University of Warsaw

Faculty of Mathematics, Informatics
and Mechanics

and

Institute of High Pressure Physics
Polish Academy of Sciences

Konrad Sakowski

Determination of the properties of the nitride laser diodes and light-emitting diodes by simulation based on the drift-diffusion model with the Discontinuous Galerkin Method

PhD dissertation

Supervisors:
dr hab. Leszek Marcinkowski
Institute of Applied Mathematics
and Mechanics
University of Warsaw

prof. dr hab. Stanisław Krukowski
Institute of High Pressure Physics
Polish Academy of Sciences

April 25, 2017

Author's declaration:

aware of legal responsibility I hereby declare that I have written this dissertation myself and all the contents of the dissertation have been obtained by legal means.

April 25, 2017

mgr Konrad Sakowski

Supervisors' declaration:

the dissertation is ready to be reviewed.

April 25, 2017

dr hab. Leszek Marcinkowski
prof. dr hab. Stanisław Krukowski

Abstract

In this dissertation we discuss numerical simulations of semiconductor devices with the drift-diffusion model. This model consists of three nonlinear elliptic differential equations known as van Roosbroeck equations. We consider two variants of discretization of this system with the Composite Discontinuous Galerkin Method, based on Symmetric Interior Penalty Galerkin method and on Weakly Over-Penalized Symmetric Interior Penalty method. We discuss these methods in context of simulations of luminescent devices based on gallium nitride and its mixed compounds. The proposed methods account for lower regularity of solutions on the interfaces between layers of different materials. It is shown that for in the equilibrium state regime the discrete problems are well posed and error estimates for the piecewise-linear finite element spaces in one- and two-dimensional domains are derived. These error bounds are then verified against results of numerical simulations, which covered both abstract settings and real semiconductor structures, in equilibrium and non-equilibrium state.

It is also demonstrated that these methods may be used in determination of physical properties of the luminescent devices. For this purpose, a linearization of the discretized van Roosbroeck equations based on the Newton method and Picard method was developed. Then the numerical code was developed, which makes use of these discretizations and the linearization. Since the physical phenomena related to operation of the gallium nitride laser diodes and light-emitting diodes are not fully understood, our application allows for modification of the underlying differential model to some extent. Examples of such modifications and their effects are presented.

Acknowledgments

I would like to thank my supervisors dr hab. Leszek Marcinkowski and prof. Stanisław Krukowski for their great help, suggestions and guidelines during the process of writing this dissertation. I am also very grateful to my colleagues, dr Paweł Kempisty, dr Paweł Siedlecki and dr Paweł Strąk for assistance and many fruitful discussions related to this work. I am indebted to dr hab. Agnieszka Kałamajska and prof. Maksymilian Dryja, whose valuable advices guided me to the path I currently follow. And I would like to thank my family for their support and encouragement, in particular to my wife Adriana, for her love and patience.

Contents

Introduction . . . 11

1 Discretization of van Roosbroeck equations **21**

1.1 Finite Element Method basics . . . 23
1.1.1 Continuous Finite Element Method . . . 23
1.1.2 Discontinuous Galerkin Method . . . 29
1.1.3 Weakly Over-Penalized Symmetric Interior Penalty . . . 31
1.2 Differential problem . . . 32
1.2.1 Drift-diffusion system . . . 32
1.2.2 Equilibrium state . . . 33
1.3 Composite Discontinuous Galerkin Method . . . 34
1.3.1 Discrete space . . . 34
1.3.2 Composite Discontinuous Galerkin variants . . . 36
1.3.3 Broken norm and the Poincare inequality . . . 39
1.3.4 Consistency . . . 41
1.4 Discretization of the equilibrium case . . . 47
1.4.1 Composite Weakly Over-Penalized Symmetric Interior Penalty (CWOPSIP) . . . 47
1.4.2 Composite Symmetric Interior Penalty Galerkin (CSIPG) . . . 48
1.4.3 Existence and uniqueness . . . 48
1.5 Interpolation operator and interpolation error . . . 52
1.5.1 One dimension . . . 53
1.5.2 Two dimensions . . . 53
1.6 Error estimates for the equilibrium case for CWOPSIP . . . 56
1.6.1 Outline of the proof . . . 56
1.6.2 Consistency . . . 57
1.6.3 Analysis . . . 58
1.6.4 Summary . . . 61
1.7 Error estimates for the equilibrium case for CSIPG . . . 63
1.7.1 Consistency . . . 63
1.7.2 Analysis . . . 64
1.7.3 Summary . . . 67

2 Numerical simulations of semiconductor devices **69**

2.1 Band structure of GaN, AlN and InN . . . 70
2.1.1 Bandgap . . . 70
2.1.2 Effective mass . . . 71
2.1.3 Current . . . 73
2.1.4 Carrier statistics . . . 73
2.1.5 Doping . . . 77

2.1.6 Energy distribution in a crystal structure . . . 81
2.2 Properties of the mixed AlGaN and InGaN crystals . . . 81
2.3 Geometry of luminescent semiconductor structures . . . 83
2.3.1 p-n homojunction . . . 83
2.3.2 Laser diodes and electroluminescent diodes . . . 84
2.4 Quantum structures: wells and barriers . . . 84
2.5 Drift-diffusion model . . . 85
2.5.1 Conservation laws and equations of motion . . . 85
2.5.2 Electric field, electrostatic potential and polarization effect . . . 87
2.5.3 Differential problem . . . 88
2.5.4 Equilibrium state and non-equilibrium state . . . 89
2.5.5 Built-in potential . . . 90
2.5.6 Boundary conditions . . . 90
2.6 Radiative and non-radiative recombination . . . 91
2.6.1 Standard recombination models . . . 91
2.6.2 Impact ionization . . . 97
2.6.3 Trap levels . . . 99
2.7 Tunneling quantum effect . . . 103
2.7.1 Trap-assisted tunneling . . . 103
2.8 P-N diode . . . 104
2.8.1 p-n homojunctions . . . 105
2.8.2 Homojunctions, p-i-n diodes and single quantum well structures . . . 109
2.8.3 Comparison with available software . . . 121
2.8.4 Computing carrier currents . . . 129
2.8.5 Trap-assisted tunneling effect on characteristics of gallium nitride diodes . . . 133
2.9 Light-emitting diodes and laser diodes . . . 138
2.9.1 Introduction . . . 138
2.9.2 Aluminum content in EBL . . . 140
2.9.3 Mg doping of p-type . . . 142
2.9.4 Number of quantum wells . . . 144
2.10 Optical excitation in quaternary alloy AlInGaN . . . 145

3 Linearization and convergence study 149
3.1 Linearization method . . . 149
3.1.1 Newton method . . . 150
3.1.2 Newton method with backtracking . . . 152
3.1.3 Comparison . . . 155
3.2 Error analysis: numerical experiments . . . 157
3.2.1 Introduction . . . 157
3.2.2 Formulation u, v, w . . . 158
3.2.3 Formulation ψ, F_n, F_p: one dimension . . . 175
3.2.4 Formulation ψ, F_n, F_p: two dimensions . . . 179
3.3 Discontinuities on interfaces . . . 181
3.3.1 Differential problem . . . 181
3.3.2 Discrete problems . . . 184
3.3.3 Simulations . . . 186

4 Appendix **193**
4.A Theorems . . . 193
4.B Lemmas . . . 195
4.C Existence of discrete solutions in one dimension . . . 199
4.C.1 Operator T . . . 200
4.C.2 Discrete operator $\tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$ . . . 201
4.C.3 Discrete operator $v_h(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h)$ . . . 210
4.C.4 Discrete operator $w_h(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h)$ . . . 216
4.C.5 Discretization of the van Roosbroeck system . . . 216
4.C.6 Analysis . . . 217
4.D List of assumptions . . . 222
4.E Physical constants . . . 223

Introduction

The digital revolution, which has been going for last 40 years, has its origins in development of the semiconductor transistor in 1960s. It was a culmination of a 30 year long research conducted in Bell Labs aiming at replacing electromechanical switching equipment components with semiconductor devices[52]. Then the transistors quickly superseded vacuum tubes. They were installed in wide range of appliances: television sets, radio receivers, communication devices, etc. First calculator based on semiconductor transistors was introduced in 1963, pocket calculators followed roughly 10 years later. At the same time minicomputers were developed, followed then in 1970s and 1980s by microcomputers — machines similar in size and design to currently widespread personal computers. The progress in this domain was generally possible due to miniaturization.

While transistors have a profound impact on live of modern society, they are not the only application of the semiconductors. An important phenomenon within the context of semiconductor material is the electroluminescence. Electroluminescence is a light emission by a material due to the electric current. In contrast to the incandescence effect, the light is not generated due to the heat. In semiconductors, it is generated by the radiative recombination of electrons and holes (described in detail in section 2.6). The electroluminescence was discovered in 1907 by H. J. Round [125], in silicon carbide (SiC). First SiC light-emitting diode (LED), was reported in 1927 by Losev. Principles of operation of these diodes were explained in 1951 by Lehovec et al. [65]. Then in 1955 Braunstein obtained infrared emission from gallium arsenide (GaAs) [16]. Shortly thereafter in 1960s commercial-grade infrared diodes was introduced by Texas Instruments. Simultaneously first red light-emitting diode was developed by Holonyak and Bevacqua [58]. In 1970s the red LED technology was mature enough so that these devices became available in many appliances.

While first gallium nitride (GaN) light-emitting diode was introduced at this time, it took twenty years to develop high-brightness blue LED by Nakamura et al. [81] in early 1990s. These diodes were composed of gallium nitride and mixed compounds: indium gallium nitride (InGaN) and aluminum gallium nitride (AlGaN). For this achievement, Nakamura was awarded the 2014 Nobel prize in physics jointly with Akasaki and Amano, who had a significant contribution in development of growth of high-quality gallium nitride [4]. They also introduced efficient p-type doping of GaN with magnesium (Mg) [3]. In early 2010s the gallium nitride devices matured and became available in commerce.

In parallel to LEDs, the laser diodes (LDs) were also developed. First GaAs infrared laser diodes were introduced in 1962 by Hall et al. [49] and Nathan et al. [83]. The same year they were followed by red gallium arsenide phosphide (GaAsP) LDs developed by Holonyak. In 1960s it was discovered that heterostructures (semiconductor structures composed of layers of different material) are generally better candidates for LDs than homojunctions (made of a single material). In particular, in early 1970s the quantum wells were introduced [34], increasing greatly the LD efficiency. In a short time semiconductor LD technology were applied to fiber optic communication and to optical data storage. The latter then evolved to compact discs and they become commercially available in early 1980s. These devices used infrared (CDs) and red lasers (DVDs), based on gallium arsenide. In 1996, few years after introduction of efficient GaN LED, Nakamura demonstrated first gallium nitride blue laser

[82]. In the meantime, the high-pressure growth technology of the low-defect gallium nitride crystal substrates was developed by Institute of High Pressure Physics of the Polish Academy of Sciences. These crystals were used by Nakamura in 1999 yielding 30 mW LDs with 3000 h lifetime — as opposed to 300 h and 15 mW at the same current for sapphire-supported GaN structures of that time [46]. Due to further studies of GaN-based LDs, prices of blue and violet semiconductor lasers dropped and these devices became available on the market. The applications followed: data storage (Blu-ray discs), digital projectors, solid state lighting, medical equipment, etc.

The technology of blue LEDs and LDs matured, but there are still many problems which remain unsolved. While red and blue LEDs reach above 30% (wall plug) efficiency, green and yellow LEDs are still under 20% [31]. To produce green light, it is more effective to use a blue diode and then convert it with a phosphorescent layer to the green light (about 22% efficiency) than to use native green LEDs. Moreover, GaN-based LEDs exhibit drop in efficiency with increasing current (*efficiency droop*) [102], which is attributed to the Auger recombination [53]. This phenomenon is absent in red GaAs-based LEDs. On the other hand, there is demand on the efficient green lasers, which could be used in wall projectors with blue and red lasers.

Numerical simulations are the important tool in development of semiconductor devices. Since our contemporary electronics relies on the semiconductors, there is strong demand on the progress in this domain. There are various approaches in simulations of such devices, depending on precision, efficiency and size of a simulated fragment. On the one hand there are so-called *ab initio* methods, which are used to investigate properties of elements composed of hundreds of thousands of atoms. These methods use fundamental laws of physics and they need days or weeks to perform a single simulation on a computational cluster. On the other hand, there are statistical methods, which approximate complex interactions of a large number of similar objects by elementary statistical laws. Among these methods, there is a drift-diffusion theory. In this case the model is straightforward and it allows to simulate whole semiconductor device on a standard desktop computer. This model describes two kinds of quasiparticles, electrons and holes, which move in the electric field present in semiconductor devices. From the mathematical point of view, the model consists of a system of three nonlinear elliptic differential equations, which are called the van Roosbroeck equations [96].

While the idea of the differential equations dates back to 17th century works of Newton and Leibniz, the concept gained much significance in the 18th and 19th centuries. Many physical phenomena was then expressed mathematically as differential equations, e.g. wave equation (d'Alembert, Euler), exponential growth model (Malthus), heat flow (Fourier), Fick's laws of diffusion, Maxwell's equations. However, exact solution methods for differential equations are available only in idealized cases, which cover only fragmentary aspects of real physical settings. Until 20th century there was no general, efficient methods of (approximate) solution of differential problems.

At the turn of 19th century and 20th century, Rayleigh and Ritz proposed to use variational formulation for approximate solutions of the boundary value problem [112]. The idea of utilizing piecewise linear functions on triangulations was introduced by Courant in 1940s. In 1950s the electronic computers become available, bringing a breakthrough and a new meaning of the numerical solution. Then, independently of the theoretical analysis, Finite Element Method (FEM) were used for engineering computations (Turner et al. [114]). On the other hand, in 1960s Finite Element Method was subject to mathematical research. In particular, piecewise polynomial approximations were studied (Birkhoff et al. [13]). In 1974, convergence analysis and error bounds for the mixed Finite Element Methods for saddle problems were introduced by Brezzi [20]. In the early 1980s p-version and hp-version of Finite Element Method were popularized (Babuska et al. [9, 8]). In classic Finite Element Method (h-version), convergence is achieved by refining a mesh, when the mesh elements diameter h goes to zero. In the p-version, the mesh is kept fixed, while the piecewise polynomial degree p is increased. The hp-version combines these two approaches.

In 1970s the Penalty method was developed to impose boundary conditions weakly [7, 35]. At the same time the Discontinuous Galerkin Method (DGM) was introduced by Reed and Hill in 1973 [93], as they discovered it is better suited for approximate solution of the neutron transport equations than continuous Finite Element Method. Error bounds was then demonstrated by Leisant and Raviart in 1974 [66]. More general result for a hyperbolic equation was then introduced by Johnson and Pitkaranta in 1986 [57]. The Interior Penalty method with discontinuous elements was introduced by Arnold [5]. At the turn of 1980s and 1990s the interest in Discontinuous Galerkin Method significantly increased. For example, it has been applied to three-dimensional incompressible steady fluid flows [24], to time-dependent hyperbolic equations [25], to the compressible Navier-Stokes equations [12] or to convection-diffusion systems [29]. In 2002, unified analysis of Discontinuous Galerkin Method for elliptic equations was presented by Arnold et al. [6].

Numerical modelling of semiconductor devices with the drift-diffusion model has been performed since 1964, when Gummel [47] proposed a numerical algorithm for simulations of silicon transistors, which was based on the simple iteration method. Various methods were then used for discretization of the van Roosbroeck equations, for example Finite Difference Method (FDM) [101], Box method [10], Finite Element Method [30]. Special variants of discretizations were developed [75].

The problem, in a broad form, is posed as follows: find u^*, v^*, w^*, such that

$$-\nabla\left(\varepsilon \nabla u^*\right) + e^{u^*-v^*} - e^{w^*-u^*} = k_1, \tag{1}$$

$$-\nabla(\mu_n e^{u^*-v^*} \nabla v^*) - Q(u^*, v^*, w^*)(e^{w^*-v^*} - 1) = 0, \tag{2}$$

$$-\nabla(\mu_p e^{w^*-u^*} \nabla w^*) + Q(u^*, v^*, w^*)(e^{w^*-v^*} - 1) = 0. \tag{3}$$

Equations (1)–(3) are called van Roosbroeck equations (or drift-diffusion equations). Functions $\varepsilon(x), \mu_n(x), \mu_p(x), k_1(x)$ are material parameters and $Q(x, u, v, w)$ is an operator depending on the semiconductor material. Here we would like to emphasize main problems with numerical solution of this system. Equation (1) is called the Poisson equation. It is the most elementary among these equations from the numerical point of view, as it is nonlinear only due to its right hand side. The remaining two equations (2), (3) additionally have nonlinear coefficients, which fluctuate strongly. It makes these equations and also the whole system very difficult to solve. A detailed discussion of the drift-diffusion equations is presented in sections 1.2 and 2.5.

An analysis of the existence of solutions for the drift-diffusion equations is presented in [54]. Homogeneous Neumann and general Dirichlet boundary conditions were considered. The proof of existence bases on Schauder fixed point theorem. A mapping with a stationary point related to the solution of the drift-diffusion system. The discrete case is also considered. The follow-up article [56] treats about properties of the mapping introduced in previous paper [54]. Authors consider both differential and discrete problem, with Finite Element Method discretization. Error bounds are presented, and the convergence is analyzed.

Additionally in [62] it is shown that under constrains on boundary values the mapping mentioned in previous articles is a contraction. Then, due to Banach theorem, the simple iteration algorithm converges to a solution of the van Roosbroeck equations. Regularity properties of the considered functions are discussed in detail.

A comprehensive information about solution of drift-diffusion system is presented in [75]. Several discretizations are considered. Existence of discrete equations are shown. The methods are considered on various meshes in one and two dimensions.

Numerical methods for simulations of semiconductor devices are described in [91]. First problem discussed is choice of unknown variables: carrier concentrations, scaled concentrations, σ-variables, quasi-Fermi levels and Slotboom variables are considered. Methods of solution the nonlinear system, i.e. Gummel method and the Newton method are discussed. Some modifications of the Newton

method are proposed, like scaling of a step and exploiting of symmetric positive definite Jacobian. Discretizations by Finite Different Method and Finite Volume Method are considered. Many examples of numerical simulations in two dimensions are given.

Algorithms for acceleration of simple iteration method are proposed in [61]. Presented ideas are devoted for increasing efficiency of Gummel's iteration. Chebyshev and Richardson accelerations determine coefficients of the three-term linear combination. Both of the methods rely on eigenvalues of Jacobian of considered transformation. On the other hand, Nonlinear Minimal Residual scheme uses similar approach as GMRES iteration and does not require estimates on eigenvalues. However the latter is more complicated for implementation.

Solution of drift-diffusion system formulated in Slotboom variables is considered in [80]. Such choice of variables leads to with three nonlinear elliptic differential equations, similar to ones based on quasi-Fermi levels. However, these variables expose strongly exponential character (see [91]). To avoid overflows and underflows, local scaling technique is used. The scaling is specific to Finite Element Method discretization. Algorithm is presented with the assumption of zero recombination. Simulation examples are also presented.

Neumann-Neumann domain decomposition scheme for decoupled drift-diffusion system is considered in [97]. Authors focus on study p-n diodes under high reverse bias, imposing impact ionization recombination. Choice of variables is discussed. The iteration comprise of the Newton method for Poisson equation and simple iteration for continuity equations. Several numerical examples are given. Also convergence slowdown of Gummel method for high recombinations is indicated.

A different kind of problem is considered in [21]. Authors present generalized Gummel scheme for optimization of a doping profile for a given device. The profile is optimized to attain certain current densities for given bias.

Application of Gummel method with electron/hole mobility dependent on electrostatic field is presented in [23]. Upwind and weighted finite difference schemes in one dimension are considered. Modifications are also proposed for higher dimensions and for Finite Element Method.

Gummel method and coupled Newton method are compared in [64]. Simulation of organic semiconductor devices are considered. Standard drift-diffusion model is enhanced by modified carrier concentration statistic, due to disorder in organic devices. Both constant and non-constant mobilities are considered. Discretization is performed by Finite Volume Method. Carrier concentrations and electrostatic potential are used as unknown functions. For both methods iteration starts from the equilibrium state, for which the initial approximation is available. Then applied voltage is increased gradually, and a result from previous potential step is an initial approximation for the next step. Numerical simulations for constant mobilities reveal that for the Newton method the iteration number for single potential step is approximately five, and is is independent of the bias. On the other hand, in Gummel method the iteration number depends linearly on applied voltage, reaching approximately 200 iterations for 10 V.

The problem we discuss is the discretization of the van Roosbroeck equations. As mentioned, Finite Difference Method and Finite Element Method discretizations are successfully used for this system since the second half of 20th century [101, 92]. However, design of the semiconductor devices has been substantially changed over time. Initially semiconductor transistors or diodes were made from a single material (e.g. silicon) divided into layers with different doping level. In terms of parameters, these conditions were reflected by variations of a possibly discontinuous k_1 function, while $\varepsilon, \mu_n, \mu_p$ were constant. On the contrary, contemporary semiconductor devices, like blue laser diodes (see figure 1), consist of layers of different semiconductor material deposited one on another. Recent designs involves also change of the material through one layer. The material parameters, like $\varepsilon, \mu_n, \mu_p$, are no longer constant, in general they are discontinuous. However, these discontinuities are localized on the layers' interfaces and inside a layer these parameters are constant or, in general, smooth functions.

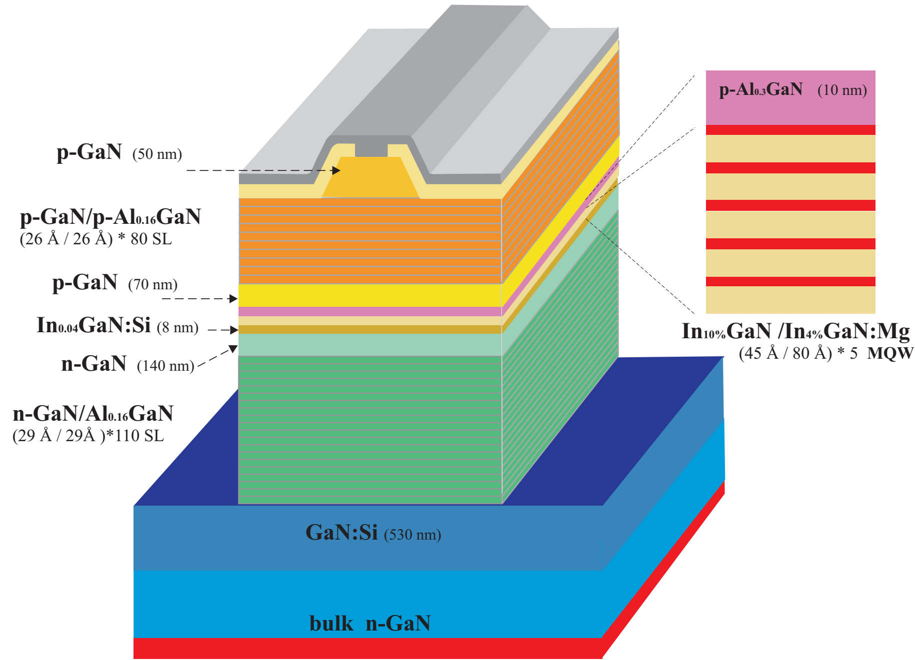

Figure 1: Example of a gallium nitride semiconductor laser structure. Device is made of planar layers of semiconductor materials. Claddings are composed of alternating GaN/AlGaN layers. Active part of the device consists of 5 InGaN quantum wells, separated by barriers with lower indium content. Layers' widths are given in angstroms: $1\,\text{Å} = 0.1\,\text{nm}$.

Thus to obtain a good precision, it would be advantageous to use a discretization which takes into account such localized lack of regularity and which allows to exploit higher regularity inside layers. A natural discretization method for such a problem would be the Discontinuous Galerkin Method [94, 90]. However, this method by its nature imposes much more degrees of freedom in the simulations to allow for discontinuities, leading to slower and more memory-consuming simulations. Since the physical layers of semiconductor devices are of regular shape, it is feasible to use the *Composite Discontinuous Galerkin Method* [37] (CDGM, see section 1.3), which is a hybrid between Continuous and Discontinuous Galerkin Method. It allows to divide the domain into subdomains, on which the standard continuous Finite Element Method is used, and on the interfaces between these subdomains it uses the Interior Penalty method, thus accounting for discontinuities there. This approach allows to greatly reduce number of additional degrees of freedom, as they are only needed on the interfaces. In addition, Composite Discontinuous Galerkin Method does not require conforming grids on the interfaces, thus allowing for nonconforming meshes between subdomains. This is beneficial in simulations of semiconductor devices, as it allows for an easy mesh adaptation in certain layers without affecting others.

The Composite Discontinuous Galerkin Method is currently successfully developed and used for various problems, for example elliptic eigenvalue problems [43], parabolic problems [73], Darcy flow in homogeneous porous media [71]. A FETI–DP-type method for Composite Discontinuous Galerkin Method in two dimensions was proposed in [38].

The Discontinuous Galerkin Method (see section 1.1.2) is closely related to Finite Element Method [19, 28], which provides a framework to numerically approximate solutions of ordinary/partial differential equations. The rough idea of the Finite Element Method is as follows. One start with a differential equation in a weak (variational) form [40], posed in some (infinite dimensional) function space. We call this a *differential problem*. Then a finite dimensional subspace is chosen and the

analogous problem is posed in this subspace. We call it a *discrete problem.* The discrete problem may utilize the same formulation as the differential problem, with infinite dimensional space replaced with the finite-dimensional subspace. The discrete problem is then solved by finding coefficients of unknown function in a given base of the discrete space. If the problem is linear, it may be reduced to a matrix equation. There are many general methods of numerical solution of such equations (see for example [45]), in particular building preconditioners adapted to Finite Element Method problems (cf. [113]). If the problem is nonlinear, some form of linearization must be used, like the Banach iteration or the Newton method [33].

This study consists of two aspects of simulations of semiconductor devices with the van Roosbroeck equations. The mathematical part is devoted to the discretization of the nonlinear Poisson equation (1) in $\mathbb{R}^d$, $d \in \{1, 2\}$. We focus on two versions of the Composite Discontinuous Galerkin Method, based on Symmetric Interior Penalty Galerkin method [94] and on Weakly Over-Penalized Symmetric Interior Penalty method [17]. To our best knowledge, the composite version of the latter method has not been introduced prior to this study.

First we introduce both methods in context of the linear elliptic differential equations. They share common meshes and discrete spaces. In spirit of Composite Discontinuous Galerkin formulation, a polygonal region Ω is divided into sub-polygons Ω_i. We use mesh which is matching in every Ω_i and nonmatching across interfaces $\partial\Omega_i \cap \partial\Omega_j$. Therefore functions of the discrete spaces are continuous in subdomains Ω_i, and discontinuous on the interfaces.

It is assumed that each Ω_i is related to a physical layer of uniform semiconductor material, such that the interfaces between physical layers are located on interfaces $\partial\Omega_i \cap \partial\Omega_j$. This assumption implies that coefficients of the van Roosbroeck equations are also discontinuous on the interfaces.

This approach is beneficial due to the following reasons. First, the physical division of a given device is determined by its geometry and thus it is known *a priori.* It is therefore natural to use this knowledge and to pick Ω_i appropriately, to confine discontinuities of the elliptic coefficients to interfaces $\partial\Omega_i \cap \partial\Omega_j$. Moreover, the physical purpose of each layer provides a hint on whether a mesh on this layer should be thick or thin. Since the mesh is nonmatching on interfaces, it may be chosen independently for each Ω_i as needed. In addition, the composite Discontinuous Galerkin discretization is superior to the standard Discontinuous Galerkin Method in terms of memory footprint, as multiple values must be stored only on the interfaces, not for the whole mesh.

The Composite Discontinuous Galerkin Method is characterized by presence of interfacial and boundary terms in the weak formulation, which are absent in the original elliptic problem. These terms emerge as the consequence of the Green formula, which cannot be used on Ω due to insufficient regularity of discrete space functions. On the other hand, it may be applied to each Ω_i, leading to these interfacial terms which do not cancel out. These terms can later be symmetrized to make use of robust algorithms for solving symmetric linear systems. Moreover, artificial penalty terms are introduced, to control discontinuities on the interfaces not present in the differential solutions.

In the Weakly Over-Penalized Symmetric Interior Penalty method, terms due to Green formula are absent for a price of higher penalty coefficients. This way the method is easier to analyze and to implement, but it has worse numerical properties.

The first problem we must face is that solutions of differential problem are in Sobolev space $H^1(\Omega)$, while discrete solutions are not. To estimate approximation error norm, we would like to embed discrete solutions and differential solutions in some more general space. Then we pass to the discrete formulation of the nonlinear Poisson equation. We adapt already established linear discrete problems to this particular nonlinear case and we prove that the resulting problems are well-posed.

Before proceeding to the error estimates, we discuss the interpolation error measured in broken norm. Generally for the norm used in composite version of Symmetric Interior Penalty Galerkin method the result is well known, but we are not aware of similar result for the Composite Weakly

Over-Penalized Interior Penalty method. We prove both estimates, as the proofs share their main part. Unfortunately the estimate for the latter method is not optimal, mostly due to increased penalty coefficients. For the original Weakly Over-Penalized Interior Penalty method, this problem was remedied by specific choice of the interpolant. This solution however cannot be generalized to nonmatching meshes.

At this moment we would have the sufficient instruments for establishing of the error estimates of discrete solutions for both considered methods. The proof of these estimates is based on the fact that broken norms related to respective discretizations is strongly bounded with the formulations of the discrete problems.

In addition to the theory, we developed a numerical software for solving the van Roosbroeck equations. This software uses the discretization methods discussed in this thesis and it allows to perform a convergence study of these discretizations as well as to perform simulations of the semiconductor devices. This application operates both in equilibrium state and in non-equilibrium state, thus it can numerically solve full van Roosbroeck system.

Our software, called `pmicro`, has two variants. Let us first discuss the one-dimensional variant. This application is written in mixed Octave/Matlab and C/C++ languages, with most computationally intensive parts written in C/C++, including the discretization. The logic is written in Octave/Matlab. The aim of such approach is to achieve a satisfactory trade-off between efficiency and code complexity. Low complexity provides ease of modification of the application, allowing alteration of the underlying mathematical/physical model. This variant is our main tool in simulations of gallium nitride luminescent devices. These devices have two contacts, located on the opposite sides of the structure, so one-dimensional model provides satisfactory results in many cases. In this application we use the van Roosbroeck equations accounting for various physical parameters of the semiconductor material (discussed in detail in section 2.5). During our study, we consequently expanded the underlying model by introducing additional physical phenomena: radiative and nonradiative recombination, trap-assisted tunneling, polarization effect and optical generation. Besides of IV characteristics (current versus voltage), which can be easily derived from the simulated electrical properties, we also estimate the light emission using radiative recombination rate, which allows us to calculate light-output, estimate optical power and losses.

Second variant of our software implements two-dimensional model. We use this application mainly in convergence study of discretizations discussed in this thesis. The code is written mostly in C/C++ and it uses PETSc library for numerical linear algebra. Unfortunately the Composite Discontinuous Galerkin Method discretizations are not widely deployed in popular Finite Element Method libraries, so a library for such discretization is implemented by us. The two-dimensional code accounts for simplified van Roosbroeck equations used in analysis of the methods as well as for full drift-diffusion equations used in physical simulations. However, the latter model does not cover full range of phenomena implemented in one-dimensional model.

During our early simulations we faced a problem of linearization of the discretized nonlinear van Roosbroeck equations. Our first choice was the Banach/Picard iteration, where the original system was substituted by decoupled linear equations. Unfortunately this method was unsatisfactory, as the number of iterations was highly dependent on device/material parameters, varying from hundreds to millions. We also tried the Gummel method, which is similar, but it consists of decoupled nonlinear equations. This method, well tested for silicon devices in 1980s, turned out to be unsatisfactory for many devices simulated by us due to strong coupling of the subsequent equations in case of gallium nitride based luminescent devices, mostly related to high recombination rate and doping. We finally turned to the Newton method for a coupled system, which give a satisfactory results, in particular the iteration number which do not vary much with physical parameters. However, we use also a variant of backtracking method. In backtracking, as well as in stopping conditions, we make use of the Banach

iteration. This approach in many cases improves the number of iteration and more importantly it often prevents divergence of the Newton method.

The second part of this research consists of simulation of the semiconductor devices based on gallium nitride. For these simulations we used one-dimensional variant of our software, as mentioned before. We start our simulations with p-n homojunctions. These simple structures are the starting points in design of complex luminescent devices like light-emitting diodes or laser diodes. These initial simulations are supposed to create basic intuition related to GaN p-n diodes as well as to fine-tune our software, as the development of `pmicro` was performed gradually, along with simulations. The gallium nitride structures are characterized by strong doping. Especially the magnesium doping in p-type region may reach concentrations $1 \times 10^{19}\,\mathrm{cm}^{-3}$ or more. Numerically this is a problem as it increases coupling between drift-diffusion equations. On the other hand, GaN is a wide bandgap semiconductor, and therefore the concentration of minority carriers in the equilibrium state is very low. This leads to vast difference in concentration of carriers, electrons or holes, between n-type region and p-type region. It poses another numerical difficulty, as then coefficients of the continuity equations vary by several orders of magnitude. Therefore simulations of such elementary gallium nitride semiconductor devices has already introduced severe numerical problems, which could only get worse with device's complexity.

Since the drift-diffusion model accounts for the electrical properties, our initial simulations focus on I-V diagrams. For gallium nitride p-n homojunctions, we observe three cases: reverse bias, when the current is more or less constant, low forward bias, when current is exponential in bias and high forward bias, when this dependence becomes linear. This situation is more complicated when more layers are introduced, like an undoped layer or quantum well between n-type region and p-type region, and also if we take into account nonradiative recombination mechanisms, like recombination on trap levels (Shockley-Read-Hall recombination, SRH) or Auger recombination. For a given structure, a dominating recombination channel may vary with the bias, altering I-V characteristics. We take a closer look on the SRH recombination. It has two aspects: contribution to the total recombination and additional electrostatic charge due to trap levels. We find the latter effect negligible. On the contrary, the SRH recombination rate may be significant. This mechanism of recombination is often dominating in the low bias regime, especially if we take into account the trap-assisted tunneling. In GaN diodes, where doping is significant and the depleted layer is narrow, there are high electrostatic field involved for low biases, which increases ratio of the carrier tunneling to the trap level. Our simulations demonstrates better agreement between theory and experiment if the trap-assisted tunneling is taken into account, especially for low/moderate forward biases.

Then we pass to simulations of light-emitting diodes (LEDs) and laser diodes (LDs). In this scope two additional aspects are important. First is the polarization effect, which is present in the gallium nitride based devices. A modification of the underlying discretization was necessary to include the polarization charges on interfaces between different material (discussed in detail in section 3.3). On the other hand, in case of LEDs and LDs we have to estimate the light-output, efficiency and energy loss. The light-output is generally proportional to the radiative recombination rate in the quantum wells. This value is predicted by the drift-diffusion model. Other channels of recombination, i.e. nonradiative recombination or even radiative recombination outside of quantum wells contributes to energy loss.

We start with a problem of aluminum content in the electron blocking layer (EBL). EBL is usually placed between the active region (quantum wells and barriers) and a p-type region. It is a barrier, which is supposed to prevent electrons form escaping the active region. Unfortunately this layer also blocks holes to some extent, thus making it too high unfavorable. In simulations we show that the efficiency of a devices is increasing with aluminum content only to certain value, then it remains constant. However, increasing the aluminum content further leads to higher resistance of the device

and smaller light-output. Then we pass to the magnesium doping level in p-type region. It is shown that insufficient doping results in high resistance due to unscreened polarization charges. We also analyze dependence of number of quantum wells on the resistance of a device.

Van Roosbroeck equations can also account for optical generation of carriers. We present such simulations in context of quaternary AlInGaN alloy.

Chapter 1

Discretization of van Roosbroeck equations

Contents

1.1 Finite Element Method basics **23**
1.1.1 Continuous Finite Element Method 23
1.1.2 Discontinuous Galerkin Method 29
1.1.3 Weakly Over-Penalized Symmetric Interior Penalty 31
1.2 Differential problem **32**
1.2.1 Drift-diffusion system 32
1.2.2 Equilibrium state 33
1.3 Composite Discontinuous Galerkin Method **34**
1.3.1 Discrete space 34
1.3.2 Composite Discontinuous Galerkin variants 36
1.3.3 Broken norm and the Poincare inequality 39
1.3.4 Consistency 41
1.4 Discretization of the equilibrium case **47**
1.4.1 Composite Weakly Over-Penalized Symmetric Interior Penalty (CWOPSIP) . . 47
1.4.2 Composite Symmetric Interior Penalty Galerkin (CSIPG) 48
1.4.3 Existence and uniqueness 48
1.5 Interpolation operator and interpolation error **52**
1.5.1 One dimension 53
1.5.2 Two dimensions 53
1.6 Error estimates for the equilibrium case for CWOPSIP **56**
1.6.1 Outline of the proof 56
1.6.2 Consistency 57
1.6.3 Analysis 58
1.6.4 Summary 61
1.7 Error estimates for the equilibrium case for CSIPG **63**
1.7.1 Consistency 63
1.7.2 Analysis 64
1.7.3 Summary 67

In this part we will derive two variants of the CDGM discretizations. First one bases on Symmetric Interior Penalty Galerkin (SIPG) method [19]. This method was introduced in [37] for the linear elliptic equations. Besides of features mentioned previously, the discrete form for this method is symmetric, what is advantageous when solving linear equations. Our equations are not linear, but still this feature will be useful in context of the linearization method we use.

Second discretization we propose is similar, but it bases on Weakly Over-Penalized Symmetric Interior Penalty (WOPSIP) [17]. This method is also symmetric, and it has very simple form, which is helpful in analysis and implementation. On the other hand, due to higher penalty parameter it may perform worse in numerical simulations.

We focus our theoretical analysis of CDGM variants on the equilibrium state solutions of the van Roosbroeck equations in $\mathbb{R}$ and $\mathbb{R}^2$. The equilibrium state is a simplified case, where only the first equation is to be solved. It corresponds to the physical state when there is no energy transfer between a device and the environment, in particular the device is disconnected from the power source. We limit our analysis to this case, as the proof framework used in this paper, which is borrowed from the DGM analysis of the Navier-Stokes problem [115], imposes the uniqueness of the solution, which is not guaranteed in the non-equilibrium state.

In our analysis, we focus on linear $\mathbb{P}_1$ element. This choice is made to simplify the analysis and implementation. While there are many computer libraries and frameworks for FEM and DGM discretizations, none that we are aware of supports CDGM out of the box. In particular, it is not possible to define separate meshes across subdomains. Therefore we develop our own framework, which currently support only linear $\mathbb{P}_1$ element.

The main problem with the van Roosbroeck system is the nonlinearity. Depending on a device composition and design, the coefficients of the latter two equations may vary by several orders of magnitude. There are various approaches for solving this system. They may involve decoupling, Banach iteration [76,], Newton method [61], etc. An important step is the choice of unknown variables. According to [92], there is a variety of possible unknown function sets for the van Roosbroeck equations. Each set provides certain balance between nonlinearity of the equations and exponential character of the unknown functions. It may seem like the choice of so-called Slotboom variables is preferable, as they provide least degree of nonlinearity. However, it also imposes enormous variations of the unknown functions (like $[1, 10^{84}]$, see [92]). We will therefore use the quasi-Fermi level formulation, which makes the equations highly nonlinear, but unknown functions do not express the exponential character.

The van Roosbroeck equations used in simulations of realistic devices depend on many material parameters, generation/recombination coefficients, technical parameters, etc. In the theoretical study, we would like to focus on the mathematical aspect of the Discontinuous Galerkin Method discretization of the underlying equations. Since there is no obvious generalization, which covers every possible case, for analysis we will focus on standard version widely used in the literature [54, 30, 61, 76]. The problem will be formulated in section 1.2.

The van Roosbroeck equations are elliptic, thus in section 1.3.2.1 we present standard weak problem in this case. This general weak problem is a basis of the Composite Discontinuous Galerkin Method (CDGM) discretizations [37], presented in section 1.3.

By a coarse grid we call the partition $\{\Omega_i\}_{i=1}^N$ of the domain Ω. Then we consider families of FEM discretizations $X_{h_i}(\Omega_i)$, parameterized with $h_i \to 0$. The CDGM discrete space $X_h(\Omega)$ is composed of $X_{h_i}(\Omega_i)$, with $h := \max h_i \to 0$. Note that the coarse partition does not change with h. In section 1.3.2 we present the discrete operators, which comprise of the part corresponding to the variational formulation inside subdomains Ω_i and penalty terms on the interfaces and on $\partial\Omega_i$.

After introduction of discretizations for linear equations, in section 1.4 we use introduced spaces and operators to discretize van Roosbroeck equation for equilibrium state. We show the existence and uniqueness of discrete solutions. Then in section 1.5 we introduce the interpolation operator and

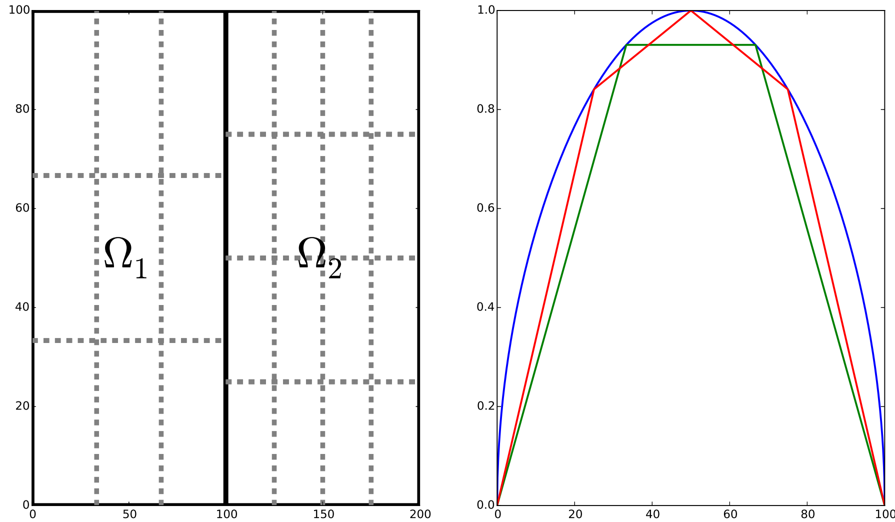

Figure 1.1: Example of a discontinuous interpolant on the interface with nonconforming grids. Blue — function to be interpolated on the interface $\partial\Omega_1 \cap \partial\Omega_2$, green — interpolation in $X_{h_1}(\Omega_1)$, red — interpolation in $X_{h_2}(\Omega_2)$.

we discuss interpolation errors. There is an important difference between $\mathbb{R}$ and $\mathbb{R}^2$ domain in this aspect. Despite using the DGM, in one dimension we can easily define an interpolation operator so that interpolants of continuous functions are also continuous. This is not the case in two dimensions in general, due to nonconformity of grids on interfaces (see figure 1.1).

In sections 1.6, 1.7 we present the error estimates for the proposed discretizations. To prove the error estimates for the discretization proposed by us, we would like to use similar approach to presented in [115] for the Navier-Stokes equation.

In addition to the analysis presented for the equilibrium state regime, we also discuss existence of the discrete solutions for the general case (section 4.C). Unfortunately these results are limited to one dimensional CWOPSIP discretization, due to the maximum principles discussed there, which are not feasible for CSIPG nor for two-dimensional CWOPSIP.

1.1 Finite Element Method basics

We start with an introduction to the discretization of differential equations with Finite Element Method. It is a foundation for the Composite Discrete Galerkin Method, which we introduce in section 1.3. In this introduction we present only selected results, which are relevant to the scope of this thesis.

1.1.1 Continuous Finite Element Method

The Finite Element Method [19, 28] provides a formalism for finding approximate solutions of the ordinary/partial differential equations in a finite dimensional space. It is a particular case of the Ritz-Galerkin methods. The basic idea of these methods is to substitute an infinite-dimensional functional domain of a given differential problem with a finite-dimensional discrete spaces and to limit the initial problem to these spaces. Upon choosing some basis of a given discrete space, the discrete problem becomes a system of algebraic equations. The characteristic feature of the continuous Finite Element Method is particular choice of the basis of the discrete space.

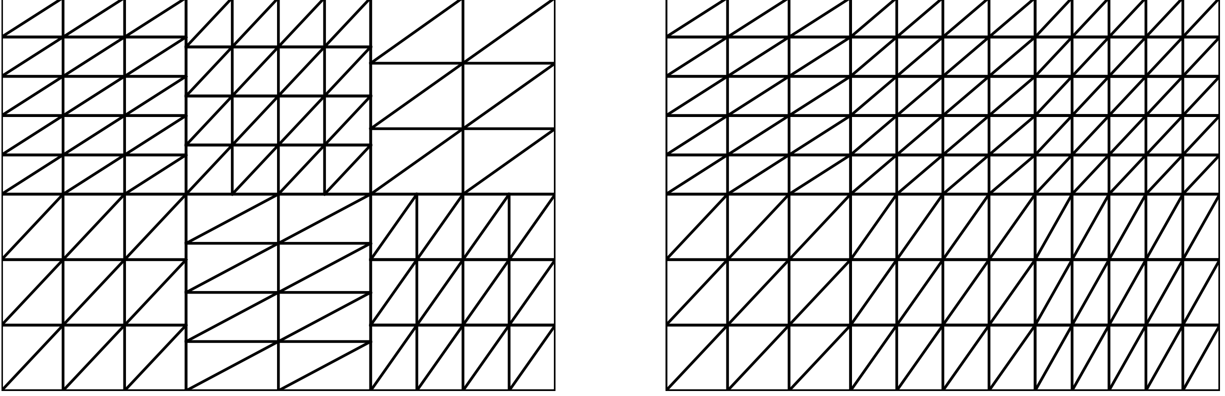

Figure 1.2: Examples of meshes composed of triangular elements: a nonconforming mesh (left) and a conforming mesh (right).

In this introduction, we focus on the standard continuous Finite Element Method, with a discrete space consisting of piecewise-linear continuous functions. We start with a definition of a *mesh*.

Definition 1.1.1. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2\}$ be a polygonal domain, and let $\mathcal{T}_h$ be a division of Ω into polygons, where h is the maximum element diameter. We say that $\mathcal{T}_h$ is a* mesh *if the intersection of any two elements of $\mathcal{T}_h$ is either empty or it is a set of measure zero.*

Definition 1.1.2. *Let $\mathcal{T}_h$ be a mesh. If intersection of any two elements of $\mathcal{T}_h$ is either empty, a vertex or an edge, then the mesh $\mathcal{T}_h$ is called a* conforming mesh. *Otherwise it is called a* nonconforming mesh.

Note that due to definition of the conforming mesh, if x is a vertex of any $\tau_1 \in \mathcal{T}_h$, and if $x \in \partial\tau_2$ for some $\tau_2 \in \mathcal{T}_h$, then x is also a vertex of τ_2. An example of a conforming/nonconforming mesh is demonstrated in figure 1.2.

Definition 1.1.3. *We say that $\mathcal{T}_h$ is a* triangulation of $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2\}$ *if:*

- *$\mathcal{T}_h$ is a conforming mesh.*
- *Every element $\tau \in \mathcal{T}_h$ is a triangle (if $d = 2$) or an interval (if $d = 1$).*

Definition 1.1.4. *Let $\mathcal{T}_h$ be a triangulation of $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2\}$. We define nodes of the mesh $\mathcal{T}_h$ as*

$$\mathcal{N}_h = \{x \in \Omega : x \text{ is a vertex of } \tau \text{ for any } \tau \in \mathcal{T}_h\}. \tag{1.1.1}$$

For a systematic study of error of the discrete approximations, we have to introduce some parameter characterizing a triangulation, and the use this parameter in estimates. A standard choice is the maximal diameter of mesh elements, denoted by h.

The intuition behind the error estimate is as follows. We have some domain Ω and a family of triangulations $\{\mathcal{T}_h\}$. We would like to bound error depending on the parameter h, and we are generally interested about asymptotics of this bound when $h \to 0$. While it is not necessary to have a mesh for every $h > 0$ in this family $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$, we would like 0 to be an accumulation point of $\mathcal{H}$, so that the limit $h \to 0$ is related to some sequence of meshes $\mathcal{T}_h$ as h approaches zero.

To take h as a reasonable characteristic of a mesh, we have to impose additional constraints. Otherwise some degenerate cases are possible, which makes it difficult to obtain useful results.

Definition 1.1.5. *Let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ be a family of meshes indexed by $h \in \mathcal{H} \subset (0,1]$, where the set $\mathcal{H}$ is a countable subset of positive real numbers having 0 as only accumulation point. For any element $\tau \in \mathcal{T}_h$, let h_τ denote the diameter of τ and let ρ_τ denote the diameter of an inscribed circle of τ.*

We say that $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ is a shape regular family of meshes, *or simply that $\mathcal{T}_h$ is a* shape regular mesh, *if there is a constant $\rho_R > 0$ such that*

$$\forall h \in \mathcal{H}\ \forall \tau \in \mathcal{T}_h \quad \frac{h_\tau}{\rho_\tau} \leq \rho_R. \tag{1.1.2}$$

Definition 1.1.6. *Let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ be a family of meshes indexed by $h \in \mathcal{H} \subset (0,1]$, where the set $\mathcal{H}$ is a countable subset of positive real numbers having 0 as only accumulation point.*

We say that $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ is a quasi-uniform family of meshes, *or simply that $\mathcal{T}_h$ is a* quasi-uniform mesh, *if there is a constant $\rho_Q > 0$ such that*

$$\forall h \in \mathcal{H}\ \forall \tau \in \mathcal{T}_h \quad \rho_\tau \geq \rho_Q h. \tag{1.1.3}$$

Conditions introduced in definitions 1.1.5, 1.1.6 assert that shapes of the mesh elements change in a systematic manner as $h \to 0$. Let us provide some examples in $\mathbb{R}^2$. In this case, shape regularity of a mesh family assures that area $|\tau|$ of any mesh element $\tau \in T_h$ is proportional to h_τ^2. Such meshes do not have "degenerate" elements, which shrink in one dimension much faster than in other dimensions. On the other hand, quasi-uniformity indicates that mesh elements cannot be arbitrarily small (in diameter) within a given h. Thus it ensures that all mesh elements have diameters proportional to h and area proportional to h^2, so they shrink rather uniformly with h. These estimates are useful in establishing interpolation error bounds or discrete solution error bounds. On the other hand, they fit into a natural concept of a mesh refinement, thus they do not pose significant problems in applications.

Now we pass to the definition of standard continuous FEM space of piecewise-linear functions.

Definition 1.1.7. *Let $\mathcal{T}_h$ be a triangulation of some polygon (or interval) Ω. We define a* continuous linear finite element space $V_h(\Omega)$ *on the triangulation $\mathcal{T}_h$ as*

$$V_h(\Omega) := \big\{ v_h \in \mathcal{C}(\overline{\Omega}) : v_h|_\tau \in \mathbb{P}_1(\tau) \quad \forall \tau \in \mathcal{T}_h \big\}. \tag{1.1.4}$$

Assume that $\{x_1, x_2, \ldots, x_J\} = \mathcal{N}_h$ are the nodes of the triangulation $\mathcal{T}_h$. Then we define a nodal base $\{\varphi_{(1)}, \varphi_{(2)}, \ldots, \varphi_{(J)}\}$ *of space $V_h(\Omega)$ as a functions which satisfy*

$$\varphi_{(j)}(x_k) := \begin{cases} 1, & \text{if } j = k, \\ 0, & \text{if } j \neq k. \end{cases} \tag{1.1.5}$$

Examples of nodal basis elements are presented on figures 1.3, 1.4.

Definition 1.1.8. *For a continuous linear finite element space $V_h(\Omega)$, we define the* interpolation operator $I_h : \mathcal{C}^0(\overline{\Omega}) \to V_h(\Omega)$ *by the following relationship: for any $u \in \mathcal{C}^0(\overline{\Omega})$ element $I_h u$ is defined as*

$$I_h u := \sum_{j=1}^{J} u(x_j)\varphi_{(j)}, \tag{1.1.6}$$

where $\{x_j\}$ are the nodes of the triangulation $\mathcal{T}_h$ and $\{\varphi_{(j)}\}$ is the nodal basis of $V_h(\Omega)$.

Note that by means of this definition, $u(x_j) = I_h u(x_j)$ for all nodal points x_j. For convenience, we will denote

$$u_I := I_h u. \tag{1.1.7}$$

We would like to define the notion of affine-equivalence of two sets.

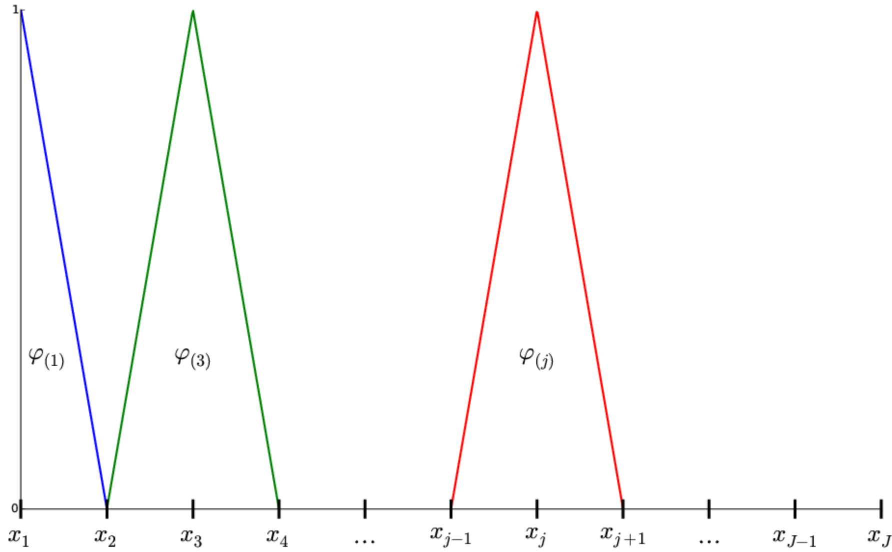

Figure 1.3: Example of a nodal basis of the one-dimensional linear continuous finite element space.

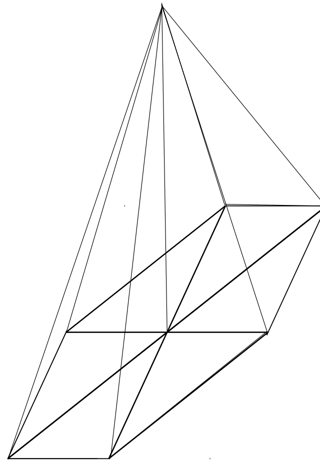

Figure 1.4: Example of a nodal basis element of the two-dimensional linear continuous finite element space.

Definition 1.1.9. *We say that two sets $\tau, \hat{\tau} \subset \mathbb{R}^d$ are* affine-equivalent *if there is an affine invertible mapping $F : \mathbb{R}^d \to \mathbb{R}^d$ such that $F(\tau) = \hat{\tau}$.*

The following result allows for establishing useful trace theorems for finite element spaces

Theorem 1.1.10. *Let $\tau, \hat{\tau} \in \mathbb{R}^d$ be affine-equivalent sets and let $F(x) = Ax + a$ be the affine invertible mapping between them.*

Then there exists some constant $C = C(s, d)$ such that for any $u \in H^s(\tau)$, $s \geq 0$, $s \in \mathbb{N}$ function $\hat{u} := u \circ F$ belongs to $H^s(\hat{\tau})$ and

$$|\hat{u}|^2_{H^s(\hat{\tau})} \leq C\|A^{-1}\|_2^{2s}|\det(A)||u|^2_{H^s(\tau)}, \tag{1.1.8}$$

$$|u|^2_{H^s(\tau)} \leq C\|A\|_2^{2s}|\det(A)|^{-1}|\hat{u}|^2_{H^s(\hat{\tau})}. \tag{1.1.9}$$

Proof. This standard result is a special case of theorem 3.1.2 of [28]. □

Establishing estimates on functions related to meshes and elements of these meshes is difficult in general case, as these elements change with h it is hard to track the dependence of various constants on their diameter and shape. The situation is much different if the mesh family has an affine-equivalent reference element, for example when it is a triangulation. Many useful properties of such finite element spaces may be demonstrated by transition to this reference element. This technique may be used to prove the following standard result [19].

Proposition 1.1.11. *Let $\{V_h\}_h$ be a family of continuous linear finite element spaces with triangulations $\mathcal{T}_h$ on a polygonal domain Ω. Let $\tau \in \mathcal{T}_h$ be some element of the triangulation and let e_τ be some edge of τ. Then for any function $u \in H^2(\tau)$ the following estimates hold*

$$\begin{aligned} \|u\|^2_{L_2(e_\tau)} &\leq C|e_\tau||\tau|^{-1}\Big(\|u\|^2_{L_2(\tau)} + h_\tau^2|u|^2_{H^1(\tau)}\Big), \\ \|\nabla u \cdot \nu\|^2_{L_2(e_\tau)} &\leq C|e_\tau||\tau|^{-1}\Big(|u|^2_{H^1(\tau)} + h_\tau^2|u|^2_{H^2(\tau)}\Big). \end{aligned} \tag{1.1.10}$$

If additionally $\mathcal{T}_h$ is quasi-uniform, then these estimates may be refined to

$$\begin{aligned} \|u\|^2_{L_2(e_\tau)} &\leq Ch^{-1}\Big(\|u\|^2_{L_2(\tau)} + h^2|u|^2_{H^1(\tau)}\Big), \\ \|\nabla u \cdot \nu\|^2_{L_2(e_\tau)} &\leq Ch^{-1}\Big(|u|^2_{H^1(\tau)} + h^2|u|^2_{H^2(\tau)}\Big). \end{aligned} \tag{1.1.11}$$

Note that in one-dimensional case, where e_τ is a point, we define $|e_\tau|$ to be equal to one.

Proof. We prove estimates on $\|u\|_{L_2(e_\tau)}$, proof of estimates on $\|\nabla u \cdot \nu\|_{L_2(e_\tau)}$ is analogous.

Let $\hat{\tau}$ be some given reference element, an interval($d = 1$) or a triangle ($d = 2$), independent of h. Provided that $\hat{\tau}$ is not degenerated, any given τ of the triangulation $\mathcal{T}_h$ is affine-equivalent to $\hat{\tau}$. Similarly any edge e_τ is affine-equivalent to $\hat{e}$, which is some edge of $\hat{\tau}$. We therefore use theorem 1.1.10 with $s = 0$ on e_τ. Taking into account that the determinant of the linear part would be equal to $|\hat{e}|/|e_\tau|$ up to a sign, we obtain

$$\|u\|^2_{L_2(e_\tau)} \leq C|e_\tau|\|\hat{u}\|^2_{L_2(\hat{e})}. \tag{1.1.12}$$

Coefficients related to $|\hat{e}|$ are included in constant C, as there is a single reference element and they do not depend on h. Then using the trace theorem on $\hat{\tau}$ we get

$$C|e_\tau|\|\hat{u}\|^2_{L_2(\hat{e})} \leq C|e_\tau|\|\hat{u}\|^2_{H^1(\hat{\tau})} = C|e_\tau|\Big(\|\hat{u}\|^2_{L_2(\hat{\tau})} + \|\nabla\hat{u}\|^2_{L_2(\hat{\tau})}\Big). \tag{1.1.13}$$

Again using theorem 1.1.10 on $\hat{\tau}$ with $s = 0$ and $s = 1$ we obtain

$$\begin{aligned} \|\hat{u}\|^2_{L_2(\hat{\tau})} &\leq C|\tau|^{-1}\|u\|^2_{L_2(\tau)}, \\ \|\nabla u\|^2_{L_2(\hat{\tau})} &\leq C|\tau|^{-1}h_\tau^2\|\nabla u\|^2_{L_2(\tau)}. \end{aligned} \tag{1.1.14}$$

Similarly as before, absolute value of determinant of the linear part is $|\hat{\tau}|/|\tau|$ and $1/|\hat{\tau}|$ is included in the constant. Moreover we may estimate $\|A^{-1}\|_2 \leq Ch_\tau$. Using these inequalities, we obtain first estimate:

$$\|u\|^2_{L_2(e_\tau)} \leq C|e_\tau||\tau|^{-1}\Big(\|u\|^2_{L_2(\tau)} + h_\tau^2\|\nabla u\|^2_{L_2(\tau)}\Big). \tag{1.1.15}$$

If additionally $\mathcal{T}_h$ is quasi-uniform, then $\rho_\tau > \rho_Q h$ and $|\tau|^{-1} \leq Ch^{-2}$. Naturally $|e_\tau| \leq h_\tau \leq h$ and the estimate reads

$$\|u\|^2_{L_2(e_\tau)} \leq Chh^{-2}\Big(\|u\|^2_{L_2(\tau)} + h^2\|\nabla u\|^2_{L_2(\tau)}\Big) = Ch^{-1}\Big(\|u\|^2_{L_2(\tau)} + h^2\|\nabla u\|^2_{L_2(\tau)}\Big). \tag{1.1.16}$$

□

If the function u in the proposition above is in the finite element space, this result may be improved due to equivalence of norms in finite-dimensional spaces used for the reference element space.

Corollary 1.1.12. *Let $\{(V_h, \mathcal{T}_h)\}_h$ be a family of continuous linear finite element spaces on a polygonal domain Ω. Let $\tau \in \mathcal{T}_h$ be some element of the triangulation and let e_τ be some edge of τ. Then for any function $u_h \in V_h$ the following estimates hold*

$$\begin{aligned} \|u_h\|^2_{L_2(e_\tau)} &\leq C|e_\tau||\tau|^{-1}\|u_h\|^2_{L_2(\tau)}, \\ \|\nabla u_h \cdot \nu\|^2_{L_2(e_\tau)} &\leq C|e_\tau||\tau|^{-1}|u_h|^2_{H^1(\tau)}. \end{aligned} \tag{1.1.17}$$

If additionally $\mathcal{T}_h$ is quasi-uniform, then these estimates may be refined to

$$\begin{aligned} \|u_h\|_{L_2(e_\tau)} &\leq Ch^{-1}\|u_h\|^2_{L_2(\tau)}, \\ \|\nabla u_h \cdot \nu\|_{L_2(e_\tau)} &\leq Ch^{-1}|u_h|^2_{H^1(\tau)}. \end{aligned} \tag{1.1.18}$$

In our analysis, we use the following standard corollary of a general interpolation estimates in finite element spaces (see for example [28], theorem 3.2.1)

Corollary 1.1.13. *Let $u \in H^2(\Omega)$ be a given function and let $V_h(\Omega)$ be a continuous linear finite element space of on a polygonal domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2\}$. Then the following estimates on the interpolation error in $V_h(\Omega)$ hold*

$$\begin{aligned} \|u - I_h u\|_{L_2(\Omega)} &\leq Ch^2|u|_{H^2(\Omega)}, \\ |u - I_h u|_{H^1(\Omega)} &\leq Ch|u|_{H^2(\Omega)}. \end{aligned} \tag{1.1.19}$$

Example We would like to present an example. We show an application of the continuous Finite Element Method to the Poisson equation on a rectangle. This is a reference for problems we introduce further. We would like to show the discretization and present the error estimate. This is a standard example, more information may be found in [19, 28].

Assume that $\Omega \subset \mathbb{R}^2$ is a rectangle. Consider Poisson's equation

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{1.1.20}$$

To establish a discrete problem, we need the weak formulation first. It is defined as follows:

Problem 1.1.14. *Let $\Omega \subset \mathbb{R}^2$ be a rectangle and let $f \in L_2(\Omega)$ be given. Find $u^* \in H^1_0(\Omega)$ such that for every $\phi \in H^1_0(\Omega)$*

$$\int_\Omega \nabla u^* \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx, \tag{1.1.21}$$

where

$$\begin{aligned} H^1_0(\Omega) &:= \textit{closure of } C^\infty_0(\overline{\Omega}) \textit{ in } H^1(\Omega), \\ C^\infty_0(\overline{\Omega}) &:= \{f \in C^\infty(\overline{\Omega}) : f|_{\partial\Omega} \equiv 0\}. \end{aligned} \tag{1.1.22}$$

Note that in the weak formulation the Dirichlet boundary condition is encapsulated in the definition of the space $H^1_0(\Omega)$. Thus it is called essential boundary condition. Then let us define a finite element space suitable for this problem. We define

$$V_{h,0}(\Omega) := \{v_h \in V_h(\Omega) : v_h|_{\partial\Omega} \equiv 0\}. \tag{1.1.23}$$

We establish the following discrete problem for the Poisson equation.

Problem 1.1.15. *Let $f \in L_2(\Omega)$ be given. Find $u_h^* \in V_{h,0}(\Omega)$ such that for every $\phi_h \in V_{h,0}(\Omega)$*

$$\int_\Omega \nabla u_h^* \cdot \nabla \phi_h \, dx = \int_\Omega f \phi_h \, dx. \tag{1.1.24}$$

The following approximation error estimate may be derived [19, 28].

Proposition 1.1.16. *Assume that V_h is a continuous linear finite element space. Assume that solution u^* of problem 1.1.14 is in $H^2(\Omega)$. Then the following estimate holds*

$$\|u^* - u_h^*\|_{H^1(\Omega)}^2 \leq Ch^2 |u|_{H^2(\Omega)}^2. \tag{1.1.25}$$

This result compared with corollary 1.1.13 shows that if the solution u^* is sufficiently regular, the estimate on approximation error of the discrete solution is similar to estimate on the interpolation error.

1.1.2 Discontinuous Galerkin Method

The continuous Finite Element Method has the following key features:

- Discrete spaces $V_{h,0}$ are subspaces of the differential problem space.
- Discrete variational problem is the restriction of a differential variational problem to a subspace, forms remain the same.

The Discontinuous Galerkin Method [94, 90] provides discretizations which is similar to continuous FEM, but they do not have these features. In particular, they may be discontinuous on interfaces between triangulation elements.

Continuing our example of the Poisson equation discretization with piecewise-linear functions, we define the linear classic DGM space by

$$X_h(\Omega) := \{u_h \in L_2(\Omega) : \forall \tau \in \mathcal{T}_h \quad u_h|_\tau \in \mathbb{P}_1(\tau)\}. \tag{1.1.26}$$

By definition $X_h(\Omega) \subset L_2(\Omega)$, but $X_h(\Omega) \not\subset H^1(\Omega)$. While functions of $X_h(\Omega)$ are smooth on the mesh elements, in general they are discontinuous. Thus, while $X_h(\Omega)$ is naturally much "smaller" than $H^1(\Omega)$, it contains a class of functions discontinuous on the triangulation edges, which is absent in $H^1(\Omega)$ as they do not have weak derivatives on Ω.

It is therefore not feasible to simply use this discrete space instead of the continuous space to obtain a discrete problem from a differential problem. It is necessary to impose some additional constrains.

The nodal basis of $X_h(\Omega)$ is associated with the basis of continuous FEM space (figure 1.3, 1.4). For any function in this basis, a restriction of this function to any mesh element contained in its support is a base function of $X_h(\Omega)$ (cf. figure 1.5). Due to this property the number of basis elements of the DGM space is higher in comparison to the continuous FEM space defined on the same mesh. Naturally, functions of $X_h(\Omega)$ have multiple candidates for values in the nodal points, coming from every mesh element adjacent to a given nodal point, and nodal basis elements are related to these values.

Let us introduce some standard notation in DGMs. Assume that e is an edge of some element $\tau \in \mathcal{T}_h$. Since we restrict ourselves to conforming meshes, then either case is possible: e is an edge between exactly two mesh elements $\tau_1 := \tau, \tau_2 \in \mathcal{T}_h$, or e lies on the boundary, i.e. $e \subset \partial\Omega$, and it is adjacent only to element τ. We therefore define two operators, for a given e: the mean value operator

$$\{u\} := \begin{cases} \frac{u|_{\tau_1} + u|_{\tau_2}}{2}, & \text{if } e = \partial\tau_1 \cap \partial\tau_2, \\ u, & \text{if } e = \partial\tau \cap \partial\Omega, \end{cases} \tag{1.1.27}$$

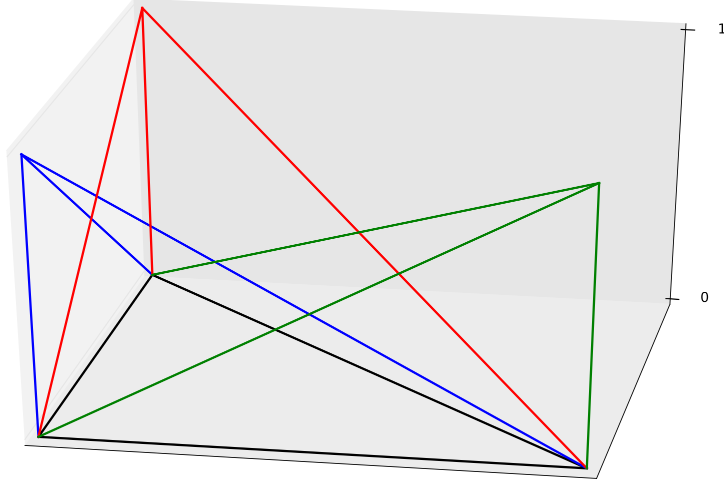

Figure 1.5: An example of nodal basis elements of a linear Discontinuous Galerkin Method space for a two-dimensional domain related to some triangle $\tau \in \mathcal{T}_h$ (black). The nodal basis elements are plotted in red, blue and green.

and the jump operator

$$[u] := \begin{cases} u|_{\tau_1} - u|_{\tau_2}, & \text{if } e = \partial\tau_1 \cap \partial\tau_2, \\ u, & \text{if } e = \partial\tau \cap \partial\Omega. \end{cases} \tag{1.1.28}$$

The order of the elements τ_1, τ_2 in second definition is not important, but it must be consequent within a given problem. For convenience, we define Γ_h to be the set of all interior edges of the mesh $\mathcal{T}_h$.

The discrete problem for DGM is given as follows.

Problem 1.1.17. *Let $f \in L_2(\Omega)$. Find $u_h^* \in X_{h,0}(\Omega)$ such that for every $\phi_h \in X_{h,0}(\Omega)$*

$$\begin{aligned} &\int_\Omega \nabla u_h^* \cdot \nabla \phi_h \, dx - \sum_{e\in\Gamma_h} \int_e \{\nabla u_h^* \cdot \nu\}[\phi_h] \, ds \\ +\xi &\sum_{e\in\Gamma_h} \int_e \{\nabla \phi_h \cdot \nu\}[u_h^*] \, ds + \sum_{e\in\Gamma_h} \frac{\sigma_e}{|e|} \int_e [u_h^*][\phi_h] \, ds = \int_\Omega f \phi_h \, dx, \end{aligned} \tag{1.1.29}$$

where

$$X_{h,0}(\Omega) := \{v_h \in X_h(\Omega) : v_h|_{\partial\Omega} \equiv 0\}, \tag{1.1.30}$$

and $\xi \in \{-1, 0, 1\}$ and $\sigma_e \geq 0$ are given, as discussed below.

Let us first discuss problem 1.1.17 with $\xi = 0$ and $\sigma_e = 0$ $\forall e \in \Gamma_h$. In this case, definition of this problem differs from the definition of problem 1.1.15 by the second term: $-\sum_{e\in\Gamma_h} \int_e \{\nabla u_h^* \cdot \nu\}[\phi_h] \, ds$. Let us elaborate on the genesis of this term. To obtain the variational form of problem 1.1.14 from equation (1.1.20), one has to multiply (1.1.20) by a test function and use the Green formula. To do so, sufficiently high regularity must be assumed on u_h^*, like $u_h^* \in H^2(\Omega)$, but then the problem may be generalized to $H^1(\Omega)$. In case of DGM space, the Green formulas may be used only on the triangulation elements. Since the test functions are discontinuous, the boundary terms do not vanish. They are included is the second term.

At this point, the relation between solutions on adjacent mesh elements it too loose. While ∇u is constrained by the second term, we can easily see that there is no restriction on function's values. For example, addition of an arbitrary constant to $u_h^*|_\tau$ if $\tau \in \mathcal{T}_h$ is not adjacent to $\partial\Omega$ do not broke the discrete solution. To remedy this problem, fourth term $\sum_{e\in\Gamma_h} \frac{\sigma_e}{|e|} \int_e [u_h^*][\phi_h]\, ds$ is added. It is called the penalty term. It constrains the jumps of the solution, provided that penalty parameters σ_e are chosen properly.

Finally we would like to comment on third term $\xi \sum_{e\in\Gamma_h} \int_e \{\nabla\phi_h \cdot \nu\}[u_h^*]\, ds$. If $\xi = -1$, then it makes the problem symmetric. The problem is then called *Symmetric Interior Penalty Galerkin* (SIPG) method. This method converges if the penalty parameters $\sigma_e > 0$ are large enough [19]. If $\xi = 0$, then the problem is then called *Incomplete Interior Penalty Galerkin* (IIPG) method. It is not symmetric, and this method also converges if the penalty parameters $\sigma_e > 0$ are large enough [19].

For continuous FEM, we presented an error estimate in $H^1(\Omega)$-norm. This approach is not feasible for DGMs, as discrete solutions do not belong to $H^1(\Omega)$. It is a standard approach to estimate the error in the *broken norm* (or *energy norm*)

$$\|u_h\|_h^2 := \sum_{\tau\in\mathcal{T}_h} \int_\tau \nabla u_h \cdot \nabla u_h\, dx + \sum_{e\in\Gamma_h} \frac{\sigma_e}{|e|} \int_e [u_h][u_h]\, ds. \tag{1.1.31}$$

The following error estimate holds [19].

Proposition 1.1.18. *Assume that $\mathcal{T}_h$ is a triangulation. Assume that the exact solution u^* of problem 1.1.14 belongs to $H^2(\mathcal{T}_h)$ (see equation 1.3.1 for the definition). Assume that the penalty parameters $\sigma_e > 0$ are large enough.*

Then there is a constant C independent of h, such that

$$\|u^* - u_h^*\|_h^2 \le Ch^2 \sum_{\tau\in\mathcal{T}_h} \|u^*\|_{H^2(\tau)}^2. \tag{1.1.32}$$

1.1.3 Weakly Over-Penalized Symmetric Interior Penalty

The *Weakly Over-Penalized Symmetric Interior Penalty* (WOPSIP) method was introduced in [17]. The idea is similar as in SIPG or IIPG method, but the problem formulation is simpler, at cost of the higher penalty term.

Problem 1.1.19. *Let $f \in L_2(\Omega)$. For any $e \in \Gamma_h$ let $\Pi^0 : L_2(e) \to \mathbb{P}_0(e)$ be the orthogonal projection. Find $u_h^* \in X_{h,0}(\Omega)$ such that for every $\phi_h \in X_{h,0}(\Omega)$*

$$\int_\Omega \nabla u_h^* \cdot \nabla\phi_h\, dx + \sum_{e\in\Gamma_h} \frac{\sigma}{|e|^3} \int_e \Pi^0[u_h^*]\Pi^0[\phi_h]\, ds = \int_\Omega f\phi_h\, dx. \tag{1.1.33}$$

As in SIPG/IIPG methods, the discrete space is not a subset of the variational problem space, so some modification of the problem formulation was necessary. This formulation is simpler, as it only contains the additional penalty term. In this case the broken norm is defined as

$$\|u_h\|_h^2 := \sum_{\tau\in\mathcal{T}_h} \int_\tau \nabla u_h \cdot \nabla u_h\, dx + \sum_{e\in\Gamma_h} |e| \int_e \nabla u_h \nabla u_h\, ds + \sum_{e\in\Gamma_h} \frac{\sigma}{|e|^3} \int_e \Pi^0[u_h]\Pi^0[u_h]\, ds. \tag{1.1.34}$$

We have the following error estimate [17].

Proposition 1.1.20. *Assume that $\mathcal{T}_h$ is a triangulation, and assume it is quasi-uniform. Assume that the exact solution u^* of problem 1.1.14 belongs to $H^2(\mathcal{T}_h)$.*

Then for any $\sigma > 0$ there is a constant C independent of h, such that

$$\|u^* - u_h^*\|_h^2 \leq Ch^2 \sum_{\tau \in \mathcal{T}_h} |u^*|^2_{H^2(\tau)}. \tag{1.1.35}$$

WOPSIP discrete solutions has therefore similar error estimates as SIPG/IIPG. However, simple form comes at a price of increased conditioning of the stiffness matrix, leading to numerical problems. For the Poisson equation, there is a preconditioner presented in [17].

1.2 Differential problem

In this section we would like to elaborate on the differential problem to be discretized. We present general drift-diffusion system first, and then we pass to the equilibrium state, which is used extensively in this chapter. Details on the physical justification of the drift-diffusion model are presented in section 2.5.

1.2.1 Drift-diffusion system

We start with the domain Ω of our problem. Luminescent semiconductor devices are generally made of planar layers deposited one on another, which vary in composition of a semiconductor material or number of impurities. At opposite ends of the device metal contacts are attached, to which the current can be applied. If this is the case, it flows through the device perpendicular to the deposited layers. We assume that $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2\}$, is an interval or a polygon and that $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$. The boundary of Ω may be either the electrical contact ($\partial\Omega_D$) or a contact with insulator ($\partial\Omega_N$). If $\Omega \subset \mathbb{R}$, then we assume that $\partial\Omega_D = \partial\Omega$, which means that the device has two electrical contacts on the opposite ends.

For analysis, we consider the following version of van Roosbroeck system [54]

$$\begin{aligned} -\nabla \cdot (\varepsilon(x)\nabla u^*(x)) + e^{u^*(x)-v^*(x)} - e^{w^*(x)-u^*(x)} &= k_1(x), \\ -\nabla \cdot (\mu_n(x)e^{u^*(x)-v^*(x)}\nabla v^*(x)) - Q(u^*(x), v^*(x), w^*(x))(e^{w^*(x)-v^*(x)} - 1) &= 0, \\ -\nabla \cdot (\mu_p(x)e^{w^*(x)-u^*(x)}\nabla w^*(x)) + Q(u^*(x), v^*(x), w^*(x))(e^{w^*(x)-v^*(x)} - 1) &= 0. \end{aligned} \tag{1.2.1}$$

We will refer the first equation in this set as the *Poisson equation* and the latter equations as *(electron/hole) continuity equations*. The weak formulation of this system is as follows.

Problem 1.2.1. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2\}$ be an interval or polygon. Let $\hat{u}, \hat{v}, \hat{w} \in H^1(\Omega) \cap L_\infty(\Omega)$ be some given functions. We say that $(u^*, v^*, w^*) \in (\hat{u}, \hat{v}, \hat{w}) + \left(H_0^1(\Omega)\right)^3$ is a weak solution of (1.2.1) if $\forall\phi \in H_0^1(\Omega)$*

$$\begin{aligned} \int_\Omega \varepsilon(x)\nabla u^*(x)\nabla\phi(x)\,dx &= \int_\Omega \left(k_1(x) - e^{u^*(x)-v^*(x)} + e^{w^*(x)-u^*(x)}\right)\phi(x)\,dx, \\ \int_\Omega \mu_n(x)e^{u^*(x)-v^*(x)}\nabla v^*(x)\nabla\phi(x)\,dx &= \int_\Omega Q(u^*(x), v^*(x), w^*(x))(e^{w^*(x)-v^*(x)} - 1)\phi(x)\,dx, \\ \int_\Omega \mu_p(x)e^{w^*(x)-u^*(x)}\nabla w^*(x)\nabla\phi(x)\,dx &= -\int_\Omega Q(u^*(x), v^*(x), w^*(x))(e^{w^*(x)-v^*(x)} - 1)\phi(x)\,dx. \end{aligned} \tag{1.2.2}$$

Let $P(u, v, w) := Q(u, v, w)(e^{w-v} - 1)$ for $u, v, w \in \mathbb{R}$. We assume the following:

Assumption A1.

1. *$\Omega \subset \mathbb{R}^d$ for $d \in \{1, 2\}$, and it is an interval ($d = 1$) or a polygon ($d = 2$).*

2. *$0 \leq Q(u, v, w) \leq Q_M$ for any $u, v, w \in \mathbb{R}$.*

3. *$P(u, v, w)$ is monotone decreasing in v for $u, v, w \in \mathbb{R}$.*

4. *$P(u, v, w)$ is monotone increasing in w for $u, v, w \in \mathbb{R}$.*

5. *P is locally Lipschitz.*

6. *$0 < \varepsilon_m \leq \varepsilon(x) \leq \varepsilon_M$ for some $\varepsilon_m, \varepsilon_M \in \mathbb{R}$.*

7. *$k_1 \in L_\infty(\Omega)$.*

8. *μ_n, μ_p are Lipschitz continuous functions.*

9. *$0 < \mu_m \leq \mu_n(x), \mu_p(x) \leq \mu_M$ for some constants $\mu_m, \mu_M \in \mathbb{R}$.*

1.2.2 Equilibrium state

Equilibrium state is a physical state of a semiconductor device, where the device is disconnected from a power source, so the current does not flow through it and the generation and recombination are in balance. Then only the Poisson equation needs to be solved and the system reduce to the following equation:

$$\begin{aligned} -\nabla \cdot \Big(\varepsilon(x)\nabla u^*\Big) + e^{u^*-\hat{v}} - e^{\hat{w}-u^*} &= k_1, \\ u^* &= \hat{u} \text{ on } \partial\Omega_D, \\ \nabla u^* \cdot \nu &= 0 \text{ on } \partial\Omega_N, \end{aligned} \tag{1.2.3}$$

where $k_1, \hat{v}, \hat{w}$ are given. A weak formulation of (1.2.3) is as follows:

Problem 1.2.2. *Let $\hat{v}, \hat{w} \in L_\infty(\Omega)$ and $k_1 \in L_2(\Omega)$ be given. Find $u^* \in \hat{u} + H^1(\Omega)$, such that*

$$a(u^*, \phi) + b(u^*, \phi) = f(\phi) \quad \forall \phi \in H^1_{0,\partial\Omega_D}(\Omega), \tag{1.2.4}$$

where

$$\begin{aligned} a(u, \phi) &:= \int_\Omega \varepsilon(x)\nabla u(x) \cdot \nabla\phi(x)\, dx, \\ b(u, \phi) &:= \int_\Omega \Big(e^{u(x)-\hat{v}(x)} - e^{\hat{w}(x)-u(x)}\Big)\phi(x)\, dx, \\ f(\phi) &:= \int_\Omega k_1(x)\phi(x)\, dx. \end{aligned} \tag{1.2.5}$$

From the physical standpoint, in the equilibrium state $\hat{v} = \hat{w} \equiv$ const. However, since existence and uniqueness of the discrete problem related to this case may be also applied to non-equilibrium case, we allow functions $\hat{v}, \hat{w}$ to be in $L_\infty(\Omega)$. A proof of the following result may be found in [54].

