# Finite-element discretization of 3D energy-transport equations for semiconductors

Dissertation
zur Erlangung des Grades
„Doktor der Naturwissenschaften”

am Fachbereich 08–Physik, Mathematik und Informatik
der Johannes Gutenberg-Universität
Mainz

Stephan Gadau,
geb. in Schönebeck (Elbe)

Mainz, 2007

Tag der mündlichen Prüfung: 22. Januar 2008

D77–Mainzer Dissertation

**Zusammenfassung**

In der vorliegenden Dissertation wird die Herleitung eines mathematischen Modells zur Beschreibung des Ladungs- und Energietransports in Halbleiterbauelementen wie Transistoren sowie die numerische Simulation dieser Prozesse thematisiert. Dabei kommen Methoden der theoretischen Physik, der Funktionalanalysis, der numerischen Mathematik und der Computerprogrammierung zum Einsatz.
Nach einer einleitenden Passage zum aktuellen Stand der Halbleitersimulationsverfahren und einem kurzen Überblick über die historische Entwicklung der mathematischen Modelle bis in die Gegenwart wird auf die Konstruktion eines Modells eingegangen, das die Grundlage aller nachfolgenden Arbeitsschritte dieser Dissertation bildet. Als Ausgangspunkt dient dabei eine fundamentale Gleichung aus der Gasdynamik. Mit Hilfe einer Reihenentwicklung wird daraus in einem mehrstufigen Prozeß das Modell spezifiziert. Diese Herleitung ist größtenteils einer Publikation entnommen und bildet nicht den Kern der Arbeit.
In der sich anschließenden Phase wird die mathematisch präzise Formulierung des Modells vorgenommen. Dabei sind u.a. funktionalanalytische Techniken nötig. Der Systemcharakter der Gleichungen stellt einen innovativen Aspekt gegenüber skalaren elliptischen Gleichungen dar, deren Theorie mittlerweile als Standard angesehen werden kann.
Im Anschluß daran ist die numerische Diskretisierung der Gleichungen Gegenstand der Betrachtungen. Eine Finite-Elemente-Methode, die speziell auf den Gleichungstyp zugeschnittenen ist, findet hierbei Verwendung. Diese Spezialisierung des Typs der Ansatzfunktionen ist notwendig, um der in der Praxis auftretenden Forderung nach physikalischer Interpretierbarkeit der Resultate Rechnung zu tragen. Mit einer Reihe von mathematischen Transformationen der so gewonnenen diskreten Gleichungen wird ein System von algebraischen Gleichungen hergeleitet, das sich zur numerischen Auswertung eignet. Vom Autor dieser Arbeit selbst entwickelte Computerprogramme kommen dabei zur numerischen Lösung der Gleichungen zum Einsatz. Diese Programme basieren teilweise auf neuen, angepaßten Iterationsalgorithmen, die im Rahmen dieser Forschungsarbeit entwickelt und getestet wurden. Ihnen ist aufgrund ihrer Bedeutung und ihrer Neuartigkeit großer Raum in der Arbeit gewidmet. Umfangreiche Analysen der Iterationsschemata und der Vergleich mit einem Standarditerationsverfahren, das mit Hilfe einer Ergänzung dem Problemkontext angepaßt wurde, sind hier als Hauptarbeitsschritte zu nennen.
Eine weitere Neuerung besteht in der numerischen Bestimmung von 3D-Resultaten unter Berücksichtigung der vorhandenen Rechnerkapazitäten. Es wurde darauf geachtet, dass die Ergebnisse bei vertretbarem Zeit- und Speicheraufwand berechnet werden können. Die Simulationsergebnisse einer Reihe von Modellen moderner Halbleiterbauelemente werden graphisch dargestellt und ausführlich kommentiert.
Den Abschluß der Arbeit bildet ein Ausblick auf Entwicklungsmöglichkeiten und Erweiterungen der vorhandenen Modelle und der Algorithmen.

**Abstract**

In this thesis a mathematical model was derived that describes the charge and energy transport in semiconductor devices like transistors. Moreover, numerical simulations of these physical processes are performed. In order to accomplish this, methods of theoretical physics, functional analysis, numerical mathematics and computer programming are applied.

After an introduction to the status quo of semiconductor device simulation methods and a brief review of historical facts up to now, the attention is shifted to the construction of a model, which serves as the basis of the subsequent derivations in the thesis. Thereby the starting point is an important equation of the theory of dilute gases. From this equation the model equations are derived and specified by means of a series expansion method. This is done in a multi-stage derivation process, which is mainly taken from a scientific paper and which does not constitute the focus of this thesis.

In the following phase we specify the mathematical setting and make precise the model assumptions. Thereby we make use of methods of functional analysis. Since the equations we deal with are coupled, we are concerned with a nonstandard problem. In contrary, the theory of scalar elliptic equations is established meanwhile.

Subsequently, we are preoccupied with the numerical discretization of the equations. A special finite-element method is used for the discretization. This special approach has to be done in order to make the numerical results appropriate for practical application. By a series of transformations from the discrete model we derive a system of algebraic equations that are eligible for numerical evaluation. Using self-made computer programs we solve the equations to get approximate solutions. These programs are based on new and specialized iteration procedures that are developed and thoroughly tested within the frame of this research work. Due to their importance and their novel status, they are explained and demonstrated in detail. We compare these new iterations with a standard method that is complemented by a feature to fit in the current context.

A further innovation is the computation of solutions in three-dimensional domains, which are still rare. Special attention is paid to applicability of the 3D simulation tools. The programs are designed to have justifiable working complexity. The simulation results of some models of contemporary semiconductor devices are shown and detailed comments on the results are given.

Eventually, we make a prospect on future development and enhancements of the models and of the algorithms that we used.

# Contents

**1 Introduction** **5**

**2 Modeling** **7**
2.1 Semiconductor fundamentals . . . . . 7
2.2 Derivation of the energy-transport model . . . . . 7
2.2.1 First macroscopic scale: the “spherical” harmonic expansion model . . . . . 10
2.2.2 Second macroscopic scale: the energy-transport model . . . . . 12
2.2.3 Properties of the diffusion matrix and the energy relaxation term . . . . . 13
2.3 The Poisson equation . . . . . 13
2.4 The stationary energy-transport model . . . . . 14

**3 The linearization and weak relation** **19**
3.1 Maximum principle for the Poisson equation . . . . . 19
3.2 The linear current continuity system . . . . . 20

**4 Numerical approximation** **25**
4.1 Mixed finite-element discretization . . . . . 25
4.2 Hybridized mixed finite elements . . . . . 27
4.3 Discretization of the Poisson equation . . . . . 28
4.4 Static condensation . . . . . 30
4.5 Evaluation of terminal currents . . . . . 31

**5 The iterations** **34**
5.1 Path following and related extended systems . . . . . 34
5.2 The full Newton method . . . . . 36
5.3 The Gummel-type iterations . . . . . 36
5.3.1 Decoupling of the current continuity system . . . . . 40
5.4 Vector extrapolation . . . . . 43
5.4.1 Reduced rank extrapolation . . . . . 44
5.4.2 Extension of the RRE to nonlinear sequences . . . . . 46
5.4.3 Application to the Gummel-type iterations . . . . . 46

**6 Numerical tests and simulations** **49**
6.1 Programming and technical notes . . . . . 49
6.2 Parameters of numerical tests . . . . . 49
6.2.1 Computation of the equilibrium solution and extension of applied voltage . 49
6.2.2 Approximation of diffusion coefficients . . . . . 50
6.2.3 Preconditioning and scaling . . . . . 50
6.2.4 Integration and smoothing of doping profile . . . . . 51
6.2.5 Continuation in the applied bias and path-following . . . . . 51
6.2.6 Grid generation and local refinement . . . . . 51
6.2.7 Postprocessing . . . . . 52
6.3 Test of 1D devices . . . . . 54
6.3.1 Ballistic diode . . . . . 54
6.4 Test of 2D devices . . . . . 58
6.4.1 Single-gate MESFET . . . . . 58
6.4.2 Double-gate MESFET . . . . . 66
6.4.3 MOSFET . . . . . 74
6.5 Simulation of quasi 1D devices and quasi 2D devices . . . . . 79
6.5.1 Ballistic diode . . . . . 79
6.5.2 Single-gate MESFET . . . . . 81
6.5.3 Double-gate MESFET . . . . . 83

6.5.4 MOSFET . . . . . . . . . . 85
6.6 Simulation of 3D devices . . . . . . . . . . 86
6.6.1 All-around gate MESFET . . . . . . . . . . 86
6.6.2 Non-Uniform single-gate MESFET . . . . . . . . . . 90

**7 Conclusion 93**

**8 Notations 94**


# 1 Introduction

Modern semiconductor devices based on hot-carrier transport are characterized by small size, and as a consequence, by strongly inhomogeneous distributions of electric field, carrier concentration, drift velocity, mean carrier energy, etc. Under these conditions, transport in sub-micrometer structures differs significantly from transport in bulk materials and/or super-micrometer devices. In order to describe these effects in sub-micron devices, simulation tools are based on new mathematical models. Because of the physical effects mentioned above, the modeling of semiconductor devices in the deep sub-micron era is a complicated and challenging procedure.

Two main classes of classical models can be distinguished, kinetic models, like the *semiconductor Boltzmann equation*, and fluid-dynamical models, like the *drift-diffusion equations* [Jer96, Jün01, MRS90]. The semiconductor Boltzmann equation gives quite accurate simulation results, but the numerical methods to solve this equations (for instance with *Monte-Carlo methods* [Wal95]) are too costly and time consuming to model real problems in semiconductor fabrication mode where simulation results are needed within hours or even minutes. Acceptable accuracy can be obtained by solving macroscopic (fluid-dynamical) model equations derived from the Boltzmann equation. The simplest such model are the well-known *drift-diffusion equations* for which very efficient numerical algorithms are available [BMP89, Jer95, JK91, JP97]. These models, however, are not as accurate as needed for sub-micron device modeling, owing to temperature effects of hot electrons, for instance.

In contrast, the *energy-transport models* are adequate to model temperature effects. They are of parabolic type such that their numerical solution needs less effort than hydrodynamic models which contain hyperbolic modes [Blø70, CCJS95, GJR89]. In the physical literature, energy-transport equations, first presented by Stratton in 1962 [Str62], are derived from hydrodynamic models usually by neglecting certain convection terms (see, e.g., [RGQ93] and references therein). This approach can be made rigorously by considering a diffusion time scaling [GN99].

Another approach is to derive the energy-transport equations from the Boltzmann equation by means of the *Hilbert expansion method* [AD96]. First, a so-called *spherical harmonic expansion* (SHE) model is derived from the Boltzmann equation in the diffusion limit, under the the assumption of dominant elastic scattering. Then, through a diffusion approximation, respectively scaling up electron-electron or phonon scattering, the energy-transport equations are derived from the SHE model. The advantage of this approach is the fact that is can be performed under quite weak assumptions on the semiconductor band structure and that explicit expressions for the diffusion coefficients and the energy relaxation term can be given [DJP00].

In the physical literature, the energy-transport equations have been investigated numerically for several years [ABC+95, CKR+92, CSP+92, SOTG94, Vis94], usually using Scharfetter-Gummel-type discretizations. Mathematicians started to pay attention to these models in the 1990s, using different numerical methods. Jerome and Shu [JS94] solved the equations employing ENO (essentially non-oscillatory) methods. Finite-difference discretizations have been used by Ringhofer [Rin01], based on an entropy-decaying property, and by Fournié [Fou99], with compact finite-difference schemes. Bosisio et al. used mixed finite-volume techniques [BSSG98]. Mixed finite-element methods applied to different formulations of the energy-transport models have been employed in [HJP03, HJP04, LC99], and in [BMM+05] an overview is given.

The energy-transport model in the *dual entropy formulation*, which will be the focus of this thesis,

has been discretized with *mixed finite elements* by Marrocco et al. [MMP96]. The same mixed finite elements have been used in this thesis. Unlike here, their approach is based on the introduction of an artificial transient problem derived from the stationary model. The semi-discretized problem is solved with the implicit Euler procedure. Another mixed finite-element discretization has been employed in [DJP00], but only in one space dimension. The most important feature of the schemes used in [MMP96] and [DJP00] is the current conservation (the current is introduced as an independent variable and continuity is directly imposed).
For devices in present ULSI technology often three-dimensional numerical tools are used to predict performance, and actually due to continuous scaling of IC structures many physical effects pose requirements for a full 3D simulation. When devices are scaled to deep sub-micron dimensions, the use of 3D simulations is essential in order to study the effects of fluctuations. Other examples of physical phenomena related to scaling are an increase of the power density, noise effects or implant shadowing.
Wong and Taur were the first to report a full 3D simulation of field-effect transistors under the influence of random discrete doping [WT93]. They used a drift-diffusion simulator, which models electron transport as incompressible fluid flow, considering the area under the gate as a checkerboard of smaller, interconnected devices, each with a different density of dopant atoms.
Another main issue consists in the geometry of modern semiconductor devices. The fabrication of real semiconductor devices relies on a planar technology. This simplifies the fabrication process, but it restricts the possibilities to built-up more powerful semiconductor devices. Three-dimensional device structures are quite promising to overcome such restrictions. For example, a possible solution to reduce the noise is the use of multi-gate field-effect transistors in which the gate contact encloses the channel region from different sides, leading to smaller no-signal currents.

## Outline of the thesis

This thesis is organized as follows. In section 2 the semiconductor Boltzmann equation is introduced as the starting point in the derivation of the model equations. Then the energy-transport equations are derived and the model assumptions are made explicit. The equations are scaled for numerical evaluation and physically motivated boundary conditions are added as well. In section 3 we make precise the function spaces for the solution and the model equations are interpreted in a weak sense. Furthermore, we state the maximum principle for the Poisson equation and the linearization of the current continuity system. Section 4 is designed to cover the main issues related to the numerical approximation of the linearized model equations. The iteration procedures used to solve the nonlinear discrete systems are considered in section 5. We start with a brief explanation of the implementation of the full Newton method, especially the combination with the path following method. The remaining part of this section is concerned with the description of two Gummel-type iterations and the implementation thereof. The combination of these Gummel-type procedures with a vector extrapolation concludes section 5. In section 6 we turn to the application of the numerical schemes to several semiconductor device models including one-dimensional and two-dimensional device models as well as uniform extensions thereof to three dimensions. Two non-uniform three-dimensional device models are also considered.
The thesis finishes with some concluding remarks and a short list of some of the notations used throughout the thesis.

# 2 Modeling

The exposition of this section on the derivation of the macroscopic semiconductor device model generally follows the lines of [AD96]. In selected parts additional information and comments are given. The proofs of the propositions are omitted.

## 2.1 Semiconductor fundamentals

The semiconductor Boltzmann equation is the origin in the derivation of many semiconductor models. This equation in turn is derived from the classical Boltzmann equation, which was developed to describe transport processes in diluted gases, where both ballistic motion of particles and collisions between them are considered.
The semiconductor Boltzmann equation is of kinetic type. Because of the semi-classical approximation, which sticks to the picture of a charge carrier as being a particle, albeit with unspecified dispersion relation, the Boltzmann equation is capable of describing charge transport in non-equilibrium states in solid materials as well. The semiconductor Boltzmann equation assumes that collision never involve more than two particles, an assumption that is adequate for solids.
The derivation of the macroscopic equations is accomplished in two steps. First, the so called *"spherical" harmonic expansion* (SHE model) is deduced from the Boltzmann equation under the assumption of dominant phonon scattering. In a second step the energy-transport model is derived from the SHE model.
Now, we introduce some preliminary notations. By $L$ we denote the Bravais lattice, which is assumed to be periodic. The Brillouin zone $B$ is the primitive cell of the dual lattice $L^*$. $L^*$ is also periodic and identified with the torus $\mathbb{R}^3/L^*$. Let $f(x,k,t)$ be the distribution function depending on the position $x \in \mathbb{R}^3$, the wave vector $k \in B$, and the time $t \geq 0$. In the following, every function of $k$ will be considered as $L^*$ periodic.

## 2.2 Derivation of the energy-transport model

The semi-classical Boltzmann equation reads as

$$\frac{\partial f}{\partial t} + \frac{1}{\hbar}\nabla_k \varepsilon(k) \cdot \nabla_x f - \frac{q}{\hbar} E_{\text{eff}}(x,t) \cdot \nabla_k f = Q_{\text{ld}}(f) + Q_{\text{e}}(f), \tag{2.1}$$

where $q$ denotes the elementary charge and $\hbar$ the reduced Planck constant. The coefficient $\varepsilon(k)$ of the Boltzmann equation is specified in order to describe the electron transport in the conduction band of a semiconductor crystal.
In order to describe the electrical phenomena of charge transport, in (2.1) the *effective electric field* $E_{\text{eff}}(x,t)$ is inserted. The semi-classical term $-\frac{q}{\hbar}E_{\text{eff}}(x,t)$ is the electrical force, acting on an electron ensemble, divided by $\hbar$. With the semiconductor crystal an energy band diagram is associated, $\varepsilon : k \in B \mapsto \varepsilon(k) \in \mathbb{R}$, to describe the energy states that the electrons can occupy in the crystal. It remains to consider collisions of the electrons. In [AD96] three types of electron scattering mechanisms due to lattice-defects are dealt with. These lattice defects are impurities, acoustical phonons, and optical phonons, which give rise to the following definition of the collision operator $Q_{\text{ld}}(f)$,

$$Q_{\text{ld}}(f) = Q_{\text{imp}}(f) + Q_{\text{ac}}(f) + Q_{\text{op}}(f). \tag{2.2}$$

By impurities foreign atoms in the solid are meant. Ionized donors and acceptors in a semiconductor are a common example of such impurities. Unlike the phonon scattering mechanisms, the impurity scattering is purely elastic and $Q_{\text{imp}}$ takes the following form, if $k$ denotes the wave vector of a particle before collision with the ion, and $k'$ is this vector after the collision,

$$Q_{\text{imp}}(f)(k) = \int_B \Phi_{\text{imp}}(x,k,k')\delta(\varepsilon' - \varepsilon)(f' - f)dk', \tag{2.3}$$

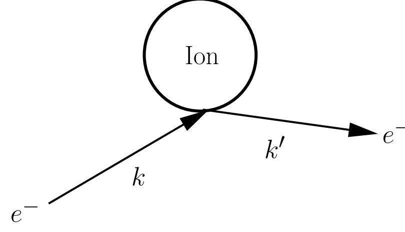



Figure 1: elastic scattering on an ion

where $\varepsilon = \varepsilon(k)$, $\varepsilon' = \varepsilon(k')$, $f = f(k)$, $f' = f(k')$, $\delta$ designates the delta measure, and $\Phi_{\rm imp}(x,k,k') = \Phi_{\rm imp}(x,k',k)$ is a collision kernel. In the literature, (2.3) is called a Boltzmann collision operator. It has the fundamental properties of conserving mass, momentum, and energy. Scattering by lattice waves, e.g. the absorption or emission of optical phonons is taken into consideration by

$$Q_{\rm op}(f)(k) = \int_B \Phi_{\rm op}(x,k,k')\{[(N_{\rm op}+1)\delta(\varepsilon-\varepsilon'+\varepsilon_{\rm op}) + N_{\rm op}\delta(\varepsilon-\varepsilon'-\varepsilon_{\rm op})]f'(1-f) - [(N_{\rm op}+1)\delta(\varepsilon'-\varepsilon+\varepsilon_{\rm op}) + N_{\rm op}\delta(\varepsilon'-\varepsilon-\varepsilon_{\rm op})]f(1-f')\}dk', \tag{2.4}$$

where $\Phi_{\rm op}(x,k,k') = \Phi_{\rm op}(x,k',k)$ is the optical phonon collision kernel. $\varepsilon_{\rm op}$ is the constant optical phonon energy, and $N_{\rm op}$ is the optical phonon occupation number,

$$N_{\rm op} = (\exp(\varepsilon_{\rm op}/(k_B T_L)) - 1)^{-1}$$

with $k_B$ the Boltzmann constant and $T_L$ the lattice temperature. The acoustical phonon collision operator is given by a corresponding expression with the index ‘op’ replaced by ‘ac’. In this case, the phonon energy is a non-constant function of $k - k'$. The inter-electronic collisions are taken into account by means of the operator $Q_e(f)$,

$$Q_{\rm e}(f)(k) = \int_{B^3} \Phi_{\rm e}(x,k,k_1,k',k_1')\delta(\varepsilon'+\varepsilon_1'-\varepsilon-\varepsilon_1)\delta_p(k'+k_1'-k-k_1)\times [f'f_1'(1-f)(1-f_1) - ff_1(1-f')(1-f_1')]dk_1dk'dk_1', \tag{2.5}$$

where

$$\delta_p(k'+k_1'-k-k_1) = \sum_{g\in L^*} \delta(k'+k_1'-k-k_1+g). \tag{2.6}$$

In order to take these different behaviors into account, in [AD96] a scaling of (2.1) is proposed that allows to separate the elastic and the inelastic scattering. The scaling is achieved by dividing the equation (2.1) by the typical kinetic energy of electrons. Let $n_0$ be the typical density of states in the structure (which can be larger than the doping density if high injection effects are present) and $\varepsilon_0$ be the typical kinetic energy that the electrons can gain in the structure ($\varepsilon_0 \approx qV_A$ if $V_A$ is the applied bias). Let $k_0$ be the typical norm of wave vectors $k$ satisfying $\varepsilon(k) = \varepsilon_0$. The dimensionless parameter

$$\eta = \frac{4\pi^3 n_0}{k_0^3}$$

is the natural distribution function scale and measures the level of degeneracy of the electron gas. The velocity scale is given by $v_0 = \varepsilon_0/\hbar k_0$ and the space scale is linked to the time scale $t_0$ through $x_0 = v_0 t_0$. Finally, the potential scale $V_0$ is connected to $\varepsilon_0$ by $qV_0 = \varepsilon_0$ (i.e., $V_0 \approx V_A$). Dimensionless parameters measuring the relative strength of the collision operators $\nu_{\rm imp}$, $\nu_{\rm op}$, $\nu_{\rm ac}$, $\nu_{\rm e}$, are given by

$$\nu_{\rm imp} = \frac{\phi_{\rm imp,0}k_0^3}{\varepsilon_0}t_0, \quad \nu_{\rm e} = \frac{4\phi_{\rm e,0}\pi^3 n_0 k_0^3}{\varepsilon_0}t_0, \tag{2.7}$$

where $\phi_{\rm imp,0}$ and $\phi_{\rm e,0}$ are common values of $\Phi_{\rm imp}$ and $\Phi_{\rm e}$. The parameters $\nu_{\rm op}$ and $\nu_{\rm op}$ are given by similar expressions as for $\nu_{\rm imp}$ associated with common values $\Phi_{\rm op,0}$ and $\Phi_{\rm ac,0}$. In this context, $\Phi_{\rm imp,0}$ is applied to scale the matrix element $\Phi_{\rm imp}$ in dimensionless form and similarly with the other scattering mechanisms.
Eventually, let also $\varepsilon_{\rm ac,0}$ be the order of magnitude of the acoustical phonon energy and introduce the dimensionless parameters

$$\alpha^2 = \frac{\varepsilon_{\rm op}}{\varepsilon_0}, \quad \beta^2 = \frac{\varepsilon_{\rm ac,0}}{\varepsilon_0}, \quad \gamma^2 = \frac{k_B T_L}{\varepsilon_0}. \tag{2.8}$$

$\alpha^2$ measures the typical energy gain or loss due to an optical collision. $\beta^2$ measures this energy balance for acoustical collisions. The parameter $\gamma^2$ quantifies how “hot” the electron gas is. $\gamma^2 \ll 1$ is a characterization of hot electron effects.
After scaling with the above units, the equations are written,

$$\frac{\partial f}{\partial t} + \nabla_k \varepsilon(k)\cdot\nabla_x f - E_{\rm eff}(x,t)\cdot\nabla_k f = \nu_{\rm imp} Q_{\rm imp}(f) + \nu_{\rm ac} Q^\beta_{\rm ac} + \nu_{\rm op} Q^\alpha_{\rm op} + \nu_{\rm e} Q_{\rm e}(f) \tag{2.9}$$

with $Q_{\rm imp}$ given by (2.3) applied with the dimensionless quantities. Here $Q^\alpha_{\rm op}(f)$ is obtained from (2.4) in the same way, and by replacing $\varepsilon_{\rm op}$ by $\alpha^2$, $(1-f)$ by $(1-\eta f)$, and $(1-f')$ by $(1-\eta f')$. In order to obtain $Q^\beta_{\rm ac}(f)$ from (2.4), it is sufficient to replace $\Phi_{\rm op}$ by the dimensionless form of $\Phi_{\rm ac}$, $N_{\rm op}$ by $N_{\rm ac}$, $\varepsilon_{\rm op}$ by $\beta^2\varepsilon_{\rm ac}$, and $(1-f)$ by $(1-\eta f)$. The dimensionless form of $Q_{\rm e}$ is obtained by this substitution, too.
The equations are posed on a dilated Brillouin zone $B/k_0$. Note that

$$N_{\rm op} = \left(e^{\alpha^2/\gamma^2} - 1\right)^{-1}, \quad N_{\rm ac} = \left(e^{\beta^2 \varepsilon_{ac}/\gamma^2} - 1\right)^{-1}.$$

The next step is to analyze the behavior of (2.9) at various time, length, and energy scales. The first case of interest is the high energy regime, i.e. for large applied voltages. There, the relative energy gain or loss of electron energy due to a phonon collision is very small. This assumption is quantified by

$$\alpha^2 \ll 1, \quad \beta^2 \ll 1. \tag{2.10}$$

Apart from (2.10) the supposition

$$\frac{\alpha^2}{\gamma^2} \approx \frac{\beta^2}{\gamma^2} \approx 1 \tag{2.11}$$

is made, which means that in the high energy regime, the acoustical phonon energy $\varepsilon_{\rm ac,0}$, the optical phonon energy $\varepsilon_{\rm op,0}$, and the lattice thermal energy $k_{\rm B}T_{\rm L}$ are all considered of the same order of magnitude. Furthermore, they are very small compared with the electron energy $\varepsilon_0$. A validation of these assumptions is given in [AD96].
By expanding $Q^\alpha_{\rm op}(f)$ and $Q^\beta_{\rm ac}(f)$ in powers of $\alpha^2$ and $\beta^2$ and using (2.10), the global lattice-defect collision operator can be written as,

$$Q_{\rm ld} = (\nu_{\rm imp} + \nu_{\rm ac} + \nu_{\rm op}) Q_0(f) + \beta^2 \nu_{\rm ac} Q^\beta_{\rm ac,1}(f) + \alpha^2 \nu_{\rm op} Q^\alpha_{\rm op,1}(f),$$

where $Q_0(f)$ is an elastic collision operator of the form,

$$\begin{aligned} Q_0(f) &= \int_{\mathbb{R}^3} \Phi_0(x,k,k')\delta(\varepsilon' - \varepsilon)(f - f')dk', \\ \Phi_0 &= \Phi_{\rm imp} + (2N_{\rm op}+1)\Phi_{\rm op} + (2N_{\rm ac}+1)\Phi_{\rm ac}, \end{aligned} \tag{2.12}$$

and $Q^\beta_{\rm ac,1}$ (resp. $Q^\alpha_{\rm op,1}$) is of order 1 when $\beta$ (resp. $\alpha$) tends to zero. The time scale is chosen so that $(\nu_{\rm imp} + \nu_{\rm ac} + \nu_{\rm op}) = 1$. According to (2.11), the choice $\alpha^2 = \beta^2$ is made and the assumption

$$\nu_{\rm e} = O(\alpha^2), \quad \text{i.e., } \nu_{\rm e} = \alpha^2 \overline{\nu}_{\rm e},\ \overline{\nu}_{\rm e} = O(1). \tag{2.13}$$

We define

$$Q_1^\alpha = \nu_{ac} Q_{ac,1}^\alpha(f) + \nu_{op} Q_{op,1}^\alpha(f) + \overline{\nu}_e Q_e(f) = O(1), \text{ as } \alpha \to 0. \tag{2.14}$$

The final form of the scaled Boltzmann equation (2.9) is

$$\frac{\partial f}{\partial t} + \nabla_k \varepsilon(k) \cdot \nabla_x f - E_{\text{eff}}(x,t) \cdot \nabla_k f = Q_0(f) + \alpha^2 Q_1^\alpha(f), \tag{2.15}$$

where $Q_0(f)$ is the elastic operator, which contains contributions of all scattering mechanisms, and the remaining inelastic part of scattering. $\alpha$ is a dimensionless parameter.

### 2.2.1 First macroscopic scale: the "spherical" harmonic expansion model

We are interested in the diffusion scaling of equation (2.15). Following [AD96], the new variables $x' = \alpha x, t' = \alpha^2 t$ are introduced. The scaled Boltzmann equation then takes the form

$$\frac{\partial f^\alpha}{\partial t} + \frac{1}{\alpha} \left( \nabla_k \varepsilon \cdot \nabla_x f^\alpha + \nabla_x V \cdot \nabla_k f^\alpha \right) = \frac{1}{\alpha^2} Q_0(f^\alpha) + Q_1^\alpha(f^\alpha). \tag{2.16}$$

The Hilbert expansion of $f^\alpha$,

$$f^\alpha = f_0 + \alpha f_1 + \alpha^2 f_2 + \cdots,$$

is inserted in (2.16) and after identification of equal powers of $\alpha$, we obtain

$$Q_0(f_0) = 0, \tag{2.17}$$

$$Q_0(f_1) = \nabla_k \varepsilon \cdot \nabla_x f_0 + \nabla_x V \cdot \nabla_k f_0, \tag{2.18}$$

$$Q_0(f_2) = \frac{\partial f_0}{\partial t} + \nabla_k \varepsilon \cdot \nabla_x f_1 + \nabla_x V \cdot \nabla_k f_1 - Q_1^\alpha(f_0). \tag{2.19}$$

The linear operator $Q_0$ needs to be examined in order to solve these equations for $f_0, f_1, f_2$. In [AD96] a thorough analysis shows that there exists a function $F(x, \varepsilon, t)$ such that

$$f_0(x,k,t) = F(x, \varepsilon(k), t), \tag{2.20}$$

where $F(x,\varepsilon,t)$ is contained in a weighted Sobolev space: let $dS_e(k)$ be the Euclidean surface element on the manifold $\varepsilon^{-1}(e)$. We denote $dN_e(k) = dS_e(k)/|\nabla\varepsilon(k)|$ and define the density of states of energy $e$

$$N(e) = \int_{\varepsilon^{-1}(e)} dN_e(k). \tag{2.21}$$

We introduce

$$Lf = Q_0\left(\frac{f}{N(\varepsilon(k))}\right) = \frac{1}{N(\varepsilon(k))} Q_0(f) \tag{2.22}$$

and the weighted space

$$L_N^2 = \left\{ f(k) \; : \; \int_B f^2(k) N(\varepsilon(k)) dk < \infty \right\}. \tag{2.23}$$

The assumption $N(\varepsilon(k)) \neq 0$ for $k$ a.e. in $B$ ensures that $L_N^2$ is a separable Hilbert space with the scalar product and norm,

$$(f,g)_N = \int_B f(k) g(k) N(\varepsilon(k)) dk, \quad \|f\|_N^2 = (f,f)_N.$$

The function $F(x,\varepsilon,t)$ is the electron distribution function depending on the position $x$, particle energy $\varepsilon$, and time $t$. $f_1$ can then be expressed in terms of this function $F$ and the unique solution $\boldsymbol{\lambda} \in (ker L)^\perp$ of

$$-L(\boldsymbol{\lambda}) = \frac{1}{N(\varepsilon(k))} \nabla_k \varepsilon. \tag{2.24}$$

The unique solution in $(kerL)^{\perp}$ of (2.18) is given by

$$f_1 = -\boldsymbol{\lambda}(x,k) \cdot \left( \nabla_x F + \nabla_x V \frac{\partial F}{\partial \varepsilon} \right). \tag{2.25}$$

With the definitions

$$S^{\alpha}(F) = \int_{\varepsilon} Q_1^{\alpha}(F) dN_{\varepsilon}(k), \quad J(x,\varepsilon,t) = \int_{\varepsilon} \nabla_k \varepsilon f_1(x,k,t) dN_{\varepsilon}(k), \tag{2.26}$$

the solvability of (2.19) is stated in the following theorem:

**Theorem 2.1.** *Let $f_0$ and $f_1$ be given by* (2.20) *and* (2.25). *Then, the solution $f_2$ of* (2.19) *exists if and only if $F(x,\varepsilon,t)$ satisfies*

$$N(\varepsilon)\frac{\partial F}{\partial t} + \nabla_x \cdot F + \nabla_x V \cdot \frac{J}{\varepsilon} - S^{\alpha}(F) = 0, \tag{2.27}$$

$$J(x,\varepsilon,t) = -D(x,\varepsilon)\left( \nabla_x F + \nabla_x V \frac{\partial F}{\partial \varepsilon} \right), \tag{2.28}$$

*where the coefficient in* (2.28) *is defined by*

$$D(x,\varepsilon) = \int_{\varepsilon} \nabla_k \varepsilon \otimes \boldsymbol{\lambda}(x,k) dN_{\varepsilon}(k).$$

The proof is given in [AD96].

**Proposition 2.1.** *$D(x,\varepsilon)$ is a symmetric non-negative $3 \times 3$ matrix. Moreover, $\exists C > 0$ such that*

$$D(x,\varepsilon) \geq \frac{C}{N(\varepsilon)} \int_{\varepsilon} \nabla_k \varepsilon \otimes \nabla_k \varepsilon dN_{\varepsilon}(k). \tag{2.29}$$

**Remark 2.1.** The right-hand side of (2.29) is a symmetric non-negative $3 \times 3$ matrix which is degenerate at the critical points of $\varepsilon$.

Now we investigate simplifying assumptions which enable us to give more explicit expressions of $D(x,\varepsilon)$. First, we assume that $\Phi_0(x,k,k') = \Phi_0(x,\varepsilon(k))$. Recall that $\Phi_0$ needs only be defined on the set $\varepsilon(k') = \varepsilon(k)$. Then $Q_0(f)$ is equivalently written in the form of a relaxation operator,

$$Q_0(f) = -\frac{1}{\tau(x,\varepsilon)}(f - Pf), \tag{2.30}$$

where $\tau(x,\varepsilon) = (\Phi_0(x,\varepsilon)N(\varepsilon))^{-1}$ and $P$ is the orthogonal projection on $kerL$. Then the solution of (2.24) can be written as

$$\boldsymbol{\lambda}(x,k) = \tau(x,\varepsilon)\nabla_k \varepsilon, \tag{2.31}$$

and can be referred to as “mean free path of particles of wave vector $k$.” This results in a another representation of $D(x,\varepsilon)$,

$$D(x,\varepsilon) = \tau(x,\varepsilon) \int_{\varepsilon} \nabla_k \varepsilon \otimes \nabla_k \varepsilon dN_{\varepsilon}(k). \tag{2.32}$$

### 2.2.2 Second macroscopic scale: the energy-transport model

The starting point now is the "spherical" harmonic expansion model (2.27) and (2.28). Taking (2.14) into consideration, the collision term $S^\alpha(f)$ is written

$$S^\alpha(F) = \frac{\nu_{\rm ac}}{\alpha^2} S^\alpha_{\rm ac}(F) + \frac{\nu_{\rm op}}{\alpha^2} S^\alpha_{\rm op}(F) + \overline{\nu}_{\rm e} S_e(F), \tag{2.33}$$

where we have

$$S^\alpha_{\rm ac}(F)(\varepsilon) = \int_\varepsilon Q^\alpha_{\rm ac}(F) dN_\varepsilon(k), \quad S^\alpha_{\rm op}(F)(\varepsilon) = \int_\varepsilon Q^\alpha_{\rm op}(F) dN_\varepsilon(k), \tag{2.34}$$

$$S^\alpha_{\rm e}(F)(\varepsilon) = \int_\varepsilon Q^\alpha_{\rm e}(F) dN_\varepsilon(k) \qquad . \tag{2.35}$$

Note that the elastic parts $Q^0_{\rm ac}$ and $Q^0_{\rm op}$ vanish on $F(\varepsilon)$. The subsequent analysis is based on the assumption that both collision operators $\nu_{\rm ac}/\alpha^2 S^\alpha_{\rm ac}(F)$ and $\nu_{\rm op}/\alpha^2 S^\alpha_{\rm op}(F)$ are of the same order of magnitude. The time unit is chosen in such a way that $\nu_{\rm ac}/\alpha^2$ and $\nu_{\rm op}/\alpha^2$ are approximately 1. Apart from that the electron-electron collision operator $\overline{\nu}_{\rm e} S_{\rm e}(F)$ is assumed to be dominant and $\overline{\nu}_{\rm e} = 1/\beta$, $\beta \ll 1$. For the physical background, see [AD96].
The evolution of a hot-electron energy distribution $F^\beta(x, \varepsilon, t)$ is governed by the equations,

$$N(\varepsilon)\frac{\partial F^\beta}{\partial t} + \nabla_x \cdot J^\beta + \nabla_x V \cdot \frac{\partial J^\beta}{\partial \varepsilon} = \frac{1}{\beta} S_{\rm e}(F^\beta) + S_1(F^\beta), \tag{2.36}$$

$$J^\beta(x, \varepsilon, t) = -D(x, \varepsilon)\left(\nabla_x F^\beta + \nabla_x V \frac{\partial F^\beta}{\partial \varepsilon}\right), \tag{2.37}$$

with $S_{\rm e}(F^\beta)$ defined by (2.35) and $S_1(F^\beta) = S_{\rm ac}(F^\beta) + S_{\rm op}(F^\beta)$ and $S_{\rm ac}, S_{\rm op}$ given by (2.34). The limit $\beta \to 0$ is of interest now. Expanding $F^\beta$ and $J^\beta$,

$$F^\beta = F_0 + \beta F_1 + \cdots, \quad J^\beta = J_0 + \beta J_1 + \cdots,$$

and identifying equal powers of $\beta$ gives

$$S_{\rm e}(F_0) = 0, \tag{2.38}$$

$$N(\varepsilon)\frac{\partial F_0}{\partial t} + \nabla_x \cdot J_0 + \nabla_x V \cdot \frac{\partial J_0}{\partial \varepsilon} - S_1(F_0) = D_{F_0} S_{\rm e}(F_1). \tag{2.39}$$

Here $D_{F_0} S_{\rm e}(F_1)$ denotes the derivative of $S_{\rm e}$ at $F_0$ applied to $F_1$. In order to solve (2.38) and (2.39), the operators $S_{\rm e}$ and $D_{F_0} S_{\rm e}(F_1)$ are analyzed. See [AD96] for details. The solution of (2.38) is given by Fermi-Dirac distribution function

$$\mathcal{F}_{\mu,T}(\varepsilon) = \frac{1}{\eta + e^{(\varepsilon-\mu)/T}}, \tag{2.40}$$

which describes the distribution of electrons in thermal equilibrium:

$$F_0(x, \varepsilon, t) = \mathcal{F}_{\mu(x,t),T(x,t)}(\varepsilon). \tag{2.41}$$

$J_0$ is given by (2.37) with $\beta = 0$,

$$J_0(x, \varepsilon, t) = -D(x, \varepsilon)\mathcal{F}(1 - \mathcal{F})\left[\left(\nabla_x\left(\frac{\mu}{T}\right) - \frac{\nabla_x V}{T}\right) - \varepsilon\nabla\left(\frac{1}{T}\right)\right]. \tag{2.42}$$

**Theorem 2.2.** *Let $F_0$ be given by (2.41). Then the solution of equation (2.39) exists (formally) if and only if $\mu(x,t)$ and $T(x,t)$ satisfy the following set of diffusion equations,*

$$\frac{\partial}{\partial t} n(\mu,T) - \nabla\cdot \left[ \mathcal{L}_{11} \left( \nabla_x \left( \frac{\mu}{T} \right) - \frac{\nabla_x V}{T} \right) + \mathcal{L}_{12} \frac{\nabla_x T}{T^2} \right] = 0, \tag{2.43}$$

$$\frac{\partial}{\partial t} (n\mathcal{E}(\mu,T)) - \nabla\cdot \left[ \mathcal{L}_{21} \left( \nabla_x \left( \frac{\mu}{T} \right) - \frac{\nabla_x V}{T} \right) + \mathcal{L}_{22} \frac{\nabla_x T}{T^2} \right] + \nabla\cdot \left[ \mathcal{L}_{11} \left( \nabla_x \left( \frac{\mu}{T} \right) - \frac{\nabla_x V}{T} \right) + \mathcal{L}_{12} \frac{\nabla_x T}{T^2} \right] = W(\mu,T,T_L), \tag{2.44}$$

*where*

$$\begin{pmatrix} n(\mu,T) \\ n\mathcal{E}(\mu,T) \end{pmatrix} = \int_{\mathbb{R}} \mathcal{F}_{\mu,T}(\varepsilon) \begin{pmatrix} 1 \\ \varepsilon \end{pmatrix} N(\varepsilon) d\varepsilon, \tag{2.45}$$

$$\mathcal{L}_{ij} = \mathcal{L}_{i,j}(x,\mu,T) = \int_{\mathbb{R}} D(x,\varepsilon)\mathcal{F}(1-\eta\mathcal{F})\varepsilon^{i+j-2} d\varepsilon, \quad i,j=1,2, \tag{2.46}$$

*and*

$$W(\mu,T,T_L) = \int_{\mathbb{R}} S_1(\mathcal{F}_{\mu,T})\varepsilon \, d\varepsilon. \tag{2.47}$$

The proof is given in [AD96].

### 2.2.3 Properties of the diffusion matrix and the energy relaxation term

The diffusion matrix is the block $6\times 6$ matrix,

$$\mathcal{L} = \begin{pmatrix} \mathcal{L}_{11} & \mathcal{L}_{12} \\ \mathcal{L}_{21} & \mathcal{L}_{22} \end{pmatrix} = \mathcal{L}(x,\mu,T).$$

In [AD96] a proof is given for the following proposition.

**Proposition 2.2.**

1. *For all $i,j=1,2$ it holds $\mathcal{L}_{ij}^T = \mathcal{L}_{ij}$.*
2. *$\mathcal{L}_{12} = \mathcal{L}_{21}$.*
3. *Assume that the six functions $\partial_{k_1}\varepsilon,\ \partial_{k_2}\varepsilon,\ \partial_{k_3}\varepsilon,\ \varepsilon\partial_{k_1}\varepsilon,\ \varepsilon\partial_{k_2}\varepsilon,\ \partial_{k_3}\varepsilon$ are linearly independent. Then the matrix $\mathcal{L}(x,\mu,T)$ is symmetric positive definite for any $\mu\in\mathbb{R}$, $T>0$.*

Moreover, there it is stated that $W$ is a relaxation term of the electron temperature $T$ towards the lattice temperature $T_L$ : $W(\mu,T,T_L)\cdot(T-T_L)\le 0$.
In the following subsection the system of diffusion equations (2.43),(2.44) is complemented by a Poisson equation for the electrostatic potential $V$ to make the system self-consistent. By means of the Poisson equation the electronic effects will be modeled.

## 2.3 The Poisson equation

Since the electrons and the background ions in the semiconductor crystal are charged particles, a force acts between them according to the Coulomb law. The electron interaction is described by an internal electric field,

$$E_{\text{int}}(x,y) = -\frac{q}{4\pi\varepsilon_s} \frac{x-y}{\|x-y\|_2^3}, \quad x,y\in\mathbb{R}^3,\ x\ne y. \tag{2.48}$$

Here, the semiconductor permittivity $\varepsilon_s$ is a constant material parameter. The minus sign indicates the negative charge of the electron located at $x$. The force

$$F(x,y) = -qE_{int}(x,y)$$

between the electrons is repulsive. The positively charged background ions that result from the doping of the semiconductor crystal, are modeled by means of a (positive) doping profile function $\mathrm{Dop} \in L^2(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$.
An external field is present due to the electrical effects of the artificial impurities. It is defined by

$$E_{\mathrm{ext}}(x,t) = \frac{q}{4\pi\varepsilon_s} \int_{\mathbb{R}^3} \mathrm{Dop}(y) \frac{x-y}{\|x-y\|_2^3} dy, \tag{2.49}$$

where $\|\cdot\|_2$ denotes the Euclidean norm in $\mathbb{R}^3$. If we combine (2.48) and (2.49), we get for the effective electric field in (2.1)

$$E_{\mathrm{eff}}(x,t) = E_{\mathrm{ext}}(x) + \int_{\mathbb{R}^3 \times B} f(y,k,t)\, E_{\mathrm{int}}(x,y)\, d(y,k). \tag{2.50}$$

Now we express (2.50) in more explicit terms. To achieve this we use the following equivalence relation,

$$\phi(x) = -\frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{f(y)}{\|x-y\|_2} dy,\ x \in \mathbb{R}^3 \quad \Longleftrightarrow \quad \Delta\phi = f \text{ in } \mathbb{R}^3,\ f \in L^2(\mathbb{R}^3). \tag{2.51}$$

Because of

$$f(x) = \Delta\phi(x) = \frac{1}{4\pi} \mathrm{div} \int_{\mathbb{R}^3} f(y) \frac{x-y}{\|x-y\|_2^3} dy,$$

we have

$$\mathrm{div}\, E_{\mathrm{eff}}(x,t) = \frac{q}{\varepsilon_s} \mathrm{Dop}(x) + \int_{\mathbb{R}^3} n(y,t)\, \mathrm{div}\, E_{\mathrm{int}}(x,y) dy. \tag{2.52}$$

By repeated use of (2.51) we conclude that

$$\int_{\mathbb{R}^3} n(y,t)\, \mathrm{div}\, E_{\mathrm{int}}(x,y) dy = -\frac{q}{\varepsilon_s} n(x,t),$$

and (2.52) constitutes Gauss' law in the special setting,

$$\mathrm{div} E_{\mathrm{eff}}(x,t) = \frac{q}{\varepsilon_s} (\mathrm{Dop}(x) - n(x,t)). \tag{2.53}$$

From (2.48) and (2.49) it follows that $E_{\mathrm{eff}}$ is a gradient field, i.e. $\exists V : E_{\mathrm{eff}} = -\nabla V$. Finally, we insert this relation into (2.53) and end up with the Poisson equation,

$$\Delta V(x,t) = \frac{q}{\varepsilon_s} (n(x,t) - \mathrm{Dop}(x)). \tag{2.54}$$

## 2.4 The stationary energy-transport model

The system of diffusion equations (2.43) and (2.44) is a macroscopic approximation of the scaled semi-classical Boltzmann equation (2.15). Since we are interested in the unscaled results, i.e. the quantities in physical units, we revert the scaling.
By $\mu, T$, and $V$ the *chemical potential*, the *electron temperature*, and the *electrostatic potential*

are denoted. The unscaled stationary unipolar energy-transport equations in the *primal entropy variables* $\mu/T$ and $-1/T$ are given by

$$-\nabla\cdot J_1 = 0, \tag{2.55}$$

$$-\nabla\cdot J_2 = -J_1\cdot\nabla V + W(n,T), \tag{2.56}$$

$$J_1 = \mathcal{L}_{11}\left(\nabla\frac{q\mu}{k_BT} - \frac{q\nabla V}{k_BT}\right) + \mathcal{L}_{12}\nabla\left(-\frac{1}{k_BT}\right), \tag{2.57}$$

$$qJ_2 = \mathcal{L}_{21}\left(\nabla\frac{q\mu}{k_BT} - \frac{q\nabla V}{k_BT}\right) + \mathcal{L}_{22}\nabla\left(-\frac{1}{k_BT}\right), \tag{2.58}$$

$$\varepsilon_s\Delta V = q(n-\mathrm{Dop}). \tag{2.59}$$

$J_1$ is the *electron current density* and $J_2$ is the *energy current density.* The system consists of the conservation law of mass (2.55) and the conservation law of (kinetic) energy (2.56) as well as the constitutive relations for the electron current density (2.57) and the energy current density (2.58). These equations are coupled to the Poisson equation (2.59) to include the electric phenomena of charge transport and to make the system self-consistent. The physical constants are the elementary charge $q$, the Boltzmann constant $k_B$, and the semiconductor permittivity $\varepsilon_s$. The electron density $n$ depends on $\mu$ and $T$. The space dependent positive function Dop is the *doping profile*, which is used to model artificial impurities. $\mathcal{L}_{ij} = \mathcal{L}_{ij}(n,T)$ are the *diffusion coefficients*, and $W = W(n,T)$ is the *energy relaxation term.* The above equations have to be solved in the bounded, polyhedral semiconductor domain $\Omega\subset\mathbb{R}^3$. The following physically motivated boundary conditions complement the problem,

$$\begin{array}{llllllll} n = n_D, & T = T_D, & V = V_D & \text{on} & \Gamma_D, \\ \nu\cdot J_1 = 0, & \nu\cdot J_2 = 0, & \nu\cdot\nabla V = 0 & \text{on} & \Gamma_N. \end{array} \tag{2.60}$$

The Dirichlet boundary $\Gamma_D$ is split into $\Gamma_{D,1}$ that represents Ohmic contacts and $\Gamma_{D,2}$ that represents Schottky contacts, where the electron density $n$ is set to an equilibrium value $n_D$. At the contacts $\Gamma_{D,1}$ we assume the *charge neutrality relation* $n_D = \mathrm{Dop}$, whereas at the Schottky contacts $\Gamma_{D,2}$ the boundary value $n_D$ represents a low doping concentration that is less than the density of states in the conduction band. The electron temperature $T$ is set to the ambient temperature $T_D$. The electrostatic potential $V_D$ at the Ohmic contacts $\Gamma_{D,1}$ is the sum of the *built-in potential* $V_{\mathrm{bi}}$ due to the doping and the applied voltage $V_A$. On $\Gamma_{D,2}$, in addition to $V_{\mathrm{bi}}$ and $V_A$ a barrier potential $V_B$ is present. The potential bias between the Ohmic contacts drives the electron current through the device, from low potential to high potential. The magnitude of this current depends on the control voltage at the Schottky contacts.
The Neumann boundary $\Gamma_N$ constitutes the insulating boundary part, i.e. the normal components of the current densities are set to zero. We assume that $\partial\Omega = \Gamma_D\cup\Gamma_N$ and $\Gamma_D\cap\Gamma_N = \emptyset$. $\nu$ denotes the exterior normal unit vector on $\Gamma_N$.
All devices under consideration will be voltage driven, i.e only Dirichlet boundary data and no (inhomogeneous) Neumann boundary condition is prescribed at the Schottky contacts $\Gamma_{D,2}$.
In this subsection we reformulate the diffusion coefficients of the energy-transport model (2.55)-(2.59), and we make precise our assumptions on the energy relaxation term ([DJP00]). In order to get more explicit expressions for the coefficients $\mathcal{L}_{ij}$ in terms of $n,T$ (or $\mu,T$), we have to impose some physical assumptions.

