5]). The nonlinear map (or family of maps) that arises from the Gauss-Seidel iteration is a compact continuous map whose fixed point characterizes the solution of the DD model, and is since then called Gummel's map. It will be the object of Section 2. The exponentially fitted difference scheme proposed in the one-dimensional case in [146] will be the object of multidimensional extension in Sections 4.1, 4.2, 5.1.1 and 6.2.

1.2 The DD system

The geometrical model of a semiconductor device consists of a bounded and simply connected portion Ω of $\mathbb{R}^d (d = 1, 2, 3)$, which is constituted by a semiconductor part, denoted by Ω_S, and, in the case of Metal-Oxide-Semiconductor (MOS) devices, also by one or more subdomains of thin oxide adjacent to Ω_S and whose union we denote by Ω_O. We also denote by $x \in \mathbb{R}^d$ the independent spatial variable and by $t \in (0, t_f)$ the time variable.

The DD model is given by the following set of equations for the electric field $\underline{E}$ $[Vcm^{-1}]$, and the electron and hole current densities $\underline{J}_n, \underline{J}_p$ $[Acm^{-2}]$, respectively:

$$\begin{cases} \operatorname{div}(\epsilon \underline{E}) = \rho & \text{in } Q := \Omega \times (0, t_f), \\ q\dfrac{\partial n}{\partial t} - \operatorname{div} \underline{J}_n = -qR & \text{in } Q_S := \Omega_S \times (0, t_f), \\ q\dfrac{\partial p}{\partial t} + \operatorname{div} \underline{J}_p = -qR & \text{in } Q_S := \Omega_S \times (0, t_f), \end{cases} \tag{1.1}$$

where q $[As]$ is the (positive) electron charge, ϵ $[AsV^{-1}cm^{-1}]$ is the dielectric constant of the materials, and the following constitutive relations hold:

$$\begin{aligned} \underline{E} &= -\underline{\nabla}\psi && \text{in } Q, \\ \rho &= \begin{cases} q(p - n + C(x)) = q(p - n + N_D^+ - N_A^-) & \text{in } Q_S, \\ 0 & \text{in } Q_O. \end{cases} \end{aligned} \tag{1.2}$$

In (1.1)-(1.2) ψ $[V]$ is the electrostatic potential, n, p $[cm^{-3}]$ are the electron and hole concentrations inside the semiconductor ($n|_{\Omega_O} \equiv p|_{\Omega_O} \equiv 0$), ρ $[Ascm^{-3}]$ is the net charge density in the device and is zero inside the oxide, which is assumed to be neutral,

$C(x) = N_D^+ - N_A^-$ $[cm^{-3}]$ is the so-called doping profile, which is assumed to be a given datum of the problem in terms of the ionized donors and acceptors concentrations N_D^+ and N_A^-, respectively. Moreover, the oxide is assumed as a perfect insulator; thus we have $\underline{J}_n|_{\Omega_O} \equiv \underline{J}_p|_{\Omega_O} \equiv 0$.

The next step towards the completion of the DD system is to provide constitutive relations for $\underline{J}_n, \underline{J}_p$ in the semiconductor and for the source term R $[cm^{-3}s^{-1}]$. This last one can be interpreted as the net generation/recombination rate of carriers in unit time and volume: $R > 0$ means net recombination, $R < 0$ means net generation. More details about the modeling of R will be given in Section 1.2.1.

Physically, the current density is the product of the elementary charge q, the carrier density, and the mean velocity (drift velocity), i.e.,

$$\underline{J}_n = -qn\underline{v}_n, \quad \underline{J}_p = qp\underline{v}_p,$$

where $\underline{v}_n, \underline{v}_p$ $[cms^{-1}]$ are the drift velocities of the carriers. We have then to determine some relations which link these velocities to the electric field and to the carrier concentrations. This can be done from the BTE. Another approach consists in carrying out a singular perturbation expansion of the current densities with respect to relaxation times. Actually, introducing the mobilities μ_n, μ_p $[cm^2V^{-1}s^{-1}]$

$$\mu_n = q\tau_n/m_n^*, \quad \mu_p = q\tau_p/m_p^*,$$

in terms of the effective masses m_n^*, m_p^* $[kg]$ and of the relaxation times τ_n, τ_p $[s]$, we obtain (see [150])

$$\begin{aligned}
\tau_n \frac{\partial \underline{J}_n}{\partial t} + \underline{J}_n &= q\mu_n n \left(\underline{E} + \frac{1}{n} \underline{\nabla} \left(nK_BT/q \right) \right), \\
\tau_p \frac{\partial \underline{J}_p}{\partial t} + \underline{J}_p &= q\mu_p p \left(\underline{E} - \frac{1}{p} \underline{\nabla} \left(pK_BT/q \right) \right),
\end{aligned}$$

where K_B $[VAsK^{-1}]$ is the Boltzmann constant and T $[K]$ denotes the lattice temperature. Given that the relaxation times are "very small" we can proceed with an expansion where the singular perturbation parameters are the relaxation times

$$\underline{J}_n(\tau_n) = \sum_{j=0}^{\infty} \underline{J}_{nj} \tau_n^j, \quad \underline{J}_p(\tau_p) = \sum_{j=0}^{\infty} \underline{J}_{pj} \tau_p^j.$$

Truncating at the first term ($j = 0$), assuming that the temperature T of the crystal is constant, and assuming Einstein's relations

$$D_n = \mu_n \frac{K_BT}{q}, \quad D_p = \mu_p \frac{K_BT}{q}, \tag{1.3}$$

for the diffusion coefficients of the carriers D_n, D_p $[cm^2 s^{-1}]$, we obtain eventually the classical DD relation for $\underline{J}_n, \underline{J}_p$

$$\begin{aligned} \underline{J}_n = q\mu_n n\underline{E} + qD_n \underline{\nabla} n &= -q\mu_n n \underline{\nabla}\psi + qD_n \underline{\nabla} n, \\ \underline{J}_p = q\mu_p p\underline{E} - qD_p \underline{\nabla} p &= -(q\mu_p p \underline{\nabla}\psi + qD_p \underline{\nabla} p), \end{aligned} \tag{1.4}$$

in which we can recognize two different contributions to the current densities: a drift term proportional to $n\underline{E}$ ($p\underline{E}$), and a diffusion term proportional to $\underline{\nabla} n$ ($\underline{\nabla} p$).
Let us recall the Maxwell-Boltzmann statistics relating the carrier concentrations to the electrostatic potential and the quasi-Fermi levels φ_n, φ_p

$$n = n_i \exp\left(\frac{\psi - \varphi_n}{V_{th}}\right), \quad p = n_i \exp\left(\frac{\varphi_p - \psi}{V_{th}}\right), \tag{1.5}$$

where $V_{th} = K_B T/q$ is the thermal voltage and n_i is the intrinsic concentration of the semiconductor (see, e.g., [157]). Substituting relations (1.5) into (1.4) we obtain the following alternative expression of the current densities:

$$\underline{J}_n = -q\mu_n n \underline{\nabla}\varphi_n, \qquad \underline{J}_p = -q\mu_p p \underline{\nabla}\varphi_p. \tag{1.6}$$

These can be interpreted as two drift currents where electron and holes are driven by the effective fields

$$\underline{E}_n = -\underline{\nabla}\varphi_n, \qquad \underline{E}_p = -\underline{\nabla}\varphi_p.$$

Looking back at (1.5), a new set of independent variables can be derived by setting

$$\rho_n = \exp\left(-\frac{\varphi_n}{V_{th}}\right), \qquad \rho_p = \exp\left(\frac{\varphi_p}{V_{th}}\right). \tag{1.7}$$

The new unknowns ρ_n and ρ_p have been first introduced by Slotboom in [154], and therefore are usually denoted as the *Slotboom* variables associated with electron and hole densities n and p. Substituting (1.7) into (1.5) we get

$$n = n_i \rho_n \exp\left(\frac{\psi}{V_{th}}\right), \qquad p = n_i \rho_p \exp\left(-\frac{\psi}{V_{th}}\right), \tag{1.8}$$

and the current densities become

$$\underline{J}_n = q\, D_n\, n_i\, \exp\left(\frac{\psi}{V_{th}}\right) \underline{\nabla}\rho_n, \qquad \underline{J}_p = -q\, D_p\, n_i\, \exp\left(-\frac{\psi}{V_{th}}\right) \underline{\nabla}\rho_p. \tag{1.9}$$

These can be interpreted as two diffusion currents of the equivalent concentrations $n_i\rho_n$ and $n_i\rho_p$ with diffusion coefficients $D_n \exp\left(\frac{\psi}{V_{th}}\right)$ and $D_p \exp\left(-\frac{\psi}{V_{th}}\right)$, respectively.

Three sets of dependent variables have been introduced so far: the primitive variables (ψ, n, p), the set including the Slotboom variables (ψ, ρ_n, ρ_p) and the set comprising the potentials (ψ, φ_n, φ_p). The first set is the most widely used in computations, although the three unknowns have different physical meaning and attain strongly varying numerical ranges. The second one is useful for analytical purposes, as will be seen in Section 2, since the current continuity equations $(1.1)_2$-$(1.1)_3$ become self-adjoint. However, the set (ψ, ρ_n, ρ_p) can be used in numerical simulation only for low-bias applications due to the enormous dynamic range required by the evaluation on the computer of the Slotboom variables. Compared with the two previous sets of variables, the set (ψ, φ_n, φ_p) has the advantage of collecting physically homogeneous quantities which have the same order of magnitude, at the price of introducing an exponentially nonlinear diffusion coefficient in the current continuity equations. Moreover, strictly positive concentrations are a priori guaranteed due to (1.5). For further details on a comparison between the various sets of dependent variables (including other possible choices that have not been addressed here) we refer to [150], Section 5.2 and to [128].

1.2.1 Boundary conditions and physical modeling

In this section we provide the DD model with appropriate boundary conditions in the case of steady state problems. Let us assume that $\Omega \equiv \Omega_S \cup \Omega_O \subset \mathbb{R}^d$ $(d \leq 2)$ is an open bounded set. A two-dimensional example of domain for semiconductor simulation is shown in Figure 1 where the cross-section of a MOS transistor is schematically represented.

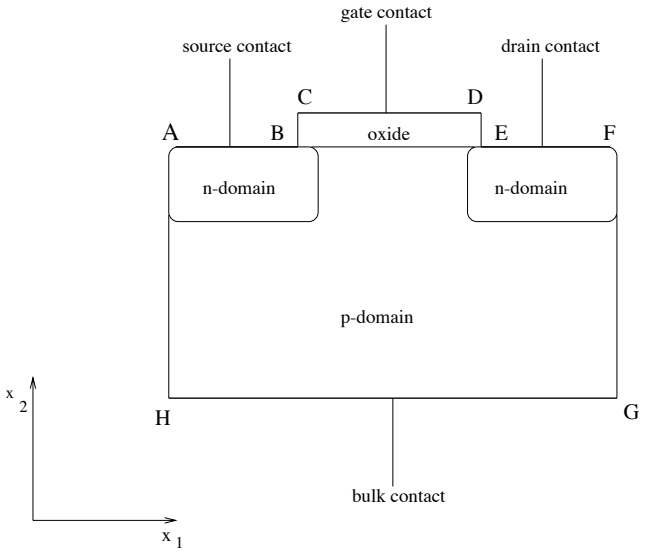


Figure 1: Cross-section of a MOS transistor.

The boundary of a semiconductor device can be subdivided in two disjoint parts $\partial\Omega_P$ and $\partial\Omega_A$, respectively. The first one is a physical boundary, e.g., the metal contacts where the external potentials are applied. The second one is an artificial boundary which separates several devices on the same chip by symmetry axes or internal interfaces, or adjacent parts of a same device but with very different properties, e.g., the oxide-semiconductor interfaces. Both parts of the device boundary can be subdivided into disjoint portions pertaining to the semiconductor and to the oxide respectively, so that we have $\partial\Omega_P \equiv \partial\Omega_{P,S} \cup \partial\Omega_{P,O}$ and $\partial\Omega_A \equiv \partial\Omega_{A,S} \cup \partial\Omega_{A,O} \cup \partial\Omega_I$, where $\partial\Omega_I$ denotes the oxide-semiconductor interface. In the example of Figure 1 we have $\partial\Omega_{P,S} \equiv \overline{AB} \cup \overline{EF} \cup \overline{GH}$, $\partial\Omega_{P,O} \equiv \overline{CD}$, $\partial\Omega_{A,S} \equiv \overline{AH} \cup \overline{FG}$, $\partial\Omega_{A,O} \equiv \overline{BC} \cup \overline{DE}$ and $\partial\Omega_I \equiv \overline{BE}$. We shall consider Dirichlet-Neumann boundary conditions, typically nonhomogeneous Dirichlet conditions for (ψ, n, p) on the segments $\partial\Omega_P$ (ideal ohmic contacts) while on the remaining parts we shall let the fluxes of the electric field and of the current densities vanish, i.e., homogeneous Neumann conditions for (ψ, n, p). For a more exhaustive survey of boundary conditions, see [102, 150, 122].

We start by describing the boundary conditions on $\partial\Omega_P$ and consider first the part $\partial\Omega_{P,S}$. This is made by ohmic contacts where external voltages are applied to electrically drive the semiconductor device. From a mathematical viewpoint, an ideal ohmic contact is a Dirichlet segment where thermodynamic equilibrium is assumed, i.e., the mass-action law

$$np_{|\partial\Omega_{P,S}} = n_i^2, \tag{1.10}$$

and the charge neutrality

$$\rho_{|\partial\Omega_{P,S}} = q(p - n + C)_{|\partial\Omega_{P,S}} = 0 \tag{1.11}$$

hold. By combining (1.10) and (1.11) and denoting by C_{oh} the restriction of the doping profile to the ohmic contact we obtain the following values n_D and p_D for the concentrations of the carriers

$$\begin{aligned} n_{|\partial\Omega_{P,S}} = n_D &:= \frac{\sqrt{C_{oh}^2 + 4n_i^2} + C_{oh}}{2}, \\ p_{|\partial\Omega_{P,S}} = p_D &:= \frac{\sqrt{C_{oh}^2 + 4n_i^2} - C_{oh}}{2}. \end{aligned} \tag{1.12}$$

Concerning the boundary conditions for the electrostatic potential ψ, the value of ψ at the contacts is obtained by summing the external applied voltage V_{ext} with the so-called built-in potential between the metal and the semiconductor

$$\psi_{|\partial\Omega_{P,S}} = \psi_{D|\partial\Omega_{P,S}} := V_{ext|\partial\Omega_{P,S}} + \psi_{bi|\partial\Omega_{P,S}}. \tag{1.13}$$

The built-in potential ψ_{bi} is computed in order that the semiconductor is in thermal equilibrium when all the applied external voltages are zero. This amounts to assuming that the energy levels of the semiconductor in the neighborhood of the contact are horizontal (flat band approximation).

From the Maxwell-Boltzmann statistics (1.5) we obtain the following relations for the built-in potential (see (1.11))

$$n_i \exp\left(-\frac{\psi_{bi}}{V_{th}}\right) - n_i \exp\left(\frac{\psi_{bi}}{V_{th}}\right) + C_{oh} = 0,$$

from which it follows

$$\psi_{bi|\partial\Omega_{P,S}} = V_{th} \sinh^{-1}\left(\frac{C_{oh}}{2n_i}\right) = V_{th} \sinh^{-1}\left(\frac{N_D^+ - N_A^-}{2n_i}|_{\partial\Omega_{P,S}}\right). \tag{1.14}$$

In the case when one of the two dopant species is dominant over the other, (1.14) simplifies into

$$\psi_{bi|\partial\Omega_{P,S}} \simeq V_{th} \ln\left(\frac{N_D^+|_{\partial\Omega_{P,S}}}{n_i}\right) \qquad N_D^+ \gg N_A^-,$$

$$\psi_{bi|\partial\Omega_{P,S}} \simeq -V_{th} \ln\left(\frac{N_A^-|_{\partial\Omega_{P,S}}}{n_i}\right) \qquad N_A^- \gg N_D^+.$$

Let us now consider the boundary conditions on the part $\partial\Omega_{P,O}$. This is made by gate contacts which are located over the oxide region Ω_O and where external voltages are applied to control the current flow between the input-output contacts of the MOS transistor. Since $n = p = 0$ within the oxide, only a boundary condition for the electrostatic potential ψ can be prescribed on $\partial\Omega_{P,O}$ and reads

$$\psi_{|\partial\Omega_{P,O}} = \psi_{D|\partial\Omega_{P,O}} := V_{ext|\partial\Omega_{P,O}} - \Phi_{ms}, \tag{1.15}$$

$\Phi_{ms} = \Phi_m - \Phi_s$ being the metal-semiconductor work function difference, referred to an intrinsic semiconductor, i.e., such that

$$\Phi_s = \chi + \frac{E_c - E_i}{q},$$

χ being the semiconductor affinity, E_c the conduction-band edge, and E_i the intrinsic Fermi level within the semiconductor (see [13]).

As for the Neumann data, we assume homogeneous conditions on $\partial\Omega_{A,S} \cup \partial\Omega_{A,O}$, i.e., vanishing fluxes of the electric field and of the current densities

$$\frac{\partial\psi}{\partial n}|_{\partial\Omega_{A,S}} \equiv \frac{\partial\psi}{\partial n}|_{\partial\Omega_{A,O}} \equiv \underline{J}_n \cdot \underline{n}|_{\partial\Omega_{A,S}} \equiv \underline{J}_p \cdot \underline{n}|_{\partial\Omega_{A,S}} \equiv 0, \tag{1.16}$$

where $\underline{n}$ is the unit outward normal vector to the boundaries.

The so-called material interfaces deserve a special treatment. We denote them by $\partial\Omega_I$, e.g., the interface between the oxide and the semiconductor, which in the example of Figure 1 is $\partial\Omega_I \equiv \overline{BE}$. In this case the Neumann condition for the electrical potential

is a consequence of the Gauss law. Assuming for simplicity that the surface charge density is zero and letting $[\cdot]$ denote the jump function, and ϵ_{ox}, ϵ_{sem} the dielectric constants of the oxide and of the semiconductor, respectively, we have

$$[\psi]_{\partial\Omega_I} = 0, \quad \left[\epsilon \frac{\partial \psi}{\partial n}\right]_{\partial\Omega_I} = 0, \quad \epsilon(x) \equiv \begin{cases} \epsilon_{sem} & x \in \Omega_S \\ \epsilon_{ox} & x \in \Omega_O. \end{cases} \tag{1.17}$$

In order to complete the description of system (1.1), we provide some constitutive relations for the function R and the mobilities μ_n, μ_p. We recall that in thermal equilibrium the mass-action law (1.10) holds. When the device works under nonequilibrium conditions, e.g., when some external voltages are applied, the carrier concentrations move away from their equilibrium values. In this case some recombination/generation mechanisms arise in order to bring the system back to equilibrium. The mechanisms typically considered in the simulations are the Shockley-Hall-Read, Auger, and Impact Ionization phenomena (for a physical description, see, e.g., [157]). The mathematical expressions of these mechanisms are

$$R_{SHR} = \frac{pn - n_i^2}{\tau_n^*(p + n_i) + \tau_p^*(n + n_i)}, \tag{1.18}$$

$$R_{AU} = (pn - n_i^2)(C_n n + C_p p), \tag{1.19}$$

$$R_{II} = -(\alpha_n(|\underline{E}|)|\underline{J}_n|/q + \alpha_p(|\underline{E}|)|\underline{J}_p|/q), \tag{1.20}$$

where τ_n^*, τ_p^* are the carrier lifetimes, C_n, C_p are the Auger coefficients, and α_n, α_p are the ionization rates for electrons and holes, respectively. The recombination/generation rate is eventually computed as

$$R = R_{SHR} + R_{AU} + R_{II}.$$

As for the carrier mobilities, the modeling has to take into account that, for high electric field, the drift velocities ($|v_{n,p}| = \mu_{n,p}|\underline{\nabla}\,\psi|$) saturate. A commonly employed model is the one suggested by Caughey and Thomas [36]

$$\mu_{n,p} = \left(\frac{\mu_{n,p}^0}{\left(1 + \left(\frac{\mu_{n,p}^0 |\underline{\nabla}\,\psi|}{v_{n,p}^{sat}}\right)^{\beta_{n,p}}\right)} \right)^{\frac{1}{\beta_{n,p}}}, \tag{1.21}$$

where $v_{n,p}^{sat}$ are the carrier saturation velocities, $\beta_n = 2$, $\beta_p = 1$ and $\mu_{n,p}^0$ are the low-field mobilities. For the details about the modeling of all the above coefficients we refer to [150], Chapter 4.

The analysis of the multi-dimensional boundary-value problem in steady-state is not trivial because of the presence of mixed boundary conditions. Actually, this is a

situation in which gradient singularities may occur at the boundary transition points. This issue has been addressed in the literature for the case of linear elliptic equations (see [166, 92, 11, 70] for asymptotic expansions and [126] for gradient integrability). A study of asymptotic behavior in the two-dimensional case has been carried out in [62], where the case of a singularity at a corner formed by the oxide region of a MOS device is also considered. As for the transient model, some developments can be found in [46, 81].

The first mathematical study of the DD model is due to Mock [121], who dealt with the steady-state case. He used a decoupling map acting on the potential (the quasi-Fermi levels are computed as an intermediate step). Since then, many variants of such an approach have been introduced in the literature, all defined by an appropriate decoupling procedure, and they are referred to as Gummel fixed point maps. Complicate geometries or parameter models affect the structure of the fixed point map. An example of such map will be discussed in Section 2. We refer the reader to the literature for existence results [149, 122, 101, 61, 80, 102]. Global uniqueness of the solution of the DD system cannot be expected in the general case, since there are devices, such as thyristors, whose performance is based explicitly on the existence of multiple steady–state solutions [157]. However, there are uniqueness results close to thermal equilibrium [122, 102]. For the analysis of the transient problem we refer to [122].

1.3 Scaling

The physical quantities in system (1.1) have different physical dimensions and, in order to compare their orders of magnitude, these quantities have to be made dimensionless first by appropriate scalings. We introduce for system (1.1) two closely related scalings, and we shall refer to these scalings as the De Mari and the Unit scalings.

1. De Mari scaling (see [54])
   - Potentials scaled by V_{th}
   - Concentrations scaled by the intrinsic concentration n_i
   - Length scaled by a characteristic Debye length $L_D = \sqrt{\epsilon V_{th}/(qn_i)}$
2. Unit scaling (see [150] and [102])
   - Potentials scaled by V_{th}
   - Concentrations scaled by $\overline{C} = \sup_{x\in\Omega} |C(x)|$
   - Length scaled by a characteristic device dimension l

After any of the above scalings, the scaled dimensionless DD system reads

$$\begin{cases} \operatorname{div}(\lambda^2 \underline{E}) = p - n + C(x), \\ \dfrac{\partial n}{\partial t} - \operatorname{div} \underline{J}_n = -R, \\ \dfrac{\partial p}{\partial t} + \operatorname{div} \underline{J}_p = -R, \\ \underline{E} = -\underline{\nabla}\,\psi, \\ \underline{J}_n = \mu_n(\underline{\nabla}\, n - n\underline{\nabla}\,\psi), \\ \underline{J}_p = -\mu_p(\underline{\nabla}\, p + p\underline{\nabla}\,\psi), \end{cases} \tag{1.22}$$

where for simplicity we used the same unscaled symbols for the variables, the mobility coefficients and the recombination/generation term. For either scalings we have

$$\lambda^2 = \frac{\epsilon V_{th}}{l_{scal}^2 q C_{scal}}, \qquad l_{scal} = L_D \text{ or } l\,, \qquad C_{scal} = n_i \text{ or } \overline{C}. \tag{1.23}$$

The entire list of scaling factors used to deduce (1.22) is reported in Table 1, where $\overline{\mu}$ denotes the maximum mobility. Moreover, numerical values of the physical quantities in the case of Silicon at 300 K are given.

In the case of the De Mari scaling $\lambda^2 = 1$, whereas in the case of the Unit scaling, $\lambda^2 \simeq 10^{-1} \div 10^{-7}$ (for instance, with the choices of Table 1 we have $\lambda^2 \simeq 10^{-6}$). With the De Mari scaling the doping is of the order $10^7 \div 10^{10}$ and the carrier concentrations are expected to have these values too. In the Unit scaling all the concentrations are expected to be maximally of order 1. The scaled Debye length λ acts as a singular perturbation parameter and the behavior of the solution of (1.22) as $\lambda \longrightarrow 0^+$ (quasi-neutral limit) can be analyzed. We refer to [101, 3, 104, 77, 106] for studies in the stationary case and to [102, 105] for an overview. The study of the quasi-neutral limit in the transient case is more intricate and still a subject of active research (see [136, 137, 64, 65], e.g.).

A different form of scaling is more appropriate for pn–junctions (which are the boundaries between p–regions and n–regions) under extreme reverse biasing conditions. In that case a so–called depletion region forms about the junction, where very few carriers exist and $n, p \simeq 0$ holds. By changing the scaling factor of the potential to $ql^2\overline{C}/\epsilon$, the singular perturbation parameter (still the scaled Debye length λ) appears in the current continuity equations, rather than in the Poisson equation, and in the depletion region $\Delta\psi \simeq -C$ holds. Consequently, the depletion region does not disappear in the limit when $\lambda \to 0$. Actually, the limiting problem is a free boundary problem where the free boundaries coincide with the edges of the depletion region. We refer to [28, 25, 34, 147, 148, 125] and to [105] for an overview.

Quantity	**De Mari factor**	**De Mari factor value**	**Unit factor**	**Unit factor value**
ψ	V_{th}	$0.0258\ V$	V_{th}	$0.0258\ V$
n, p	n_i	$1.482 \times 10^{10}\ cm^{-3}$	$\overline{C}$	$10^{18}\ cm^{-3}$
x	L_D	$3.357 \times 10^{-3}\ cm$	l	$10^{-4}\ cm$
μ_n, μ_p	$D_0 V_{th}^{-1}$	$38.68\ cm^2 V^{-1} s^{-1}$	$\overline{\mu}$	$1000\ cm^2 V^{-1} s^{-1}$
D_n, D_p	D_0	$1\ cm^2 s^{-1}$	$\overline{\mu} V_{th}$	$25.8\ cm^2 s^{-1}$
$\underline{J}_n, \underline{J}_p$	$qD_0 n_i / L_D$	$0.71 \times 10^{-6}\ A cm^{-2}$	$qV_{th}\overline{\mu}\overline{C}/l$	$4.13 \times 10^6 A cm^{-2}$
R	$D_0 n_i / L_D^2$	$1.314 \times 10^{15}\ cm^{-3} s^{-1}$	$V_{th}\overline{\mu}\overline{C}/l^2$	$2.58 \times 10^{29}\ cm^{-3} s^{-1}$
t	L_D^2 / D_0	$1.127 \times 10^{-5}\ s$	$l^2/(V_{th}\overline{\mu})$	$4 \times 10^{-10}\ s$

Table 1: De Mari and Unit scaling factors.

1.4 The Energy–Transport Model

As already pointed out in Section 1.1, the DD model provides a good description of the electrical behavior of the semiconductor devices only close to thermal equilibrium, but it is not accurate enough for sub-micron device modeling, owing to the rapidly changing fields and temperature effects. In the last years, various models, either of kinetic or of macroscopic type, have been derived in order to improve the physical description of the transport in semiconductor devices. The semiconductor Boltzmann equation gives quite accurate simulation results, but the numerical methods to solve this equation (for instance, with Monte-Carlo methods) are too costly and time consuming to model real problems in semiconductor production mode where simulation results are needed in hours or minutes. Extended drift-diffusion models able to describe temperature effects in sub-micron devices are the so-called Energy-Transport (or Energy-Balance) models. They consist of the conservation laws of mass and energy, together with constitutive relations for the particle and energy currents.

The first Energy-Transport model has been presented by Stratton in 1962 [156]. In the physical literature, Energy-Transport equations have been derived from Hydrody-

namic (HD) models usually by neglecting the inertia terms in the momentum transport equation (see, e.g., [141], [155] and references therein). This approach can be made mathematically rigorous by considering a diffusion time scaling [66]. For HD models, which can be regarded as the Euler equations of gas and fluid dynamics for a gas of charged and colliding particles in an electric field, we refer to Chapter V where they are extensively discussed.

Another approach is to derive the Energy-Transport model from the semiconductor Boltzmann equation in the diffusive limit, by means of the Hilbert expansion method [20], [21]. In this derivation, the dominant scattering mechanisms are assumed to be electron-electron and elastic electron–phonon scattering. The Energy-Transport model for electrons reads as follows:

$$\begin{cases} \operatorname{div}(\epsilon \underline{E}) = q(C-n), \\ q\dfrac{\partial n}{\partial t} - \operatorname{div} \underline{J}_n = 0, \\ \dfrac{\partial U}{\partial t} - \operatorname{div} \underline{S}_n = \underline{E}\cdot \underline{J}_n + W(n,T_n), \\ \underline{E} = -\underline{\nabla}\,\psi, \\ \underline{J}_n = L_{11}\left(\dfrac{\underline{\nabla} n}{n} - \dfrac{q\underline{\nabla}\psi}{K_B T_n}\right) + \left(\dfrac{L_{12}}{K_B T_n} - \dfrac{3}{2}L_{11}\right)\dfrac{\underline{\nabla} T_n}{T_n}, \\ q\underline{S}_n = L_{21}\left(\dfrac{\underline{\nabla} n}{n} - \dfrac{q\underline{\nabla}\psi}{K_B T_n}\right) + \left(\dfrac{L_{22}}{K_B T_n} - \dfrac{3}{2}L_{21}\right)\dfrac{\underline{\nabla} T_n}{T_n}. \end{cases} \tag{1.24}$$

The variables are the electron density n, the electron temperature T_n, and the electrostatic potential ψ. Furthermore, $U = \frac{3}{2}nK_BT_n$ is the internal energy, $\underline{J}_n$, $\underline{S}_n$ are the particle current and energy flux densities, respectively. The diffusion coefficients $L_{ij} = L_{ij}(n,T_n)$ and the energy relaxation term $W(n,T_n)$ are non–linear functions of n and T_n, depending on the physical assumptions (distribution function *ansatz*, energy band diagram and so on). For instance, if we assume the parabolic band approximation, non-degenerate Boltzmann statistics (the distribution function being a Maxwellian) and a special form of momentum relaxation time, the diffusion matrix takes the form

$$L = (L_{ij}) = \mu_n K_B T^{eq} n \begin{pmatrix} 1 & \frac{3}{2}K_BT_n \\ \frac{3}{2}K_BT_n & \frac{15}{4}(K_BT_n)^2 \end{pmatrix}, \tag{1.25}$$

and the energy relaxation term is given by

$$W(n,T_n) = -\frac{3}{2}nK_B(T_n - T^{eq})/\tau_{w_n}, \tag{1.26}$$

where T^{eq} denotes the lattice temperature (assumed to be constant) and τ_{w_n} denotes the energy relaxation time. This corresponds to the model introduced in [37]. More

general diffusion coefficients, including non–parabolic energy band cases, are computed in [20, 52] and corresponding numerical simulations are shown in Section 7.4.

When the diffusion matrix is taken as

$$L = (L_{ij}) = \mu_n K_B T_n n \begin{pmatrix} 1 & \frac{5}{2} K_B T_n \\ \frac{5}{2} K_B T_n & (\frac{25}{4} + \kappa_0)(K_B T_n)^2 \end{pmatrix}, \tag{1.27}$$

with $\kappa_0 > 0$, system (1.24) corresponds to the model known in the literature as Energy-Balance model. It reads

$$\begin{cases} \text{div } (\epsilon \underline{E}) = q\,(C - n)\,, \\ q\dfrac{\partial n}{\partial t} - \text{ div } \underline{J}_n = 0, \\ \dfrac{\partial U}{\partial t} - \text{ div } \underline{S}_n = \underline{E} \cdot \underline{J}_n - \dfrac{3}{2} n K_B \dfrac{(T_n - T^{eq})}{\tau_{w_n}}, \\ \underline{E} = -\underline{\nabla}\, \psi, \\ \underline{J}_n = q D_n \underline{\nabla}\, n - q \mu_n n \underline{\nabla} \left(\psi - \dfrac{K_B T_n}{q} \right), \\ \underline{S}_n = \kappa_n \underline{\nabla}\, T_n + \dfrac{5}{2} \dfrac{K_B T_n}{q} \underline{J}_n, \end{cases} \tag{1.28}$$

where D_n is the non–constant diffusion coefficient given by $D_n = \mu_n K_B T_n q^{-1}$, and κ_n is the non–constant heat conductivity coefficient corresponding to the Wiedemann–Franz law $\kappa_n = \kappa_0 K_B^2 \mu_n q^{-1} n T_n$.

We complement the equations with physically motivated mixed Dirichlet-Neumann boundary conditions

$$n = n_D, \quad T_n = T_D, \quad \psi = \psi_D \qquad \text{on } \Gamma_D, \tag{1.29}$$

$$\underline{J}_n \cdot \underline{n} = \underline{S}_n \cdot \underline{n} = \underline{\nabla}\, \psi \cdot \underline{n} = 0 \qquad \text{on } \Gamma_N, \tag{1.30}$$

modeling the contacts Γ_D and the insulating boundary parts Γ_N. We have assumed that $\partial\Omega = \Gamma_D \cup \Gamma_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$. The boundary conditions are usually taken as for the DD model: at ohmic contacts, n_D is defined by (1.12), ψ_D is given by (1.13) and $T_D = T^{eq}$. Moreover, initial conditions for n and T_n are prescribed.

We point out that these models reduce to the classical DD model under the simplifying assumption of vanishing relaxation time (i.e. $\tau_{w_n} \to 0$) which amounts to considering the carriers in thermal equilibrium with the crystal lattice, $T_n = T^{eq}$, yielding

$$\underline{J}_n = q D_n \underline{\nabla}\, n - q \mu_n n \underline{\nabla}\, \psi. \tag{1.31}$$

The mathematical analysis of the equations (1.24) has been recently carried out in [49, 50], under the assumption of uniformly bounded diffusion coefficients. The existence and uniqueness of weak solutions to both the stationary and the time-dependent

(initial) boundary-value problems have been proved. For an overview, we refer also to [89]. Existence results under different assumptions (for instance, near-equilibrium situations) have been shown in [4, 69, 86].

2 A nonlinear block iterative solution of the semiconductor device equations: the Gummel map

In this section we introduce the so-called *Gummel map*, a nonlinear block iterative algorithm that splits the DD semiconductor device equations into the successive solution of a nonlinear Poisson equation for the electric potential ψ and two linearized continuity equations for the electron and hole densities n and p.

The basic form of the iterative map was first proposed by Gummel in [72]. Since then, the map has become the most established approach in computer simulations due to its good convergence properties (even in case of a badly chosen initial guess), ease of implementation, and reduced computational effort (compared, for instance, to a fully coupled solution of the DD system using Newton's method).

Due to its successful use in semiconductor device simulation, a considerable amount of mathematical work has been carried out to analyze the convergence of Gummel's iteration (see, e.g., the papers [93, 94] and the books by Markowich [102], Markowich et al. [105] and Jerome [82]). In the following, we review the basic formalism and notation needed for a mathematically consistent presentation of the algorithm.

With this aim, the scaled DD system is formulated in Section 2.1 in terms of the Slotboom variables introduced in Section 1.1. Using these new unknowns, the linearized carrier continuity equations assume a self-adjoint form which is convenient in view of the analysis.

The Sections 2.2 and 2.3 are devoted to the presentation and discussion of the Gummel map as a fixed-point iteration. In Section 2.2 the Gummel map is used as an abstract tool for constructively proving the existence of a (weak) solution to the DD system in the stationary case. In Section 2.3 it is shown how to utilize the Gummel map as an iteration scheme for the approximate computation of the solution of the semiconductor device problem. In particular, we summarize the main theoretical convergence results for the iteration and briefly address the delicate matter of its acceleration, referring for more details on this subject to Chapter 7 of this book.

Finally, in Section 2.4 the differential subproblems involved in the decoupled iterative solution of the DD equations are cast into the unified framework of a reaction-diffusion model problem. This will be the object of the discretization techniques discussed in the forthcoming sections of this chapter.

2.1 The drift-diffusion equations in self-adjoint form

We will use henceforth the Unit scaling introduced in Section 1.3. Under this assumption the scaled Slotboom variables are

$$\rho_n = e^{-\varphi_n}, \qquad \rho_p = e^{\varphi_p}, \tag{2.1}$$

while the scaled Maxwell-Boltzmann statistics and the current densities become

$$n = \delta^2 \rho_n e^{\psi}, \qquad p = \delta^2 \rho_p e^{-\psi}, \tag{2.2}$$

and

$$\underline{J}_n = \delta^2 \, \mu_n \, e^{\psi} \, \underline{\nabla} \, \rho_n, \qquad \underline{J}_p = -\delta^2 \, \mu_p \, e^{-\psi} \, \underline{\nabla} \, \rho_p, \tag{2.3}$$

where $\delta^2 = n_i/\overline{C}$. Notice that in thermal equilibrium conditions (i.e., zero external applied biases) $\varphi_n = \varphi_p = 0$ and correspondingly $\rho_n = \rho_p = 1$ which clearly implies $\underline{J}_n = \underline{J}_p = \underline{0}$. Moreover, by definition ρ_n and ρ_p are strictly positive quantities.

Using the triplet (ψ, ρ_n, ρ_p), and assuming for simplicity that $\Omega_O = \emptyset$, the DD system (1.22), can be written in $Q_S \equiv Q$ as

$$\begin{cases} -\mathrm{div}\,(\lambda^2 \underline{E}) = \delta^2 \rho_p e^{-\psi} - \delta^2 \rho_n e^{\psi} + C(x), \\ \delta^2 \dfrac{\partial(\rho_n e^{\psi})}{\partial t} - \mathrm{div}\,\underline{J}_n = -R, \\ \delta^2 \dfrac{\partial(\rho_p e^{-\psi})}{\partial t} + \mathrm{div}\,\underline{J}_p = -R, \\ \underline{E} = -\underline{\nabla}\,\psi, \\ \underline{J}_n = \delta^2 \mu_n \, e^{\psi} \, \underline{\nabla}\,\rho_n, \\ \underline{J}_p = -\delta^2 \mu_p \, e^{-\psi} \, \underline{\nabla}\,\rho_p, \\ \psi = \psi_D, \quad \rho_n = \rho_{nD}, \quad \rho_p = \rho_{pD} \qquad \text{on } \Gamma_D, \\ \underline{E} \cdot \underline{n} = \underline{\nabla}\,\rho_n \cdot \underline{n} = \underline{\nabla}\,\rho_p \cdot \underline{n} = 0 \qquad \text{on } \Gamma_N, \end{cases} \tag{2.4}$$

where Γ_D and Γ_N denote the Dirichlet and Neumann portions of the boundary, respectively, with $\partial\Omega =: \Gamma = \Gamma_D \cup \Gamma_N$, and the strictly positive Dirichlet boundary data ρ_{nD} and ρ_{pD} are related to the corresponding data for n and p through (1.8) as

$$n_D = \delta^2 \rho_{nD} e^{\psi_D}, \qquad p_D = \delta^2 \rho_{pD} e^{-\psi_D}, \tag{2.5}$$

where n_D, p_D are computed according to (1.12). The advantage of working with the Slotboom variables is that the spatial part of the continuity equations is, for given R, in self-adjoint form, which greatly facilitates the mathematical analysis. On the other hand, for given ρ_n and ρ_p the Poisson equation becomes nonlinear in the unknown ψ, so that a suitable linearization technique (for example, Newton's method) is required to carry out its abstract analysis as well as its discretization.

2.2 The Gummel map in the stationary case

In this section we introduce the Gummel map for the solution of the semiconductor device equations. Without loss of generality, only the stationary case will be considered since a semidiscretization in time (using for instance the Backward Euler method) modifies the continuity equations by the addition of a zero-order term proportional to the reciprocal of the time-step $\triangle t$ in the right and left-hand sides. The presentation follows the guidelines of [102], Chapter 3, and [82], Chapter 4, to which we refer the interested reader for more details and for the proofs of the results. We also refer to Section 3.1 for the definition of the functional spaces that will be used in the sequel.

Assume that there exist two positive constants ρ_{min}, ρ_{max} such that

$$0 < \rho_{min} \leq \rho_n(x)_{|\Gamma_D}, \, \rho_p(x)_{|\Gamma_D} \leq \rho_{max} \qquad \forall x \in \Gamma_D,$$

and let

$$\mathcal{K} = \max\left\{|\log(\rho_{min})|, \, |\log(\rho_{max})|\right\}.$$

Let $(\rho_n^{(0)}, \rho_p^{(0)}) \in (L^\infty(\Omega))^2$ be a given pair of positive essentially bounded functions such that

$$e^{-\mathcal{K}} \leq \rho_n^{(0)}(x), \, \rho_p^{(0)}(x) \leq e^{\mathcal{K}} \qquad a.e. \text{ in } \Omega. \tag{2.6}$$

The Gummel abstract fixed point iteration consists of:

(1) solving the semilinear Poisson equation

$$\begin{cases} -\text{div}\left(\lambda^2 \underline{\nabla}\, \psi\right) = \delta^2 \rho_p^{(0)} e^{-\psi} - \delta^2 \rho_n^{(0)} e^{\psi} + C(x) & \text{in } \Omega, \\ \psi = \psi_D \text{ on } \Gamma_D, \qquad \underline{\nabla}\, \psi \cdot \underline{n} = 0 \text{ on } \Gamma_N, \end{cases} \tag{2.7}$$

for $\psi = \psi^{(1)}$:

(2) solving the two decoupled continuity equations

$$\begin{cases} -\text{div}\left(\mu_n e^{\psi^{(1)}} \underline{\nabla}\, \rho_n\right) = -R(x, \psi^{(1)}, \rho_n, \rho_p^{(0)}) & \text{in } \Omega, \\ \rho_n = \rho_{n\,D} \text{ on } \Gamma_D, \qquad \underline{\nabla}\, \rho_n \cdot \underline{n} = 0 \text{ on } \Gamma_N, \end{cases} \tag{2.8}$$

for $\rho_n = \rho_n^{(1)}$, and

$$\begin{cases} -\text{div}\left(\mu_p e^{-\psi^{(1)}} \underline{\nabla}\, \rho_p\right) = -R(x, \psi^{(1)}, \rho_n^{(1)}, \rho_p) & \text{in } \Omega, \\ \rho_p = \rho_{p\,D} \text{ on } \Gamma_D, \qquad \underline{\nabla}\, \rho_p \cdot \underline{n} = 0 \text{ on } \Gamma_N, \end{cases} \tag{2.9}$$

for $\rho_p = \rho_p^{(1)}$.

Let

$$N = \left\{(u, v) \in (L^2(\Omega))^2 \,|\, e^{-\mathcal{K}} \leq u(x), v(x) \leq e^{\mathcal{K}} \quad a.e. \text{ in } \Omega\right\}.$$

It can be checked that $\rho_n^{(1)}, \rho_p^{(1)} \in N$. Steps **(1)-(2)** implicitly define an operator

$$T : N \to N$$

such that

$$T(\rho_n^{(0)}, \rho_p^{(0)}) = (\rho_n^{(1)}, \rho_p^{(1)}).$$

Under suitable assumptions on the doping profile C, on the mobilities μ_n and μ_p, and on the net recombination rate R (see [102], p. 34), it can be shown that T admits a fixed point $(\rho_n^*, \rho_p^*) \in (H^1(\Omega) \cap L^\infty(\Omega))^2$ with $\rho_n^*, \rho_p^* \in N$. This in turn allows us to prove that there exists a solution $w^* = (\psi^*, \rho_n^*, \rho_p^*) \in (H^1(\Omega) \cap L^\infty(\Omega))^3$ of the DD system (2.4). Explicit bounds for ψ^* can be found in [102]. The decoupling abstract procedure described above is the basic Gummel iteration that is commonly employed in semiconductor device simulation. Its actual implementation is the object of the next section.

2.3 A fixed-point iteration for the approximate solution of the drift-diffusion system

In the following we use the Gummel fixed-point iteration introduced in the previous section to construct a sequence of approximate solutions of the semiconductor DD system (2.4) in the stationary case. For ease of presentation the algorithm will be described "on the continuous level", although it can be applied to discrete schemes or, conversely, the "continuous" iterative scheme can be discretized appropriately (see, e.g., [16] or [82], Chapter 5).

Let $k \geq 0$ be a fixed integer; for given $(\psi^{(k)}, \rho_n^{(k)}, \rho_p^{(k)})$ the Gummel map consists of solving successively the following boundary-value subproblems:

$$\begin{cases} -\mathrm{div}\,(\lambda^2 \underline{\nabla} \psi) = \delta^2 \rho_p^{(k)} e^{-\psi} - \delta^2 \rho_n^{(k)} e^{\psi} + C(x) & \text{in } \Omega, \\ \psi = \psi_D \text{ on } \Gamma_D, \qquad \underline{\nabla} \psi \cdot \underline{n} = 0 \text{ on } \Gamma_N, \end{cases} \tag{2.10}$$

for $\psi = \psi^{(k+1)}$;

$$\begin{cases} -\mathrm{div}\left(\mu_n e^{\psi^{(k+1)}} \underline{\nabla} \rho_n\right) = -R(x, \psi^{(k+1)}, \rho_n, \rho_p^{(k)}) & \text{in } \Omega, \\ \rho_n = \rho_{nD} \text{ on } \Gamma_D, \qquad \underline{\nabla} \rho_n \cdot \underline{n} = 0 \text{ on } \Gamma_N, \end{cases} \tag{2.11}$$

for $\rho_n = \rho_n^{(k+1)}$, and

$$\begin{cases} -\mathrm{div}\left(\mu_p e^{-\psi^{(k+1)}} \underline{\nabla} \rho_p\right) = -R(x, \psi^{(k+1)}, \rho_n^{(k+1)}, \rho_p) & \text{in } \Omega, \\ \rho_p = \rho_{pD} \text{ on } \Gamma_D, \qquad \underline{\nabla} \rho_p \cdot \underline{n} = 0 \text{ on } \Gamma_N, \end{cases} \tag{2.12}$$

for $\rho_p = \rho_p^{(k+1)}$. Steps (2.10)–(2.12) can be simply written as $T(\rho_n^{(k)}, \rho_p^{(k)}) = (\rho_n^{(k+1)}, \rho_p^{(k+1)})$. We prefer to use the triplet $(\psi^{(k)}, \rho_n^{(k)}, \rho_p^{(k)})$, instead of the pair $(\rho_n^{(k)}, \rho_p^{(k)})$, in order to

emphasize the fact that $\psi^{(k)}$ is used as an initial guess for solving the nonlinear problem (2.10).

Looking at the structure of the Gummel map, we recognize a nonlinear block Gauss-Seidel iteration which can be subdivided into two main loops: (i) a nonlinear inner iteration for solving the semilinear Poisson equation (2.10); (ii) two iterations for solving the continuity equations (2.11) and (2.12), which will be suitably linearized in the following. Typically, a damped Newton method is used for dealing with (i) (see [15]); this leads to the following inner nonlinear iteration:

(A) set $u^{(0)} = \psi^{(k)}$;

then, for $j = 0, 1, \ldots$ until convergence execute the following steps **(B)**-**(D)**:

(B) set

$$n^{(j)} = \delta^2 \rho_n^{(k)} e^{u^{(j)}}, \qquad p^{(j)} = \delta^2 \rho_p^{(k)} e^{-u^{(j)}};$$

(C) solve

$$\begin{cases} -\mathrm{div}\,(\lambda^2 \underline{\nabla}\,\phi) + (n^{(j)} + p^{(j)})\phi = -\mathcal{R}_\psi(u^{(j)}) & \text{in } \Omega, \\ \phi = 0 \text{ on } \Gamma_D, \qquad \underline{\nabla}\,\phi \cdot \underline{n} = 0 \text{ on } \Gamma_N, \end{cases} \tag{2.13}$$

(D) and set

$$u^{(j+1)} = u^{(j)} + t_j \phi, \qquad t_j \in (0, 1].$$

The right-hand side in (2.13) is the residual of the Poisson equation at the j-th step of Newton's iteration and is defined as

$$\mathcal{R}_\psi(V) = -\mathrm{div}\,(\lambda^2 \underline{\nabla}\, V) + \delta^2 \rho_n^{(k)} e^V - \delta^2 \rho_p^{(k)} e^{-V} - C(x).$$

The damping coefficients t_j can be chosen as suggested in [15] in such a way to ensure a monotonical reduction of the norm of the residual. As for the solution of step (ii) of Gummel's iteration, suitable exponentially-fitted box-like schemes are usually employed for a stable and accurate discretization (see, e.g., [150], Chapter 6).

A convergence proof of iteration (2.10)-(2.12) has been first given in [93] under the assumption of vanishing recombination/generation rate (i.e., $R = 0$). Using contraction-type arguments, and always assuming $R = 0$, it is possible to prove the convergence of the Gummel map only close to thermal equilibrium, i.e. for small values of the applied external biases ψ_D (see [102], Theorem 3.6.5 and [82], Theorem 4.1.1). Convergence results under less restrictive conditions do not exist, although numerical experiments show that the iteration is always rapidly converging (even if a bad initial guess is chosen) whenever the magnitude of the current densities and of the recombination-generation rate is not too large (see [102], formula (3.6.56)). On the contrary, the convergence of the decoupled iteration becomes very slow in the case of

high injection, or in the presence of impact ionization phenonema. This prompts for devising suitable acceleration techniques that improve the performance of Gummel's map.

This latter issue will be more extensively addressed in Chapter 7. Here we just mention the vectorial acceleration methods proposed in [71] that produce a superlinear asymptotic rate of convergence without excessively increasing the CPU time and the memory resources. A different approach has been pursued in [113] where a variant of Gummel's map based on the use of BI-CGSTAB method has been developed. Acceleration methods based on Newton-Krylov subspace iterations have been recently proposed in [95], and an extension of their techniques to the simulation of optoelectronic semiconductor devices under high voltage operation conditions has been carried out in [23].

2.4 The decoupled drift-diffusion system

Each differential subproblem that has to be solved at each step of the Gummel iterative procedure (2.10)-(2.12) can be cast under the following general form of a reaction-diffusion problem

$$\begin{cases} -\mathrm{div}\,(a\underline{\nabla} u) + \gamma u = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma_D, \\ a\underline{\nabla} u \cdot \underline{n} = 0 & \text{on } \Gamma_N. \end{cases} \tag{2.14}$$

Actually, the linearized Poisson equation that has to be solved within the Newton inner iteration (2.13) can be recovered from (2.14) (after dividing by λ^2) by letting $u = \phi$ and taking

$$a = 1, \quad \gamma = (n^{(j)} + p^{(j)})/\lambda^2, \quad f = -\mathcal{R}_\psi(u^{(j)})/\lambda^2, \quad g = 0.$$

As for the solution of the linearized continuity equations, assume, as in Section 1.2.1, that the recombination-generation term R can be written as

$$R(x, \psi, \rho_n, \rho_p) = F(x, \psi, \rho_n, \rho_p)(\rho_n\rho_p - 1) - G(x, |\nabla\psi|, |\underline{J}_n|, |\underline{J}_p|)$$

for a positive function $F(\cdot,\cdot,\cdot,\cdot)$ in $(\Omega \times \mathbb{R} \times (0,+\infty)^2)$ and a nonnegative function $G(\cdot,\cdot,\cdot,\cdot)$ in $(\Omega \times (0,+\infty)^3)$. The first term at right-hand side models the Shockley-Read-Hall and Auger recombination rates while the second one is a net generation rate modelling impact ionization phenomena (see (1.18)–(1.20)). Namely, $-G$ is the (scaled) term defined in (1.20), and

$$F_{SHR} = \frac{\delta^2}{\left(\tau_n^*(\rho_p e^{-\psi} + 1) + \tau_p^*(\rho_n e^{\psi} + 1)\right)}, \tag{2.15}$$

$$F_{AU} = \delta^2(C_n\rho_n e^{\psi} + C_p\rho_p e^{-\psi}). \tag{2.16}$$

Finally, we have

$$F(x,\psi,\rho_n,\rho_p) = F_{SHR} + F_{AU}.$$

Then we can recover problems (2.11) and (2.12) from (2.14) by setting

$$u = \rho_n, \quad a = \mu_n e^{\psi^{(k+1)}}, \quad \gamma = F(x,\psi^{(k+1)},\rho_n^{(k)},\rho_p^{(k)})\rho_p^{(k)}$$
$$f = F(x,\psi^{(k+1)},\rho_n^{(k)},\rho_p^{(k)}) + G(x,|\nabla\psi^{(k+1)}|,|\underline{J}_n^{(k)}|,|\underline{J}_p^{(k)}|), \quad g = \rho_{nD},$$

and

$$u = \rho_p, \quad a = \mu_p e^{-\psi^{(k+1)}}, \quad \gamma = F(x,\psi^{(k+1)},\rho_n^{(k+1)},\rho_p^{(k)})\rho_n^{(k+1)},$$
$$f = F(x,\psi^{(k+1)},\rho_n^{(k+1)},\rho_p^{(k)}) + G(x,|\nabla\psi^{(k+1)}|,|\underline{J}_n^{(k+1)}|,|\underline{J}_p^{(k)}|), \quad g = \rho_{pD}.$$

As anticipated in Section 1.2, the use of the Slotboom variables in numerical computation is typically restricted to low-bias applications due to possible overflow/underflow problems. For this reason it is convenient to formulate the Gummel map in terms of the variables (ψ, n, p). By doing so, the linearized Poisson equation remains in the form of the model problem (2.14), while the linearized current continuity equations can be cast in the form of linear advection-diffusion-reaction problems, where the advective term is represented by the electric field (see for the details [102], Sect. 3.6). The discretization of these latter problems using mixed methods will be addressed in Sections 4.1 and 6.2. Next section will address the mixed finite element discretization of the reaction-diffusion problem (2.14).