Lemma 1.2.3. *Solution u^* of problem 1.2.2 is bounded.*

1.3 Composite Discontinuous Galerkin Method

In this section we will introduce the Composite Discontinuous Galerkin Method variants, used in this research in theoretical analysis and numerical simulations. First we describe the discrete functional space, and then we introduce the discrete problems. These problems are formulated for a linear elliptic equation first. Adjustment of these discretizations to our nonlinear problem is presented in section 1.4.

1.3.1 Discrete space

Let open set $\Omega \subset \mathbb{R}^2$ ($\mathbb{R}^1$) be a rectangle (interval), divided to disjoint rectangles (intervals) $\{\Omega_i\}_{i=1}^N =: \mathcal{E}$ in such a manner that $\mathcal{E}$ is a conforming mesh of Ω. We will call this division a coarse mesh and we assume that each external edge of any Ω_i is contained either in $\partial\Omega_D$ or in $\partial\Omega_N$.

On every Ω_i we would like to introduce a mesh, which is in general independent of the meshes of neighboring elements Ω_j, $i \neq j$. Let us define $\mathcal{T}_{h_i} := \mathcal{T}_{i,h_i}(\Omega_i)$ to be triangulations of Ω_i, where $h_i := \max\{h_\tau : \tau \in \mathcal{T}_{h_i}\}$, and $h_\tau := \operatorname{diam}(\tau)$. By $\mathcal{N}_{h_i}$ we denote the nodes of the mesh $\mathcal{T}_{h_i}$.

Assumption A2. *$\{\mathcal{T}_{i,h_i}(\Omega)\}_{h_i}$ is a quasi-uniform family of meshes (see definition 1.1.6).*

We will define $\mathcal{T}_h := \bigcup_{i=1}^N \mathcal{T}_{h_i}$. Note that $\mathcal{T}_h$ is a nonconforming mesh of Ω.

For $s > 0$, we define the broken Sobolev spaces $H^s(\mathcal{E})$ and $H^s(\mathcal{T}_h)$ (see [94]) as

$$\begin{aligned} H^s(\mathcal{E}) &:= \{v \in L_2(\Omega) : \forall i \in \{1, \dots, N\} \;\; v_i := v|_{\Omega_i} \in H^s(\Omega_i)\} \subset L_2(\Omega), \\ H^s(\mathcal{T}_h) &:= \{v \in L_2(\Omega) : \forall \tau \in \mathcal{T}_h \;\; v|_\tau \in H^s(\tau)\} \subset L_2(\Omega), \end{aligned} \tag{1.3.1}$$

equipped with the broken Sobolev norms

$$\|v\|_{H^s(\mathcal{E})} := \left(\sum_{\Omega_i \in \mathcal{E}} \|v\|^2_{H^s(\Omega_i)} \right)^{1/2}, \qquad \|v\|_{H^s(\mathcal{T}_h)} := \left(\sum_{\tau \in \mathcal{T}_h} \|v\|^2_{H^s(\tau)} \right)^{1/2}, \tag{1.3.2}$$

and seminorms

$$|v|_{H^s(\mathcal{E})} := \left(\sum_{\Omega_i \in \mathcal{E}} |v|^2_{H^s(\Omega_i)} \right)^{1/2}, \qquad |v|_{H^s(\mathcal{T}_h)} := \left(\sum_{\tau \in \mathcal{T}_h} |v|^2_{H^s(\tau)} \right)^{1/2}. \tag{1.3.3}$$

Then on every Ω_i we define a discrete space $X_{h_i}(\Omega_i) \subset \mathcal{C}(\overline{\Omega_i})$ of piecewise linear functions on the triangulation $\mathcal{T}_{h_i}$:

$$X_{h_i} := X_{h_i}(\Omega_i) := \left\{ u_{h,i} \in \mathcal{C}(\overline{\Omega}_i) : \forall \tau \in \mathcal{T}_{h_i} \;\; u_{h,i}\big|_\tau \in \mathbb{P}_1(\tau) \right\}. \tag{1.3.4}$$

Space $X_{h_i}(\Omega_i)$ is a continuous linear finite element space (see definition 1.1.8) on Ω_i.

Finally we define $X_h(\Omega)$ as

$$X_h(\Omega) = \left\{ (u_{h,1}, \dots, u_{h,N}) : u_{h,i} \in X_{h_i}(\Omega_i), i \in \{1, \dots, N\} \right\} \subset L_2(\Omega). \tag{1.3.5}$$

Thus functions from $X_h(\Omega)$ are not continuous in general. In particular, while $X_h(\Omega)$ is a product of continuous linear finite element spaces, it is not a continuous linear finite element space. Note that $X_h(\Omega) \not\subset H^1(\Omega)$ and $X_h(\Omega) \not\subset H^2(\mathcal{E})$, but $X_h(\Omega) \subset H^1(\mathcal{E})$, $H^1(\Omega) \subset H^1(\mathcal{E})$ and $X_h(\Omega) \subset H^2(\mathcal{T}_h)$.

By Γ we denote the set of all internal and boundary edges of subdomains $\Omega_i \in \mathcal{E}$. We assume the following:

Assumption A3. *The coarse mesh $\mathcal{E}$ is chosen in such a manner so that Γ is a sum of disjoint sets Γ_D, Γ_N and Γ_I, where*

$$\begin{aligned} \Gamma_D &:= \{e \in \Gamma : e \subset \partial\Omega_D\}, \\ \Gamma_N &:= \{e \in \Gamma : e \subset \partial\Omega_N\}, \\ \Gamma_I &:= \{e \in \Gamma : e \subset \mathrm{int}(\Omega)\}. \end{aligned} \tag{1.3.6}$$

Therefore Γ_D (resp. Γ_N) contains edges lying on the boundary where Dirichlet (resp. Neumann) boundary conditions are imposed and in Γ_I there are all internal edges, which we call interfaces, as they frequently correspond to the physical interfaces between different semiconductor materials. We also define

$$\Gamma_{DI} := \Gamma_D \cup \Gamma_I, \quad \Gamma_i := \{e \in \Gamma : e \subset \partial\Omega_i\}. \tag{1.3.7}$$

Let $e \in \Gamma$. Then two cases are possible. Either $e \in \Gamma_D \cup \Gamma_N$, so there is a unique $\Omega_i \in \mathcal{E}$ such that e is an edge of Ω_i, or $e \in \Gamma_I$ and there are exactly two sets $\Omega_i, \Omega_j \in \mathcal{E}$ such that e is their common edge. We will often refer to these two cases. We also define the set of neighboring domains $\mathrm{nb}(\Omega_i) := \{\Omega_k \in \mathcal{E} : \Gamma_i \cap \Gamma_k \neq \emptyset\}$.

Note that in one dimension, every $e \in \Gamma$ is a point. In $\mathbb{R}^2$ we may consider functions $f : e \to \mathbb{R}$ for any $e \in \Gamma$ and integrals $\int_e f\, ds$. However for one dimension any $e \in \Gamma$ is a singleton of a point from $\overline{\Omega}$.

Thus if $\Omega \subset \mathbb{R}$ and $e \in \Gamma$, then for convenience we denote

$$\int_e f\, ds := f(e). \tag{1.3.8}$$

For $s > 1/2$ we define operators $[\cdot] := [\cdot]_e : H^s(\mathcal{E}) \to L_2(e)$, $\{\cdot\} := \{\cdot\}_e : H^s(\mathcal{E}) \to L_2(e)$ as

$$\begin{aligned} [u] &:= \begin{cases} u_i - u_j, & \text{if } e \subset \Gamma_I, e = \partial\Omega_i \cap \partial\Omega_j, i < j, \\ u_i, & \text{if } e \subset \Gamma_D \cup \Gamma_N, e = \partial\Omega_i \cap \partial\Omega, \end{cases} \\ \{u\} &:= \begin{cases} \frac{1}{2}\big(u_i + u_j\big), & \text{if } e \subset \Gamma_I, e = \partial\Omega_i \cap \partial\Omega_j, \\ u_i, & \text{if } e \subset \Gamma_D \cup \Gamma_N, e = \partial\Omega_i \cap \partial\Omega. \end{cases} \end{aligned} \tag{1.3.9}$$

Similar notation is used in Discontinuous Galerkin Method (see section 1.1.2). Note that in this case, these operators are related to interfaces of the coarse mesh $\mathcal{E}$, which do not change with h.

For convenience, we will also use an analogous notation for triangulation parameters, i.e.

$$\{h^{-r}\} := \left\{\frac{1}{h^r}\right\} := \begin{cases} \frac{1}{2}\Big(\frac{1}{h_i^r} + \frac{1}{h_j^r}\Big), & \text{if } e = \partial\Omega_i \cap \partial\Omega_j, \\ \frac{1}{h_i^r}, & \text{if } e = \partial\Omega_i \cap \partial\Omega. \end{cases} \tag{1.3.10}$$

For further analysis, we introduce broken norm $\|\cdot\|_{h,\Sigma_r}$ in $X_h(\Omega)$ [37] as

$$\|u_h\|^2_{h,\Sigma_r} := \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon(x)\Big(\nabla u_{h,i}\Big)^2 dx + \sum_{e \in \Gamma_{DI}} \eta_{r,e} \int_e [u_h]^2\, ds. \tag{1.3.11}$$

Here $\eta_{r,e}$ is a penalty coefficient for e. It depends on the triangulation parameters and penalty parameters $\sigma_e > 0$ [94]:

$$\eta_{r,e} := 2\sigma_e\{h^{-r}\} = \begin{cases} \sigma_e\Big(h_i^{-r} + h_j^{-r}\Big), & \text{if } e \in \Gamma_I, e \subset \Omega_i \cap \Omega_j, \\ 2\sigma_e h_i^{-r}, & \text{if } e \in \Gamma_D, e \subset \Omega_i. \end{cases} \tag{1.3.12}$$

Depending on the method, we will use $r = 1$ or $r = 2$.

To simplify the analysis, we assume the following:

Assumption A4.

- $\Gamma_D \neq \emptyset$.
- $\mathcal{T}_h$ *is a shape regular mesh (see definition 1.1.5).*

We will use the following standard result for finite element spaces:

Lemma 1.3.1. *For any* $u_h \in X_h(\Omega)$, $\Omega_i \in \mathcal{E}$ *and* $e \in \Gamma_i$, *the following estimates hold*

$$\begin{aligned} \|u_{h,i}\|^2_{L_2(e)} &\leq Ch_i^{-1}\|u_h\|^2_{L_2(\Omega_i)}, & (1.3.13)\\ \|\nabla u_{h,i}\cdot\nu\|^2_{L_2(e)} &\leq Ch_i^{-1}|u_h|^2_{H^1(\Omega_i)}. & (1.3.14)\end{aligned}$$

Constant C *does not depend on* h_i.

Proof. These estimates are a consequence of the trace theorem for finite element spaces (see corollary 1.1.12) used for $X_{h_i}(\Omega_i)$ and applied to each $\partial\tau \cap e$, where $e \subset \partial\Omega_i$, $\tau \in \mathcal{T}_{h,i}$. Let $\mathcal{T}_{h,e} := \{\tau \in \mathcal{T}_h : \tau \text{ has an edge on } e\}$. Then

$$\begin{aligned} \|u_{h,i}\|^2_{L_2(e)} &= \sum_{\tau\in\mathcal{T}_{h,e}\cap\mathcal{T}_{h,i}} \|u_{h,i}\|^2_{L_2(\tau\cap e)} \leq \sum_{\tau\in\mathcal{T}_{h,e}\cap\mathcal{T}_{h,i}} h_i^{-1}\|u_{h,i}\|^2_{L_2(\tau)} \\ &\leq h_i^{-1}\sum_{\tau\in\mathcal{T}_{h,e}\cap\mathcal{T}_{h,i}} \|u_{h,i}\|^2_{L_2(\tau)} \leq h_i^{-1}\|u_{h,i}\|^2_{L_2(\Omega_i)}, \end{aligned} \tag{1.3.15}$$

and analogously

$$\begin{aligned} \|\nabla u_{h,i}\cdot\nu\|^2_{L_2(e)} &= \sum_{\tau\in\mathcal{T}_{h,e}\cap\mathcal{T}_{h,i}} \|\nabla u_{h,i}\cdot\nu\|^2_{L_2(\tau\cap e)} \leq h_i^{-1}\sum_{\tau\in\mathcal{T}_{h,e}\cap\mathcal{T}_{h,i}} |u_{h,i}|^2_{H^1(\tau)} \\ &\leq h_i^{-1}|u_{h,i}|^2_{H^1(\Omega_i)}. \end{aligned} \tag{1.3.16}$$

□

1.3.2 Composite Discontinuous Galerkin variants

We propose two variants of the Composite Discontinuous Galerkin discretization. First approach is based on Weakly Over-Penalized Symmetric Interior Penalty (WOPSIP) method (cf. [17, 11] or section 1.1.3). Second approach is derived from Symmetric Interior Penalty Galerkin (SIPG) method (cf. [94, 90] or section 1.1.2). In each case we use the composite scheme (cf. [37]), i.e. inside every Ω_i we use a standard continuous linear Finite Element Method on the triangulation $\mathcal{T}_{h_i}$, while on boundaries $e \in \Gamma_{DI}$ we use the respective variant of the Discontinuous Galerkin Method. Thus we call these methods *Composite Weakly Over-Penalized Symmetric Interior Penalty* (CWOPSIP) method and *Composite Symmetric Interior Penalty Galerkin* (CSIPG) method, respectively.

1.3.2.1 Linear differential problem

The discretizations will be first constructed for the auxiliary linear problem.

Problem 1.3.2. *Let* $\Omega \subset \mathbb{R}^d$, $d \in \{1,2\}$ *be a rectangle (interval) and let* $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, *where* $\partial\Omega_D \cap \partial\Omega_N = \emptyset$. *Let* $f \in L_2(\Omega)$. *Find* $u^* \in \hat{u} + H^1_{0,\partial\Omega_D}(\Omega)$ *such that*

$$a(u^*,\phi) = f(\phi) \quad \forall\phi \in H^1_{0,\partial\Omega_D}(\Omega), \tag{1.3.17}$$

where

$$a(u,\phi) := \int_\Omega \varepsilon(x)\nabla u(x)\cdot\nabla\phi(x)\,dx, \quad f(\phi) := \int_\Omega f(x)\phi(x)\,dx, \tag{1.3.18}$$

and

$$H^1_{0,\partial\Omega_D}(\Omega) := \{u \in H^1(\Omega) : u|_{\partial\Omega_D} = 0\}. \tag{1.3.19}$$

We assume that ε is positive and bounded, as indicated in assumptions A1.

Assumption A5. $\hat{u} \in H^1(\Omega) \cap L_\infty(\Omega)$.

1.3.2.2 Composite Weakly Over-Penalized Symmetric Interior Penalty (CWOPSIP)

This discretization has a simpler formulation of the two methods we introduce. The discrete problem is defined as follows.

Problem 1.3.3. *Find* $u_h^* \in X_h(\Omega)$ *such that*

$$a_{h,2}(u_h^*,\phi_h) = f_{h,2}(\phi_h), \quad \forall\phi_h \in X_h(\Omega), \tag{1.3.20}$$

where

$$\begin{aligned} a_{h,2}(u_h,\phi_h) =& \sum_{i=1}^N \int_{\Omega_i} \varepsilon\nabla u_{h,i}\cdot\nabla\phi_{h,i}\,dx + \sum_{e\in\Gamma_{DI}} \eta_{2,e}\int_e [u_h]\cdot[\phi_h]\,ds, \\ f_{h,2}(\phi_h) =& \int_\Omega f\phi_h\,dx + \sum_{e\in\Gamma_D} \eta_{2,e}\int_e [\hat{u}]\cdot[\phi_h]\,ds. \end{aligned} \tag{1.3.21}$$

In the theoretical analysis, it is helpful that the broken norm can be expressed in terms of operator $a_{h,2}$, i.e.

$$\|u_h\|^2_{h,\Sigma_2} = a_{h,2}(u_h,u_h). \tag{1.3.22}$$

1.3.2.3 Composite Symmetric Interior Penalty Galerkin (CSIPG)

This discrete problem is defined as follows.

Problem 1.3.4. *Find* $u_h^* \in X_h(\Omega)$ *such that*

$$a_{h,1}(u_h^*,\phi_h) = f_{h,1}(\phi_h), \quad \forall\phi_h \in X_h(\Omega), \tag{1.3.23}$$

where

$$\begin{aligned} a_{h,1}(u_h,\phi_h) =& \sum_{i=1}^N \int_{\Omega_i} \varepsilon\nabla u_{h,i}\cdot\nabla\phi_{h,i}\,dx - \sum_{e\in\Gamma_{DI}} \int_e \{\varepsilon\nabla u_h\cdot\nu\}[\phi_h]\,ds \\ & - \sum_{e\in\Gamma_{DI}} \int_e \{\varepsilon\nabla\phi_h\cdot\nu\}[u_h]\,ds + \sum_{e\in\Gamma_{DI}} \eta_{1,e}\int_e [u_h]\cdot[\phi_h]\,ds, \\ f_{h,1}(\phi_h) =& \int_\Omega f\phi_h\,dx - \sum_{e\in\Gamma_D} \int_e \{\varepsilon\nabla\phi_h\cdot\nu\}[\hat{u}]\,ds \\ & + \sum_{e\in\Gamma_D} \eta_{1,e}\int_e [\hat{u}][\phi_h]\,ds. \end{aligned} \tag{1.3.24}$$

In this case, we cannot relate operator $a_{h,1}$ to the broken norm $\|\cdot\|_{h,\Sigma_1}$ (defined in (1.3.11)) in such a simple manner as in (1.3.22), because

$$\|u_h\|^2_{h,\Sigma_1} = a_{h,1}(u_h,u_h) + 2\sum_{e\in\Gamma_{DI}}\int_e \{\varepsilon\nabla u_h\cdot\nu\}[u_h]\,ds. \tag{1.3.25}$$

Instead we can establish the following lemma.

Lemma 1.3.5. *For any $\alpha\in(0,1)$ there exist $\sigma_m>0$ and $c>0$, such that for every $\sigma_e\geq\sigma_m$ and $u_h, v_h\in X_h(\Omega)$*

$$2\sum_{e\in\Gamma_{DI}}\Big|\int_e\{\varepsilon\nabla u_h\cdot\nu\}[v_h]\,ds\Big| \leq \alpha\|u_h\|_{h,\Sigma_1}\|v_h\|_{h,\Sigma_1}, \tag{1.3.26}$$

and

$$(1-\alpha)\|u_h\|^2_{h,\Sigma_1}\leq a_{h,1}(u_h,u_h), \tag{1.3.27}$$

where $\|\cdot\|_{h,\Sigma_1}$ is defined in (1.3.11).

Proof. First we prove estimate (1.3.26). Let us take any $e=\Omega_j\cap\Omega_k$. By the Schwarz inequality and a triangle inequality

$$\begin{aligned}\Big|2\int_e\{\varepsilon\nabla u_h\cdot\nu\}[v_h]\,ds\Big| \leq \Big(&\big\|\varepsilon_{|\Omega_j}\nabla u_{h,j}\cdot\nu\big\|_{L_2(e)}\\ &+\big\|\varepsilon_{|\Omega_k}\nabla u_{h,k}\cdot\nu\big\|_{L_2(e)}\Big)\big\|[v_h]\big\|_{L_2(e)}.\end{aligned} \tag{1.3.28}$$

Taking $\Omega_i\in\{\Omega_j,\Omega_k\}$, we use lemma 1.3.1

$$\begin{aligned}\big\|\varepsilon_{|\Omega_i}\nabla u_{h,i}\cdot\nu\big\|_{L_2(e)}\big\|[v_h]\big\|_{L_2(e)} &\leq \varepsilon_M\big\|\nabla u_{h,i}\cdot\nu\big\|_{L_2(e)}\big\|[v_h]\big\|_{L_2(e)}\\ &=\varepsilon_M\sqrt{h_i}\big\|\nabla u_{h,i}\cdot\nu\big\|_{L_2(e)}\frac{1}{\sqrt{h_i}}\big\|[v_h]\big\|_{L_2(e)}\\ &\leq\varepsilon_M\big\|\nabla u_{h,i}\big\|_{L_2(\Omega_i)}\frac{1}{\sqrt{h_i}}\big\|[v_h]\big\|_{L_2(e)}.\end{aligned} \tag{1.3.29}$$

Therefore we get

$$\Big|2\int_e\{\varepsilon\nabla u_h\cdot\nu\}[u_h]\,ds\Big|\leq\varepsilon_M\big\|\nabla u_{h,i}\big\|_{L_2(\Omega_i)}\frac{1}{\sqrt{h_i}}\big\|[v_h]\big\|_{L_2(e)}+\varepsilon_M\big\|\nabla u_{h,j}\big\|_{L_2(\Omega_i)}\frac{1}{\sqrt{h_j}}\big\|[v_h]\big\|_{L_2(e)}. \tag{1.3.30}$$

On the other hand, if $e\in\Gamma_D$ then $e\in\partial\Omega_i\cap\partial\Omega$ and by similar arguments we have

$$\Big|2\int_e\{\varepsilon\nabla u_h\cdot\nu\}[u_h]\,ds\Big|\leq\varepsilon_M\big\|\nabla u_{h,i}\big\|_{L_2(\Omega_i)}\frac{1}{\sqrt{h_i}}\big\|[v_h]\big\|_{L_2(e)}. \tag{1.3.31}$$

Summing these results up and using Cauchy's inequality, for any $\alpha>0$

$$\begin{aligned}2\sum_{e\in\Gamma_{DI}}\Big|\int_e\{\varepsilon\nabla u_h\cdot\nu\}[u_h]\,ds\Big| &\leq\varepsilon_M\sum_{e\in\Gamma_{DI}}\sum_{\Gamma_i\ni e}\big\|\nabla u_{h,i}\big\|_{L_2(\Omega_i)}\frac{1}{\sqrt{h_i}}\big\|[v_h]\big\|_{L_2(e)}\\ &\leq\varepsilon_M\Bigg(\sum_{i=1}^N\big\|\nabla u_{h,i}\big\|^2_{L_2(\Omega_i)}\Bigg)^{1/2}\Bigg(\sum_{e\in\Gamma_{DI}}2\{h_i^{-1}\}\big\|[v_h]\big\|^2_{L_2(e)}\Bigg)^{1/2}\\ &\leq\alpha\|u_h\|_{h,\Sigma_1}\Bigg(\sum_{e\in\Gamma_{DI}}\frac{2\varepsilon_M^2}{\varepsilon_m\alpha^2}\{h_i^{-1}\}\big\|[v_h]\big\|^2_{L_2(e)}\Bigg)^{1/2}\\ &\leq\alpha\|u_h\|_{h,\Sigma_1}\Bigg(\sum_{e\in\Gamma_{DI}}\sigma_e\{h_i^{-1}\}\big\|[v_h]\big\|^2_{L_2(e)}\Bigg)^{1/2},\end{aligned} \tag{1.3.32}$$

where the last inequality is true if we take for example $\sigma_m(\alpha) := 2\varepsilon_M^2/\varepsilon_m\alpha^2$. Thus we get

$$2\sum_{e\in\Gamma_{DI}}\Big|\int_e\{\varepsilon\nabla u_h\cdot\nu\}[v_h]\,ds\Big| \le \alpha\|u_h\|_{h,\Sigma_1}\|v_h\|_{h,\Sigma_1}. \tag{1.3.33}$$

Thus estimate (1.3.26) is proven for any $\alpha > 0$.

We now restrict to $\alpha \in (0,1)$ and we pass to estimate (1.3.27). Due to definition of broken norm, cf. equation (1.3.11), we have

$$a_{h,1}(u_h,u_h) = \|u_h\|^2_{h,\Sigma_1} - 2\sum_{e\in\Gamma_{DI}}\int_e\{\varepsilon\nabla u_h\cdot\nu\}[u_h]\,ds. \tag{1.3.34}$$

Using (1.3.26) we get

$$\begin{aligned} a_{h,1}(u_h,u_h) &= \|u_h\|^2_{h,\Sigma_1} - 2\sum_{e\in\Gamma_{DI}}\int_e\{\varepsilon\nabla u_h\cdot\nu\}[u_h]\,ds \\ &\ge \|u_h\|^2_{h,\Sigma_1} - \alpha\|u_h\|^2_{h,\Sigma_1} \ge (1-\alpha)\|u_h\|^2_{h,\Sigma_1}. \end{aligned} \tag{1.3.35}$$

□

1.3.3 Broken norm and the Poincare inequality

We would like to have an analogue of the Poincare inequality for the $H^1(\mathcal{E})$ space. To do so, we would like to use the following result, which was proven in [18]

Lemma 1.3.6. *There is some constant $C > 0$ such that for any $u \in H^1(\mathcal{E})$*

$$\|u\|^2_{L_2(\Omega)} \le C\Big[\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}(\nabla u)^2\,dx + \sum_{e\in\Gamma_I}|e|^{-1}\int_e[u]^2\,ds + \int_{\partial\Omega_D}u^2\,ds\Big]. \tag{1.3.36}$$

Proof. This result is proven in [18]. We use results (8.1), (1.8) of [18] for one dimension and two dimensions, respectively. □

The Poincare inequality analogue is as follows.

Lemma 1.3.7. *Let $u \in H^s(\mathcal{E})$, $s \ge 1$. Then for sufficiently small $h = \max\{h_1,\dots,h_N\}$ we have $\|u\|_{L_2(\Omega)} \le c\|u\|_{h,\Sigma_r}$, where c is independent of h.*

Proof. By definition of the broken norm (1.3.11), we have

$$\|u\|^2_{h,\Sigma_r} := \sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}\varepsilon\Big(\nabla u\Big)^2\,dx + \sum_{e\in\Gamma_{DI}}\eta_{r,e}\int_e[u]^2\,ds. \tag{1.3.37}$$

Using lemma 1.3.6

$$\|u\|^2_{L_2(\Omega)} \le C\Big[\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}(\nabla u)^2\,dx + \sum_{e\in\Gamma_I}|e|^{-1}\int_e\big([u]\big)^2\,ds + \int_{\partial\Omega_D}u^2\,ds\Big]. \tag{1.3.38}$$

Note that $|e|$ does not depend on h and $\eta_{r,e}\to\infty$ as $h\to 0$. Thus we can find $h_M > 0$, such that $\eta_{r,e} \ge |e|^{-1}$ and $\eta_{r,e} \ge 1$ for any $0 < h < h_M \le 1$ and then

$$\begin{aligned} &C\Big[\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}(\nabla u)^2\,dx + \sum_{e\in\Gamma_I}|e|^{-1}\int_e[u]^2\,ds + \sum_{e\in\Gamma_D}\int_e[u]^2\,ds\Big] \\ &\quad\le C\Big[\varepsilon_m^{-1}\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}\varepsilon(\nabla u)^2\,dx + \sum_{e\in\Gamma_{DI}}\eta_{r,e}\int_e[u]^2\,ds\Big] \le C\|u\|^2_{h,\Sigma_r}. \end{aligned} \tag{1.3.39}$$

□

Now we show that $\|\cdot\|_{h,\Sigma_r}$ is indeed a norm in $H^1(\mathcal{E})$. In fact, we will prove slightly more general lemma:

Lemma 1.3.8. *The space $H^1(\mathcal{E})$ equipped with a scalar product*

$$\langle u|v\rangle_{h,\Sigma_r} := \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon \nabla u \cdot \nabla v \, dx + \sum_{e\in\Gamma_{DI}} \eta_{r,e} \int_e [u][v] \, ds, \tag{1.3.40}$$

is a Hilbert space.

Proof. Linearity and symmetry of this form is obvious, so we only demonstrate that

$$\langle u|u\rangle_{h,\Sigma_r} = 0 \quad \Leftrightarrow \quad u \equiv 0, \tag{1.3.41}$$

for any $u \in H^1(\mathcal{E})$. Note that

$$\langle u|u\rangle_{h,\Sigma_r} := \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon(x)\Big(\nabla u(x)\Big)^2 dx + \sum_{e\in\Gamma_{DI}} \eta_{e,r} \int_e \Big([u]\Big)^2 ds \geq 0, \tag{1.3.42}$$

as $\varepsilon(x) > 0$ and $\eta_{e,r} > 0$. If $u \equiv 0$, then $\nabla u \equiv 0$ and by definition (1.3.9) $[u] = 0$ for every $e \in \Gamma_{DI}$, thus $\langle u|u\rangle_{h,\Sigma_r} = 0$.

On the other hand, assume that $\langle u|u\rangle_{h,\Sigma_r} = 0$. Then for any $\Omega_i \in \mathcal{E}$ we have $|u|_{H^1(\Omega_i)} = 0$, so $u_{h,i}(x) := c_i = \text{const}$. For any adjacent $\Omega_i, \Omega_j \in \mathcal{E}$ with the common edge e

$$\eta_{e,r}|e|(c_i - c_j)^2 \leq \sum_{\tilde{e}\in\Gamma_{DI}} \eta_{\tilde{e},r} \int_{\tilde{e}} \Big([u]\Big)^2 ds = 0, \tag{1.3.43}$$

and for any Ω_i with an edge $e \in \Gamma_D$

$$\eta_{e,r}|e|c_i^2 \leq \sum_{\tilde{e}\in\Gamma_{DI}} \eta_{\tilde{e},r} \int_{\tilde{e}} \Big([u]\Big)^2 ds = 0. \tag{1.3.44}$$

Combining these two results, we obtain that $c_i = 0$ for every $\Omega_i \in \mathcal{E}$ and thus $u \equiv 0$.

Still we have to prove that $H^1(\mathcal{E})$ with a broken norm is a complete space. Let $\{u_{(n)}\}_n$ be a Cauchy sequence in the broken norm. Then for every $v \in H^1(\mathcal{E})$

$$\sum_{i=1}^{N} |v_i|^2_{H^1(\Omega_i)} \leq \varepsilon_M \|v\|^2_{h,\Sigma_r}. \tag{1.3.45}$$

Also by lemma 1.3.7 we have

$$\|v\|_{L_2(\Omega)} \leq c\|v\|^2_{h,\Sigma_r}. \tag{1.3.46}$$

These estimates imply that for any $i \in \{1,\dots,N\}$ the sequence $\{u_{(n),i}\}_n$ is a Cauchy sequence in $H^1(\Omega_i)$. Therefore $u_{(n),i} \to u_i$ in $H^1(\Omega_i)$ for some $u_i \in H^1(\Omega_i)$, as it is a closed space.

We will show that $u_{(n)} \to u := (u_1,\dots,u_N)$ in $H^1(\mathcal{E})$. We have

$$\|u_{(n)} - u\|^2_{h,\Sigma_r} \leq \varepsilon_M \sum_{i=1}^{N} |u_{(n),i} - u_i|^2_{H^1(\Omega_i)} \, dx + \sum_{e\in\Gamma_{DI}} \eta_{r,e} \|[u_{(n)} - u]\|^2_{L_2(e)} \, ds. \tag{1.3.47}$$

It is therefore clear that both elements of this sum goes to zero with n, as convergence if $H^1(\Omega_i)$ implies convergence in seminorm and convergence of traces in $L^2(e)$ on any $e \in \Gamma_{DI}$. Therefore $\|u_{(n)} - u\|_{h,\Sigma_r} \to 0$, so $H^1(\mathcal{E})$ with the broken norm is closed.

Then $\big(H^1(\mathcal{E}), \langle\cdot|\cdot\rangle_{h,\Sigma_r}\big)$ is a Hilbert space. □

1.3.4 Consistency

At this point we have two kind of problems. There is a differential problem 1.3.2:

Problem 1.3.2. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{1,2\}$ be a rectangle (interval) and let $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, where $\partial\Omega_D \cap \partial\Omega_N = \emptyset$. Let $f \in L_2(\Omega)$. Find $u^* \in \hat{u} + H^1_{0,\partial\Omega_D}(\Omega)$ such that*

$$a(u^*,\phi) = f(\phi) \quad \forall \phi \in H^1_{0,\partial\Omega_D}(\Omega), \tag{1.3.17}$$

where

$$a(u,\phi) := \int_\Omega \varepsilon(x)\nabla u(x)\cdot\nabla\phi(x)\,dx, \quad f(\phi) := \int_\Omega f(x)\phi(x)\,dx, \tag{1.3.18}$$

and

$$H^1_{0,\partial\Omega_D}(\Omega) := \{u \in H^1(\Omega) : u|_{\partial\Omega_D} = 0\}. \tag{1.3.19}$$

We have also two related discrete problems 1.3.3, 1.3.4. There is a crucial difference in incorporating of the Dirichlet boundary conditions in these problems. In the differential problem, they are imposed strongly, by appropriate constrains on the problem's domain and the test space. On the other hand, the discrete problems impose Dirichlet boundary conditions weakly, by penalty terms in forms $a_{h,r}, f_{h,r}$.

For further analysis, we would like to formulate a variational problem, which is an analogue of the differential problem 1.3.2 with Dirichlet boundary conditions imposed weakly. We therefore define the following problem:

Problem 1.3.9. *Find $u^* \in H^2(\mathcal{E})$, such that $\forall\phi \in H^1(\mathcal{E}) \cap H^2(\mathcal{T}_h)$*

$$\begin{aligned}\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}\varepsilon\nabla u^*\cdot\nabla\phi\,dx &- \sum_{e\in\Gamma_{DI}}\int_e\Big\{\varepsilon\nabla u^*\cdot\nu\Big\}[\phi]\,ds\\ &+\xi_r\sum_{e\in\Gamma_{DI}}\int_e\Big\{\varepsilon\nabla\phi\cdot\nu\Big\}[u^*]\,ds+\sum_{e\in\Gamma_{DI}}\eta_{r,e}\int_e[u^*][\phi]\,ds\\ &=\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}f\phi\,dx+\xi_r\sum_{e\in\Gamma_D}\int_e\Big\{\varepsilon\nabla\phi\cdot\nu\Big\}[\hat{u}]\,ds+\sum_{e\in\Gamma_D}\eta_{r,e}\int_e[\hat{u}][\phi]\,ds,\end{aligned} \tag{1.3.48}$$

where $\xi_1 := -1$ and $\xi_2 := 0$.

Coefficients ξ_1, ξ_2 correspond to CSIPG and CWOPSIP, respectively.

To use problem 1.3.9 instead of problem 1.3.2, we have to show that under certain regularity assumptions, these problems have the same solutions. At this point, we have to introduce some assumptions on the domain Ω.

Assumption A6.

- *$\Omega \subset \mathbb{R}^d$, $d \in \{1,2\}$, is an interval ($d = 1$) or a polygon ($d = 2$).*
- *$\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$.*
- *$\partial\Omega_D$ has nonzero measure.*

We would like to prove the following result, an analogue of proposition 2.9 of [94].

Theorem 1.3.10. *Under assumptions A3, A6, suppose that the solution u of problem 1.3.2 belongs to $H^1(\Omega)\cap H^2(\mathcal{E})$ and $\varepsilon\nabla u \in H^1(\mathcal{E})$. Then u is a solution of problem 1.3.9. Conversely, if $u \in H^2(\mathcal{E})\cap H^1(\Omega)$ is a solution of problem 1.3.9 and $\varepsilon\nabla u \in H^1(\mathcal{E})$, then it is also a solution of problem 1.3.2.*

Before we prove this theorem, we would like to establish the following two lemmas.

Lemma 1.3.11. *Under assumptions A3, A6, let* $u \in H^1(\Omega) \cap H^2(\mathcal{E})$, $\varepsilon \in L_\infty(\Omega)$, $\varepsilon \nabla u \in H^1(\mathcal{E})$, $0 < \varepsilon_m \le \varepsilon \le \varepsilon_M$ *and* $f \in L_2(\Omega)$. *Moreover let* u *satisfy:*

$$\int_\Omega \varepsilon \nabla u \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx, \quad \forall \phi \in H^1_{0,\partial\Omega_D}(\Omega). \tag{1.3.49}$$

Then for every $e \in \Gamma_I \cup \Gamma_N$ *we have*

$$\Big[\varepsilon \nabla u \cdot \nu\Big]\Big|_e = 0. \tag{1.3.50}$$

Proof. Take $e \in \Gamma_I$, $e = \partial\Omega_i \cap \partial\Omega_j$. Then take any $\bar{\phi} \in C_0^\infty(e)$ ($\bar{\phi} \in \mathbb{R}$ for one dimension). Since $\mathcal{E}$ consists of rectangles (resp. intervals), we may extend $\bar{\phi}$ to $\phi \in C_0^\infty(\overline{\Omega_i \cup \Omega_j})$. The lemma conditions imply

$$\int_\Omega \varepsilon \nabla u \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx, \tag{1.3.51}$$

Then by Green's formula (theorem 4.A.3) and assumptions on u and $\varepsilon \nabla u$ we have on Ω_i, Ω_j

$$\begin{aligned} \int_{\Omega_i} \varepsilon \nabla u \cdot \nabla \phi &= -\int_{\Omega_i} \nabla \cdot \Big(\varepsilon \nabla u\Big) \phi \, dx + \int_{\partial\Omega_i} \varepsilon \nabla u \cdot \nu_{\Omega_i} \phi \, ds, \\ \int_{\Omega_j} \varepsilon \nabla u \cdot \nabla \phi &= -\int_{\Omega_j} \nabla \cdot \Big(\varepsilon \nabla u\Big) \phi \, dx + \int_{\partial\Omega_j} \varepsilon \nabla u \cdot \nu_{\Omega_j} \phi \, ds. \end{aligned} \tag{1.3.52}$$

Let $\nu := \nu_{\Omega_i} = -\nu_{\Omega_j}$ on e. Summing up these equations and noting that boundary integrals are nonzero only on e and $\mathrm{supp}(\phi) \subset \overline{\Omega_i \cup \Omega_j}$, we obtain

$$\int_\Omega f \phi \, dx = \int_\Omega \varepsilon \nabla u \cdot \nabla \phi \, dx = -\int_\Omega \nabla \cdot \Big(\varepsilon \nabla u\Big) \phi \, dx + \int_e \Big[\varepsilon \nabla u \cdot \nu\Big] \phi \, ds. \tag{1.3.53}$$

Now take sequence $\{\phi_\varepsilon\}_\varepsilon \subset C_0^\infty(\overline{\Omega_i \cup \Omega_j})$, such that $\phi_\varepsilon|_e = \bar{\phi}$ and $\|\phi_\varepsilon\|_{L_2(\Omega)} \xrightarrow[\varepsilon \to 0]{} 0$. Then we have

$$\left| \int_\Omega f \phi_\varepsilon \, dx \right| \le \|f\|_{L_2(\Omega)} \|\phi_\varepsilon\|_{L_2(\Omega)} \xrightarrow[\varepsilon \to 0]{} 0, \tag{1.3.54}$$

and

$$\left| \int_\Omega \nabla \cdot \Big(\varepsilon \nabla u\Big) \phi_\varepsilon \, dx \right| \le \Big\| \nabla \cdot \Big(\varepsilon \nabla u\Big) \Big\|_{L_2(\Omega)} \|\phi_\varepsilon\|_{L_2(\Omega)} \xrightarrow[\varepsilon \to 0]{} 0, \tag{1.3.55}$$

since $\nabla \cdot \Big(\varepsilon \nabla u\Big) \in L_2(\Omega)$. Thus due to

$$\int_\Omega f \phi_\varepsilon \, dx = -\int_\Omega \nabla \cdot \Big(\varepsilon \nabla u\Big) \phi_\varepsilon \, dx + \int_e \Big[\varepsilon \nabla u \cdot \nu\Big] \bar{\phi} \, ds, \tag{1.3.56}$$

we obtain

$$\int_e \Big[\varepsilon \nabla u \cdot \nu\Big] \bar{\phi} \, ds = 0. \tag{1.3.57}$$

Since this is true for any $\bar{\phi} \in C_0^\infty(e)$, we have that

$$\Big[\varepsilon \nabla u \cdot \nu\Big] = 0 \quad \text{in } L_2(e). \tag{1.3.58}$$

Then let $e \in \Gamma_N$, $e \subset \partial\Omega_i$ for some $\Omega_i \in \mathcal{E}$. Again we fix $\bar{\phi} \in C_0^\infty(e)$ (in one dimension $\Gamma_N = \emptyset$, so this step is omitted) and we extend it to $\phi \in C^\infty(\Omega)$, such that $\mathrm{supp}(\phi) \in \overline{\Omega_i}$ and $\phi_{\partial\Omega_i \setminus e} \equiv 0$. Then by Green's theorem, we obtain

$$\int_\Omega f\phi\,dx = \int_\Omega \varepsilon\nabla u \cdot \nabla\phi\,dx = -\int_\Omega \nabla \cdot \Big(\varepsilon\nabla u\Big)\phi\,dx + \int_e \varepsilon\nabla u \cdot \nu\phi\,ds. \tag{1.3.59}$$

Taking sequence $\{\phi_\varepsilon\}_\varepsilon \subset C^\infty(\Omega)$, such that $\phi_\varepsilon|_e = \bar{\phi}$, $\mathrm{supp}(\phi) \in \overline{\Omega_i}$, $\phi_{\partial\Omega_i\setminus e} \equiv 0$ and $\|\phi_\varepsilon\|_{L_2(\Omega)} \xrightarrow[\varepsilon\to 0]{} 0$, using the above equation we get

$$\int_\Omega f\phi_\varepsilon\,dx = -\int_\Omega \nabla \cdot \Big(\varepsilon\nabla u\Big)\phi_\varepsilon\,dx + \int_e \varepsilon\nabla u \cdot \nu\bar{\phi}\,ds. \tag{1.3.60}$$

By the same estimations as for $e \in \Gamma_I$, we obtain

$$\int_e \varepsilon\nabla u \cdot \nu\bar{\phi}\,ds = 0. \tag{1.3.61}$$

Since this is true for any $\bar{\phi} \in C_0^\infty(e)$, we have that

$$[\varepsilon\nabla u \cdot \nu]|_e = \varepsilon\nabla u \cdot \nu|_e = 0. \tag{1.3.62}$$

□

Lemma 1.3.12. *Under assumptions A3, A6, let $u \in H^1(\Omega) \cap H^2(\mathcal{E})$, $\varepsilon \in L_\infty(\Omega)$, $\varepsilon\nabla u \in H^1(\mathcal{E})$, $0 < \varepsilon_m \le \varepsilon \le \varepsilon_M$ and $f \in L_2(\Omega)$. The following conditions are equivalent:*

(1). u satisfy:

$$\int_\Omega \varepsilon\nabla u \cdot \nabla\phi\,dx = \int_\Omega f\phi\,dx, \quad \forall\phi \in H^1_{0,\partial\Omega_D}(\Omega). \tag{1.3.63}$$

(2). u satisfy:

$$\begin{gathered} -\sum_{\Omega_i\in\mathcal{E}} \int_{\Omega_i} \nabla \cdot \Big(\varepsilon\nabla u\Big)\phi\,dx = \int_\Omega f\phi\,dx, \quad \forall\phi \in L_2(\Omega), \\ \Big[\varepsilon\nabla u \cdot \nu\Big]\Big|_e = 0 \quad \forall e \in \Gamma_I, \\ \nabla u \cdot \nu = 0 \quad \text{on } \partial\Omega_N. \end{gathered} \tag{1.3.64}$$

Proof. (2) ⇒ (1)

Take any $\phi \in H^1_{0,\partial\Omega_D}(\Omega)$. We have

$$-\sum_{\Omega_i\in\mathcal{E}} \int_{\Omega_i} \nabla \cdot \Big(\varepsilon\nabla u\Big)\phi\,dx = \int_\Omega f\phi\,dx. \tag{1.3.65}$$

For a given $\Omega_i \in \mathcal{E}$, conditions of this lemma imply $u|_{\Omega_i} \in H^2(\Omega_i)$, $\varepsilon\nabla u|_{\Omega_i} \in H^1(\Omega_i)$ and we have $\phi|_{\Omega_i} \in H^1(\Omega_i)$, thus we may use Green's formula (theorem 4.A.3) on Ω_i

$$-\int_{\Omega_i} \nabla \cdot \Big(\varepsilon\nabla u\Big)\phi\,dx = \int_{\Omega_i} \varepsilon\nabla u \cdot \nabla\phi\,dx - \int_{\partial\Omega_i} \varepsilon\nabla u \cdot \nu\phi\,ds. \tag{1.3.66}$$

Summing up over $\Omega_i \in \mathcal{E}$

$$\sum_{\Omega_i \in \mathcal{E}} \int_{\Omega_i} \varepsilon \nabla u \cdot \nabla \phi \, dx - \sum_{e \in \Gamma} \int_e [\varepsilon \nabla u \cdot \nu] \phi \, ds = \int_\Omega f \phi \, dx. \tag{1.3.67}$$

If $e \in \Gamma_D$, then by definition $\phi|_e = 0$. On the other hand, for $e \in \Gamma_I$ we have $[\varepsilon \nabla u \cdot \nu] = 0$ and for $e \in \Gamma_N$ we have $\nabla u \cdot \nu = 0$, so $[\varepsilon \nabla u \cdot \nu] = \varepsilon \nabla u \cdot \nu = 0$. Thus the second sum is zero and we obtain the result

$$\sum_{e \in \mathcal{E}} \int_{\Omega_i} \varepsilon \nabla u \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx. \tag{1.3.68}$$

(1) $\Rightarrow$ (2)
Take any $\phi \in C_0^\infty(\Omega)$. Since $C_0^\infty(\Omega) \subset H^1_{0,\partial\Omega_D}(\Omega)$, then by (1.3.63) we have

$$\int_\Omega \varepsilon \nabla u \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx. \tag{1.3.69}$$

By Green's formula, as above

$$\int_\Omega f \phi \, dx = \sum_{e \in \mathcal{E}} \int_{\Omega_i} \varepsilon \nabla u \cdot \nabla \phi \, dx = - \sum_{\Omega_i \in \mathcal{E}} \int_{\Omega_i} \nabla \cdot \Big(\varepsilon \nabla u \Big) \phi \, dx + \sum_{e \in \Gamma} \int_e [\varepsilon \nabla u \cdot \nu] \phi \, ds. \tag{1.3.70}$$

Since ϕ is zero on $\partial\Omega$, we may rewrite last sum

$$\int_\Omega f \phi = - \sum_{\Omega_i \in \mathcal{E}} \int_{\Omega_i} \nabla \cdot \Big(\varepsilon \nabla u \Big) \phi \, dx + \sum_{e \in \Gamma_I} \int_e [\varepsilon \nabla u \cdot \nu] \phi \, ds. \tag{1.3.71}$$

Then by lemma 1.3.11 applied to (1.3.63), we have

$$\forall e \in \Gamma_I \quad [\varepsilon \nabla u \cdot \nu] = 0, \tag{1.3.72}$$

and we obtain

$$- \sum_{\Omega_i \in \mathcal{E}} \int_{\Omega_i} \nabla \cdot \Big(\varepsilon \nabla u \Big) \phi \, dx = \int_\Omega f \phi \, dx. \tag{1.3.73}$$

But this is true for $\phi \in C_0^\infty(\Omega)$. To obtain this result for $\phi \in L_2(\Omega)$, we use lemma 4.B.2 with $f := f$, $g := \nabla \cdot (\varepsilon \nabla u)$. Note that we assume that $\varepsilon \nabla u \in H^1(\mathcal{E})$, so $(\varepsilon \nabla u)|_{\Omega_i} \in H^1(\Omega_i)$ for every $\Omega_i \in \mathcal{E}$ and thus $\nabla \cdot (\varepsilon \nabla u)|_{\Omega_i} \in L_2(\Omega_i)$. Therefore $\nabla \cdot (\varepsilon \nabla u) \in L_2(\Omega)$ and first statement of (1.3.64) is shown.

To shown remaining statements, we again use lemma 1.3.11

$$\forall e \in \Gamma_I \cup \Gamma_N \quad [\varepsilon \nabla u \cdot \nu] = 0. \tag{1.3.74}$$

For $e \in \Gamma_N$ we conclude that since $\varepsilon > 0$, and

$$[\varepsilon \nabla u \cdot \nu] \overset{\text{def}}{=} \varepsilon \nabla u \cdot \nu = 0, \tag{1.3.75}$$

then

$$\nabla u \cdot \nu = 0. \tag{1.3.76}$$

$\square$

We now give the proof of theorem 1.3.10.

Proof. *(Theorem 1.3.10)*

Problem 1.3.2 $\Rightarrow$ problem 1.3.9.

First assume that u is a solution of problem 1.3.2 and that it belongs to $H^1(\Omega)\cap H^2(\mathcal{E})$. We have by definition

$$\int_\Omega \varepsilon\nabla u\cdot\nabla\phi\,dx = \int_\Omega f\phi\,dx \quad \forall\phi\in H^1_{0,\partial\Omega_D}(\Omega). \tag{1.3.77}$$

We use lemma 1.3.12 and we obtain that for any $\phi\in L_2(\Omega)$

$$-\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}\nabla\cdot\Big(\varepsilon\nabla u\Big)\phi\,dx = \int_\Omega f\phi\,dx. \tag{1.3.78}$$

Let us restrict to $\phi\in H^1(\mathcal{E})\cap H^2(\mathcal{T}_h)$. Then $\phi_i\in H^1(\Omega_i)$ and by Green's theorem we have

$$\int_{\Omega_i}\nabla\cdot\Big(\varepsilon\nabla u\Big)\phi\,dx = \int_{\Omega_i}\varepsilon\nabla u\cdot\nabla\phi\,dx - \int_{\partial\Omega_i}\varepsilon\nabla u\cdot\nu\phi\,dx. \tag{1.3.79}$$

Summing up these results over Ω_i, we get

$$\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}\varepsilon\nabla u\cdot\nabla\phi\,dx - \sum_{\Omega_i\in\mathcal{E}}\int_{\partial\Omega_i}\varepsilon\nabla u\cdot\nu\phi\,dx = \int_\Omega f\phi\,dx. \tag{1.3.80}$$

By lemma 1.3.12, we have that $[\varepsilon\nabla u\cdot\nu]=0$ on every $e\in\Gamma_I$, thus $\{\varepsilon\nabla u\cdot\nu\}=\varepsilon\nabla u\cdot\nu$ on any $\partial\Omega_i$ and we have

$$\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}\varepsilon\nabla u\cdot\nabla\phi\,dx - \sum_{e\in\Gamma}\int_e\Big\{\varepsilon\nabla u\cdot\nu\Big\}[\phi]\,dx = \int_\Omega f\phi\,dx. \tag{1.3.81}$$

By homogeneous Neumann boundary condition (lemma 1.3.12) on $e\in\Gamma_N$ we have $\{\varepsilon\nabla u\cdot\nu\}=0$ and

$$\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}\varepsilon\nabla u\cdot\nabla\phi\,dx - \sum_{e\in\Gamma_{DI}}\int_e\Big\{\varepsilon\nabla u\cdot\nu\Big\}[\phi]\,dx = \int_\Omega f\phi\,dx. \tag{1.3.82}$$

Since $u\in H^1(\Omega)$, then $[u]=0$ for any $e\in\Gamma_I$ and, by assumption, on $e\in\Gamma_D$ we have $u=\hat{u}$, so for any $\phi\in H^1(\mathcal{E})$ we get

$$\begin{aligned}
&\sum_{e\in\Gamma_{DI}}\eta_{r,e}\int_e[u][\phi]\,ds + \xi_r\sum_{e\in\Gamma_{DI}}\int_e\Big\{\varepsilon\nabla\phi\cdot\nu\Big\}[u]\,ds\\
&\qquad = \sum_{e\in\Gamma_I}\eta_{r,e}\int_e 0\cdot[\phi]\,ds + \xi_r\sum_{e\in\Gamma_I}\int_e\Big\{\varepsilon\nabla\phi\cdot\nu\Big\}0\,ds\\
&\qquad + \sum_{e\in\Gamma_D}\eta_{r,e}\int_e[\hat{u}][\phi]\,ds + \xi_r\sum_{e\in\Gamma_D}\int_e\Big\{\varepsilon\nabla\phi\cdot\nu\Big\}[\hat{u}]\,ds\\
&\qquad = \sum_{e\in\Gamma_D}\eta_{r,e}\int_e[\hat{u}][\phi]\,ds + \xi_r\sum_{e\in\Gamma_D}\int_e\Big\{\varepsilon\nabla\phi\cdot\nu\Big\}[\hat{u}]\,ds.
\end{aligned} \tag{1.3.83}$$

By adding this result side-by-side to (1.3.82) we obtain

$$\begin{aligned}
&\sum_{\Omega_i\in\mathcal{E}}\int_{\Omega_i}\varepsilon\nabla u\cdot\nabla\phi\,dx - \sum_{e\in\Gamma_{DI}}\int_e\Big\{\varepsilon\nabla u\cdot\nu\Big\}[\phi]\,dx + \xi_r\sum_{e\in\Gamma_{DI}}\int_e\Big\{\varepsilon\nabla\phi\cdot\nu\Big\}[u]\,ds + \sum_{e\in\Gamma_{DI}}\eta_{r,e}\int_e[u][\phi]\,ds\\
&\qquad = \int_\Omega f\phi\,dx + \xi_r\sum_{e\in\Gamma_D}\int_e\Big\{\varepsilon\nabla\phi\cdot\nu\Big\}[\hat{u}]\,ds + \sum_{e\in\Gamma_D}\eta_{r,e}\int_e[\hat{u}][\phi]\,ds.
\end{aligned} \tag{1.3.84}$$

Since this is true for any $\phi \in H^1(\mathcal{E}) \cap H^2(\mathcal{T}_h)$, we have (1.3.48).

Problem 1.3.9 $\Rightarrow$ problem 1.3.2.

Conversely assume (1.3.48) is true for some $u \in H^2(\mathcal{E}) \cap H^1(\Omega)$, i.e. $\forall \phi \in H^1(\mathcal{E}) \cap H^2(\mathcal{T}_h)$

$$\begin{aligned} \sum_{\Omega_i \in \mathcal{E}} \int_{\Omega_i} \varepsilon \nabla u \cdot \nabla \phi \, dx - \sum_{e \in \Gamma_{DI}} \int_e \Big\{ \varepsilon \nabla u \cdot \nu \Big\} [\phi] \, ds + \xi_r \sum_{e \in \Gamma_{DI}} \int_e \Big\{ \varepsilon \nabla \phi \cdot \nu \Big\} [u] \, ds + \sum_{e \in \Gamma_{DI}} \eta_{r,e} \int_e [u][\phi] \, ds \\ = \int_\Omega f \phi \, dx + \xi_r \sum_{e \in \Gamma_D} \int_e \Big\{ \varepsilon \nabla \phi \cdot \nu \Big\} [\hat{u}] \, ds + \sum_{e \in \Gamma_D} \eta_{r,e} \int_e [\hat{u}][\phi] \, ds, \end{aligned} \tag{1.3.85}$$

and that $\varepsilon \nabla u \in H^1(\mathcal{E})$.

First we recover the Dirichlet boundary conditions. Take any $e \in \Gamma_D$, such that $e \subset \partial\Omega_i$, and $\bar{\phi} \in C_0^\infty(e)$. Then let $\{\phi_\epsilon\}$ be a sequence of functions, such that

$$\begin{aligned} \phi_\epsilon \in C^\infty(\Omega), \quad \phi_\epsilon|_e = \bar{\phi}, \quad \operatorname{supp}(\phi_\epsilon) \subset \Omega_i \cup e, \quad \phi_\epsilon|_{\partial\Omega_i \setminus e} \equiv 0, \\ \nabla \phi_\epsilon \cdot \nu \Big|_{\partial\Omega_i} = 0, \quad \|\phi_\epsilon\|_{L_2(\Omega)} \xrightarrow[\epsilon \to 0]{} 0. \end{aligned} \tag{1.3.86}$$

We can get such a functions by expanding functions obtained from lemma 4.B.4 by 0 to whole Ω. Then $\phi \in H^1(\mathcal{E})$ and (1.3.48) becomes

$$\begin{aligned} \int_{\Omega_i} \varepsilon \nabla u \cdot \nabla \phi_\epsilon \, dx - \int_e \varepsilon \nabla u \cdot \nu \bar{\phi} \, ds + \xi_r \int_e \varepsilon \nabla \phi_\epsilon \cdot \nu u \, ds + \eta_{r,e} \int_e u \bar{\phi} \, ds = \\ \int_{\Omega_i} f \phi_\epsilon \, dx + \xi_r \int_e \varepsilon \nabla \phi_\epsilon \cdot \nu \hat{u} \, ds + \eta_{r,e} \int_e \hat{u} \bar{\phi} \, ds. \end{aligned} \tag{1.3.87}$$

Due to definition of ϕ_ϵ, $\nabla \phi_\epsilon \cdot \nu|_e = 0$ and we have

$$\int_{\Omega_i} \varepsilon \nabla u \cdot \nabla \phi_\epsilon \, dx - \int_e \varepsilon \nabla u \cdot \nu \bar{\phi} \, ds + \eta_{r,e} \int_e u \bar{\phi} \, ds = \int_{\Omega_i} f \phi_\epsilon \, dx + \eta_{r,e} \int_e \hat{u} \bar{\phi} \, ds. \tag{1.3.88}$$

By the Green formula

$$\int_{\Omega_i} \nabla \cdot \Big(\varepsilon \nabla u \Big) \phi_\epsilon \, dx + \eta_{r,e} \int_e u \bar{\phi} \, ds = \int_{\Omega_i} f \phi_\epsilon \, dx + \eta_{r,e} \int_e \hat{u} \bar{\phi} \, ds. \tag{1.3.89}$$

Passing to the limit $\epsilon \to 0$

$$\eta_{r,e} \int_e u \bar{\phi} \, ds = \eta_{r,e} \int_e \hat{u} \bar{\phi} \, ds. \tag{1.3.90}$$

Since $\bar{\phi} \in C_0^\infty(e)$ and $e \in \Gamma_D$ are arbitrary, we get

$$u|_{\partial\Omega_D} = \hat{u}|_{\partial\Omega_D}, \tag{1.3.91}$$

and the Dirichlet boundary conditions are satisfied.

Then take any $\phi \in \mathcal{C}^\infty_{0,\partial\Omega_D}(\Omega)$. Thus

$$\sum_{e \in \Gamma_{DI}} \eta_{r,e} \int_e [u][\phi] \, ds = \sum_{e \in \Gamma_D} \eta_{r,e} \int_e [\hat{u}][\phi] \, ds = 0. \tag{1.3.92}$$

as $[\phi] = 0$ for any $e \in \Gamma_I$ since $\phi \in \mathcal{C}^\infty_{0,\partial\Omega_D}(\Omega)$ and on $e \in \Gamma_D$ we have $[\phi] = \phi \equiv 0$. By the same argument

$$- \sum_{e \in \Gamma_{DI}} \int_e \Big\{ \varepsilon \nabla u \cdot \nu \Big\} [\phi] \, ds = 0. \tag{1.3.93}$$

Then $u \in H^1(\Omega)$, so $[u] = 0$ for any $e \in \Gamma_I$ while as we have already shown that $u = \hat{u}$ for $e \in \Gamma_D$, so

$$\xi_r \sum_{e \in \Gamma_{DI}} \int_e \Big\{ \varepsilon \nabla \phi \cdot \nu \Big\} [u] \, ds = \xi_r \sum_{e \in \Gamma_D} \int_e \Big\{ \varepsilon \nabla \phi \cdot \nu \Big\} [\hat{u}] \, ds. \tag{1.3.94}$$

Thus we obtain

$$\sum_{\Omega_i \in \mathcal{E}} \int_{\Omega_i} \varepsilon \nabla u \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx. \tag{1.3.95}$$

Since choice of $\phi \in \mathcal{C}^\infty_{0,\partial\Omega_D}(\Omega)$ is arbitrary, this statement is also true for any $\phi \in H^1_{0,\partial\Omega_D}(\Omega)$ (see lemma 4.B.3), so u satisfies equation (1.3.17).

□

1.4 Discretization of the equilibrium case

We would like to establish a discrete analogue of equation (1.2.4) using discretizations introduced in sections 1.3.2. Then we will show existence and uniqueness of these nonlinear problems.

We would like to introduce additional assumptions, useful in context of error estimates.

Assumption A7.

- *There is some $0 < h_M \leq 1$ such that for any $0 < h < h_M$ and for any $e \in \Gamma_{DI}$ we have $\eta_{e,r} \geq |e|^{-1}$ and $\eta_{e,r} \geq 1$ (cf. (1.3.12)).*
- *Constant h_M is sufficiently small, so that for any $0 < h < h_M$ lemma 1.3.7 holds.*
- (CSIPG only) *Constant $\sigma_m > 0$ is sufficiently large such that lemma 1.3.5 holds with $\alpha = 1/2$.*
- $\varepsilon|_{\Omega_i} \in \mathcal{C}^1(\overline{\Omega_i})$ *for every $\Omega_i \in \mathcal{E}$* (this assumption could be weakened, but in semiconductor simulations this function is normally constant or linear on Ω_i).
- *$u^* \in H^1(\Omega) \cap H^2(\mathcal{E})$, where u^* is a solution of problem 1.2.2.*
- *$\hat{v}, \hat{w} \in L_2(\Omega) \cap L_\infty(\Omega)$, where $\hat{v}, \hat{w}$ are defined in problem 1.2.2.*

1.4.1 Composite Weakly Over-Penalized Symmetric Interior Penalty (CWOPSIP)

We start with CWOPSIP discretization, as it is simpler. We modify problem 1.3.3 by including the nonlinear part of (1.2.4).

Problem 1.4.1. *Find $u_h^* \in X_h(\Omega)$ such that*

$$a_{h,2}(u_h^*, \phi_h) + b(u_h^*, \phi_h) = f_{h,2}(\phi_h), \quad \forall \phi_h \in X_h(\Omega), \tag{1.4.1}$$

where

$$\begin{aligned} a_{h,2}(u_h, \phi_h) &= \sum_{i=1}^N \int_{\Omega_i} \varepsilon \nabla u_{h,i} \cdot \nabla \phi_{h,i} \, dx + \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [u_h] \cdot [\phi_h] \, ds, \\ b(u_h, \phi_h) &:= \int_\Omega \Big(e^{u_h(x) - \hat{v}(x)} - e^{\hat{w}(x) - u_h(x)} \Big) \phi_h(x) \, dx, \\ f_{h,2}(\phi_h) &= \int_\Omega k_1 \phi_h \, dx + \sum_{e \in \Gamma_D} \eta_{2,e} \int_e [\hat{u}] \cdot [\phi_h] \, ds. \end{aligned} \tag{1.4.2}$$

1.4.2 Composite Symmetric Interior Penalty Galerkin (CSIPG)

Analogously as in case of CWOPSIP, we extend problem 1.3.4 by the nonlinear part.

Problem 1.4.2. *Find $u_h^* \in X_h(\Omega)$ such that*

$$a_{h,1}(u_h^*, \phi_h) + b(u_h^*, \phi_h) = f_{h,1}(\phi_h), \quad \forall \phi_h \in X_h(\Omega), \tag{1.4.3}$$

where

$$\begin{aligned} a_{h,1}(u_h, \phi_h) =& \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon \nabla u_{h,i} \cdot \nabla \phi_{h,i} \, dx - \sum_{e \in \Gamma_{DI}} \int_e \{\varepsilon \nabla u_h \cdot \nu\}[\phi_h] \, ds \\ &- \sum_{e \in \Gamma_{DI}} \int_e \{\varepsilon \nabla \phi_h \cdot \nu\}[u_h] \, ds + \sum_{e \in \Gamma_{DI}} \eta_{1,e} \int_e [u_h] \cdot [\phi_h] \, ds, \\ f_{h,1}(\phi_h) =& \int_\Omega k_1 \phi_h \, dx - \sum_{e \in \Gamma_D} \int_e \{\varepsilon \nabla \phi_h \cdot \nu\}[\hat{u}] \, ds \\ &+ \sum_{e \in \Gamma_D} \eta_{1,e} \int_e [\hat{u}][\phi_h] \, ds, \end{aligned} \tag{1.4.4}$$

and b is defined as in problem 1.4.1.

1.4.3 Existence and uniqueness

We would like to show that problems 1.4.1 and 1.4.2 are well-posed.

Proposition 1.4.3. *Under assumptions A1 to A7, problems 1.4.1 and 1.4.2 have solutions and these solutions are unique.*

In the remainder of this section, we will prove this proposition.

We define $P : X_h(\Omega) \to X_h^*(\Omega)$ as

$$P(u_h)\phi_h := a_{h,r}\big(u_h, \phi_h\big) + b\big(u_h, \phi_h\big) - f_{h,r}\big(\phi_h\big). \tag{1.4.5}$$

We would like to use the following consequence of the Brouwer theorem [44, 69]:

Theorem 1.4.4. *Let $P : X \to X^*$ be a continuous function on a finite-dimensional normed real vector space X, such that for suitable $\rho > 0$ we have*

$$P(x)x \geq 0 \quad \forall \|x\| \geq \rho. \tag{1.4.6}$$

Then there exists $x \in X$ such that

$$P(x) = 0. \tag{1.4.7}$$

Also we would like to use the following result:

Lemma 1.4.5. *Let $\Omega \subset \mathbb{R}^d$ be bounded. Let $f \in \mathcal{C}^1(\mathbb{R})$, $g \in L_\infty(\Omega)$. Let $P : X_h(\Omega) \to X_h^*(\Omega)$ be defined as*

$$P(u_h)\phi_h := \int_\Omega g(x) f\big(u_h(x)\big)\phi_h(x) \, dx. \tag{1.4.8}$$

Then P is continuous.

For the sake of completeness, proofs of these results are presented in Appendix (see pages 194, 198).

1.4.3.1 Existence for CWOPSIP method

By definition (1.4.5)

$$P(u_h)u_h := a_{h,2}\big(u_h,u_h\big) + b\big(u_h,u_h\big) - f_{h,2}\big(u_h\big). \tag{1.4.9}$$

Then we have that

$$a_{h,2}\big(u_h,u_h\big) = \|u_h\|^2_{h,\Sigma_2}. \tag{1.4.10}$$

Using Schwarz inequality for $f_{h,2}(u_h)$, defined in equation (1.4.2), and then lemma 1.3.7 we get

$$\Big|\int_\Omega k_1(x)u_h(x)\,dx\Big| \le \|k_1\|_{L_2(\Omega)}\|u_h\|_{L_2(\Omega)} \le c\|k_1\|_{L_2(\Omega)}\|u_h\|_{h,\Sigma_r}, \tag{1.4.11}$$

and

$$\begin{aligned}\Big|\sum_{e\in\Gamma_D}\eta_{r,e}\int_e[\hat u]\cdot[u_h]\,ds\Big| &= \Big|\sum_{e\in\Gamma_D}\int_e\sqrt{\eta_{r,e}}[\hat u]\cdot\sqrt{\eta_{r,e}}[u_h]ds\Big| = \Big|\int_{\partial\Omega_D}\sqrt{\eta_{r,e}}[\hat u]\cdot\sqrt{\eta_{r,e}}[u_h]ds\Big|\\ &\le \|\sqrt{\eta_{r,e}}[\hat u]\|_{L_2(\partial\Omega_D)}\|\sqrt{\eta_{r,e}}[u_h]\|_{L_2(\partial\Omega_D)}\\ &= \sqrt{\sum_{e\in\Gamma_D}\eta_{r,e}\int_e[\hat u]^2\,ds}\sqrt{\sum_{e\in\Gamma_D}\eta_{r,e}\int_e[u_h]^2\,ds}\\ &\le \|\hat u\|_{h,\Sigma_r}\|u_h\|_{h,\Sigma_r}.\end{aligned} \tag{1.4.12}$$

Thus

$$-f_{h,2}\big(u_h\big) \ge -c(\hat u,k_1,h)\|u_h\|_{h,\Sigma_2}, \tag{1.4.13}$$

where

$$c(\hat u,k_1,h) = \tilde c\times(\|\hat u\|_{h,\Sigma_r} + \|k_1\|_{L_2(\Omega)}). \tag{1.4.14}$$

Then let $C := \max\{\|\hat v\|_{L_\infty(\Omega)}, \|\hat w\|_{L_\infty(\Omega)}\}$. We may decompose $b\big(u_h,u_h\big)$ to

$$\begin{aligned}b\big(u_h,u_h\big) &= \int_\Omega\Big(e^{u_h-\hat v} - e^{\hat w-u_h}\Big)u_h dx\\ &= \int_\Omega\Big(e^{u_h-\hat v} - e^{\hat w-u_h}\Big)u_h\mathbb{1}_{\{x\in\Omega:|u_h(x)|>C\}}dx\\ &\quad+\int_\Omega\Big(e^{u_h-\hat v} - e^{\hat w-u_h}\Big)u_h\mathbb{1}_{\{x\in\Omega:|u_h(x)|\le C\}}dx.\end{aligned} \tag{1.4.15}$$

The first integral is non-negative, and the latter we can estimate from below

$$\int_\Omega\Big(e^{u_h(x)-\hat v(x)} - e^{\hat w(x)-u_h(x)}\Big)u_h(x)\mathbb{1}_{\{x\in\Omega:|u_h(x)|\le C\}}(x)dx \ge -|\Omega|2e^{2C}C. \tag{1.4.16}$$

In conclusion, we may use these estimations to obtain

$$P(u_h)u_h \ge \|u_h\|^2_{h,\Sigma_2} - c_1\|u_h\|_{h,\Sigma_2} - c_2. \tag{1.4.17}$$

Note that constants c_i in this inequality depend on h. It is therefore clear that for $\|u_h\|_{h,\Sigma_2}$ large enough we have $P(u_h)u_h \ge 0$.

Still we must show that P is continuous. The proof is elementary, we present it for the sake of completeness. We decompose $P(u_h)$ into a sum $P(u_h) = P_a(u_h) + P_b(u_h) + P_f(u_h)$. We start with linear part, which we denote by P_a. By Schwarz inequality

$$\begin{aligned}|P_a(u_h)\phi_h| := |a_{h,2}(u_h, \phi_h)| = \Big| \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon \nabla u_{h,i} \cdot \nabla \phi_{h,i} \, dx + \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [u_h] \cdot [\phi_h] \, ds \Big| \\ \leq \|u_h\|_{h,\Sigma_2} \|\phi_h\|_{h,\Sigma_2}.\end{aligned} \tag{1.4.18}$$

Thus

$$\|P_a(u_h)\| = \sup_{\|\phi_h\|_{h,\Sigma_2}=1} |P_a(u_h)\phi_h| \leq \sup_{\|\phi_h\|_{h,\Sigma_2}=1} \|u_h\|_{h,\Sigma_2} \|\phi_h\|_{h,\Sigma_2} = \|u_h\|_{h,\Sigma_2}, \tag{1.4.19}$$

so P_a is bounded and thus it is continuous. Then we have

$$P_f(u_h)\phi_h := -f_{h,2}(\phi_h), \tag{1.4.20}$$

which is trivially continuous, as it does not depend on u_h. Finally we have

$$\begin{aligned}P_b(u_h)\phi_h := b(u_h, \phi_h) &= \int_{\Omega} \left(e^{u_h(x) - \hat{v}(x)} - e^{\hat{w}(x) - u_h(x)} \right) \phi_h(x) \, dx \\ &= \int_{\Omega} e^{u_h(x) - \hat{v}(x)} \phi_h(x) \, dx - \int_{\Omega} e^{\hat{w}(x) - u_h(x)} \phi_h(x) \, dx.\end{aligned} \tag{1.4.21}$$

P_b is not linear, so we use lemma 1.4.5 with $f(x) := e^x$, $g(x) := e^{-\hat{v}(x)}$ and $f(x) := e^{-x}$, $g(x) := e^{\hat{w}(x)}$. Conditions of the lemma are then satisfied as f is smooth and $\hat{v}, \hat{w} \in L_\infty(\Omega)$ due to assumptions A7. Thus P_b is continuous.

Then by theorem 1.4.4 we have that there exists some u_h^*, such that $P(u_h^*) = 0$.

1.4.3.2 Existence for CSIPG method

We proceed analogously to the CWOPSIP case. For $b(u_h, \phi_h)$ the argumentation exactly the same. Then $f_{h,1}(u_h)$ has one additional term, which may be estimated using lemma 1.3.1 and the trace inequality

$$\begin{aligned}\Big| \sum_{e \in \Gamma_D} \int_e \{\varepsilon \nabla u_h \cdot \nu\} [\hat{u}] \, ds \Big| &\leq \sum_{e \in \Gamma_D} \|\{\varepsilon \nabla u_h \cdot \nu\}\|_{L_2(e)} \|[\hat{u}]\|_{L_2(e)} \\ &\leq c\varepsilon_M \sum_{i=1}^{N} h_i^{-1/2} \|\nabla u_h\|_{L_2(\Omega_i)} \|\hat{u}\|_{H^1(\Omega_i)} \\ &\leq C \|u_h\|_{h,\Sigma_1} \|\hat{u}\|_{H^1(\Omega)},\end{aligned} \tag{1.4.22}$$

where C depends on ε_M and h.

Therefore

$$-f_{h,1}(u_h) \geq -c(\hat{u}, k_1, h) \|u_h\|_{h,\Sigma_1}. \tag{1.4.23}$$

Then estimating $a_{h,1}(u_h, u_h)$ by lemma 1.3.5 with $\alpha = 1/2$ (cf. assumption A7), we have

$$P(u_h)u_h \geq \frac{1}{2} \|u_h\|_{h,\Sigma_1}^2 - c_1 \|u_h\|_{h,\Sigma_1} - c_2. \tag{1.4.24}$$

Finally we show that P is continuous. For $P_b(u_h)\phi_h := b(u_h, \phi_h)$ and $P_f(u_h)\phi_h := f_{h,1}(u_h, \phi_h)$ we use analogous argumentation as for CWOPSIP. Still we have to show that the linear operator $P_a(u_h)\phi_h := a_{h,1}(u_h, \phi_h)$ is continuous.

Using lemma 1.3.5 with $\alpha = 1/2$, lemma 1.3.8 and the Schwarz inequality we get

$$\begin{aligned}|a_{h,1}(u_h, \phi_h)| &= \Big| \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon \nabla u_{h,i} \cdot \nabla \phi_{h,i}\, dx - \sum_{e \in \Gamma_{DI}} \int_e \{\varepsilon \nabla u_h \cdot \nu\}[\phi_h]\, ds \\ &\quad - \sum_{e \in \Gamma_{DI}} \int_e \{\varepsilon \nabla \phi_h \cdot \nu\}[u_h]\, ds + \sum_{e \in \Gamma_{DI}} \eta_{1,e} \int_e [u_h] \cdot [\phi_h]\, ds \Big| \\ &= \Big| \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon \nabla u_{h,i} \cdot \nabla \phi_{h,i}\, dx + \sum_{e \in \Gamma_{DI}} \eta_{1,e} \int_e [u_h] \cdot [\phi_h]\, ds \\ &\quad - \sum_{e \in \Gamma_{DI}} \int_e \{\varepsilon \nabla u_h \cdot \nu\}[\phi_h]\, ds - \sum_{e \in \Gamma_{DI}} \int_e \{\varepsilon \nabla \phi_h \cdot \nu\}[u_h]\, ds \Big| \\ &\leq 2 \|u_h\|_{h,\Sigma_1} \|\phi_h\|_{h,\Sigma_1}.\end{aligned} \tag{1.4.25}$$

Thus P_a is bounded and continuous.

The existence of u_h^* is now proven.

1.4.3.3 Uniqueness

The uniqueness can be shown by contradiction for both cases. Assume that there exist two solutions $u_h^*, u_h^\dagger \in X_h(\Omega)$ of problem 1.4.1 or problem 1.4.2. Thus we have

$$\begin{aligned} a_{h,r}(u_h^*, \phi_h) + b(u_h^*, \phi_h) &= f_{h,r}(\phi_h), \\ a_{h,r}(u_h^\dagger, \phi_h) + b(u_h^\dagger, \phi_h) &= f_{h,r}(\phi_h). \end{aligned} \tag{1.4.26}$$

Then by taking $\phi_h := u_h^* - u_h^\dagger$ and subtracting these equations we obtain

$$a_{h,r}(u_h^* - u_h^\dagger, u_h^* - u_h^\dagger) = b(u_h^\dagger, u_h^* - u_h^\dagger) - b(u_h^*, u_h^* - u_h^\dagger). \tag{1.4.27}$$

Thus expanding operator b

$$\begin{aligned} a_{h,r}(u_h^* - u_h^\dagger, u_h^* - u_h^\dagger) = &\sum_{i=1}^{N} \int_{\Omega_i} e^{-\hat{v}} \left(e^{u_h^\dagger} - e^{u_h^*}\right) \left(u_h^* - u_h^\dagger\right) dx \\ &+ \sum_{i=1}^{N} \int_{\Omega_i} e^{\hat{w}} \left(e^{-u_h^*} - e^{-u_h^\dagger}\right) \left(u_h^* - u_h^\dagger\right) dx. \end{aligned} \tag{1.4.28}$$

By monotonicity of the exponential function, the right hand side is nonpositive.

Therefore for CWOPSIP we simply have

$$0 < \|u_h^* - u_h^\dagger\|_{h,\Sigma_2}^2 = a_{h,2}(u_h^* - u_h^\dagger, u_h^* - u_h^\dagger) \leq 0, \tag{1.4.29}$$

while for CSIPG we use lemma 1.3.5 (cf. assumption A7)

$$0 < \frac{1}{2} \|u_h^* - u_h^\dagger\|_{h,\Sigma_1}^2 \leq a_{h,1}(u_h^* - u_h^\dagger, u_h^* - u_h^\dagger) \leq 0. \tag{1.4.30}$$

Thus $0 < \|u_h^* - u_h^\dagger\|_{h,\Sigma_r}^2 \leq 0$ and we have the contradiction since $u_h^* \neq u_h^\dagger$. Therefore the uniqueness is proven.

1.5 Interpolation operator and interpolation error

In this section we would like to discuss interpolation error in discrete spaces $X_h(\Omega)$. While these results are not specific to our problem, they are technical tools required by our convergence study in sections 1.6 and 1.7.

First let us take any $\Omega_i \in \mathcal{E}$ and let us define interpolation operator $I_{h_i} : H^2(\Omega_i) \to X_{h_i}(\Omega_i) \subset \mathcal{C}^0(\overline{\Omega}_i)$ as follows

$$\forall x \in \mathcal{N}_{h_i} \qquad I_{h_i} u_i(x) = u_i(x). \tag{1.5.1}$$

Note that for $\Omega_i \subset \mathbb{R}^d, d \in \{1,2\}$ we have $H^2(\Omega_i) \subset \mathcal{C}^0(\Omega_i)$ (see [94]), so this definition is not ambiguous. Then we define $I_h : H^2(\mathcal{E}) \to X_h$ by

$$\forall \Omega_i \in \mathcal{E} \quad I_h u\Big|_{\Omega_i} := I_{h_i} u_i. \tag{1.5.2}$$

For convenience, we define

$$u_I := I_h u, \qquad u_I^* := I_h u^*. \tag{1.5.3}$$

We would like to establish an estimate on $\|u - u_I\|_{h,\Sigma_r}$, $r \in \{1,2\}$, where $u \in H^2(\mathcal{E})$. We consider $\Omega \subset \mathbb{R}^d$ for $d \in \{1,2\}$.

Theorem 1.5.1. *Under assumptions A1 to A7, let $u \in H^1(\Omega) \cap H^2(\mathcal{E})$, $\Omega \subset \mathbb{R}^d$ be a given function and let the interpolation operator $I_h : H^2(\mathcal{E}) \to X_h$ be defined as in equation (1.5.2). The following interpolation error estimates hold:*

- *If $d = 1$, $r \in \{1,2\}$*

$$\|u - I_h u\|^2_{h,\Sigma_r} \le C h^2 \sum_{\Omega_i \in \mathcal{E}} |u|^2_{H^2(\Omega_i)}. \tag{1.5.4}$$

- *If $d = 2$*

$$\begin{aligned} \|u - u_I\|^2_{h,\Sigma_1} &\le C \sum_{\Omega_i \in \mathcal{E}} \left(h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j} \right) |u|^2_{H^2(\Omega_i)}, \\ \|u - u_I\|^2_{h,\Sigma_2} &\le C \sum_{\Omega_i \in \mathcal{E}} \left(h_i + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j^2} \right) |u|^2_{H^2(\Omega_i)}. \end{aligned} \tag{1.5.5}$$

Note that this estimate is poor if $h_i/h_j \to \infty$ as $h \to 0$ for some adjacent $\Omega_i, \Omega_j \in \mathcal{E}$. Such a situation may occur if for example meshes $\mathcal{T}_{i,h_i}(\Omega_i)$ and $\mathcal{T}_{j,h_j}(\Omega_j)$ increase density disproportionately with h. However if we increase destiny proportionally, i.e. $h_i := c_i h$, we obtain the following estimates.