(H1) The energy-band diagram $\varepsilon$ of the semiconductor crystal is spherically symmetric and a strictly monotone function of the modulus $|k|$ of the wave vector $k$. In such a case, the relation $\varepsilon:\mathbb{R}\to\mathbb{R}$, $k\mapsto\varepsilon(|k|)$ can be inverted, and we can define a function $\gamma$ such that $|k|^2 = \gamma(\varepsilon)$. Furthermore, the Brillouin zone equals $\mathbb{R}^3$.

(H2) A momentum relaxation time can be defined by

$$\tau(\varepsilon) = \left(\phi_0(2N_0+1)\varepsilon^\beta N(\varepsilon)\right)^{-1}, \quad \beta > -2,\ \phi_0 > 0, \tag{2.61}$$

where $N(\varepsilon) = 4\pi|k|^2/|\varepsilon(|k|)|$ is the density of states of energy $\varepsilon = \varepsilon(|k|)$ [AD96, (III.31)] and $N_0$ is the phonon occupation number [ADMS01, section 3].

(H3) The electron density $n$ and the internal energy $n\mathcal{E}$ are given by non-degenerate Boltzmann statistics. This also means $\mathcal{F}(1-\eta\mathcal{F}) \to \exp(-(\varepsilon-\mu)/T)$ when $\eta \to 0$ in (2.46).

The assumptions (H1) and (H2) are made in order to get simpler expressions for the variables. The non-degeneracy assumption (H3) is valid for semiconductor devices with a doping concentration which is below $10^{19}\,\mathrm{cm}^{-3}$. Many devices in practical applications satisfy this condition.
Under these assumptions, the diffusion coefficients are given by

$$\mathcal{L}_{ij} = L_{ij} Id, \quad L_{ij} = L_{ij}(\mu, T) = e^{\mu/T} \int_0^\infty d(\varepsilon)\varepsilon^{i+j-2} e^{-\varepsilon/T} d\varepsilon, \tag{2.62}$$

where $Id$ is the identity matrix in $\mathbb{R}^{3\times 3}$ and

$$d(\varepsilon) = \frac{4\pi}{3}\tau(\varepsilon)|\varepsilon'(|k|)\,|k|^2 \quad \text{and} \quad \varepsilon = \varepsilon(|k|),$$

(see [AD96, (IV.17),(III.333)]).
Due to assumption (H3), we have further [AD96, (IV.16)]

$$n = n(\mu, T) = e^{\mu/T} \int_0^\infty e^{-\varepsilon/T} N(\varepsilon) d\varepsilon, \tag{2.63}$$

$$n\mathcal{E} = n\mathcal{E}(\mu, T) = e^{\mu/T} \int_0^\infty \varepsilon e^{-\varepsilon/T} N(\varepsilon) d\varepsilon. \tag{2.64}$$

We further simplify the model equations by imposing the hypotheses,

(H4) The collision operator matrix element $\Phi_0 = \Phi_0(x)$ in (2.12) is independent of $k, k'$, which is true for acoustical and nonpolar optical phonons and the energy band diagram $\varepsilon$ is parabolic, i.e., $\gamma(\varepsilon) = 2m_*\varepsilon$ with $m_*$ being the effective mass. (see [AD96, III.D])

(H5) The energy relaxation term $W(n,T)$ is given by a Fokker-Planck approximation (see [DJP00, section 2.2]).

Under these hypotheses it is shown in [DJP00] that the following relations hold:

$$\begin{aligned}
n &= n_i T^{\frac{3}{2}} \exp\left(\frac{\mu}{T}\right), \quad \text{where } n_i > 0,\\
L_{ij} &= \mu_0 \frac{2\Gamma(i+j-\beta)}{\sqrt{\pi}} T^{i+j-\beta} n,\ i,j = 1,2,\\
W(n,T) &= -\frac{3}{2}\frac{n(T-T_L)}{\tau_\beta(T)},\\
\tau_\beta(T) &= \tau_0 \frac{3\sqrt{\pi}}{4\Gamma(\beta+2)} T^{\frac{1}{2}-\beta}.
\end{aligned}$$

Here, $\tau_\beta(T)$ is the *energy relaxation time*, and the constant $\beta > -2$ is a parameter in the energy relaxation model. For the precise definition of the parameters $n_i, \tau_0$ and $\mu_0$ we refer to [DJP00, section 2.4]. The symbol $\Gamma$ denotes the Gamma function. In the physical literature, the values $\beta = 0$ (see [LPSV92]) and $\beta = 1/2$ (see [CKR+92]) have been used in the case of a parabolic energy band. We set $\beta = 1/2$. Since $\Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}$, $\Gamma\left(\frac{5}{2}\right) = \frac{3\sqrt{\pi}}{4}$, $\Gamma\left(\frac{7}{2}\right) = \frac{15\sqrt{\pi}}{8}$, and $T_L = 1$ we have

$$(L_{ij}) = \mu_0 n \begin{pmatrix} 1 & \frac{3T}{2} \\ \frac{3T}{2} & \frac{15T^2}{4} \end{pmatrix}$$

and

$$W(\mu, T) = \frac{3}{2}\frac{n(T_L - T)}{\tau_0}.$$

We refer to an energy-transport model with $\beta = \frac{1}{2}$ as the *Chen model*.
The entries of the diffusion matrix $L_{ij}$ allow for a physical interpretation. $L_{11}/\mu_0$ is the electron density, and $L_{12}/\mu_0$ is the internal (or thermal) energy of the electron gas. Obviously, the internal energy is proportional to the electron temperature $T$ and independent of the volume that the electron gas occupies. The scaled diffusion coefficient $L_{22}/\mu_0$ also has a physical meaning, since $3/2nT^2$ is the variance of the thermal energy.
Now the equations (2.55)-(2.59) are transformed into a scaled and dimensionless form, where the scaling slightly differs from that used to derive the model equations. The second scaling is done to make the equations more appropriate for numerical evaluation and to identify the model parameters. Let $C_m$ be the maximum value of the doping profile function, $l^*$ the diameter of the device, $\mu_0$ the low-field mobility of the charge carriers, $T_L$ the lattice temperature, and $U_T = k_B T_L/q$ the thermal voltage. We apply the scaling,

$$n \mapsto C_m\, n, \quad \text{Dop} \mapsto C_m\text{Dop}, \quad T \mapsto T_L T, \quad \mu \mapsto U_T\mu, \quad x \mapsto l^* x,$$
$$J_1 \mapsto (q\mu_0 U_T C_m/l^*)J_1, \quad J_2 \mapsto (q\mu_0 U_T^2 C_m/l^*)J_2,$$
$$L_{ij} \mapsto [(qU_T)^{i+j-1}\mu_0 C_m]L_{ij}, \quad W \mapsto \left(q\mu_0 U_T^2 C_m/{l^*}^2\right) W$$

to obtain the scaled system,

$$-\nabla\cdot J_1 = 0, \tag{2.65}$$
$$-\nabla\cdot J_2 = -J_1\cdot\nabla V + W(n,T), \tag{2.66}$$
$$J_1 = L_{11}\left(\nabla\left(\frac{\mu}{T}\right) - \frac{\nabla V}{T}\right) + L_{12}\nabla\left(-\frac{1}{T}\right), \tag{2.67}$$
$$J_2 = L_{21}\left(\nabla\left(\frac{\mu}{T}\right) - \frac{\nabla V}{T}\right) + L_{22}\nabla\left(-\frac{1}{T}\right), \tag{2.68}$$
$$\lambda^2\Delta V = n - \text{Dop}, \tag{2.69}$$

where $\lambda = \sqrt{\varepsilon_s U_T/(qC_m {l^*}^2)}$ is the scaled *Debye length.* For the density scaling $n_0$ the value $C_m$ is used here, since high injection effects are not to be expected in the simulations.
The space scale is also different to take into account the extensions of the device.
By introducing the *dual entropy variables*,

$$w_1 = \frac{\mu - V}{T}, \quad w_2 = -\frac{1}{T}$$

we attain to the following (scaled) stationary unipolar semiconductor device model for the electrostatic potential $V$ and the *quasi-Fermi levels* $w_1$ and $w_2$ in the semiconductor domain $\Omega$:

$$-\nabla\cdot I_1 = 0, \tag{2.70}$$
$$-\nabla\cdot I_2 = W(n, w_2), \tag{2.71}$$
$$I_1 = D_{11}(n,V,w_2)\nabla w_1 + D_{12}(n,V,w_2)\nabla w_2, \tag{2.72}$$
$$I_2 = D_{21}(n,V,w_2)\nabla w_1 + D_{22}(n,V,w_2)\nabla w_2, \tag{2.73}$$
$$\lambda^2\Delta V = n(V,w_1,w_2) - \text{Dop}, \tag{2.74}$$

where $n$ denotes the scaled electron density and $W$ the scaled energy relaxation term,

$$n = n_i(-w_2)^{-\frac{3}{2}}e^{w_1 - Vw_2} \quad \text{and} \quad W = \frac{3}{2}\frac{1+w_2^{-1}}{\tau_0}n. \tag{2.75}$$

The scaled diffusion coefficients $D_{ij}$ are given by

$$D_{11} = n,\ D_{12} = D_{21} = -\left(V + \frac{3}{2}w_2^{-1}\right)n,\ D_{22} = \left[\left(V + \frac{3}{2}w_2^{-1}\right)^2 + \frac{3}{2}w_2^{-2}\right]n. \tag{2.76}$$

Dop is the scaled doping profile function. In (2.75), the constant parameter $n_i$ is the scaled *intrinsic density* of the semiconductor material and $\tau_0$ denotes the scaled *energy relaxation time* (constant). If we set $T_D = T_L$ on all of $\Gamma_D$ in (2.60), then the following Dirichlet boundary conditions for $V$, $w_1$ and $w_2$ can be derived,

$$V = \log(n_D) + V_A, \quad w_1 = -\log(n_i) - V_A, \quad \text{and} \quad w_2 = -1 \quad \text{on} \quad \Gamma_D. \tag{2.77}$$

Below, we refer to the equations (2.70) through (2.73) as the current continuity system.
In this formulation, the Joule heating term $-J_1 \cdot \nabla V$ is eliminated and the problem is symmetrized. Hence, the resulting model does not contain a dominating convection term, which is one of its advantages.
The formulation of the ET model in dual entropy variables was the basis of analytical studies [Gri99, DGJ97] and was employed for numerical approximations in one space dimension in [MMP96].

# 3 The linearization and weak relation

A common technique to solve the nonlinear system (2.70)-(2.74) consists in its successive linearization in appropriate function spaces and the subsequent solution of the linear problems. Here we use the method of *freezing the coefficients*. The diffusion coefficients (2.76) of the current continuity system (2.70)-(2.73) depend on the electrostatic potential $V$. Thus the quasi-Fermi levels $w_1$ and $w_2$ also depend on $V$. First we establish an existence result for the (nonlinear) Poisson equation, where we impose boundedness restrictions on $w_1$ and $w_2$. These restrictions are needed, since the right hand side of the Poisson equation (2.74) depends on $w_1$ and $w_2$. Then a maximum principle for $V$ in terms of these bounds for $w_1$ and $w_2$ will follow. Solving the Poisson equation is considered as a fractional step in solving the whole system. Then this result is substituted in the coefficients of the current continuity system, which in turn is solved for new functions $w_1$ and $w_2$. In this way, one can define a system map for the entire system (2.70)-(2.74), which we assume is a fixed point map under additional conditions. The definition of this fixed point map is a preliminary step for the construction of iteration methods.
In the following subsection we focus on the Poisson equation and the related maximum principle.

## 3.1 Maximum principle for the Poisson equation

The following lemma along with its proof is a slight modification of the ideas presented in [Jer96, section 4.3].
We restrict the domain of $w_1$ and $w_2$ to be the convex subset $\mathcal{M}$ in $L^2(\Omega)\times L^2(\Omega)$, where

$$\mathcal{M} := \{(\varphi_1,\varphi_2)\in L^2(\Omega)\times L^2(\Omega)\ :\ \gamma_1\le\varphi_1\le\delta_1, 0<\gamma_2\le-\varphi_2\le\delta_2 \text{ a.e.}\}. \tag{3.1}$$

The bounds $\gamma_1,\gamma_2,\delta_1,\delta_2$ may depend on the minimum and maximum values of the doping profile Dop and the (fixed) boundary data $V_D, w_{1D}$ and $w_{2D}$. We take $(v_1,v_2)\in\mathcal{M}$, where $v_1$ and $v_2$ fulfill the boundary conditions for $w_1$ and $w_2$, respectively.

In the following, we assume that $\mathrm{Dop}\in L^2(\Omega)$ is bounded. The test function space in the following relation, denoted $Y_0 := H^1_{0,\Gamma_D}(\Omega)$, consists of $H^1$ functions with zero trace on $\Gamma_D$. The definition of the fractional step $\mathbf{U}[v_1,v_2]$ entails the solution of the problem below,

**Lemma 3.1.** *Given a pair $(v_1,v_2)\in\mathcal{M}$, the image under the map $\mathbf{U}$ is the unique element $V$ satisfying the weak relation,*

$$\lambda^2\int_\Omega \nabla V\cdot\nabla\varphi dx = \int_\Omega (n(V,v_1,v_2)-\mathrm{Dop})\varphi dx \quad \forall\,\varphi\in Y_0, \tag{3.2}$$

*subject to the boundary conditions*

$$V = V_D \text{ on } \Gamma_D, \quad \nu\cdot\nabla V = 0 \text{ on } \Gamma_N.$$

*Moreover, $V$ satisfies the maximum principle,*

$$V\ge\gamma \;=\; \min(\inf_{\Gamma_D} V_D,\gamma'), \tag{3.3}$$

$$V\le\delta \;=\; \max(\sup_{\Gamma_D} V_D,\delta'), \tag{3.4}$$

*where $\gamma'$ and $\delta'$ are uniquely defined by*

$$n_i\gamma_2^{-\frac{3}{2}}\exp(\delta_1+\delta_2\gamma') - \inf_\Omega \mathrm{Dop} \;=\; 0 \tag{3.5}$$

$$n_i\delta_2^{-\frac{3}{2}}\exp(\gamma_1+\gamma_2\delta') - \sup_\Omega \mathrm{Dop} \;=\; 0. \tag{3.6}$$

*Proof.* We first demonstrate the validity of the bounds (3.3), (3.4). Thus, if $V$ is a solution in $H^1(\Omega)$, with $\exp(V)$ in $L^2(\Omega)$, we select for the admissible test function, $\varphi = (V-\delta)^+$, where $t^+ = \max\{t, 0\}$. The restriction of the trace of $\varphi$ to $\Gamma_D$ is zero, hence $\varphi$ is admissible, since the positive part is sub-additive, and this property may be applied to $V - \delta = U_0 + (V_D - \delta)$, for some admissible $U_0$. Now the substitution of $\varphi$ into (3.2) reduces integrations to the set $\{V > \delta\}$. For the terms involving the gradients, this uses the chain rule for the composition of the positive part and an $H^1$ function [GT77]. Once this reduction is achieved, one uses the non-negativity of $n(V, v_1, v_2) - \text{Dop}$ on the set $\{V > \delta\}$, to conclude $\nabla(V-\delta)^+ = 0$, and hence $(V-\delta)^+ = 0$. Note that we have used the fact that the gradient of a $Y_0$ function $f$ cannot vanish without $f$ vanishing a.e. The non-negativity above follows from the inequalities,

$$
\begin{aligned}
&n_i(-v_2)^{-\frac{3}{2}} \exp(v_1 - Vv_2) - \text{Dop} \geq n_i\delta_2^{-\frac{3}{2}} \exp(v_1 - Vv_2) - \text{Dop} \geq \\
&n_i\delta_2^{-\frac{3}{2}} \exp(\gamma_1 - Vv_2) - \text{Dop} \geq n_i\delta_2^{-\frac{3}{2}} \exp(\gamma_1 + \delta_2\delta) - \text{Dop} \geq \\
&n_i\delta_2^{-\frac{3}{2}} \exp(\gamma_1 + \delta_2\delta') - \sup_{\Omega} \text{Dop} = 0.
\end{aligned}
$$

Note that we have used the definitions of $\gamma_1, \delta$ and $\delta'$ above. In a similar way, one shows that $(V-\gamma)^- = \varphi$ is the zero function, where $t^- = -(-t)^+$. The use of $\varphi$ as a test function, coupled with the inequality, $n_i(-v_2)^{-\frac{3}{2}} \exp(w_1 - Vw_2) - \text{Dop} \leq 0$ on $\{V \leq \gamma\}$, leads to this result. If we combine these to observations, we obtain the inequalities, (3.3), (3.4).

□

## 3.2 The linear current continuity system

We insert $(v_1, v_2) \in \mathcal{M}$ and $V = \boldsymbol{U}[v_1, v_2]$ into the diffusion coefficients and the energy relaxation term. Under the above assumptions and by the preceeding lemma, $\mathbf{D}(x) = (D_{ij}(x))_{i,j} \in [L^\infty(\Omega)]^{2\times 2}$ is a bounded, symmetric and uniformly positive definite matrix-valued function. Furthermore, it holds $W \in L^2(\Omega)$. Now we give a weak relation of the current continuity system (2.70)-(2.73). The current densities are considered the primal unknowns here, since in our application $I_1$ and $I_2$ rather than the potentials $w_1$ and $w_2$ are the interesting variables. They are treated as independent variables and hence can be approximated more accurately. This necessitates the introduction of a new function space for $I_1$ and $I_2$. We seek $I_1, I_2 \in H(\text{div}; \Omega)$, where

$$
H(\text{div};\Omega) = \left\{\tau = (\tau^1, \tau^2, \tau^3) \in [L^2(\Omega)]^3 : \text{div}\,\tau \in L^2(\Omega)\right\}, \quad \|\tau\|^2_{\text{div}} = \sum_{k=1}^{3} \|\tau^k\|^2_{0,2,\Omega} + \|\text{div}\,\tau\|^2_{0,2,\Omega}.
$$

From the surjectivity of the divergence operator $\text{div} : H(\text{div};\Omega) \to L^2(\Omega)$ on domains with smooth boundary (for instance Lipschitzian, see, e.g., [BF91]), it follows that (2.70) and (2.71) have solutions $I_1, I_2$ in $H(\text{div};\Omega)$. It can be shown (see, e.g.,[Tem77]), by the methods of Lions and Magenes [LM68], that vectors of $H(\text{div};\Omega)$ admit a well defined *normal trace* on $\Gamma = \partial\Omega$. For $\tau \in H(\text{div};\Omega)$ the normal trace $\nu \cdot \tau$ lies in $H^{-1/2}(\Gamma)$, and Green's formula holds on $\Omega$ [BF91],

$$
\int_\Omega \nabla \cdot \tau v \, dx + \int_\Omega \tau \cdot \nabla v \, dx = \langle \nu \cdot \tau, v \rangle, \quad \forall v \in H^1(\Omega). \tag{3.7}
$$

Here, the angular brackets denote the duality between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$. The trace operator implicitly defined above also satisfies a surjectivity property, i.e.

$$
\gamma_n : \tau \in H(\text{div};\Omega) \to \nu \cdot \tau_{|\Gamma} \in H^{-1/2}(\Gamma) \tag{3.8}
$$

is surjective. The proof is given in [BF91].

**Remark 3.1.** The surjectivity of this trace operator is important in the case, where non-homogeneous Neumann boundary conditions are imposed. This situation occurs, for instance, in bipolar transistors, where a control current instead of a control voltage is applied. But we only consider homogeneous Neumann boundary conditions throughout the thesis.

**Remark 3.2.** The norm $\|\cdot\|_{\text{div}}$ is weaker than the standard norm on $[H^1(\Omega)]^3$. Therefore, we have the inclusion $[H^1(\Omega)]^3 \subset H(\text{div};\Omega)$; but still, the divergence operator div : $[H^1(\Omega)]^3 \to L^2(\Omega)$ is surjective (for smooth boundary see [BS02], for polygonal boundary see [GR86]).

In order to impose the boundary conditions on $I_1$ and $I_2$, we have to restrict the trace operator $\gamma_n$ to $\Gamma_N$. We define the spaces

$$H^1_{0,\Gamma_D}(\Omega) = \{\varphi \in H^1(\Omega) : \varphi_{|\Gamma_D} = 0\}$$

and

$$H_{0,\Gamma_N} = \{\tau \in H(\text{div};\Omega) : \langle \nu \cdot \tau, \varphi\rangle = 0,\ \forall \varphi \in H^1_{0,\Gamma_D}(\Omega)\}. \tag{3.9}$$

The space $H_{0,\Gamma_N}$ contains functions of $H(\text{div};\Omega)$ whose normal traces vanish on $\Gamma_N$. For reasons related to pathological properties of $H^{1/2}(\Gamma_D)$ and $H^{-1/2}(\Gamma_N)$, it is necessary to use definition (3.9) and not an expression such as $\nu \cdot \tau_{|\Gamma_N} = 0$ in $H^{-1/2}(\Gamma_N)$.
The current densities are actually sought in the subspace $H_{0,\Gamma_N}$, and we formally write

$$\nu \cdot I_1 = 0 \quad \text{and} \quad \nu \cdot I_2 = 0 \quad \text{on } \Gamma_N.$$

We seek the quasi-Fermi levels $w_1$ and $w_2$ in $L^2(\Omega)$. This means that for $w_1$ and $w_2$ less smoothness is required than in the standard weak relation. Moreover, we assume that the corresponding boundary data satisfy $w_{1D}, w_{2D} \in H^{1/2}(\Gamma_D)$.
We define $\mathbf{A}(x) = (\mathbf{D}(x))^{-1}$ almost everywhere in $\Omega$. The entries of $\mathbf{A}$ are given by

$$A_{11} = \frac{2}{3}v_2^2 n^{-2} D_{22}, \quad A_{12} = A_{21} = -\frac{2}{3}v_2^2 n^{-2} D_{21}, \quad A_{22} = \frac{2}{3}v_2^2 n^{-2} D_{11}, \tag{3.10}$$

where we have used the abbreviations $n = n(V, v_1, v_2)$ and $D_{ij} = D_{ij}(V, v_1, v_2)$. The definition of the current densities can be reformulated as

$$\begin{aligned} \nabla w_1 &= A_{11}(x) I_1 + A_{12}(x) I_2, &(3.11)\\ \nabla w_2 &= A_{21}(x) I_1 + A_{22}(x) I_2. &(3.12) \end{aligned}$$

The subsequent analysis is concerned with the solution of the *linear* problem,

$$\begin{aligned} -\nabla\cdot I_1 &= 0, &(3.13)\\ -\nabla\cdot I_2 &= W(x), &(3.14)\\ \nabla w_1 &= A_{11}(x) I_1 + A_{12}(x) I_2, &(3.15)\\ \nabla w_2 &= A_{21}(x) I_1 + A_{22}(x) I_2. &(3.16) \end{aligned}$$

Now, we state an abstract result [GR94]. Later, we will identify the abstract quantities. Let $\mathcal{V}, \mathcal{W}$ be two real-valued Hilbert spaces endowed with the inner products, $(\cdot,\cdot)_\mathcal{V}, (\cdot,\cdot)_\mathcal{W}$ and the corresponding norms $\|\cdot\|_\mathcal{V}, \|\cdot\|_\mathcal{W}$. Let $a(\cdot,\cdot) : \mathcal{V}\times\mathcal{V} \to \mathbb{R}$ and $b(\cdot,\cdot) : \mathcal{V}\times\mathcal{W} \to \mathbb{R}$ given continuous bilinear forms, i.e. there are constants $\alpha, \beta > 0$ such that

$$\begin{aligned} |a(u,v)| &\le \alpha \|u\|_\mathcal{V} \|v\|_\mathcal{V} \quad \forall\, u, v \in \mathcal{V},\\ |b(u,v)| &\le \beta \|v\|_\mathcal{V} \|w\|_\mathcal{W} \quad \forall\, v \in \mathcal{V},\ \forall\, w \in \mathcal{W}. \end{aligned}$$

Define

$$Z := \{v \in \mathcal{V} \ :\ b(u,w) = 0 \quad \forall\, w \in \mathcal{W}\}.$$

$Z$ is a closed subspace of $\mathcal{V}$ due to the continuity of $b(\cdot,\cdot)$. Hence, $(Z,(\cdot,\cdot)_{\mathcal{V}})$ is a Hilbert space, too. Apart from that we assume that $a(\cdot,\cdot)$ is coercive with respect to $Z$, i.e. there is a constant $\gamma > 0$ such that

$$\gamma \|z\|_{\mathcal{V}}^2 \le a(z,z) \quad \forall\, z \in Z.$$

To be complete, we take linear, continuous functionals $f \in \mathcal{V}^*, g \in \mathcal{W}^*$ and define

$$G := \{v \in \mathcal{V} \;:\; b(v,w) = g(w) \quad \forall\, w \in \mathcal{W}\}.$$

The abstract variational problem is then given by

$$a(u,z) = f(z) \quad \forall\, z \in Z. \tag{3.17}$$

Since we incorporated side conditions on $u$ into $G$, we call (3.17) a restricted variational equation. There is no unique relationship between $\mathcal{V}$ and $\mathcal{W}$, but it is assumed that there is a continuous operator between them that provides a link.
In the following lemma a solvability condition is given.

**Lemma 3.2.** *Let $G \neq \emptyset$. Then the restricted variational equation* (3.17) *has a unique solution $u \in G$, which satisfies the estimate,*

$$\|u\|_{\mathcal{V}} \le \frac{1}{\gamma}\|f\|_* + \left(\frac{\alpha}{\gamma} + 1\right)\|v\|_{\mathcal{V}} \quad \forall\, v \in G.$$

The proof is given in [GR94].
Now, we identify the elements in the lemma and state the corresponding weak relation of (3.13)-(3.16). To this end we define the space,

$$H_0(\mathrm{div};\Omega) := \{\tau \in H_{0,\Gamma_N}(\mathrm{div};\Omega) \;:\; \nabla\cdot\tau = 0\},$$

and set

$$\begin{aligned}
\mathcal{V} &= H_{0,\Gamma_N}(\mathrm{div};\Omega) \times H_{0,\Gamma_N}(\mathrm{div};\Omega), & \|(\tau_1,\tau_2)\|_{\mathrm{div}}^2 &= \|\tau_1\|_{\mathrm{div}}^2 + \|\tau_2\|_{\mathrm{div}}^2,\\
\mathcal{W} &= L^2(\Omega) \times L^2(\Omega), & \|(\varphi_1,\varphi_2)\|_{0,2,\Omega}^2 &= \|\varphi_1\|_{0,2,\Omega}^2 + \|\varphi_2\|_{0,2,\Omega}^2.
\end{aligned}$$

In [GR86] a proof is given for the statement $\ker(\gamma_n) = H_{0,\Gamma_N}(\mathrm{div};\Omega)$.
For the identifications, we define

$$\begin{aligned}
a_i(I_1,I_2,\tau) &:= \int_\Omega (A_{i1}(x)I_1 + A_{i2}(x)I_2)\cdot\tau dx \quad \forall\,\tau \in H_{0,\Gamma_N}(\mathrm{div};\Omega),\ \ i = 1,2,\\
a(I_1,I_2,\tau_1,\tau_2) &:= a_1(I_1,I_2,\tau_1) + a_2(I_1,I_2,\tau_2) \quad \forall\,\tau_1,\tau_2 \in H_{0,\Gamma_N}(\mathrm{div};\Omega),\\
f(\tau_1,\tau_2) &:= \langle \nu\cdot\tau_1, w_{1D}\rangle + \langle \nu\cdot\tau_2, w_{2D}\rangle \quad \forall\,\tau_1,\tau_2 \in H_{0,\Gamma_N}(\mathrm{div};\Omega),\\
b(I_1,I_2,\varphi_1,\varphi_2) &:= \int_\Omega (\nabla\cdot I_1\varphi_1 + \nabla\cdot I_2\varphi_2)dx \quad \forall\,\varphi_1,\varphi_2 \in L^2(\Omega),\\
g(\varphi) &:= \int_\Omega W(x)\varphi(x)dx \quad \forall\,\varphi \in L^2(\Omega),
\end{aligned}$$

**Remark 3.3.** We shall sometimes write formally, $\int_{\Gamma_D}(w_{1D}\nu\cdot\tau_1 + w_{2D}\nu\cdot\tau_2)\,ds$ instead of $\langle \nu\cdot\tau_1, w_{1D}\rangle + \langle \nu\cdot\tau_2, w_{2D}\rangle$ to denote duality between $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$.

We continue the identifications by defining the solution sets,

$$\begin{aligned}
G &:= \{(\tau_1,\tau_2) \in H_{0,\Gamma_N}(\mathrm{div};\Omega) \times H_{0,\Gamma_N}(\mathrm{div};\Omega) \;:\; -b(\tau_1,\tau_2,\varphi_1,\varphi_2) = g(\varphi_2)\ \forall\,\varphi_1,\varphi_2 \in L^2(\Omega)\},\\
Z &:= \{(\tau_1,\tau_2) \in H_{0,\Gamma_N}(\mathrm{div};\Omega) \times H_{0,\Gamma_N}(\mathrm{div};\Omega) \;:\; b(\tau_1,\tau_2,\varphi_1,\varphi_2) = 0 \quad \forall\,\varphi_1,\varphi_2 \in L^2(\Omega)\}.
\end{aligned}$$

It follows that $Z = H_0(\text{div};\Omega) \times H_0(\text{div};\Omega)$.
$a(\cdot,\cdot)$ and $f(\cdot)$ are defined by multiplying (3.11) and (3.12) by test functions $\tau_1, \tau_2 \in H_{0,\Gamma_N}(\text{div};\Omega)$, respectively, and subsequent integration. By adding these equations, we obtain $a(\cdot,\cdot)$ immediately from the right hand side. In order to obtain $f(\cdot)$, we apply Green's formula (3.7), which gives the following relation,

$$\int_\Omega (\tau_1 \cdot \nabla w_1 + \tau_2 \cdot \nabla w_2) dx = \int_{\partial\Omega} (w_1 \nu \cdot \tau_1 + w_2 \nu \cdot \tau_2) ds - \int_\Omega (w_1 \nabla\cdot \tau_1 + w_2 \nabla\cdot \tau_2) dx, \tag{3.18}$$

where $(\tau_1,\tau_2) \in Z$. The first term on the right hand side of (3.18) reduces to $f(\tau_1,\tau_2)$, since $\nu \cdot \tau = 0$ on $\Gamma_N$ for all $\tau \in H_{0,\Gamma_N}(\text{div};\Omega)$. The second term on the right hand side of (3.18) vanishes, since $\nabla\cdot \tau_1 = \nabla\cdot \tau_2 = 0 \quad \forall (\tau_1,\tau_2) \in Z$.
The bilinear form $a(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ is continuous,

$$\begin{aligned} |a(I_1, I_2, \tau_1, \tau_2)| &\leq \sum_{i,j=1}^{2} \|A_{ij}\|_{0,\infty,\Omega} \|(I_1, I_2)\|_{\text{div}} \|(\tau_1,\tau_2)\|_{\text{div}} \\ &\leq 2\left(\|A_{11}\|_{0,\infty,\Omega} + \|A_{22}\|_{0,\infty,\Omega}\right) \|(I_1, I_2)\|_{\text{div}} \|(\tau_1,\tau_2)\|_{\text{div}}, \end{aligned} \tag{3.19}$$

and coercive on $Z$, since $\boldsymbol{A}(x)$ is positive definite almost everywhere in $\Omega$,

$$\forall (\xi_1,\xi_2)^T \in \mathbb{R}^2 : \sum_{i,j=1}^{2} A_{ij}(x)\, \xi_i \xi_j \geq \frac{\xi_1^2 + \xi_2^2}{D_{11}(x) + D_{22}(x)}, \tag{3.20}$$

and integration over $\Omega$ leads to

$$a(\tau_1,\tau_2,\tau_1,\tau_2) \geq \sum_{k=1}^{3} \frac{\|\tau_1^k\|_{0,2,\Omega}^2 + \|\tau_2^k\|_{0,2,\Omega}^2}{\|D_{11}\|_{0,\infty,\Omega} + \|D_{22}\|_{0,\infty,\Omega}} \quad \forall \tau_1,\tau_2 \in H_{0,\Gamma_N}(\text{div};\Omega). \tag{3.21}$$

The estimate (3.20) immediately follows from the definition of the inverse matrix coefficients (3.10) by computation of the eigenvalues of $\mathbf{A}(x)$. The diffusion coefficients $D_{11}(x)$ and $D_{22}(x)$ in (2.76) are positive functions and the bounds for $V$ in (3.3) and (3.4) are independent of $x \in \Omega$. Obviously, the bilinear form $b(\cdot,\cdot)$ is continuous on $\mathcal{V} \times \mathcal{W}$. $f$ is a continuous linear functional on $H_{0,\Gamma_N}(\text{div};\Omega) \times H_{0,\Gamma_N}(\text{div};\Omega)$, provided that $w_{1D}, w_{2D} \in H^{1/2}(\Gamma_D)$, since the trace operator 3.8 is continuous, and it holds

$$\|f\|_* \leq C_0 \left(\|w_{1D}\|_{1/2,2,\Gamma_D} + \|w_{2D}\|_{1/2,2,\Gamma_D}\right) \leq C_0 \|(w_{1D}, w_{2D})\|_{1/2,2,\Gamma_D}, \tag{3.22}$$

where we use the abbreviation

$$\|(w_{1D}, w_{2D})\|_{1/2,2,\Gamma_D} = \left(\|w_{1D}\|_{1/2,2,\Gamma_D}^2 + \|w_{2D}\|_{1/2,2,\Gamma_D}^2\right)^{\frac{1}{2}}.$$

The norm $\|\cdot\|_*$ is the dual norm on $(H_{0,\Gamma_N}(\text{div};\Omega))' \times (H_{0,\Gamma_N}(\text{div};\Omega))'$. $g$ is a continuous linear functional on $L^2(\Omega)$,

$$\|g\|_* \leq \|W\|_{0,2,\Omega} \leq \frac{3}{2\tau_0} \|n v_2^{-1}(1 + v_2^{-1})\|_{0,\infty,\Omega} \|(v_1, v_2)\|_{0,2,\Omega}. \tag{3.23}$$

Recall the definition of the energy relaxation term $W$ in (2.75). The norm $\|\cdot\|_*$ is the dual norm on $(L^2(\Omega))'$.

*(restricted weak formulation)* Find $(I_1, I_2) \in G$ such that

$$a(I_1, I_2, \tau_1, \tau_2) = f(\tau_1,\tau_2) \quad \forall (\tau_1,\tau_2) \in Z. \tag{3.24}$$

Note that the boundary conditions for $I_1, I_2$ are incorporated into $G$.

**Remark 3.4.** The weak formulation (3.24) is a Dirichlet principle for the minimization problem

$$J(I_1, I_2) = \frac{1}{2} a(I_1, I_2, I_1, I_2) - f(I_1, I_2) \to \min!,\ (I_1, I_2) \in G.$$

Since $G \neq \emptyset$, from the abstract lemma 3.2 it follows that the problem (3.24) has a unique solution and the following estimate holds,

$$\begin{aligned} \|(I_1, I_2)\|_{\mathrm{div}} \leq\ & C_0 \|(D_{11}, D_{22})\|_{0,\infty,\Omega} \|(w_{1D}, w_{2D})\|_{1/2,2,\Gamma_D} + \\ & (1 + 2\|(D_{11}, D_{22})\|_{0,\infty,\Omega} \|(A_{11}, A_{22})\|_{0,\infty,\Omega})\, \|(\tau_1, \tau_2)\|_{\mathrm{div}} \end{aligned} \tag{3.25}$$

for all $(\tau_1, \tau_2) \in G,$ where we use the abbreviation,

$$\|(D_{11}, D_{22})\|_{0,\infty,\Omega} = \|D_{11}\|_{0,\infty,\Omega} + \|D_{22}\|_{0,\infty,\Omega}. \tag{3.26}$$

An advantage of the formulation (3.24) consists in the close relation to the minimization of a convex functional. But it is restrictive in view of the underlying sets $G$ and $Z$ that include side relations. We assign to (3.24) an extended weak formulation with no side relations.

*(weak formulation)* Find $I_1, I_2 \in H_{0,\Gamma_N}(\mathrm{div};\Omega)$ and $w_1, w_2 \in L^2(\Omega)$ such that

$$a(I_1, I_2, \tau_1, \tau_2) + b(\tau_1, \tau_2, w_1, w_2) = f(\tau_1, \tau_2), \tag{3.27}$$

$$-b(I_1, I_2, \varphi_1, \varphi_2) = g(\varphi_2) \tag{3.28}$$

for all $\tau_1, \tau_2 \in H_{0,\Gamma_N}(\mathrm{div};\Omega)$ and $\varphi_1, \varphi_2 \in L^2(\Omega)$.

**Remark 3.5.** The system (3.27),(3.28) is generally indefinite. The Lagrange functional associated with (3.27), (3.28) is given by

$$L(I_1, I_2, w_1, w_2) = J(I_1, I_2) + b(I_1, I_2, w_1, w_2) - g(w_2) \quad \forall\, I_1, I_2 \in H(\mathrm{div};\Omega), \forall\, w_1, w_2 \in L^2(\Omega).$$

**Lemma 3.3.** *If* $(I_1, I_2) \in H_{0,\Gamma_N}(\mathrm{div};\Omega) \times H_{0,\Gamma_N}(\mathrm{div};\Omega),\ w_1, w_2 \in L^2(\Omega)$ *is a solution to the weak variational equation* (3.27)-(3.28)*, then* $(I_1, I_2)$ *is a solution to* (3.24).

In [GR94] a proof is given.
Hence, it is left to show that the system (3.27), (3.28) has a unique solution $I_1, I_2 \in H_{0,\Gamma_N}(\mathrm{div};\Omega)$ and $w_1, w_2 \in L^2(\Omega)$. The well-posedness of this problem follows from a new kind of coercivity [BS02]. But this is again a consequence of the surjectivity of the divergence operator:

$$\exists c_0 > 0\ :\ \sup_{(\tau_1,\tau_2)\in\mathcal{V}} \frac{b(\tau_1, \tau_2, \varphi_1, \varphi_2)}{\|(\tau_1, \tau_2)\|_{\mathrm{div}}} \geq c_0 \|(\varphi_1, \varphi_2)\|_{0,2,\Omega} \quad \forall\, (\varphi_1, \varphi_2) \in L^2(\Omega) \times L^2(\Omega).$$

This follows from [BS02, lemma 11.2.3] or [GR86].
The unique solution to the system (3.27),(3.28) can be estimated in terms of the data and (3.22), (3.23), as well as (3.2),

$$\begin{aligned} \|(I_1, I_2)\|_{\mathrm{div}} \leq\ & C_0\, \|(D_{11}, D_{22})\|_{0,\infty,\Omega} \|(w_{1D}, w_{2D})\|_{0,2,\Gamma_D} + \\ & c_0^{-1}\, (1 + 2\|(D_{11}, D_{22})\|_{0,\infty,\Omega} \|(A_{11}, A_{22})\|_{0,\infty,\Omega})\, \|W\|_{0,2,\Omega}, \end{aligned} \tag{3.29}$$

$$\begin{aligned} \|(w_1, w_2)\|_{0,2,\Omega} \leq\ & c_0^{-1}\, (1 + 2\|(D_{11}, D_{22})\|_{0,\infty,\Omega} \|(A_{11}, A_{22})\|_{0,\infty,\Omega})\, \|(w_{1D}, w_{2D})\|_{0,2,\Gamma_D} + \\ & c_0^{-2}\, (1 + 2\|(D_{11}, D_{22})\|_{0,\infty,\Omega} \|(A_{11}, A_{22})\|_{0,\infty,\Omega})\, \|(A_{11}, A_{22})\|_{0,\infty,\Omega} \|W\|_{0,2,\Omega} \end{aligned} \tag{3.30}$$

Finally, to the equations (2.70)-(2.74) the continuous system map

$$\Phi\ :\ \mathcal{M} \to L^2(\Omega) \times L^2(\Omega), \qquad \Phi(v_1, v_2) := (w_1, w_2), \tag{3.31}$$

is associated. The continuity of $\Phi$ follows from the continuity of $\mathbf{U}\ :\ (v_1, v_2) \to V$ and the estimate (3.30), since the diffusion coefficients and the energy relaxation term are smooth functions of $V, v_1, v_2$. In section 5.3, $\Phi$ will be used to define a fixed point map that is the basis for the construction of iteration methods that find approximate solutions to the system (2.70)-(2.74).

# 4 Numerical approximation

## 4.1 Mixed finite-element discretization

In the current continuity system (3.27)-(3.28) no convection dominant term and no zero order term is present. Hence, we can apply lowest order Raviart-Thomas finite elements for the discretization of the current continuity system [RT77]. Let $\{\mathcal{T}_h\}_{h\in(0,h_0]}$, $h_0 > 0$, be a family of simplicial finite element partitions of $\Omega$ and $K \in \mathcal{T}_h$ for some fixed $h$. The following vector-valued approximation space is used to discretize the current densities $I_1, I_2$ locally. Define $\tau_i = e_i$, $i = 1, 2, 3$, where $\{e_i\}_{i=1}^3$ is the canonical basis of $\mathbb{R}^3$ and $\tau_0 = x - x_B$, where $x_B \in \mathbb{R}^3$ denotes the barycentre of $K$. Then the local space is defined as

$$\Sigma(K) = \text{span}\{\tau_0, \tau_1, \tau_2, \tau_3\}.$$

The space of Raviart-Thomas finite elements of lowest order is defined as

$$RT_0(\mathcal{T}_h) = \{\tau_h \in [L^2(\Omega)]^3 \ : \ \tau_h|_K \in \Sigma(K)\ \forall\, K \in \mathcal{T}_h,\ \nu \cdot \tau_h \text{ continuous on } e,\ \forall\, e \in \mathcal{E}_h'\}.$$

Here, $\nu$ is the unit outward normal on $\partial K$ and $\mathcal{E}_h'$ is the set of all inner faces of the partition elements $K$. The discretization by means of the Raviart-Thomas finite elements is a conforming method, i.e. $RT_0(\mathcal{T}_h) \subset H(\text{div}; \Omega)$ (see, e.g. [GR94, lemma 4.3]). Now we derive a characterization of the continuity requirement. Let $K, K' \in \mathcal{T}_h$ be neighboring triangles, i.e. $K \cap K' = e \in \mathcal{E}'_h$. Since $\tau_{h|K} \in \mathcal{P}^1(K)\ \forall\, K \in \mathcal{T}_h$, the following limits exist on $e$,

$$\forall\, x \in e \ : \ \nu \cdot \tau_h^K(x) := \lim_{t\to 0} \nu \cdot \tau_h(x - t\nu), \quad \nu \cdot \tau_h^{K'}(x) := \lim_{t\to 0} \nu \cdot \tau_h(x + t\nu),$$

and the continuity of the normal component of $\tau_h$ requires $\nu \cdot \tau_h^K(x) - \nu \cdot \tau_h^{K'}(x) = 0\ \forall\, x \in e$. This can also be written as

$$\int_e [\nu \cdot \tau_h^K(x) - \nu \cdot \tau_h^{K'}(x)] ds = 0.$$

Now we give an equivalent representation of the continuity in terms of a weak relation. By $\mathcal{E}_h$ the set of all faces of $K \in \mathcal{T}_h$ are denoted and by $\mathcal{E}(\Gamma_D)$ the faces lying on the Dirichlet boundary. Here we assume that $\Gamma_D$ consists only of entire faces of $K \in \mathcal{T}_h$. We introduce the space of functions that are constant on the faces,

$$\Lambda_{h,\xi} \ = \ \left\{\mu \in L^2(\mathcal{E}_h) : \mu_{|e} = const.\ \forall\, e \in \mathcal{E}_h,\ \int_e (\mu - \xi) ds = 0\ \ \forall\, e \in \mathcal{E}(\Gamma_D)\right\}, \quad \xi \in L^2(\Gamma_D).$$

Then the continuity on the inner faces is characterized by

$$\sum_{K\in\mathcal{T}_h} \sum_{e\in\mathcal{E}(K)} \int_e \nu \cdot \tau_h \mu ds = 0 \quad \forall\, \mu \in \Lambda_{h,0}.$$

The continuity requirement for $I_{1h}, I_{2h}$ is taken into account by use of an additional bilinear form,

$$c(I_{1h}, I_{2h}, \mu_{1h}, \mu_{2h}) = \sum_{K\in\mathcal{T}_h} \sum_{e\in\mathcal{E}(K)} \int_e (\mu_{1h} I_{1h} \cdot \nu + \mu_{2h} I_{2h} \cdot \nu) ds, \quad \mu_{1h}, \mu_{2h} \in \Lambda_{h,0}.$$

We define the space

$$X_h := \{\tau_h \in [L^2(\Omega)]^3 \ : \ \tau_h|_K \in \Sigma(K)\ \forall\, K \in \mathcal{T}_h\},$$

and

$$Z_h := \{(\tau_{1h}, \tau_{2h}) \in \ X_h \times X_h \ : \ c(\tau_{1h}, \tau_{2h}, \mu_{1h}, \mu_{2h}) = 0\ \ \forall\, \mu_{1h}, \mu_{2h} \in \Lambda_{h,0}\}.$$

Obviously, it holds $Z_h = RT_0(\mathcal{T}_h) \times RT_0(\mathcal{T}_h)$. The quasi-Fermi levels $w_1, w_2 \in L^2(\Omega)$ are discretized using the space

$$Y_h = \left\{\varphi \in L^2(\Omega) : \varphi_{|K} = const. \,\forall\, K \in \mathcal{T}_h\right\} \subset L^2(\Omega).$$

Before we state the weak discrete relation, we define a discrete bilinear form $a_h(\cdot,\cdot)$ and discrete linear forms $f_h(\cdot)$, $g_h(\cdot)$. We make use of the assumption on the regularity of the solution that ensures the consistency of these discrete forms. We assume that for all $K \in \mathcal{T}_h$ we have $V, w_1, w_2 \in W^{1,\infty}(K) \cap L^2(K)$ and that for all faces $e$ in $\Gamma_D$ we have $w_{1D}, w_{2D} \in H^{1/2}(e)$. Since $\gamma_1 \leq w_1 \leq \delta_1$ and $0 < \gamma_2 \leq -w_2 \leq \delta_2$ a.e. (recall the definition of the set $\mathcal{M}$ in 3.1), and $V$ fulfills the maximum principles (3.3), (3.4) it follows that $n(V, w_1, w_2) \in W^{1,\infty}(K) \cap L^2(K)$. With the same argument we show that $A_{ij}(n, w_1, w_2) \in W^{1,\infty}(K)$. Furthermore we have $W(n, w_1, w_2) \in L^2(K)$.
We introduce an approximation operator as follows (see.,e.g. [Cia91]). For arbitrary $K \in \mathcal{T}_h$, let

$$\Pi_K : W^{1,\infty}(K) \to \mathcal{P}_0(K) \subset L^\infty(K) \tag{4.1}$$

be a continuous approximation operator preserving constants,

1. $\forall\, v \in W^{1,\infty}(K) \;:\; \|\Pi_K v\|_{0,\infty,K} \leq C_\Pi \|v\|_{1,\infty,K}$,
2. $\forall\, v \in W^{1,\infty}(K) \;:\; \|(Id - \Pi_K)v\|_{0,\infty,K} \leq C_\Pi h_K \|v\|_{1,\infty,K}$,

where the constant $C_\Pi$ does not depend on $K$. Then we set, for any $K \in \mathcal{T}_h$,

$$V_K := \Pi_K V_{|K}, \quad w_{1K} := \Pi_K w_{1|K}, \quad w_{2K} := \Pi_K w_{2|K}, \tag{4.2}$$

and

$$A_{ij,K} := A_{ij}(V_K, w_{1K}, w_{2K}), \quad W_K := W(V_K, w_{1K}, w_{2K}).$$

The discrete forms are defined by

$$a_h(I_{1h}, I_{2h}, \tau_{1h}, \tau_{2h}) := \sum_{K \in \mathcal{T}_h} \sum_{i,j=1}^{2} \int_K A_{ij,K} I_{ih} \cdot \tau_{jh} dx, \quad g_h(\varphi_h) := \sum_{K \in \mathcal{T}_h} \int_K W_K \varphi_h dx,$$

for all $\tau_{1h}, \tau_{2h} \in X_h$ and $\varphi_h \in Y_h$. Moreover, let $\mathsf{w}_{1D}, \mathsf{w}_{2D}$ be the piecewise constant approximations of $w_{1D}, w_{2D}$ on $\Gamma_D$ that follow from the definition of the spaces $\Lambda_{h,w_{1D}}, \Lambda_{h,w_{2D}}$, respectively. Then we define

$$f_h(\tau_{1h}, \tau_{2h}) := \int_{\Gamma_D} (\mathsf{w}_{1D}\nu \cdot \tau_{1h} + \mathsf{w}_{2D}\nu \cdot \tau_{2h}) ds.$$

With these definitions the discretization of (3.27)-(3.28) reads as

*(discrete weak formulation)* Find $I_{1h}, I_{2h} \in Z_h$ and $w_{1h}, w_{2h} \in Y_h$ such that

$$a_h(I_{1h}, I_{2h}, \tau_{1h}, \tau_{2h}) + b(\tau_{1h}, \tau_{2h}, w_{1h}, w_{2h}) = f_h(\tau_{1h}, \tau_{2h}) \quad \forall\, \tau_{1h}, \tau_{2h} \in Z_h, \tag{4.3}$$

$$-b(I_{1h}, I_{2h}, \varphi_{1h}, \varphi_{2h}) = g_h(\varphi_{2h}) \quad \forall\, \varphi_{1h}, \varphi_{2h} \in Y_h. \tag{4.4}$$

Since $Z_h \neq \emptyset$ and $b(\cdot,\cdot)$ satisfies an inf-sup condition on $X_h \times Y_h$, this discrete variational equation has a unique solution. In [AB85] it is a proved that a discretization of a scalar elliptic equation applying the Raviart-Thomas finite elements fulfills this condition. This result can be extended to $b(\cdot,\cdot)$.
The $Z_h$-ellipticity of $a_h(\cdot,\cdot)$ immediately follows from the fact that $Z_h \subset Z$.