3 Mixed Formulation of Second Order Elliptic Problems

We shall introduce here the mixed formulation of the second order elliptic model problem (2.14). From now on, Ω is assumed to be a convex polygonal domain in $\mathbb{R}^2$, with unitary measure, and f, g are given functions, with $f \in L^2(\Omega)$, and $g \in H^{1/2}(\Gamma_D)$. Moreover, $a = a(x)$ and $\gamma = \gamma(x)$ are given regular functions on $\overline{\Omega}$, bounded from above and below, i.e.,

$$\exists\, a_0,\, a_M \quad \text{such that} \quad a_M \geq a(x) \geq a_0 > 0, \tag{3.1}$$
$$\exists\, \gamma_0,\, \gamma_M \quad \text{such that} \quad \gamma_M \geq \gamma(x) \geq \gamma_0 \geq 0. \tag{3.2}$$

The key point for deriving the mixed formulation of (2.14) is to introduce the flux $\underline{\sigma} = a\underline{\nabla} u$ as an independent variable, so that (2.14) becomes

$$\begin{cases} a^{-1}\underline{\sigma} - \underline{\nabla} u = 0 & \text{in } \Omega, \\ -\mathrm{div}\,\underline{\sigma} + \gamma u = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma_D, \\ \underline{\sigma}\cdot\underline{n} = 0 & \text{on } \Gamma_N. \end{cases} \tag{3.3}$$

When treating the (scaled) Poisson equation (2.13), the flux $\underline{\sigma}$, from the physical point of view, will represent the (scaled) electric displacement. When taking into account the (scaled) continuity equations (2.11) and (2.12) , $\underline{\sigma}$ will represent the (scaled) current density for electrons and holes, respectively.

We have now to introduce the proper functional setting for writing the variational formulation of (3.3). For that, let us first set a few notation. Some of the functional spaces have been previously used without precise definitions.

3.1 Notation

We shall constantly use Sobolev spaces on $\Omega \subset \mathbb{R}^2$, for which we refer to [1], [96], [127]. They are based on

$$L^2(\Omega) := \left\{ v \mid \int_\Omega |v|^2 \, dx = ||v||^2_{L^2(\Omega)} < +\infty \right\}, \tag{3.4}$$

the space of square integrable (and measurable) functions on Ω. The scalar product in $L^2(\Omega)$ is denoted by

$$(u, v) = \int_\Omega u \, v \, dx. \tag{3.5}$$

We then define, for m integer ≥ 0,

$$H^m(\Omega) := \left\{ v \mid D^\alpha v \in L^2(\Omega), \ \forall \, |\alpha| \leq m \right\}, \tag{3.6}$$

where

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}}, \quad |\alpha| = \alpha_1 + \alpha_2, \tag{3.7}$$

these derivatives being taken in the sense of distributions. On this space we shall use the seminorm

$$|v|^2_{m,\Omega} = \sum_{|\alpha|=m} ||D^\alpha v||^2_{L^2(\Omega)}, \tag{3.8}$$

and the norm

$$||v||^2_{m,\Omega} = \sum_{k \leq m} |v|^2_{k,\Omega}. \tag{3.9}$$

The space $L^2(\Omega)$ is then $H^0(\Omega)$, and we shall usually write $||v||_{0,\Omega}$ to denote its norm $||v||_{L^2(\Omega)}$. We shall also use the space of traces on $\Gamma = \partial\Omega$ of functions in $H^1(\Omega)$:

$$H^{1/2}(\Gamma) := (H^1(\Omega))_{|\Gamma}, \tag{3.10}$$

with the norm

$$||g||_{1/2,\Gamma} = \inf_{v \in H^1(\Omega), \ v_{|\Gamma} = g} ||v||_{1,\Omega}. \tag{3.11}$$

The space $H_0^1(\Omega)$ will denote, as usual, the space of functions in $H^1(\Omega)$ vanishing on the boundary:

$$H_0^1(\Omega) := \left\{ v \,|\, v \in H^1(\Omega), \quad v_{|\Gamma} = 0 \right\}. \tag{3.12}$$

For $v \in H_0^1(\Omega)$, or $v \in H^1(\Omega)$ and vanishing on a part of the boundary, the Poincaré inequality holds

$$||v||_{0,\Omega} \leq C(\Omega)|v|_{1,\Omega}, \tag{3.13}$$

and the seminorm $|.|_{1,\Omega}$ is therefore a norm in $H_0^1(\Omega)$, equivalent to the $||.||_{1,\Omega}$ norm. We denote, for $1 \leq p < +\infty$,

$$L^p(\Omega) := \left\{ v \,|\, \int_\Omega |v|^p \, dx =: ||v||^p_{L^p(\Omega)} < +\infty \right\}. \tag{3.14}$$

As usual, $L^\infty(\Omega)$ denotes the space of essentialy bounded functions, with norm

$$||v||_{\infty,\Omega} = sup\ ess_{x\in\Omega}|v(x)|. \tag{3.15}$$

For $1 \leq p \leq +\infty$ let

$$W^{1,p}(\Omega) := \{ v \,|\, D^\alpha v \in L^p(\Omega),\ \forall\, |\alpha| \leq 1 \}, \tag{3.16}$$

equipped with the norm

$$||v||^p_{W^{1,p}(\Omega)} = \sum_{|\alpha|\leq 1} ||D^\alpha v||^p_{L^p(\Omega)}. \tag{3.17}$$

Finally, we denote by $H(\mathrm{div}\,;\Omega)$ the space of vector-valued functions

$$H(\mathrm{div}\,;\Omega) := \left\{ \underline{\tau} \in (L^2(\Omega))^2 \,|\, \mathrm{div}\,\underline{\tau} \in L^2(\Omega) \right\}, \tag{3.18}$$

equipped with the graph norm

$$||\underline{\tau}||^2_{H(\mathrm{div}\,;\Omega)} = ||\underline{\tau}||^2_{0,\Omega} + ||\mathrm{div}\,\underline{\tau}||^2_{0,\Omega}, \tag{3.19}$$

where the symbol $||.||_{0,\Omega}$ denotes as well the L^2-norm for vector valued functions.

3.2 Mixed formulation

We are now ready to write the variational formulation of system (3.3). For that, define the spaces

$$\Sigma = \left\{ \underline{\tau} \in (L^2(\Omega))^2 \,|\, \mathrm{div}\,\underline{\tau} \in L^2(\Omega),\ \underline{\tau}\cdot\underline{n} = 0 \text{ on } \Gamma_N \right\} \subset H(\mathrm{div}\,;\Omega), \tag{3.20}$$

$$V = L^2(\Omega), \tag{3.21}$$

with norms

$$\|\underline{\tau}\|_{\Sigma}^2 = \|\underline{\tau}\|_{H(\mathrm{div};\Omega)}^2, \tag{3.22}$$

$$\|v\|_V = \|v\|_{0,\Omega}. \tag{3.23}$$

Next, we introduce the following bilinear forms

$$a(\underline{\sigma},\underline{\tau}) = \int_\Omega a^{-1}\underline{\sigma}\cdot\underline{\tau}\,dx, \qquad \underline{\sigma},\underline{\tau}\in\Sigma, \tag{3.24}$$

$$b(v,\underline{\tau}) = \int_\Omega v\,\mathrm{div}\,\underline{\tau}\,dx, \qquad v\in V,\underline{\tau}\in\Sigma, \tag{3.25}$$

$$c(u,v) = \int_\Omega \gamma\,u\,v\,dx, \qquad u,v\in V. \tag{3.26}$$

Then, the mixed formulation of (2.14) is

$$\begin{cases} \text{Find } (\underline{\sigma},u)\in\Sigma\times V \text{ such that} \\ a(\underline{\sigma},\underline{\tau})+b(u,\underline{\tau}) = <g,\underline{\tau}\cdot\underline{n}>_{|\Gamma_D} & \forall\,\underline{\tau}\in\Sigma, \\ b(v,\underline{\sigma})-c(u,v) = -(f,v) & \forall\,v\in V, \end{cases} \tag{3.27}$$

where the bracket $<.,.>_{|\Gamma_D}$ denotes the duality between $H^{1/2}(\Gamma_D)$ and its dual space $H^{-1/2}(\Gamma_D)$. Following [24] and [27], in order to prove existence, uniqueness and stability for problem (3.27) the following properties are needed

(i) the bilinear form $a(.,.)$ is bounded on $\Sigma\times\Sigma$ and coercive on $\mathrm{Ker}\,B$, that is,

$$\exists\,M_a>0 \text{ such that } \quad |a(\underline{\sigma},\underline{\tau})|\le M_a\|\underline{\sigma}\|_\Sigma\|\underline{\tau}\|_\Sigma, \qquad \forall\,\underline{\tau}\,\underline{\sigma}\in\Sigma, \tag{3.28}$$

$$\exists\ \alpha>0 \text{ such that } \quad a(\underline{\tau},\underline{\tau})\ge\alpha\|\underline{\tau}\|_\Sigma^2 \qquad \forall\,\underline{\tau}\in\mathrm{Ker}B\,, \tag{3.29}$$

where $B:\Sigma\longrightarrow V'$ is the operator associated with the bilinear form $b(.,.)$, and

$$\mathrm{Ker}\,B = \{\underline{\tau}\in\Sigma\mid b(v,\underline{\tau})=0 \quad \forall\,v\in V\}\,; \tag{3.30}$$

(ii) the bilinear form $b(.,.)$ is bounded on $V\times\Sigma$, and satisfies the *inf-sup* condition, that is,

$$\exists\,M_b>0 \text{ such that } \quad |b(v,\underline{\tau})|\le M_b\|v\|_V\|\underline{\tau}\|_\Sigma, \quad \forall\,v\in V,\ \underline{\tau}\in\Sigma, \tag{3.31}$$

$$\exists\,\beta>0 \text{ such that } \quad \inf_{v\in V}\sup_{\underline{\tau}\in\Sigma}\frac{b(v,\underline{\tau})}{\|v\|_V\,\|\underline{\tau}\|_\Sigma}\ge\ \beta; \tag{3.32}$$

(here and in the following the inf is obviously taken over functions v not identically zero, and we will omit to specify $\|v\|_V\neq 0$);

(*iii*) the bilinear form $c(.,.)$ is symmetric, bounded on $V \times V$ and positive semidefinite, that is

$$\exists\, M_c > 0 \;\text{ such that }\quad |c(u,v)| \le M_c ||u||_V ||v||_V, \quad \forall\, u,\, v \in V, \tag{3.33}$$

$$c(v,v) \ge 0 \qquad \forall\, v \in V\,. \tag{3.34}$$

Let us recall the abstract Theorem 1.2 of Chapt. II in [27].

Theorem 3.1 *Let Σ, V be two Hilbert spaces with norms $||\cdot||_\Sigma$, $||\cdot||_V$, and let $f \in V'$, $g \in \Sigma'$, (V', Σ' being the dual spaces of V, Σ, respectively). Consider the problem*

$$\begin{cases} \text{Find } (\underline{\sigma}, u) \in \Sigma \times V \;\text{ such that} \\ a(\underline{\sigma}, \underline{\tau}) + b(u, \underline{\tau}) = < g, \underline{\tau} >_{\Sigma' \times \Sigma} \qquad \forall\, \underline{\tau} \in \Sigma, \\ b(v, \underline{\sigma}) - c(u,v) = < f, v >_{V' \times V} \qquad \forall\, v \in V, \end{cases} \tag{3.35}$$

where $< \cdot, \cdot >$ denotes the duality brackets. Under assumptions (3.28)–(3.34)*, problem* (3.35) *has a unique solution for all $f \in V'$, $g \in \Sigma'$. Moreover, there exists a positive constant $\mathcal{M}$, depending nonlineraly on M_a, M_b, M_c, α, β, such that*

$$||\underline{\sigma}||_\Sigma + ||u||_V \le \mathcal{M}\left(||g||_{\Sigma'} + ||f||_{V'}\right), \tag{3.36}$$

where

$$||g||_{\Sigma'} = \sup_{\underline{\tau} \in \Sigma} \frac{\langle g, \underline{\tau} \rangle}{||\underline{\tau}||_\Sigma}, \qquad ||f||_{V'} = \sup_{v \in V} \frac{\langle f, v \rangle}{||v||_V}. \tag{3.37}$$

Let us now check that the abstract hypotheses of Theorem 3.1 are verified for problem (3.27). Boundedness of the bilinear form (3.24) follows trivially from (3.1) with $M_a = a_0^{-1}$. Next, notice that $\mathrm{Ker}\, B$ is characterized by elements of $\Sigma \subset H(\mathrm{div}; \Omega)$ with null divergence, so that $||\underline{\tau}||_\Sigma \equiv ||\underline{\tau}||_{0,\Omega}$ for $\underline{\tau} \in \mathrm{Ker}\, B$. Then property (3.29) holds with $\alpha = a_M^{-1}$ (see (3.1)).
Symmetry, boundedness, and property (3.34) for the bilinear form (3.26) follow immediately from definition and assumptions (3.2); in particular (3.33) holds with $M_c = \gamma_M$. Boundedness of the bilinear form (3.25) holds trivially, with $M_b = 1$, by definition. In order to prove (3.32), we consider, $\forall\, v \in L^2(\Omega)$, the following auxiliary problem:

$$\begin{cases} -\Delta\varphi & = v & \text{in}\, \Omega \\ \varphi & = 0 & \text{on}\, \Gamma_D \\ \underline{\nabla}\,\varphi \cdot \underline{n} & = 0 & \text{on}\, \Gamma_N. \end{cases} \tag{3.38}$$

Problem (3.38) has a unique solution $\varphi \in H^1(\Omega)$, and $||\varphi||_{1,\Omega} \le C||v||_{0,\Omega}$. Then, the vector $\underline{\tau} = -\underline{\nabla}\,\varphi$ verifies $\underline{\tau} \in \Sigma$, $\mathrm{div}\,\underline{\tau} = v$ and $||\underline{\tau}||_\Sigma \le \sqrt{C^2+1}||v||_V$. Therefore, the *inf-sup* condition (3.32) holds with $\beta = (C^2+1)^{-1/2}$.

Remark 3.1 *The solution φ of the auxiliary problem* (3.38) *actually belongs to $W^{1,p}(\Omega)$ for some $p > 2$, and*

$$||\varphi||_{W^{1,p}(\Omega)} \leq C||v||_{0,\Omega}. \tag{3.39}$$

(See, e.g., [70]). Hence, setting

$$\Sigma^* = \left\{ \underline{\tau} \in (L^p(\Omega))^2 \mid \operatorname{div} \underline{\tau} \in L^2(\Omega), \underline{\tau} \cdot \underline{n} = 0 \text{ on } \Gamma_N \right\}, \tag{3.40}$$

we see that the inf-sup *condition* (3.32) *holds with Σ replaced by Σ^*:*

$$\exists\, \beta^* > 0 \quad \text{such that} \quad \inf_{v \in V} \sup_{\underline{\tau} \in \Sigma^*} \frac{b(v, \underline{\tau})}{||v||_V \; ||\underline{\tau}||_{\Sigma^*}} \geq \beta^*. \tag{3.41}$$

We also notice that the inf-sup *condition* (3.32) *implies that*

$$\text{the operator } B \text{ is surjective in } V \;\; (\forall\, v \in V, \exists\, \underline{\tau} \in \Sigma \text{ such that } \operatorname{div} \underline{\tau} = v). \tag{3.42}$$

Moreover, (3.41) *implies that*

$$\text{the operator } B \text{ has a continuous lifting from } V \text{ into } \Sigma^*. \tag{3.43}$$

We can then restate Theorem 3.1 applied to problem (3.27).

Theorem 3.2 *For every $g \in H^{1/2}(\Gamma_D)$, and for every $f \in L^2(\Omega)$, there exists a unique $(\underline{\sigma}, u) \in \Sigma^* \times V$ solution of problem* (3.27). *Moreover, the following bound holds*

$$||\underline{\sigma}||_\Sigma \; + \; ||u||_V \leq \mathcal{M}(||g||_{H^{1/2}(\Gamma_D)} \; + \; ||f||_{L^2(\Omega)}) \tag{3.44}$$

with $\mathcal{M}$ dependent nonlinearly on $a_0, \gamma_M, \alpha, \beta$.

Remark 3.2 *We explicitly point out that the solution of* (3.27) *is such that u coincides with the solution of* (2.14), *and $\underline{\sigma} \equiv a \underline{\nabla} u$. Therefore u, a priori seeked in $L^2(\Omega)$, is actually more regular (at least $u \in H^1(\Omega)$), according to the regularity of the solution of* (2.14).

3.3 Discretization schemes: an abstract framework

Let $\{\mathcal{T}_h\}_h$ be a family of regular decompositions of $\overline{\Omega}$ into elements K, with boundary ∂K, made in such a way that there is always a vertex of $\mathcal{T}_h$ on the interface between Γ_D and Γ_N. Following [41], a family of decompositions is regular if there exists a constant $\mathcal{K}^*$ such that

$$\frac{D_K}{\rho_K} \leq \mathcal{K}^* \qquad \forall\, K \in \mathcal{T}_h, \tag{3.45}$$

where D_K is the diameter of the smallest circumscribed circle of K and ρ_K is the diameter of the largest inscribed circle in K. Inequality (3.45) immediately implies that

$$D_K \leq \mathcal{K}^* \rho_K \leq \mathcal{K} h_K \qquad \forall\, K \in \mathcal{T}_h\,, \tag{3.46}$$

where h_K denotes the diameter of K, and $\mathcal{K}$ denotes, here and in the sequel, a generic constant depending only on $\mathcal{K}^*$. Finally, we set $h = \max_K h_K$. For every triangulation $\mathcal{T}_h$ we define

$$\widetilde{\Sigma} = \left\{ \underline{\tau} \in L^2(\Omega) \,|\, \underline{\tau} \in \prod_K H(\mathrm{div}\,; K)\,,\, \underline{\tau} \cdot \underline{n} = 0 \text{ on } \partial K \cap \Gamma_N\,, \forall\, K \in \mathcal{T}_h \right\}, \tag{3.47}$$

with the norm

$$||\underline{\tau}||^2_{\widetilde{\Sigma}} = ||\underline{\tau}||^2_{0,\Omega} + \sum_K ||\mathrm{div}\,\underline{\tau}||^2_{0,K}, \tag{3.48}$$

and we notice that

$$||\underline{\tau}||_{\widetilde{\Sigma}} \equiv ||\underline{\tau}||_{\Sigma} \qquad \forall\, \underline{\tau} \in \Sigma. \tag{3.49}$$

For all $K \in \mathcal{T}_h$, we introduce finite dimensional spaces $Q(K)$ for vector functions and $P(K)$ for scalar functions. For the sake of simplicity, we might assume that there exists an integer $k \geq 0$ such that, on each K, $Q(K)$ and $P(K)$ consist of polynomials of degree $\leq k$. Let us define the following discrete spaces

$$\widetilde{\Sigma}_h = \left\{ \underline{\tau}_h \in \widetilde{\Sigma} \,|\, \underline{\tau}_{h|K} \in Q(K)\,, \forall\, K \in \mathcal{T}_h \right\}, \tag{3.50}$$

$$V_h = \left\{ v_h \in V \,|\, v_{h|K} \in P(K)\,, \forall\, K \in \mathcal{T}_h \right\}. \tag{3.51}$$

As we shall see, the discrete solution will be seeked in a more regular space

$$\Sigma_h \subset \widetilde{\Sigma}_h, \tag{3.52}$$

still equipped with the norm $||\cdot||_{\widetilde{\Sigma}}$. Σ_h will include 'some' continuity of the normal component across the interelements but, in general, a nonconforming approximation will be allowed, with Σ_h not included in Σ. The abstract formulation which we introduce here includes the well known Raviart–Thomas [135] and Brezzi–Douglas–Marini [26] elements, and the elements introduced in [98]. Actual choices of $Q(K)$ and $P(K)$ will be given in Section 3.5.
Definition (3.47) makes the bilinear form (3.25) meaningless on $\widetilde{\Sigma}$. For this reason we introduce the discrete bilinear form

$$b_h(v, \underline{\tau}) = \sum_K \int_K v\, \mathrm{div}\,\underline{\tau} dx \qquad v \in V,\ \underline{\tau} \in \widetilde{\Sigma}, \tag{3.53}$$

and we notice that

$$b_h(v, \underline{\tau}) \equiv b(v, \underline{\tau}) \qquad v \in V,\ \underline{\tau} \in \Sigma. \tag{3.54}$$

Accordingly, instead of (3.30) the following weaker definition will be used from now on

$$\mathrm{Ker}\,\widetilde{B} = \left\{ \underline{\tau} \in \widetilde{\Sigma} \,|\, b_h(v, \underline{\tau}) = 0 \quad \forall\, v \in V \right\}, \tag{3.55}$$

$\widetilde{B}$ being the operator $\widetilde{\Sigma} \longrightarrow V'$ associated with the bilinear form $b_h(.,.)$. Definition (3.55) implies that $\mathrm{Ker}\,\widetilde{B}$ is made of vectors which are divergence-free element by element

$$\underline{\tau} \in \mathrm{Ker}\,\widetilde{B} \Longrightarrow \mathrm{div}\,\underline{\tau} = 0 \ \text{ in } K, \quad \forall K \in \mathcal{T}_h. \tag{3.56}$$

Similarly we define

$$\mathrm{Ker}\,B_h = \left\{ \underline{\tau}_h \in \widetilde{\Sigma}_h \,|\, b_h(v_h, \underline{\tau}_h) = 0 \quad \forall\, v_h \in V_h \right\}. \tag{3.57}$$

The discrete formulation of (3.27) is then:

$$\begin{cases} \text{Find } (\underline{\sigma}_h, u_h) \in \Sigma_h \times V_h \ \text{ such that} \\ a(\underline{\sigma}_h, \underline{\tau}_h) + b_h(u_h, \underline{\tau}_h) = < g, \underline{\tau}_h \cdot \underline{n} >_{|\Gamma_D} \quad \forall \underline{\tau}_h \in \Sigma_h, \\ b_h(v_h, \underline{\sigma}_h) - c(u_h, v_h) = -(f, v_h) \qquad\quad \forall\, v_h \in V_h. \end{cases} \tag{3.58}$$

In order to prove existence and uniqueness of the solution of (3.58), and optimal error bounds, we need to state, at this very abstract level, the assumptions on the spaces Σ_h, V_h that allow to derive optimal error estimates. The following two assumptions are crucial guidelines to introduce proper discretizations:

$$\mathrm{Ker}B_h \subset \mathrm{Ker}\widetilde{B}; \tag{3.59}$$

there exists an operator $\Pi_h : \Sigma^* \to \Sigma_h$ such that

$$b_h(v_h, \underline{\tau} - \Pi_h \underline{\tau}) = 0 \quad \forall\, v_h \in V_h, \tag{3.60}$$

and

$$||\Pi_h \underline{\tau}||_{\widetilde{\Sigma}} \leq C ||\underline{\tau}||_{\Sigma^*} \quad \forall\, \underline{\tau} \in \Sigma^*, \tag{3.61}$$

with C a constant independent of h.

We note that (3.60)-(3.61) imply that the finite dimensional spaces Σ_h and V_h verify a discrete *inf-sup* condition (see [59]):

$$\exists\, \beta > 0 : \quad \inf_{v_h \in V_h} \sup_{\underline{\tau}_h \in \Sigma_h} \frac{b_h(v_h, \underline{\tau}_h)}{||v_h||_V \, ||\underline{\tau}_h||_{\widetilde{\Sigma}}} \geq \ \beta\,. \tag{3.62}$$

Indeed, as a consequence of the *inf-sup* condition (3.41), which holds for the continuous problem, for $v_h \in V_h \subset V$ there exists $\underline{\tau} \in \Sigma^*$ such that

$$\frac{b(v_h, \underline{\tau})}{||v_h||_V \, ||\underline{\tau}||_{\Sigma^*}} \geq \ \beta^*. \tag{3.63}$$

Due to (3.60), (3.61), and (3.54), from (3.63) we obtain

$$\frac{b_h(v_h, \Pi_h \underline{\tau})}{||v_h||_V \, ||\Pi_h \underline{\tau}||_{\widetilde{\Sigma}}} \geq C^{-1} \frac{b_h(v_h, \underline{\tau})}{||v_h||_V \, ||\underline{\tau}||_{\Sigma^*}} = C^{-1} \frac{b(v_h, \underline{\tau})}{||v_h||_V \, ||\underline{\tau}||_{\Sigma^*}} \geq \ C^{-1}\beta^*\,. \tag{3.64}$$

Hence, (3.62) holds with $\beta = C^{-1}\beta^*$.

We have the following result.

Theorem 3.3 *Under assumptions* (3.59) *and* (3.62)*, problem* (3.58) *has a unique solution.*

Proof. For the proof we could apply the abstract theorem 3.1. However, the problem being finite dimensional, uniqueness implies existence, and the proof can be simplified. For the convenience of the reader we report it here. Let $(\underline{\sigma}_h^*, u_h^*) \in \Sigma_h \times V_h$ be the solution of the homogeneous discrete problem associated with (3.58). Taking $\underline{\tau}_h = \underline{\sigma}_h^*$ in the first equation of (3.58), and using the second equation with $v_h = u_h^*$ we obtain

$$a(\underline{\sigma}_h^*, \underline{\sigma}_h^*) + c(u_h^*, u_h^*) = 0. \tag{3.65}$$

This implies that $\underline{\sigma}_h^* = 0$, and, if $\gamma_0 > 0$, $u_h^* = 0$. Instead, if $\gamma_0 = 0$, we make use of the first equation of (3.58), which reduces to

$$b_h(u_h^*, \underline{\tau}_h) = 0 \quad \forall\, \underline{\tau}_h \in \Sigma_h. \tag{3.66}$$

Thanks to (3.62), this implies $u_h^* = 0$. □

Remark 3.3 *A direct consequence of* (3.59) *and* (3.56) *is that*

$$\int_K v_h \operatorname{div} \underline{\tau}_h dx = 0 \quad \forall\, v_h \in P(K) \Longrightarrow \operatorname{div} \underline{\tau}_{h|K} = 0 \quad \forall\, K \in \mathcal{T}_h, \tag{3.67}$$

so that $dim(P(K)) \geq dim(\operatorname{div}(Q(K)))$*. On the other hand, from* (3.62) *it follows*

$$\int_K v_h \operatorname{div} \underline{\tau}_h dx = 0 \quad \forall\, \underline{\tau}_h \in Q(K) \Longrightarrow v_{h|K} = 0 \quad \forall\, K \in \mathcal{T}_h, \tag{3.68}$$

and consequently $dim(\operatorname{div}(Q(K))) \geq dim(P(K))$*. Therefore, assumptions* (3.59) *and* (3.62) *imply*

$$dim(\operatorname{div}(Q(K))) \equiv dim(P(K)). \tag{3.69}$$

Moreover, for any choice of basis functions $\{v^1, \cdots, v^r\}$*, and* $\{d^1, \cdots, d^r\}$ *in* $P(K)$ *and* $\operatorname{div}(Q(K))$ *respectively, the matrix*

$$\int_K d^i\, v^j\, dx \quad i, j = 1, r \tag{3.70}$$

is nonsingular.

3.4 Error Estimates in the Abstract Framework

In order to derive error estimates we introduce an interpolation operator P_h from V to V_h, which can be defined locally by

$$\int_K (v - P_h v) \operatorname{div} \underline{\tau}_h dx = 0 \qquad \forall\, \underline{\tau}_h \in Q(K). \tag{3.71}$$

In order to see that P_h is well defined, thanks to (3.69) it is sufficient to check uniqueness, that is,

$$\int_K P_h v \,\mathrm{div}\,\underline{\tau}_h dx = 0 \quad \forall \underline{\tau}_h \in Q(K) \Longrightarrow P_h v = 0. \tag{3.72}$$

This is an immediate consequence of the (3.70). Next, for any $v \in H^1(\Omega)$ we introduce

$$N_h(v, \underline{\tau}_h) = \sum_K < v, \underline{\tau}_h \cdot \underline{n} >_{|\partial K} - < v, \underline{\tau}_h \cdot \underline{n} >_{|\Gamma_D} \quad \underline{\tau}_h \in \Sigma_h, \tag{3.73}$$

and we notice that for u solution of (3.27) , which belongs to $H^1(\Omega)$ (see Remark 3.2), we have

$$N_h(u, \underline{\tau}_h) = \sum_K < u, \underline{\tau}_h \cdot \underline{n} >_{|\partial K} - < g, \underline{\tau}_h \cdot \underline{n} >_{|\Gamma_D} \quad \underline{\tau}_h \in \Sigma_h. \tag{3.74}$$

It is easy to see that the solution $(\underline{\sigma}, u)$ of (3.27) verifies

$$\begin{cases} a(\underline{\sigma}, \underline{\tau}_h) + b_h(u, \underline{\tau}_h) - N_h(u, \underline{\tau}_h) =< g, \underline{\tau}_h \cdot \underline{n} >_{|\Gamma_D} & \forall \underline{\tau}_h \in \Sigma_h, \\ b_h(v_h, \underline{\sigma}) - c(u, v_h) = -(f, v_h) & \forall\, v_h \in V_h. \end{cases} \tag{3.75}$$

The term $N_h(u, \underline{\tau}_h)$ is a measure of the possible nonconformity of the space Σ_h, and it will be zero, due to the regularity of u, for conforming choices of Σ_h. By subtracting (3.58) from (3.75), and using (3.71) and (3.60), we obtain the error equations

$$\begin{cases} a(\underline{\sigma} - \underline{\sigma}_h, \underline{\tau}_h) + b_h(P_h u - u_h, \underline{\tau}_h) - N_h(u, \underline{\tau}_h) = 0 & \forall \underline{\tau}_h \in \Sigma_h, \\ b_h(v_h, \Pi_h \underline{\sigma} - \underline{\sigma}_h) - c(u - u_h, v_h) = 0 & \forall\, v_h \in V_h. \end{cases} \tag{3.76}$$

We can prove the following result.

Theorem 3.4 *Let* $(\underline{\sigma}, u)$ *be the solution of* (3.27) *and* $(\underline{\sigma}_h, u_h)$ *that of* (3.58). *Under assumptions* (3.28)-(3.34), (3.59) *and* (3.62) *there exists a constant* C *independent of* h, *such that the following estimate holds*

$$||\Pi_h \underline{\sigma} - \underline{\sigma}_h||_{\widetilde{\Sigma}} + ||P_h u - u_h||_{0,\Omega} \leq C\Big(||\underline{\sigma} - \Pi_h \underline{\sigma}||_{0,\Omega} + ||u - P_h u||_{0,\Omega} + \sup_{\underline{\tau}_h \in \Sigma_h} \frac{N_h(u, \underline{\tau}_h)}{||\underline{\tau}_h||_{\widetilde{\Sigma}}} \Big).$$

Proof. We add and subtract $a(\Pi_h \underline{\sigma}, \underline{\tau}_h)$, with $\Pi_h \underline{\sigma}$ defined in (3.60)-(3.61), in the first equation of (3.76), and $c(P_h u, v_h)$, with $P_h u$ defined in (3.71), in the second equation of (3.76), thus obtaining

$$\begin{cases} a(\Pi_h \underline{\sigma} - \underline{\sigma}_h, \underline{\tau}_h) + b_h(P_h u - u_h, \underline{\tau}_h) = a(\Pi_h \underline{\sigma} - \underline{\sigma}, \underline{\tau}_h) + N_h(u, \underline{\tau}_h) & \forall \underline{\tau}_h \in \Sigma_h, \\ b_h(v_h, \Pi_h \underline{\sigma} - \underline{\sigma}_h) - c(P_h u - u_h, v_h) = c(u - P_h u, v_h) & \forall\, v_h \in V_h. \end{cases} \tag{3.77}$$

Problem (3.77) has the form (3.35) and verifies the hypotheses of Theorem 3.1. Hence, from (3.36) we get

$$\begin{aligned} ||\Pi_h\underline{\sigma} - \underline{\sigma}_h||_{\widetilde{\Sigma}} + ||P_h u - u_h||_{0,\Omega} &\leq \mathcal{M}\Big(\sup_{\underline{\tau}_h \in \Sigma_h} \frac{a(\Pi_h\underline{\sigma} - \underline{\sigma}, \underline{\tau}_h)}{||\underline{\tau}_h||_{\widetilde{\Sigma}}} + \sup_{\underline{\tau}_h \in \Sigma_h} \frac{N_h(u, \underline{\tau}_h)}{||\underline{\tau}_h||_{\widetilde{\Sigma}}} \\ &+ \sup_{v_h \in V_h} \frac{c(u - P_h u, v_h)}{||v_h||_{0,\Omega}} \Big). \end{aligned}$$

Then, the result follows. □

Remark 3.4 *Under the same assumptions of Theorem 3.4, we immediately deduce, via triangle inequality and the obvious observation that* $||\cdot||_{0,\Omega} \leq ||\cdot||_{\widetilde{\Sigma}}$,

$$||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + ||u - u_h||_{0,\Omega} \leq C\Big(||\underline{\sigma} - \Pi_h\underline{\sigma}||_{0,\Omega} + ||u - P_h u||_{0,\Omega} + \sup_{\underline{\tau}_h \in \Sigma_h} \frac{N_h(u, \underline{\tau}_h)}{||\underline{\tau}_h||_{\widetilde{\Sigma}}} \Big). \quad (3.78)$$

Similarly, whenever an estimate for $||\underline{\sigma} - \Pi_h\underline{\sigma}||_{\widetilde{\Sigma}}$ *is available, we also deduce*

$$||\underline{\sigma} - \underline{\sigma}_h||_{\widetilde{\Sigma}} + ||u - u_h||_{0,\Omega} \leq C\Big(||\underline{\sigma} - \Pi_h\underline{\sigma}||_{\widetilde{\Sigma}} + ||u - P_h u||_{0,\Omega} + \sup_{\underline{\tau}_h \in \Sigma_h} \frac{N_h(u, \underline{\tau}_h)}{||\underline{\tau}_h||_{\widetilde{\Sigma}}} \Big). \quad (3.79)$$

3.5 Examples of Mixed Finite Elements

We shall provide here examples of finite elements which fit the abstract framework introduced above. The first four examples refer to well known families of mixed finite elements, namely, the Raviart–Thomas (*RT*) and Brezzi–Douglas–Marini (*BDM*) families, while the last two examples were introduced in [98].

Example 1 The Raviart–Thomas elements [135] over a triangular decomposition of Ω. For any integer $k \geq 0$, and for any triangle $K \in \mathcal{T}_h$, define

$$\begin{aligned} Q(K) &= (P_k(K))^2 + \underline{x} P_k(K), & (3.80) \\ P(K) &= P_k(K), & (3.81) \end{aligned}$$

where $\underline{x} = (x_1, x_2)$, and $P_k(K)$ denotes the set of polynomials of degree $\leq k$ in K.

Example 2 The Raviart–Thomas elements [135] over a rectangular decomposition of Ω. For any integer $k \geq 0$, and for any rectangle $K \in \mathcal{T}_h$, define

$$\begin{aligned} Q(K) &= (Q_k(K) + x_1 Q_k(K)) \times (Q_k(K) + x_2 Q_k(K)), & (3.82) \\ P(K) &= Q_k(K), & (3.83) \end{aligned}$$

where $Q_k(K)$ denotes the set of polynomials of degree $\leq k$ in each variable x_1 and x_2.

Example 3 The Brezzi–Douglas–Marini elements [26] over a triangular decomposition of Ω. For any integer $k \geq 1$, and for any triangle $K \in \mathcal{T}_h$, define

$$Q(K) = (P_k(K))^2, \tag{3.84}$$
$$P(K) = P_{k-1}(K). \tag{3.85}$$

Example 4 The Brezzi–Douglas–Marini elements [26] over a rectangular decomposition of Ω. For any integer $k \geq 1$, and for any rectangle $K \in \mathcal{T}_h$ define

$$Q(K) = (P_k(K))^2 \oplus \left\{\underline{curl}(x_1 x_2^{k+1})\right\} \oplus \left\{\underline{curl}(x_2 x_1^{k+1})\right\}, \tag{3.86}$$
$$P(K) = P_{k-1}(K), \tag{3.87}$$

where $\underline{curl}\,\varphi = (\partial_2\varphi, -\partial_1\varphi)$, for $\varphi \in H^1(\Omega)$.

The choice of $P(K)$ identifies the space of scalars V_h defined in (3.51), while the choice of $Q(K)$ identifies only the space of vectors $\widetilde{\Sigma}_h$. In order to define the space $\Sigma_h \subset \widetilde{\Sigma}_h$, where the solution of (3.75) is seeked, the regularity assumption must be specified. We shall show that, with the above choices of polynomial spaces, it is possible to define degrees of freedom, and consequently to construct local bases, that guarantee continuity of the normal component of vectors in Σ_h across interelement boundaries, thus giving rise to a conforming approximation:

$$\Sigma_h = \left\{\underline{\tau}_h \in H(\mathrm{div}\,;\Omega) \mid \underline{\tau}_{h|K} \in Q(K)\ \forall K \in \mathcal{T}_h,\ \underline{\tau}_h \cdot \underline{n} = 0 \text{ on } \Gamma_N\right\}. \tag{3.88}$$

Indeed, for each k that identifies the pair $(Q(K), P(K))$ in the above families, let us introduce the spaces of polynomials on the edges of K

$$R(e) = P_k(e) \text{ for each edge } e \text{ of } K, \quad R(\partial K) = \prod_{e \in \partial K} R(e). \tag{3.89}$$

Notice that, in all the previous examples,

$$\underline{\tau} \in Q(K) \Longrightarrow \underline{\tau} \cdot \underline{n}_{|\partial K} \in R(\partial K). \tag{3.90}$$

Moreover,

$$\mathrm{div}\ Q(K) = P(K). \tag{3.91}$$

Next, let D be the space

$$D = \left\{\underline{q} \in Q(K) \mid \mathrm{div}\,\underline{q}_{|K} = 0,\ \underline{q} \cdot \underline{n}_{|\partial K} = 0\right\}, \tag{3.92}$$

and notice that $D = \{0\}$ for the low elements of the four families.

Proposition 3.1 *For any $\underline{q} \in Q(K)$, the following relations imply $\underline{q} = 0$:*

$$\int_e \mu\, \underline{q} \cdot \underline{n}\, ds = 0 \quad \forall \text{ edge } e \text{ of } K\,,\ \forall \mu \in R(e)\,, \tag{3.93}$$

$$\int_K \underline{q} \cdot \underline{\nabla} v dx = 0 \quad \forall v \in P(K)\,, \tag{3.94}$$

$$\int_K \underline{q} \cdot \underline{\varphi}\, dx = 0 \quad \forall \underline{\varphi} \in D\,. \tag{3.95}$$

Proof. For any fixed k, we have that $\underline{q} \cdot \underline{n}_{|e} \in R(e)$. Hence, (3.93) implies $\underline{q} \cdot \underline{n}_{|\partial K} = 0$. Moreover, the trace of a function $v \in P(K)$ belongs to $R(\partial K)$ in all the examples. Hence, integration by parts and (3.93)–(3.94) give

$$\int_K \operatorname{div} \underline{q}\, v\, dx = -\int_K \underline{q} \cdot \underline{\nabla} v dx + \int_{\partial K} v\, \underline{q} \cdot \underline{n} ds = 0 \ \forall v \in P(K), \tag{3.96}$$

which implies $\operatorname{div} \underline{q} = 0$, due to (3.91). Then, (3.93)-(3.94) imply $\underline{q} \in D$. Reciprocally, it is obvious that (3.93)-(3.94) hold for $\underline{\varphi} \in D$. Hence, (3.93)-(3.94) are equivalent to $\underline{q} \in D$, and (3.95) gives $\underline{q} = 0$. □

To see that (3.93)–(3.95) can be used as degrees of freedom, after choosing local bases, it remains to check that (3.93)-(3.94) are linearly independent. For this we refer, e.g., to [27] Section III.3, Lemma 3.1. It is then possible to use $\underline{\tau} \cdot \underline{n}$ among the degrees of freedom, and then to impose continuity at the interelements, thus constructing a conforming approximation. In particular, for the lowest order Raviart–Thomas triangular elements (RT_0) of *Example 1* a local basis for $Q(K)$ is uniquely defined by the following degrees of freedom

$$\int_{e^j} \underline{\tau}^i \cdot \underline{n}^j\, ds = \delta_{ij} \qquad i, j = 1, 3\,. \tag{3.97}$$

Conformity, together with relation (3.91) (which is a stronger property than (3.69)), will allow to simplify (and somewhat improve) the abstract results. In particular, conformity implies that the discrete bilinear form $b_h(.,.)$ defined in (3.53) is not needed here, since it will always be applied to elements of Σ. Hence, it can be replaced by the bilinear form $b(.,.)$ defined in (3.25):

$$b_h(v_h, \underline{\tau}_h) \equiv b(v_h, \underline{\tau}_h) \qquad \forall v_h \in V_h,\ \underline{\tau}_h \in \Sigma_h \subset \Sigma. \tag{3.98}$$

From property (3.91) it is immediate to check that the first abstract hypothesis (3.59) is fulfilled. An important consequence of (3.91) is that the operator P_h defined from V to V_h in (3.71) coincides with as the usual L^2–projection

$$\int_K (v - P_h v)\, w dx = 0 \qquad \forall w \in P(K)\,. \tag{3.99}$$

It remains to define uniquely the element $\Pi_h \underline{\tau} \in Q(K)$ verifying (3.60)–(3.61). Proposition 3.1 implies that, $\forall \underline{\tau} \in \Sigma^*$, $\Pi_h \underline{\tau} \in \Sigma_h$ can be uniquely defined locally through (3.93)-(3.94)-(3.95) as:

$$\int_e \mu \, (\underline{\tau} - \Pi_h \underline{\tau}) \cdot \underline{n} \, ds \;=\; 0 \quad \forall \, \text{edge } e \text{ of } K \,, \; \forall \mu \in R(e) \,, \tag{3.100}$$

$$\int_K (\underline{\tau} - \Pi_h \underline{\tau}) \cdot \underline{\nabla} \, v \, dx \;=\; 0 \quad \forall v \in P(K) \,, \tag{3.101}$$

$$\int_K (\underline{\tau} - \Pi_h \underline{\tau}) \cdot \underline{q} \, dx \;=\; 0 \quad \forall \underline{q} \in D \,. \tag{3.102}$$

Conditions (3.100)–(3.102) define an interpolation operator Π_h from Σ^* to Σ_h which is uniformly bounded from Σ^* to $\Sigma_h \subset \Sigma$, that is,

$$||\Pi_h \underline{\tau}||_\Sigma \leq C ||\underline{\tau}||_{\Sigma^*} \,, \tag{3.103}$$

with C independent of h and of $\underline{\tau}$, as shown, e.g., in [27]. Moreover, since the trace of a function $v \in P(K)$ belongs to $R(\partial K)$, integration by parts, and (3.100)–(3.101) imply, for all $v \in P(K)$,

$$\int_K \mathrm{div} \, (\underline{\tau} - \Pi_h \underline{\tau}) \, v \, dx = - \int_K (\underline{\tau} - \Pi_h \underline{\tau}) \cdot \underline{\nabla} \, v \, dx + \int_{\partial K} (\underline{\tau} - \Pi_h \underline{\tau}) \cdot \underline{n} \, v \, ds = 0. \tag{3.104}$$

Collecting the various properties (3.104), (3.98), and (3.99) we deduce that the commuting diagram property [55] holds

$$\begin{array}{ccccc} \Sigma^* & \xrightarrow{\mathrm{div}} & V & \longrightarrow & 0 \\ {\scriptstyle \Pi_h}\downarrow & & \downarrow {\scriptstyle P_h} & & \\ \Sigma_h & \xrightarrow{\mathrm{div}} & V_h & \longrightarrow & 0 \end{array} \tag{3.105}$$

This means, in other words, that for every $\underline{\tau} \in \Sigma^*$ one has

$$\mathrm{div} \, (\Pi_h \underline{\tau}) \;=\; P_h \mathrm{div} \, \underline{\tau} \,. \tag{3.106}$$

therefore, the discrete *inf-sup* condition (3.62) holds. This, together with (3.59) allows to apply the abstract theory of Sections 3.3 and 3.4. In particular, error estimates (3.78)–(3.79) hold with $N_h \equiv 0$, due to the regularity of u and to the continuity of the normal component of elements in Σ_h.

Since in the next sections the lowest order Raviart-Thomas element RT_0 will be extensively used, we report here, for the reader convenience, the definition of the discrete spaces

$$V_h = \left\{ v_h \in L^2(\Omega) \,|\, v_{h|K} \in P_0(K) \,, \forall \, K \in \mathcal{T}_h \right\} , \tag{3.107}$$

$$\Sigma_h = \left\{ \underline{\tau}_h \in H(\mathrm{div} \,; \Omega) \,|\, \underline{\tau}_{h|K} \in Q(K) \; \forall K \in \mathcal{T}_h, \, \underline{\tau}_h \cdot \underline{n} = 0 \text{ on } \Gamma_N \right\} , \tag{3.108}$$

with $Q(K) = (P_0(K))^2 + \underline{x}P_0(K)$. Notice that $dim(V_h) = \#$ of elements in $\mathcal{T}_h$, and $dim(\Sigma_h) = \#$ of edges not belonging to Γ_N. According to (3.97), a basis function $\underline{\tau}^r \in \Sigma_h$ is defined requiring that, for each edge e^s not belonging to Γ_N,

$$\int_{e^s} \underline{\tau}^r \cdot \underline{n}^s \, ds = \delta_{rs}, \tag{3.109}$$

where $\underline{n}^s$ is the normal unit vector to e^s, whose orientation is chosen once and for all. Consequently, considering the two elements having e^s in common, $\underline{n}^s$ is outward for one triangle and inward for the other one. Error estimate (3.79) gives

$$||\underline{\sigma} - \underline{\sigma}_h||_\Sigma + ||u - u_h||_{0,\Omega} \leq C\, h(|\underline{\sigma}|_{1,\Omega} + |\mathrm{div}\, \underline{\sigma}|_{1,\Omega} + |u|_{1,\Omega}). \tag{3.110}$$

Indeed, the interpolation errors are (see [41, 27])

$$||\underline{\sigma} - \Pi_h \underline{\sigma}||_\Sigma \leq C\, h(|\underline{\sigma}|_{1,\Omega} + |\mathrm{div}\, \underline{\sigma}|_{1,\Omega}), \quad ||u - P_h u||_{0,\Omega} \leq C\, h|u|_{1,\Omega}. \tag{3.111}$$

The key properties to obtain the commuting diagram property are conformity and (3.91). We provide now an example, taken from [98], of conforming approximation which violates (3.91), though remaining in the abstract framework.

Example 5 Consider a triangular decomposition of Ω. For each triangle $K \in \mathcal{T}_h$, we choose the finite–dimensional polynomial sets as follows

$$Q(K) = span\left\{\underline{\tau}^1, \underline{\tau}^2, \underline{\tau}^3\right\}, \tag{3.112}$$

$$P(K) = P_0(K), \tag{3.113}$$

where

$$\underline{\tau}^1 = (1,0)\,,\ \underline{\tau}^2 = (0,1)\,,\ \underline{\tau}^3 = (\psi_1, \psi_2)\,. \tag{3.114}$$

The choice for ψ_1 and ψ_2 is

$$\psi_1\,,\ \psi_2 \in P_2(K)\,, \tag{3.115}$$

and, having chosen an edge $\tilde{e}$ of K, $\underline{\tau}^3 = (\psi_1, \psi_2)$ is defined through the following degrees of freedom

$$\begin{cases} \underline{\tau}^3 \cdot \underline{n}_{|\tilde{e}} & = 1\,, \\ \underline{\tau}^3 \cdot \underline{n}_{|e} & = 0 \quad e \neq \tilde{e}, \\ \int_K \psi_1 dx = \int_K \psi_2 dx & = 0\,, \\ \underline{\tau}^3 \cdot \underline{t}\,(\tilde{m}) & = 0\,, \end{cases} \tag{3.116}$$

where $\tilde{m}$ is the midpoint of $\tilde{e}$, and $\underline{t}$ denotes the unit tangent vector.

Proposition 3.2 *The degrees of freedom* (3.116) *uniquely define* $\underline{\tau}^3$.

Proof. Since $\underline{\tau}^3$ is seeked in a space of dimension 12 (see (3.115)), and the number of degrees of freedom (3.116) is precisely 12, it is sufficient to check uniqueness. Let then $\underline{\tau} = (\tau_1, \tau_2)$ be a vector verifying

$$i)\ \ \underline{\tau}\cdot\underline{n}_{|e} = 0 \quad \forall e, \qquad ii)\ \ \int_K \tau_1\, dx = \int_K \tau_2\, dx = 0, \qquad iii)\ \ \underline{\tau}\cdot\underline{t}(\tilde{m}) = 0.$$

From $i)$ we deduce $\underline{\tau}\cdot\underline{n}_{|\partial K} = 0$. From $ii)$, after integration by parts, we have

$$0 = \int_K \underline{\tau}\cdot\underline{\nabla} p_1\, dx = -\int_K \mathrm{div}\,\underline{\tau}\, p_1\, dx \qquad \forall p_1 \in P_1(K), \tag{3.117}$$

which gives $\mathrm{div}\,\underline{\tau} = 0$. Hence,

$$\underline{\tau} = \underline{curl}\,\varphi, \text{ with } \varphi \in P_3(K) \text{ and } \frac{\partial\varphi}{\partial s} = 0 \text{ on } \partial K, \tag{3.118}$$

where $\partial\varphi/\partial s$ denotes the tangential derivative of φ along ∂K. Consequently, $\varphi =$ constant on ∂K, i.e., $\varphi = a(\lambda_1\lambda_2\lambda_3 + c)$, and $\underline{\tau} = a\,\underline{curl}(\lambda_1\lambda_2\lambda_3)$, where λ_i, $i = 1,3$ being the barycentric coordinates. Condition $iii)$ becomes $a(\partial(\lambda_1\lambda_2\lambda_3)/\partial n)(\tilde{m}) = 0$, implying $a = 0$, since $(\partial(\lambda_1\lambda_2\lambda_3)/\partial n)(\tilde{m}) \neq 0$. □

With this choice, condition (3.91) is violated, since div $\underline{\tau}^3$ is not constant on K. However, we have

$$dim(\mathrm{div}\,(Q(K))) = 1 = dim(P(K))\,, \tag{3.119}$$

and

$$\int_K \mathrm{div}\,\underline{\tau}^3 dx = \int_{\partial K} \underline{\tau}^3\cdot\underline{n} ds = 1\,. \tag{3.120}$$

Hence $\int_K \mathrm{div}\,\underline{\tau}_h dx = 0$, with $\underline{\tau}_h \in Q(K)$, implies $\mathrm{div}\,\underline{\tau}_h = 0$, and the first abstract hypothesis (3.59) is fulfilled.
The space Σ_h can be taken as in (3.88) , and an interpolation operator Π_h from Σ^* to Σ_h verifying (3.60)–(3.61) can be defined locally by

$$\int_e (\underline{\tau} - \Pi_h\underline{\tau})\cdot\underline{n}\, ds = 0 \qquad \forall\, e \text{ edge of } K\,. \tag{3.121}$$

Π_h is well defined, since the matrix $\int_{e^i} \underline{\tau}^j\cdot\underline{n}^i\, ds$, with $i,j = 1,2,3$, is nonsingular, as can be easily seen by construction. Moreover, it can be proved that

$$||\Pi_h\underline{\tau}||_{\tilde{\Sigma}} \leq C||\underline{\tau}||_{\Sigma^*}. \tag{3.122}$$

The second abstract hypothesis is then fulfilled and the abstract theory applies. In particular, since we are using a conforming approximation, error estimate (3.78) holds true with $N_h = 0$. In this case, (3.79) is not applicable since only an interpolation estimate in $L^2(\Omega)$ can be proved

$$||\underline{\tau} - \Pi_h\underline{\tau}||_{0,\Omega} \leq C\, h||\underline{\tau}||_{1,\Omega}. \tag{3.123}$$

We explicitly point out that, in contrast with the previous examples, in this case the operator P_h, as defined in (3.71), is not the L^2-projection. Consequently, the commuting diagram property (3.105) does not hold true for this element, but the interpolation estimate for u in (3.111) still applies. Hence, for this element (3.78) gives

$$||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + ||u - u_h||_{0,\Omega} \leq C\, h(|\underline{\sigma}|_{1,\Omega} + |u|_{1,\Omega}). \tag{3.124}$$

The last example we present here has also been introduced in [98] and it is an example of nonconforming mixed finite element: the continuity of the normal component of the vector variable is imposed only in weak form and the inclusion of Σ_h in Σ is violated. In contrast with the previous case, for this example property (3.91) will be satisfied.