Remark 1.5.2. *If in theorem 1.5.1 we additionally assume that $h_i := c_i h$ for $i \in \{1, \ldots, N\}$, then for $d = 2$ estimates may be improved to*

$$\begin{aligned} \|u - u_I\|^2_{h,\Sigma_1} &\le C h^2 \sum_{\Omega_i \in \mathcal{E}} |u|^2_{H^2(\Omega_i)}, \\ \|u - u_I\|^2_{h,\Sigma_2} &\le C h \sum_{\Omega_i \in \mathcal{E}} |u|^2_{H^2(\Omega_i)}. \end{aligned} \tag{1.5.6}$$

The remaining of this section is devoted to proofs of theorem 1.5.1 and remark 1.5.2. We start with the following proposition.

Proposition 1.5.3. *Let $u_i \in H^2(\Omega_i)$ be a given function, where $\Omega_i \in \mathcal{E}$. Then the following estimates on the interpolation error in $X_{h_i}(\Omega_i)$ hold*

$$\begin{aligned} \|u_i - I_{h_i} u_i\|_{L_2(\Omega_i)} &\leq C h_i^2 |u_i|_{H^2(\Omega_i)}, \\ |u_i - I_{h_i} u_i|_{H^1(\Omega_i)} &\leq C h_i |u_i|_{H^2(\Omega_i)}, \end{aligned} \tag{1.5.7}$$

where the interpolation operator I_{h_i} defined as in (1.5.1).

Proof. By definition (see equation (1.3.4)), $X_{h_i}(\Omega_i)$ is a continuous linear finite element space. Thus this proposition is a direct consequence of corollary 1.1.13.

□

1.5.1 One dimension

We start the proof of theorem 1.5.1 in one dimension. This part is much simpler as the two-dimensional case, as then u and $I_h u$ are continuous on interfaces.

We see that for $u_I := I_h u$

$$\begin{aligned} \|u - u_I\|_h^2 &= \sum_{i=1}^N \int_{\Omega_i} \varepsilon(x) \Big(\nabla u_i - \nabla I_{h_i} u_i \Big)^2 \, dx + \sum_{e \in \Gamma_{DI}} \eta_{r,e} \int_e [u - I_h u]^2 \, ds \\ &\leq \varepsilon_M \sum_{\Omega_i \in \mathcal{E}} \|u_i - I_{h_i} u_i\|_{H^1(\Omega_i)}^2 + \sum_{e \in \Gamma_{DI}} \eta_{r,e} \int_e [u - I_h u]^2 \, ds \\ &\leq \varepsilon_M \|u - u_I\|_{H^1(\Omega)}^2 + \sum_{e \in \Gamma_{DI}} \eta_{r,e} \int_e [u - I_h u]^2 \, ds, \end{aligned} \tag{1.5.8}$$

for $u \in H^1(\Omega) \cap H^2(\mathcal{E})$. In this case, by proposition 1.5.3 used for every $\Omega_i \in \mathcal{E}$, we have

$$\|u - u_I\|_{H^1(\mathcal{E})}^2 \leq C h_i^2 \sum_{\Omega_i \in \mathcal{E}} |u|_{H^2(\Omega_i)}^2 \leq C h^2 \sum_{\Omega_i \in \mathcal{E}} |u|_{H^2(\Omega_i)}^2. \tag{1.5.9}$$

Then for any $e \in \Gamma_{DI}$, $e \subset \partial \Omega_i$ is a real number, so we have

$$\int_e u - I_h u \, ds = u(e) - I_h u(e) = u(e) - u(e) = 0. \tag{1.5.10}$$

Thus the latter element of (1.5.8) is zero for any $u \in H^1(\mathcal{E})$. Therefore we can estimate

$$\|u - u_I\|_{h,\Sigma_r}^2 \leq c \varepsilon_M \|u - u_I\|_{H^1(\Omega)}^2 \leq C \varepsilon_M h^2 \sum_{\Omega_i \in \mathcal{E}} |u|_{H^2(\Omega_i)}^2. \tag{1.5.11}$$

1.5.2 Two dimensions

The two-dimensional case is more problematic, as then even if u is continuous, $I_h u$ may be discontinuous across interfaces (see figure 1.1). Due to this fact, we start jointly, but then we have to separate cases $r = 1$ (CSIPG) and $r = 2$ (CWOPSIP).

For any $\Omega_i \in \mathcal{E}$ proposition 1.5.3 yields that

$$\|u - I_h u\|_{H^1(\mathcal{E})}^2 = \sum_{\Omega_i \in \mathcal{E}} \|u_i - I_{h_i} u_i\|_{H^1(\Omega_i)}^2 \leq \sum_{\Omega_i \in \mathcal{E}} C_i^2 h_i^2 |u_i|_{H^2(\Omega_i)}^2. \tag{1.5.12}$$

Therefore

$$\|u - u_I\|_{H^1(\mathcal{E})}^2 \le C \sum_{\Omega_i \in \mathcal{E}} h_i^2 |u_i|_{H^2(\Omega_i)}^2. \tag{1.5.13}$$

Note that

$$\|u - u_I\|_{h,\Sigma_r}^2 = \sum_{\Omega_i \in \mathcal{E}} \int_{\Omega_i} \varepsilon \Big(\nabla u - \nabla u_I \Big)^2 dx + \sum_{e \in \Gamma_{DI}} \eta_{r,e} \int_e [u - u_I]^2 \, ds, \tag{1.5.14}$$

so the first element of this sum is estimated and we have to estimate the latter.

Then let us take any $e \in \Gamma_{DI}$. For $e \in \Gamma_I$ we assume that $e = \partial\Omega_i \cap \partial\Omega_j$ for some $\Omega_i, \Omega_j \in \mathcal{E}$ and we have

$$\begin{aligned} \int_e [u - u_I]^2 \, ds = \|[u - u_I]\|_{L_2(e)}^2 &= \Big\| u_i - u_{I,i} - \big(u_j - u_{I,j}\big) \Big\|_{L_2(e)}^2 \\ &\le 2\|u_i - u_{I,i}\|_{L_2(e)}^2 + 2\|u_j - u_{I,j}\|_{L_2(e)}^2, \end{aligned} \tag{1.5.15}$$

while for $e \in \Gamma_D$ we have $e \subset \partial\Omega_i$ for some $\Omega_i \in \mathcal{E}$ and simply

$$\int_e [u - u_I]^2 \, ds = \int_e \Big(u_i - u_{I,i} \Big)^2 ds = \|u_i - u_{I,i}\|_{L_2(e)}^2. \tag{1.5.16}$$

Therefore it is sufficient to estimate $\|u_i - u_{I,i}\|_{L_2(e)}^2$ for any $e \in \Gamma_{DI}$, $e \subset \partial\Omega_i$. First using the trace theorem for finite element functions (proposition 1.1.11), taking into account that $\mathcal{T}_{h_i}(\Omega_i)$ is a quasi-uniform mesh (assumptions A2), and then error estimates of proposition 1.5.3, we get

$$\begin{aligned} \|u_i - u_{I,i}\|_{L_2(e)}^2 &\le C h_i^{-1} \Big(\|u_i - u_{I,i}\|_{L_2(\tau)}^2 + h_i^2 |u_i - u_{I,i}|_{H^1(\tau)}^2 \Big) \\ &\le c h_i^{-1} \Big(h_i^4 |u_i|_{H^2(\Omega_i)}^2 + h_i^4 |u_i|_{H^2(\Omega_i)}^2 \Big) = C h_i^3 |u_i|_{H^2(\Omega_i)}^2. \end{aligned} \tag{1.5.17}$$

To proceed further, we need to consider different form of $\eta_{r,e}$. Therefore we will distinguish two cases: CSIPG ($r = 1$) and CWOPSIP ($r = 2$).

1.5.2.1 CSIPG

In this case, we have by definition (1.3.12)

$$\eta_{1,e} = \sigma_e \{h^{-1}\}. \tag{1.5.18}$$

Therefore if $e \in \Gamma_D$, we simply have

$$\begin{aligned} \eta_{1,e} \int_e \Big(u_i - u_{I,i} \Big)^2 ds = \eta_{1,e} \|u_i - u_{I,i}\|_{L_2(e)}^2 &= \sigma_e h_i^{-1} \|u_i - u_{I,i}\|_{L_2(e)}^2 \\ &\le C\sigma_e h_i^{-1} h_i^3 |u|_{H^2(\Omega_i)}^2 = C\sigma_e h_i^2 |u|_{H^2(\Omega_i)}^2. \end{aligned} \tag{1.5.19}$$

On the other hand, if $e \in \Gamma_I$ then

$$\begin{aligned} \eta_{1,e} \int_e \Big(u_i - u_{I,i} \Big)^2 ds = \eta_{1,e} \|u_i - u_{I,i}\|_{L_2(e)}^2 &= 0.5\sigma_e (h_i^{-1} + h_j^{-1}) \|u_i - u_{I,i}\|_{L_2(e)}^2 \\ &\le C\sigma_e (h_i^{-1} + h_j^{-1}) h_i^3 |u|_{H^2(\Omega_i)}^2 = C\sigma_e \Big(h_i^2 + \frac{h_i^3}{h_j} \Big) |u|_{H^2(\Omega_i)}^2. \end{aligned} \tag{1.5.20}$$

Then if we sum up over $e \in \Gamma_{DI}$

$$\sum_{e\in\Gamma_{DI}} \eta_{1,e} \int_e \Big([u-u_I]\Big)^2 ds \le \sum_{\Omega_i\in\mathcal{E}} C\left(h_i^2 + \sum_{\Omega_j\in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j}\right)|u|^2_{H^2(\Omega_i)}. \tag{1.5.21}$$

Thus taking into account this estimate and the estimate for $H^1(\Omega)$ interpolation error (1.5.12), we get

$$\begin{aligned}\|u-u_I\|^2_{h,\Sigma_1} &= \sum_{\Omega_i\in\mathcal{E}} \varepsilon \int_{\Omega_i} \Big(\nabla u - u_I\Big)^2 dx + \sum_{e\in\Gamma_{DI}} \eta_{1,e} \int_e \Big([u-u_I]\Big)^2 ds \\ &\le C \sum_{\Omega_i\in\mathcal{E}} \left(h_i^2 + \sum_{\Omega_j\in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j}\right)|u|^2_{H^2(\Omega_i)}.\end{aligned} \tag{1.5.22}$$

Constants σ_e are included in constant C, as they do not change with h. If we increase destiny proportionally ($h_i := c_i h$), the result is as follows

$$\|u-u_I\|^2_{h,\Sigma_1} \le Ch^2 \sum_{\Omega_i\in\mathcal{E}} |u|^2_{H^2(\Omega_i)}. \tag{1.5.23}$$

1.5.2.2 CWOPSIP

In this case, we have by definition (1.3.12)

$$\eta_{2,e} = \sigma_e\{h^{-2}\}. \tag{1.5.24}$$

We proceed as in previous case, but with increased penalty term. For $e \in \Gamma_D$

$$\begin{aligned}\eta_{2,e} \int_e \Big(u_i - u_{I,i}\Big)^2 ds &= \sigma_e h_i^{-2} \|u_i - u_{I,i}\|^2_{L_2(e)} \\ &\le C\sigma_e h_i^{-2} h_i^3 |u|^2_{H^2(\Omega_i)} = C\sigma_e h_i |u|^2_{H^2(\Omega_i)},\end{aligned} \tag{1.5.25}$$

and for $e \in \Gamma_I$

$$\begin{aligned}\eta_{2,e} \int_e \Big(u_i - u_{I,i}\Big)^2 ds &= 0.5\sigma_e (h_i^{-2} + h_j^{-2}) \|u_i - u_{I,i}\|^2_{L_2(e)} \\ &\le C\sigma_e (h_i^{-2} + h_j^{-2}) h_i^3 |u|^2_{H^2(\Omega_i)} = C\sigma_e \Big(h_i + \frac{h_i^3}{h_j^2}\Big)|u|^2_{H^2(\Omega_i)}.\end{aligned} \tag{1.5.26}$$

Finally taking into account (1.5.12)

$$\begin{aligned}\|u-u_I\|^2_{h,\Sigma_2} &= \sum_{\Omega_i\in\mathcal{E}} \varepsilon \int_{\Omega_i} \Big(\nabla(u - u_I)\Big)^2 dx + \sum_{e\in\Gamma_{DI}} \eta_{2,e} \int_e \Big([u-u_I]\Big)^2 ds \\ &\le C \sum_{\Omega_i\in\mathcal{E}} \left(h_i^2 + h_i + \sum_{\Omega_j\in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j^2}\right)|u|^2_{H^2(\Omega_i)} \\ &\le C \sum_{\Omega_i\in\mathcal{E}} \left(h_i + \sum_{\Omega_j\in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j^2}\right)|u|^2_{H^2(\Omega_i)},\end{aligned} \tag{1.5.27}$$

as $0 < h_i \le 1$.

Assuming additionally that $h_i := c_i h$ we obtain

$$\|u-u_I\|^2_{h,\Sigma_2} \le Ch \sum_{\Omega_i\in\mathcal{E}} |u|^2_{H^2(\Omega_i)}. \tag{1.5.28}$$

Thus this estimate is weaker than for CSIPG.

1.6 Error estimates for the equilibrium case for CWOPSIP

In this section we will present the analysis of discretization of the equilibrium state (problem 1.2.2) with Composite Weakly Over-Penalized Interior Penalty method (see problem 1.4.1). We would like to prove the following theorem

Theorem 1.6.1. *Under assumptions A1 to A7, let $u^* \in H^1(\Omega) \cap H^2(\mathcal{E})$ be a solution of differential problem 1.2.2 and let $u^* \in X_h(\Omega)$ be solutions of problem 1.4.1. Then the following error estimate holds:*

$$\|u^* - u_h^*\|_{h,\Sigma_2} \leq C\left(\sum_{i=1}^{N}\Big(h_i + \sum_{\Omega_k \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_k^2}\Big)|u_{h,i}^*|^2_{H^2(\Omega_i)}\right)^{1/2} + Ch\|u^*\|_{H^2(\mathcal{E})}. \tag{1.6.1}$$

Remark 1.6.2. *If additionally we assume that mesh parameters are proportional, i.e. $h_i := c_i h$ for every $\Omega_i \in \mathcal{E}$, then estimate (1.6.1) reduces to*

$$\|u^* - u_h^*\|_{h,\Sigma_2} \leq Ch^{1/2}\Big(\sum_{i=1}^{N} |u_{h,i}^*|^2_{H^2(\Omega_i)}\Big)^{1/2} + Ch\|u^*\|_{H^2(\mathcal{E})}. \tag{1.6.2}$$

Remark 1.6.3. *If $\Omega \subset \mathbb{R}$, the estimate of remark 1.6.2 can be improved to*

$$\|u^* - u_h^*\|_{h,\Sigma_2} \leq Ch\|u^*\|_{H^2(\mathcal{E})}. \tag{1.6.3}$$

In the remainder of this section we prove these results.

1.6.1 Outline of the proof

The general idea standing behind this proof is as follows. We base on fact, as noted in section 1.3.2.2, that the broken norm may be expressed in terms of $a_{h,2}$, i.e.

$$\|u_h\|^2_{h,\Sigma_2} = a_{h,2}(u_h, u_h). \tag{1.6.4}$$

Our goal is to estimate the discrete solution error by an interpolation error.

Let us assume, for the purpose of this sketch, that the differential problem may be written as

$$a(u^*, \phi) = f(\phi), \tag{1.6.5}$$

and the discrete problems

$$a(u_h^*, \phi_h) = f(\phi_h), \tag{1.6.6}$$

where $u^*, \phi \in V$, $u_h^*, \phi_h \in V_h$ and $V_h \subset V$. We also assume that a is bilinear and elliptic and f is linear and that

$$a(u, u) = \|u\|^2, \tag{1.6.7}$$

for an appropriate norm. Example of such a problem is presented in section 1.1.1.

Then we could proceed as follows

1. Take $\phi := \phi_h := I_h u^* - u_h^* = u_I^* - u_h^*$ and subtract (1.6.6) from (1.6.5) to get

$$a(u^* - u_h^*, u_I^* - u_h^*) = 0. \tag{1.6.8}$$

2. Transform this result to

$$a(u_I^* - u_h^*, u_I^* - u_h^*) = a(u^* - u_I^*, u_h^* - u_I^*). \tag{1.6.9}$$

3. Use boundedness property of a on the right-hand side to get

$$\|u_I^* - u_h^*\|^2 = a(u_I^* - u_h^*, u_I^* - u_h^*) = a(u^* - u_I^*, u_h^* - u_I^*) \leq C\|u^* - u_I^*\|\|u_h^* - u_I^*\|. \tag{1.6.10}$$

4. Divide by $\|u_h^* - u_I^*\|$ to obtain the estimate

$$\|u_I^* - u_h^*\| \leq C\|u^* - u_I^*\|. \tag{1.6.11}$$

5. Use this estimate and the triangle inequality to get

$$\|u^* - u_h^*\| \leq \|u^* - u_I^*\| + \|u_I^* - u_h^*\| \leq (1 + C)\|u^* - u_I^*\|. \tag{1.6.12}$$

In our case, discrete spaces $X_h(\Omega)$ do not lie in the differential problem space $H^1(\Omega)$, so it is not feasible to take $I_h u^* - u_h^*$ as a test function. Also discrete problems and the differential problem do not share the same forms. To overcome these problems, we formulate another differential problem 1.6.4, which is consistent with problem 1.2.4 under additional assumptions (section 1.6.2). At the cost of the regularity of solutions, problem 1.6.4 accounts for more general test functions.

At this moment, it is possible to get result similar to (1.6.8). Still the right-hand side will not be zero due to differences in discrete/differential forms, and it must be estimated as well in (1.6.10) along with ellipticity property of a. These estimates are demonstrated in section 1.6.3. The remaining steps are performed in section 1.6.4.

1.6.2 Consistency

Our starting point is the problem defined in section 1.2.2:

Problem 1.2.2. *Let $\hat{v}, \hat{w} \in L_\infty(\Omega)$ and $k_1 \in L_2(\Omega)$ be given. Find $u^* \in \hat{u} + H^1(\Omega)$, such that*

$$a(u^*, \phi) + b(u^*, \phi) = f(\phi) \quad \forall \phi \in H^1_{0,\partial\Omega_D}(\Omega), \tag{1.2.4}$$

where

$$\begin{aligned} a(u, \phi) &:= \int_\Omega \varepsilon(x)\nabla u(x) \cdot \nabla \phi(x)\, dx, \\ b(u, \phi) &:= \int_\Omega \left(e^{u(x)-\hat{v}(x)} - e^{\hat{w}(x)-u(x)}\right)\phi(x)\, dx, \\ f(\phi) &:= \int_\Omega k_1(x)\phi(x)\, dx. \end{aligned} \tag{1.2.5}$$

For convenience, we define the following operators

$$
\begin{aligned}
A(u,\phi) &:= \sum_{\Omega_i\in\mathcal{E}} \int_{\Omega_i} \varepsilon\nabla u\cdot\nabla\phi\,dx,\\
B(u,\phi) &:= \sum_{\Omega_i\in\mathcal{E}} \int_{\Omega_i} u\phi\,dx,\\
C(\phi) &:= \sum_{\Omega_i\in\mathcal{E}} \int_{\Omega_i} k_1\phi\,dx,\\
D(u,\phi) &:= -\sum_{e\in\Gamma_{DI}} \int_e \{\varepsilon\nabla u\cdot\nu\}[\phi]\,ds,\\
F(\phi) &:= -\sum_{e\in\Gamma_D} \int_e \{\varepsilon\nabla\phi\cdot\nu\}[\hat{u}]\,ds,\\
I_r(\phi) &:= \sum_{e\in\Gamma_D} \eta_{r,e}\int_e [\hat{u}]\cdot[\phi]\,ds,\\
J_r(u,\phi) &:= \sum_{e\in\Gamma_{DI}} \eta_{r,e}\int_e [u]\cdot[\phi]\,ds.
\end{aligned}
\tag{1.6.13}
$$

Note that for any $u_h \in X_h(\Omega)$, $r\in\{1,2\}$

$$
\|u_h\|^2_{h,\Sigma_r} = A(u_h,u_h) + J_r(u_h,u_h). \tag{1.6.14}
$$

Using operators defined above, we may rewrite this problem as: find $u^*\in H^1(\Omega)$ such that

$$
\begin{aligned}
A(u^*,\phi) + B(e^{u^*-\hat{v}},\phi) - B(e^{\hat{w}-u^*},\phi) = C(\phi) \quad \forall\phi\in H^1_{0,\partial\Omega_D}(\Omega),\\
u^* = \hat{u} \quad \text{on } \partial\Omega.
\end{aligned}
\tag{1.6.15}
$$

Second problem is a nonlinear variant of abstract problem 1.3.9 with $r=2$:

Problem 1.6.4. *Find $u^*\in H^2(\mathcal{E})$, such that $\forall\phi\in H^1(\mathcal{E})\cap H^2(\mathcal{T}_h)$*

$$
A(u^*,\phi) + B(e^{u^*-\hat{v}},\phi) - B(e^{\hat{w}-u^*},\phi) + D(u^*,\phi) + J_2(u^*,\phi) = C(\phi) + I_2(\phi). \tag{1.6.16}
$$

As in general case covered by theorem 1.3.10, if the solution is sufficiently regular these problems are consistent, i.e.

Theorem 1.6.5. *Under assumptions A3, A6, A7, if u^* is a solution of problem 1.2.2 then u^* is a solution of problem 1.6.4. Conversely, if u^* is a solution of problem 1.6.4 then it is also a solution of problem 1.2.2.*

Proof. We use theorem 1.3.10 with $r:=2$ and $f:=k_1-e^{u^*-\hat{v}}+e^{\hat{w}-u^*}$. By assumptions A7 we have $u^*\in H^1(\Omega)\cap H^2(\mathcal{E})$ and $\varepsilon|_{\Omega_i}\in C^1(\overline{\Omega_i})$. They imply $\varepsilon\nabla u^*\in H^1(\mathcal{E})$. Theorem 1.3.10 then concludes the result. □

1.6.3 Analysis

We would like to estimate a difference between solutions of two problems. A solution u^* of differential problem 1.6.4 satisfies $\forall\phi\in H^1(\mathcal{E})\cap H^2(\mathcal{T}_h)$

$$
A(u^*,\phi) + B(e^{u^*-\hat{v}},\phi) - B(e^{\hat{w}-u^*},\phi) + D(u^*,\phi) + J_2(u^*,\phi) = C(\phi) + I_2(\phi). \tag{1.6.17}
$$

On the other hand, a solution u_h^* of discrete problem 1.4.1 depending on parameter h satisfies $\forall \phi_h \in X_h$

$$A(u_h^*, \phi_h) + B(e^{u_h^* - \hat{v}}, \phi_h) - B(e^{\hat{w} - u_h^*}, \phi_h) + J_2(u_h^*, \phi_h) = C(\phi_h) + I_2(\phi_h). \tag{1.6.18}$$

We take

$$\phi := \phi_h := u_I^* - u_h^*, \tag{1.6.19}$$

where $u_I^* := I_h u^*$, and then we subtract equation (1.6.18) from equation (1.6.17). We obtain

$$\begin{aligned} A(u^* - u_h^*, u_I^* - u_h^*) + B(e^{u^* - \hat{v}} - e^{u_h^* - \hat{v}}, u_I^* - u_h^*) - B(e^{\hat{w} - u^*} - e^{\hat{w} - u_h^*}, u_I^* - u_h^*) \\ + J_2(u^* - u_h^*, u_I^* - u_h^*) = -D(u^*, u_I^* - u_h^*). \end{aligned} \tag{1.6.20}$$

We will estimate every element of this equation.

As accordance with assumptions A7, in this section $u^* \in H^1(\Omega) \cap H^2(\mathcal{E})$.

1.6.3.1 Estimate of $A(u^* - u_h^*, \phi_h)$

We have

$$A(\underline{u^* - u_h^*}, u_I^* - u_h^*) = A(\underline{u^* - u_I^* + u_I^* - u_h^*}, u_I^* - u_h^*) = A(u^* - u_I^*, u_I^* - u_h^*) + A(u_I^* - u_h^*, u_I^* - u_h^*). \tag{1.6.21}$$

By the Schwarz inequality, we have

$$|A(u^* - u_I^*, u_I^* - u_h^*)| = \left| \int_\Omega \varepsilon \nabla (u^* - u_I^*) \cdot \nabla (u_I^* - u_h^*) \, dx \right| \leq \|u^* - u_I^*\|_{h,\Sigma_r} \|u_I^* - u_h^*\|_{h,\Sigma_r}. \tag{1.6.22}$$

1.6.3.2 Estimate of $B(e^{u^* - \hat{v}} - e^{u_h^* - \hat{v}}, \phi_h)$

Similarly as in the previous case, we have

$$\begin{aligned} B(e^{u^* - \hat{v}} - e^{u_h^* - \hat{v}}, u_I^* - u_h^*) =& B(\underline{e^{u^* - \hat{v}} - e^{u_I^* - \hat{v}} + e^{u_I^* - \hat{v}} - e^{u_h^* - \hat{v}}}, u_I^* - u_h^*) \\ =& B(\underline{e^{u_I^* - \hat{v}} - e^{u_h^* - \hat{v}}}, u_I^* - u_h^*) + B(\underline{e^{u^* - \hat{v}} - e^{u_I^* - \hat{v}}}, u_I^* - u_h^*) \\ =& B(e^{-\hat{v}} \left[e^{u_I^*} - e^{u_h^*} \right], u_I^* - u_h^*) + B(e^{-\hat{v}} \left[e^{u^*} - e^{u_I^*} \right], u_I^* - u_h^*). \end{aligned} \tag{1.6.23}$$

First element of this sum is positive due to the definition of B and monotonicity of exponential function. Note that the solution of the differential problem u^* is bounded (lemma 1.2.3), also $\hat{v}, \hat{w} \in L_\infty(\Omega)$. Due to our definition of interpolation operator, u_I^* is also bounded by the same values as u^*. Then let us define L_e to be a Lipschitz constant of exponential function in sufficiently large bounded set, i.e

$$\begin{aligned} L_e := \inf \Big\{ L > 0 : |e^x - e^y| \leq L|x - y| \quad \forall x, y \in [-M, M], \\ \text{where } M := \|u^*\|_{L_\infty(\Omega)} + \|\hat{v}\|_{L_\infty(\Omega)} + \|\hat{w}\|_{L_\infty(\Omega)} \Big\}. \end{aligned} \tag{1.6.24}$$

This definition allows for a bigger set than necessary, as we would like to use the constant L_e also in sections to follow. Then we may estimate

$$\begin{aligned} B(e^{u^* - \hat{v}} - e^{u_I^* - \hat{v}}, u_I^* - u_h^*) \leq& B(|e^{u^* - \hat{v}} - e^{u_I^* - \hat{v}}|, |u_I^* - u_h^*|) \\ \leq& L_e B \left(|u^* - \hat{v} - u_I^* + \hat{v}|, |u_I^* - u_h^*| \right) \\ =& L_e B \left(|u^* - u_I^*|, |u_I^* - u_h^*| \right) \\ \leq& L_e \|u^* - u_I^*\|_{L_2(\Omega)} \|u_I^* - u_h^*\|_{L_2(\Omega)}. \end{aligned} \tag{1.6.25}$$

1.6.3.3 Estimate of $-B(e^{\hat{w}-u^*} - e^{\hat{w}-u_h^*}, \phi_h)$

As in previous section

$$\begin{aligned}
-B(e^{\hat{w}-u^*} - e^{\hat{w}-u_h^*}, u_I^* - u_h^*) =& - B(\underline{e^{\hat{w}-u^*} - e^{\hat{w}-u_I^*} + e^{\hat{w}-u_I^*} - e^{\hat{w}-u_h^*}}, u_I^* - u_h^*) \\
=& - B(\underline{e^{\hat{w}-u^*} - e^{\hat{w}-u_I^*}}, u_I^* - u_h^*) - B(\underline{e^{\hat{w}-u_I^*} - e^{\hat{w}-u_h^*}}, u_I^* - u_h^*) \quad (1.6.26) \\
=& B(e^{\hat{w}}[e^{-u_I^*} - e^{-u^*}], u_I^* - u_h^*) + B(e^{\hat{w}}[e^{-u_h^*} - e^{-u_I^*}], u_I^* - u_h^*).
\end{aligned}$$

Second element of this sum is positive due to the definition of B and anti-monotonicity of function $\exp(-x)$. First element we may estimate as

$$\begin{aligned}
B(e^{\hat{w}}\left[e^{-u_I^*} - e^{-u^*}\right], u_I^* - u_h^*) \leq& L_e B\left(|u^* - u_I^*|, |u_I^* - u_h^*|\right) \\
\leq& L_e \|u^* - u_I^*\|_{L_2(\Omega)} \|u_I^* - u_h^*\|_{L_2(\Omega)}.
\end{aligned} \quad (1.6.27)$$

1.6.3.4 Estimate of $-D(u^*, \phi_h)$

This element is in some way special, as it is absent in the discrete formulation. To estimate it, we use Schwarz inequality.

$$\begin{aligned}
\left|D(u^*, u_I^* - u_h^*)\right| =& \left|\sum_{e\in\Gamma_{DI}} \int_e \{\varepsilon\nabla u^*\cdot\nu\}[u_I^* - u_h^*]\, ds\right| \leq \sum_{e\in\Gamma_{DI}} \int_e |\{\varepsilon\nabla u^*\cdot\nu\}[u_I^* - u_h^*]|\, ds \\
=& \sum_{e\in\Gamma_{DI}} \int_e \left|\left(\eta_{2,e}^{-1/2}\{\varepsilon\nabla u^*\cdot\nu\}\right)\left(\eta_{2,e}^{1/2}[u_I^* - u_h^*]\right)\right| ds \\
\leq& \left[\sum_{e\in\Gamma_{DI}} \int_e \eta_{2,e}^{-1}\{\varepsilon\nabla u^*\cdot\nu\}^2\, ds\right]^{1/2} \left[\sum_{e\in\Gamma_{DI}} \int_e \eta_{2,e}\left([u_I^* - u_h^*]\right)^2 ds\right]^{1/2} \\
\leq& \left[\sum_{e\in\Gamma_D} \int_e 0.5\sigma_e^{-1}h_i^2\{\varepsilon\nabla u^*\cdot\nu\}^2\, ds + \sum_{e\in\Gamma_I} \int_e \sigma_e^{-1}\frac{1}{h_i^{-2}+h_j^{-2}}\{\varepsilon\nabla u^*\cdot\nu\}^2\, ds\right]^{1/2} \|u_I^* - u_h^*\|_{h,\Sigma_2} \\
\leq& \left[\sum_{e\in\Gamma_D} \int_e 0.5\sigma_e^{-1}h_i^2\{\varepsilon\nabla u^*\cdot\nu\}^2\, ds + \sum_{e\in\Gamma_I} \int_e \sigma_e^{-1}\left(h_i^2+h_j^2\right)\{\varepsilon\nabla u^*\cdot\nu\}^2\, ds\right]^{1/2} \|u_I^* - u_h^*\|_{h,\Sigma_2} \quad (1.6.28) \\
\leq& \left[\sum_{\Omega_i\in\mathcal{E}} h_i^2 \sum_{e\in\Gamma_{DI}\cap\Gamma_i} \sigma_e^{-1}\int_e \{\varepsilon\nabla u^*\cdot\nu\}^2\, ds\right]^{1/2} \|u_I^* - u_h^*\|_{h,\Sigma_2} \\
\leq& h\left[\sum_{\Omega_i\in\mathcal{E}} \sum_{e\in\Gamma_{DI}\cap\Gamma_i} \sigma_e^{-1}\int_e \{\varepsilon\nabla u^*\cdot\nu\}^2\, ds\right]^{1/2} \|u_I^* - u_h^*\|_{h,\Sigma_2} \\
\leq& Ch\left[\sum_{e\in\Gamma_{DI}} \sigma_e^{-1}\int_e \{\varepsilon\nabla u^*\cdot\nu\}^2\, ds\right]^{1/2} \|u_I^* - u_h^*\|_{h,\Sigma_2},
\end{aligned}$$

where the constant C depends on maximal number of edges of coarse mesh elements $\mathcal{E}$.

Then by trace inequality we have

$$\begin{aligned}\left|\sum_{e\in\Gamma_{DI}}\sigma_e^{-1}\int_e\{\varepsilon\nabla u^*\cdot\nu\}^2\,ds\right| &\le \sigma_m^{-1}\sum_{e\in\Gamma_{DI}}\|\{\varepsilon\nabla u^*\cdot\nu\}\|^2_{L_2(e)}\,ds\\ &\le \sigma_m^{-1}\varepsilon_M\left(\sum_{e\in\Gamma_D}\|\{\nabla u^*\cdot\nu\}\|^2_{L_2(e)}+\sum_{e\in\Gamma_I}0.5\sum_{\{\Omega_i:e\in\Gamma_i\}}\|\nabla u^*\cdot\nu|_{\Omega_i}\|^2_{L_2(e)}\right)\\ &\le C\sigma_m^{-1}\varepsilon_M\sum_{\Omega_i\in\mathcal{E}}\|u^*\|^2_{H^2(\Omega_i)}=C\sigma_m^{-1}\varepsilon_M\|u^*\|^2_{H^2(\mathcal{E})}.\end{aligned}\tag{1.6.29}$$

Therefore we have

$$\left|D(u^*,u_I^*-u_h^*)\right|\le Ch\|u^*\|_{H^2(\mathcal{E})}\|u_I^*-u_h^*\|_{h,\Sigma_2}.\tag{1.6.30}$$

1.6.3.5 Estimate of $J_r(u^*-u_h^*,\phi_h)$

Taking $r=2$ has no advantage here, so we consider general case. We have

$$\begin{aligned}J_r(u^*-u_h^*,u_I^*-u_h^*)=&J_r(u^*-u_I^*+u_I^*-u_h^*,u_I^*-u_h^*)\\ =&J_r(u^*-u_I^*,u_I^*-u_h^*)+J_r(u_I^*-u_h^*,u_I^*-u_h^*).\end{aligned}\tag{1.6.31}$$

Second element is nonnegative, and first one we may estimate as

$$\begin{aligned}\left|J_r(u^*-u_I^*,u_I^*-u_h^*)\right|=&\Big|\sum_{e\in\Gamma_{DI}}\int_e\eta_{r,e}[u^*-u_I^*][u_I^*-u_h^*]\Big|\\ \le&\left(\sum_{e\in\Gamma_{DI}}\int_e\eta_{r,e}[u^*-u_I^*]^2\right)^{1/2}\cdot\left(\sum_{e\in\Gamma_{DI}}\int_e\eta_{r,e}[u_I^*-u_h^*]^2\right)^{1/2}\\ \le&\|u^*-u_I^*\|_{h,\Sigma_r}\|u_I^*-u_h^*\|_{h,\Sigma_r}.\end{aligned}\tag{1.6.32}$$

1.6.4 Summary

Using results of the previous subsections, we may rewrite equation

$$\begin{aligned}A(u^*-u_h^*,u_I^*-u_h^*)+B(e^{u^*-\hat v}-e^{u_h^*-\hat v},u_I^*-u_h^*)-B(e^{\hat w-u^*}-e^{\hat w-u_h^*},u_I^*-u_h^*)\\ +J_2(u^*-u_h^*,u_I^*-u_h^*)=-D(u^*,u_I^*-u_h^*),\end{aligned}\tag{1.6.33}$$

as

$$\text{LHS}=\text{RHS},\tag{1.6.34}$$

where

$$\begin{aligned}\text{LHS}=&A(u_I^*-u_h^*,u_I^*-u_h^*)+B(e^{-\hat v}\left[e^{u_I^*}-e^{u_h^*}\right],u_I^*-u_h^*)+B(e^{\hat w}[e^{-u_h^*}-e^{-u_I^*}],u_I^*-u_h^*)\\ &+J_2(u_I^*-u_h^*,u_I^*-u_h^*),\\ \text{RHS}=&-A(u^*-u_I^*,u_I^*-u_h^*)-B(e^{-\hat v}\left[e^{u^*}-e^{u_I^*}\right],u_I^*-u_h^*)-B(e^{\hat w}\left[e^{-u_I^*}-e^{-u^*}\right],u_I^*-u_h^*)\\ &-J_2(u^*-u_I^*,u_I^*-u_h^*)-D(u^*,u_I^*-u_h^*).\end{aligned}\tag{1.6.35}$$

We can estimate LHS from below by omitting nonnegative elements with operator B (cf. (1.6.23), and (1.6.26)) to obtain

$$\text{LHS}\ge A(u_I^*-u_h^*,u_I^*-u_h^*)+J_2(u_I^*-u_h^*,u_I^*-u_h^*)=\|u_I^*-u_h^*\|^2_{h,\Sigma_2}.\tag{1.6.36}$$

On the other hand, for RHS we may use estimates (1.6.22), (1.6.25), (1.6.27), (1.6.30) and (1.6.32) of this section to get

$$\begin{aligned}\text{RHS} \leq & \|u^* - u_I^*\|_{h,\Sigma_2} \|u_I^* - u_h^*\|_{h,\Sigma_2} + L_e \|u^* - u_I^*\|_{L_2(\Omega)} \|u_I^* - u_h^*\|_{L_2(\Omega)} \\ & + L_e \|u^* - u_I^*\|_{L_2(\Omega)} \|u_I^* - u_h^*\|_{L_2(\Omega)} + \|u^* - u_I^*\|_{h,\Sigma_2} \|u_I^* - u_h^*\|_{h,\Sigma_2} \\ & + Ch \|u^*\|_{H^2(\mathcal{E})} \|u_I^* - u_h^*\|_{h,\Sigma_2}.\end{aligned} \tag{1.6.37}$$

Therefore using LHS = RHS we obtain

$$\begin{aligned}\|u_I^* - u_h^*\|_{h,\Sigma_2}^2 \leq & 2\|u^* - u_I^*\|_{h,\Sigma_2} \|u_I^* - u_h^*\|_{h,\Sigma_2} + 2L_e \|u^* - u_I^*\|_{L_2(\Omega)} \|u_I^* - u_h^*\|_{L_2(\Omega)} \\ & + Ch \|u^*\|_{H^2(\mathcal{E})} \|u_I^* - u_h^*\|_{h,\Sigma_2}.\end{aligned} \tag{1.6.38}$$

Then if we divide both sides of this inequality by $\|u_I^* - u_h^*\|_{h,\Sigma_2} > 0$ we will have

$$\|u_I^* - u_h^*\|_{h,\Sigma_2} \leq 2\|u^* - u_I^*\|_{h,\Sigma_2} + 2L_e \|u^* - u_I^*\|_{L_2(\Omega)} \frac{\|u_I^* - u_h^*\|_{L_2(\Omega)}}{\|u_I^* - u_h^*\|_{h,\Sigma_2}} + Ch \|u^*\|_{H^2(\mathcal{E})}. \tag{1.6.39}$$

Then by lemma 1.3.7, we have $\|u_I^* - u_h^*\|_{L_2(\Omega)} \leq c \|u_I^* - u_h^*\|_{h,\Sigma_2}$ as $X_h(\Omega) \subset H^1(\mathcal{E})$, so

$$\frac{\|u_I^* - u_h^*\|_{L_2(\Omega)}}{\|u_I^* - u_h^*\|_{h,\Sigma_2}} \leq \frac{c\|u_I^* - u_h^*\|_{h,\Sigma_2}}{\|u_I^* - u_h^*\|_{h,\Sigma_2}} = c. \tag{1.6.40}$$

Thus

$$\begin{aligned}\|u_I^* - u_h^*\|_{h,\Sigma_2} \leq & 2\|u^* - u_I^*\|_{h,\Sigma_2} + 2L_e \|u^* - u_I^*\|_{L_2(\Omega)} \frac{\|u_I^* - u_h^*\|_{L_2(\Omega)}}{\|u_I^* - u_h^*\|_{h,\Sigma_2}} + Ch \|u^*\|_{H^2(\mathcal{E})} \\ \leq & 2\|u^* - u_I^*\|_{h,\Sigma_2} + 2cL_e \|u^* - u_I^*\|_{H^1(\Omega)} + Ch \|u^*\|_{H^2(\mathcal{E})} \\ \leq & (2 + 2cL_e) \|u^* - u_I^*\|_{h,\Sigma_2} + Ch \|u^*\|_{H^2(\mathcal{E})}.\end{aligned} \tag{1.6.41}$$

Therefore we obtain for some constant C

$$\|u_I^* - u_h^*\|_{h,\Sigma_2} \leq C\Big(\|u^* - u_I^*\|_{h,\Sigma_2} + h\|u^*\|_{H^2(\mathcal{E})}\Big). \tag{1.6.42}$$

On the other hand, by the triangle inequality we have

$$\|u^* - u_h^*\|_{h,\Sigma_2} \leq \|u^* - u_I^*\|_{h,\Sigma_2} + \|u_I^* - u_h^*\|_{h,\Sigma_2}. \tag{1.6.43}$$

Then we get in general in two dimensions for $u^* \in H^2(\mathcal{E})$ the estimate

$$\begin{aligned}\|u^* - u_h^*\|_{h,\Sigma_2} & \leq \|u^* - u_I^*\|_{h,\Sigma_2} + \|u_I^* - u_h^*\|_{h,\Sigma_2} \\ & \leq \|u^* - u_I^*\|_{h,\Sigma_2} + C\Big(\|u^* - u_I^*\|_{h,\Sigma_2} + h\|u^*\|_{H^2(\mathcal{E})}\Big) \\ & \leq (1 + C)\|u^* - u_I^*\|_{h,\Sigma_2} + Ch \\ & \leq c\left(\sum_{\Omega_i \in \mathcal{E}} \Big(h_i + \sum_{\Omega_j \in \text{nb}(\Omega_i)} \frac{h_i^3}{h_j^2}\Big) |u^*|_{H^2(\Omega_i)}^2\right)^{1/2} + Ch\|u^*\|_{H^2(\mathcal{E})}.\end{aligned} \tag{1.6.44}$$

With the additional assumption $h_i := c_i h$ for every $\Omega_i \in \mathcal{E}$ we can simplify this expression to

$$\|u^* - u_h^*\|_{h,\Sigma_2} \leq C\Big(h^{1/2}|u^*|_{H^2(\mathcal{E})} + h\|u^*\|_{H^2(\mathcal{E})}\Big). \tag{1.6.45}$$

In one dimension, this estimate may be improved due to better interpolation error (1.5.11)

$$\|u^* - u_h^*\|_{h,\Sigma_2} \leq Ch\|u^*\|_{H^2(\mathcal{E})}. \tag{1.6.46}$$

1.7 Error estimates for the equilibrium case for CSIPG

In this section we will present the analysis of discretization of the drift-diffusion system 1.2.1 with the Composite Symmetric Interior Penalty Galerkin method [37]. Our analysis is similar to the CWOPSIP case.

Theorem 1.7.1. *Under assumptions A1 to A7, let $u^* \in H^1(\Omega) \cap H^2(\mathcal{E})$ be a solution of differential problem 1.2.2 and let $u^* \in X_h(\Omega)$ be solutions of problem 1.4.2. Then the following error estimate holds:*

$$\|u^* - u_h^*\|_{h,\Sigma_1} \leq C \sum_{\Omega_i \in \mathcal{E}} \left(\left[h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j} \right]^{1/2} \right) |u^*|_{H^2(\Omega_i)}. \tag{1.7.1}$$

Remark 1.7.2. *If additionally we assume that mesh parameters are proportional, i.e. $h_i := c_i h$ for every $\Omega_i \in \mathcal{E}$, then estimate (1.7.1) reduces to*

$$\|u^* - u_h^*\|_{h,\Sigma_1} \leq Ch \sum_{\Omega_i \in \mathcal{E}} |u^*|_{H^2(\Omega_i)}. \tag{1.7.2}$$

The estimate for CSIPG is therefore optimal.

The rest of this section is devoted to proof of these estimates. The proof is analogous as in CWOPSIP case. It follows the outline presented in section 1.6.1, with the consistency result concluded in theorem 1.7.4. Additionally lemma 1.3.5 is used as analogue of equation (1.6.10). We will use operators A, B, C, D, F, I_1, J_1, introduced in section 1.6.2 and some estimates of section 1.6.3.

1.7.1 Consistency

In a similar manner as before, we would like to relate the following problems. First is already defined:

Problem 1.2.2. *Let $\hat{v}, \hat{w} \in L_\infty(\Omega)$ and $k_1 \in L_2(\Omega)$ be given. Find $u^* \in \hat{u} + H^1(\Omega)$, such that*

$$a(u^*, \phi) + b(u^*, \phi) = f(\phi) \quad \forall \phi \in H^1_{0,\partial\Omega_D}(\Omega), \tag{1.2.4}$$

where

$$\begin{aligned} a(u, \phi) &:= \int_\Omega \varepsilon(x) \nabla u(x) \cdot \nabla \phi(x)\, dx, \\ b(u, \phi) &:= \int_\Omega \left(e^{u(x) - \hat{v}(x)} - e^{\hat{w}(x) - u(x)} \right) \phi(x)\, dx, \\ f(\phi) &:= \int_\Omega k_1(x) \phi(x)\, dx. \end{aligned} \tag{1.2.5}$$

Second problem is specific to CSIPG:

Problem 1.7.3. *Find $u^* \in H^2(\mathcal{E})$, such that $\forall \phi \in H^1(\mathcal{E}) \cap H^2(\mathcal{T}_h)$*

$$A(u^*, \phi) + B(e^{u^* - \hat{v}}, \phi) - B(e^{\hat{w} - u^*}, \phi) + D(u^*, \phi) + D(\phi, u^*) + J_1(u^*, \phi) = C(\phi) + F(\phi) + I_1(\phi). \tag{1.7.3}$$

As for CWOPSIP method, we introduce on the following consistency result:

Theorem 1.7.4. *Under assumptions A3, A6, A7, if u^* is a solution of problem 1.2.2 then u^* is a solution of problem 1.7.3. Conversely, if u^* is a solution of problem 1.7.3 then it is also a solution of problem 1.2.2.*

Proof. The result follows directly from theorem 1.3.10 with $r = 1$ and $f := k_1 - e^{u^* - \hat{v}} + e^{\hat{w} - u^*}$. ☐

1.7.2 Analysis

The differential problem satisfies: for every $\phi \in H^1(\mathcal{E}) \cap H^2(\mathcal{T}_h)$

$$A(u^*,\phi)+B(e^{u^*-\hat{v}},\phi)-B(e^{\hat{w}-u^*},\phi)+D(u^*,\phi)+D(\phi,u^*)+J_1(u^*,\phi)=C(\phi)+F(\phi)+I_1(\phi). \quad (1.7.4)$$

On the other hand, the family of discrete problems depending on parameter h is defined as: for every $\phi_h \in X_h$

$$\begin{aligned} A(u_h^*,\phi_h)+B(e^{u_h^*-\hat{v}},\phi_h)-B(e^{\hat{w}-u_h^*},\phi_h)+D(u_h^*,\phi_h)& \\ +D(\phi_h,u_h^*)+J_1(u_h^*,\phi_h)=C(\phi_h)+F(\phi_h)+I_1(\phi_h).& \end{aligned} \quad (1.7.5)$$

We proceed as in CWOPSIP case. As before, we subtract these equations from each other taking

$$\phi := \phi_h := u_I^* - u_h^*. \quad (1.7.6)$$

C, F and I vanish, as they depend on the test function only, and we get

$$\begin{aligned} A(u^*-u_h^*,u_I^*-u_h^*)+B(e^{u^*-\hat{v}}-e^{u_h^*-\hat{v}},u_I^*-u_h^*)-B(e^{\hat{w}-u^*}-e^{\hat{w}-u_h^*},u_I^*-u_h^*)& \\ +D(u^*-u_h^*,u_I^*-u_h^*)+D(u_I^*-u_h^*,u^*-u_h^*)& \\ +J_1(u^*-u_h^*,u_I^*-u_h^*)=0.& \end{aligned} \quad (1.7.7)$$

We will discuss every element of the resulting equation separately.

For elements $A(u^*-u_h^*,\phi_h)$, $B(e^{u^*-\hat{v}}-e^{u_h^*-\hat{v}},\phi_h)$, $-B(e^{\hat{w}-u^*}-e^{\hat{w}-u_h^*},\phi_h)$, and $J_1(u^*-u_h^*,\phi_h)$ we use estimates established already in section 1.6.3.

1.7.2.1 Estimate of $D(u^*-u_h^*,\phi_h)$

Here we have

$$D(u^*-u_h^*,u_I^*-u_h^*) = D(u^*-u_I^*,u_I^*-u_h^*)+D(u_I^*-u_h^*,u_I^*-u_h^*). \quad (1.7.8)$$

We start with the second element of this sum

$$D(u_I^*-u_h^*,u_I^*-u_h^*) = -\sum_{e\in\Gamma_{DI}} \int_e \Big\{\varepsilon\nabla(u_I^*-u_h^*)\cdot\nu\Big\}[u_I^*-u_h^*]\,ds. \quad (1.7.9)$$

Using lemma 1.3.5 with $\alpha = 1/2$ (cf. assumption A7) we obtain

$$-\sum_{e\in\Gamma_{DI}} \int_e \Big\{\varepsilon\nabla(u_I^*-u_h^*)\cdot\nu\Big\}[u_I^*-u_h^*]\,ds \geq -\frac{1}{4}\|u_I^*-u_h^*\|_{h,\Sigma_1}^2. \quad (1.7.10)$$

Then we have

$$D(u^*-u_I^*,u_I^*-u_h^*) = -\sum_{e\in\Gamma_{DI}} \int_e \Big\{\varepsilon\nabla(u^*-u_I^*)\cdot\nu\Big\}[u_I^*-u_h^*]\,ds. \quad (1.7.11)$$

Let us take any $e \in \Gamma_I$, $e \in \partial\Omega_i \cap \partial\Omega_j$. Then

$$\begin{aligned} \int_e \Big\{\varepsilon\nabla(u^*-u_I^*)\cdot\nu\Big\}[u_I^*-u_h^*]\,ds &\leq \varepsilon_M \|\{\nabla(u^*-u_I^*)\cdot\nu\}\|_{L_2(e)}\|[u_I^*-u_h^*]\|_{L_2(e)} \\ &\leq \varepsilon_M \eta_{1,e}^{-1/2}\|\{\nabla(u^*-u_I^*)\cdot\nu\}\|_{L_2(e)}\eta_{1,e}^{1/2}\|[u_I^*-u_h^*]\|_{L_2(e)}. \end{aligned} \quad (1.7.12)$$

Next we have

$$\begin{aligned}\|\{\nabla(u^*-u_I^*)\}\|_{L_2(e)} &= \left\| \frac{\nabla(u^*-u_I^*)\cdot\nu}{2}\Big|_{\Omega_i} + \frac{\nabla(u^*-u_I^*)\cdot\nu}{2}\Big|_{\Omega_j} \right\|_{L_2(e)} \\ &\le \left\| \frac{\nabla(u^*-u_I^*)\cdot\nu}{2}\Big|_{\Omega_i} \right\|_{L_2(e)} + \left\| \frac{\nabla(u^*-u_I^*)\cdot\nu}{2}\Big|_{\Omega_j} \right\|_{L_2(e)}. \end{aligned} \tag{1.7.13}$$

For any $\Omega_i \in \mathcal{E}$, $e \subset \partial\Omega_i$, we can estimate $\|\nabla(u^*-u_I^*)\cdot\nu\big|_{\Omega_i}\|_{L_2(e)}$ in the following manner. Let us define

$$\mathcal{T}_{i,e} = \{\tau \in \mathcal{T} : \tau \subset \Omega_i, |e \cap \partial\tau| > 0\}. \tag{1.7.14}$$

Then using proposition 1.1.11 on every $\tau \in \mathcal{T}_{i,e}$ and taking into account assumption A2 we obtain

$$\begin{aligned}\left\| \nabla u^*\cdot\nu - \nabla u_I^*\cdot\nu \Big|_{\Omega_i} \right\|_{L_2(e)}^2 &= \sum_{\tau\in\mathcal{T}_{i,e}} \left\| \nabla u^*\cdot\nu - \nabla u_I^*\cdot\nu \Big|_{\Omega_i} \right\|_{L_2(e\cap\partial\tau)}^2 \\ &\le Ch_i^{-1} \sum_{\tau\in\mathcal{T}_{i,e}} \big(|u^*-u_I^*|_{H^1(\tau)}^2 + h_i^2|u^*-u_I^*|_{H^2(\tau)}^2\big). \end{aligned} \tag{1.7.15}$$

Note that also since $u_I^*\big|_\tau \in \mathbb{P}_1(\tau)$, then $\nabla^2 u_I^*\big|_\tau \equiv 0$ and

$$|u^*-u_I^*|_{H^2(\tau)} = |u^*|_{H^2(\tau)}. \tag{1.7.16}$$

With aid of proposition 1.5.3 we obtain

$$\begin{aligned}\left\| \nabla u^*\cdot\nu - \nabla u_I^*\cdot\nu \Big|_{\Omega_i} \right\|_{L_2(e)}^2 &\le Ch_i^{-1} \sum_{\tau\in\mathcal{T}_{i,e}} \big(|u^*-u_I^*|_{H^1(\tau)}^2 + h_i^2|u^*-u_I^*|_{H^2(\tau)}^2\big) \\ &= Ch_i^{-1} \sum_{\tau\in\mathcal{T}_{i,e}} \big(|u^*-u_I^*|_{H^1(\tau)}^2 + h_i^2|u^*|_{H^2(\tau)}^2\big) \\ &= Ch_i^{-1}\big(|u^*-u_I^*|_{H^1(\Omega_i)}^2 + h_i^2|u^*|_{H^2(\Omega_i)}^2\big) \\ &\le Ch_i^{-1}\big(h_i^2|u^*|_{H^2(\Omega_i)}^2 + h_i^2|u^*|_{H^2(\Omega_i)}^2\big) \\ &= 2Ch_i|u^*|_{H^2(\Omega_i)}^2. \end{aligned} \tag{1.7.17}$$

Therefore

$$\begin{aligned}\left\| \{\nabla(u^*-u_I^*)\cdot\nu\} \right\|_{L_2(e)} &\le \left\| \frac{\nabla(u^*-u_I^*)\cdot\nu}{2}\Big|_{\Omega_i} \right\|_{L_2(e)} + \left\| \frac{\nabla(u^*-u_I^*)\cdot\nu}{2}\Big|_{\Omega_j} \right\|_{L_2(e)} \\ &\le c(h_i^{1/2}|u^*|_{H^2(\Omega_i)} + h_j^{1/2}|u^*|_{H^2(\Omega_j)}) \\ &\le c(h_i^{1/2}+h_j^{1/2})(|u^*|_{H^2(\Omega_i)} + |u^*|_{H^2(\Omega_j)}) \\ &\le 2c(h_i+h_j)^{1/2}(|u^*|_{H^2(\Omega_i)} + |u^*|_{H^2(\Omega_j)}). \end{aligned} \tag{1.7.18}$$

Thus we have

$$\begin{aligned}\eta_{1,e}^{-1}\|\{\nabla(u^*-u_I^*)\}\|_{L_2(e)}^2 &= \sigma_e^{-1}2\big(h_i^{-1}+h_j^{-1}\big)^{-1}\|\{\nabla(u^*-u_I^*)\}\|_{L_2(e)}^2 \\ &\le \sigma_e^{-1}2\big(h_i^{-1}+h_j^{-1}\big)^{-1}4c(h_i+h_j)(|u^*|_{H^2(\Omega_i)} + |u^*|_{H^2(\Omega_j)})^2 \\ &= C\sigma_e^{-1}\frac{h_ih_j}{h_i+h_j}(h_i+h_j)(|u^*|_{H^2(\Omega_i)} + |u^*|_{H^2(\Omega_j)})^2 \\ &= C\sigma_e^{-1}h_ih_j(|u^*|_{H^2(\Omega_i)} + |u^*|_{H^2(\Omega_j)})^2. \end{aligned} \tag{1.7.19}$$

If $e \in \Gamma_D$, $e \in \partial\Omega_i$, then analogously

$$\eta_{1,e}^{-1}\|\{\nabla(u^* - u_I^*)\}\|_{L_2(e)}^2 \leq C\sigma_e^{-1}h_i^2|u^*|_{H^2(\Omega_i)}^2. \tag{1.7.20}$$

Therefore by Cauchy-Schwarz inequality and the inequalities derived above

$$\begin{aligned}
\big|D(u^* - u_I^*, u_I^* - u_h^*)\big| &= \sum_{e\in\Gamma_{DI}} \int_e \Big\{\varepsilon\nabla(u^* - u_I^*)\Big\}[u_I^* - u_h^*]\, ds \\
&\leq \varepsilon_M \sum_{e\in\Gamma_{DI}} \eta_{1,e}^{-1/2}\|\{\nabla(u^* - u_I^*)\}\|_{L_2(e)}\eta_{1,e}^{1/2}\|[u_I^* - u_h^*]\|_{L_2(e)} \\
&\leq \varepsilon_M \Big(\sum_{e\in\Gamma_{DI}} \eta_{1,e}^{-1}\|\{\nabla(u^* - u_I^*)\}\|_{L_2(e)}^2\Big)^{1/2}\Big(\sum_{e\in\Gamma_{DI}} \eta_{1,e}\|[u_I^* - u_h^*]\|_{L_2(e)}^2\Big)^{1/2} \\
&\leq \varepsilon_M \Big(\sum_{e\in\Gamma_{DI}} \eta_{1,e}^{-1}\|\{\nabla(u^* - u_I^*)\}\|_{L_2(e)}^2\Big)^{1/2}\|u_I^* - u_h^*\|_{h,\Sigma_1} \\
&\leq \varepsilon_M \Big(\sum_{e\in\Gamma_{DI}} C\sigma_e^{-1}h_ih_j(|u^*|_{H^2(\Omega_i)} + |u^*|_{H^2(\Omega_j)})^2\Big)^{1/2}\|u_I^* - u_h^*\|_{h,\Sigma_1} \\
&\leq C\varepsilon_M \Big(\sum_{e\in\Gamma_{DI}} C\sigma_e^{-1}h_ih_j\Big)^{1/2}\sum_{\Omega_i\in\mathcal{E}} |u^*|_{H^2(\Omega_i)}\|u_I^* - u_h^*\|_{h,\Sigma_1} \\
&\leq C\varepsilon_M\sigma_m^{-1/2}h\sum_{\Omega_i\in\mathcal{E}} |u^*|_{H^2(\mathcal{E})}\|u_I^* - u_h^*\|_{h,\Sigma_1},
\end{aligned} \tag{1.7.21}$$

as number of elements of Γ_{DI} does not depend on h. Here by Ω_i,Ω_j we denote elements of $\mathcal{E}$ adjacent to $e \in \Gamma$, noting that $\Omega_i = \Omega_j$ if $e \subset \partial\Omega$.

1.7.2.2 Estimate of $D(\phi_h, u^* - u_h^*)$

As before

$$D(u_I^* - u_h^*, u^* - u_h^*) = D(u_I^* - u_h^*, u^* - u_I^*) + D(u_I^* - u_h^*, u_I^* - u_h^*). \tag{1.7.22}$$

We have

$$D(u_I^* - u_h^*, u_I^* - u_h^*) \geq -\frac{1}{4}\|u_I^* - u_h^*\|_{h,\Sigma_1}^2. \tag{1.7.23}$$

Then we have

$$\begin{aligned}
|D(u_I^* - u_h^*, u^* - u_I^*)| &\leq \sum_{e\in\Gamma_{DI}} \int_e \Big|\Big\{\varepsilon\nabla(u_I^* - u_h^*)\cdot\nu\Big\}\Big|\Big|[u^* - u_I^*]\Big|\, ds, \\
&\leq \varepsilon_M \sum_{e\in\Gamma_{DI}} \|\{\nabla(u_I^* - u_h^*)\cdot\nu\}\|_{L_2(e)}\|[u^* - u_I^*]\|_{L_2(e)}.
\end{aligned} \tag{1.7.24}$$

We proceed in a similar way as for $D(u^* - u_h^*, \phi_h)$. Thus splitting this sum up, we have

$$\Big\|\{\nabla(u_I^* - u_h^*)\cdot\nu\}\Big\|_{L_2(e)} \leq \Big\|\nabla(u_I^* - u_h^*)\cdot\nu\Big|_{\Omega_i}\Big\|_{L_2(e)} + \Big\|\nabla(u_I^* - u_h^*)\cdot\nu\Big|_{\Omega_j}\Big\|_{L_2(e)}. \tag{1.7.25}$$

Then using corollary 1.1.13 with assumption A2 we get

$$\begin{aligned}
\Big\|\nabla(u_I^* - u_h^*)\cdot\nu\Big|_{\Omega_i}\Big\|_{L_2(e)}^2 &\leq ch_i^{-1}\sum_{\tau\in\mathcal{T}_{i,e}} |u_I^* - u_h^*)|_{H^1(\tau)}^2 \\
&\leq ch_i^{-1}|u_I^* - u_h^*)|_{H^1(\Omega_i)}^2 \leq ch_i^{-1}\|u_I^* - u_h^*\|_{h,\Sigma_1}^2.
\end{aligned} \tag{1.7.26}$$

On the other hand we have

$$\|[u^* - u_I^*]\|_{L_2(e)} \leq \left\| u^* - u_I^*|_{\Omega_i} \right\|_{L_2(e)} + \left\| u^* - u_I^*|_{\Omega_j} \right\|_{L_2(e)}. \tag{1.7.27}$$

Then using proposition 1.1.11 with assumption A2 and proposition 1.5.3 we get

$$\begin{aligned}
\|u^* - u_I^*|_{\Omega_i}\|_{L_2(e)}^2 &= \sum_{\tau \in \mathcal{T}_{i,e}} \left\| u^* - u_I^*\Big|_{\Omega_i} \right\|_{L_2(e \cap \partial\tau)}^2 \\
&\leq c h_i^{-1} \sum_{\tau \in \mathcal{T}_{i,e}} \left(\|u^* - u_I^*\|_{L_2(\tau)}^2 + h_i^2 |u^* - u_I^*|_{H^1(\tau)}^2 \right) \\
&\leq c h_i^{-1} \left(\|u^* - u_I^*\|_{L_2(\Omega_i)}^2 + h_i^2 |u^* - u_I^*|_{H^1(\Omega_i)}^2 \right) \\
&\leq c h_i^{-1} \left(h_i^4 |u^*|_{H^2(\Omega_i)}^2 + h_i^4 |u^*|_{H^2(\Omega_i)}^2 \right) \\
&= 2c h_i^3 |u^*|_{H^2(\Omega_i)}^2
\end{aligned} \tag{1.7.28}$$

Thus for any $e = \partial\Omega_i \cap \partial\Omega_j \in \Gamma_I$

$$\begin{aligned}
&\|\{\varepsilon \nabla(u_I^* - u_h^*) \cdot \nu\}\|_{L_2(e)}^2 \|[u^* - u_I^*]\|_{L_2(e)}^2 \\
&\quad \leq C\varepsilon_M^2 (h_i^{-1} + h_j^{-1}) \|u_I^* - u_h^*\|_{h,\Sigma_1}^2 \left(h_i^3 |u^*|_{H^2(\Omega_i)}^2 + h_j^3 |u^*|_{H^2(\Omega_j)}^2 \right) \\
&\quad = C\varepsilon_M^2 \left[\left(h_i^2 + \frac{h_i^3}{h_j} \right) |u^*|_{H^2(\Omega_i)}^2 + \left(h_j^2 + \frac{h_j^3}{h_i} \right) |u^*|_{H^2(\Omega_j)}^2 \right] \|u_I^* - u_h^*\|_{h,\Sigma_1}^2.
\end{aligned} \tag{1.7.29}$$

So finally

$$\begin{aligned}
|D(u_I^* - u_h^*, u^* - u_I^*)|^2 &\leq C\varepsilon_M^2 \sum_{e \in \Gamma_{DI}} \|\{\nabla(u^* - u_I^*) \cdot \nu\}\|_{L_2(e)}^2 \|[u_I^* - u_h^*]\|_{L_2(e)}^2 \\
&\leq C\varepsilon_M^2 \sum_{\Omega_i \in \mathcal{E}} \left(h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j} \right) |u^*|_{H^2(\Omega_i)}^2 \|u_I^* - u_h^*\|_{h,\Sigma_1}^2,
\end{aligned} \tag{1.7.30}$$

and

$$|D(u_I^* - u_h^*, u^* - u_I^*)| \leq C\varepsilon_M \left[\sum_{\Omega_i \in \mathcal{E}} \left(h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j} \right) |u^*|_{H^2(\Omega_i)}^2 \right]^{1/2} \|u_I^* - u_h^*\|_{h,\Sigma_1}. \tag{1.7.31}$$

1.7.3 Summary

After re-arrangement of elements of (1.7.7) we obtain

$$LHS = RHS, \tag{1.7.32}$$

where

$$\begin{aligned}
LHS = {} & A(u_I^* - u_h^*, u_I^* - u_h^*) + B(e^{-\hat{v}} \left[e^{u_I^*} - e^{u_h^*} \right], u_I^* - u_h^*) + B(e^{\hat{w}}[e^{-u_h^*} - e^{-u_I^*}], u_I^* - u_h^*) \\
& + 2D(u_I^* - u_h^*, u_I^* - u_h^*) + J_1(u_I^* - u_h^*, u_I^* - u_h^*),
\end{aligned} \tag{1.7.33}$$

$$\begin{aligned}
RHS = {} & -A(u^* - u_I^*, u_I^* - u_h^*) - B(e^{-\hat{v}} \left[e^{u^*} - e^{u_I^*} \right], u_I^* - u_h^*) - B(e^{\hat{w}} \left[e^{-u_I^*} - e^{-u^*} \right], u_I^* - u_h^*) \\
& - D(u^* - u_I^*, u_I^* - u_h^*) - D(u_I^* - u_h^*, u^* - u_I^*) - J_1(u^* - u_I^*, u_I^* - u_h^*).
\end{aligned} \tag{1.7.34}$$

As we noted, the elements of LHS with the operator B are nonnegative (cf. (1.6.23), and (1.6.26)), and since

$$A(u_I^* - u_h^*, u_I^* - u_h^*) + J_1(u_I^* - u_h^*, u_I^* - u_h^*) = \|u_I^* - u_h^*\|_{h,\Sigma_1}^2, \tag{1.7.35}$$

then by lemma 1.3.5 with $\alpha = 1/2$ (cf. assumption A7)

$$\begin{aligned} LHS &\geq A(u_I^* - u_h^*, u_I^* - u_h^*) + 2D(u_I^* - u_h^*, u_I^* - u_h^*) + J_1(u_I^* - u_h^*, u_I^* - u_h^*) \\ &= \|u_I^* - u_h^*\|_{h,\Sigma_1}^2 + 2D(u_I^* - u_h^*, u_I^* - u_h^*) \\ &\geq \|u_I^* - u_h^*\|_{h,\Sigma_1}^2 - \frac{1}{2}\|u_I^* - u_h^*\|_{h,\Sigma_1}^2 = \frac{1}{2}\|u_I^* - u_h^*\|_{h,\Sigma_1}^2. \end{aligned} \tag{1.7.36}$$

On the other hand, using estimates (1.6.22), (1.6.25), (1.6.27), (1.6.32), (1.7.21) and (1.7.31) we obtain

$$\begin{aligned} |RHS| \leq \Bigg(&(2 + 2L_e)\|u^* - u_I^*\|_{h,\Sigma_1} + ch \sum_{\Omega_i \in \mathcal{E}} |u^*|_{H^2(\Omega_i)} \\ &+ c\Bigg[\sum_{\Omega_i \in \mathcal{E}} \Big(h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j}\Big)|u^*|_{H^2(\Omega_i)}^2\Bigg]^{1/2}\Bigg)\|u_I^* - u_h^*\|_{h,\Sigma_1} \\ &\leq C \sum_{\Omega_i \in \mathcal{E}} \Bigg[\Big(h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j}\Big)^{1/2} + h\Bigg]|u^*|_{H^2(\Omega_i)}. \end{aligned} \tag{1.7.37}$$

Thus

$$\frac{1}{2}\|u_I^* - u_h^*\|_{h,\Sigma_1}^2 \leq LHS = RHS \leq C \sum_{\Omega_i \in \mathcal{E}} \Bigg(\Big[h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j}\Big]^{1/2} + h\Bigg)|u^*|_{H^2(\Omega_i)}\|u_I^* - u_h^*\|_{h,\Sigma_1}. \tag{1.7.38}$$

Then dividing by $\frac{1}{2}\|u_i^* - u_h^*\|_{h,\Sigma_1}$ we finally have

$$\|u_I^* - u_h^*\|_{h,\Sigma_1} \leq C \sum_{\Omega_i \in \mathcal{E}} \Bigg(\Big[h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j}\Big]^{1/2} + h\Bigg)|u^*|_{H^2(\Omega_i)}. \tag{1.7.39}$$

Then we may estimate using this result and properties of the interpolation operator

$$\begin{aligned} \|u^* - u_h^*\|_{h,\Sigma_1} &\leq \|u^* - u_I^*\|_{h,\Sigma_1} + \|u_I^* - u_h^*\|_{h,\Sigma_1} \\ &\leq C \sum_{\Omega_i \in \mathcal{E}} \Bigg(\Big[h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j}\Big]^{1/2} + h\Bigg)|u^*|_{H^2(\Omega_i)} \\ &\leq C \sum_{\Omega_i \in \mathcal{E}} \Bigg(\Big[h_i^2 + \sum_{\Omega_j \in \mathrm{nb}(\Omega_i)} \frac{h_i^3}{h_j}\Big]^{1/2}\Bigg)|u^*|_{H^2(\Omega_i)}. \end{aligned} \tag{1.7.40}$$

If we assume that $h_i = c_i h$ for every $\Omega_i \in \mathcal{E}$, then this expression simplifies to

$$\|u^* - u_h^*\|_{h,\Sigma_1} \leq Ch \sum_{\Omega_i \in \mathcal{E}} |u^*|_{H^2(\Omega_i)}. \tag{1.7.41}$$

Chapter 2

Numerical simulations of semiconductor devices

Contents

2.1 Band structure of GaN, AlN and InN 70
2.1.1 Bandgap 70
2.1.2 Effective mass 71
2.1.3 Current 73
2.1.4 Carrier statistics 73
2.1.5 Doping 77
2.1.6 Energy distribution in a crystal structure 81
2.2 Properties of the mixed AlGaN and InGaN crystals 81
2.3 Geometry of luminescent semiconductor structures 83
2.3.1 p-n homojunction 83
2.3.2 Laser diodes and electroluminescent diodes 84
2.4 Quantum structures: wells and barriers 84
2.5 Drift-diffusion model 85
2.5.1 Conservation laws and equations of motion 85
2.5.2 Electric field, electrostatic potential and polarization effect 87
2.5.3 Differential problem 88
2.5.4 Equilibrium state and non-equilibrium state 89
2.5.5 Built-in potential 90
2.5.6 Boundary conditions 90
2.6 Radiative and non-radiative recombination 91
2.6.1 Standard recombination models 91
2.6.2 Impact ionization 97
2.6.3 Trap levels 99
2.7 Tunneling quantum effect 103
2.7.1 Trap-assisted tunneling 103
2.8 P-N diode 104
2.8.1 p-n homojunctions 105
2.8.2 Homojunctions, p-i-n diodes and single quantum well structures 109
2.8.3 Comparison with available software 121
2.8.4 Computing carrier currents 129
2.8.5 Trap-assisted tunneling effect on characteristics of gallium nitride diodes 133
2.9 Light-emitting diodes and laser diodes 138
2.9.1 Introduction 138
2.9.2 Aluminum content in EBL 140
2.9.3 Mg doping of p-type 142

2.9.4 Number of quantum wells 144
2.10 Optical excitation in quaternary alloy AlInGaN 145

In this chapter we discuss simulations of semiconductor devices with the drift-diffusion method, with particular emphasis on structures based on gallium nitride and its alloys with aluminum nitride and indium nitride. Due to nature of this model, we focus on the electrical properties of these devices.

This chapter is organized as follows. In sections 2.1 and 2.2 we discuss elementary physical properties of the semiconductor material. We also present the Vegard rule, which allows to estimate material parameters of alloys. Then in sections 2.3 and 2.4 elementary information on the luminescent semiconductor structures is given. The drift-diffusion model in formulation accounting for physical and material properties is discussed in detail in section 2.5. In sections 2.6 and 2.7.1 we present recombination channels important form the standpoint of these simulations, along with appropriate formulas to be used with the van Roosbroeck equations. To conclude, we present simulations of semiconductor structures in sections 2.8 to 2.10.

2.1 Band structure of GaN, AlN and InN

2.1.1 Bandgap

The Free Electron Model provides basic understanding of electronic phenomena in solid metals. This theory has however its limitations. In particular, it does not explain well all aspects of electron transport in the semiconductors. Thus this theory is generalized to take into account periodic crystal structures to the Near-Free Electron model. According to this model, electrons in crystal structures are distributed in the energy bands, which are separated by the bandgaps with no energy levels. If for some material in a given temperature all these bands are either full or empty, this material is an insulator, as no electron transport is possible. On the other hand, if any energy band is filled only partially, the material is an electric conductor.

Generally two bands are important in the conductance of the semiconductor material. In the absolute zero temperature, these bands are characterized as follows: a lower one, called a valence band, is the topmost fully occupied band and then the one right above, the first empty band, is called a conduction band. There in no intermediate, partially occupied band in between, as in the absolute zero temperature any pure semiconductor material is insulating. Then the energy gap between these bands is called the bandgap of the semiconductor.

The focus of our interest are three semiconductor materials: gallium nitride (GaN), aluminum nitride (AlN) and indium nitride (InN). All these semiconductors have so-called direct band gap, which means that in the k-space the minimal energy of the conduction band and the maximal energy of the valence band are for the same value of quasi-momentum vector. Thus in general the radiative recombination does not need to be phonon-assisted, which allows photon emission to be quite effective and makes these materials to be good candidates for a base of luminescent devices.

Material	Bandgap [eV]	Reference
AlN	6.1–6.2	[105]
GaN	3.4–3.5	[78, 122]
InN	0.7	[124]

Table 2.1: Approximate values for bandgaps of AlN, InN and GaN between absolute zero and room temperature.

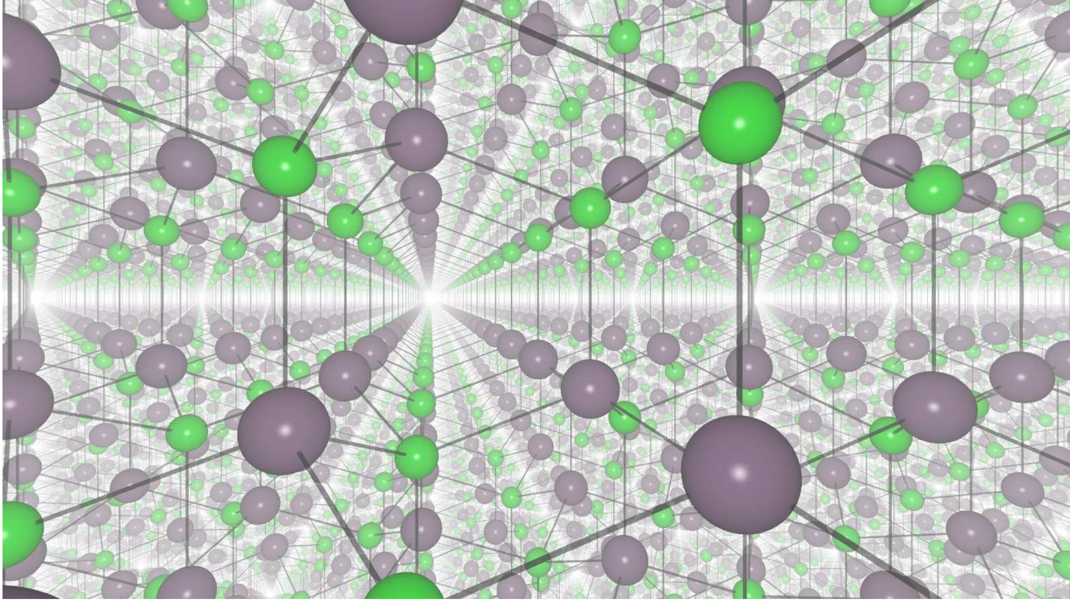

Figure 2.1: Gallium nitride wurtzite structure schema. Green spheres — nitrogen atoms, violet spheres — gallium atoms.

Bandgaps of AlN, GaN and InN are listed in table 2.1. They do not change considerably with temperature. Relative change of these parameters between the absolute zero and the room temperature (300 K) is less than 2 % [123].

2.1.2 Effective mass

Assume that $\Omega \in \mathbb{R}^3$ is some bounded, measurable set. Then according to the Free Electron Model, a wavefunction of a free electron may be represented as

$$\psi(x,t) = C \exp\Big(i\big(kx - \omega t\big)\Big), \tag{2.1.1}$$

where $x \in \Omega$, $t \in \mathbb{R}$ is the time, $\omega \in \mathbb{R}$ is the angular frequency of this particle and $k \in \mathbb{R}^3$ is its wavenumber. C is a normalizing constant, such that

$$\int_\Omega \psi(x,t)\psi^*(x,t)dx = 1. \tag{2.1.2}$$

Angular frequency depends on k by the dispersion relation

$$\omega(k) = \frac{\hbar k^2}{2m}, \tag{2.1.3}$$

where m is the mass of electron. Such a particle can be found in every part of Ω with the same probability, as its probability density of being found is constant:

$$\psi(x,t)\psi^*(x,t) = C^2 \exp\Big(i\big(kx - \omega t\big) - i\big(kx - \omega t\big)\Big) = C^2. \tag{2.1.4}$$

Therefore the wavefunction ψ of a free electron is determined by the wavenumber k. Knowing k, we can determine many corpuscular properties of the particle, for example

- energy: $E(k) = \hbar\omega(k)$,
- momentum: $p(k) = \hbar k$,
- velocity: $v(k) = \frac{\hbar k}{m}$,

and wave properties, for example

- wavelength: $\lambda(k) = \frac{2\pi}{k}$.

Note that the momentum p is proportional to the wavenumber k, so we will call both p and k the momentum unless it leads to confusion.

We emphasize a kinetic energy of a free electron is a parabolic function of the momentum:

$$E(k) = \frac{\hbar^2 k^2}{2m}. \tag{2.1.5}$$

In a crystal structure, the situation is much more complex as quasiparticles do interact with atoms and such a simple dependence does not exist. On the other hand, we are interested in wide bandgap semiconductors. We moreover assume that they are not degenerate, which means that the Fermi level is in the bandgap. Then most of occupied electron states in a conduction band are distributed near the energy minimum, and vacant states in the valence band are mostly near the energy maximum. Only these states contribute to the current, as others are either fully vacant or occupied.

The approach to deal with this problem is to treat the mass of a quasiparticle as a function dependent on k. The mass $m(k)$ is chosen such that the classical motion equation is satisfied:

$$a = m^{-1}(k)F. \tag{2.1.6}$$

Here a is the acceleration vector and F denotes the force. If some force acts on the quasiparticle, its state changes with time. Thus $k = k(t)$. Assume that k is continuously twice differentiable. First we note that

$$F = \frac{dp}{dt}(t) = \hbar \frac{dk}{dt}(t). \tag{2.1.7}$$

Thus using dependencies introduced above ($D = D_k$)

$$\begin{aligned} a(t) &= \frac{d}{dt} v\big(k(t)\big) = \frac{d}{dt} D\omega\big(k(t)\big) = \hbar^{-1} \frac{d}{dt} DE\big(k(t)\big) = \hbar^{-1} D^2 E\big(k(t)\big) \frac{dk}{dt}(t) \\ &= \hbar^{-2} D^2 E\big(k(t)\big) F(t). \end{aligned} \tag{2.1.8}$$

Thus by comparison to (2.1.6) we define

$$m(k) := \hbar^2 \Big[D^2 E\big(k\big)\Big]^{-1}, \qquad m^{-1}(k) := \hbar^{-2} D^2 E\big(k\big). \tag{2.1.9}$$

Now we would like to establish some analogy to equation (2.1.5). First we assume that the band is *parabolic*, i.e.

$$E(k) := C_1 k_1^2 + C_2 k_2^2 + C_3 k_3^2 = \sum_{i=1}^{3} \frac{\hbar^2}{2m_i^*} k_i^2, \tag{2.1.10}$$

where we define $m_i^* := \frac{\hbar^2}{2C_i}$. Then

$$D^2 E(k) = \begin{bmatrix} \frac{\hbar^2}{m_1^*} & 0 & 0 \\ 0 & \frac{\hbar^2}{m_2^*} & 0 \\ 0 & 0 & \frac{\hbar^2}{m_3^*} \end{bmatrix}, \tag{2.1.11}$$

and we get

$$a(t) = m_1^* F_1(t) + m_2^* F_2(t) + m_3^* F_3(t). \tag{2.1.12}$$

We call $m^* = [m_1^*, m_2^*, m_3^*]$ the *effective mass*. Note that m_i^* may be nonpositive. If $m_1^* = m_2^* = m_3^*$, then we say that the particle is *isotropic*. Then we treat m^* as a real number and we obtain

$$E(k) = \frac{\hbar^2 k^2}{2m^*}. \tag{2.1.13}$$

Otherwise we say that the particle is *anisotropic* and it behaves as it would have different mass in respective directions.

2.1.3 Current

Assume that some electric potential difference is applied to a piece of a semiconductor material. Then the electrostatic force acts on electrons and their momenta change. As we mentioned already, two kinds of electrons contribute to the electron flow: conduction band electrons being near the energy minimum and valence band electrons near the energy maximum.

The conduction band is full of energy states, which are mostly empty as electrons are scarce. Thus the electrons in this band can easily change their momenta and energy levels. On the other hand, in the valence band most of the states are already occupied. Only small part of topmost energy states in this band can be unoccupied and only there any movement of electrons is possible. In this regime it is convenient to treat these scarce unoccupied states as virtual particles instead of electrons to obtain similar behavior of carriers in both bands.