The discrete weak formulation (4.3), (4.4) leads to a linear system of the form

$$\begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \boldsymbol{0} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}^T & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{1h} \\ \boldsymbol{I}_{2h} \\ \boldsymbol{w}_{1h} \\ \boldsymbol{w}_{2h} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ -F \end{pmatrix}. \tag{4.5}$$

Here, $\boldsymbol{I}_{1h}$ and $\boldsymbol{I}_{2h}$ are the nodal values of the current densities pertaining to the local approximation space $\Sigma(K)$. $\boldsymbol{w}_{1h}$ and $\boldsymbol{w}_{2h}$ are constant on $K$. The matrices $\boldsymbol{A}_{ij}, \boldsymbol{B}$ and the vector $F$ correspond to

$$\int_K \tau_i \cdot \tau_j \, dx, \quad \int_K \nabla \cdot \tau_i \varphi_j \, dx, \quad \int_K W_K \varphi_i \, dx,$$

respectively, where $\tau_i$, $\varphi_i$ and $\mu_i$ are the basis functions of the respective spaces.

**Remark 4.1.** There are alternative numerical schemes that ensure the continuity of the normal component of the current densities. Since (2.70) and (2.71) are a conservation principle and a balance equation, resp., a natural approach is to construct numerical methods based on the divergence theorem, e.g. *finite volume methods*. They are especially useful on unstructured grids. A mixed finite volume method for semiconductor device simulation can be found in [SS97].

**Remark 4.2.** Note that the bilinear form $a_h(\cdot,\cdot)$ and the linear functionals $f_h$, $g_h$ depend on the discrete functions $(V_h, v_{1h}, v_{2h})$ and so they depend on the mesh parameter $h$.

The choice of a numerical method to solve the system (4.5) is restricted by the fact that it is indefinite. Although one can arrive at a positive definite system by elimination of $\boldsymbol{I}_{1h}$ and $\boldsymbol{I}_{2h}$ from the equations, this requires inversion of $(\boldsymbol{A}_{ij})$ and maintaining $(\boldsymbol{A}_{ij})^{-1}$ which is typically a full matrix. However, for the Raviart-Thomas finite elements this drawback can be circumvented by an implementational technique which leads to a positive definite system [dV65].

## 4.2 Hybridized mixed finite elements

The technique proposed in [dV65] applies, essentially, whenever the finite elements approximating the current densities $\boldsymbol{I}_{1h}$ and $\boldsymbol{I}_{2h}$ are subject to continuity constraints only at points interior to the inter-element boundaries, but *not* at element vertices. Then one may eliminate the continuity requirement of the normal component from the definition of the finite element space $RT_0(\mathcal{T}_h)$, and the partition elements are no longer coupled to each other. In order to achieve this, we extend the discrete mixed formulation (4.3) again. We attain to a non-conforming method, where the discrete current densities are sought in the space $X_h$. We call this a *hybridization procedure* because this discretization involves two different spaces for $w_1$ and $w_2$, namely $Y_h$ and $\Lambda_{h,\xi}$. The continuity constraint is then enforced through Lagrange multipliers defined on the inter-element boundaries. They enter the discrete system as additional unknowns sought in the space $\Lambda_{h,\xi}$. The resulting weak formulation is given by

*(hybridized discrete weak formulation)*

Find $I_{1h}, I_{2h} \in X_h$, $w_{1h}, w_{2h} \in Y_h$ and $\mathsf{w}_{1h}, \mathsf{w}_{2h} \in \Lambda_{h,0}$ such that

$$\begin{aligned} a_h(I_{1h}, I_{2h}, \tau_{1h}, \tau_{2h}) + b(\tau_{1h}, \tau_{2h}, w_{1h}, w_{2h}) + c(\tau_{1h}, \tau_{2h}, \mathsf{w}_{1h}, \mathsf{w}_{2h}) &= f_h(\tau_{1h}, \tau_{2h}), \\ -b(I_{1h}, I_{2h}, \varphi_{1h}, \varphi_{2h}) &= g_h(\varphi_{2h}), \\ -c(I_{1h}, I_{2h}, \mu_{1h}, \mu_{2h}) &= 0, \end{aligned}$$

$\forall\, \tau_{1h}, \tau_{2h} \in X_h$, $\forall\, \varphi_{1h}, \varphi_{2h} \in Y_h$ and $\forall\, \mu_{1h}, \mu_{2h} \in \Lambda_{h,0}$.

Written in more explicit terms the mixed hybrid formulation reads as,

*Find* $I_{1h}, I_{2h} \in X_h\,,\ w_{1h}, w_{2h} \in Y_h\,,\ \mathsf{w}_{1h} \in \Lambda_{h,\mathsf{w}_{1D}}\,,\ \mathsf{w}_{2h} \in \Lambda_{h,\mathsf{w}_{2D}}$ *such that*

$$\sum_{K\in\mathcal{T}_h}\int_K (A_{11,K}I_{1h} + A_{12,K}I_{2h})\cdot\tau_{1h}\,dx + \int_K w_{1h}\nabla\cdot\tau_{1h}\,dx - \sum_{e\in\mathcal{E}(K)}\int_e \mathsf{w}_{1h}\nu\cdot\tau_{1h}ds = 0, \tag{4.6}$$

$$\sum_{K\in\mathcal{T}_h}\int_K (A_{21,K}I_{1h} + A_{22,K}I_{2h})\cdot\tau_{2h}\,dx + \int_K w_{2h}\nabla\cdot\tau_{2h}\,dx - \sum_{e\in\mathcal{E}(K)}\int_e \mathsf{w}_{2h}\nu\cdot\tau_{2h}ds = 0, \tag{4.7}$$

$$\sum_{K\in\mathcal{T}_h}\int_K \nabla\cdot I_{1h}\varphi_{1h}dx = 0, \tag{4.8}$$

$$\sum_{K\in\mathcal{T}_h}\int_K \nabla\cdot I_{2h}\varphi_{2h}\,dx + \int_K W_K\varphi_{2h}\,dx = 0, \tag{4.9}$$

$$\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}(K)}\int_e \mu_{1h}I_{1h}\cdot\nu ds = 0, \tag{4.10}$$

$$\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}(K)}\int_e \mu_{2h}I_{2h}\cdot\nu ds = 0, \tag{4.11}$$

where $\tau_{1h}, \tau_{2h} \in X_h\,,\ \varphi_{1h}, \varphi_{2h} \in Y_h\,,\ \mu_{1h}, \mu_{2h} \in \Lambda_{h,0}$. The first and the second equations are weak discrete versions of (2.72) and (2.73). The third and fourth equations correspond to the discrete version of (2.70) and (2.71). The last two equations impose the continuity requirement of the normal component of $I_1$ and $I_2$ at the inter-element boundaries. It follows from these equations that the mean values of the jumps $I_1\cdot\nu$ and $I_2\cdot\nu$ are equal to zero, since the test functions $\mu$ are constant on every edge. Hence, conservation of the flux densities in a weak sense is guaranteed. The linear system associated with the mixed-hybrid formulation (4.6)-(4.11) is given by

$$\begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \boldsymbol{B} & \mathbf{0} & -C & \mathbf{0} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \mathbf{0} & \boldsymbol{B} & \mathbf{0} & -C \\ \boldsymbol{B}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{C}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{1h} \\ \boldsymbol{I}_{2h} \\ \boldsymbol{w}_{1h} \\ \boldsymbol{w}_{2h} \\ \mathbf{w}_{1h} \\ \mathbf{w}_{2h} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ -F \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}. \tag{4.12}$$

Here, $\boldsymbol{I}_{1h}$ and $\boldsymbol{I}_{2h}$ are the nodal values of the current densities pertaining to the local approximation space $\Sigma(K)$. $\boldsymbol{w}_{1h}$ and $\boldsymbol{w}_{2h}$ are the nodal values associated with the constant basis function on $K$; $\mathbf{w}_{1h} = (w_{1i})$ and $\mathbf{w}_{2h} = (w_{2i})$ are the nodal values belonging to the Lagrange multipliers on the faces of the tetrahedron $K$. The matrices $\boldsymbol{A}_{ij}, \boldsymbol{B}, \boldsymbol{C}$ and the vector $F$ correspond to

$$\int_K \tau_i\cdot\tau_j\,dx,\quad \int_K \nabla\cdot\tau_i\varphi_j\,dx,\quad \int_{\partial K}\tau_i\,\mu_j\,ds,\quad \int_K W_K\varphi_i\,dx,$$

respectively, where $\tau_i$, $\varphi_i$ and $\mu_i$ are the basis functions of the respective spaces.

## 4.3 Discretization of the Poisson equation

It remains to discretize the Poisson equation (2.74). Again, the mixed hybrid finite element discretization is used. On $\Omega$, (2.74) is equivalent to the system,

$$\nabla\cdot E = n - \mathrm{Dop}, \tag{4.13}$$

$$E = \lambda^2\nabla V. \tag{4.14}$$

The variable $-E$ is the *electric displacement.* The mixed-hybrid formulation of this system of equations reads as

*Find* $E_h \in X_h$ , $V_h \in Y_h$ , $\mathsf{V}_h \in \Lambda_{h,V_D}$ *such that*

$$0 = \sum_{K\in\mathcal{T}_h} \left( \lambda^{-2} \int_K E_h \cdot \tau \, dx + \int_K V_h \nabla\cdot \tau \, dx - \sum_{e\in\mathcal{E}(K)} \int_e \mathsf{V}_h \nu \cdot \tau ds \right), \tag{4.15}$$

$$0 = \sum_{K\in\mathcal{T}_h} \left( \int_K \nabla\cdot E_h \varphi \, dx + \int_K [\mathrm{Dop} - n_K]\varphi \, dx \right), \tag{4.16}$$

$$0 = \sum_{K\in\mathcal{T}_h} \sum_{e\in\mathcal{E}(K)} \int_e \mu E_h \cdot \nu ds, \tag{4.17}$$

where $\tau \in X_h$ , $\varphi \in Y_h$ , $\mu \in \Lambda_{h,0}$ and $n_K := n(V_K, w_{1K}, w_{2K})$.
The linear system associated with the mixed-hybrid formulations (4.15)-(4.17) is

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} & -\boldsymbol{C} \\ \boldsymbol{B}^T & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{C}^T & \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} E_h \\ V_h \\ \mathsf{V}_h \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ -\widetilde{F} \\ \boldsymbol{0} \end{pmatrix}, \tag{4.18}$$

where $E_h$ , $V_h$ and $\mathsf{V}_h$ also denote the vectors of nodal values. The matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$ and the vector $\widetilde{F}$ correspond to

$$\lambda^{-2} \int_K \tau_i \cdot \tau_j \, dx, \quad \int_K \nabla\cdot \tau_i \varphi_j \, dx, \quad \int_{\partial K} \tau_i \, \mu_j \, ds, \quad \int_K [\mathrm{Dop} - n_K]\varphi_i \, dx,$$

respectively, where $\tau_i$ , $\varphi_i$ and $\mu_i$ are the basis functions of the respective spaces. The solution $E_h$ and $V_h$ can be expressed in terms of the Lagrange multiplier $\mathsf{V}_h$ [CG04]. To this end, two mappings are introduced. They are defined using the equations (4.15) and (4.16).
One of these mappings lifts functions on edges of the triangulation $\mathcal{T}_h$ to functions on $\Omega$. Let $\mathcal{E}_h$ be the set of edges of the triangulation $\mathcal{T}_h$. We shall denote the set of all square integrable functions on the union of all edges of $\mathcal{E}_h$ by $L^2(\mathcal{E}_h)$. The lifting associates to each $\mathsf{m} \in L^2(\mathcal{E}_h)$, the pair of functions $(E_h, V_h)_\mathsf{m} \equiv (E_{h,\mathsf{m}}, V_{h,\mathsf{m}}) \in X_h \times Y_h$ defined by requiring that

$$\sum_{K\in\mathcal{T}_h} \left( \lambda^{-2} \int_K E_{h,\mathsf{m}} \cdot \tau \, dx + \int_K V_{h,\mathsf{m}} \nabla\cdot \tau \, dx \right) = \sum_{K\in\mathcal{T}_h} \sum_{e\in\mathcal{E}(K)} \int_e \mathsf{m}\nu \cdot \tau ds, \tag{4.19}$$

$$\sum_{K\in\mathcal{T}_h} \int_K \nabla\cdot E_{h,\mathsf{m}} \varphi \, dx = 0, \tag{4.20}$$

hold for all $(\tau, \varphi) \in X_h \times Y_h$. The other mapping associates to the function $n - \mathrm{Dop} \in L^2(\Omega)$ the element $(E_h, V_h)_n \equiv (E_{h,n}, V_{h,n}) \in X_h \times Y_h$ and is defined by requiring that

$$\sum_{K\in\mathcal{T}_h} \left( \lambda^{-2} \int_K E_{h,n} \cdot \tau \, dx + \int_K V_{h,n} \nabla\cdot \tau \, dx \right) = 0, \tag{4.21}$$

$$\sum_{K\in\mathcal{T}_h} \int_K \nabla\cdot E_{h,n} \varphi \, dx = \int_K [n_K - \mathrm{Dop}]\varphi \, dx, \tag{4.22}$$

hold for all $(\tau, \varphi) \in X_h \times Y_h$.
These mappings can be computed in an element-by-element fashion. Indeed, they are uniquely defined on each element because of the surjectivity of the map $(\nabla\cdot) : X_h \to Y_h$ restricted to an element. Moreover, on each element $K \in \mathcal{T}_h$, the lifting $(E_h, V_h)_\mathsf{m}$ can be thought of as the result of a one element discretization of the following boundary value problem

$$\begin{aligned} \nabla\cdot E_\mathsf{m} &= 0 && \text{in } K, \\ \lambda^{-2} E_\mathsf{m} &= \nabla V_\mathsf{m} && \text{in } K, \\ V_\mathsf{m} &= \mathsf{m} && \text{on } \partial K, \end{aligned}$$

and that the mapping $(E_h, V_h)_n$ is an approximation to the solution of

$$\begin{aligned} \nabla\cdot E_n &= n - \mathrm{Dop} && \text{in } K, \\ \lambda^{-2} E_n &= \nabla V_n && \text{in } K, \\ V_n &= 0 && \text{on } \partial K. \end{aligned}$$

Before stating the characterization of the solution, the following convention is introduced: the extension by zero of the function $\eta \in L^2(\mathcal{F}_h)$, where $\mathcal{F}_h$ is a subset of $\mathcal{E}_h$, to $\mathcal{E}_h$ is also denoted by $\eta$. In this way, if $\mathsf{m} = \mathsf{V}_h$ on $\mathcal{E}'_h$ and $\mathsf{m} = \mathsf{V}_D$ on $\Gamma_D$, with $\mathsf{V}_D$ being the piecewise constant approximation of $V_D$ we simply write $\mathsf{m} = \mathsf{V}_h + \mathsf{V}_D$, and as a consequence we also write

$$(E_h, V_h)_{\mathsf{m}} = (E_h, V_h)_{\mathsf{V}_h} + (E_h, V_h)_{\mathsf{V}_D}.$$

Let $(E_h, V_h, \mathsf{w}_h)$ be the solution of the hybridized Raviart-Thomas method (4.15), (4.16) and (4.17). Then

$$(E_h, V_h) = (E_h, V_h)_{\mathsf{V}_h} + (E_h, V_h)_{\mathsf{V}_D} + (E_h, V_h)_n.$$

The Lagrange multiplier $\mathsf{V}_h \in \Lambda_{h,V_D}$ is the unique solution of

$$\tilde{a}(V_h, \mu) = \tilde{b}(\mu), \qquad \mu \in \Lambda_{h,V_D}, \tag{4.23}$$

where

$$\tilde{a}(\mathsf{V}_h, \mu) = \int_\Omega \lambda^{-2} E_{h,\mathsf{V}_h} \cdot E_{h,\mu} dx \quad \text{and} \quad \tilde{b}(\mu) = \int_\Omega [n - \mathrm{Dop}] V_{h,\mu} dx.$$

**Remark 4.3.** Just like for classical finite-element methods, the variational formulation (4.23) gives rise to a matrix equation for the degrees of freedom of the multiplier $\mathsf{V}_h$.

## 4.4 Static condensation

The static condensation procedure adopted from [HJP03] eliminates $\boldsymbol{E}_h$ and $\boldsymbol{V}_h$ from the system (4.18) leading to the final algebraic system in the variable $\mathsf{V}_h$. This can be done on every single partition element $K \in \mathcal{T}_h$, since the matrix $\boldsymbol{A}$ is *block-diagonal.* Hence it may be easily and inexpensively inverted at the element level, leading to the linear system

$$\boldsymbol{M}\,\mathsf{V}_h = \boldsymbol{G} \tag{4.24}$$

with

$$\boldsymbol{M} = \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{C} - \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{C}, \tag{4.25}$$

$$\boldsymbol{G} = \boldsymbol{C}^T\boldsymbol{A}^{-1}\boldsymbol{B}\left(\boldsymbol{B}^T\boldsymbol{A}^{-1}\boldsymbol{B}\right)^{-1}\boldsymbol{F}. \tag{4.26}$$

The matrix $\boldsymbol{M}$ is an M-matrix, provided $\mathcal{T}_h$ is weakly acute. For details on the properties of the finite element grid we refer to section 6.
That the matrices in (4.25) exist is due to the regularity of the approximation spaces. With the notation

$$G_K = -\frac{1}{36|K|}\left(|K|\, n_K - \int_K \mathrm{Dop}\, dx\right),$$

on $K$, the linear system (4.24) has the form

$$\frac{\lambda^2}{|K|}\begin{pmatrix} n_1\cdot n_1 & n_1\cdot n_2 & n_1\cdot n_3 & n_1\cdot n_4 \\ n_2\cdot n_2 & n_2\cdot n_2 & n_2\cdot n_3 & n_2\cdot n_4 \\ n_3\cdot n_1 & n_3\cdot n_2 & n_3\cdot n_3 & n_3\cdot n_4 \\ n_4\cdot n_1 & n_4\cdot n_2 & n_4\cdot n_3 & n_4\cdot n_4 \end{pmatrix}\begin{pmatrix} \mathsf{V}_1 \\ \mathsf{V}_2 \\ \mathsf{V}_3 \\ \mathsf{V}_4 \end{pmatrix} = G_K \begin{pmatrix} n_1\cdot(+\tau_{12} + \tau_{13} + \tau_{14}) \\ n_2\cdot(-\tau_{12} + \tau_{23} + \tau_{24}) \\ n_3\cdot(-\tau_{13} - \tau_{23} + \tau_{34}) \\ n_4\cdot(-\tau_{14} - \tau_{24} - \tau_{34}) \end{pmatrix} \tag{4.27}$$

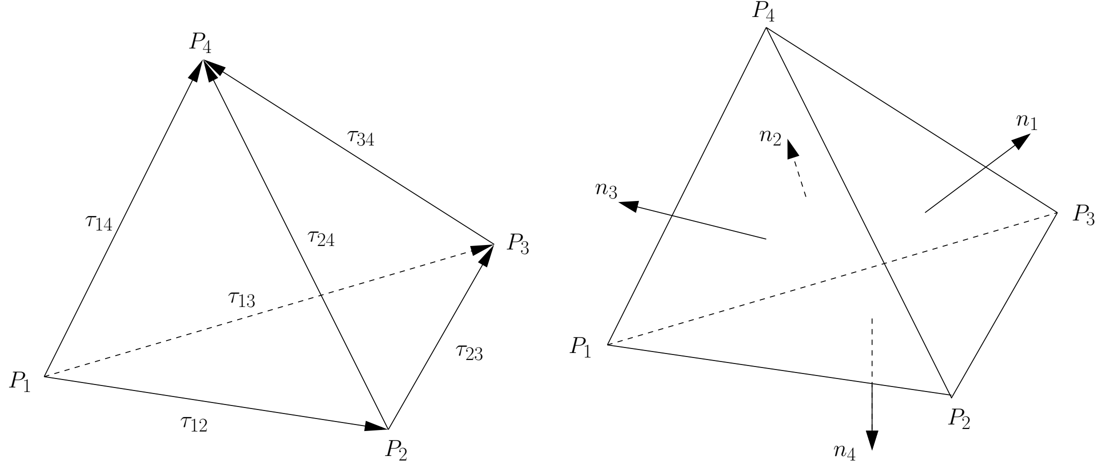



Figure 2: Directional vectors of the edges of $K$ and outward normal vectors

where $n_i = |e_i|\nu_i$, $i = 1, \dots, 4$, are the outward normal vectors on the face $e_i$ scaled with the area $|e_i|$ of that face and $\tau_{ij}$ are the directional vectors of the edges of $K$ with the corresponding edge length (see figure 2),

$$\tau_{ij} = \overrightarrow{OP_j} - \overrightarrow{OP_i}, \quad i = 1 \dots 3,\ j = i+1 \dots 4.$$

By static condensation from (4.12) we obtain a linear system acting only on the Lagrange multipliers $\mathbf{w}_{1h}$ and $\mathbf{w}_{2h}$. In an element-by-element post-processing, we can easily reconstruct the nodal values of the current densities $\boldsymbol{I}_{1h}$ and $\boldsymbol{I}_{2h}$ from the Lagrange multipliers $\mathbf{w}_{1h}$ and $\mathbf{w}_{2h}$,

$$(\boldsymbol{I}_{2h})_j = \begin{cases} \dfrac{D_{21,K}}{|K|} \displaystyle\sum_{i=1}^{4} |e_i|(-\nu_i)_j \mathsf{w}_{1i} + \dfrac{D_{22,K}}{|K|} \displaystyle\sum_{i=1}^{4} |e_i|(-\nu_i)_j \mathsf{w}_{2i}, & \text{if } j = 1,2,3, \\ \dfrac{1}{3}\dfrac{1}{|K|} W_K, & \text{if } j = 4, \end{cases} \tag{4.28}$$

where we have set $D_{2i,K} = D_{2i}(V_K, w_{1K}, w_{2K})$, $i = 1,2$, and $W_K = W(V_K, w_{1K}, w_{2K})$. The approximation $I_{1h}$ is constant on each tetrahedron,

$$(\boldsymbol{I}_{1h})_j = \begin{cases} \dfrac{D_{11,K}}{|K|} \displaystyle\sum_{i=1}^{4} |e_i|(-\nu_i)_j \mathsf{w}_{1i} + \dfrac{D_{12,K}}{|K|} \displaystyle\sum_{i=1}^{4} |e_i|(-\nu_i)_j \mathsf{w}_{2i}, & \text{if } j = 1,2,3, \\ 0, & \text{if } j = 4, \end{cases} \tag{4.29}$$

where we have set $D_{1i,K} = D_{1i}(V_K, w_{1K}, w_{2K})$, $i = 1,2$. The constants $V_K, w_{1K}$ and $w_{2K}$ are defined by (4.2).

**Remark 4.4.** This technique may be seen as an implementational trick only, i.e, as a computationally convenient way to determine the solution of the original system (4.5). But the Lagrange multipliers may offer additional information. According to [AB85] it is possible to use the multipliers in the reconstruction of approximate current densities which are asymptotically *more accurate* than the approximation furnished by the original current densities $I_{1h}$ and $I_{2h}$. But this requires the diffusion coefficients to be sufficiently smooth, which can not be assumed in the present context.

## 4.5 Evaluation of terminal currents

The terminal currents at the Ohmic contacts are an essential output of device simulations. In this subsection, we make precise how these currents are computed. This discussion is similar to that

in [FMWW90]. Since the device has a finite number of Ohmic contacts, the boundary part $\Gamma_D$ is a finite set of separated contacts. We comment that, in the case of different types of boundary conditions, the definition of a regular sequence of meshes includes the property that any connected component of $\Gamma_D$ or $\Gamma_N$ is a union of faces of tetrahedra from the partition. This is the same assumption as in [AW99].
From the definition of Ohmic contacts we know that the potential drop within a contact is negligible (cf. [Sze81, p. 304]). Thus $V$ is constant on each Ohmic contact.
Let $c \subset \Gamma_D$ be a contact (separated from other contacts by a positive distance) and let $\phi_c$ be a smooth function on $\Omega$, fulfilling

$$\phi_c(x) = \begin{cases} 1, & x \in c, \\ 0, & x \in \Gamma_D \setminus c. \end{cases}$$

Multiplying (2.70) by $\phi_c$ and integration by parts gives

$$0 = -\int_\Omega \nabla \cdot I_1 \phi_c \, dx = -\int_{\partial\Omega} I_1 \cdot \nu \phi_c ds + \int_\Omega I_1 \cdot \nabla \phi_c \, dx = -\int_c I_1 \cdot \nu \, ds + \int_\Omega I_1 \cdot \nabla \phi_c \, dx.$$

This suggests to define the corresponding terminal current via

$$I_{1,c} := \int_c I_1 \cdot \nu \, ds = \int_\Omega (D_{11} \nabla w_1 + D_{12} \nabla w_2) \cdot \nabla \phi_c \, dx. \tag{4.30}$$

Equivalently, we have

$$I_{1,c} = \sum_{K \in \mathcal{T}_h} \int_K (D_{11} \nabla w_1 + D_{12} \nabla w_2) \cdot \nabla \phi_c \, dx. \tag{4.31}$$

In order to show that the terminal current (4.30) is independent of the actual choice of $\phi_c$, take $\phi_c \in H_0^1(\Omega)$. It follows immediately that

$$0 = \int_\Omega (D_{11} \nabla w_1 + D_{12} \nabla w_2) \cdot \nabla \phi_c \, dx,$$

since $\phi_c$ vanishes on $c$ and hence the integral over the contact $c$ vanishes.
Now, we proceed to the definition of the discrete electron current. Recall that by $\mathcal{E}_h$ we denote the set of all faces of the partition elements, and by $\mathcal{E}_h'$ all inner faces. Moreover, on $K$ and the contact $c$ we define

$$\mathcal{E}(K) = \{e \in \mathcal{E}_h \, : \, e \subset \partial K\}, \quad \mathcal{E}(c) = \{e \in \mathcal{E}_h \, : \, e \subset c\}.$$

The discrete electron current at $c$ is given by

$$I_{1,c}^h := \sum_{\tilde{e} \in \mathcal{E}(c)} \int_{\tilde{e}} I_{1h} \cdot \nu \, ds, \tag{4.32}$$

which corresponds to (4.30). The continuity of the normal components of $I_{1h}$ across inter-element faces allows for an expansion of (4.32) that involves all faces. In order to show this, we take $\mu \in \Lambda_{h,\xi}$, where $\xi \in L^2(\Gamma_D)$ is defined by

$$\xi(x) = \begin{cases} 1, & x \in c, \\ 0, & x \in \Gamma_D \setminus c, \end{cases}$$

and apply the continuity requirement,

$$I_{1,c}^h = \sum_{\tilde{e} \in \mathcal{E}(c)} \int_{\tilde{e}} \mu I_{1h} \cdot \nu \; ds = \sum_{\tilde{e} \in \mathcal{E}_h} \int_{\tilde{e}} \mu I_{1h} \cdot \nu \; ds - \sum_{\tilde{e} \in \mathcal{E}_h'} \int_{\tilde{e}} \mu I_{1h} \cdot \nu \; ds.$$

Hence, the discrete electron current takes the form

$$I_{1,c}^h = \sum_{\tilde{e}\in\mathcal{E}_h} \int_{\tilde{e}} \mu I_{1h} \cdot \nu \, ds = \sum_{K\in\mathcal{T}_h} \sum_{\tilde{e}\in\mathcal{E}(K)} \int_{\tilde{e}} \mu I_{1h} \cdot \nu \, ds.$$

Finally, by means of (4.29) and the fact that $\mu$ is piecewise constant on the faces we get

$$I_{1,c}^h = -\sum_{K\in\mathcal{T}_h} \left( \frac{D_{11,K}}{|K|} \sum_{e,\tilde{e}\in\mathcal{E}(K)} (n_e \cdot n_{\tilde{e}}) \mathsf{w}_{1e} \mu_{\tilde{e}} + \frac{D_{12,K}}{|K|} \sum_{e,\tilde{e}\in\mathcal{E}(K)} (n_e \cdot n_{\tilde{e}}) \mathsf{w}_{2e} \mu_{\tilde{e}} \right), \tag{4.33}$$

where $w_{1e}$ denotes the value of $w_{1h}$ associated with the face $e$ and $n_e = |e|\nu_e$ is the scaled outward normal on $e$. In the numerical tests we choose $\mu$ with $\mu_{\tilde{e}} = 0$, if $\tilde{e} \in \mathcal{E}_h'$, and (4.33) reduces to

$$I_{1,c}^h = -\sum_{\tilde{e}\in\mathcal{E}(c)} \mu_{\tilde{e}} \sum_{K\in\mathcal{T}_h} \left( \frac{D_{11,K}}{|K|} \sum_{e\in\mathcal{E}(K)} (n_e \cdot n_{\tilde{e}}) \mathsf{w}_{1e} + \frac{D_{12,K}}{|K|} \sum_{e\in\mathcal{E}(K)} (n_e \cdot n_{\tilde{e}}) \mathsf{w}_{2e} \right).$$

The discrete electron current (4.33) is an approximation of (4.31).

# 5 The iterations

The system of equations (2.70)-(2.74) is nonlinear, since the diffusion coefficients (2.76) and the electron density (2.75) depend on the solution. Therefore, the mixed-hybrid discretization (4.18),(4.12) is a nonlinear system. In this section two iteration methods for solving the nonlinear finite-dimensional systems are discussed, the full Newton method and a Gummel-type method. In addition to the nonlinearity, the system is parameter dependent, where the applied voltage $V_A$ in (2.77) is a parameter in some sense. More precisely, the boundary function $V_A$ is scaled with a parameter $\rho \in [0,1]$, and by varying $\rho$ different values of the voltage can be applied to the device. The parameter dependence is taken into account by a *path following method* described in [STL96]. It provides a step control for the parameter $\rho$. This method is explained in detail in the following subsection.

## 5.1 Path following and related extended systems

In order to solve the system (2.70)-(2.74), (2.77) by iteration, we need to increase the applied voltage $V_A$ in small steps. Once the solution is obtained for a fixed voltage, a small bias is added and the corresponding solution is computed. This is repeated until the maximum voltage is reached. We call this technique the *continuation in the applied bias.* A simple strategy is to choose steps of equal size, but this is too costly in many simulations. A more advanced strategy is that of a step control, where the step size is automatically adjusted to the solution. The path following described in this section is an appropriate tool to control the step size. In a preliminary phase we transform the equations (2.70)-(2.74), (2.77) by introducing a boundary data parameter, $\rho$.
We define an extension $U$ of the applied voltage $V_A$ at the Ohmic contacts $\Gamma_{D,1}$ into $\Omega$ by

$$\Delta U = 0 \text{ in } \Omega, \quad U = V_A \text{ on } \Gamma_{D,1}, \quad U = 0 \text{ on } \Gamma_{D,2}, \quad \nu \cdot \nabla U = 0 \text{ on } \Gamma_N, \tag{5.1}$$

and consider the following boundary value problem for $\widetilde{V}$, $\widetilde{w}_1$, $\widetilde{w}_2$,

$$\begin{aligned}
-\nabla \cdot \widetilde{I}_1 &= 0, && (5.2)\\
-\nabla \cdot \widetilde{I}_2 &= W(n, \widetilde{w}_2), && (5.3)\\
\widetilde{I}_1 &= D_{11}(n, \widetilde{V} + \rho U, \widetilde{w}_2)\nabla(\widetilde{w}_1 + \rho U \widetilde{w}_2) + D_{12}(n, \widetilde{V} + \rho U, \widetilde{w}_2)\nabla \widetilde{w}_2, && (5.4)\\
\widetilde{I}_2 &= D_{21}(n, \widetilde{V} + \rho U, \widetilde{w}_2)\nabla(\widetilde{w}_1 + \rho U \widetilde{w}_2) + D_{22}(n, \widetilde{V} + \rho U, \widetilde{w}_2)\nabla \widetilde{w}_2, && (5.5)\\
\nabla \cdot \widetilde{E} &= n(\widetilde{V}, \widetilde{w}_1, \widetilde{w}_2) - \mathrm{Dop}, && (5.6)\\
\widetilde{E} &= \lambda^2 \nabla \widetilde{V}, && (5.7)
\end{aligned}$$

subject to the boundary conditions

$$\begin{array}{llll}
\widetilde{V} = \log n_D, & \widetilde{w}_1 = -\log n_i, & \widetilde{w}_2 = -1 & \text{on } \Gamma_{D,1}, \\
\widetilde{V} = \log n_D + V_A + V_B, & \widetilde{w}_1 = -\log n_i, & \widetilde{w}_2 = -1 & \text{on } \Gamma_{D,2}, \\
\nu \cdot \widetilde{E} = 0, & \nu \cdot \widetilde{I}_1 = 0, & \nu \cdot \widetilde{I}_2 = 0 & \text{on } \Gamma_N.
\end{array} \tag{5.8}$$

We solve the boundary value problem (5.2)-(5.7), (5.8) instead of (2.70)-(2.74), (2.77) with the mixed-hybrid finite elements, yielding a discrete solution $\widetilde{V}_h(\rho)$, $\widetilde{w}_{1h}(\rho)$, $\widetilde{w}_{2h}(\rho)$. If $U_h$ denotes the mixed-hybrid solution of (5.1), then the discrete solution to the original system is obtained as

$$V_h = \widetilde{V}_h(\rho = 1) + U_h, \quad w_{1h} = \widetilde{w}_{1h}(\rho = 1) + U_h \widetilde{w}_{2h}(\rho = 1), \quad w_{2h} = \widetilde{w}_{2h}(\rho = 1).$$

In order to solve (5.2)-(5.7),(5.8) with one of the iterations described below, we start from the equilibrium solution, which is given by $\widetilde{w}_1 = -\log n_i$, $\widetilde{w}_2 = -1$ and the built-in potential $V_{\mathsf{bi}}$,

$$\nabla \cdot E_{\mathsf{bi}} = e^{V_{\mathsf{bi}}} - \mathrm{Dop}, \quad E_{\mathsf{bi}} = \lambda^2 \nabla V_{\mathsf{bi}} \text{ in } \Omega, \tag{5.9}$$

subject to the boundary conditions

$$V_{\mathsf{bi}} = \log n_D \text{ on } \Gamma_{D,1}, \quad V_{\mathsf{bi}} = \log n_D + V_A + V_B \text{ on } \Gamma_{D,2}, \quad \nu \cdot E_{\mathsf{bi}} = 0 \text{ on } \Gamma_N, \tag{5.10}$$

The barrier height $V_B$ is applied to avoid that the electron current flows across the Schottky contact $\Gamma_{D,2}$.
The boundary value problem (5.9), (5.10) is obtained by setting $\widetilde{w}_1 = -\log n_i$, $\widetilde{w}_2 = -1$ and $\rho = 0$ in (5.2)-(5.7), i.e. no voltage is applied to the Ohmic contacts $\Gamma_{D,1}$ and a fixed control voltage is applied to the Schottky contacts $\Gamma_{D,2}$.

**Remark 5.1.** The extension of $V_A$ into $\Omega$ has the effect that the convergence of the iterations is improved, since the boundary data $V_A$ are imposed through the smooth solution $U$. Moreover, $w_1$ in (5.4) and (5.5) is updated by $\rho U \widetilde{w}_2$ instead of $-\rho U$ (which would be straightforward because of the boundary conditions (2.77)). The factor $\widetilde{w}_2$ instead of -1 also improves the convergence.

**Remark 5.2.** The definition of the extension $U$ of $V_A$ is tailored to the so called *open state* of the semiconductor device. In this state, we fix the control voltage at the Schottky contacts $\Gamma_{D,2}$ and the voltage applied to the Ohmic contacts $\Gamma_{D,1}$ is determined by the path-following parameter $\rho$. An increase of the applied voltage results in a significant terminal current at the drain contact, thus the device is in open state. In the so called *closed state* we fix the applied voltage at the Ohmic contacts and vary the control voltage at the Schottky contacts using the path-following method. To this end we define the extension

$$\Delta U = 0 \text{ in } \Omega, \quad U = 0 \text{ on } \Gamma_{D,1}, \quad U = V_A \text{ on } \Gamma_{D,2}, \quad \nu \cdot \nabla U = 0 \text{ on } \Gamma_N. \tag{5.11}$$

The iteration is then started from the solution of the open state instead of the built-in potential $V_{\mathsf{bi}}$, and $\rho U$ is added to the electrostatic potential $V$. In this way, the device is gradually closed and the drain current is reduced.

With the abbreviation $u = (\widetilde{V}_h, \widetilde{w}_{1h}, \widetilde{w}_{2h})$, the mixed-hybrid formulation of (5.2)-(5.7), (5.8) has the form

$$G(u,\rho) = 0, \quad G : D \subseteq \mathbb{R}^{3N} \times \mathbb{R} \to \mathbb{R}^{3N}, \tag{5.12}$$

where $N$ denotes the degrees of freedom. The solution set

$$\mathcal{L} = \{(u,\rho) \in D \,:\, G(u,\rho) = 0\},$$

of (5.12) generically consists of one-dimensional paths. In order to approximate the component $\mathcal{L}_0$ of $\mathcal{L}$ which is defined by an initial solution $(u^{(0)}, \rho^{(0)}) \in \mathcal{L}$, so-called *predictor-corrector path following methods* are used. For tracing the solution path $\mathcal{L}_0$ of (5.12), the *pseudo arclength* approach is used [STL96]. It defines the successor $(u^{(k+1)}, \rho^{(k+1)})$ to the current point $(u^{(k)}, \rho^{(k)})$ on $\mathcal{L}_0$ as the solution of the extended system

$$\begin{aligned} G(u,\rho) &= 0, & (5.13)\\ (c^{(k)})^T (u - u^{(k)}) + \gamma^{(k)} (\rho - \rho^{(k)}) &= \tau^{(k)}. & (5.14) \end{aligned}$$

In (5.14), $\tau^{(k)} > 0$ is a step size, and $c^{(k)} \in \mathbb{R}^N, \gamma^{(k)} \in \mathbb{R}$ have to be chosen such that

$$\begin{pmatrix} c^{(k)} \\ \gamma^{(k)} \end{pmatrix}^T \begin{pmatrix} \varphi^{(k)} \\ \omega^{(k)} \end{pmatrix} \neq 0, \tag{5.15}$$

where $\varphi^{(k)}$ and $\omega^{(k)}$ are defined by $\ker G'(u^{(k)}, \rho^{(k)}) = \operatorname{span}\left\{ \begin{pmatrix} \varphi^{(k)} \\ \omega^{(k)} \end{pmatrix} \right\}, \left\| \begin{pmatrix} \varphi^{(k)} \\ \omega^{(k)} \end{pmatrix} \right\|_2 = 1$. We used the abbreviation

$$G' = \left( \frac{\partial G}{\partial u}, \frac{\partial G}{\partial \rho} \right)^T.$$

Applying Newton's method to (5.13), (5.14) requires solving linear systems with matrices

$$\begin{pmatrix} \partial_1 G(u,\rho) & \partial_2 G(u,\rho) \\ c^T & \gamma \end{pmatrix}, \tag{5.16}$$

where the block $\partial_1 G(u,\rho)$ is a square $3N$-dimensional matrix. The first Newton step determines a so-called predictor point in which the tangent of $\mathcal{L}$ at $(u^{(k)},\rho^{(k)})$ meets the hyperplane defined by the constraint (5.14). In this sense it corresponds to an Euler step. The subsequent Newton steps generate points in the hyperplane (5.14) which, for sufficiently small step size $\tau^{(k)}$, will converge to the successor $(u^{(k)},\rho^{(k)}) \in \mathcal{L}$. Hence, these steps are denoted as Newton corrector iteration, and the whole algorithm as Euler-Newton continuation.
In section 6 numerical tests show how this path following method works on our semiconductor device models.

## 5.2 The full Newton method

The system (5.13), (5.14) can be written as $F(u,\rho)=0$, where $F : D \subseteq \mathbb{R}^{3N+1} \to \mathbb{R}^{3N+1}$. The local convergence of the Newton iterates,

$$\begin{aligned} &F'(u^{(k)},\rho^{(k)})\, s^{(k)} = -F(u^{(k)},\rho^{(k)}), \\ &(u^{(k+1)},\rho^{(k+1)}) = (u^{(k)},\rho^{(k)}) + (s_1^{(k)}, s_2^{(k)}), \quad k \in \mathbb{N}_0, \quad (u^{(0)},\rho^{(0)}) \in D, \end{aligned} \tag{5.17}$$

to a solution $(u^*,\rho^*) \in \operatorname{int} D$ is often examined under the conditions that

- the mapping $F$ is Fréchet differentiable on $\operatorname{int} D$ and $F'$ is continuous at $(u^*,\rho^*)$,
- the Jacobian $F'(u^*,\rho^*)$ is invertible.

The discrete system (5.2)-(5.6) fulfills these conditions.
The full Newton method is attractive because of its rapid convergence from a sufficiently good initial guess, which in our model problem is given by the thermal equilibrium solution,

$$u^{(0)} = (V_{\mathsf{bi}}, \log n_D, -1),\ \rho^{(0)} = 0,$$

provided that the voltage steps are sufficiently small. One drawback of Newton's method is having to solve the Newton equations (5.17) at each stage, which is expensive, since the number of unknowns is large in the numerical tests. Moreover, the Jacobian matrix of the entire system is neither symmetric nor definite. For solving the linear systems arising in the cycles of the Newton method we use direct elimination. By reduction of the band width, the fill-in is reduced.
We use the Newton method as standard with which to compare other iteration methods. The Newton iteration is stopped if $\|s^{(k)}\|_\infty$ is smaller than a prescribed tolerance, with $\|\cdot\|_\infty$ being the discrete maximum norm on $\mathbb{R}^{3N+1}$.

## 5.3 The Gummel-type iterations

In section 3 a linearization of the current continuity system (2.70)-(2.73) is described and the nonlinearity is incorporated into the Poisson equation (2.74). In this section a new iteration method is presented that rests on this linearization and decoupling of the entire system. Starting from an appropriate initial guess $v_1, v_2$, the nonlinear Poisson equation is solved for the electrostatic potential $V = \mathbf{U}[v_1,v_2]$ first. This can be done by means of the full Newton method. The result is put in the diffusion coefficients and the right hand side of the current continuity system, which in turn is solved as linear system for the quasi-Fermi levels $w_1$ and $w_2$. Then the nonlinear Poisson equation is updated from these new values for $w_1$ and $w_2$ and solved again for $V$. This procedure is continued until convergence of the outer iteration is reached. This method is of the type that is widely known as *secant modulus* or *Kačanov's method*. It has some advantages over the Newton method in that it makes the algorithm more flexible to changes of the model parameters

or the semiconductor band structure, since no derivative of the diffusion coefficients need to be computed. Moreover, the method modifies the current continuity system in a way that the diffusion matrix remains symmetric positive definite, provided the initial guess $v_1$ and $v_2$ is properly chosen. Therefore, a preconditioned conjugate gradient method can be applied to solve the linear systems.
The method uses the special structure of the diffusion matrix of the system (2.70)-(2.73) based on the following new idea. Just like for iteration methods used to solve linear systems, the diffusion matrix is split

$$\mathbf{D} = \mathbf{M} - \mathbf{N},$$

where

$$\mathbf{M} = n \begin{pmatrix} 1 & -\left(V + \frac{3}{2}w_2^{-1}\right) \\ -\left(V + \frac{3}{2}w_2^{-1}\right) & \left(V + \frac{3}{2}w_2^{-1}\right)^2 + \frac{3}{2}(1+\alpha)w_2^{-2} \end{pmatrix} \tag{5.18}$$

and

$$\mathbf{N} = n \begin{pmatrix} 0 & 0 \\ 0 & \frac{3}{2}\alpha w_2^{-2} \end{pmatrix}. \tag{5.19}$$

Here, $\alpha$ is a nonnegative parameter. Note that (5.18) is a symmetric positive definite matrix for all $\alpha \geq 0$, since the electron density $n = n_i(-w_2)^{-\frac{3}{2}}e^{w_1 - V w_2}$ and $-w_2$ are positive functions. In some sense, the ellipticity is increased since the determinant of $\mathbf{M}$ is $\frac{3}{2}n^2(1+\alpha)w_2^{-2}$.
If the current continuity system (2.70)-(2.73) is split as suggested, it takes the form

$$\begin{aligned} -\nabla\cdot I_1 &= 0, && (5.20)\\ -\nabla\cdot I_2 &= W - \alpha\,\nabla\cdot\left(\frac{3}{2}n\,w_2^{-2}\,\nabla w_2\right), && (5.21)\\ I_1 &= D_{11}\nabla w_1 + D_{12}\nabla w_2, && (5.22)\\ I_2 &= D_{21}\nabla w_1 + \left(D_{22} + \alpha\,\frac{3}{2}n w_2^{-2}\right)\nabla w_2, && (5.23) \end{aligned}$$

The relaxation term (2.75) can be split in a similar way as the diffusion matrix,

$$W = \frac{3}{2}\frac{n}{\tau_0}w_2^{-1}w_2 + \frac{3}{2}\frac{n}{\tau_0}w_2^{-1} = \frac{3}{2}(1+\alpha)\frac{n}{\tau_0}w_2^{-1}w_2 + \frac{3}{2}\frac{n}{\tau_0}(1-\alpha w_2)w_2^{-1}. \tag{5.24}$$

This splitting adds a zero order term with positive coefficient to (5.21),

$$-\nabla\cdot I_2 - \frac{3}{2}(1+\alpha)\frac{n}{\tau_0}w_2^{-1}\,w_2 = \frac{3}{2}\frac{n}{\tau_0}(1-\alpha w_2)w_2^{-1} - \alpha\,\nabla\cdot\left(\frac{3}{2}n\,w_2^{-2}\,\nabla w_2\right). \tag{5.25}$$

For brevity, the solution of the system (5.20)-(5.23) again is designated $I_1, I_2$ and $w_1, w_2$.

**Remark 5.3.** For the zero order term as introduced above the Raviart-Thomas finite elements are not really appropriate, since they cannot theoretically guarantee the positivity of the discrete solution. In the numerical experiments we have always obtained positive solutions. Equivalently, an equation without zero order term could be solved, guaranteeing the positivity property, but the formulation including the zero order term improves the properties of the stiffness matrix.

We associate the following system map to the parameter-dependent equations (5.20)-(5.23).

1. Choose $v_1, v_2 \in \mathcal{M}$ and insert these functions into the electron density function of the right hand side of the Poisson equation. The set $\mathcal{M}$ is defined in (3.1).

2. Solve the nonlinear Poisson equation for $V$ subject to the boundary conditions $V = V_D$ on $\Gamma_D$ and $\nu\cdot\nabla V = 0$ on $\Gamma_N$.