Example 6 Consider a triangular decomposition of Ω. For each triangle $K \in \mathcal{T}_h$ we choose the finite–dimensional polynomial sets as follows

$$Q(K) = span\left\{\underline{\tau}^1, \underline{\tau}^2, \underline{\tau}^3\right\}, \tag{3.125}$$
$$P(K) = P_0(K), \tag{3.126}$$

where

$$\underline{\tau}^1 = (1,0)\,,\ \underline{\tau}^2 = (0,1)\,,\ \underline{\tau}^3 = (\psi_1, \psi_2)\,. \tag{3.127}$$

The choice for ψ_1 and ψ_2 is

$$\psi_1\,,\ \psi_2 \in P_1(K)\,, \tag{3.128}$$

and, for a chosen edge $\tilde{e}$ of K, $\underline{\tau}^3 = (\psi_1, \psi_2)$ is defined through the following degrees of freedom

$$\begin{cases} \underline{\tau}^3 \cdot \underline{n}_{|\tilde{e}} & = 1/|\tilde{e}|\,, \\ \int_e \underline{\tau}^3 \cdot \underline{n}\, ds & = 0\,, \qquad e \neq \tilde{e} \\ \int_K \psi_1 dx = \int_K \psi_2 dx & = 0\,. \end{cases} \tag{3.129}$$

Proposition 3.3 *The degrees of freedom* (3.129) *uniquely define* $\underline{\tau}^3$.

Proof. Since $\underline{\tau}^3$ is seeked in a space of dimension 6 (see (3.128)), and the number of degrees of freedom (3.129) is precisely 6, it is sufficient to check uniqueness. To fix ideas, let e^1 be the special edge $\tilde{e}$, and let e^2, e^3 be the other two edges (see Fig.2). Let then $\underline{\tau} = (\tau_1, \tau_2)$ be a vector verifying

$$i)\ \ \underline{\tau} \cdot \underline{n}^1_{|e^1} = 0, \qquad ii)\ \ \int_{e^j} \underline{\tau} \cdot \underline{n}^j\, ds = 0\ j \neq 1, \qquad iii)\ \ \int_K \tau_1\, dx = \int_K \tau_2\, dx = 0.$$

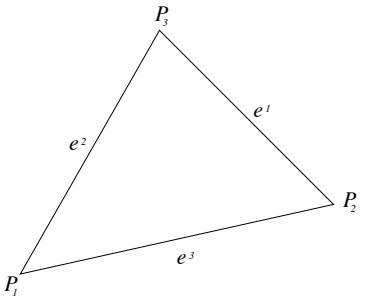


Figure 2: Local numbering of edges and vertices

Since $\operatorname{div}\underline{\tau}$ is constant, from *i*)-*ii*) we deduce $\operatorname{div}\underline{\tau} = 0$. From this and *iii*), after integration by parts, we have, for all $p_1 \in P_1(K)$,

$$0 = \int_K \underline{\tau} \cdot \underline{\nabla} p_1 \, dx = \int_{\partial K} \underline{\tau} \cdot \underline{n} \, p_1 \, ds = \int_{e^2} \underline{\tau} \cdot \underline{n}^2 \, p_1 \, ds + \int_{e^3} \underline{\tau} \cdot \underline{n}^3 \, p_1 \, ds, \tag{3.130}$$

where in the last step we used *i*). Taking $p_1 = \lambda_2$ in (3.130) (λ_2 being the barycentric coordinates verifying $\lambda_{2|e^2} \equiv 0$, $\lambda_2(P_2) = 1$) we obtain

$$\int_{e^3} \underline{\tau} \cdot \underline{n}^3 \, \lambda_2 \, ds = 0. \tag{3.131}$$

Using Simpson's rule and *ii*) we deduce $\underline{\tau} \cdot \underline{n}^3(P_2) = 0$. This, together with *ii*), implies $\underline{\tau} \cdot \underline{n}^3 = 0$ on e^3. With the same argument we deduce $\underline{\tau} \cdot \underline{n}^2 = 0$ on e^2. Hence, $\underline{\tau} \cdot \underline{n} = 0$ on ∂K, together with (3.128), ends the proof. □

Due to the choice of $Q(K)$, $dim(Q(K)) = 3$, but $\underline{\tau} \cdot \underline{n}_{|e} \in P_1(e)$; hence, for $\underline{\tau} \in Q(K)$ continuity of $\underline{\tau} \cdot \underline{n}$ on the interelement edges cannot be imposed, as for the previous examples. However, a weak continuity can be required. More precisely, denoting by $\mathcal{E}_h$ the set of edges of $\mathcal{T}_h$, we define

$$\Sigma_h = \left\{ \underline{\tau}_h \in (L^2(\Omega))^2 \mid \underline{\tau}_{h|K} \in Q(K) \, \forall K \in \mathcal{T}_h, \int_e [\underline{\tau}_h \cdot \underline{n}] \, ds = 0 \, \forall e \in \mathcal{E}_h \setminus \Gamma_D \right\}, \tag{3.132}$$

where $[\underline{\tau}_h \cdot \underline{n}]$ denotes the jump of $\underline{\tau}_h \cdot \underline{n}$ across the edge e when e is an internal edge, and, when e belongs to Γ_N, $[\underline{\tau}_h \cdot \underline{n}]$ simply denotes $\underline{\tau}_h \cdot \underline{n}$.
Note that (3.129) implies that

$$\int_K \operatorname{div}\underline{\tau}^3 \, dx = \int_{\tilde{e}} \underline{\tau}^3 \cdot \underline{n} \, ds = 1 \quad \Longrightarrow \quad \operatorname{div}\underline{\tau}^3|_K = 1/|K|, \tag{3.133}$$

where $|K|$ denotes the area of K. Therefore, (3.91) is satisfied, and the first abstract hypothesis (3.59) is fulfilled; moreover, the operator P_h defined from V to V_h in (3.71) is the usual orthogonal L^2-projection.

An interpolation operator Π_h from Σ^* to Σ_h, verifying (3.60)–(3.61), can be defined locally by

$$\int_e (\underline{\tau} - \Pi_h \underline{\tau}) \cdot \underline{n}\, ds = 0 \qquad \forall\, e \text{ edge of } K\,. \tag{3.134}$$

Π_h is well defined, since the matrix $\int_{e^i} \underline{\tau}^j \cdot \underline{n}^i\, ds$, with $i, j = 1, 2, 3$, is nonsingular, as can be easily seen by construction. We notice that (3.134) implies

$$\operatorname{div} \Pi_h \underline{\tau}_{|K} = P_h \operatorname{div} \underline{\tau}_{|K} \qquad \forall\, \underline{\tau} \in \Sigma^*. \tag{3.135}$$

Therefore,

$$||\Pi_h \underline{\tau}||_{\tilde{\Sigma}} \leq C ||\underline{\tau}||_{\Sigma^*}, \tag{3.136}$$

and the second abstract hypothesis is satisfied. Moreover, the following interpolation estimates hold

$$||\underline{\tau} - \Pi_h \underline{\tau}||_{0,\Omega} \leq C\, h |\underline{\tau}|_{1,\Omega}\,, \tag{3.137}$$

$$\sum_K ||\operatorname{div} (\underline{\tau} - \Pi_h \underline{\tau})||^2_{0,K} \leq C\, h^2 |\operatorname{div} \underline{\tau}|^2_{1,\Omega}\,. \tag{3.138}$$

In contrast with the previous examples, the term N_h is no longer zero, due to the nonconformity of the discrete space Σ_h, and has to be bounded properly.

Proposition 3.4 *Let Σ_h be the space defined in* (3.132)*, and let $w \in H^1(\Omega)$. Then*

$$\sup_{\underline{\tau}_h \in \Sigma_h} \frac{N_h(w, \underline{\tau}_h)}{||\underline{\tau}_h||_{\tilde{\Sigma}}} \leq C h |w|_{1,\Omega}, \tag{3.139}$$

where N_h is defined in (3.73)*.*

Proof. Let $\mathcal{E}'$ be the set of edges of $\mathcal{T}_h$ not belonging to Γ_D. We can rewrite (3.73) as

$$N_h(w, \underline{\tau}_h) = \sum_{e \in \mathcal{E}'} \int_e w\, [\underline{\tau}_h \cdot \underline{n}]\, ds. \tag{3.140}$$

Let P_h^e be the piecewise constant interpolant of w defined as

$$\int_e (w - P_h^e w)\, ds = 0 \qquad \forall e \in \mathcal{E}'. \tag{3.141}$$

According to (3.132), we have

$$N_h(w, \underline{\tau}_h) = \sum_{e \in \mathcal{E}'} \int_e (w - P_h^e w)\, [\underline{\tau}_h \cdot \underline{n}]\, ds. \tag{3.142}$$

Any $\underline{\tau}_h \in \Sigma_h$ can be split as $\underline{\tau}_h = \underline{\tau}' + \underline{\tau}''$, with $\underline{\tau}' \in span\left\{\underline{\tau}^1, \underline{\tau}^2\right\}$, and $\underline{\tau}'' \in span\left\{\underline{\tau}^3\right\}$ in $K, \forall K \in \mathcal{T}_h$. Therefore, using (3.141), (3.142) reads

$$N_h(w, \underline{\tau}_h) = \sum_{e \in \mathcal{E}'} \int_e (w - P_h^e w)\, [\underline{\tau}'' \cdot \underline{n}]\, ds. \tag{3.143}$$

Since $\operatorname{div} \underline{\tau}'' = 0$ in K if and only if $\underline{\tau}'' = 0$ in K due to (3.133), then $\|\operatorname{div} \underline{\tau}''\|_{0,K}$ is a norm. A simple scaling argument in (3.143), classical interpolation estimates, and Cauchy-Schwarz inequality give

$$\begin{aligned} N_h(w, \underline{\tau}_h) &\leq C \sum_K \|w - P_h^e w\|_{0,\partial K} \|\operatorname{div} \underline{\tau}''\|_{0,K} h_K^{1/2} && (3.144) \\ &\leq C \sum_K h_K |w|_{1,K} \|\operatorname{div} \underline{\tau}''\|_{0,K} && (3.145) \\ &\leq C h |w|_{1,\Omega} \|\underline{\tau}_h\|_{\widetilde{\Sigma}}, && (3.146) \end{aligned}$$

and the proof is concluded. □

Using (3.137), (3.138), the interpolation estimate for u in (3.111), and (3.139) in (3.79) we deduce

$$\|\underline{\sigma} - \underline{\sigma}_h\|_{\widetilde{\Sigma}} + \|u - u_h\|_{0,\Omega} \leq C\, h(|\underline{\sigma}|_{1,\Omega} + |\operatorname{div} \underline{\sigma}|_{1,\Omega} + |u|_{1,\Omega}). \tag{3.147}$$

3.6 Hybridization of the mixed formulation

It is well known that problem (3.58) leads to a final linear system whose matrix can be indefinite (this is the case, for instance, when the zeroth–order term is not present in (2.14)). A way to circumvent this problem is to relax the continuity of the normal component of the vectors at the interelement boundaries (even the weak one of the nonconforming case) and to enforce it back through the use of Lagrange multipliers. This technique, introduced in [60] in a different context, was used successfully as a trick to deal with the algebraic system. In [10] it was also studied from the theoretical point of view in the context of mixed formulations for the Raviart-Thomas elements. Error estimates were derived for the Lagrange multipliers, which proved to give an approximation of the scalar variable at the interelements better than that given directly by u_h. In [26] error estimates were proved for the *BDM*–family, while in [98] the elements of *Examples 5-6* of Section 3.5 were analyzed. In order to describe this procedure, we then introduce a new space of vectors, made of discontinuous piecewise polynomial functions (without boundary conditions) defined as

$$\widehat{\Sigma}_h = \left\{ \underline{\tau}_h \in (L^2(\Omega))^2 \mid \underline{\tau}_{h|K} \in Q(K)\,, \forall K \in \mathcal{T}_h \right\}, \tag{3.148}$$

equipped with the norm

$$\|\underline{\tau}\|_h^2 = \|\underline{\tau}\|_{0,\Omega}^2 + \sum_K h_K^2 \, \|\operatorname{div} \underline{\tau}\|_{0,K}^2, \qquad \underline{\tau} \in \widehat{\Sigma}_h. \tag{3.149}$$

To enforce continuity across interelement boundaries of the normal component of vectors in $\widehat{\Sigma}_h$ we need to define a space for the multipliers. Let $\mathcal{E}_h$ be the set of all edges of the decomposition $\mathcal{T}_h$, and let $\mathcal{E}'_h$ be the set of edges not belonging to Γ_D. For every edge $e \in \mathcal{E}_h$ we introduce a finite dimensional space of scalars $R(e)$. For the sake of simplicity, we assume that, on each edge e, $R(e)$ consists of polynomials of degree $\leq k$. We set

$$\Lambda_h = \left\{ \mu_h \in L^2(\mathcal{E}_h) \mid \mu_{h|e} \in R(e),\ \forall e \in \mathcal{E}_h \right\}, \tag{3.150}$$

and

$$\Lambda_{h,0} = \left\{ \mu_h \in \Lambda_h \mid \mu_{h|e} = 0,\ \forall e \in \mathcal{E}_h \cap \Gamma_D \right\}, \tag{3.151}$$

equipped with the norm

$$||\mu_h||^2_{1/2,h} = \sum_{e \in \mathcal{E}_h} h_e^{-1} \, ||\mu_h||^2_{0,e}, \tag{3.152}$$

where h_e denotes the length of the edge e. The norm (3.152) is a sort of $H^{1/2}$-norm, which is natural for a space "of traces". We also have to introduce a bilinear form on $\widehat{\Sigma}_h \times \Lambda_h$:

$$d_h(\mu_h, \underline{\tau}_h) = \sum_K \int_{\partial K} \mu_h \, \underline{\tau}_h \cdot \underline{n} \, ds, \quad \mu_h \in \Lambda_h,\ \underline{\tau}_h \in \widehat{\Sigma}_h. \tag{3.153}$$

We denote by $[\underline{\tau} \cdot \underline{n}]_{|e}$ the jump of $\underline{\tau} \cdot \underline{n}$ across the edge e when e is internal; for boundary edges, $[\underline{\tau} \cdot \underline{n}]_{|e}$ simply denotes $\underline{\tau} \cdot \underline{n}$. Then, (3.153) can be equivalently written as

$$d_h(\mu_h, \underline{\tau}_h) = \sum_e \int_e \mu_h \, [\underline{\tau} \cdot \underline{n}] \, ds \qquad \mu_h \in \Lambda_h,\ \underline{\tau}_h \in \widehat{\Sigma}_h. \tag{3.154}$$

We define

$$\mathrm{Ker}\, D_h = \left\{ \underline{\tau}_h \in \widehat{\Sigma}_h \mid d_h(\mu_h, \underline{\tau}_h) = 0 \quad \forall \mu_h \in \Lambda_{h,0} \right\}, \tag{3.155}$$

having denoted by D_h the operator from $\widehat{\Sigma}_h \longrightarrow (\Lambda_{h,0})'$ associated with the bilinear form $d_h(.,.)$. Using (3.154) and the definition of $\Lambda_{h,0}$ it is easy to see that

$$\underline{\tau} \in Ker\, D_h \implies \int_e \mu_h \, [\underline{\tau} \cdot \underline{n}] \, ds = 0 \quad \forall \mu_h \in R(e) \quad \forall\, e \in \mathcal{E}'_h. \tag{3.156}$$

Together with the abstract assumptions (3.59) and (3.62), we need two additional hypotheses:

$$\Sigma_h = \mathrm{Ker}\, D_h, \tag{3.157}$$

and a discrete *inf-sup* condition for the bilinear form $d_h(.,.)$

$$\exists\, \delta > 0 : \quad \inf_{\mu_h \in \Lambda_h} \sup_{\underline{\tau}_h \in \widehat{\Sigma}_h} \frac{d_h(\mu_h, \underline{\tau}_h)}{||\mu_h||_{1/2,h} \, |||\underline{\tau}_h|||_h} \geq \delta. \tag{3.158}$$

A direct consequence of (3.158) is that, for all $e \in \mathcal{E}_h$,

$$\int_e \mu_h \underline{\tau}_h \cdot \underline{n}\, ds = 0 \quad \forall \underline{\tau}_h \in Q(K) \text{ with } e \subset \partial K \Longrightarrow \mu_{h|e} = 0. \tag{3.159}$$

For every function $\xi \in L^2(\Gamma_D)$, define now

$$\Lambda_{h,\xi} = \left\{ \mu_h \in \Lambda_h \,|\, \int_e (\mu_h - \xi)\underline{\tau}_h \cdot \underline{n}\, ds = 0, \forall \underline{\tau}_h \in \widehat{\Sigma}_h, \;\; \forall e \in \mathcal{E}_h \cap \Gamma_D \right\}. \tag{3.160}$$

Note that, thanks to (3.159), condition $\int_e (\mu_h - \xi)\underline{\tau}_h \cdot \underline{n}\, ds = 0 \;\forall \underline{\tau}_h \in \widehat{\Sigma}_h$ determines uniquely μ_h on e. The affine manifold $\Lambda_{h,\xi}$ can be seen as the subset of Λ_h made of functions μ_h which satisfy $\mu_h = \xi$ on Γ_D in a weak sense.
The discrete formulation of (3.27) is then:

$$\begin{cases} \text{Find } (\widehat{\underline{\sigma}}_h, \widehat{u}_h, \lambda_h) \in \widehat{\Sigma}_h \times V_h \times \Lambda_{h,g} \;\text{ such that} \\ a(\widehat{\underline{\sigma}}_h, \underline{\tau}_h) + b_h(\widehat{u}_h, \underline{\tau}_h) - d_h(\lambda_h, \underline{\tau}_h) = 0 & \forall \underline{\tau}_h \in \widehat{\Sigma}_h, \\ b_h(v_h, \widehat{\underline{\sigma}}_h) - c(\widehat{u}_h, v_h) = -(f, v_h) & \forall v_h \in V_h, \\ d_h(\mu_h, \widehat{\underline{\sigma}}_h) = 0 & \forall \mu_h \in \Lambda_{h,0}. \end{cases} \tag{3.161}$$

Theorem 3.5 *Problem* (3.161) *has a unique solution* $(\widehat{\underline{\sigma}}_h, \widehat{u}_h, \lambda_h)$, *and* $(\widehat{\underline{\sigma}}_h, \widehat{u}_h)$ *coincides with* $(\underline{\sigma}_h, u_h)$, *solution of Problem* (3.58).

Proof. Let $(\underline{\sigma}_h^*, u_h^*, \lambda_h^*) \in \widehat{\Sigma}_h \times V_h \times \Lambda_{h,0}$ be the solution of the homogeneous discrete problem associated with (3.161). Taking $\underline{\tau}_h = \underline{\sigma}_h^*$ in the first equation of (3.161), and using the second equation with $v_h = u_h^*$ and the third equation with $\mu_h = \lambda_h^*$ we obtain

$$a(\underline{\sigma}_h^*, \underline{\sigma}_h^*) + c(u_h^*, u_h^*) = 0\,. \tag{3.162}$$

This implies that $\underline{\sigma}_h^* = 0$, and, if $\gamma_0 > 0$, $u_h^* = 0$. Instead, if $\gamma_0 = 0$, the first equation of (3.161) with $\underline{\tau}_h \in \operatorname{Ker} D_h$ reduces to

$$b_h(u_h^*, \underline{\tau}_h) = 0 \quad \forall \underline{\tau}_h \in \operatorname{Ker} D_h\,. \tag{3.163}$$

Thanks to (3.62) and (3.157), this implies $u_h^* = 0$. Hence, we are left with the equation

$$d_h(\lambda_h^*, \underline{\tau}_h) = 0 \qquad \forall \underline{\tau}_h \in \widehat{\Sigma}_h\,, \tag{3.164}$$

which implies $\lambda_h^* = 0$, due to (3.158). Therefore, uniqueness is proved.
The third equation of problem (3.161) gives $\widehat{\underline{\sigma}}_h \in \operatorname{Ker} D_h$, and hence, due to (3.157),

$$\widehat{\underline{\sigma}}_h \in \Sigma_h\,. \tag{3.165}$$

Notice that for $\underline{\tau}_h \in \Sigma_h$ the bilinear form $d_h(\lambda_h, \underline{\tau}_h)$ simplifies to

$$d_h(\lambda_h, \underline{\tau}_h) = \sum_{e \subset \Gamma_D} \int_e \lambda_h \underline{\tau}_h \cdot \underline{n}\, ds = < g, \underline{\tau}_h \cdot \underline{n} >_{|\Gamma_D}\,, \tag{3.166}$$

since $\lambda_h \in \Lambda_{h,g}$. Hence, the first equation of (3.161) with $\underline{\tau}_h \in \Sigma_h$ reads

$$a(\widehat{\underline{\sigma}}_h, \underline{\tau}_h) + b_h(\widehat{u}_h, \underline{\tau}_h) =< g, \underline{\tau}_h \cdot \underline{n} >_{|\Gamma_D} \quad \forall \underline{\tau}_h \in \Sigma_h \,. \tag{3.167}$$

Collecting (3.165), (3.167) and the second equation of problem (3.161) we see that $(\widehat{\underline{\sigma}}_h, \widehat{u}_h)$ is solution of problem (3.58). □

From now on the superscript $\widehat{\ }$ will be dropped from the variables.

A projection P_h^e from $L^2(\mathcal{E}_h)$ to Λ_h can be defined locally as the usual orthogonal L^2-projection

$$\int_e \mu_h (u - P_h^e u)\, ds = 0, \qquad \forall e \subset \partial K, \quad \forall \mu_h \in R(e)\,. \tag{3.168}$$

Let us give now the abstract error estimate for the Lagrange multipliers.

Theorem 3.6 *Let $(\underline{\sigma}, u)$ be the solution of* (3.27)*, and $(\underline{\sigma}_h, u_h, \lambda_h)$ that of* (3.161)*. The following estimate holds*

$$||\lambda_h - P_h^e u||_{1/2,h} \le C \left(||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + ||P_h u - u_h||_{1,h} + \sup_{\underline{\tau} \in \widehat{\underline{\Sigma}}_h} \frac{d_h(u - P_h^e u, \underline{\tau})}{||\underline{\tau}||_h} \right), \tag{3.169}$$

where $||\cdot||_{1,h}$ is defined as

$$||v_h||_{1,h}^2 = \sum_K h_K^{-2} ||v_h||_{0,K}^2, \qquad \forall v_h \in V_h, \tag{3.170}$$

and P_h is defined in (3.71)*.*

Proof. Taking $\mu = \lambda_h - P_h^e u$, from (3.158) we deduce

$$\begin{aligned} \delta ||\lambda_h - P_h^e u||_{1/2,h} &\le \sup_{\underline{\tau} \in \widehat{\underline{\Sigma}}_h} \frac{d_h(\lambda_h - P_h^e u, \underline{\tau})}{||\underline{\tau}||_h} \\ &\le \sup_{\underline{\tau} \in \widehat{\underline{\Sigma}}_h} \frac{d_h(\lambda_h - u, \underline{\tau})}{||\underline{\tau}||_h} + \sup_{\underline{\tau} \in \widehat{\underline{\Sigma}}_h} \frac{d_h(u - P_h^e u, \underline{\tau})}{||\underline{\tau}||_h}. \end{aligned} \tag{3.171}$$

Next, let us observe that $(\underline{\sigma}, u)$ solution of (3.27) verifies

$$a(\underline{\sigma}, \underline{\tau}) + b_h(u, \underline{\tau}) - d_h(u, \underline{\tau}) = 0 \qquad \forall \underline{\tau} \in \widehat{\underline{\Sigma}}_h. \tag{3.172}$$

Then, subtracting the first equation of (3.161) from (3.172) and using definition (3.71) we obtain

$$d_h(u - \lambda_h, \underline{\tau}) = a(\underline{\sigma} - \underline{\sigma}_h, \underline{\tau}) + b_h(P_h u - u_h, \underline{\tau}) \qquad \forall \underline{\tau} \in \widehat{\underline{\Sigma}}_h. \tag{3.173}$$

Using (3.53) and the Cauchy-Schwarz inequality

$$\begin{aligned} b_h(P_h u - u_h, \underline{\tau}) &\le \sum_K h_K ||\mathrm{div}\, \underline{\tau}||_{0,K}\, h_K^{-1} ||P_h u - u_h||_{0,K} \\ &\le ||\underline{\tau}||_h \Big(\sum_K h_K^{-2} ||P_h u - u_h||_{0,K}^2\Big)^{1/2} = ||\underline{\tau}||_h ||P_h u - u_h||_{1,h}\,. \end{aligned} \tag{3.174}$$

Hence, from (3.173), (3.174) we have

$$\sup_{\underline{\tau}\in\widehat{\Sigma}_h} \frac{d_h(\lambda_h - u, \underline{\tau})}{||\underline{\tau}||_h} \leq C(||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + ||P_h u - u_h||_{1,h}). \tag{3.175}$$

Using (3.175) in (3.171) gives (3.169). □

Corollary 3.1 *Under the same assumptions as in Theorem 3.6, if the decomposition $\mathcal{T}_h$ is quasi-uniform, we have*

$$||\lambda_h - P_h^e u||_{0,\mathcal{E}_h} \leq C(h^{1/2}||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + h^{-1/2}||P_h u - u_h||_{0,\Omega}) + h^{1/2} \sup_{\underline{\tau}\in\widehat{\Sigma}_h} \frac{d_h(u - P_h^e u, \underline{\tau})}{||\underline{\tau}||_h}). \tag{3.176}$$

Proof. The result follows multiplying (3.169) by $h^{1/2}$ and using the quasi-uniformity assumption. □

Remark 3.5 *In all practical cases, the space $\widehat{\Sigma}_h$ will be made of piecewise polynomials of some given degree. This implies that the following inequality will easily hold*

$$||\underline{\tau} \cdot \underline{n}||_{0,e}^2 \leq C h_K^{-1} ||\underline{\tau}||_{0,K}^2 \qquad \forall \underline{\tau} \in \widehat{\Sigma}_h,\ \forall K,\ \forall e \subset \partial K, \tag{3.177}$$

where C denotes a constant depending on the degree of polynomials. In these cases we have, using (3.177) *and Cauchy-Schwarz inequality,*

$$\frac{d_h(u - P_h^e u, \underline{\tau})}{||\underline{\tau}||_h} \leq C||u - P_h^e u||_{1/2,h}, \tag{3.178}$$

so that (3.169) *becomes*

$$||\lambda_h - P_h^e u||_{1/2,h} \leq C(||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + ||P_h u - u_h||_{1,h} + ||u - P_h^e u||_{1/2,h}). \tag{3.179}$$

Moreover, if the decomposition is quasi-uniform, (3.176) *becomes*

$$||\lambda_h - P_h^e u||_{0,\mathcal{E}_h} \leq C(h^{1/2}||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + h^{-1/2}||P_h u - u_h||_{0,\Omega} + h^{1/2}||u - P_h^e u||_{1/2,h}). \tag{3.180}$$

Remark 3.6 *The above estimates on the multipliers, as* (3.169) *and* (3.176)*, or* (3.179) *and* (3.180)*, could be used in order to estimate the error for suitable lifting of λ_h. For instance, if λ_h is piecewise constant on the edges, one can introduce a new approximation u_h^* to u defined as the P_1-nonconforming function that, on each edge e, verifies*

$$\int_e (u_h^* - \lambda_h)\, ds = 0. \tag{3.181}$$

Similarly, one can introduce the P_1-nonconforming interpolant of u, that we denote by u_I^*, defined as

$$\int_e (u_I^* - u)\, ds = 0. \tag{3.182}$$

It is immediate to check that, for each $K \in \mathcal{T}_h$

$$||u_h^* - u_I^*||_{0,K} \leq C \sum_{e \subset \partial K} h_e^{1/2} ||\lambda_h - P_h^e u||_{0,e}\,, \tag{3.183}$$

that allows to use the above estimates in order to obtain suitable bounds for $||u_h^* - u_I^*||_{0,\Omega}$ *(and then for* $||u - u_h^*||_{0,\Omega}$ *by the triangle inequality).*

We shall discuss now how the hybridization of the mixed formulation, presented above in the abstract framework, can be applied to the examples of Section 3.5. For this, we associate with each pair $(Q(K), P(K))$ suitable spaces of polynomials defined on the edges $\mathcal{E}_h$.

Example 1 For any integer $k \geq 0$, and for any triangle $K \in \mathcal{T}_h$, define

$$k \geq 0 \qquad \left\{ \begin{array}{lcl} Q(K) & = & (P_k(K))^2 + \underline{x} P_k(K)\,, \\ P(K) & = & P_k(K)\,, \\ R(e) & = & P_k(e)\,, \quad \forall e \text{ edge of } K. \end{array} \right. \tag{3.184}$$

Example 2 For any integer $k \geq 0$, and for any rectangle $K \in \mathcal{T}_h$, define

$$k \geq 0 \qquad \left\{ \begin{array}{lcl} Q(K) & = & (Q_k(K) + x_1 Q_k(K)) \times (Q_k(K) + x_2 Q_k(K))\,, \\ P(K) & = & Q_k(K)\,, \\ R(e) & = & P_k(e)\,, \quad \forall e \text{ edge of } K. \end{array} \right. \tag{3.185}$$

Example 3 For any integer $k \geq 1$, and for any triangle $K \in \mathcal{T}_h$, define

$$k \geq 1 \qquad \left\{ \begin{array}{lcl} Q(K) & = & (P_k(K))^2\,, \\ P(K) & = & P_{k-1}(K)\,, \\ R(e) & = & P_k(e)\,, \quad \forall e \text{ edge of } K. \end{array} \right. \tag{3.186}$$

Example 4 For any integer $k \geq 1$, and for any rectangle $K \in \mathcal{T}_h$, define

$$k \geq 1 \qquad \left\{ \begin{array}{lcl} Q(K) & = & (P_k(K))^2 \oplus \left\{ \underline{curl}(x_1 x_2^{k+1}) \right\} \oplus \left\{ \underline{curl}(x_2 x_1^{k+1}) \right\}\,, \\ P(K) & = & P_{k-1}(K)\,, \\ R(e) & = & P_k(e)\,, \quad \forall e \text{ edge of } K. \end{array} \right. \tag{3.187}$$

Example 5 For each triangle $K \in \mathcal{T}_h$, define

$$\left\{ \begin{array}{lcl} Q(K) & = & span\,\{\underline{\tau}^1, \underline{\tau}^2, \underline{\tau}^3\}\,, \text{ (see (3.114)-(3.116)),} \\ P(K) & = & P_0(K)\,, \\ R(e) & = & P_0(e)\,, \quad \forall e \text{ edge of } K. \end{array} \right. \tag{3.188}$$

Example 6 For each triangle $K \in \mathcal{T}_h$, define

$$\left\{ \begin{array}{lcl} Q(K) & = & span\,\{\underline{\tau}^1, \underline{\tau}^2, \underline{\tau}^3\}\ , \text{ (see (3.127)-(3.129))}, \\ P(K) & = & P_0(K)\,, \\ R(e) & = & P_0(e)\,, \quad \forall\, e \text{ edge of } K. \end{array} \right. \tag{3.189}$$

Denoting by $R(\partial K) = \prod_{e \in \partial K} R(e)$, we see that all the triplets $(Q(K),\ P(K),\ R(\partial K))$ of the examples above are such that Ker D_h coincides with the space Σ_h used in Problem (3.58), so that assumption (3.157) is fulfilled. Indeed, for *Example 6* definition (3.132) of Σ_h clearly coincides with that of (3.155), as it can be easily seen using (3.156). For *Examples 1-5* the space Σ_h is defined in (3.88), and $\underline{\tau}_h \in \Sigma_h$ *iff* $\underline{\tau}_h \in \widehat{\Sigma}_h$ and $[\underline{\tau} \cdot \underline{n}]_{|e} = 0$, $\forall\, e \in \mathcal{E}_h'$, that is, $\underline{\tau} \cdot \underline{n}$ is continuous across the internal edges and $\underline{\tau} \cdot \underline{n} = 0$ on Γ_N. Moreover, we have that $\underline{\tau} \in Q(K) \Longrightarrow \underline{\tau} \cdot \underline{n}_{|\partial K} \in R(\partial K)$, and in particular, $[\underline{\tau} \cdot \underline{n}]_{|e} \in R(e)$. Then, according to (3.156), $\underline{\tau} \in Ker\, D_h$ implies that $[\underline{\tau} \cdot \underline{n}]_{|e} = 0$, $\forall\, e \in \mathcal{E}_h'$. Therefore, $Ker\, D_h = \Sigma_h$.

We check now that the *inf-sup* condition (3.158) holds for the bilinear form $d_h(.,.)$ in all the examples.

Proposition 3.5 *For the mixed finite elements of Examples 1-6 it holds*

$$\exists\ \delta > 0 : \quad \inf_{\mu_h \in \Lambda_h} \sup_{\underline{\tau}_h \in \widehat{\Sigma}_h} \frac{d_h(\mu_h, \underline{\tau}_h)}{||\mu_h||_{1/2,h}\ ||\underline{\tau}_h||_h} \geq\ \delta. \tag{3.190}$$

Proof. Let us consider $\mu^* \in \Lambda_h$. Let e be an edge in $\mathcal{E}_h$ and let $K = K(e)$ be an element of $\mathcal{T}_h$ having e as an edge. Take $\underline{\tau}^e \in \widehat{\Sigma}_h$ such that $\underline{\tau}^e = 0$ in $\Omega \setminus K$, and in K is the vector $\underline{\tau}^e \in Q(K)$ which satisfies

$$\int_e \mu\, \underline{\tau}^e \cdot \underline{n}\, ds \ = \ h_e^{-1} \int_e \mu^*\, \mu\, ds \quad \forall \mu \in R(e)\,, \tag{3.191}$$

$$\int_{\partial K \setminus e} \mu\, \underline{\tau}^e \cdot \underline{n}\, ds \ = \ 0 \quad \forall \mu \in R(\partial K)\,, \tag{3.192}$$

$$\int_K \underline{\tau}^e \cdot \underline{\nabla}\, v dx \ = \ 0 \quad \forall v \in P(K)\,, \tag{3.193}$$

$$\int_K \underline{\tau}^e \cdot \underline{\varphi}\, dx \ = \ 0 \quad \forall \underline{\varphi} \in D\,, \tag{3.194}$$

where $D = \left\{ \underline{q} \in Q(K) \mid \mathrm{div}\, \underline{q} = 0,\ \underline{q} \cdot \underline{n}_{|\partial K} = 0 \right\}$ as defined in (3.92). Such a vector $\underline{\tau}^e$ exists and is unique in $Q(K)$ since (3.191)-(3.194) are a set of degrees of freedom for all $Q(K)$ considered here, as discussed in Section 3.5. Notice that the degrees of freedom (3.191), (3.192) are always present and allow us to impose conditions on the normal component of vectors in $Q(K)$; (3.193) and (3.194) are needed for higher order elements.
A usual scaling argument gives

$$||\underline{\tau}^e||^2_{0,K} + h_K^2 ||\mathrm{div}\, \underline{\tau}^e||^2_{0,K} \leq C h_e^{-1} ||\mu^*||^2_{0,e}\,. \tag{3.195}$$

Define now $\underline{\tau}^* = \sum_e \underline{\tau}^e$. Since $\underline{\tau}^*$ on a single element K is the sum of at most three $\underline{\tau}^e$, summing (3.195) over all $K \in \mathcal{T}_h$ gives

$$||\underline{\tau}^*||_h \leq C(\sum_e h_e^{-1}||\mu^*||_{0,e}^2)^{1/2} = C||\mu^*||_{1/2,h}\,. \tag{3.196}$$

Using the definition of $\underline{\tau}^*$ and (3.191) we obtain

$$\int_e \mu\,[\underline{\tau}^* \cdot \underline{n}]\,ds = \int_e \mu\,[\underline{\tau}^e \cdot \underline{n}]\,ds = \int_e \mu\,\underline{\tau}^e \cdot \underline{n}\,ds = h_e^{-1}\int_e \mu^*\,\mu\,ds, \tag{3.197}$$

and consequently we have

$$d_h(\underline{\tau}^*, \mu) = \sum_e \int_e \mu\,[\underline{\tau}^* \cdot \underline{n}]\,ds = \sum_e h_e^{-1}\int_e \mu^*\,\mu\,ds\,. \tag{3.198}$$

Hence, taking $\mu = \mu^*$ in (3.198) and using (3.196) we have

$$\frac{d_h(\underline{\tau}^*, \mu^*)}{||\underline{\tau}^*||_h} \geq C^{-1}\frac{\sum_e h_e^{-1}||\mu^*||_{0,e}^2}{(\sum_e h_e^{-1}||\mu^*||_{0,e}^2)^{1/2}} = C^{-1}||\mu^*||_{1/2,h}\,, \tag{3.199}$$

and the proof is concluded with $\delta = C^{-1}$. □

Due to Proposition 3.5 the abstract error estimate (3.169) holds, (together with (3.176), if the decomposition is quasi–uniform). Notice that, when conforming approximations are considered, the term $d_h(u - P_h^e u, \underline{\tau})$, with $\underline{\tau} \in \widehat{\Sigma}_h$, vanishes. This is the case for all the *Examples 1-5*. In particular, for the RT_0 element, using (3.110) in (3.176) we obtain

$$||\lambda_h - P_h^e u||_{0,\mathcal{E}_h} \leq C(h^{1/2}||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + h^{-1/2}||P_h u - u_h||_{0,\Omega}) \leq C\,h^{1/2}. \tag{3.200}$$

The same estimate applies to the element of *Example 5*, using (3.124) in (3.176). For the nonconforming element of *Example 6* we give the following

Proposition 3.6 *Let* $\widehat{\Sigma}_h$ *and* Λ_h *be the spaces defined in* (3.148) *and* (3.150) *associated with the element* (3.189). *For* $w \in H^1(\Omega)$ *it holds*

$$\sup_{\underline{\tau} \in \widehat{\Sigma}_h} \frac{d_h(w - P_h^e w, \underline{\tau})}{||\underline{\tau}||_h} \leq C|w|_{1,\Omega}, \tag{3.201}$$

where P_h^e *is defined in* (3.168).

Proof. We can use a scaling argument on each edge $e \in \mathcal{E}_h$ and obtain for $\underline{\tau} \in \widehat{\Sigma}_h$

$$\int_e (w - P_h^e w)\,\underline{\tau} \cdot \underline{n}\,ds \leq C h_K^{1/2}|w|_{1,K}||\underline{\tau} \cdot \underline{n}||_{0,e} \leq C|w|_{1,K}||\underline{\tau}||_{0,K}\,. \tag{3.202}$$

Then, (3.201) easily follows. □

Consequently, using (3.147) and (3.201) in (3.176) we deduce

$$||\lambda_h - P_h^e u||_{0,\mathcal{E}_h} \leq C(h^{1/2}||\underline{\sigma} - \underline{\sigma}_h||_{0,\Omega} + h^{-1/2}||P_h u - u_h||_{0,\Omega} + h^{1/2}|u|_{1,\Omega}) \leq C\,h^{1/2}. \tag{3.203}$$

Remark 3.7 *The order* $\mathcal{O}(h^{1/2})$ *in estimates* (3.200), (3.203) *is not in disagreement with the estimates for the scalar variable* u. *Indeed,* (3.200) *(or* (3.203)*) in* (3.183) *gives back the* $\mathcal{O}(h)$ *order of convergence proved for these elements. Moreover, for the* RT_0 *element we can use the superconvegence result (see [55])*

$$||P_h u - u_h||_{0,\Omega} \leq Ch^2, \tag{3.204}$$

thus obtaining in (3.200)

$$||\lambda_h - P_h^e u||_{0,\mathcal{E}_h} \leq C\, h^{3/2}. \tag{3.205}$$

This, in turns, using (3.183), *gives*

$$||u_h^* - u_I^*||_{0,\Omega} \leq C\, h^2, \tag{3.206}$$

which is an optimal error bound. Estimate (3.204) *(and then, in the end,* (3.206)*) can also be proved for the lowest order elements of the families of the first four examples presented here. Actually, one can consider other values of* k *as well, obtaining better convergence estimates, analogous to* (3.204), (3.205), *and finally* (3.206), *for suitably defined higher order interpolants (instead of* u_I^**) and liftings (instead of* u_h^**). We refer for that e.g. to [10], [26]. Finally, we recall that for all the examples presented here estimates for* $||\lambda_h - P_h^e u||_{0,\mathcal{E}_h}$ *can be obtained directly, case by case, dropping the quasi–uniformity assumption on the mesh. We refer to [10], [26] and [98] for the proofs.*

3.7 Algebraic treatment of problem (3.161)

Problem (3.161) is easier to deal with than (3.75). Indeed, the linear system associated with (3.161) takes the matrix form

$$\begin{pmatrix} A & B & -D \\ B^t & -C & 0 \\ -D^t & 0 & 0 \end{pmatrix} \begin{pmatrix} \underline{\sigma}_h \\ u_h \\ \lambda_h \end{pmatrix} = \begin{pmatrix} 0 \\ -F \\ 0 \end{pmatrix}, \tag{3.207}$$

with obvious meaning of the notation. The matrix A is now a block-diagonal matrix, each block being a matrix of dimension equal to the dimension of $Q(K)$, easy to invert. Hence, the variable $\underline{\sigma}_h$ can be eliminated by static condensation,

$$\underline{\sigma}_h = A^{-1}(D\lambda_h - Bu_h), \tag{3.208}$$

leading to the new system

$$\begin{pmatrix} B^tA^{-1}B + C & -B^tA^{-1}D \\ -D^tA^{-1}B & D^tA^{-1}D \end{pmatrix} \begin{pmatrix} u_h \\ \lambda_h \end{pmatrix} = \begin{pmatrix} F \\ 0 \end{pmatrix}. \tag{3.209}$$

Since no continuity assumption is made on V_h, the matrix $B^tA^{-1}B + C$ is also block-diagonal, so that the variable u_h can be eliminated by static condensation,

$$u_h = (B^tA^{-1}B + C)^{-1}(F + B^tA^{-1}D\lambda_h). \tag{3.210}$$

This leads to a final system, acting on the unknown λ_h only, of the form

$$\mathcal{M}\lambda_h = \mathcal{G}, \tag{3.211}$$

where $\mathcal{M}$ and $\mathcal{G}$ are given by

$$\mathcal{M} = D^tA^{-1}D - D^tA^{-1}B(B^tA^{-1}B + C)^{-1}B^tA^{-1}D, \tag{3.212}$$

$$\mathcal{G} = D^tA^{-1}B(B^tA^{-1}B + C)^{-1}F. \tag{3.213}$$

We present in detail the structure of the matrix $\mathcal{M}$ and of the right–hand side $\mathcal{G}$ defined in (3.212) and (3.213). Since for applications to semiconductor device simulation the low regularity of the solution makes the use of high order elements unsuitable, we discuss here only the case of lowest order elements, namely, the lowest order Raviart–Thomas element on triangles (RT_0) and the two elements of *Example 5-6*. We construct the element matrix $\mathcal{M}^K$ and the element right–hand side $\mathcal{G}^K$ associated with the current element K. In order to give a compact presentation, we set some common notation and recall the definition of the triplet $(Q(K), P(K), R(\partial K))$ for the three cases under consideration.

$$\begin{aligned} Q(K) &= span\left\{\underline{\tau}^1, \underline{\tau}^2, \underline{\tau}^3\right\}, \\ P(K) &= P_0(K), \\ R(\partial K) &= \prod_{e\in\partial K} P_0(e), \end{aligned} \tag{3.214}$$

where

$$\underline{\tau}^1 = (1,0)\,,\ \underline{\tau}^2 = (0,1)\,,\ \underline{\tau}^3 = (\psi_1, \psi_2)\,. \tag{3.215}$$

The choice for ψ_1 and ψ_2 makes the difference of the three cases. For RT_0 we take

$$\underline{\tau}^3 = \underline{x} - \underline{x}_B\,, \qquad \underline{x}_B = \text{ coordinates of the centroid of } K. \tag{3.216}$$

The reason for this choice is that $\int_K \psi_1\,dx = \int_K \psi_2\,dx = 0$, and therefore each block of A becomes diagonal. We refer to the definition (3.115)–(3.116) for the element of *Example 5* and to (3.128)–(3.129) for the element of *Example 6*.

As basis function in $P_0(K)$ we make the natural choice $v = 1$ in K, and as basis function in $R(\partial K)$ we take $\mu = 1$ on one edge e and $\mu = 0$ on the others.

The functions $a^{-1}(x)$ and $\gamma(x)$ appearing in the bilinear forms (3.24) and (3.26) are approximated by piecewise constant functions defined in each element K by

$$\overline{\alpha} := \frac{1}{|K|}\left(\int_K a^{-1}(x)\,dx\right), \tag{3.217}$$

$$\overline{\gamma} := \frac{1}{|K|}\left(\int_K \gamma(x)\,dx\right). \tag{3.218}$$

We introduce the following notation

$$\underline{\nu}^i = \underline{n}^i |e^i| \qquad i = 1,3, \tag{3.219}$$

$$\delta = \int_K (\psi_1^2 + \psi_2^2)\, dx, \tag{3.220}$$

$$\beta = \int_K \operatorname{div} \underline{\tau}^3 dx, \tag{3.221}$$

$$\eta_i = \int_{e^i} \underline{\tau}^3 \cdot \underline{n}\, ds \qquad i = 1,3. \tag{3.222}$$

Then the element matrices are

$$A^K = \overline{\alpha} \begin{pmatrix} |K| & 0 & 0 \\ 0 & |K| & 0 \\ 0 & 0 & \delta \end{pmatrix} \qquad B^K = \begin{pmatrix} 0 \\ 0 \\ \beta \end{pmatrix} \tag{3.223}$$

$$D^K = \begin{pmatrix} \nu_1^1 & \nu_1^2 & \nu_1^3 \\ \nu_2^1 & \nu_2^2 & \nu_2^3 \\ \eta_1 & \eta_2 & \eta_3 \end{pmatrix} \qquad C^K = \overline{\gamma}|K| \qquad F^K = \int_K f\, dx. \tag{3.224}$$

The coefficients of the matrix $\mathcal{M}^K$ are then given by

$$m_{ij}^K = (\overline{\alpha})^{-1} \frac{\underline{\nu}^i \cdot \underline{\nu}^j}{|K|} + \frac{\overline{\gamma}|K|}{\beta^2 + \overline{\alpha}\,\delta\,\overline{\gamma}|K|} \eta_i \eta_j \qquad i,j = 1,3, \tag{3.225}$$

and the right-hand side by

$$g_i^K = \frac{\beta \eta_i}{\beta^2 + \overline{\alpha}\,\delta\,\overline{\gamma}|K|} \int_K f\, dx \qquad i = 1,3. \tag{3.226}$$

If $\gamma(x) = 0$ in (2.14), the coefficients of $\mathcal{M}^K$ reduce to

$$m_{ij}^K = (\overline{\alpha})^{-1} \frac{\underline{\nu}^i \cdot \underline{\nu}^j}{|K|}. \tag{3.227}$$

The matrix $\mathcal{M}$ thus corresponds to a nonconforming piecewise linear approximation of (2.14), where the function $a(x)$ is approximated by its harmonic average:

$$a(x)_{|K} \simeq \frac{|K|}{\int_K a^{-1}(x)\, dx}. \tag{3.228}$$

The final matrix $\mathcal{M}$ is then symmetric and positive definite. Moreover, if the decomposition is of weakly acute type (i.e., every angle θ of every triangle is $\theta \leq \pi/2$), then

$$m_{ii}^K > 0, \qquad m_{ij}^K \leq 0, \qquad i \neq j,\ i,j = 1,3. \tag{3.229}$$

Hence, in this case, $\mathcal{M}$ is an $M-$matrix. In the general case $\gamma(x) \geq 0$, we see that the quantity $\overline{\gamma}|K|/(\beta^2 + \overline{\alpha}\,\delta\,\overline{\gamma}|K|)$ is always nonnegative. As far as the η_i are concerned, we have to discuss separately what happens in the three cases considered above. For the RT_0 element, a simple computation shows that

$$\eta_i = \frac{2|K|}{3} > 0 \qquad i = 1, 3. \tag{3.230}$$

Therefore, the off-diagonal coefficients of $\mathcal{M}^K$ might be positive even if the triangulation is weakly acute, and the final matrix is not, in general, an $M-$matrix.
Instead, the other two elements are designed to guarantee a final $M-$matrix. Indeed, to fix ideas, let e^1 be the special edge ($\widetilde{e}$ in the definitions (3.116) or (3.129)). Then,

$$\widetilde{e} = e^1 \quad \Longrightarrow \qquad \eta_1 > 0, \qquad \eta_2 = \eta_3 = 0, \tag{3.231}$$

with $\eta_1 = |e^1|$ for *Example 5*, and $\eta_1 = 1$ for *Example 6*. Incidentally we also notice that, for both elements, definition (3.221) gives $\beta = \eta_1$. Hence, the coefficients of $\mathcal{M}^K$ are given by

$$m_{ij}^K = \begin{cases} (\overline{\alpha})^{-1}\dfrac{\underline{\nu}^1 \cdot \underline{\nu}^1}{|K|} + \dfrac{\overline{\gamma}|K|}{\beta^2 + \overline{\alpha}\,\delta\,\overline{\gamma}|K|}\eta_1^2 & \text{for } i = j = 1, \\ (\overline{\alpha})^{-1}\dfrac{\underline{\nu}^i \cdot \underline{\nu}^j}{|K|} & \text{otherwise,} \end{cases} \tag{3.232}$$

and the right-hand side is given by

$$g_i^K = \begin{cases} \dfrac{\eta_1^2}{\beta^2 + \overline{\alpha}\,\delta\,\overline{\gamma}|K|}\displaystyle\int_K f\,dx & \text{for } i = 1 \\ 0 & \text{otherwise .} \end{cases} \tag{3.233}$$

It is then clear that the final matrix is, for both elements, always an $M-$matrix, if the decomposition is of weakly acute type, since the zeroth–order term gives contribution, with the positive sign, only to the coefficient m_{11}.

Another approach which yields the $M-$matrix property for the RT_0 element, even in the case $\gamma(x) \geq 0$, will be examined in the following Section 3.8.

3.8 Numerical quadrature: towards finite volumes

In this section we analyze the family of mixed finite volume methods proposed in [115] for the approximation of the reaction-diffusion problem (2.14). All of the methods are a suitable discretization of the dual mixed formulation (3.27) and employ the lowest-order Raviart-Thomas (RT_0) finite element spaces (3.107)-(3.108) plus a suitable quadrature formula for the matrix corresponding to $a(\underline{\sigma}_h, \underline{\tau}_h)$. This allows the use of different averages of the inverse diffusion coefficient a^{-1} to enforce the constitutive law $(3.3)_1$ for the fluxes at the interelement boundaries in a finite volume fashion.