Then an unoccupied energy state in the valence band is called a hole. Such a quasi-particle is in fact a virtualization of an unoccupied state and an ensemble of electrons, which are responsible of its movement other physical properties. Thus charge of a hole is equal to the charge of an electron up to a sign, which is positive. The same is true for momentum, energy and effective mass.

2.1.4 Carrier statistics

2.1.4.1 Electrons

Since electrons and holes contribute to the current, it is necessary to estimate their concentration. First let us focus on electrons. In the absolute zero temperature, they tend to fill the lowest possible energy states. On the other hand they are fermions, thus due to the Pauli exclusion principle, only one electron can occupy a single state. Thus in absolute zero electrons fill up energy levels up to a valence band edge. Then, above the valence band, there is a forbidden zone, where no energy states are possible (in a pure semiconductor), and above there is a conduction band, which is full of empty states. When the temperature increases, some electrons attain higher energy, which allows them to cross the forbidden zone and occupy empty states in the conduction band.

It is important to evaluate the number of electrons, which are in the conduction band, as they contribute to . To estimate the electron concentration, we can use a distribution of electrons depending on energy level. To do so, we may use the *Fermi-Dirac distribution*:

$$f_{e,T}(E) := \frac{1}{1 + \exp\left(\frac{E-\mu}{k_B T}\right)}. \tag{2.1.14}$$

This is a probability density of the energy level E being occupied by some electron in the perfect electron gas in thermodynamic equilibrium. k_B is Boltzmann constant. μ is the Fermi level. This

formula for $T = 0$ gives

$$f_{e,0}(E) := \begin{cases} 1 & \text{if } E \leq \mu, \\ 0 & \text{if } E > \mu. \end{cases} \tag{2.1.15}$$

Function $f_{e,T}$ allows us to estimate the probability of an electron to be in some interval $[E_0, E_1]$, but there is one more concept to be introduced. In general, the number of available energy states is not uniform in E. As we mentioned before, there is for example a forbidden zone, where there are no energy states in pure semiconductor. Thus there is some distribution of states, which we call $dN(E)$. It is a probability measure, i.e.

$$\int_{-\infty}^{\infty} dN(E) = 1. \tag{2.1.16}$$

Two particular examples of density of states measure are:

$$\begin{aligned} dN(E) &:= \mathcal{D}(E)dE, \\ dN(E) &:= \sum_{i=0}^{\infty} \mathcal{D}(E)\delta(E - E_i)dE, \end{aligned} \tag{2.1.17}$$

where $g : \mathbb{R} \to [0, \infty)$ is a measurable function. In the latter case, we say that we have discrete energy levels. Thus the probability of an electron to have energy between E_0 and E_1 is

$$\int_{E_0}^{E_1} cf_{e,T}(E)dN(E), \tag{2.1.18}$$

where c is a normalization constant, such that

$$\int_{-\infty}^{\infty} cf_{e,T}(E)dN(E) = 1. \tag{2.1.19}$$

Let us mention an important simplification. Assume that available energy states are bounded from below by some E_0 and that $E_0 > \mu$. Then if $\exp\left(\frac{E-\mu}{k_B T}\right) \gg 1$, we can simplify

$$f_{e,T}(E) = \frac{1}{1 + \exp\left(\frac{E-\mu}{k_B T}\right)} \approx \exp\left(\frac{\mu - E}{k_B T}\right) =: b_{e,T}(E). \tag{2.1.20}$$

Function $b_{e,T}$ is called the *Boltzmann distribution*. To use the Boltzmann statistics effectively, we must assume that we deal with the *non-degenerate semiconductor*, i.e. that the Fermi level μ is in the bandgap, between the conduction band and the valence band. Then indeed $E - \mu > 0$ for the energy states available to electrons, and thus $\exp\left(\frac{E-\mu}{k_B T}\right) > 1$. On the other hand, a *degenerate semiconductor* has the Fermi level in the conduction band or in the valence band, and it behaves more like a metal than a semiconductor.

If a distribution of an ensemble of particles is governed by Boltzmann function, we say that they are described by *Boltzmann statistics*. In the other case, we say they are described by *Fermi-Dirac statistics*.

We would like to use these results in simulations of the semiconductor material. Schema of the energetic bands in the semiconductor material is in figure 2.2. We assume that valence and conduction band edges are denoted by $E_v(x)$ and $E_c(x)$, respectively. Let us denote by $n(x)$ the concentration of the electrons in conduction band in a given point $x \in \Omega$, where Ω corresponds to the space inside the semiconductor.

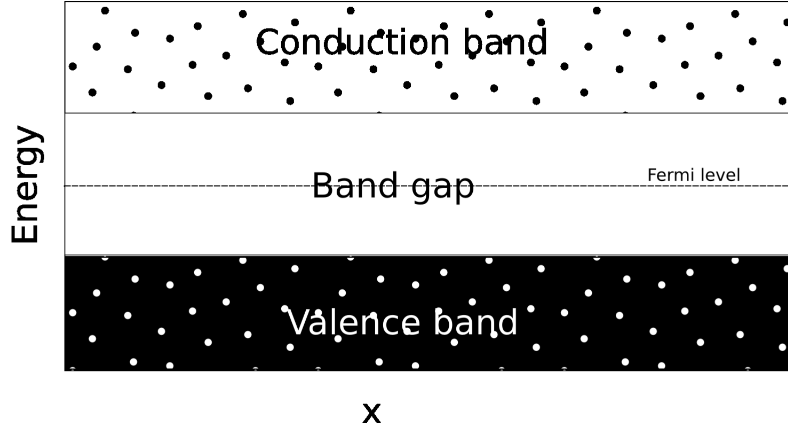

Figure 2.2: Schema of the energetic bands in the semiconductor material in nonzero temperature.

The density of states in the conduction band is given by (see [100])

$$\mathcal{D}_e(x,E) = \frac{1}{2\pi^2}\Big(\frac{2m_e}{\hbar^2}\Big)^{3/2}\sqrt{E-E_c(x)}, \tag{2.1.21}$$

where m_e is the effective electron mass, $\hbar$ is the reduced Planck constant. Thus, using the distributions above, we may calculate the concentration with the Boltzmann statistics

$$n(x) = \int_{E_c(x)}^{\infty} b_{e,T}(E)\mathcal{D}_e(x,E)dE = 2\Big(\frac{m_e k_B T}{2\pi\hbar^2}\Big)^{3/2}\exp\Big(\frac{\mu - E_c(x)}{k_B T}\Big) =: N_c \exp\Big(\frac{\mu - E_c(x)}{k_B T}\Big), \tag{2.1.22}$$

where $N_c := 2\Big(\frac{m_e k_B T}{2\pi\hbar^2}\Big)^{3/2}$ is called the *effective density of states in the conduction band*. In this derivation, we could use Fermi-Dirac statistics to get more precise results. Unfortunately, in that case the formula on $n(x)$ is more complicated. Also we will use the formula presented above in derivation of the drift-diffusion equations, what is not possible when Fermi-Dirac statistics are used.

Starting from equation (2.1.22), we would like to derive a form suitable for semiconductor simulations. In equation (2.1.22), parameter μ is the Fermi level. Under certain conditions (equilibrium state, see section 2.5.4), Fermi level for electrons and for holes is the same. If this is not the case, there are two *quasi-Fermi levels*, for electrons F_n and for holes F_p. For convenience, we will use the latter approach universally, even if $F_n \equiv F_p$. Thus we may substitute the Fermi level μ in (2.1.22) with the electron quasi-Fermi level F_n.

Then we must also take into account the contribution of the electrostatic potential to the band energies. Previously we generally ignored this contribution, which is equivalent to assuming $\psi \equiv 0$. Thus let us assume that E_c (E_v) is the conduction (valence) band edge for $\psi \equiv 0$. The contribution of ψ to the band edges is then as follows (see figure 2.3)

$$E_{c,\text{eff}}(x) := E_c(x) - q\psi(x), \qquad E_{v,\text{eff}}(x) := E_v(x) - q\psi(x). \tag{2.1.23}$$

In conclusion, we obtain the following formula for the electron concentrations, by substituting (μ, E_c) by $(F_n, E_{c,\text{eff}})$ in (2.1.22)

$$n(x) := N_c \exp\left(\frac{F_n(x) - E_c(x) + q\psi(x)}{k_B T}\right). \tag{2.1.24}$$

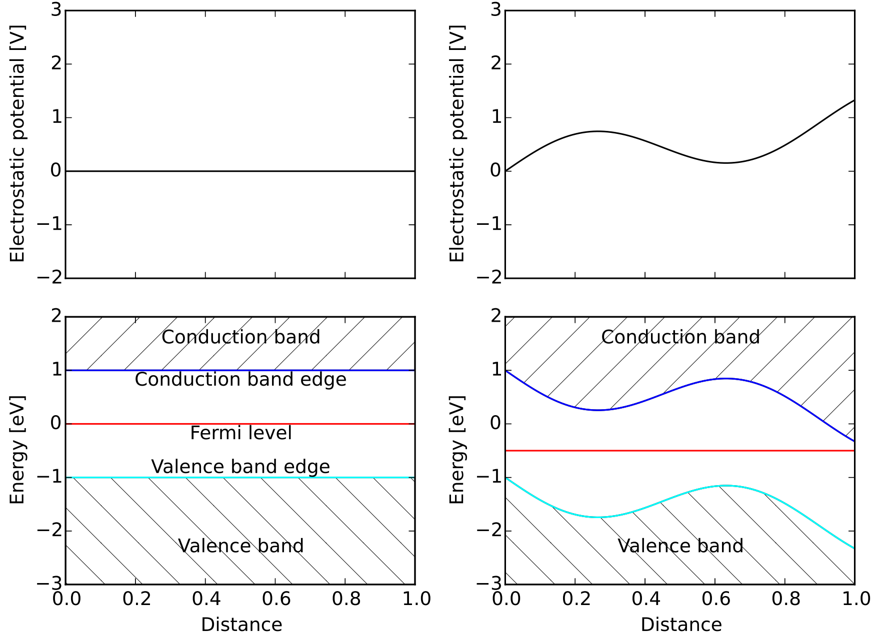

Figure 2.3: Schema of the energetic bands in the uniform semiconductor material subject to variation of the electrostatic potential ψ. Left: zero potential, right: arbitrary non-zero potential. Band edges (E_c, E_v) bend proportionally to $-\psi$.

2.1.4.2 Holes

By definition, a hole is a quasi-particle corresponding to empty electron state in the valence band. Therefore we have

$$f_{h,T}(E) := 1 - \frac{1}{1+\exp\left(\frac{E-\mu}{k_BT}\right)} = \frac{\exp\left(\frac{E-\mu}{k_BT}\right)}{1+\exp\left(\frac{E-\mu}{k_BT}\right)} = \frac{1}{\exp\left(\frac{\mu-E}{k_BT}\right)+1}. \tag{2.1.25}$$

This distribution corresponds to Fermi-Dirac statistics, while for Boltzmann we have, analogously as for electrons

$$b_{h,T}(E) := \exp\left(\frac{E-\mu}{k_BT}\right). \tag{2.1.26}$$

Distribution of energy states is also similar, but here it goes from $E_v(x)$ downwards

$$\mathcal{D}_h(x,E) = \frac{1}{2\pi^2}\left(\frac{2m_h}{\hbar^2}\right)^{3/2}\sqrt{E_v(x)-E}, \tag{2.1.27}$$

where m_h is the effective hole mass. Thus we obtain formula on concentration of holes $p(x)$ analogous to $n(x)$

$$p(x) = \int_{E_c(x)}^{\infty} b_{h,T}(E)\mathcal{D}_h(x,E)dE = 2\left(\frac{m_hk_BT}{2\pi\hbar^2}\right)^{3/2}\exp\left(\frac{E_v(x)-\mu}{k_BT}\right) =: N_v\exp\left(\frac{E_v(x)-\mu}{k_BT}\right), \tag{2.1.28}$$

where $N_v := 2\left(\frac{m_hk_BT}{2\pi\hbar^2}\right)^{3/2}$ is called the *effective density of states in the valence band.*

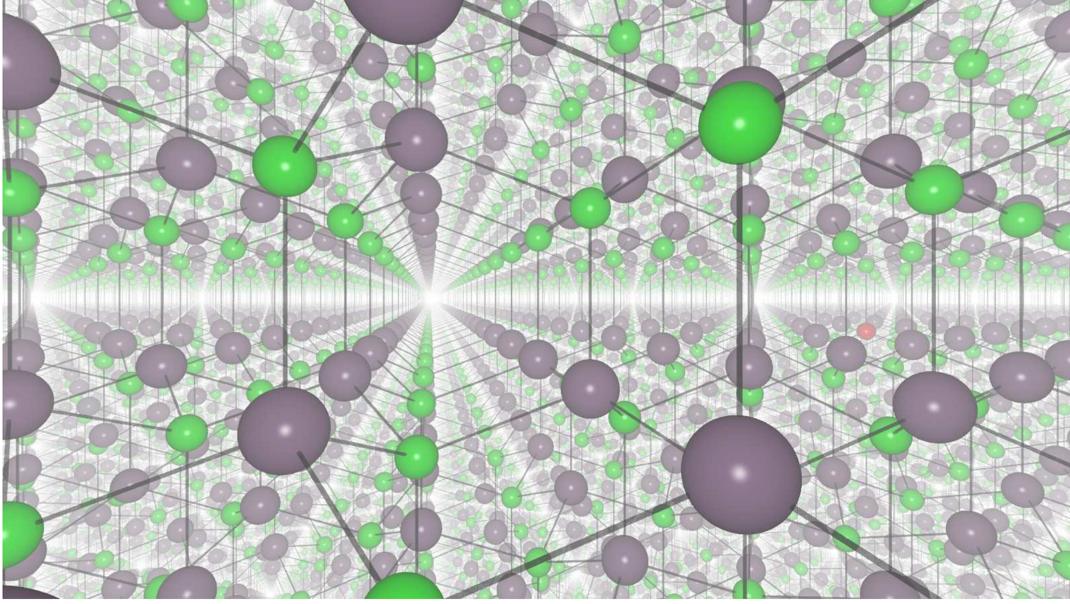

Figure 2.4: Schema of gallium nitride doped with magnesium (red sphere).

To derive a form used in simulations, we proceed analogously as for electrons, so substituting (μ, E_v) by $(F_p, E_{v,\mathrm{eff}})$ we obtain

$$p(x) = N_v \exp\left(\frac{E_v(x) - F_p(x) - q\psi(x)}{k_B T}\right). \tag{2.1.29}$$

2.1.5 Doping

In an intrinsic semiconductor, in absence of bias, concentration of electrons n_i and holes p_i are equal. Using equations (2.1.22), (2.1.28) we have

$$n_i(x)p_i(x) = 4(m_h m_e)^{3/2}\Big(\frac{k_B T}{2\pi\hbar^2}\Big)^3 \exp\Big(\frac{E_v(x) - E_c(x)}{k_B T}\Big). \tag{2.1.30}$$

Thus

$$\begin{aligned} n_i(x) = p_i(x) &= 2(m_h m_e)^{3/4}\Big(\frac{k_B T}{2\pi\hbar^2}\Big)^{3/2} \exp\Big(\frac{E_v(x) - E_c(x)}{2k_B T}\Big) \\ &= 2(m_h m_e)^{3/4}\Big(\frac{k_B T}{2\pi\hbar^2}\Big)^{3/2} \exp\Big(\frac{-E_g(x)}{2k_B T}\Big), \end{aligned} \tag{2.1.31}$$

where E_g is a *bandgap* defined as

$$E_g(x) := E_c(x) - E_v(x). \tag{2.1.32}$$

Using these formula and GaN parameters from tables 2.1, 2.2, we may estimate the intrinsic electron and hole concentration in room temperature as $n_i = p_i \approx 3 \times 10^{-10}$ cm^{-3}. For comparison, free electron concentration in copper is approximately 10^{23} cm^{-3}.

Thus we may easily conclude that pure gallium nitride is almost a perfect insulator (in room temperature). In that case, there are another ways of introducing current carriers into semiconductor material. One can increase the temperature, but this is a rather inefficient way. For example, in

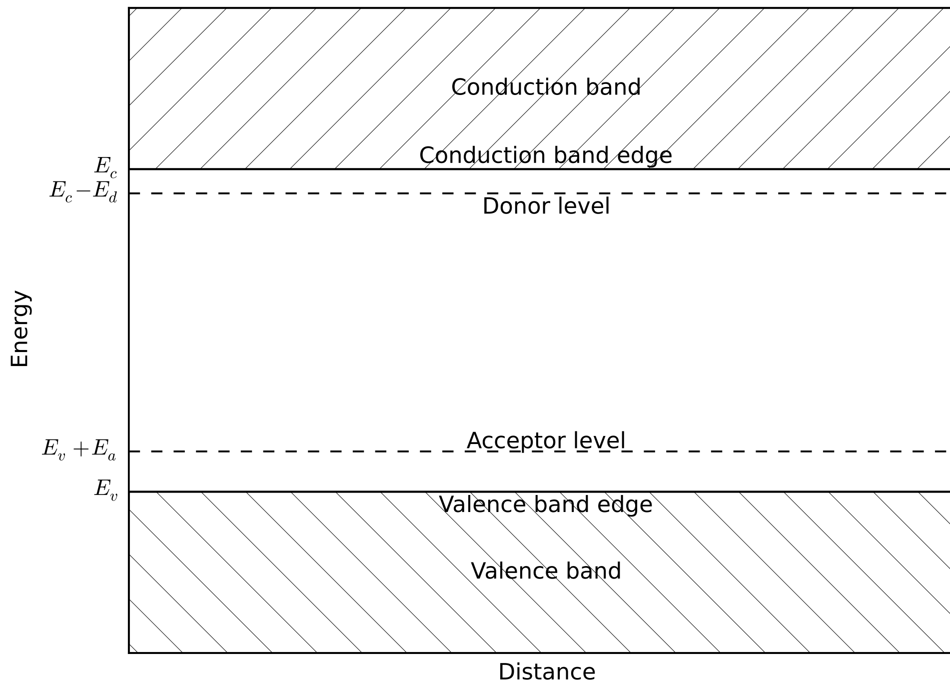

Figure 2.5: Schema of the band diagram of a doped semiconductor.

1000 °C, we have that $n_i = p_i \approx 10^{13}$ cm^{-3}, which is still too small amount to provide well-enough conductance, leaving aside for a moment that such temperature is unacceptable for a real device.

Much more efficient is to introduce defects in the crystal structure. These defects introduce additional possible states in the bandgap. For example, in gallium nitride, some gallium atoms can be replaced by silicon or magnesium. This method is called *doping*. In general, if we replace an atom with an atom with one more valence electron, this electron can be easily detached and contribute to the conduction band. Such dopant we call *donor*, as it gives an additional electron. On the other hand, if the replaced atom has one less valence electron, it can easily bind some electron from the valence band, introducing a hole in the valence band. This kind of impurity is called *acceptor*, as it accepts an electron.

Atom substitution (doping) is only an example of acceptor or donor defect. Other crystal impurities can also play similar role. Typical example is an atom vacancy. However, doping is widely used as it can be easily controlled by special growth techniques. It must be noted that atom doping is only possible up to certain concentration, approximately up to 0.1% of total substance (see figure 2.4), but this limit strongly depends on impurity type and growth technique. Then the substance behaves more like a mixed crystal (see section 2.2) or it can degenerate. We must emphasize that in a non-degenerated semiconductor, the energy states associated with the impurities are local. It is very hard or impossible for electrons to jump between these states, in contradiction to conduction band and valence band, where electrons may move freely in the semiconductor material (unless the density of defects is very high, which causes scattering of electrons). If the impurity concentration is high enough, the energy states introduced by impurities delocalize and they can act as an additional band. It causes a semiconductor to act more as a metal than a semiconductor, thus it is a degenerate semiconductor.

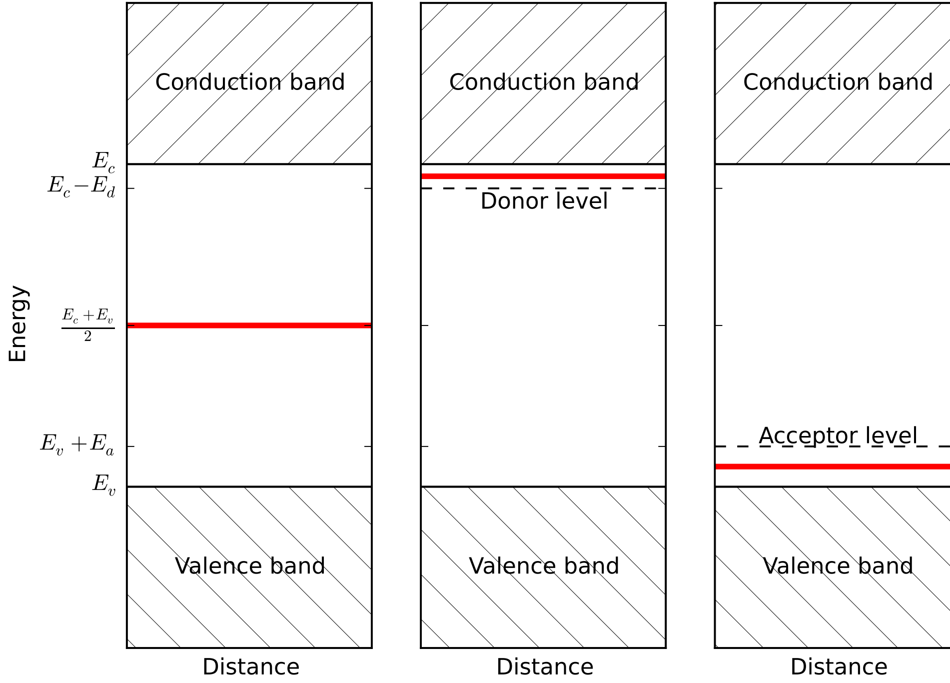

Figure 2.6: Fermi level (red line) in pure semiconductor (left), donor-doped semiconductor (center), acceptor-doped semiconductor (right).

Also not every atom makes a desired impurity. Generally proper impurities introduce new, discrete energy levels in the bandgap. This level should be close to the conduction band edge for a donor and close to the valence band edge for an acceptor (figure 2.5). Electrons from this level do not contribute to overall current, as due to low concentration these impurities are distant from each other and they form spatially localized states.

Let us focus on the effect of doping on the Fermi level in a semiconductor material in $T > 0$ (see figure 2.6). As we noted, in pure semiconductor, concentrations of electrons n_i and of holes p_i are equal. Thus we may easily calculate Fermi level μ comparing formulas (2.1.22), (2.1.28)

$$\mu = \frac{E_c + E_v}{2} + \frac{3}{4} k_B T \log\left(\frac{m_h}{m_e}\right) \approx \frac{E_c + E_v}{2}. \tag{2.1.33}$$

In room temperature (300 K), the latter element is relatively small, so μ lies in vicinity of the middle of the bandgap. It is in agreement with a fact, that for sufficiently low temperature, Fermi level value is approximately equal to the arithmetic mean of last occupied level and first unoccupied level in absolute zero.

Let us now discuss the position of Fermi level of semiconductor with donor doping. Donor-doped semiconductor is called *n-type* semiconductor. Again assume that $T > 0$ and that the concentration of doping is N_d. To estimate concentration of ionized donors N_d^+, we may use Fermi-Dirac function (2.1.14). First we calculate the concentration of donor states occupied by electrons N_d^0

$$N_d^0 := N_d \frac{1}{1 + g_d^{-1} \exp\left(\frac{E_c - E_d - \mu}{k_B T}\right)}, \tag{2.1.34}$$

where E_d is *ionization energy*. This is a difference between the conduction state edge and the donor level. Additional parameter g_d is called *donor degeneracy level*. If we take $g_d = 1$, then the formula above agrees with Fermi-Dirac function. This parameter allows to take into account certain devotion of the donor level. Typical values are $0.5, 1, 2$.

Then the concentration of donors which contribute their electron to the conduction band is

$$\begin{aligned} N_d^+ &:= N_d - N_d^0 = N_d \frac{g_d^{-1} \exp\left(\frac{E_c - E_d - \mu}{k_B T}\right)}{1 + g_d^{-1} \exp\left(\frac{E_c - E_d - \mu}{k_B T}\right)} = N_d \frac{1}{g_d \exp\left(\frac{\mu - E_c + E_d}{k_B T}\right) + 1} \\ &\approx N_d g_d^{-1} \exp\left(\frac{E_c - E_d - \mu}{k_B T}\right). \end{aligned} \tag{2.1.35}$$

This approximation is not essential for further analysis, but we will use it to approximate value of Fermi level. Assume that the temperature $T > 0$ is sufficiently low, such that the thermal excitation is negligible. Then $n = N_d^+$. Thus

$$2\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2} \exp\left(\frac{\mu - E_c(x)}{k_B T}\right) = N_d g_d^{-1} \exp\left(\frac{E_c - E_d - \mu}{k_B T}\right). \tag{2.1.36}$$

Let $n_0 := 2\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2}$. Then we have

$$\mu = E_c - 0.5E_d + k_B T \log\sqrt{\frac{N_d}{n_0 g_d}} \approx E_c - 0.5E_d. \tag{2.1.37}$$

This result is consistent with our previous result for pure semiconductor.

Note that if a donor atom loses its electron, it gains positive charge q. Thus ionized donors contribute also to total charge concentration.

Similar analysis may be performed for acceptors. In this case, ionized acceptor is simply acceptor level with an electron, so

$$N_a^- := \frac{N_a}{1 + g_a^{-1} \exp\left(\frac{E_v + E_a - \mu}{k_B T}\right)} \approx N_a g_a \exp\left(\frac{\mu - E_v - E_a}{k_B T}\right), \tag{2.1.38}$$

where E_a is the acceptor ionization energy and N_a is the acceptor concentration. As for donors, we assume that $N_a^- = p$ and using the above approximation

$$p_0 \exp\left(\frac{E_v(x) - \mu}{k_B T}\right) = N_a g_a \exp\left(\frac{\mu - E_v - E_a}{k_B T}\right), \tag{2.1.39}$$

where $p_0 := 2\left(\frac{m_h k_B T}{2\pi\hbar^2}\right)^{3/2}$. Thus

$$\mu = E_v + 0.5E_a + k_B T \log\sqrt{\frac{p_0}{N_a g_a}} \approx E_v + 0.5E_a. \tag{2.1.40}$$

This is similar to analogous result for donors. Also note that ionized acceptors have charge of value $-q$.

To obtain forms to be used in modelling, we have to substitute (μ, E_c) by $(F_n, E_{c,\mathrm{eff}})$ in (2.1.35) and (μ, E_v) by $(F_p, E_{v,\mathrm{eff}})$ in (2.1.38), as explained in section 2.1.4. Then we obtain

$$\begin{aligned} N_d^+ &= \frac{N_d}{1 + g_d \exp\left(\frac{F_n - E_c + q\psi + E_d}{k_B T}\right)}, \\ N_a^- &= \frac{N_a}{1 + g_a^{-1} \exp\left(\frac{E_v - q\psi + E_a - F_p}{k_B T}\right)}. \end{aligned} \tag{2.1.41}$$

Parameter	Symbol	AlN	GaN	InN
Relative permittivity	ε_r	8.5 [2]	8.9 [68]	15.3 [27]
Acceptor degeneracy level	g_a	2 [79]	2 [79]	2 [79]
Donor degeneracy level	g_d	2 [79]	2 [79]	2 [79]
Band gap	$E_c - E_v$	6.2 eV [121]	3.4 eV [77]	0.7 eV [119]
Acceptor level (Mg)	E_a	0.78 eV [72]	0.17 eV [68]	0.2 eV [107]
Donor level (Si, hydrogen-like)	E_d	0.064 eV [87]	0.02 eV [68]	0.013 eV [32]
Electron effective mass	m_n	0.33 [109]	0.2 [89, 68]	0.12 [42]
Hole effective mass	m_p	3.53 [109]	1.7 [41]	1.51 [120]
Electron mobility	μ_n	300 $\frac{\mathrm{cm}^2}{\mathrm{Vs}}$ [27]	200 $\frac{\mathrm{cm}^2}{\mathrm{Vs}}$ [95]	250 $\frac{\mathrm{cm}^2}{\mathrm{Vs}}$ [50]
Hole mobility	μ_p	14 $\frac{\mathrm{cm}^2}{\mathrm{Vs}}$ [39]	5 $\frac{\mathrm{cm}^2}{\mathrm{Vs}}$ [70]	39 $\frac{\mathrm{cm}^2}{\mathrm{Vs}}$ [26]

Table 2.2: Material parameters of aluminum nitride, gallium nitride and indium nitride in room temperature (300 K).

2.1.6 Energy distribution in a crystal structure

In nonzero temperature, energy of a crystal is divided into several degrees of freedom, including electrons on their energy levels and vibrations of the crystal lattice. Quantum of energy of the lattice vibration is called a *phonon*. Thus we consider electrons and phonons in the crystal and photons outside of the crystal. All of them can carry some portion of energy.

The most simple energy transfer is involved in radiative generation/recombination process. In the most simple case, during radiative recombination, an electron from the conduction band loses its energy landing in the valence band. All the energy is transferred to a new photon, which is emitted during the process. The opposite effect, where a photon is absorbed and its energy is transferred to some valence band electron, which is then raised to the conduction band, is called a radiative generation. The term *generation* here refers to an electron/hole pair, not to the photon.

In the nonradiative recombination process, some electron also loses its energy, but this energy is transferred to phonons. From the point of view of the efficiency of a luminescent device, these phenomena are harmful, as they increase the temperature of a device.

2.2 Properties of the mixed AlGaN and InGaN crystals

Blue and green optoelectronics is generally based on aluminum nitride, indium nitride and gallium nitride. Selected physical properties of these materials is presented in table 2.2. These materials crystallize in wurtzite structure (figure 2.7) with lattice parameters as in table 2.3.

However, almost every device contains also mixed compounds: $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{N}$ or $\mathrm{In}_x\mathrm{Ga}_{1-x}\mathrm{N}$ (see

Material	a [nm]	c [nm]
AlN	0.31	0.50
GaN	0.32	0.52
InN	0.35	0.57

Table 2.3: Lattice parameters of wurtzite structure for AlN, InN and GaN [67].

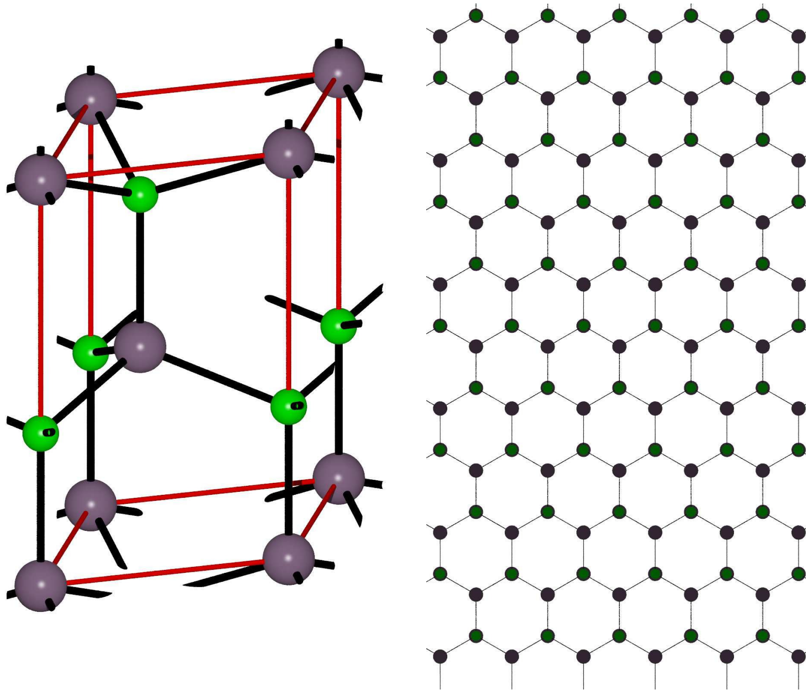

Figure 2.7: Primitive cell of a wurtzite structure (left) and projection of a wurtzite crystal to the plane parallel to the base of the cell. Dimensions of the primitive cell are determined by length a of base of the cell, which is a rhombus, and height c of the cell.

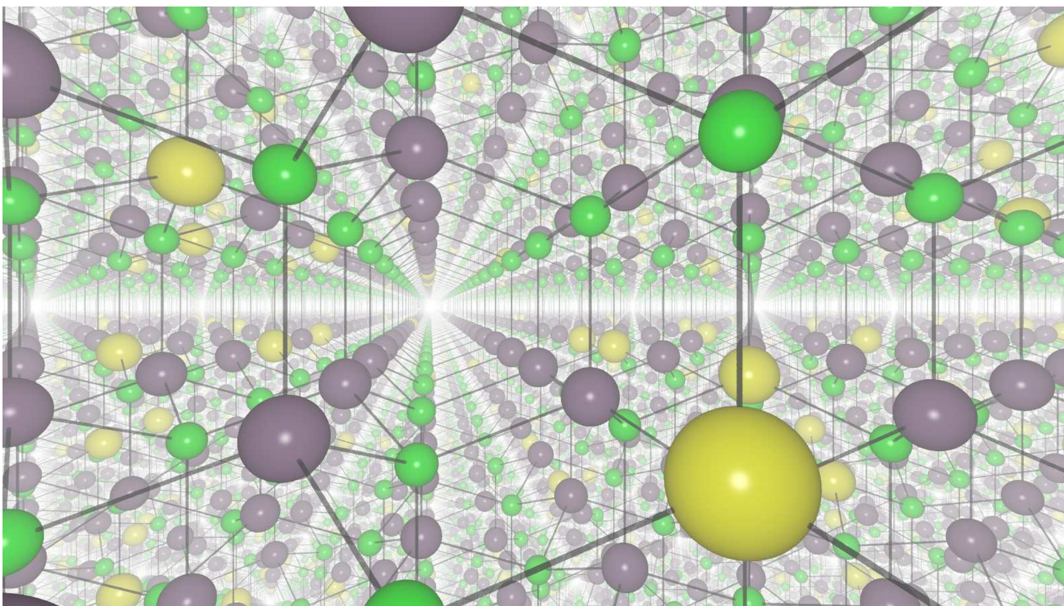

Figure 2.8: Schema of indium-gallium nitride (10% In). Yellow spheres — indium atoms.

figure 2.8). To perform simulations of these devices, it is necessary to provide properties of such materials. These properties may be obtained via physical experiments or numerical simulations.

While obtaining certain physical parameters of pure AlN, GaN and InN itself may be a hard problem, it is even more complicated in case of mixed compounds. For example, it is not easy to obtain uniform concentrations of components, as they often tend to cluster in some regions, changing macroscopic properties of the material. On the other hand, it is experimentally observed that many attributes of such compounds satisfy the following elementary rule. Let $f : \{\mathrm{Al}_x\,\mathrm{In}_y\,\mathrm{Ga}_{1-x-y}\,\mathrm{N}\}_{x,y\geq 0, x+y\leq 1} \to \mathbb{R}$ be some physical property. Let us fix $\mathrm{Al}_x\,\mathrm{In}_y\,\mathrm{Ga}_{1-x-y}\,\mathrm{N}$ for a given x, y. Then $f(\mathrm{Al}_x\,\mathrm{In}_y\,\mathrm{Ga}_{1-x-y}\,\mathrm{N})$ is approximately equal to a convex combination of respective properties of pure materials, i.e.

$$f(\mathrm{Al}_x\,\mathrm{In}_y\,\mathrm{Ga}_{1-x-y}\,\mathrm{N}) \approx x f(\mathrm{AlN}) + y f(\mathrm{InN}) + (1-x-y) f(\mathrm{GaN}). \tag{2.2.1}$$

This result is called the Vegard rule. This general rule is related to the fact that mixed compounds, Al or In atoms substitute Ga atoms. It is assumed that the concentration of compounds is big enough so that the substitutions do not form isolated states, but they contribute to conduction band and valence band.

If we assume that the Vegard rule is approximately true for the band gap, we conclude that any energy gap in range 0.7–6.2 eV may be obtained for appropriate $\mathrm{Al}_x\,\mathrm{Ga}_{1-x}\,\mathrm{N}$ or $\mathrm{In}_y\,\mathrm{Ga}_{1-y}\,\mathrm{N}$ compound. Thus the nitrides seems to be good candidates for optoelectronic devices, as they cover full visible spectrum range.

It is sometimes beneficial to improve the Vegard rule with terms of higher order. For example, for $\mathrm{In}_y\,\mathrm{Ga}_{1-y}\,\mathrm{N}$ we can approximate

$$f(\mathrm{In}_y\,\mathrm{Ga}_{1-y}\,\mathrm{N}) \approx y f(\mathrm{InN}) + (1-y) f(\mathrm{GaN}) + y(1-y) C_{\mathrm{InGaN}}, \tag{2.2.2}$$

where C_{InGaN} is fitted to the experimental data. Parameter C_{InGaN} is called a *bowing parameter*.

Another problem is the temperature dependence of material parameters. For nitrides, however, such dependence is often limited. For example, difference in band gap for 0 K and 300 K for AlN, GaN and InN is less that 2% [123].

We must emphasize that in a realistic device there is always certain amount of impurities, both acceptors and donors, which come from limitations of growth techniques, strain of material, environment, etc. These impurities may also act as recombination centers. More details will be presented in section 2.6.3.

2.3 Geometry of luminescent semiconductor structures

2.3.1 p-n homojunction

This is a very simple device. It consists of two layers, which differ only by the doping type. Material composition is the same for both layers. First layer is donor-doped and second layer is acceptor-doped. Thus the device is divided into two regions, called *n-type region* and *p-type region*.

In vicinity of the interface between n-type region and p-type region, there is so-called *depletion region*. It is formed in the following manner. Initially in the n-type region there are two types of charge: negative mobile electrons and positive immobile ionized donors. Their concentrations are equal and thus the net charge is zero. Analogously in p-type region there are positive mobile holes and negative immobile ionized acceptors, and their concentrations are initially equal.

Then the interface between n-type region and p-type region is formed and mobile carriers diffuse. Thus electrons approach p-type region, leaving uncompensated donors, and analogously holes enter n-type region, leaving uncompensated acceptors. Finally they recombine and annihilate, so their

concentrations descend. Therefore positive charge in n-type region and negative charge in p-type region emerges. This charge creates an electric field, which acts on the carriers contrarily to the diffusion. These effects continue until the drift due to electric field balances the diffusion. It leads to depletion of the interface vicinity from the mobile carriers.

Then if some bias is applied to the device, two situations are possible. Under *forward bias*, holes are injected to the p-type region and electrons are injected to the n-type region. These carriers contribute to the charge concentration by balancing uncompensated immobile charge in the depletion zone and they decrease the field in the depletion region. The overflow carriers recombine, mostly near the interface between regions.

On the contrary, under *reverse bias*, number of the majority carriers descend and the depletion region increases, as well as the field near the interface. It acts as an insulating layer, so the current under the reverse bias is very small.

2.3.2 Laser diodes and electroluminescent diodes

In our simulations we focus our attention on optoelectronic devices. We have two types of such devices. Electroluminescent diodes (LEDs) emit light simultaneously in wide range of directions. They operate in relatively low voltage (up to 5 V). As in p-n homojunctions, they generally consist of two regions, n-type region and p-type region.

Most of the light emitted by a device comes from the *active region* placed on a border between the n-region and p-region. The active region structure is explained in detail in section 2.4.

LEDs consist of layers of semiconductor material deposited one on piled upon each other. These layers are mostly cuboids, with a rectangular base. The base dimensions (e.g. $300\,\mu m \times 300\,\mu m$) are greater than the cuboid's height ($1\,nm$–$100\,\mu m$). Due to technological requirements LEDs may have sophisticated shape, obtained by cutting, etching or cleaving of the deposited layers, but generally the current flows mostly perpendicular to bases of cuboids-layers, parallel to the growth direction. Such a structure make it feasible to use a one-dimensional model for simulation of LEDs.

On the other hand, laser diodes (LDs) operate in higher voltage (5–8 V) and they emit focused beam in a given direction only. Their structure is similar to LED's, but there are important differences. There are two cladding layers, which confine the light to the active region of these devices. Also their cross-section is a prolonged rectangle (e.g. $10\,\mu m \times 300\,nm$). In comparison to LEDs, where the structure must be optimized to get good transport properties of electrons and holes, LDs' structures must also form an optical resonator when combined with mirrors situated on the faces.

2.4 Quantum structures: wells and barriers

The crucial part of a luminescent optoelectronic device is the *active region*. The active region is a part of a luminescent device, which is dedicated to generate the light. It is located in the vicinity of the border between the n-type region and the p-type region and, depending on a device, it may be undoped or partially doped.

Due to proximity of both regions, it is possible to inject both electrons and holes to the active region, when they recombine, preferably radiatively, emitting photons. To make the radiative recombination most likely, and to obtain a given photon energy, this part of a device must be designed in a special manner. In a direct bandgap material, an electron and a hole must have the same pseudo-momentum value to recombine radiatively. In AlN, GaN and InN the conduction/valence band extremes are achieved for zero pseudo-momentum. Thus to increase the radiative recombination rate, it is favorable to stop the carriers in a small area.

Thus quantum wells (QWs) are introduced in the active region of luminescent devices. These wells are narrow to localize the carriers in a possibly small area and they made of material with the smaller band gap than surrounding layers. Then the band edges form potential wells for electrons and holes, where separating layers act as potential barriers. In such a structure, two kinds of carrier transport are possible: classic ballistic transport of highly energetic carriers over the barriers and quantum tunneling through the barriers. The latter is unlikely in modern devices, as the barriers are designed wide enough ($\approx 8\,\mathrm{nm}$) to prohibit quantum tunneling. The layers surrounding the quantum wells/barriers region also act as barriers, making this region some kind of trap for the electrons and holes.

In an efficient device, most of the recombination should occur in the quantum wells, and it should be the radiative recombination. To promote the radiative recombination, quantum well layers should be made of possibly high quality material. To improve the quality, it is often the case that these layers are not doped as well as the rest of the active region.

2.5 Drift-diffusion model

2.5.1 Conservation laws and equations of motion

Let us focus on movement of the charge carriers in the semiconductor material. Take electrons as an example. On the microscopic level, an electron travelling through the semiconductor material is subjected to the periodic variation of the potential due to atomic cores superimposed on the macroscopic, slowly-varying potential. On the basis of quantum mechanics, we may mimic this microscopic periodic changes by introducing an *effective mass* of the quasiparticle (see section 2.1.2) . Then we may consider the macroscopic potential only, replacing mass of the particle with this effective mass.

In the drift-diffusion model, we are interested in statistical approach, thus instead of individual particles, we would like to describe the evolution of concentrations of the quasiparticles in time.

It is convenient to express movement of current carriers in terms of *electric current density.* In general, it is a vector J denoting the amount of charge flowing through the infinitesimal surface during the infinitesimal time duration. In other words, having a given surface S and time interval $[t_1, t_2]$, the total amount of charge flowing through this surface in that time may be expressed as

$$\int_{t_1}^{t_2} \int_S J(x) \cdot \nu dS dt, \tag{2.5.1}$$

where ν is a normal vector to the surface S for a given $x \in S$. In the semiconductor, we may distinguish two contributions to the total current density: *electron current density* J_n and *hole current density* J_p ($J = J_n + J_p$).

First let us focus on electrons. In the semiconductor material, electrons accelerate due to electric field. On the other hand, moving electrons collide with phonons and impurities and they lose momentum. In presence of certain concentration of electrons, this phenomenon may be treated statistically by introducing of *mean free path* between collisions and *group velocity*, defined as mean velocity of all present electrons. Electrons accelerate during free path, then they collide and lose velocity, and then they accelerate again. This argumentation leads to the conclusion that given an *electric field* E, we have some group velocity $v_{g,n}$ proportional to E, i.e. $v_g = \mu_n E$. The proportionality constant μ_n is called the *electron mobility* and it depends on the semiconductor material and temperature.

This effect is called a drift of electrons in the electric field. We may express it in therms of the electron current density as a product of the concentration of electron charge concentration $qn(x)$ by the electron velocity, i.e.

$$J_{n,\mathrm{drift}}(x) := qn(x)\mu_n E(x). \tag{2.5.2}$$

On the other hand, electrons undergo the diffusion, which is proportional to their negative concentration gradient (Fick's first law). Thus the *electron diffusive flux* is $-D_n n$, where D_n is called the *electron effective diffusivity.* Then the diffusive part of the electron current density is

$$J_{n,\text{diff}} := qD_n \nabla n(x). \tag{2.5.3}$$

Note that the positive sign is due to negative charge of an electron.

Combining both drift and diffusion, we obtain

$$J_n = J_{n,\text{drift}} + J_{n,\text{diff}} = qn(x)\mu_n E(x) + qD_n \nabla n(x). \tag{2.5.4}$$

By analogous argumentation for holes we obtain

$$J_p = J_{p,\text{drift}} + J_{p,\text{diff}} = qp(x)\mu_p E(x) - qD_p \nabla p(x), \tag{2.5.5}$$

where μ_p is the *hole mobility* and D_p is the *hole effective diffusivity.* Note that there is a minus sign before the latter element, which is due to positive charge of a hole. Positive sign before the first element is due to both opposite drift direction of holes and the positive charge.

Let us simplify expressions for J_n and J_p. First note that $E := -\nabla\psi$, where ψ is an electric potential. Then by Einstein relations

$$D_n = \mu_n \frac{k_B T}{q}, \qquad D_p = \mu_p \frac{k_B T}{q}. \tag{2.5.6}$$

We assume that the semiconductor material is homogeneous, so the material parameters are constant. Then note that

$$\begin{aligned} \nabla n(x) &= N_c \nabla \exp\Big(\frac{F_n(x) - E_c + q\nabla\psi(x)}{k_B T}\Big) \\ &= \frac{1}{k_B T} N_c \exp\Big(\frac{F_n(x) - E_c + q\psi(x)}{k_B T}\Big)\Big(\nabla F_n(x) + q\nabla\psi(x)\Big) \\ &= \frac{1}{k_B T} n(x)\Big(\nabla F_n(x) + q\nabla\psi(x)\Big). \end{aligned} \tag{2.5.7}$$

Analogously

$$\nabla p(x) = N_v \nabla \exp\Big(\frac{E_v - F_p(x) - q\psi(x)}{k_B T}\Big) = -\frac{1}{k_B T} p(x)\Big(\nabla F_p(x) + q\nabla\psi(x)\Big). \tag{2.5.8}$$

Therefore we obtain

$$\begin{aligned} J_n &= -qn(x)\mu_n \nabla\psi(x) + \mu_n n(x)\Big(\nabla F_n(x) + q\nabla\psi(x)\Big) = \mu_n n(x)\nabla F_n(x), \\ J_p &= -qp(x)\mu_p \nabla\psi(x) + \mu_p p(x)\Big(\nabla F_p(x) + q\nabla\psi(x)\Big) = \mu_p p(x)\nabla F_p(x). \end{aligned} \tag{2.5.9}$$

By virtue of Maxwell's equations, if there is no magnetic field, then

$$\nabla \cdot J + \frac{\partial \rho}{\partial t} = 0, \tag{2.5.10}$$

which represents the simple fact that in and out current densities are not in balance, the electrostatic charge accumulates. In stationary case the latter term is zero, so we obtain

$$\nabla \cdot J = 0. \tag{2.5.11}$$

This argumentation cannot be extended to electron current density J_n and hole current density J_p, as these currents are affected by variation of carrier concentrations n, p due to the generation/recombination effect.

Since $J = J_n + J_p$, then

$$\nabla \cdot J_n = -\nabla \cdot J_p. \tag{2.5.12}$$

Change of the electron current density is proportional to the generation/recombination rate, so

$$\nabla \cdot J_n = qR, \tag{2.5.13}$$

and thus using (2.5.12)

$$-\nabla \cdot J_p = qR. \tag{2.5.14}$$

Combining these results with (2.5.9), we obtain two so-called *continuity equations* for electrons and for holes:

$$\begin{aligned} -\nabla \cdot \left(\mu_n n \nabla F_n\right) &= -qR, \\ -\nabla \cdot \left(\mu_p p \nabla F_p\right) &= qR. \end{aligned} \tag{2.5.15}$$

2.5.2 Electric field, electrostatic potential and polarization effect

Equations (2.5.9) serve as a starting point to two of the three van Roosbroeck equations, called the continuity equations. Still we have to elaborate on the remaining equation, called the Poisson equation.

We start with the Gauss law for the *electric displacement field* D:

$$\nabla \cdot D = \rho. \tag{2.5.16}$$

Here ρ is the *free electric charge*. In case of semiconductors, the free electric charge consists of electrons and holes, introduced in section 2.1.4, and the ionized impurities described in detail in section 2.1.5. While the ionized impurities are not really "free" in the sense of spatial movement, they can change their state during operation of a device and they do not contribute significantly to the electric polarization. This brings us to the notion of the *bound charge* ρ_{bd}. The bound charge consists of core protons and valence electrons, which do not contribute to the electrical conductivity. On the other hand, they do contribute to the electric polarization, which we would like to briefly discuss.

Let us start with isotropic ("uniform in all orientations") dielectric (insulating) material. If this is the case, then the bound charge is distributed uniformly and the polarization is zero. If we, however, apply some electric field to that material, positive atom cores and negative electrons will slightly shift in opposite directions, leading to non-uniform distribution of the bound charge. This kind of polarization is proportional to the applied electrical field.

Such a non-uniform distribution of the bound charge may occur also due to different reasons, like temperature, strain, or it may be characteristic to a given material without any external stimulus (*spontaneous polarization* P_o).

The electric displacement is a sum of the electric field E scaled by the *permittivity of vacuum* ε_0 and the polarization P, i.e.

$$D = \varepsilon_0 E + P. \tag{2.5.17}$$

We assume that the polarization P, in case of semiconductor material, consists of a part proportional to the electric field and of the part independent of the electric field:

$$P := \varepsilon_0 \chi E + P_o. \tag{2.5.18}$$

Parameter χ is called the *electric susceptibility* of the medium and it indicates the response of the material to the external electric field. Thus we have

$$\rho = \nabla \cdot D = \nabla \cdot (\varepsilon_0 E + \varepsilon_0 \chi E + P_o) = \nabla \cdot (\varepsilon_r \varepsilon_0 E + P_o) = \nabla \cdot (\varepsilon E + P_o), \tag{2.5.19}$$

where $\varepsilon_r := 1 + \chi$ is called a *relative permittivity.* We also define

$$\varepsilon := \varepsilon_r \varepsilon_0. \tag{2.5.20}$$

Under stationary conditions, the electric potential satisfies

$$-\nabla \psi := E. \tag{2.5.21}$$

Combining this definition with equation (2.5.19), we obtain

$$\nabla \cdot (-\varepsilon \nabla \psi + P_o) = \rho. \tag{2.5.22}$$

Finally we pass the polarization term to the right hand side and we obtain

$$-\nabla \cdot (\varepsilon \nabla \psi) = \rho - \nabla \cdot P_o. \tag{2.5.23}$$

We will call this equation *Poisson's equation.*

In this derivation, we assumed χ and ε to be real numbers. It is possible that these parameters are different on different directions due to low symmetry of crystalic structure. Then these parameters are represented by matrices instead of scalars, but the analysis is similar. Such materials are called the *anisotropic* materials.

2.5.3 Differential problem

From the modelling standpoint, the drift-diffusion model consists of three elliptic nonlinear equations: Poisson's Equation (2.5.23) and two continuity equations (2.5.15).

$$\begin{aligned} -\nabla \cdot (\varepsilon \nabla \psi) &= \rho - \nabla \cdot P_o, \\ -\nabla \cdot \Big(\mu_n n \nabla F_n\Big) &= -qR, \\ -\nabla \cdot \Big(\mu_p p \nabla F_p\Big) &= qR. \end{aligned} \tag{2.5.24}$$

This system is also known as *van Roosbroeck equations* [96]. In this form, the unknown functions are: the electrostatic potential ψ, the electron quasi-Fermi level F_n and the hole quasi-Fermi level F_p. If by $\Omega \in \mathbb{R}^d$, $d \in \{1, 2, 3\}$ we denote the domain of a modelled semiconductor device, then $\psi, F_n, F_p : \Omega \to \mathbb{R}$. Functions $\mu_n, \mu_p : \Omega \to \mathbb{R}$ are the electron and hole mobilities. They are the material parameters described earlier, and we assume they are piecewise-constant. Function $P_o : \Omega \to \mathbb{R}$, the polarization parameter, is somewhat special. In some devices it is constant and then it is irrelevant. If it is piecewise constant, then $\nabla \cdot P$ should be interpreted as a distributional derivative, which corresponds to the concept of interfacial polarization charge, occurring at interfaces between materials with different polarization.

Functions $n, p : \Omega \to \mathbb{R}_+$ are the concentration of electrons and the concentration of holes. In this form of van Roosbroeck equations, they are dependent functions.

Function $\rho : \Omega \to \mathbb{R}$, the charge concentration, is a sum of contributions of localized charge, like ionized donors and ionized acceptors (section 2.1.5), ionized traps (section 2.6.3), and delocalized

charge, which consists of electrons and holes (section 2.1.4). Thus if trap contribution is negligible, it reads

$$\rho = p - n + N_d^+ - N_a^-, \tag{2.5.25}$$

and if we take traps into account

$$\rho = p - n + N_d^+ - N_a^- + \sum_{t \in T_L} \pm N_t^\pm, \tag{2.5.26}$$

where T_L is a set of indices of trap levels and the sign for a given $t \in T_L$ depends on the trap's occupation.

Finally the generation/recombination rate $R : \Omega \to \mathbb{R}$ is a sum of elements corresponding to different physical effects. More details on the recombination models are presented in section 2.6.

The drift-diffusion equations may be presented in several sets of unknown functions. Above we presented the formulation using unknowns (ψ, F_n, F_p). Alternatively we could use carrier concentrations (ψ, n, p) instead of the quasi-Fermi levels. Then using definitions (2.5.4), (2.5.4), $E = -\nabla\psi$, we get the following formulation:

$$\begin{aligned} -\nabla \cdot (\varepsilon \nabla \psi) &= \rho - \nabla \cdot P_o, \\ -\nabla \cdot \Big(D_n \nabla n - \mu_n n \nabla \psi\Big) &= -R, \\ -\nabla \cdot \Big(D_p \nabla p + \mu_p p \nabla \psi\Big) &= -R. \end{aligned} \tag{2.5.27}$$

Different choices of unknown functions are possible, a detailed discussion is presented in [92]. Generally ψ is present in every described set, while other two unknowns are subject to change. The difference between these options is mostly in exponential character of the solutions and nonlinearity of the equations.

In this study, the main emphasis in on the quasi-Fermi level formulation (2.5.24), as it keeps the equations possibly simple and the unknowns (ψ, F_n, F_p) do not express the exponential character. On the other hand, the carrier concentrations n, p are clearly exponential, e.g they may vary between 10^{-40}–10^{20}. This behavior poses a significant problem with numerical solution of these equations.

2.5.4 Equilibrium state and non-equilibrium state

In simulation on semiconductor devices, we distinguish between two states of a device.

The equilibrium state corresponds to thermodynamic equilibrium, with no net flow of energy within a device or between a device and the environment. Under these conditions, the net generation/recombination is zero, so the recombination is in balance with generation. Otherwise there would be some photons absorbed or generated, heat would be generated, etc. Zero recombination implies the quasi-Fermi levels to be equal, so in fact there is a single Fermi level for both types of carriers. These conditions are physically realized to some extent by a device disconnected from the power source, in the dim light.

On the other hand, if we apply some voltage to such device, illuminate it, etc., then there would be additional charge carriers injected through the contacts or generated, throwing the generation/recombination out of balance. The Fermi level would split up into quasi-Fermi levels, separate for electrons and for holes. This is the non-equilibrium state, a natural state for operating devices.

From the modelling standpoint, the latter state is corresponding to a solution of the drift-diffusion equations (2.5.24). Conversely, in the equilibrium state $F_n \equiv F_p = \text{const}$ and only unknown is ψ. Thus only Poisson's equation (2.5.23) is necessary in this case.

2.5.5 Built-in potential

Before the discussion of the boundary conditions, we would like to introduce the concept of built-in potential. The physical phenomena occurring in semiconductor material may be classified as a bulk property or a boundary property. A bulk property is associated with the interior of the material, and it generally occur in the major part of the volume of the material and they are independent of the boundary, while boundary properties are strongly correlated with the state of the boundary and they do not penetrate deeply into the interior. The latter type of phenomena may also be present on the interfaces between different semiconductor materials.

The bulk semiconductor material, in the natural state, without external forces, is charge-neutral (in the macroscopic scale). The charged state over the substantial volume is hard to maintain, as it generates large forces, attracting carriers of opposite value, which will accumulate with time and balance the charge.

Note that while maintaining large volume of charged material is energetically costly, it is often possible that existence of charge on interfaces of boundaries is energetically favorable.

Assume that we have some device in the equilibrium state, and assume that it is built of layers of possibly different semiconductor material, homogeneous within every layer. Then there is single constant Fermi level for electrons and holes in this structure, as explained in section 2.5.4. Then we may define the *built-in potential* $\psi_b : \Omega \to \mathbb{R}$ by condition of charge-neutrality:

$$\rho(x, \psi) = 0. \tag{2.5.28}$$

We would like not to go into details whether this definition is well-posed or not, and we assume this equation can be solved uniquely for almost every $x \in \Omega$. Note that ψ_b is generally discontinuous, and it is piecewise constant when the layers are homogeneous.

2.5.6 Boundary conditions

In this section, we would like to briefly discuss certain aspects of boundary conditions, in particular ohmic contacts and contacts with insulators. A detailed analysis is presented in [101].

2.5.6.1 Contact with insulator

The basic aspect of the contact with insulator is that current does not flow through it. This statement is equivalent to

$$\begin{aligned} J_n \cdot \nu &= 0, \\ J_p \cdot \nu &= 0. \end{aligned} \tag{2.5.29}$$

In the above equations we also assumed that there is no surface recombination, as otherwise it would be present on the right hand side of both equations (see [101]). By definition (2.5.9) of J_n and J_p, we have

$$\begin{aligned} 0 = J_n \cdot \nu = \mu_n n(x) \nabla F_n(x) \cdot \nu = \mu_n n(x) \frac{\partial F_n}{\partial \nu}(x), \\ 0 = J_p \cdot \nu = \mu_p p(x) \nabla F_p(x) \cdot \nu = \mu_p p(x) \frac{\partial F_p}{\partial \nu}(x). \end{aligned} \tag{2.5.30}$$

Since $\mu_n, \mu_p, n, p > 0$, then $\frac{\partial F_n}{\partial \nu} = \frac{\partial F_p}{\partial \nu} = 0$, and we get homogeneous Neumann boundary conditions on F_n, F_p.

When it is also assumed that the electric field vanish in the insulator. When there is no boundary charge, we can assume that $\frac{\partial \psi}{\partial \nu} = 0$.

Thus on the contact with insulator, we have homogeneous Neumann boundary conditions:

$$\frac{\partial \psi}{\partial \nu} = \frac{\partial F_n}{\partial \nu} = \frac{\partial F_p}{\partial \nu} = 0. \tag{2.5.31}$$

2.5.6.2 Ohmic contact

Ohmic contact is an edge of a semiconductor device to which some metal is attached, which allows the device to be connected to an electrical circuit. In simulations it is frequently assumed that the ohmic contact is ideally conducting.

To derive boundary conditions, we must discuss some properties of a semiconductor device and introduce some additional assumptions.

Let a device be given with ohmic contacts denoted as $\partial\Omega_{D,1}, \partial\Omega_{D,2}, \ldots$. We start from the equilibrium state. In this case, the quasi-Fermi levels are equal to the Fermi level ($F_n = F_p = \mu \equiv \text{const}$), and the bias is zero. We assume that there are no potential differences within a single ohmic contact.

Let us consider an ohmic contact $\partial\Omega_{D,1}$. In general, the electrostatic potential is relative and we can only measure the difference. Thus without loss of generality, we may assume that $\psi|_{\partial\Omega_{D,1}} = 0$. To determine Fermi level μ, we assume that the charge $\rho_{\partial\Omega_{D,1}} = 0$. To justify this assumption, we would like to argue that it is a natural state of the bulk semiconductor material, as discussed in section 2.5.5.

Thus, having $\psi|_{\partial\Omega_{D,1}} = 0$ we may determine $F_n = F_p = \mu$, such that $\rho|_{\partial\Omega_{D,1}} = 0$. The Fermi level is determined for the whole device, but due to different doping, other contacts in general will not be charge-neutral for zero electrostatic potential. Thus again from the charge-neutrality condition, we can determine values $\psi_{b,i}$, such that for $\psi|_{\partial\Omega_{D,i}} = \psi_{b,i}$ we have $\rho_{\partial\Omega_{D,i}} = 0$ for $i > 1$. In general, $\psi_{b,i}$ coincide with the built-in potential, introduced in section 2.5.5. If there are only two contacts, we will denote $\psi_b := \psi_{b,2}$. We will also define $\psi_{b,1} := 0$ to simplify the notation.

Thus we have the following conditions for the equilibrium state:

$$\begin{aligned} \psi|_{\partial\Omega_{D,i}} &= \psi_{b,i}, \\ F_n|_{\partial\Omega_{D,i}} &= \mu, \\ F_p|_{\partial\Omega_{D,i}} &= \mu. \end{aligned} \tag{2.5.32}$$

To derive analogous conditions in the general case, we want to emphasize two effects. First, the quasi-Fermi levels should be equal within a single contact, as in metal there is a single Fermi level. All these levels are equal in a given contact. Then note that the bias is not equal to differences between the electrostatic potential ψ on the contacts, but it is proportional to difference between Fermi levels in contacts.

Therefore let ψ_{D_i} for $i > 1$ denote the bias on contact i. For convenience we define $\psi_{D_1} = 0$. Then the boundary conditions for ohmic contacts read

$$\begin{aligned} \psi|_{\partial\Omega_{D,i}} &= \psi_{b,i} + \psi_{D,i}, \\ F_n|_{\partial\Omega_{D,i}} &= \mu - q^{-1}\psi_{D,i}, \\ F_p|_{\partial\Omega_{D,i}} &= \mu - q^{-1}\psi_{D,i}. \end{aligned} \tag{2.5.33}$$

2.6 Radiative and non-radiative recombination

2.6.1 Standard recombination models

In pure semiconductor material, electron energy levels are splitted into two bands, valence band with lower energies and conduction band with higher energies. In between lies so-called forbidden zone,

where no energy levels exist. When temperature goes to zero, all electrons tend to occupy low energy states, in valence band, which is filled completely. When temperature is raised, energies of electrons increase and some of them jump to conduction band.

Every electron jump to conduction band leaves an empty place in the valence band. Then, the electron can move spatially through almost empty conduction band. On the other hand, in valence band, another electron may move to the empty place left, creating another empty place. This is how a hole can move spatially in valence band. A movement of electron or holes may then, under application of voltage, create the current flowing through a device. Thus they are called carriers.

The case is slightly different when doped semiconductor is considered. Then, electron and hole numbers are increased due to contribution of donors and acceptors respectively. The conduction of current may be then due to mainly one type of carrier. Introduction of impurities may also create additional allowed energy states in the forbidden zone.

Generation-recombination phenomena in semiconductor denotes two physical processes of energy change of electrons:

Generation A jump of an electron from valence band to conduction band. Then, a hole in valence band is created.

Recombination A jump of an electron from conduction band to valence band. There must be a hole available, which is annihilated.

We may imagine that these phenomena could consist of single transfer of an electron from level in one band to level in another one, what is called direct generation/recombination, or several steps through energy states in forbidden zone, called indirect generation/recombination.

Note that generation involves an increase of energy of a given electron, therefore the energy must be taken from an incident photon or phonon. On the contrary, in the recombination process electron looses its energy, which may be emitted as photon or phonon. Since we are interested in simulations of luminescent devices, we are concerned about recombination mechanism.

There is a set of usually exploited recombination models, available in many computer simulators of semiconductor structures. Physical phenomena are represented rather coarsely, but resulting formulae are simple and computationally cheap. Therefore we present an explanations and description of basic recombination set.

For every generation/recombination mechanism short physical explanation and the formulae would be given. An important fact, which leads to elimination of excess coefficients, is the the detailed balance principle, which states that at equilibrium each process is balanced by its reverse process.

Presented analysis is based on the section 4.2 of the book [101]. For convenience, functions $n_0(x)$, $p_0(x)$ denote concentrations of electrons and holes in the equilibrium state. Then intrinsic concentration $n_i(x)$ is defined as

$$n_i(x) := \sqrt{n_0(x)p_0(x)}. \tag{2.6.1}$$

If Boltzmann statistics for carrier concentrations are valid, i.e. for nondegenerate case, we may rewrite n_i as

$$n_i(x) = \sqrt{N_c(x)N_v(x)} \exp\left(\frac{E_v(x) - E_c(x)}{2kT}\right). \tag{2.6.2}$$

The derivation of formulae is carried out in the phenomenological way.

It must be explained that we will use the term *electron* in two different meanings. When the electron is said to change its energy (jump between energy levels), the physical particle is meant. But when we say that electron is generated or annihilated, we tend to use semiconductor nomenclature, where electrons denote quasiparticles from the conduction band.

2.6.1.1 Radiative recombination

Radiative generation/recombination is a direct mechanism, i.e. creation/annihilation of electron-hole pair assisted only by photons. It consist of two processes. During radiative recombination, called also band to band or photon recombination, the excess energy is emitted as a photon. During generation, called also optical generation, necessary energy is taken from incident photon.

For radiative recombination to occur, there must be both electron and hole present in a given position of space. Therefore it is proportional to the product of concentrations $n(x)p(x)$. Thus the recombination rate is

$$r^{\mathrm{rad}}(x,n,p) := C^{\mathrm{rad}}(x)n(x)p(x). \tag{2.6.3}$$

We denote C^{rad} as a *capture coefficient.*

Then we assume that in the valence band there are always electrons available and there are empty states in the conduction band everywhere. Then optical generation process does not depend on any carrier concentration. Therefore we assume the rate is constant

$$g^{\mathrm{rad}}(x,n,p) := C_g^{\mathrm{rad}}(x), \tag{2.6.4}$$

where C_g^{rad} is called a *emission coefficient.*

Under equilibrium conditions we have $g^{\mathrm{rad}}(x,n_0,p_0) = r^{\mathrm{rad}}(x,n_0,p_0)$, thus

$$C_g^{\mathrm{rad}}(x) = C^{\mathrm{rad}}(x)n_0(x)p_0(x) = C^{\mathrm{rad}}(x)n_i^2(x). \tag{2.6.5}$$

Thus the generation/recombination rate is

$$R^{\mathrm{rad}}(x,n,p) = r^{\mathrm{rad}}(x,n,p) - g^{\mathrm{rad}}(x,n,p) = C^{\mathrm{rad}}(x)[n(x)p(x) - n_i^2(x)]. \tag{2.6.6}$$

Then capture coefficient for radiative recombination C^{rad} would be just called radiative recombination coefficient.

2.6.1.2 Shockley-Read-Hall recombination

Shockley-Read-Hall generation/recombination phenomena involve phonons, so may lead to change of the temperature of a device. It is indirect mechanism, occurring when trap levels are available. Trap levels are additional energy levels of energy in the bandgap. For simplicity it is assumed that there is only one important level given, and other ones may be neglected. If for a given position an electron occupies the trap level, we will say the trap is occupied, otherwise it is empty.

In the process, electrons move between the bands in two steps: first they jump onto the trap level, and then jump again to the second band. We will therefore consider four types of the phenomena:

Electron emission Jump of an electron to the conduction band from occupied trap. In other words, an electron in the conduction band is generated, and the trap becomes unoccupied.

Electron capture Jump of an electron to unoccupied trap from the conduction band. Thus an electron in the conduction band is annihilated and the trap becomes occupied.

Hole emission Jump of electron to unoccupied trap from the valence band. A hole in the valence band is generated and the trap becomes occupied.

Hole capture Jump of an electron into a hole in the valence band from occupied trap. A hole in the valence band is annihilated and the trap becomes unoccupied.

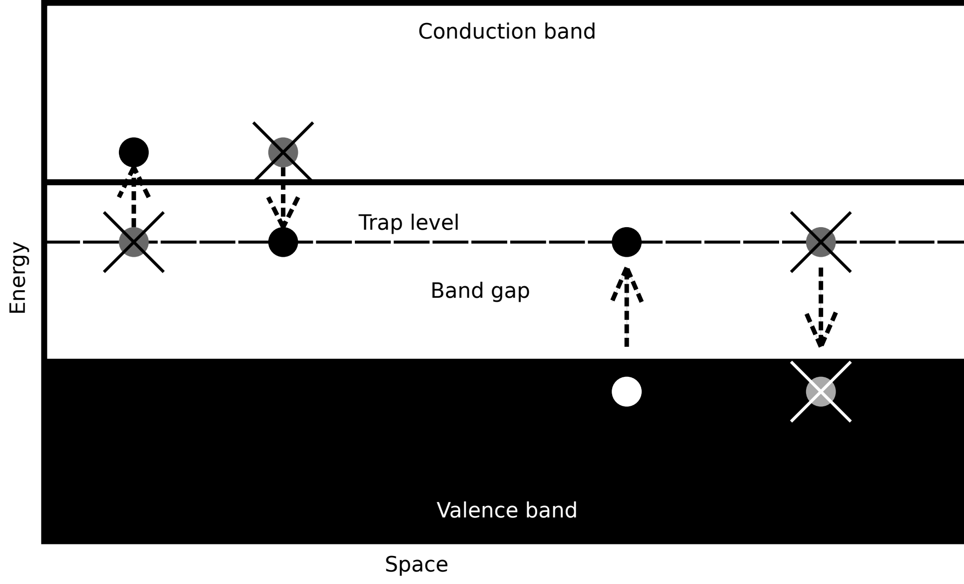

Figure 2.9: Schema of Shockley-Read-Hall generation/recombination. From the left: an electron emission and a capture, a hole emission and a capture.

In this section we derive simple mathematical formulae describing the given process. More sophisticated approach is presented in section 2.6.3.

Note that none of the above partial processes is a direct generation/recombination. For a generation to occur, both an electron emission and a hole emission must arise. Similarly a recombination means that the electron annihilation and hole annihilation occurred. The case then become complicated, as we cannot easily write partial recombination forms, like we did before for radiative or Auger recombination. However, we may try to establish rates for emissions and captures, what we denote by letters E and P, respectively.

Let f_t denote the fraction of occupied traps (by electrons). Then the fraction of unoccupied traps is $1 - f_t$. The rate of electron emission is proportional to rate of occupied traps f_t only, as we assume there are always unoccupied states in conduction band

$$E_n^{\mathrm{SRH}}(x, n, p, f_t) = C_{en}^{\mathrm{SRH}}(x) f_t(x). \tag{2.6.7}$$

Analogously the hole emission is proportional to the fraction of unoccupied traps

$$E_p^{\mathrm{SRH}}(x, n, p, f_t) = C_{ep}^{\mathrm{SRH}}(x)[1 - f_t(x)]. \tag{2.6.8}$$

For electron capture, both electron in the conduction band and unoccupied trap must be available. It is thus proportional to product $n(x)[1 - f_t(x)]$

$$P_n^{\mathrm{SRH}}(x, n, p, f_t) = C_{cn}^{\mathrm{SRH}}(x) n(x)[1 - f_t(x)]. \tag{2.6.9}$$

Similarly for the hole capture we may write

$$P_p^{\mathrm{SRH}}(x, n, p, f_t) = C_{cp}^{\mathrm{SRH}}(x) p(x) f_t(x). \tag{2.6.10}$$

Now we have four coefficients $C_{..}^{\mathrm{SRH}}$ and additional ratio f_t. We would like to reduce the number to two, and get rid of the ratio f_t.

The trap level states are so-called localized states [117], what means that the carriers present there cannot move in space. Now we can estimate the flux of the electrons arising from the conduction band

into the trap level in a given point as a difference between electron capture and electron emission rates $P_n^{\rm SRH}(x,n,p,f_t) - E_n^{\rm SRH}(x,n,p,f_t)$. Analogously for flux of electrons from trap level to valence band we have the difference $P_p^{\rm SRH}(x,n,p,f_t) - E_p^{\rm SRH}(x,n,p,f_t)$. Any of the above amounts being nonzero influences the fraction of occupied traps. In stationary case, the ratio f_t must be constant in time. Since the spatial movement of trapped electrons is not possible, both fluxes must balance. Therefore we have

$$P_n^{\rm SRH}(x,n,p,f_t) - E_n^{\rm SRH}(x,n,p,f_t) = P_p^{\rm SRH}(x,n,p,f_t) - E_p^{\rm SRH}(x,n,p,f_t). \tag{2.6.11}$$

From now on we will omit the arguments of functions. For transitional case, we should assume for example that time of the stabilization of the fraction f_t is much smaller than the time of essential changes of concentrations n,p to get similar conclusion.

The point of equation (2.6.11) is that if the electron jumps from conduction band to trap level, and do not go back, it will proceed to valence band. The same goes to electrons from valence band entering trap level. The sense of above statement is statistical, because in fact electrons stay on the trap level over nonzero time. It does not mean that time is diminished, but that one electron is going in, and another out, in approximately the same moment. However, since in drift-diffusion theory the electrons are not distinguishable, we may assume it was the same electron.

Then, since the expression $P_n^{\rm SRH} - E_n^{\rm SRH}$ controls all the electrons, as well as the expression $P_p^{\rm SRH} - E_p^{\rm SRH}$, we may write

$$R^{\rm SRH}(x,n,p,f_t) := P_n^{\rm SRH} - E_n^{\rm SRH} = P_p^{\rm SRH} - E_p^{\rm SRH}. \tag{2.6.12}$$

Under the equilibrium condition, the recombination is zero and we have

$$\begin{aligned}(P_n^{\rm SRH} - E_n^{\rm SRH})(x,n_0,p_0,f_{t0}) &\equiv 0, & (2.6.13)\\ (P_p^{\rm SRH} - E_p^{\rm SRH})(x,n_0,p_0,f_{t0}) &\equiv 0, & (2.6.14)\end{aligned}$$

where the index 0 means the function in the equilibrium state. Thus we may compute constants $C_{e\cdot}^{\rm SRH}$ to be

$$\begin{aligned}C_{en}^{\rm SRH} &= C_{cn}^{\rm SRH} n_0 \frac{1-f_{t0}}{f_{t0}} =: C_{cn}^{\rm SRH} n_1, & (2.6.15)\\ C_{ep}^{\rm SRH} &= C_{cp}^{\rm SRH} p_0 \frac{f_{t0}}{1-f_{t0}} =: C_{cp}^{\rm SRH} p_1. & (2.6.16)\end{aligned}$$

The functions n_1, p_1 are introduced to simplify the notation. Using the above we may calculate recombination rate twofold

$$\begin{aligned}R^{\rm SRH}(x,n,p,f_t) &= P_n^{\rm SRH} - E_n^{\rm SRH} = C_{cn}^{\rm SRH} n(1-f_t) - C_{cn}^{\rm SRH} n_1 f_t, & (2.6.17)\\ R^{\rm SRH}(x,n,p,f_t) &= P_p^{\rm SRH} - E_p^{\rm SRH} = C_{cp}^{\rm SRH} p f_t - C_{cp}^{\rm SRH} p_1(1-f_t). & (2.6.18)\end{aligned}$$

Comparing both terms allows to determine function f_t:

$$C_{cn}^{\rm SRH} n f_t + C_{cn}^{\rm SRH} n_1 f_t + C_{cp}^{\rm SRH} p f_t + C_{cp}^{\rm SRH} p_1 f_t = C_{cn}^{\rm SRH} n + C_{cp}^{\rm SRH} p_1, \tag{2.6.19}$$

$$f_t(x,n,p) = \frac{C_{cn}^{\rm SRH} n + C_{cp}^{\rm SRH} p_1}{C_{cn}^{\rm SRH}(n+n_1) + C_{cp}^{\rm SRH}(p+p_1)}. \tag{2.6.20}$$

Inserting f_t into equation (2.6.17) we obtain, after calculations

$$R^{\rm SRH}(x,n,p) = C_{cn}^{\rm SRH} C_{cp}^{\rm SRH} \frac{np - n_1 p_1}{C_{cn}^{\rm SRH}(n+n_1) + C_{cp}^{\rm SRH}(p+p_1)}, \tag{2.6.21}$$

or in the simplified form

$$R^{\mathrm{SRH}}(x,n,p) = \frac{np - n_1p_1}{\frac{n+n_1}{C_{cp}^{\mathrm{SRH}}} + \frac{p+p_1}{C_{cn}^{\mathrm{SRH}}}}. \tag{2.6.22}$$

Note that by definitions of functions n_1, p_1 (2.6.15),(2.6.16)

$$n_1p_1 = n_0p_0 = n_i^2. \tag{2.6.23}$$

Then, defining electron lifetime τ_n^{SRH} and hole lifetime τ_p^{SRH} as a reciprocals of respective capture rates

$$\tau_n^{\mathrm{SRH}} = \frac{1}{C_{cn}^{\mathrm{SRH}}}, \qquad \tau_p^{\mathrm{SRH}} = \frac{1}{C_{cp}^{\mathrm{SRH}}}. \tag{2.6.24}$$

we obtain the formula

$$R^{\mathrm{SRH}}(x,n,p) = \frac{np - n_0p_0}{\tau_p^{\mathrm{SRH}}(n+n_1) + \tau_n^{\mathrm{SRH}}(p+p_1)}. \tag{2.6.25}$$

Having this formula, one still have to compute f_{t0} to obtain fictitious concentrations n_1, p_1. On the other hand, for p-n diode structures, one may use the observation that in the depleted region, the SRH recombination is the highest and increasing bias increases concentrations n, p substantially, so n_1, p_1 can be neglected for the bias high enough. Thus previous formula reduces to

$$R^{\mathrm{SRH}}(x,n,p) = \frac{np - n_0p_0}{\tau_p^{\mathrm{SRH}}n + \tau_n^{\mathrm{SRH}}p}. \tag{2.6.26}$$

We will return to the form (2.6.25) in the section 2.6.3.

To derive expressions on n_1, p_1, let us use the Fermi-Dirac distribution to estimate the trap level occupation f_{t0}

$$f_{t0} := \frac{1}{1 + g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right)}, \tag{2.6.27}$$

where ψ_0 is an electric potential in the equilibrium state, E_t is the trap level energy and $g > 0$ is a degeneracy coefficient.

Thus using equation (2.6.15) we have

$$\begin{aligned} n_1 &= n_0\frac{1-f_{t0}}{f_{t0}} = n_0\frac{g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right)}{1+g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right)}\left[1+g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right)\right] \\ &= n_0 g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right). \end{aligned} \tag{2.6.28}$$

Using definitions (2.1.24), (2.1.29) we obtain

$$n_0 = N_c\exp\left(\frac{\mu - E_c + q\psi_0}{k_BT}\right), \quad p_0 = N_v\exp\left(\frac{E_v-\mu-q\psi_0}{k_BT}\right), \tag{2.6.29}$$

so

$$n_1 = gN_c\exp\left(\frac{\mu-E_c+q\psi_0+E_t-\mu-q\psi_0}{k_BT}\right) = gN_c\exp\left(\frac{E_t-E_c}{k_BT}\right). \tag{2.6.30}$$

Analogously using equation (2.6.16) we obtain

$$\begin{aligned} p_1 &= p_0\frac{f_{t0}}{1-f_{t0}} = p_0\frac{1}{1+g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right)}\frac{1+g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right)}{g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right)} = p_0\frac{1}{g\exp\left(\frac{E_t-\mu-q\psi_0}{k_BT}\right)} \\ &= p_0g^{-1}\exp\left(\frac{\mu-E_t+q\psi_0}{k_BT}\right) = g^{-1}N_v\exp\left(\frac{E_v-E_t}{k_BT}\right). \end{aligned} \tag{2.6.31}$$

To sum up, we have the following formulas on n_1, p_1:

$$n_1(x) := g(x)N_c(x)\exp\Big(\frac{E_t(x)-E_c(x)}{k_BT}\Big), \quad p_1(x) := g^{-1}(x)N_v(x)\exp\Big(\frac{E_v(x)-E_t(x)}{k_BT}\Big), \tag{2.6.32}$$

where g is the degeneracy coefficient and E_t is the energy of the trap level (absolute).

2.6.2 Impact ionization

The impact ionization phenomena we will describe in limited length, as it is not so important mechanism in luminous structures. However, devices exploiting that effect, like thyristors, are typical example for experimental evidence of a nonuniqueness of stationary solutions of the drift-diffusion system. Thus the mechanism is worth of mentioning.

The impact ionization denotes the phenomena of jumping electrons from valence to conduction band due to heavy electric field $\nabla\psi$ and currents J_n, J_p. There are two cases considered:

Electron assisted generation An electron from valence band gains energy and jumps into conduction band, energy is taken from incidental electron laying in conduction band.

Hole assisted generation Like above, but energy is taken from incidental hole in valence band.

Above descriptions are analogous as for Auger recombination, but there is no recombination mechanisms. The effect consists of generation only. The physical explanation is that under strong enough electric field and heavy current, new carriers are generated, what leads to lower resistance.