3. Insert the solution $V$ of step 2 and the initial guess $v_1, v_2$ of step 1 into the diffusion coefficients and the right hand side of the modified current continuity system (5.20)-(5.23) and solve the resulting *linear* system for $w_1$ and $w_2$,

$$\begin{aligned} -\nabla\cdot I_1 &= 0, \\ -\nabla\cdot I_2 - \frac{3}{2}(1+\alpha)\frac{n}{\tau_0}v_2^{-1}\,w_2 &= \frac{3}{2}\frac{n}{\tau_0}(1-\alpha v_2)v_2^{-1} - \alpha\,\nabla\cdot\left(\frac{3}{2}n\,v_2^{-2}\,\nabla v_2\right), \\ I_1 &= D_{11}\nabla w_1 + D_{12}\nabla w_2, \\ I_2 &= D_{21}\nabla w_1 + (D_{22} + \alpha\,\frac{3}{2}n v_2^{-2})\nabla w_2, \end{aligned}$$

subject to the boundary conditions

$$w_1 = w_{1D},\; w_2 = w_{2D} \;\text{ on } \Gamma_D \quad\text{and}\quad \nu\cdot I_1 = 0,\; \nu\cdot I_2 = 0 \;\text{ on } \Gamma_N.$$

Step 2 is the solution map $\mathbf{U}[v_1, v_2]$ defined in the section 3.1. Step 3 is the second fractional step, which corresponds to a map $\mathbf{W}(V; v_1, v_2) := (w_1, w_2)$. We define a function $\Phi_1 : \mathcal{M} \to L^2(\Omega)\times L^2(\Omega)$ by $\Phi_1(v_1, v_2) := (w_1, w_2)$, which is the composition of $\mathbf{U}$ and $\mathbf{W}$. In the following, we shall assume that $\Phi_1 : \mathcal{M} \to \mathcal{M}$ is a continuous fixed point map. This is needed for the existence of a fixed point.
Now we interpret the system (5.20)-(5.23) in the following sense: first, the unchanged system (3.13)-(3.16) is discretized with the mixed finite-elements, then hybridized, then reduced to the Lagrange multipliers by static condensation and then the splitting is applied. This interpretation allows to consider the map $\Phi_1$ as a discrete version of the system map $\Phi : \mathcal{M} \to L^2(\Omega)\times L^2(\Omega)$ in (3.31).
That this fixed point map is consistent with the system (2.70)-(2.73), i.e. that every solution to this system is a fixed point of $\Phi_1$, is easily checked by replacing $(v_1, v_2)$ with a solution $(w_1, w_2)$. In this case, both terms containing the parameter $\alpha$ cancel and the system (2.70)-(2.73) is retrieved.
Next we show that $\Phi_1$ is even a contraction in a simplified situation and under the boundedness conditions on the quasi-Fermi levels $w_1$ and $w_2$, and we derive iteration schemes that are based on $\Phi_1$. To this end we define the following iteration matrix, neglecting the differential operators of the current continuity system and, for the sake of simplicity, the artificial zero order term,

$$\mathbf{I}(V, w_2) := \mathbf{M}^{-1}\mathbf{N} = \frac{\alpha}{1+\alpha}\begin{pmatrix} 0 & V + \frac{3}{2}w_2^{-1} \\ 0 & 1 \end{pmatrix}.$$

It is obvious that large values of $\alpha$ shift the convergence rate $\alpha/(1+\alpha)$ close to 1. Consequently, $\alpha$ must be kept as small as possible. On the other hand, in the simulations it is required that this damping parameter $\alpha$ has to be chosen sufficiently large to guarantee convergence of the outer iteration. Hence, an optimal value for $\alpha$ has to be found. Of course, the particular values may vary, depending on the semiconductor device geometry, the doping profile function and the voltage applied to the device.
Let $(V^{(l)}, {w_1}^{(l)}, {w_2}^{(l)})$, $l = 0, \ldots, m$, be some iterates generated by $\Phi_1$ and $\mathbf{U}$,

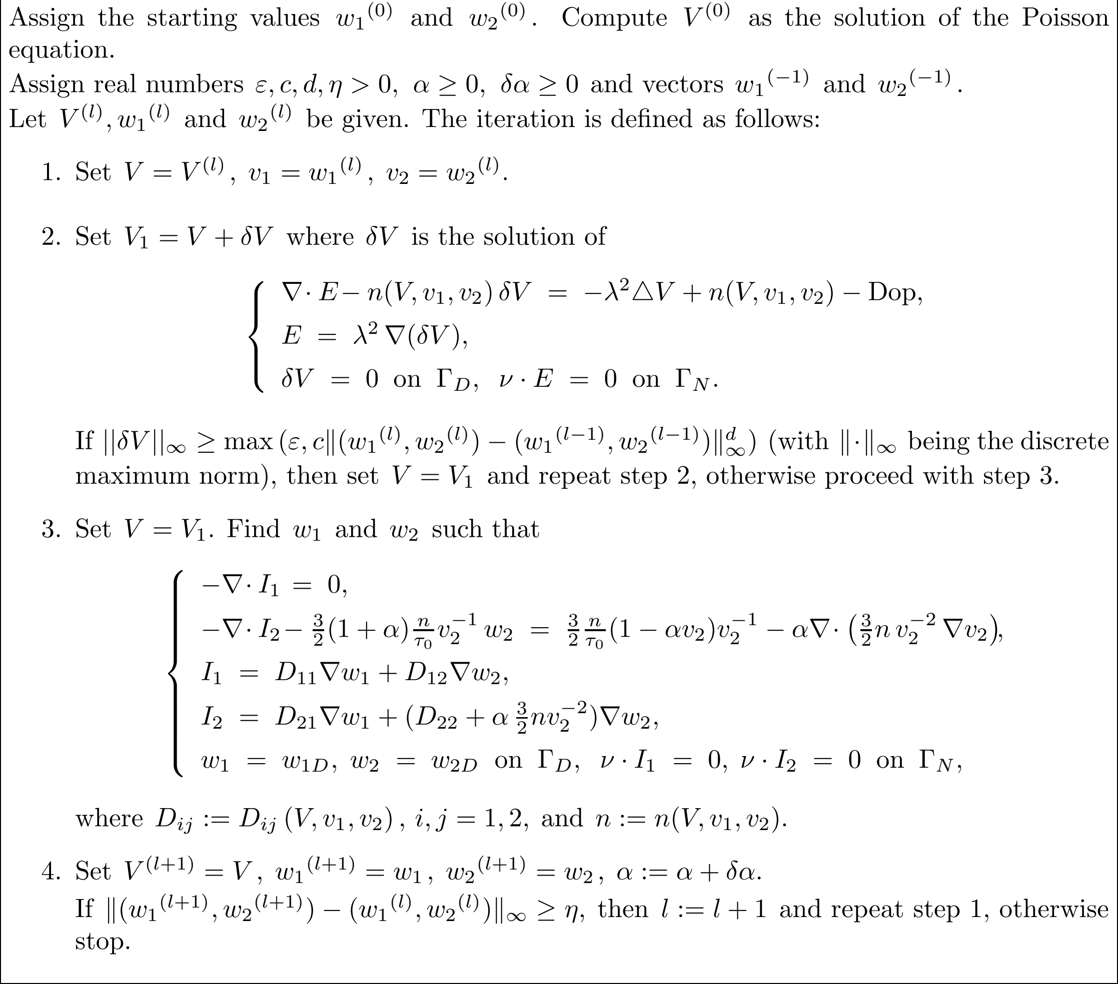
$$({w_1}^{(l)}, {w_2}^{(l)}) = \Phi({w_1}^{(l-1)}, {w_2}^{(l-1)}) \quad\text{and}\quad V^{(l)} = \mathbf{U}[{w_1}^{(l)}, {w_2}^{(l)}].$$

Then it holds,

$$\mathbf{I}(V^{(m)}, {w_2}^{(m)}) \times \cdots \times \mathbf{I}(V^{(0)}, {w_2}^{(0)}) = \left(\frac{\alpha}{1+\alpha}\right)^m \begin{pmatrix} 0 & V^{(m)} + \frac{3}{2}/{w_2}^{(m)} \\ 0 & 1 \end{pmatrix}. \tag{5.26}$$

Furthermore, since $V$ is bounded by $\gamma$ and $\delta$ as stated in (3.3) and (3.4), we have

$$\|\mathbf{I}\|_2 = \frac{\alpha}{1+\alpha}\sqrt{1 + \left(V + \frac{3}{2}w_2^{-1}\right)^2} \le \frac{\alpha}{1+\alpha}\left(1 + \max\{|\gamma|, |\delta|\} + \frac{3}{2}\gamma_2^{-1}\right),$$



Assign the starting values ${w_1}^{(0)}$ and ${w_2}^{(0)}$. Compute $V^{(0)}$ as the solution of the Poisson equation.
Assign real numbers $\varepsilon, c, d, \eta > 0,\ \alpha \geq 0,\ \delta\alpha \geq 0$ and vectors ${w_1}^{(-1)}$ and ${w_2}^{(-1)}$.
Let $V^{(l)}, {w_1}^{(l)}$ and ${w_2}^{(l)}$ be given. The iteration is defined as follows:

1. Set $V = V^{(l)},\ v_1 = {w_1}^{(l)},\ v_2 = {w_2}^{(l)}$.

2. Set $V_1 = V + \delta V$ where $\delta V$ is the solution of

$$\begin{cases} \nabla\cdot E - n(V, v_1, v_2)\,\delta V \ = \ -\lambda^2 \triangle V + n(V, v_1, v_2) - \mathrm{Dop}, \\ E \ = \ \lambda^2\,\nabla(\delta V), \\ \delta V \ = \ 0 \ \text{ on } \Gamma_D,\ \ \nu\cdot E \ = \ 0 \ \text{ on } \Gamma_N. \end{cases}$$

   If $||\delta V||_\infty \geq \max\left(\varepsilon, c\|({w_1}^{(l)}, {w_2}^{(l)}) - ({w_1}^{(l-1)}, {w_2}^{(l-1)})\|_\infty^d\right)$ (with $\|\cdot\|_\infty$ being the discrete maximum norm), then set $V = V_1$ and repeat step 2, otherwise proceed with step 3.

3. Set $V = V_1$. Find $w_1$ and $w_2$ such that

$$\begin{cases} -\nabla\cdot I_1 \ = \ 0, \\ -\nabla\cdot I_2 - \frac{3}{2}(1+\alpha)\frac{n}{\tau_0}v_2^{-1}\,w_2 \ = \ \frac{3}{2}\frac{n}{\tau_0}(1-\alpha v_2)v_2^{-1} - \alpha\nabla\cdot\left(\frac{3}{2}n\,v_2^{-2}\,\nabla v_2\right), \\ I_1 \ = \ D_{11}\nabla w_1 + D_{12}\nabla w_2, \\ I_2 \ = \ D_{21}\nabla w_1 + (D_{22} + \alpha\,\frac{3}{2}nv_2^{-2})\nabla w_2, \\ w_1 \ = \ w_{1D},\ w_2 \ = \ w_{2D} \ \text{ on } \Gamma_D,\ \ \nu\cdot I_1 \ = \ 0,\ \nu\cdot I_2 \ = \ 0 \ \text{ on } \Gamma_N, \end{cases}$$

   where $D_{ij} := D_{ij}\left(V, v_1, v_2\right),\ i,j = 1,2,$ and $n := n(V, v_1, v_2)$.

4. Set $V^{(l+1)} = V,\ {w_1}^{(l+1)} = w_1,\ {w_2}^{(l+1)} = w_2,\ \alpha := \alpha + \delta\alpha$.
   If $\|({w_1}^{(l+1)}, {w_2}^{(l+1)}) - ({w_1}^{(l)}, {w_2}^{(l)})\|_\infty \geq \eta$, then $l := l+1$ and repeat step 1, otherwise stop.

Figure 3: Implementation of the discrete fixed point iteration $\Phi_1$

and this shows that $\Phi_1$ is a contraction with the contraction number $\alpha/(1+\alpha)$.
For the discrete system, we implement $\Phi_1$ in terms of the algorithm shown in figure 3.
The increment $\delta\alpha$ in $\Phi_1$ may also change with each cycle. In section 6 we give the values for $\varepsilon, c, d, \eta, \alpha$ and $\delta\alpha$ actually used in the computations. By virtue of the stopping criterion for $\delta V$ the Poisson equation is solved the more accurate the smaller the error of the quasi-Fermi levels $w_1$ and $w_2$ is, provided that the latter is above $\varepsilon$.
The step 2 can be seen as a linearized Newton method, where the derivative of the electron density w.r.t. $V$, $\partial_V n = -w_2 n$, is replaced by $n$. (Recall the definition of $n$, (2.75).)
At the start of a simulation, no voltage is applied to the device. Then initial values are given by

$${w_1}^{(-1)} = {w_1}^{(0)} = -\log n_i \quad \text{and} \quad {w_2}^{(-1)} = {w_2}^{(0)} = -1 \quad \text{on} \quad \overline{\Omega}.$$

In this case $V^{(0)}$ is the electrostatic potential in thermal equilibrium determined only by the doping profile.
The above algorithm is new and presented here for the first time. It belongs to the Gummel-type iterations. In [HJP03] a similar algorithm is described, which is used to solve energy-transport equations given in the so-called *drift-diffusion formulation*. There, the Gummel iteration procedure is based on iteration methods used to solve the classical drift-diffusion model, and this Gummel iteration procedure can be regarded as an approximate Newton method, where the infor-

mation about the strong coupling of the unknowns is incorporated into the Poisson equation and the Jacobian matrix of the system is replaced by a diagonal matrix [Jer96]. The current continuity system is symmetrized by a transformation of the unknowns. For the algorithm described in this thesis no such transform is needed, since the system is generically symmetric. The Jacobian matrix of the system is replaced by a block-diagonal matrix (the current continuity system is still coupled), and the nonlinearity is also incorporated into the Poisson equation, but more straightforward than it is done for the algorithm in [HJP03]. On the other hand, that algorithm does not require a damping parameter.
As it is well known that a Gummel-type iteration scheme is very sensitive to the choice of initial data, in particular far from thermal equilibrium, the above algorithm is coupled to a continuation in the applied bias. In our simulations, we used uniform bias steps for simplicity.
To keep the algorithm simple, the damping parameter $\alpha$ has been chosen independently of the solution. But we expect faster convergence for $\alpha$ controlled by the solution.
In numerical tests we observed that the error tolerance in step 2 of the above algorithm must be chosen relatively small. The global convergence is very sensitive to the solution of the Poisson equation, especially far from thermal equilibrium, i.e. for high voltages applied to the device.
In the above algorithm the current continuity system (5.20)-(5.23) is still coupled. In the following subsection, we show that this system can be decoupled by backward substitution, which corresponds to a nonlinear Gauss-Seidel iteration for the entire system.

### 5.3.1 Decoupling of the current continuity system

The idea of the decoupling is as follows. Multiplying the energy balance equation (5.21) and the energy current density (5.23) by $(1+\alpha)^{-1}$ leads to the following new current continuity system,

$$
\begin{aligned}
-\nabla\cdot I_1 &= 0 && (5.27)\\
-\nabla\cdot I_2 &= (1+\alpha)^{-1}\,W - \frac{\alpha}{1+\alpha}\,\nabla\cdot\left(\frac{3}{2}n\,w_2^{-2}\,\nabla w_2\right) && (5.28)\\
I_1 &= D_{11}\nabla w_1 + D_{12}\nabla w_2 && (5.29)\\
I_2 &= (1+\alpha)^{-1}D_{21}\nabla w_1 + \widetilde{D}_{22}\nabla w_2, && (5.30)
\end{aligned}
$$

where $\widetilde{D}_{22} := [(1+\alpha)^{-1}\left(V+\frac{3}{2}w_2^{-1}\right)^2 + \frac{3}{2}w_2^{-2}]\,n$. Again, for the sake of brevity we denote the solution by $I_1, I_2$ and $w_1, w_2$.
For large values of $\alpha$ we have reduced the coupling of the balance equations, i.e., the new diffusion matrix is close to a triangular matrix.
In order to obtain a fixed point iteration, we perform the following steps:

1. Choose $v_1, v_2 \in \mathcal{M}$ and insert these functions into the electron density function of the right hand side of the Poisson equation.

2. Solve the nonlinear Poisson equation for $V$ subject to the boundary conditions $V = V_D$ on $\Gamma_D$ and $\nu\cdot\nabla V = 0$ on $\Gamma_N$.

3. Linearize the energy balance equation (5.28), i.e. insert the solution $V$ of step 2 and $v_1, v_2$ of step 1 into the electron density function and the energy relaxation term in (5.28) as well as the diffusion coefficients in (5.30) and consider the following *linear* boundary value problem for a function $w_2$,

$$
\begin{aligned}
-\nabla\cdot I_2 &= (1+\alpha)^{-1}\,\nabla\cdot\left[-n\left(V+\frac{3}{2}v_2^{-1}\right)\nabla v_1 + n\left(V+\frac{3}{2}v_2^{-1}\right)^2\nabla v_2\right]\\
&- \frac{\alpha}{1+\alpha}\,\nabla\cdot\left(\frac{3}{2}n\,v_2^{-2}\nabla v_2\right) + (1+\alpha)^{-1}\,W,\\
I_2 &= \frac{3}{2}n\,v_2^{-2}\nabla w_2,
\end{aligned}
$$

subject to the boundary conditions $w_2 = w_{2D}$ on $\Gamma_D$ and $\nu\cdot I_2 = 0$ on $\Gamma_N$.

4. Insert $v_1$ of step 1 and $w_2$ of step 2 into the electron density function of the right hand side of the Poisson equation and solve it for $V$ subject to the boundary conditions $V = V_D$ on $\Gamma_D$ and $\nu \cdot \nabla V = 0$ on $\Gamma_N$.

5. Linearize the electron equation (5.27), i.e. insert the solution $V$ of step 4, $v_1$ of step 1 and the solution $w_2$ of step 3 into the diffusion coefficients in (5.29) and consider the following *linear* boundary value problem for a function $w_1$,

$$\begin{aligned} -\nabla\cdot I_1 &= 0, \\ I_1 &= D_{11}(V, v_1, w_2)\,\nabla w_1 + D_{12}(V, v_1, w_2)\,\nabla w_2, \end{aligned}$$

subject to the boundary conditions $w_1 = w_{1D}$ on $\Gamma_D$ and $\nu \cdot I_1 = 0$ on $\Gamma_N$.

Again, we can define a fixed point map, $\Phi_2 : \mathcal{M} \to L^2(\Omega) \times L^2(\Omega)$ by $\Phi_2(v_1, v_2) := (w_1, w_2)$.

**Remark 5.4.** We note here that the convergence of this fixed point iteration depends on the order in which we solve the linear problems. If we interchange step 3 with step 5 we observe that the algorithm does not work.
In the steps 2 and 4 the solution $V$ must be computed with high precision, i.e. it is not sufficient only to perform a single iteration step. Moreover, it is necessary to update the electrostatic potential in step 4 from the new solution $w_2$ in order to improve the convergence of the outer iteration.

**Remark 5.5.** We can relate the solution $w_2$ of the linear boundary value problem of step 3 to $v_2$ by $w_2 = v_2 + (1+\alpha)^{-1}\delta w_2$, where $\delta w_2$ is the solution to the linear boundary value problem,

$$-\nabla\cdot(\delta I_2) = \nabla\cdot(D_{21}\nabla v_1 + D_{22}\nabla v_2) + W, \tag{5.31}$$

$$\delta I_2 = \frac{3}{2} n\, v_2^{-2}\nabla(\delta w_2), \tag{5.32}$$

subject to the boundary conditions $\delta w_2 = 0$ on $\Gamma_D$ and $\nu\cdot(\delta I_2) = 0$ on $\Gamma_N$. The electron density, the energy relaxation term and the diffusion coefficients depend on $v_1, v_2$ and $V = \mathbf{U}[v_1, v_2]$. The equation (5.31) is nothing but the energy balance equation in residual form, and $\delta w_2$ is a correction term for the quasi-Fermi level $v_2$. Since $\frac{3}{2} n v_2^{-2} > 0$, a vanishing residuum of the energy balance equation implies a zero correction because of discrete maximum principles.

Splitting of the relaxation term (5.24) as well as inserting $v_2$ and the relation $w_2 = v_2 + (1+\alpha)^{-1}\delta w_2$ in selected instances of $w_2$ leads to the term

$$\frac{3}{2}\frac{n}{\tau_0} v_2^{-1} v_2 + \frac{3}{2}(1+\alpha)^{-1}\frac{n}{\tau_0} v_2^{-1}\delta w_2 + \frac{3}{2}\frac{n}{\tau_0} v_2^{-1} = \frac{3}{2}(1+\alpha)^{-1}\frac{n}{\tau_0} v_2^{-1}\delta w_2 + W(V, v_1, v_2),$$

and to a linear boundary value problem with an artificial zero order term,

$$-\nabla\cdot(\delta I_2) - \frac{3}{2}(1+\alpha)^{-1}\frac{n}{\tau_0} v_2^{-1}\,\delta w_2 = \nabla\cdot(D_{21}\nabla v_1 + D_{22}\nabla v_2) + W, \tag{5.33}$$

$$\delta I_2 = \frac{3}{2} n\, v_2^{-2}\nabla(\delta w_2), \tag{5.34}$$

subject to the boundary conditions $\delta w_2 = 0$ on $\Gamma_D$ and $\nu\cdot(\delta I_2) = 0$ on $\Gamma_N$. Again, the electron density, the energy relaxation term and the diffusion coefficients depend on $v_1, v_2$ and $V = \mathbf{U}[v_1, v_2]$. In order to make the zero order term more relevant, we omit the factor $(1+\alpha)^{-1}$ in front of the artificial zero order term and replace the boundary value problem (5.33),(5.34) by the following,

$$\begin{aligned} -\nabla\cdot(\delta I_2) - \frac{3}{2}\frac{n}{\tau_0} v_2^{-1}\,\delta w_2 &= \nabla\cdot(D_{21}\nabla v_1 + D_{22}\nabla v_2) + W, \\ \delta I_2 &= \frac{3}{2} n\, v_2^{-2}\nabla(\delta w_2), \end{aligned}$$

subject to the boundary conditions $\delta w_2 = 0$ on $\Gamma_D$ and $\nu\cdot(\delta I_2) = 0$ on $\Gamma_N$. This boundary value problem is the one actually used in the computations.

Assign the starting values ${w_1}^{(0)}$ and ${w_2}^{(0)}$. Compute $V^{(0)}$ as the solution of the Poisson equation.
Assign real numbers $\varepsilon, c, d, \eta > 0,\ \alpha \geq 0,\ \delta\alpha \geq 0$ and vectors ${w_1}^{(-1)}$ and ${w_2}^{(-1)}$.
Let $V^{(l)}, {w_1}^{(l)}$ and ${w_2}^{(l)}$ be given. The iteration is defined as follows:

1. Set $V = V^{(l)},\ v_1 = {w_1}^{(l)}$ and $v_2 = {w_2}^{(l)}$.
2. Set $V_1 = V + \delta V$ where $\delta V$ is the solution of
$$\begin{cases} \nabla\cdot E - n(V, v_1, v_2)\,\delta V \;=\; -\lambda^2 \triangle V + n(V, v_1, v_2) - \text{Dop}, \\ E \;=\; \lambda^2\, \nabla(\delta V), \\ \delta V \;=\; 0 \text{ on } \Gamma_D,\ \ \nu\cdot E \;=\; 0 \text{ on } \Gamma_N. \end{cases}$$
If $||\delta V||_\infty \geq \max\left(\varepsilon, c\|({w_1}^{(l)}, {w_2}^{(l)}) - ({w_1}^{(l-1)}, {w_2}^{(l-1)})\|_\infty^d\right)$ (with $\|\cdot\|_\infty$ being the discrete maximum norm) then set $V = V_1$ and repeat step 2, otherwise proceed with step 3.
3. Set $V = V_1$ and find $\delta w_2$ such that
$$\begin{cases} -\nabla\cdot(\delta I_2) - \frac{3}{2}\frac{n}{\tau_0} v_2^{-1}\,\delta w_2 \;=\; \nabla\cdot(D_{21}\nabla v_1 + D_{22}\nabla v_2) + W(V, v_1, v_2), \\ \delta I_2 \;=\; \frac{3}{2} n\, v_2^{-2}\nabla(\delta w_2), \\ \delta w_2 \;=\; 0 \text{ on } \Gamma_D,\ \ \nu\cdot(\delta I_2) \;=\; 0 \text{ on } \Gamma_N, \end{cases}$$
where $D_{2j} := D_{2j}\,(V, v_1, v_2)\,,\ j = 1, 2,$ and $n = n(V, v_1, v_2)$, set $w_2 = v_2 + (1+\alpha)^{-1}\delta w_2$.
4. Set $V_1 = V + \delta V$ where $\delta V$ is the solution of
$$\begin{cases} \nabla\cdot E - n(V, v_1, w_2)\,\delta V \;=\; -\lambda^2 \triangle V + n(V, v_1, w_2) - \text{Dop}, \\ E \;=\; \lambda^2\, \nabla(\delta V), \\ \delta V \;=\; 0 \text{ on } \Gamma_D,\ \ \nu\cdot E \;=\; 0 \text{ on } \Gamma_N. \end{cases}$$
If $||\delta V||_\infty \geq \max\left(\varepsilon, c\|({w_1}^{(l)}, w_2) - ({w_1}^{(l-1)}, {w_2}^{(l-1)})\|_\infty^d\right)$ then set $V = V_1$ and repeat step 4, otherwise proceed with step 5.
5. Set $V = V_1$ and find $w_1$ such that
$$\begin{cases} -\nabla\cdot I_1 \;=\; 0, \\ I_1 \;=\; D_{11}\,(V, v_1, w_2)\ \nabla w_1 + D_{12}\,(V, v_1, w_2)\ \nabla w_2, \\ w_1 \;=\; w_{1D} \text{ on } \Gamma_D,\ \ \nu\cdot I_1 \;=\; 0 \text{ on } \Gamma_N. \end{cases}$$
6. Set $V^{(l+1)} = V,\ {w_1}^{(l+1)} = w_1,\ {w_2}^{(l+1)} = w_2,\ \alpha := \alpha + \delta\alpha$.
If $\|({w_1}^{(l+1)}, {w_2}^{(l+1)}) - ({w_1}^{(l)}, {w_2}^{(l)})\|_\infty \geq \eta$, then $l := l+1$ and repeat step 1, otherwise stop.

Figure 4: Implementation of the discrete fixed point iteration $\Phi_2$

For the discrete system, $\Phi_2$ is implemented in terms of the iteration method shown in figure 4. Again, $\delta\alpha$ may change in the cycles (iterations), and in section 6 we give the values for $\varepsilon, c, d, \eta, \alpha$ and $\delta\alpha$ actually used in the computations. Note that the stopping criterion that finishes the loop over step 4 depends on the current solution $w_2$ instead of the previous iterate ${w_2}^{(l)}$. Thus we take the new information into account.

The choice of the unknowns for the Gummel-type algorithm described in [HJP03] completely decouples the current continuity system, i.e. the resulting system is diagonal. The Gummel iteration procedure $\Phi_2$ only partially decouples the equations, so that the resulting system is triangular.
There is a third version of the Gummel map. We can solve the Poisson equation and the particle equation sequentially while keeping $w_2$ fixed. More precisely, we perform a Newton step (or its linearized equivalent) and put the solution in the coefficients of the particle equation, which in turn is solved as a linear equation for $v_1$. Again, we apply this iteration procedure until convergence with respect to a prescribed error tolerance. This way we perform a fractional step in this Gummel map yielding a new pair $(V, w_1)$. Subsequently, we can either proceed with the energy balance equation as for the map $\Phi_2$ to attain to a correction $\delta w_2$ for the temperature. But a different approach works well, too. We assume $\delta w_2 = 0$ which entails

$$0 = \nabla \cdot (D_{12}\nabla v_1 + D_{22}\nabla \widetilde{v}_2) + W \tag{5.35}$$

for some function $\widetilde{v}_2$. Note that the diffusion coefficients and the energy relaxation term in (5.35) depend on $(V, v_1)$ and the previous iterate $v_2$. We consider (5.35) as a boundary value problem for $\widetilde{v}_2$ and define the next iterate $w_2$ by

$$w_2 = \frac{\alpha}{1+\alpha}\, v_2 + \frac{1}{1+\alpha}\, \widetilde{v}_2. \tag{5.36}$$

Solving (5.35) along with the definition (5.36) completes a step of the Gummel iteration. The whole procedure is repeated until we attain to a sufficiently accurate approximation for $(V, w_1, w_2)$.

**Remark 5.6.** The first fractional step resembles the well-known Gummel map for the drift-diffusion model, where $w_2$ is not present. Since the Gummel map for the drift-diffusion model provides a means to prove the existence of solutions, the close relation of both algorithms suggests to base an existence proof for the ET model on this third variant of our Gummel map. Other then $\Phi_1$ and $\Phi_2$, this Gummel iteration is too slow in the numerical tests to be of practical interest.

The advantage of the iterations $\Phi_1$ and $\Phi_2$ is that the system is decoupled and the resulting linear systems are smaller and symmetric positive definite. For suitably chosen parameter values of $\alpha$, we have convergence for both fixed point iterations, where numerical experiments suggest that $\alpha$ has to be larger for $\Phi_2$ than for $\Phi_1$. In order to accelerate the convergence, we use an extrapolation technique described in the next subsection.

## 5.4 Vector extrapolation

A simple trick to improve the convergence is a vector extrapolation, called the *reduced rank extrapolation*. In [SFS87] the details of this method are described. This method accelerates convergent vector sequences or transforms divergent vector sequences to convergent ones. It does not require explicit knowledge of the "sequence generator" but is computed directly from the terms of the sequence. In our particular case, the base sequences are given by

$$({w_1}^{(l+1)}, {w_2}^{(l+1)}) = \Phi_i({w_1}^{(l)}, {w_2}^{(l)}),\ i = 1, 2, \tag{5.37}$$

and no derivative of the fixed point maps are computed. When applied iteratively to a sequence generated nonlinearly, they tend to converge quadratically ([SFS87]).
The reduced rank extrapolation pertains to the class of polynomial extrapolation methods. It is nonlinear, in that the coefficients of the extrapolating polynomials are functions of the terms of the sequence. The basic idea is the exact solution for a fixed point in the case of a linear generator $\Phi$. This is done without the equivalent of solving a system of equations in dimension $N$ (the total number of degrees of freedom). The polynomial methods find the exact solution as a weighted average of $k+1$ terms of the base sequence, where the $k$ independent weights are found by solving a linear system of that size. $k$ is the "essential degree" of the problem. We always have $k \leq N$,

and sometimes $k$ is much smaller then $N$.
In practical situations $k$ is not known, but the method may be applied with a possibly much smaller "effective degree", which is essentially the number of "dominant" eigenvalues of the sequence generator. This method works well for sequences of vectors that converge linearly in some norm. Although this assumption is not numerically verified in our special situation, the method provides super-linear convergence. Now the general setting of the reduced rank extrapolation is described in detail [SFS87].

### 5.4.1 Reduced rank extrapolation

Suppose that a sequence of vectors in real $N$-space is generated linearly from a starting point $\mathbf{x}_0$

$$\mathbf{x}_{j+1} = A\mathbf{x}_j + \mathbf{b}, \quad j = 0, 1, 2, \ldots, \tag{5.38}$$

where $A$ is a fixed matrix and $\mathbf{b}$ is a fixed vector. We do not assume that either $A$ or $\mathbf{b}$ is known; only the sequence $\{\mathbf{x}_j\}$ or a means of generating it is given. We do assume that 1 is not an eigenvalue of $A$, so the iteration (5.38) has a unique fixed point $\mathbf{s}$, namely

$$\mathbf{s} = (I - A)^{-1}\mathbf{b}. \tag{5.39}$$

If $|\lambda| < 1$ for every eigenvalue $\lambda$ of $A$, then

$$\mathbf{s} = \lim_{j\to\infty} \mathbf{x}_j; \tag{5.40}$$

if the sequence diverges, $\mathbf{s}$, defined by (5.39), still exists and is called the *anti-limit*, and we may still expect to determine $\mathbf{s}$ from a finite number of the terms of the sequence. The objective is to do that from as few terms as possible, without requiring any additional information about $A$, and without inverting an $N \times N$ matrix. In a typical application, $N$ may be quite large relative to the number of eigenvalues $\lambda$ with $|\lambda|$ near 1. (causing slow convergence) or greater than 1 (usually causing divergence).
The extrapolation method to be derived is based on differences, and it will be convenient to have abbreviated notation for first and second differences of the vectors $\mathbf{x}_j$. We write

$$\mathbf{u}_j = \Delta\mathbf{x}_j = \mathbf{x}_{j+1} - \mathbf{x}_j, \tag{5.41}$$

$$\mathbf{v}_j = \Delta^2\mathbf{x}_j = \mathbf{u}_{j+1} - \mathbf{u}_j. \tag{5.42}$$

For a fixed integer $k$ (to be determined), we define $N \times k$ matrices whose columns are the vectors of differences

$$U \equiv \mathbf{U}_k = [\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}], \tag{5.43}$$

$$V \equiv \mathbf{V}_k = [\mathbf{v}_0, \mathbf{v}_1, \ldots, \mathbf{v}_{k-1}]. \tag{5.44}$$

Note that

$$\mathbf{u}_{j+1} = A\mathbf{u}_j = A^{j+1}\mathbf{u}_0 \tag{5.45}$$

and

$$\mathbf{v}_j = (A - I)\mathbf{u}_j \tag{5.46}$$

for each $j$.
The method to be considered calculates $\mathbf{s}$ as a weighted average of the $\mathbf{x}_j$'s, with weights determined by the coefficients of the *minimal polynomial* $P(\lambda)$ of $A$ with respect to $\mathbf{u}_0$, i.e. the unique monic polynomial of least degree such that

$$P(A)\mathbf{u}_0 = 0. \tag{5.47}$$

For the present, $k$ is taken to be the degree of $P$. Then $k \le N$, and it may happen that $k$ is much smaller than $N$. We observe that

$$(A - I)(\mathbf{x}_j - \mathbf{s}) = \mathbf{u}_j, \tag{5.48}$$

which follows immediately from the definitions.
If $k$ is taken to be $N$ in (5.41) and (5.42), and if $V$ is invertible, then it follows from (5.46) that

$$(I-A)^{-1} = -UV^{-1},$$

and hence from (5.48) that

$$\mathbf{s} = \mathbf{x}_j - UV^{-1}\mathbf{u}_j \tag{5.49}$$

for any $j$, in particular for $j = 0$. This is a "full rank extrapolation"; it requires inversion of a rank $N$ matrix, and its only advantage over (5.39) is that $A$ and $\mathbf{b}$ still need not be known explicitly. However, (5.49) may be rewritten as a pair of equations

$$\mathbf{s} = \mathbf{x}_0 + U\xi, \tag{5.50}$$

$$\mathbf{0} = \mathbf{u}_0 + V\xi. \tag{5.51}$$

Now we may ask if there is a simultaneous solution $\xi$ of (5.50) and (5.51) when $U$ and $V$ have $k$ columns, where $k$ may be less than $N$. The answer is "yes", and for $k = \deg P$. In fact, if we set

$$\xi_j = b_j/P(1), \quad 0 \leq j \leq k-1,$$

then it follows (see [SFS87]) that $\xi$ satisfies (5.50), and from (5.46) and (5.48) we have

$$V\xi = (A-I)U\xi = (A-I)(\mathbf{s}-\mathbf{x}_0) = -\mathbf{u}_0,$$

so (5.51) is also satisfied. Now $\xi$ may be computed directly from (5.51), consistency having been demonstrated, and the result substituted in (5.50) to get the "reduced rank" equivalent of (5.49)

$$\mathbf{s} = \mathbf{x}_0 - UV^{+}\mathbf{u}_0, \tag{5.52}$$

where $U^{+}$ is the pseudo-inverse (or Moore-Penrose generalized inverse) of $U$ (see [LH74] or [Str76]). Note that this requires the equivalent of an inversion of a $k \times k$ matrix. We summarize the algorithm as follows:

*Reduced Rank Extrapolation* (RRE). Given a sequence generator of the form (2.1) and a starting vector $\mathbf{x}_0$,

1. generate $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k+1}$;
2. compute $U, \mathbf{u}_k$ as in (5.41) and (5.43);
3. compute $V$ as in (5.42) and (5.44);
4. compute the generalized inverse $V^{+}$;
5. compute $\mathbf{s}$ from (5.52).

Note that $\mathbf{v}_k$ is not used, so $\mathbf{x}_{k+2}$ is not needed. One might expect that second order differences of a slowly convergent sequence would cause more numerical problems than first order differences, and this is sometimes the case.
If we write $\Delta X$ for $U$ and $\Delta^2 X$ for $V$ (where $X$ is a matrix whose columns are the iterates $\mathbf{x}_j$, and $\Delta$ is applied to columns), then (5.52) becomes

$$\mathbf{s} = \mathbf{x}_0 - \Delta X(\Delta^2 X)^{+}\Delta\mathbf{x}_0.$$

In this form it is clear that RRE is a natural extension of Aitken's $\Delta^2$ method [Ait37] to vector problems.

### 5.4.2 Extension of the RRE to nonlinear sequences

We turn our attention to a sequence $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots$ generated by

$$\mathbf{x}_{j+1} = F(\mathbf{x}_j), \tag{5.53}$$

where $F$ is a vector-valued function of real vectors, defined on an open and connected domain $D$ in $N$-space, and which has a Lipschitz continuous derivative. If $\mathbf{s} = F(\mathbf{s})$ is a fixed point in $D$, and $F'(\mathbf{s})$ is the Jacobian matrix of $F$ at $\mathbf{s}$, then

$$F(\mathbf{x}) - \mathbf{s} = F'(\mathbf{s})(\mathbf{x} - \mathbf{s}) + O\left(\|\mathbf{x} - \mathbf{s}\|^2\right), \tag{5.54}$$

for all $\mathbf{x} \in D$. Not that this context includes that of the linearly generated sequences studied in the previous section, as well as sequences generated by more general iterations of the form,

$$\mathbf{x}_{j+1} = A\mathbf{x}_j + \mathbf{b} + \mathbf{e}_j, \tag{5.55}$$

where $\mathbf{e}_j$ is any small "error" that approaches $\mathbf{0}$ quadratically as $j \to \infty$. If $\mathbf{e}_j$ is the actual error in the linearly generated sequence (5.38), for example, from use of finite precision arithmetic, there is no reason to think that $\mathbf{e}_j \to \mathbf{0}$. Nevertheless, we will see that the technique to be described here for nonlinear sequences has some practical application to linear sequences as well. That is, it has the effect of "squeezing out" small errors in the linear computations in many situations that are not covered by the theoretical development.
We assume as before that the Jacobian matrix $F'(\mathbf{s})$ does not have 1 as an eigenvalue, and we let $k$ denote the degree of the minimal polynomial of $F'(\mathbf{s})$ with respect to $\mathbf{x}_0 - \mathbf{s}$. The RRE may now be extended by "cycling" to generate a sequence $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \dots$ of approximations to $\mathbf{s}$, in much the same way that scalar iterations by iterating Aitken's $\Delta^2$ method (Steffenson's method).

(*Nonlinear extrapolation algorithm.*)

1. Set $\mathbf{s}_0 = \mathbf{x}_0$, the given starting vector, and $i = 1$.
2. Generate $k+1$ of vectors $\mathbf{x}$ by (5.53)
3. Apply the RRE to these iterates to compute a vector $\mathbf{s}^*$.
4. Set $\mathbf{s}_i = \mathbf{s}^*$, increase $i$ by 1, replace $\mathbf{x}_0$ by $\mathbf{s}^*$, and return to step 2.

Each time through this loop is called a *cycle*. The principal "result" concerning this cycled extrapolation algorithm is that it is quadratically convergent, in the following sense.

**"Theorem" 5.1.** *Under the assumptions stated above on $F$ and $\mathbf{s}$, if $k$ is chosen on the $i$th cycle to be the degree of the minimal polynomial of $F'(\mathbf{s})$ with respect to $\mathbf{s}_{j-1} - \mathbf{s}$, and if $\mathbf{s}_0$ is sufficiently close to $\mathbf{s}$ in $D$, then*

$$\|\mathbf{s}_{j+1} - \mathbf{s}\| = O\left(\|\mathbf{s}_i - \mathbf{s}\|^2\right). \tag{5.56}$$

In [SFS87] some implementation strategies are recommended to apply the RRE to the "base sequence" $\{\mathbf{x_j}\}$, which in our particular case is given by (5.37). In the following subsection it is shown, how the vector extrapolation works on the sequences generated by the Gummel-maps $\Phi_1$ and $\Phi_2$.

### 5.4.3 Application to the Gummel-type iterations

Let $\mathbf{x}_0 = (w_1^{(l)}, w_2^{(l)})$ be a discrete iterate solution defined in step 4 of the Gummel-type iteration $\Phi_1$ shown in figure 3. We generate a sequence of vectors $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \dots$ by

$$\mathbf{x}_j = (w_1^{(l+j)}, w_2^{(l+j)}) = \Phi_1(w_1^{(l+j-1)}, w_2^{(l+j-1)}). \tag{5.57}$$

(*Nonlinear extrapolation algorithm for* $\Phi_1$ .)

1. Set $\mathbf{s}_0 = (w_1^{(l)}, w_2^{(l)})$, the given starting vector, and set $i$ and $j$ to 1.

2. Generate a vector $(w_1^{(l+j)}, w_2^{(l+j)})$ by (5.57) and increase $j$ by 1.
   If $\|(w_1^{(l+j)}, w_2^{(l+j)}) - (w_1^{(l+j-1)}, w_2^{(l+j-1)})\|_\infty < \eta$ then stop, otherwise, if $j$ equals $k$ proceed with step 3, otherwise return to 2.

3. Apply the RRE to the iterates to compute a vector $(w_1^*, w_2^*)$.

4. Set $\mathbf{s}_i = (w_1^*, w_2^*)$, increase $i$ by 1, set $j$ to 1, replace $(w_1^{(l)}, w_2^{(l)})$ by $(w_1^*, w_2^*)$, and return to step 2.

Figure 5: pseudo code notation of the nonlinear extrapolation algorithm for $\Phi_1$

One application of $\Phi_1$ means to run through the global loop of the algorithm in figure 3 *once*. After $k$ cycles of the global loop, which generates the $k+1$ base sequence vectors

$$(\mathsf{x}_0, \ldots, \mathsf{x}_k) = ((w_1^{(l)}, w_2^{(l)}), \ldots, (w_1^{(l+k)}, w_2^{(l+k)})),$$

we apply the RRE to these vectors to compute a vector $\mathbf{s}^* = (w_1^*, w_2^*)$. The result is then used in step 1 instead of the last iterate $(w_1^{(l+k)}, w_2^{(l+k)})$ and another $k$ cycles of the global loop are performed (if needed). Again, the RRE is applied, and the base iteration is restarted from this result and so on. The error

$$\|(w_1^{(l+j)}, w_2^{(l+j)}) - (w_1^{(l+j-1)}, w_2^{(l+j-1)})\|_\infty$$

in the global loop is monitored in each cycle and not only in the cycles, where the RRE is applied. Therefore, exiting the global loop may happen between two applications of the RRE. We summarize this in figure 5.

If the discrete Poisson equation is not solved sufficiently accurate or if the parameter $\alpha$ is set to values too small for effective damping, then the base sequence appears to diverge. In this case no super-linear convergence of the transformed sequence can be expected and only a few base sequence vectors are generated to avoid numerical difficulties with data ranging over many orders of magnitude.

Up to now nothing was said about determination of $k$, the degree of the minimal polynomial. In fact, there is no practical way to determine $k$ in advance. Fortunately, it is not necessary to do so [SFS87]. Even poor approximations to $k$ can lead to good approximations to $\mathbf{s}$.

An appropriate way to implement RRE on the first cycle is to extrapolate to $\mathbf{s}_{m,k}$ from $\mathbf{x}_m, \mathbf{x}_{m+1}, \ldots, x_{m-k+1}$ with $k = 1, 2, 3, \ldots$, and to stop when the least square residuals are acceptably small. The decline in residuals with increasing $k$ is gradual. (When there is a strong separation between the "dominant" and the "small" eigenvalues, there is often a precipitous drop in the magnitudes of these residuals when $k$ reaches the number of dominant ones (multiplicities included)). In the simulations a value of $k$ is accepted if the least square residual is smaller than $10^{-6} \ldots 10^{-4}$, some values proved to be useful. The actual values of $k$ are given in section 6 for every device model.

The pseudo-inverse $V^+$ in (5.52) is not actually computed. Instead, we solve (5.51) as a least-square problem, where the solution $\xi$ minimizes $\|\mathbf{u}_0 + V\xi\|_2$, and by substitution of this minimizer $\xi$ into (5.50) we obtain an approximation of the fixed point $\mathbf{s}$. The iteration is stopped, if the relative residual is smaller than a given tolerance,

$$\frac{\|\mathbf{u}_0 + V\xi\|_2}{\|\mathbf{u}_0\|_2} < 10^{-12} \ldots 10^{-10}.$$

The RRE can also be applied to the vector sequences generated in the inner iteration for solving the nonlinear Poisson equation, step 2 of the iteration shown in figure 3. Such a modification is needed for the algorithm to be implementable. The global convergence of this modified iteration is accelerated, if carefully designed. Since the convergence of the global iteration is very sensitive to the convergence of this inner iteration, one must generate enough base sequence vectors to ensure rapid convergence of the extrapolated values. On the other hand the number of base sequence vectors should be tightly limited, since the convergence is only linear and in every inner cycle a linear system of size $N \times N$ must be solved.
The above comments on the use of the RRE combined with the Gummel-type iteration $\Phi_1$ apply to the iteration $\Phi_2$ as well. The REE is used in both inner loops, represented by step 2 and 4 of the iteration shown in figure 4, yielding much more rapid convergence of the global iteration.
The actual value of the number of base sequence vectors $k_1$ in the inner iteration is also given in section 6 for every device model under investigation.

**Remark 5.7.** It is not assured that the reduced rank extrapolation always produces positive discrete solutions, i.e. it may introduce negative values even if all vectors involved in the extrapolation have positive components. In the comprehensive numerical tests, however, no such difficulties occured, which justifies the application of RRE to our problems.

# 6 Numerical tests and simulations

In this section the focus is shifted to the actual implementation of the finite-element method and the iterative solvers. We start this section with some brief remarks on the programming. Some pieces of source code are shown to make clear how the finite elements and the iteration schemes are actually implemented. Then the description of the parameters of the numerical tests and a presentation of the test devices follows. All tests are accomplished with one-dimensional and two-dimensional device models. The results are used to validate the simulation results that are done with three-dimensional device models. We apply the full Newton method with step control and the Gummel-type iteration method combined with the reduced rank vector extrapolation.
In the following subsection we mention some of the details of the source code that was used to produce the numerical results presented here.

## 6.1 Programming and technical notes

The tests and simulations are accomplished with programs written by the author of this thesis. The source code is programmed in MATLAB® . All of the algorithms presented here (step control, decoupling procedures, Newton iteration etc.) are vectorized as far as possible by use of the MATLAB® matrix operations to make the programs run faster. The most often used predefined MATLAB® commands are listed in table 1.
The matrix assembling and the solution of the linear systems have to be done numerous times.

| command | description or purpose |
|---|---|
| `accumarray` | construct an array by accumulation |
| `cholinc` | incomplete Cholesky decomposition, produce preconditioner |
| `find` | find indices of nonzero elements |
| `kron` | Kronecker tensor product |
| `mldivide` or \ | left matrix divide, LU factorization or Cholesky factorization |
| `pcg` | preconditioned conjugate gradients method |
| `sparse` | create sparse matrix from index vectors |

Table 1: Useful predefined MATLAB® commands

The assembling is based on low level commands as `accumarray`, `sum` or `sort`, which results in efficient source code.
As an example of source code in figure 6 the implementation of the vector extrapolation using least squares solution command `lsqr` is shown. If the statement in line 1 is true, then the base sequence vector is kept in `X` (line 2), otherwise the else branch is entered. There, `U` and `V` are the first and second order difference matrices (5.43) and (5.44), resp. `X` is a $N \times k$ matrix of base sequence vectors. In line 7 the method `lsqr` gets these matrices in its first and second argument, followed by the error tolerance and a maximum number of iterations allowed. In line 8 the equation (5.49) is solved and finally the error is computed in line 9.

## 6.2 Parameters of numerical tests

In this section we state the common parameters used in all of the numerical tests and simulations. The section concludes with a flow chart (see figure 7) that shows the main steps of the algorithms used in the simulations.