The mixed finite volume formulation (MFV) addressed in this section is to be viewed as an alternative approach to the hybridization procedure discussed in Section 3.6, and has been the object of several researches in the recent literature. The central issue of the MFV formulation is to perform a *lumping* of the matrix A associated with $a(\underline{\sigma}_h, \underline{\tau}_h)$ in (3.58) through some suitable quadrature formula. In the case of rectangular grids, this strategy has been first proposed in [129] in the case $a(x) = 1$. Theoretical and implementational issues that link "nodal" finite elements, mesh-centered finite differences and mixed-hybrid finite elements have been addressed in [75, 76]. In [124] a theoretical analysis shows that, under appropriate smoothness assumptions, the quadrature error (in the evaluation of A and of the right-hand side in (3.58)) does not spoil the accuracy of the mixed method with exact integration.

In the case of triangular elements, similar conclusions have been drawn in [18, 19], where a quadrature formula to diagonalize the matrix A is proposed and examined in the case of the Laplace operator on triangular grids. An extension of this latter lumping procedure has been carried out in [2, 143, 144, 114] for diffusion and convection-diffusion problems. In particular, in [2] $a(x)$ is approximated by its harmonic average over each triangle. In this section we analyze a family of averages, all characterized by being piecewise constant over the dual tessellation of the domain. Moreover we also include in the analysis the zeroth-order term.

The case of an elliptic problem where $a(x)$ is a symmetric positive definite tensor has been studied in [9, 35], where RT_0 mixed finite elements with numerical integration are considered on both rectangular and logically rectangular grids. Another numerical scheme on this subject has been proposed in [8], where an expanded mixed finite element method capable to handle the case of a discontinuous tensor $a(x)$ and general shape elements and geometry is derived. It can be checked that in the special case of a reference equilateral triangle and RT_0 finite elements, the quadrature rule proposed in [8] to diagonalize the mass matrix yields the same result as the lumping formula of [18]. However, the resulting method in [8] gives a ten-point finite difference stencil while the method we are going to analyze in this section gives a four-point finite difference stencil.

The common feature of the approaches based on a lumping procedure is that the mixed formulation using RT_0 finite elements can be interpreted as a finite volume method acting on the scalar unknown u_h. This connection has been also recently investigated in [57], where cell-centered finite difference schemes are constructed on triangular grids of regular shape (equilateral, and isosceles right triangles). In this respect we also mention [162] where finite difference schemes on triangular cell-centered grids are derived under the assumption of weakly acute triangulation. In this case, and using the harmonic average of $a(x)$ along the Voronoi edge, the methods of [162] and of the present section coincide, although the error analysis in [162] is carried out only in the case $a(x) = 1$.

Let us now give a brief outline of the contents of this section. The dual tessellation $\mathcal{L}_h$ associated with $\mathcal{T}_h$ is first introduced in Section 3.8.1. Notice that $\mathcal{T}_h$ is only re-

quired to be a Delaunay triangulation, while previous schemes in the literature assume that $\mathcal{T}_h$ must be a weakly acute triangulation (see [30, 98, 17, 162]). The Delaunay property allows for the presence of obtuse triangles in the mesh while in a weakly acute triangulation all the angles are required to be $\leq \pi/2$. Next, the mixed finite volume schemes are derived in Section 3.8.2. The methods are based on the combined use of a piecewise constant approximation $\overline{\alpha}$ of $\alpha \equiv a^{-1}$ over $\mathcal{L}_h$ and of the quadrature formula proposed in [18]. The resulting discretization is a cell-centered finite volume scheme where the degrees of freedom for u_h are taken at the circumcenters of each element of $\mathcal{T}_h$, while the interelement fluxes are computed using the values of u_h at the circumcenters of two neighboring triangles of $\mathcal{T}_h$.

The error analysis is carried out in Section 3.8.3, where we proved the (optimal) $\mathcal{O}(h)$ convergence of this family of methods with respect to the $H(\mathrm{div};\Omega) \times L^2(\Omega)$-norm, which is the standard norm for the analysis of dual mixed methods. It is worth noting that the derivation of the methods, as well as their convergence analysis, allows us in every respect to call them "mixed finite volume schemes" or, equivalently, "mixed finite element schemes with numerical integration". In Section 3.8.4 we consider three choices of $\overline{\alpha}$: two averages are suitable approximations of the *harmonic average* of a (see [12]) while the third one is the trapezoidal rule.

3.8.1 Geometry and notation

In view of the mixed finite volume discretization of problem (3.27) we introduce the following

Definition 3.1 *$\mathcal{T}_h$ is a Delaunay triangulation if, for every $K \in \mathcal{T}_h$, the closed circumcircle of K contains no other vertices than those belonging to K. Moreover, $\mathcal{T}_h$ is a degenerate Delaunay mesh when the above property only holds for the open circumcircle. (cf. [53]).*

We assume henceforth that $\mathcal{T}_h$ is a Delaunay triangulation. Moreover, we shall indicate by $\mathtt{N_E}$ and $\mathtt{N_T}$ the total number of edges and triangles of $\mathcal{T}_h$, respectively. Throughout this section, any geometrical entity will be always understood as being an open bounded subset of $\mathbb{R}^2$ or $\mathbb{R}$.

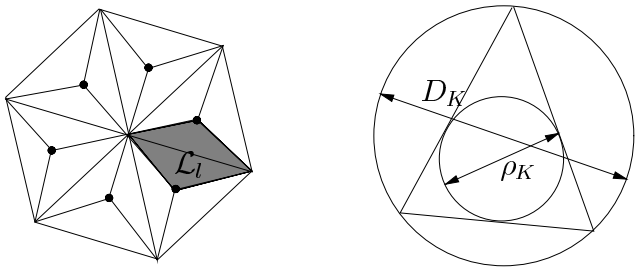


Figure 3: Primal triangulation $\mathcal{T}_h$ with the corresponding lumping regions $\mathcal{L}^l$ (left), mesh parameters (right).

For any pair of neighboring triangles K_i and K_j of $\mathcal{T}_h$, $i, j = 1, \ldots, \mathtt{N_T}$, let l be the common edge. We also consider the dual tessellation $\mathcal{L}_h$ of $\mathcal{T}_h$ and denote its elements by "lumping regions" (see Figure 3, left). The lumping region $\mathcal{L}^l$ corresponding to edge l is obtained by joining the common vertices and the two circumcenters C_i and C_j (see Figure 4). We set also $\mathcal{L}^l_k = \mathcal{L}^l \cap K_k$, for $k = i,\ j$.

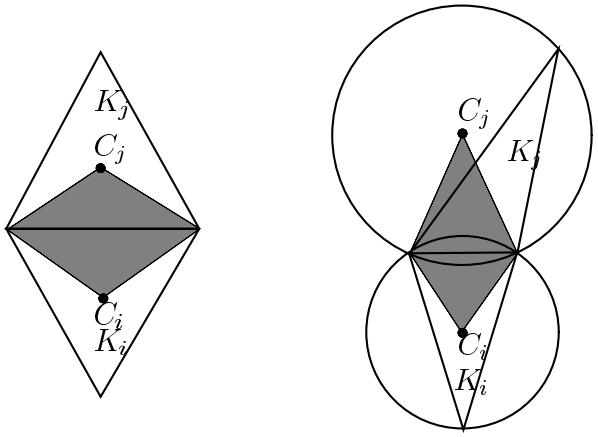


Figure 4: Examples of lumping regions for acute (left) and obtuse (right) triangles.

Throughout this section it is understood that any geometrical entity referred to by a superscript, say r, and/or a subscript, say k, has the following meaning: the superscript and the subscript refer to an edge and a triangle, respectively. When both the superscript and the subscript appear, we are referring to an entity depending on the edge e^r that shares the triangle K_k. Moreover, $T(r)$ is the index set of the triangles shared by the edge e^r, and it is understood that the cardinality of $T(r)$ is one or two depending on whether or not the edge $e^r \in \partial\Omega$. By $E(k)$ we denote the index set of the edges of the triangle K_k. For any triangle K_k we have: three indices of the edges of K_k (say r, i, j), three corresponding edges e^r, e^i, e^j and the vectors $\underline{e}^r_k, \underline{e}^i_k, \underline{e}^j_k$

obtained by orienting the boundary of K_k counterclockwise. Observe that $\underline{c}^r_{k'} = -\underline{c}^r_k$ for $k', k \in T(r)$.
We assume the regularity assumptions (3.45)-(3.46) on $\mathcal{T}_h$, namely (see Figure 3, right):

$$\frac{D_K}{\rho_K} \leq \mathcal{K}^* \quad \forall K \in \mathcal{T}_h, \qquad D_K \leq \mathcal{K}^* \rho_K \leq \mathcal{K} h_K \quad \forall K \in \mathcal{T}_h, \tag{3.234}$$

where we recall that $\mathcal{K}$ denotes, here and in the sequel, a generic constant depending only on $\mathcal{K}^*$. It follows that the triangulation $\mathcal{T}_h$ satisfies

$$h_{\mathcal{L}^l} \leq D \leq \mathcal{K} h \qquad \forall \mathcal{L}^l \in \mathcal{L}_h,$$

where $D = \max_K D_K$, $h = \max_K h_K$ and $h_{\mathcal{L}^l}$ is the diameter of $\mathcal{L}^l$ (see Figure 3, left).

3.8.2 A family of finite volume methods

Let us now deal with the saddle-point problem (3.58) using the lowest-order Raviart-Thomas approximation spaces (see (3.107)-(3.109)). For any $K_k \in \mathcal{T}_h$ and for any integrable function φ, define its mean value as

$$\varphi_k = \frac{1}{|K_k|} \int_{K_k} \varphi \, dx \qquad \forall K_k \in \mathcal{T}_h. \tag{3.235}$$

Taking v_h equal to the characteristic function χ_k of the triangle K_k in the second equation of (3.58) we obtain

$$\sum_{r \in E(k)} \int_{e^r} \underline{\sigma}_h \cdot \underline{n}^r_k \, ds - \int_{K_k} \gamma u_h \, dx = - \int_{K_k} f \, dx \qquad \forall K_k \in \mathcal{T}_h, \tag{3.236}$$

where $\underline{n}^r_k$ is the outward unit normal vector to edge e^r. Equation (3.236) can be rewritten as

$$\sum_{r \in E(k)} \Phi^r_k - u_k \gamma_k |K_k| = -f_k |K_k| \qquad \forall K_k \in \mathcal{T}_h, \tag{3.237}$$

where Φ^r_k is the outward flux of $\underline{\sigma}_h$ from the triangle K_k through the edge e^r:

$$\Phi^r_k := \int_{e^r} \underline{\sigma}_h \cdot \underline{n}^r_k \, ds, \quad \text{with } \Phi^r_k = 0 \text{ on } e^r \in \Gamma_N. \tag{3.238}$$

Clearly, (3.237) is a genuine finite volume discretization for (2.14), provided that we are able to express each flux Φ^r_k in terms of the values $u_j, j \in T(r)$ only. With this aim, let us consider the first equation in (3.58), and denote by e^r an internal edge of K_k and by K_j the element of $\mathcal{T}_h$ such that $k, j \in T(r)$, i.e., $\partial K_k \cap \partial K_j = e^r$. Let $\underline{\tau}^r_h$ be the basis function defined in (3.109), where the normal is taken outward to K_k, so

that $\mathrm{div}\,\underline{\tau}^r_{h|K_k} = 1/|K_k|$ and $\mathrm{div}\,\underline{\tau}^r_{h|K_j} = -1/|K_j|$. Taking $\underline{\tau}_h = \underline{\tau}^r_h$ in the first equation of problem (3.58), we get

$$\int_{K_k} \alpha \underline{\sigma}_h \cdot \underline{\tau}^r_h \, dx + \int_{K_j} \alpha \underline{\sigma}_h \cdot \underline{\tau}^r_h \, dx + u_k - u_j = 0, \tag{3.239}$$

where we have set $\alpha := a^{-1}$. For any $K_k \in \mathcal{T}_h$ let now $r, r' \in E(k)$; we introduce the following exact and approximate bilinear forms restricted over K_k

$$\begin{aligned} a^{K_k}(\underline{\tau}^{r'}_h, \underline{\tau}^r_h) &= \left(\alpha \underline{\tau}^{r'}_h, \underline{\tau}^r_h\right)_{0,K_k} = \int_{K_k} \alpha \underline{\tau}^{r'}_h \cdot \underline{\tau}^r_h \, dx \\ a^{K_k}_h(\underline{\tau}^{r'}_h, \underline{\tau}^r_h) &= \overline{\alpha}^r \left(\underline{\tau}^{r'}_h, \underline{\tau}^r_h\right)_{h,0,K_k} = \overline{\alpha}^r \delta_{rr'} \omega^r_k \end{aligned} \tag{3.240}$$

where the expression for ω^r_k is provided in the next Proposition, and $\overline{\alpha}^r$ is a suitable average of α over $\mathcal{L}^r$.

Proposition 3.7 *Let $\underline{\tau}^r_h$, $\underline{\tau}^{r'}_h$ be two basis functions in Σ_h associated with two edges of $K_k \in \mathcal{T}_h$. The quadrature formula*

$$\int_{K_k} \underline{\tau}^r_h \cdot \underline{\tau}^{r'}_h \, dx \sim \delta_{rr'} \omega^r_k \tag{3.241}$$

is exact on constant vectors iff *the ω^r_k's are chosen as*

$$\omega^r_k = -\frac{\underline{e}^i_k \cdot \underline{e}^j_k}{4|K_k|} \qquad i, j \in E(k)\,,\ i \neq j \neq r. \tag{3.242}$$

Proof. We provide in the following an alternative proof to the original one given in [18]. For ease of exposition let us switch to a local numbering of the indices of the geometrical quantitities. In particular, we shall use the local indices $1, 2, 3$ in place of the global ones. Then let K be a generic triangle with edges e^1, e^2, e^3, and let $\underline{t}^i$ and $\underline{n}^i$ $(i = 1, 2, 3)$ be the unit tangent and normal vectors to edge e^i (see Fig. 5 for the local numbering and orientation).

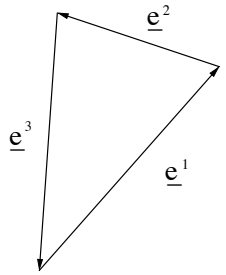


Figure 5: Local numbering and orientation of the edges of a triangle

Finally, let $\underline{e}^i = \underline{t}^i|e^i|$, $\underline{\nu}^i = \underline{n}^i|e^i|$ $(i = 1,2,3)$, where $|e^i|$ denotes the length of edge e^i. We require the numerical quadrature to be exact for constant vectors, i.e., we require the formula to integrate exactly

$$\int_K \underline{\tau} \cdot \underline{\sigma}\, dx \quad \forall \text{ constant vectors } \underline{\tau}, \underline{\sigma}. \tag{3.243}$$

For this, let us first consider $\underline{\sigma} = \underline{\tau} = \underline{c}$, $\underline{c}$ being any constant vector. Let $\underline{\tau}^i$, $i = 1,3$ be the local basis for RT_0 elements, defined by the following choice of degrees of freedom (see (3.97)) :

$$\int_{e^j} \underline{\tau}^i \cdot \underline{n}^j\, ds = \delta_{ij}, \quad i,j = 1,3\,. \tag{3.244}$$

Then $\underline{c}$ can be expressed as

$$\underline{c} = \sum_{i=1}^{3} c_i \underline{\tau}^i\,,$$

where the coefficients c_i can be easily obtained using (3.244):

$$(\underline{c}, \underline{\nu}^j) = (\underline{c}, \underline{n}^j)|e^j| = \int_{e^j} \underline{c} \cdot \underline{n}^j\, ds = \int_{e^j} (\sum_{i=1}^{3} c_i \underline{\tau}^i) \cdot \underline{n}^j\, ds = c_j\,.$$

By imposing the quadrature formula to be exact on constant vectors $\underline{\tau} = \underline{c}$ we obtain

$$|\underline{c}|^2|K| \equiv \int_K \underline{c} \cdot \underline{c}\, dx = \int_K (\sum_{i=1}^{3} c_i \underline{\tau}^i) \cdot (\sum_{j=1}^{3} c_j \underline{\tau}^j)\, dx = \sum_{j=1}^{3} c_j^2 \omega^j = \sum_{j=1}^{3} (\underline{c}, \underline{\nu}^j)^2 \omega^j\,. \tag{3.245}$$

Choosing now $\underline{c} = \underline{e}^i$, $i = 1,2,3$, and observing that $|(\underline{e}^i, \underline{\nu}^j)| = 2(1 - \delta_{ij})|K|$, from (3.245) we obtain

$$|e^i|^2|K| = 4|K|^2 \sum_{j=1}^{3} (1 - \delta_{ij})^2 \omega^j, \quad i = 1,2,3,$$

that is, the linear system

$$\sum_{j=1}^{3} (1 - \delta_{ij}) \omega^j = \frac{|e^i|^2}{4|K|}, \quad i = 1,2,3\,,$$

using the obvious fact that $(1 - \delta_{ij})^2 \equiv 1 - \delta_{ij}$. The solution to this system, given by

$$\omega^i = \frac{\sum_{j=1}^{3} (1 - 2\delta_{ij}) |e^j|^2}{8|K|}\,, \tag{3.246}$$

can be further simplified by noting that, with the orientation of Fig. 5, $\underline{e}^1 + \underline{e}^2 + \underline{e}^3 = \underline{0}$, so that, for instance, $|e^1|^2 = |e^2|^2 + |e^3|^2 + 2(\underline{e}^2, \underline{e}^3)$, and analogous relations hold cyclically. Therefore, from (3.246) we have

$$\omega^1 = -\frac{(\underline{e}^2, \underline{e}^3)}{4|K|}, \quad \omega^2 = -\frac{(\underline{e}^3, \underline{e}^1)}{4|K|}, \quad \omega^3 = -\frac{(\underline{e}^1, \underline{e}^2)}{4|K|}\,. \tag{3.247}$$

Notice that these quantities are not necessarily all positive, as their sign depends on the angles of the triangle. It remains to prove that (3.247) gives a solution to (3.245) also for all possible choices of the constant vectot $\underline{c}$. First, since (3.245) is homogeneous in $\underline{c}$, this can be taken of length 1. Second, letting $\underline{c} = (\cos\theta, \sin\theta)$, $\theta \in \mathbb{R}$ and substituting in (3.245) leads to a homogeneous polynomial of degree 2 in $\cos\theta, \sin\theta$ that must vanish identically in θ. Thus, it suffices to require that this occurs only for three different values of θ. Since the triangle is not degenerate, $\underline{e}^i$ $(i = 1, 2, 3)$ are three independent vectors which correspond to three different values for θ.

To conclude, we have to check that the formula is exact for any pair of constant vectors $\underline{\tau} = \underline{c}^1 \neq \underline{\sigma} = \underline{c}^2$. To this end, we observe that the bilinear form

$$\Phi(\underline{c}^1, \underline{c}^2) := (\underline{c}^1, \underline{c}^2)|K| - \sum_{j=1}^{3} (\underline{c}^1, \underline{\nu}^j)(\underline{c}^2, \underline{\nu}^j)\omega^j \tag{3.248}$$

is symmetric and verifies, for ω^j as in (3.247), $\Phi(\underline{c}, \underline{c}) = 0$. Using this, for every $\underline{c}^1$ and $\underline{c}^2$ we can write $\Phi(\underline{c}^1 + \underline{c}^2, \underline{c}^1 + \underline{c}^2) = 0$, and derive $\Phi(\underline{c}^1, \underline{c}^2) \equiv 0$.

□

Remark 3.8 *For each triangle K_k and for each edge e^r, the quantities ω_k^r can also be computed using the formula*

$$\omega_k^r = \frac{d_k^r}{|e^r|}, \tag{3.249}$$

where d_k^r is the "distance" between C_k and the edge e^r, in a sense that will be made clear in a while. For the moment we notice that these quantities are positive when the angle opposite to the edge e^r is acute, and negative when it is obtuse. This expression is very important in view of the finite volume interpretation of the numerical method obtained with the quadrature formula $(3.240)_2$. *However, we point out that expression* (3.242) *is easier to compute, and it is actually used for the implementation of the method.*

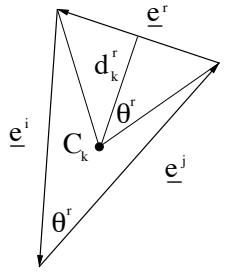


Figure 6: Geometrical quantities of a triangle

To prove (and explain) (3.249) *note that, referring to Fig. 6, the scalar product in the numerator of* (3.242) *can be written as $|e^i||e^j|\cos\theta^r$, where $0 < \theta^r < \pi$ is the angle (opposite to the edge e^r) between edges e^i and e^j; on the other hand, $2|K_k| =$*

$|e^i||e^j|\sin\theta^r$. *This gives* $2\omega_k^r = \cot\theta^r$ *and, since* C_k *is the circumcenter of the triangle,* θ^r *is also half the angle opposite to the edge* e^r *seen by* C_k. *This finally yields* $\cot\theta^r = 2\,d_k^r/|e^r|$ *and thus relation* (3.249), *where it is understood that* d_k^r *is positive when* θ^r *is acute* (C_k *inside* K_k) *and it is negative when* θ^r *is obtuse* (C_k *outside of* K_k).

The construction of the piecewise constant function $\overline{\alpha}^r$ will be fully discussed in Sections 3.8.3 and 3.8.4; here we just emphasize the two basic properties of the average:

i) $\overline{\alpha}^r$ is constant on each lumping region $\mathcal{L}^r$;

ii) $\overline{\alpha}^r$ is some average of α on a suitable subset of each lumping region $\mathcal{L}^r$.

We notice that in the case $\alpha = 1$, the approximate bilinear form a_h^K coincides with the quadrature formula proposed in [18], where it is shown that the quadrature error is $\mathcal{O}(h_K)$. Let then $(\underline{\sigma}_h^*, u_h^*) \in \Sigma_h \times V_h$ be the solution of the dual mixed system (3.58) in the presence of some quadrature error.

Using $(3.240)_2$ in (3.239) and recalling that $\overline{\alpha}^r$ is constant over the *whole* lumping region $\mathcal{L}^r$, we end up with the following equation for the approximate interelement flux through the edge $e^r \notin \Gamma_N$,

$$\Phi_k^{r,*} = (\overline{\alpha}^r)^{-1}\left(\frac{u_j^* - u_k^*}{d^r}\right)|e^r|, \qquad d^r = d_k^r + d_j^r, \qquad k, j \in T(r). \tag{3.250}$$

Equation (3.250) holds for any edge e^r in the interior of Ω; if e^r lies on Γ_D, it is understood that the value u_j^* is set equal to the average g^r of the Dirichlet datum g over e^r and that $d^r = d_k^r$ is the "distance" between the circumcenter of K_k and e^r (see [23]).
Substituting the exact fluxes Φ_k^r in (3.237) with the corresponding approximations (3.250), we finally obtain the family of cell-centered finite volume schemes in the new unknown u_h^*

$$\sum_{\substack{r\in E(k)\\ e^r\notin\Gamma_N}} (\overline{\alpha}^r)^{-1}\left(\frac{u_k^* - u_{j(r)}^*}{d^r}\right)|e^r| + u_k^*\gamma_k|K_k| = f_k|K_k| \quad \forall K_k \in \mathcal{T}_h \tag{3.251}$$

where, for an internal e^r, $j(r) \in T(r)$ is the index of the triangle shared by the edge e^r opposite to the triangle K_k, and $u_{j(r)}^*$ is the unknown on such a triangle. For an edge $e^r \in \Gamma_D$, we set $u_{j(r)}^* = g^r$. In this last case, g^r is the L^2-projection of g on the space of the constant functions over e^r, as in (3.160). Finally, we recall that $\Phi_k^{r,*} = 0$ on $e^r \in \Gamma_N$.

The set of linear algebraic equations (3.251) can be written in matrix form as

$$W^*\mathbf{u}^* = \mathbf{f}^* \tag{3.252}$$

where the i-th component of $\mathbf{f}^*$ is $f_i|K_i|$ and the ij-th nonzero entries of the $\mathtt{N_T} \times \mathtt{N_T}$ matrix W^* are

$$W^*_{ij} = \begin{cases} \displaystyle\sum_{\substack{r \in E(i) \\ e^r \notin \Gamma_N}} (\overline{\alpha}^r)^{-1} \frac{|e^r|}{d^r} + \gamma_i |K_i|, & \text{if } i = j \\ -(\overline{\alpha}^{r(j)})^{-1} \dfrac{|e^{r(j)}|}{d^{r(j)}}, & \text{if } j \in T(E(i)), \quad j \neq i \end{cases} \tag{3.253}$$

where $r(j) = E(j) \cap E(i)$ refers to the edge shared by the triangles K_i and K_j.

Lemma 3.1 *Assume that $\mathcal{T}_h$ is a Delaunay triangulation. Then, the matrix W^* in* (3.252) *is a symmetric, positive definite and irreducibly diagonally dominant* M-matrix.

Proof. We first notice that the Delaunay property for $\mathcal{T}_h$ implies that the quantities d^r's in (3.253) are positive. As a consequence, $W^*_{ii} > 0$ and $W^*_{ij} \leq 0$ in (3.253), for $i = 1, \ldots, \mathtt{N_T}$ and $j \in T(E(i)), j \neq i$. The expressions (3.253) show that W^* is a symmetric matrix with at most four nonzero entries on each row. Let s_i denote the sum of the entries of each row i of W^*; then

$$s_i = \begin{cases} \gamma_i |K_i|, & \text{if } \partial K_i \cap \Gamma_D = \emptyset \\ \displaystyle\sum_{\substack{r \in E(i) \\ e^r \in \Gamma_D}} (\overline{\alpha}^r)^{-1} \frac{|e^r|}{d^r} + \gamma_i |K_i|, & \text{if } \partial K_i \cap \Gamma_D \neq \emptyset. \end{cases}$$

We see that the sum of the entries of the rows corresponding to triangles intersecting Γ_D is strictly positive while it is nonnegative for the rows corresponding to internal triangles. Thus, W^* is a symmetric, positive definite irreducibly diagonally dominant M-matrix (see [161], Corollary 1, p. 85). □

Remark 3.9 *The* M-matrix *property ensures that system* (3.252) *is uniquely solvable. This property along with the fact that W^* is also positive definite and irreducibly diagonally dominant ensures that the family of finite volume schemes* (3.251) *verifies a discrete maximum principle (see [161, 43, 140]). In particular, provided $\mathbf{f}^* \geq 0$, the solution $\mathbf{u}^*$ of system* (3.252) *turns out to be nonnegative.*

The M-matrix property is very desirable in applications. As an example, consider the case where (2.14) is the linearized current continuity equation using the drift-diffusion or energy-balance transport models. In such a case, the unknown u is the scaled concentration and therefore it must be nonnegative.

Remark 3.10 (Extension to the 3D case) *Let us briefly comment about the extension of the finite volume formulation at hand to the case where Ω is a polygonal domain in $\mathbb{R}^3$.*

The extension is straightforward if Ω is of the form $\Omega = \Omega_{xy} \times (0, H)$ with $\Omega_{xy} \subset \mathbb{R}^2$ and $H > 0$, and $\mathcal{T}_h$ is a partition of Ω into prisms obtained as "tensor product" of a decomposition of Ω_{xy} and a decomposition of $[0, H]$. We refer to [7] for the definition and to [116] for the use of mixed approximations with Raviart-Thomas finite elements on prismatic triangulations.

The extension of the mixed finite volume method analyzed in this section to the case where $\mathcal{T}_h$ is a decomposition of Ω into simplices is less straightforward. In particular, if the finite element mesh is made of tetrahedra a consistent diagonalization of the mass matrix has been provided in [19], in the case of the Laplace operator, under restrictive conditions on the mesh. Under the same assumptions, the present mixed finite volume scheme can be formulated also in three dimensions.

In the case of a general 3D triangulation the method of this section as well as the related approaches proposed in [18, 19, 2] cannot be extended, as pointed out in [158].

3.8.3 Analysis of the finite volume scheme

In this section we analyze the finite volume scheme defined in (3.251) by setting it into an abstract framework of Galerkin methods with numerical quadratures for the approximation of saddle point problems of the form (3.35). Concerning this issue, error estimates and analysis can be found in [27], sect. II.2.4 and [138], Ch. 3, sect. 11. For the case under consideration, the choice of RT_0 finite element spaces for Σ_h, V_h is such that the family of finite volume schemes (3.251) can be written in the form:

$$\begin{cases} \text{Find } \underline{\sigma}_h \in \Sigma_h \text{ and } u_h \in V_h \text{ such that:} \\ a_h(\underline{\sigma}_h, \underline{\tau}_h) + b_h(u_h, \underline{\tau}_h) = \langle g, \underline{\tau}_h \rangle_h \qquad \forall \underline{\tau}_h \in \Sigma_h, \\ b_h(v_h, \underline{\sigma}_h) - c_h(u_h, v_h) = -\langle f, v_h \rangle_h \qquad \forall v_h \in V_h, \end{cases} \tag{3.254}$$

where $a_h(\cdot,\cdot)$, $b_h(\cdot,\cdot)$, $c_h(\cdot,\cdot)$ and $\langle \cdot, \cdot \rangle_h$ are suitable approximations of the corresponding continuous counterparts through the use of numerical quadratures. In this case we have

$$\begin{aligned} &a_h(\underline{\sigma}_h, \underline{\tau}_h) = \sum_{K_k \in \mathcal{T}_h} a_h^{K_k}(\underline{\sigma}_h, \underline{\tau}_h), && b_h(v_h, \underline{\tau}_h) = b(v_h, \underline{\tau}_h), \\ &c_h(u_h, v_h) = \sum_{K_k \in \mathcal{T}_h} \gamma_k \, (u_h, v_h)_{0,K_k} \equiv c(u_h, v_h), \\ &\langle f, v_h \rangle_h = \sum_{K_k \in \mathcal{T}_h} f_k \, (v_h, 1)_{0,K_k} \equiv (f, v_h), && \langle g, \underline{\tau}_h \rangle_h = \langle g, \underline{\tau}_h \cdot \underline{n} \rangle_{|\Gamma_D}. \end{aligned} \tag{3.255}$$

(See (3.25)–(3.27) and Section 3.8.2 for notation and definitions). Hypotheses (3.31)-(3.34) are all verified, while (3.28) and (3.29) for $a_h(\cdot,\cdot)$ need to be checked in order to apply the abstract Theorem 3.1. For that we recall that definition (3.57) of $\operatorname{Ker} B_h$ reads in this case

$$\operatorname{Ker} B_h \, = \, \{ \underline{\tau}_h \in \Sigma_h \, | \, b(v_h, \underline{\tau}_h) = 0 \quad \forall \, v_h \in V_h \} \, . \tag{3.256}$$

Lemma 3.2 *There exist positive constants $\overline{\mathcal{A}}$ and $\underline{\mathcal{A}}$, independent of h, such that*

$$\begin{aligned} &|a_h(\underline{\sigma}_h, \underline{\tau}_h)| \le \overline{\mathcal{A}} ||\underline{\sigma}_h||_\Sigma ||\underline{\tau}_h||_\Sigma, \qquad && \forall \underline{\sigma}_h,\, \underline{\tau}_h \in \Sigma_h \\ &a_h(\underline{\tau}_h, \underline{\tau}_h) \ge \underline{\mathcal{A}} ||\underline{\tau}_h||_\Sigma^2, && \forall \underline{\tau}_h \in \mathrm{Ker} B_h. \end{aligned} \tag{3.257}$$

Proof. To prove $(3.257)_1$, let us first observe that, $a_h(\underline{\sigma}_h, \underline{\tau}_h)$ being symmetric and positive semidefinite, we have

$$a_h(\underline{\sigma}_h, \underline{\tau}_h) \le \Big(a_h(\underline{\sigma}_h, \underline{\sigma}_h)\, a_h(\underline{\tau}_h, \underline{\tau}_h) \Big)^{1/2} \qquad \forall\, \underline{\sigma}_h, \underline{\tau}_h \in \Sigma_h. \tag{3.258}$$

Then it is sufficient to prove that

$$a_h(\underline{\tau}_h, \underline{\tau}_h) \le \overline{\mathcal{A}} ||\underline{\tau}_h||_\Sigma^2, \qquad \underline{\tau}_h \in \Sigma_h. \tag{3.259}$$

Consider a triangle $K_k \in \mathcal{T}_h$ and set $\underline{\tau}_{h|K_k} = \sum_{i \in E(k)} \Phi^i_{\underline{\tau}_h} \underline{\tau}^i_h$ and $\omega^i_k = d^i_k / |e^i|$, where $\Phi^i_{\underline{\tau}_h}$ is the flux of $\underline{\tau}_h$ across the edge e^i, and d^i_k is the "distance" between the circumcenter C_k of K_k and the edge e^i (see Section 3.8.2). We also set $\mathcal{R}_{K_k} = \cup_{i \in E(k)} \mathcal{L}^i$. For brevity, the subscript k will be dropped. On the element K the approximate bilinear form can be written as

$$a^K_h(\underline{\tau}_h, \underline{\tau}_h) = \sum_{i \in E(k)} \overline{\alpha}^i (\Phi^i_{\underline{\tau}_h})^2 \omega^i_k.$$

Notice that, using (3.234), we have

$$\max_{i \in E(k)} |\omega^i_k| \le \frac{D_K}{\rho_K} \le \mathcal{K}^*, \tag{3.260}$$

which holds both for acute and obtuse triangles. Therefore, since $\overline{\alpha}$ is an average of α

$$|a^K_h(\underline{\tau}_h, \underline{\tau}_h)| \le ||\alpha||_{\infty, \mathcal{R}_K} \sum_{i \in E(k)} |\Phi^i_{\underline{\tau}_h}|^2 \, |\omega^i_k| \le \mathcal{K}^* ||\alpha||_{\infty, \Omega} \sum_{i \in E(k)} |\Phi^i_{\underline{\tau}_h}|^2. \tag{3.261}$$

Recalling the definition of flux through an edge e^i of $\underline{\tau}_h \in \Sigma_h$

$$\Phi^i_{\underline{\tau}_h} = \int_{e^i} \underline{\tau}_h \cdot \underline{n}^i \, ds,$$

Cauchy–Schwarz inequality gives

$$|\Phi^i_{\underline{\tau}_h}|^2 \le |e^i| \, ||\underline{\tau}_h \cdot \underline{n}^i||^2_{0, e^i}. \tag{3.262}$$

Using (3.177), i.e. $||\underline{\tau}_h \cdot \underline{n}^i||^2_{0,e^i} \le C h_K^{-1} ||\underline{\tau}_h||^2_{0,K}$, we deduce

$$|\Phi^i_{\underline{\tau}_h}|^2 \le C ||\underline{\tau}_h||^2_{0,K} \qquad \forall \underline{\tau}_h \in \Sigma_h. \tag{3.263}$$

It can be easily checked that, in the present case of RT_0 elements, $C \le 6$. Substituting in (3.261) gives

$$|a^K_h(\underline{\tau}_h, \underline{\tau}_h)| \le 18 \mathcal{K}^* ||\alpha||_{\infty, \Omega} ||\underline{\tau}_h||^2_{0,K}, \tag{3.264}$$

from which, summing over all the triangles of $\mathcal{T}_h$, the estimate $(3.257)_1$ follows with $\overline{\mathcal{A}} = 18\mathcal{K}^*||\alpha||_{\infty,\Omega}$.

Let us now prove $(3.257)_2$. We remark that, in this case, it is not possible to work at the element level since the geometric quantities ω_k^r's associated with an element K_k are not, a priori, both positive. Actually, we are dealing with a Delaunay mesh and this guarantees only that $\sum_{k\in T(r)} \omega_k^r \geq 0$, the equal sign holding in the case of a degenerate Delaunay mesh. Thus, we are led to working with the full bilinear form. Since any function $\underline{\tau}_h \in \mathrm{Ker}B_h$ is such that $\underline{\tau}_{h|K} \in (P_0(K))^2$, $\forall\, K \in \mathcal{T}_h$, we have $||\underline{\tau}_h||_\Sigma = ||\underline{\tau}_h||_{0,\Omega}$. Moreover, Proposition 3.7 ensures that the quadrature formula is exact on piecewise constant vectors, that is,

$$\sum_{i\in E(k)} (\Phi^i_{\underline{\tau}_h})^2 \omega_k^i = ||\underline{\tau}_h||^2_{0,K}, \qquad \underline{\tau}_h \in \mathrm{Ker}B_h.$$

Thus

$$\begin{aligned} a_h(\underline{\tau}_h, \underline{\tau}_h) &= \sum_{r=1}^{\mathtt{N_E}} \sum_{k\in T(r)} \overline{\alpha}^r (\Phi^r_{\underline{\tau}_h})^2 \omega_k^r \geq \alpha_0 \sum_{r=1}^{\mathtt{N_E}} \sum_{k\in T(r)} (\Phi^r_{\underline{\tau}_h})^2 \omega_k^r \\ &= \alpha_0 \sum_{k=1}^{\mathtt{N_T}} \sum_{r\in E(k)} (\Phi^r_{\underline{\tau}_h})^2 \omega_k^r = \alpha_0 \sum_{k=1}^{\mathtt{N_T}} ||\underline{\tau}_h||^2_{0,K} = \alpha_0 ||\underline{\tau}_h||^2_{0,\Omega}, \end{aligned}$$

where $\alpha_0 = \inf_{x\in\Omega} \alpha(x)$. We finally get $(3.257)_2$ with $\underline{\mathcal{A}} = \alpha_0$. □

We are in position to state the main result of this section.

Theorem 3.7 *For every $g \in H^{1/2}(\Gamma_D)$, and for every $f \in L^2(\Omega)$ problem* (3.254) *has a unique solution $(\underline{\sigma}_h, u_h)$, and the following a priori error bound holds*

$$\begin{aligned} ||\underline{\sigma} - \underline{\sigma}_h||_\Sigma + ||u - u_h||_V \;\leq\; & \mathcal{M} \inf_{\substack{\underline{w}_h\in\Sigma_h \\ v_h\in V_h}} \Big\{ ||\underline{\sigma} - \underline{w}_h||_\Sigma + ||u - v_h||_V \\ & + \sup_{\underline{\tau}_h\in\Sigma_h} \frac{|a(\underline{w}_h, \underline{\tau}_h) - a_h(\underline{w}_h, \underline{\tau}_h)|}{||\underline{\tau}_h||_\Sigma} \Big\}, \end{aligned} \tag{3.265}$$

with $(\underline{\sigma}, u)$ solution of problem (3.27)*, and $\mathcal{M}$ a constant independent of h.*

Proof. It is an extension of the proof of Proposition 2.11, [27], which holds in the case of exact integration and it is here omitted. The interested reader can consult [115]. □

Let us deal now with the quadrature errors occurring in the approximate bilinear form $a_h(\cdot,\cdot)$. We recall that $\mathcal{K}$ denotes a generic constant depending only on the local geometrical properties of the mesh, namely on $\mathcal{K}^*$ (see (3.45)-(3.46)). With a slight change of notation, we let K be a generic triangle and C_K its circumcenter. We shall use the following result:

Proposition 3.8 *Let $\mathcal{R}_K$ be the union of K and of the three lumping regions associated with its edges. Then under assumptions* (3.45)-(3.46) *on the mesh $\mathcal{T}_h$, for every $\overline{x} \in \overline{\mathcal{R}}_K$ we have*

$$\sup_{x \in \mathcal{R}_K} |v(x) - v(\overline{x})| \leq \mathcal{K} h_K |v|_{1,\infty,\mathcal{R}_K} \qquad \forall v \in C^1(\overline{\mathcal{R}}_K). \tag{3.266}$$

Inequality (3.266) follows immediately from the mean-value theorem. For more general inequalities of this type see, e.g., [42], Section 15, e.g., Theorem 15.3.

Let us now give a bound for the quadrature error associated with $a_h(\cdot,\cdot)$. With this aim, let $\underline{w}_h$ and $\underline{\tau}_h$ be any two functions in Σ_h. For any $K_k \in \mathcal{T}_h$ we set $\underline{w}_{h|K_k} = \sum_{r \in E(k)} \Phi^r_{\underline{w}_h} \underline{\tau}^r_h$ and $\underline{\tau}_{h|K_k} = \sum_{r \in E(k)} \Phi^r_{\underline{\tau}_h} \underline{\tau}^r_h$. Dropping subscript k, we recall the result proved in [18]

$$\left| (\underline{w}_h, \underline{\tau}_h)_{0,K} - (\underline{w}_h, \underline{\tau}_h)_{h,0,K} \right| \leq \mathcal{K} h_K \|\underline{w}_h\|_{H(\mathrm{div};K)} \|\underline{\tau}_h\|_{H(\mathrm{div};K)}. \tag{3.267}$$

Let $\widetilde{\alpha}_K$ be a suitable average of the function α over K, for instance equal to the value of α at the centroid of K. This choice clearly satisfies (3.266) where $v = \alpha$. The quadrature error $a^K(\underline{w}_h, \underline{\tau}_h) - a^K_h(\underline{w}_h, \underline{\tau}_h)$ can then be split as

$$a^K(\underline{w}_h, \underline{\tau}_h) - a^K_h(\underline{w}_h, \underline{\tau}_h) = \mathcal{E}_1(\underline{w}_h, \underline{\tau}_h) + \mathcal{E}_2(\underline{w}_h, \underline{\tau}_h) + \mathcal{E}_3(\underline{w}_h, \underline{\tau}_h)$$

where

$$\mathcal{E}_1(\underline{w}_h, \underline{\tau}_h) = (\alpha \underline{w}_h, \underline{\tau}_h)_{0,K} - (\widetilde{\alpha}_K \underline{w}_h, \underline{\tau}_h)_{0,K} = \int_K (\alpha - \widetilde{\alpha}_K) \underline{w}_h \cdot \underline{\tau}_h dx$$

$$\mathcal{E}_2(\underline{w}_h, \underline{\tau}_h) = (\widetilde{\alpha}_K \underline{w}_h, \underline{\tau}_h)_{0,K} - (\widetilde{\alpha}_K \underline{w}_h, \underline{\tau}_h)_{h,0,K} = \int_K \widetilde{\alpha}_K \underline{w}_h \cdot \underline{\tau}_h dx - \sum_{r \in E(k)} \widetilde{\alpha}_K \Phi^r_{\underline{w}_h} \Phi^r_{\underline{\tau}_h} \omega^r_k$$

$$\mathcal{E}_3(\underline{w}_h, \underline{\tau}_h) = (\widetilde{\alpha}_K \underline{w}_h, \underline{\tau}_h)_{h,0,K} - a^K_h(\underline{w}_h, \underline{\tau}_h) = \sum_{r \in E(k)} (\widetilde{\alpha}_K - \overline{\alpha}^r) \Phi^r_{\underline{w}_h} \Phi^r_{\underline{\tau}_h} \omega^r_k$$

Using the Cauchy-Schwarz inequality and bounding the term $\alpha - \widetilde{\alpha}_K$ over $K \subset \mathcal{R}_K$ using (3.266) we get

$$|\mathcal{E}_1(\underline{w}_h, \underline{\tau}_h)| \leq C h_K |\alpha|_{1,\infty,\mathcal{R}_K} \|\underline{w}_h\|_{H(\mathrm{div};K)} \|\underline{\tau}_h\|_{H(\mathrm{div};K)}$$

while (3.267) yields

$$|\mathcal{E}_2(\underline{w}_h, \underline{\tau}_h)| \leq \widetilde{\alpha}_K \mathcal{K} h_K \|\underline{w}_h\|_{H(\mathrm{div};K)} \|\underline{\tau}_h\|_{H(\mathrm{div};K)}.$$

Recalling that $\widetilde{\alpha}_K$ and $\overline{\alpha}^r$ are averages of α over $K \subset \mathcal{R}_K$ and $\mathcal{L}^r \subset \mathcal{R}_K$, respectively, we can immediately conclude that

$$\begin{aligned} |\mathcal{E}_3(\underline{w}_h, \underline{\tau}_h)| &\leq \left(\max_{x \in \mathcal{R}_K} \alpha(x) - \min_{x \in \mathcal{R}_K} \alpha(x) \right) \sum_{r \in E(k)} |\Phi^r_{\underline{w}_h}| \, |\Phi^r_{\underline{\tau}_h}| \, |\omega^r_k| \\ &\leq |\alpha|_{1,\infty,\mathcal{R}_K} \mathcal{K} h_K \sum_{r \in E(k)} |\Phi^r_{\underline{w}_h}| \, |\Phi^r_{\underline{\tau}_h}| \, |\omega^r_k| \leq \mathcal{K} |\alpha|_{1,\infty,\mathcal{R}_K} h_K \|\underline{w}_h\|_{H(\mathrm{div};K)} \|\underline{\tau}_h\|_{H(\mathrm{div};K)}, \end{aligned}$$

where we have used (3.266), (3.260), Cauchy–Schwarz inequality and (3.263). Gathering the previous estimates yields, $\forall K \in \mathcal{T}_h$

$$|a^K(\underline{w}_h, \underline{\tau}_h) - a_h^K(\underline{w}_h, \underline{\tau}_h)| \leq \mathcal{K}\left(|\alpha|_{1,\infty,\Omega} + ||\alpha||_{\infty,\Omega}\right) h_K ||\underline{w}_h||_{H(\mathrm{div};K)} ||\underline{\tau}_h||_{H(\mathrm{div};K)}.$$

Summing over all triangles and using the Cauchy-Schwarz inequality finally gives

$$\sup_{\underline{\tau}_h \in \Sigma_h} \frac{|a(\underline{w}_h, \underline{\tau}_h) - a_h(\underline{w}_h, \underline{\tau}_h)|}{||\underline{\tau}_h||_\Sigma} \leq Ch ||\underline{w}_h||_{H(\mathrm{div};\Omega)}, \qquad \forall \underline{w}_h \in \Sigma_h, \tag{3.268}$$

with $C = \mathcal{K}\left(|\alpha|_{1,\infty,\Omega} + ||\alpha||_{\infty,\Omega}\right)$. In particular, taking in (3.268) $\underline{w}_h = \Pi_h \underline{\sigma}$,as defined in (3.100), and recalling that (see, [138], p. 583)

$$||\Pi_h \underline{\sigma}||_\Sigma \leq C\left(h|\underline{\sigma}|_{1,\Omega} + ||\underline{\sigma}||_\Sigma\right),$$

we get

$$\sup_{\underline{\tau}_h \in \Sigma_h} \frac{|a(\Pi_h \underline{\sigma}, \underline{\tau}_h) - a_h(\Pi_h \underline{\sigma}, \underline{\tau}_h)|}{||\underline{\tau}_h||_\Sigma} \leq Ch\left(h|\underline{\sigma}|_{1,\Omega} + ||\underline{\sigma}||_\Sigma\right). \tag{3.269}$$

Taking in (3.265) $v_h = P_h u$, as defined in (3.99), $\underline{w}_h = \Pi_h \underline{\sigma}$, and using the interpolation estimates (3.110) and (3.269), we finally obtain the convergence result

$$||\underline{\sigma} - \underline{\sigma}_h||_\Sigma + ||u - u_h||_V \leq Ch\left(|u|_{1,\Omega} + |\underline{\sigma}|_{1,\Omega} + |\mathrm{div}\underline{\sigma}|_{1,\Omega}\right). \tag{3.270}$$

We have proved the following

Theorem 3.8 *Assume that the solution* $(\underline{\sigma}, u)$ *of problem* (3.27) *is such that* $(\underline{\sigma}, u) \in (H^1(\Omega))^2 \times H^1(\Omega)$ *and* $\mathrm{div}\underline{\sigma} \in H^1(\Omega)$, *and that* $a \in W^{1,\infty}(\Omega)$. *Then, there exists a positive constant* $\mathcal{C}$, *independent of* h, *such that the solution* $(\underline{\sigma}_h, u_h)$ *of problem (3.254), with the choices* (3.255), *satisfies*

$$||\underline{\sigma} - \underline{\sigma}_h||_\Sigma + ||u - u_h||_V \leq \mathcal{C}h. \tag{3.271}$$

We emphasize that Theorem 3.8 holds irrespectively of the choice of the average $\overline{\alpha}^r$ in (3.251), the only requirement being that

$$\min_{x \in \mathcal{L}^r} \alpha(x) \leq \overline{\alpha}^r \leq \max_{x \in \mathcal{L}^r} \alpha(x) \qquad \forall \mathcal{L}^r \in \mathcal{L}_h. \tag{3.272}$$

Remark 3.11 *We remark that the* $\mathcal{O}(h)$ *convergence proved above for the mixed finite volume methods* (3.251) *is optimal and that it has been obtained without giving up the M-*matrix *property of the schemes. This feature, together with the reduced computational cost, makes these methods quite attractive and competitive with respect to the standard dual mixed approaches with exact integration analyzed in Section 3.5 and to dual mixed schemes with numerical integration recently proposed for the approximation of problem* (3.27) *in the case* $\gamma = 0$ *[18, 19]. Finally, the genuine finite volume flavor of these methods allows for the basic conservation properties (mass and interelement fluxes) to be satisfied, even in the presence of jumps in the coefficient of the problem (as happens for instance in porous media flows governed by Darcy's law [57]) or steep internal/boundary layers in the scalar unknown* u *(as happens in semiconductor device simulation using the drift-diffusion model [82, 102, 124, 143]).*

3.8.4 Several choices of the averages

In this section we characterize the choice of the average $\overline{\alpha}$ of the inverse diffusion coefficient α. Being this matter one-dimensional, we restrict our attention on the affine-equivalent interval $\mathcal{I} = [0, L]$, for any $L > 0$.

The obvious candidate for $\overline{\alpha}$ is the mean value of α over $\mathcal{I}$. Since we are interested in the inverse of $\overline{\alpha}$ (see (3.251)) and $\alpha = a^{-1}$, it follows that

$$\overline{\alpha}^{-1} = \left(\frac{\int_0^L a^{-1}(x)dx}{L} \right)^{-1} =: \mathcal{H}_{\mathcal{I}}(a) \tag{3.273}$$

where $\mathcal{H}_{\mathcal{I}}(a)$ denotes the *harmonic average* of a over the interval $\mathcal{I}$.
Use of harmonic averaging for the diffusion coefficient a is quite natural in mixed methods (see [12, 27]) and has been proved in one dimension to provide better results than the mean value, in particular when a exhibits sharp variations on $\mathcal{I}$ or is even discontinuous. Typical instances of such a behavior are flows in porous media (see [57] and the references cited therein) or electron and hole carrier flow in a semiconductor device.

It is clear that, except for special cases, the evaluation of (3.273) cannot be carried out. This, of course, demands for the use of a suitable quadrature formula. With this aim, define for any function $\phi : \mathcal{I} \to \mathbb{R}^+$ the *exponential interpolant* to ϕ as

$$\mathcal{E}_{\mathcal{I}}\phi(x) := \exp\{[\ln(\phi(x))]_I\} = \phi_0 \left(\frac{\phi_L}{\phi_0}\right)^{\frac{x}{L}}, \qquad x \in \mathcal{I} \tag{3.274}$$

where $\phi_0 = \phi(0)$, $\phi_L = \phi(L)$ and v_I is the P_1-interpolant to v. The quadrature formula approximating (3.273) can then be defined as

$$\begin{aligned} \overline{\alpha}^{-1} \simeq \left(\frac{\int_0^L \mathcal{E}_{\mathcal{I}} a^{-1}(x)dx}{L} \right)^{-1} &= \frac{\ln(1/a_L) - \ln(1/a_0)}{1/a_L - 1/a_0} \\ &= \left(\frac{\ln(a_0) - \ln(a_L)}{a_0 - a_L} \right) a_0 a_L \end{aligned} \tag{3.275}$$

which clearly satisfies (3.272). Assuming $\alpha \in W^{1,\infty}(\mathcal{I})$, it is easy to prove the following bound for the interpolation error

$$||\alpha - \mathcal{E}_{\mathcal{I}}\alpha||_{\infty,\mathcal{I}} \leq Ch\frac{\alpha_M}{\alpha_m}|\alpha|_{1,\infty,\mathcal{I}}$$

where α_M and α_m are the maximum and the minimum values of α over $\mathcal{I}$, respectively. We remark that (3.275) is *exact* if $a(x) = e^{lx+m}$, $l, m \in \mathbb{R}$, $x \in \mathcal{I}$, which is the case of

the numerical approximation of the drift-diffusion semiconductor device equations when the self-adjoint form (1.9) is used for the current densities and when a linear variation of the (scaled) electric potential ψ is assumed over $\mathcal{I}$. The resulting discretization scheme will be addressed in Section 4.2.