The rates are

$$\begin{aligned} G_n^{\mathrm{II}}(x, J_n, E) &= \alpha_n^{\mathrm{II}}(x)\frac{|J_n(x)|}{q}\exp\Big(-\frac{E_{n,crit}^{\mathrm{II}}(x)}{E(x)}\Big), \\ G_p^{\mathrm{II}}(x, J_p, E) &= \alpha_p^{\mathrm{II}}(x)\frac{|J_p(x)|}{q}\exp\Big(-\frac{E_{p,crit}^{\mathrm{II}}(x)}{E(x)}\Big), \end{aligned} \tag{2.6.33}$$

where parameters $\alpha_{n,p}^{\mathrm{II}}$ are called *maximal ionization rates* and *critical field strengths* $E_{n,p}^{\mathrm{II}}$. The function $E(x)$ denotes the electric field component in the direction of current flow. The formulation above is so-called *lucky drift model* [74, 103]. More sophisticated formulations can be found in [101].

Therefore the formula for the recombination component is

$$\begin{aligned} R^{\mathrm{II}}(x, E, J_n, J_p) = -G_n^{\mathrm{II}} - G_p^{\mathrm{II}} &= -\alpha_n^{\mathrm{II}}\frac{|J_n|}{q}\exp\Big(-\frac{E_{n,crit}^{\mathrm{II}}}{E}\Big) - G_p^{\mathrm{II}}(x, J_p, E) \\ &= \alpha_p^{\mathrm{II}}\frac{|J_p|}{q}\exp\Big(-\frac{E_{p,crit}^{\mathrm{II}}}{E}\Big). \end{aligned} \tag{2.6.34}$$

As there is no recombination, we cannot eliminate constants like we did for another recombinations. For equilibrium conditions the value is zero, since carrier currents $J_{n0}, J_{p0} \equiv 0$.

In above notation we indicated the dependence of recombination R^{II} on variables E, J_n, J_p. These functions can be calculated from variables ψ, n, p, but that leads to differentiation and solution of linear differential equation. Thus, computationally it is much more costly than other recombinations presented in the document.

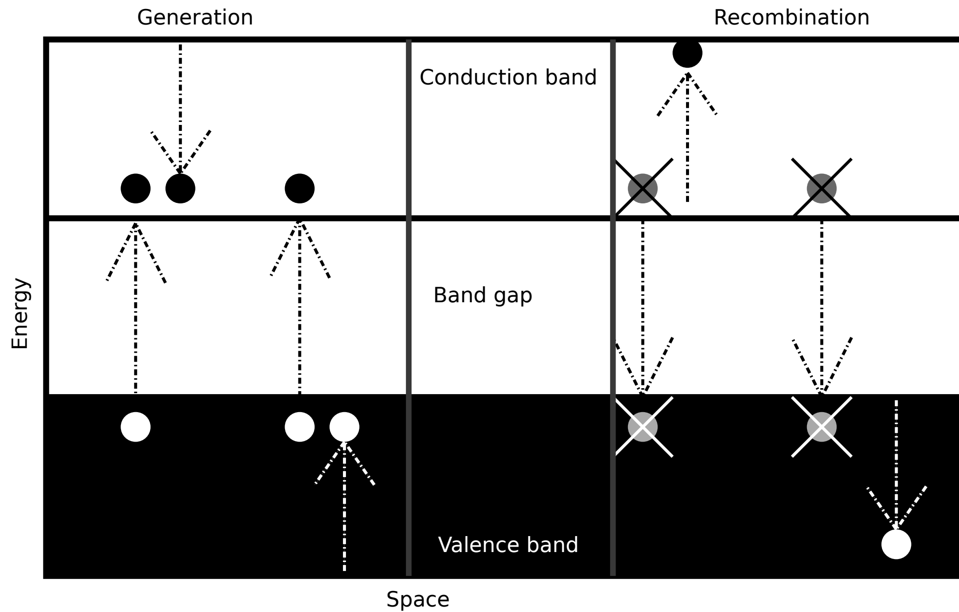

Figure 2.10: Schema of Auger generation/recombination. From the left: an electron and a hole assisted generations, then respective recombinations.

2.6.2.1 Auger recombination

The nature of the Auger effect is as follows. Assume that there is a electron vacancy in the inner shell of a given atom. Normally this vacancy may be filled by an electron falling from a higher energy level, where the energy difference is released as a photon. On the other hand, this excessive energy may be transferred to the outer-orbit electron, which is then ejected from the atom. The latter phenomena is known as the *Auger effect* [22].

The term *Auger recombination* describes the phenomena of jump of a electron through the band gap, where the excessive energy is transferred to the third carrier, energy or hole (see figure 2.10). This carrier then thermalizes generating phonons.

Therefore we will consider four possibilities:

Electron assisted generation An electron from valence band gains energy and jumps into conduction band, energy is taken from some electron in the conduction band.

Hole assisted generation Like above, but energy is taken from some hole in the valence band.

Electron assisted recombination An electron from conduction band comes down to valence band, where a hole is annihilated. Excess energy is passed on incidental electron in conduction band.

Hole assisted recombination As before, but excess energy is passed on incident hole in valence band.

It must be noted that the transition of a hole in the energy landscape is counter-intuitive. A hole gaining the energy in fact means, that some valence band electron gained the energy and occupied empty space above its initial position, leaving new empty space. Thus the hole energy diminish, even though it have gained the energy. Similarly a hole losing energy means an electron losing energy, so the hole is going up on the energy axis.

The cases are now similar to radiative recombination, but slightly more complicated. Let us begin from *electron assisted recombination.* To occur, there must be an electron and a hole to recombine,

and additional electron to take the energy. The rate is therefore proportional to $n^2(x)p(x)$. Similarly for *hole assisted recombination* there must be two holes and an electron, so the rates are:

$$R_n^{\mathrm{Aug}}(x,n,p) = C_n^{\mathrm{Aug}}(x)n^2(x)p(x), \qquad R_p^{\mathrm{Aug}}(x,n,p) = C_p^{\mathrm{Aug}}(x)n(x)p^2(x). \tag{2.6.35}$$

For generation the cases are more simple, because of the assumption that there is always an electron available in the valence band and empty space in conduction band, too. Thus only necessary component is the carrier which loses the energy. We have then

$$G_n^{\mathrm{Aug}}(x,n,p) = C_{gn}^{\mathrm{Aug}}(x)n(x), \qquad G_p^{\mathrm{Aug}}(x,n,p) = C_{gp}^{\mathrm{Aug}}(x)p(x). \tag{2.6.36}$$

Due to the detailed balance principle, in the equilibrium case generations must be in balance with respective recombinations, thus

$$C_{gn}^{\mathrm{Aug}}(x) = C_n^{\mathrm{Aug}}(x)n_i^2(x), \qquad C_{gp}^{\mathrm{Aug}}(x) = C_p^{\mathrm{Aug}}(x)n_i^2(x). \tag{2.6.37}$$

Therefore by summing up all the rates we obtain the formula for Auger recombination

$$\begin{aligned} R^{\mathrm{Aug}}(x,n,p) &= (R_n^{\mathrm{Aug}} - G_n^{\mathrm{Aug}} + R_p^{\mathrm{Aug}} - G_p^{\mathrm{Aug}})(x,n,p) \\ &= [C_n^{\mathrm{Aug}}(x)n(x) + C_p^{\mathrm{Aug}}(x)p(x)][n(x)p(x) - n_i^2(x)], \end{aligned} \tag{2.6.38}$$

where C_n^{Aug} and C_p^{Aug} are called *Auger capture coefficients* for electrons and holes, respectively.

2.6.3 Trap levels

The generation-recombination on trap levels, called also Shockley-Read-Hall recombination, is substantial for modeling optoelectronic devices. The mechanism was described by Shockley and Read [104], and by Hall [48]. The derivation presented in section 2.6.1.2 leads to most simple modifications of drift-diffusion system. We want, however, to gain deeper insight of the physical phenomena and to grasp extensions of classic formulae, available in the literature.

Better understanding of mechanisms governing the processes of SRH recombination should lead to vital improvement of simulation results, as the mentioned recombination seems to be the biggest in magnitude for certain devices. Nonetheless more complex formulae causes loss of computational efficiency and increase of simulation's time. Therefore it is advantageous to know which effects have to be considered and in what extend.

2.6.3.1 Occupation of trap levels

Let us start from the pure semiconductor situation, far from the boundary. We have then crystalline structure with given number of electrons and holes, contributing to the charge. Now we replace an atom with impurity atom. We consider two cases, most frequently used. Introduced atom may have one valence electron more or less than the original one. Therefore we have two possibilities

Electron trap Introduced atom has additional electron. If the electron is present on its level, so the trap is occupied, the atom remains neutral, as original atom would be. Otherwise, if the trap is unoccupied, the core charge is unbalanced and adds q to the total charge. Such a impurity is called a donor, as it may give additional electron, and it is of neutral or positive charge.

Hole trap Introduced atom has one electron less. If the electron is absent, the atom is neutral, but still there is a level for additional electron. If it is present, the trap level contributes $-q$ to the charge. The impurity is called an acceptor, as it takes electrons. It may be of neutral or negative charge.

The latter case may be explained also using the concept of hole. The atom has a level unoccupied by electron, we say that it is occupied by a hole, and so on. Then, analogously as for the first case, if trap is occupied by a hole, it is neutral, and it becomes negative if unoccupied.

In the section 2.6.1.2 we have derived the occupation rate of a given trap level to be

$$f_t(x,n,p) = \frac{C_{cn}^{SRH} n + C_{cp}^{SRH} p_1}{C_{cn}^{SRH}(n+n_1) + C_{cp}^{SRH}(p+p_1)}. \tag{2.6.39}$$

The above value is defined as fraction of traps which are occupied by electrons. Thus the number of unoccupied traps, or traps occupied by holes is

$$1 - f_t(x,n,p) = \frac{C_{cp}^{SRH} p + C_{cn}^{SRH} n_1}{C_{cn}^{SRH}(n+n_1) + C_{cp}^{SRH}(p+p_1)}. \tag{2.6.40}$$

Then, the standard nomenclature is as follows. When a hole (or electron) trap is said to be occupied, it is occupied by hole (or electron, respectively). The term "occupation" is also considered in the sense of hole (electron, respectively). Thus the occupation rate for hole trap level is $1 - f_t$, and for electron level it is f_t.

Using the above reasoning we may write the number of unoccupied traps of both types, giving the contribution to the charge:

$$N_t^+(x,n,p) := N_t(x)(1 - f_t(x,n,p)), \qquad N_t^-(x,n,p) := N_t(x) f_t(x,n,p). \tag{2.6.41}$$

Using equations (2.6.39), (2.6.40) we finally obtain

$$N_t^+(x,n,p) = N_t \frac{C_{cp}^{SRH} p + C_{cn}^{SRH} n_1}{C_{cn}^{SRH}(n+n_1) + C_{cp}^{SRH}(p+p_1)}, \tag{2.6.42}$$

$$N_t^-(x,n,p) = N_t \frac{C_{cn}^{SRH} n + C_{cp}^{SRH} p_1}{C_{cn}^{SRH}(n+n_1) + C_{cp}^{SRH}(p+p_1)}. \tag{2.6.43}$$

By plus sign we denote electron traps, and hole traps by minus, up to the sign of a charge contribution they give.

2.6.3.2 Shallow and deep levels

Now when we have precise formulae for concentration of so-called ionized traps (2.6.41), we would like to establish possible simplifications. Let us begin form the case of electron traps. By equations (2.6.39), (2.6.41), the concentration of ionized electron traps read

$$N_t^+(x,n,p) := \frac{C_{cp}^{SRH} p + C_{cn}^{SRH} n_1}{C_{cn}^{SRH}(n+n_1) + C_{cp}^{SRH}(p+p_1)}. \tag{2.6.44}$$

When a trap level is near the middle of the forbidden zone, the probability of an electron jump between the level and any of bands is approximately the same. Thus the capture coefficients are similar. However, when the level gets closer to the conduction band, the electron capture becomes more plausible, on the contrary to hole capture. Thus the coefficient C_{cn}^{SRH} increases, while the coefficient C_{cp}^{SRH} is getting small. For electron trap level close enough, we may assume $C_{cp}^{SRH} \approx 0$. Thus we have

$$N_t^+(x,n,p) \approx \frac{C_{cn}^{SRH} n_1}{C_{cn}^{SRH}(n+n_1)} = \frac{1}{n \cdot n_1^{-1} + 1}. \tag{2.6.45}$$

Then, by use of definition of concentration n and n_1 (2.6.32), we obtain

$$\begin{aligned} N_t^+(x,n,p) &\approx \frac{1}{1+g^{-1}(x)N_c^{-1}(x)\exp\left(\frac{-E_t^{SRH}(x)+E_c(x)}{kT}\right)\cdot N_c(x)\exp\left(\frac{F_n(x)-E_c(x)+q\psi(x)}{kT}\right)} \\ &= \frac{1}{1+g^{-1}(x)\exp\left(\frac{F_n(x)-E_t^{SRH}(x)+q\psi(x)}{kT}\right)}. \end{aligned} \tag{2.6.46}$$

Analogous derivation we may present for hole traps, when the level is close to the valence band. Then $C_{cn}^{SRH} \approx 0$ and we obtain

$$N_t^-(x,n,p) := N_t(x)\frac{C_{cn}^{SRH}n + C_{cp}^{SRH}p_1}{C_{cn}^{SRH}(n+n_1)+C_{cp}^{SRH}(p+p_1)} \approx \frac{C_{cp}^{SRH}p_1}{C_{cp}^{SRH}(p+p_1)} = \frac{1}{p\cdot p_1^{-1}+1}. \tag{2.6.47}$$

By definitions of functions p and n_1 (2.6.32) then

$$\begin{aligned} N_t^-(x,n,p) &\approx N_t(x)\frac{1}{1+g(x)N_v^{-1}(x)\exp\left(\frac{-E_v(x)+E_t^{SRH}(x)}{kT}\right)\cdot N_v(x)\exp\left(\frac{E_v(x)-F_p(x)-q\psi(x)}{kT}\right)} \\ &= N_t(x)\frac{1}{1+g(x)\exp\left(\frac{E_t^{SRH}(x)-F_p(x)-q\psi(x)}{kT}\right)}. \end{aligned} \tag{2.6.48}$$

Now we may explain the origin of formulae for singly ionized donor and acceptor concentrations N_d^+, N_a^-, by substitution $E_t \leftarrow E_c - E_d$ in equation (2.6.46) and $E_t \leftarrow E_v + E_a$ in equation (2.6.46), respectively.

The above reasoning also reveals, that the recombination for shallow trap levels is small and insignificant, since jump between any band and the level should occur for generation/recombination to take place.

2.6.3.3 Estimation of carrier lifetimes

In section 2.6.1.2 we introduced, for a given trap level, electron lifetime τ_n^{SRH} and hole lifetime τ_p^{SRH} as reciprocals of capture rates of these carriers. In this section, we would like to estimate these lifetimes for a given trap concentration N_t and temperature T.

Let us assume that electrons and holes behave more or less like particles of the ideal gas. Due to the law of equipartition, the kinetic energy E_{kin} of a particle of the ideal gas of mass m and velocity v is given by

$$E_{\mathrm{kin}} = \frac{m|v|^2}{2} = \frac{3}{2}k_BT. \tag{2.6.49}$$

Thus

$$|v| = \sqrt{\frac{k_BT}{m}}. \tag{2.6.50}$$

Let us fix some time τ and assume that κ is a capture cross section. Then the volume traversed by the particle during time τ is equal to $\kappa|v|\tau$. If the time τ is small enough, we may assume that these volumes for the ideal gas particles do not overlap. Assume that there is N particles in the ideal gas. Thus the total volume traversed by all particles is given by

$$V := \kappa|v|\tau N. \tag{2.6.51}$$

If the concentration of traps is N_t, then the number of collisions of particles with the traps may be roughly estimated as

$$N_{\text{coll}} := N_t V = N_t \kappa |v| \tau N. \tag{2.6.52}$$

Now let us fix on electron quasiparticles. If we fix some small volume $\tilde{V}$, then we can estimate number of electrons in this volume as $n\tilde{V}$. Thus we can estimate number of electron captures in volume $\tilde{V}$ in time τ as

$$P_n^{\text{SRH}} \tilde{V} \tau = N_{\text{coll}}[1 - f_t]. \tag{2.6.53}$$

To capture an electron, a trap must be unoccupied before a collision, thus we have to multiply number of collisions by the coefficient $1 - f_t$. Here we assume that the time τ and the volume $\tilde{V}$ are both small enough so that re-emission and change of trap occupancy are negligible. Therefore we have

$$P_n^{\text{SRH}} = \frac{N_{\text{coll}}[1 - f_t]}{\tilde{V}\tau} = \frac{N_t \kappa_n |v_n| \tau n \tilde{V}[1 - f_t]}{\tilde{V}\tau} = N_t \kappa_n |v_n| n [1 - f_t]. \tag{2.6.54}$$

If we compare this result with the estimate of electron capture rate (2.6.9), we get

$$P_n^{\text{SRH}} = C_{cn}^{\text{SRH}} n[1 - f_t] = N_t \kappa_n |v_n| n [1 - f_t]. \tag{2.6.55}$$

Thus C_{cn}^{SRH} is given by

$$C_{cn}^{\text{SRH}} = N_t \kappa_n |v_n|. \tag{2.6.56}$$

By analogous analysis for holes and by estimate (2.6.9) we obtain

$$P_p^{\text{SRH}} = C_{cp}^{\text{SRH}} p f_t = N_t \kappa_p |v_p| p f_t, \tag{2.6.57}$$

so

$$C_{cp}^{\text{SRH}} = N_t \kappa_p |v_p|. \tag{2.6.58}$$

We would like to roughly estimate coefficients $C_{cn}^{\text{SRH}}, C_{cp}^{\text{SRH}}$ for a gallium nitride material in room temperature (300 K). Let us assume that a trap can capture a carrier within distance comparable to the primitive cell diameter. Let us assume this distance to be equal to a lattice parameter (see table 2.3). Then the cross section is given as

$$\kappa_n = \pi a^2 = \pi (0.32\,\text{nm})^2 \approx 3.22 \times 10^{-19}\,\text{m}^2. \tag{2.6.59}$$

Also using (2.6.50)

$$|v_n| = \sqrt{\frac{k_B T}{m_n m_0}} \approx \sqrt{\frac{1.38 \times 10^{-23}\,\text{JK}^{-1} \times 300\,\text{K}}{0.2 \times 9.11 \times 10^{-31}\,\text{kg}}} \approx 1.51 \times 10^5\,\text{ms}^{-1}. \tag{2.6.60}$$

Assume that the concentration of traps is $N_t = 1 \times 10^{16}\,\text{cm}^{-3} = 1 \times 10^{22}\,\text{m}^{-3}$. Then

$$C_{cn}^{\text{SRH}} = N_t \kappa_n |v_n| \approx 1 \times 10^{22}\,\text{m}^{-3} \times 3.22 \times 10^{-19}\,\text{m}^2 \times 1.51 \times 10^5\,\text{ms}^{-1} \approx 4.86 \times 10^8\,\text{s}^{-1}. \tag{2.6.61}$$

Therefore

$$\tau_n^{\text{SRH}} = \frac{1}{C_{cn}^{\text{SRH}}} = 2.06 \times 10^{-9}\,\text{s}. \tag{2.6.62}$$

For holes we assume the cross section is the same, i.e. $\kappa_n = \kappa_p$. Then

$$|v_p| = \sqrt{\frac{k_B T}{m_p m_0}} \approx \sqrt{\frac{1.38 \times 10^{-23}\,\text{JK}^{-1} \times 300\,\text{K}}{1.7 \times 9.11 \times 10^{-31}\,\text{kg}}} \approx 5.17 \times 10^4\,\text{ms}^{-1}. \tag{2.6.63}$$

Then

$$C_{cp}^{\text{SRH}} = N_t \kappa_p |v_p| \approx 1 \times 10^{22}\,\text{m}^{-3} \times 3.22 \times 10^{-19}\,\text{m}^2 \times 5.17 \times 10^4\,\text{ms}^{-1} \approx 1.66 \times 10^8\,\text{s}^{-1}, \tag{2.6.64}$$

and

$$\tau_p^{\text{SRH}} = \frac{1}{C_{cp}^{\text{SRH}}} = 6.01 \times 10^{-9}\,\text{s}. \tag{2.6.65}$$

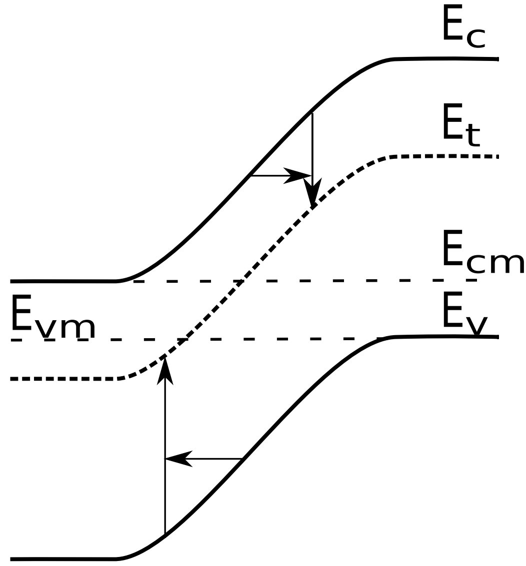

Figure 2.11: Schema of tunneling to the trap level. Transmission of electrons in the space proceeds along horizontal arrows, and then in the energy along vertical arrows.

2.7 Tunneling quantum effect

2.7.1 Trap-assisted tunneling

The case of a tunneling described here quite particular, as it only concerns modification of trap tunneling in strong electric field of certain shape. Such an approach is presented in article by Hurkx et al. [51]. The derivation of results presented there is not so straightforward, so here we just describe, how to modify already defined formulae to include tunneling effect.

The conditions under which an local modification of recombination form is valid, are severe and some of them are obviously broken, like linearity of electrostatic potential and constant quasi-Fermi levels. However, we assume that in the extent of a usage of tunneling modification, such conditions are at least approximately good.

The theory covers a case similar to shown on picture 2.11. It is assumed that quasi-Fermi levels and potential are approximately such as for basic p-n diode, at least locally. By E_{cm} and E_{vm} we denote boundary energy levels from which tunneling is possible.

Tunneling to the trap level may be then considered as a enhancement of standard SRH formula (2.6.25), with electron and hole lifetimes τ_n^{SRH}, τ_p^{SRH} multiplied by coefficients:

$$\tau_n^{tun}(x) = \frac{\tau_n^{SRH}(x)}{1+\Gamma_n(x)}, \qquad \tau_p^{tun}(x) = \frac{\tau_p^{SRH}(x)}{1+\Gamma_p(x)}. \tag{2.7.1}$$

Then functions Γ_n and Γ_p are defined as follows

$$\Gamma_{n,p}(x) = \frac{D_{n,p}(x)}{kT} \int_0^1 \exp\left(\frac{D_{n,p}(x)}{kT} u - K_{n,p}(x) u^{\frac{3}{2}}\right) \mathrm{d}u, \tag{2.7.2}$$

where

$$K_{n,p}(x) = \frac{4\sqrt{2m_{n,p}^{tun}(x) D_{n,p}^3(x)}}{3q\hbar|\psi'(x)|}. \tag{2.7.3}$$

The functions K_n and K_p are defined for one-dimensional case. In above equation $m_n^{tun}(x)$ and $m_p^{tun}(x)$ denote effective masses of electrons and holes for tunneling effect, which may be different than respective masses in general. Function ψ' denotes the electrostatic field. Functions D_n and D_p represent the size of a range of energies, from which the tunneling is possible. For an electron, it reads (compare Pic. 2.11)

$$D_n(x) = \begin{cases} E_c(x) - E_{cm} & E_t(x) \leq E_{cm}, \\ E_c(x) - E_t(x) & E_t(x) > E_{cm}. \end{cases} \tag{2.7.4}$$

Therefore it is assumed that the tunneling is possible only from energies laying in the conduction band, above the trap level. Similarly for holes we have

$$D_p(x) = \begin{cases} E_{vm} - E_v(x) & E_t(x) > E_{vm}, \\ E_t(x) - E_v(x) & E_t(x) \leq E_{vm}. \end{cases} \tag{2.7.5}$$

Having relations (2.7.1) and (2.6.24), we may also compute capture rates for tunneling modification:

$$C_{n,p}^{tun} := \frac{1}{\tau_{n,p}^{tun}} = \frac{1 + \Gamma_{n,p}(x)}{\tau_n^{SRH}(x)} = C_{n,p}^{SRH}(1 + \Gamma_{n,p}(x)). \tag{2.7.6}$$

Therefore we may write occupied trap concentration

$$N_t^+(x, n, p) = N_t \frac{C_{cp}^{tun} p + C_{cn}^{tun} n_1}{C_{cn}^{tun}(n + n_1) + C_{cp}^{tun}(p + p_1)}, \tag{2.7.7}$$

$$N_t^-(x, n, p) = N_t \frac{C_{cn}^{tun} n + C_{cp}^{tun} p_1}{C_{cn}^{tun}(n + n_1) + C_{cp}^{tun}(p + p_1)}, \tag{2.7.8}$$

where sign index denotes sign of unoccupied trap. Then recombination rate is

$$R^{tun}(x, n, p) = \frac{np - n_0 p_0}{\tau_p^{tun}(n + n_1) + \tau_n^{tun}(p + p_1)}. \tag{2.7.9}$$

For shallow donors, the tunneling approach described here does not change any formula, as recombination is assumed to be zero and neither capture rates nor lifetimes are present in occupied trap concentrations.

2.8 P-N diode

The p-n homojunction (see section 2.3.1) is an elementary device in semiconductor electronics. While it mostly does not generate the light alone, it is a basis for design of electroluminescent diodes or laser diodes. In these devices, the luminescence comes from the radiative recombination of the electrons and holes. This recombination takes place in active region, in quantum wells (see section 2.4), which confine these carriers on small volume and allow them to recombine.

Through their way to the active zone, electrons and holes are subject to nonradiative recombination, to rogue radiative recombination and to energy loss. These effects can significantly reduce efficiency of semiconductor devices. In particular, electrons (resp. holes) traversing regions abundant in holes (resp. electrons) generate big losses due to recombination.

It is therefore beneficial to put an active region of a device between n-type region and p-type region. Then the recombining carriers can be provided through the appropriate regions to minimize the rogue recombination, i.e. electrons through n-type region and holes through p-type region.

Therefore to explain many physical mechanisms governing operation of the LEDs and LDs, the phenomena taking place in the p-n homojunctions must be carefully studied. Therefore we start the simulations with the p-n homojunctions. Then we move on to more complicated setting.

Apart from section 2.8.3, all simulations are performed with software `pmicro`, developed by us.

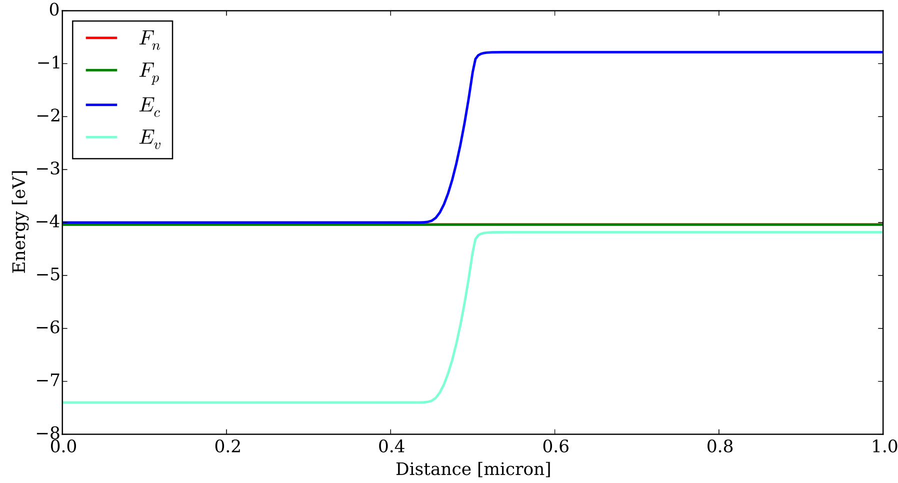

Figure 2.12: Band diagram of a GaN p-n homojunction in equilibrium case. Quasi-Fermi electron level F_n and quasi-Fermi hole level F_p coincide to the Fermi level, thus they are equal.

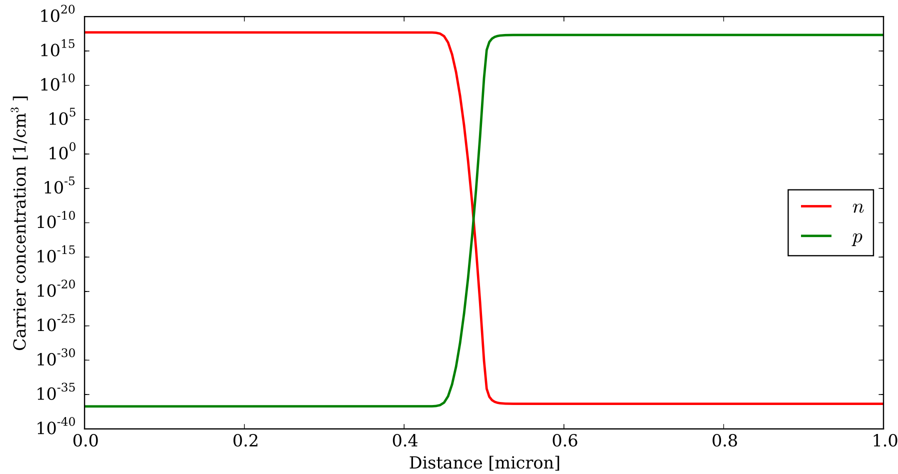

Figure 2.13: Concentration of electrons n and holes p in a GaN p-n homojunction in equilibrium case.

2.8.1 p-n homojunctions

Simulations of the p-n homojunctions reveal basic features of the semiconductor devices. We focus on the materials based on nitrides, as they are main compound of the blue and green optoelectronic devices.

We start with the p-n homojunction, composed of 500 nm GaN n-type region doped with $1 \times 10^{18}\,\mathrm{cm}^{-3}$ donors, and 500 nm GaN p-type region doped with $1 \times 10^{19}\,\mathrm{cm}^{-3}$.

In figures 2.12, 2.13 we see the band diagram and carrier concentrations of this device in equilibrium state. The potential barrier in the middle of the band diagram coincides with the depletion region. It is clearly visible, as the total carrier concentration in this region is low. As shown in figure 2.14, width of a depleted region is dependent on the level of the doping. Lower doping results in wider depleted region.

The main property of a diode is that generally it conducts only in one direction. In semiconductor diodes, we distinguish between *forward bias*, which corresponds to that direction, and *reverse bias* in

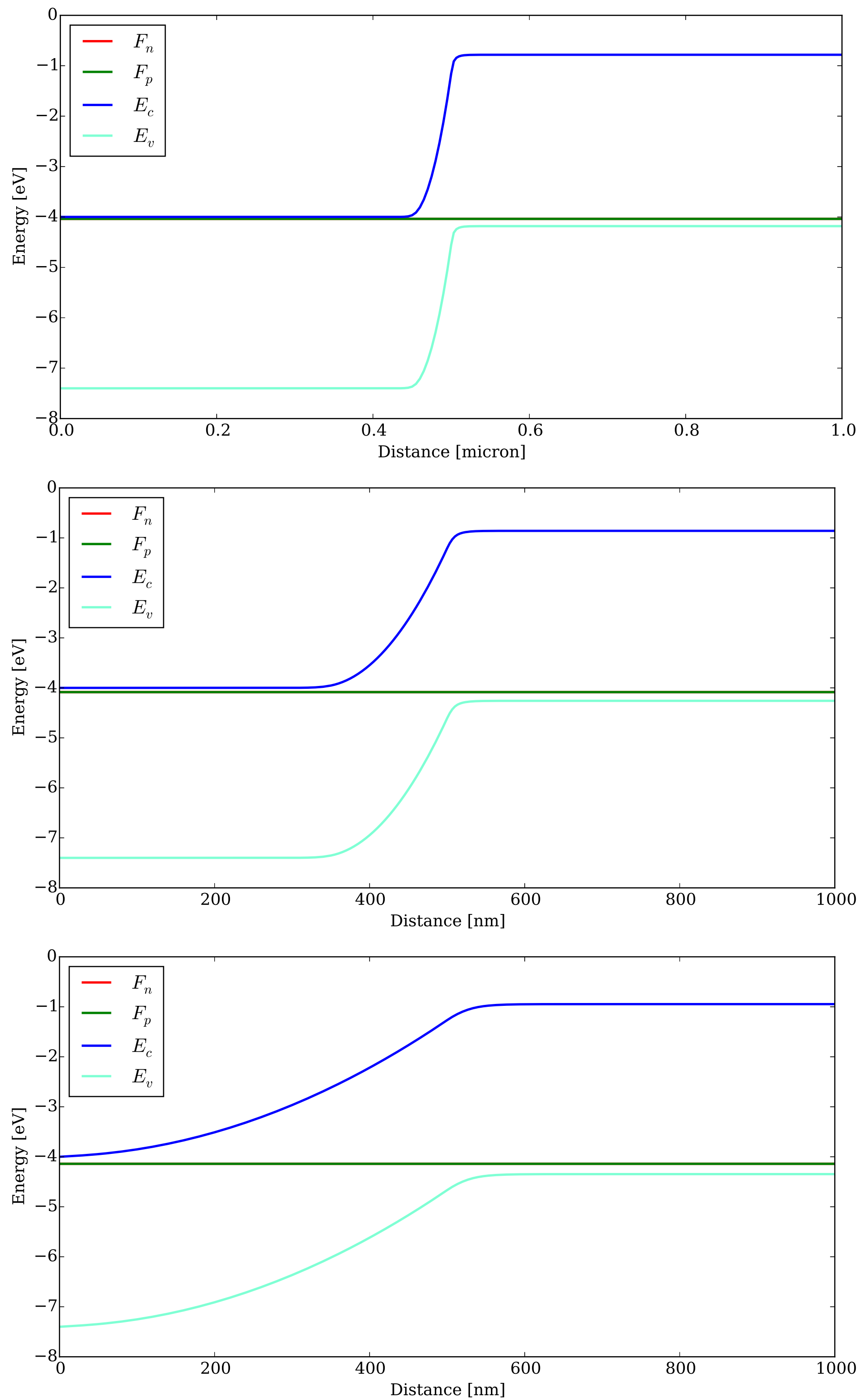

Figure 2.14: Comparison of band diagram from figure 2.12 with band diagrams of a GaN p-n homojunctions in equilibrium case with doping levels 10 times lower and 100 times lower. Quasi-Fermi electron level F_n and quasi-Fermi hole level F_p coincide to the Fermi level, thus they are equal.

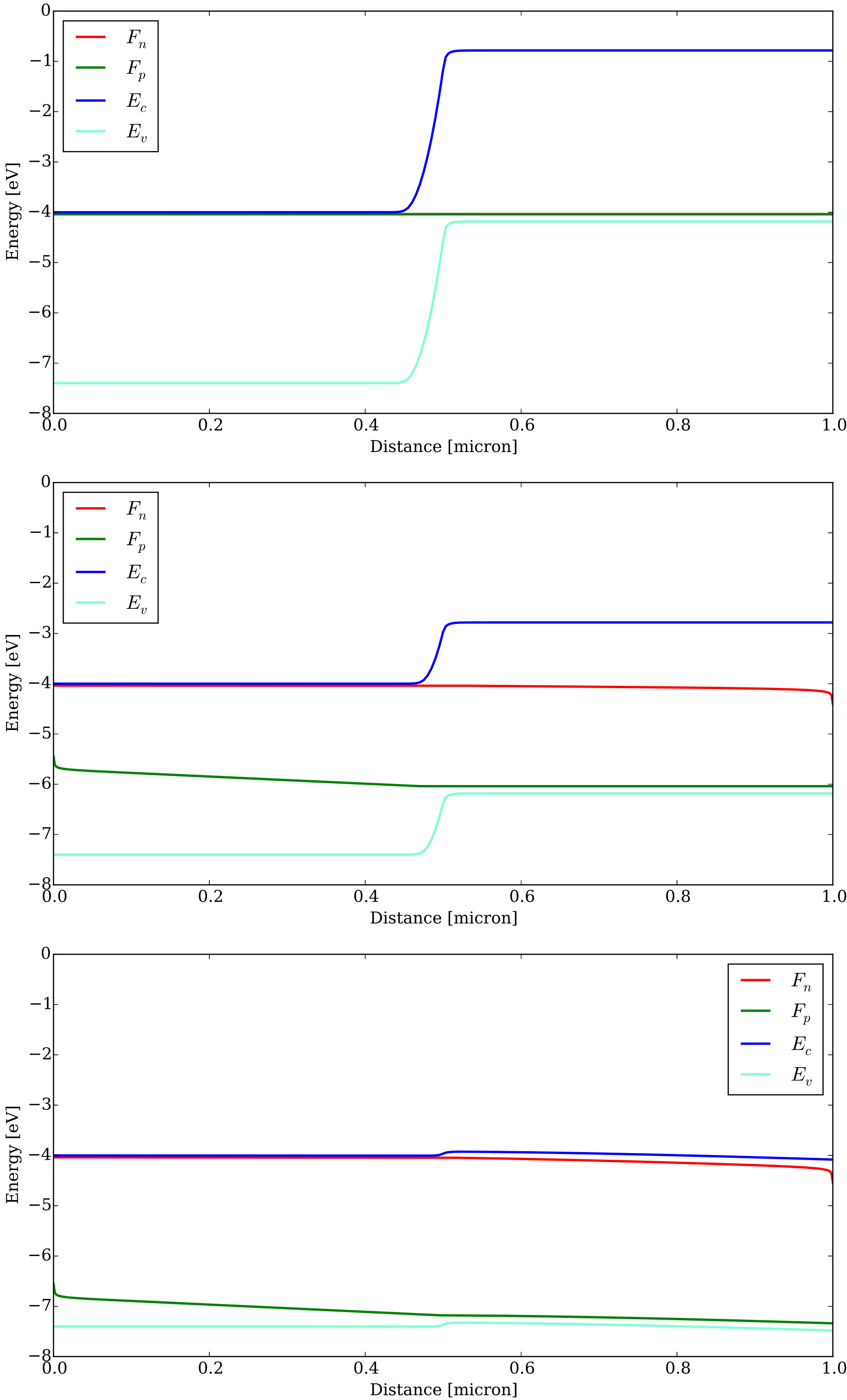

Figure 2.15: Comparison of band diagram of p-n homojunction for forward bias: 0 V, 2 V and 3.3 V.

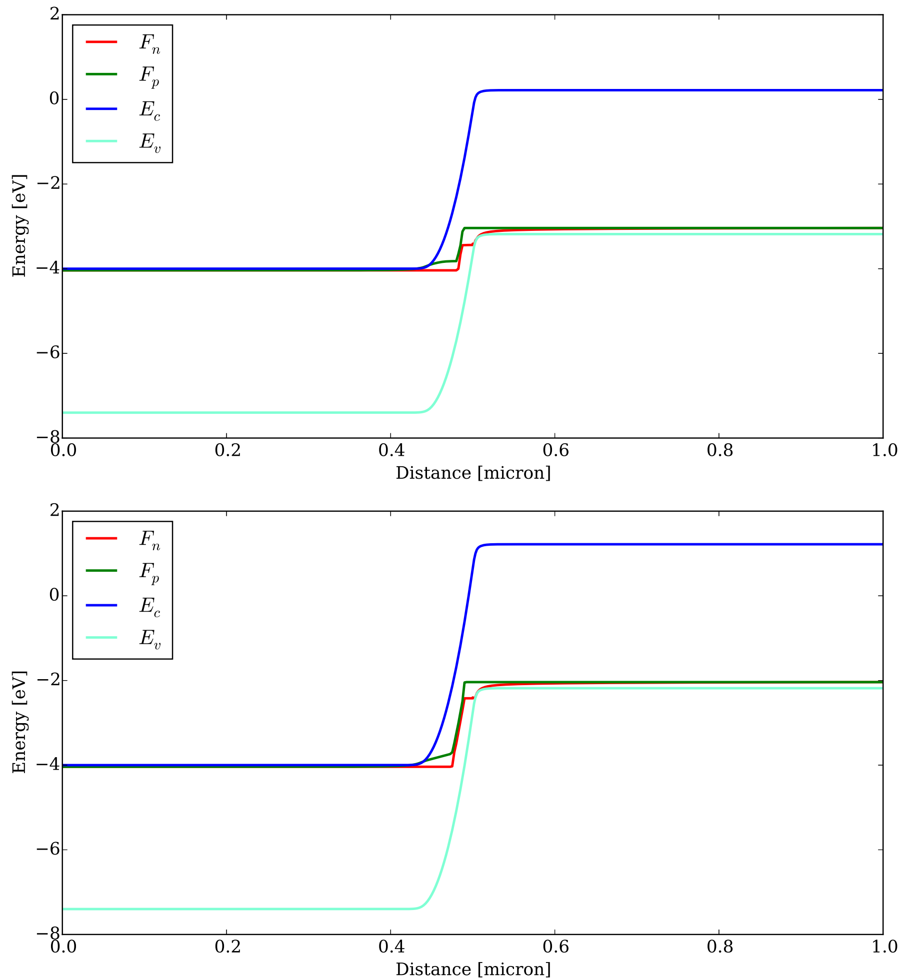

Figure 2.16: Comparison of band diagram of p-n homojunction for reverse bias: $-1\,\mathrm{V}$ and $-2\,\mathrm{V}$.

the opposite direction.

This phenomena is clearly visible in simulations. Comparison of band diagrams for increasing forward bias is presented in figure 2.15. Injection of electrons to the p-type region and holes to the n-type region leads to decrease of the potential barrier in the depleted region. When the forward bias (in volts) is roughly close to the band gap (in electronvolts), then the potential barrier vanishes almost completely, there is no depleted region anymore and the device behaves more or less as a linear resistor. On the other hand, under the reverse bias the potential barrier increases, and the n-type region and the p-type region become isolated from each other.

This behavior is clearly indicated by the I-V characteristic presented in figure 2.17. This figure reveals three operating modes of a semiconductor diode. Under reverse bias (bias $< 0\,\mathrm{V}$), the current is very small and generally it does not increase with voltage. Then for the forward bias, there is an exponential mode, when the current is small, but it increases exponentially in bias, and the linear mode, when the current (and resistance) becomes linear in bias.

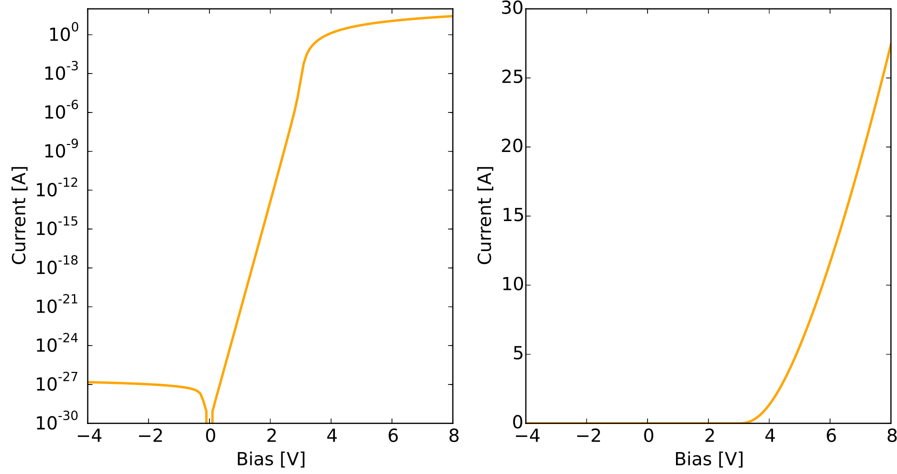

Figure 2.17: Current-voltage (I-V) characteristic of a GaN p-n homojunction. Logarithmic and linear scales are presented. Absolute value of current is plotted.

2.8.2 Homojunctions, p-i-n diodes and single quantum well structures

In this section, we simulate several GaN-based semiconductor devices, with different level of complexity. We would like to determine basic properties of these devices, before proceeding to more complicated structures.

2.8.2.1 Devices

The basic structure to simulate is a p-n homojunction made of GaN, referred by abbreviation *N-P*. The doping is assumed to be $N_d = 5 \cdot 10^{18}\mathrm{cm}^{-3}$ in the n-type region and $N_a = 5 \cdot 10^{19}\mathrm{cm}^{-3}$ in the p-type region. The compensation is 2% of doping concentrations. Both regions are of 500 nm length.

The second structure we consider is a p-n junction with insulating layer in the middle. We consider two lengths of the insulating layer: 20 nm and 200 nm. These devices are referred by abbreviations *N-I-P 20nm* and *N-I-P 200nm*. It is assumed that in insulating layer, there are in fact donors with concentration $5 \cdot 10^{16}$ cm^{-3}, as such impurity level is expected to be present in real devices due to the nature of growth processes.

Next device is a single quantum well (QW) structure. The quantum well is 3 nm length, and we assume n-type region and p-type region to be of length 499 nm and 498 nm, respectively. QW region is made of $\mathrm{In}_{0.1}\mathrm{Ga}_{0.9}\mathrm{N}$. As in *N-I-P* case, the quantum well is assumed to be donor-doped on level $5 \cdot 10^{16}$ cm^{-3}. This device will be abbreviated as *N-W-P*.

The most complex device in this simulation is the two quantum well and a barrier structure. The layers of the device are as follows: a GaN n-type region 495 nm length, 3 nm quantum well, 5 nm $\mathrm{In}_{0.015}\mathrm{Ga}_{0.985}\mathrm{N}$ donor doped barrier, 3 nm quantum well and 494 nm p-type region. Quantum wells' properties are the same as in previous device. The barrier is doped with $5 \cdot 10^{18}\mathrm{cm}^{-3}$ shallow donors, as n-type region. This device would be referred as *N-W-B-W-P*.

2.8.2.2 Dopants in insulating layers

First we would like to study the impact of low impurity concentration in insulating layers and a quantum wells. In theoretical considerations one can assume these parts of device to be perfectly pure, but during the real semiconductor growth it is always possible to introduce some level of impurities. From the computational point of view, removing any doping in certain layer results not only in zero

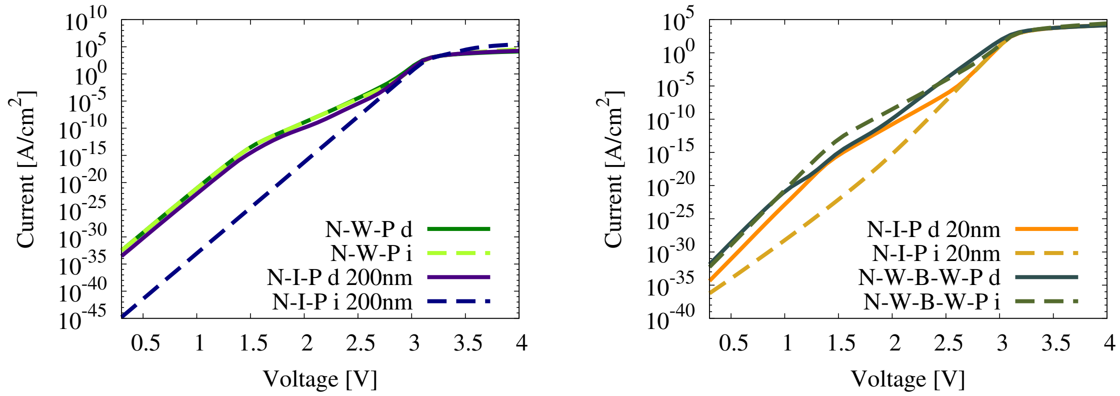

Figure 2.18: Comparison of the I-V characteristics of devices with doped (impure, d) and undoped (pure, i) insulating layers and quantum wells. Tunneling to trap level was not included into the simulation.

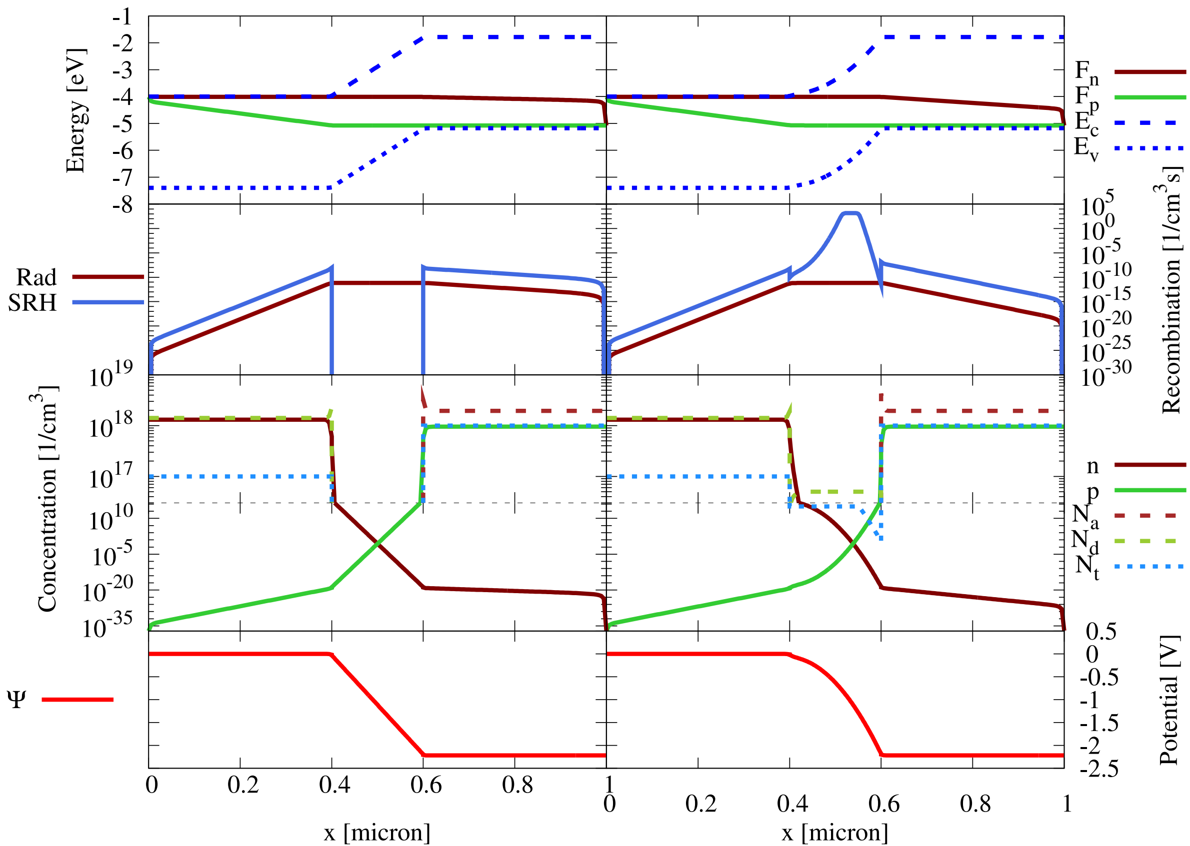

Figure 2.19: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-I-P 200nm* device for pure (left) and doped (right) insulating layer. The applied potential is 1.06 V. Tunneling to trap level was not included into the simulation.

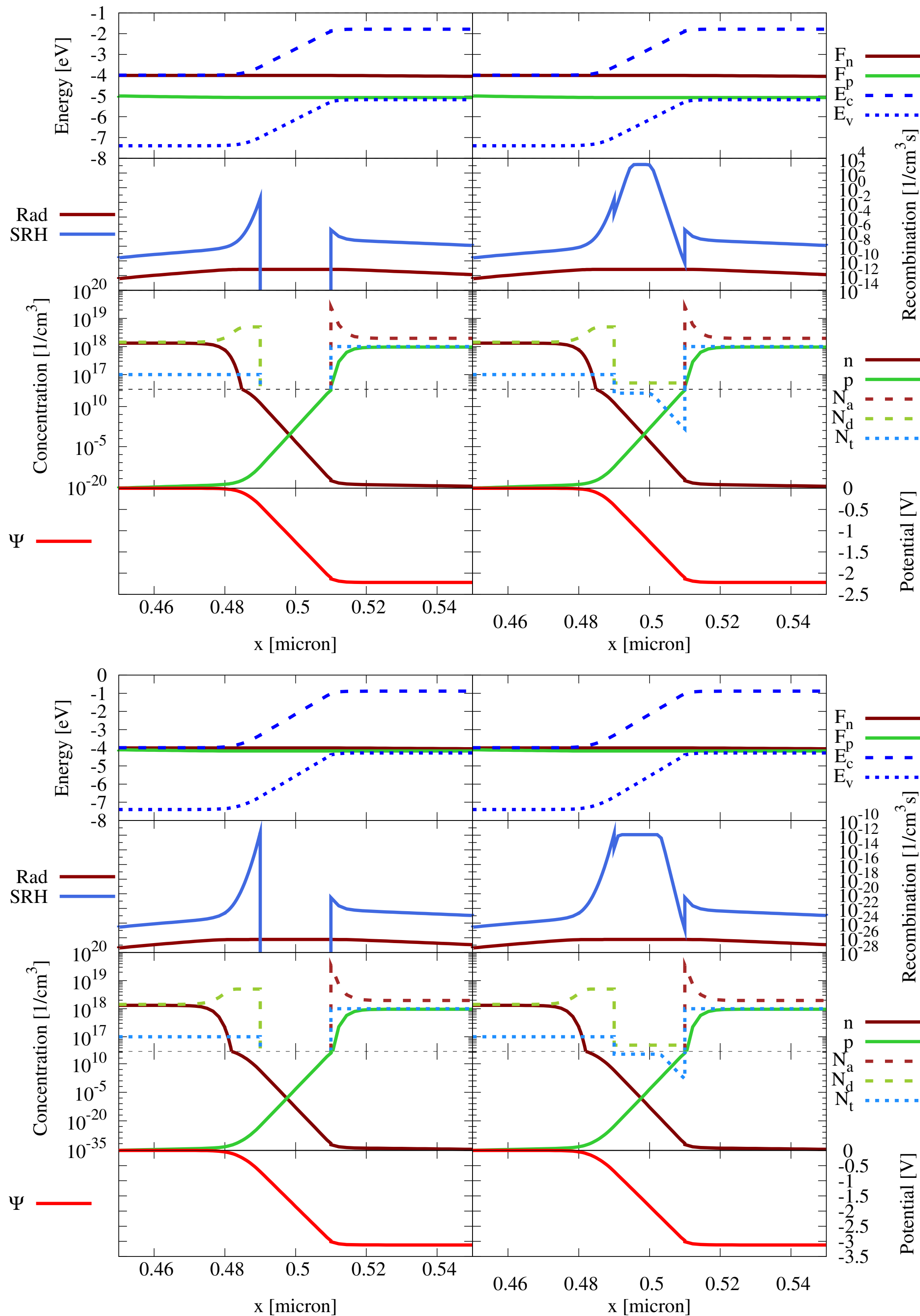

Figure 2.20: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-I-P 20nm* device for pure (left) and doped (right) insulating layer. Only the middle of the device is shown. The applied potential is 0.16 V and 1.06 V for lower and upper plot, respectively. Tunneling to trap level was not included into the simulation.

E_t[eV]	-6.4		-4.5
N_t[cm^{-3}]	10^{15}	10^{17}	10^{18}
τ_n^{SRH}[s]	10^{-7}	10^{-9}	10^{-12}
τ_p^{SRH}[s]	10^{-9}	10^{-11}	10^{-10}

Table 2.4: The electron and hole lifetimes for trap-level recombination for selected trap concentrations and the energy level of the trap. It is assumed that energy band boundaries are similar as for GaN material: the conduction band is $E_c \approx -4\,\text{eV}$ and valence band is $E_v \approx -7.4\,\text{eV}$. Values presented in the table are tentative.

acceptor/donor concentration, but also no recombination on trap level, thus may lead to considerable change of I-V characteristic.

Since the recombination terms play vital role in a simulation of luminescent devices, we must assume some parameters for it. For all devices the radiative recombination coefficient are $C^{rad} = 1.1 \times 10^{-10}\,\text{cm}^3/\text{s}$. The case is more complicated for SRH recombination. We assume that for a donor-doped part, a corresponding acceptor compensating level is $E_t = -6.4\,\text{eV}$, and for an acceptor-doped part compensating donor level is $E_t = -4.5\,\text{eV}$. The energy level and impurity concentration affect carrier lifetimes. The assumed values are presented in table 2.4. Relative effective carrier masses are assumed to be $m_n = 0.2$ for electrons and $m_p = 1.7$ for holes. In this section, we do not include the trap-assisted tunneling, but level occupation and full SRH formula is used.

Having the parameters covered, we proceed to the simulation results. There are four devices considered, as described in section 2.8.2.1, with insulating layers doped and undoped, what would be denoted by additional letter *d* or *i*, respectively. We would like to verify to what extent small concentration of impurities affects a device. Thus we assume the doping of insulating layers to be of donor type, of magnitude 1% of the regular n-type region doping.

The comparison of I-V characteristics is presented in figure 2.18. We will start with *N-I-P 200 nm* (figure 2.19). For bias in proximity of 3 V the currents are similar, for higher voltages the *d*-device has bigger current, due to lower resistance. For bias lower than 2.5 V however, the characteristics differ substantially. An explanation of such a behavior is lack of the SRH recombination for pure insulating layer in the depleted region (see figure 2.19). The SRH recombination increases total current, as it may be treated as a mean of electron transport through the potential barrier in the depleted region. The recombination on the trap level rate, even with no tunneling effect, is significant there, so even for small trap concentration it is important. For higher biases, the potential barrier is small and this effect is negligible for the electric conductance, so the I-V characteristics are similar above 3 V bias.

The case of 20 nm length insulating layer is similar, however for bias around zero the difference of currents diminish. Such a distinction results from the *I*-layer being too narrow to enclose the depleted region of the device. When an applied potential is low, maximal rate of the SRH recombination is reached in the n-type region, and the impurities in the undoped layer (figure 2.20). Nevertheless, for moderate bias, still quite large difference of currents is observed.

For the device with one quantum well (*N-W-P*) the characteristics are almost the same. In this case reason is that the region of maximal recombination is not in the middle of the device, but it is slightly shifted, into the n-type region (figure 2.21). Thus, a doping in the quantum well does not increase the recombination much and the difference in the charge is so small so it does not change the behavior of the device.

In the case of the device with two quantum wells (*N-W-B-W-P*), there are differences in current magnitudes in the range 1.5–2 V. The analysis is not so straightforward as before, as the recombination maxima are in different positions for pure and doped quantum wells (figure 2.22). It seems that for

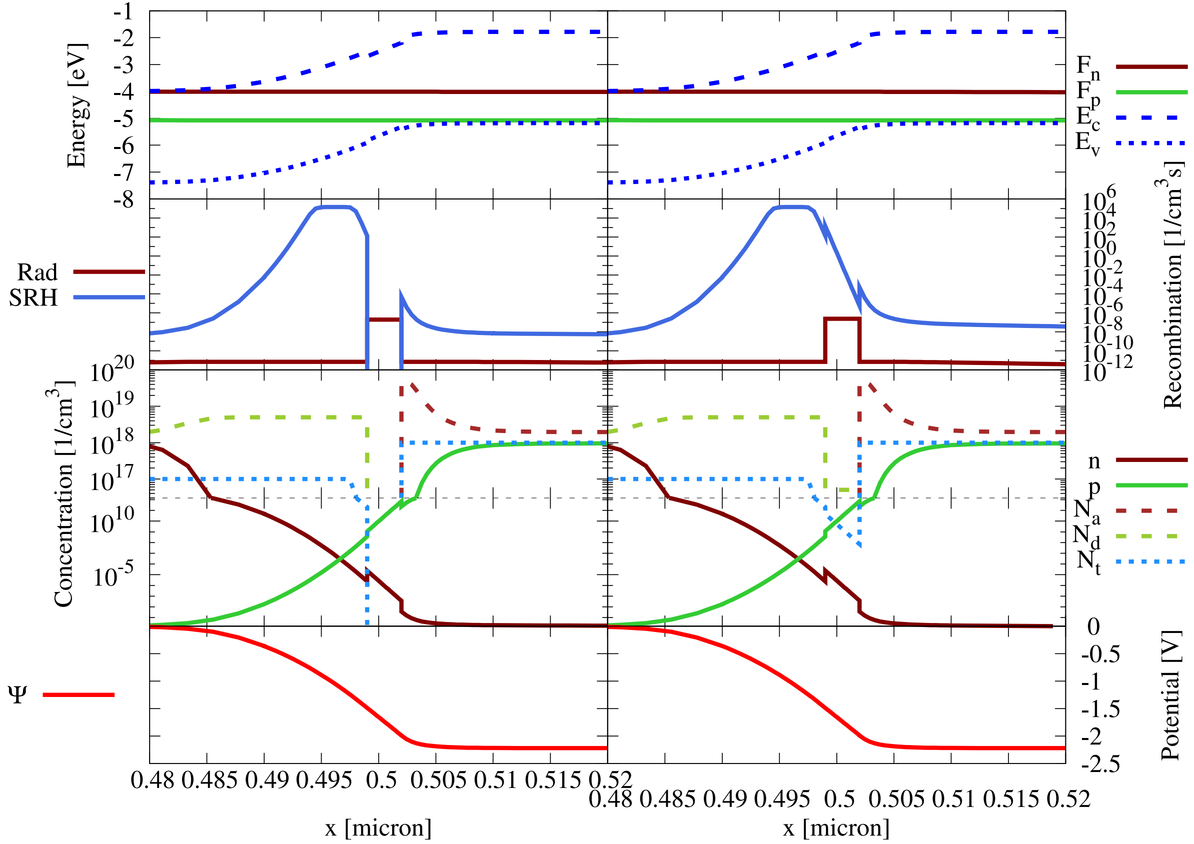

Figure 2.21: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-W-P* device for pure (left) and doped (right) quantum well. The applied potential is 1.06 V. Tunneling to trap level was not included into the simulation.

lower potentials, the current is approximately the same. On the other hand, for medium voltage, the device with pure quantum wells has slightly bigger current. This effect is due to SRH recombination in the barrier, which is smaller than SRH recombination in QWs, when it is present, but if there is no trap levels in QWs, it becomes much larger.

The conclusion is that the impurities in insulating layers may change the I-V characteristic, but it depends on the device. If an insulating layer is located in the region where there is significant SRH recombination, then impact of impurities may be important. Otherwise it is negligible, as the charge generated by additional concentrations of ionized traps is too small to affect the device. The impact is observed for small and medium biases, up to 2.5–3 V. Above 3 V the rate of the SRH recombination is not considerably larger near the p-n junction, as there is no potential barrier, thus its impact on the I-V characteristic is negligible.

Nevertheless the above simulations reveal, that in general we cannot neglect the impact of impurities in the active region on a device behavior, despite of their low concentration.

2.8.2.3 Tunneling

The the trap-assisted tunneling (see section 2.7.1) might be an important phenomenon governing operation of a p-n junction. Therefore we would like to determine the impact of this mechanism on I-V characteristic of the device. To perform simulations, one can utilize formulae from section 2.7.1. Derivation of these formulae require rigorous assumptions, which are satisfied by an idealized device,

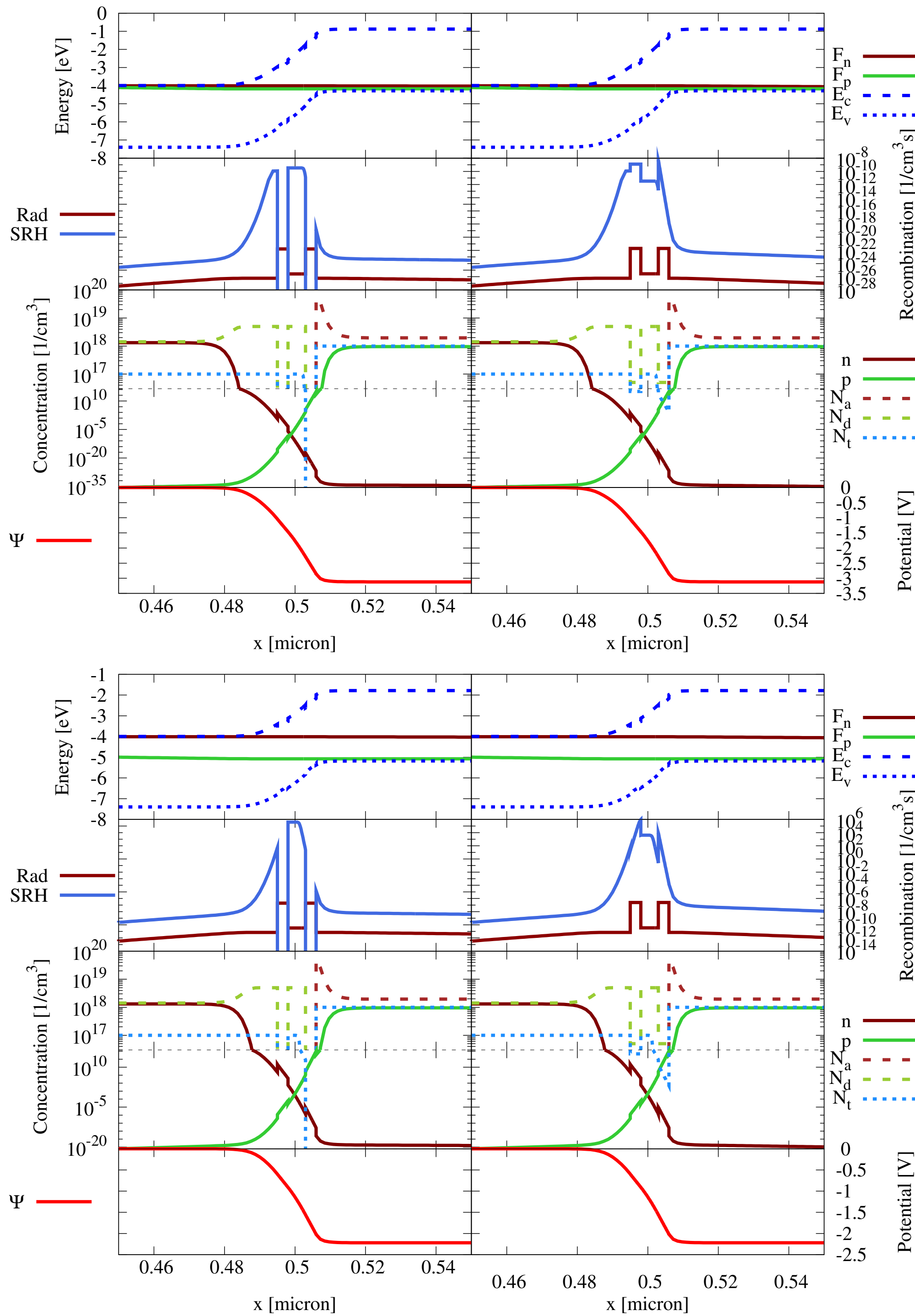

Figure 2.22: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N*-*W*-*B*-*W*-*P* device for pure (left) and doped (right) quantum wells. The applied potential is 0.16 V and 1.06 V for lower and upper plot, respectively. Tunneling to trap level was not included into the simulation.

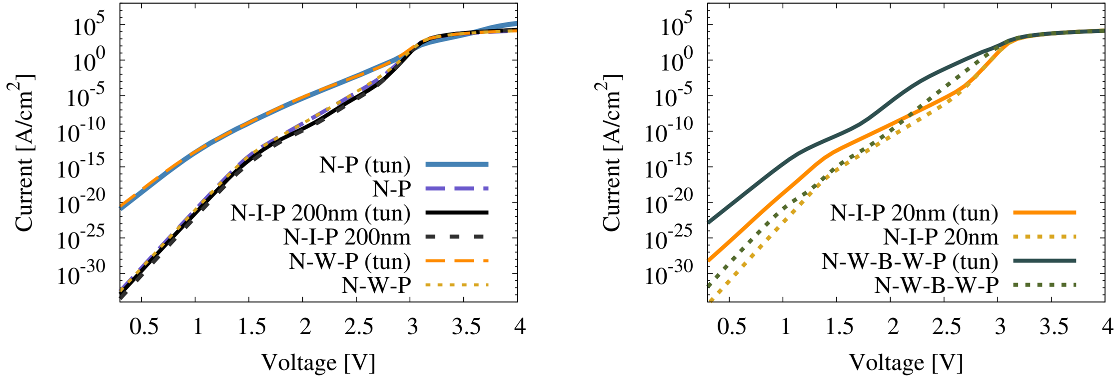

Figure 2.23: Comparison of I-V characteristics of devices with trap-assisted tunneling *(tun)* and with standard SRH recombination.

but they could fail in a real situation. Still, one can treat results of such simulations as some kind of approximation.

We performed simulations of devices from section 2.8.2.1 with and without trap-assisted tunneling and we compared the results. The I-V characteristics are shown on figure 2.23. The tunneling effect has increased the current for all devices for biases below 3 V. For higher voltages, the differences are negligible. Such a behavior is expected, as a high electric field is present for a small bias applied to a device.

The smallest change we observe for the *N-I-P 200nm* device, and the greatest for *N-P* diode. The explanation is that the latter device has the the shortest depleted region and it contains largest concentration of impurities. On the contrary, *N-I-P 200nm* structure has the longest depleted region and with low impurity concentration. Also the single quantum well structure *N-W-P* might be considered as an external case instead of *N-P*, since as shown in section 2.8.2.2, its insulating layer with reduced recombination rate does not lie in the depleted region. For two other devices, *N-I-P 20nm* and *N-W-B-W-P*, we also observe a considerable difference in the current magnitude. However it is smaller, due to a longer depleted region and lower concentration of impurities present there.

The analysis of band diagrams reveals that the tunneling effect has considerable impact on devices (see figures 2.24, 2.28). The electron quasi-Fermi level, which previously was almost constant in the active part of the device, now varies rapidly in the depleted region and it aligns to hole quasi-Fermi level. Thus the electron concentration in p-type region is much lower than the hole concentration in the n-type region. This effect is observed for low potentials, and vanishes for high ones. The cause of such a behavior is much greater recombination rate for the devices with tunneling effects.

The devices for which the mentioned effect is less significant are the *N-I-P* structures, where the electrostatic field is lower than in other cases (see figures 2.26, 2.25). For *N-I-P 200nm* structure an large field zone is about 200 nm long, as it enclosed in the insulating layer. It is much longer than in other devices, where the depleted regions are of 20–30 nm length. Thus it slightly exceeds the insulating layer for the diode *N-I-P 20nm*. In these cases, the tunneling effect is not sufficiently significant for the device operation, as the potential in the depleted region is almost linear, so the electric field is distributed almost equally, on contrary to previous examples, where it is stronger near p-type region interface.

The observed result reveals also significant problem with the theory. Formulas presented in [51], used in simulations, are valid under the assumption of constant quasi-Fermi levels in the active part of the device. This assumption is obviously not valid, due to application of the very theory. Thus, the results should probably be verified with more subtle approach.

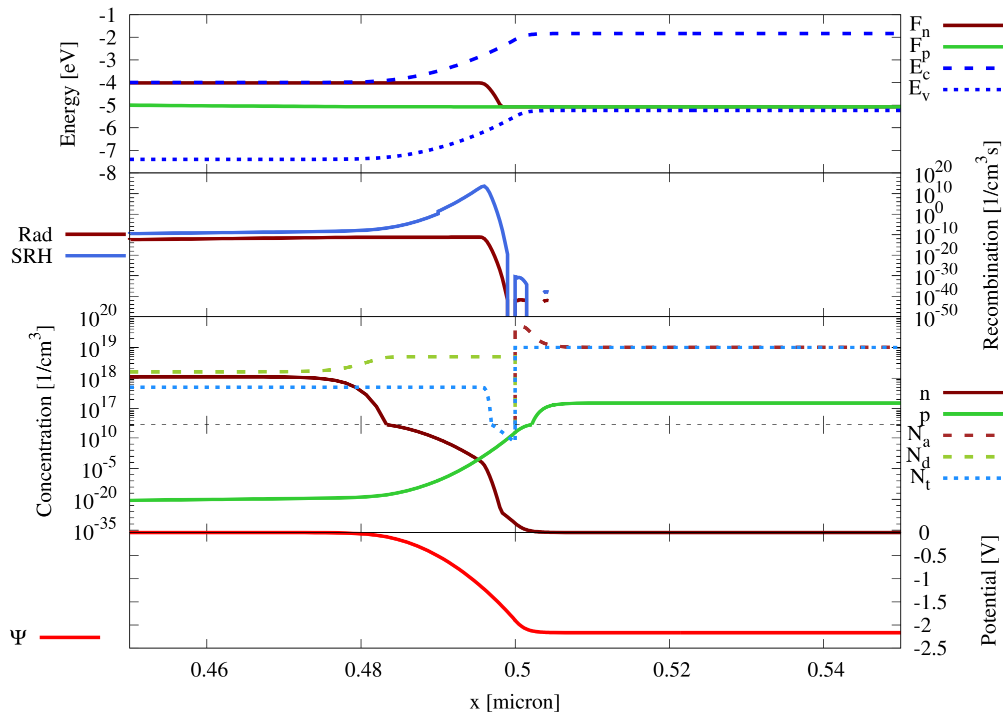

Figure 2.24: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-P* device. The middle of the device is shown. The applied potential is 1.06 V. Trap-assisted tunneling was included into the simulation.

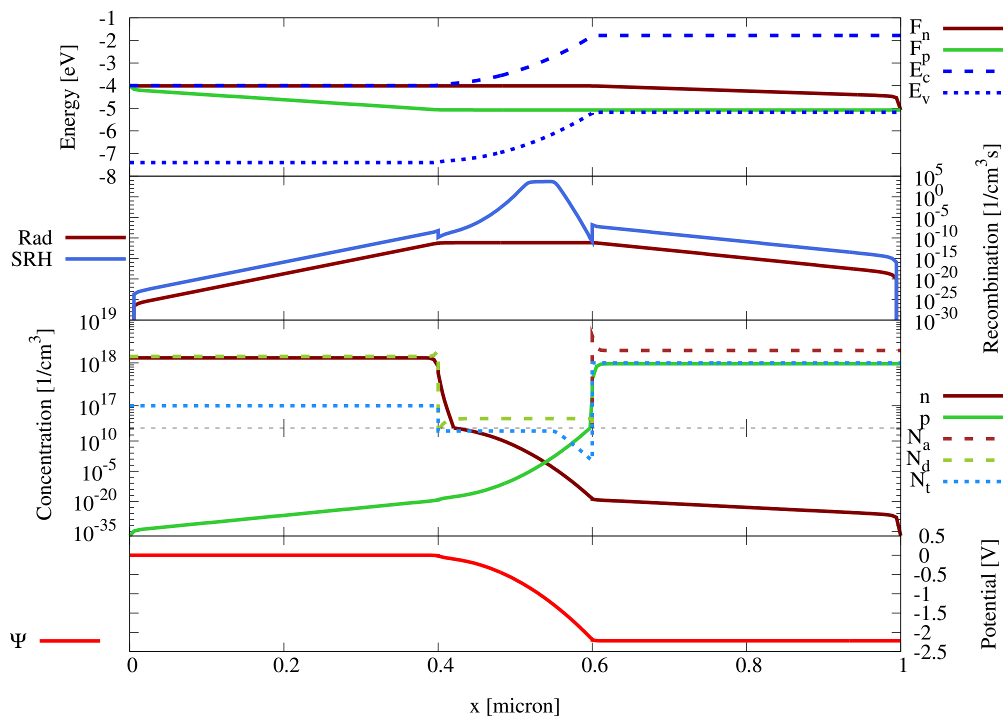

Figure 2.25: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-I-P 200nm* device. The applied potential is 1.06 V. Trap-assisted tunneling was included into the simulation.

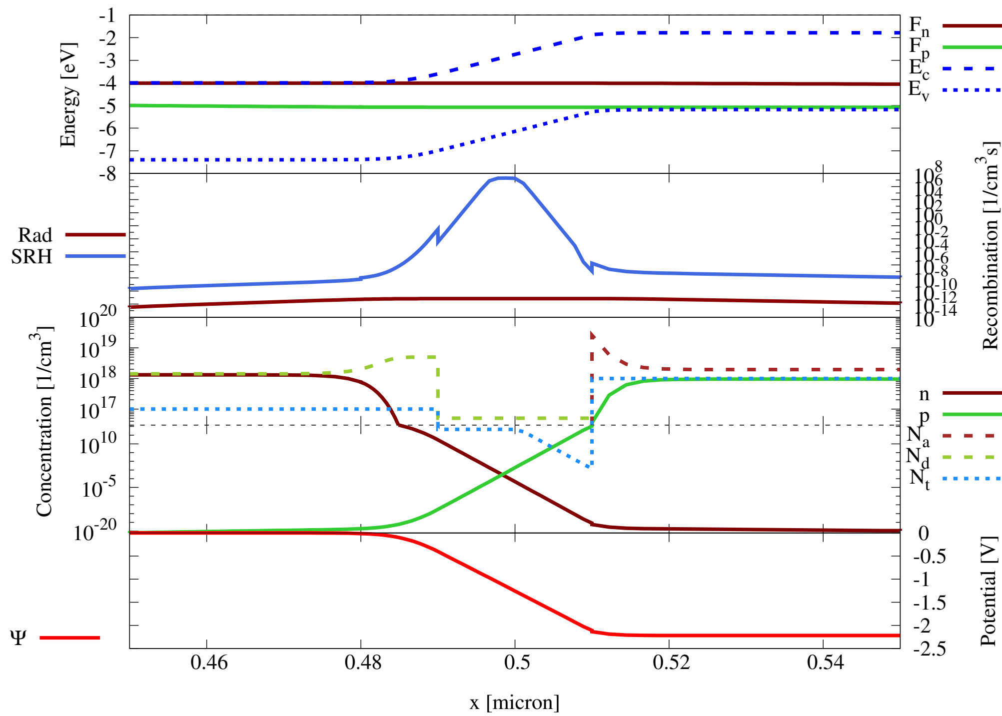

Figure 2.26: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-I-P 20nm* device. Only the middle of the device is shown. The applied potential is 1.06 V. Trap-assisted tunneling was included into the simulation.

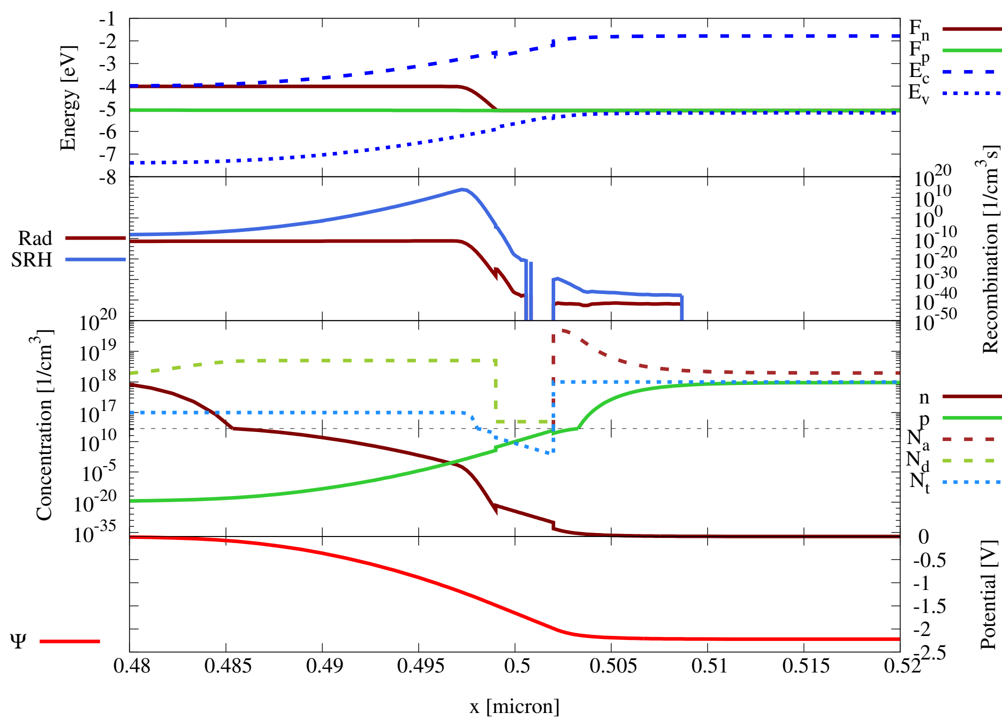

Figure 2.27: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-W-P* device. Only the middle of the device is shown. The applied potential is 1.06 V. Trap-assisted tunneling was included into the simulation.

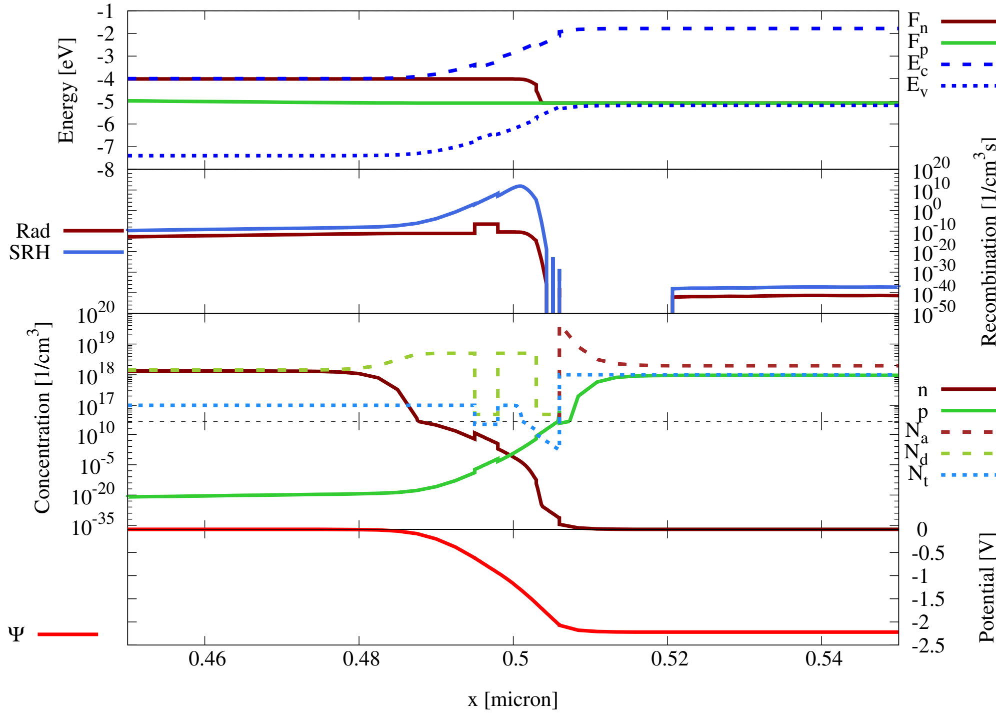

Figure 2.28: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-W-B-W-P* device. Only the middle of the device is shown. The applied potential is 1.06 V. Trap-assisted tunneling was included into the simulation.