### 6.2.1 Computation of the equilibrium solution and extension of applied voltage

In all tests and simulations the nonlinear Poisson equation must be solved first to get the equilibrium solution, which is used as a starting point of the global iteration. The nonlinear Poisson

```
% ================== reduced rank vector extrapolation ==================
1  if mod(l,k+1) < k
2     X(:,mod(l,k+1)+1) = [w1;w2];
3  else
4     X = [X(:,2:end),[w1;w2]];
5     U = diff(X(:,1:k-1),1,2);
6     V = diff(X,2,2);
7     y = lsqr(V,-U(:,1),1e-12,1e3);
8     s = X(innerFacesTwo,1)+U(innerFacesTwo,:)*y;
9     theError = norm(s-[w1(innerFaces);w2(innerFaces)],inf);
10 end
% =======================================================================
```

Figure 6: A source code example

equation is solved by means of the full Newton method or by means of the linearized Newton method, where the linear problems are solved with the preconditioned conjugate gradient method. An incomplete Cholesky factorization with a problem dependent drop tolerance is used as preconditioner. The MATLAB® command `cholinc` implements this factorization.
The extension $U$ of the applied voltage $V_A$ that was defined as the solution of the boundary value problem (5.1) is computed by means of the preconditioned conjugate gradient method. Thereby, we make use of the same preconditioner.

### 6.2.2 Approximation of diffusion coefficients

Until now nothing is said about the local approximation operator (4.1) actually used in the numerical simulations. In this subsection we specify this.
For linear problems with sufficiently smooth diffusion coefficients it is shown in [AB85] that the Lagrange multipliers on the faces approximate the solution more accurate if a post-processing is applied. We also expect this for the Lagrange multipliers $\mathsf{V}_h$, $\mathsf{w}_{1h}$, $\mathsf{w}_{2h}$ and therefore use the following local approximation operator in place of (4.1),

$$\tilde{\Pi}_K \,:\, L^2(\mathcal{E}(K)) \to \mathcal{P}_0, \quad \tilde{\Pi}_K \mu = \left( \sum_{e \in \mathcal{E}(K)} |e|^2 \right)^{-1} \sum_{e \in \mathcal{E}(K)} |e|^2 \mu_{|e}.$$

It takes the geometry of the elements into account, where faces of large area do contribute above average. This choice is not straightforward, instead of averaging over faces averaging over the edges is also possible. Even simple arithmetic mean values do work, and the results differ only slightly.

### 6.2.3 Preconditioning and scaling

The matrices arising in both the Newton method and the Gummel-type methods are generally badly conditioned, especially for three-dimensional domains $\Omega$. A simple scaling of rows and columns improves the condition of the systems. The Jacobian matrix is scaled by dividing each line by the corresponding diagonal entry, which is assumed to be nonzero. There is no guarantee for this assumption to hold in general, but the method worked in all simulations presented here.
For the Gummel-type methods the situation is more complicated, since we aim at retaining the positive definiteness and the symmetry of the matrices. We multiply the rows and the columns by the square root of the entry of the corresponding diagonal elements, which is positive in general. The modified system is then solved by means of the Cholesky factorization.

### 6.2.4 Integration and smoothing of doping profile

For the integration of the doping profile function Dop on the partition elements $K \in \mathcal{T}_h$ the MATLAB® command `triplequad` has been used. It implements an adaptive Simpson quadrature. Since it evaluates the triple integral of a function over three dimensional rectangular regions only, we integrate over the unit cube $(0,1)^3$ and map from $(0,1)^3$ into $\Delta^3$ and then $\Delta^3$ into $K$. Here $\Delta^3$ denotes the three-dimensional (open) standard simplex

$$\Delta^3 = \{(\hat{x},\hat{y},\hat{z}) \in (0,1)^3 : \hat{x}+\hat{y}+\hat{z} < 1\}.$$

To map the reference elements we make use of the Jacobi transform,

$$(0,1)^3 \ni (\hat{x},\hat{y},\hat{z}) \mapsto (\hat{x}(1-\hat{y})(1-\hat{z}), \hat{x}\hat{y}(1-\hat{z}), \hat{x}\hat{z}) \in \Delta^3,$$

$$\int\limits_{\Delta^3} \mathrm{Dop}(\hat{x},\hat{y},\hat{z})\, d(\hat{x},\hat{y},\hat{z}) = \int\limits_{(0,1)^3} \hat{x}^2(1-\hat{z})\mathrm{Dop}(\hat{x}(1-\hat{y})(1-\hat{z}), \hat{x}\hat{y}(1-\hat{z}), \hat{x}\hat{z})\, d(\hat{x},\hat{y},\hat{z}).$$

To improve the convergence of the iterative methods we smooth the doping profile function. More precisely, we average for all $K \in \mathcal{T}_h$ the approximations of $\mathrm{Dop}_K = \int_K \mathrm{Dop}(x,y,z)\, d(x,y,z)$ with those of the neighboring tetrahedrons. This ensures that the smoothing depends on the mesh parameter $h$, such that for the values resulting from this modification, $\overline{\mathrm{Dop}}_K$, it holds

$$\sum_{K \in \mathcal{T}_h} \overline{\mathrm{Dop}}_K \to \int\limits_{\Omega} \mathrm{Dop}(x,y,z)\, d(x,y,z), \quad (h \to 0).$$

### 6.2.5 Continuation in the applied bias and path-following

In section 5 a path following method was proposed to trace the solution path. In the numerical simulations, a fixed step size $\tau$ is used, and its value depends on the device to be simulated and the dimension of the domain. The coefficients $c$ and $\gamma$ of system (5.13),(5.14) are set to the values

$$c = -(\underbrace{\overline{c}/N,\ldots,\overline{c}/N}_{2N \text{ entries}}, \underbrace{0,\ldots,0}_{N \text{ entries}})^T, \quad \gamma = 1,$$

in all tests and simulations, where $N$ denotes the number of degrees of freedom and $\overline{c}$ is a non-negative constant independent of $N$. The actual value of $\overline{c}$ varies with the device to be simulated. This choice of the coefficient vector $c$ has the effect of taking the nodal values of $\widetilde{V}_h$ and $\widetilde{w}_{1h}$ in the system (5.2)-(5.7) into account by averaging these values, while the nodal values of $\widetilde{w}_{2h}$ do not enter. If $\overline{c}$ is set to zero, then the step control reduces to the continuation in the applied bias with an increment that is only controlled by the step length $\tau$.

### 6.2.6 Grid generation and local refinement

The properties of the underlying grid determine the quality of the discrete solution. Often, the preservation of maximum principles is demanded. The M-matrix property guarantees the respect of the discrete maximum principle. A matrix is said to have the M-matrix property, if it is a nonsingular matrix with $m_{ii} > 0$ and $m_{ij} \leq 0$. The matrix of the resulting system obtained with mixed hybrid finite element (4.25) is symmetric and positive definite but is not, in general, an M-matrix [RT77]. The lowest order Raviart-Thomas space defined over triangles (resp. tetrahedrons) is equivalent to a modified $P1$-nonconforming Galerkin method [BF91]. In this case, the obtained matrix with mixed-finite elements is an M-matrix if the triangles and tetrahedrons are weakly acute, i.e. all angles and dihedral angles, resp., are less than $\pi/2$.
The two-dimensional reference solutions have been computed on triangular grids that have been generated with the MATLAB® command `initmesh`, which has to be given a geometry description file. The quality of a triangle $K$ is measured by

$$q = \frac{4\sqrt{3}|K|}{l_{12}^2 + l_{13}^2 + l_{23}^2},$$

where $l_{12}$, $l_{13}$ and $l_{23}$ denote the lengths of the triangle edges. If $q > 0.6$ the triangle is of acceptable quality. If $q \leq 0.6$ then the triangle $K$ is further refined.
The function `initmesh` implements a Delaunay triangulation algorithm [Geo91]. The Delaunay triangulation is a relaxation of weakly acute triangulations, since it allows for the presence of obtuse angles but it maximizes the minimum angle. Compared to any other triangulation of the points, the smallest angle in the Delaunay triangulation is at least as large as the smallest angle in any other. A Delaunay triangulation is conforming (or admissible), i.e. no hanging nodes are present. We use this fact later to define globally continuous, piecewise linear interpolants.
The three-dimensional domain $\Omega$ is decomposed into tetrahedrons by a grid generator implemented in FEMLAB® . The quality of a tetrahedron $K$ is measured by

$$q = \frac{72\sqrt{3}|K|}{(l_{12}^2 + l_{13}^2 + l_{14}^2 + l_{23}^2 + l_{24}^2 + l_{34}^2)^{3/2}}, \quad l_{ij} = \|\tau_{ij}\|_2,\ i = 1,2,3,\ j = i+1,\dots 4,$$

and $K$ is accepted, if $q > 0.1$. As in the two-dimensional case, this is a weaker condition than the non-obtuseness of $K$.

### 6.2.7 Postprocessing

The discrete solutions $w_{1h}, w_{2h}$ in (4.6) and (4.7) are piecewise constant approximations of $w_1$ and $w_2$. In a postprocessing step we can increase the approximation order by using the Lagrange multipliers $\mathsf{w}_{1h}$ and $\mathsf{w}_{2h}$, which are only defined on the union of all edges of the triangles (of all faces of the tetrahedrons). They are piecewise constant on the triangle edges (tetrahedron faces), and from these values we compute values at the vertices of the triangles (tetrahedrons) in the following way: we fix a face $e$, and a vertex $p$ and define the real number

$$m_{ep} = \sum_{\{p\} \subsetneq e \cap e'} |e'|^2,$$

i.e. we sum up the squares of the areas of all faces that share an edge that contains the vertex $p$. Further we set for the values $\mathsf{w}_{1e}$ of the Lagrange-multiplier $\mathsf{w}_{1h}$ at the face $e$

$$b_e = \mathsf{w}_{1e} \sum_p m_{ep}.$$

The problem to be solved is

$$(m_{ep})(w_{1p}) = (b_e),$$

which gives the least square solution

$$(w_{1p}) = (m_{ep})^+(b_e),$$

With these values at the vertices $p$ we define a global discrete solution by means of the Lagrange interpolation for $\mathsf{w}_1$, and in the same way for $\mathsf{w}_2$.

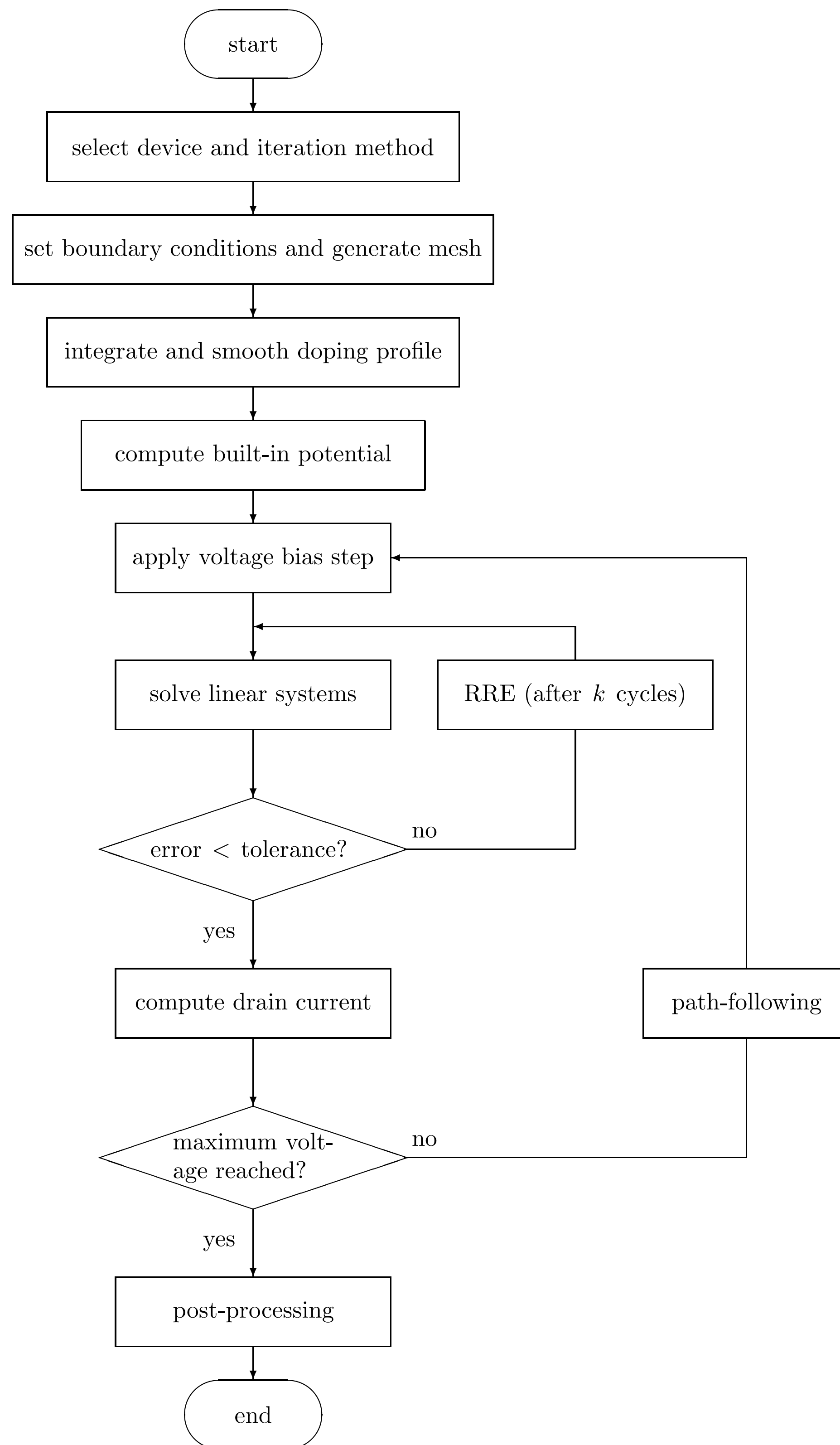



Figure 7: flow chart of the device simulations

## 6.3 Test of 1D devices

Now we turn to the presentation of the one-dimensional examples. They are simplifications of the quasi 1D diode of section 6.5.1. For the calculation of the terminal currents we use the width and the height of these three-dimensional devices.
The numerical results are obtained with MATLAB® programs similar to those used in the three-dimensional case. The same finite elements have been used.
In this section, all numerical errors have been computed with respect to the discrete maximum norm, unless otherwise specified.

### 6.3.1 Ballistic diode

The following one-dimensional $n^+nn^+$ ballistic silicon diode is used as a benchmark for unipolar devices. It is a simple model for the channel of a MOS transistor. The semiconductor domain is $\Omega = (0, l_x)$, where $l_x = 0.6\,\mu$m. The channel length is equal to $0.4\,\mu$m. The highly-doped regions are $\Omega_L = (0, 0.1\,\mu\text{m})$ and $\Omega_R = (0.5\mu\text{m}, l_x)$. The intermediate part $\Omega \setminus (\Omega_L \cup \Omega_R)$ is the lightly-doped $n$ region. The device has two Ohmic contacts, the source contact and the drain contact,

$$\Gamma_{\mathsf{Source}} = \{0\}, \quad \Gamma_{\mathsf{Drain}} = \{l_x\}.$$

The doping profile function is given by

$$\text{Dop}(x) = \begin{cases} 5 \cdot 10^{17}\,\text{cm}^{-3}, & x \in \Omega_L \cup \Omega_R, \\ 2 \cdot 10^{15}\,\text{cm}^{-3}, & \text{else.} \end{cases}$$

At the source contact 0 V are applied, at the drain contact 1.5 V are applied. These values are taken from [DJP00] and this allows for a comparison with the results presented in this paper. For this diode model the scaled Debye length is $\lambda = 9.4832 \cdot 10^{-3}$.
The tests are accomplished on uniform grids with a maximum grid parameter $h \approx \lambda$ in order to resolve the small scaled Debye length. The matrices $\mathbf{A}_{ij}$ in (4.12) are tri-diagonal and hence the linear systems can be efficiently solved by direct elimination. The results presented in the figures are the output of the Gummel-type method $\Phi_1$ combined with the vector extrapolation described in section 5.4.3 on a uniform grid with parameter $h = 5 \cdot 10^{-3}$. Moreover, the iteration $\Phi_2$ have been tested on this device model. The values of the parameters for $\Phi_1$ and $\Phi_2$ are stated in table 2. We briefly recall the meaning of $k$ and $k_1$. $k$ is the number of base sequence vectors generated in the global loop, while $k_1$ is the number of sequence vectors from the inner loop (linearized Newton iteration for solving the Poisson equation). The full Newton iteration is used to compare the convergence of both schemes. The Newton method is applied without step control in this situation, since no beneficial effect has been observed. In figure 8 and 9 we see the super-linear convergence of the Newton method and the Gummel-type iterations, resp. For the Newton method, a bias step size of $\tau = 0.33$ is chosen. The iteration is stopped, if the norm of the Newton correction $\|s^{(k)}\|_\infty$ is smaller than $\eta = 1 \cdot 10^{-6}$. By increasing the voltage, the device is "driven away" from thermal equilibrium, and this causes a monotone increase in the number of steps in all of the iterations to reach at the given error tolerance.
Now we turn to the numerical results. The current-voltage characteristic is shown in figure 10. For the calculation of the current we take $0.2\mu m$ for the device width and $0.1\mu m$ for the device height. The dependence of the terminal current on the applied voltage is almost linear.
The distribution of the electron temperature depicted in figure 11 shows a maximum value of approx. $T_{\mathsf{max}} = 2340\,\text{K}$ at $x = 0.44\,\mu$m and a minimum value of approx. $T_{\mathsf{min}} = 280\,\text{K}$ at $x = 0.12\,\mu$m, i.e. when the electrons are injected from the highly-doped region to the lightly-doped region, the electron temperature is slightly lower than the lattice temperature. Because of the boundary conditions imposed, the electron temperature at the contacts is the same as the lattice temperature. The electrons emanate from the source contact and pass the highly-doped region. In the lightly-doped channel they move more freely and gain energy from the electrical forces. The velocity rapidly increases to a value of approx. $1.2 \cdot 10^7\,\text{cm}\,\text{s}^{-1}$. The velocity overshoot

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 0.0 | $0.5\ldots1$ | $5\ldots85$ | $5\ldots15$ | $2.5\cdot10^{-2}$ | $1\cdot10^{-8}$ | $1\cdot10^{-4}$ | 0.7 | $1\cdot10^{-6}$ |
| $\Phi_2$ | 3.0 | $0.5\ldots1$ | $10\ldots50$ | $10\ldots15$ | $2.5\cdot10^{-2}$ | $1\cdot10^{-8}$ | $1\cdot10^{-4}$ | 0.7 | $1\cdot10^{-6}$ |

Table 2: Parameter values in the test of the ballistic diode ($k$ and $k_1$ are the numbers of base sequence vectors)

| $h$ | RE for $V_h^I$ | RE for $w_{1h}^I$ | RE for $w_{2h}^I$ | RE for $T_h^I$ | RE for $n_h^I$ | RE for $g_h^I$ |
|---|---|---|---|---|---|---|
| $10^{-2}$ | $9.9\cdot10^{-4}$ | $1.8\cdot10^{-3}$ | $1.7\cdot10^{-3}$ | $2.5\cdot10^{-3}$ | $6.9\cdot10^{-3}$ | $7.9\cdot10^{-3}$ |
| $5\cdot10^{-3}$ | $3.8\cdot10^{-4}$ | $6.6\cdot10^{-4}$ | $7.0\cdot10^{-4}$ | $9.2\cdot10^{-4}$ | $2.8\cdot10^{-3}$ | $3.1\cdot10^{-3}$ |
| $2.5\cdot10^{-3}$ | $1.8\cdot10^{-4}$ | $2.9\cdot10^{-4}$ | $3.2\cdot10^{-4}$ | $4.1\cdot10^{-4}$ | $1.2\cdot10^{-3}$ | $1.3\cdot10^{-3}$ |

Table 3: Relative errors from simulation of the one-dimensional ballistic diode model

at $x = 0.46\,\mu$m with a maximum value of $1.43\cdot10^7\,\mathrm{cm\,s^{-1}}$ is spurious and is introduced by the model. No such velocity overshoot has been observed in Monte-Carlo simulations that are based on the Boltzmann equation. The electron density is shown in figure 12.

The electrostatic potential shown in figure 13 has a slight decrease along the junction where the electrons are injected from the highly-doped region to the lightly-doped region. Then it increases nonlinearly and reaches the saturation value at the drain contact. When the bias field is larger than the built-in field which results from the distribution of electron density, the slight decrease disappears and the electrostatic potential increases directly and nonlinearly from the source to the drain.

It remains to calculate the relative errors. We use piecewise linear interpolation. By means of this interpolation, we obtain discrete solutions $(V_h^I, w_{1h}^I, w_{2h}^I)$, where the superscript $I$ denotes the interpolation. $T_h^I$ is the piecewise linear interpolant of the nodal values of $T_h = -w_{2h}^{-1}$. The discrete electron density $n_h^I$ is the piecewise linear interpolant of the nodal values of $V_h, w_{1h}, w_{2h}$ according to the definition (2.75). The discrete thermal energy $g_h^I$ is the piecewise linear interpolant of the nodal values of $\frac{3}{2}n_hT_h$. The relative error is defined by

$$\mathrm{RE}(V_h^I) = \frac{||V_h^I - V^*||_{0,2,\Omega}}{||V^*||_{0,2,\Omega}}.$$

$V^*$ is a reference solution on a fine grid with $h = 10^{-4}$.

In table 3 the relative errors for the solution obtained with the Gummel-type iteration $\Phi_1$ are shown. It is clear from the relative errors that the discrete electron temperature $T_h^I$, the discrete electron density $n_h^I$ and the discrete thermal energy $g_h^I$ are less good approximations compared to $V_h^I, w_{1h}^I, w_{2h}^I$, since they are obtained by a simple post-processing.

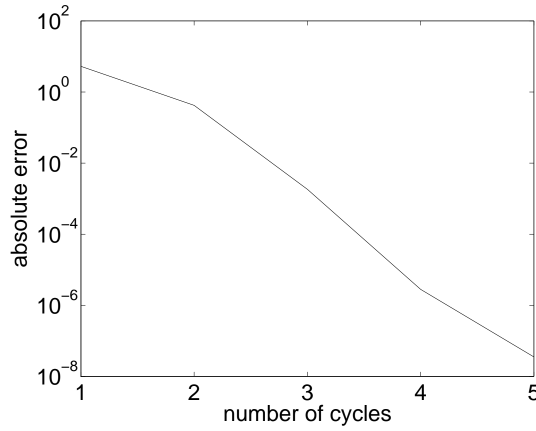



Figure 8: convergence of the full Newton method, $V_{\mathsf{Drain}} = 1.5V$

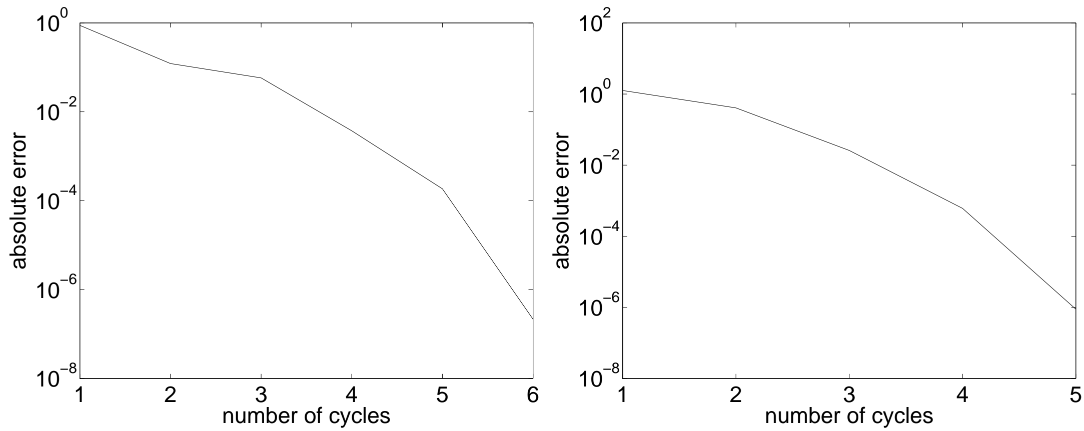



Figure 9: convergence of the fixed-point methods $\Phi_1$ (left) and $\Phi_2$ combined with vector extrapolation, $V_{\mathsf{Drain}} = 1.5V$

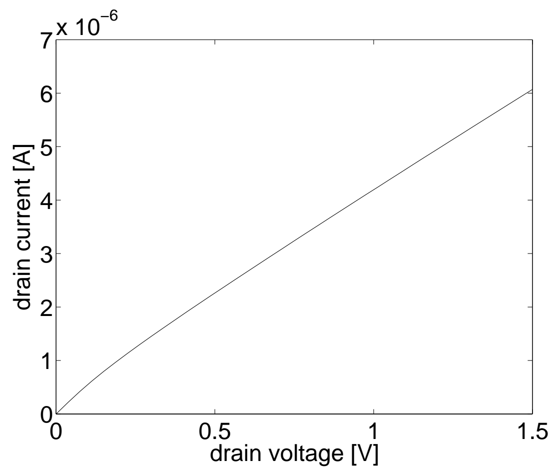



Figure 10: current-voltage characteristic of a one-dimensional ballistic diode

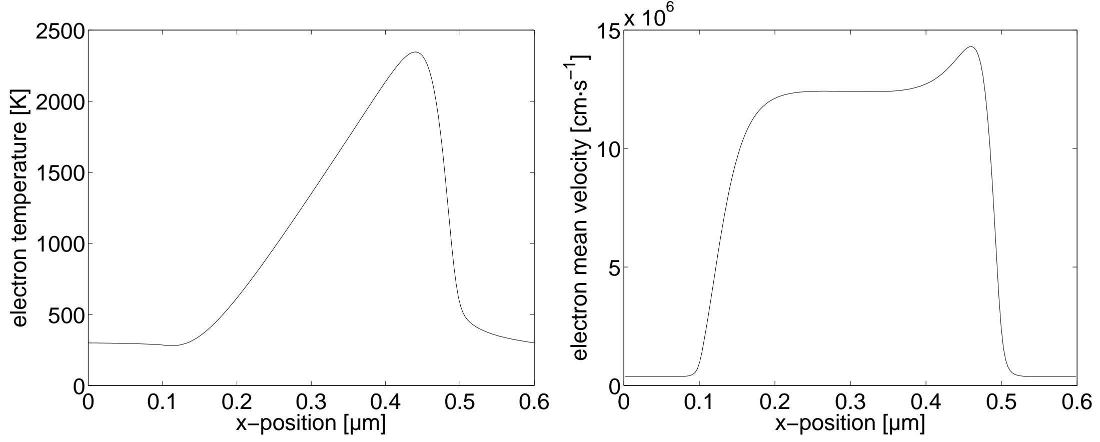



Figure 11: electron temperature and electron mean velocity in the one-dimensional ballistic diode

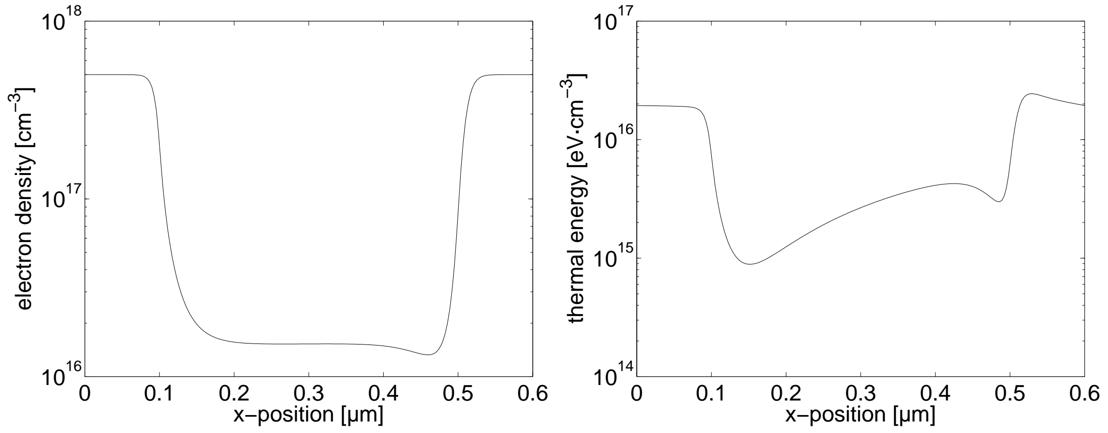



Figure 12: electron density and thermal energy in the one-dimensional ballistic diode

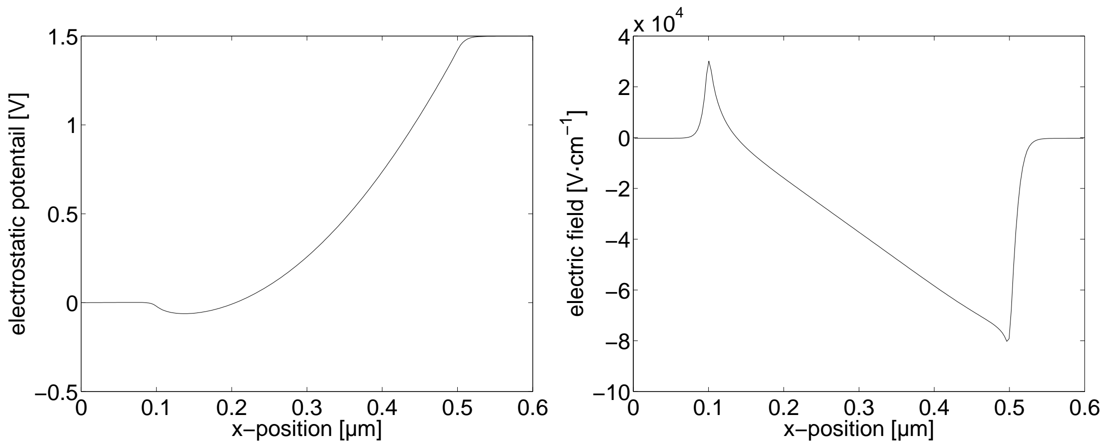



Figure 13: electrostatic potential $V$ and electric field $-\nabla V$ in the one-dimensional ballistic diode

## 6.4 Test of 2D devices

### 6.4.1 Single-gate MESFET

The MESFET (**me**tal-**s**emiconductor **f**ield **e**ffect **t**ransistor) consists of a conducting channel positioned between a source and drain contact region as shown in the figure 14. The carrier flow from source to drain is controlled by a Schottky metal gate. The control of the channel is obtained by varying the depletion layer width underneath the metal contact which modulates the thickness of the conducting channel and thereby the current between source and drain. This device is used as a switch or amplifier [Sze81].
The semiconductor domain is $\Omega = (0, l_x) \times (0, l_y)$, where $l_x = 0.6\,\mu$m and $l_y = 0.2\,\mu$m. The device consists of two highly-doped $n^+$ regions near the Ohmic contacts, the source contact and the drain contact. The source and drain contact lengths are $0.1\,\mu$m, the gate contact length is $0.2\,\mu$m.

PSfrag replacements

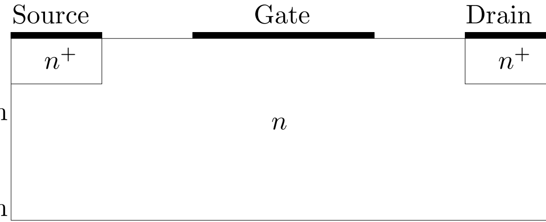



Figure 14: Geometry of the two-dimensional single-gate MESFET

The doping profile function is given by (in $\mu$m)

$$\mathrm{Dop}(x,y) = \begin{cases} 3 \cdot 10^{17}\,\mathrm{cm}^{-3} & (x,y) \in [0,0.1] \times [0.15,0.2] \cup [0.5,0.6] \times [0.15,0.2] \\ 1 \cdot 10^{17}\,\mathrm{cm}^{-3} & \text{else.} \end{cases}$$

With an applied voltage $V_A = 2\,\mathrm{V}$ at the drain and $V_A = 0$ (open state) at the gate as well as a barrier voltage $V_B = -0.8\,\mathrm{V}$ at the gate, the boundary data is given as follows,

- at the source: $n = 3 \cdot 10^{17}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}}$;
- at the drain: $n = 3 \cdot 10^{17}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}} + 2\,\mathrm{V}$;
- at the gate:
  open state: $n = 3.9 \cdot 10^{5}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}} - 0.8\,\mathrm{V}$;
  close state: $n = 2.4 \cdot 10^{5}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = -1.2\,\mathrm{V} + V_{\mathsf{bi}} - 0.8\,\mathrm{V}$;
- at the remaining boundary segments, homogeneous Neumann boundary conditions for $I_1, I_2$, and $\nabla V$ are used.

These are the same values as in [CCJS95, JS94], but we do not prescribe the velocity on the contact parts. The values for $n$ at the gate contact has been computed from the formula (5.1-19) in [Sel84]. The scaled Debye length for this device model is $\lambda = 1.1614 \cdot 10^{-2}$.
We use the Gummel-type iterations combined with the vector extrapolation method. The values of the parameters for $\Phi_1$ and $\Phi_2$ are stated in table 4. In figure 15 the convergence of the Gummel-type iterations are shown on a grid with $h = 5 \cdot 10^{-3}$. The convergence of $\Phi_1$ is only linear, while that of $\Phi_2$ is super-linear. For most of the values of the applied voltage, the convergence of $\Phi_2$ is even quadratic. Recall that $\Phi_1$ is the fixed-point map based on the fully coupled current continuity system, and this may cause the number $k$ of dominant eigenvalues to vary over a wide range. But a number of base sequence vectors close to $k$ is necessary to have quadratic convergence. The situation is different for $\Phi_2$. In this case, only scalar systems need to be solved, and $k$ might remain almost constant from one extrapolation to the next. Once a proper value of $k$ is found for

the first cycle of the RRE, it also works on the next cycles. However, $\Phi_1$ takes less runtime than $\Phi_2$, since we need less steps of the base sequence for extrapolation.
For comparison, the Newton method combined with path following is applied. In the left diagram of figure 16 the effect of the path following is shown (solid line). For small values of the applied voltage the step control reduces the step width, while the step size is increased for higher voltages. The dashed line indicates the value of the parameter where no step control is active, i.e. the voltage is increased uniformly and $(c^{(k)})^T = 0$ in (5.14). The uniform step size in this example is $\tau^{(k)} = 5 \cdot 10^{-2}$. In the right diagram the number of Newton steps needed to achieve the prescribed error tolerance $10^{-6}$ are plotted against the values of $\rho$ determined by the step control. Unlike in the one-dimensional ballistic diode model, the dependence is not monotone.
Now we turn to the numerical results. The current-voltage characteristic of the MESFET is shown in figure 17. For the given gate voltage, the drain current initially increases linearly with drain voltage, indicating that the conductive channel acts as a constant resistor. As the drain voltage increases, however, the cross-sectional area of the conductive channel is reduced, causing an increase in the channel resistance. As a result, the current increases at a slower rate and eventually saturates. In the right picture the dependence of the drain current on the gate voltage is shown, which is almost linear.
In figure 18 the electron concentration in open state is depicted. We see a drop-off in the region just below the gate, the so-called *depletion region* of the metal-semiconductor contact. In closed state, where the gate voltage is decreased with resect to the source contact, the depletion region expands, see figure 24.
In the figures 19 and 25 the electron temperatures, which contribute to the particles total kinetic energy, show similar behavior in open and closed state. The steep temperature gradients near the gate contact indicates that different boundary conditions may be physically more adequate. One possibility is to employ Robin boundary conditions at the contacts, since they are second-order approximations of boundary conditions for the Boltzmann equation [DS99] and the energy-transport equations are itself derived from the Boltzmann equation [AD96].
For the illustrations, a mesh with parameter $h = 5 \cdot 10^{-3}\, l^*$, is used with $l^* = 6.3246 \cdot 10^{-5}\,\text{cm}$ being the device diameter. The number of triangles is 11712. The number of unknowns is 17699. The results are in good accordance with that in [HJP03].
The relative errors in the tables 5 and 6 indicate that the discretization error is rapidly reduced in the regions where the solution is smooth; these error components dominate the global error for large values of $h$, for smaller values of $h$ the error components pertaining to regions where the solution is not smooth dominates the overall discretization error. These regions are, for instance, in the vicinity of the jumps of the doping profile function or near the transition points of the Dirichlet boundary data and the Neumann boundary data. Further reduction of the mesh parameter $h$ does not substantially diminishes the relative errors, especially in closed state. In [HJP04] an adaptive mixed scheme is proposed that refines the grid locally depending on the regularity of the solution. By means of this adaptive mixed scheme the error can be reduced more uniformly. The reference solution used in the calculation of the relative errors is given on a fine grid with maximum element size $h = 2.5 \cdot 10^{-3} l^*$.
The computations of the closed state of the device is started from the numerical solution of the open state. That is the gate contact voltage of -1.2V is applied in small bias steps and the drain voltage is kept at 2V, which results in a drain current of approx. $1.4 \cdot 10^{-5}\,\text{A}$. This means that the current is reduced only to the half compared with the current produced in open state.
The convergence of the Gummel-type iterations in closed state is comparable to that in open state. The path following has no improving effect on the convergence of the Newton iteration when the device is operated in closed state.

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 5.0 | $0.5\ldots1$ | $10\ldots160$ | $10\ldots15$ | $1.33\cdot10^{-2}$ | $1\cdot10^{-9}$ | $1\cdot10^{-5}$ | 0.7 | $1\cdot10^{-6}$ |
| $\Phi_2$ | 15 | 2.0 | 100 | 15 | $1.33\cdot10^{-2}$ | $1\cdot10^{-9}$ | $1\cdot10^{-5}$ | 0.7 | $1\cdot10^{-6}$ |

Table 4: Parameter values in the test of the single-gate MESFET (open state), $k$ and $k_1$ are the numbers of base sequence vectors

| $h/l^*$ | RE for $V_h^I$ | RE for $w_{1h}^I$ | RE for $w_{2h}^I$ | RE for $T_h^I$ | RE for $n_h^I$ | RE for $g_h^I$ |
|---|---|---|---|---|---|---|
| $8\cdot10^{-2}$ | $3.2\cdot10^{-1}$ | $3.7\cdot10^{-1}$ | $2.4\cdot10^{-1}$ | $4.4\cdot10^{-1}$ | $5.1\cdot10^{-1}$ | $2.4\cdot10^{0}$ |
| $4\cdot10^{-2}$ | $4.6\cdot10^{-2}$ | $2.2\cdot10^{-1}$ | $9.5\cdot10^{-2}$ | $9.4\cdot10^{-2}$ | $1.8\cdot10^{-1}$ | $3.0\cdot10^{-1}$ |
| $2\cdot10^{-2}$ | $1.3\cdot10^{-2}$ | $7.9\cdot10^{-2}$ | $5.0\cdot10^{-2}$ | $4.1\cdot10^{-2}$ | $4.8\cdot10^{-2}$ | $5.2\cdot10^{-2}$ |
| $1\cdot10^{-2}$ | $1.1\cdot10^{-2}$ | $2.6\cdot10^{-2}$ | $3.1\cdot10^{-2}$ | $2.2\cdot10^{-2}$ | $2.4\cdot10^{-2}$ | $1.8\cdot10^{-2}$ |

Table 5: Relative errors from the simulation of the 2D single-gate MESFET model (open state)

| $h/l^*$ | RE for $V_h^I$ | RE for $w_{1h}^I$ | RE for $w_{2h}^I$ | RE for $T_h^I$ | RE for $n_h^I$ | RE for $g_h^I$ |
|---|---|---|---|---|---|---|
| $8\cdot10^{-2}$ | $1.7\cdot10^{-1}$ | $5.1\cdot10^{-1}$ | $2.3\cdot10^{-1}$ | $3.2\cdot10^{-1}$ | $4.1\cdot10^{-1}$ | $7.9\cdot10^{-1}$ |
| $4\cdot10^{-2}$ | $3.7\cdot10^{-2}$ | $2.1\cdot10^{-1}$ | $8.5\cdot10^{-2}$ | $6.7\cdot10^{-2}$ | $5.3\cdot10^{-2}$ | $4.2\cdot10^{-2}$ |
| $2\cdot10^{-2}$ | $1.8\cdot10^{-2}$ | $1.4\cdot10^{-1}$ | $5.0\cdot10^{-2}$ | $3.7\cdot10^{-2}$ | $2.4\cdot10^{-2}$ | $1.8\cdot10^{-2}$ |
| $1\cdot10^{-2}$ | $1.8\cdot10^{-2}$ | $9.5\cdot10^{-2}$ | $3.4\cdot10^{-2}$ | $2.3\cdot10^{-2}$ | $1.8\cdot10^{-2}$ | $2.1\cdot10^{-2}$ |

Table 6: Relative errors from the simulation of the 2D single-gate MESFET model (closed state)

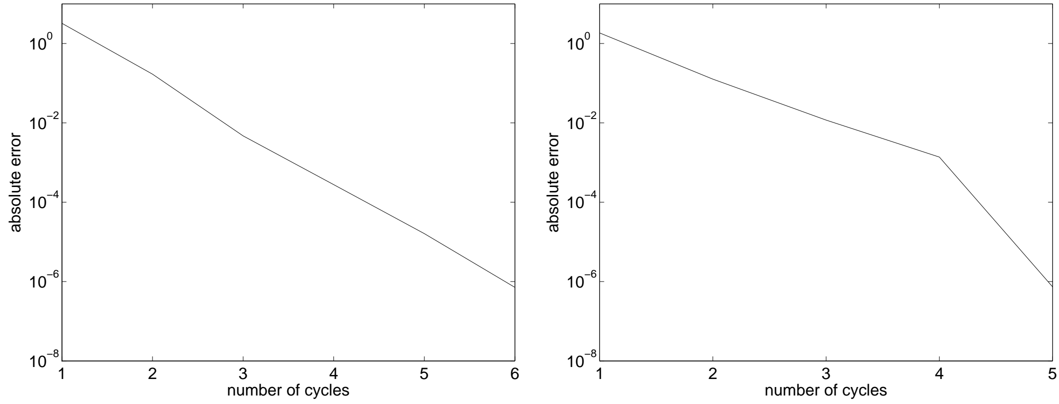



Figure 15: convergence of the fixed-point methods $\Phi_1$ (left) and $\Phi_2$ combined with vector extrapolation, $V_{\mathsf{Drain}} = 2V$ (open state)

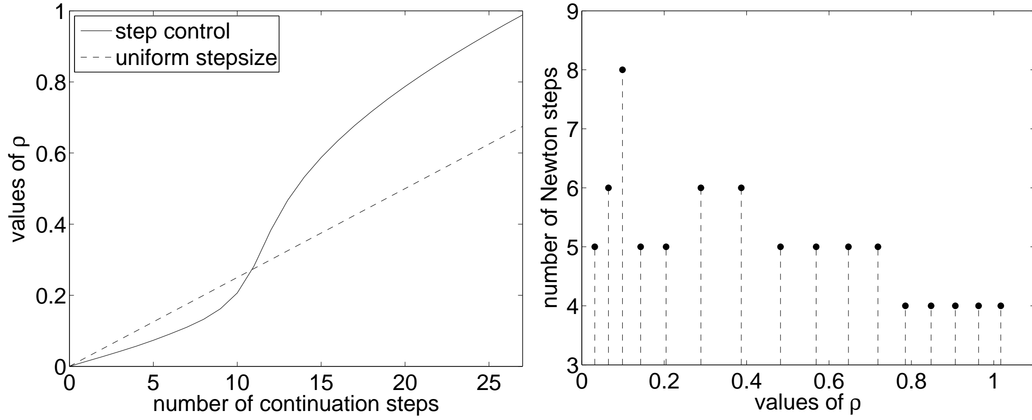



Figure 16: effect of the path following (left) and number of Newton cycles versus $\rho^{(k)}$ for single-gate MESFET (open state)

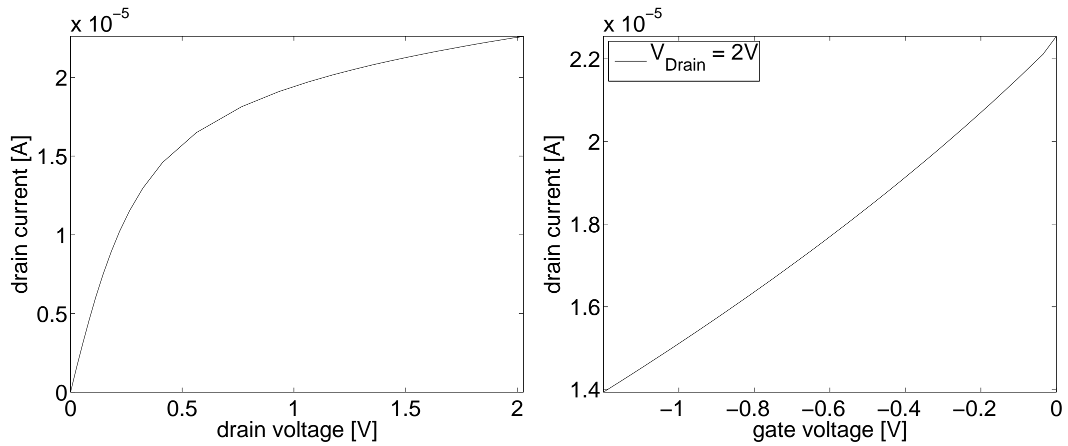



Figure 17: current-voltage characteristics of the two-dimensional single-gate MESFET in open state (left) and in closed state

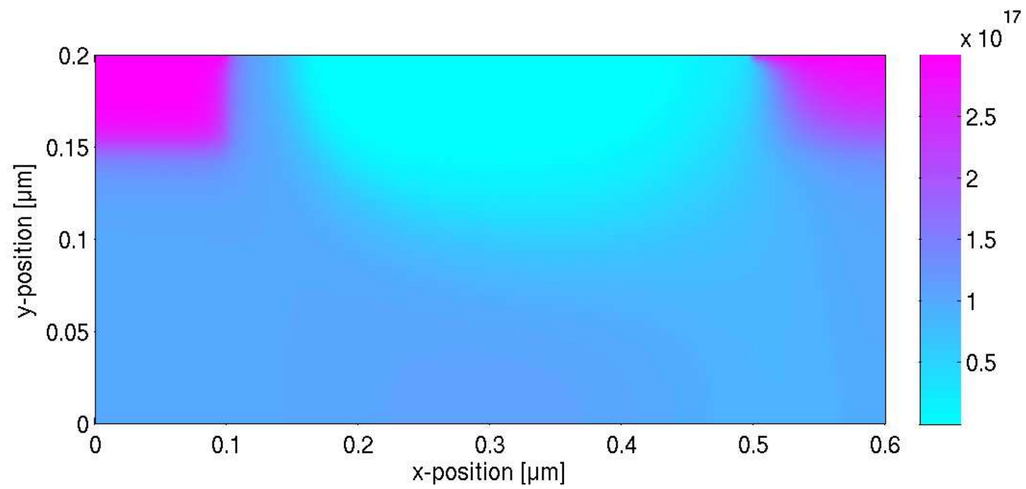



Figure 18: Electron density $[\mathrm{cm}^{-3}]$ in the single-gate MESFET (open state)

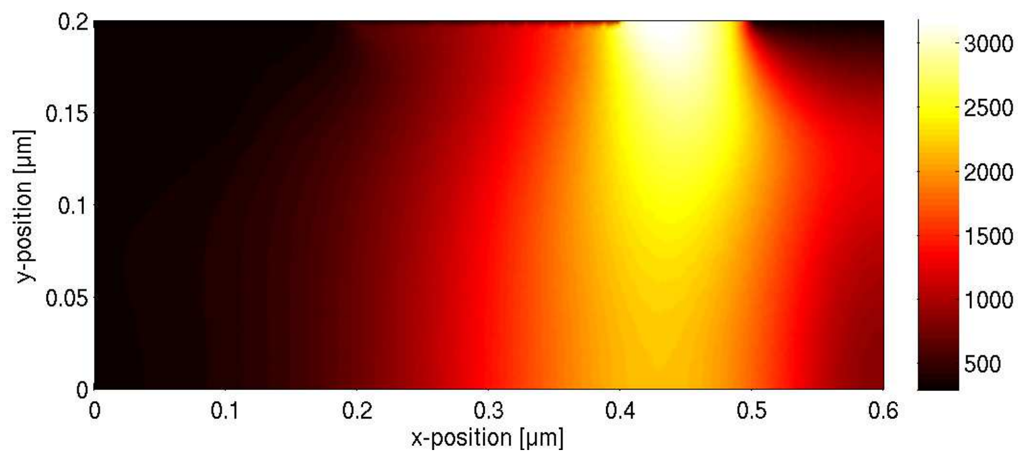



Figure 19: Electron temperature [K] in the single-gate MESFET (open state)