Our second choice for $\overline{\alpha}^{-1}$ is the harmonic average of the piecewise constant extension of a over $\mathcal{I}$

$$\overline{\alpha}^{-1} \simeq \left(\frac{\displaystyle\int_0^{\overline{x}} a_0^{-1} dx + \int_{\overline{x}}^{L} a_L^{-1} dx}{L} \right)^{-1} = \frac{a_0 a_L}{a_L \overline{x}/L + a_0(1 - \overline{x}/L)} \tag{3.276}$$

where $\overline{x} \in \mathcal{I}$. In our application, $\mathcal{I}$ will be the segment joining the circumcenters of two adjacent triangles K and K' and $\overline{x}$ will be the intersection between $\mathcal{I}$ and the edge common to K and K'. The average (3.276) seems to be quite promising in presence of discontinuities of a.

The last choice that we consider for $\overline{\alpha}^{-1}$ employs linear interpolation for a. This leads to the trapezoidal quadrature formula

$$\overline{\alpha}^{-1} \simeq \frac{\displaystyle\int_0^{L} a_I(x) dx}{L} = \frac{a_0 + a_L}{2}. \tag{3.277}$$

Numerical experiments using the three averages introduced above will be presented in Section 7.

4 Application to continuity equations

We shall describe in this section mixed discretizations of the scaled current continuity equations $(1.22)_2$, $(1.22)_3$. For simplicity, we shall deal only with the scaled continuity equation $(1.22)_3$ for the positive charge density p. Moreover, we shall consider the stationary case and a constant mobility $\mu_p \equiv 1$. The problem under investigation has the form

$$\begin{cases} \text{Find } p \in H^1(\Omega) \text{ such that} & \\ -\text{div}\,(\underline{\nabla} p + p \underline{\nabla} \psi) = -R(p,n) & \text{in } \Omega \subset \mathbb{R}^2, \\ p = g := p_D & \text{on } \Gamma_D \subset \partial\Omega, \\ (\underline{\nabla} p + p \underline{\nabla} \psi) \cdot \underline{n} = 0 & \text{on } \Gamma_N \subset \partial\Omega\,, \end{cases} \tag{4.1}$$

and the current is given by

$$\underline{J} = -\,(\underline{\nabla} p + p \underline{\nabla} \psi)\,. \tag{4.2}$$

In the simulation of the DD model the solution of (4.1) is an intermediate step inside an iterative process, as presented in Section 2, and we shall assume that ψ and n are known. Moreover, during the iterative solution procedure, equation $(4.1)_1$ is usually linearized (see Section 2.4) so that problem (4.1) becomes

$$\begin{cases} \text{Find } p \in H^1(\Omega) \text{ such that} & \\ -\text{div}\,(\underline{\nabla} p + p\underline{\nabla}\,\psi) + cp = f & \text{in } \Omega, \\ p = g & \text{on } \Gamma_D, \\ (\underline{\nabla} p + p\underline{\nabla}\,\psi)\cdot \underline{n} = 0 & \text{on } \Gamma_N\,, \end{cases} \tag{4.3}$$

where $f = f(x)$ is a function independent of p, and $c = c(x)$ is a nonnegative function independent of p, which can be assumed piecewise constant. Finally, ψ is assumed to be piecewise linear (resulting from a discretization of the Poisson equation). Problem (4.3) is not directly suited to the application of the schemes described in Section 3 because it is not self–adjoint. Using the Slotboom variable ρ (see (1.8))

$$p = \rho\, e^{-\psi}, \tag{4.4}$$

(4.3) can be written in the symmetric form as

$$\begin{cases} \text{Find } \rho \in H^1(\Omega) \text{ such that} & \\ -\text{div}\,\left(e^{-\psi}\underline{\nabla}\,\rho\right) + c\,e^{-\psi}\rho = f & \text{in } \Omega, \\ \rho = \chi := e^{\psi} g & \text{on } \Gamma_D, \\ \underline{\nabla}\,\rho\cdot\underline{n} = 0 & \text{on } \Gamma_N\,, \end{cases} \tag{4.5}$$

and the current is now given by

$$\underline{J} = -e^{-\psi}\underline{\nabla}\,\rho. \tag{4.6}$$

The symmetric problem (4.5) is now of the form (2.14) with $a = e^{-\psi}$ and $\gamma = c\,e^{-\psi}$. We point out that other approaches, based on the use of the quasi–Fermi potentials defined in (1.5), also lead to self–adjoint problems. However, in this case the equations are exponentially nonlinear. For mixed approximations of the quasi–Fermi potential formulation we refer to [74, 73].

Problem (4.5) will not be discretized directly because the presence of the exponential functions can be a source of numerical troubles in many relevant situations. The idea of the schemes presented here is to discretize the symmetric equation (4.5) with a mixed finite element scheme in the hybridization form presented in Section 3.6 or with a mixed finite volume schemes presented in Section 3.8, to write the system in matrix form as in (3.211)-(3.213) or in (3.251)–(3.253), to go back to the original variable p by using a suitable discrete version of transformation (4.4), and then to solve for p.

4.1 Exponential fitting mixed finite elements

We shall describe in this section a mixed approximation to the scaled current continuity equations (4.3). For the case $c = 0$ a mixed scheme based on the lowest order Raviart–Thomas element RT_0 has been introduced in [29] and extensively discussed in [30, 31]. For the case $c \neq 0$ mixed schemes based on the elements of *Examples 5-6*, introduced and analyzed in [98], have been applied and discussed in [99]. In the following we shall give a compact presentation which includes the three different elements.

The presence of the exponentials in the definition of a and γ requires special care. For the approximation of $a^{-1} = e^{\psi}$, following (3.217), we define $\overline{\psi}$ as the piecewise constant function given in each element K by

$$e^{\overline{\psi}}{}_{|K} := \frac{1}{|K|}\int_K e^{\psi} dx, \tag{4.7}$$

and we take

$$\overline{\alpha} = e^{\overline{\psi}}. \tag{4.8}$$

The approximation of $\gamma = ce^{-\psi}$ suggested in (3.218) is not suited here for reasons which will be made clear at the end of this section. We proceed as follows. Let $\widetilde{\psi}$ denote the piecewise constant function defined in each element K via a harmonic average on a special edge $\tilde{e}$. In order to define $\tilde{e}$, let V_{max}, and V_{min} be the vertices of K where ψ assumes maximum and minimum value respectively, and let V_{med} be the third vertex. The edge $\tilde{e}$ is taken as the edge connecting V_{max} and V_{med}, and $\widetilde{\psi}$ is computed by

$$e^{-\widetilde{\psi}}{}_{|K} := |\tilde{e}|/(\int_{\tilde{e}} e^{\psi} ds)\,, \qquad \tilde{e} = V_{max}V_{med}, \tag{4.9}$$

or, with the notation of (3.273), by $e^{-\widetilde{\psi}}{}_{|K} := \mathcal{H}_{\tilde{e}}(e^{\psi})$. Therefore, instead of (3.218), we have

$$\overline{\gamma} = ce^{-\widetilde{\psi}}. \tag{4.10}$$

For the reader's convenience, we recall the definition of the finite dimensional spaces already introduced in (3.148), (3.51), (3.160)

$$\begin{aligned}
\widehat{\Sigma}_h &= \left\{\underline{\mathcal{I}}_h \in (L^2(\Omega))^2 \mid \underline{\mathcal{I}}_{h|K} \in Q(K)\,, \forall K \in \mathcal{T}_h\right\}, \\
V_h &= \left\{v_h \in V \mid v_{h|K} \in P_0(K)\,, \forall\, K \in \mathcal{T}_h\right\}, \\
\Lambda_{h,\xi} &= \left\{\mu_h \in L^2(\mathcal{E}_h) \mid \mu_{h|e} \in P_0(e) \forall e \in \mathcal{E}_h, \int_e (\mu_h - \xi) ds = 0, \forall e \in \mathcal{E}_h \cap \Gamma_D\right\},
\end{aligned}$$

where, as usual, $P_0(D)$ denotes the set of constant functions in the domain D. Moreover, $Q(K)$ denotes here a set of polynomial vectors with $Q(K) = span\,\{\underline{\mathcal{I}}^1, \underline{\mathcal{I}}^2, \underline{\mathcal{I}}^3\}$. In the three cases $\underline{\mathcal{I}}^1 = (1, 0)\,,\ \underline{\mathcal{I}}^2 = (0, 1)$; instead $\underline{\mathcal{I}}^3$ is defined by (3.216) for the RT_0

element, by (3.115)–(3.116) for the element of *Example 5* and by (3.128)–(3.129) for the element of *Example 6*.

Specializing (3.161) to this case, the discrete formulation of (4.5), with $\underline{J}$ defined as in (4.6), becomes

$$\begin{cases} \text{Find } (\underline{J}_h, \rho_h, \lambda_h) \in \widehat{\Sigma}_h \times V_h \times \Lambda_{h,\chi} \text{ such that} \\ \displaystyle\int_\Omega e^{\overline{\psi}} \underline{J}_h \cdot \underline{\tau}_h dx - \sum_K \int_K \rho_h \mathrm{div}\, \underline{\tau}_h \, dx + \sum_K \int_{\partial K} \lambda_h \underline{\tau}_h \cdot \underline{n} \, ds = 0 & \forall \underline{\tau}_h \in \widehat{\Sigma}_h, \\ \displaystyle -\sum_K \int_K v_h \mathrm{div}\, \underline{J}_h \, dx - \sum_K \int_K c e^{-\widetilde{\psi}} \rho_h \, v_h \, dx = -\int_\Omega f \, v_h \, dx & \forall v_h \in V_h, \\ \displaystyle \sum_K \int_{\partial K} \mu_h \underline{J}_h \cdot \underline{n} \, ds = 0 & \forall \mu_h \in \Lambda_{h,0}. \end{cases} \tag{4.11}$$

Performing the elimination of $\underline{J}_h$ and ρ_h by static condensation as in Section 3.7, we obtain a final matrix $\mathcal{M}$ acting only on the Lagrange multipliers λ_h, which we recall to be an approximation of ρ on the edges. In the present case, the coefficients of the element matrix $\mathcal{M}^K$ defined in (3.225) take the form

$$m_{ij}^K = e^{-\overline{\psi}} \frac{\underline{\nu}^i \cdot \underline{\nu}^j}{|K|} + \frac{c e^{-\widetilde{\psi}} |K|}{\beta^2 + \delta \, c \, e^{\overline{\psi} - \widetilde{\psi}} |K|} \eta_i \eta_j \qquad i,j = 1,3, \tag{4.12}$$

and the right-hand side takes the form

$$g_i^K = \frac{\beta \eta_i}{\beta^2 + \delta \, c \, e^{\overline{\psi} - \widetilde{\psi}} |K|} \int_K f \, dx \qquad i = 1,3, \tag{4.13}$$

where δ, β, η_i are defined in (3.220)-(3.222). We are now ready to introduce an approximation of the original variable p. Since λ_h lives on the edges, a discrete version of the inverse transformation of (4.4) must be defined edge by edge. For this, for every $\zeta \in L^2(\mathcal{E}_h)$, we define ζ^I to be the L^2-projection of ζ onto $\Lambda_{h,\zeta}$, that is,

$$\zeta^I{}_{|e} = \frac{1}{|e|} \int_e \zeta \, ds \qquad \forall e \in \mathcal{E}_h \, . \tag{4.14}$$

The discrete change of variable is then

$$\lambda_h = (e^{\psi})^I \, p_h \, , \tag{4.15}$$

and (4.11) becomes

$$\begin{cases} \text{Find } (\underline{J}_h, \rho_h, p_h) \in \widehat{\Sigma}_h \times V_h \times \Lambda_{h,g} \text{ such that} \\ \displaystyle\int_\Omega e^{\overline{\psi}} \underline{J}_h \cdot \underline{\tau}_h dx - \sum_K \int_K \rho_h \mathrm{div}\, \underline{\tau}_h \, dx + \sum_K \int_{\partial K} (e^{\psi})^I p_h \underline{\tau}_h \cdot \underline{n} \, ds = 0 & \forall \underline{\tau}_h \in \widehat{\Sigma}_h, \\ \displaystyle -\sum_K \int_K v_h \mathrm{div}\, \underline{J}_h \, dx - \sum_K \int_K c e^{-\widetilde{\psi}} \rho_h \, v_h \, dx = -\int_\Omega f \, v_h \, dx & \forall v_h \in V_h, \\ \displaystyle \sum_K \int_{\partial K} \mu_h \underline{J}_h \cdot \underline{n} \, ds = 0 & \forall \mu_h \in \Lambda_{h,0}, \end{cases} \tag{4.16}$$

where p_h is an approximation of p on the edges. We point out that (4.15) is intended on $\mathcal{E}_h \backslash \Gamma_D$ only, unless g is piecewise constant on Γ_D (which is the case in most applications). In that case, $\lambda_h \in \Lambda_{h,\chi}$ and $p_h \in \Lambda_{h,g}$ are equivalent.

Performing transformation (4.15) at the matrix level (for the nodes not belonging to Γ_D) amounts to multiplying the matrix $\mathcal{M}$ columnwise by the value of $(e^{\psi})^I$ on the corresponding edge, thus giving rise to the matrix $\widetilde{\mathcal{M}}$ acting on the variable p_h. The algebraic system to be solved is then

$$\widetilde{\mathcal{M}} p_h = \mathcal{G}\,. \tag{4.17}$$

The matrix $\widetilde{\mathcal{M}}$ is not symmetric anymore (as expected, since it corresponds to the non symmetric problem (4.1)), but if $\mathcal{M}$ is an M–matrix this property is preserved in $\widetilde{\mathcal{M}}$, since $(e^{\psi})^I$ is always positive. The coefficients of the stiffness matrix $\widetilde{\mathcal{M}}$ take the form

$$\widetilde{m}_{ij}^K = (e^{\psi})^I{}_{|e^j}\, e^{-\overline{\psi}}\, \frac{\underline{\nu}^i \cdot \underline{\nu}^j}{|K|} + (e^{\psi})^I{}_{|e^j}\, e^{-\widetilde{\psi}} \frac{c|K|}{\beta^2 + \delta\, c\, e^{\overline{\psi}-\widetilde{\psi}}|K|} \eta_i \eta_j \qquad i,j = 1,3. \tag{4.18}$$

In the case of the elements of *Examples 5-6*, formula (4.18) can be simplified. Indeed, taking the special edge $\tilde{e}$ (denoted also by e^1 in (3.231)) appearing in the definition of these elements ((3.116), (3.129)) as the edge $V_{max}V_{med}$ (see (4.9)), one can see that

$$(e^{\psi})^I{}_{|e^1}\, e^{-\widetilde{\psi}} = 1\,, \tag{4.19}$$

and (4.18) reduces to (see also (3.232))

$$\widetilde{m}_{ij}^K = \begin{cases} (e^{\psi})^I{}_{|e^1}\, e^{-\overline{\psi}}\, \dfrac{\underline{\nu}^1 \cdot \underline{\nu}^1}{|K|} + \dfrac{c|K|}{\beta^2 + \delta\, c\, e^{\overline{\psi}-\widetilde{\psi}}|K|} \eta_1^2 & \text{for } i = j = 1, \\ (e^{\psi})^I{}_{|e^j}\, e^{-\overline{\psi}}\, \dfrac{\underline{\nu}^i \cdot \underline{\nu}^j}{|K|} & \text{otherwise,} \end{cases} \tag{4.20}$$

with $\eta_1 = |e^1|$ for *Example 5*, and $\eta_1 = 1$ for *Example 6*. Using the results presented at the end of Section 3.7, if the triangulation is of weakly acute type, the two elements of *Examples 5-6* provide a final matrix $\widetilde{\mathcal{M}}$ which is an M-matrix for all $c \geq 0$. Instead, the element RT_0 guarantees the M-matrix property only if $c = 0$.

Remark 4.1 *Few remarks on numerical "tricks". For the computation of* $\int_K f\,dx$ *in* (4.13) *a quadrature formula which is exact for constant* f *can be used. Exact integration can be used for computing* $e^{\overline{\psi}}$, $e^{\widetilde{\psi}}$ *defined in* (4.7), (4.9), *since* ψ *is piecewise linear. The choice of* $\tilde{e}$ *as the edge which connects the vertex with the largest potential value and the vertex with the second largest potential value takes care of all possible cases for the potential:* $\psi \equiv$ *constant on* K, $\psi =$ *constant* $= \psi^M$ *on one edge, and* ψ *having the maximum in one vertex only.*

As already pointed out, the electric field $\underline{E}(= -\underline{\nabla}\,\psi)$ can be quite large in a portion of the domain, so that the presence of exponentials in the coefficients might be a source of numerical problems. A (rough) analysis of the behaviour of the coefficients when the electric field becomes larger and larger will be performed. It is more convenient to set

$$\psi = \frac{\psi_0}{l}\,, \tag{4.21}$$

and assume that $\underline{\nabla}\,\psi_0$ is smooth everywhere and l is a small number. Accordingly, equation (4.3) becomes

$$-\mathrm{div}\left(\underline{\nabla}\,p + p\frac{\underline{\nabla}\,\psi_0}{l}\right) + c\,p = f\,. \tag{4.22}$$

The nature of equation (4.22) is such that, as $l \longrightarrow 0$, the differential term behaves like l^{-1}, while the zeroth-order term is of order 1. Hence, for very small l (say $l << |\underline{\nabla}\,\psi_0|h_K$), our discrete scheme must reproduce the behavior of the continuous equation (4.22). To check that, recall that ψ (and then ψ_0) is assumed piecewise linear, and denote by ψ^M the maximum of ψ on K and by ψ^{M_j} the maximum of ψ on the edge e^j. We only consider the generic case where the maximum is reached at one vertex. When $l << |\underline{\nabla}\,\psi_0|h_K$, a simple computation shows that

$$e^{\overline{\psi}} \;=\; \frac{1}{|K|}\int_K e^{\psi_0/l}dx \simeq l^2 e^{\psi_0^M/l} = l^2\, e^{\psi^M}, \tag{4.23}$$

$$(e^{\psi})^I{}_{|e^j} \;=\; \frac{1}{|e^j|}\int_{e^j} e^{\psi_0/l}ds \simeq l e^{\psi_0^{M_j}/l} = l\, e^{\psi^{M_j}}. \tag{4.24}$$

Moreover, recalling that $\tilde{e} = V_{max}V_{med}$, we have

$$e^{\widetilde{\psi}} \simeq l\, e^{\psi^M}\,. \tag{4.25}$$

Then, we obtain

$$(e^{\psi})^I{}_{|e^j}\, e^{-\overline{\psi}} \simeq \begin{cases} l^{-1} & \text{if } \psi^{M_j} = \psi^M\,, \\ 0 & \text{otherwise,} \end{cases} \tag{4.26}$$

$$(e^{\psi})^I{}_{|e^j}\, e^{-\widetilde{\psi}} \simeq \begin{cases} 1 & \text{if } \psi^{M_j} = \psi^M\,, \\ 0 & \text{otherwise,} \end{cases} \tag{4.27}$$

$$e^{\overline{\psi}}\, e^{-\widetilde{\psi}} \simeq l\,. \tag{4.28}$$

Hence, coefficients (4.18) behave as

$$\widetilde{m}_{ij}^K \simeq \begin{cases} l^{-1}\,\dfrac{\underline{\nu}^i \cdot \underline{\nu}^j}{|K|} + \dfrac{c|K|}{\beta^2 + \delta\, c\, l\, |K|}\eta_i\eta_j & \text{if } \psi^{M_j} = \psi^M\,, \\ 0 & \text{otherwise.} \end{cases} \tag{4.29}$$

The reason for the choice (4.9) is now clear. The expected behaviour in terms of the order of magnitude with respect to l is preserved and, moreover, no bad blow–up occurs. Different (and maybe more natural) choices for $e^{-\widetilde{\psi}}$ could lead to a coefficient for the zeroth-order term in which $(e^{\psi})^I{}_{|e^j}\, e^{-\widetilde{\psi}}$ is not of order 1 when $\psi^{M_j} = \psi^M$. Then, the presence of this factor could give rise to schemes whose structure does not fit the structure of the continuous problem and which produce poor results, unless the mesh size is very small.

To conclude this section, we summarize the main features of the discretization schemes presented here.

- Current conservation.

  The discretization schemes presented here are based on mixed finite elements and enforce some continuity of the normal component of the current at the interelements. As extensively discussed in Section 3, strong continuity is imposed when conforming approximations of $H(\mathrm{div}\,;\Omega)$ are considered (RT_0 or the element of *Example 5*), weak continuity (in the sense that the jump of the normal component of the current has zero mean value at the interelements) is imposed for the element of *Example 6*. This property is particularly relevant here, since the current is possibly the most important output of device simulation.

- Automatic upwinding effects.

  The expression (4.29) tells us that whenever $|\underline{\nabla}\,\psi|$ is large, the coefficient corresponding to the node on the edge where ψ does not reach its maximum is zero (with respect to the machine precision). Such a node can be regarded as downwind node (wind$= -\underline{\nabla}\,\psi$) and the scheme as an upwind scheme. In a sense, the scheme automatically adapts to the changed nature of the problem when advection becomes bigger than diffusion, and chooses the upwind nodes with no extra computational cost. The numerical tests presented in Section 7.3 will show that the numerical diffusion introduced by the schemes is very small in applications to semiconductor device problems.

- M-matrix property.

  The two elements of *Examples 5-6* provide a final matrix $\widetilde{\mathcal{M}}$ which is an M-matrix for all $c \geq 0$. Instead, the element RT_0 guarantees the M-matrix property only if $c = 0$. The M-matrix property implies a sort of discrete maximum principle, and this entails, in particular, that the discrete solution is nonnegative if the boundary data are nonnegative. This is highly desirable in this case since, on the one hand, the variables we are dealing with (charge densities) are intrinsecally nonnegative. On the other hand, the possible development of spurious oscillations might heavily pollute the nonlinear Gauss–Seidel iterations, thus seriously compromising the results.

4.2 MFV approximation of the continuity equation

In this section we address the discretization of the current continuity equation (4.3) using the MFV method discussed in Section 3.8.2. Using the symmetric form (4.5) of (4.3), the MFV discrete formulation (3.251) reads in the present case

$$\begin{cases} \displaystyle\sum_{r\in E(k)} e^{-\overline{\psi}_r}\left(\frac{\rho_k^*-\rho_{j(r)}^*}{d^r}\right)|e^r| + \rho_k^* c_k e^{-\psi_k}|K_k| = f_k|K_k| \quad \forall K_k\in\mathcal{T}_h \\ \rho_{j(r)}^* = \chi^r = e^{\psi_r} g^r \qquad\qquad \forall e^r\in\Gamma_D, \end{cases} \tag{4.30}$$

where $j(r) = T(r)\backslash k$. We note that in the formulation (3.251) v_k denotes the mean value of the function v over the triangle K_k. However, in practical implementation it is easier to use, instead of the mean value, the value of v at the circumcenter C_k of every $K_k\in\mathcal{T}_h$. Similarly, $\chi^r = e^{\psi_r}g^r$ is taken as the value of the function $e^{\psi}g$ at the midpoint of each edge $e^r\in\Gamma_D$. The quantity $e^{-\overline{\psi}_r}$ could be set equal to one of the three averages discussed in Section 3.8.4. In the context of semiconductor device simulation, the more appropriate choice is the harmonic average (3.275) of e^{ψ} along the segment d_r joining the circumcenters C_k and $C_{j(r)}$. With this choice we obtain

$$e^{-\overline{\psi}_r} = e^{-\psi_k}\frac{\psi_{j(r)}-\psi_k}{e^{\psi_{j(r)}-\psi_k}-1} \equiv e^{-\psi_k}\mathrm{B}(\psi_{j(r)}-\psi_k) \qquad \forall r\in E(k), \tag{4.31}$$

where

$$\mathrm{B}(t) = \begin{cases} \dfrac{t}{e^t-1} & t\neq 0 \\ 1 & t=0 \end{cases} \tag{4.32}$$

is the Bernoulli function. Substituting (4.31) into (4.30) leaves us with solving the following symmetric and positive definite linear system acting on the variable $\boldsymbol{\rho}^*$

$$W^*\boldsymbol{\rho}^* = \mathbf{f}^* \tag{4.33}$$

whose nonzero matrix entries are

$$W_{ij}^* = \begin{cases} \displaystyle\sum_{\substack{k\in T(E(i))\\ k\neq i}} e^{-\psi_i}\mathrm{B}(\psi_k-\psi_i)\frac{|e^{r(k)}|}{d^{r(k)}} + c_i e^{-\psi_i}|K_i|, & \text{if } i=j \\ -e^{-\psi_i}\mathrm{B}(\psi_{r(j)}-\psi_i)\dfrac{|e^{r(j)}|}{d^{r(j)}}, & \text{if } j\in T(E(i)),\quad j\neq i, \end{cases} \tag{4.34}$$

where $r(k) = E(k)\cap E(i)$ and $r(j) = E(j)\cap E(i)$. As already pointed out in the previous section, the presence of the exponential terms in (4.34) is a source of trouble in numerical computations due to potential occurrence of overflow/underflow problems. Therefore, we go back to the primitive variable p by applying (4.4) at the discrete level, i.e.

$$\rho_m^* = p_m e^{\psi_m} \qquad \forall K_m\in\mathcal{T}_h. \tag{4.35}$$

Using (4.35) into system (4.33)-(4.34) amounts to multiplying the matrix W^* columnwise by the value of e^{ψ_m} on the corresponding triangle K_m and leads to solving the following linear system acting on the variable $\mathbf{p}^*$

$$\widetilde{W}^*\mathbf{p}^* = \mathbf{f}^*. \tag{4.36}$$

Noting that

$$e^t \mathrm{B}(t) = \mathrm{B}(-t) \qquad \forall\, t \in \mathbb{R}, \tag{4.37}$$

the nonzero entries of matrix $\widetilde{W}^*$ are given by

$$\widetilde{W}^*_{ij} = \begin{cases} \displaystyle\sum_{\substack{k \in T(E(i)) \\ k \neq i}} \mathrm{B}(\psi_k - \psi_i)\frac{|e^{r(k)}|}{d^{r(k)}} + c_i|K_i|, & \text{if } i = j \\ -\mathrm{B}(-(\psi_{r(j)} - \psi_i))\dfrac{|e^{r(j)}|}{d^{r(j)}}, & \text{if } j \in T(E(i)), \quad j \neq i. \end{cases} \tag{4.38}$$

The matrix $\widetilde{W}^*$ is no longer symmetric and positive definite; however, under the assumption that $\mathcal{T}_h$ is a Delaunay triangulation, $\widetilde{W}^*$ turns out to be an M-matrix with strictly positive inverse. As a consequence, a sort of discrete maximum principle holds for the discrete formulation and $\mathbf{p}^* > 0$ provided that $\mathbf{f}^* > 0$, as it is desirable since the variable p has the physical meaning of a (scaled) hole density.

Remark 4.2 *The mixed finite volume* (4.36)–(4.38) *can be regarded as a two-dimensional generalization of the exponentially fitted Scharfetter-Gummel method [146] for the discretization of the current continuity equation* (4.3)*, which will be object of investigation of Section 5. To check this assertion, let us integrate* (4.3) *over a single element* $K_k \in \mathcal{T}_h$*; using the divergence theorem, we obtain*

$$\int_{\partial K_k} \underline{J} \cdot \underline{n}\, ds + \int_{K_k} c\, p\, dx = \int_{K_k} f\, dx,$$

where the current $\underline{J}$ *is defined in* (4.2)*. The approximate evaluation of the net flux throughout the control volume* K_k *can be performed as*

$$\sum_{r \in E(k)} \mathcal{J}_{k,j(r)}|e^r| + c_k p_k |K_k| = f_k|K_k|, \qquad \forall K_k \in \mathcal{T}_h, \tag{4.39}$$

where, for any couple of adjoining triangles K_k*,* $K_{j(r)}$*,* $\mathcal{J}_{k,j(r)}$ *is the constant hole current density flowing along the "pipeline" linking the circumcenters* C_k*,* $C_{j(r)}$ *of the two elements and evaluated according to the classical Scharfetter-Gummel formula [146]*

$$\mathcal{J}_{k,j(r)} = \frac{p_k \mathrm{B}(\psi_{j(r)} - \psi_k) - p_{j(r)}\mathrm{B}(-(\psi_{j(r)} - \psi_k))}{d_r}. \tag{4.40}$$

Substituting (4.40) *into* (4.39) *we exactly get the mixed-finite volume discretization* (4.36)*.*

Remark 4.3 *The mixed-finite volume scheme* (4.36) *can be proved to recover the* exact *solution (p, $\underline{J}$) at the circumcenters of $\mathcal{T}_h$ when $c = 0$, $f = 0$, ψ is linear in Ω, and suitable Dirichlet-Neumann boundary conditions are assumed in problem* (4.3) *in such a way that $\underline{J} = \underline{const}$ (see for the proof [144] and [160] in the case of triangles and rectangles, respectively). This nice property is a special instance of the "patch-test" (see [138], Chap. V, Sect. 34 and [76]) and turns out to be a sound indication for good behavior of a numerical scheme to deal with advection-dominated flow problems, as previously remarked in [160, 142].*

Remark 4.4 *The expression* (4.31) *used to compute the average of e^{ψ} has been chosen by observing that its use in the one-dimensional case leads to a proper exponential fitting interpolation for the carrier concentration p. This gives rise to difference schemes of optimal order (see [140] for a complete survey of this subject). At the same time, the Bernoulli weights in* (4.38) *ensure robustness and stability of the approximation irrespectively of the potential drop $\psi_{r(j)} - \psi_k$ across e^r and provide automatically the suitable upwinding effect, exactly as happens for the standard mixed formulation discussed in Section 4.1.*

5 Other approaches

In this section we shall discuss a number of other finite element methods for the semiconductor device equations. We shall start with the Scharfetter-Gummel box scheme which is very popular in the community of semiconductor device modelling. We shall then discuss some extensions of the box scheme and other stable finite element methods.

5.1 The Scharfetter-Gummel box/finite volume scheme

The key idea of the Scharfetter-Gummel box or finite volume method is to approximate the flux of a boundary value problem along each edge in a mesh by a constant, which yields an exponential approximation to the potential function of the problem. Therefore, it is also called an exponentially fitted method. This method was proposed by Scharfetter and Gummel for a one-dimensional problem [146]. The same idea was earlier introduced by Allen and Southwell [5] in a different context. The one-dimensional Scharfetter-Gummel scheme approximation has been combined with higher dimensional box scheme by several authors to form the so-called Scharfetter-Gummel box method. (cf., for example, [33, 111, 17, 119]). In what follows we discuss the method in two dimensions. The extension to three dimensions is straightforward.

Let us consider the decoupled and linearized equations in the Slotboom variables given in Section 2. These equations are of the form (2.14) in which $a = \lambda^2$ for the Poisson equation, $a = \mu_n e^{\psi}$ for the continuity equation for ρ_n, and $a = \mu_p e^{-\psi}$ for the continuity equation for ρ_p. Without loss of generality, we assume that $g = 0$ on Γ_D. The case of non-homogeneous Dirichlet boundary conditions can be transformed

into this one by subtracting from both sides of (2.14) a known function satisfying a given non-homogeneous boundary condition. We now discuss the discretization of this boundary value problem by the box method. We shall first formulate it as a finite element method and then we present a stability and error analysis for the method. For brevity, we only consider the case that $a = e^{\psi}$, and we let $\underline{\sigma} = e^{\psi}\underline{\nabla} u$. The discretization for the other two cases and for the case that μ_n is not constant are similar to this one.

5.1.1 Formulation of the method

Classically, the Scharfetter-Gummel box method can be formulated as a finite volume method with a constant approximation to the flux projected on each of the mesh edges. This formulation can be found, for example, in [33, 111, 17]. Though the formulation is easy to understand, it is not very suitable for stability and error analysis. In what follows, we shall formulate it as a non-conforming Petrov-Galerkin finite element method as described in [118]. This can be viewed as a generalized finite element method and is closely related to a mixed finite element formulation (cf. [12]). It can also be viewed as a lumped form of a primal mixed finite element method proposed in [117] and [120]. Alternative analysis can be found in [123].

To discuss the method we first define primal and dual decompositions of Ω. Let $\mathcal{T}_h$ be a Delaunay triangulation of Ω (see Definition 3.1). We also let $\mathcal{X}_h = \{x_i\}_1^N$ denote the vertices of $\mathcal{T}_h$ and $\mathcal{E}_h = \{e_i\}_1^M$ the edges of $\mathcal{T}_h$. We assume that the nodes in $\mathcal{X}_h$ and the edges in $\mathcal{E}_h$ are numbered in such a way that $\mathcal{X}_h' = \{x_i\}_1^{N'}$ and $\mathcal{E}_h' = \{e_i\}_1^{M'}$ are the set of nodes in $\mathcal{X}_h$ not on Γ_D and the set of edges in $\mathcal{E}_h$ not on Γ_D, respectively . The first nontriangular mesh, dual to $\mathcal{T}_h$, is the tessellation denoted by $\mathcal{L}_h$ in Section 3.8.1 (see Fig.3, left). For convenience, we repeat here the defintion with a slight change of notation more suited for our purposes. With each edge $e_{i,j} \in \mathcal{E}_h$ connecting two vertices x_i, x_j we associate an open box $B_{i,j}$ which is the interior of the polygon having as its vertices x_i, x_j, and the circumcentres of the triangles having $e_{i,j}$ as a common edge. If $e_{i,j}$ is not on $\partial\Omega$ the region $B_{i,j}$ consists of two triangles, otherwise of only one triangle. A second nontriangular mesh, also dual to $\mathcal{T}_h$, is defined as follows.

Definition 5.1 The Dirichlet tessellation $\mathcal{D}_h$, corresponding to the triangulation $\mathcal{T}_h$, is defined by $\mathcal{D}_h = \cup D_i$, $i = 1, N$, where, for each $x_i \in \mathcal{X}_h$, the tile D_i is given by

$$D_i = \{x \in \Omega : |x - x_i| < |x - x_j|,\ \forall\, x_j \in \mathcal{X}_h, j \neq i\}\,. \tag{5.1}$$

We remark that for each $x_i \in \mathcal{X}_h$, the boundary ∂D_i of the tile D_i is the polygon having as its vertices the circumcentres of all triangles with the common vertex x_i. Each segment of ∂D_i is perpendicular to one of the edges sharing the vertex x_i. The subset of $\mathcal{D}_h$ corresponding to X_h' is denoted by $\mathcal{D}_h' = \cup D_i$, for $i = 1, N'$.

With the two meshes $\mathcal{T}_h$ and $\mathcal{D}_h$ we associate a trial space $V_h \subset L^2(\Omega)$ and a test space $Q_h \subset L^2(\Omega)$, respectively, each of dimension N'.

To construct Q_h we define a set of piecewise constant basis functions ξ_i, $(i = 1, 2, ..., N)$ corresponding to the mesh $\mathcal{D}_h$ as follows

$$\xi_i = \begin{cases} 1 & \text{on } D_i, \\ 0 & \text{otherwise.} \end{cases}$$

We then define $Q_h = \text{span}\,\{\xi_i\}_1^{N'}$. To construct V_h we proceed as follows.

Definition 5.2 *For each edge $e_{i,j} \in \mathcal{E}'_h$, let v be a regular function on $e_{i,j}$. We define $\widehat{v}$ the extension of v to the box $B_{i,j}$ taken constant along all the perpendiculars to $e_{i,j}$.*

For each edge $e_{i,j} \in \mathcal{E}'_h$ connecting the vertices x_i and x_j we define an exponential function $\phi_{i,j}$ on $e_{i,j}$ solution of

$$\begin{cases} \dfrac{d}{ds}(e^{\psi}\dfrac{d\phi_{i,j}}{ds}) = 0 \qquad \text{on} \quad e_{i,j} \\ \\ \phi_{i,j}(x_i) = 1, \quad \phi_{i,j}(x_j) = 0 \end{cases}$$

where d/ds denotes the derivative along the edge $e_{i,j}$ from x_i to x_j . We then extend $\phi_{i,j}$ to $B_{i,j}$ as in definition 5.2. Finally, the basis functions $\phi_i(x)$'s in V_h are taken as

$$\phi_i(x) = \begin{cases} \widehat{\phi}_{i,j} \quad \text{on } B_{i,j}, \ \forall\, j \in I_i \\ \\ 0 \qquad \text{elsewhere} \end{cases}$$

where

$$I_i = \{j \neq i : \ \exists\, e_{i,j} \in \mathcal{E}'_h \text{ connecting } x_i \text{ and } x_j\} \tag{5.2}$$

denotes the index set of all neighbouring vertices of x_i. The support of ϕ_i is star-shaped (see Fig.3, left). Setting $V_h = \text{span}\,\{\phi_i\}_1^{N'}$, we obviously have $V_h \subset L^2(\Omega)$.

For simplicity we make the following assumption.

Assumption 5.1 *For every $K \in \mathcal{T}_h$ the function ψ is linear on K.*

Assumption 5.1 is true in practice. Indeed, the system of equations is normally solved iteratively using the Gummel method discussed in Section 2 of this chapter, and thus ψ is the numerical solution of the decoupled Poisson equation. Therefore, we can always use the piecewise linear interpolant of this numerical solution.
We notice that any $v_h \in V_h$ has the form $v_h(x) = \sum \phi_i(x) v_i$, with $v_i = v_h(x_i)$. Consequently v_h satisfies

$$\begin{cases} \dfrac{d}{ds}(e^{\psi}\dfrac{dv_h}{ds}) = 0 \qquad \text{on } e_{i,j} \\ v_h(x_i) = v_i, \quad v_h(x_j) = v_j. \end{cases}$$

Solving this two-point boundary value problem, and thanks to Assumption 5.1, we have the following flux representation on $e_{i,j}$

$$\sigma_{i,j}(v_h) := e^{\psi}\frac{dv_h}{ds} = e^{\psi_i}\mathrm{B}(\psi_i - \psi_j)\frac{v_j - v_i}{|e_{i,j}|}, \tag{5.3}$$

where $\psi_i = \psi(x_i)$, and $\mathrm{B}(x)$ denotes the Bernoulli function defined in (4.32). The constant value (5.3) is then extended to the whole box $B_{i,j}$. For any sufficiently smooth function w we can define the V_h-interpolant w_I of w, given by $w_I(x) = \sum \phi_i(x)w(x_i)$. Let $\underline{\sigma}(w) := e^{\psi}\underline{\nabla} w$ be the flux associated with w, and let $\underline{\sigma}_{i,j}(w_I)$ the flux associated with w_I as in (5.3). Denoting by $\underline{e}_{i,j}$ the unit tangent vector on $e_{i,j}$, oriented from x_i to x_j, it is easy to see that $\sigma_{i,j}(w_I)$ is the projection of $\underline{\sigma}(w)\cdot\underline{e}_{i,j} \in L^2(e_{i,j})$ onto the space of constant polynomials on $e_{i,j}$ with respect to the weighted inner product $\int_{e_{i,j}} e^{-\psi} f g ds$, $f, g \in L^2(e_{i,j})$. Thus, using Taylor expansion we easily obtain

$$||\underline{\sigma}(w)\cdot\underline{e}_{i,j} - \sigma_{i,j}(w_I)||_{\infty,e_{i,j}} \leq C|e_{i,j}||\underline{\sigma}(w)\cdot\underline{e}_{i,j}|_{1,\infty,e_{i,j}}.$$

Moreover, since $e_{i,j} \subset B_{i,j}$, a continuity argument gives

$$||\underline{\sigma}(w)\cdot\underline{e}_{i,j} - \sigma_{i,j}(w_I)||_{\infty,B_{i,j}} \leq Ch|\underline{\sigma}(w)|_{1,\infty,B_{i,j}}, \tag{5.4}$$

where C is a positive constant, independent of h and w.

We introduce the mass lumping operator $P : C^0(\bar{\Omega}) \mapsto Q_h$ such that

$$P(v)(x) = \sum_{x_i \in \mathcal{X}_h} v(x_i)\xi_i(x) \quad \text{for all} \quad x \in \overline{\Omega}. \tag{5.5}$$

Using the trial and test spaces V_h and Q_h, we now define the following Petrov–Galerkin problem corresponding to (2.14)

Problem 5.1 *Find $u_h \in V_h$ such that for all $q_h \in Q_h$*

$$a(u_h, q_h) + (P(\gamma u_h), q_h) = (\hat{f}, q_h) \tag{5.6}$$

where $\hat{f}$ is an approximation to f and $a(\cdot,\cdot)$ denotes the bilinear form on $V_h \times Q_h$ defined by

$$a(u_h, q_h) = -\sum_{i=1}^{N'} \int_{\partial D_i \setminus \partial\Omega} e^{\widehat{\psi}}\underline{\nabla} u_h \cdot \underline{n}\, q_h \, ds, \tag{5.7}$$

where $\widehat{\psi}$ is taken as in definition 5.2, and $\underline{n}$ is the unit outward normal vector on ∂D_i.

We shall prove that Problem 5.1 has a unique solution by showing that the associated linear system is nonsingular. The solution we are looking for will have the form $u_h(x) = \sum_{i=1}^{N'} u_i\phi_i(x)$. Taking $q_h = \xi_j$ in (5.6), we get

$$-\int_{\partial D_j \setminus \partial\Omega} e^{\widehat{\psi}}\underline{\nabla} u_h \cdot \underline{n}\, ds + \gamma_j u_j |D_j| = \int_{D_j} \hat{f} dx \qquad j = 1, 2, \ldots, N',$$

where $\gamma_j = \gamma(x_j)$. Let the line segment $l_{j,k} = \partial D_j \cap \partial D_k$, so that $\partial D_j = \cup_{k\in I_j} l_{j,k}$, where I_j is the index set defined in (5.2). It is easy to check that

$$|l_{j,k}| = \frac{2|B_{j,k}|}{|e_{j,k}|} \qquad j,k = 1,\ldots N',\ j \neq k. \tag{5.8}$$

Therefore, we have from the above equality

$$-\sum_{k\in I_j} \int_{l_{j,k}} e^{\widehat{\psi}} \underline{\nabla} u_h \cdot \underline{n} ds + \gamma_j u_j |D_j| = \int_{D_j} \hat{f} dx \qquad j = 1,2,\ldots,N'.$$

Noticing that $\underline{n} = \underline{e}_{j,k}$, and $e^{\widehat{\psi}}_{|l_{j,k}} = \textit{constant}$ = value of e^{ψ} in the midpoint of $e_{j,k}$, we can use (5.3) to obtain

$$\sum_{k\in I_j} e^{\psi_j} \mathrm{B}(\psi_j - \psi_k) \frac{u_j - u_k}{|e_{j,k}|} |l_{j,k}| + \gamma_j u_j |D_j| = \int_{D_j} \hat{f} dx \qquad j = 1,2,...,N'. \tag{5.9}$$

The coefficient matrix of this linear system is a symmetric and positive-definite M-matrix, since it is diagonally dominant with positive diagonal elements and negative off-diagonal elements (cf., for example, [161], p.85). Each element of this coefficient matrix depends exponentially on ψ_i for some i. Therefore, the matrix may be computationally very stiff as the values may vary by several orders of magnitude across an element. This drawback can be overcome by transforming the Slotboom variables back to the original ones (electron density in this case) at the discrete level, as done in Section 4. Setting

$$u_i = e^{-\psi_i} w_i \qquad i = 1,2,\ldots,N,$$

and substituting into (5.9) we have

$$\Big(\sum_{k\in I_j} \mathrm{B}(\psi_j - \psi_k) \frac{|l_{j,k}|}{|e_{j,k}|} + \gamma_j e^{-\psi_j} |D_j|\Big) w_j - \sum_{k\in I_j} e^{(\psi_j - \psi_k)} \mathrm{B}(\psi_j - \psi_k) \frac{|l_{j,k}|}{|e_{j,k}|} w_k = \int_{D_j} \hat{f} dx.$$

Thanks to (4.37) the above reduces to

$$\Big(\sum_{k\in I_j} \mathrm{B}(\psi_j - \psi_k) \frac{|l_{j,k}|}{|e_{j,k}|} + \gamma_j e^{-\psi_j} |D_j|\Big) w_j - \sum_{k\in I_j} \mathrm{B}(\psi_k - \psi_j) \frac{|l_{j,k}|}{|e_{j,k}|} w_k = \int_{D_j} \hat{f} dx,$$

for $j = 1,2,\ldots,N'$. Obviously the entries of the coefficient matrix of the above are more balanced than those of (5.9), although the matrix is not symmetric anymore, unless ψ is constant. However, it is diagonally dominant with respect to its columns. Furthermore, it can be shown that the system matrix is an M-matrix (cf.[118, 119]), and thus can be solved by a preconditioned conjugate gradient method, for example the CGS method or the Bi-CGSTAB as discussed in Chapter 9 of this handbook.

5.1.2 Convergence of the approximate solution

We now show stability and convergence of the method just described with respect to a suitable discrete norm defined on V_h. For this it is convenient to rewrite Problem 5.1 in an equivalent form where trial and test spaces coincide. Let $b(\cdot,\cdot)$ be a bilinear form on $V_h \times V_h$ defined by

$$b(u_h, v_h) = a(u_h, P(v_h)) + (P(\gamma u_h), P(v_h)). \tag{5.10}$$

Consider the following Bubnov-Galerkin problem:

Problem 5.2 *Find $u_h \in V_h$ such that for all $v_h \in V_h$*

$$b(u_h, v_h) = (\hat{f}, P(v_h)). \tag{5.11}$$

It is easy to prove the following lemma:

Lemma 5.1 *The mass lumping operator defined in (5.5) is surjective from V_h to Q_h.*

It is obvious from this Lemma that Problem 5.2 is equivalent to Problem 5.1, hence we shall concentrate on Problem 5.2. On V_h we define the discrete energy norm

$$||v_h||^2 = ||v_h||_h^2 + \sum_{i=1}^{N'} \gamma_i v_i^2 |D_i| \qquad \forall\ v_h = \sum_{i=1}^{N'} v_i \phi_i \in V_h, \tag{5.12}$$

where

$$||v_h||_h^2 = \sum_{e_{i,j} \in \mathcal{E}_h'} \left(\frac{v_j - v_i}{|e_{i,j}|} \right)^2 |B_{i,j}| \qquad \forall\ v_h \in V_h. \tag{5.13}$$

It is trivial to check that $||\cdot||_h$ is a norm on V_h, and so is $||\cdot||$. Let

$$\beta = \min_{e_{i,j} \in \mathcal{E}_h'} \frac{|e_{i,j}|}{\int_{e_{i,j}} e^{-\psi} ds}. \tag{5.14}$$

Since $|\psi|$ is bounded, there exists a $\beta_0 > 0$ such that $\beta \geq \beta_0 > 0$. It is also easy to verify that $\beta = \min_{e_{i,j} \in \mathcal{E}_h'} e^{\psi_i} \mathrm{B}(\psi_i - \psi_j)$ because of Assumption 5.1. The following theorem shows the coercivity of the bilinear form $b(\cdot,\cdot)$ with respect to the norm (5.12).

Theorem 5.1 *For all $v_h \in V_h$ we have*

$$b(v_h, v_h) \geq \min\{1, 2\beta\}\, ||v_h||^2. \tag{5.15}$$

Proof. If $v_h = 0$, then (5.15) holds. Let $v_h \neq 0$. Using the method for the derivation of (5.9) we have

$$\begin{aligned} a(v_h, P(v_h)) &= -\sum_{i=1}^{N'} \int_{\partial D_i \setminus \partial\Omega} e^{\widehat{\psi}} \underline{\nabla} v_h \cdot \underline{n} P(v_h)\, ds \\ &= \sum_{e_{i,j} \in \mathcal{E}_h'} (v_j - v_i) \int_{l_{i,j}} e^{\widehat{\psi}} \underline{\nabla} v_h \cdot \underline{e}_{i,j} ds \\ &= \sum_{e_{i,j} \in \mathcal{E}_h'} (v_j - v_i) e^{\psi_i} \mathrm{B}(\psi_i - \psi_j) \frac{v_j - v_i}{|e_{i,j}|} |l_{i,j}| \geq 2\beta \|v_h\|_h^2, \end{aligned}$$

where in the last step we used the relation (5.8). From (5.10) and (5.12) we finally have

$$\begin{aligned} b(v_h, v_h) &= a(v_h, P(v_h)) + (P(\gamma v_h), P(v_h)) \\ &\geq 2\beta \|v_h\|_h^2 + \sum_{i=1}^{N'} \gamma_i v_i^2 |D_i| \geq \min\{1, 2\beta\} \, \|v_h\|^2. \end{aligned}$$

□

Theorem 5.1 implies that the solution to Problem 5.2 is stable with respect to the norm $\|\cdot\|$.

Lemma 5.2 *For any $v_h \in V_h$, there is a constant $C > 0$, independent of h and v_h, such that*

$$\|P(v_h)\|_{0,\Omega} \leq C \|v_h\|_h. \tag{5.16}$$

Proof. The proof of this can be found in [117], Lemma 3.4. □

For any $\underline{p} \in (W^{1,\infty}(\Omega))^2$ we define

$$|\underline{p}|_{1,\infty,h} = \left(\sum_{e_{i,j} \in \mathcal{E}_h'} |\underline{p}|_{1,\infty,B_{i,j}}^2 |B_{i,j}| \right)^{1/2}. \tag{5.17}$$

Obviously $|\cdot|_{1,\infty,h}$ is only a seminorm on $(W^{1,\infty}(\Omega))^2$. The following theorem establishes the convergence of the approximate solution u_h to the V_h-interpolant of u.

Theorem 5.2 *Let u_h be the solution of Problem 5.2 and let u_I be the V_h-interpolant of the solution of problem* (2.14)*. Then there is a constant $C > 0$, independent of h, u and ψ, such that*

$$\|u_h - u_I\| \leq \frac{C}{\min\{1, 2\beta\}} (h |\underline{\sigma}|_{1,\infty,h} + \|\gamma u - P(\gamma u)\|_0 + \|f - \hat{f}\|_0). \tag{5.18}$$

Proof. For the proof we refer the interested reader to [118]. □

We remark that depending on the decoupling technique used for the two continuity equations, we may have $\gamma = 0$ in the above (as the case in Gummel's original work). In such cases the error bound (5.18) depends only on the seminorm of the flux $\underline{\sigma}$ and the approximation error of the source term, while error bounds in the energy norm for classical linear finite element methods depend on $||u||_{2,\Omega}$.

Finally we remark that the approximate flux $\underline{\sigma}_h := e^{\widehat{\psi}} \underline{\nabla} u_h$ does not converge to the exact flux $\underline{\sigma} = e^{\psi} \underline{\nabla} u$. This is because in each box $B_{i,j}$, $\underline{\sigma}_h \equiv \sigma_{i,j}(u_h) \underline{e}_{i,j}$ which converges locally only to $\underline{\sigma} \cdot \underline{e}_{i,j}$ (see (5.4)). However, by post processing it is easy to define an approximate flux which converges to the exact one. For example, we can define

$$\underline{\sigma}_{h|B_{i,j}} = \sigma_{i,j}(u_h) \underline{e}_{i,j} + \frac{\int_{l_{i,j}} \underline{\nabla} u_h \cdot \underline{l}_{i,j} ds}{\int_{l_{i,j}} e^{-\psi} ds} \underline{l}_{i,j}$$

for all $e_{i,j} \in \mathcal{E}'_h$, where $\underline{l}_{i,j}$ denotes the unit tangential vector along $l_{i,j}$. Moreover the computed ohmic contact currents are convergent, as is shown in the next subsection.

5.1.3 The Evaluation of the Ohmic Contact Currents

The ultimate goal of device simulation is to find the terminal currents. We now consider the evaluation of the ohmic contact currents. For simplicity, we restrict our attention to a device with a finite number of ohmic contacts, and so Γ_D is a finite set of separated contacts Γ_c's. For any $\Gamma_c \subset \Gamma_D$, let $\{x_i^c\}_1^{N_c}$ denote the mesh nodes on Γ_c.