2.8.2.4 Occupation of a trap level

As we saw already in the section 2.8.2.2, small concentrations of impurities introduced unintentionally may lead to a relevant change of a I-V characteristic of a device. The doping introduces two effects: the SRH recombination and the charge generated by the trap occupation. We would like to determine the impact of this charge.

Therefore we will compare two kinds of devices with insulating layers: with unintentionally doped *W* and *I* layers, denoted as *d*-devices in section 2.8.2.2, and devices with recombination rate adjusted as for doped insulating layers, but zero trap and dopant concentrations, what we denote as *ir*.

The I-V characteristics computed for structures *N-I-P 200nm* and *N-W-P* are similar for both cases (figure 2.29). Let us study the *N-I-P 200nm* device more closely (figure 2.30). The quasi-Fermi levels are quite similar, but the potentials in the insulating layer differ slightly. Thus the carrier concentrations are different and the maximum of the SRH recombination is shifted a bit. However, the magnitude of the recombination rate is almost the same, so the characteristics do not differ much. For the *N-W-P* device, the differences are hardly noticeable (figure 2.31). The introduced charge is too low to impact these devices, as it is relatively small in comparison with doping of n-type region and p-type region.

The above simulations suggest the charge generated by the ionized traps in the insulating layer to be a factor of low importance. The main impact is due to a recombination term, on which a I-V characteristic is directly dependent.

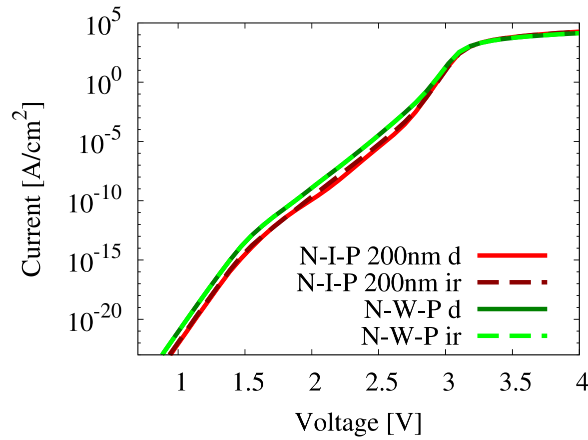

Figure 2.29: Comparison of I-V characteristics of devices with impure insulating layers: with no charge generated by impurities in insulating layers, but a recombination taken into account (ir) and with charge included (d). Tunneling to trap level was not simulated.

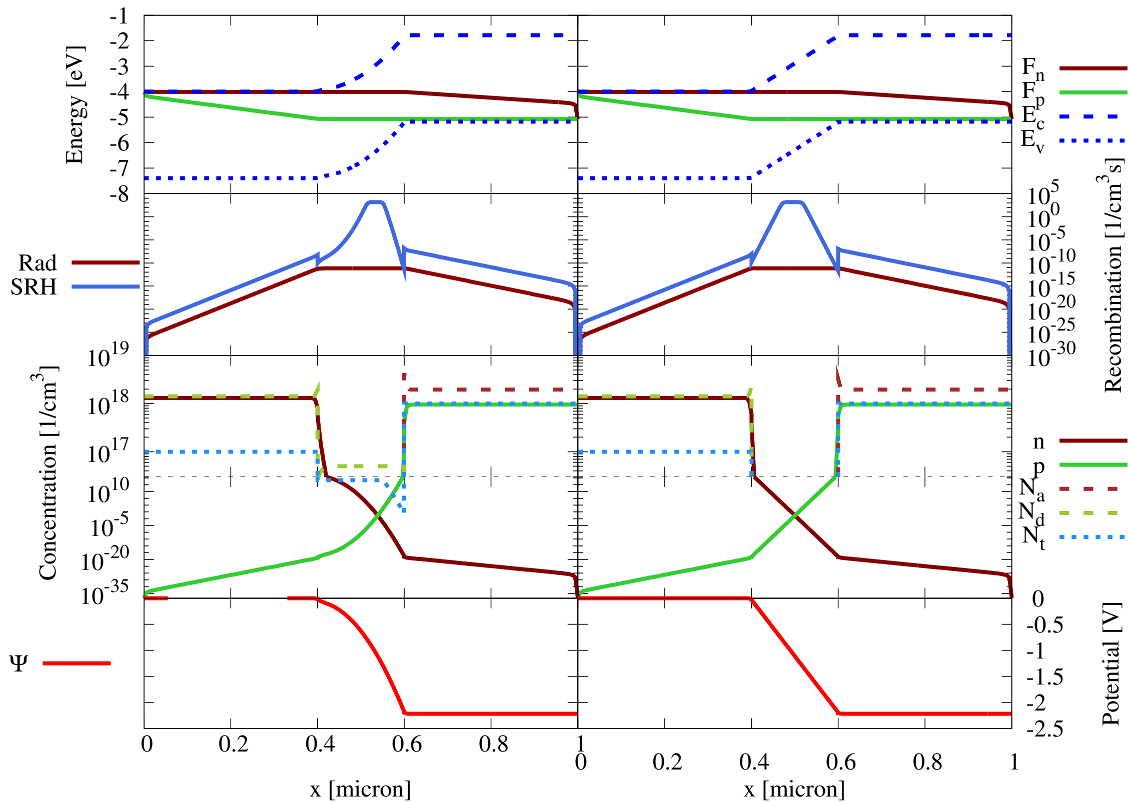

Figure 2.30: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of *N-I-P 200nm* device for full trap level occupation (left,d) and recombination only (right,ir) in the insulating layer. The applied potential is 1.06 V.

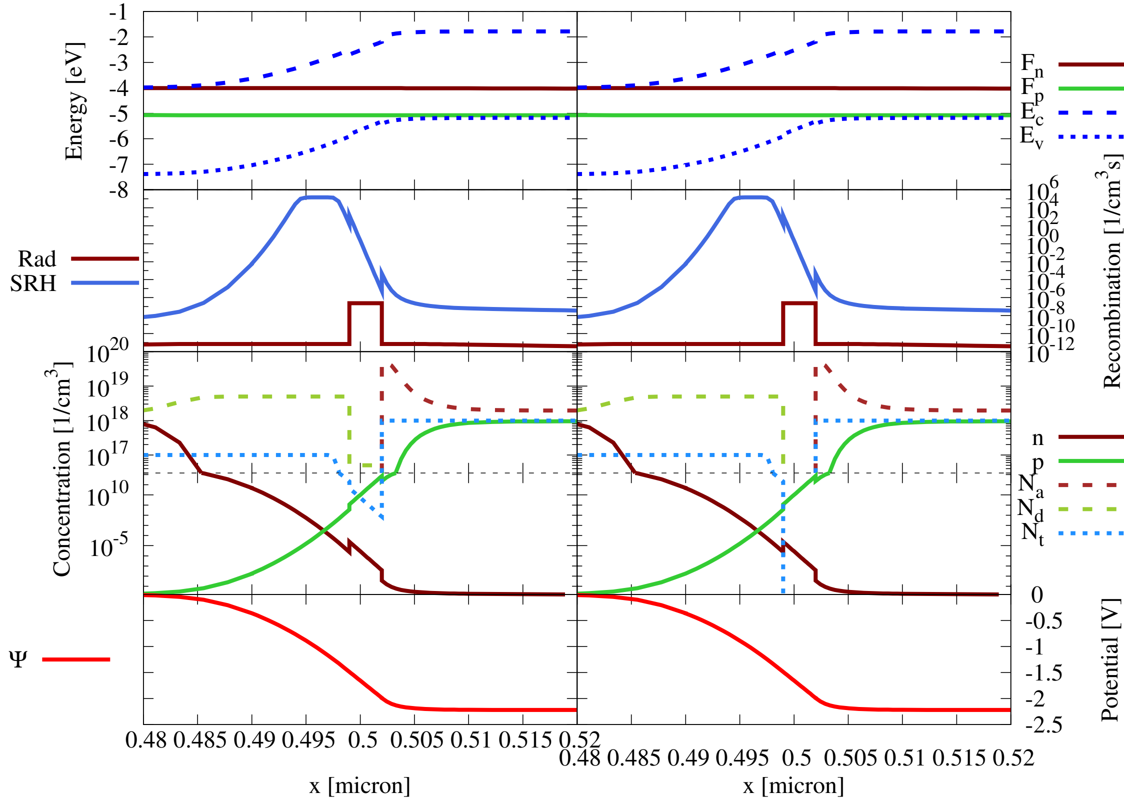

Figure 2.31: Plots of the band diagram, the recombination rates, the charge carrier concentrations, and the potential of N-W-P device for full trap level occupation (left,d) and recombination only (right,ir) in insulating layer. The applied potential is 1.06 V.

2.8.2.5 Dopant compensation

Since the SRH recombination term seems to have considerable impact on the device operation, and the charge concentration due to ionized traps in insulating layers are negligible, then maybe the charge of ionized traps may be completely neglected? Perhaps one should compute only the recombination term? Calculation of the ionized traps charge is not particularly computationally costly, but still omitting it simplifies the model and it improves simulation time.

We consider an p-n diode in two variants: with the trap occupation term included (N-P) and omitted (N-P oto). In both cases, the recombination generated by traps is taken into account.

The resulting characteristics are displayed on the figure 2.32. They generally agree, differences can be hardly seen for voltages above 3 V. The same is true not only for the characteristics, but for the potential and quasi-Fermi levels. Maximal difference potentials were $|\psi(x) - \psi_{oto}(x)| < 0.02$ V, $|F_n(x) - F_{n,oto}(x)| < 0.03$ eV and $|F_p(x) - F_{p,oto}(x)| < 0.03$ eV. The simulation has been performed also for more complex N-W-B-W-P device, and similar estimates hold.

The conclusion is then that the compensation at level of 2% of doping is important mainly due to a SRH recombination it generates, as the ionized traps charge is negligible.

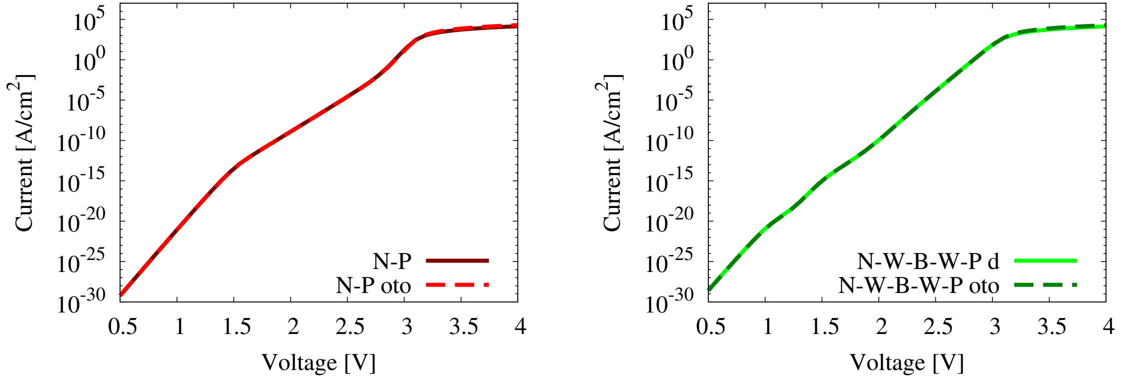

Figure 2.32: Comparison of I-V characteristics of devices with the ionized traps charge omitted (*oto*) and included. The recombination generated by traps are included into both cases. Differences are hardly noticeable.

2.8.3 Comparison with available software

Simulations presented in other sections of this study were performed with our own software. There are some free or commercial programs, which accounts for simulations of semiconductor devices either with the drift-diffusion model or with more subtle methods. However, these applications come with a certain set of features, like the physical phenomena accounted for. While this is natural for any modelling setting, introducing additional effects is limited without a source code and appropriate license.

The drift-diffusion model is not very sophisticated, thus to experiment with new features it is quite viable to develop one's own computer code. In our research we focus on GaN luminescent devices, and therefore we are interested in additional effects affecting operation of such structures, not included in standard drift-diffusion model. The available software offer simulations of wide range of devices. It is, however, hard or impossible to add specific improvement, if it is not implemented already.

Before proceeding to the full-blown simulations of semiconductor devices, we would like to compare our program with other software performing similar tasks, which are available to us. For comparison we have chosen two programs: *SimWindows* (free) and *SiLENSe* (commercial). In these tests we do not want to show effects of improvements of the model, but to compare results of simulations accounting for similar physical phenomena.

2.8.3.1 SimWindows

SimWindows [118] is a free program for a numerical solution of the drift-diffusion system for one-dimensional structures. This program is lightweight and fast, configurable with graphic user interface. It incorporates some standard recombination models (SRH, radiative, Auger) and it is configurable to some extent. However, there is no built-in support for extending the model. In particular, scope of modeling SRH recombination is rather narrow. It is possible to use standard SRH formula with fictitious concentrations n_1, p_1 assumed to be zero, thus independent of a trap energy. Extension beyond that is not possible.

For the comparison we use already simulated devices: *N-P* and *N-W-P d*. Since we want to compare the results, we simulate the same physical conditions in our code and in SimWindows. Thus we modified the devices in such a way: we assume simple *SRH* recombination form from equation 2.6.26, and we assume that occupation of trap levels is like for shallow traps, but with respect of their energies, and no trap-assisted tunneling. We take into account the incomplete ionization of acceptors

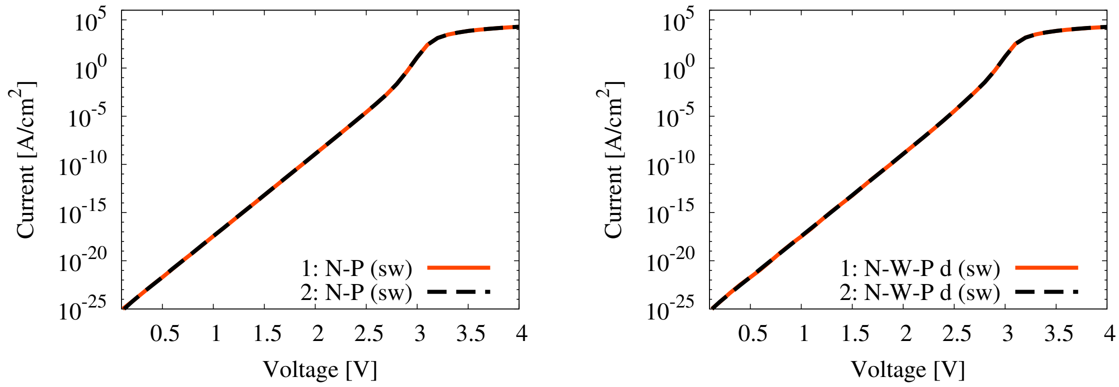

Figure 2.33: Comparison of I-V characteristics of p-n diode and single quantum well structure computed with our code(1) and SimWindows(2). The results are in good agreement.

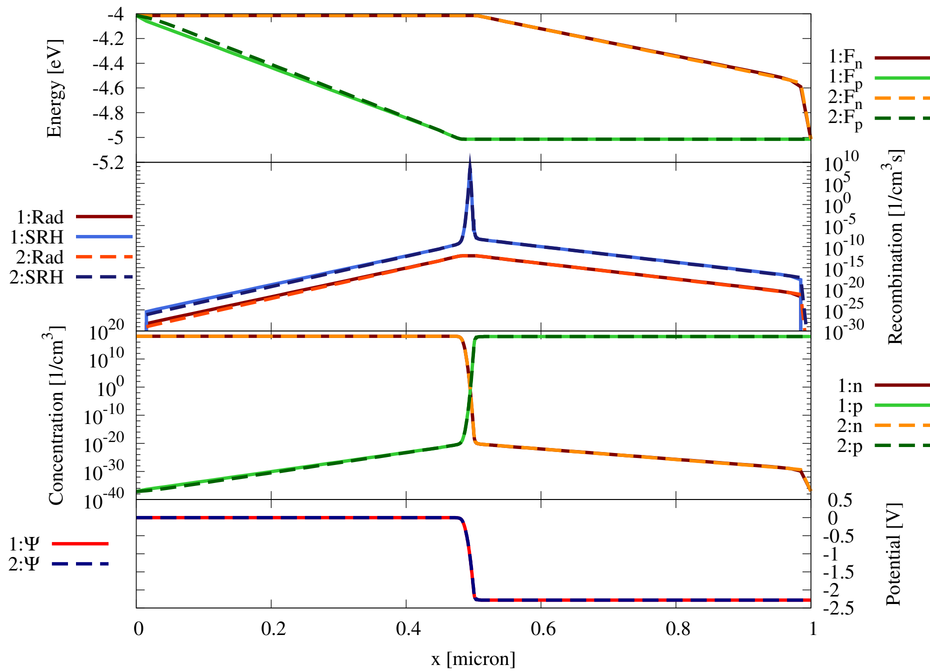

Figure 2.34: Comparison of quasi-Fermi levels, the recombination rates, the charge carrier concentrations, and the potential of *N-P (sw)* device simulated in our code(1) and in SimWindows(2). The applied potential is 1.0V. Results are in good agreement.

and donors. The Auger recombination is neglected. Devices simulated in such conditions would be denoted as *(sw)*.

The I-V characteristics of the devices *N-P* and *N-W-P d* are compared on figure 2.33. These results generally agree.

The comparison of electrostatic potential, quasi-Fermi levels and other functions describing the devices on figure 2.34,2.35, for applied potential 1.0 V. In both cases values from simulations performed by our code and SimWindows are similar. In particular, good agreement in the depleted region of the first device and near quantum well of the latter one is obtained, what implies the recombinations are similar, and so are the currents.

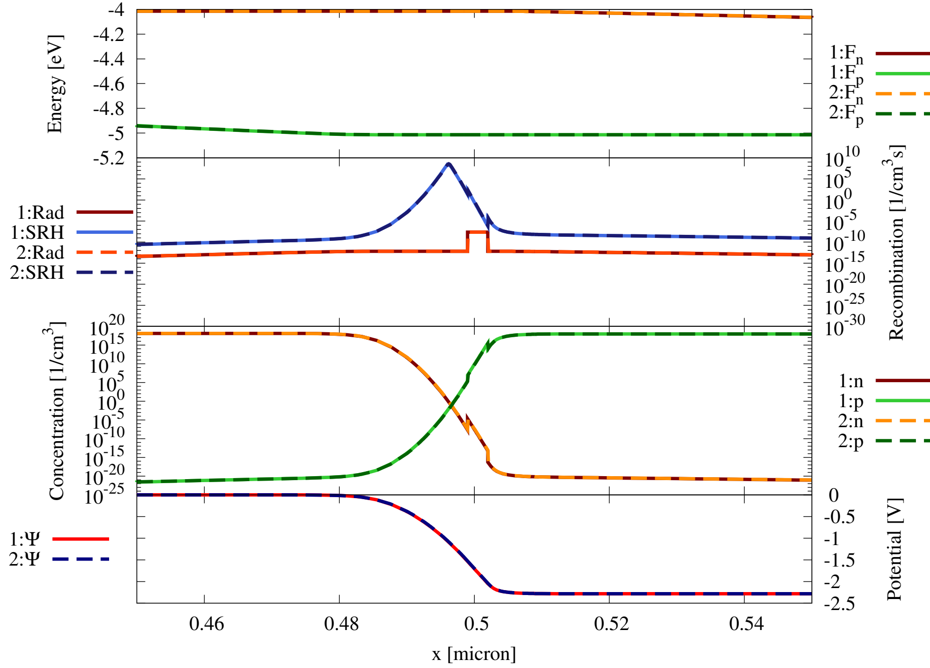

Figure 2.35: Comparison of quasi-Fermi levels, the recombination rates, the charge carrier concentrations, and the potential of *N-W-P d (sw)* device in the proximity of quantum well simulated in our code(1) and in SimWindows(2). The applied potential is 1.0V. Results are in good agreement.

2.8.3.2 SiLENSe

The second program we compare our code with is SiLENSe [1]. Full meaning of SiLENSe acronym is *Simulator of Light Emitters based on Nitride Semiconductors*. It is a commercial software, which in addition to simulation of electric properties, has also an optical module. It is supplied with the database of common material parameters, in particular for GaN, InN and AlN crystals. Parameters for $In_xGa_{1-x}N$ mixed crystals are computed automatically by the program. SiLENSe uses one-dimensional drift-diffusion model for simulations of electrical properties of semiconductor devices.

We will focus on the results for the following devices: *N-P*, *N-I-P 200nm* and *N-W-P*. SiLENSe uses simple form of SRH recombination (2.6.26), like SimWindows. Also the SiLENSe database uses slightly different material parameters. Thus, to perform a comparison, we have adjusted certain parameters to be consistent with SiLENSe defaults. Such modified devices are denoted by *(si)* suffix.

The comparison of I-V characteristics are presented on figures 2.36,2.37. Differences between our code and SiLENSe are marginal. However, when we compare band diagrams, the results do not agree in general. For structure *N-P (si)* and low bias (figure 2.38) the electrostatic potentials agree, but we observe considerable differences for quasi-Fermi levels: for the hole level in the n-type region and for the electron level in the p-type region. Such inconsistency introduces also differences in the carrier concentrations on respective parts, but since they are minority carriers, they do not affect the charge. Nevertheless, the band diagrams are quite different. Note however, that in near the depleted region the differences between quasi-Fermi functions are much smaller. Since it is a region of maximal recombination rate, the total current is similar for both simulations, which explains why the I-V characteristics are in good agreement despite this inconsistency in quasi-Fermi levels. It must be noted that this inconsistency is observed only for low bias. For high bias, these differences become rather small (see figure 2.39). Similar behavior is observed also for the device *N-I-P 200nm* (figure 2.40).

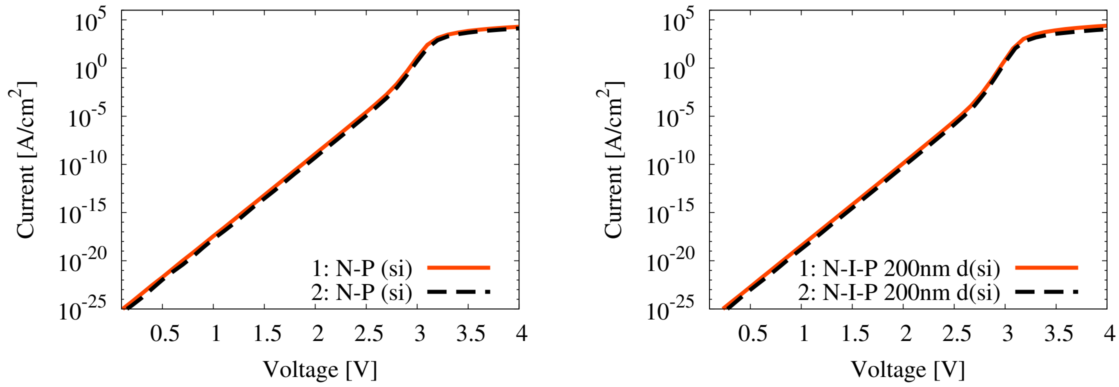

Figure 2.36: Comparison of I-V characteristics of p-n diodes without and with 200 nm insulating layer in the depleted region, computed with our code(1) and SiLENSe(2). The results are in good agreement.

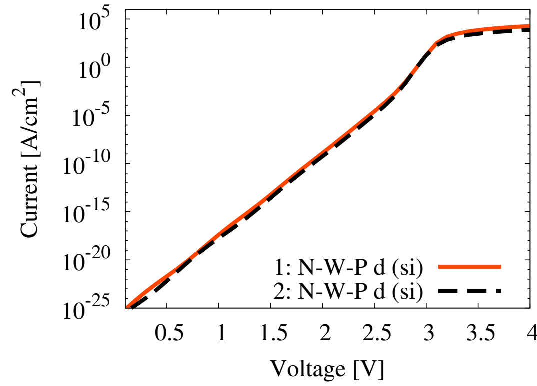

Figure 2.37: Comparison of I-V characteristics of the single quantum well structure computed with our code(1) and SiLENSe(2). The results are in good agreement.

Similar behavior is observed for single quantum well structure N-W-P. While the quasi-Fermi levels do not agree in general, the differences become small near the quantum well (figure 2.41). Thus the I-V characteristics agree (figure 2.37). The examination of the electrostatic potential ψ in the quantum well indicates polarization charges on the quantum well interfaces, which leads to non-smooth potentials. These charges are present in our code and in SiLENSe.

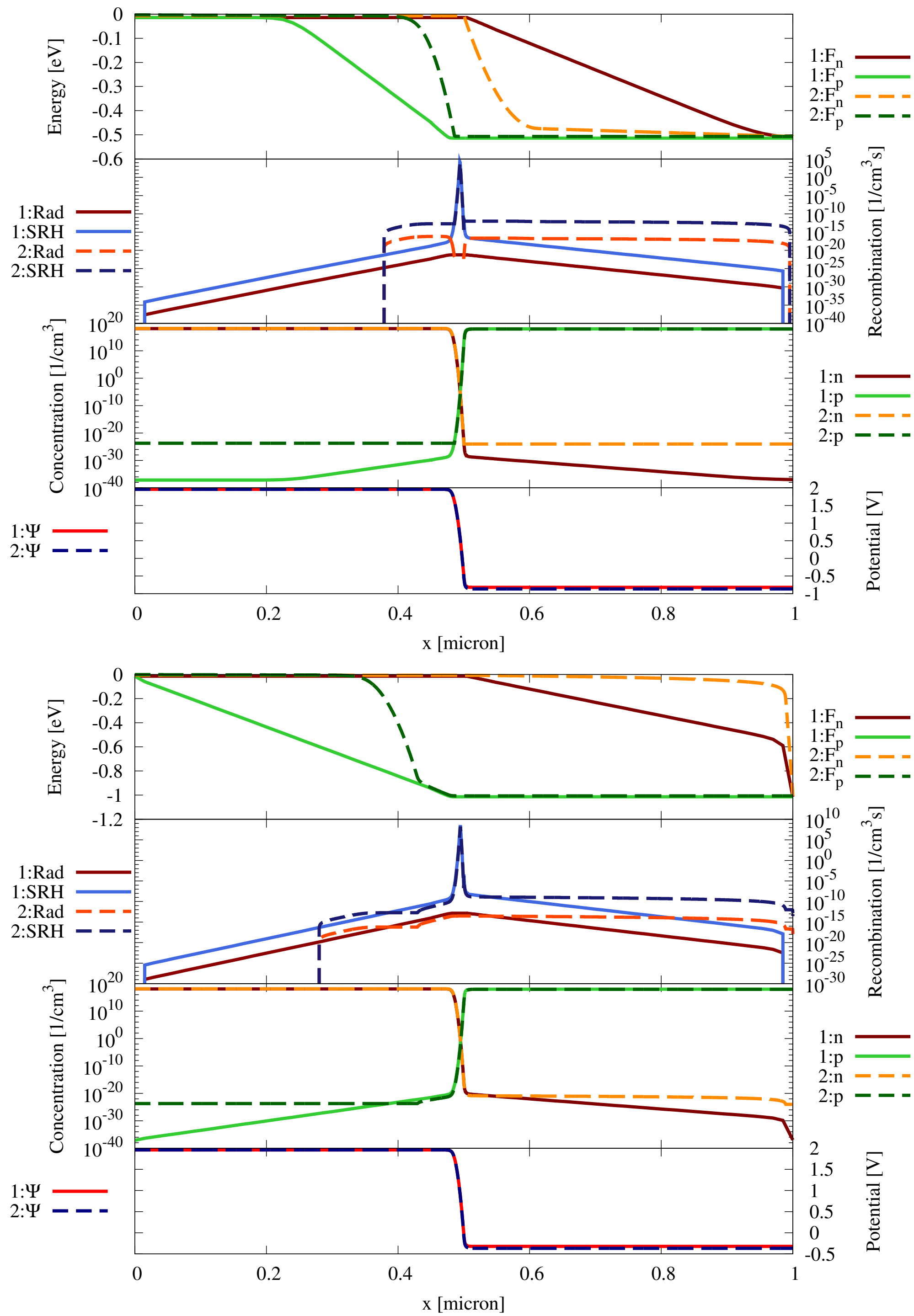

Figure 2.38: Comparison of quasi-Fermi levels, recombination rates, charge carrier concentrations, and the potential of *N-P (si)* device simulated with our code(1) and with SiLENSe(2). The bias is 0.5 V for the upper plot and 1.0 V for the lower plot. The maximal SRH recombination rates in (1) and (2) simulations are similar, but the results do not agree in general.

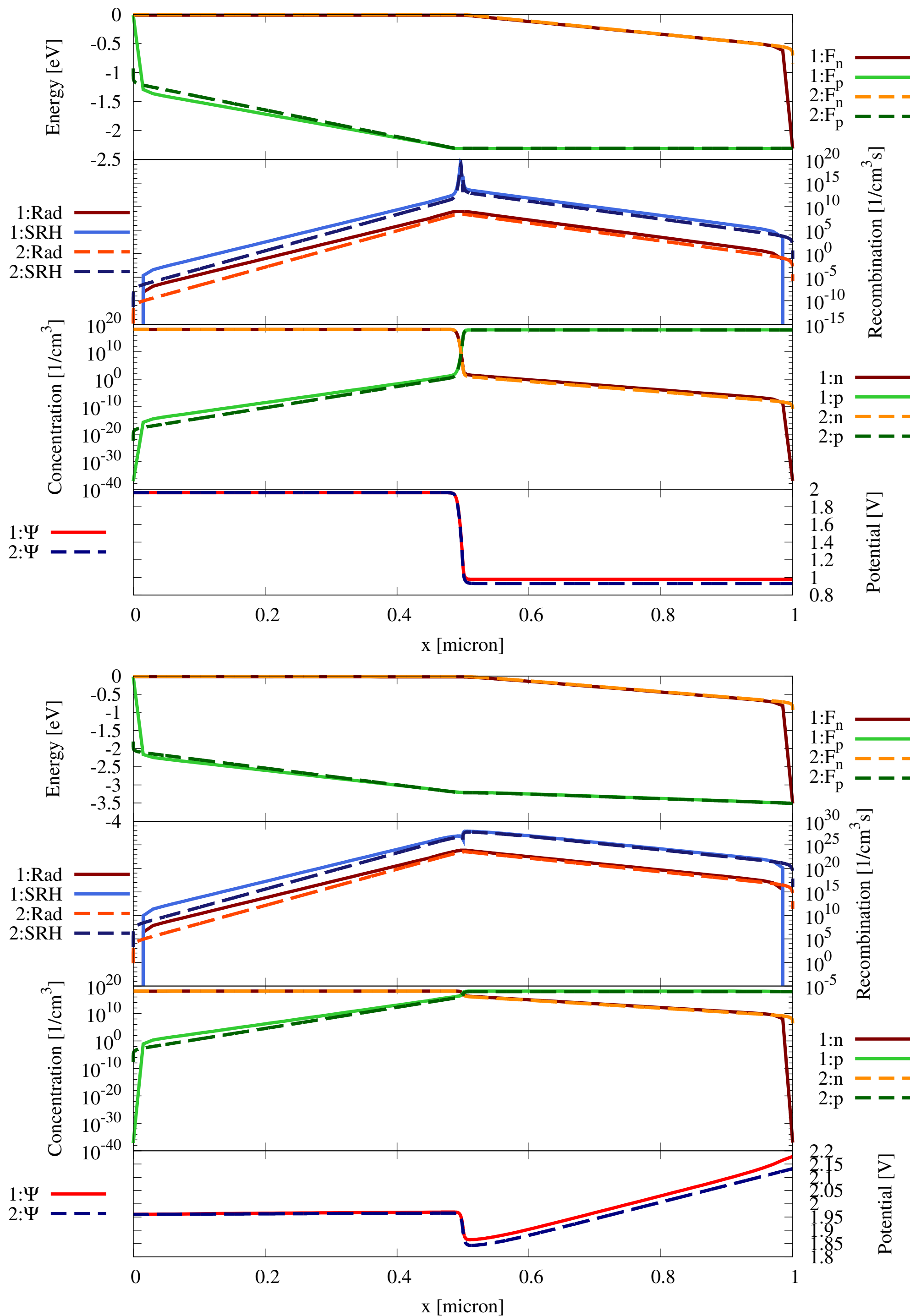

Figure 2.39: Comparison of quasi-Fermi levels, recombination rates, charge carrier concentrations, and the potential of *N-P (si)* device simulated with our code(1) and with SiLENSe(2). The bias is 2.3 V for the upper plot and 3.5 V for the lower plot. The results are in good agreement.

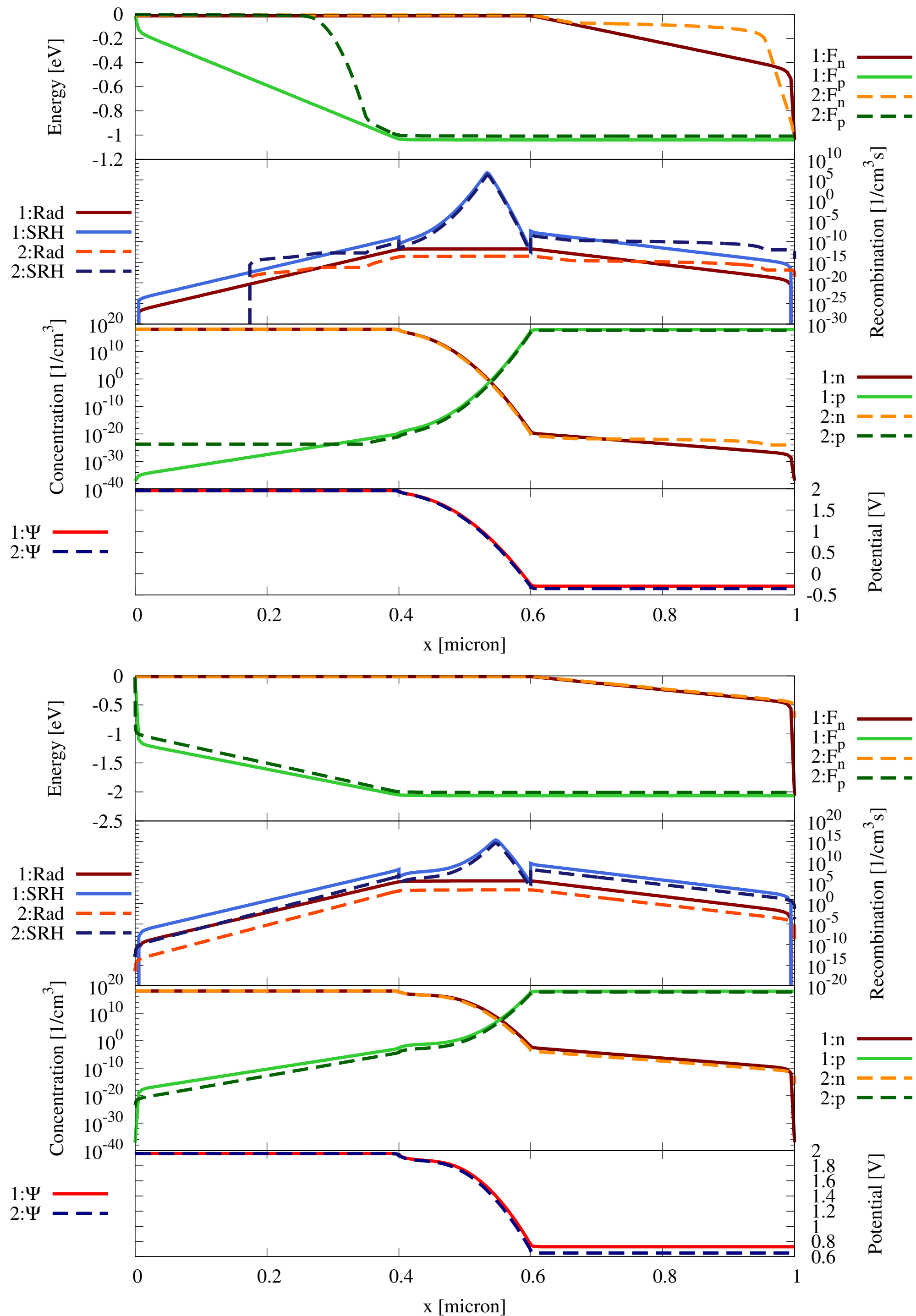

Figure 2.40: Comparison of quasi-Fermi levels, recombination rates, charge carrier concentrations, and the potential of *N-I-P 200nm d (si)* device simulated with our code(1) and with SiLENSe(2). The bias is 1 V for the upper plot and 2 V for the lower one. The results are in good agreement for 2 V, while for 1 V they do not agree. Still the unknown functions are similar in the depleted region.

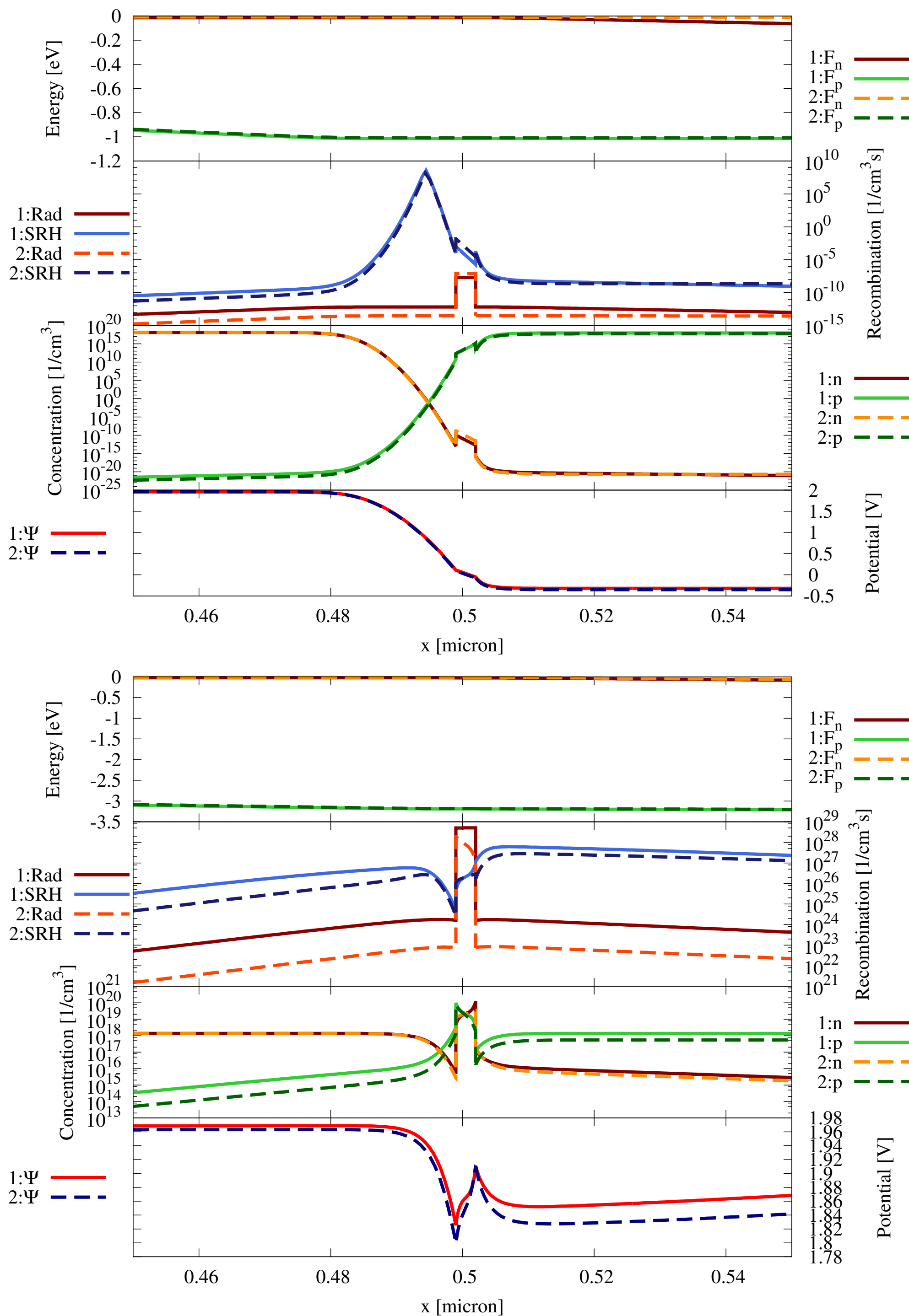

Figure 2.41: Comparison of quasi-Fermi levels, recombination rates, charge carrier concentrations, and the potential near the quantum well of *N-W-P d (si)* device simulated with our code(1) and with SiLENSe(2). The bias is 1 V for the upper plot and 3.5 V for the lower plot. The results are similar in the presented region for lower voltage, but differences arise when the bias increases, especially for the potential.

2.8.4 Computing carrier currents

In simulations of semiconductor luminescent devices, it is important to estimate precisely the current densities J_n, J_p, which contribute to the current density J. Current density is an important property of a device, as it may be used to obtain I-V characteristics or L-I (light-current) characteristics, which are measured experimentally. These characteristics allows to compare numerical results with real devices, and they are important factor in design of the semiconductor devices.

In this section, we discuss main problems related to calculation of current densities J_n, J_p. As an example, we simulate a p-n homojunction with 100 nm GaN n-type region, doped moderately with $1 \times 10^{18}\,\mathrm{cm}^{-3}$ donors and 100 nm GaN p-type region, doped with $1 \times 10^{19}\,\mathrm{cm}^{-3}$ acceptors. We will apply forward bias to this simple structure.

Before analysis of the simulation results, we would like to briefly discuss our expectations, basing on physical grounds. For low bias (figure 2.42), the built-in potential prohibits carriers from spatial traverse across the device, but electrons and holes recombine in the depleted regions. These carriers are constantly supplemented by carriers injected via contacts (see figure 2.43). Thus in the n-type region we expect some electron current and negligible hole current, and quite the reverse in the p-type region. Under high bias, the depleted region disperses and there is no potential barrier, thus the current carriers can traverse across the device.

2.8.4.1 Using the definition

The most natural approach in calculating currents is to directly use definitions (2.5.9). Unfortunately this simple method does not give satisfactory results, as it is demonstrated in figure 2.44). Let us focus on electron current density first. These results indicate, that it is only present in the depleted region and in the p-type region. In the n-type region it is nonexistent. This is in contradiction with our discussion. Analogous effect is observed for the hole current density. Moreover the total current density J is not constant, as it varies heavily between the depleted region and the rest of the device. This effect is completely nonphysical.

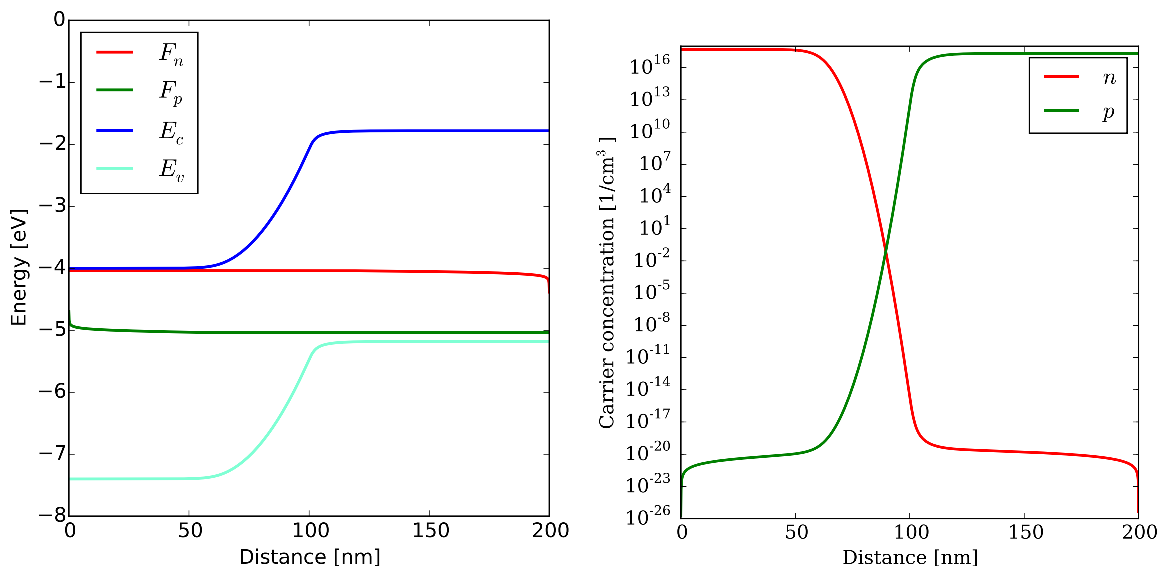

Figure 2.42: Band diagram and concentrations of carriers for a GaN diode, used in the simulations. Bias is $\psi_D = 1\,\mathrm{V}$.

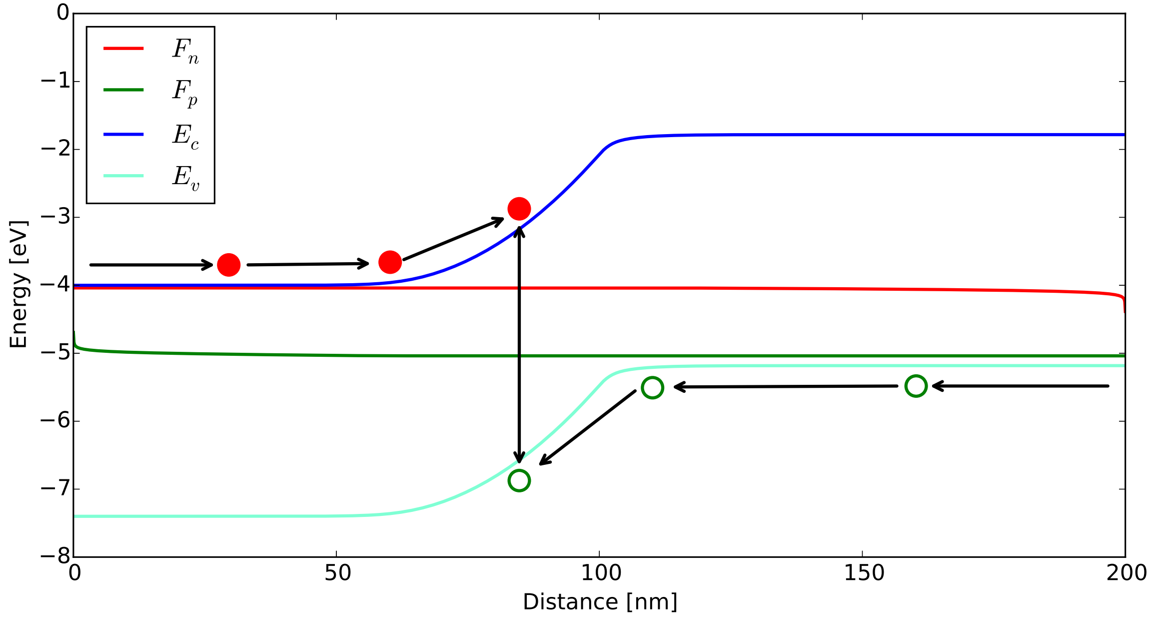

Figure 2.43: Schema of carrier transport in p-n junction under low bias. The meaning of symbols is as follows: electrons — red circles, holes — green circles, recombination — vertical arrow, spatial movement of carriers — horizontal arrows.

These numerical artifacts may be explained if we take into account the floating-point arithmetic. To compute concentration of currents with definitions (2.5.9), the derivative of quasi-Fermi levels F_n, F_p must be available. In our simulations, they were computed by finite differences from functions F_n, F_p. On the n-type region, the electron quasi-Fermi level F_n is almost constant, as well as the hole quasi-Fermi level on the p-type region (figure 2.42). In spite of huge variations of the carrier concentrations n, p, it appears that relative difference of quasi-Fermi levels can be smaller that the floating-point number precision, which leads to zero derivative on large regions of the device.

Therefore we cannot compute currents form definition, as this method fails even for such a simple case as p-n junction.

2.8.4.2 Using continuity equations

Instead of definitions of J_n and J_p, we may use continuity equations (2.5.15). Then we obtain a first order equations on J_n, J_p

$$\begin{aligned} \nabla \cdot J_n(x) &= qR(x), \\ \nabla \cdot J_p(x) &= -qR(x). \end{aligned} \tag{2.8.1}$$

We assume that the van Roosbroeck equations are numerically solved and we have approximations of ψ, F_n, F_p. Thus R is a known function.

Still we need to use definition (2.5.9) for boundary conditions. In figure 2.44 we clearly see, that on large parts of the device these formulas give reasonable results. However, near the boundary there are often some artifacts related to boundary values, as shown in figure 2.44. In one dimension, we can easily solve equations (2.8.1 starting form an arbitrary $x \in \Omega$. The solutions are given by

$$J_n(x) = J_n(x_0) + q\int_{x_0}^{x} R(\xi)\mathrm{d}\xi, \qquad J_p(x) = J_p(x_1) - q\int_{x_1}^{x} R(\xi)\mathrm{d}\xi, \tag{2.8.2}$$

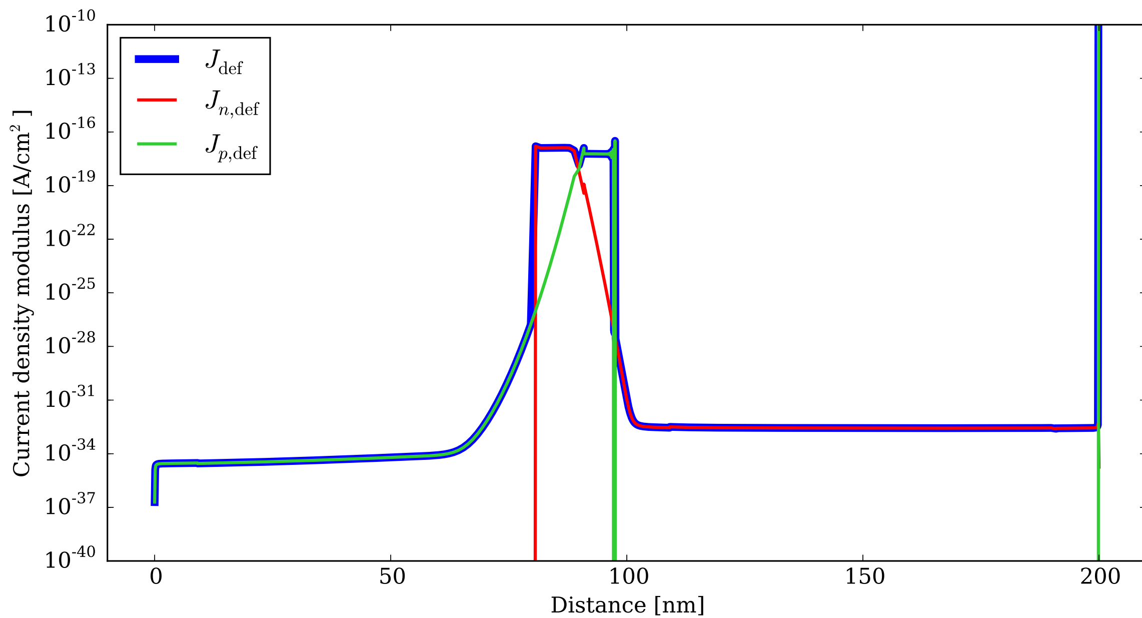

Figure 2.44: Currents calculated from definition (2.5.9). Note subsets where currents become zero. Total current J is definitely not constant.

where the choice of $x_0, x_1 \in \Omega$ is arbitrary.

In figure 2.45 currents calculated by the above method are presented. In comparison with results of previous method, this result is more regular and it corresponds to our physical intuition (see figure 2.43), i.e. there is almost constant electron current in n-type region and hole current in p-type region, and they switch in the depleted region due to recombination. Moreover the total current density J is constant.

If we compare both methods (figure 2.46), we clearly see that both methods give similar results on significant part of device. Differences emerge due to limitation of floating-point arithmetic, as explained in section 2.8.4.1.

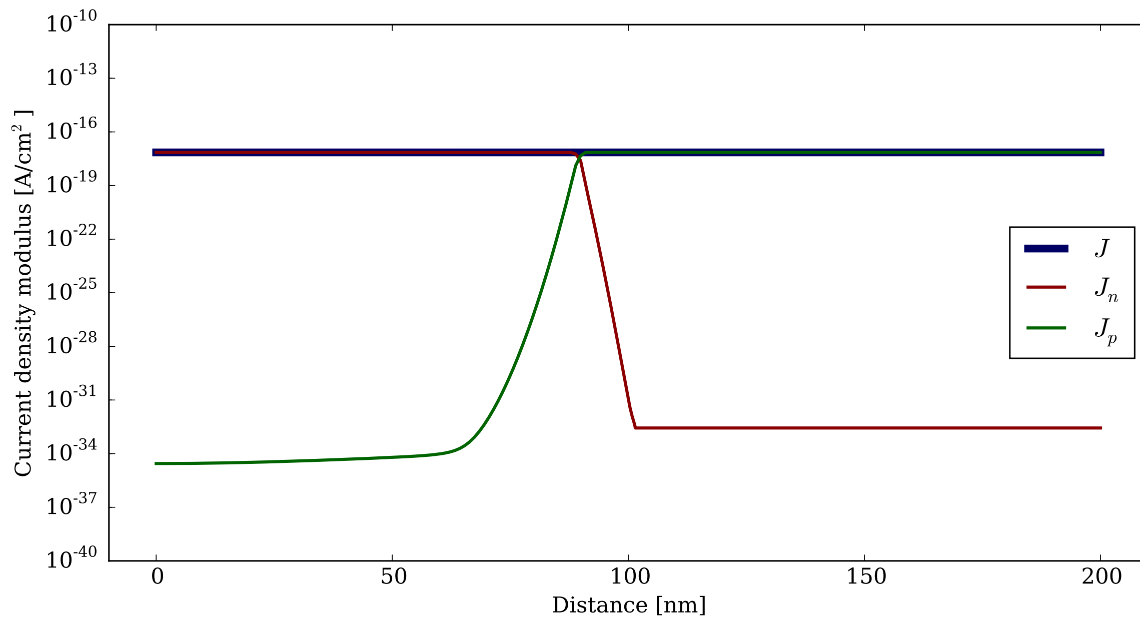

Figure 2.45: Currents calculated by integration of the recombination.

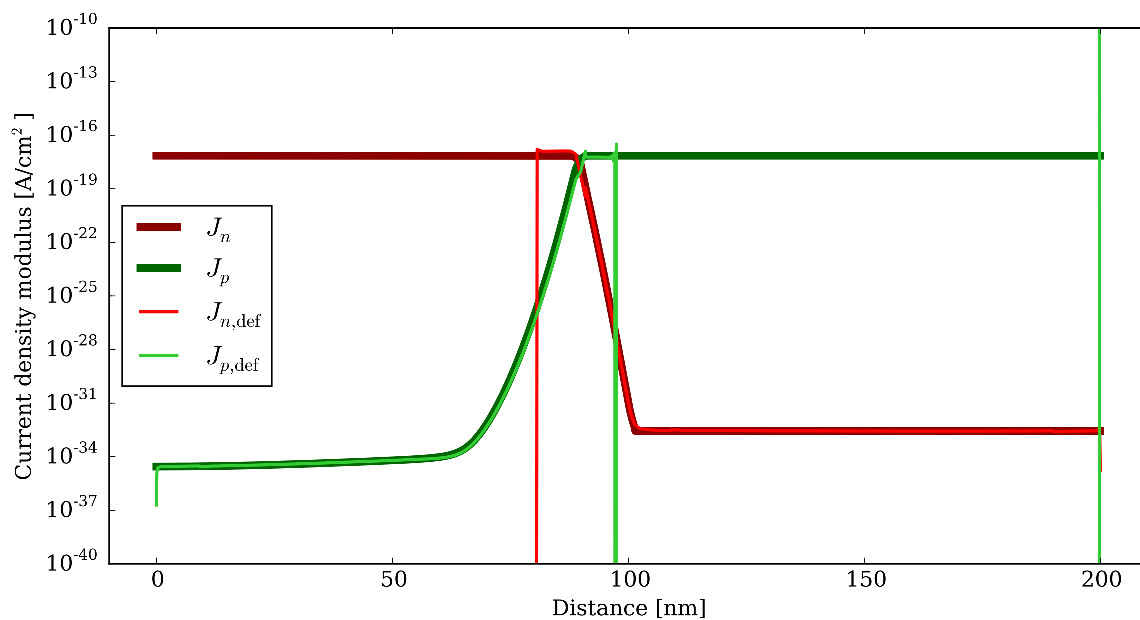

Figure 2.46: Comparison of currents calculated from definition (2.5.9) $J_{n,\mathrm{def}}$, $J_{p,\mathrm{def}}$, and by integration of recombination J_n, J_p.

2.8.5 Trap-assisted tunneling effect on characteristics of gallium nitride diodes

In section 2.7.1 we described the trap-assisted tunneling effect, which increases the SRH recombination. Then in section 2.8.2.3 we presented simulations, which indicates that depending on the length of the depleted region, this effect may be more or less significant.

In this section, we would like to study in detail the impact of the trap-assisted tunneling on the operation of the GaN p-n homojunction. In particular, we would like to present a comparison of simulations with experiments.

2.8.5.1 Comparison with experiments

For the comparison, a realistic device suggested by Smalc-Koziorowska et al. was used [106]. The diode was divided into three layers: 60 micron thick n-GaN substrate with donor concentration $N_d = 2 \times 10^{18}$ cm^{-3}, 0.5 micron thick n-GaN layer with $N_d = 1.4 \times 10^{19}$ cm^{-3} and 0.5 micron thick p-GaN layer with acceptor concentration $N_a = 5 \times 10^{19}$ cm^{-3}. Computations with and without the trap-assisted tunneling modification were performed.

In this set of simulations we used material parameters from tables 2.5, 2.6. We assume that the p-GaN is doped with shallow Mg acceptors, which do not play important role in the non-radiative recombination. We expect the p-GaN deep trap level in Mg-doped GaN to be related to the nitrogen

Parameter	Symbol	Value	Details
Relative permittivity	ε_r	8.9	Ref. [68]
Acceptor degeneracy level	g_a	2	Ref. [110, 99]
Donor degeneracy level	g_d	2	Ref. [79]
Band gap	$E_c - E_v$	3.4 eV	Ref. [77]
Acceptor level (Mg, shallow)	E_a	0.17 eV	Ref. [68]
Donor level (Si, shallow)	E_d	0.02 eV	Ref. [68]
Electron effective mass	m_n, m_n^{tun}	0.2	Ref. [89, 68]
Hole effective mass	m_p, m_p^{tun}	1.7	Ref. [41]
Electron mobility	μ_n	$200 \ \frac{\mathrm{cm}^2}{\mathrm{Vs}}$	Ref. [95]
Hole mobility	μ_p	$5 \ \frac{\mathrm{cm}^2}{\mathrm{Vs}}$	Ref. [70]
Radiative recombination constant	C^{rad}	$1.1 \times 10^{-10} \ \frac{\mathrm{cm}^3}{\mathrm{s}}$	fitting param.
Temperature	T	300 K	room temp.
Acceptor level (deep, n-GaN)	$E_t - E_v$	1.0 eV	Sec. 2.8.5.1
Donor level (deep, p-GaN)	$E_c - E_t$	0.5 eV	Sec. 2.8.5.1

Table 2.5: Material parameters of gallium nitride, used in our numerical simulations.

Impurity concentration	N_t	$\tau_n = \left(C_n^{SRH}\right)^{-1}$	$\tau_p = \left(C_p^{SRH}\right)^{-1}$
$N_d = 10^{18}\mathrm{cm}^{-3}$	$5 \times 10^{16}\mathrm{cm}^{-3}$	5×10^{-8}s	5×10^{-10}s
$N_d = 5 \times 10^{19}\mathrm{cm}^{-3}$	$2.5 \times 10^{18}\mathrm{cm}^{-3}$	10^{-9}s	5×10^{-11}s
$N_a = 5 \times 10^{19}\mathrm{cm}^{-3}$	$10^{18}\mathrm{cm}^{-3}$	7×10^{-12}s	7×10^{-10}s

Table 2.6: Trap concentrations and carrier lifetimes based on impurity concentrations, used in our simulations. The values τ_n, τ_p were the fitting parameters.

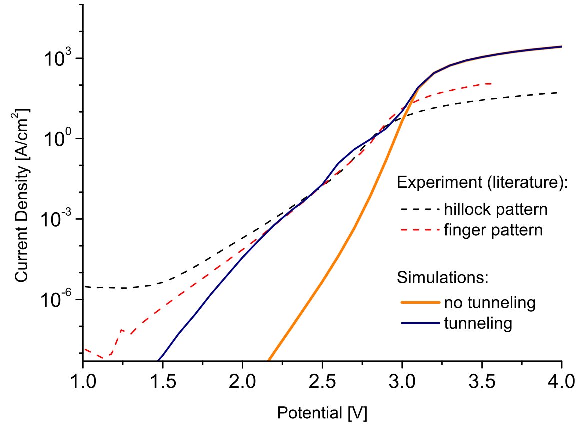

Figure 2.47: Comparison of the experimentally obtained I-V characteristics of p-n junctions from the article [106] and simulation results with and without the trap-assisted tunneling. In the simulations, the n-GaN donor concentration $N_d = 1.4 \times 10^{19}$ cm^{-3} with the trap concentration $N_t = 7 \times 10^{17}$ cm^{-3}, and the p-GaN acceptor concentration $N_a = 5 \times 10^{19}$ cm^{-3} with the trap concentration $N_t = 10^{18}$ cm^{-3} were assumed near the space charge region.

vacancies with energy level around 0.43 eV[85, 88] or nitrogen antisites at around 0.8 eV[111, 108]. Thus, the deep donor energy level $E_c - E_t = 0.5$ eV is chosen. In the n-GaN the deep trap level may be caused by the V_{Ga}-O_N and V_{Ga}-Si_{Ga} complexes[86] with energy levels 1.1 eV and 0.9 eV, respectively, or by Si[14] at 1.2 eV. Thus the deep acceptor energy level $E_t - E_v = 1.0$ eV is taken in the n-GaN. We assume trap concentrations to be about 2% of the p-GaN acceptor concentration and 5% of the n-GaN donor concentration.

The comparison of the I-V characteristics of the simulated devices with the experimental data from [106] is presented in figure 2.47. Both computed characteristics agree for bias above 3 V. On the other hand, for lower bias the curves present substantially different behavior. The total current is up to five orders of magnitude higher when the tunneling is taken into account. The agreement with the experiment is much better for the tunneling case. In the comparison, we used the interval [1.5 V, 3 V]. The precision of the measurement is too low below about 1.5 V to get reliable values. Above 3 V the resistance of contacts dominates over the resistance of the structure, that is not taken into account in the simulations.

To identify the physical factors for the difference between tunneling included or not, the recombination must be examined. The recombination rates for various potentials for both simulations are presented in figure 2.48. Note that the recombination in the depleted region for the tunneling included is much higher, especially for lower bias. When the potential increases, the difference is smaller. In fact, this is what we should expect. The tunneling of carriers to the trap level increases the number of electrons and holes participating in the non-radiative recombination, and therefore it increases SRH recombination rate. Tunneling distances are larger for higher voltages, because of straightening the bands. Thus a contribution of the tunneled carriers to the recombination decreases, as the tunneling over the barrier becomes less probable.

2.8.5.2 Impact of n type doping level on I-V characteristic of a diode

The example presented in the previous section demonstrates that the tunneling to the trap level may accelerate the non-radiative recombination, at least for low and moderate potentials. The effect may be neglected for devices which are supposed to work in high biases, above 3 V, but it is very

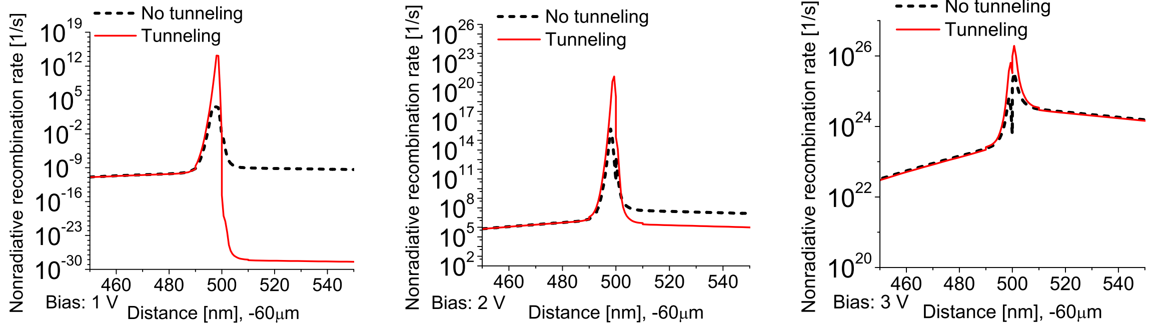

Figure 2.48: Rates of the non-radiative recombination on trap levels in the space charge region for applied potentials 1 V, 2 V and 3 V taken from the simulations with and without the trap-assisted tunneling. Increase of the bias leads to diminish of the effect of the tunneling. The concentration of the impurities are the same as for figure 2.47.

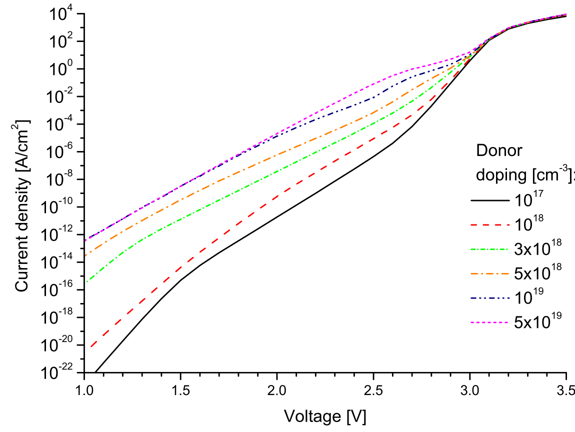

Figure 2.49: Simulation of the impact of a donor doping of one micron p-n junction with the p-GaN acceptor concentration $N_a = 5 \times 10^{19}$ cm^{-3} with the trap concentration of $N_t = 10^{18}$ cm^{-3}. The trap concentration in the n-GaN is assumed to be 5% of the donor concentration. The trap-assisted tunneling was active in the simulation. Increase of the donor doping significantly affects the current for low biases.

disadvantageous for devices operating in low voltages.

As we have shown, the mentioned phenomena has a large impact on currents when a depleted region is narrow and heavily-doped. We would like to verify whether an extension of the depleted region, which renders the tunneling of electrons to be unlikely, leads to a decrease in the trap-assisted recombination rate.

For simplicity, we would like to focus on a short p-n homojunction which has two layers: 0.5 micron thick n-GaN and 0.5 micron thick p-GaN. To extend the space charge region, one can decrease the doping of the layers. However, this is rather unfavorable in the p-GaN, as the common GaN acceptor, magnesium, is not so shallow. Then the resistance of the p-GaN would increase greatly. This is not a problem with the n-GaN, where the silicon donor is very shallow.

Therefore we performed a simulation of devices with the constant acceptor doping $N_a = 5 \times 10^{19}$ cm^{-3} and the donor concentration varying from 10^{18} cm^{-3} to 5×10^{19} cm^{-3}. The resulting I-V curves are presented in figure 2.49. We see that increase in the donor concentration leads to a

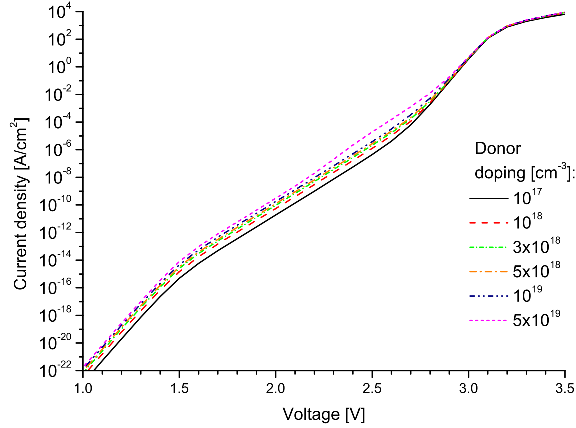

Figure 2.50: Simulation of the impact of a donor doping of one micron p-n junction with the p-GaN acceptor concentration $N_a = 5 \times 10^{19}$ cm^{-3} and the trap concentration of $N_t = 10^{18}$ cm^{-3}. The trap concentration in n-GaN is assumed to be 5% of the donor concentration. The trap-assisted tunneling was omitted in the simulation. There is no significant influence of the donor doping on the current.

significant increase in the current density for low and moderate voltages. For potentials above 3 V the differences are barely visible.

The simulation results are a numerical evidence of the already presented theoretical considerations about the tunneling distance. The simulations leads to the following conclusion: in order to reduce the non-radiative recombination in a p-n junction, one should extend the space charge region. For example, by reducing concentration of donors in the n-GaN.

We would like to ensure that the presented result is an effect of the depleted region narrowing and the trap-assisted tunneling, not just an increase of the donor concentration. Therefore these simulations were repeated with the tunneling inactive (figure 2.50). As we see, in this case the current density do not change considerably with the doping.

For the sake of comparison, band diagrams of the devices with the donor concentration 10^{18} cm^{-3} and 5×10^{19} cm^{-3} are shown in figures 2.51, 2.52. Note that for low donor concentrations or high currents, there is no significant difference between the case with the tunneling effect active or not. Thus the potential in the depleted region remains almost parabolic and quasi-Fermi levels are almost constant. The difference in the current densities follows just from the magnitude of the non-radiative recombination.

On the other hand, for low biases and high donor doping, the recombination becomes so strong that the quasi-Fermi level varies rapidly near the junction. This fluctuation poses a problem with the theory presented by Hurkx et al. [51], as the expressions presented in the article and used in the simulations are valid under the assumption of constant quasi-Fermi levels in the space charge region. Therefore the calculations for this case should be revised with a more subtle approach in the future. Nevertheless, such a high recombination is caused by the trap-assisted tunneling, not just the narrow space charge region (cf. figure 2.52). Thus the increase of the current in comparison with no tunneling case still is expected, but it may be misvalued by the model.

2.8.5.3 Discussion

Our simulations prove the trap-assisted tunneling plays an important role in the Shockley-Read-Hall recombination. The results reveal that this effect may modify considerably I-V characteristics and band diagrams of devices. The change of the total current density is natural, as it is directly dependent

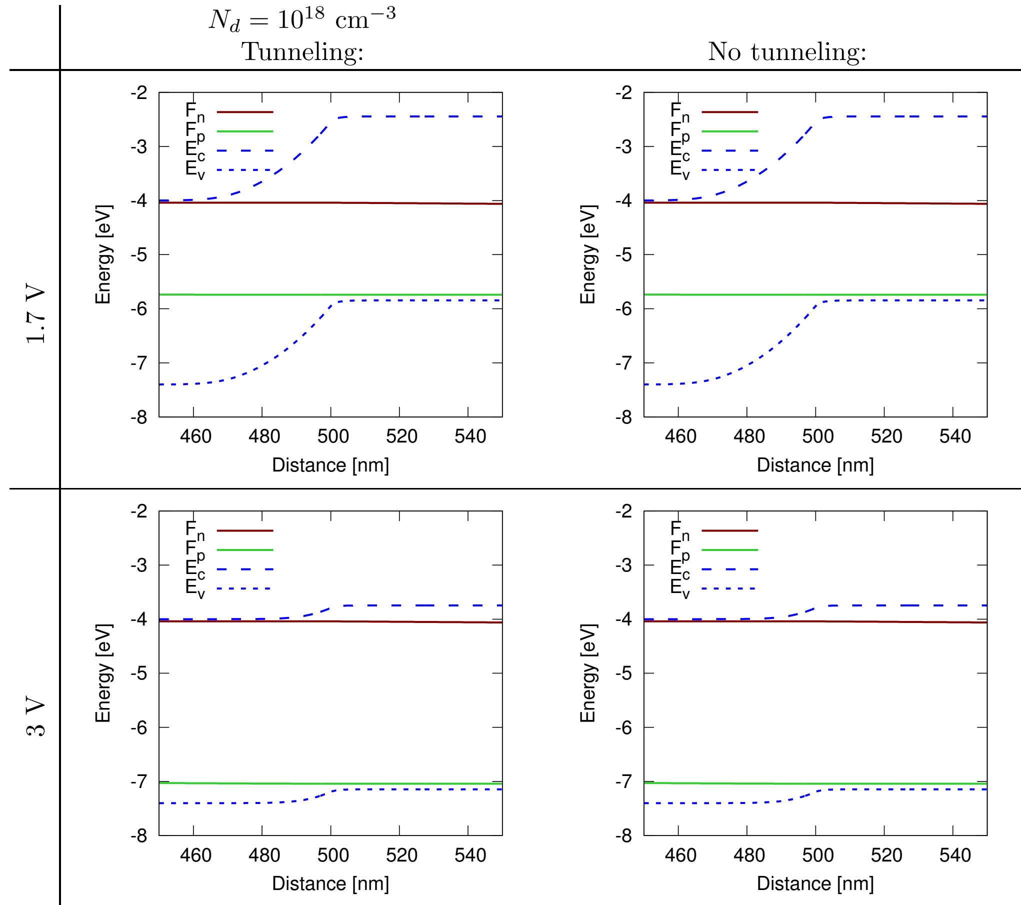

Figure 2.51: Comparison of band diagrams of the one micron p-n junction for potentials 1.7 V, 3 V, with the n-GaN donor doping concentration $N_d = 10^{18}$ cm^{-3} and the p-GaN acceptor doping concentration $N_a = 5 \times 10^{19}$ cm^{-3}. The respective trap concentrations are $N_t = 5 \times 10^{16}$ cm^{-3} and $N_t = 10^{18}$ cm^{-3}. The simulations were performed in two variants: with the trap-assisted tunneling active or not. In both cases changes of the band diagram are hardly noticeable.

on the recombination rate. However, the generated recombination is so big that the electron quasi-Fermi level varies rapidly in the proximity of a recombination peak. It happens in a significant part of luminescent devices.

The result implies that the assumptions made by Hurkx et al. [51] could be invalid, thus the simulation with a more general method should be performed. On the other hand, the effect cannot be considered as a small perturbation. It influences not only the current, but also the band diagram and should be included in the computations.

The performed simulations of a p-n GaN junction demonstrates that a high doping of the junction (above 10^{19} cm^{-3}) leads to a significant increase of the non-radiative recombination. As the results suggest, the issue may be settled by lengthening the space charge region by decreasing the doping of the n-GaN layer.

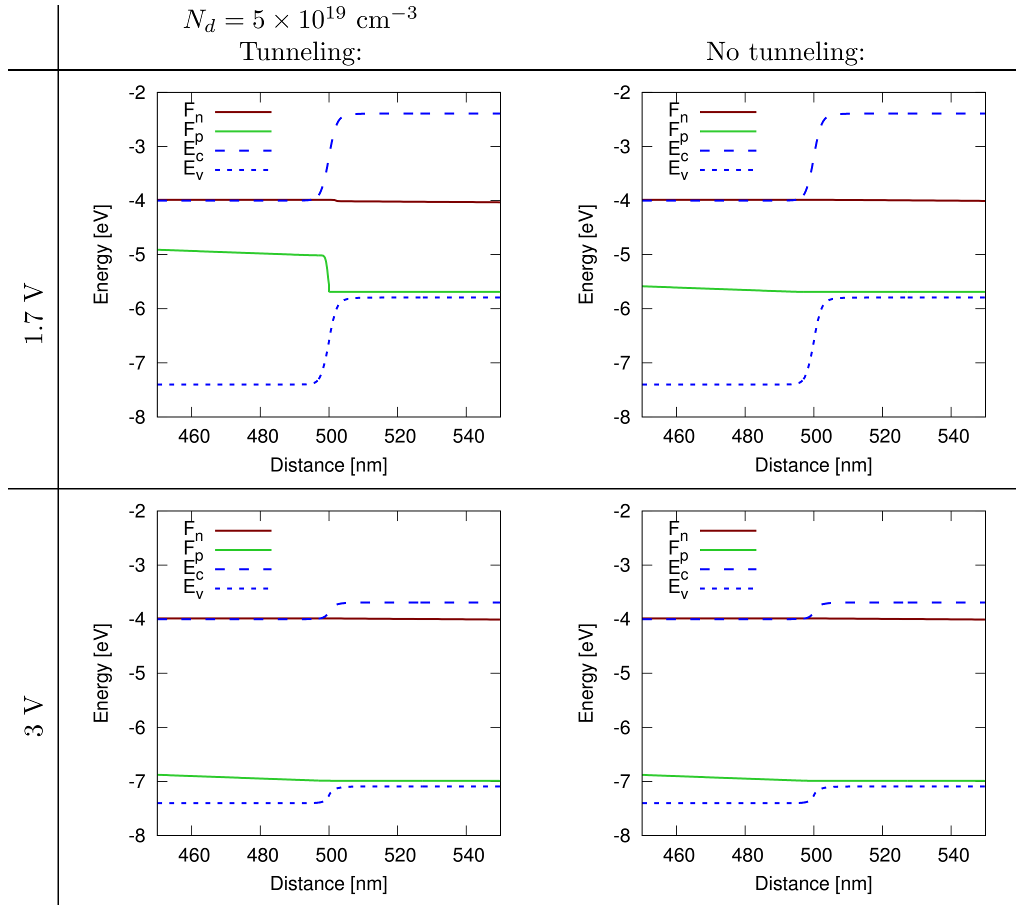

Figure 2.52: Comparison of the band diagrams of the one micron p-n junction for potentials 1.7 V, 3 V, with the n-GaN donor doping concentration $N_d = 5 \times 10^{19}$ cm^{-3} and the p-GaN acceptor doping concentration $N_a = 5 \times 10^{19}$ cm^{-3}. The respective trap concentrations are $N_t = 2.5 \times 10^{18}$ cm^{-3} and $N_t = 10^{18}$ cm^{-3}. The simulations were performed in two variants: with the trap-assisted tunneling active or not. Note a hole quasi-Fermi fluctuation for the tunneling included and the low bias.

2.9 Light-emitting diodes and laser diodes

2.9.1 Introduction

A *light-emitting diode* (LED) is a semiconductor light source. It is a device based on a p-n diode, where upon application of a sufficient voltage the light is emitted, which is generated due to radiative recombination of electrons and holes.

Emission of the photons occurs mainly in the quantum wells. As explained in section 2.4, QWs are supposed to localize the quasiparticles of both species in large concentrations, to make the radiative recombination efficient. Thus they are located generally in the proximity of the depleted region.

In theory this simple schema is sufficient, in practice structure of modern GaN-based LEDs is much more complicated due to several issues. While holes generally do not escape the active region due to low mobility, this is not the case for electrons. To prevent electron leakage, the *electron blocking layer* (EBL) is introduced. It is a quantum barrier, localized typically between quantum wells and a p-type region, made of AlGaN. This material is used, as the band offset between GaN and AlN is realized mostly in the conduction band edge. Thus EBL mostly blocks electrons. For holes the barrier is much

smaller.

On the other hand, quantum wells are made of InGaN. Its forbidden zone energy can be chosen to match wide range of the visible light spectrum (violet, blue, green, ...) upon appropriate choice the indium-to-gallium rate. The problem of growing quantum wells followed by the EBL is that while InGaN is grown in approximately 800 °C, AlGaN is grown above 1400 °C. It is likely that neighboring InGaN layers will be destroyed during the latter process. Thus intermediate layer is often introduced.

Blue optoelectronics is based on gallium nitride grown mostly in polar directions. In this case, every interface between different materials (GaN, InGaN, AlGaN) introduces polarization charges, which affect the band structure of a device. In particular, an electric field in quantum wells is present, which leads to separation of electrons and holes to opposite sides of a quantum well, reducing the probability of radiative recombination (*quantum-confined Stark effect*, QCSE).

An important problem in LEDs design is to prevent the nonradiative recombination. Nonradiative recombination may occur due to material impurity in areas of large concentration of electrons and holes. Main countermeasure is to grow possibly pure material. On the other hand, additional care must be taken to eliminate areas of both carrier species aside from the active region.

In simulations of LEDs it is important to compare the optical power of devices. The drift diffusion model simulates only electric properties, we do not simulate optical properties. However, we can estimate the photon emission rate. A radiative recombination process results in emission of a photon with energy approximately equal to the bandgap energy. The emitted photon may be reabsorbed by the device, but we assume that reabsorption is already included in the radiative coefficient C_{rad} (see section 2.6.1.1). Then we should only account for photons emitted from the active region of a device, from quantum wells. Photons emitted from outside of the quantum wells are are ignored, as they typically have different wavelength, and thus they do not contribute to target light spectrum.

Thus assume that the recombination rate R of (2.5.24) decomposes to

$$R(x) = R_L(x) + R_N(x), \tag{2.9.1}$$

where R_L accounts for the radiative recombination rate in the quantum wells contributing to the emitted light, while R_N is the non-radiative recombination and radiative recombination outside of the quantum wells.