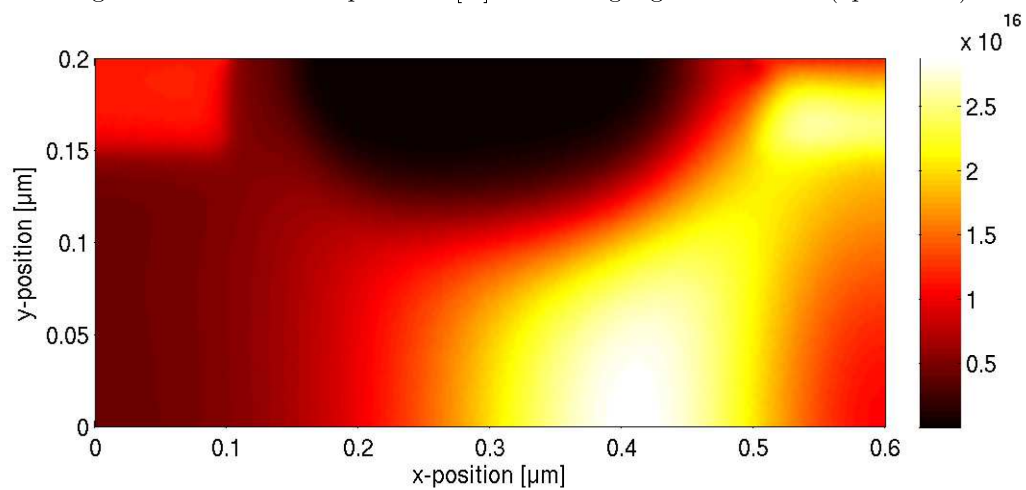



Figure 20: thermal energy $[\mathrm{eV} \cdot \mathrm{cm}^{-3}]$ in the single-gate MESFET (open state)

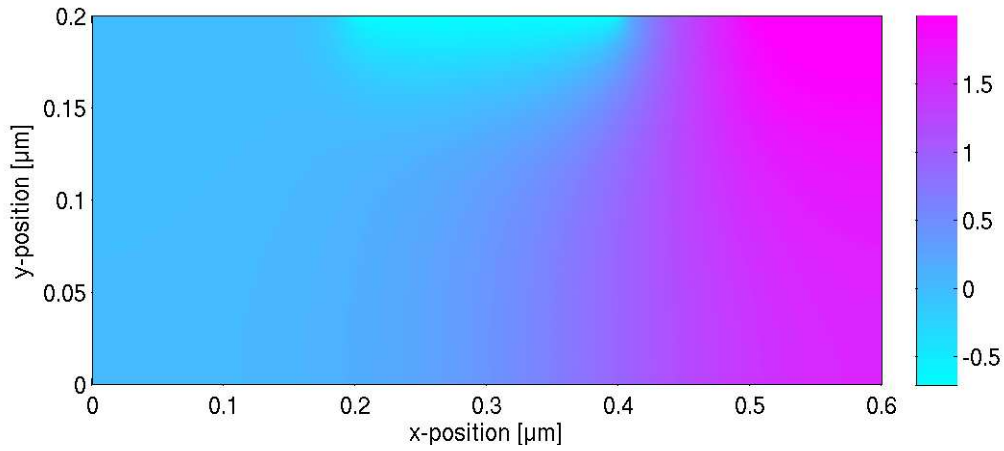



Figure 21: Electrostatic potential [V] in the single-gate MESFET (open state)

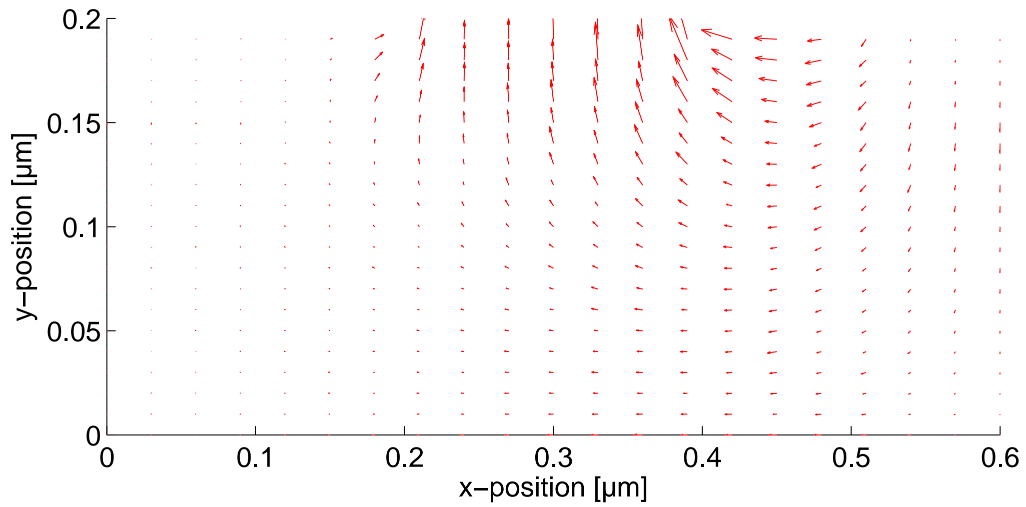



Figure 22: Electric field $-\nabla V$ [V/cm] in the single-gate MESFET (open state)

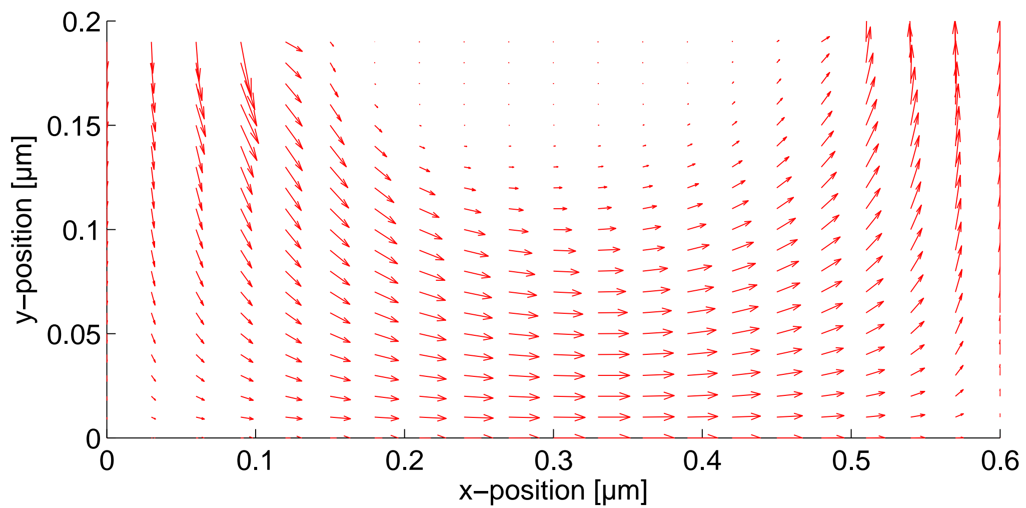



Figure 23: Electron current density in the single-gate MESFET (open state)

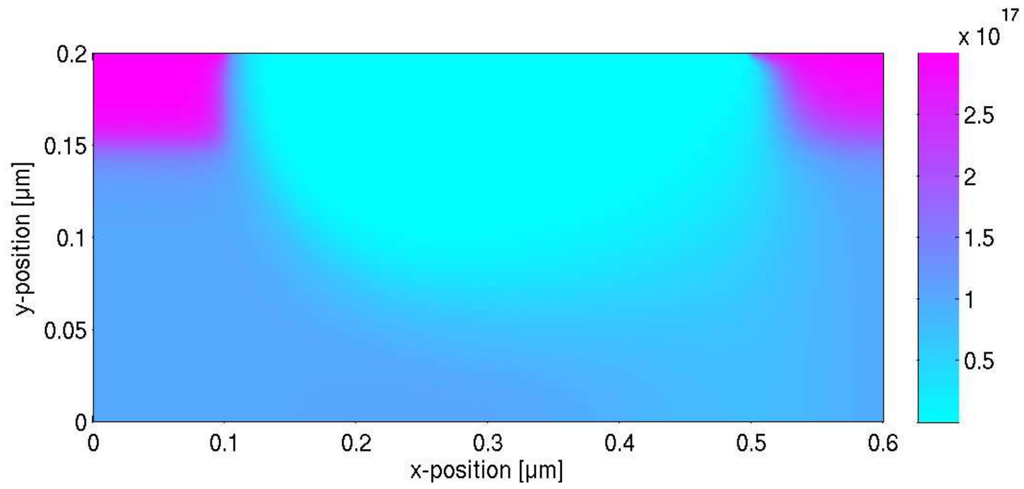



Figure 24: Electron density [$\mathrm{cm}^{-3}$] in the single-gate MESFET (closed state)

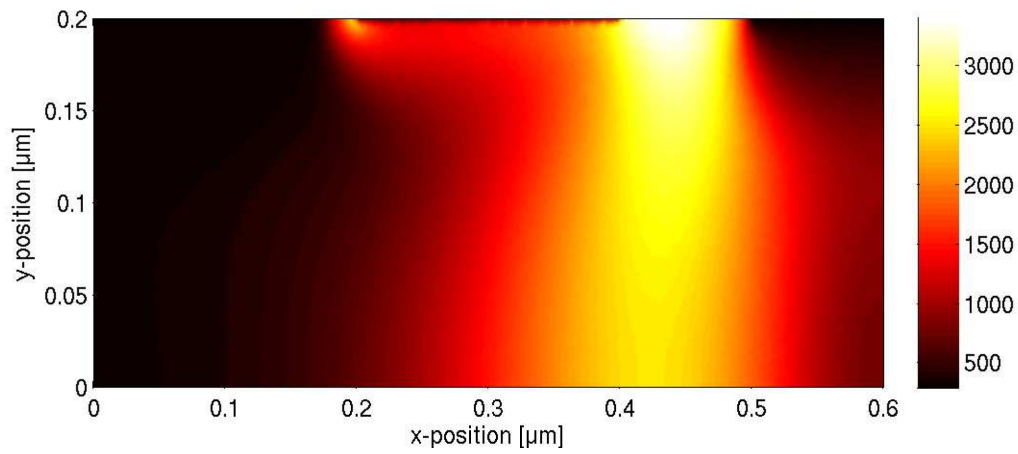



Figure 25: Electron temperature [K] in the single-gate MESFET (closed state)

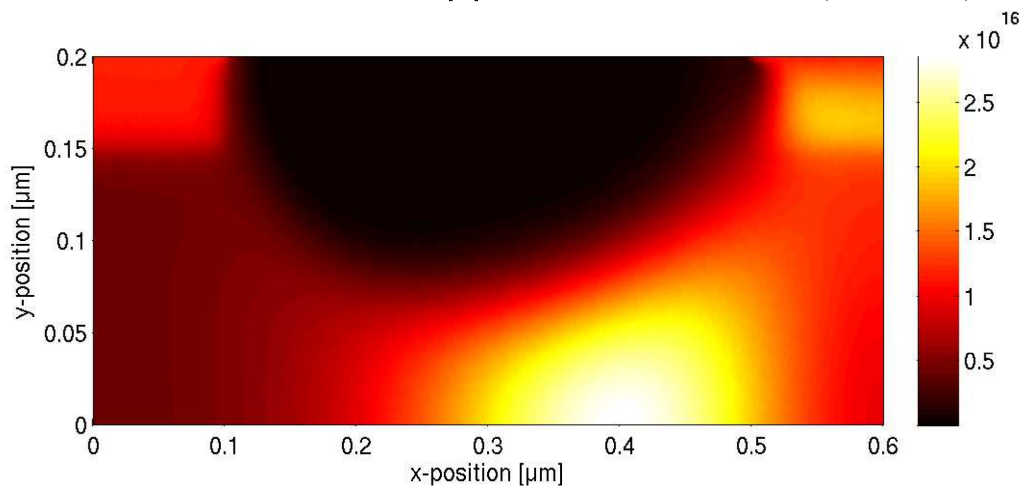



Figure 26: thermal energy [$\mathrm{eV} \cdot \mathrm{cm}^{-3}$] in the single-gate MESFET (closed state)

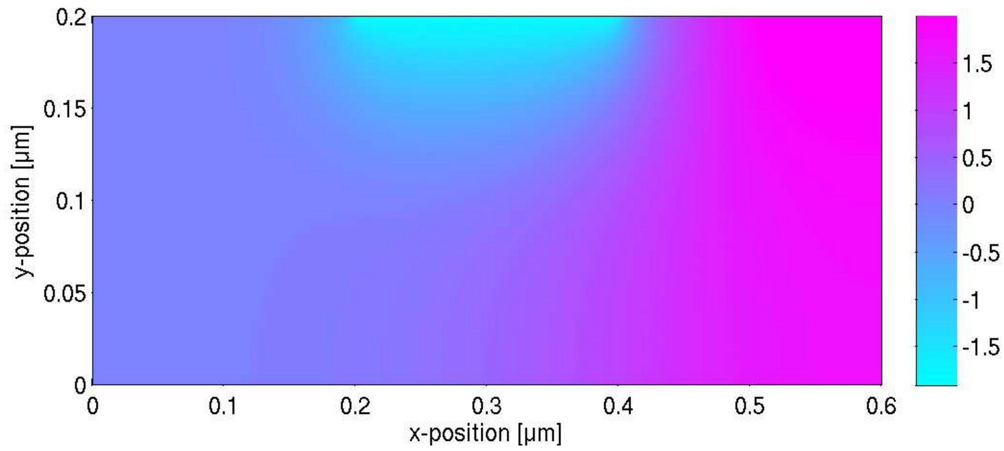



Figure 27: Electrostatic potential [V] in the single-gate MESFET (closed state)

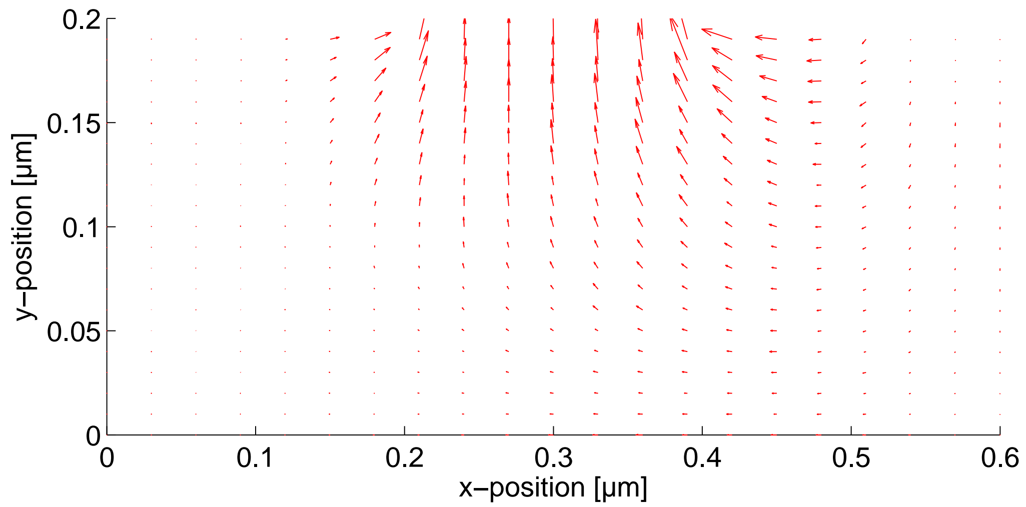



Figure 28: Electric field $-\nabla V$ [V/cm] in the single-gate MESFET (closed state)

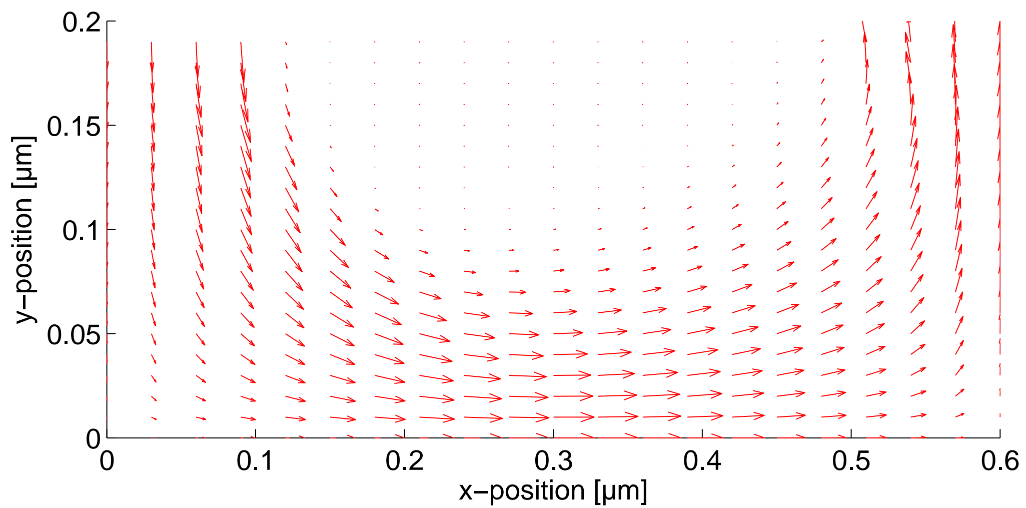



Figure 29: Electron current density in the single-gate MESFET (closed state)

### 6.4.2 Double-gate MESFET

A double-gate MESFET is a voltage-driven device which is used as a switch or amplifier [Sze81]. The model parameters of the silicon MESFET presented here are taken from [HJP03]. The device consists of two highly-doped $n^+$ regions near the Ohmic contacts, the source and the drain, and an $n$ region with an upper and a lower Schottky contact, the gate, in a sandwich configuration (see figure 30). The device behavior is mainly governed by the size of the depletion region (i.e., a region with very low electron density) that develops around the Schottky contacts. This depletion region enlarges if the gate voltage is decreased, and therefore diminishes the channel width which results in a reduced current for a fixed applied voltage (close state). For larger gate voltage, the depletion region becomes smaller and a significant current can flow (open state).
The semiconductor domain is $\Omega = (0, 0.6\,\mu\text{m}) \times (0, 0.24\,\mu\text{m})$. The source and drain contact lengths are $0.24\,\mu$m, the gate contact length is $0.2\,\mu$m. Moreover, the length of the low doped (chanel) region region is $0.36\,\mu$m.

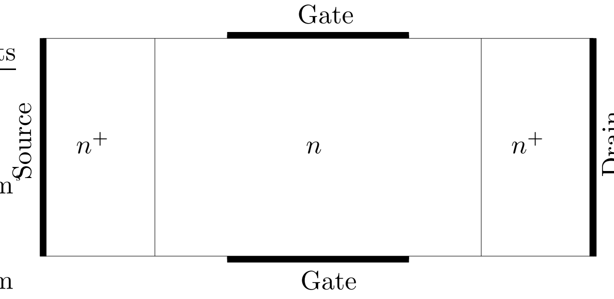



Figure 30: Geometry of the two-dimensional double-gate MESFET

The doping profile function is given by

$$\text{Dop}(x, y) = \begin{cases} 1 \cdot 10^{17}\,\text{cm}^{-3}, & x \in [0.12\,\mu\text{m}, 0.48\,\mu\text{m}],\ y \in [0, 0.24\,\mu\text{m}], \\ 3 \cdot 10^{17}\,\text{cm}^{-3}, & \text{else.} \end{cases}$$

At the source and drain contacts the data for the particle density is equal to the equilibrium values [MRS90]. The boundary values for the potential at the Ohmic contacts are the sum of the built-in potential $V_{\text{bi}}$ and the applied voltage. At the Schottky contacts, the Schottky barrier height is added, $V_{|\text{Gate}} = V_A + V_{\text{bi}} + V_B$. A barrier height of $-0.8\,$V is used as a typical value for a n-type silicon/metal contact. The temperature at the contacts is equal to the ambient temperature $T_0$. The particle density at the Schottky gates is computed from formula (5.1-19) in [Sel84]. In the following the boundary conditions are summarized,

- at the source: $n = 3 \cdot 10^{17}\,\text{cm}^{-3},\ T = 300\,\text{K},\ V = V_{\text{bi}}$;
- at the drain: $n = 3 \cdot 10^{17}\,\text{cm}^{-3},\ T = 300\,\text{K},\ V = V_{\text{bi}} + 2\,\text{V}$;
- at the gate:
  open state: $n = 3.9 \cdot 10^{5}\,\text{cm}^{-3},\ T = 300\,\text{K},\ V = V_{\text{bi}} - 0.8\,\text{V}$;
  close state: $n = 2.4 \cdot 10^{5}\,\text{cm}^{-3},\ T = 300\,\text{K},\ V = -1.2\,\text{V} + V_{\text{bi}} - 0.8\,\text{V}$;
- at the remaining boundary segments, homogeneous Neumann boundary conditions for $I_1, I_2$, and $\nabla V$ are used.

The scaled Debye length in this model is $\lambda = 1.1367 \cdot 10^{-2}$.
The reference solution used in the calculation of the relative errors listed in table 9 is given on a fine grid with $h = 2.5 \cdot 10^{-3} l^*$, where $l^* = 6.4622 \cdot 10^{-5}\,$cm. It is obvious that after a stage of rapid convergence stagnation in the error reduction occurs and even a pickup of the relative error

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 5.0 | 0.5 | 50 | 15 | $2.0 \cdot 10^{-2}$ | $1 \cdot 10^{-9}$ | $1 \cdot 10^{-5}$ | 0.7 | $1 \cdot 10^{-6}$ |
| $\Phi_2$ | 20 | 0.5 | 80 | 15 | $2.0 \cdot 10^{-2}$ | $1 \cdot 10^{-9}$ | $1 \cdot 10^{-5}$ | 0.7 | $1 \cdot 10^{-6}$ |

Table 7: Parameter values in the test of the 2D double-gate MESFET (open state), $k$ and $k_1$ are the numbers of base sequence vectors

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 5.0 | 0.5 | 50 | 15 | $8.0 \cdot 10^{-3}$ | $1 \cdot 10^{-9}$ | $1 \cdot 10^{-5}$ | 0.7 | $1 \cdot 10^{-6}$ |
| $\Phi_2$ | 20 | 0.5 | 80 | 15 | $8.0 \cdot 10^{-3}$ | $1 \cdot 10^{-9}$ | $1 \cdot 10^{-5}$ | 0.7 | $1 \cdot 10^{-6}$ |

Table 8: Parameter values in the test of the 2D double-gate MESFET (closed state)

| $h/l^*$ | RE for $V_h^I$ | RE for $w_{1h}^I$ | RE for $w_{2h}^I$ | RE for $T_h^I$ | RE for $n_h^I$ | RE for $g_h^I$ |
|---|---|---|---|---|---|---|
| $8 \cdot 10^{-2}$ | $7.1 \cdot 10^{-2}$ | $6.5 \cdot 10^{-2}$ | $1.2 \cdot 10^{-1}$ | $1.1 \cdot 10^{-1}$ | $1.0 \cdot 10^{-1}$ | $1.8 \cdot 10^{-1}$ |
| $4 \cdot 10^{-2}$ | $3.3 \cdot 10^{-2}$ | $3.9 \cdot 10^{-2}$ | $8.1 \cdot 10^{-2}$ | $5.8 \cdot 10^{-2}$ | $3.9 \cdot 10^{-2}$ | $7.2 \cdot 10^{-2}$ |
| $2 \cdot 10^{-2}$ | $3.5 \cdot 10^{-2}$ | $2.1 \cdot 10^{-2}$ | $5.1 \cdot 10^{-2}$ | $4.7 \cdot 10^{-2}$ | $2.7 \cdot 10^{-2}$ | $7.3 \cdot 10^{-2}$ |
| $1 \cdot 10^{-2}$ | $6.2 \cdot 10^{-3}$ | $1.1 \cdot 10^{-2}$ | $3.2 \cdot 10^{-2}$ | $2.4 \cdot 10^{-2}$ | $5.2 \cdot 10^{-3}$ | $1.2 \cdot 10^{-2}$ |

Table 9: Relative errors from the simulation of the two-dimensional double-gate MESFET model (open state)

is visible. This might be related to round-off errors in the calculation.
Again, we apply the full Newton method combined with the path-following as well as the Gummel-type iteration $\Phi_1$ and $\Phi_2$ combined with the RRE. The results are shown in the figures 31 and 32. The convergence is similar to the convergence in the test of the single-gate MESFET, i.e. the iteration $\Phi_1$ shows only a suboptimal linear convergence, while $\Phi_2$ tends to converge super-linear and even quadratic.
In the left diagram of figure 33 the effect of the path following in open state is shown (solid line). For small values of the applied voltage the step control reduces the step width, while the step size is increased for higher voltages. The dashed line indicates the value of the parameter where no step control is active, i.e. the voltage is increased uniformly and $c^{(k)} = 0$ in (5.14). The uniform step size in this example is $\tau^{(k)} = 2.5 \cdot 10^{-2}$, for the coefficient we take $\overline{c} = 0.125$. In the right diagram the effect of the path following in closed state is shown (solid line). For small values of the applied voltage the step control increases the step width, while the step size is reduced for higher voltages. The uniform step size in this example is $\tau^{(k)} = 1 \cdot 10^{-2}$, for the coefficient we take $\overline{c} = 0.05$.
In table 10 some CPU times are stated, which show that the path following has an improving effect in the calculation of the open state solution on coarse grids. On a fine grid the CPU times are almost equal.
In figure 34 two current-voltage curves of the double-gate MESFET are presented. The current-voltage curve in open state is similar to that of the single-gate MESFET , where in the double-gate MESFET the saturation is reached more rapidly. One can easily see that the saturation start around $V_{\mathsf{Drain}} = 0.3$V, and this point is detected by the path following method. Beyond this point, no significant increase in the terminal current comes up and hence the step length is increased. In the right picture one can see the dependence of the Drain current on the gate voltage with a drain voltage of 2V. This dependence is approximately quadratic which confirms the result mentioned in [Sze81]. In [HJP03] a similar result is presented.

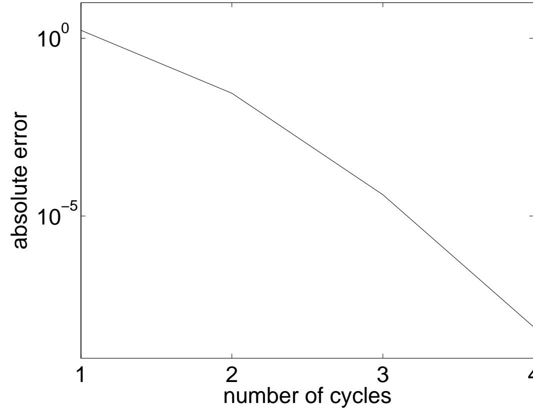



Figure 31: convergence of the full Newton method, $V_{\mathsf{Drain}} = 2V$, (open state)

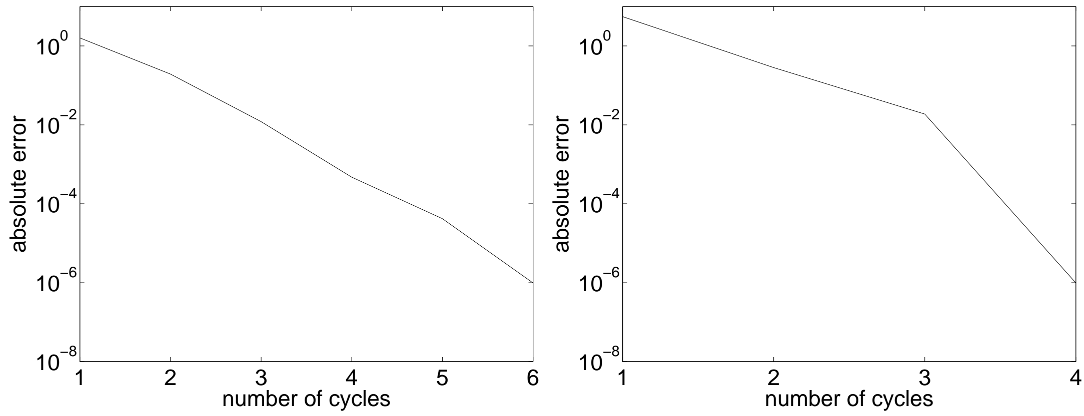



Figure 32: convergence of the fixed-point methods $\Phi_1$ (left) and $\Phi_2$ combined with vector extrapolation, $V_{\mathsf{Drain}} = 2V$ (open state)

It turns out that the double-gate MESFET has a more pronounced switching behavior than the single-gate MESFET. For the double-gate MESFET in closed state the maximum drain current is $6.3 \cdot 10^{-8}$ A, while for the single-gate MESFET the maximum drain current is $7.6 \cdot 10^{-5}$ A.
We now turn to the numerical results. In figure 35 the electron density in open state is shown, where the data are computed on a grid with $h = 5 \cdot 10^{-3} l^*$. One can see the two depletion region near the gate contacts, and between them a channel is formed. The temperature in figure 36 attains to its maximum at the right ends of the gate contact, which leads to steep temperature gradients. Between the two peaks the temperature reduces. The thermal energy shows a more complicated distribution in the device, see figure 37. Near the gate contacts the thermal energy energy values are small as expected due to the small carrier density. In the tune of the right gate contact ends a small area of very high energy is formed, centered with respect to the y-axis. A second peak is formed further to the right directly behind the jump of the doping profile. There, relatively high values of the electron temperature and the high doping meet, which may explain the massive increase in thermal energy.

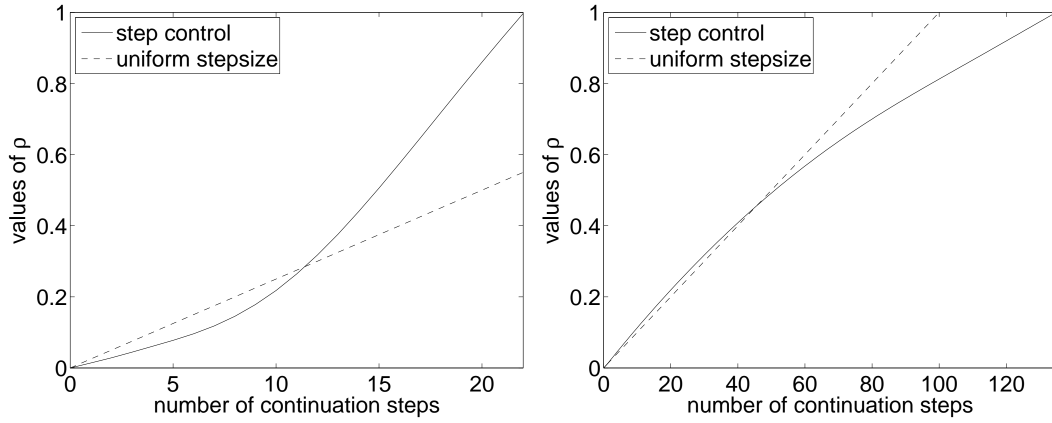



Figure 33: effect of the path following for the double-gate MESFET in open state (left) and in closed state

| $h/l^*$ | without path-follow. | with path-follow. | rel. time saved |
|---|---|---|---|
| $4 \cdot 10^{-2}$ | 28 | 20 | 28.6 |
| $2 \cdot 10^{-2}$ | 133 | 96 | 27.8 |
| $1 \cdot 10^{-2}$ | 704 | 533 | 24.3 |
| $5 \cdot 10^{-3}$ | 3670 | 3658 | 0.3 |

Table 10: CPU times [s] and relative time savings [%] for the double-gate MESFET (open state)

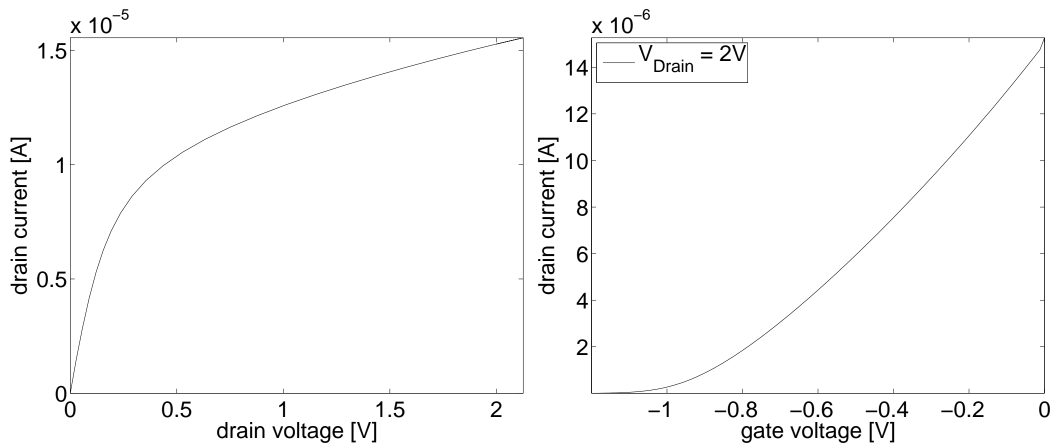



Figure 34: current-voltage characteristics of the two-dimensional double-gate MESFET in open state (left) and in closed state

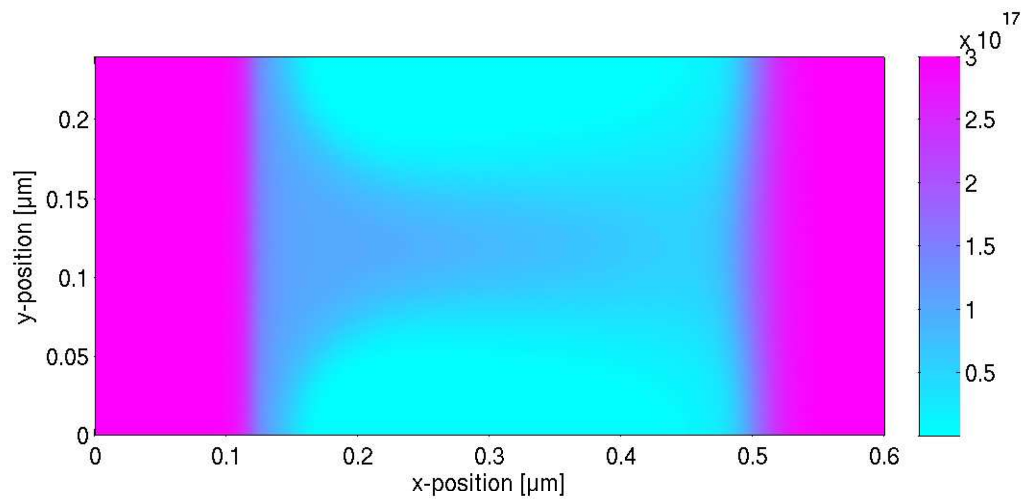



Figure 35: Electron density [$\mathrm{cm}^{-3}$] in the double-gate MESFET (open state)

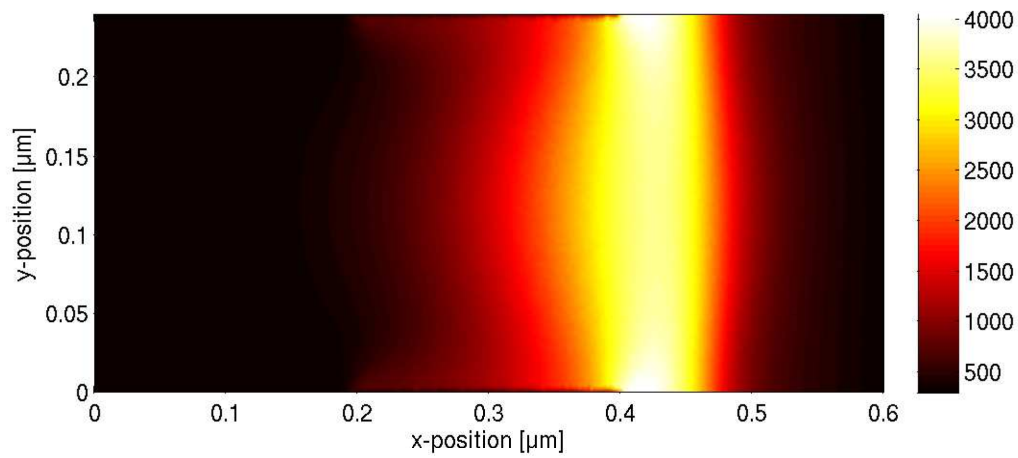



Figure 36: Electron temperature [K] in the double-gate MESFET (open state)

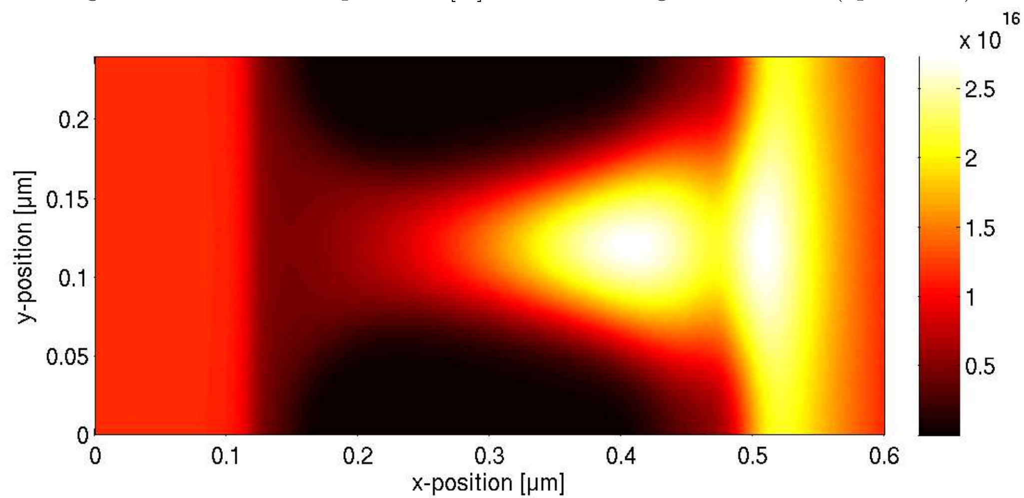



Figure 37: thermal energy [$\mathrm{eV} \cdot \mathrm{cm}^{-3}$] in the double-gate MESFET (open state)

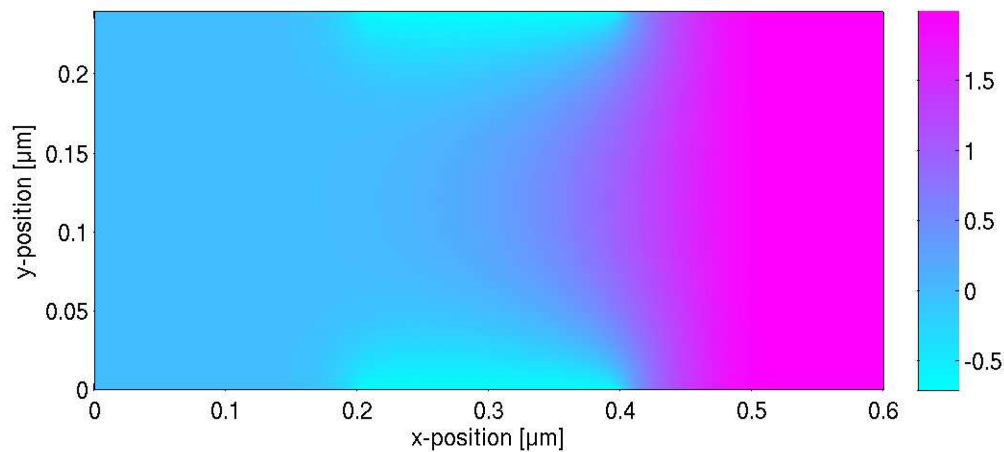



Figure 38: Electrostatic potential [V] in the double-gate MESFET (open state)

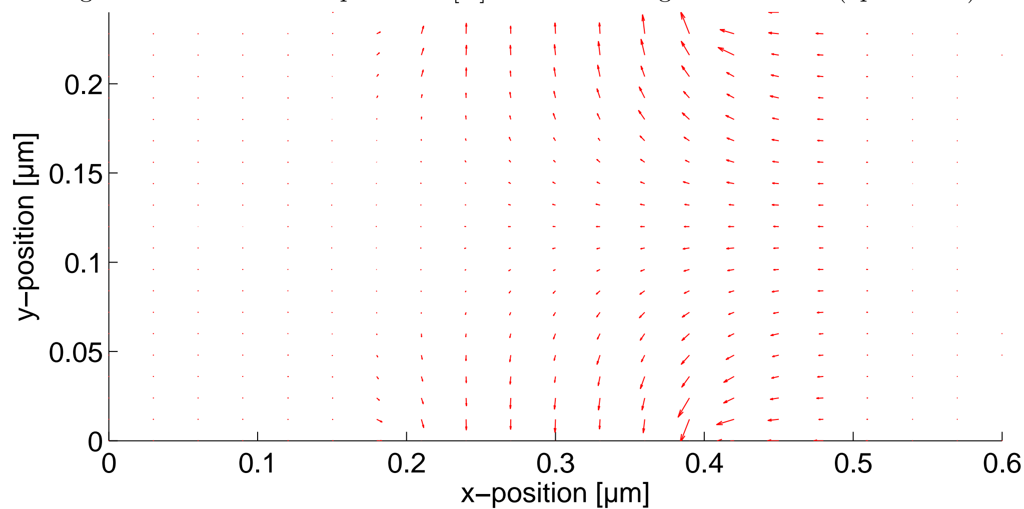



Figure 39: Electric field $-\nabla V$ [V/cm] in the double-gate MESFET (open state)

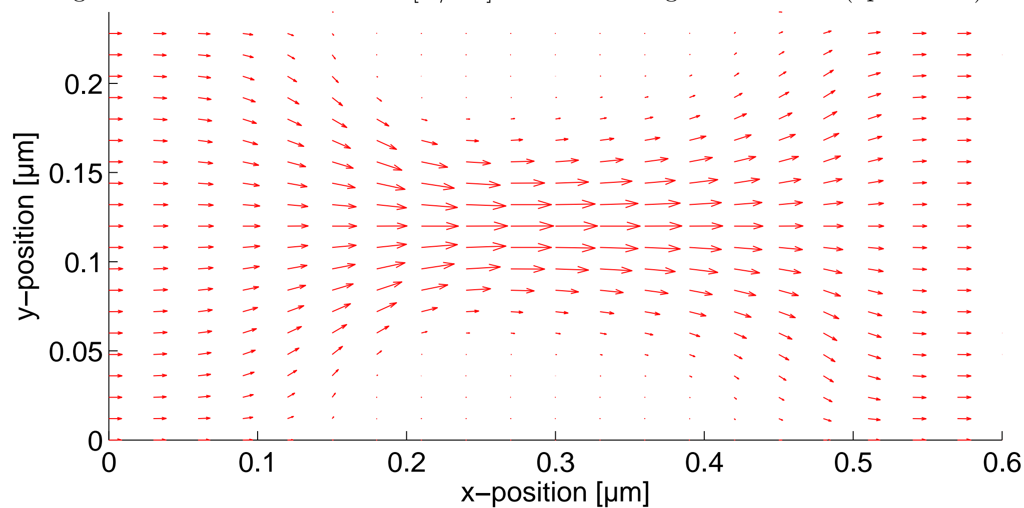



Figure 40: Electron current density in the double-gate MESFET (open state)

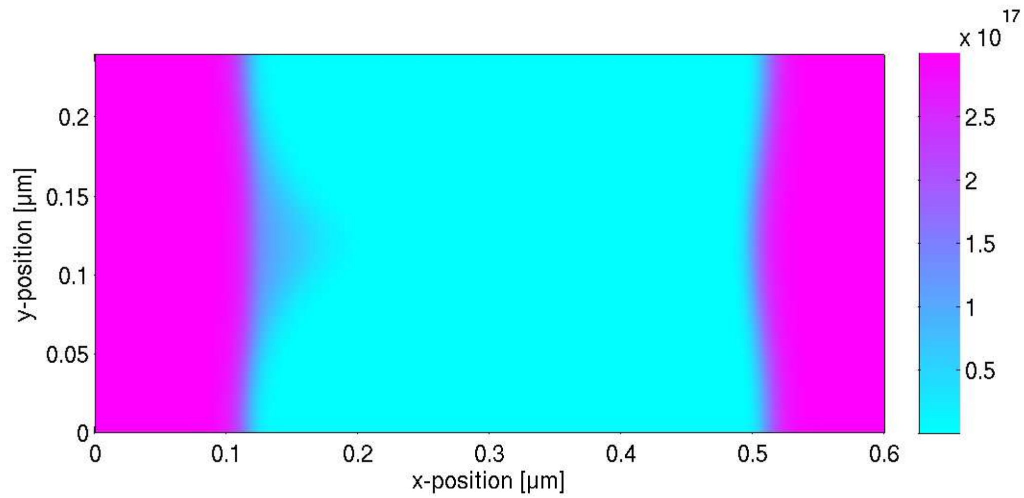



Figure 41: Electron density $[\mathrm{cm}^{-3}]$ in the double-gate MESFET (closed state)

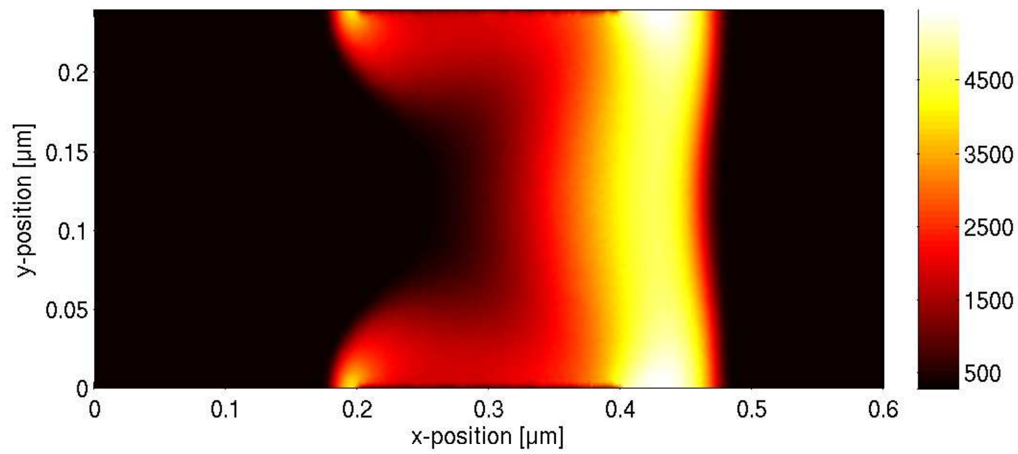



Figure 42: Electron temperature [K] in the double-gate MESFET (closed state)

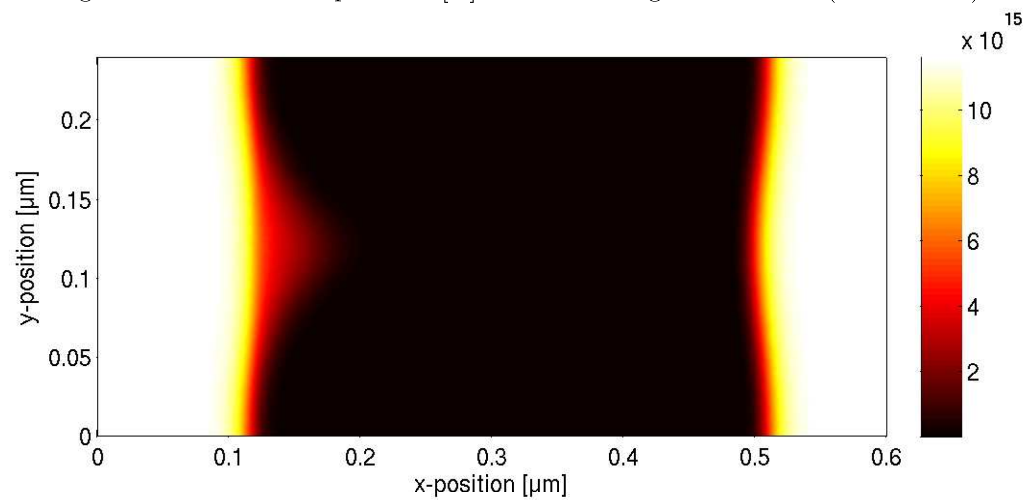



Figure 43: thermal energy $[\mathrm{eV} \cdot \mathrm{cm}^{-3}]$ in the double-gate MESFET (closed state)

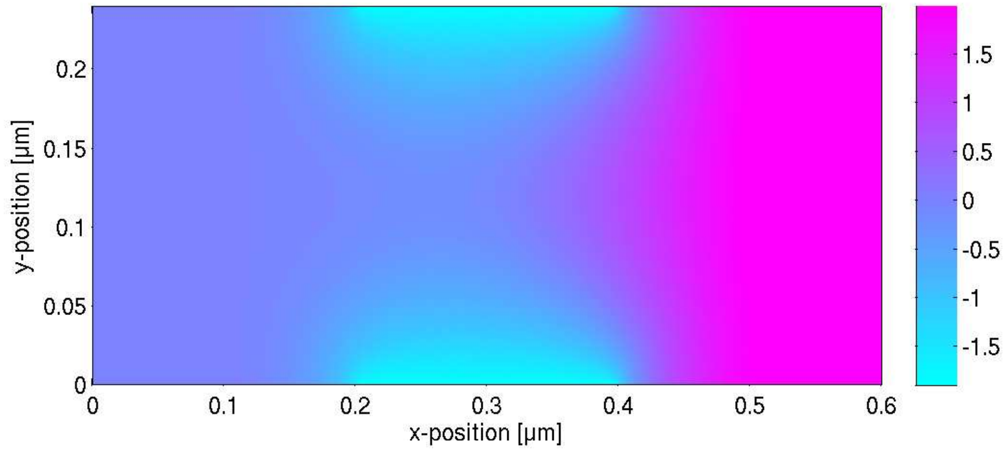



Figure 44: Electrostatic potential [V] in the double-gate MESFET (closed state)

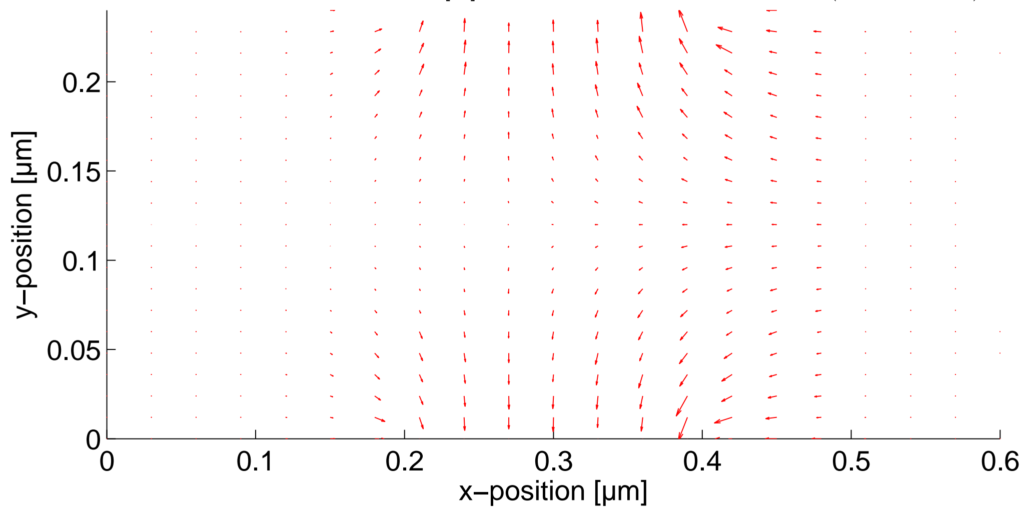



Figure 45: Electric field $-\nabla V$ [V/cm] in the double-gate MESFET (closed state)

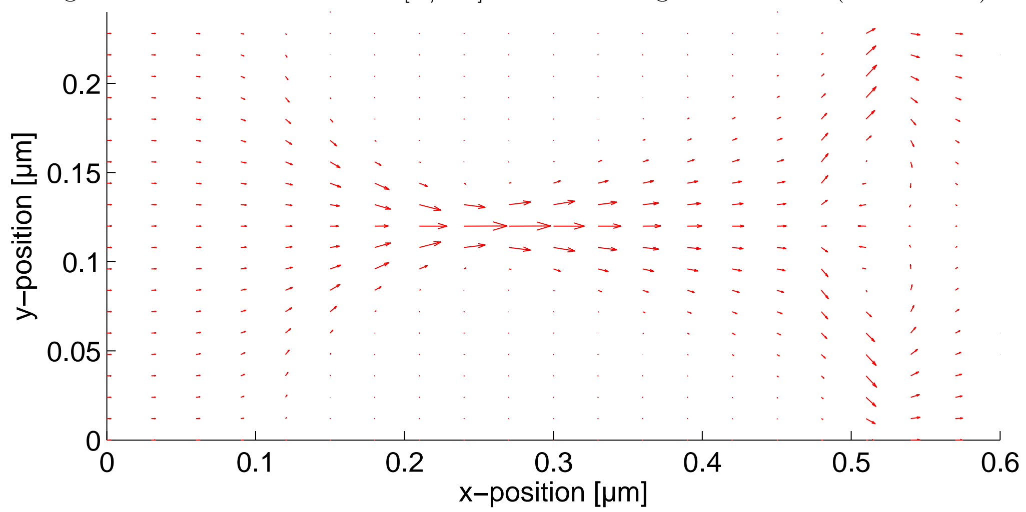



Figure 46: Electron current density in the double-gate MESFET (closed state)

### 6.4.3 MOSFET

A MOSFET (**m**etal-**o**xide **s**emiconductor **f**ield **e**ffect **t**ransistor) device can be used as a voltage-driven switch and is the most common device in computer technology. We simulate a transistor of size 420nmx210nm with an effective channel length of 70nm and an oxide thickness of 1.5nm. the length of the source S and drain D contacts is 30nm (see figure 47). In [HJP04] a similar model has been used, but with a different formulation of the energy-transport equations. The doping profile is given by

$$\mathrm{Dop}(x,y) = \begin{cases} 5 \cdot 10^{17}\,\mathrm{cm}^{-3}, & x \in [0, 175\,\mathrm{nm}] \cup [245\,\mathrm{nm}, 420\,\mathrm{nm}],\ y \in [180\,\mathrm{nm}, 210\,\mathrm{nm}], \\ 1 \cdot 10^{15}\,\mathrm{cm}^{-3}, & \text{else}. \end{cases}$$

The geometry and data of this device are adapted from the work of Cassan et al. [CGDH00].