Let ξ^c be a piecewise constant function satisfying

$$\xi^c(x) = \begin{cases} 1 & x \in \cup_{i=1}^{N_c} D_i^c \\ 0 & \text{otherwise} \end{cases} \tag{5.19}$$

where D_i^c denotes the element in $\mathcal{D}_h$ containing x_i^c. Taking, for simplicity, $\gamma = 0$ in (2.14), multiplying by ξ^c and integrating by parts we have

$$-\int_{\Gamma_c} \underline{\sigma} \cdot \underline{n} ds - \sum_{i=1}^{N_c} \int_{\partial D_i^c \setminus \Gamma_c} \underline{\sigma} \cdot \underline{n} ds = (f, \xi^c).$$

Thus, the (scalar) outflow current through Γ_c is

$$J_c := \int_{\Gamma_c} \underline{\sigma} \cdot \underline{n} ds = -\sum_{i=1}^{N_c} \int_{\partial D_i^c \setminus \Gamma_c} \underline{\sigma} \cdot \underline{n} ds - (f, \xi^c). \tag{5.20}$$

Replacing $\underline{\sigma}$ by the approximate flux $\underline{\sigma}_h = e^{\widehat{\psi}} \underline{\nabla} u_h$ and f by the approximation $\hat{f}$, we define the following approximate outflow current through Γ_c

$$J_{c,h} := -\sum_{i=1}^{N_c} \int_{\partial D_i^c \setminus \Gamma_c} \underline{\sigma}_h \cdot \underline{n} ds - (\hat{f}, \xi^c). \tag{5.21}$$

From (5.19), (5.3) and the argument used in the derivation of (5.9), we obtain

$$\begin{aligned} J_{c,h} &= -\sum_{j=1}^{N_c}\left[\int_{\partial D_j^c\setminus\Gamma_c}\underline{\sigma}_h\cdot\underline{n}ds+\int_{D_j^c}\hat{f}dx\right] \\ &= \sum_{j=1}^{N_c}\left[\sum_{k\in I_j,x_k\notin\Gamma_c}e^{\psi_j}\mathrm{B}(\psi_j-\psi_k)\frac{2|B_{j,k}|}{|e_{j,k}|}\frac{u_j-u_k}{|e_{j,k}|}-\int_{D_j^c}\hat{f}dx\right] \end{aligned}$$

where I_j is the index set of neighbouring nodes of x_j as defined in (5.2).

The convergence and the conservation of the computed ohmic contact currents are established in the following theorem.

Theorem 5.3 *Let J_c and $J_{c,h}$ be the exact and the computed outflow currents through $\Gamma_c\subset\Gamma_D$, respectively. Then, there exists a constant $C>0$, independent of h, ψ and u, such that*

$$|J_c-J_{c,h}|\le C(h|\underline{\sigma}|_{1,\infty,\mathcal{B}_h}+||f-\hat{f}||_0).$$

Furthermore

$$\sum_{\Gamma_c\subset\Gamma_D}J_{c,h}=-\int_\Omega\hat{f}dx.$$

Proof. The proof is omitted and we refer the interested reader to [118]. □

5.2 Finite element methods based on ad hoc chosen basis functions

The key idea of Scharfetter-Gummel box method is to approximate the flux by a constant on each mesh line. This is in contrast to classic approaches in which a potential function is approximated by a piecewise polynomial. Physically, a flux is better behaved than the corresponding potential function. It is often bounded with respect to the singular perturbation parameter, though this is not proved. Thus, approximation of a flux by piecewise polynomials such as piecewise constant provides a more stable and accurate numerical scheme for the semiconductor device equations. From the construction of the basis functions in the previous section we see that, in each box element in $\mathcal{L}_h$, the Scharfetter-Gummel method is based on a 'divergence-free' approximation of the flux along the mesh edge and its perpendiculars. A more general choice is to seek divergence-free basis functions on the elements, i.e., seek solutions to $\mathrm{div}\,\underline{\sigma}=0$ in each element with appropriate boundary condition. However, solving this problem is equivalent to solving the original problem (2.14). Note that if we only require that

such a basis function is unity at one node and zero at all other nodes of the element, it is rather arbitrary. So, we may restrict ourselves to the case that

$$\frac{\partial \sigma_x}{\partial x} = 0 = \frac{\partial \sigma_y}{\partial y}$$

where σ_x and σ_y denote the two components of $\underline{\sigma}$. This also courses the problem that there are only two degrees of freedom so that only two of the three vertices in each triangle are needed to determine this basis function. Therefore, on each triangle one can define three hat functions which are unity at one vertex and zero at the other two vertices of the triangle, though the difference between any two of them is small. Of course we may choose only one hat function or the average of the three as our basis. In [17] the authors proposed such a divergence-free flux approximation on triangular elements. An improvement of this divergence-free basis function on triangles is given in [145]. All these basis functions may be non-conforming, i.e., they are not continuous across element boundaries. A similar approach can also be found in [108] in which a rigorous analysis of the method is also given. A novel flux approximation technique based on that of the Scharfetter-Gummel was proposed in [151], which makes use of three degrees of freedom on each triangle. This technique has the property that, at each point in a triangle, the approximate flux is constant along each of the lines connecting the vertices of the triangle and the point. The same idea was also proposed independently in [142]. Based on this idea, a set of conforming basis functions is proposed in [164] and a rigorous analysis for the method applied to semiconductor device equations is given in [165]. This analysis is based on a weighted $L^2(\Omega)$ inner product weak formulation which coincides with that in [67], while the finite element basis functions used in [165] coincide with those in [142]. It is interesting to note that the method can also be formulated as a mixed finite element method. In what follows we shall discuss the formulation and analysis of the method. For brevity, we shall omit most of the proofs and refer the reader to the relevant references.

5.2.1 The weak formulation

For brevity we only consider the discretization of the scaled continuity equation for the electron density. The corresponding convection-diffusion problem is of the form

$$\begin{cases} -\mathrm{div}\,\underline{\sigma} + \gamma u = f & \text{in } \Omega \\ u = g & \text{on } \Gamma_D \\ \underline{\sigma} \cdot \underline{n} = 0 & \text{on } \Gamma_N. \end{cases} \tag{5.22}$$

with the flux defined by

$$\underline{\sigma} = \underline{\nabla} u - u \underline{\nabla} \psi.$$

Without loss of generality, we assume that $g = 0$ as in the previous section. Now, we define a weighted inner product $(\cdot,\cdot)_\psi$ on $L^2(\Omega)$ and on $(L^2(\Omega))^2$ by

$$(v, w)_\psi = (e^{-\psi} v, w). \tag{5.23}$$

The $L^2(\Omega)$-norm corresponding to this weighted inner product is denoted by $||\cdot||_{0,\psi}$. Using this inner product we define the following variational problem corresponding to (5.22)

Problem 5.3 *Find* $u \in H^1_D(\Omega) := \{v \in H^1(\Omega) |\ v_{|\Gamma_D} = 0\}$ *such that for all* $v \in H^1_D(\Omega)$

$$A_\psi(u, v) = (f, v)_\psi, \tag{5.24}$$

where $A_\psi(\cdot,\cdot)$ *is the bilinear form on* $(H^1(\Omega))^2$ *defined by*

$$A_\psi(u, v) = (\underline{\nabla} u - \underline{\nabla}\psi u, \underline{\nabla} v - \underline{\nabla}\psi v)_\psi + (\gamma u, v)_\psi. \tag{5.25}$$

Since $A_\psi(\cdot,\cdot)$ is symmetric and positive definite on $H^1_D(\Omega)$, we can associate with A_ψ the norm

$$|||v|||^2_{1,\psi} = A_\psi(v, v), \quad v \in H^1_D(\Omega). \tag{5.26}$$

Existence and uniqueness of the solution of (5.24) follow immediately from (5.26).

5.2.2 The finite element method

Let $\{\mathcal{T}_h\}_h$ be a regular sequence of decompositions of $\overline{\Omega}$ into triangles (see (3.45)). As in the previous section we denote by $\mathcal{X}_h = \{x_i\}_1^N$ the set of vertices of $\mathcal{T}_h$, numbered in such a way that $\{x_i\}_1^{N'}$ is the set of vertices not on Γ_D. As before, we make Assumption 5.1, so that the vector $\underline{a} = \underline{\nabla}\psi$ is constant on each triangle $K \in \mathcal{T}_h$.

Corresponding to the mesh $\mathcal{T}_h$, we now construct a space $S_h \subset H^1_D(\Omega)$ of dimension N' using the basis functions $\{\phi_i\}_1^{N'}$ defined below. Let $K \in \mathcal{T}_h$ be a triangle with vertices x_i, x_j and x_k. We define a local function $\phi_i(x)$ on K associated with x_i as follows. For any point $x \in K$ we denote by $\underline{l}_m$ $(m = i, j, k)$ the vector of length $|l_m|$ connecting x_m to x, and by $\underline{e}_m := \underline{l}_m/|l_m|$ $(m = i, j, k)$ the unit vector from x_m to x (cf. Figure 7). We now consider the following two-point boundary value problem on the segment l_m: find $g_i(s)$ such that

$$\begin{cases} \dfrac{d}{ds}(\underline{p}_i \cdot \underline{e}_m) := \dfrac{d}{ds}\Big(\dfrac{dg_i(s)}{ds} - \underline{a} \cdot \underline{e}_m g_i(s)\Big) = 0 \quad s \in (0, |l_m|) \\ g_i(0) = \delta_{im}, \quad g_i(|l_m|) = \phi_i(x) \end{cases}$$

for $m = i, j, k$, (i.e. $\underline{p}_i \cdot \underline{e}_m$ is constant on l_m), where δ_{im} denotes the Kronecker delta and $\phi_i(x)$ is yet to be determined. Solving the above boundary value problem yields

$$\underline{p}_i \cdot \underline{e}_m = \frac{1}{|l_m|}\Big[\mathrm{B}(a_m)\phi_i(x) - \mathrm{B}(-a_m)\delta_{im}\Big], \ \ \forall\, x \in l_m, \ \ m = i, j, k,$$

where $a_m = \underline{a} \cdot \underline{l}_m$ and $\mathrm{B}(z)$ denotes the Bernoulli function defined in (4.32). The above equation motivates us to define the following problem.

Problem 5.4 *Find, for all* $x \in \overline{K}$, $\phi_i(x)$ *and* $\underline{p}_i = (p_{i,1}, p_{i,2})$ *such that*

$$D(x) \begin{pmatrix} p_{i,1} \\ p_{i,2} \\ \phi_i(x) \end{pmatrix} = \begin{pmatrix} -\mathrm{B}(-a_i) \\ 0 \\ 0 \end{pmatrix} \tag{5.27}$$

where $D(x)$ *is a* 3×3 *matrix defined by*

$$D(x) = \begin{pmatrix} l_{i,1} & l_{i,2} & -\mathrm{B}(a_i) \\ l_{j,1} & l_{j,2} & -\mathrm{B}(a_j) \\ l_{k,1} & l_{k,2} & -\mathrm{B}(a_k) \end{pmatrix}. \tag{5.28}$$

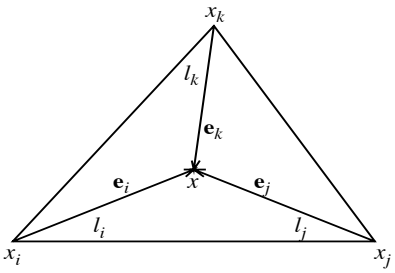


Figure 7: Notation associated with the triangle K.

Solving Problem 5.4 for all $x \in \overline{K}$ defines the point values of the function ϕ_i and an auxiliary flux $\underline{p}_i$. Similarly we can define functions ϕ_j and ϕ_k associated with x_j and x_k respectively. The following theorem shows that Problem 5.4 is uniquely solvable for all $x \in \overline{K}$, and that ϕ_i, ϕ_j and ϕ_k form a system of local basis functions.

Theorem 5.4 *Let* $K \in \mathcal{T}_h$. *Then, for any* $x \in \overline{K}$, *there exists a unique solution to Problem 5.4. Furthermore, we have*

$$\phi_i(x_i) = 1, \quad \phi_i(x) = 0 \ \ \forall x \in \overline{x_j x}_k, \tag{5.29}$$

$$\phi_i + \phi_j + \phi_k = 1, \quad \underline{p}_i + \underline{p}_j + \underline{p}_k = -\underline{a} \qquad \text{in } \overline{K}, \tag{5.30}$$

where $\overline{x_j x}_k$ *denotes the edge of* K *connecting* x_j *and* x_k.

Proof. To prove that Problem 5.4 is uniquely solvable we need only to show that for any $x \in \overline{K}$ the system matrix $D(x)$ is non-singular, or $\det D(x) \neq 0$. From (5.28) we have, by direct computation,

$$\begin{aligned} \det D(x) &= \mathrm{B}(a_k)(l_{j,1}l_{i,2} - l_{j,2}l_{i,1}) + \mathrm{B}(a_i)(l_{k,1}l_{j,2} - l_{k,2}l_{j,1}) + \mathrm{B}(a_j)(l_{i,1}l_{k,2} - l_{i,2}l_{k,1}) \\ &= -\Big[\mathrm{B}(a_k)\underline{e}_z \cdot (\underline{l}_i \times \underline{l}_j) + \mathrm{B}(a_i)\underline{e}_z \cdot (\underline{l}_j \times \underline{l}_k) + \mathrm{B}(a_j)\underline{e}_z \cdot (\underline{l}_k \times \underline{l}_i)\Big] \end{aligned} \tag{5.31}$$

with $\underline{e}_z = (0,0,1)$ the unit vector perpendicular to K. From the orientations of $\underline{l}_i, \underline{l}_j, \underline{l}_k$ and $\underline{e}_z$ (cf. Figure 7) we see that $\underline{e}_z \cdot (\underline{l}_i \times \underline{l}_j), \underline{e}_z \cdot (\underline{l}_j \times \underline{l}_k)$ and $\underline{e}_z \cdot (\underline{l}_k \times \underline{l}_i)$ are all nonnegative, and at least two of them are positive. Furthermore, since $\mathrm{B}(\cdot)$ is always positive and at least two of $|l_i|, |l_j|$ and $|l_k|$ are not zero, we have $\det D(x) \neq 0$.

Solving (5.27) we obtain, in particular, the explicit expression for $\phi_i(x)$:

$$\phi_i(x) = \frac{-\mathrm{B}(-a_i)\ \underline{e}_z \cdot (\underline{l}_j \times \underline{l}_k)}{\det D(x)}. \tag{5.32}$$

When $x = x_i$ we have $|l_i| = 0$, $a_i = 0$, and thus $\det D(x_i) = -\mathrm{B}(0)\underline{e}_z \cdot (\underline{l}_j \times \underline{l}_k)$. Hence, from (5.32) we deduce $\phi_i(x_i) = 1$. When $x \in \overline{x_j x_k}$, we have $\underline{l}_j \times \underline{l}_k = 0$ and (5.32) gives $\phi_i(x) = 0$. Thus, (5.29) is proved. In order to show that (5.30) is verified, let $\phi = \phi_i + \phi_j + \phi_k$ and $(p_1, p_2) := \underline{p} = \underline{p}_i + \underline{p}_j + \underline{p}_k$. Since ϕ_m and $\underline{p}_m$ satisfy (5.27) for $m = i, j, k$, summing the three linear systems yields

$$D(x) \begin{pmatrix} p_1 \\ p_2 \\ \phi \end{pmatrix} = \begin{pmatrix} -\mathrm{B}(-a_i) \\ -\mathrm{B}(-a_j) \\ -\mathrm{B}(-a_k) \end{pmatrix}.$$

We now show that $\phi = 1$ and $\underline{p} = -\underline{a}$ satisfy the above linear system. Using (4.37) and substituting $\phi = 1$ into the above linear system we have, for $m = i, j, k$,

$$\underline{l}_m \cdot \underline{p} = \mathrm{B}(a_m) - \mathrm{B}(-a_m) = \mathrm{B}(a_m)\,(1 - e^{a_m}) = -\underline{l}_m \cdot \underline{a}.$$

Therefore we have $\underline{p} = -\underline{a}$, proving (5.30). □

For each triangle having x_i as a vertex we have defined a local function ϕ_i and an auxiliary flux $\underline{p}_i$ associated with ϕ_i as above. Combining all the local functions associated with x_i we obtain a hat function ϕ_i defined on the union of all the triangles sharing x_i, denoted by Ω_i. From Theorem 5.4 we see that this ϕ_i is unity at x_i and 0 on $\partial\Omega_i$. This hat function ϕ_i can then be extended to Ω by defining $\phi_i(x) = 0$ for all $x \in \Omega \setminus \Omega_i$. If we can show that ϕ_i is continuous across inter-element boundaries in Ω_i, then $\phi_i \in C^0(\overline{\Omega}) \cap H^1_D(\Omega)$. On the edge $\overline{x_i x_j}$ we have $\underline{l}_i \times \underline{l}_j = 0$ and $\underline{l}_i = -|l_i|\underline{e}_j$. Hence, for $x \in \overline{x_i x_j}$ (5.31) becomes

$$\det D(x) = -(\mathrm{B}(a_i)|l_j| + \mathrm{B}(a_j)|l_i|)\,\underline{e}_z \cdot (\underline{e}_j \times \underline{l}_k), \tag{5.33}$$

which, substituted in (5.32), gives

$$\phi_i(x) = \frac{\mathrm{B}(-a_i)\ |l_j|}{\mathrm{B}(a_i)|l_j| + \mathrm{B}(a_j)|l_i|}. \tag{5.34}$$

Hence the function $\phi_i(x)$ on the edge $\overline{x_i x_j}$ depends only on the edge and not on the triangle. We comment that the continuity of the basis function ϕ_i does not depend on the continuity of $\underline{a}$, but depends on the continuity of its tangent component along each edge having x_i as a vertex. Although $\underline{a} = \underline{\nabla}\,\psi$ is not continuous across element edges,

its tangent component along each edge is continuous, since ψ is assumed piecewise linear and continuous. To visualize this kind of hat functions, we divide $[0,1]\times[0,1]$ into four triangles by the two diagonals of this square and solve (5.27) on these triangles. The computed hat functions associated with the mid-point of the square corresponding to $\underline{a}=(5,1)$ and $\underline{a}=(1,10)$ are shown in Figure 8. From this we see that the hat functions are 1 at the mid-point and zero along the boundary. They are also continuous across the inter-element boundaries.

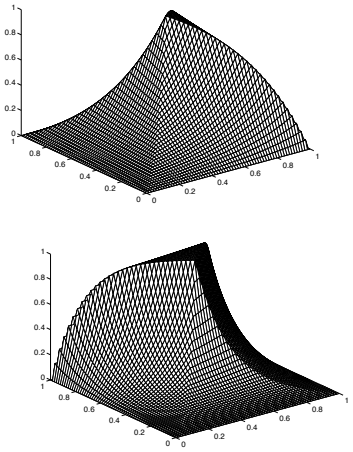


Figure 8: Hat functions for different values of $\underline{a}$.

We remark that when $\underline{a}\equiv 0$, the basis function ϕ_i reduces to the standard piecewise linear basis function. We comment that although (5.27) defines an auxiliary flux $\underline{p}_i$, in general, $\underline{p}_i \neq \underline{\sigma}_{\phi_i} := \underline{\nabla}\phi_i - \underline{a}\phi_i$ on a triangle K having x_i as one vertex. Nevertheless, it can be shown that that $\underline{p}_i = \underline{\sigma}_{\phi_i}$ at the three vertices of K ([165]).

We now set $S_h = \text{span}\,\{\phi_i\}_1^{N'}$; from the above discussion we see that $S_h \subset C^0(\overline{\Omega})\cap H_D^1(\Omega)$. Using the finite element space S_h we define the following discrete problem.

Problem 5.5 *Find $u_h \in S_h$ such that for all $v_h \in S_h$*

$$A_\psi(u_h, v_h) = (f, v_h)_\psi, \tag{5.35}$$

where $A_\psi(\cdot,\cdot)$ is the bilinear form defined by (5.25).

Problem 5.5 is a discrete problem corresponding to Problem 5.3. Since $S_h \subset H_D^1(\Omega)$, existence and uniqueness of the solution of this problem follows immediately.

We define a seminorm $|\cdot|_{1,\infty,\psi,h}$ on $(W^{1,\infty}(\Omega))^2$ by

$$|\underline{p}|_{1,\infty,\psi,h} = \left(\sum_{K\in\mathcal{T}_h}\int_K e^{-\psi}dx|\underline{p}|^2_{1,\infty,K}\right)^{1/2}.$$

The convergence of the solution to Problem 5.5 to that of Problem 5.3 is established in the following theorem.

Theorem 5.5 *Let u and u_h be the solutions to Problem 5.5 and Problem 5.3 respectively, and let $\underline{\sigma}$ and $\underline{\sigma}_{u_h}$ be the respective associated fluxes ($\underline{\sigma} = \underline{\nabla} u - \underline{\nabla}\psi\, u$ and $\underline{\sigma}_{u_h|K} = (\underline{\nabla} u_h - \underline{\nabla}\psi\, u_h)_{|K}$). Then there exists a constant $C > 0$, independent of h and u, such that*

$$\begin{aligned} ||u - u_h||_{1,\psi} &\leq Ch|\underline{\sigma}|_{1,\infty,\psi,h}, \\ ||\underline{\sigma} - \underline{\sigma}_{u_h}||_{0,\psi} &\leq Ch|\underline{\sigma}|_{1,\infty,\psi,h}. \end{aligned}$$

Proof. The proof of this theorem can be found in [165]. □

This theorem shows that the solution of Problem 5.3 and its associated flux converge to the exact ones with the convergence rate of order $\mathcal{O}(h)$. The error bounds depend only on the weighted first order seminorms of the exact flux, in contrast with the standard piecewise linear finite element method in which the error bound depends on $||u||_{2,\Omega}$. Also, the variable used in this analysis is the electron or hole concentration rather than one of the Slotboom variables. The latter is physically less interesting than the former.

5.2.3 Evaluation of terminal currents

We now consider the evaluation of the ohmic contact currents, which is often the final goal of device simulation. For simplicity, we restrict again our attention to a device with a finite number of ohmic contacts, and so Γ_D is a finite set of separated contacts. We assume that the mesh $\mathcal{T}_h$ is such that the end-points of any contact are mesh nodes of $\mathcal{T}_h$. From the definition of ohmic contacts we know that the potential drop within a contact is negligible (cf. [157], p.304). Thus ψ is constant on each ohmic contact. Let $V_h := \text{span}\,\{\phi_i\}_1^N \subset C^0(\overline{\Omega}) \cap H^1(\Omega)$. Obviously, if $v \in V_h$ and $v_{|\Gamma_D} = 0$, then $v \in S_h$. For any $\Gamma_c \subset \Gamma_D$, we choose $\phi_c \in V_h$ satisfying

$$\phi_c(x) = \begin{cases} e^{\psi_c} & x \in \Gamma_c, \\ 0 & x \in \Gamma_D \setminus \Gamma_c, \end{cases}$$

where ψ_c denotes the (constant) value of ψ on Γ_c. Taking $\gamma = 0$ in (5.22) (with $\underline{\sigma} = \underline{\nabla} u - \underline{\nabla}\psi\, u$), multiplying by $e^{-\psi}\phi_c$ and integrating by parts we have

$$-\int_{\Gamma_c} \underline{\sigma} \cdot \underline{n}\, ds + (\underline{\sigma}, \underline{\sigma}_{\phi_c})_\psi = (f, \phi_c)_\psi,$$

where $\underline{\sigma}_{\phi_c} = \underline{\nabla}\phi_c - \underline{\nabla}\psi\,\phi_c$. Thus, the outflow current through Γ_c is

$$J_c := \int_{\Gamma_c} \underline{\sigma} \cdot \underline{n}\, ds = (\underline{\sigma}, \underline{\sigma}_{\phi_c})_\psi - (f, \phi_c)_\psi.$$

Replacing $\underline{\sigma}$ by the approximate flux $\underline{\sigma}_{u_h}$ we obtain the following approximate outflow current through Γ_c

$$J_{c,h} := (\underline{\sigma}_{u_h}, \underline{\sigma}_{\phi_c})_\psi - (f, \phi_c)_\psi.$$

The convergence and the conservation of the computed ohmic contact currents are established in the following theorem.

Theorem 5.6 *Let J_c and $J_{c,h}$ be respectively the exact and the computed outflow currents through $\Gamma_c \subset \Gamma_D$ defined above. Then, there exists a constant $C > 0$, independent of h and u, such that*

$$|J_c - J_{c,h}| \leq Ch \; |\underline{\sigma}|_{1,\infty,\psi,h} \; ||\phi_c||_{1,\psi}.$$

Furthermore,

$$\sum_{\Gamma_c \subset \Gamma_D} J_{c,h} = -\int_\Omega f dx.$$

Proof. For the proof we refer to [165]. □

5.3 Stabilized finite element methods

Solutions to the boundary value problem (2.14) normally show sharp interior layers so that the application of a conventional finite element method to the problem often yields numerical solutions with non-physical spurious oscillations. This is because classical methods are numerically not stable, or the discretized diffusion term is too small so that the corresponding system matrix is almost singular. To overcome this difficulty, stabilized finite elements have been developed. The simplest case is to add to the system an artificial diffusion term with a coefficient of the order $\mathcal{O}(h)$. However, although this technique stabilizes the discretization, it introduces much artificial/computational diffusion, especially in the direction perpendicular to the characteristic or streamline direction (called cross-wind direction), and thus yields a substantial cross-wind dissipation so that a sharp interior layer is smeared out. Furthermore, the resulting scheme is of first order accuracy at most. A better approach is to add a diffusion term with a positive coefficient δ to the streamline direction only. This forms the base of the so called streamline diffusion method. This approach, of course, introduces less cross-wind diffusion. The artificial streamline diffusion term is rather arbitrary and can be conforming or non-conforming. It can also be used in a Galerkin or Petrov-Galerkin formulation. Several methods have been developed for semiconductor device equations. For example, [17], [152], [153] and [112]. In particular, the streamline upwind Petrov-Galerkin (SUPG) method used in [152] and originally proposed in [79] and [32] has been very successful for solving fluid flow problems with layers. We now give a brief

account of the method for problem (5.22). For discussion brevity we assume $\gamma = 0$, and, as before, $g = 0$.

Multiplying equation (5.22) by $v \in H^1_D(\Omega)$, integrating by parts, and using $\underline{\sigma} = \underline{\nabla} u - \underline{\nabla}\psi u$ gives the usual variational formulation: find $u \in H^1_D(\Omega)$ such that

$$\sum_{K\in\mathcal{T}_h} \int_K (\underline{\nabla} u - \underline{\nabla}\psi u)\cdot \underline{\nabla} v\, dx = \sum_{K\in\mathcal{T}_h} \int_K f\, v\, dx \qquad \forall\, v \in H^1_D(\Omega). \tag{5.36}$$

We then add a stabilyzing term, which does not alter consistency, and write

$$\sum_{K\in\mathcal{T}_h} \int_K (\underline{\nabla} u - \underline{\nabla}\psi u)\cdot \underline{\nabla} v\, dx + \sum_{K\in\mathcal{T}_h} \delta \int_K (f + \mathrm{div}\,(\underline{\nabla} u - \underline{\nabla}\psi u))\underline{\nabla}\psi\cdot\underline{\nabla} v dx = \sum_K \int_K f\, v\, dx,$$

for all $v \in H^1_D(\Omega)$, where δ is a positive stabilyzing parameter to be chosen. At the continuous level we are not modifying the formulation, as we are adding a term containing the residual on the exact solution, which is zero. Let $V_h \subset H^1_D(\Omega)$ be the finite element space with the conventional piecewise linear basis functions constructed on $\mathcal{T}_h$. The approximate problem reads find $u_h \in V_h$ such that

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} \int_K (\underline{\nabla} u_h - \underline{\nabla}\psi u_h)\cdot \underline{\nabla} v_h dx - \sum_{K\in\mathcal{T}_h} \delta \int_K (\underline{\nabla}\psi\cdot\underline{\nabla} u_h)(\underline{\nabla}\psi\cdot\underline{\nabla} v_h) dx \\ = \sum_{K\in\mathcal{T}_h} \int_K f(v_h - \delta\underline{\nabla}\psi\cdot\underline{\nabla} v_h) dx.\end{aligned}$$

for all $v_h \in V_h$. Indeed, since u_h and ψ are piecewise linear, $\Delta u_h = 0$, and $\mathrm{div}\,(\underline{\nabla}\psi u_h) = \underline{\nabla}\psi\underline{\nabla} u_h$ in K, $\forall\ K \in \mathcal{T}_h$.

This approach has the merit that it can reduce numerical cross-wind dissipation, and thus it is expected that it can capture sharp interior layers. However, the discretized system does not satisfy the maximum principle, or the system matrix is no longer an M-matrix. So, non-physical oscillations may occur near the layers. One way to improve the numerical stability of this formulation is to add a so-called shock-capturing term which may be non-linear. Analysis of this type methods can be found in [79], [32] and [87], just to name a few. An overview of streamline diffusion type methods for convection-diffusion equations is given in [88] in which a list of further readings can also be found.

5.4 Combinations of various numerical methods

Several combinations of finite element and other methods have been developed by various authors for the semiconductor device equations. In what follows we give a brief review of these methods.

In [45] a numerical method for one-dimensional time-dependent semiconductor device equations with zero diffusion coefficient was proposed. This method combines a

mixed finite element method for the electric field with an explicit upwind finite element method for the electron continuity equation. A rigorous analysis of the method is given. The same method is analyzed in [39] for the case that the electron concentration does not have discontinuity, and an improved error bound was obtained. The relevance of these methods is mostly theoretical, as the system under consideration is over-simplified. In [40] the author proposed a combined numerical method for a two-dimensional system containing the Poisson equation for ψ and the continuity equation for n. The method is a combination of a mixed finite element discretization for the Poisson equation and a discontinuous upwind finite element discretization for the continuity equation. A mathematical analysis was given and some numerical results were presented. Similar combinations of methods, using the finite element method only for the discretization of Poisson's equation to obtain a global representation of the electric field, have also been described for some particular types of semiconductor device simulations. For example, in [74], a combination of a mixed finite element method with an alternating direction implicit (ADI) method is presented for the modelling of semiconductor heterojunction structures.

6 Discretization schemes for Energy–Transport and Energy–Balance Models

In the physical literature, the Energy–Transport equations have been investigated numerically for several years [6, 37, 38, 155, 163], usually using Scharfetter-Gummel-type discretizations. Jerome and Shu [85] solved the equations employing ENO (essentially non-oscillatory) methods, while Marrocco et al. [109, 110] used mixed finite element discretization for the dual entropy formulation. In Section 6.1 we briefly present a mixed exponential fitting discretization for Energy–Transport models. These schemes are an extension of the schemes presented in Sections 3 and 4.1 for the Drift–Diffusion continuity equation, referring to [52, 78] for a complete discussion. In Section 6.2 we address the extension of the MFV scheme discussed in Section 4.2 to the case of the Energy-Balance transport model (1.28).

6.1 Mixed Finite Element Discretization for Energy–Transport Models

We consider here the stationary Energy–Transport equations in a scaled and dimensionless form. Moreover, in order to simplify the subsequent presentation, we drop the subscript n appearing in equations (1.24) and we redefine the current and energy densities as

$$\underline{J}^1 := \underline{J}_n\,; \qquad\qquad \underline{J}^2 := \underline{S}_n\,. \tag{6.1}$$

We use the Unit scaling of Table 1 of Section 1.3 and the lattice temperature T^{eq} as scaling factor for the electron temperature

$$n \to \overline{C}n, \quad C \to \overline{C}C, \quad T \to T^{eq}T, \quad \psi \to V_{th}\psi, \quad \mu \to \overline{\mu}\mu, \quad x \to lx,$$
$$\underline{J}^1 \to (q\overline{\mu}V_{th}\overline{C}/l)\underline{J}^1, \quad \underline{J}^2 \to (q\overline{\mu}V_{th}^2\overline{C}/l)\underline{J}^2,$$
$$L_{ij} \to ((qV_{th})^{i+j-1}\overline{\mu}\overline{C})L_{ij}, \quad W \to (q\overline{\mu}V_{th}^2\overline{C}/l^2)W.$$

System (1.24) in the stationary case takes now the form

$$\begin{cases} \lambda^2\Delta\psi &= n - C. \\ -\operatorname{div}\underline{J}^1 &= 0, \\ -\operatorname{div}\underline{J}^2 &= -\underline{\nabla}\psi\cdot\underline{J}^1 + W(n,T), \\ \underline{J}^1 &= L_{11}\left(\dfrac{\underline{\nabla}n}{n} - \dfrac{\underline{\nabla}\psi}{T}\right) + \left(\dfrac{L_{12}}{T} - \dfrac{3}{2}L_{11}\right)\dfrac{\underline{\nabla}T}{T}, \\ \underline{J}^2 &= L_{21}\left(\dfrac{\underline{\nabla}n}{n} - \dfrac{\underline{\nabla}\psi}{T}\right) + \left(\dfrac{L_{22}}{T} - \dfrac{3}{2}L_{21}\right)\dfrac{\underline{\nabla}T}{T}. \end{cases} \tag{6.2}$$

where $\lambda^2 = \epsilon V_{th}/(q\overline{C}l^2)$ denotes, as usual, the square of the scaled Debye length.
The starting point for the numerical discretization of the Energy–Transport system (6.2) is the observation that a suitable choice of variables allows to write the current and the energy densities $(6.2)_4$–$(6.2)_5$ in a Drift–Diffusion form in a very general context. More precisely, the following physical assumptions are imposed

- The energy band diagram is spherically symmetric and monotone with respect to the modulus of the wave vector $\vec{k}$.
- The electron density is given by non-degenerate Boltzmann statistics.
- The energy relaxation term is given by a Fokker-Planck approximation (see [52, Section 2.2]).

Under these assumptions, explicit expressions for the diffusion matrix $L = (L_{ij})$ and the energy relaxation term W in terms of n, T can be given. We refer to [52] for details of the computation, which goes through the derivation of a so-called *spherical harmonic expansion* (SHE) model (derived from the Boltzmann equation in the diffusion limit, under the assumption of dominant elastic scattering) and, afterwards, through a diffusion approximation, making electron-electron or electron-phonon scattering large. *Any* Energy–Transport model derived in this way allows for a Drift-Diffusion formulation of the form

$$\underline{J}^i = \underline{\nabla}g^i(n,T) - g^i(n,T)\frac{\underline{\nabla}\psi}{T}, \quad i = 1,2, \tag{6.3}$$

where g^1 and g^2 are nonlinear functions of n and T. (In fact, $g^1 = L_{11}$ and $g^2 = L_{21}$). For constant temperature, this expression reduces to the standard drift-diffusion current definition.

Moreover, the energy relaxation term can be written in the form

$$W = c^1 g^1 - c^2 g^2, \quad \text{with } c^i = c^i(g^1, g^2) \geq 0, \tag{6.4}$$

and the continuity equation $(6.2)_3$, (6.3) in the variables g^1 and g^2 reads as follows

$$-\mathrm{div}\, \underline{J}^2 + c^2 g^2 = f, \quad \underline{J}^2 = \underline{\nabla} g^2 - g^2 \frac{\underline{\nabla}\psi}{T(g^1, g^2)},$$

where f contains the Joule heating term $\underline{J}^1 \cdot \underline{\nabla}\psi$ and the term $c^1 g^1$. Finally, T can be defined in terms of g^1, g^2, whenever the diffusion matrix $L = (L_{ij})$ is positive definite (see [52], Lemma 2.1). Notice that this property has to be satisfied in order to get a well-posed mathematical problem.

The particular expression of g^1, g^2, c^1 and c^2 clearly depends on the actual choice of the energy band diagram and of the time relaxation model. For the parabolic band case and a special choice of the relaxation term (corresponding to the model studied in [37]), the diffusion matrix $L = (L_{ij})$ is given by the scaled version of (1.25) and we have

$$g^1(n, T) = n, \quad g^2(n, T) = \frac{3}{2} nT. \tag{6.5}$$

The energy relaxation term is of the form (see (1.26))

$$W = -\frac{3}{2} \frac{n(T - T^{eq})}{\tau_0} = \frac{1}{\tau_0}\Big(\frac{3}{2} g^1 - g^2\Big). \tag{6.6}$$

Here, τ_0 is the (scaled) energy relaxation time. In the general case, τ_0 depends on the temperature T (see [52]).

In the following, we shall consider the discretization of the Energy–Transport model in the (g^1, g^2, ψ) variables:

$$\begin{cases} \lambda^2 \Delta\psi = n(g^1, g^2) - C(x) & \\ -\mathrm{div}\, \underline{J}^1 = 0, & \\ -\mathrm{div}\, \underline{J}^2 + c^2(g^1, g^2) g^2 = c^1(g^1, g^2) g^1 - \underline{J}^1 \cdot \underline{\nabla}\psi, & \text{in } \Omega, \\ \underline{J}^i = \underline{\nabla} g^i - g^i \dfrac{\underline{\nabla}\psi}{T(g^1, g^2)}, \quad i = 1, 2, & \\ g^1 = g^1_D, \quad g^2 = g^2_D, \quad \psi = \psi_D & \text{on } \Gamma_D, \\ \underline{J}^1 \cdot \underline{n} = \underline{J}^2 \cdot \underline{n} = \underline{\nabla}\psi \cdot \underline{n} = 0 & \text{on } \Gamma_N, \end{cases} \tag{6.7}$$

where $g_D^i = g^i(n_D, T_D)$, $i = 1, 2$, and

$$c^1(g^1, g^2) = \frac{3}{2\tau_0}, \quad c^2(g^1, g^2) = \frac{1}{\tau_0}, \tag{6.8}$$

$$T(g^1, g^2) = \frac{2g^2}{3g^1}, \quad n(g^1, g^2) = g^1, \tag{6.9}$$

when (6.5) and (6.6) are chosen. Although in the example given here g^1 coincides with n and the relaxation time τ_0 does not depend on T, we prefer to write (6.7) in the general setting, which includes other choices of energy band diagrams and relaxation time models. In particular, when a non-parabolic band in the sense of Kane [91] is chosen, the dependence on T is non-local, but the discretization scheme presented here can be used without changes (the computation of $c^1(g^1, g^2)$, $c^2(g^1, g^2)$, $T(g^1, g^2)$, and $n(g^1, g^2)$ requires, of course, more effort than in the case (6.8) and (6.9)).

The Drift–Diffusion form (6.3) of the fluxes in the Energy–Transport system suggests that numerical schemes developed for the linear (in the charge variable) Drift–Diffusion continuity equation might be employed here. In [52, 78] extensions of the exponential fitting mixed finite element methods presented in Section 4.1 have been developed (in the one-dimensional, and the two-dimensional case, respectively). We refer also to [90] for a variant of these schemes in the case of a nonlinear (in the charge variable) Drift-Diffusion model [89]. In the following, we shall describe in detail the discretization of the energy flux continuity equation $(6.7)_3$. The discretization of equations $(6.7)_2$ is similar but simpler (since the zeroth–order term and the right-hand side of $(6.7)_2$ are zero). For the Poisson equation $(6.7)_1$, P_1–nonconforming elements are used (see [47]). Therefore, in the forthcoming, we assume ψ to be linear on each element.

We recall that the numerical scheme of Section 4.1 is based on the following ingredients: transform the problem by means of the Slotboom change of variable to a symmetric form, then discretize the symmetric form with mixed finite elements, and, finally, use a suitable discrete change of variable to return to the original density variable. Due to the non-constant electron temperature, a *global* Slotboom variable does not exist in the present case. However, we can define a *local* "Slotboom variable" ρ, assuming that the temperature $T = T(g^1, g^2)$ is a prescribed piecewise constant function, called $\overline{T}$, and defined in the global iteration process which solves the entire system (6.7). Therefore, $\underline{\nabla}\,\psi/\overline{T}$ is constant on each element $K \in \mathcal{T}_h$, and we can define (see (4.4))

$$\rho = e^{-\psi/\overline{T}} g \quad \text{in } K\,, \tag{6.10}$$

under the assumption $\overline{T} > 0$. For simplicity of notation, here and in the following the superscript ^{2} is dropped. Equations $(6.7)_4$ and $(6.7)_3$ can be rewritten on each triangle K as follows:

$$e^{-\psi/\overline{T}} \underline{J} - \underline{\nabla}\,\rho = 0, \tag{6.11}$$

$$-\mathrm{div}\,\underline{J} + \overline{c}\, e^{\psi/\overline{T}} \rho = -\underline{J}^1 \cdot \underline{\nabla}\,\psi + \overline{c^1} g^1, \tag{6.12}$$

where $\overline{c^1} = c^1(\overline{T})$, and $\overline{c} = c^2(\overline{T})$. In order to approximate the exponential functions in (6.11), (6.12), following (4.7) and (4.9), we define $\overline{\psi}$ as the piecewise constant function given in each element K by

$$e^{-\overline{\psi}/\overline{T}}{}_{|K} := \frac{1}{|K|}\int_K e^{-\psi/\overline{T}}dx, \tag{6.13}$$

and we define $\widetilde{\psi}$ as the piecewise constant function defined in each element K by

$$e^{\widetilde{\psi}/\overline{T}}{}_{|K} := |\tilde{e}|/(\int_{\tilde{e}} e^{-\psi/\overline{T}_{|K}}ds)\,, \qquad \tilde{e} = V_{min}V_{med}, \tag{6.14}$$

where the special edge $\tilde{e}$ is the edge connecting the vertices with the smallest values of ψ in K.

Since $\overline{T}$ is piecewise constant and it might take different values in two triangles having an edge e in common, $\int_e e^{-\psi/\overline{T}}ds$ is not uniquely defined on e. Therefore, formula (4.14) cannot be used in the present case. Here, for each element K, we introduce on each edge $e \subset \partial K$ the following constant function

$$\left(e^{-\psi/\overline{T}}\right)^{I_K}_{|e} := \frac{1}{|e|}\int_e e^{-\psi/\overline{T}_{|K}}\,ds\,. \tag{6.15}$$

Notice that the change of sign of ψ with respect to the analogous formulas of Section 4.1 is due to the different sign of ψ in equation $(6.7)_4$ (given for negative charges) and in equation (4.2) (given for positive charges).

Due to the presence of the zeroth-order term in $(6.7)_3$ we shall use, instead of the well known RT_0 element, the mixed finite element of *Example 5* (Section 3.5), which guarantees positivity of the solution, and, in contrast with the element of *Example 6*, gives a conforming approximation of $H(\mathrm{div};\Omega)$, i.e. the discrete current vector has continuous normal component across the interelement boundaries.

For the reader's convenience, we recall the definition of the finite dimensional spaces already introduced in (3.148), (3.51), (3.160),

$$\begin{aligned}
\widehat{\Sigma}_h &= \left\{\underline{\tau}_h \in (L^2(\Omega))^2 \mid \underline{\tau}_{h|K} \in Q(K)\,, \forall K \in \mathcal{T}_h\right\},\\
V_h &= \left\{v_h \in V \mid v_{h|K} \in P_0(K)\,, \forall\, K \in \mathcal{T}_h\right\},\\
\Lambda_{h,\xi} &= \left\{\mu_h \in L^2(\mathcal{E}_h)|\ \mu_{h|e} \in P_0(e) \forall e \in \mathcal{E}_h, \int_e (\mu_h - \xi)ds = 0, \forall e \in \mathcal{E}_h \cap \Gamma_D\right\},
\end{aligned}$$

where, as usual, $P_0(D)$ denotes the set of constant functions in the domain D. Moreover, $Q(K)$ denotes here the set of polynomial vectors with $dim(Q(K)) = 3$ defined by (3.112), (3.114)-(3.116), where the special edge $\tilde{e}$ is the edge connecting the vertices with the smallest values of ψ in K (see also (6.14)).

Specializing (4.16) to this case, the discrete formulation of $(6.7)_3$, $(6.7)_4$, becomes

$$\begin{cases} \text{Find } (\underline{J}_h, \rho_h, g_h) \in \widehat{\Sigma}_h \times V_h \times \Lambda_{h,g_D} \text{ such that} \\ \displaystyle\sum_K \left(\int_K (e^{-\overline{\psi}/\overline{T}} \underline{J}_h \cdot \underline{\tau}_h + \rho_h \mathrm{div}\, \underline{\tau}_h) dx - \int_{\partial K} \left(e^{-\psi/\overline{T}} \right)^{I_K} g_h \underline{\tau}_h \cdot \underline{n}\, ds \right) = 0\,, \\ \displaystyle -\sum_K \int_K (v_h \mathrm{div}\, \underline{J}_h\, dx + \overline{c} e^{\widetilde{\psi}/\overline{T}} \rho_h\, v_h)\, dx = \sum_K \int_K (-\underline{J}_h^1 \cdot \underline{\nabla}\, \psi + \overline{c^1} \overline{g}_h^1)\, v_h\, dx\,, \\ \displaystyle\sum_K \int_{\partial K} \mu_h \underline{J}_h \cdot \underline{n}\, ds = 0\,, \end{cases} \tag{6.16}$$

for all $\underline{\tau}_h \in \widehat{\Sigma}_h$, $v_h \in V_h$, and $\mu_h \in \Lambda_{h,0}$, respectively. The vector $\underline{J}_h^1 \in \Sigma \cap \widehat{\Sigma}_h$ is the approximation of the current density $\underline{J}^1$ (Σ being defined in (3.20)) and $\overline{g}_h^1 \in V_h$ is a piecewise constant approximation of g^1 provided by an analogous discretization of $(6.7)_2$, $(6.7)_4$. The first equation in (6.16) is a weak discrete version of (6.11), the second equation corresponds to a discrete version of (6.12), and the third equation imposes a continuity requirement of the normal component of $\underline{J}_h$ at the interelement boundaries. Notice that ρ has been introduced as a "trick" and its computation is of no interest (actually it will be eliminated by static condensation). The "density" variable g is approximated by g_h on the edges.

Performing the elimination of $\underline{J}_h$ and ρ_h by static condensation as in Section 3.7 and in Section 4.1, we obtain an algebraic system in the variable g_h of the form

$$\widetilde{\mathcal{M}}\, g_h = \mathcal{G}\,. \tag{6.17}$$

When computing the element matrix $\widetilde{\mathcal{M}}^K$, the edges of K are numbered counter–clockwise starting from e^1, chosen as the special edge $\tilde{e} = V_{min} V_{med}$, used in (6.14) and in the definition of $Q(K)$ (see (3.116)). As in (4.19) one can see that

$$\left(e^{-\psi/\overline{T}} \right)_{|e^1}^{I_K} e^{\widetilde{\psi}/\overline{T}} = 1\,. \tag{6.18}$$

The coefficients of $\widetilde{\mathcal{M}}^K$ have the form (see also (3.232) and (4.18))

$$\widetilde{m}_{ij}^K = \begin{cases} \left(e^{-\psi/\overline{T}} \right)_{|e^1}^{I_K} e^{\overline{\psi}/\overline{T}}\, \dfrac{\underline{\nu}^1 \cdot \underline{\nu}^1}{|K|} + \dfrac{\overline{c}|K|}{\beta^2 + \delta\, \overline{c}\, e^{(\widetilde{\psi} - \overline{\psi})/\overline{T}} |K|} \gamma_1^2 & \text{for } i = j = 1, \\ \left(e^{-\psi/\overline{T}} \right)_{|e^j}^{I_K} e^{\overline{\psi}/\overline{T}}\, \dfrac{\underline{\nu}^i \cdot \underline{\nu}^j}{|K|} & \text{otherwise,} \end{cases} \tag{6.19}$$

for $i, j = 1, 3$, and the coefficients of the element right-hand side $\mathcal{G}^K$ are given by (see also (3.233))

$$g_i^K = \begin{cases} \dfrac{\beta \gamma_1}{\beta^2 + \delta\, \overline{c}\, e^{(\widetilde{\psi} - \overline{\psi})/\overline{T}} |K|} \displaystyle\int_K f\, dx & \text{for } i = 1 \\ 0 & \text{otherwise}\,, \end{cases} \tag{6.20}$$

with $f = -\underline{J}_h^1 \cdot \underline{\nabla}\psi + \overline{c^1}\overline{g}_h^1$. The definition of δ, β, γ_1 is given in (3.220)-(3.222), and it is related to $Q(K)$. As for (4.17) with (4.20), the matrix $\widetilde{\mathcal{M}}$ is an $M-$matrix, if the decomposition is of weakly acute type.

For the definition of $\overline{T}$ and a description of the global iteration process we refer to [78], where a generalization of the Gummel map presented in Section 2 has been proposed, together with a Newton–type algorithm.

6.2 MFV discretization of the Energy-Balance model

In this section we shall consider the discretization of the Energy-Balance (EB) model (1.28) introduced in Section 1.4.

The fundamental constraints for the discretization are the conservation of electric field, current and energy fluxes, and the nonnegativity of the concentrations and of the temperatures of the carriers. With this aim, we propose an efficient and accurate solver for the EB equations in steady-state conditions. The Gummel decoupled algorithm introduced in Section 2 is employed to solve iteratively the full system, which consists of a linearized Poisson equation and of a linearized current continuity and energy-balance equations. The discretization of these linearized equations is based on the use of the cell-centered MFV method discussed in Section 3.8.

In the case of the linearized Poisson equation, this leads to solving a linear system whose coefficient matrix is symmetric, positive definite and diagonally dominant, while in the case of the linearized current continuity and energy-balance equations a stabilization procedure is developed by adding a suitable artificial diffusion term to the discretization of the constitutive laws $(1.28)_5$ and $(1.28)_6$. The artificial diffusion is associated with each edge of the finite element triangulation and can be written in terms of the jumps of the approximate scalar unknown across the edge and of the convective flux.

A special choice of the artificial diffusion term is considered which yields the exponentially-fitted upwinded Scharfetter-Gummel (SG) scheme [146]. This method provides an *optimal* upwinding approximation of the interelement fluxes and is employed in numerical computations. The resulting matrices in the discretization of $(1.28)_2$ and $(1.28)_3$ turn out to be diagonally dominant M-matrices with respect to the columns.

The outline of the section is as follows. The dual mixed formulation of a convection-diffusion model problem is considered in Section 6.2.1. The approximation of the model problem using RT_0 finite elements of lowest degree is then carried out in Section 6.2.2, while a detailed description of the stabilization procedure and of the stability analysis of the method are addressed in Sections 6.2.3 and 6.2.4, respectively. In this latter section, sufficient conditions for the stiffness matrix of the scheme to be a nonsingular M-matrix and a coercivity result in a discrete energy norm are provided. In Section 6.2.5 we specialize the general stabilization approach to both the DD and EB transport models, providing the expressions of the interelement fluxes which extend the SG discretization to the two-dimensional case. The section is concluded with a summary of the linear

algebraic systems that must be solved in the case of a bipolar model (comprising electrons and holes) at each step of the Gummel algorithm.

6.2.1 The Mixed Formulation

In this section we address the dual mixed formulation of the differential subproblems obtained applying the iterative Gummel method described in Section 2 to the EB system (1.28) in the stationary case. We shall focus our attention on the discretization of the electron current continuity and energy equations only, since the treatment of the linearized Poisson equation has already been dealt with in Section 3.8. Moreover, since similar equations hold for the holes, we shall supplement the current equation $(1.28)_2$ and the energy–balance equation $(1.28)_3$ with nonvanishing right-hand sides.

At each step of the Gummel loop the following problems must be solved for $\hat{n} = qn$ and $\phi_n = K_B T_n/q$:

$$-\mathrm{div}\,(D_n \underline{\nabla}\,\hat{n} - \mu_n \hat{n} \underline{\nabla}\,(\psi - \phi_n)) + \hat{p}\overline{R}\hat{n} = \overline{G} \tag{6.21}$$

and

$$\begin{aligned} &-\mathrm{div}\left(\lambda_n \underline{\nabla}\,\phi_n + \underline{J}_n \frac{5}{2}\phi_n\right) + \frac{3}{2}\left(\hat{p}\hat{n}\overline{R} + \frac{\hat{n}}{\tau_{w_n}}\right)\phi_n \\ &= \underline{E}\cdot\underline{J}_n + \frac{3}{2}\left(\overline{G}\chi_n + \frac{\hat{n}}{\tau_{w_n}}V_{th}\right), \end{aligned} \tag{6.22}$$

where $\hat{p} = qp$ and $\lambda_n = \frac{q}{K_B}\kappa_n$ is the modified thermal conductivity. In equation (6.21) $\hat{n}$ is the unknown whereas ψ, ϕ_n, $\hat{p}$, $\overline{R}$ and $\overline{G}$ are given functions corresponding to the previous Gummel iteration, namely

$$\begin{aligned} \overline{R} &= \frac{1}{\tau_n^*(\hat{p} + \hat{n}_i) + \tau_p^*(\hat{n} + \hat{n}_i)} + \frac{1}{q^2}(C_n\hat{n} + C_p\hat{p}) \\ \overline{G} &= \hat{n}_i^2\overline{R} + \alpha_n(|\underline{E}|)|\underline{J}_n| + \alpha_p(|\underline{E}|)|\underline{J}_p|, \end{aligned}$$

where $\hat{n}_i = qn_i$. Notice that $\overline{R}$ and $\overline{G}$ are strictly positive functions since they are the recombination and the generation terms, respectively, constructed according to the procedure discussed in Section 2.4, provided that the n, p variables are employed.