Functions R_L and R_N may be extracted from simulation's output. If we perform a simulation for a given bias, we may then estimate the photon emission rate as

$$\text{photon emission rate} := \int_\Omega R_L(x)dx. \tag{2.9.2}$$

The optical power may be then estimated as

$$L = \int_\Omega E_g(x)R_L(x)dx. \tag{2.9.3}$$

Similarly we can estimate power loss due to rogue recombination

$$\text{recombination power loss} = \int_\Omega E_g(x)R_N(x)dx. \tag{2.9.4}$$

The power loss in LEDs is not only due to rogue recombination. Other causes are the electron (and hole) thermalization and recombination of leakage carriers on the contacts. Power loss due to these factors (combined) may be estimated by subtracting optical power and recombination loss from the total power:

$$\text{total power} = VI, \tag{2.9.5}$$

Table 2.7: Simulation results of the laser structure (section 2.9.2) versus aluminum content in the electron blocking layer for 5 V bias. I stands for the current. Optical power is estimated as the radiative recombination amount in the QW multiplied by the band gap. Power loss accounts for the radiative recombination outside the QW, non-radiative recombination on the whole device, recombination of overflow carriers on contacts and the resistance.

EBL Al	I [A]	Power loss [W]	Opt. power [W]	Eff.
0%	0.364	1.713	0.108	6%
3%	0.528	2.471	0.167	6%
6%	0.373	1.551	0.313	17%
9%	0.235	0.740	0.435	37%
12%	0.169	0.411	0.435	51%
15%	0.155	0.349	0.424	55%
20%	0.147	0.328	0.408	55%
30%	0.134	0.298	0.369	55%

where I is the current and V is the voltage (bias). We define optical wall-plug efficiency as

$$\text{efficiency} := \frac{L}{\text{total power}}. \tag{2.9.6}$$

In this section, we also would like to present simulations of *laser diodes* (LDs). In contrast to LEDs, light emission in laser diodes is due to stimulated emission effect. For this effect to occur, the active part of a device must be supplemented with mirrors and additional reflecting layers, called claddings. Model used in our work does not account for the optical properties. However, we are still interested in electrical properties of these heterostructures. To simulate these devices to some extent without subtle optic model, we assume that the stimulated emission increases greatly the radiative recombination rate.

2.9.2 Aluminum content in EBL

Our first problem is to determine the optimal level of aluminum doping in the electron blocking layer. Simulations presented in this section are based on the following example of the laser heterostructure. It starts with one micron n-GaN layer, followed by 550 nm n-$Al_{8\%}Ga_{92\%}N$ cladding and 100 nm n-GaN waveguide. Then the active part consists of a 20 nm n-$In_{1.5\%}Ga_{98.5\%}N$ layer, a 4 nm undoped $In_{17\%}Ga_{83\%}N$ quantum well and a 22 nm n-$In_{1\%}Ga_{99\%}N$ cap. It is followed by a 20 nm p-AlGaN EBL, a 100 nm p-GaN waveguide, a 400 nm p-$Al_{8\%}Ga_{92\%}N$ cladding and a 30 nm p-GaN contact layer. We performed simulations of this with aluminum content in the EBL varying from 0 to 30%. Results are presented in tables 2.7, 2.8. Increasing aluminum content generally also increases efficiency of the device, but only up to a certain level. For example, under 5 V bias we reach the maximum of 55% efficiency for 15% Al in the EBL.

The explanation of this phenomena is as follows. Lower Al content decreases the barrier for electrons, allowing them to escape from the active zone to the p-type. Then they either recombine with abundant holes in p-type (recombination loss), or they reach the contact. In the latter case the electrons generates energy loss not only at the contact, but also along its way through the p-type, as the potential differences for biases > 4 V are quite significant there (conduction loss).

Table 2.8: Simulation results of the laser structure (section 2.9.2) versus aluminum content in the electron blocking layer for 6 V bias.

EBL Al	I [A]	Power loss [W]	Opt. power [W]	Eff.
0%	0.689	3.949	0.185	4%
3%	0.679	3.846	0.230	6%
6%	0.650	3.424	0.475	12%
9%	0.481	2.125	0.762	26%
12%	0.296	1.039	0.739	42%
15%	0.266	0.868	0.727	46%
20%	0.255	0.824	0.709	46%
30%	0.240	0.776	0.665	46%

Table 2.9: Simulation results for the laser structure (section 2.9.2) versus aluminum content in the electron blocking layer for 5 V. Indium content in the cap was increased to 8%.

EBL Al	I [A]	Power loss [W]	Opt. power [W]	Eff.
0%	0.408	1.602	0.439	22%
3%	0.212	0.557	0.501	47%
6%	0.183	0.424	0.491	54%
9%	0.177	0.404	0.482	54%
12%	0.174	0.396	0.473	54%

It is also interesting why the conduction loss is much greater than the recombination loss for 5 V and 6 V bias. Intuitively we could expect any escaped electron to recombine in p-type due to abundance of holes. However due to the electric field, the concentration of electrons is several orders of magnitude smaller that the hole concentration and they move fast. And even if they recombine, they still travel to the contact losing its energy, forming holes traveling in the opposite way.

For the optical power we observe a different pattern: it also increases rapidly up to a maximum level, but then it slightly drops with the Al content. For example, under 5 V bias, the maximal optical power is reached for 9-12% Al. This is an additional problem with a high Al content in EBL, as it also creates a mild barrier for holes. It is disadvantageous, as it prohibits holes from reaching the active zone. Note that in general increasing the EBL barrier leads to lower total current.

We have performed also simulations of the laser structure with 8% In in the cap (table 2.9), to study a more general setting. Then the barrier on the cap/EBL interface is higher for the same aluminum content in the EBL. The general trend is similar as for the previous case, but then the maximal efficiency/optical power is reached for 3-6% Al in EBL.

These two examples leads to the conclusion that the aluminum content in EBL should be chosen in accordance to the material in the n-type cap preceding the EBL. Higher indium content allows to decrease Al in EBL.

Table 2.10: Simulation results for the laser structure (section 2.9.3) versus magnesium concentration in the p-type 5 V and 6 V bias.

Mg conc. [cm^{-3}]	U [V]	I [A]	Power loss [W]	Opt. power [W]	Eff.	Resist. [Ohm]
1×10^{18}	5.0	0.004	0.008	0.010	54%	1355
5×10^{18}	5.0	0.046	0.105	0.125	54%	109
1×10^{19}	5.0	0.075	0.171	0.204	54%	67
3×10^{19}	5.0	0.137	0.313	0.374	54%	36
5×10^{19}	5.0	0.175	0.399	0.476	54%	29
Mg conc. [cm^{-3}]	U [V]	I [A]	Power loss [W]	Opt. power [W]	Eff.	Resist. [Ohm]
1×10^{18}	6.0	0.016	0.052	0.043	45%	376
5×10^{18}	6.0	0.099	0.323	0.269	45%	61
1×10^{19}	6.0	0.155	0.509	0.423	45%	39
3×10^{19}	6.0	0.284	0.932	0.774	45%	21
5×10^{19}	6.0	0.365	1.198	0.995	45%	16

2.9.3 Mg doping of p-type

In this section we would like to discuss the magnesium acceptor doping on operation of laser structures. In simulations we use the structure described in section 2.9.2 with variable Mg concentration in the p-waveguide and p-cladding. Length of n-cap was also shortened to 2 nm to reduce impact of the recombination loss on the simulation. The indium content in the quantum well is 25% and the aluminum content in the EBL is 20%.

Results are presented in table 2.10. We observe that the optical power of the device increases with the Mg doping. Such a result is expected, as generally it is a consequence of the high activation energy of the magnesium acceptor. However, a more subtle reason is related to polarization charges on interfaces of laser structures. Note the resistance in function of the Mg concentration. We observe the decrease of the resistance of the structure when we increase the Mg concentration, but the most spectacular difference is between concentrations 1×10^{18} cm^{-3} and 5×10^{18} cm^{-3}. High resistance for 1×10^{18} cm^{-3} is caused by the polarization charge on the waveguide/cladding interface (see figure 2.53). It pushes out holes away from the interface deep into the waveguide and increasing greatly the resistance of the depleted fragment. As a consequence, most of the potential difference is located in the waveguide. If we increase the concentration of holes, this polarization charge will be screened and its impact on the resistance can be significantly reduced.

Note that the resistance decrease has no noticeable influence on the efficiency of the device, as it do not increase the carrier overflow.

The conclusion of this results is to use the highest possible magnesium concentration in the p-type. Please note, however, that the numerical model used in this study do not account for increased absorption in such case, which could make the device less efficient in real experiment. Also growth of highly Mg-doped layers, depending on a growth method, may lead to severe problems, like the polarization inversion or acceptor passivation. Therefore we suggest to make the concentration possibly high, but the actual concentration level should be adjusted experimentally and may vary depending

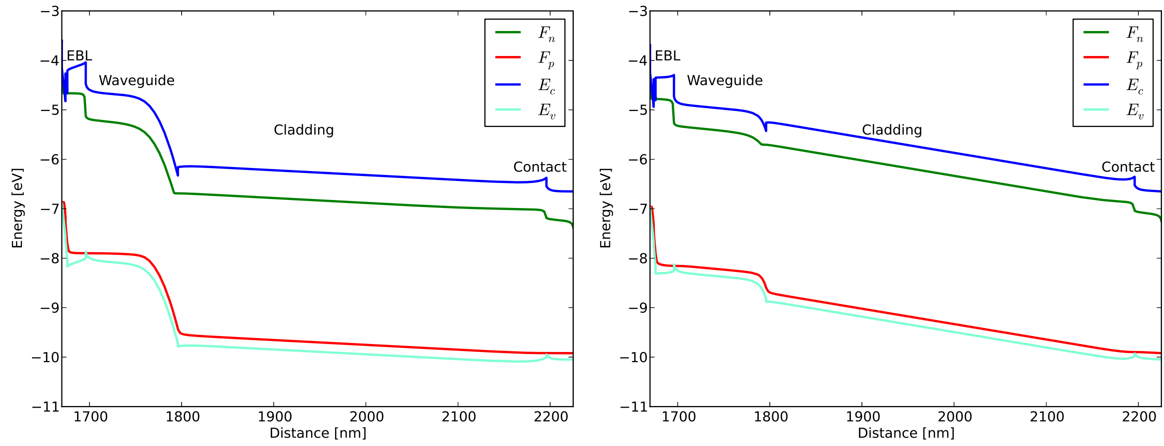

Figure 2.53: Band diagram of the p-type of the laser structure for Mg doping 1×10^{18} cm^{-3} (left) and 1×10^{19} cm^{-3} (right) for 6 V bias.

on the growth method.

2.9.4 Number of quantum wells

We would like to discuss effect of number of quantum wells on the operation of the blue laser diodes. The quantum wells are often very narrow, as they should localize quasiparticles. Also in gallium nitride devices, quantum-confined Stark effect leads to spatial separation of electrons and holes due to electric field, so the quantum wells cannot be too wide. Thus to increase the recombination volume, most straightforward method is to simply increase number of quantum wells. We are interested whether increasing number of QWs will improve the efficiency and optical power of a laser diode.

As before, in simulations we use our model device from section 2.9.2 with 17% indium content in the quantum well is and 20% aluminum content in the EBL. Number of QWs is between 1 and 7.

The intuition behind increasing number of QWs suggests that increased volume should increase

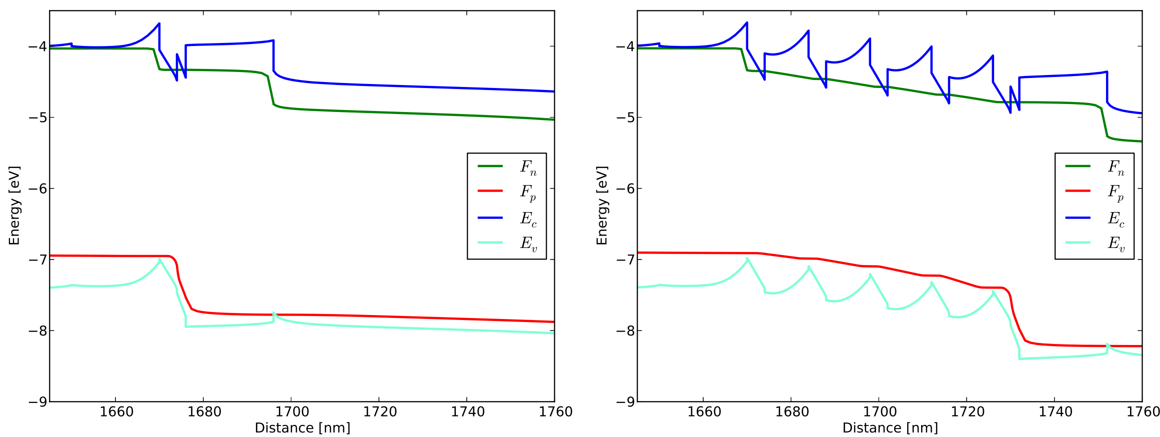

Figure 2.54: Comparison of band diagram of a laser structure with a single quantum well (left) against a laser structure with 5 QWs for 5 V bias.

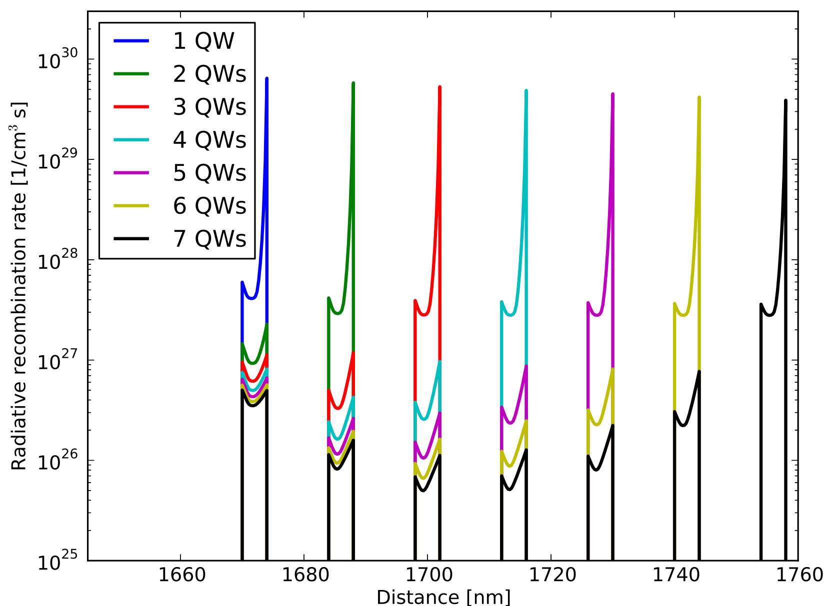

Figure 2.55: Comparison of radiative recombination rates in quantum wells for laser structures with number of QWs between 1 and 7 for 5 V bias.

Table 2.11: Optical power of a laser structure for 5 V bias versus number of quantum wells.

QWs	I [A]	Power loss [W]	Opt. power [W]	Eff.	Resist. [Ohm]
1	0.063	0.130	0.185	59%	79
2	0.058	0.120	0.171	59%	86
3	0.055	0.113	0.160	59%	92
4	0.052	0.107	0.151	59%	97
5	0.049	0.101	0.143	59%	102
6	0.047	0.096	0.136	59%	107
7	0.044	0.092	0.130	59%	113

the optical output, at least to some extent. Our simulations indicate that this must not be the case. In case of many quantum wells, the majority of the recombination takes place in a single quantum well (see figure 2.55). In this case, it is always a quantum well closest to p-type region, but this behavior depends on the balance between electron conductivity in n-type region and hole conductivity in p-type region.

We computed efficiency and optical power of these structures for bias 5 V (table 2.11). These results indicate that efficiency of laser diodes does not increase or degrade substantially on the number of quantum wells. However, increased amount of quantum wells increases resistance of a heterostructure.

We must emphasize that these simulations focus on the electrical properties of laser heterostructures. From this standpoint, a single QW is most favorable. However, increased number of quantum wells may be beneficial to the optical properties. Some of the QWs may also degrade during EBL growth, as mentioned earlier. Thus multiple QWs may be more practical in real experiment, as the resistance does not grow considerably.

2.10 Optical excitation in quaternary alloy AlInGaN

Simulations discussed in this section were performed in collaboration with experiments. The aim of this study was to determine the basic physical properties of quaternary AlInGaN alloys. AlInGaN layers, embedded in InGaN layers, are a potential material for construction of the electron blocking layers (EBLs) in LEDs and LDs. Results of this study may be found in [15].

We would like to focus on the simulations performed in the context of this research. Two structures were simulated, sample 1839 with higher indium content in the quaternary alloy and sample 1844 with

Layer name	Thickness [nm]	Sample 1839	Sample 1844
GaN 1	110	GaN	GaN
AlGaN	21	$Al_{14\%}Ga_{86\%}N$	$Al_{15\%}Ga_{85\%}N$
GaN 2	10	GaN	GaN
InGaN 1	40	$In_{10.8\%}Ga_{89.2\%}N$	$In_{6.8\%}Ga_{93.2\%}N$
AlInGaN	18	$Al_{18\%}In_{5.5\%}Ga_{76.5\%}N$	$Al_{15\%}In_{2.5\%}Ga_{82.5\%}N$
InGaN 2	18	$In_{0.6\%}Ga_{99.4\%}N$	$In_{7\%}Ga_{93\%}N$

Table 2.12: Schemata of devices used for study of the quaternary AlInGaN alloys. All layers are n-doped with $2 \times 10^{18}\,\mathrm{cm}^{-3}$ donor concentration.

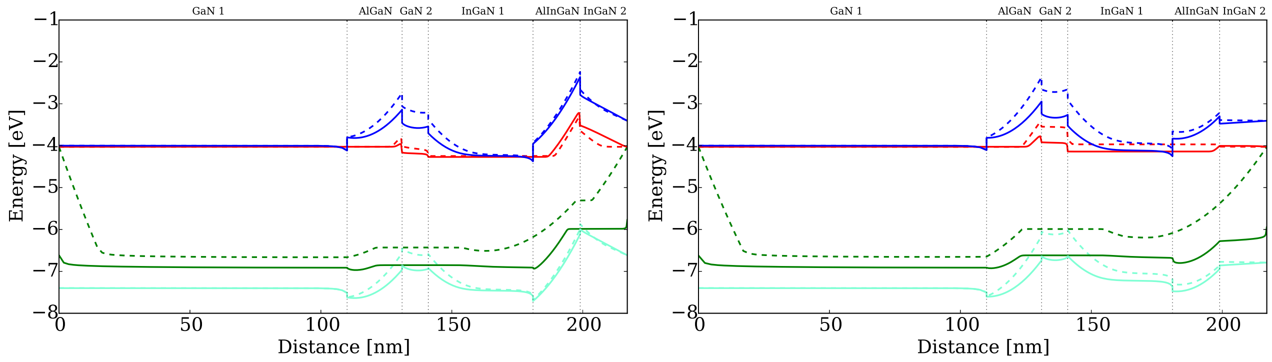

Figure 2.56: Band diagrams of the low-In (1844 - upper diagram) and high-In (1839 - lower diagram) samples under low (dashed lines) and high (solid lines) optical excitations. The blue and cyan lines represent conduction and valence bands respectively. The red and green lined represent electron and hole Fermi levels.

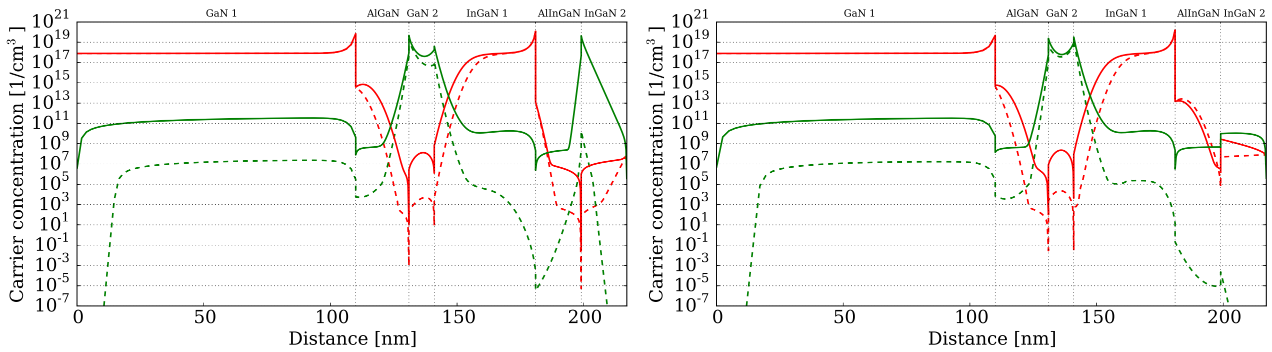

Figure 2.57: Carrier concentration in the low-In (1844 - upper diagram) and high-In (1839 - lower diagram) samples under low (dashed lines) and high (solid lines) optical excitations. The red and green lines represent electrons and holes, respectively.

lower indium content (table 2.12).

The optical emission is proportional to combined density of the electrons and holes. In the classical statistics used here, the carrier's density is the inverse exponential function of the band energy and the Fermi energy difference. Thus it is the lowest for the largest difference. As shown in figure 2.56, this energy difference is relatively small for the uniform regions and it changes drastically in the structure regions suggesting huge density changes there, thus it is useful to plot the density of carriers across the samples. It is expected that the electron density dominates in n-doped bulk GaN. The relation between densities in other regions may be compared using diagrams presented in figure 2.57.

The above diagrams indicate that in GaN bulk, the electron concentration dominates, and that even relatively high excitation cannot compensate the difference, though the hole concentration is relatively much higher for higher excitation. The electron concentration is virtually unchanged. In the structural part, in the AlGaN and GaN layers, the hole concentration dominates, while the electrons are swept away. The concentration difference is relatively weakly affected by the optical excitation. The most complex behavior is observed for In-containing part. It is shown that in InGaN layers, the electron concentration is relatively high, induced by polarization charges at InGaN/AlGaN interfaces. The concentration is unaffected by the optical excitation. Naturally, as expected the holes are swept away to the opposite side of the InGaN layer. These concentrations are much different in

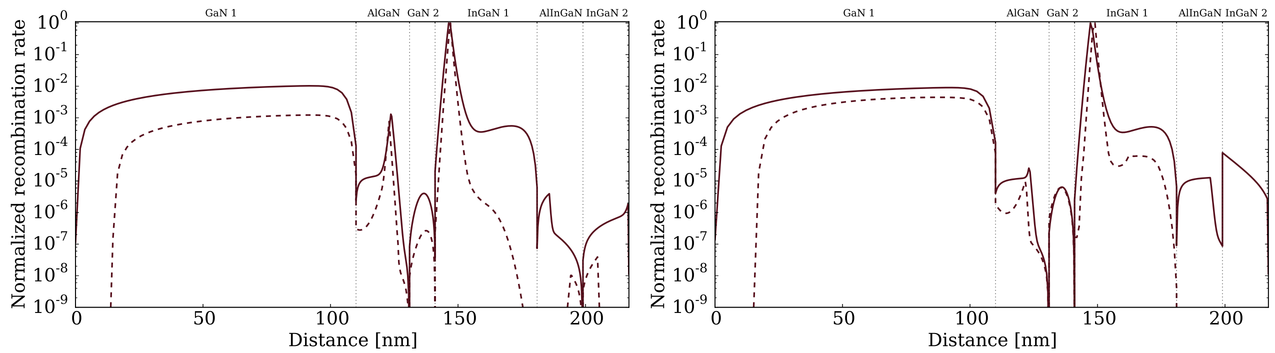

Figure 2.58: Optical recombination intensity normalized to unity, in the low-In (1844 - upper diagram) and high-In (1839 - lower diagram) samples under low (dashed lines) and high (solid lines) optical excitations.

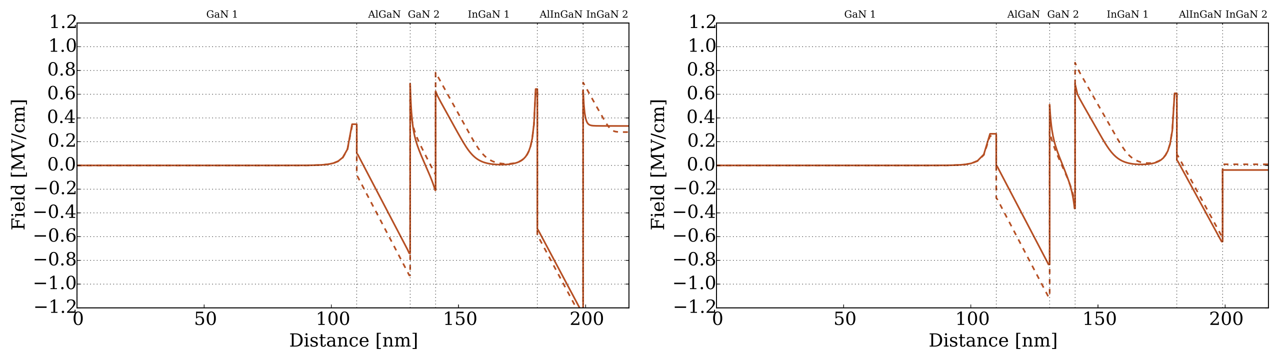

Figure 2.59: Electric field in the low-In (1844 - upper diagram) and high-In (1839 - lower diagram) samples under low (dashed lines) and high (solid lines) optical excitations.

the majority of the sample volume. Thus it is expected that the optical recombination is controlled by the minority carriers, with the majority carriers abundant. The recombination rate was therefore calculated assuming that the minority carriers control the recombination rate. The recombination intensity was calculated assuming classical model which could be affected by the quantum effects that are not directly incorporated into the model used.

From these results it follows that the two regions contribute significantly to the optical emission form the sample. The first is the GaN bulk, where the recombination is lower by two orders of magnitude than the maximal but the extent of the region compensates that leading to considerable emission form that region. The second, the maximum originates from the InGaN layer, close to GaN. The emission region is shifted in function of the optical excitation power. It is unaffected by the indium content.

The other features may be deduced from the distribution of the electric field within the samples, presented in figure 2.59. As it is shown, the electric field is zero in the GaN bulk. Thus, the emission should be intensive and short lived. The emission energy should be close to the bandgap, corrected by the exciton energy. The second peak originates from the InGaN layer. Its energy should be lower due to presence of indium, reducing the bandgap and also due to the electric field. The electric field is considerable and strongly depends on the excitation (see also figure 2.60). Thus the emission is characterized by long times, and also by considerable change with time. For longer time the redshift should be observed due to increase of the electric field. The third emission source is located within the InGaN layer, adjacent to the surface. The simulation result indicate that the main emission is form

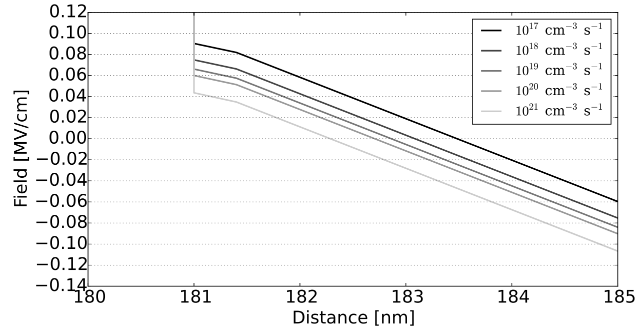

Figure 2.60: Electric field in the high-In (1839) samples near the InGaN1/AlInGaN interface for optical excitations between $1 \times 10^{17}\,\mathrm{cm}^{-3}\mathrm{s}^{-1}$ and $1 \times 10^{21}\,\mathrm{cm}^{-3}\mathrm{s}^{-1}$.

the region close to AlGaInN, but this conclusion may be affected by the presence of surface related field which may shift the carriers close to the surface. The emission is from the considerable field region in case of low In content. In case of high-In content, the emission is from reduced field. Thus the emission should be relatively shorter, but still quite long lived. The emission energy will show moderate redshift for low In content. For the high-In sample the emission energy should not change.

From these data it follows that the light is emitted from GaN and InGaN layers. The quaternary AlGaInN layer will not emit light in any of the studied cases. Thus such layer is not suitable for quantum wells but it may be successfully used for construction of the EBL. These results are in agreement with the results already suggested in [116].

Chapter 3

Linearization and convergence study

Contents

3.1 Linearization method 149
3.1.1 Newton method 150
3.1.2 Newton method with backtracking 152
3.1.3 Comparison 155
3.2 Error analysis: numerical experiments 157
3.2.1 Introduction 157
3.2.2 Formulation u, v, w 158
3.2.3 Formulation ψ, F_n, F_p: one dimension 175
3.2.4 Formulation ψ, F_n, F_p: two dimensions 179
3.3 Discontinuities on interfaces 181
3.3.1 Differential problem 181
3.3.2 Discrete problems 184
3.3.3 Simulations 186

3.1 Linearization method

The van Roosbroeck equations (or drift-diffusion equations, see section 2.5.3), constitute a nonlinear system. Therefore some kind of linearization must be used to numerically find approximate solutions of these equations. The Banach iteration scheme for drift-diffusion equations is proposed in [76]. To obtain linear equations, the algorithm uses the Picard method. Additionally convex combinations are used to improve convergence, i.e. for a given approximation u_i of i-th iteration, let $T(u_i)$ denote approximate solution of a given equation with the Picard method (see (3.1.8) and (3.1.9)). Then u_{i+1} is given by

$$u_{i+1} = (1 - \omega)u_i + \omega T(u_i), \tag{3.1.1}$$

where $\omega \in (0, 1]$. The existence of fixed point is proven, and then it is shown that for sufficiently small ω, the simple iteration scheme will converge under additional assumptions on boundary values.

In [30] similar method were proposed, which involves solving of decoupled system with quasi-Newton method for Poisson equation (2.5.23) and simple iteration scheme for continuity equations (2.5.15). The convergence is proven for small bias and zero recombination (see sections 2.3, 2.6 for physical details). Numerical simulations are presented, which demonstrate convergence of the algorithm. In these simulations, abstract devices are used.

In early simulations, our first choice of a linearization was the Picard method. It is most straight-forward choice, as to use it it is only necessary to implement a solver for linear elliptic equations, which

is also needed for more sophisticated methods. Unfortunately our simulations with the Picard method did not achieved satisfactory results. The method was successful, but to obtain convergence we used ω as in (3.1.1) close to zero. The number of iterations was highly dependent on the length of device. For example, for a short GaN p-n homojunction, 200 nm long, the iteration number was of order 10^3, while for 2 µm p-n homojunction it is of order 10^6. Similar effect is discussed in [55]. We also observed the dependence of iteration number on material parameters, like for example SRH quasiparticle lifetimes, but with no clear pattern. These numbers of iterations are generally prohibitive. While this method may be used to some extent in one dimension with low number of nodes, it is unfeasible otherwise.

In our numerical code, unknown functions and computations are performed in SI units, i.e. the electrostatic potential in volts, quasi-Fermi levels in joules, distance in meters, concentrations in m^{-3}, etc. Naturally, some of the causes of the big iteration number may be mitigated by appropriate scaling of the unknown functions or specific discretization schemes. However, this approach is feasible when the equations are given, i.e. they shall not be altered. In our case, our fundamental assumption was that the drift-diffusion equations should be altered to incorporate additional physical effects, specific to gallium nitride devices, which would be gradually included. Examples of such effects, not initially incorporated in our model, are: trap-assisted tunneling, ionization of traps, Auger recombination, polarization charges, carrier generation due to illumination.

It is obvious that the discretization or linearization cannot account for an arbitrary generalization of the model, but our aim was to make it possibly broad. The assumption here was that whether the equations can change, they should remain elliptic differential equations, generally with the same unknown variables. On the other hand, additional physical phenomena are often represented by modifications of underlying equations. These modifications can be found in the literature, but there is additional effort necessary to make appropriate scaling, if the scaled equations are used. This approach, while simple in theory, is a tedious and error-prone task. There is no clear method of debugging the modified code, especially due to lack of closed-form solutions of the van Roosbroeck equations. Thus to diminish chances of introducing errors, we decided to use SI units for physical variables, and we do not use scaling of variables. We also used quite general discretizations of elliptic differential equations, discussed in detail in section 1.3.

3.1.1 Newton method

A natural choice for solution of a nonlinear system of algebraic equations is the Newton method [33, 59]. This method is originally intended to finding roots of the nonlinear functions. Let $G : \mathbb{R}^J \to \mathbb{R}^J$. Vector $\xi^* \in \mathbb{R}^J$, such that $G(\xi^*) = 0$ may be approximated by Newton method using the following iterative process:

$$\begin{aligned} &\xi_0 \text{ — a given initial approximation of } \xi^*, \\ &\xi_i = \xi_{i-1} - [\mathrm{D}G(\xi_{i-1})]^{-1} G(\xi_{i-1}). \end{aligned} \tag{3.1.2}$$

The main advantage of this method is fast, quadratic local convergence. On the other hand, good initial approximation must be provided, otherwise the method does not have to converge at all.

To use the Newton iteration to our problem, we do the following. Let $\psi, F_n, F_p \in X_h(\Omega)$ be some approximations of the potential and quasi-Fermi levels respectively. Let us denote

$$\psi = [\psi_1, \ldots, \psi_J], \quad F_n = [F_{n,1}, \ldots, F_{n,J}], \quad F_p = [F_{p,1}, \ldots, F_{p,J}], \tag{3.1.3}$$

where $\psi_j, F_{n,j}, F_{p,j} \in \mathbb{R}$ are the coefficients of $\psi, F_n, F_p \in X_h(\Omega)$ in a basis of the discrete space $X_h(\Omega)$ and J is a dimension of this space. We define ξ as

$$\xi = [\psi, F_n, F_p]. \tag{3.1.4}$$

p-GaN
$N_d = 0$
$N_a = 5 \times 10^{19}$ cm^{-3}
d = 300 nm

n-GaN
$N_d = 5 \times 10^{18}$ cm^{-3}
$N_a = 0$
d = 300 nm

n-GaN
$N_d = 5 \times 10^{18}$ cm^{-3}
$N_a = 0$
d = 499 nm

QW - $In_{0.1}Ga_{0.9}N$
$N_d = 5 \times 10^{16}$ cm^{-3}
$N_a = 0$
d = 3 nm

B - $In_{0.015}Ga_{0.9}N$
$N_d = 5 \times 10^{18}$ cm^{-3}
$N_a = 0$
d = 5 nm

QW - $In_{0.1}Ga_{0.9}N$
$N_d = 5 \times 10^{16}$ cm^{-3}
$N_a = 0$
d = 3 nm

p-GaN
$N_d = 0$
$N_a = 5 \times 10^{19}$ cm^{-3}
d = 498 nm

Figure 3.1: Devices used in testing linearization algorithms: a p-n homojunction and a two quantum well heterostructure.

We would like to use the Newton method to find an approximate solution of the discretized problem (1.2.1). Therefore let $a_\psi, f_\psi, a_n, f_n, a_p, f_p$ denote discrete problem operators (Composite Discontinuous Galerkin Method operators as in section 1.3.2) for left hand sides and right hand sides of the equations (1.2.1). Then let us define residual functions

$$\begin{aligned} G_{\psi,j}(\psi, F_n, F_p) &:= a_\psi(\psi, F_n, F_p, \varphi_{(j)}) - f_\psi(\psi, F_n, F_p, \varphi_{(j)}), \\ G_{n,j}(\psi, F_n, F_p) &:= a_n(\psi, F_n, \varphi_{(j)}) - f_n(\psi, F_n, F_p, \varphi_{(j)}), \\ G_{p,j}(\psi, F_n, F_p) &:= a_p(\psi, F_p, \varphi_{(j)}) - f_p(\psi, F_n, F_p, \varphi_{(j)}), \end{aligned} \tag{3.1.5}$$

where $\{\varphi_{(j)}\}_{j=1}^J$ is the base of $X_h(\Omega)$. Note that operators $f_\psi, a_n, f_n, a_p, f_p$ are nonlinear in ψ, F_n, F_p, and linear in $\varphi_{(j)}$.

Then we define coupled residual function G as:

$$G(\xi) := [G_\psi(\xi), G_n(\xi), G_p(\xi)], \tag{3.1.6}$$

where

$$\begin{aligned} G_\psi(\xi) &:= [G_{\psi,1}(\xi), \dots, G_{\psi,J}(\xi)], \\ G_n(\xi) &:= [G_{n,1}(\xi), \dots, G_{n,J}(\xi)], \\ G_p(\xi) &:= [G_{p,1}(\xi), \dots, G_{p,J}(\xi)]. \end{aligned} \tag{3.1.7}$$

If $G(\xi)$ is zero, then ξ is a discrete solution. We may then pick some initial approximation ξ_0 and use the Newton method to find the approximate solution.

The Newton method is very sensitive to the initial approximation. Unfortunately good initial approximations for the drift-diffusion model are available only for devices in the equilibrium state (see section 2.5.4). Thus the idea is to start a simulation from the equilibrium state, and then gradually increase the bias to the given value, which corresponds to change of the boundary conditions. The sketch of the algorithm is as follows:

$\xi_0 :=$ initial_approximation();
$i := 1$;

for bias:=0 **to** bias_max **step** bias_step **do**
  while $\|G(\xi_{i-1})\|$ is not small **do**
    $s_i := -[\mathrm{D}G(\xi_{i-1})]^{-1}G(\xi_{i-1});$
    $\xi_i := \xi_{i-1} + s_i;$
    $i := i + 1;$
  end while
end for

In the above schema the *bias_max* denotes the target voltage of the device. In this method we clearly have two iterations. The inner iteration is based on the Newton method. The outer iteration advances the bias. For example, to perform a simulation of device's operation under 3 V bias, we may take 0.1 V *bias_step* and simulate series of biases: 0 V, 0.1 V, 0.2 V, ..., 3 V. We assume that for no bias there is a good initial approximation, and then the successive solutions are initial approximations for steps to follow. Thus the outer loop is a sort of homotopy method.

As mentioned in section 2.5.6, the bias is introduced to van Roosbroeck equations by boundary conditions. It is not explicitly denoted above, but change of the bias modifies the operator G. From this perspective we see that the inner loop is the Newton method for the operator G and a given bias, while the outer loop changes slightly the operator G by advancing the bias, i.e. setting appropriate boundary conditions.

While this outer loop may seem as an unnecessary additional effort, in fact it is often beneficial. For example, to simulate a current-voltage characteristic or a light-current characteristic, it is necessary to have simulations for a range of biases.

Unfortunately this straightforward algorithm does not perform well for the drift-diffusion simulations of GaN-based devices. While the iteration number (per inner loop) is mostly invariant in a device length or material parameters, the step change in outer iteration must be small to prevent divergence. The divergence is often a direct consequence of overflows or underflows, which emerge easily due to the exponential character of coefficients of the continuity equations. Also taking very small *bias_step* increases the simulation time.

3.1.2 Newton method with backtracking

The convergence by may be improved by use the backtracking method for the Newton iteration [33]. The idea is to scale the Newton method step by a coefficient $0 < \lambda \leq 1$ in every iteration to ensure decrease of the norm $\|G(\xi)\|$ for some norm $\|\cdot\|$. It can be shown [33] that if the Jacobian is nonsingular, then it is possible to find λ small enough to reduce the norm $\|G(\xi)\|$. When the approximation ξ is sufficiently close to the solution, λ equal to one may be taken and from then the convergence is as good as for the standard Newton method.

The algorithm with very simple strategy of choosing λ may be written as follows (we omit the outer iteration, as it does not change):

while $\|G(\xi_{i-1})\|$ is not small **do**
  $s_i := -[\mathrm{D}G(\xi_{i-1})]^{-1}G(\xi_{i-1});$
  $\lambda_i := 1;$
  $\xi_i := \xi_{i-1} + s_i;$
  while $\|G(\xi_i)\| > \|G(\xi_{i-1})\|$ **do**
    $\lambda_i := \lambda_i/2;$
    $\xi_i := \xi_{i-1} + \lambda_i \cdot s_i;$
  end while
  $i := i + 1;$
end while

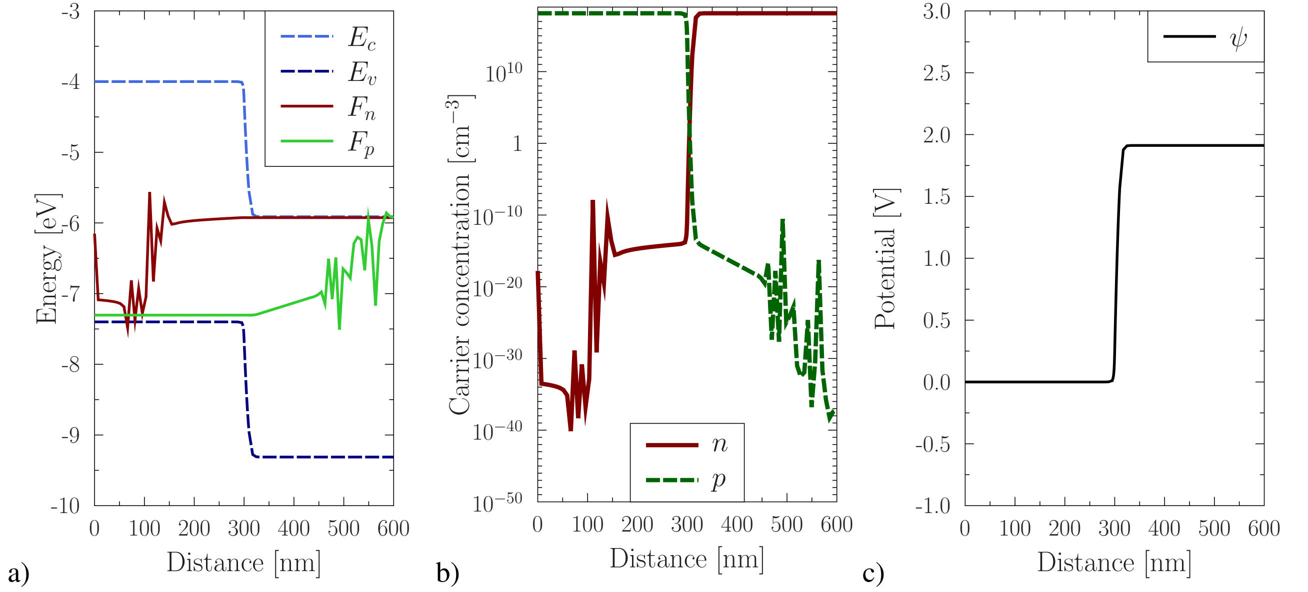

Figure 3.2: An example of a simulation of the 600 nm GaN p-n diode simulated using the Newton method with linear backtracking. The bias is 1.38 V.

There are naturally more subtle backtracking methods. For example, they look for a minimum of function $g(\lambda) := \|G(\xi_{i-1} + \lambda s_i)\|$ approximating it by a polynomial function. In general, smaller λ_i coefficients results in slower convergence.

The following example demonstrates that this modification still needs improvement. We present a simulation of a p-n GaN diode, which is a fairly simple device (figure 3.1). The result of 11th outer step of the simulation is shown in figure 3.2. Note the nonphysical fluctuations of the functions F_n, F_p. However, the residuum size $\|G(\xi_i)\|$ does not indicate any problems. For the initial approximation (from the previous bias-step), the residuum was 1.8×10^{11}, then four steps of the Newton method were performed with $\lambda = 1$, which reduced the residuum as follows: 3.4×10^6, 6.5×10^3, 2.7×10^{-2}, 6.6×10^{-4}. Thus the magnitude of the residuum was reduced by 14 orders of magnitude, which is close to the machine precision.

Unfortunately this iteration diverged few outer steps later due to underflow. However, this is not a general behavior, sometimes such fluctuations vanish for a large bias and the iteration does not diverge.

Still the question remains why such an nonphysical behavior may be present when the residuum is so small. The reason is that n and p are the coefficients of the continuity equations, which formally correspond to F_n and F_p, respectively. On the left part of this device, where the fluctuations of F_n emerged, the coefficient n is very small, more than a 20 orders of magnitude smaller than on the other side of the device. Therefore this error is completely neglected by this algorithm due to the precision of the floating point arithmetic. Similar effect is observed for F_p and p.

This example is not a isolated case. We observed such an effect for many devices, varying from simple p-n homojunctions to laser heterostructures. This behaviour is repetitive, as our code is deterministic. Changing the discretization, number of steps, or device's parameters may affect this problem, either introducing or ceasing it.

This phenomenon indicates also an additional problem — due to the nature of floating-point arithmetic, the residual norm alone is inadequate to evaluate whether a given approximation is close to a discrete solution in practice. We will return to this problem in next section.

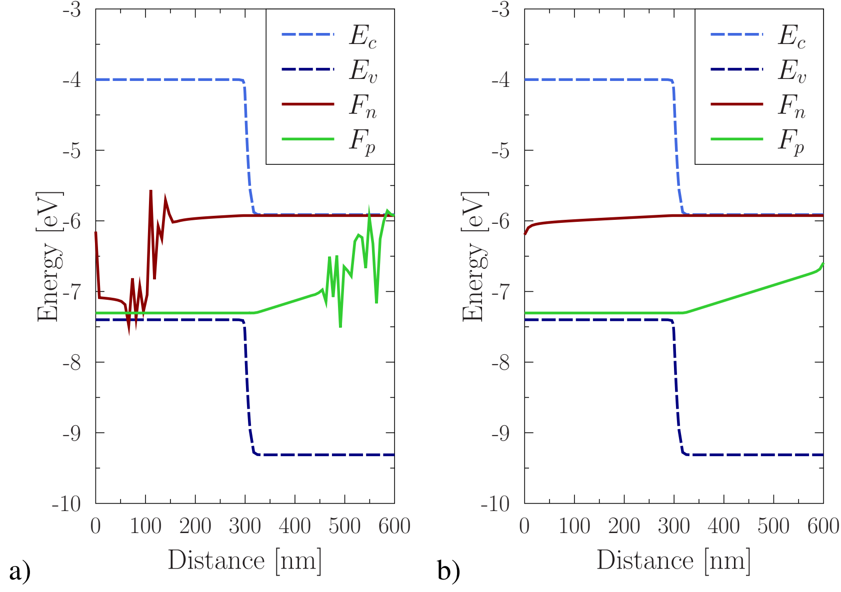

Figure 3.3: Comparison of simulation results of the 600 nm GaN p-n diode simulated using: a) the Newton method with linear backtracking, b) the Newton method with our modification. The bias is 1.38 V.

3.1.2.1 Modification of the Newton method

As we demonstrated, the weakness of the backtracking algorithm presented is the lack of good estimates of the quality of discrete approximations, as the residuum alone is insufficient. In this section we would like to show how to get such an estimate. Inspired by [56], we would like to rewrite the problem in the Banach iteration manner.

Let $(\psi_0, F_{n,0}, F_{p,0})$ be some initial approximation. Let us define function T as

$$\xi_i = T(\xi_{i-1}), \tag{3.1.8}$$

where $\xi_i = (\psi_i, F_{n,i}, F_{p,i})$ is a solution of the discrete version of a following system of differential equations

$$\begin{aligned} \nabla \cdot \Big(\varepsilon_0 \varepsilon \nabla \psi_i\Big) &= -qC(\psi_{i-1}, F_{n,i-1}, F_{p,i-1}), \\ \nabla \cdot \big(\mu_n n(\psi_{i-1}, F_{n,i-1}) \nabla F_{n,i}\big) &= qR(\psi_{i-1}, F_{n,i-1}, F_{p,i-1}), \\ \nabla \cdot \big(\mu_p p(\psi_{i-1}, F_{p,i-1}) \nabla F_{p,i}\big) &= -qR(\psi_{i-1}, F_{n,i-1}, F_{p,i-1}), \end{aligned} \tag{3.1.9}$$

where $\xi_{i-1} = (\psi_{i-1}, F_{n,i-1}, F_{p,i-1})$. If $\xi_i = \xi_{i-1}$, then ξ_i is a solution of the discrete problem. Note that (3.1.9) is a system of three independent linear differential equations, so $T(\xi_{i-1})$ may be computed easily.

We do not aim at finding a solution by Banach iteration for T, as generally it is not a contraction. We would like to use T for estimate of the quality of solutions in a following manner. Let us define H as

$$H(\xi) := T(\xi) - \xi. \tag{3.1.10}$$

Assume that ξ_{i-1} is close to the solution. Then $\xi_i = T(\xi_{i-1}) \approx \xi_{i-1}$. Unlike n and p, functions ψ, F_n and F_p do not express exponential behavior, and they are of similar order of magnitude in appropriate choice of units (ψ in volts; F_n, F_p in electronvolts). Therefore elements of the vector $H(\xi_{i-1})$ do not

vary by orders of magnitude and $\|H(\xi_{i-1})\|$ may be used as an estimate of an approximation ξ_{i-1}, because it has no drawbacks of $\|G(\xi_{i-1})\|$.

Therefore we propose the following modification of the inner loop:

while $\|H(\xi_{i-1})\|$ is not small **do**
  $s_i := -[DG(\xi_{i-1})]^{-1}G(\xi_{i-1})$;
  $\lambda_i := 1$;
  $\xi_i := \xi_{i-1} + s_i$;
  while $\|H(\xi_i)\| > (1+c)\|H(\xi_{i-1})\|$ **do**
    $\lambda_i := \lambda_i/2$;
    $\xi_i := \xi_{i-1} + \lambda_i \cdot s_i$;
  end while
  $i := i + 1$;
end while

Generally for $c = 0$ this modification tends to minimize $\|H(\xi_i)\|$. Our observations show that it is often favorable to allow limited growth of $\|H(\xi_i)\|$ by setting $c > 0$. In contrast to the backtracking method, here we have no guarantee that $\|H(\xi_{i-1})\| > \|H(\xi_i)\|$ for any choice of the parameter λ.

To illustrate usefulness of the function H, we revisit our example from the previous section (figure 3.2). The nonphysical solution had a residuum norm 6.6×10^{-4}. However, in this case $\|H(\xi)\| = 1.1 \times 10^{14}$. Therefore without question $\xi \not\approx T(\xi)$, which is the information we expect to get.

We have therefore repeated this simulation, using the algorithm proposed in this section. The result is presented on the figure 3.3b). In this case there are no fluctuations of F_n, F_p, $\|G(\xi)\| \approx 3.2 \times 10^{-5}$ and $\|H(\xi)\| \approx 4.6 \times 10^{-6}$. Therefore the residuum is similar as for the nonphysical case, but the latter value is much lower, which corresponds to better quality of this approximation.

3.1.3 Comparison

As we pointed out in Section 3.1.1, the backtracking strategy proposed in this paper may prevent divergence and lead to the approximations, which are physically more favorable. Still we would like to show that it is also more efficient than the standard Newton method in terms of iteration number and computational time.

Therefore we compare simulation results for a two quantum well heterostructure presented on figure 3.1. We take into account the classic Newton method, backtracking linesearch [33], and our backtracking strategy (section 3.1.2.1). Simulations account for radiative recombination, Shockley-Read-Hall recombination with trap-assisted tunneling [51], ionization of impurities and polarization charges.

A goal of these simulations were to find an approximate solution of the drift-diffusion equations for 4 V bias. We have to point out that it is not feasible to compute the solution for nonzero bias with the Newton method alone, or using the inner loop of the presented algorithms, as the initial approximations are only available for so-called steady state of a device, when bias is zero. So to perform our simulations we set `bias_max` to 4 V. Every consecutive solution is used as initial approximation for next inner loop. It is generally not a waste, as normally they are also used to compute a IV characteristic on [0, `bias_max`], which are used to compare results with physical experiments in real simulations. To obtain a fine IV characteristic, it is enough to have 10–20 steps, as it generally should not fluctuate much.

Since the Newton method is sensitive to an initial approximation, generally more steps should improve the convergence of the considered methods (number of steps = 1 + `bias_max/bias_step`). However, too much steps would increase the total iteration number and it is not very beneficial to the IV characteristic.

Results of these simulations are presented in table 3.1. For each method we performed few simulations with an outer iteration number varying from 21 to 331. Every method considered in this study diverged if the number of steps was below 21.

In this setting, the most efficient was the Newton method with our backtracking strategy. The simulation took 284 seconds and 174 iterations in 21 steps. Next one was the linear backtracking on $\|G(\xi)\|$, with a stop condition on $\|H(\xi)\|$, which took 482 seconds and 377 iterations in 101 steps. For lower number of steps, the latter method generally did not return satisfactory results. In comparison, results of standard linear backtracking were acceptable for 161 steps (714 iterations, 898 s). Note that the classic Newton method with no backtracking was more efficient, so the linear backtracking did not help.

Generally setting the number of outer steps to a number high enough leads to convergence of every tested method. Then the number of iterations become similar, as methods need not backtrack due to good initial approximations. Our simulations also reveal that imposing a stop condition on $\|H(\xi)\|$ alone leads to slightly better efficiency.

We must mention also about another possibility, which is the application of the Newton method directly to a problem $H(\xi) = 0$. However in this case Jacobian $\mathrm{D}H$ is dense [61], and the method is inefficient.

Table 3.1: Comparison of efficiency of our modification with linear backtracking and the classic Newton method, with stop conditions imposed on $\|G(\xi)\|$ or $\|H(\xi)\|$. In this table we present an outer iterations number (steps), total iteration number of the Newton method, computation time and average number of iterations per one step. Simulations were performed on a standard desktop PC. In every simulation, the Newton method is used for function G

Our backtracking strategy allows slight increase of $\|H(\xi_i)\|$

Steps	Iter.	Iter./steps	Time (s)	Result
21	174	8.29	284	Good
41	227	5.54	360	Good
91	378	4.15	610	Good
101	409	4.05	666	Good
161	566	3.52	919	Good
331	908	2.74	1516	Good

Linear backtracking stop condition on $\|G(\xi_i)\|$

Steps	Iter.	Iter./steps	Time (s)	Result
21	—	—	—	Diverged
41	317	7.73	377	Nonphysical
91	544	5.98	670	Nonphysical
101	465	4.60	581	Nonphysical
161	714	4.43	898	Good
331	870	2.63	1189	Good

Linear backtracking stop condition on $\|H(\xi_i)\|$

Steps	Iter.	Iter./steps	Time (s)	Result
21	—	—	—	Diverged
41	234	5.71	288	Nonphysical
91	354	3.89	452	Nonphysical
101	377	3.73	482	Good
161	519	3.22	676	Good
331	870	2.63	1190	Good

Classic Newton method stop condition on $\|H(\xi_i)\|$

Steps	Iter.	Iter./steps	Time (s)	Result
21	—	—	—	Diverged
41	—	—	—	Diverged
91	—	—	—	Diverged
101	378	3.74	484	Nonphysical
161	510	3.17	665	Good
331	870	2.63	1189	Good

Classic Newton method stop condition on $\|G(\xi_i)\|$

Steps	Iter.	Iter./steps	Time (s)	Result
21	—	—	—	Diverged
41	—	—	—	Diverged
91	—	—	—	Diverged
101	398	3.94	496	Nonphysical
161	612	3.80	762	Good
331	1197	3.62	1497	Good

3.2 Error analysis: numerical experiments

3.2.1 Introduction

In sections 1.6 and 1.7 we derived error estimates for the CWOPSIP discretization and CSIPG discretization. These estimates are shown for the equilibrium state in formulation presented in section 1.2.2, for one- and two-dimensional domains.

In this section, we would like to verify whether these theoretical estimates hold in simulations. We start with the model problem as in section 1.2.2. Our simulations, however, are extended also to non-equilibrium state, presented in section 1.2.1. These simulations are referred to as formulation u, v, w. Parameters and device schemata used in these simulations are artificial, used for the purpose of error analysis.

Then we pass to the physical-oriented formulation presented in section 2.5. In this case, we

must provide simulation parameters, i.e. material parameters, temperature, physical constants, in appropriate units. Here we present simulations of realistic semiconductor devices based on gallium nitride. In general, our numerical code carries out calculations in SI units, but input and output is provided in more natural units for semiconductor simulations, for example electronvolts instead of joules, centimeters instead of meters, etc. We refer to this series of simulations as formulation ψ, F_n, F_p.

In either case, we do not have closed-form solutions of the drift-diffusion equations except for the trivial cases. Thus as a reference solution, we take a discrete approximation computed for a fine discretization. In general, we use the following scheme. We take some parameter K, for example $K \in \{1, 2, 4, 8, \dots\}$, and we choose a discretization such that $h_i(K) := c_i K^{-1}$, $h(K) := \max\{h_i(K)\}_{i=1}^{N}$. Then we perform simulations for some range of parameters K.

For example, if simulations were performed for $K \in \{1, 2, 4, 8, 16, 32\}$, then $K = 32$ is treated as a reference solution, and $L_2(\Omega)$- and $H^1(\mathcal{E})$-errors of u_h are defined as

$$\text{error}_{K,L_2(\Omega)} := \|u_K - u_{32}\|_{L_2(\Omega)}, \quad \text{error}_{K,H^1(\mathcal{E})} := \|u_K - u_{32}\|_{H^1(\mathcal{E})}, \tag{3.2.1}$$

where $u_K := u_{h(K)}$ for the grid parameter $K \in \{1, 2, 4, 8, 16\}$. For other functions (v_h, w_h, ψ_h, ...) errors are defined analogously. However, in ψ, F_n, F_p formulation, due to difference in magnitudes of discussed functions it is more favorable to use relative errors defined as

$$\text{error}_{K,L_2(\Omega)} := \frac{\|\psi_K - \psi_{32}\|_{L_2(\Omega)}}{\|\psi_{32}\|_{L_2(\Omega)}}, \quad \text{error}_{K,H^1(\mathcal{E})} := \frac{\|\psi_K - \psi_{32}\|_{H^1(\mathcal{E})}}{\|\psi_{32}\|_{H^1(\mathcal{E})}}. \tag{3.2.2}$$

In either case, it is convenient to analyze rate of convergence, which we define as

$$\text{conv_rate}_{2K,L_2(\Omega)} := \frac{\text{error}_{K,L_2(\Omega)}}{\text{error}_{2K,L_2(\Omega)}}, \quad \text{conv_rate}_{2K,H^1(\mathcal{E})} := \frac{\text{error}_{K,H^1(\mathcal{E})}}{\text{error}_{2K,H^1(\mathcal{E})}}. \tag{3.2.3}$$

3.2.2 Formulation u, v, w

We would like to check whether the error estimate derived in sections 1.6 and 1.7 can be achieved in numerical simulations. Therefore we present some examples. These examples are not directly related to any specific semiconductor material. We present simulations of abstract devices mimicking semiconductor p-n diodes. Our first example is a p-n diode consisting of two layers Ω_1, Ω_2 (see figure 3.4), corresponding to n-type layer and p-type layer of a p-n homojunction. It has two contacts with metal electrodes, left and right, denoted by $\partial\Omega_{D,1}$ and $\partial\Omega_{D,2}$. Horizontal boundaries correspond to the contact with insulator (e.g. air). Parameters of the device are presented in table 3.2. For $K = 1$, we divide both layers into two pieces in horizontal direction, while in vertical direction Ω_1 is divided into two pieces, while Ω_2 is divided into four pieces.

In these simulations we assume that the operator Q of problem 1.2.1 is some given piecewise-constant function:

$$Q(x, u, v, w) := C_{\text{rad}}(x). \tag{3.2.4}$$

This form corresponds to the radiative recombination (see section 2.6.1.1).

We start with the equilibrium state. Then the boundary conditions are as follows: $\hat{u}|_{\partial\Omega_{D,1}} = 0$ and $\hat{u}|_{\partial\Omega_{D,2}} = u_b$, where u_b is a built-in potential (see section 2.5.5). In context of the u, v, w-formulation, functions n, p, ρ are defined as

$$n(x) := e^{u(x)-v(x)}, \quad p(x) := e^{w(x)-u(x)}, \quad \rho(x) := k_1(x) - n(x) + p(x). \tag{3.2.5}$$

Functions v, w are constant, such that $\rho|_{\partial\Omega_{D,1}} = 0$.

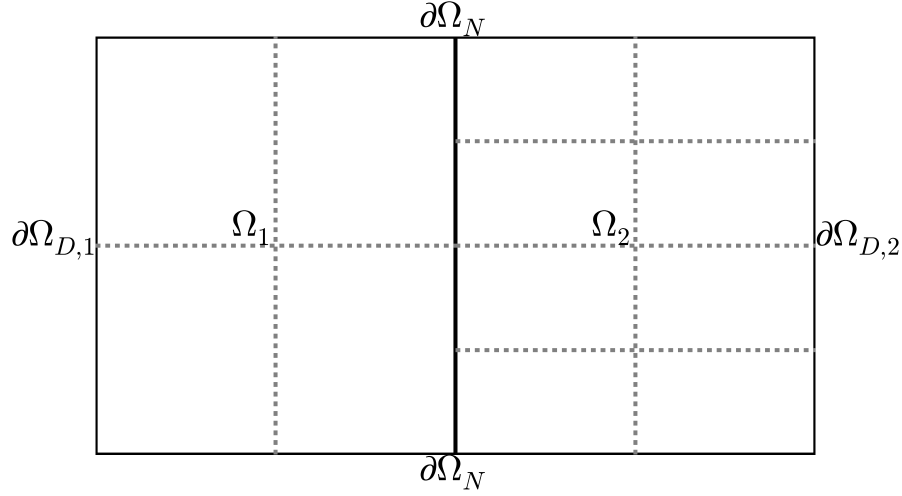

Figure 3.4: Schema of the first device used in simulations. It has two layers, corresponding to n-type layer and p-type layer. Grid for $K = 1$ is presented.

Table 3.2: Parameters of the first device used in simulations. N_x and N_y denote number of nodes in horizontal and vertical direction, depending on parameter K.

Param.	Ω_1	Ω_2
Length	1×10^{-2}	1×10^{-2}
Width	1×10^{-2}	1×10^{-2}
N_x	$2K + 1$	$2K + 1$
N_y	$2K + 1$	$4K + 1$
ε	3×10^{-3}	1×10^{-3}
μ_n	1×10^{3}	3×10^{3}
μ_p	1×10^{2}	3×10^{2}
k_1	3×10^{2}	-3×10^{2}
C_{rad}	1×10^{-3}	2×10^{-3}

Table 3.3: $L_2(\Omega)$ and $H^1(\mathcal{E})$-error of u_h in function of grid density parameter K for the first device in equilibrium state. Numbers in brackets denote the rate of convergence. Solution for CSIPG method with $K = 32$ is taken as a reference function.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
1	4.4×10^{-3}		3.3		4.4×10^{-3}		3.3	
2	1.1×10^{-3}	*(4.0)*	1.6	*(2.0)*	1.1×10^{-3}	*(4.0)*	1.6	*(2.0)*
4	2.7×10^{-4}	*(4.0)*	8.2×10^{-1}	*(2.0)*	2.7×10^{-4}	*(4.0)*	8.2×10^{-1}	*(2.0)*
8	6.5×10^{-5}	*(4.1)*	4.0×10^{-1}	*(2.0)*	6.5×10^{-5}	*(4.2)*	4.0×10^{-1}	*(2.0)*
16	1.4×10^{-5}	*(4.7)*	1.8×10^{-1}	*(2.2)*	1.4×10^{-5}	*(4.7)*	1.8×10^{-1}	*(2.2)*

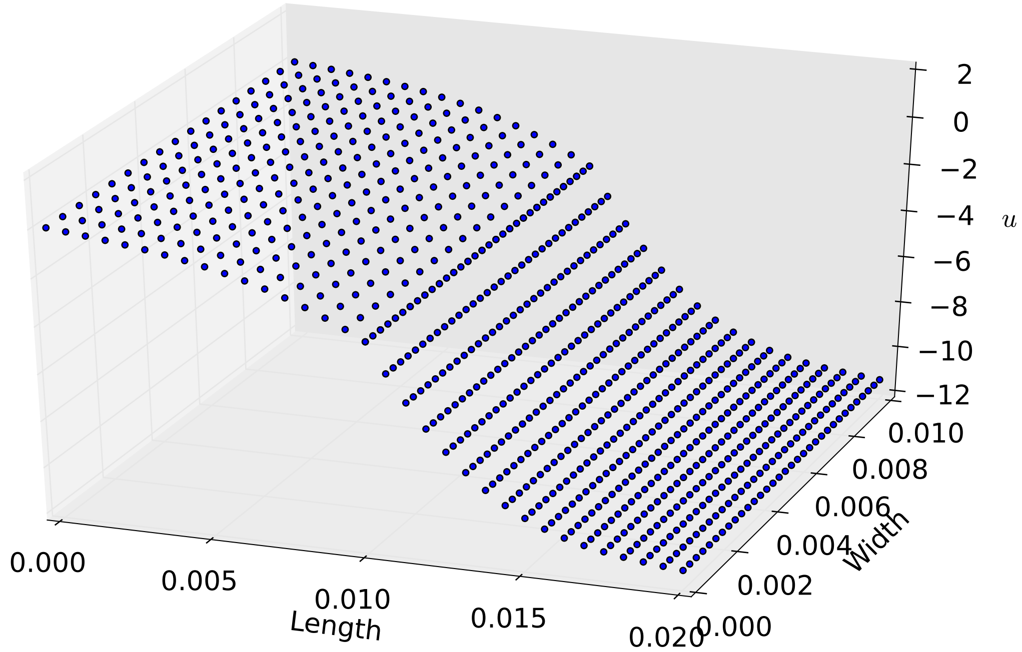

Figure 3.5: Function u_h for the first example in the equilibrium state for $K = 8$. Note one-dimensional character of the solution.

Results of these simulations are presented in table 3.3. We observe a linear reduction of the H^1-error, which is consistent with our theoretical result, as the H^1-norm is bounded by the broken norm up to a constant factor. We also note the quadratic L_2-norm convergence rate. This behavior for CSIPG method and CWOPSIP method is consistent. The error values are generally very close for both methods.

The theory presented in sections 1.6 and 1.7 covers only the equilibrium state. Nevertheless we performed the simulations also for the non-equilibrium state, to verify whether the discussed discretizations are feasible for simulations of semiconductor devices in general.

Boundary conditions on the function u are similar as before, i.e. $\hat{u}|_{\partial\Omega_{D,1}} = 0$ and $\hat{u}|_{\partial\Omega_{D,2}} = u_b + u_{\text{bias}}$, where u_{bias} is a nonzero difference potential between the electrodes. On functions v, w we impose two implicit conditions on $\partial\Omega_D$: $v|_{\partial\Omega_D} = w|_{\partial\Omega_D}$ and $\rho|_{\partial\Omega_D} = 0$, cf. (3.2.5). On Ω_N we impose homogeneous Neumann boundary condition.

Results of this simulation are presented in tables 3.4, 3.5, 3.6. Let us start from the bias of magnitude 6. For the approximation of function u, results are similar to the equilibrium state. For the functions v, w, the convergence is much worse. We may roughly estimate that the L_2-error reduces linearly, while the H^1-error convergence rate is sublinear, hard to estimate precisely without the exact solution. In the comparison, we included also the functions n, p. The van Roosbroeck equations may be formulated in terms of functions u, v, w, but from the physical point of view there are other logical choices possible [92]. Another choice is u, n, p, as the charge ρ and many recombination models (radiative, Shockley-Read-Hall, Auger) can be easily expressed in terms of these functions (see sections 2.5 and 2.6).

We observe that the error convergence for n, p is faster than for v, w, and it is similar as for the function u, although it starts slower for n. Thus simulation of many physical parameters may rely on better precision of functions n, p despite the slow convergence of functions v, w.

Table 3.4: $L_2(\Omega)$ and $H^1(\mathcal{E})$-error of u_h, v_h, w_h and dependent functions n, p in function of grid density parameter K for the first device under bias of magnitude 6. Numbers in brackets denote the rate of convergence. Solution for CSIPG method with $K = 32$ is taken as a reference function.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
Function: u								
1	6.2×10^{-2}		4.3×10^{-1}		6.2×10^{-2}		4.3×10^{-1}	
2	1.6×10^{-2}	*(3.9)*	2.2×10^{-1}	*(2.0)*	1.6×10^{-2}	*(3.9)*	2.2×10^{-1}	*(2.0)*
4	4.0×10^{-3}	*(4.0)*	1.1×10^{-1}	*(2.0)*	4.0×10^{-3}	*(4.0)*	1.1×10^{-1}	*(2.0)*
8	9.6×10^{-4}	*(4.1)*	5.4×10^{-2}	*(2.0)*	9.7×10^{-4}	*(4.2)*	5.4×10^{-2}	*(2.0)*
16	2.1×10^{-4}	*(4.7)*	2.4×10^{-2}	*(2.2)*	2.1×10^{-4}	*(4.7)*	2.4×10^{-2}	*(2.2)*
Function: v								
1	1.9×10^{-1}		9.3×10^{-1}		1.9×10^{-1}		9.3×10^{-1}	
2	1.1×10^{-1}	*(1.8)*	8.8×10^{-1}	*(1.1)*	1.1×10^{-1}	*(1.8)*	8.8×10^{-1}	*(1.1)*
4	6.0×10^{-2}	*(1.8)*	8.1×10^{-1}	*(1.1)*	6.0×10^{-2}	*(1.8)*	8.1×10^{-1}	*(1.1)*
8	3.1×10^{-2}	*(2.0)*	7.0×10^{-1}	*(1.2)*	3.1×10^{-2}	*(2.0)*	7.0×10^{-1}	*(1.2)*
16	1.2×10^{-2}	*(2.5)*	5.2×10^{-1}	*(1.3)*	1.2×10^{-2}	*(2.5)*	5.2×10^{-1}	*(1.3)*
Function: w								
1	1.4		9.3×10^{-1}		1.4		9.3×10^{-1}	
2	8.1×10^{-1}	*(1.8)*	8.8×10^{-1}	*(1.1)*	8.1×10^{-1}	*(1.8)*	8.8×10^{-1}	*(1.1)*
4	4.5×10^{-1}	*(1.8)*	8.1×10^{-1}	*(1.1)*	4.5×10^{-1}	*(1.8)*	8.1×10^{-1}	*(1.1)*
8	2.3×10^{-1}	*(2.0)*	7.0×10^{-1}	*(1.2)*	2.3×10^{-1}	*(2.0)*	7.0×10^{-1}	*(1.2)*
16	9.0×10^{-2}	*(2.5)*	5.2×10^{-1}	*(1.3)*	9.0×10^{-2}	*(2.5)*	5.2×10^{-1}	*(1.3)*
Function: n								
1	6.6×10^{-2}		3.8×10^{-1}		6.6×10^{-2}		3.9×10^{-1}	
2	3.0×10^{-2}	*(2.2)*	3.0×10^{-1}	*(1.3)*	3.0×10^{-2}	*(2.2)*	3.0×10^{-1}	*(1.3)*
4	1.1×10^{-2}	*(2.8)*	1.9×10^{-1}	*(1.5)*	1.1×10^{-2}	*(2.8)*	1.9×10^{-1}	*(1.5)*
8	3.0×10^{-3}	*(3.6)*	1.1×10^{-1}	*(1.8)*	3.0×10^{-3}	*(3.6)*	1.1×10^{-1}	*(1.8)*
16	6.5×10^{-4}	*(4.6)*	5.0×10^{-2}	*(2.2)*	6.5×10^{-4}	*(4.6)*	5.0×10^{-2}	*(2.2)*
Function: p								
1	1.9×10^{-1}		5.5×10^{-1}		1.9×10^{-1}		5.5×10^{-1}	
2	3.7×10^{-2}	*(5.3)*	2.3×10^{-1}	*(2.4)*	3.7×10^{-2}	*(5.3)*	2.3×10^{-1}	*(2.4)*
4	8.9×10^{-3}	*(4.1)*	1.1×10^{-1}	*(2.1)*	8.8×10^{-3}	*(4.2)*	1.1×10^{-1}	*(2.1)*
8	2.1×10^{-3}	*(4.1)*	5.3×10^{-2}	*(2.1)*	2.1×10^{-3}	*(4.1)*	5.3×10^{-2}	*(2.1)*
16	4.5×10^{-4}	*(4.8)*	2.4×10^{-2}	*(2.2)*	4.4×10^{-4}	*(4.8)*	2.4×10^{-2}	*(2.2)*

Table 3.5: $L_2(\Omega)$ and $H^1(\mathcal{E})$-error of u_h, v_h, w_h and dependent functions n, p in function of grid density parameter K for the first device under bias of magnitude 10. Numbers in brackets denote the rate of convergence. Solution for CSIPG method with $K = 32$ is taken as a reference function.

	CSIPG		CWOPSIP	
K	$L_2(\Omega)$	$H^1(\Omega)$	$L_2(\Omega)$	$H^1(\Omega)$
		Function: u		
1	1.4×10^{-1}	6.3×10^{-1}	1.2×10^{-1}	6.3×10^{-1}
2	5.4×10^{-2} *(2.7)*	3.6×10^{-1} *(1.7)*	4.2×10^{-2} *(3.0)*	3.6×10^{-1} *(1.7)*
4	2.3×10^{-2} *(2.3)*	1.9×10^{-1} *(1.9)*	1.8×10^{-2} *(2.3)*	1.9×10^{-1} *(1.9)*
8	9.1×10^{-3} *(2.5)*	9.3×10^{-2} *(2.0)*	7.7×10^{-3} *(2.4)*	9.3×10^{-2} *(2.0)*
16	2.7×10^{-3} *(3.4)*	4.1×10^{-2} *(2.2)*	2.3×10^{-3} *(3.4)*	4.1×10^{-2} *(2.2)*
		Function: v		
1	3.8×10^{-1}	9.7×10^{-1}	3.8×10^{-1}	9.7×10^{-1}
2	2.4×10^{-1} *(1.6)*	9.4×10^{-1} *(1.0)*	2.4×10^{-1} *(1.6)*	9.4×10^{-1} *(1.0)*
4	1.5×10^{-1} *(1.6)*	9.0×10^{-1} *(1.0)*	1.5×10^{-1} *(1.6)*	9.0×10^{-1} *(1.0)*
8	9.0×10^{-2} *(1.7)*	8.2×10^{-1} *(1.1)*	8.9×10^{-2} *(1.7)*	8.2×10^{-1} *(1.1)*
16	4.0×10^{-2} *(2.2)*	6.6×10^{-1} *(1.2)*	4.0×10^{-2} *(2.2)*	6.6×10^{-1} *(1.2)*
		Function: w		
1	5.1×10^{-1}	9.7×10^{-1}	5.1×10^{-1}	9.7×10^{-1}
2	3.3×10^{-1} *(1.6)*	9.5×10^{-1} *(1.0)*	3.3×10^{-1} *(1.6)*	9.5×10^{-1} *(1.0)*
4	2.1×10^{-1} *(1.6)*	9.1×10^{-1} *(1.0)*	2.1×10^{-1} *(1.6)*	9.1×10^{-1} *(1.0)*
8	1.2×10^{-1} *(1.7)*	8.3×10^{-1} *(1.1)*	1.2×10^{-1} *(1.7)*	8.3×10^{-1} *(1.1)*
16	5.5×10^{-2} *(2.2)*	6.7×10^{-1} *(1.2)*	5.5×10^{-2} *(2.2)*	6.7×10^{-1} *(1.2)*
		Function: n		
1	1.5×10^{-1}	5.5×10^{-1}	1.3×10^{-1}	5.8×10^{-1}
2	9.4×10^{-2} *(1.6)*	4.1×10^{-1} *(1.3)*	8.1×10^{-2} *(1.6)*	4.2×10^{-1} *(1.4)*
4	4.9×10^{-2} *(1.9)*	2.6×10^{-1} *(1.6)*	4.4×10^{-2} *(1.8)*	2.6×10^{-1} *(1.6)*
8	2.0×10^{-2} *(2.4)*	1.4×10^{-1} *(1.9)*	1.9×10^{-2} *(2.3)*	1.4×10^{-1} *(1.9)*
16	6.0×10^{-3} *(3.4)*	6.3×10^{-2} *(2.2)*	5.6×10^{-3} *(3.3)*	6.3×10^{-2} *(2.2)*
		Function: p		
1	1.6×10^{-1}	6.2×10^{-1}	1.4×10^{-1}	6.1×10^{-1}
2	8.6×10^{-2} *(1.9)*	3.1×10^{-1} *(2.0)*	7.6×10^{-2} *(1.9)*	3.0×10^{-1} *(2.0)*
4	4.6×10^{-2} *(1.9)*	1.8×10^{-1} *(1.7)*	4.3×10^{-2} *(1.8)*	1.8×10^{-1} *(1.6)*
8	2.0×10^{-2} *(2.3)*	1.1×10^{-1} *(1.7)*	1.9×10^{-2} *(2.3)*	1.1×10^{-1} *(1.7)*
16	5.9×10^{-3} *(3.3)*	5.0×10^{-2} *(2.1)*	5.7×10^{-3} *(3.3)*	5.0×10^{-2} *(2.1)*

Table 3.6: $L_2(\Omega)$ and $H^1(\mathcal{E})$-error of u_h, v_h, w_h and dependent functions n, p in function of grid density parameter K for the first device under bias of magnitude 16. Numbers in brackets denote the rate of convergence. Solution for CSIPG method with $K = 32$ is taken as a reference function.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
Function: u								
1	9.2×10^{-2}		4.2×10^{-1}		1.0×10^{-1}		4.2×10^{-1}	
2	1.3×10^{-1}	*(0.7)*	5.4×10^{-1}	*(0.8)*	1.3×10^{-1}	*(0.8)*	5.2×10^{-1}	*(0.8)*
4	8.7×10^{-2}	*(1.5)*	4.2×10^{-1}	*(1.3)*	8.5×10^{-2}	*(1.5)*	4.1×10^{-1}	*(1.3)*
8	3.4×10^{-2}	*(2.6)*	1.8×10^{-1}	*(2.4)*	3.4×10^{-2}	*(2.5)*	1.8×10^{-1}	*(2.3)*
16	9.2×10^{-3}	*(3.7)*	5.1×10^{-2}	*(3.5)*	9.4×10^{-3}	*(3.6)*	5.2×10^{-2}	*(3.4)*
Function: v								
1	8.4×10^{-1}		9.7×10^{-1}		7.9×10^{-1}		9.7×10^{-1}	
2	6.4×10^{-1}	*(1.3)*	9.6×10^{-1}	*(1.0)*	5.7×10^{-1}	*(1.4)*	9.6×10^{-1}	*(1.0)*
4	4.4×10^{-1}	*(1.5)*	9.2×10^{-1}	*(1.0)*	4.1×10^{-1}	*(1.4)*	9.2×10^{-1}	*(1.0)*
8	2.5×10^{-1}	*(1.8)*	8.5×10^{-1}	*(1.1)*	2.4×10^{-1}	*(1.7)*	8.5×10^{-1}	*(1.1)*
16	1.1×10^{-1}	*(2.3)*	6.9×10^{-1}	*(1.2)*	1.1×10^{-1}	*(2.3)*	6.9×10^{-1}	*(1.2)*
Function: w								
1	3.2×10^{-1}		9.7×10^{-1}		3.2×10^{-1}		9.7×10^{-1}	
2	2.3×10^{-1}	*(1.4)*	9.6×10^{-1}	*(1.0)*	2.3×10^{-1}	*(1.4)*	9.6×10^{-1}	*(1.0)*
4	1.6×10^{-1}	*(1.4)*	9.3×10^{-1}	*(1.0)*	1.6×10^{-1}	*(1.4)*	9.3×10^{-1}	*(1.0)*
8	9.4×10^{-2}	*(1.7)*	8.6×10^{-1}	*(1.1)*	9.4×10^{-2}	*(1.7)*	8.6×10^{-1}	*(1.1)*
16	4.3×10^{-2}	*(2.2)*	7.0×10^{-1}	*(1.2)*	4.3×10^{-2}	*(2.2)*	7.0×10^{-1}	*(1.2)*
Function: n								
1	2.7		4.3		9.5×10^{-1}		2.0	
2	2.0	*(1.4)*	4.4	*(1.0)*	1.1	*(0.8)*	2.7	*(0.7)*
4	9.3×10^{-1}	*(2.2)*	2.7	*(1.6)*	7.0×10^{-1}	*(1.6)*	2.2	*(1.2)*
8	3.0×10^{-1}	*(3.1)*	1.2	*(2.3)*	2.6×10^{-1}	*(2.7)*	1.1	*(2.0)*
16	7.6×10^{-2}	*(3.9)*	4.7×10^{-1}	*(2.5)*	6.9×10^{-2}	*(3.7)*	4.7×10^{-1}	*(2.4)*
Function: p								
1	2.7		3.6		1.1		2.0	
2	2.2	*(1.2)*	5.2	*(0.7)*	1.4	*(0.8)*	3.9	*(0.5)*
4	1.1	*(1.9)*	4.6	*(1.1)*	9.3×10^{-1}	*(1.5)*	4.2	*(0.9)*
8	3.8×10^{-1}	*(2.9)*	2.5	*(1.9)*	3.6×10^{-1}	*(2.6)*	2.4	*(1.7)*
16	1.0×10^{-1}	*(3.8)*	1.1	*(2.3)*	9.8×10^{-2}	*(3.6)*	1.1	*(2.3)*

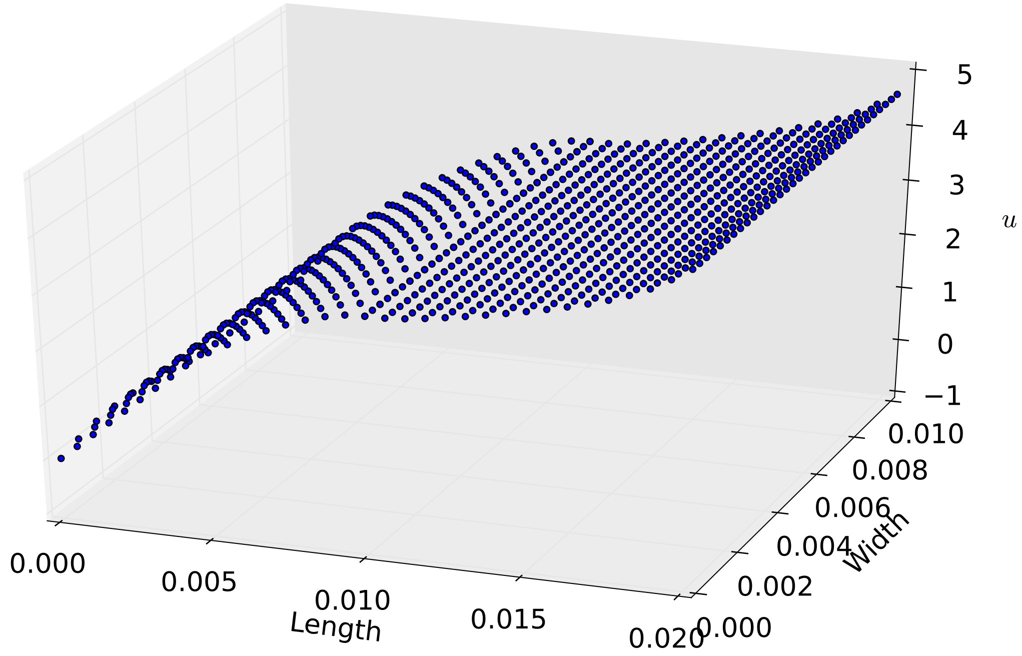

Figure 3.6: Function u_h for the first example for the bias of magnitude 16. Note one-dimensional character of the solution.