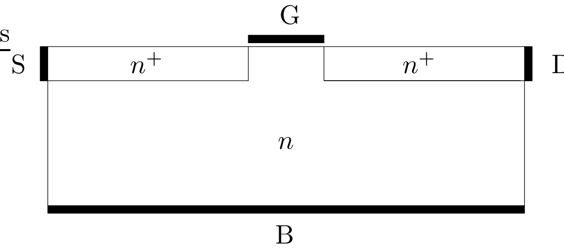



Figure 47: Geometry of the two-dimensional MOSFET

The current-voltage characteristics of the device are mainly influenced by the electric field at the semiconductor oxide junction. To model the influence of the oxide we assume that the particles do not penetrate the oxide region. We denote the semiconductor region by $\Omega_S$, the oxide region is denoted by $\Omega_O$,

$$\Omega_S = [0, 420\,\mathrm{nm}] \times [0, 210\,\mathrm{nm}], \quad \Omega_O = [175\,\mathrm{nm}, 254\,\mathrm{nm}] \times [210\,\mathrm{nm}, 211.5\,\mathrm{nm}].$$

$\Gamma_{S/O} = \partial\Omega_S \cap \partial\Omega_O$ is the silicon/silicon oxide interface, $\Gamma_{\mathsf{Gate}}$ is the gate contact part of $\partial\Omega_o$, and $\Gamma_{N,O} = \partial\Omega_O \setminus (\Gamma_G \cup \Gamma_{S/O})$ are the remaining boundary parts of $\Omega_O$. We choose Dirichlet boundary conditions at the source, drain and bulk B, i.e., the densities are set to their equilibrium values. Since we assume that the particles do not penetrate the oxide, the particle and energy densities need not be computed in $\Omega_O$. We impose homogeneous Neumann boundary conditions $I_1 \cdot \nu = 0$ and $I_2 \cdot \nu = 0$ on $\Gamma_{S/O}$ and on the remaining parts of the boundary of $\Omega_S$. The Poisson equation is solved in the domain $\Omega_S \cup \Omega_O$ with a space-dependent permittivity which is constant in both $\Omega_S$ and $\Omega_O$. We specify homogeneous Neumann boundary conditions on $\Gamma_{N,O}$ and a Dirichlet condition on $\Gamma_{\mathsf{Gate}}$ for the electrostatic potential $V$.
The scaled Debye length for this model is $\lambda = 2.7095 \cdot 10^{-3}$.
In table 11 the relative errors are specified, where the reference solution has been computed on a grid with parameter $h = 5 \cdot 10^{-3} l^*$.
Again, we use the Gummel-type iterations to solve the nonlinear systems. The convergence of $\Phi_1$ and $\Phi_2$ combined with the RRE is shown in the figure 48. And again we see that the iteration $\Phi 1$ converges only linearly, and the convergence of $\Phi_2$ is quadratic. In the figure 49, one can see that the number of cycles for $\Phi_1$ varies more than for $\Phi$, as shown in figure 49. In a way, the iteration $\Phi_2$ behaves more like the Newton method than $\Phi_1$ does.
In figure 52 a finite-element mesh is depicted that was used in the numerical tests. One can see the strong mesh grading between the bulk region and the region in the vicinity of the gate contact ends. The refinement of the triangles is necessary to guarantee that the mesh is quasi-uniform. But this inflates the number of triangles and therefore the number of degrees of freedom.
In [HJP04] an adaptive mixed scheme is proposed that can improve the accuracy of the numerical solution. It refines the triangles near the discontinuities of the doping profile function.

| $h/l^*$ | RE for $V_h^I$ | RE for $w_{1h}^I$ | RE for $w_{2h}^I$ | RE for $T_h^I$ | RE for $n_h^I$ | RE for $g_h^I$ |
|---|---|---|---|---|---|---|
| $8 \cdot 10^{-2}$ | $2.1 \cdot 10^{-2}$ | $6.5 \cdot 10^{-2}$ | $5.2 \cdot 10^{-2}$ | $6.8 \cdot 10^{-1}$ | $1.5 \cdot 10^{-1}$ | $7.9 \cdot 10^{-2}$ |
| $4 \cdot 10^{-2}$ | $1.4 \cdot 10^{-2}$ | $2.8 \cdot 10^{-2}$ | $4.0 \cdot 10^{-2}$ | $4.1 \cdot 10^{-2}$ | $7.2 \cdot 10^{-2}$ | $4.9 \cdot 10^{-2}$ |
| $2 \cdot 10^{-2}$ | $1.1 \cdot 10^{-2}$ | $1.4 \cdot 10^{-2}$ | $2.4 \cdot 10^{-2}$ | $2.3 \cdot 10^{-2}$ | $3.1 \cdot 10^{-2}$ | $2.3 \cdot 10^{-2}$ |
| aver. ord. | 0.46 | 1.11 | 0.55 | 0.80 | 1.13 | 0.88 |

Table 11: Relative errors and average convergence order from the simulation of the two-dimensional MOSFET model (open state)

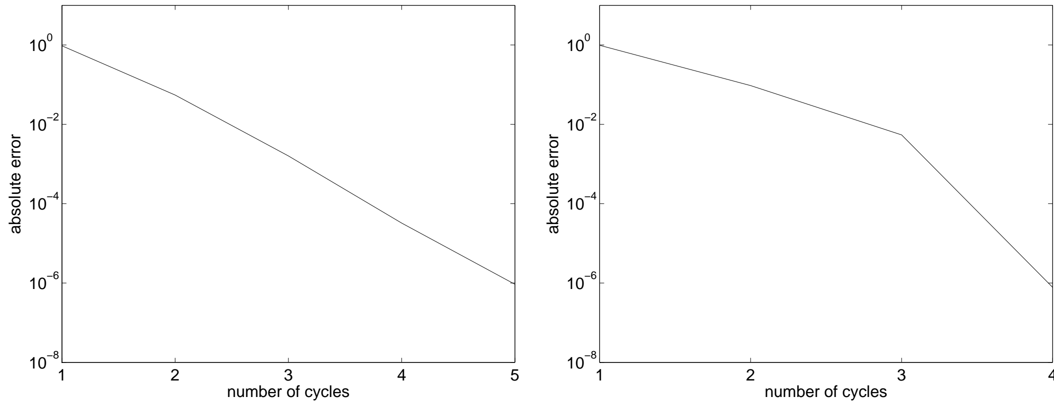



Figure 48: convergence of the fixed-point methods $\Phi_1$ (left) and $\Phi_2$ combined with vector extrapolation, $V_{\mathsf{Drain}} = 2V$ (open state)

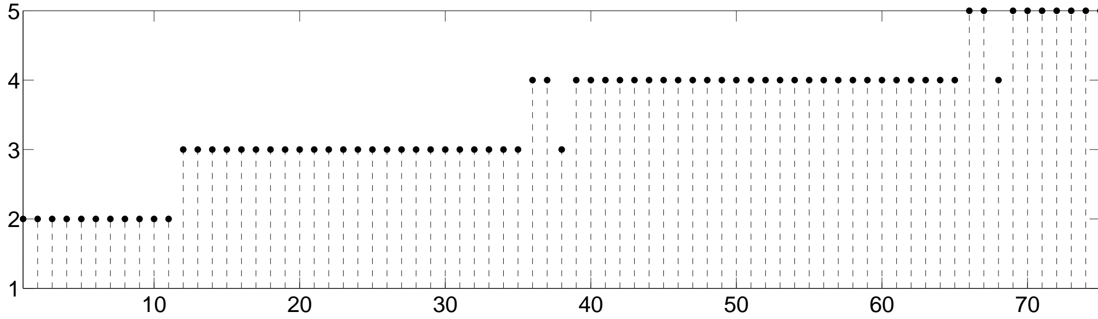


Figure 49: Number of cycles of $\Phi_1$ in every step of the continuation method

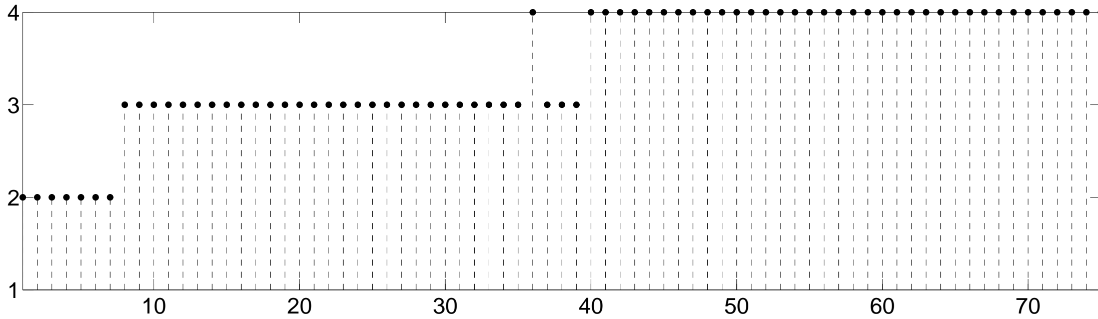


Figure 50: Number of cycles of $\Phi_2$ in every step of the continuation method

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 5.0 | 0.5 | 50 | 15 | $1.33 \cdot 10^{-2}$ | $1 \cdot 10^{-9}$ | $1 \cdot 10^{-5}$ | 0.7 | $1 \cdot 10^{-6}$ |
| $\Phi_2$ | 15 | 2.0 | 100 | 15 | $1.33 \cdot 10^{-2}$ | $1 \cdot 10^{-9}$ | $1 \cdot 10^{-5}$ | 0.7 | $1 \cdot 10^{-6}$ |

Table 12: Parameter values in the test of the MOSFET, $k$ and $k_1$ are the numbers of base sequence vectors

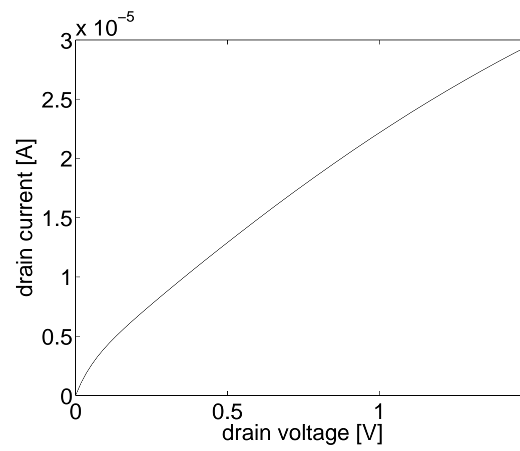



Figure 51: current-voltage characteristics of the two-dimensional MOSFET (open state)

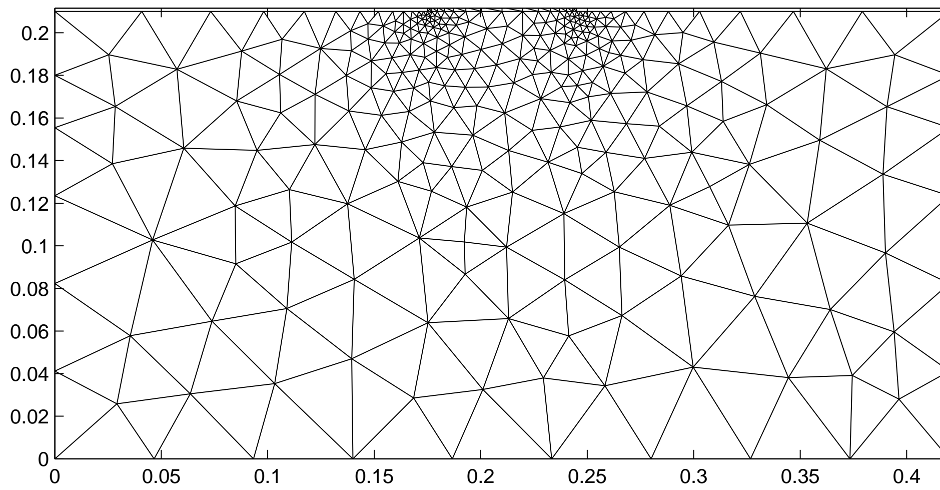



Figure 52: triangulation with maximum element size 0.1 in the MOSFET

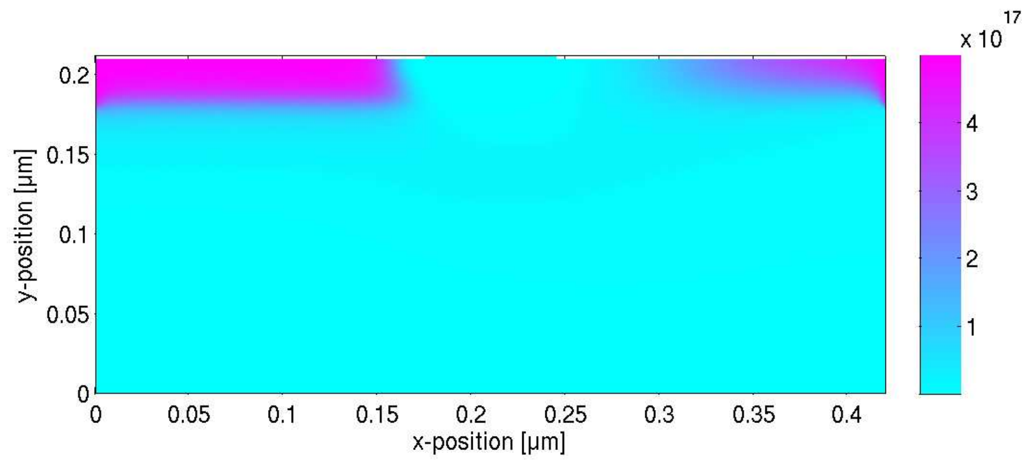



Figure 53: Electron density $[\mathrm{cm}^{-3}]$ in the MOSFET (open state)

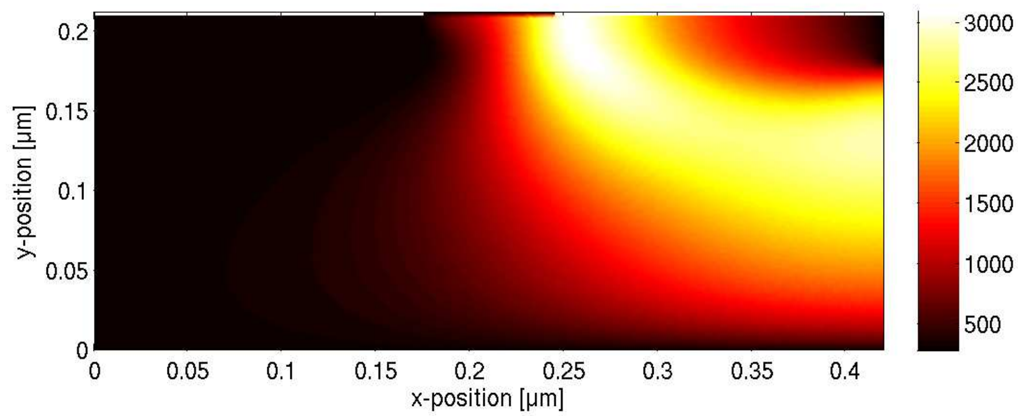



Figure 54: Electron temperature [K] in the MOSFET (open state)

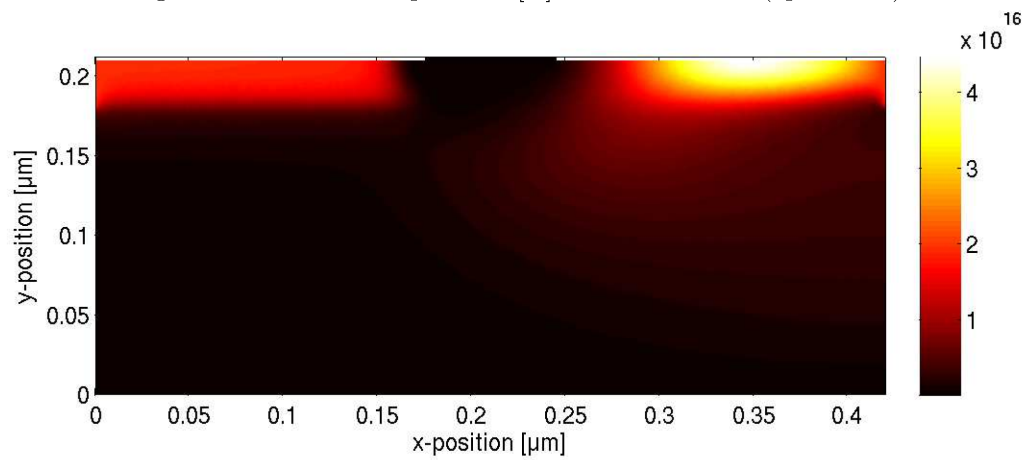



Figure 55: Thermal energy $[\mathrm{eV} \cdot \mathrm{cm}^{-3}]$ in the MOSFET (open state)

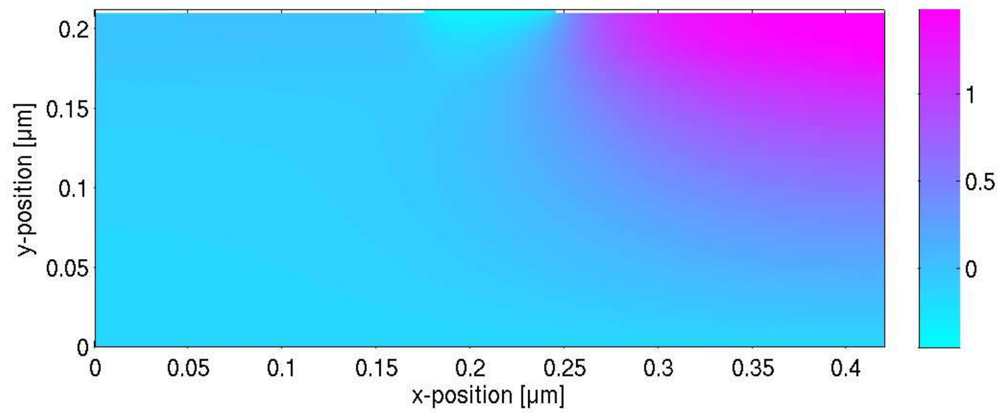



Figure 56: Electrostatic potential [V] in the MOSFET (open state)

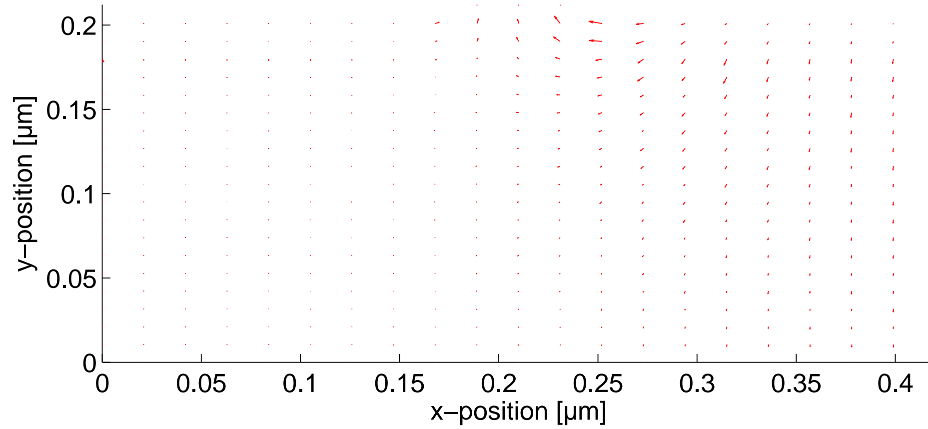



Figure 57: Electric field $-\nabla V$ [V/cm] in the MOSFET (open state)

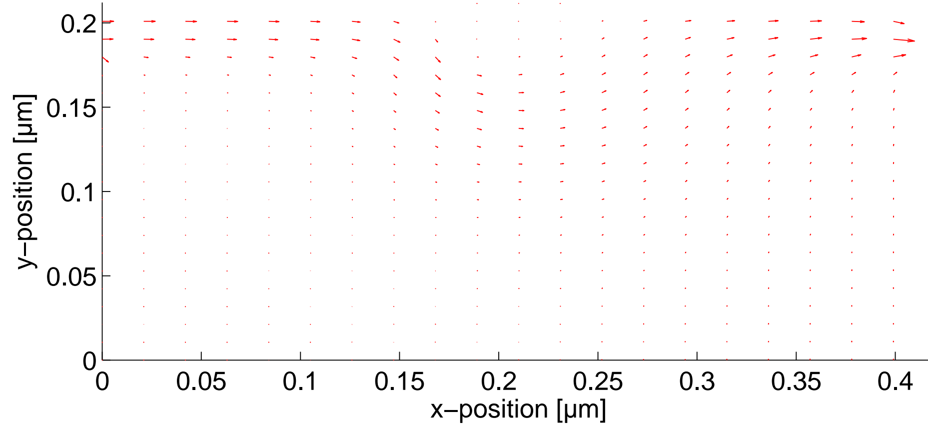



Figure 58: Electron current density in the MOSFET (open state)

## 6.5 Simulation of quasi 1D devices and quasi 2D devices

The devices presented in this section are uniform extensions of the 1D devices and 2D devices of the preceeding sections.
All device simulations have been performed with the full Newton method combined with the path following, and the results are shown in the pictures. The Gummel-type iterations have also been applied to some of the device models presented here. If the Gummel-type iterations are applied, then the corresponding parameter values are stated.

### 6.5.1 Ballistic diode

The device is three-dimensional and uniform in two space dimensions. Along the direction of x it equals the one-dimensional ballistic diode model. The semiconductor domain is $\Omega = (0, l_x) \times (0, l_y) \times (0, l_z)$, where $l_x = 0.6\mu$m, $l_y = 0.2\mu$m and $l_z = 0.1\mu$m. The channel length is equal to $0.4\mu$m. The highly-doped regions are $\Omega_L = (0, 0.1\mu\text{m}) \times (0, l_y) \times (0, l_z)$ and $\Omega_R = (0.5\mu\text{m}, l_x) \times (0, l_y) \times (0, l_z)$. The intermediate part $\Omega \setminus (\Omega_L \cup \Omega_R)$ is the lightly-doped $n$ region. The device has two Ohmic contacts, the source contact and the drain contact

$$\Gamma_{D1} = \{0\} \times (0, l_y) \times (0, l_z), \quad \Gamma_{D2} = \{l_x\} \times (0, l_y) \times (0, l_z).$$

The remaining boundary parts are insulating, i.e. we impose homogeneous Neumann data. On the Dirichlet boundary parts $\Gamma_{D1}$ and $\Gamma_{D2}$ the following data are given

$$\begin{array}{llll} n = C_m, & T = T_0, & V = V_{\mathsf{bi}} & \text{on} \quad \Gamma_{D1}, \\ n = C_m, & T = T_0, & V = U_0 + V_{\mathsf{bi}} & \text{on} \quad \Gamma_{D2}, \end{array}$$

where $U_0 = 1.5$V is the applied voltage. The *built-in potential* $V_{\mathsf{bi}}$ is the solution of the Poisson equation,

$$\lambda^2 \Delta V_{\mathsf{bi}} = e^{V_{\mathsf{bi}}} - \text{Dop} \quad \text{in } \Omega,$$

subject to the boundary conditions

$$V_{\mathsf{bi}} = \log(\text{n}) \quad \text{on } \Gamma_{D1} \cup \Gamma_{D2}.$$

The doping profile function is

$$\text{Dop}(x, y, z) = \begin{cases} 5 \cdot 10^{17}\text{cm}^{-3}, & (x, y, z) \in \Omega_L \cup \Omega_R, \\ 2 \cdot 10^{15}\text{cm}^{-3}, & \text{else.} \end{cases}$$

In figure 60 the electron temperature is depicted, where the data are computed on a grid with maximum diameter $h_{max} = 0.03\, l^*$ with $l^*$ denoting the device diameter. The number of partition elements is ca. 12100, the number of degrees of freedom is ca. 25300. The minimum temperature is 280 K at $x = 0.49\mu$m, the maximum temperature is 2440 K at $x = 0.16\mu$m. The results almost coincide with those of the 1D ballistic diode simulation.

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 5.0 | 1.0 | 150 | 15 | $2.5 \cdot 10^{-2}$ | $1 \cdot 10^{-6}$ | $1 \cdot 10^{-2}$ | 0.7 | $1 \cdot 10^{-3}$ |
| $\Phi_2$ | 5.0 | 3.0 | 150 | 15 | $2.0 \cdot 10^{-2}$ | $1 \cdot 10^{-6}$ | $1 \cdot 10^{-2}$ | 0.7 | $1 \cdot 10^{-3}$ |

Table 13: Parameter values in the test of the 3D ballistic diode

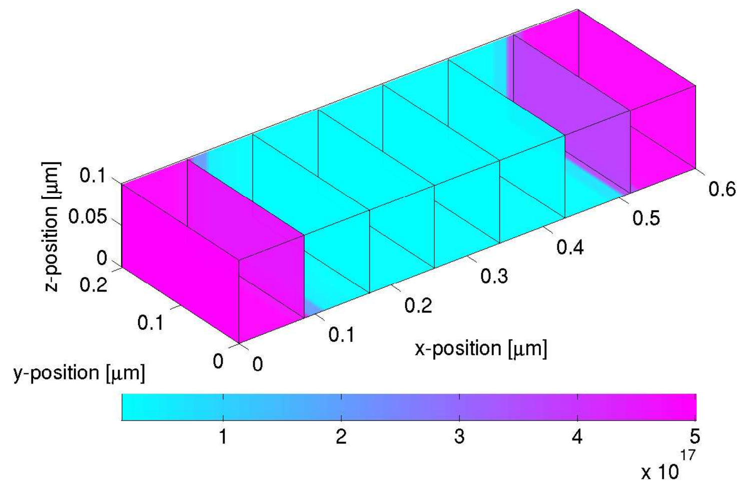



Figure 59: Electron density [$\mathrm{cm}^{-3}$] in the ballistic diode

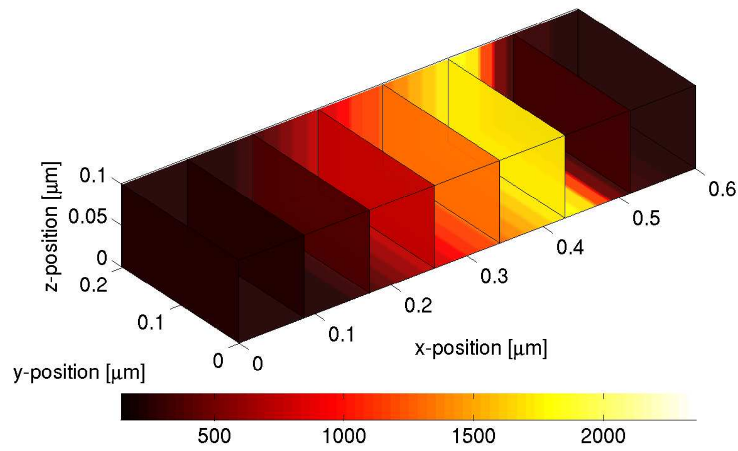



Figure 60: Electron temperature [K] in the ballistic diode

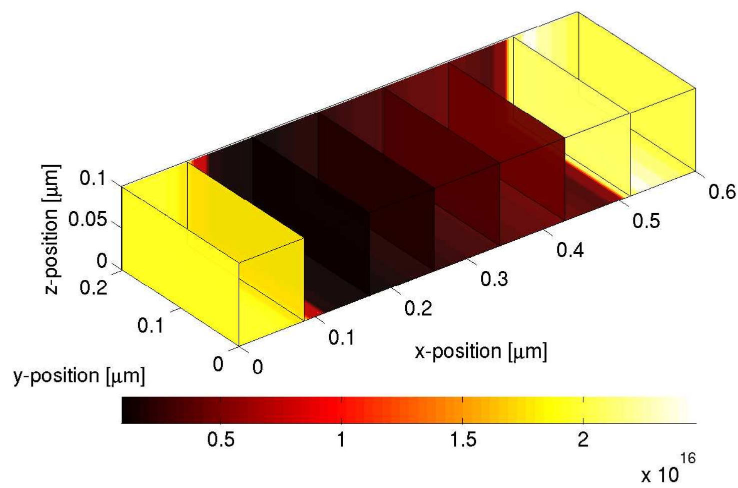



Figure 61: thermal energy [$\mathrm{eV} \cdot \mathrm{cm}^{-3}$] in the ballistic diode

### 6.5.2 Single-gate MESFET

This device is a uniform extension of the 2D single-gate MESFET of section 6.4.1. The highly doped regions are $\Omega_L = [0, 0.1] \times [0.15, 0.2] \times [0, 0.1]$ and $\Omega_R = [0.5, 0.6] \times [0.15, 0.2] \times [0, 0.1]$.

PSfrag replacements

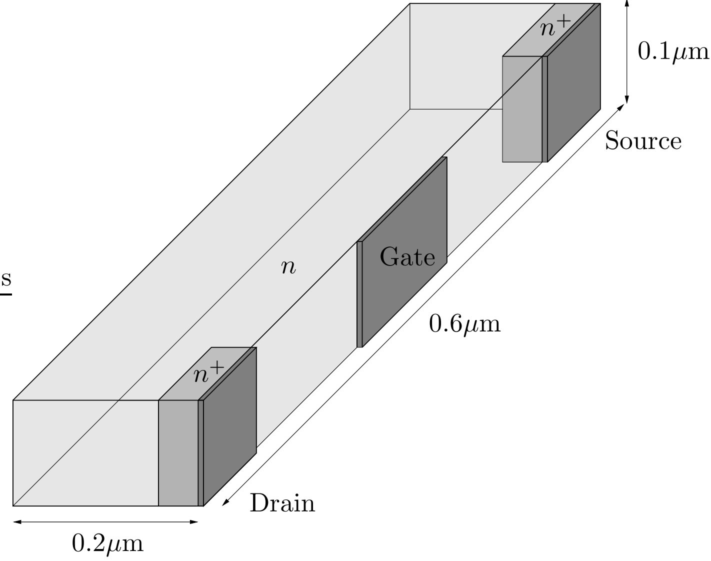



Figure 62: Schematic layout of the simulated single-gate MESFET

The doping profile function is given by (in $\mu$m)

$$\mathrm{Dop}(x, y, z) = \begin{cases} 3 \cdot 10^{17}\,\mathrm{cm}^{-3} & (x, y, z) \in \Omega_L \cup \Omega_R, \\ 1 \cdot 10^{17}\,\mathrm{cm}^{-3} & \text{else.} \end{cases}$$

With an applied voltage $V_A = 2\,\mathrm{V}$ at the drain and $V_A = 0$ (open state) at the gate as well as a barrier voltage $V_B = -0.8\,\mathrm{V}$ at the gate, the boundary data is given as follows,

- at the source: $n = 3 \cdot 10^{17}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}}$;
- at the drain: $n = 3 \cdot 10^{17}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}} + 2\,\mathrm{V}$;
- at the gate:$n = 3.9 \cdot 10^{5}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}} - 0.8\,\mathrm{V}$;
- at the remaining boundary segments, homogeneous Neumann boundary conditions for $I_1, I_2$, and $\nabla V$ are used.

In table 14 the Gummel-type iteration parameters used in the simulations are given.

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 2.0 | 2.0 | 150 | 15 | $2.0 \cdot 10^{-2}$ | $1 \cdot 10^{-6}$ | $1 \cdot 10^{-2}$ | 0.7 | $1 \cdot 10^{-3}$ |
| $\Phi_2$ | 15 | 3.0 | 150 | 15 | $1.33 \cdot 10^{-2}$ | $1 \cdot 10^{-6}$ | $1 \cdot 10^{-2}$ | 0.7 | $1 \cdot 10^{-3}$ |

Table 14: Parameter values in the test of the 3D single-gate MESFET

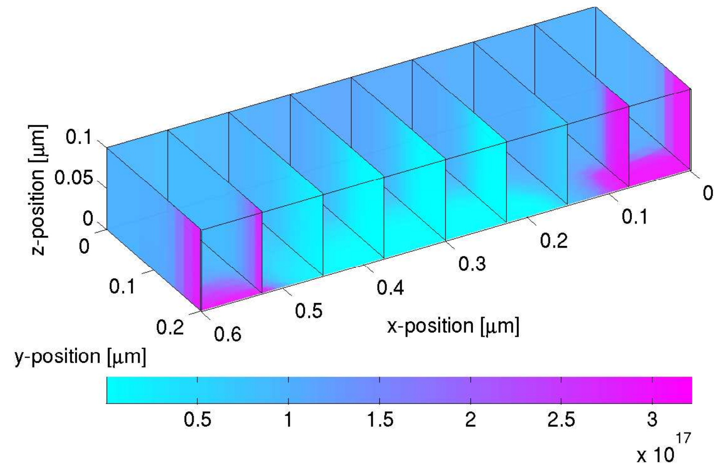



Figure 63: Electron density $[\mathrm{cm}^{-3}]$ in the single-gate MESFET (open state)

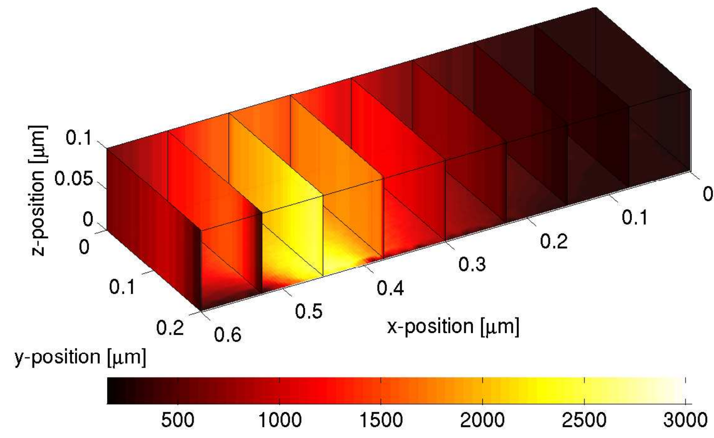



Figure 64: Electron temperature [K] in the single-gate MESFET (open state)

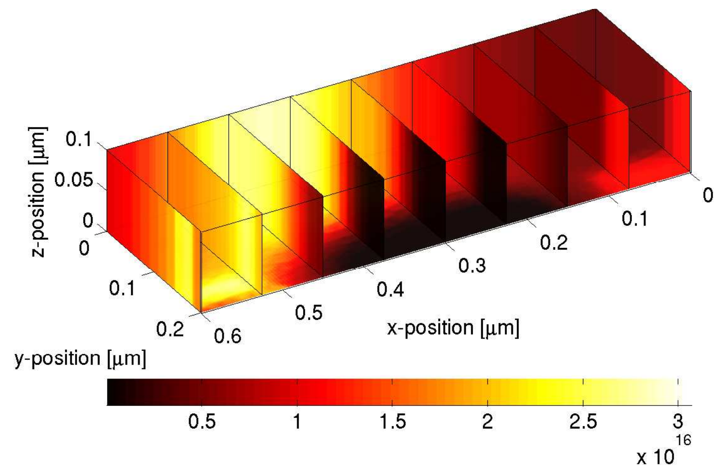



Figure 65: thermal energy $[\mathrm{eV} \cdot \mathrm{cm}^{-3}]$ in the single-gate MESFET (open state)

### 6.5.3 Double-gate MESFET

This device is a uniform extension of the 2D double-gate MESFET of section 6.4.2.

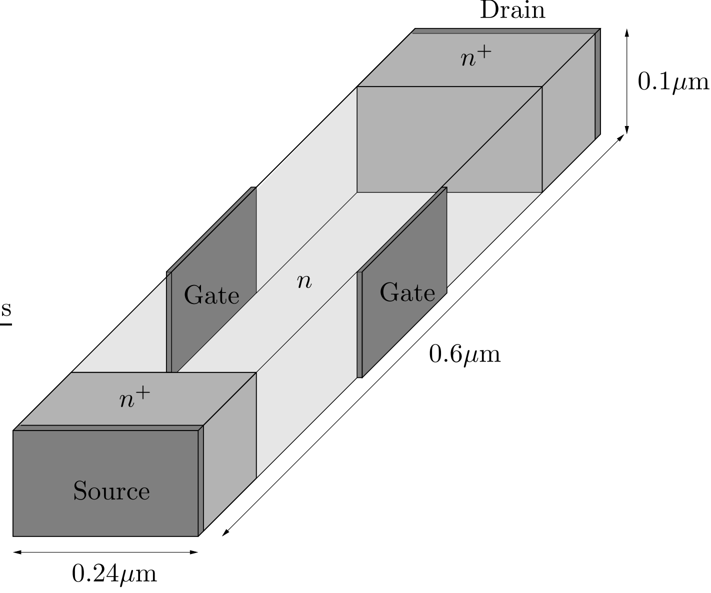



Figure 66: Schematic layout of the simulated double-gate MESFET

The doping profile function is given by (in $\mu$m)

$$\text{Dop}(x,y,z) = \begin{cases} 1\cdot 10^{17}\,\text{cm}^{-3}, & x \in [0.12\,\mu\text{m}, 0.48\,\mu\text{m}],\ y \in [0, 0.24\,\mu\text{m}], \\ 3\cdot 10^{17}\,\text{cm}^{-3}, & \text{else.} \end{cases}$$

With an applied voltage $V_A = 2\,\text{V}$ at the drain and $V_A = 0$ (open state) at the gate as well as a barrier voltage $V_B = -0.8\,\text{V}$ at the gate, the boundary data is given as follows,

- at the source: $n = 3\cdot 10^{17}\,\text{cm}^{-3},\ T = 300\,\text{K},\ V = V_{\text{bi}}$;
- at the drain: $n = 3\cdot 10^{17}\,\text{cm}^{-3},\ T = 300\,\text{K},\ V = V_{\text{bi}} + 2\,\text{V}$;
- at the gate: $n = 3.9\cdot 10^{5}\,\text{cm}^{-3},\ T = 300\,\text{K},\ V = V_{\text{bi}} - 0.8\,\text{V}$;
- at the remaining boundary segments, homogeneous Neumann boundary conditions for $I_1, I_2$, and $\nabla V$ are used.

In table 14 the Gummel-type iteration parameters used in the simulations are given.

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 2.0 | 2.0 | 150 | 15 | $2.0\cdot 10^{-2}$ | $1\cdot 10^{-6}$ | $1\cdot 10^{-2}$ | 0.7 | $1\cdot 10^{-3}$ |
| $\Phi_2$ | 15 | 3.0 | 150 | 15 | $1.33\cdot 10^{-2}$ | $1\cdot 10^{-6}$ | $1\cdot 10^{-2}$ | 0.7 | $1\cdot 10^{-3}$ |

Table 15: Parameter values in the test of the 3D double-gate MESFET

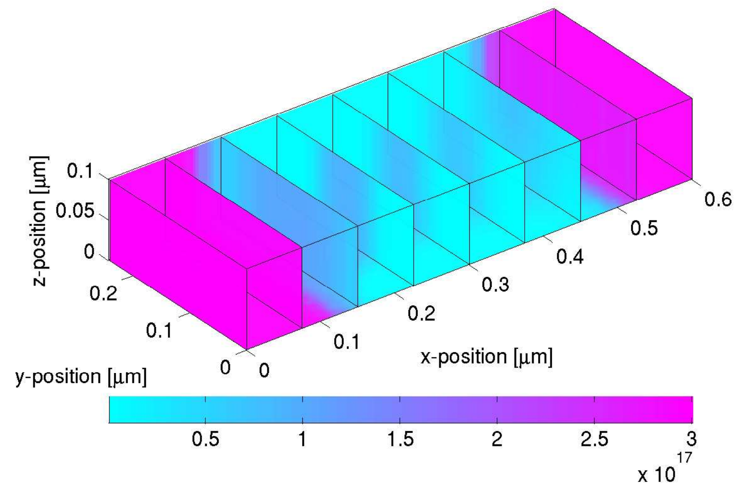



Figure 67: Electron density [cm$^{-3}$] in the double-gate MESFET (open state)

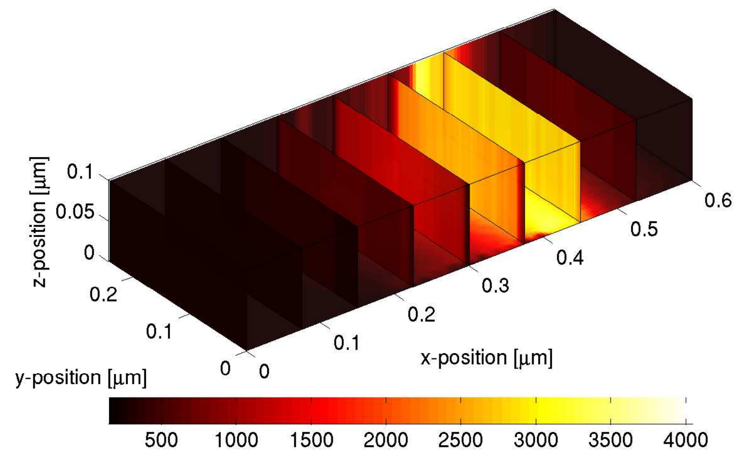



Figure 68: Electron temperature [K] in the double-gate MESFET (open state)

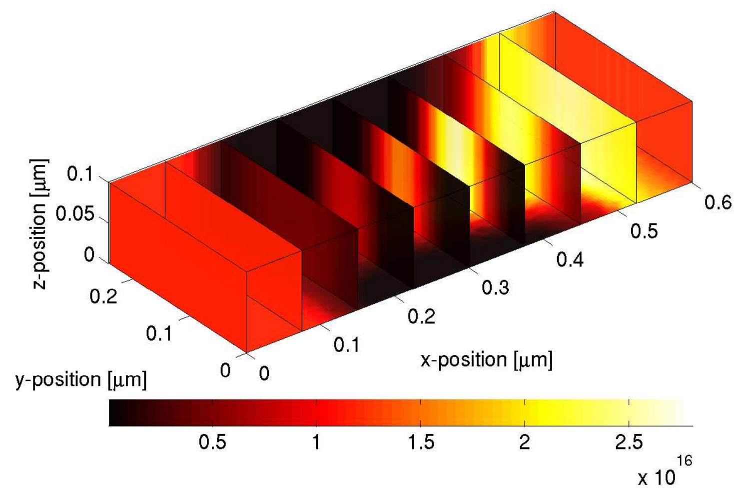



Figure 69: thermal energy [eV $\cdot$ cm$^{-3}$] in the double-gate MESFET (open state)

### 6.5.4 MOSFET

This device is the uniform extension of the 2D MOSFET. The results of the numerical tests in 2D coincide. The mesh parameter is $h = 0.04$, and the mesh consists of approx. 36300 tetrahedrons, where most of them are located in the vicinity of the oxide. The number of faces, which is the number of degrees of freedom, is approx. 75000. The numerical results shown here are computed by means of the full Newton method combined with the path following.

PSfrag replacements

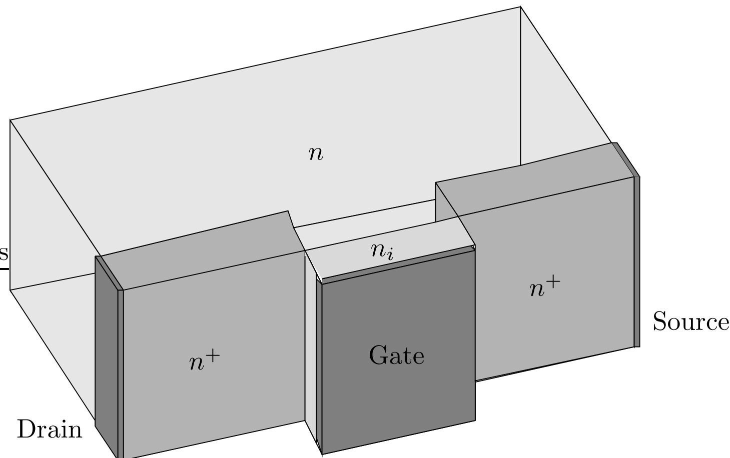



Figure 70: Schematic layout of the simulated MOSFET

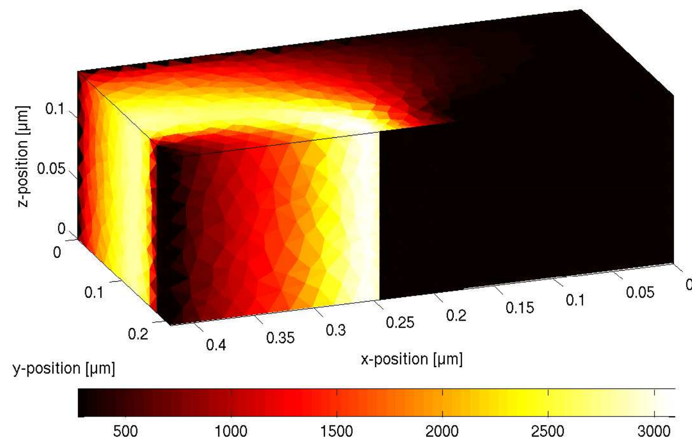



Figure 71: Electron temperature [K] in the MOSFET (open state)

## 6.6 Simulation of 3D devices

### 6.6.1 All-around gate MESFET

The all-around gate MESFET is a more advanced device. The semiconductor domain is given by $\Omega = (0, l_x) \times (0, l_y) \times (0, l_z)$, where $l_x = 0.6\mu\text{m}$, $l_y = 0.24\mu\text{m}$ and $l_z = 0.1\mu\text{m}$. The channel length is equal to $0.4\mu\text{m}$. The highly-doped regions are $\Omega_L = (0, 0.1\mu\text{m}) \times (0, l_y) \times (0, l_z)$ and $\Omega_R = (0.5\mu\text{m}, l_x) \times (0, l_y) \times (0, l_z)$. The intermediate part $\Omega \setminus (\Omega_L \cup \Omega_R)$ is the lightly-doped $n$ region. The doping profile function is

$$\text{Dop}(x, y, z) = \begin{cases} 3 \cdot 10^{18}\text{cm}^{-3}, & (x, y, z) \in \Omega_L \cup \Omega_R, \\ 1 \cdot 10^{18}\text{cm}^{-3}, & \text{else.} \end{cases}$$

The device is non-uniform, since it has a wrapped-around gate contact centered with respect to the x-direction, which is shown in figure 72. The gate contact length is $0.2\mu\text{m}$.