In equation (6.22) the only unknown is ϕ_n, while χ_n denotes the function ϕ_n at the previous Gummel step.

Equations (6.21) and (6.22) can be cast in the form of a stationary convection-diffusion-reaction problem

$$\begin{cases} -\mathrm{div}\underline{\sigma} + \gamma u = f \\ \underline{\sigma} = a\underline{\nabla}\,u - \underline{\beta}u \end{cases} \tag{6.23}$$

where $\underline{\sigma}$ and u are the vector and scalar unknowns, respectively. System (6.23) is supplemented with suitable boundary conditions, depending on the problem we are dealing with, of the type

$$u = g \quad \text{on } \Gamma_D, \qquad \underline{\sigma} \cdot \underline{n} = 0 \quad \text{on } \Gamma_N. \tag{6.24}$$

It is clear that the model problem (2.14) extensively discussed in previous sections can be recovered from (6.23) by simply setting $\underline{\beta} = \underline{0}$. Comparing (6.23) with (6.21) we see that $u = \hat{n}$, $a = D_n$, $\underline{\beta} = \mu_n \underline{\nabla}(\psi - \phi_n)$, $\gamma = \hat{p}\overline{R}$, $f = \overline{G}$, and the boundary condition is $g = qn_D$ (see (1.29)). Comparing (6.23) with (6.22) yields $u = \phi_n$, $a = \lambda_n$, $\underline{\beta} = -(5/2)\underline{J}_n$, $\gamma = (3/2)(\hat{p}\hat{n}\overline{R} + \hat{n}/\tau_{w_n})$, $f = \underline{E} \cdot \underline{J}_n + (3/2)(\overline{G}\chi_n + \hat{n}V_{th}/\tau_{w_n})$, and the boundary condition is $g = K_B T_D/q$ (see (1.29)).

Referring to the model problem (6.23), we remark that the choice of the decoupling splitting in (6.21) and (6.22) aims at ensuring that u is positive through guaranteeing that γ and f are positive. It is worth noticing that the sign of the first term at the right-hand side in (6.22) is not a priori established, although from a physical standpoint it is expected to be nonnegative since it represents the dissipative Joule effect. In analogy with the presentation of Section 3.2, the mixed formulation of (6.23) reads

$$\begin{cases} \text{Find } (\underline{\sigma}, u) \in \Sigma \times V \text{ such that:} \\ (\alpha\underline{\sigma}, \underline{\tau}) + (\alpha u\underline{\beta}, \underline{\tau}) + (\operatorname{div}\underline{\tau}, u) = < g, \underline{\tau} \cdot \underline{n} >_{|\Gamma_D} & \forall \underline{\tau} \in \Sigma, \\ (\operatorname{div}\underline{\sigma}, v) - (\gamma u, v) = -(f, v), & \forall v \in V, \end{cases} \tag{6.25}$$

where the spaces Σ and V are defined in (3.20)-(3.21), and $\alpha := a^{-1}$. Notice that setting $\underline{\beta} = \underline{0}$ we recover the mixed formulation (3.27) that has been thoroughly analyzed in Section 3. A way to show existence and uniqueness of the solution of problem (6.25) is to check that a solution of (6.23) (in the distributional sense) is a solution of (6.25) and viceversa. Uniqueness then follows from uniqueness of the solution of (6.23), which is guaranteed if the usual coercivity conditions hold

$$\gamma + \frac{1}{2}\operatorname{div}\underline{\beta} \geq b_0 \geq 0 \qquad a.e. \text{ in } \Omega, \quad \underline{\beta} \cdot \underline{n} \leq 0 \quad \text{on } \Gamma_N. \tag{6.26}$$

A sufficient condition (though not always applicable) to obtain uniqueness for (6.25) is $\|\underline{\beta}\|_\infty^2 < 4\gamma_0 a_0/a_M$ where a_0, a_M, and γ_0 are defined in (3.1) and (3.2).

6.2.2 Mixed finite element discretization

The discrete form of (6.25) reads:

$$\begin{cases} \text{Find } (\underline{\sigma}_h, u_h) \in \Sigma_h \times V_h \text{ such that:} \\ (\alpha\underline{\sigma}_h, \underline{\tau}_h) + (\alpha u_h\underline{\beta}, \underline{\tau}_h) + (\operatorname{div}\underline{\tau}_h, u_h) = < g, \underline{\tau}_h \cdot \underline{n} >_{|\Gamma_D} & \forall \underline{\tau}_h \in \Sigma_h, \\ (\operatorname{div}\underline{\sigma}_h, v_h) - (\gamma u_h, v_h) = -(f, v_h), & \forall v_h \in V_h, \end{cases} \tag{6.27}$$

where (Σ_h, V_h) is the RT_0 mixed finite element space given in (3.107)-(3.108). For future purposes it is convenient to assume that the convective field $\underline{\beta}$ in (6.27) has continuous normal components across each edge of the triangulation. We therefore assume that $\underline{\beta}$ is itself an RT_0 finite element vector field. The algebraic form of (6.27) reads

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} \Phi_h \\ U_h \end{pmatrix} = \begin{pmatrix} G_h \\ F_h \end{pmatrix} \tag{6.28}$$

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

$$(D - CA^{-1}B)U_h = F_h - CA^{-1}G_h.$$

The matrix $M \equiv D - CA^{-1}B$ is full and, in general, neither symmetric nor positive definite, so that solving this system can be quite expensive. Moreover, it is well known that M is not an M-matrix for any value of γ, as pointed out in [27, 99] in the case of the numerical approximation of the DD current continuity equations.
The stabilization procedure developed in the forthcoming sections will allow us to circumvent the drawbacks of the RT_0 approximation, leading to a family of stable cell-centered finite volume methods that preserve the good approximation properties for $\underline{\sigma}_h$ provided by the mixed approach, though at a reduced computational cost.

6.2.3 Mixed finite volume stabilization

In this section we apply the MFV formulation considered in Section 3.8 to the convection-diffusion-reaction problem (6.23). With this aim, we shall extend the stabilization procedure for convection-diffusion problems proposed in [144] to the case of a varying diffusion coefficient a. Without loss of generality, from now on we shall assume $g \equiv 0$ in order to simplify the exposition.

Throughout, we shall adopt the same notation as in Section 3.8, with the following extension: for each $K_k \in \mathcal{T}_h$ and for any edge e^r, with $r \in E(k)$, we define $\underline{n}^r$ as the unit normal vector chosen to ensure that

$$\beta^r := \int_{e^r} \underline{\beta} \cdot \underline{n}^r \, ds \geq 0. \tag{6.29}$$

Let us consider the second equation in (6.27); taking $v_h = \chi_k$ (the characteristic function of triangle K_k) and proceeding as in Section 3.8.2, we get the discrete conservation law (see (3.237))

$$\sum_{\substack{r \in E(k) \\ e^r \notin \Gamma_N}} \Phi_k^r - \gamma_k u_k |K_k| = -f_k |K_k| \qquad \forall K_k \in \mathcal{T}_h. \tag{6.30}$$

We recall that $\Phi_k^r = 0$ on $e^r \in \Gamma_N$. To end up with a finite volume scheme we must write the flux Φ_k^r as a function of the values u_j, $j \in T(r)$, only. With this aim, we diagonalize the first two bilinear forms in $(6.27)_1$ using the quadrature formula $(3.240)_2$. In particular, in order to apply to the convective term the same diagonalization procedure used for the diffusive term, a *unique* value for u_h needs to be defined at each edge. For this purpose, for any edge e^r we define

$$u^r = \frac{u_k + u_j}{2} \quad \text{on } e^r \notin \Gamma, \qquad u^r = \frac{u_k}{2} \quad \text{on } e^r \in \Gamma_D, \tag{6.31}$$

where the index $j \in T(r) \backslash k$. Taking in the first equation of (6.27) $\underline{\tau}_h = \underline{\tau}_h^r$ as the RT_0 basis function associated with edge e^r and such that $\underline{\tau}_h^r \cdot \underline{n}^r > 0$, we get the following constitutive equation for the edge advective flux

$$\sum_{m \in T(r)} \int_{K_m} \alpha u_h \underline{\beta} \cdot \underline{\tau}_h^r \, dx \simeq \overline{\alpha}^r u^r \sum_{m \in T(r)} \int_{K_m} \underline{\beta} \cdot \underline{\tau}_h^r \, dx \simeq \overline{\alpha}^r u^r \beta^r \frac{d^r}{|e^r|}. \tag{6.32}$$

After diagonalizing the diffusive term as done in Section 3.8.2 and using (6.32) we obtain

$$\Phi_k^r = (\overline{\alpha}^r)^{-1} \frac{u_j - u_k}{d^r} |e^r| - \beta^r u^r S_k^r \qquad \forall e^r \notin \Gamma_N, \quad \Phi_k^r = 0 \quad \text{on } e^r \in \Gamma_N, \tag{6.33}$$

where we set

$$S_k^r = \underline{n}_k^r \cdot \underline{n}^r.$$

Substituting the above expression into (6.30) yields a finite volume scheme for u_h of cell-centered type. This is easily checked to become unstable in the advection-dominated case, so that a stabilization procedure is required.

In order to stabilize (6.27) or, equivalently, to introduce flux upwinding, we consider a general *artificial diffusion* function $\rho_h : \mathcal{L}_h \to \mathbb{R}$, such that ρ_h is piecewise constant over $\mathcal{L}_h$ and for every lumping region $\mathcal{L}^r \in \mathcal{L}_h$, $\rho_h^r \geq 0$ and $\lim_{h \to 0} \rho_h^r = 0$, $(\rho_h^r := \rho_{h|\mathcal{L}^r})$. Next, we introduce the stabilized dual mixed discretization

$$\begin{cases} \text{Find } \underline{\sigma}_h^* \in \Sigma_h, \ u_h^* \in V_h \text{ such that } \forall (\underline{\tau}_h, v_h) \in \Sigma_h \times V_h \\ (\alpha \underline{\sigma}_h^*, \underline{\tau}_h) + (\alpha u_h^* \underline{\beta}, \underline{\tau}_h) + (\operatorname{div} \underline{\tau}_h, u_h^*) + \displaystyle\sum_{K_k \in \mathcal{T}_h} \int_{\partial K_k} \rho_h u_h^* \underline{\tau}_h \cdot \underline{n}_k \, ds = 0 \\ (\operatorname{div} \underline{\sigma}_h^*, v_h) - (\gamma u_h^*, v_h) = -(f, v_h) \end{cases} \tag{6.34}$$

where $\underline{n}_k$ is the unit outward normal vector along ∂K_k.

Our goal is to choose ρ_h in such a way that problem (6.34) becomes stable irrespectively of the strength of the *local Péclet number*

$$\mathbb{Pe}^r = \frac{1}{2} \overline{\alpha}^r \widehat{\beta}^r d^r, \tag{6.35}$$

where $\widehat{\beta}^r := \beta^r/|e^r| \geq 0$. Proceeding analogously as in the nonstabilized case, we get the following expression for the fluxes

$$\Phi_k^r = (\overline{\alpha}^r)^{-1}(1+\rho_h^r)\frac{u_j - u_k}{d^r}|e^r| - \beta^r u^r S_k^r \qquad \forall e^r \notin \Gamma_N, \quad \Phi_k^r = 0 \quad \text{on } e^r \in \Gamma_N. \tag{6.36}$$

Substituting the above expression into (6.30) yields the following cell-centered finite volume scheme for u_h^*

$$\begin{cases} \displaystyle\sum_{\substack{r\in E(k)\\ e^r\notin\Gamma_N}} (\overline{\alpha}^r)^{-1}(1+\rho_h^r)\left(\frac{u_k^* - u_{j(r)}^*}{d^r}\right)|e^r| + \beta^r u^r S_k^r + u_k^*\gamma_k|K_k| = f_k|K_k| & \forall K_k \in \mathcal{T}_h, \\ u_{j(r)}^* = 0 & \forall e^r \in \Gamma_D, \end{cases} \tag{6.37}$$

where $j(r) = T(r)\backslash k$. The set of linear algebraic equations (6.37) can be written in matrix form as

$$W^*\mathbf{u}^* = \mathbf{f}^* \tag{6.38}$$

where the i-th component of $\mathbf{f}^*$ is $f_i|K_i|$ and the ij-th nonzero entries of the $\mathtt{N_T} \times \mathtt{N_T}$ matrix W^* are

$$W_{ij}^* = \begin{cases} \displaystyle\sum_{\substack{r\in E(i)\\ e^r\notin\Gamma_N}} (\overline{\alpha}^r)^{-1}(1+\rho_h^r)\frac{|e^r|}{d^r} + \frac{1}{2}\beta^r S_i^r + \gamma_i|K_i|, & \text{if } i = j, \\ -(\overline{\alpha}^{r(j)})^{-1}(1+\rho_h^{r(j)})\dfrac{|e^{r(j)}|}{d^{r(j)}} + \dfrac{1}{2}\beta^r S_i^r, & \text{if } j \in T(E(i)), \quad j \neq i, \end{cases} \tag{6.39}$$

where $r(j) = E(j) \cap E(i)$ refers to the edge shared between triangles K_i and K_j.

6.2.4 Stability Analysis

In this section we prove stability of the MFV method (6.38)-(6.39). With this aim, let us write system (6.38) in the abstract form

$$\begin{cases} \text{Find } u_h \in V_h \text{ such that} \\ a_h^S(u_h, v_h) = (f, v_h), \qquad \forall v_h \in V_h, \end{cases}$$

where the discrete stabilized bilinear form $a_h^S(u_h, v_h) : V_h \times V_h \to \mathbb{R}$ is defined as

$$\begin{cases} a_h^S(u_h, v_h) &= \displaystyle\sum_{K_k\in\mathcal{T}_h} \int_{K_k} (-\operatorname{div} \underline{\sigma}_h^S(u_h) + \gamma u_h)v_h \, dx, \\ \underline{\sigma}_h^S(u_h)_{|K_k} &= \displaystyle\sum_{r\in E(k)} \Phi_k^r \underline{\tau}_k^r, \end{cases} \tag{6.40}$$

where the fluxes Φ_k^r are given in (6.36), and $\underline{\tau}_k^r = S_k^r \underline{\tau}_{h|K_k}^r$. Moreover, we denote by $\mathcal{E}_h''$ the set of all edges not belonging to Γ_N.

Theorem 6.1 *For each $\rho_h^r \geq 0$, we have*

$$a_h^S(v_h, v_h) \geq \sum_{e^r \in \mathcal{E}_h''} \kappa^r |\mathcal{L}^r| \left(\frac{v_k - v_j}{d^r} \right)^2 + b_0 ||v_h||_{0,\Omega}^2, \tag{6.41}$$

with

$$\kappa^r = 2\,(\overline{\alpha}^r)^{-1}\,(1 + \rho_h^r)\,. \tag{6.42}$$

Proof. Let v_h be a function in V_h. With a (minor) abuse of notation we denote, for every edge $e^r \in \mathcal{E}_h''$, by v_k and v_j (instead of $v_{k(r)}$, $v_{j(r)}$) the values of v_h in the triangles sharing edge e^r, with the convention that $v_{j(r)} = 0$ when $e^r \in \Gamma_D$ (as in (6.37)). Multiplying the left-hand-side of (6.37) by v_k and summing over k, we have:

$$\begin{aligned}
a_h^S(v_h, v_h) &= \sum_{K_k \in \mathcal{T}_h} \Big(\sum_{\substack{r \in E(k) \\ e^r \notin \Gamma_N}} \Big[(\overline{\alpha}^r)^{-1} (1 + \rho_h^r) \left(\frac{v_k - v_j}{d^r} \right) |e^r| + \beta^r v^r S_k^r \Big] v_k \Big) \\
&+ \sum_{K_k \in \mathcal{T}_h} \gamma_k v_k^2 |K_k| \\
&= \sum_{e^r \in \mathcal{E}_h''} (1 + \rho_h^r) \frac{(\overline{\alpha}^r)^{-1} |e^r|}{d^r} \left[(v_k - v_j) v_k + (v_j - v_k) v_j \right] \\
&+ \sum_{K_k \in \mathcal{T}_h} \sum_{\substack{r \in E(k) \\ e^r \notin \Gamma_N}} \beta^r v^r S_k^r v_k + \sum_{K_k \in \mathcal{T}_h} \gamma_k v_k^2 |K_k| \\
&= \underbrace{\sum_{e^r \in \mathcal{E}_h''} (1 + \rho_h^r) \frac{(\overline{\alpha}^r)^{-1} |e^r|}{d^r} (v_k - v_j)^2}_{I} + \underbrace{\sum_{K_k \in \mathcal{T}_h} \sum_{\substack{r \in E(k) \\ e^r \notin \Gamma_N}} \beta^r v^r S_k^r v_k}_{II} + \underbrace{\sum_{K_k \in \mathcal{T}_h} \gamma_k v_k^2 |K_k|}_{III.}
\end{aligned} \tag{6.43}$$

Recalling definition (6.31) of v^r, and observing that $\beta^r S_k^r = \int_{e^r} \underline{\beta} \cdot \underline{n}_k^r$, we deduce

$$\begin{aligned}
II &= \frac{1}{2} \sum_{K_k \in \mathcal{T}_h} v_k^2 \sum_{\substack{r \in E(k) \\ e^r \notin \Gamma_N}} \int_{e^r} \underline{\beta} \cdot \underline{n}_k^r \, ds + \frac{1}{2} \sum_{K_k \in \mathcal{T}_h} v_k \sum_{\substack{r \in E(k) \\ e^r \notin \Gamma}} v_j \int_{e^r} \underline{\beta} \cdot \underline{n}_k^r \, ds \\
&= \frac{1}{2} \sum_{K_k \in \mathcal{T}_h} v_k^2 \int_{\partial K_k} \underline{\beta} \cdot \underline{n}_k \, ds - \frac{1}{2} \sum_{K_k \in \mathcal{T}_h} v_k^2 \sum_{\substack{r \in E(k) \\ e^r \in \Gamma_N}} \int_{e^r} \underline{\beta} \cdot \underline{n}_k^r \, ds \\
&\geq \frac{1}{2} \sum_{K_k \in \mathcal{T}_h} \int_{K_k} v_k^2 \operatorname{div} \underline{\beta} \, dx,
\end{aligned} \tag{6.44}$$

where in the last step we used the fact that $\underline{\beta} \cdot \underline{n} \leq 0$ on Γ_N (see (6.26)). Then, using again (6.26) we obtain

$$II + III = \sum_{K_k \in \mathcal{T}_h} \int_{K_k} (\gamma + \frac{1}{2} \operatorname{div} \underline{\beta}) v_k^2 \, dx \geq b_0 ||v_h||_{0,\Omega}^2. \tag{6.45}$$

By multiplying and dividing by d^r each term in I, and using $d^r|e^r| = 2|\mathcal{L}^r|$ we obtain

$$I = 2 \sum_{e^r \in \mathcal{E}_h''} (1 + \rho_h^r)(\overline{\alpha}^r)^{-1} |\mathcal{L}^r| \left(\frac{v_k - v_j}{d^r} \right)^2 . \tag{6.46}$$

Finally, using (6.46), and (6.45) in (6.43) we deduce

$$a_h^S(v_h, v_h) \geq 2 \sum_{e^r \in \mathcal{E}_h''} (1 + \rho_h^r)(\overline{\alpha}^r)^{-1} |\mathcal{L}^r| \left(\frac{v_k - v_j}{d^r} \right)^2 + b_0 \|v_h\|_{0,\Omega}^2. \tag{6.47}$$

□

As a consequence of Theorem 6.1 it is easily seen that the diagonal entries of matrix (6.39) are strictly positive. Therefore, by requiring the off-diagonal entries to be non-positive, an M-matrix is obtained. The conditions on ρ_h which enforce such a property are stated in the following proposition, which extends the analogous result given in [144].

Proposition 6.1 *Let the edge artificial viscosity ρ_h^r be chosen in such a way that for each $\mathcal{L}^r \in \mathcal{L}_h$ we have*

$$\rho_h^r \geq \max\{0, \mathbb{Pe}^r - 1\} . \tag{6.48}$$

Then the stiffness matrix W^* of the stabilized dual mixed finite volume scheme turns out to be an irreducible diagonally dominant M-matrix with respect to its columns.

A consequence of Proposition 6.1 is that the elemental values U_k are nonnegative, provided $\mathbf{f}^* \geq 0$, irrespectively of the strength of the local Péclet number.
We notice that, by taking

$$\rho_h^r = \max\{0, \mathbb{Pe}^r - 1\} , \tag{6.49}$$

(6.42) becomes

$$\kappa^r = \max\left\{ 2(\overline{\alpha}^r)^{-1}, \widehat{\beta}^r d^r \right\} . \tag{6.50}$$

On the other hand, by taking

$$\rho_h^r = \mathbb{Pe}^r = \frac{1}{2} \overline{\alpha}^r \widehat{\beta}^r d^r , \tag{6.51}$$

we have

$$\kappa^r = 2(\overline{\alpha}^r)^{-1} + \widehat{\beta}^r d^r , \tag{6.52}$$

which, inserted into (6.41), produces an estimate similar to those that are typically obtained for stabilized formulations of advection-diffusion problems (see, for instance [140]).

Remark 6.1 *We note that the choice* (6.51) *gives the same scheme as the one that would be obtained, starting from the non stabilized MFV scheme (i.e.,* (6.34) *with* $\rho_h = 0$*), using classical upwind for the convective term. Indeed, let us rewrite the non stabilized fluxes* (6.33)

$$\Phi_k^r = (\overline{\alpha}^r)^{-1} \frac{u_j - u_k}{d^r} |e^r| - \beta^r u^r S_k^r \qquad \forall e^r \notin \Gamma_N. \tag{6.53}$$

Taking, instead of (6.31)*, the upwind choice*

$$u_{uw}^r = \begin{cases} u_k & \text{if } S_k^r = 1 \\ u_j & \text{if } S_k^r = -1 \end{cases} \tag{6.54}$$

it is easy to see that

$$\beta^r u_{uw}^r S_k^r = \beta^r u^r S_k^r + \beta^r \frac{u_k - u_j}{2}. \tag{6.55}$$

Hence, taking u_{uw}^r *instead of* u^r *in* (6.53) *amounts to adding the term*

$$\beta^r \frac{u_j - u_k}{2}.$$

On the other hand, taking instead of (6.53)*:*

$$\Phi_k^r = (\overline{\alpha}^r)^{-1}(1 + \mathbb{Pe}^r) \frac{u_j - u_k}{d^r} |e^r| - \beta^r u^r S_k^r \qquad \forall e^r \notin \Gamma_N \tag{6.56}$$

corresponds to adding to (6.53) *the term*

$$(\overline{\alpha}^r)^{-1} \mathbb{Pe}^r \frac{u_j - u_k}{d^r} |e^r| = (\overline{\alpha}^r)^{-1} \frac{1}{2} \overline{\alpha}^r \widehat{\beta}^r d^r \frac{u_j - u_k}{d^r} |e^r| = \beta^r \frac{u_j - u_k}{2}.$$

Remark 6.2 *The presence of the factor 2 multiplying* $|\mathcal{L}^r|$ *in* (6.46) *should not be surprising. Indeed, it can be checked that, taking for the sake of simplicity a uniform mesh made of equilateral triangles of edge* h*, we have, for a smooth enough function* v*,*

$$\sum_{e^r \in \mathcal{E}_h} \left(\frac{v_k - v_j}{d^r} \right)^2 2 |\mathcal{L}^r| \simeq \int_\Omega |\nabla v|^2 dx \qquad \text{for } h \to 0.$$

6.2.5 The Scharfetter-Gummel Stabilization

We provide in this section a choice of ρ_h that fulfils the stability requirement (6.48) and extends to the two-dimensional case the Scharfetter-Gummel (SG) exponentially fitted difference scheme [146] that is widely employed in contemporary semiconductor device simulation. We shall consider in the following both the DD and EB cases (in unscaled forms).

The DD case

The classical SG method, hereafter denoted by SG-MFV, can be recovered by setting in the flux expression in (6.40)

$$\rho_h^r = \mathbb{Pe}^r - 1 + \mathrm{B}(2\mathbb{Pe}^r) \qquad \forall \mathcal{L}^r \in \mathcal{L}_h, \tag{6.57}$$

where for any $z \in \mathbb{R}$, $\mathrm{B}(z)$ is the Bernoulli function defined in (4.32). It is worth noting that for high Péclet numbers the SG mixed finite volume scheme degenerates into the standard Engquist-Osher (EO) upwinding procedure, which is well-known to be first-order accurate. However, as $\mathbb{Pe}^r \to 0$, the amount of extra viscosity introduced by the SG flux approximation is $\mathcal{O}(h^2)$, whilst the EO one is $\mathcal{O}(h)$. After some algebra the flux across edge e^r reads

$$\Phi_k^r = qD_n \frac{u_j \mathrm{B}(\Delta\widehat{\psi}) - u_k \mathrm{B}(-\Delta\widehat{\psi})}{d^r} |e^r|, \tag{6.58}$$

where $\Delta z = z_j - z_k$ and $\widehat{\psi} = \psi / V_{th}$. As for the analysis of the SG-MFV method, the convergence estimate (3.271) has been proved in [115] under the assumption $\mathrm{curl}\underline{\beta} = \underline{0}$, which is actually the case here, since $\underline{\beta}$ is the gradient of a potential. We finally mention that the SG-MFV scheme gives the *exact solution* at the circumcenters when $a = const$, $\gamma = f = 0$, ψ is linear in Ω, and suitable boundary conditions are assumed for u and $\underline{\sigma}$. In this case the method passes the Constant-Current Patch-Test [144], which is a sound indication for a good behaviour of the numerical method in presence of steep layers arising in advection-dominated flows (see also [160]).

The EB case

Here we address the generalization of the SG method to the EB model. In particular we introduce two different expressions for the edge fluxes of the electron current (see also [58]).

As usual for this approach, we consider a one-dimensional reference interval $[0, L]$ (which can be mapped into the segment joining the circumcenters K_k and K_j) and set $\overline{R} = \overline{G} = 0$, so that the continuity equation reads

$$\frac{d}{dx}\left[q\mu_n\phi_n\left(\frac{dn}{dx} - \frac{n}{\phi_n}\frac{d(\psi - \phi_n)}{dx}\right)\right] = 0.$$

Supposing μ_n constant and both ψ and ϕ_n linearly varying, after some algebra we obtain

$$J_n = q\mu_n\left(\frac{\Delta\phi_n - \Delta\psi}{L}\right)\left[\frac{n_0\left(\frac{\phi_n(L)}{\phi_n(0)}\right)^{(\Delta\psi/\Delta\phi_n - 1)} - n_L}{\left(\frac{\phi_n(L)}{\phi_n(0)}\right)^{(\Delta\psi/\Delta\phi_n - 1)} - 1}\right], \tag{6.59}$$

where J_n is the one-dimensional electron current density and $n_L = n(L)$, $n_0 = n(0)$. Similarly as what is done in the heat exchanger theory, we introduce the *average logarithmic temperature*

$$T_n^{\ell m} = \frac{T_n(L) - T_n(0)}{\ln\left(\dfrac{T_n(L)}{T_n(0)}\right)} = \frac{\Delta T_n}{\ln\left(\dfrac{T_n(L)}{T_n(0)}\right)},$$

and the *average logarithmic thermal potential* $\phi_n^{\ell m} = K_b T_n^{\ell m}/q$. Using these definitions we can rewrite the expression of the current as

$$J_n = \frac{q\mu_n \phi_n^{\ell m}}{L} \left[n_L \, \mathrm{B}\left(\frac{\Delta\psi - \Delta\phi_n}{\phi_n^{\ell m}}\right) - n_0 \, \mathrm{B}\left(-\frac{\Delta\psi - \Delta\phi_n}{\phi_n^{\ell m}}\right)\right]. \tag{6.60}$$

Comparing this expression with (6.58) we note that:

- the thermal potential V_{th} has been replaced by the average logarithmic thermal potential $\phi_n^{\ell m}$;
- the argument of the Bernoulli functions is the difference between the electric and thermal potentials instead of being simply the potential drop across the interval.

This guarantees that if the electron temperature is constant then we exactly recover the DD expression.

An alternative form of the EB current can be provided by assuming the validity of the mobility model as in [141]. It turns out that the diffusion coefficient is independent of T_n since

$$D_n = \mu_n(T_n)\, \phi_n = \left(\mu_{n0}\frac{T_0}{T_n}\right) \phi_n = \mu_{n0} V_{th} = D_{n0}. \tag{6.61}$$

The electron current can thus be written as

$$J_n = qD_{n0}\left(\frac{dn}{dx} - \frac{n}{\phi_n}\frac{d(\psi - \phi_n)}{dx}\right). \tag{6.62}$$

As above, assuming that J_n is constant on $[0, L]$ and that both potentials are linearly varying we eventually get

$$J_n = \frac{qD_{n0}}{L}\phi_n^{\ell m} \left[\frac{n_L}{\phi_n(L)} \, \mathrm{B}\left(\frac{\Delta\psi - 2\Delta\phi_n}{\phi_n^{\ell m}}\right) - \frac{n_0}{\phi_n(0)} \, \mathrm{B}\left(-\frac{\Delta\psi - 2\Delta\phi_n}{\phi_n^{\ell m}}\right)\right]. \tag{6.63}$$

Comparing this expression with (6.60) we note three main differences:

- the diffusion coefficient has replaced the product between the mobility and the average logarithmic thermal potential. This in turn implies that with the same electric potential, the current is smaller when the temperature raises due to the decrease of μ_n with respect to an increase of T_n;

- in the argument of the Bernoulli functions in (6.63) the variation of the thermal potential is multiplied by a factor 2 and guarantees the correct behaviour of (6.63) in all the limiting cases;
- each of the two concentrations n_i $(i = 0, L)$ is also multiplied by the ratio $\phi_n^{\ell m}/\phi_n(i)$, that is, we have another weight which takes into account the variation of the temperature across the interval.

Also in this case if the electron temperature is constant equation (6.63) recovers the DD limit since $\phi_n^{\ell m}$ tends to V_{th} as $\Delta T_n \to 0$.

We conclude this section by casting the flux expressions corresponding to (6.60) and (6.63) within the general stabilization framework summarized in (6.40). Indeed, with regard to the former we have

$$\begin{cases} (\overline{\alpha}^r)^{-1} = q\mu_n \phi_n^{\ell m}, \\ \beta^r = \dfrac{\Delta(\psi - \phi_n)}{d^r} |e^r|, \\ \rho_h^r = \mathbb{P}\mathrm{e}^r - 1 + \mathrm{B}(2\mathbb{P}\mathrm{e}^r), \end{cases} \tag{6.64}$$

where the local Péclet number is evaluated according to (6.35). About expression (6.63) we note that it can be obtained in a straightforward manner from the following equivalent form of the current (6.62)

$$J_y = qD_{n0} \left(\frac{d(y\phi_n)}{dx} - y \frac{d(\varphi_n)}{dx} \right),$$

where $y = n/\phi_n$, $\varphi_n = \psi - \phi_n$ and applying (6.60) to the new variables. Therefore we get, with respect to y

$$\begin{cases} (\overline{\alpha}^r)^{-1} = qD_{n0} \phi_n^{\ell m}, \\ \beta^r = \dfrac{\Delta(\psi - \phi_n)}{d^r} |e^r|, \\ \rho_h^r = \mathbb{P}\mathrm{e}^r - 1 + \mathrm{B}(2\mathbb{P}\mathrm{e}^r). \end{cases} \tag{6.65}$$

Scharfetter-Gummel Approximation to the Energy Fluxes

Recalling $(1.28)_6$ and the definition of ϕ_n, the electron energy flux can be written as

$$\underline{S}_n = \lambda_n \underline{\nabla} \phi_n + \frac{5}{2} \phi_n \underline{J}_n \,, \tag{6.66}$$

where $\lambda_n = q\kappa_n/K_B$. As before, we consider a one-dimensional reference interval $[0, L]$; assuming λ_n to be constant and replacing n with a suitable average (see [58]), we get

$$S_n = \lambda_n \left(\frac{d\phi_n}{dx} + \frac{5J_n}{2\lambda_n} \phi_n \right). \tag{6.67}$$

This flux expression is the analogue of the hole current in the DD case where the electric field is replaced by $\dfrac{5J_n}{2\lambda_n}$. Assuming as usual that S_n is constant yields

$$S_n = \frac{\lambda_n}{L}\left[\phi_n(L)\,\mathrm{B}\left(-\frac{5J_n}{2\lambda_n}L\right) - \phi_n(0)\,\mathrm{B}\left(\frac{5J_n}{2\lambda_n}L\right)\right]. \tag{6.68}$$

Exactly as happens in the DD case for the hole current, if $|(5J_nL)/(2\lambda_n)| \gg 1$ then an upwinding effect is obtained otherwise the finite-difference approximation of the temperature gradient is recovered.

Discretization of the EB Equations

We summarize below all the systems of equations obtained applying the MFV method to discretize each of the EB equations in the Gummel procedure. For any $K_k \in \mathcal{T}_h$, we let as usual $r \in E(k)$, $j \in \{T(E(k))\backslash k\}$, $\gamma^r = |e^r|/d^r$, $\Delta\psi = \psi_j - \psi_k$, $\Delta\phi_\nu = \phi_{\nu j} - \phi_{\nu k}$ $(\nu = n, p)$ and define

$$\theta_n = \frac{\Delta\psi - 2\Delta\phi_n}{\phi_n^{\ell m}}, \qquad \theta_p = \frac{\Delta\psi + 2\Delta\phi_p}{\phi_p^{\ell m}}, \tag{6.69}$$

where the average logarithmic thermal potentials $\phi_\nu^{\ell m}$ are evaluated between K_k and K_j. The discrete momentum and energy conservation equations read

$$\sum_{j\in\{T(E(k))\backslash k\}} D_{n0}\gamma^r\phi_n^{\ell m}\left[\frac{\hat{n}_k}{\phi_{nk}}\mathrm{B}(-\theta_n) - \frac{\hat{n}_j}{\phi_{nj}}\mathrm{B}(\theta_n)\right] + \hat{n}_k\widehat{p}_k R_k^{(1)}|K_k| = G_k^{(1)}|K_k|, \tag{6.70}$$

$$\sum_{j\in\{T(E(k))\backslash k\}} D_{p0}\gamma^r\phi_p^{\ell m}\left[\frac{\hat{p}_k}{\phi_{pk}}\mathrm{B}(\theta_p) - \frac{\hat{p}_j}{\phi_{pj}}\mathrm{B}(-\theta_p)\right] + \hat{n}_k\widehat{p}_k R_k^{(2)}|K_k| = G_k^{(2)}|K_k|, \tag{6.71}$$

$$\begin{aligned}&\sum_{j\in\{T(E(k))\backslash k\}} \lambda_n\gamma^r\left[\phi_{n_k}\mathrm{B}\left(\frac{5j_n^r}{2\lambda_n}|d^r|\right) - \phi_{n_j}\mathrm{B}\left(-\frac{5j_n^r}{2\lambda_n}|d^r|\right)\right] + \frac{3}{2}\left(R_k^{(1)}\hat{p}_k + \frac{1}{\tau_{w_n}}\right)\hat{n}_k|K_k|\phi_{n_k}\\ &= \underline{E}\cdot\underline{J}_n|K_k| + \frac{3}{2}\left(G_k^{(1)}\phi_{n_k} + \frac{n}{\tau_{w_n}}V_{th}\right)|K_k|,\end{aligned} \tag{6.72}$$

$$\begin{aligned}&\sum_{j\in\{T(E(k))\backslash k\}} \lambda_p\gamma^r\left[\phi_{p_k}\mathrm{B}\left(-\frac{5j_p^r}{2\lambda_p}|d^r|\right) - \phi_{p_j}\mathrm{B}\left(\frac{5j_p^r}{2\lambda_p}|d^r|\right)\right] + \frac{3}{2}\left(R_k^{(2)}\hat{n}_k + \frac{1}{\tau_{w_p}}\right)\hat{p}_k|K_k|\phi_{p_k}\\ &= \underline{E}\cdot\underline{J}_p|K_k| + \frac{3}{2}\left(G_k^{(2)}\phi_{p_k} + \frac{p}{\tau_{w_p}}V_{th}\right)|K_k|,\end{aligned} \tag{6.73}$$

where $R_k^{(l)}$ and $G_k^{(l)}$ $(l = 1, 2)$ are the recombination and generation terms computed using the values of ψ, $\hat{n}$, $\hat{p}$, T_n and T_p of the previous Gummel step $(l = 1)$ and the

values of $\hat{n}$ computed by solving (6.70) ($l = 2$), respectively. The electron current flowing between K_k and K_j given by (6.63) is denoted by j_n^r and the hole current is

$$j_p^r = \frac{qD_{p0}}{d^r}\phi_p^{\ell m}\left[\frac{p_k}{\phi_{p_k}}\,\mathrm{B}(\theta_p) - \frac{p_j}{\phi_{p_j}}\,\mathrm{B}(-\theta_p)\right]. \tag{6.74}$$

Each of the above systems shares the feature that the corresponding coefficient matrix is a diagonally dominant M-matrix with respect to its columns.

7 Numerical results

The first set of numerical experiments, collected in Section 7.1, is concerned with the discretization of the symmentric problem (2.14) with the methods discussed in Section 3.8. Next, Section 7.2 aims at showing the performances of the discretization schemes presented in Sections 4.1–4.2 for the linear continuity equation (4.3). Section 7.3 presents simulations of a *pn*–diode with the DD model. Finally, Section 7.4 collects results obtained with the ET and the EB models, both for one and two dimensional geometry.

7.1 Mixed Finite Volumes

We assume that, in (2.14), Ω is the unit square; five unstructured grids of triangular elements have been employed, with average mesh sizes of 2^{-k}, for $k = 3, \ldots, 7$. Suitably accurate integration formulae have been used to evaluate the $L^2(\Omega)$ and $H(\mathrm{div};\Omega)$ norms of the discretization errors $u - u_h$ and $\underline{\sigma} - \underline{\sigma}_h$. Precisely, a symmetric formula using 10 nodes on each element of the triangular grid has been employed. We emphasize that all the numerical results presented below have been obtained using the midpoint quadrature formula to approximate the right-hand side in (3.237). In all the reported figures the symbols have the following meaning: '*' refers to the scheme using the average (3.277), 'o' refers to the use of the average (3.275) and '□' refers to use of the average (3.276). We aim at checking the theoretical convergence properties of the method of Section 3.8. Therefore, we take as exact solution the smooth function

$$u(x_1, x_2) = 16x_1x_2(x_1 - 1)(x_2 - 1),$$

with $||u||_{\infty,\Omega} = 1$. The diffusion coefficient a is chosen as

$$a(x_1, x_2) = \left(\frac{\nu_a + \nu_b}{2}\right)\left(1 + \frac{\nu_b - \nu_a}{\nu_a + \nu_b}\tanh(K(x_2 - x_1))\right),$$

where $\nu_b > \nu_a > 0$ and K is a suitable constant. Notice that for $\nu_b/\nu_a \gg 1$ and $K \gg 1$, a tends to a step function with jump equal to $\nu_b - \nu_a$ across the main diagonal of the unit square.

The first set of numerical results refers to the case $\nu_a = 1$, $\nu_b = 10$ and $K = 2$. The maximum values of $|\underline{\sigma}|$ and $\mathrm{div}\underline{\sigma}$ are 36.05 and 129.6, respectively. Fig.9 shows the L^2-norm (left) and the discrete L^∞-norm (right) of the error $u - u_h$, this latter norm being evaluated as

$$||u - u_h||_{\infty,C} = \max_{K_k \in \mathcal{T}_h} |u(\underline{x}_{C_k}) - u_k|.$$

Linear convergence with respect to the mesh size h is clearly observed for all of the considered averages in the case of the L^2-norm while second-order convergence is exhibited by the method when the discrete L^∞-norm is used to measure convergence. This appears to be a (nodal) superconvergence of the mixed finite volume schemes and is to be ascribed to the special choice of the point of measurement of the convergence (i.e., the circumcenter of the triangle) which coincides with the center of the stencil of the method. Fig.10 shows the L^2-norm of the discretization error $\underline{\sigma} - \underline{\sigma}_h$ (left) and the graph norm of the error $U - U_h \equiv (u - u_h, \underline{\sigma} - \underline{\sigma}_h)$ (right) defined as

$$||U - U_h||_{\Sigma \times V} = \left(||u - u_h||^2_{0,\Omega} + ||\underline{\sigma} - \underline{\sigma}_h||^2_{H(\mathrm{div};\Omega)}\right)^{1/2}.$$

Linear convergence is clearly observed for all of the three averages.

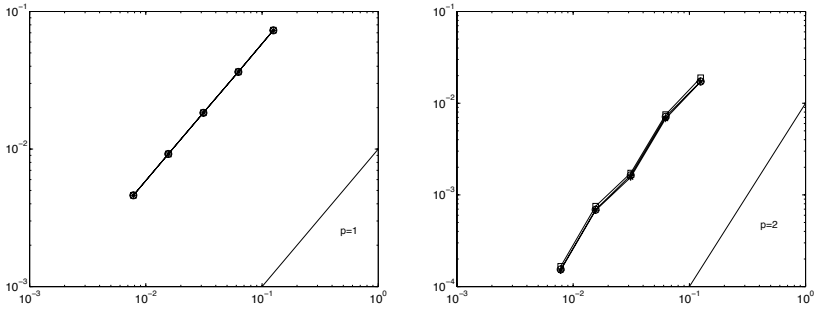


Figure 9: Discretization error $u - u_h$: L^2-norm (left) and discrete L^∞-norm (right) ($\nu_a = 1$, $\nu_b = 10$, $K = 2$).

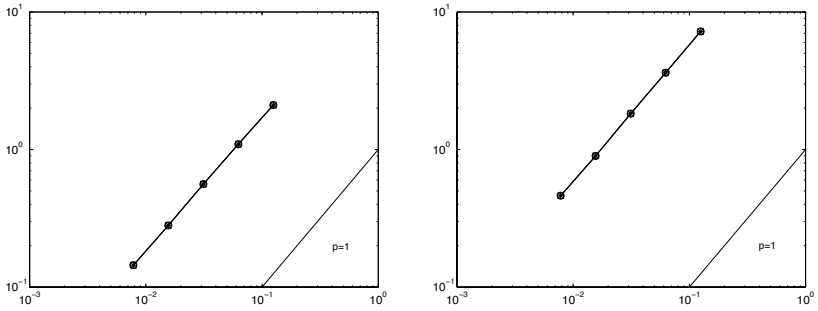


Figure 10: L^2-norm of the discretization error $\underline{\sigma} - \underline{\sigma}_h$ (left) and graph norm of the discretization error $U - U_h = (u - u_h, \underline{\sigma} - \underline{\sigma}_h)$ (right) ($\nu_a = 1$, $\nu_b = 10$, $K = 2$).

The second set of numerical results refers to the case $\nu_a = 1$, $\nu_b = 1000$ and $K = 100$. The maximum values of $|\underline{\sigma}|$ and $\mathrm{div}\underline{\sigma}$ are 4000 and 16308.8, respectively (clearly much bigger than in the previous case). Fig.11 shows that linear convergence in the L^2-norm is exhibited by all the considered averages, while the schemes appear to be second-order accurate in the discrete L^∞-norm. Linear convergence is also observed for the discretization errors $\underline{\sigma} - \underline{\sigma}_h$ and $U - U_h$, as shown in Fig. 12.

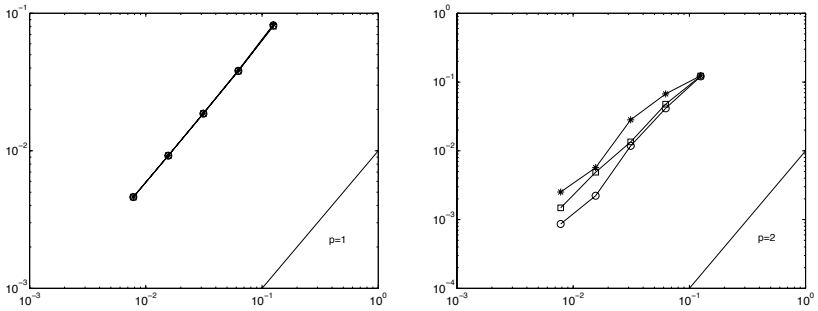


Figure 11: Discretization error $u - u_h$: L^2-norm (left) and discrete L^∞-norm (right) ($\nu_a = 1$, $\nu_b = 1000$, $K = 100$).

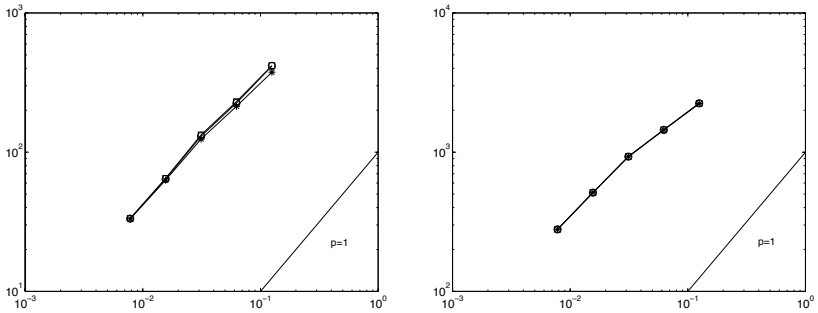


Figure 12: L^2-norm of the discretization error $\underline{\sigma} - \underline{\sigma}_h$ (left) and graph norm of the discretization error $U - U_h = (u - u_h, \underline{\sigma} - \underline{\sigma}_h)$ (right) ($\nu_a = 1$, $\nu_b = 1000$, $K = 100$).

7.2 Linear continuity equation

The examples presented here intend to compare the mixed finite element/finite volume methods described in Section 4 for the linear continuity equation (4.3), with ψ, c and f prescribed. In particular, the case $c \neq 0$ will be explored, in situations where the solution exhibits boundary and/or internal layers. The domain Ω is the unit square.
Test 1 The first example deals with a case of a boundary layer produced by a reaction term dominant with respect to diffusion and convection: ψ has a constant gradient $\underline{\nabla}\psi = -(1, 0.5)$, $f \equiv c \equiv 10^6$, and homogeneous Dirichlet boundary conditions for p

are taken on the entire border $\partial\Omega$. In Figs.13 and 14 the solution has been computed with the lowest order Raviart–Thomas elements (RT_0) and with the two elements of *Examples 5–6* of Section 3.5, respectively. Fig.15 shows the solution obtained using the MFV scheme (4.38) and its piecewise linear interpolant (we recall that the MFV scheme computes a piecewise constant density).

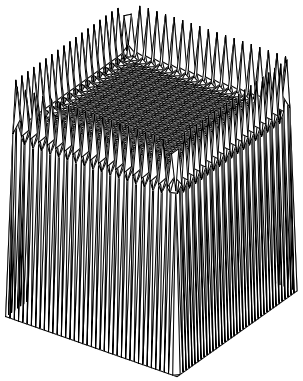


Figure 13: Test 1 with RT_0 elements.

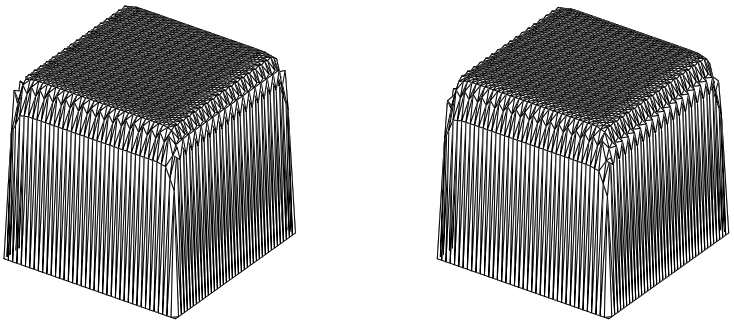


Figure 14: Test 1 with elements of *Example 5* (left) and of *Example 6* (right).

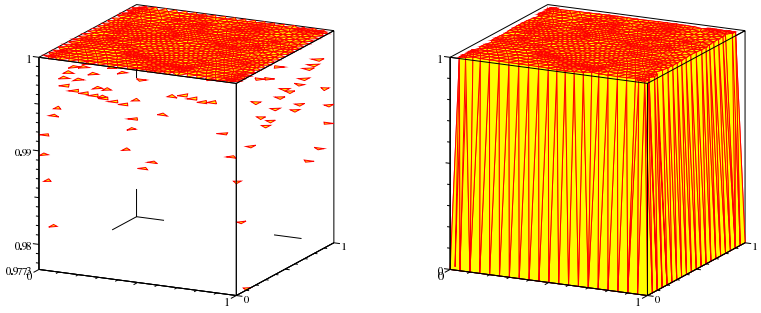


Figure 15: Test 1 with the MFV scheme (left) and after interpolation (right).

Test 2 In the second test the problem is partly advection dominated, and partly reaction dominated, and the solution exhibits internal and boundary layers. The potential ψ has a large constant gradient $\underline{\nabla}\psi = -10^4 \times (1, 0.5)$, $f \equiv 0$ and c changes abruptly: $c = 10^{10}$ in the quarter of the circle centered in $B = (1, 0)$ and with radius 0.5, $c = 0$ elsewhere. The boundary conditions for p are: $p = 0$ on $EB \cup BC$ and $p = 1$ on the rest of $\partial\Omega$ (see Fig.16 (left)). Figs.16 (right) and 17 report the results obtained by applying the RT_0 elements and the two elements of *Examples 5–6*, respectively. Fig.18 shows the solution of the MFV scheme (4.38) and its piecewise linear interpolant.

As expected, in both test problems the Raviart–Thomas element (with exact integration) provides sharp fronts, but the lack of monotonicity properties shows up in heavy overshooting and undershooting. A remedy to this is the use of a proper numerical quadrature, as done with the MFV schemes. The elements of *Examples 5-6* are both monotonic, and show little differences, though the element of *Example 5* seems to be slightly less diffusive.

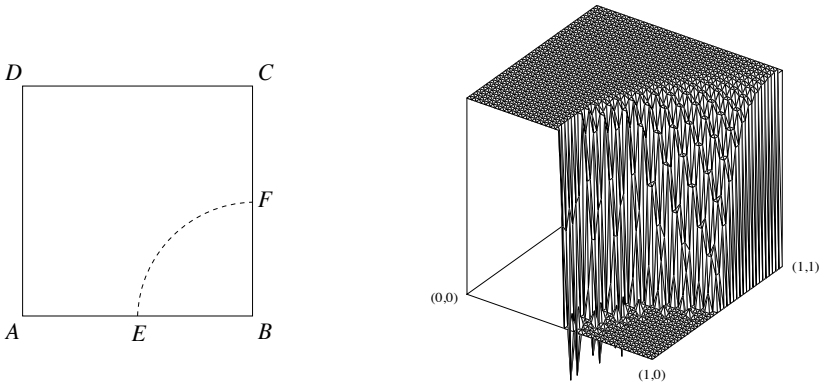


Figure 16: Test 2: computational domain (left) and RT_0-solution (right).

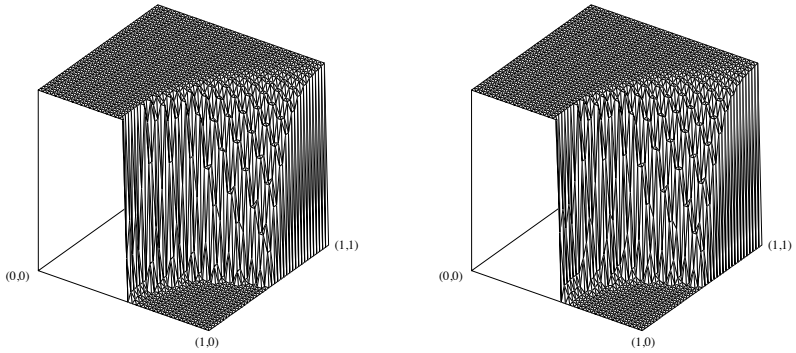


Figure 17: Test 2 with elements of *Example 5* (left) and of *Example 6* (right).

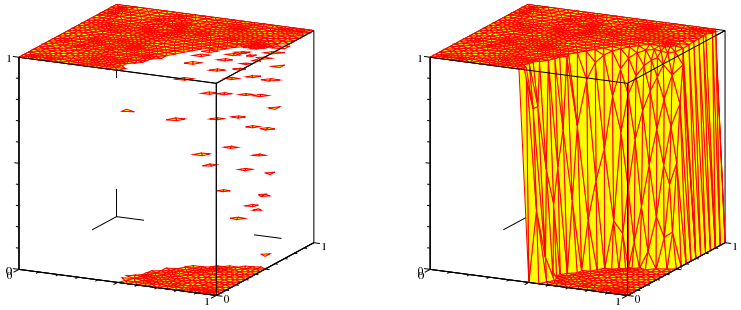


Figure 18: Test 2 with the MFV scheme (left) and after interpolation (right).