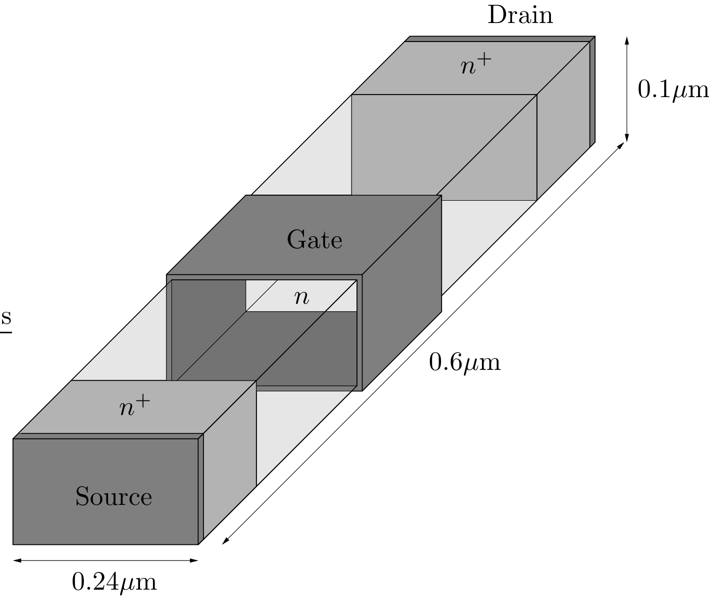



Figure 72: Schematic layout of the simulated all-around gate MESFET

In the following the boundary conditions are summarized,

- at the source: $n = 3 \cdot 10^{18}\,\text{cm}^{-3}$, $T = 300\,\text{K}$, $V = V_{\text{bi}}$;
- at the drain: $n = 3 \cdot 10^{18}\,\text{cm}^{-3}$, $T = 300\,\text{K}$, $V = V_{\text{bi}} + 2\,\text{V}$;
- at the gate: $n = 3.9 \cdot 10^{5}\,\text{cm}^{-3}$, $T = 300\,\text{K}$, $V = V_{\text{bi}} - 0.8\,\text{V}$;
- at the remaining boundary segments, homogeneous Neumann boundary conditions for $I_1, I_2$, and $\nabla V$ are used.

We employ both the Newton method and the Gummel-type iterations. The parameters stated in table 16 are used in the computations. For this device geometry, the Gummel-type iterations $\Phi_1$ and $\Phi_2$ converge very slowly. The reason may be the complex structure of the temperature profile, in particular the large gradients near the gate contact. For this reason, the full Newton scheme combined with path-following has been used. In table 17 some CPU times are given to demonstrate the effect of the path following. The step width is $\tau = 5 \cdot 10^{-3}$, the coefficient is $\overline{c} = 5 \cdot 10^{-2}$ (cf. section 6.2.5 for the definition of $\overline{c}$).

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 5.0 | 2.0 | 150 | 15 | $1.33 \cdot 10^{-2}$ | $1 \cdot 10^{-6}$ | $1 \cdot 10^{-2}$ | 0.7 | $1 \cdot 10^{-3}$ |
| $\Phi_2$ | 15 | 3.0 | 150 | 15 | $1.33 \cdot 10^{-2}$ | $1 \cdot 10^{-6}$ | $1 \cdot 10^{-2}$ | 0.7 | $1 \cdot 10^{-3}$ |

Table 16: Parameter values in the test of the all-around gate MESFET

| $h/l^*$ | without path-follow. | with path-follow. |
|---|---|---|
| 0.08 | 431 | 345 |
| 0.07 | 763 | 653 |
| 0.06 | 1247 | 1330 |

Table 17: CPU times [s] for the all-around gate MESFET (open state)

The current-voltage curve in figure 73 shows a linear increase of the drain current for small voltages which is followed by an abrupt saturation. The magnitude of the current is larger than that of the 3D double-gate MESFET due to the higher doping.
The electron current density through the device is depicted in figure 74. It is almost parallel to the x-direction except for the region under the gate contact.
The electron density in figure 75 shows a steep gradient under the wrapped-around gate contact. A region of very low concentration can be found directly underneath the gate contact, which surrounds a small part of the device with much higher concentration. This indicates that the channel is relatively small, and it explains the abrupt saturation shown in the current-voltage curve plot. It is also clear from the picture that the concentration decreases towards the drain contact in the surrounded part.
The electron temperature shown in figure 76 is different from that of the double-gate MESFET not only in magnitude (which is higher here), but also in its distribution under the gate contact. The thermal energy attains to its maximum value under the gate, where the region of high values of the thermal energy are surrounded by a region of very low energy due to the influence the very low particle density under the Schottky contact. This is shown in figure 77. Furthermore, the thermal energy increases under the gate contact starting from left left side of the channel (source) to the right.

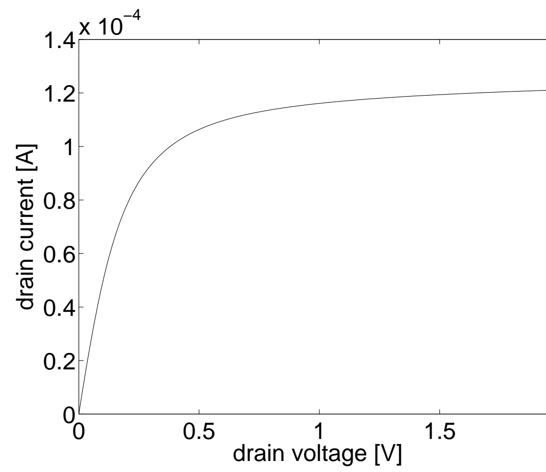



Figure 73: Current-voltage characteristic of the all-around gate MESFET (open state)

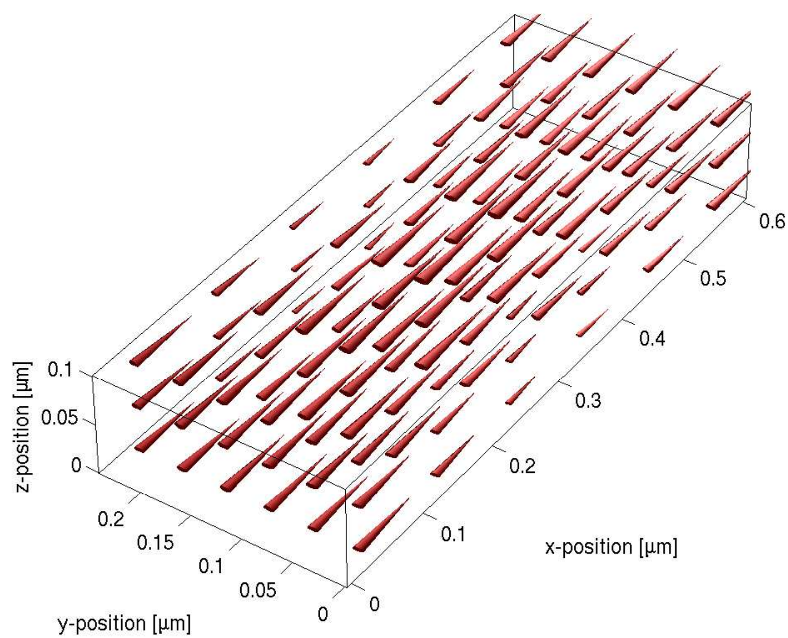



Figure 74: Electron current density in the all-around gate MESFET (open state)

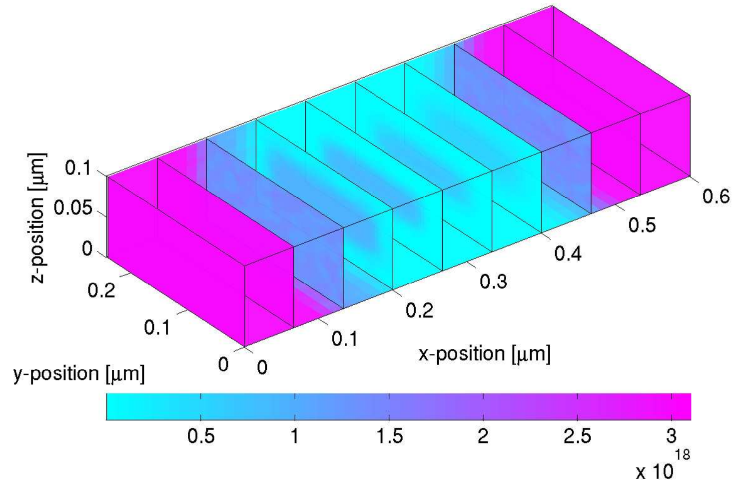



Figure 75: Electron density $[\mathrm{cm}^{-3}]$ in the all-around gate MESFET (open state)

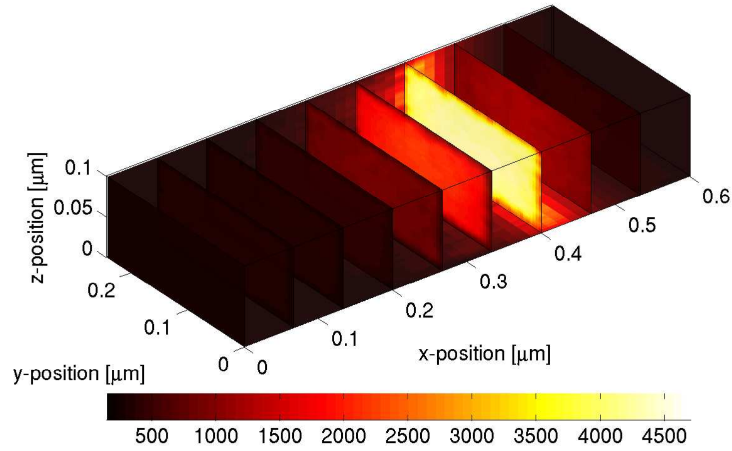



Figure 76: Electron temperature [K] in the all-around gate MESFET (open state)

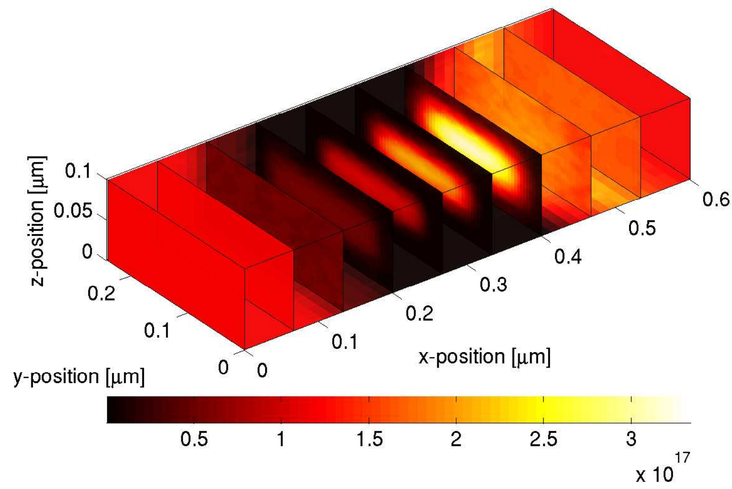



Figure 77: thermal energy $[\mathrm{eV} \cdot \mathrm{cm}^{-3}]$ in the all-around gate MESFET (open state)

### 6.6.2 Non-Uniform single-gate MESFET

The geometry of this MESFET is adopted from [ABGS03], where a GaAs device structure is considered. Here we assume that the device is made from silicon. In figure 78 the schematic layout of the device is shown. The semiconductor domain is $\Omega = (0, l_x) \times (0, l_y) \times (0, l_z)$, where $l_x = 1.0\mu\mathrm{m}$, $l_y = 0.55\mu\mathrm{m}$ and $l_z = 0.42\mu\mathrm{m}$. The source contact and the drain contact length is $0.167\mu\mathrm{m}$. The electron concentration in the highly doped region is $n^+ = 5 \cdot 10^{17}\mathrm{cm}^{-3}$. This region is surrounded by a lightly doped substrate of $0.25\,\mu\mathrm{m}$ on the front and back, with a concentration of $n = 5 \cdot 10^{15}\mathrm{cm}^{-3}$.

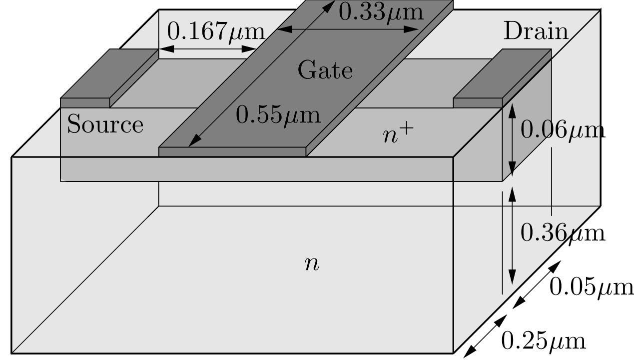



Figure 78: Schematic layout of the simulated single-gate MESFET

In the following the boundary conditions are summarized,

- at the source: $n = 5 \cdot 10^{17}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}}$;
- at the drain: $n = 5 \cdot 10^{17}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}} + 2\,\mathrm{V}$;
- at the gate: $n = 3.9 \cdot 10^{5}\,\mathrm{cm}^{-3},\ T = 300\,\mathrm{K},\ V = V_{\mathsf{bi}} - 0.8\,\mathrm{V}$;
- at the remaining boundary segments, homogeneous Neumann boundary conditions for $I_1, I_2$, and $\nabla V$ are used.

The scaled Debye length of this device is $8.8861 \cdot 10^{-3}$. The current-voltage curve in figure 79 shows no abrupt saturation. It initially increases linearly, while gradually the saturation begins. In this picture no complete saturation is visible. The electron current density in this device is shown in figure 80. The relative low doping in the bulk material allows for a ballistic flow of the electrons through this part of the device. And as expected, almost no current is there in the depletion region under the gate contact.
In figure 81 the electron density in the device is shown, where the data are computed on a grid with maximum diameter $h_{max} = 0.04\,l^*$ with $l^*$ denoting the device diameter. The typical reduction of the electron concentration under the gate contact and in the vicinity of the drain contact is visible.
In figure 82 the electron temperature is depicted. It attains its maximum value at $x = 0.4\mu\mathrm{m}$, which is the location of the gate contact edge near the drain contact. In the vicinity of the drain contact, one can see a rapid reduction of the electron temperature, especially in the highly-doped region.
In the last three pictures one can see strong variations of the physical parameters near the contacts and in the high doping region, while in the bulk material no such strong variations are detectable.

| | $\alpha^{(0)}$ | $\delta\alpha$ | $k$ | $k_1$ | $\tau$ | $\varepsilon$ | $c$ | $d$ | $\eta$ |
|---|---|---|---|---|---|---|---|---|---|
| $\Phi_1$ | 5.0 | 2.0 | 150 | 25 | $1.33 \cdot 10^{-2}$ | $1 \cdot 10^{-6}$ | $1 \cdot 10^{-2}$ | 0.7 | $1 \cdot 10^{-3}$ |
| $\Phi_2$ | 15 | 3.0 | 150 | 25 | $1.33 \cdot 10^{-2}$ | $1 \cdot 10^{-6}$ | $1 \cdot 10^{-2}$ | 0.7 | $1 \cdot 10^{-3}$ |

Table 18: Parameter values in the test of the non-uniform MESFET

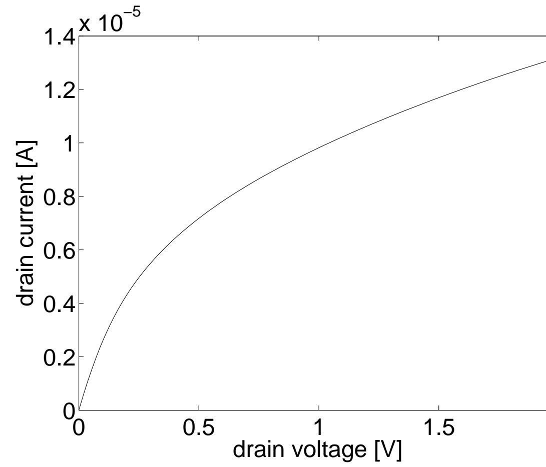



Figure 79: Current-voltage characteristic of the single-gate MESFET (open state)

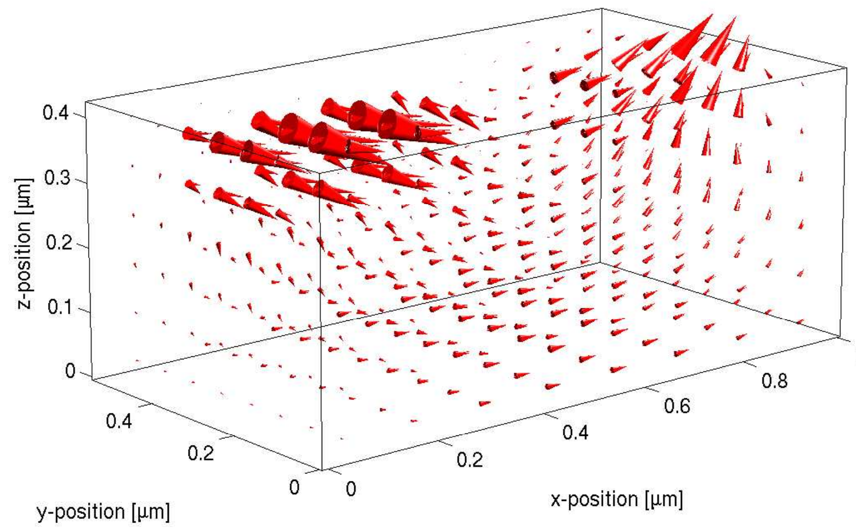



Figure 80: Electron current density in the single-gate MESFET (open state)

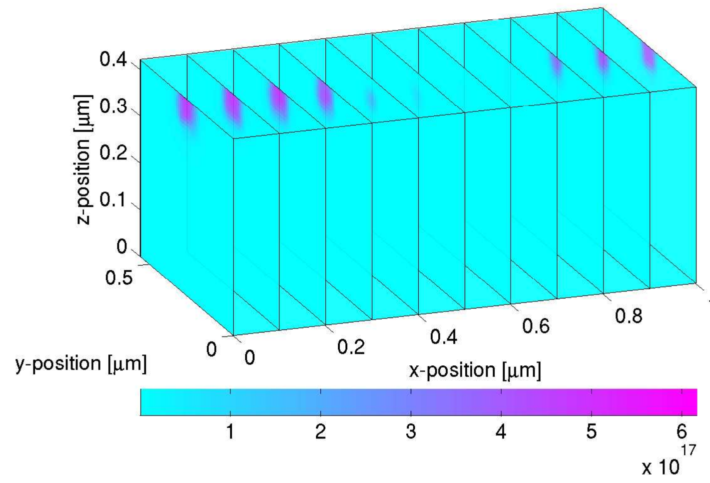



Figure 81: Electron density $[\mathrm{cm}^{-3}]$ in the single-gate MESFET (open state)

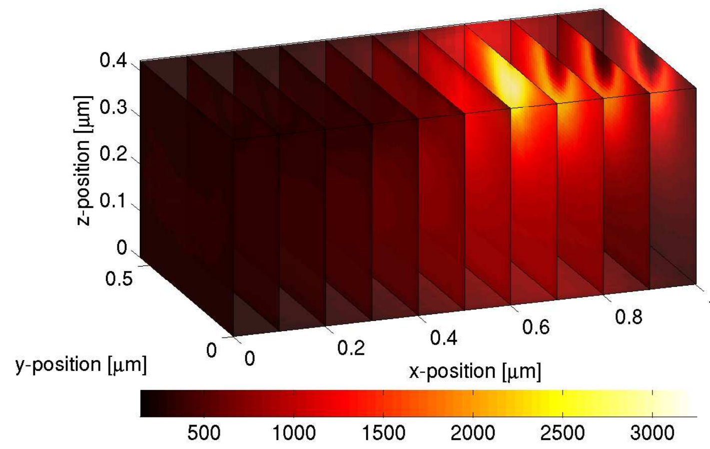



Figure 82: Electron temperature [K] in the single-gate MESFET (open state)

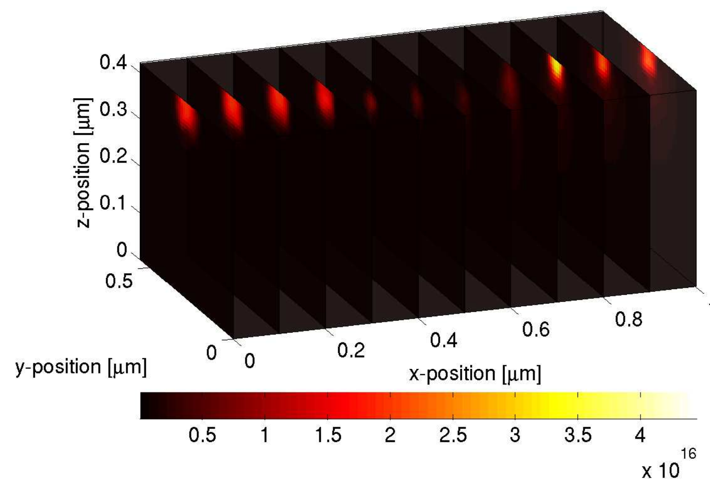



Figure 83: thermal energy $[\mathrm{eV} \cdot \mathrm{cm}^{-3}]$ in the single-gate MESFET (open state)

| parameter | physical meaning | numerical value |
|---|---|---|
| $k_B$ | Boltzmann constant | $1.3806 \cdot 10^{-23}\,\mathrm{VAsK}^{-1}$ |
| $q$ | elementary charge | $1.6 \cdot 10^{-19}$ As |
| $\varepsilon_s$ | permittivity constant (silicon) | $10^{-12}\mathrm{AsV}^{-1}\mathrm{cm}^{-1}$ |
| $\varepsilon_r(Si)$ | relative permittivity of silicon | 11.7 |
| $\mu_0$ | low-field mobility of electrons | $1500\,\mathrm{cm}^2\mathrm{V}^{-1}\mathrm{s}^{-1}$ |
| $n_i$ | intrinsic density (silicon) | $1.4 \cdot 10^{10}\,\mathrm{cm}^{-3}$ |
| $T_0$ | ambient temperature | 300 K |
| $U_T$ | thermal voltage at $T_0$ | 0.0259 V |
| $\tau_0$ | energy relaxation time | $10^{-13}$ s |

Table 19: material and model parameters

# 7 Conclusion

In this thesis an energy-transport model in the dual entropy formulation for the description of hot carrier transport in semiconductor devices was presented. The equations were deduced from the semiconductor Boltzmann equation by means of a Hilbert expansion and a diffusion scaling. In an intermediate step, the spherical harmonics expansion model (SHE model) was derived. The resulting system is of parabolic type and hence could be examined in a standard frame. For the diffusion coefficients and the energy relaxation term explicit expressions were given under the assumption that the energy band of the semiconductor material is parabolic. The Poisson equation for the electrostatic potential, which models the electric phenomena, was added to the system of continuity equations to produce a self-consistent mathematical model. Finally, the system was complemented by physically motivated mixed boundary conditions.
Then, a weak formulation was derived, where special function spaces have been used for the current densities. The discretization in finite-dimensional subspaces was designed to preserve the continuity of the current densities across finite-element boundaries. This has been accomplished by using mixed finite elements, namely Raviart-Thomas finite elements. Then a hybridization of the discrete system was introduced. By this procedure, so called Lagrange multipliers entered the system. They enforce the continuity of the normal components of the current densities.
The resulting finite-dimensional algebraic systems have been reduced by means of a static condensation procedure that produced a system acting only on the Lagrange multipliers. These finite-dimensional non-linear algebraic systems have to be solved using iteration schemes. The full Newton method combined with a path following method based on pseudo arclength continuation was applied to obtain a solution. These results and the convergence of the full Newton method were compared with those of two Gummel-type iterations that decouple the system partially or even completely. It turned out that the convergence of the Gummel-type iterations can be improved by means of a vector extrapolation, the so called reduced rank extrapolation. It was appropriate to achieve super-linear convergence and hence to built an alternative to the full Newton method.
Numerical examples have been presented in one, two and three spatial dimensions, where special attention was paid to the case of 3D devices. Applicability and accuracy of the algorithms have been tested on several numerical simulations involving the simulation of standard devices like ballistic diodes, single-gate MESFETs and single-gate MOSFETs, but also more advanced structures like an all-around gate MESFET. The research made here can be extended in several direction, where the simulation of even more complex 3D device structures is of special importance. The 3D simulation results obtained so far support the expectation that the mathematical model and the numerical algorithms can also be applied to such complex device structures. The improvement of the iteration schemes is also subject matter of further analysis.

# 8 Notations

| | |
|---|---|
| $\mathbf{A}$ | inverse diffusion matrix; matrix corresponding to $a_h(\cdot,\cdot)$ |
| $\mathbf{B}$ | matrix corresponding to $b(\cdot,\cdot)$ |
| $\mathbf{C}$ | matrix corresponding to $c(\cdot,\cdot)$ |
| $C_0$ | boundedness constant |
| $D$ | domain of (nonlinear) base sequence generator |
| Dop | doping profile function |
| $E$ | electric field, electric displacement |
| $\mathcal{E}_h, \mathcal{E}'_h$ | set of all faces of $\mathcal{T}_h$ resp. set of all inner faces of $\mathcal{T}_h$ |
| $F$ | distribution function; right hand side of discrete system; nonlinear base sequence generator |
| $G$ | solution set of the restricted weak formulation |
| $\Gamma(x)$ | Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt,\ x > 0$ |
| $\Gamma_D, \Gamma_N$ | Dirichlet boundary and Neumann boundary, resp. |
| $\Gamma_{D,1}, \Gamma_{N,1}$ | Ohmic contacts and Schottky contacts, resp. |
| $I_1, I_2$ | electron and energy current density, resp., of dual entropy formulation |
| $\widetilde{I}_1, \widetilde{I}_2$ | electron and energy current density, resp., of the modified dual entropy formulation (5.2)-(5.7) |
| $J_1, J_2$ | electron and energy current density, resp., of primal entropy formulation |
| $K$ | partition element, i.e. interval, triangle, tetrahedron |
| $L^*$ | dual lattice |
| $L^2_N$ | weighted Sobolev space |
| $\mathcal{M}$ | closed solution set |
| $N$ | number of degrees of freedom |
| $N(\varepsilon)$ | density of energy states |
| $\Omega$ | semiconductor domain; bounded, open, connected set in $\mathbb{R}^3$ |
| $P$ | minimal polynomial |
| $\mathcal{P}_0$ | subspace of $L^\infty(K)$ of constant functions on $K$ |
| $\Phi, \Phi_1, \Phi_2$ | fixed point maps |
| $Q_0$ | elastic collision operator |
| $\mathrm{RE}(V_h)$ | relative error of $V_h$ |
| $\mathcal{T}_h$ | simplicial partition of the domain $\Omega \subset \mathbb{R}^3$ |
| $T_L, T_0$ | lattice temperature and ambient temperature, resp. |
| $U$ | first differences of vectors (RRE); extension of applied voltage |
| $U^+$ | pseudo-inverse of matrix $U$ |
| $\mathbf{U}$ | solution operator of non-linear Poisson equation |
| $V$ | second differences of vectors (RRE); electrostatic potential |
| $V_A, V_B, V_{bi}$ | applied voltage, barrier voltage, built-in potential, resp. |
| $V_h^I$ | Lagrange finite element interpolation of $V_h$ |
| $\mathcal{V}^*$ | dual space of space $\mathcal{V}$ |
| $V^*$ | reference solution for calculation of relative errors |
| $\mathbf{W}$ | solution operator of linearized current continuity system |
| $Z, Z_h$ | spaces of vector valued test functions |

| | |
|---|---|
| $a(\cdot,\cdot)$ | bilinear form on $\mathcal{V}\times\mathcal{V}$ |
| $a_h(\cdot,\cdot)$ | discrete bilinear form on $X_h^2\times X_h^2$, discretization of $a(\cdot,\cdot)$ |
| $\alpha$ | typical energy gain; boundedness constant; damping parameter |
| $b(\cdot,\cdot)$ | bilinear form on $\mathcal{V}\times\mathcal{W}$ |
| $\beta$ | energy balance of acoustical collisions; model parameter; boundedness constant |
| $c$ | Ohmic contact; parameter vector in path following; error constant in Gummel-type iterations |
| $c(\cdot,\cdot)$ | bilinear form on $\mathcal{V}\times\Lambda_{h,0}^2$ |
| $c_0$ | coercivity constant |
| $d$ | error constant in Gummel-type iterations |
| $\delta$ | Dirac measure |
| $\delta_1,\delta_2$ | upper bound for quasi-Fermi levels $w_1$, $w_2$ |
| $\varepsilon$ | energy band diagram; error constant in Gummel-type iterations |
| $\varepsilon_0$ | typical kinetic energy of electrons |
| $\eta$ | natural distribution function scale; error constant in Gummel-type iterations |
| $f$ | distribution function |
| $f(\cdot)$ | linear form on $\mathcal{V}$ |
| $g(\cdot)$ | linear form on $L^2(\Omega)$ |
| $\gamma$ | quantifier of hot electron effects; inverse relation of energy band diagram; lower bound for electrostatic potential; coercivity constant; parameter in path following |
| $\gamma_n$ | normal trace operator |
| $\gamma_1,\gamma_2$ | lower bounds for quasi-Fermi levels $w_1$, $w_2$ |
| $h$ | discretization parameter |
| $k_0$ | modulus of a typical wave vector |
| $k,k_1\,k',k_1'$ | wave vector before and after collision |
| $k,k_1$ | number of base sequence vectors |
| $l:=l+1$ | reassignment |
| $\lambda$ | scaled Debye length; eigenvalue of matrix in vector extrapolation |
| $n,n^+$ | lightly doped region, highly doped region |
| $n_0$ | typical density of states (for scaling) |
| $n_1,\ldots,n_4$ | scaled outward normal vectors |
| $n_i$ | intrinsic density |
| $\nu_1,\ldots\,\nu_4$ | unit outward normal vectors |
| $\mu$ | chemical potential; test function; |
| $\mu_0$ | electron mobility |
| $\partial\Omega$ | boundary of $\Omega$ |
| $q$ | elementary charge; quality measure of tetrahedrons |
| $\rho$ | boundary value parameter in path following |
| $\mathsf{s}^*$ | fixed point of base sequence generator |
| $\mathsf{s}_0,\mathsf{s}_1,\mathsf{s}_2,\ldots$ | RRE approximations to $\mathsf{s}^*$ |
| $\tau$ | test function; step width in path following |
| $\tau_0$ | energy relaxation time; basis function of local space |
| $\tau_{ij}$ | directional vectors of edges |
| $\mathsf{x}_0,\mathsf{x}_1,\mathsf{x}_2,\ldots$ | vectors produced by means of base sequence generator |

| | |
|---|---|
| $H^1_{0,\Gamma_D}(\Omega)$ | Sobolev space of $H^1(\Omega)$ functions with zero trace on Dirichlet boundary $\Gamma_D$ |
| $H_{0,\Gamma_N}$ | space of functions of $H(\text{div};\Omega)$ with normal traces vanishing on $\Gamma_N$ |
| $H_0(\text{div};\Omega)$ | space of divergence-free functions of $H_{0,\Gamma_N}$ |
| $\|\cdot\|_{\text{div}}$ | norms in $H(\text{div});\Omega)$ or $H(\text{div};\Omega)\times H(\text{div};\Omega)$ depending on the context |
| $\|\cdot\|_{0,2,\Omega}$ | norms in $L^2(\Omega)$ or $L^2(\Omega)\times L^2(\Omega)$ depending on the context |
| $\|\cdot\|_{0,\infty,\Omega}$ | norms in $L^\infty(\Omega)$ or $L^\infty(\Omega)\times L^\infty(\Omega)$ depending on the context |
| $\|\cdot\|_{1/2,2,\Gamma_D}$ | norms in $H^{1/2}(\Gamma_D)$ or $H^{1/2}(\Gamma_D)\times H^{1/2}(\Gamma_D)$ depending on the context |
| $\|\cdot\|_*$ | norm in dual space depending on the context |
| $\|\cdot\|_2$ | Euclidean norm |
| $\|\cdot\|_\infty$ | discrete maximum norm |
| $\|\cdot\|_N$ | norm in weighted space $L^2_N$ |
| $\Delta$ | differential expression $\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}$ (Laplace operator) |
| $\Delta$, $\Delta^2$ | first order and second order differences |
| $\Delta^3$ | three-dimensional (open) standard simplex |
| $\nabla\cdot,\text{div}$ | divergence of a vector field |
| $\nabla$, $\nabla_x$ | gradient of a scalar function (with respect to $x$) |
| log | natural logarithm |
| $\|K\|$ | area or volume of a partition element $K$ |
| $\otimes$ | tensor product |
| $(kerL)^\perp$ | orthogonal complement of subspace $kerL$ |
| $\mathbb{R}^3/L^*$ | factor space, factorization of Euclidean space w.r.t. dual lattice |
| $G'$ | total derivative of $G$ |
| $\partial_1 G,\partial_2 G$ | partial derivative of $G$ w.r.t. the first argument and second argument |
| $u^{(k)}$ | $k$th iterative solution |
| $O$ | Landau symbol |

# References


[AB85] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. $M^2$ *AN*, 19,1:7–32, 1985.

[ABC+95] 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.

[ABGS03] S. Aboud, J. Brandlard, S. Goodnick, and M. Saraniti. Frequency analysis of 3D GaAs MESFET structures using full-band particle-based simulations. *Journal of Computational Electronics*, 2:213–217, 2003.

[AD96] N. Ben Abdallah and P. Degond. On a hierarchy of macroscopic models for semiconductors. *J. Math. Phys.*, 37:3308–3333, 1996.

[ADMS01] N. Ben Abdallah, P. Degond, P. Markovich, and C. Schmeiser. High field approximations of the spherical harmonics expansion model for semiconductors. *ZAMP*, 52, no. 2:201–230, 2001.

[Ait37] A. C. Aitken. The evaluation of the latent roots and latent vectors of a matrix. *Proc. Roy. Soc. Edinburgh*, 57:269–304, 1936-1937.

[AW99] L. Angermann and S. Wang. An exponentially fitted conforming tetrahedral finite element method for the semiconductor continuity equations. Technical report, preprint No.10, "Fakultät für Mathematik, Otto-Von-Guericke-Universität Magdeburg", 1999.

[BF91] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods.* Springer, Berlin, 1991.

[Blø70] K. Bløtekjær. Transport equations for electrons in two-valley semiconductors. *IEEE Trans. El. Dev.*, 17:38–47, 1970.

[BMM+05] F. Brezzi, L. Marini, Micheletti, P. Pietra, R. Sacco, and S. Wang. Discretization of semiconductor device problems. In W. Schilders and E. ter Maten, editors, *Handbook of Numerical Analysis. Numerical Methods for Electrodynamical Problems.* North-Holland-Elsevier Science Publishers, 2005.

[BMP89] F. Brezzi, L. Marini, and P. Pietra. Two-dimensional exponential fitting and applications to drift-diffusion models. *SIAM J. Num. Anal.*, 26:1342–1355, 1989.

[BS02] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods.* Springer, New York, 2nd edition, 2002.

[BSSG98] F. Bosisio, R. Sacco, F. Saleri, and E. Gatti. Exponentially fitted mixed finite volumes for energy balance models in semiconductor device simulation. In H. Bock et al., editor, *Proceedings of ENUMATH 97*, pages 188–197, Singapore, 1998. World Scientific.

[CCJS95] Z. Chen, B. Cockburn, J. Jerome, and C. Shu. Finite element computation of the hydrodynamic model of semiconductor devices. *VLSI Design*, 3:145–158, 1995.

[CG04] B. Cockburn and J. Gopalakrishnan. A characterization of hybridized mixed methods for second order elliptic equations. *SIAM J. Numer. Anal.*, 42:283–301, 2004.

[CGDH00] E. Cassan, S. Galdin, P. Dollfus, and P. Hesto. Comparison between device simulators for gate current calculation in ultra-thin gate oxide n-MOSFETs. *IEEE Trans. Elect. Dev.*, 83:1194–1202, 2000.

[Cia91] P. G. Ciarlet. *Basic error estimates for elliptic problems.* in: P.G. Ciarlet, J.L. Lions (Eds.), Finite Element Methods, Part 1, North-Holland, Amsterdam, 1991.

[CKR+92] D. Chen, E. Kan, U. Ravaioli, C. Shu, and R. Dutton. An improved energy transport model including nonparabolicity and non-Maxwellian distribution effects. *IEEE Electr. Dev. Letters*, 13:26–28, 1992.

[CSP+92] D. Chen, E. Sangiorgi, M. Pinto, E. Kan, U. Ravaioli, and R. Dutton. Analysis of spurious velocity overshoot in hydrodynamic simulations. *NUPAD IV*, pages 109–114, 1992.

[DGJ97] 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.

[DJP00] P. Degond, A. Jüngel, and P. Pietra. Numerical discretization of energy-transport models for semiconductors with non-parabolic band structure. *SIAM J. Sci. Comp.*, 22:986–1007, 2000.

[DS99] P. Degond and C. Schmeiser. Kinetic boundary layers and fluid-kinetic coupling in semiconductors. *Transp. Theory Stat. Phys.*, 28:31–55, 1999.

[dV65] B. Fraejis de Veubeke. Displacement and equilibrium models in the finite element method. In O. C. Zienkiewicz and G. Holister, editors, *Stress Analysis*, New York, 1965. Wiley.

[FMWW90] C.J. Fitzsimons, J.J.H. Miller, S. Wang, and C.H. Wu. Hexahedral finite elements for the stationary semiconductor device equations. *Comp. Meth. Appl. Mech. Engrg.*, 84(5):43–57, 1990.

[Fou99] M. Fournié. *Construction et analyse de schémas compacts d'ordre élevé pour des problèmes fortement convectifs. Application à la simulation de semi-conducteurs.* PhD thesis, Université Raul Sabatier, Toulouse, France, 1999.

[Geo91] P. L. George. *Automatic Mesh Generation - Application to Finite Element Methods.* Wiley, 1991.

[GJR89] C. Gardner, J. Jerome, and D. Rose. Numerical methods for the hydrodynamic device model: subsonic flow. *IEEE Trans. CAD*, 8:501–507, 1989.

[GN99] 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.

[GR86] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations, theory and algorithms.* Springer, Berlin, New York, 1986.

[GR94] Ch. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen.* Teubner, Stuttgart, 1994.

[Gri99] 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.

[GT77] D. Gilbarg and N. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Springer, New York, 1977.

[HJP03] S. Holst, A. Jüngel, and P. Pietra. A mixed-finite-element discretization of the energy-transport model for semiconductors. *SIAM J. Sci. Comp.*, 24:2058–2075, 2003.

[HJP04] S. Holst, A. Jüngel, and P. Pietra. An adaptive mixed scheme for energy-transport simulators of field-effect transistors. *SIAM J. Sci. Comp.*, 25:1698–1716, 2004.

[Jer95] J. Jerome. The approximation problem for drift-diffusion systems. *SIAM Review*, 37:552–572, 1995.

[Jer96] J. Jerome. *Analysis of Charge Transport. A mathematical study of semiconductor devices.* Springer, 1996.

[JK91] J. Jerome and T. Kerkhoven. A finite-element approximation theory for the drift-diffusion semiconductor model. *SIAM J. Num. Anal.*, 28:403–422, 1991.

[JP97] A. Jüngel and P. Pietra. A discretization scheme of a quasi-hydrodynamic semiconductor model. *Math. Models Meth. Appl. Sci.*, 7:935–955, 1997.

[JS94] J. Jerome and C.-W. Shu. Energy transport 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*, pages 185–207, New York, 1994. Springer.

[Jün01] A. Jüngel. *Quasi-hydrodynamic Semiconductor Equations.* Progress in Nonlinear Differential Equations. Birkhäuser, Basel, 2001.

[LC99] C. Lab and P. Caussignac. An energy-transport model for semiconductor heterostructure devices: application to AlGaAs/GaAs MOSFETs. *COMPEL*, 18:61–76, 1999.

[LH74] C. L. Lawson and R. J. Hanson. *Solving least square problems.* Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[LM68] J.L. Lions and E. Magenes. *Problèmes aux limites non-homogènes et applications.* Dunod, Paris, 1968.

[LPSV92] E. Lyumkis, B. Polsky, A. Shur, and P. Visocky. Transient semiconductor device simulation including energy balance equation. *COMPEL*, 11:311–325, 1992.

[MMP96] 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*, London, 1996. John Wiley.

[MRS90] P. A. Markovich, C. A. Ringhofer, and C. Schmeiser. *Semiconductor equations.* Springer, Vienna, 1990.

[RGQ93] M. Rudan, A. Gnudi, and W. Quade. A generalized approach to the hydrodynamic model of semiconductor equations. In G. Baccarani, editor, *Process and Device Modeling for Microelectronics*, pages 109–154, Amsterdam, 1993. Elsevier.

[Rin01] C. Ringhofer. An entropy-based finite difference method for the energy transport system. *Math. Models Meth. Appl. Sci.*, 11:769–796, 2001.

[RT77] P. Raviart and J. Thomas. A mixed finite element method for second order elliptic equations. In *Mathematical Aspects of the Finite Element Method*, volume 606 of *Lecture Notes in Mathematics*, pages 292–315. Springer, 1977.

[Sel84] S. Selberherr. *Analysis and Simulation of Semiconductor Devices.* Springer, Vienna, 1984.

[SFS87] D. A. Smith, W. F. Ford, and A. Sidi. Extrapolation methods for vector sequences. *SIAM Review*, 29:199–233, 1987.

[SOTG94] K. Souissi, F. Odeh, H. Tang, and A. Gnudi. Comparative studies of hydrodynamic and energy transport models. *COMPEL*, 13:439–453, 1994.

[SS97] R. Sacco and F. Saleri. Mixed finite volume methods for semiconductor device simulation. *Numer. Meth. Part. Diff. Eq.*, 13:215–236, 1997.

[STL96] H. Schwetlick, G. Timmermann, and Ralf Lösche. Path following for large nonlinear equations by implicit block elimination based on recursive projections. *Lectures in Applied Mathematics*, 32:715–732, 1996.

[Str62] R. Stratton. Diffusion of hot and cold electrons in semiconductor barriers. *Phys. Rev.*, 126:2002–2014, 1962.

[Str76] G. Strang. *Linear algebra and its applications.* Academic Press, New York, 1976.

[Sze81] S. Sze. *Physics of Semiconductor Devices.* John Wiley, New York, 1981.

[Tem77] R. Temam. *Navier-Stokes equations.* North-Holland, Amsterdam, 1977.

[Vis94] P. Visocky. A method for transient semiconductor device simulation using hot-electron transport equations. In J. Miller, editor, *Proc. of the Nasecode X Conf.*, Dublin, 1994. Boole Press.

[Wal95] W. Walus. Computational methods for the Boltzmann equation. In N. Bellomo, editor, *Lecture Notes on the Mathematical Theory of the Boltzmann Equation*, pages 179–223, Singapore, 1995. World Scientific.

[WT93] H.-S. Wong and Y. Taur. Three-Dimensional 'Atomistic' Simulation of Discrete Microscopic Random Dopant Distributions Effects in Sub-0.1 $\mu$m MOSFETs. *IEDM Tech. Digest*, pages 705–708, 1993.

# Index

Numbers written in italic refer to the page where the corresponding entry is described; numbers underlined refer to the definition; numbers in roman refer to the pages where the entry is used.


**A**

all-around gate MESFET . . 87

**B**

ballistic diode 54–57, 59, 79, 80
barrier height . . . *see* barrier potential
barrier potential . . . . . . 15, 66
barrier voltage . . . . . . *see* barrier potential, 81, 83
bias continuation . . 34, 40, 51
Boltzmann collision operator 8
Boltzmann equation . . . . . 5, 7, 10, 14, 55, 59
boundary condition 6, 15, 18, 19
Brillouin zone . . . . . . . 7, 9, 15

**C**

Cholesky factorization . . 49, 50
conjugate gradient method . . . . . . . . . . . 37, 49, 50
current continuity system 6, 18–20, 25, 36–38, 40, 43
current continuity system, decoupled . . . . . . . . . 40

**D**

Debye length, scaled . . . . . 17, 54, 58, 66, 74, 90
diffusion coefficients . . . 5, 15–17, 19, 20, 23, 24, 31, 34, 36–38, 41, 43, 50
diffusion equations . . . . 13, 14
diffusion scaling . . . . . . . . . 10
double-gate MESFET . . . . 66, 67, 69–73, 83, 84
drain current . . . . . . . . . . . 59
drain voltage . . . . . . . . . . . 59
drift-diffusion model . 5, 39, 43

**E**

eigenvalue . . . 23, 44, 46, 47, 58
elastic scattering . . . . . . . . 5, 8
electron density . . . . . . . . 15–17, 34, 37, 39–41, 55
electron temperature . . . . . 14, 15, 17, 54, 55
electron-electron scattering . 5
electrostatic potential . 15, 17, 19, 36, 39, 41, 55, 57
energy balance equation . . . . . . . . . . . . . 40, 41, 43
energy relaxation term . . . . 5, 15–17, 20, 23, 24, 43
energy state . . . . . . . . . . . . 7
energy-transport equations . . . . 5, 6, 15, 39, 59, 74
energy-transport model . . . . . . . 5, 7, 12, 14, 15, 17

**F**

fixed point map . . . . . . . . 19, 24, 38, 39, 41, 43
Fokker-Planck approximation . . . . . . . . . . . . 16

**G**

Gauss' law . . . . . . . . . . . . . 14
grid generator . . . . . . . . . . . 52
grid parameter . . . . . . . . . . 54
Gummel-type iteration . . . . . . 6, 34, 39, 40, 43, 46, 48–50, 54, 55, 58, 59

**H**

Hilbert expansion . . . . . . 5, 10

**I**

impurity scattering . . . . . . . 7
inelastic scattering . . . . . 8, 10

**J**

Joule heating term . . . . . . . 18

**L**

lattice temperature . 13, 17, 54

**M**

macroscopic model . . . . . . . 7
matrix assembling . . . . . . . . 49
maximum principle . 19, 41, 51
mesh parameter . . . . . . . . *see* grid parameter, 51
mixed finite elements . . . . . 6
mixed hybrid finite elements . . . . . . . . . . . 27, 28, 51
MOSFET . . . . . . 74–76, 78, 85

**N**

Newton iteration . 6, 34, 36, 39, 49, 50, 54, 59, 67, 79
non-uniform MESFET . . . . 91

**O**

Ohmic contact . . . . 15, 31, 32, 34, 35, 54, 58, 66, 79

**P**

path following method . . . . . 6, 34, 35, 51, 59, 67, 79
phonon . . . . . . . . . . . . . . . 7, 8
phonon scattering . . . . . . . 5, 7
Planck constant . . . . . . . . . 7
Poisson equation . . . . . . 6, 13–15, 19, 36, 37, 39, 40, 42, 43, 47–50, 74, 79
post-processing . . . . 31, 50, 55
preconditioner . . . . . . . 49, 50

**Q**

quasi-Fermi level . . . . . 17, 19, 21, 26, 36, 38, 39, 41

**R**

Raviart-Thomas finite elements . . . . . . 25–27, 51
reduced rank extrapolation (RRE) 43, 44, 49, 50, 93
relative errors . . . . . . . . . . . 55

**S**

Schottky contact . . . 15, 35, 66
SHE model . . . . . . . . . . . . 5, 7
single-gate MESFET . . . . . 58, 60–65, 67, 78, 81, 82
sparse matrix . . . . . . . . . . . 49
static condensation . 30, 31, 38

**T**

terminal current . . . 31, 32, 54
thermal energy . . . . . . . . . . 92
thermal equilibrium . . . . . . . . . . . . . 12, 36, 40, 54

**U**

uniform devices . . . . . . . . . . 79

**V**

vector extrapolation . . . . . *see* reduced rank extrapolation, 43, 46, 49, 58

**W**

wave vector . . . . . . . . . . . . 7

**Z**

zero order term . . 25, 37, 38, 41