Test 3 The third example deals with the solution of the electron continuity equation (see (4.3) with ψ changed in $-\psi$, and p changed in n). The domain Ω is the unit square. The Dirichlet boundary conditions for n are:

$n = 10^3$ for $0 \le x_1 \le 0.2$, $x_2 = 0$ and $x_1 = 0$, $0 \le x_2 \le 0.2$

$n = 10^{17}$ for $x_1 = 1$, $0.8 \le x_2 \le 1$ and $0.8 \le x_1 \le 1$, $x_2 = 1$.

On the rest of the boundary homogeneous Neumann conditions are given. The reaction coefficient is $c = 0$ and the given electrostatic potential is $\psi(x_1, x_2) = \psi_0(x_1, x_2)/\ell$, with

$$\psi_0(\rho) = \begin{cases} 0 & 0 \le \rho \le 0.8, \\ 2(\rho - 0.8) & 0.8 \le \rho \le 0.9, \\ 0.2 & \rho \ge 0.9, \end{cases} \tag{7.1}$$

where $\rho = (x_1^2 + x_2^2)^{1/2}$, and $\ell = 10^{-4}$. The right hand side is $f = 0.5 \times 10^{16}$ in the annular region $0.82 \le \rho \le 0.88$, and $f = 0$ elsewhere. This choice of ψ, f, as well as

that of the Dirichlet boundary values for n, aims at reproducing the physical behavior of a curved pn–junction diode subject to a high reverse bias.

Fig.19 shows (in semilogarithmic scale) the solution computed with the MFV scheme on an unstructured mesh and its piecewise linear continuous interpolant. The computed current field $\underline{J}_h$ is displayed in Fig. 20. Fig.21 shows the result obtained with the mixed exponential fitting scheme of Section 4.1 (based on the RT_0 elements).

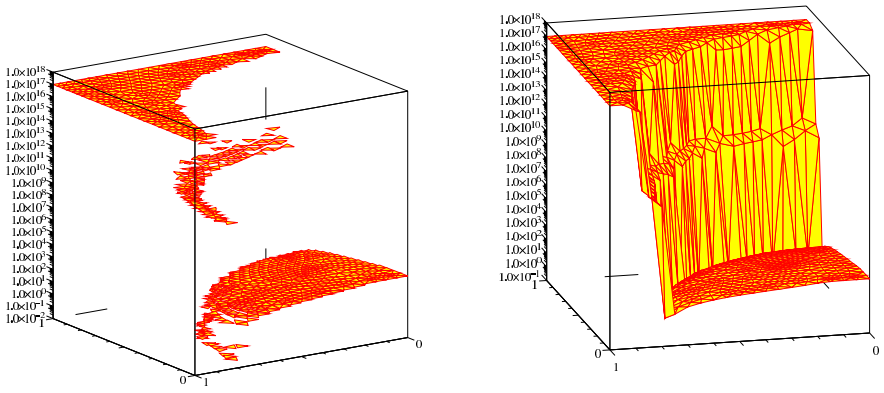


Figure 19: Electron concentration as computed with the MFV scheme (left) and after interpolation (right).

The net generation effect on the electron concentration is very well reproduced, as confirmed by the constant behavior within the anular region and by the net generation current lines collected at the upper right corner of the domain. Notice also the effectiveness of the schemes in capturing the internal layers without any spurious oscillations, as well as in tracking the current lines throughout the semiconductor, especially along the Neumann boundary where they turn out to be rigorously tangential. We point out that the undershooting about the arc $\rho = 0.8$ is an actual feature of the solution and it is not due to numerical pollution.

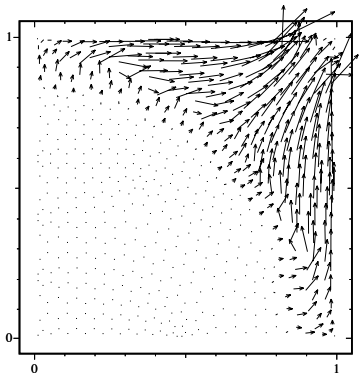


Figure 20: Electron current field computed with the MFV scheme.

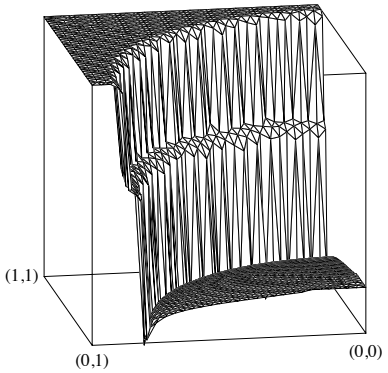


Figure 21: Electron concentration computed with RT_0 elements.

7.3 Drift–Diffusion Model: a pn–diode

We present here simulations of a *pn*–diode, neglecting recombination-generation effects. The DD system is solved in the (ψ, n, p) variables, the Poisson equation is discretized with P_1–nonconforming elements (see [47]) and the continuity equations are approximated with the mixed explonential fitting scheme based on the RT_0 elements of Section 4.1. The semiconductor material is silicon and the device is assumed to be a square of size $10^{-3}cm \times 10^{-3}cm$. The numerical values of physical parameters are given as in Table 2. The electron and hole mobilities are taken constant, with $\mu_n = \mu_p = 10^3$ $cm^2V^{-1}s^{-1}$. The geometry of the device and the triangular mesh are shown in Fig.22. The choice of a rather coarse quasi–uniform decomposition will emphasize the good performances of the numerical method which produces little numerical diffusion and does not develop any spurious oscillation.

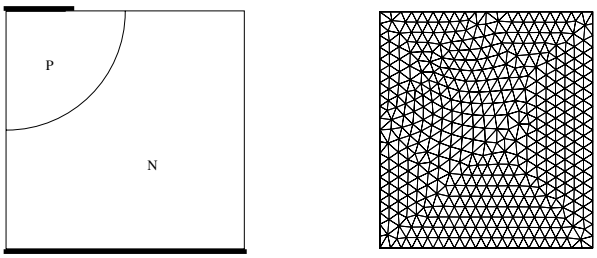


Figure 22: *pn*-junction diode: geometry (left) and mesh (right).

First, we consider a strongly reverse biased diode with an applied bias of $U = -2.0V$, defined as the difference of the applied potential at the contacts of the n-region, p-region, respectively. A moderate doping profile $C(x) = -10^{15}\ cm^{-3}$ in the p-region and $C(x) = 10^{15}\ cm^{-3}$ in the n-region (abrupt junction) is chosen. Then, the squared scaled Debye length defined in (1.23) is $\lambda^2 = 2.56 \cdot 10^{-4}$. The corresponding electron and hole densities are depicted in Fig.23. Fig.24 represents their level curves and it shows, about the junction, the so-called depletion region where both the electron and hole densities are negligible.

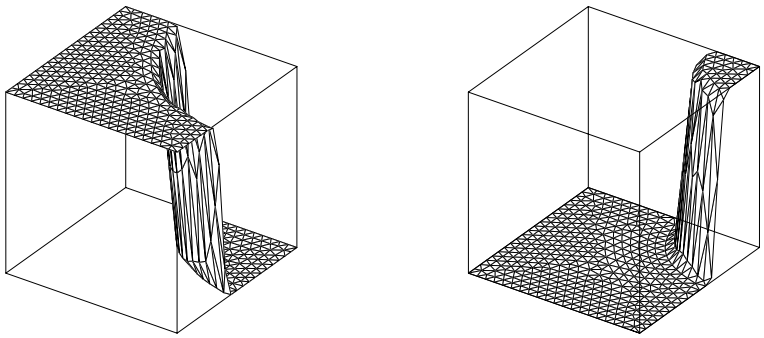


Figure 23: Reverse biased diode ($U = -2.0V, |C(x)| = 10^{15}$): electron density (left) and hole density (right).

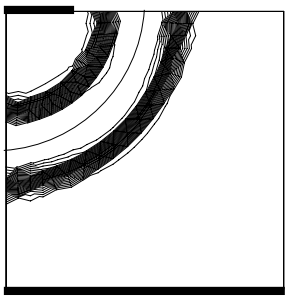


Figure 24: Level curves of the electron and hole densities of Fig.23.

In the case of a forward biased diode, the computations are carried out with a doping profile $C(x) = -10^{16}\ cm^{-3}$ in the p-region and $C(x) = 10^{16}\ cm^{-3}$ in the n-region, and with an applied voltage $U = 0.8V$. The corresponding squared scaled Debye length is $\lambda^2 = 2.56 \cdot 10^{-5}$. The carrier densities are shown in Fig. 25. The total current density flow from the anode to the cathode is reported in Fig.26 ($\underline{J}_{tot} = \underline{J}_n + \underline{J}_p$).

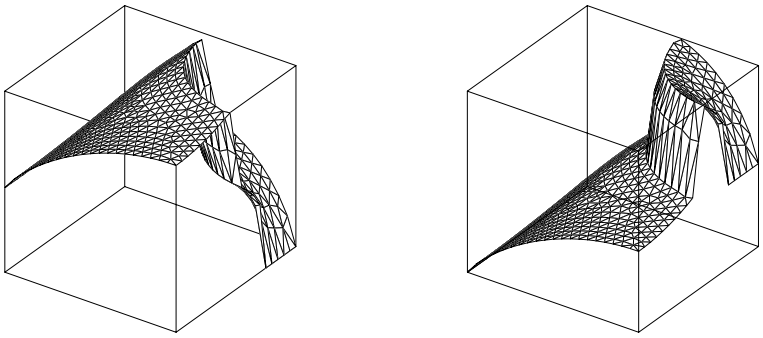


Figure 25: Forward biased diode ($U = 0.8V, |C(x)| = 10^{16}$): electron density (left) and hole density (right).

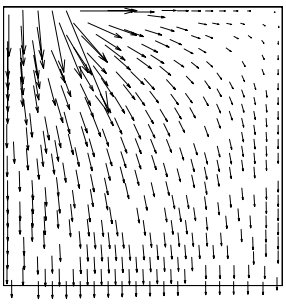


Figure 26: Forward biased diode ($U = 0.8V, |C(x)| = 10^{16}$): total current density.

7.4 Energy–Transport and Energy–Balance Models

We present first the simulation of a one-dimensional n^+nn^+ ballistic silicon diode which is a simple model for the channel of a MOS transistor. The spherically symmetric energy-band diagram ε of the semiconductor crystal, function of the modulus $k = |\vec{k}|$ of the wave vector $\vec{k}$, is given by the Kane model [91]:

$$\varepsilon(1 + \alpha\varepsilon) = \frac{\hbar^2}{2m_0}|\vec{k}|^2,$$

where $\hbar$ is the reduced Planck constant, m_0 the effective electron mass, and $\alpha > 0$ the non-parabolicity parameter. When $\alpha = 0$ the energy–band is parabolic and we take the diffusion matrix L as in (1.25). For $\alpha > 0$ the coefficients L_{ij} of the diffusion matrix depend nonlocally on the temperature T and we refer to [52] for their explicit expression. The semiconductor domain is given by the interval $\Omega = (0, \ell)$ with $\ell > 0$. In the n^+-regions the maximal doping concentration is $C_1 = 5 \cdot 10^{17}\ cm^{-3}$; in the n-channel the minimal doping profile is $C_0 = 2 \cdot 10^{15}\ cm^{-3}$. The doping profile is shown in Figure 27. The length of each n^+-region is $0.1\,\mu$m, whereas the length of the channel region equals $0.4\,\mu$m.

Parameter	Physical meaning	Numerical value
q	elementary charge	$1.6 \cdot 10^{-19}\, As$
ε_s	permittivity constant	$10^{-12}\ AsV^{-1}\ cm^{-1}$
μ_0	(low field) mobility constant	$1.5 \cdot 10^{3}\ cm^2\ V^{-1}\ s^{-1}$
V_{th}	thermal voltage at $T^{eq} = 300\, K$	$0.026\, V$
ℓ	length of the device	$0.6\, \mu$m
ℓ_0	length of the n^+ region	$0.1\, \mu$m
C_0	doping concentration in the n region	$2 \cdot 10^{15}\ cm^{-3}$
C_1	doping concentration in the n^+ region	$5 \cdot 10^{17}\ cm^{-3}$
τ_0	energy relaxation time	$0.4 \cdot 10^{-12}$ s
α	non-parabolicity parameter	$0.5\ (eV)^{-1}$

Table 2: Physical parameters.

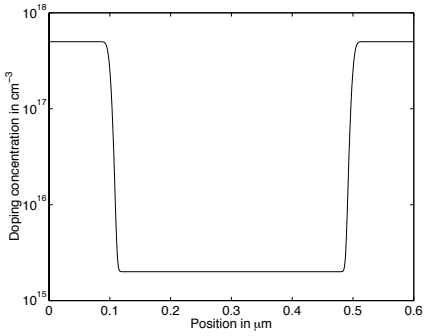


Figure 27: Doping concentration in the n^+nn^+ diode.

The numerical values of the physical parameters (for a silicon diode) are given in Table 2. On the boundary points $x = 0$ and $x = \ell$ we assume that the (unscaled) total space charge $C - n$ vanishes and that the (unscaled) temperature has the ambient temperature:

$$n(0) = n(\ell) = C_1, \qquad T(0) = T(\ell) = T^{eq}, \qquad \psi(0) = 0, \quad \psi(\ell) = V_{ext},$$

where $V_{ext} > 0$ is the applied voltage. We take the value $V_{ext} = 1.5\, V$.

We perform numerical tests on a uniform mesh of 100 nodes with the mixed exponential fitting scheme of Section 6.1. In Figure 28 we present the electron temperature for vanishing and non-vanishing non-parabolicity parameter α. As expected, the temperature in the n-channel is high, i.e. the electrons are *hot*. The maximal temperature values are $T = 2330\, K$ for $\alpha = 0$ and $T = 1610\, K$ for $\alpha = 0.5\ (eV)^{-1}$. Therefore, the temperature is reduced due to the non-parabolicity effects. In Figure 29 the electron mean velocity is shown. The mean velocity u is defined by $u = J^1/(qn)$. The spurious velocity overshoot spike at the left junction almost vanishes for $\alpha = 0.5\ (eV)^{-1}$. The

maximal velocities are $u = 1.44 \cdot 10^7\, cms^{-1}$ for $\alpha = 0$ and $u = 1.25 \cdot 10^7\, cms^{-1}$ for $\alpha = 0.5\ (eV)^{-1}$.

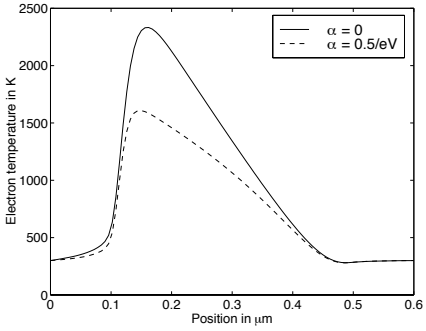


Figure 28: Electron temperature in a ballistic diode.

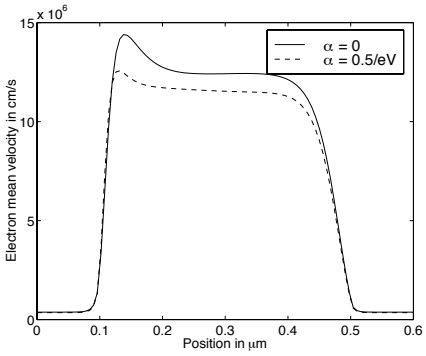


Figure 29: Electron mean velocity in a ballistic diode.

Finally, we present the simulation of two realistic semiconductor devices carried out by applying the SG-MFV method of Section 6.2 for the discretization of the EB equations for electrons and holes.
The first device is a *pn*–diode similar to that studied in [133]. The structure of the device is one-dimensional, corresponding to the interval $0 \leq x \leq \ell$ with $\ell = 1.23\, \mu m$. The doping profile C is piecewise constant with $C(x) = 10^{17}\, cm^{-3}$ for $0 \leq x < \ell/2$, and $C(x) = -10^{17}\, cm^{-3}$ for $\ell/2 < x \leq \ell$. Constant carrier mobilities have been assumed, namely $\mu_n = 800\, cm^2 V^{-1} s^{-1}$ and $\mu_p = 400\, cm^2 V^{-1} s^{-1}$. The SHR recombination/generation model has been considered with $\tau_n^* = \tau_p^* = 1 \cdot 10^{-6}\, s$ and the Wiedemann-Franz law described in Section 1.4 has been used to model the thermal conductivity with $\kappa_0 = 1/2$, while the carrier energy relaxation times have been set equal to $\tau_{w_\nu} = 0.5 \cdot 10^{-12}\, s$. The diode has been simulated under reverse-bias conditions,

with the applied voltage varying from $0V$ to $50V$. Fig. 30 shows the carrier concentrations n (electrons) and p (holes) and the corresponding temperatures T_n and T_p at an applied bias of $28V$. The comparison with the results of [133] is quite satisfactory. In particular, we point out the strong carrier heating in the middle of the device due to the presence of a high electric field.

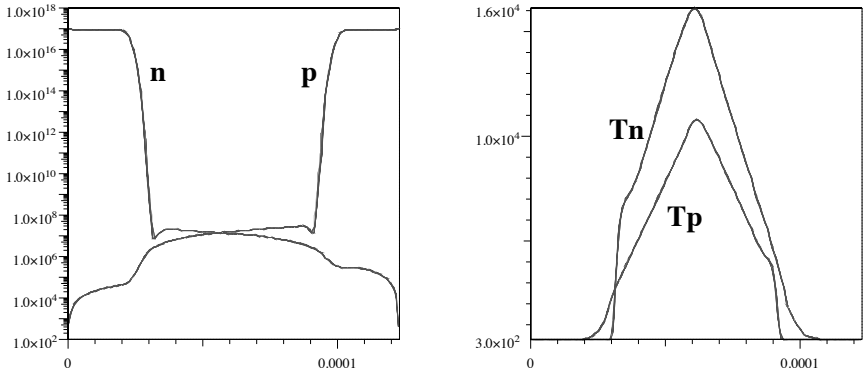


Figure 30: Hole and electron concentrations (left) and hole and electron temperatures (right).

The next example refers to a realistic 1μm channel-length nMOS transistor operating at room temperature $T^{eq} = 300\,K$. The MOS structure is composed by a silicon semiconductor wafer Ω_S coated with a thin layer of insulating oxide Ω_O, as shown in Fig.31 (left). The ohmic contacts (marked in bold) are respectively denoted by Source (S), Gate (G), Drain (D) and Back (B). The silicon-oxide interface is $\partial\Omega_I$; Ω_S is the rectangle $0 \leq x_1 \leq 2.5\ \mu m$, $0 \leq x_2 \leq 2\ \mu m$ and Ω_O is the rectangle $0.5 \leq x_1 \leq 2\ \mu m$, $2 \leq x_2 \leq 2.025\ \mu m$ (oxide thickness $t_{ox} = 25\ nm$).

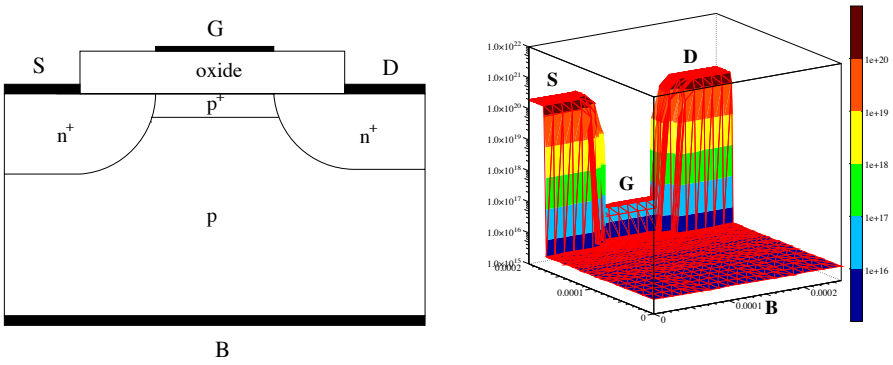


Figure 31: MOS structure (left) and doping profile (right).

The absolute value of the doping profile is plotted in logarithmic scale in Fig.31 (right), with a p-type substrate of $3 \cdot 10^{15}\ cm^{-3}$ (the *semiconductor bulk*), two n^+-type regions (source and drain) of $2 \cdot 10^{20}\ cm^{-3}$ and a p^+ implantation of $3 \cdot 10^{16}\ cm^{-3}$ in the bulk region close to the silicon-oxide interface. The following model has been used for carrier mobilities

$$\mu_\nu = \frac{\mu_{\nu,0}}{\left(1 + \left(\frac{\mu_{\nu,0} E_\nu^{||}}{v_\nu^{sat}}\right)^{\beta_\nu}\right)^{1/\beta_\nu}} \frac{1}{\sqrt{1 + \frac{E_\nu^\perp}{E_\nu^{crit}}}}, \quad \nu = n, p,$$

where $\mu_{n,0} = 800\, cm^2 V^{-1} s^{-1}$, $\mu_{p,0} = 400\, cm^2 V^{-1} s^{-1}$, $\beta_n = 2$, $\beta_p = 1$, $v_n^{sat} = 1.1 \cdot 10^7\, cms^{-1}$, $v_p^{sat} = 9.5 \cdot 10^6\, cms^{-1}$, $E_n^{crit} = 6.49 \cdot 10^4\, Vcm^{-1}$, $E_p^{crit} = 1.87 \cdot 10^4\, Vcm^{-1}$, and

$$E_\nu^{||} = \frac{|\underline{E} \cdot \underline{J}_\nu|}{|\underline{J}_\nu|}, \qquad E_\nu^\perp = \frac{|\underline{E} \times \underline{J}_\nu|}{|\underline{J}_\nu|}.$$

As for the recombination/generation mechanisms, only the SHR model has been applied using the following doping-dependent model for the carrier lifetimes

$$\tau_\nu^* = \frac{\tau_{\nu,0}}{1 + \frac{N_D^+ + N_A^-}{N_\nu^{ref}}},$$

with $\tau_{\nu,0} = 5 \cdot 10^{-5}\, s$ and $N_\nu^{ref} = 5 \cdot 10^{16}\, cm^{-3}$. The Wiedemann-Franz law described in Section 1.4 has been used to model the thermal conductivity with $\kappa_0 = 0.5$, while the carrier energy relaxation times have been set equal to $\tau_{w_\nu} = 0.5 \cdot 10^{-12}\, s$.

The device has been simulated with $5V$ applied between the gate and the source and with a drain-source voltage varying from $0V$ to $5V$. In Figs.32-33 the electron concentration, electric field and carrier temperatures are shown at a drain-source voltage of $4.8V$. The carrier heating around the drain region, where a high peak of electric field exists, is clearly visible. This, in turn, is the reason for the large flooding of electrons near the drain end of the channel, due to their increased thermal diffusivity.

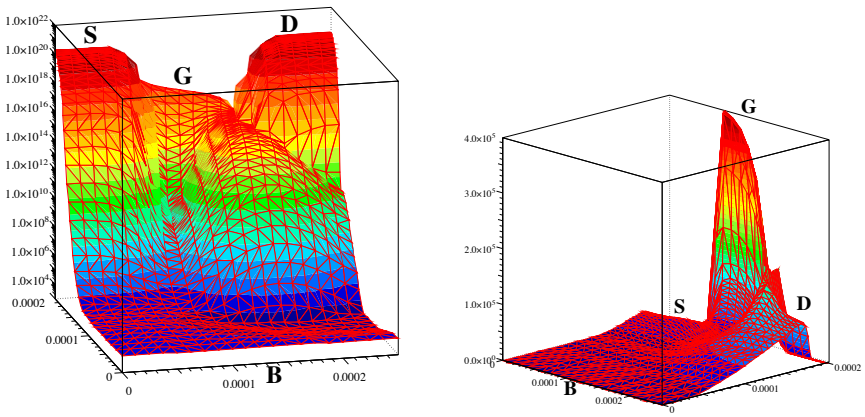


Figure 32: Electron concentration (left) and electric field (right).

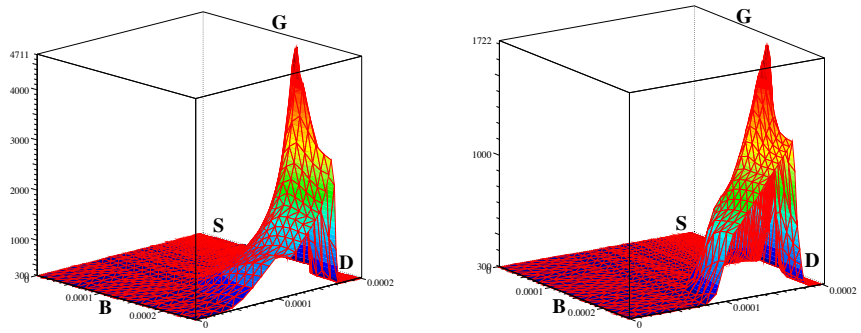


Figure 33: Electron (left) and hole (right) temperatures.

References

[1] D.A. Adams, *Sobolev Spaces*, Academic Press New York (1975).

[2] A. Agouzal, J. Baranger, J. F. Maitre, F. Oudin, Connection between finite volume and mixed finite element methods for a diffusion problem with nonconstant coefficients. Application to a convection-diffusion problem, East-West J. Numer. Math., **3** (4) 237–254 (1995).

[3] F. Alabeau, A singular perturbation analysis of the semiconductor device and the electrochemistry equations, (1984).

[4] W. Allegretto and H. Xie, Nonisothermal semiconductor systems. In X. Liu and D. Siegel, editors, *Comparison methods and stability theory*, volume 162 of *Lecture Notes in Pure and Applied Mathematics*, Marcel Dekker New York (1994).

[5] D.N. de G. Allen, R.V. Southwell, Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder, Quart. J. Mech. Appl. Math., **8**, 129–145 (1955).

[6] Y. Apanovich, P. Blakey, R. Cottle, E. Lyumkis, B. Polsky, A. Shur, and A. Tcherniaev, Numerical simulations of submicrometer devices including coupled nonlocal transport and nonisothermal effects. IEEE Trans. El. Dev. **42**, 890–897 (1995).

[7] T. Arbogast, Z. Chen, On the implementation of mixed methods as nonconforming methods for second-order elliptic problems, Math. Comp., **64** (211), 943–972 (1995).

[8] T. Arbogast, C.N. Dawson, P.T. Keenan, M. F. Wheeler, I. Yotov, Enhanced cell-centered finite differences for elliptic equations on general geometry, SIAM J. Sci. Comput., **19** (2), 404–425 (1998).

[9] 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** (2), 826–852 (1997).

[10] D.N. Arnold, F. Brezzi, Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates, RAIRO Modél. Math. Anal. Numér., **19**, 7–32 (1985).

[11] A. Azzam, E. Kreyszig, On solutions of elliptic equations satisfying mixed boundary conditions, SIAM J. Math. Anal. **13** 254–262 (1982).

[12] I. Babuška, J.E. Osborn, Generalized Finite Element Methods: Their Performance and Their Relation to Mixed Methods, SIAM J. Numer. Anal., **20** (3), 510–536 (1983).

[13] G. Baccarani, M. Rudan, R. Guerrieri, P. Ciampolini, Physical models for Numerical Device Simulation, *Process and Device Modeling*, W. H. Engl Ed., Elsevier, North-Holland, 107–158 (1986).

[14] G. Baccarani, M.R. Wordeman, An Investigation of Steady-State Velocity Overshoot Effects in Si and GaAs Devices, Solid State Electr., **28**, 407–416 (1985).

[15] R.E. Bank, D.J. Rose, Global Approximate Newton Methods, Numer. Math., **37**, 279–295 (1981).

[16] R.E. Bank, D.J. Rose, W. Fichtner, Numerical Methods for Semiconductor Device Simulation, IEEE Trans. Electr. Dev., **ED-30**, 1031–1041 (1983).

[17] R.E. Bank, J.F. Bürgler, W. Fichtner, R.K. Smith, Some upwinding techniques for finite element approximations of convection-diffusion equations, Numer. Math., **58**, 185–202 (1990).

[18] 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, **316**, série I 509–512 (1993).

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

[20] N. Ben Abdallah and P. Degond, On a hierarchy of macroscopic models for semiconductors, J. Math. Phys. **37**, 3308–3333 (1996).

[21] N. Ben Abdallah, P. Degond and S. Génieys, An energy-transport model for semiconductors derived from the Boltzmann equation, J.Stat.Phys. **84**, 205–231 (1996).

[22] K. Blotekjaer, Transport equations for electrons in two-valley semiconductors, IEEE Trans. on Electr. Dev., **ED-17**, 38–47 (1970).

[23] F. Bosisio, S. Micheletti, R. Sacco, A Discretization Scheme for an Extended Drift-Diffusion Model including Trap-Assisted Phenomena, J. Comp. Phys., **159**, 197–212 (2000).

[24] F. Brezzi, On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers, R.A.I.R.O. Anal. Numer. **8**, 129–151 (1974).

[25] F. Brezzi, A.C.S. Capelo and L. Gastaldi, A singular perturbation analysis of reversed-biased semiconductor diodes, SIAM J. Math. Anal. **20**, 372-387 (1989).

[26] F. Brezzi, J. Douglas,jr., and L.D. Marini, Two families of mixed finite elements for second order elliptic problems, Numer.Math. **47**, 217–235 (1985).

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

[28] F. Brezzi and L. Gastaldi, Mathematical properties of one-dimensional semiconductors, Matem. Apl. e Comput. **5**, 123-137 (1986).

[29] F. Brezzi, L.D. Marini, P. Pietra, Méthodes d'éléments finis mixtes et schéma de Scharfetter-Gummel. C.R.Acad.Sci. Paris, **305**, Série I, 599–604 (1987).

[30] F. Brezzi, L.D. Marini, P. Pietra, Two-dimensional exponential fitting and applications to semiconductor device equations, SIAM J. Numer. Anal., **26**, 1342–1355 (1989).

[31] F. Brezzi,L.D. Marini and P. Pietra, Numerical simulation of semiconductor devices, Comp.Meths.Appl.Mech.Engrg. **75**, 493–514 (1989).

[32] A. Brooks, T.J.R. Hughes, Streamline upwind Petrov-Galerkin formulation for convection-dominated flows with particular emphasis on incompressible Navier-Stokes equations, Comp. Meth. Appl. Mech. Eng., **32**, 199–259 (1982).

[33] E. Buturla, P. Cottrell, B.M. Grossman, K.A. Salsburg, Finite Element Analysis of Semiconductor Devices: The FIELDAY Program, IBM J. Res. Develop., **25** (4), 218–231 (1981).

[34] L. Caffarelli and A. Friedman, A singular perturbation problem for semiconductors, Boll UMI **7**, 409–421 (1987).

[35] Z. Cai., J.E. Jones, S.F. McCormick, T.F. Russell, Control-volume mixed finite element methods, UCD/CCM Report No. 89, University of Colorado, Denver (1996).

[36] D.M. Caughey, R.E. Thomas, Carrier Moblities in in Silicon Empirically related to Doping and Field, Proc. IEEE **52**, 2192–2193 (1967).

[37] D. Chen, E. Kan, U. Ravaioli, C. Shu, R. Dutton, An improved energy transport model including non-parabolic and non-Maxwellian distribution effects. IEEE Elec. Dev. Lett. **13**, 26–28 (1992).

[38] D. Chen, E. Sangiorgi, M. Pinto, E. Kan, U. Ravaioli, and R. Dutton, Analysis of spurious velocity overshoot in hydrodynamic simulations, NUPAD IV, 109–114 (1992).

[39] Z. Chen, B. Cockburn, Error estimates for a finite element method for the drift-diffusion semiconductor device equations, SIAM J. Numer. Anal., **31** 1062–1089 (1994).

[40] Z. Chen, B. Cockburn, Analysis of a finite element method for the drift-diffusion semiconductor device equations: the multidimensional case, Numer. Math., **71**, 1–28 (1995).

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

[42] Ph.G. Ciarlet, "Basic Error Estimates", in *Handbook of Numerical Analysis*, (Ph. G. Ciarlet, J. L. Lions, eds.), Vol. II, Finite Element Methods (Part I), North-Holland, Amsterdam (1991).

[43] Ph.G. Ciarlet, P.A. Raviart, Maximum principle and uniform convergence for the finite element method, Comput. Methods Appl. Mech. Engrg., **2**, 17–31 (1973).

[44] P. Clément, Approximation by finite element functions using local regularization, RAIRO Sér. Rouge Anal. Numér., **R-2**, 77–84 (1975).

[45] B. Cockburn, I. Triandaf, Error estimates for a finite element method for the drift-diffusion semiconductor device equations: The zero diffusion case. Math. Comp., **63**, 1-28 (1994).

[46] W.M. Coughran, J.W. Jerome, Modular algorithms for transient semiconductor device simulation, Part I: Analysis of the outer iteration. In: R.E. Bank (ed.) *Computational Aspects of VLSI Design with an Emphasis on Semiconductor Device Simulation. Lecture in Applied Mathematics*, vol. 25. American Mathematical Society, Providence, R.I. 107–149 (1990).

[47] M. Crouzeix and P.A. Raviart, Conforming and nonconforming finite element methods for solving the stationary Stokes equation, RAIRO **7**, 33–76 (1973).

[48] E.M. Cowell, *High-field transport in semiconductor*, Solid State Physics Vol.g, Academic Press, New York (1967).

[49] P. Degond, S. Génieys, and A. Jüngel, A system of parabolic equations in nonequilibrium thermodynamics including thermal and electrical effects, J. Math. Pures Appl. **76**, 991–1015 (1997).

[50] P. Degond, S. Génieys, and A. Jüngel, A steady-state system in nonequilibrium thermodynamics including thermal and electrical effects, Math. Meth. Appl. Sci. **21**, 1399–1413 (1998).

[51] P. Degond, F. Guyot-Delaurens, F.J. Mustieles, F. Nier, Semiconductor modelling via the Boltzmann equation . In: R.E. Bank, R. Bulirsch, K. Merten (eds.). *Mathematical Modelling and Simulation of Electrical Circuits and Semiconductor Devices*. Birkäuser, Basel, 153–167 (1990)

[52] P. Degond, A. Jüngel, and P. Pietra, Numerical discretization of energy-transport model for semiconductors with non-parabolic band structure. SIAM J. Sci. Comp. **22**, 986–1007 (2000).

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

[54] A. DeMari, An Accurate Numerical Steady State One-Dimensional Solution of the P-N Junction, Solid-State Electron., **11**, 33–58 (1968).

[55] J. Douglas,jr., J.E. Roberts, "Global estimates for mixed methods for second order elliptic equations", Math. Comp., **44** (169), 39–52 (1985).

[56] R.E. Ewing, M.M. Liu and J. Wang, Superconvergence of mixed finite element approximation over quadrilaterals, SIAM J. Numer. Anal. **36** (3), 772–787 (1999).

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

[58] A. Forghieri, R. Guerrieri, P. Ciampolini, A. Gnudi, M. Rudan, G. Baccarani, "A new discretization strategy of the semiconductor equations comprising momentum and energy balance", IEEE Trans. on CAD, **7**, 231–241 (1988).

[59] M. Fortin, An analysis of the convergence of mixed finite element methods, R.A.I.R.O. Anal. Numer. **11**, 341–354 (1977).

[60] B.X. Fraeijs de Veubeke, Displacement and equilibrium models in the finite element method, in Stress Analysis, O.C. Zienkiewicz and G. Hollister, eds., New York (1965).

[61] H. Gajewski, On existence, uniqueness and asymptotic behavior of solutions of the basic equations for carrier transport in semiconductors, ZAMM **65**, 101–108 (1985).

[62] I.M. Gamba, Asymptotic behavior at the boundary of a semiconductor device in two space dimensions. Annali Mat. Pura Apppl. **163** 43–91 (1993).

[63] I. Gasser, The initial time layer problem and the quasineutral limit in a nonlinear drift diffusion model for semiconductors, NoDEA , 237–249 (2001).

[64] I. Gasser, L. Hsiao, P.A. Markowich and S. Wang, Quasi–neutral limit of a nonlinear drift–diffusion model for semiconductors, to appear in J.Math.Anal.Appl. (2002).

[65] I. Gasser, D. Levermore, P.A. Markowich and C. Schmeiser, The initial time layer problem and the quasi–neutral limit in the semiconductor drift–diffusion model, Euro. J. Appl.Math **12**, 497–512 (2001).

[66] I. Gasser and R. Natalini, The energy-transport and the drift-diffusion equations as relaxation limits of the hydrodynamic model for semiconductors, Quart. Appl. Math. **57**, 269–282 (1999).

[67] E. Gatti, S. Micheletti, R. Sacco, A new Galerkin framework for the drift-diffusion equation in semiconductors, East-West J. Numer. Math., **6**(2) (1998).

[68] F. Golse and F. Poupaud, Limite fluide des equations de Boltzmann des semiconducteuts pour une statistique de fermi–Dirac, Asympt. Analysis **6**, 135–160 (1992).

[69] J. Griepentrog, An application of the implicit function theorem to an energy model of the semiconductor theory, Z. Angew. Math. Mech. **79**, 43–51 (1999).

[70] P. Grisvard, *Elliptic Problems in Nonsmooth Domains. Monographs and Studies in Mathematics*, Vol. 24, Pitman, London (1985).

[71] R. Guerrieri, M. Rudan, P. Ciampolini, G. Baccarani, Vectorial Convergence Acceleration Techniques in Semiconductor Device Analysis, in *Proceedings of the NASECODE IV Conference*, edited by J. J. H. Miller (Dublin Boole Press), pp. 293-298 (1985).

[72] H.K. Gummel, A Self-Consistent Iterative Scheme for One-Dimensional Steady-State Transistor Calculations, IEEE Trans. El. Dev., **ED-11**, 455–465 (1964).

[73] F. Hecht, A. Marrocco, Mixed finite element simulation of heterojunction structures including a boundary layer model for the quasi-Fermi levels, COMPEL **13**, 757–770 (1994).

[74] F. Hecht, A. Marrocco, E. Caquot, M. Filoche, Semiconductor device modelling for heterojunctions structures with mixed finite elements, COMPEL **10**, 425–438 (1991).

[75] J.P. Hennart, E. del Valle, On the relationship between nodal schemes and mixed-hybrid finite elements, Numer. Meth. Part. Diff. Eq., **9**, 411–430 (1993).

[76] J.P. Hennart, E. del Valle, Mesh-centered finite differences from nodal finite elements, INRIA, Rapport de recherche n. 2979, (1996).

[77] J. Henri and B. Louro, Singular perturbation theory applied to electrochemistry equations in the case of electroneutrlity, Nonlinear Analysis TMA **13**, 787–801 (1989).

[78] S. Holst, A. Jüngel and P. Pietra, A mixed finite element discretization of the Energy–Transport model for semiconductors, submitted.

[79] T.J.R. Hughes, A. Brooks, A theoretical framework for Petrov-Galerkin methods with discontinuous weighting functions. Application to the streamline upwind procedure, in *Finite Elements in Fluids* (eds. Gallagher et. al.), Vol.4, 47–65 (1982).

[80] J.W. Jerome, Consistency of semiconductor modeling: an existence/stability analysis for the stationary Van Roosbroeck system, Siam J.Appl.Math. **45**, 565–590 (1985).

[81] J.W. Jerome, Evolution systems in semiconductor device modeling: A cyclic uncoupled line analysis for the Gummel map. Math. Methods in the Appl. Sc. **9** 455–492 (1987).

[82] J.W. Jerome, *Analysis of Charge Transport*, Springer-Verlag Berlin Heidelberg (1996).

[83] J.W. Jerome, The approximation problem for drift-diffusion systems. SIAM Review **37**, 552–572 (1995).

[84] J.W. Jerome and T. Kerkhoven, A finite-element approximation theory for the drift-diffusion semiconductor model. SIAM J. Num. Anal. **28**, 403–422 (1991).

[85] J.W. Jerome and C.-W. Shu, Energy models for one-carrier transport in semiconductor devices. In W. Coughran, J. Colde, P. Lloyd, and J. White, editors, *Semiconductors, Part II*, volume 59 of *IMA Volumes in Mathematics and its Applications*, 185–207, Springer New York (1994).

[86] J.W. Jerome and C.-W. Shu. Energy transport systems for semiconductors: Analysis and simulation. In V. Lakshmikantham, editor, *Proccedings of the First World Congress of Nonlinear Analysts*, 3835–3846, Walter de Gruyter Berlin 1996.

[87] C. Johnson, U. Navert, J. Pitkarana, Finite element methods for linear hyperbolic problems, Comput. Meth. Appl. Mech., **45**, 285–312 (1984).

[88] C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Element Method*, Cambridge:Cambridge University Press (1987).

[89] A. Jüngel, *Quasi-Hydrodynamic Semiconductor Equations*, Birkhauser Basel-Boston-Berlin (2001).

[90] A. Jüngel and P. Pietra, A discretization scheme of a quasi-hydrodynamic semiconductor model, Math. Models Meth. Appl. Sci. **7**, 935–955 (1997).

[91] E. Kane. *J. Phys. Chem. Solids*, 1–249 (1957).

[92] R.B. Kellogg, Higher order singularities for interface problems. In: A.K. Aziz (ed.) *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*. Academic Press, New York, 589–602 (1972).

[93] T. Kerkhoven, A Proof of Convergence of Gummel's Algorithm for Realistic Device Geometries, SIAM J. Numer. Anal., **23**, 1121–1137 (1986).

[94] T. Kerkhoven, A Spectral Analysis of the Decoupling Algorithm for Semiconductor Simulation, SIAM J. Numer. Anal., **25**, 1299–1312 (1988).

[95] T. Kerkhoven, Y. Saad, On Acceleration Methods for Coupled Nonlinear Elliptic Systems, Numer. Math., **60**, 525–548 (1992).

[96] J.L. Lions and E. Magenes, *Problèmes aux limites non-homogènes et applications*, Dunod Paris (1968).

[97] E. Lyumkis, B. Polsky, A. Shur, and P. Visocky, Transient semiconductor device simulation including energy balance equation, Compel **11**, 311–325 (1992).

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

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

[100] L.D. Marini, P. Pietra, A monotonic mixed finite element on rectangles for second order elliptic problems, Mat. Aplic. Comp. **10**, 263–277 (1991).

[101] P.A. Markowich, A singular perturbation analysis for the fundamental semiconductor device equations, SIAM J.Appl.Math. **44**, 896–928 (1984).

[102] P.A. Markowich, *The Stationary Semiconductor Device Equations*, Springer-Verlag, Wien-New York (1986).

[103] P.A. Markowich, F. Poupaud and C. Schmeiser, Diffusion approximation of nonlinear electron phonon collision mechanisms, RAIRO, M^2AN **29**, 857-869 (1995).

[104] P.A. Markowich and C.A. Ringhofer, A singularly perturbed boundary value problem modelling a semiconductor device, SIAM J.Appl.Math. **44**, 231–256 (1984).

[105] P.A. Markowich, C.A. Ringhofer, C. Schmeiser, *Semiconductor equations*, Springer-Verlag, Wien New York, 1990.

[106] P.A. Markowich and C. Schmeiser, Uniform asymptotic representations of the basic semiconductor device equations, IMA J.Appl.Math. **36**, 43–57 (1986).

[107] P.A. Markowich and C. Schmeiser, The drift-diffusion limit for electron-phonon interaction in semiconductors, M^3AS **7**, 707-729 (1997).

[108] P.A. Markowich, M. Zlámal, Inverse–Average–Type Finite Element Discretisations of Selfadjoint Second–Order Elliptic Problems, Math. Comp., **51**(184), 431–449 (1988).

[109] A. Marrocco, P. Montarnal, Simulation des modèles energy-transport à l'aide des éléments finis mixtes , C.R. Acad. Sci. Paris **323**, Série I, 535–541 (1996).

[110] A. Marrocco, P. Montarnal and B. Perthame, Simulation of the energy-transport and simplified hydrodynamic models for semiconductor devices using mixed finite elements. In *Proceedings ECCOMAS 96*, John Wiley London (1996).

[111] B.J. McCartin, Discretization of the Semiconductor Device Equations, from *New Problems and New Solutions for Device and Process Modelling*, ed. J.J.H. Miller, Boole Press, Dublin (1985).

[112] S. Micheletti, Stabilized finite elements for semiconductor device simulation, Computing & Visualization in Science, **3** (4) 177–183 (2001).

[113] S. Micheletti, A. Quarteroni, R. Sacco, Current-Voltage Characteristics Simulation of Semiconductor Devices using Domain Decomposition, J. Comp. Phys., **119**, 46–61 (1995).

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

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

[116] E. Miglio, A. Quarteroni, F. Saleri, Finite element approximation of quasi-3D shallow water equations, Comp. Meth. Appl. Mech. Engrg., **174**, 355–369 (1999).

[117] J.J.H. Miller, S. Wang, A Triangular Mixed Finite Element Method for the Stationary Semiconductor Device Equations, RAIRO Modél. Math. Anal. Numér., **25** (2), 441–463 (1991).

[118] J.J.H. Miller, S. Wang, An analysis of the Scharfetter-Gummel box method for the stationary semiconductor device equations, RAIRO Modél. Math. Anal. Numér., **28** (2), 123–140 (1994).

[119] J.J.H. Miller, S. Wang, A new non-conforming Petrov-Galerkin finite-element method with triangular elements for a singularly perturbed advection-diffusion problem, IMA J. Numer. Anal., **14**, 257–276 (1994).

[120] J.J.H. Miller, S. Wang, A tetrahedral mixed finite element method for the stationary semiconductor continuity equations, SIAM J. Numer. Anal., **31** (1), 196–216 (1994).

[121] M.S. Mock, On equations describing steady-state carrier distribution in a semiconductor device. Comm. Pure Appl. Math. **25**, 781–792 (1972).

[122] M.S. Mock, *Analysis of Mathematical Models of Semiconductor Devices*, Dublin, Bool Press, (1983).

[123] M.S. Mock, Analysis of a Discretization Algorithm for Stationary Continuity Equations in Semiconductor Device Models, COMPEL, **2** (4), 117–139 (1983).

[124] J. Molenaar, Adaptive multigrid applied to a bipolar transistor problem, Appl. Numer. Math., **17**, 61–83 (1995).

[125] P. Montarnal and B. Perthame, Asymptotic analysis of the drift-diffusion equations and Hamilton-Jacobi equations, M^3AS **7**, 61–80 (1997).

[126] M.K.V. Murthy, G. Stampacchia, A variational inequality with mixed boundary conditions. Israel J. Math. **13** 188–224 (1972).

[127] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson Paris (1967).

[128] S.J. Polak, C. den Heijer, W.H.A. Schilders, P.A. Markowich, "Semiconductor device modelling from the numerical point of view", Int. Jour. Numer. Meth. Engrg., **24**, 763–838 (1987).

[129] S.J. Polak, W.H.A. Schilders, H.D. Couperus, A finite element method with current conservation, in Proc. SISDEP-88, (G. Baccarani and M. Rudan, eds.), Bologna 453–462 (1988).

[130] F. Poupaud, Diffusion approximation of the linear semiconductor equation: analysis of boundary layers, Asympt. Analysis **4**, 293–317 (1991).

[131] F. Poupaud, Runaway phenomena and fluid approximation under high fields in semiconductor kinetic theory, ZAMM **72**, 359–372 (1992).

[132] F. Poupaud and C. Schmeiser, Charge transport in semiconductors with degeneracy effects, Math. Methods Appl. Sci. **14**, 301-318 (1991).

[133] W. Quade, M. Rudan, E. Schöll, 'Hydrodynamic simulation of impact ionization effects in P-N junctions", IEEE Trans. on CAD, **10**, 1287–1294 (1991).

[134] A. Quarteroni, A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer-Verlag, Berlin, (1994).

[135] 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).

[136] C.A. Ringhofer, An asymptotic analysis of a transient P–N junction model, SIAM J.Appl.Math. **47**, 624–642 (1987).

[137] C.A. Ringhofer, A singular perturbation analysis for the transient semiconductor device equations in one space dimension, IMA J.Appl.Math. **39**, 17–32 (1987).

[138] J.E. Roberts, J.M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, (Ph.G. Ciarlet, J.L. Lions, eds.), Vol. II, Finite Element Methods (Part I), North-Holland, Amsterdam, (1991).

[139] D.L. Rode, Low–field electron transport, in *Semiconductors and Semimetals* **10** 1–52, Academic Press, New York (1995).

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

[141] M. Rudan, A. Gnudi, W. Quade, "A generalized approach to the hydrodynamic model of semiconductor equations", in *Precess and Device Modeling for Microelectronics*, G. Baccarani Ed., North-Holland Elsevier, 109–154 (1993).

[142] R. Sacco, E. Gatti, L. Gotusso, The patch-test as a validation of a new finite element for the solution of convection-diffusion equations, Comput. Methods Appl. Mech. Engrg., **124**, 113–124 (1995).

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

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

[145] R. Sacco, M. Stynes, Finite element methods for convection-diffusion problems using exponential splines on triangles, Comp. Math. Appl., **35** (3), 35–45 (1998).

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

[147] C. Schmeiser, On strongly reverse biased semiconductor diodes, SIAM J. Appl.Math. **49**, 1734-1748 (1989).

[148] C. Schmeiser, A singular perturbation analysis of reverse biased pn-junctions, SIAM J. Math. Anal. **21**, 313-326 (1990).

[149] T.I. Seidman, Steady state solutions of reaction–diffusion systems with electrostatic convection, Nonlinear Analysis **4**, 632–637 (1980).

[150] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer-Verlag, Wien-New York (1984).

[151] M. Sever, Discretization of time-dependent continuity equations, *Proceedings of the 6th International NASECODE Conference*, ed. J.J.H. Miller, Boole Press, Dublin, 71–83 (1988).

[152] M. Sharma and G. Carey, Semiconductor device simulation using adaptive refinement and flux upwinding, IEEE Trans. on CAD, **8**, 590–598 (1989).

[153] M. Sharma and G. Carey, Semiconductor device modeling using flux upwinding finite elements, COMPEL, **8** (4), 219–224 (1989).

[154] J.W. Slotboom, Computer-Aided Two-Dimensional Analysis of Bipolar Transistors, IEEE Trans. Electr. Dev., **ED-20**, 669–679 (1973).

[155] K. Souissi, F. Odeh, H. Tang, and A. Gnudi, Comparative studies of hydrodynamic and energy transport models, COMPEL **13**, 439–453 (1994).

[156] R. Stratton, Diffusion of hot and cold electrons in semiconductor barriers, Phys. Rev. **126**, 2002–2014 (1962).

[157] S.M. Sze, *The physics of semiconductor devices (2nd Ed.)*, John Wiley & Sons, New York (1981).

[158] J.M. Thomas, D. Trujillo, Finite volume methods for elliptic problems: convergence on unstructured meshes, in Numerical Methods in Mechanics (C. Conca and G. Gatica eds.), Addison Wesley Longman Ltd., 163–174 (1997).

[159] W. Van Roosbroeck, Theory of flow of electrons and holes in germanium and other semiconductors. Bell System Tech, J. **29**, 560–607 (1950).

[160] R.R.P. van Nooyen, A Petrov-Galerkin mixed finite element method with exponential fitting, Numer. Meth. Part. Diff. Eq., **11**, 501–524 (1995).

[161] R.S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, New Jersey (1962).

[162] P.S. Vassilevski, S.I. Petrova, R.D. Lazarov, Finite difference schemes on triangular cell-centered grids with local refinement, SIAM J. Sci. Stat. Comput., **13** (6), 1287–1313 (1992).

[163] P. Visocky, A method for transient semiconductor device simulation using hot-electron transport equations, In J. Miller, editor, *Proc. of the Nasecode X Conf.*, Boole Press Dublin (1994).

[164] S. Wang, A novel exponentially fitted triangular finite element method for an advection-diffusion problem with boundary layers, J. Comp. Phys., **134**, 253–260 (1997).

[165] S. Wang, A new exponentially fitted triangular finite element method for the continuity equations in the drift-diffusion model of semiconductor devices, RAIRO Modél. Math. Anal. Numér., **33** (1), 99–112 (1999).

[166] N.M. Wigley, Mixed boundary value problems in plane domains with corners. Math. Z. **115** 33–52 (1970).