For higher biases (tables 3.5, 3.6) the trend is similar, however we observe slower start for u, n, p. This effect may be due to the fact, that generally higher bias increases coupling between the van Roosbroeck equations (1.2.1):

$$\begin{aligned} -\nabla \cdot (\varepsilon \nabla u^*) &= k_1 - n + p, \\ -\nabla \cdot (\mu_n e^{u^* - v^*} \nabla v^*) &= P(u^*, v^*, w^*), \\ -\nabla \cdot (\mu_p e^{w^* - u^*} \nabla w^*) &= -P(u^*, v^*, w^*). \end{aligned} \tag{3.2.6}$$

Let us take a closer look on this system. For small bias, the majority carrier concentrations (n in n-type regions, p in p-type regions) do not increase considerably, thus the right-hand side of the Poisson equation does not change much in comparison with the equilibrium case. Main difference in u is due to the Dirichlet boundary conditions accounting for increased bias. The recombination rate P is also small, so while v, w depends strongly on u due to the exponent in the coefficients, this is not the case in the other direction. Also coupling between v and w is loose.

On the other hand, under high bias the recombination rate P is big, and the concentrations n, p are both of similar order of magnitude. Thus coupling between unknown functions is strong, coming both from the coefficients as well as from the right-hand sides of the respective equations. Errors of discrete solutions thus accumulate and the convergence rate is worse.

As can be observed in figures 3.5, 3.6, in this case the solution has a one-dimensional nature. To study more sophisticated behavior, we introduce a second device with a more complex structure (figure 3.7, see table 3.7 for “material” parameters). This is also a p-n diode, but it has contacts attached to the horizontal edges of the device.

Plots of the functions u_h, v_h, w_h for simulations with $K = 8$ are presented in figures 3.8, 3.9. For the equilibrium state, the solutions u_h preserve one-dimensional character despite of position of the

Table 3.7: Parameters of second device used in simulations.

Param.	Ω_1	Ω_2	Ω_3	Ω_4	Ω_5	Ω_6
Length	5×10^{-3}	5×10^{-3}	1×10^{-2}	1×10^{-2}	5×10^{-3}	5×10^{-3}
Width	1×10^{-2}	1×10^{-2}	1×10^{-2}	1×10^{-2}	1×10^{-2}	1×10^{-2}
N_x	$4K+1$	$4K+1$	$2K+1$	$2K+1$	$4K+1$	$4K+1$
N_y	$4K+1$	$2K+1$	$2K+1$	$2K+1$	$2K+1$	$4K+1$
ε	3×10^{-3}	3×10^{-3}	1×10^{-3}	1×10^{-3}	3×10^{-3}	3×10^{-3}
μ_n	1×10^{3}	1×10^{3}	3×10^{3}	3×10^{3}	1×10^{3}	1×10^{3}
μ_p	1×10^{2}	1×10^{2}	3×10^{2}	3×10^{2}	1×10^{2}	1×10^{2}
k_1	3×10^{2}	3×10^{2}	5×10^{2}	-5×10^{2}	-3×10^{2}	-3×10^{2}
C_{rad}	1×10^{-3}	1×10^{-3}	2×10^{-3}	2×10^{-3}	1×10^{-3}	1×10^{-3}

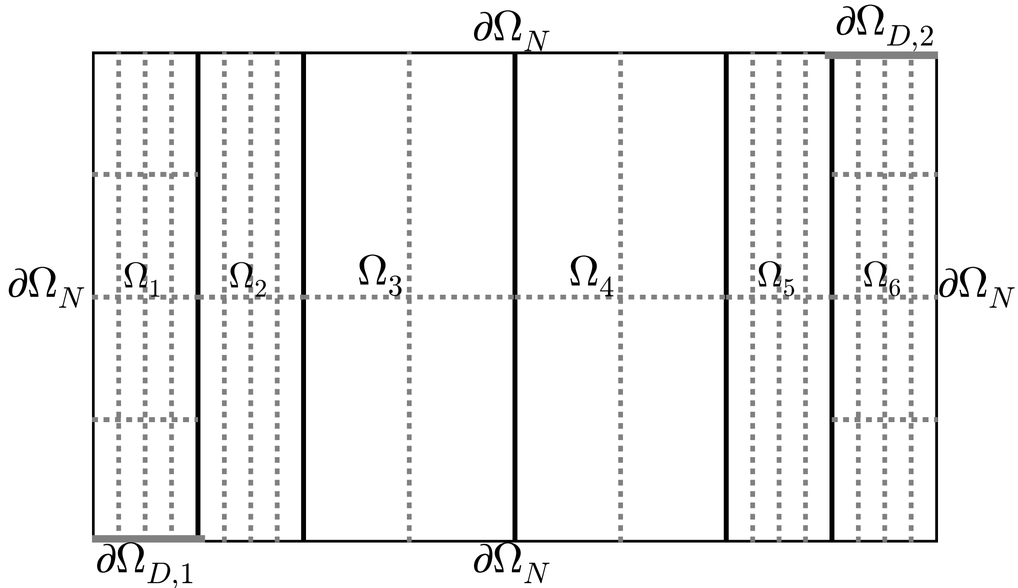

Figure 3.7: Schema of the second device used in simulations. Left contact is attached to bottom edge of Ω_1 (gray color) and right contact is attached to top edge of Ω_6. Grid for $K = 1$ is presented.

contacts. This behavior is generally preserved for small bias. For high bias, however, we clearly see that this is not the case, especially near the contacts. On the other hand, functions v_h, w_h do not vary much, besides of proximity of the contacts. Even for high biases we do not observe significant variations.

As we see in table 3.8, the convergence rates are similar as for the first device for the equilibrium state: linear convergence of $H^1(\mathcal{E})$-error and quadratic convergence of $L^2(\Omega)$-error for functions u, n, p and sublinear convergence of L_2-error for functions v, w. We performed simulations also for bias 6, 10 and 16 (tables 3.9, 3.10, 3.11). We conclude that these results are in agreement with our previous observations, i.e. the convergence starts slower if bias is higher.

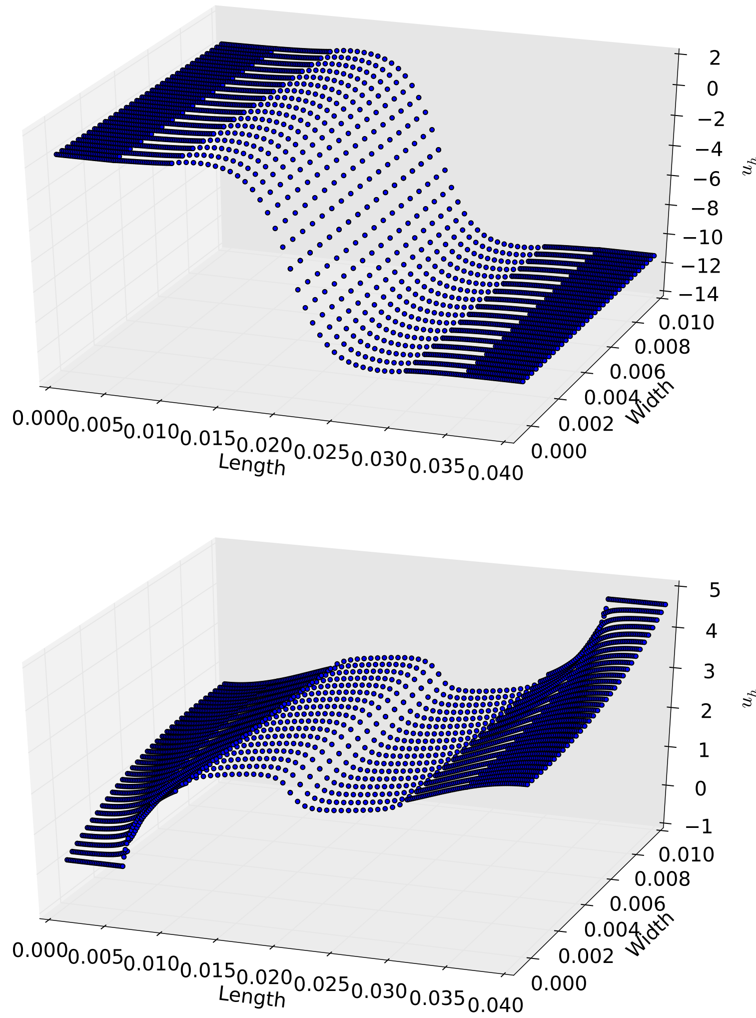

Figure 3.8: Function u_h for the second example in the equilibrium state (left) and for bias equal to 16 (right) for $K = 8$.

Table 3.8: $L_2(\Omega)$ and $H^1(\mathcal{E})$-error of u_h in function of grid density parameter K for the second device in equilibrium state. Numbers in brackets denote the rate of convergence.

	CSIPG		CWOPSIP	
K	$L_2(\Omega)$	$H^1(\Omega)$	$L_2(\Omega)$	$H^1(\Omega)$
1	7.4×10^{-3}	5.9	7.4×10^{-3}	5.9
2	2.0×10^{-3} *(3.8)*	3.0 *(2.0)*	2.0×10^{-3} *(3.8)*	3.0 *(2.0)*
4	5.0×10^{-4} *(3.9)*	1.5 *(2.0)*	5.0×10^{-4} *(3.9)*	1.5 *(2.0)*
8	1.2×10^{-4} *(4.1)*	7.4×10^{-1} *(2.0)*	1.2×10^{-4} *(4.1)*	7.4×10^{-1} *(2.0)*
16	2.6×10^{-5} *(4.7)*	3.3×10^{-1} *(2.2)*	2.6×10^{-5} *(4.7)*	3.3×10^{-1} *(2.2)*

Table 3.9: $L_2(\Omega)$ and $H^1(\mathcal{E})$-error of u_h, v_h, w_h and dependent functions n, p in function of grid density parameter K for the second device for $u_{\text{bias}} = 6$. Numbers in brackets denote the rate of convergence.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
	Function: u							
1	6.4×10^{-2}		5.7×10^{-1}		6.4×10^{-2}		5.7×10^{-1}	
2	1.8×10^{-2}	*(3.5)*	3.0×10^{-1}	*(1.9)*	1.8×10^{-2}	*(3.5)*	3.0×10^{-1}	*(1.9)*
4	4.7×10^{-3}	*(3.9)*	1.5×10^{-1}	*(2.0)*	4.7×10^{-3}	*(3.9)*	1.5×10^{-1}	*(2.0)*
8	1.2×10^{-3}	*(4.1)*	7.5×10^{-2}	*(2.0)*	1.2×10^{-3}	*(4.1)*	7.5×10^{-2}	*(2.0)*
16	2.5×10^{-4}	*(4.7)*	3.4×10^{-2}	*(2.2)*	2.5×10^{-4}	*(4.7)*	3.4×10^{-2}	*(2.2)*
	Function: v							
1	7.9×10^{-2}		9.9×10^{-1}		7.9×10^{-2}		9.9×10^{-1}	
2	3.9×10^{-2}	*(2.0)*	9.1×10^{-1}	*(1.1)*	3.9×10^{-2}	*(2.0)*	9.1×10^{-1}	*(1.1)*
4	1.9×10^{-2}	*(2.1)*	8.0×10^{-1}	*(1.1)*	1.9×10^{-2}	*(2.1)*	8.0×10^{-1}	*(1.1)*
8	8.4×10^{-3}	*(2.2)*	6.6×10^{-1}	*(1.2)*	8.4×10^{-3}	*(2.2)*	6.6×10^{-1}	*(1.2)*
16	3.0×10^{-3}	*(2.8)*	4.6×10^{-1}	*(1.4)*	3.0×10^{-3}	*(2.8)*	4.6×10^{-1}	*(1.4)*
	Function: w							
1	8.1×10^{-1}		9.9×10^{-1}		8.1×10^{-1}		9.9×10^{-1}	
2	4.0×10^{-1}	*(2.0)*	9.1×10^{-1}	*(1.1)*	4.0×10^{-1}	*(2.0)*	9.1×10^{-1}	*(1.1)*
4	1.9×10^{-1}	*(2.1)*	8.0×10^{-1}	*(1.1)*	1.9×10^{-1}	*(2.1)*	8.0×10^{-1}	*(1.1)*
8	8.6×10^{-2}	*(2.2)*	6.6×10^{-1}	*(1.2)*	8.6×10^{-2}	*(2.2)*	6.6×10^{-1}	*(1.2)*
16	3.0×10^{-2}	*(2.8)*	4.6×10^{-1}	*(1.4)*	3.0×10^{-2}	*(2.8)*	4.6×10^{-1}	*(1.4)*
	Function: n							
1	2.8×10^{-1}		1.0		2.8×10^{-1}		1.0	
2	4.2×10^{-2}	*(6.6)*	3.5×10^{-1}	*(2.9)*	4.2×10^{-2}	*(6.6)*	3.5×10^{-1}	*(2.9)*
4	9.5×10^{-3}	*(4.4)*	1.7×10^{-1}	*(2.1)*	9.5×10^{-3}	*(4.4)*	1.7×10^{-1}	*(2.1)*
8	2.3×10^{-3}	*(4.2)*	8.1×10^{-2}	*(2.1)*	2.3×10^{-3}	*(4.2)*	8.1×10^{-2}	*(2.1)*
16	4.7×10^{-4}	*(4.8)*	3.6×10^{-2}	*(2.2)*	4.7×10^{-4}	*(4.8)*	3.6×10^{-2}	*(2.2)*
	Function: p							
1	2.8×10^{-1}		1.0		2.8×10^{-1}		1.0	
2	4.2×10^{-2}	*(6.6)*	3.5×10^{-1}	*(2.9)*	4.2×10^{-2}	*(6.6)*	3.5×10^{-1}	*(2.9)*
4	9.5×10^{-3}	*(4.4)*	1.7×10^{-1}	*(2.1)*	9.5×10^{-3}	*(4.4)*	1.7×10^{-1}	*(2.1)*
8	2.3×10^{-3}	*(4.2)*	8.1×10^{-2}	*(2.1)*	2.3×10^{-3}	*(4.2)*	8.1×10^{-2}	*(2.1)*
16	4.7×10^{-4}	*(4.8)*	3.6×10^{-2}	*(2.2)*	4.7×10^{-4}	*(4.8)*	3.6×10^{-2}	*(2.2)*

Table 3.10: $L_2(\Omega)$ and $H^1(\mathcal{E})$-error of u_h, v_h, w_h and dependent functions n, p in function of grid density parameter K for the second device for $u_{\text{bias}} = 10$. Numbers in brackets denote the rate of convergence.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
Function: u								
1	1.1×10^{-1}		7.0×10^{-1}		1.1×10^{-1}		7.0×10^{-1}	
2	3.7×10^{-2}	*(3.1)*	4.1×10^{-1}	*(1.7)*	3.7×10^{-2}	*(3.1)*	4.1×10^{-1}	*(1.7)*
4	1.0×10^{-2}	*(3.7)*	2.1×10^{-1}	*(1.9)*	1.0×10^{-2}	*(3.7)*	2.1×10^{-1}	*(1.9)*
8	2.8×10^{-3}	*(3.7)*	1.1×10^{-1}	*(2.0)*	2.8×10^{-3}	*(3.6)*	1.1×10^{-1}	*(2.0)*
16	7.0×10^{-4}	*(3.9)*	4.8×10^{-2}	*(2.2)*	7.9×10^{-4}	*(3.6)*	4.8×10^{-2}	*(2.2)*
Function: v								
1	1.6×10^{-1}		1.0		1.6×10^{-1}		1.0	
2	8.9×10^{-2}	*(1.8)*	9.8×10^{-1}	*(1.0)*	8.9×10^{-2}	*(1.8)*	9.8×10^{-1}	*(1.0)*
4	5.1×10^{-2}	*(1.7)*	9.3×10^{-1}	*(1.1)*	5.1×10^{-2}	*(1.7)*	9.3×10^{-1}	*(1.1)*
8	2.8×10^{-2}	*(1.8)*	8.5×10^{-1}	*(1.1)*	2.8×10^{-2}	*(1.8)*	8.5×10^{-1}	*(1.1)*
16	1.2×10^{-2}	*(2.3)*	6.8×10^{-1}	*(1.3)*	1.2×10^{-2}	*(2.3)*	6.8×10^{-1}	*(1.3)*
Function: w								
1	2.1×10^{-1}		1.0		2.1×10^{-1}		1.0	
2	1.2×10^{-1}	*(1.8)*	9.8×10^{-1}	*(1.0)*	1.2×10^{-1}	*(1.8)*	9.8×10^{-1}	*(1.0)*
4	7.0×10^{-2}	*(1.7)*	9.3×10^{-1}	*(1.1)*	7.0×10^{-2}	*(1.7)*	9.3×10^{-1}	*(1.1)*
8	3.9×10^{-2}	*(1.8)*	8.5×10^{-1}	*(1.1)*	3.9×10^{-2}	*(1.8)*	8.5×10^{-1}	*(1.1)*
16	1.7×10^{-2}	*(2.3)*	6.8×10^{-1}	*(1.3)*	1.7×10^{-2}	*(2.3)*	6.8×10^{-1}	*(1.3)*
Function: n								
1	1.8×10^{-1}		8.8×10^{-1}		1.8×10^{-1}		8.8×10^{-1}	
2	3.5×10^{-2}	*(5.1)*	3.5×10^{-1}	*(2.6)*	3.4×10^{-2}	*(5.2)*	3.5×10^{-1}	*(2.6)*
4	1.0×10^{-2}	*(3.4)*	1.9×10^{-1}	*(1.9)*	1.0×10^{-2}	*(3.3)*	1.9×10^{-1}	*(1.9)*
8	3.4×10^{-3}	*(3.0)*	9.7×10^{-2}	*(1.9)*	3.4×10^{-3}	*(3.0)*	9.7×10^{-2}	*(1.9)*
16	9.5×10^{-4}	*(3.5)*	4.5×10^{-2}	*(2.2)*	1.0×10^{-3}	*(3.4)*	4.5×10^{-2}	*(2.2)*
Function: p								
1	1.8×10^{-1}		8.8×10^{-1}		1.7×10^{-1}		8.8×10^{-1}	
2	3.4×10^{-2}	*(5.1)*	3.5×10^{-1}	*(2.6)*	3.4×10^{-2}	*(5.2)*	3.4×10^{-1}	*(2.5)*
4	1.0×10^{-2}	*(3.4)*	1.8×10^{-1}	*(1.9)*	1.0×10^{-2}	*(3.3)*	1.8×10^{-1}	*(1.9)*
8	3.4×10^{-3}	*(3.0)*	9.7×10^{-2}	*(1.9)*	3.4×10^{-3}	*(3.0)*	9.7×10^{-2}	*(1.9)*
16	9.6×10^{-4}	*(3.5)*	4.5×10^{-2}	*(2.2)*	1.0×10^{-3}	*(3.4)*	4.5×10^{-2}	*(2.2)*

Table 3.11: $L_2(\Omega)$ and $H^1(\mathcal{E})$-error of u_h, v_h, w_h and dependent functions n, p in function of grid density parameter K for the second device for $u_{\text{bias}} = 16$. Numbers in brackets denote the rate of convergence.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
Function: u								
1	3.5×10^{-3}		2.2		3.8×10^{-3}		2.2	
2	2.0×10^{-3}	*(1.7)*	1.6	*(1.4)*	1.9×10^{-3}	*(2.0)*	1.6	*(1.4)*
4	1.1×10^{-3}	*(1.9)*	9.8×10^{-1}	*(1.6)*	1.0×10^{-3}	*(1.8)*	9.8×10^{-1}	*(1.6)*
8	4.8×10^{-4}	*(2.2)*	5.5×10^{-1}	*(1.8)*	4.5×10^{-4}	*(2.2)*	5.6×10^{-1}	*(1.8)*
16	1.6×10^{-4}	*(3.1)*	2.8×10^{-1}	*(2.0)*	1.5×10^{-4}	*(3.0)*	2.8×10^{-1}	*(2.0)*
Function: v								
1	2.3×10^{-2}		7.5×10^{1}		2.3×10^{-2}		7.5×10^{1}	
2	1.5×10^{-2}	*(1.6)*	7.4×10^{1}	*(1.0)*	1.4×10^{-2}	*(1.6)*	7.4×10^{1}	*(1.0)*
4	9.3×10^{-3}	*(1.6)*	7.1×10^{1}	*(1.0)*	9.3×10^{-3}	*(1.5)*	7.1×10^{1}	*(1.0)*
8	5.4×10^{-3}	*(1.7)*	6.6×10^{1}	*(1.1)*	5.5×10^{-3}	*(1.7)*	6.6×10^{1}	*(1.1)*
16	2.4×10^{-3}	*(2.2)*	5.3×10^{1}	*(1.2)*	2.5×10^{-3}	*(2.2)*	5.4×10^{1}	*(1.2)*
Function: w								
1	2.5×10^{-2}		7.6×10^{1}		2.4×10^{-2}		7.6×10^{1}	
2	1.6×10^{-2}	*(1.6)*	7.4×10^{1}	*(1.0)*	1.5×10^{-2}	*(1.5)*	7.4×10^{1}	*(1.0)*
4	9.7×10^{-3}	*(1.6)*	7.2×10^{1}	*(1.0)*	9.7×10^{-3}	*(1.6)*	7.2×10^{1}	*(1.0)*
8	5.5×10^{-3}	*(1.8)*	6.6×10^{1}	*(1.1)*	5.5×10^{-3}	*(1.8)*	6.6×10^{1}	*(1.1)*
16	2.5×10^{-3}	*(2.3)*	5.4×10^{1}	*(1.2)*	2.5×10^{-3}	*(2.3)*	5.4×10^{1}	*(1.2)*
Function: n								
1	3.5		1.5×10^{3}		1.9		1.0×10^{3}	
2	2.3	*(1.5)*	1.2×10^{3}	*(1.2)*	1.9	*(1.0)*	1.0×10^{3}	*(1.0)*
4	1.3	*(1.8)*	9.0×10^{2}	*(1.3)*	1.3	*(1.5)*	9.1×10^{2}	*(1.2)*
8	5.7×10^{-1}	*(2.3)*	6.0×10^{2}	*(1.5)*	6.1×10^{-1}	*(2.1)*	6.2×10^{2}	*(1.5)*
16	1.8×10^{-1}	*(3.1)*	3.4×10^{2}	*(1.8)*	2.1×10^{-1}	*(2.9)*	3.5×10^{2}	*(1.8)*
Function: p								
1	3.6		1.8×10^{3}		2.1		1.5×10^{3}	
2	2.4	*(1.5)*	1.4×10^{3}	*(1.3)*	2.1	*(1.0)*	1.3×10^{3}	*(1.1)*
4	1.3	*(1.8)*	9.8×10^{2}	*(1.4)*	1.3	*(1.6)*	9.8×10^{2}	*(1.3)*
8	5.8×10^{-1}	*(2.3)*	6.3×10^{2}	*(1.6)*	6.1×10^{-1}	*(2.2)*	6.3×10^{2}	*(1.5)*
16	1.8×10^{-1}	*(3.1)*	3.5×10^{2}	*(1.8)*	2.1×10^{-1}	*(2.9)*	3.5×10^{2}	*(1.8)*

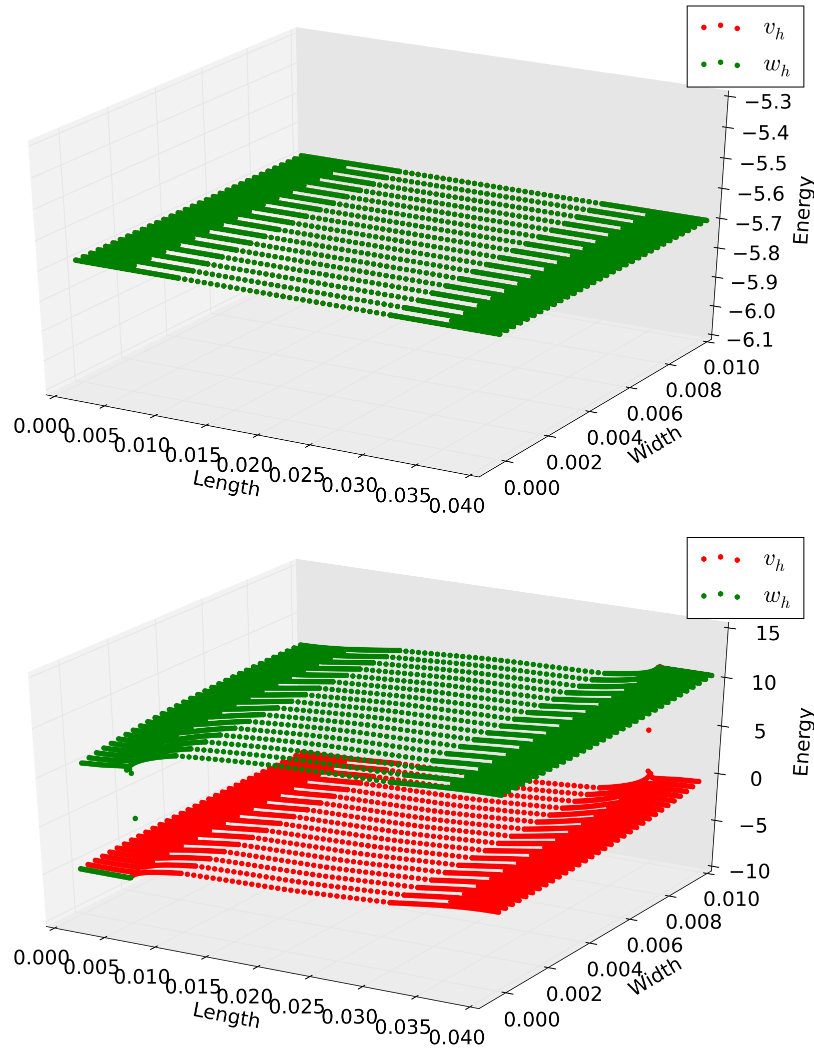

Figure 3.9: Functions v_h, w_h for the second example in the equilibrium state (left) and for bias equal to 16 (right) for $K = 8$.

To this moment it is still not clear whether the slow start emerges mostly due to the coupling of the equations, or do it also arise due to the two-dimensional character of the solutions. To study the latter possibility, we introduce a device presented in figure 3.10 (see also table 3.12). This is also a p-n diode, but this time the interface between a n-type region and a p-type region is not perpendicular to any axis. The n-type region consists of subdomains $\Omega_7, \Omega_4, \Omega_1, \Omega_2, \Omega_3$, thus it forms a L-shape, and the remaining subdomains belong to the square-shaped p-type region. The depleted region goes along the internal boundary of the L-shape.

Simulation results are presented in figures 3.11, 3.12. In this case, function u_h clearly exhibits two-dimensional behaviour even in the equilibrium case.

Error and convergence rates for the equilibrium state is presented in tables 3.13, 3.14. These results clearly indicate that there is no slow start of convergence for $H^1(\mathcal{E})$-error in this case, despite of the two-dimensional character of the unknown function.

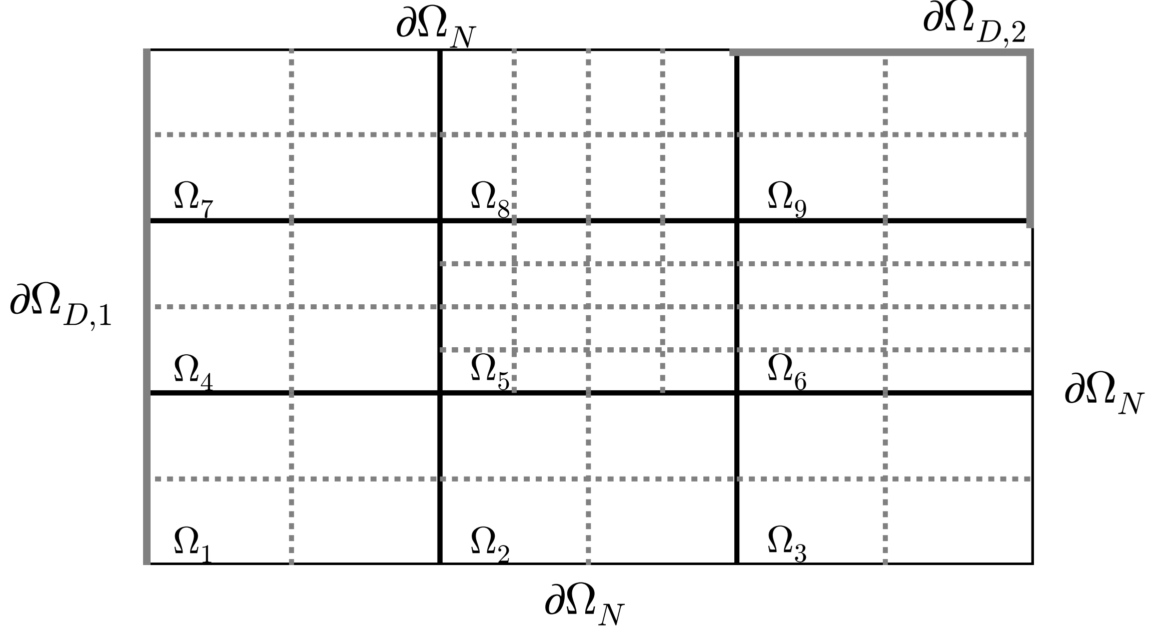

Figure 3.10: Schema of the second device used in simulations. Layers $\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_7$ correspond to the n-type region, while the remainder correspond to the p-type region. Left contact is attached to whole left edge, while right contact is attached to the boundary of Ω_9. Grid for $K = 1$ is presented.

Table 3.12: Parameters of third device used in simulations.

Param.	$\Omega_1, \Omega_2, \Omega_3, \Omega_4, \Omega_7$	$\Omega_5, \Omega_6, \Omega_8, \Omega_9$	Grid	N_x	N_y
Length	1×10^{-2}	1×10^{-2}	Ω_1	$2K+1$	$2K+1$
Width	1×10^{-2}	1×10^{-2}	Ω_2	$2K+1$	$2K+1$
ε	3×10^{-3}	1×10^{-3}	Ω_3	$2K+1$	$2K+1$
μ_n	1×10^{3}	3×10^{3}	Ω_4	$2K+1$	$2K+1$
μ_p	1×10^{2}	3×10^{2}	Ω_5	$4K+1$	$4K+1$
k_1	3×10^{2}	-3×10^{2}	Ω_6	$2K+1$	$4K+1$
C_{rad}	1×10^{-3}	2×10^{-3}	Ω_7	$2K+1$	$2K+1$
			Ω_8	$4K+1$	$2K+1$
			Ω_9	$2K+1$	$2K+1$

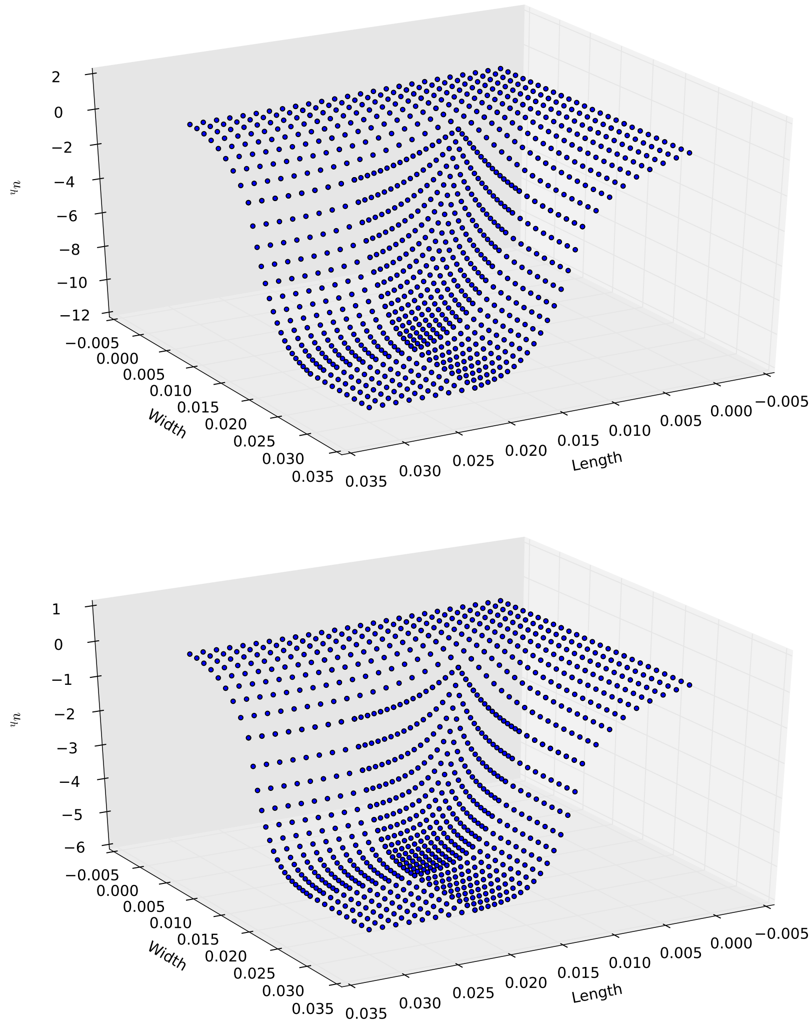

Figure 3.11: Function u_h for the third example in the equilibrium state (left) and for bias of magnitude 8 (right) for $K = 4$.

Table 3.13: $L_2(\Omega)$- and $H^1(\Omega)$-error of u_h in function of grid density parameter K for the third device in equilibrium state. Numbers in brackets denote the estimated rate of convergence.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
1	2.0×10^{-2}		1.9×10^{-1}		2.0×10^{-2}		1.9×10^{-1}	
2	5.2×10^{-3}	*(3.8)*	9.5×10^{-2}	*(2.0)*	5.2×10^{-3}	*(3.8)*	9.5×10^{-2}	*(2.0)*
4	1.3×10^{-3}	*(3.9)*	4.8×10^{-2}	*(2.0)*	1.3×10^{-3}	*(3.9)*	4.8×10^{-2}	*(2.0)*
8	3.5×10^{-4}	*(3.8)*	2.4×10^{-2}	*(2.0)*	3.5×10^{-4}	*(3.8)*	2.4×10^{-2}	*(2.0)*
16	9.4×10^{-5}	*(3.7)*	1.2×10^{-2}	*(2.0)*	9.4×10^{-5}	*(3.7)*	1.2×10^{-2}	*(2.0)*
32	2.3×10^{-5}	*(4.0)*	5.4×10^{-3}	*(2.2)*	2.3×10^{-5}	*(4.0)*	5.4×10^{-3}	*(2.2)*

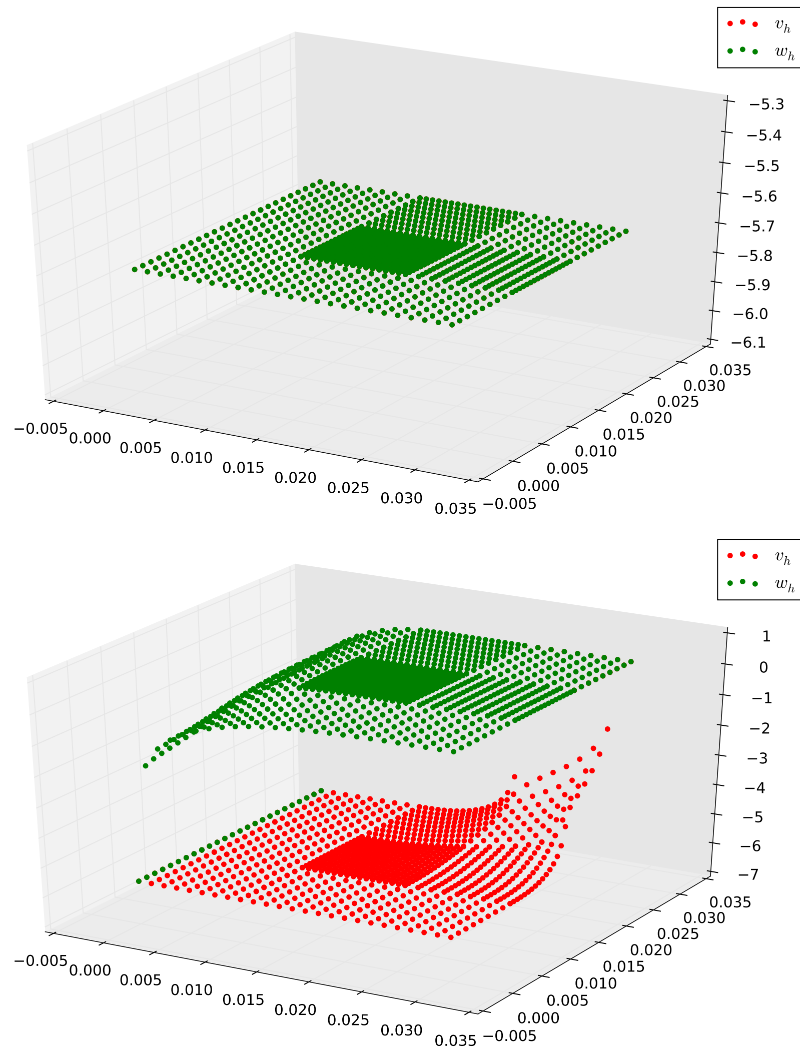

Figure 3.12: Functions v_h, w_h for the third example in the equilibrium state (left) and for bias of magnitude 8 (right) for $K = 4$.

Table 3.14: $L_2(\Omega)$- and $H^1(\Omega)$-error of u_h, v_h, w_h and dependent functions n, p in function of grid density parameter K for the third device for $u_{\text{bias}} = 8$. Numbers in brackets denote the estimated rate of convergence.

	CSIPG		CWOPSIP	
K	$L_2(\Omega)$	$H^1(\Omega)$	$L_2(\Omega)$	$H^1(\Omega)$
Function: u				
1	1.9×10^{-3}	2.2	2.0×10^{-3}	2.2
2	5.3×10^{-4} *(3.6)*	1.2 *(1.9)*	5.2×10^{-4} *(3.8)*	1.2 *(1.9)*
4	1.6×10^{-4} *(3.3)*	5.8×10^{-1} *(2.0)*	1.5×10^{-4} *(3.5)*	5.8×10^{-1} *(2.0)*
8	6.0×10^{-5} *(2.7)*	2.9×10^{-1} *(2.0)*	5.7×10^{-5} *(2.6)*	2.9×10^{-1} *(2.0)*
16	2.4×10^{-5} *(2.5)*	1.4×10^{-1} *(2.0)*	2.2×10^{-5} *(2.5)*	1.4×10^{-1} *(2.0)*
32	8.0×10^{-6} *(3.1)*	6.4×10^{-2} *(2.2)*	4.4×10^{-6} *(5.0)*	6.4×10^{-2} *(2.2)*
Function: v				
1	2.7×10^{-2}	4.6×10^{1}	2.7×10^{-2}	4.6×10^{1}
2	1.6×10^{-2} *(1.7)*	4.5×10^{1} *(1.0)*	1.6×10^{-2} *(1.7)*	4.5×10^{1} *(1.0)*
4	9.2×10^{-3} *(1.7)*	4.3×10^{1} *(1.0)*	9.2×10^{-3} *(1.7)*	4.3×10^{1} *(1.0)*
8	5.2×10^{-3} *(1.8)*	4.0×10^{1} *(1.1)*	5.2×10^{-3} *(1.8)*	4.0×10^{1} *(1.1)*
16	2.7×10^{-3} *(1.9)*	3.5×10^{1} *(1.1)*	2.7×10^{-3} *(1.9)*	3.5×10^{1} *(1.1)*
32	1.1×10^{-3} *(2.4)*	2.7×10^{1} *(1.3)*	1.1×10^{-3} *(2.4)*	2.7×10^{1} *(1.3)*
Function: w				
1	3.6×10^{-2}	6.9×10^{1}	3.6×10^{-2}	6.9×10^{1}
2	2.2×10^{-2} *(1.6)*	6.7×10^{1} *(1.0)*	2.2×10^{-2} *(1.6)*	6.7×10^{1} *(1.0)*
4	1.3×10^{-2} *(1.6)*	6.5×10^{1} *(1.0)*	1.3×10^{-2} *(1.6)*	6.5×10^{1} *(1.0)*
8	7.8×10^{-3} *(1.7)*	6.1×10^{1} *(1.1)*	7.8×10^{-3} *(1.7)*	6.1×10^{1} *(1.1)*
16	4.2×10^{-3} *(1.9)*	5.4×10^{1} *(1.1)*	4.2×10^{-3} *(1.9)*	5.4×10^{1} *(1.1)*
32	1.8×10^{-3} *(2.4)*	4.2×10^{1} *(1.3)*	1.8×10^{-3} *(2.4)*	4.2×10^{1} *(1.3)*
Function: n				
1	2.5×10^{-1}	2.7×10^{2}	2.5×10^{-1}	2.7×10^{2}
2	8.3×10^{-2} *(3.1)*	1.6×10^{2} *(1.7)*	7.8×10^{-2} *(3.2)*	1.6×10^{2} *(1.7)*
4	2.7×10^{-2} *(3.0)*	8.8×10^{1} *(1.8)*	2.5×10^{-2} *(3.1)*	8.8×10^{1} *(1.8)*
8	9.5×10^{-3} *(2.9)*	4.5×10^{1} *(2.0)*	9.2×10^{-3} *(2.8)*	4.5×10^{1} *(2.0)*
16	3.3×10^{-3} *(2.9)*	2.2×10^{1} *(2.0)*	3.1×10^{-3} *(3.0)*	2.2×10^{1} *(2.0)*
32	9.3×10^{-4} *(3.5)*	1.0×10^{1} *(2.2)*	8.2×10^{-4} *(3.8)*	1.0×10^{1} *(2.2)*
Function: p				
1	1.6×10^{-1}	1.5×10^{2}	1.4×10^{-1}	1.5×10^{2}
2	5.3×10^{-2} *(3.0)*	7.0×10^{1} *(2.1)*	4.7×10^{-2} *(2.9)*	7.0×10^{1} *(2.1)*
4	2.2×10^{-2} *(2.4)*	3.5×10^{1} *(2.0)*	2.1×10^{-2} *(2.3)*	3.5×10^{1} *(2.0)*
8	9.2×10^{-3} *(2.4)*	1.8×10^{1} *(2.0)*	9.1×10^{-3} *(2.3)*	1.8×10^{1} *(2.0)*
16	3.4×10^{-3} *(2.7)*	8.6 *(2.0)*	3.3×10^{-3} *(2.7)*	8.6 *(2.0)*
32	9.6×10^{-4} *(3.5)*	3.8 *(2.3)*	7.6×10^{-4} *(4.4)*	3.8 *(2.2)*

3.2.3 Formulation ψ, F_n, F_p: one dimension

After the study of theoretical setting in section 3.2.2, we would like to check whether the error estimates holds also in simulations of realistic semiconductor devices.

We start from a simple device: a p-n diode. It consists of two physical layers (table 3.15). We additionally divide these layers to introduce additional narrow layers near the interface of the n-type, p-type and contacts of the device to improve the convergence. Then in every layer we setup K equidistant nodes. Simulation is in one-dimension.

We start with the equilibrium case (table 3.16), where we present relative errors of CSIPG and CWOPSIP numerical solutions for the potential ψ. These results indicate clearly that errors of all these methods converge linearly to zero in $H^1(\Omega)$ norm as $h \to 0$. For $L_2(\Omega)$ norm, the errors drop quadratically in h. Also note that for given K errors are similar for both discretization methods.

Then we pass to non-equilibrium simulations for 1 V bias (table 3.17). For the potential ψ, the conclusion is as in equilibrium case. On the other hand, for the quasi-Fermi levels the situation is much worse. For CSIPG discretization, we observe sublinear convergence on both norms, and the $H^1(\Omega)$

Table 3.15: Schemata of devices used in the simulations.

p-n diode				
Layer	Material	Donor doping	Acceptor doping	Length
n-type	GaN	2×10^{18} cm^{-3}	0	100 nm
p-type	GaN	0	2×10^{19} cm^{-3}	100 nm

Blue laser				
Layer	Material	Donor doping	Acceptor doping	Length
n-base	GaN	3×10^{18} cm^{-3}	0	1000 nm
n-cladding	$Al_{0.1}Ga_{0.9}N$	3×10^{18} cm^{-3}	0	500 nm
n-waveguide	GaN	3×10^{18} cm^{-3}	0	100 nm
quantum well	$In_{0.2}Ga_{0.8}N$	0	0	4 nm
p-EBL	$Al_{0.2}Ga_{0.8}N$	0	2×10^{19} cm^{-3}	20 nm
p-cladding	$Al_{0.1}Ga_{0.9}N$	0	1×10^{19} cm^{-3}	500 nm

Table 3.16: Relative errors of the potential ψ. Simulation were performed for the p-n diode in the equilibrium state.

	CSIPG		CWOPSIP	
K	$L_2(\Omega)$	$H^1(\Omega)$	$L_2(\Omega)$	$H^1(\Omega)$
2	8.7×10^{-2}	4.6×10^{-1}	8.7×10^{-2}	4.6×10^{-1}
4	1.5×10^{-2} *(5.8)*	2.3×10^{-1} *(2.0)*	1.5×10^{-2} *(5.8)*	2.3×10^{-1} *(2.0)*
8	4.0×10^{-3} *(3.8)*	1.2×10^{-1} *(1.9)*	4.0×10^{-3} *(3.8)*	1.2×10^{-1} *(1.9)*
16	1.2×10^{-3} *(3.3)*	5.8×10^{-2} *(2.0)*	1.2×10^{-3} *(3.3)*	5.8×10^{-2} *(2.0)*
32	3.0×10^{-4} *(4.1)*	2.9×10^{-2} *(2.0)*	3.0×10^{-4} *(4.1)*	2.9×10^{-2} *(2.0)*
64	7.4×10^{-5} *(4.0)*	1.4×10^{-2} *(2.0)*	7.4×10^{-5} *(4.0)*	1.4×10^{-2} *(2.0)*
128	1.8×10^{-5} *(4.0)*	7.2×10^{-3} *(2.0)*	1.8×10^{-5} *(4.0)*	7.2×10^{-3} *(2.0)*
256	4.4×10^{-6} *(4.1)*	3.5×10^{-3} *(2.0)*	4.4×10^{-6} *(4.2)*	3.5×10^{-3} *(2.0)*

Table 3.17: Relative error of ψ, F_n and F_p. Simulation were performed for the p-n diode under 1 V bias.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
Function: ψ								
2	7.8×10^{-2}		4.8×10^{-1}		7.9×10^{-2}		4.8×10^{-1}	
4	3.0×10^{-2}	*(2.6)*	2.7×10^{-1}	*(1.8)*	3.0×10^{-2}	*(2.6)*	2.7×10^{-1}	*(1.8)*
8	7.1×10^{-3}	*(4.2)*	1.3×10^{-1}	*(2.0)*	7.1×10^{-3}	*(4.2)*	1.3×10^{-1}	*(2.0)*
16	1.5×10^{-3}	*(4.9)*	6.7×10^{-2}	*(2.0)*	1.5×10^{-3}	*(4.9)*	6.7×10^{-2}	*(2.0)*
32	3.7×10^{-4}	*(4.0)*	3.4×10^{-2}	*(2.0)*	3.7×10^{-4}	*(4.0)*	3.4×10^{-2}	*(2.0)*
64	9.2×10^{-5}	*(4.0)*	1.7×10^{-2}	*(2.0)*	9.2×10^{-5}	*(4.0)*	1.7×10^{-2}	*(2.0)*
128	2.3×10^{-5}	*(4.0)*	8.3×10^{-3}	*(2.0)*	2.3×10^{-5}	*(4.0)*	8.3×10^{-3}	*(2.0)*
256	5.5×10^{-6}	*(4.1)*	4.1×10^{-3}	*(2.0)*	5.5×10^{-6}	*(4.1)*	4.1×10^{-3}	*(2.0)*
Function: F_n								
2	3.0×10^{-3}		1.0		1.3×10^{-2}		1.0	
4	2.2×10^{-3}	*(1.4)*	1.0	*(1.0)*	9.8×10^{-3}	*(1.4)*	1.0	*(1.0)*
8	1.5×10^{-3}	*(1.4)*	1.0	*(1.0)*	7.1×10^{-3}	*(1.4)*	1.0	*(1.0)*
16	1.1×10^{-3}	*(1.4)*	9.8×10^{-1}	*(1.0)*	5.1×10^{-3}	*(1.4)*	1.0	*(1.0)*
32	7.5×10^{-4}	*(1.4)*	9.6×10^{-1}	*(1.0)*	3.6×10^{-3}	*(1.4)*	1.1	*(1.0)*
64	5.1×10^{-4}	*(1.5)*	9.3×10^{-1}	*(1.0)*	2.6×10^{-3}	*(1.4)*	1.1	*(0.9)*
128	3.3×10^{-4}	*(1.6)*	8.7×10^{-1}	*(1.1)*	1.8×10^{-3}	*(1.4)*	1.3	*(0.9)*
256	1.9×10^{-4}	*(1.7)*	7.6×10^{-1}	*(1.1)*	1.3×10^{-3}	*(1.4)*	1.7	*(0.8)*
Function: F_p								
2	2.0×10^{-3}		1.0		1.0×10^{-2}		1.0	
4	1.6×10^{-3}	*(1.3)*	1.0	*(1.0)*	7.6×10^{-3}	*(1.3)*	1.0	*(1.0)*
8	1.2×10^{-3}	*(1.3)*	1.0	*(1.0)*	5.6×10^{-3}	*(1.4)*	1.0	*(1.0)*
16	8.5×10^{-4}	*(1.4)*	9.8×10^{-1}	*(1.0)*	4.1×10^{-3}	*(1.4)*	1.0	*(1.0)*
32	6.0×10^{-4}	*(1.4)*	9.6×10^{-1}	*(1.0)*	2.9×10^{-3}	*(1.4)*	1.1	*(1.0)*
64	4.0×10^{-4}	*(1.5)*	9.3×10^{-1}	*(1.0)*	2.1×10^{-3}	*(1.4)*	1.1	*(0.9)*
128	2.6×10^{-4}	*(1.5)*	8.7×10^{-1}	*(1.1)*	1.5×10^{-3}	*(1.4)*	1.3	*(0.9)*
256	1.5×10^{-4}	*(1.7)*	7.6×10^{-1}	*(1.1)*	1.0×10^{-3}	*(1.4)*	1.7	*(0.8)*

Table 3.18: Relative error of the carrier concentrations n and p. Simulation were performed for the p-n diode under 1 V bias.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
Function: n								
2	2.0×10^{-1}		8.3×10^{-1}		2.0×10^{-1}		8.3×10^{-1}	
4	1.5×10^{-1}	*(1.3)*	7.9×10^{-1}	*(1.1)*	1.5×10^{-1}	*(1.3)*	7.9×10^{-1}	*(1.1)*
8	6.7×10^{-2}	*(2.3)*	5.7×10^{-1}	*(1.4)*	6.7×10^{-2}	*(2.3)*	5.7×10^{-1}	*(1.4)*
16	2.0×10^{-2}	*(3.4)*	3.1×10^{-1}	*(1.9)*	2.0×10^{-2}	*(3.4)*	3.1×10^{-1}	*(1.9)*
32	5.0×10^{-3}	*(3.9)*	1.5×10^{-1}	*(2.1)*	5.0×10^{-3}	*(3.9)*	1.5×10^{-1}	*(2.1)*
64	1.3×10^{-3}	*(3.9)*	7.3×10^{-2}	*(2.0)*	1.3×10^{-3}	*(3.9)*	7.3×10^{-2}	*(2.0)*
128	3.2×10^{-4}	*(4.0)*	3.7×10^{-2}	*(2.0)*	3.2×10^{-4}	*(4.0)*	3.7×10^{-2}	*(2.0)*
256	7.6×10^{-5}	*(4.2)*	1.8×10^{-2}	*(2.0)*	7.6×10^{-5}	*(4.2)*	1.8×10^{-2}	*(2.0)*
Function: p								
2	9.7×10^{-2}		7.4×10^{-1}		9.8×10^{-2}		7.5×10^{-1}	
4	4.9×10^{-2}	*(2.0)*	5.8×10^{-1}	*(1.3)*	4.9×10^{-2}	*(2.0)*	5.8×10^{-1}	*(1.3)*
8	1.8×10^{-2}	*(2.7)*	3.2×10^{-1}	*(1.8)*	1.8×10^{-2}	*(2.7)*	3.2×10^{-1}	*(1.8)*
16	5.2×10^{-3}	*(3.5)*	1.5×10^{-1}	*(2.2)*	5.2×10^{-3}	*(3.5)*	1.5×10^{-1}	*(2.2)*
32	1.4×10^{-3}	*(3.7)*	6.8×10^{-2}	*(2.2)*	1.4×10^{-3}	*(3.7)*	6.8×10^{-2}	*(2.2)*
64	3.6×10^{-4}	*(3.9)*	3.3×10^{-2}	*(2.1)*	3.6×10^{-4}	*(3.9)*	3.3×10^{-2}	*(2.1)*
128	8.9×10^{-5}	*(4.0)*	1.6×10^{-2}	*(2.0)*	8.9×10^{-5}	*(4.0)*	1.6×10^{-2}	*(2.0)*
256	2.1×10^{-5}	*(4.2)*	7.9×10^{-3}	*(2.1)*	2.1×10^{-5}	*(4.2)*	7.9×10^{-3}	*(2.1)*

convergence is much slower. For CWOPSIP, we observe $L_2(\Omega)$ convergence only. Having in mind that we do not have exact solution, it is hard to determine whether there is any $H^1(\Omega)$ convergence at all in any case. This behavior is analogous to slower convergence of v, w approximations reported in section 3.2.2. Therefore we also present convergence results for derived functions n, p. For these functions we observe that the convergence is linear in $\|\cdot\|_{H^1(\Omega)}$ and quadratic in $\|\cdot\|_{L_2(\Omega)}$. Errors are similar for all the methods taken into account. This observation also explains, how could convergence of ψ be as good as in equilibrium case while F_n, F_p convergence rates is much worse, as generally the drift-diffusion equations' coefficients and right hand sides are directly dependent on n, p, not on F_n, F_p. This effect is consistent with analogous behavior for formulation u, v, w presented in section 3.2.2.

In second approach we proceed to more complex device - blue InGaN laser. The structure used in this simulation (table 3.15) is simplified a little in comparison with the real laser structure, but it resembles its essential features: a GaN base, AlGaN claddings, an InGaN quantum well and an electron blocking layer.

The results are presented in table 3.19. Generally they agree with the conclusions drawn before, i.e. quadratic $L_2(\Omega)$ convergence and linear $H^1(\Omega)$ convergence of ψ, n, p, but the methods start slower (from K above 64). Errors of both discretizations are similar for a given K.

Therefore we conclude that it is possible to achieve L_2-convergence of the electrostatic potential and carrier concentrations is quadratic and H^1-convergence linear. We also observed sublinear L_2-convergence for the quasi-Fermi levels, while the H^1-convergence is very slow and hard to estimate

Table 3.19: Relative error of the potential ψ, n and p. Simulation were performed for the laser under 2 V bias.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
Function: ψ								
16	1.2×10^{-3}		1.1×10^{-1}		1.4×10^{-3}		1.1×10^{-1}	
32	4.5×10^{-4}	*(2.7)*	5.8×10^{-2}	*(1.9)*	4.5×10^{-4}	*(3.1)*	5.8×10^{-2}	*(1.9)*
64	1.4×10^{-4}	*(3.2)*	2.9×10^{-2}	*(2.0)*	1.4×10^{-4}	*(3.2)*	2.9×10^{-2}	*(2.0)*
128	3.9×10^{-5}	*(3.6)*	1.4×10^{-2}	*(2.0)*	3.9×10^{-5}	*(3.6)*	1.4×10^{-2}	*(2.0)*
256	9.6×10^{-6}	*(4.1)*	7.0×10^{-3}	*(2.1)*	9.6×10^{-6}	*(4.1)*	7.0×10^{-3}	*(2.1)*
Function: n								
16	6.8×10^{-2}		6.1×10^{-1}		6.6×10^{-2}		6.0×10^{-1}	
32	2.9×10^{-2}	*(2.3)*	4.4×10^{-1}	*(1.4)*	2.9×10^{-2}	*(2.2)*	4.4×10^{-1}	*(1.4)*
64	1.0×10^{-2}	*(2.9)*	2.7×10^{-1}	*(1.6)*	1.0×10^{-2}	*(2.9)*	2.7×10^{-1}	*(1.6)*
128	2.8×10^{-3}	*(3.5)*	1.4×10^{-1}	*(1.9)*	2.8×10^{-3}	*(3.5)*	1.4×10^{-1}	*(1.9)*
256	7.3×10^{-4}	*(3.9)*	7.2×10^{-2}	*(2.0)*	7.3×10^{-4}	*(3.9)*	7.2×10^{-2}	*(2.0)*
Function: p								
16	1.1×10^{-3}		1.8×10^{-1}		1.2×10^{-3}		1.8×10^{-1}	
32	3.1×10^{-4}	*(3.7)*	8.8×10^{-2}	*(2.0)*	3.0×10^{-4}	*(3.9)*	8.8×10^{-2}	*(2.0)*
64	7.8×10^{-5}	*(4.0)*	4.4×10^{-2}	*(2.0)*	7.7×10^{-5}	*(4.0)*	4.4×10^{-2}	*(2.0)*
128	1.9×10^{-5}	*(4.0)*	2.2×10^{-2}	*(2.0)*	1.9×10^{-5}	*(4.0)*	2.2×10^{-2}	*(2.0)*
256	4.7×10^{-6}	*(4.1)*	1.1×10^{-2}	*(2.0)*	4.6×10^{-6}	*(4.1)*	1.1×10^{-2}	*(2.0)*

without a closed-form solution.

3.2.4 Formulation ψ, F_n, F_p: two dimensions

Simulations presented in sections 3.2.2 and 3.2.3 shown that the methods CSIPG and CWOPSIP are capable of performing simulations with van Roosbroeck equations. In this section, we would like to conclude convergence testing with two-dimensional simulations of realistic semiconductor devices.

We start with a simulation of a p-n diode (figure 3.13). We use uniform mesh inside layers of the device, which is nonmatching on the interface of the layers. It is formed by division of the layers to K parts in the horizontal direction and then by dividing the first layer into $3K$ parts and dividing the second part into $2K$ parts in the perpendicular direction. We start with initial grid for $K = 2$, presented in figure 3.14. Since the exact solution is not known, as a reference we use the result of one-dimensional simulation for $K = 1024$.

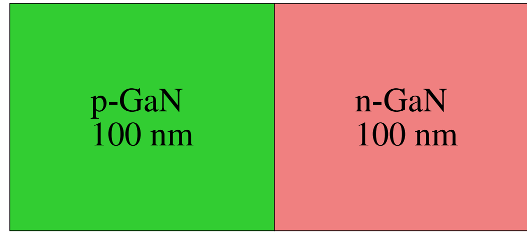

Figure 3.13: Schema of the n-p diode used in numerical simulations.

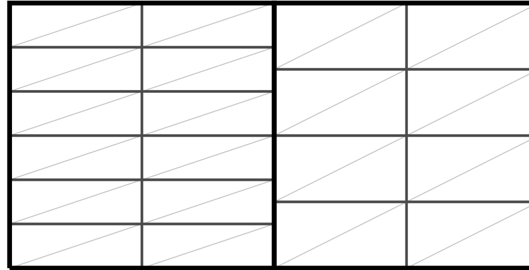

Figure 3.14: Initial grid used in numerical simulations of p-n diode.

Table 3.20: Relative errors of the potential ψ. Simulation were performed for the p-n diode in the equilibrium state, in two dimensions.

	CSIPG		CWOPSIP	
K	$L_2(\Omega)$	$H^1(\Omega)$	$L_2(\Omega)$	$H^1(\Omega)$
2	1.8×10^{-1}	6.9×10^{-1}	2.3×10^{-1}	9.8×10^{-1}
4	4.5×10^{-2} *(4.0)*	4.2×10^{-1} *(1.6)*	1.2×10^{-1} *(1.9)*	5.9×10^{-1} *(1.7)*
8	2.3×10^{-2} *(1.9)*	3.0×10^{-1} *(1.4)*	3.8×10^{-2} *(3.1)*	3.2×10^{-1} *(1.9)*
16	9.7×10^{-3} *(2.4)*	2.1×10^{-1} *(1.4)*	7.7×10^{-3} *(4.9)*	2.1×10^{-1} *(1.5)*
32	2.8×10^{-3} *(3.4)*	1.2×10^{-1} *(1.8)*	1.5×10^{-3} *(5.1)*	1.2×10^{-1} *(1.8)*
64	6.1×10^{-4} *(4.6)*	5.7×10^{-2} *(2.0)*	4.8×10^{-4} *(3.1)*	5.7×10^{-2} *(2.0)*
128	1.5×10^{-4} *(4.2)*	2.9×10^{-2} *(2.0)*	1.3×10^{-4} *(3.8)*	2.9×10^{-2} *(2.0)*
256	3.5×10^{-5} *(4.2)*	1.4×10^{-2} *(2.0)*	3.3×10^{-5} *(3.9)*	1.4×10^{-2} *(2.0)*

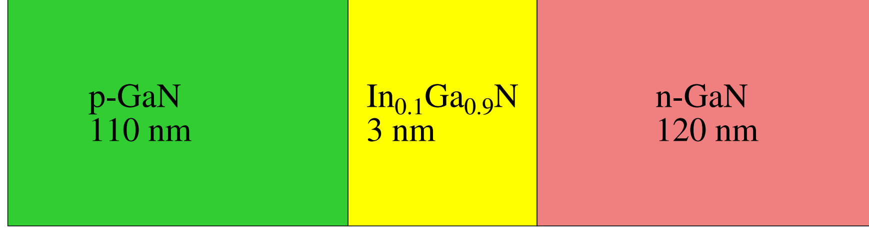

Figure 3.15: Schema of the quantum well structure used in numerical simulations.

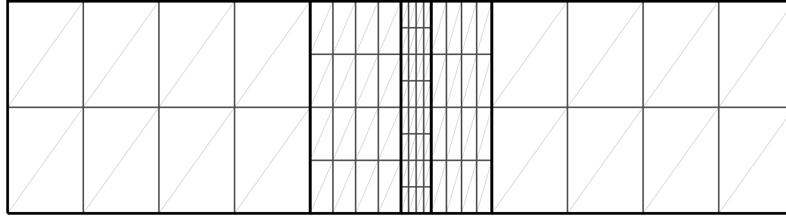

Figure 3.16: Initial grid used in numerical simulations of the quantum well. To improve the convergence, p-GaN and n-GaN were splitted into two layers to thicken the grid near the quantum well.

Table 3.21: Relative errors of the potential ψ. Simulation were performed for the single quantum well in the equilibrium state, in two dimensions.

	CSIPG				CWOPSIP			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
2	5.2×10^{-2}		3.8×10^{-1}		7.5×10^{-2}		4.6×10^{-1}	
4	2.0×10^{-2}	*(2.6)*	2.3×10^{-1}	*(1.7)*	2.4×10^{-2}	*(3.1)*	2.6×10^{-1}	*(1.8)*
8	5.2×10^{-3}	*(3.8)*	1.1×10^{-1}	*(2.0)*	7.3×10^{-3}	*(3.3)*	1.2×10^{-1}	*(2.2)*
16	1.3×10^{-3}	*(3.9)*	5.8×10^{-2}	*(1.9)*	1.8×10^{-3}	*(4.0)*	5.9×10^{-2}	*(2.0)*
32	3.4×10^{-4}	*(4.0)*	2.9×10^{-2}	*(2.0)*	4.5×10^{-4}	*(4.0)*	3.0×10^{-2}	*(2.0)*
64	8.5×10^{-5}	*(4.0)*	1.5×10^{-2}	*(2.0)*	1.1×10^{-4}	*(4.0)*	1.5×10^{-2}	*(2.0)*

Results of these simulations are presented in table 3.20. Both methods start slowly, for $K \leq 32$ error rate is smaller than 2. From $K = 64$ we observe convergence rate of approximately 2. The error norms for CSIPG and CWOPSIP are similar.

Then we repeated our simulations for a single quantum well structure (figure 3.15). It is similar to the p-n diode, but it has a narrow layer between n-type region and p-type region, the quantum well (see section 2.4). In this case, we introduce five layers in our mesh, while two additional layers are used to improve the grid in the n-type region and p-type region near the quantum well (figure 3.16).

Results of this simulation is presented in table 3.21 and they generally agree with our previous simulations. Note that the slow start is absent in this case. This is due to additional layers introduced near the quantum well, where the function changes are crucial.

3.3 Discontinuities on interfaces

In this section we would like to generalize discretizations presented in section 1.3 to account for discontinuities of unknown functions on interfaces. This generalization is necessary to introduce interfacial charges. These charges emerge for example due to polarization effect, present in GaN-based heterostructures.

Since we discuss the Discontinuous Galerkin Method, the discrete space accounts for discontinuous functions, and we can introduce discontinuities in a natural way using techniques already established for imposing continuity.

In this section, we derive one-dimensional discrete problems and we present numerical experiments of one-dimensional model. However, these problems can be extended to two dimensional domains in a natural way.

3.3.1 Differential problem

3.3.1.1 Classic formulation

We would like to focus on the one-dimensional problem. Let Ω denote an open interval. We define nodes $\{d_i\}_{i=0}^{N}$ such that $d_0 \leq d_1 \leq \ldots \leq d_N$ and $\Omega = (d_0, d_N)$, and subsets $\Omega_i := (d_{i-1}, d_i)$. Note that $\overline{\Omega} = \overline{\Omega}_1 \cup \ldots \cup \overline{\Omega}_N$.

Consider the following differential problem

$$\begin{cases} -\dfrac{d}{dx}\Big(\rho_i(x)\dfrac{d}{dx}u_i(x)\Big) = f(x) \qquad \forall x \in \Omega_i, \\ u_1(d_0) = \bar{u}_0, \\ u_N(d_N) = \bar{u}_N, \\ u_{i+1}(d_i) - u_i(d_i) = \bar{u}_i \qquad i \in \{1, \ldots, N-1\}, \\ \rho_{i+1}(d_i)\dfrac{du_{i+1}}{dx}(d_i) - \rho_i(d_i)\dfrac{du_i}{dx}(d_i) = r_i \qquad i \in \{1, \ldots, N-1\}, \end{cases} \tag{3.3.1}$$

where $f \in \mathcal{C}(\overline{\Omega})$, $\rho_i \in \mathcal{C}^1(\overline{\Omega_i})$ and $u \in \mathcal{C}^2(\overline{\Omega})$. We may interpret $u = (u_1, \ldots, u_N)$ as a piecewise continuous function with undefined values at nodes. Then $\bar{u}_i$ for $i \in \{0, N\}$ can be identified with boundary values and for $i \in \{1, \ldots, N-1\}$ with jumps of the function at the nodal points.

3.3.1.2 Derivation of weak formulation

We define spaces

$$\begin{aligned} \mathcal{B} &= \Big\{(v_1, \ldots, v_n) : v_i \in \mathcal{C}^1(\overline{\Omega}_i), i \in \{1, \ldots, N\}\Big\}, \\ \mathcal{B}_0 &= \Big\{v \in B : v_1(d_0) = 0 = v_N(d_N)\Big\}. \end{aligned} \tag{3.3.2}$$

Let $v \in \mathcal{B}$. We then identify v with a function $v : \Omega \backslash \{d_0, \ldots, d_N\} \mapsto \mathbb{R}$ defined as

$$v(x) = \begin{cases} v_i(x) & x \in \Omega_i, \\ \text{undefined} & x \in \{d_0, \ldots, d_N\}. \end{cases} \tag{3.3.3}$$

Note that values $v(d_i)$ are not defined, but the right and left limit is well-posed

$$v(d_i^-) := \lim_{x \to d_i^-} v(x) = v_i(d_i), \quad v(d_i^+) := \lim_{x \to d_i^+} v(x) = v_{i+1}(d_i). \tag{3.3.4}$$

Now we would like to derive a weak formulation of the problem (3.3.1). For every Ω_i we have

$$-\frac{d}{dx}\Big(\rho_i(x)\frac{d}{dx}u_i(x)\Big) = f(x) \qquad \forall x \in \Omega_i. \tag{3.3.5}$$

We multiply the above equation by $v \in B$ and integrate it on Ω_i

$$-\int_{\Omega_i} \frac{d}{dx}\Big(\rho_i(x)\frac{d}{dx}u_i(x)\Big)v(x)dx = \int_{\Omega_i} f(x)v(x)dx. \tag{3.3.6}$$

Then using the integration by parts

$$\int_{\Omega_i} \rho_i(x)\frac{du_i}{dx}(x)\frac{dv}{dx}(x)dx - \rho_i(d_i)\frac{du_i}{dx}(d_i)v_i(d_i) + \rho_i(d_{i-1})\frac{du_i}{dx}(d_{i-1})v_i(d_{i-1}) = \int_{\Omega_i} f(x)v(x)dx. \tag{3.3.7}$$

If we add left hand sides of the equations (3.3.7) we obtain

$$\begin{aligned}\sum_{i=1}^{N}\int_{\Omega_i} \rho_i(x)\frac{du_i}{dx}(x)\frac{dv}{dx}(x)dx + \sum_{i=1}^{N}\Big(- \rho_i(d_i)\frac{du_i}{dx}(d_i)v_i(d_i) + \rho_i(d_{i-1})\frac{du_i}{dx}(d_{i-1})v_i(d_{i-1})\Big)\\ = \int_{\Omega} f(x)v(x)dx.\end{aligned} \tag{3.3.8}$$

After reordering of the components we obtain a new equation

$$\begin{aligned}\sum_{i=1}^{N}\int_{\Omega_i} \rho_i(x)\frac{du_i}{dx}(x)\frac{dv}{dx}(x)dx + \rho_1(d_0)\frac{du_1}{dx}(d_0)v_1(d_0) - \rho_N(d_N)\frac{du_N}{dx}(d_N)v_N(d_N)\\ + \sum_{i=1}^{N-1}\Big(\rho_{i+1}(d_i)\frac{du_{i+1}}{dx}(d_i)v_{i+1}(d_i) - \rho_i(d_i)\frac{du_i}{dx}(d_i)v_i(d_i)\Big)\\ = \int_{\Omega} f(x)v(x)dx.\end{aligned} \tag{3.3.9}$$

We would like to use the internal boundary values to modify the components

$$\rho_{i+1}(d_i)\frac{du_{i+1}}{dx}(d_i)v_{i+1}(d_i) - \rho_i(d_i)\frac{du_i}{dx}(d_i)v_i(d_i). \tag{3.3.10}$$

From now on, to improve readability, we would extract the argument out of expressions if it is the same for all functions. Thus the internal boundary condition from the problem (3.3.1) reads

$$\Big(\rho_{i+1}\frac{du_{i+1}}{dx} - \rho_i\frac{du_i}{dx}\Big)(d_i) = r_i, \quad i \in \{1,\dots,N-1\}. \tag{3.3.11}$$

Therefore

$$\Big(\rho_i\frac{du_i}{dx}\Big)(d_i) = -r_i + \Big(\rho_{i+1}\frac{du_{i+1}}{dx}\Big)(d_i). \tag{3.3.12}$$

We may then use these identities and rewrite the equation (3.3.10) as

$$\begin{aligned}&\Big(\rho_{i+1}\frac{du_{i+1}}{dx}v_{i+1} - \rho_i\frac{du_i}{dx}v_i\Big)(d_i) = \frac{1}{2}\Big(\big[2\rho_{i+1}\frac{du_{i+1}}{dx}\big]v_{i+1} - \big[2\rho_i\frac{du_i}{dx}\big]v_i\Big)(d_i) =\\ &= \frac{1}{2}\Big(\big[\rho_{i+1}\frac{du_{i+1}}{dx} + r_i + \rho_i\frac{du_i}{dx}\big]v_{i+1} - \big[\rho_i\frac{du_i}{dx} - r_i + \rho_{i+1}\frac{du_{i+1}}{dx}\big]v_i\Big)(d_i)\\ &= \frac{1}{2}\Big(\underbrace{\rho_{i+1}\frac{du_{i+1}}{dx}v_{i+1}}_{1} + \underbrace{r_iv_{i+1}}_{3} + \underbrace{\rho_i\frac{du_i}{dx}v_{i+1}}_{2} - \underbrace{\rho_i\frac{du_i}{dx}v_i}_{2} + \underbrace{r_iv_i}_{3} - \underbrace{\rho_{i+1}\frac{du_{i+1}}{dx}v_i}_{1}\Big)(d_i).\end{aligned} \tag{3.3.13}$$

Therefore we obtain

$$\begin{aligned}\Big(\rho_{i+1}\frac{du_{i+1}}{dx}v_{i+1}-\rho_i\frac{du_i}{dx}v_i\Big)(d_i)&=\frac{1}{2}\Big(\rho_{i+1}\frac{du_{i+1}}{dx}(v_{i+1}-v_i)+\rho_i\frac{du_i}{dx}(v_{i+1}-v_i)+r_i(v_{i+1}+v_i)\Big)(d_i)\\&=\frac{1}{2}\Big(\Big[\rho_{i+1}\frac{du_{i+1}}{dx}+\rho_i\frac{du_i}{dx}\Big](v_{i+1}-v_i)+r_i(v_{i+1}+v_i)\Big)(d_i).\end{aligned}\tag{3.3.14}$$

Then we may reformulate equation (3.3.9) to obtain

$$\begin{aligned}\sum_{i=1}^{N}\int_{\Omega_i}\rho_i(x)\frac{du_i}{dx}(x)\frac{dv}{dx}(x)dx+\Big(\rho_1\frac{du_1}{dx}v_1\Big)(d_0)-\Big(\rho_N\frac{du_N}{dx}v_N\Big)(d_N)&\\+\frac{1}{2}\sum_{i=1}^{N-1}\Big(\Big[\rho_{i+1}\frac{du_{i+1}}{dx}+\rho_i\frac{du_i}{dx}\Big](v_{i+1}-v_i)+r_i(v_{i+1}+v_i)\Big)(d_i)&\\=\int_{\Omega}f(x)v(x)dx.&\end{aligned}\tag{3.3.15}$$

Moving the second component under the sum on the right hand side we obtain the equation

$$\begin{aligned}\sum_{i=1}^{N}\int_{\Omega_i}\rho_i(x)\frac{du_i}{dx}(x)\frac{dv}{dx}(x)dx+\Big(\rho_1\frac{du_1}{dx}v_1\Big)(d_0)-\Big(\rho_N\frac{du_N}{dx}v_N\Big)(d_N)&\\+\frac{1}{2}\sum_{i=1}^{N-1}\Big(\Big[\rho_{i+1}\frac{du_{i+1}}{dx}+\rho_i\frac{du_i}{dx}\Big](v_{i+1}-v_i)\Big)(d_i)&\\=\int_{\Omega}f(x)v(x)dx-\frac{1}{2}\sum_{i=1}^{N-1}r_i\Big(v_{i+1}+v_i\Big)(d_i).&\end{aligned}\tag{3.3.16}$$

We define

$$A(u,v):=\sum_{i=1}^{N}A_i(u,v),\quad A_i(u,v):=\int_{\Omega_i}\rho_i(x)\frac{du_i}{dx}(x)\frac{dv}{dx}(x)dx,\tag{3.3.17}$$

$$B(v):=\sum_{i=1}^{N}B_i(v),\qquad B_i(v):=\int_{\Omega_i}f(x)v_i(x)dx,\tag{3.3.18}$$

$$C(u,v):=\sum_{i=0}^{N}C_i(u,v),\quad C_i(u,v):=\begin{cases}\Big(\rho_1\frac{du_1}{dx}v_1\Big)(d_0), & i=0,\\ \frac{1}{2}\Big(\Big[\rho_{i+1}\frac{du_{i+1}}{dx}+\rho_i\frac{du_i}{dx}\Big](v_{i+1}-v_i)\Big)(d_i), & 0<i<N,\\ -\Big(\rho_N\frac{du_N}{dx}v_N\Big)(d_N), & i=N,\end{cases}\tag{3.3.19}$$

$$D(v):=\sum_{i=1}^{N-1}D_i(v),\qquad D_i(v):=-\frac{1}{2}r_i\Big(v_{i+1}+v_i\Big)(d_i).\tag{3.3.20}$$

3.3.1.3 Weak formulation

The derivation from the section 3.3.1.2 suggests the following weak formulation of the problem (3.3.1). Let us define spaces $\mathcal{D}$, $\mathcal{D}_0$ as

$$\begin{aligned}\mathcal{D}&=H^1(\mathcal{E}),\\ \mathcal{D}_0&=\Big\{v\in\mathcal{D}:v_1(d_0)=0=v_N(d_N)\Big\}.\end{aligned}\tag{3.3.21}$$

Then the weak problem is to find $u \in \mathcal{D}$ such that

$$\begin{aligned} & A(u,v) + C(u,v) = B(v) + D(v) \quad \forall v \in \mathcal{D}_0, \\ & u_1(d_0) = \bar{u}_0, \\ & u_{i+1}(d_i) - u_i(d_i) = \bar{u}_i \qquad i \in \{1, \dots, N-1\}, \\ & u_N(d_N) = \bar{u}_N. \end{aligned} \tag{3.3.22}$$

Thus in this formulation the discontinuities of the flux are imposed weakly, while the discontinuities of the unknown function are still imposed strongly. Proceeding to discrete problems, we must take care of the latter.

3.3.2 Discrete problems

We would derive discrete problems gradually. We start with the Composite Incomplete Interior Penalty Galerkin method (CIIPG), based on Incomplete Interior Penalty Galerkin method [94], which is derived from the weak problem (3.3.22) by adding the penalty terms. Then we proceed to the CSIPG method by symmetrizing CIIPG, and then to CWOPSIP.

3.3.2.1 CIIPG

We would like to transform the weak problem into most straightforward Discontinuous Galerkin formulation. In the weak form (3.3.22) the flux conditions are already present in the operators C and D, but still we must impose the internal and external Dirichlet boundary conditions. To do so, we will introduce the penalty term. Let $\sigma_i > 0$ be a parameter, called the penalty parameter for Dirichlet boundary conditions. We define operators $G_i : (X_h(\Omega))^2 \mapsto \mathbb{R}$, $H_i : X_h(\Omega) \to \mathbb{R}$ as

$$G_i(u,v) = \begin{cases} \sigma_1 \frac{\rho_1(d_0)}{h_1} u_1(d_0) v_1(d_0), & i = 0, \\ \sigma_i \Big(\frac{\rho_i(d_i)}{h_i} + \frac{\rho_{i+1}(d_i)}{h_{i+1}} \Big) \Big(u_{i+1}(d_i) - u_i(d_i) \Big) \Big(v_{i+1}(d_i) - v_i(d_i) \Big), & i = 1, \dots, N-1, \\ \sigma_N \frac{\rho_N(d_N)}{h_N} u_N(d_N) v_N(d_N), & i = N, \end{cases} \tag{3.3.23}$$

$$H_i(v) = \begin{cases} \sigma_1 \frac{\rho_1(d_0)}{h_1} \bar{u}_0 v_1(d_0), & i = 0, \\ \sigma_i \Big(\frac{\rho_i(d_i)}{h_i} + \frac{\rho_{i+1}(d_i)}{h_{i+1}} \Big) \bar{u}_i \Big(v_{i+1}(d_i) - v_i(d_i) \Big), & i = 1, \dots, N-1, \\ \sigma_N \frac{\rho_N(d_N)}{h_N} \bar{u}_N v_N(d_N), & i = N. \end{cases} \tag{3.3.24}$$

for any $u, v \in X_h$.

Then we define the operators $G : (X_h)^2 \to \mathbb{R}$, $H : X_h \to \mathbb{R}$ as

$$G(u,v) := \sum_{i=0}^{N} G_i(u,v), \qquad H(v) := \sum_{i=0}^{N} H_i(v). \tag{3.3.25}$$

Then the plain formulation would be posed as follows. Find $u \in X_h$ such that

$$\forall v \in X_h \quad b_{\mathrm{CIIPG}}(u,v) = f_{\mathrm{CIIPG}}(v), \tag{3.3.26}$$

where

$$b_{\mathrm{CIIPG}}(u,v) := A(u,v) + C(u,v) + G(u,v), \tag{3.3.27}$$

$$f_{\mathrm{CIIPG}}(v) := B(v) + D(v) + H(v). \tag{3.3.28}$$

3.3.2.2 CSIPG

In this section we would like to use the approach similar to [37]. As a starting point we take the operator b_{CIIPG}, which we want to symmetrize. According to the definition (3.3.27)

$$b_{\mathrm{CIIPG}}(u,v) = A(u,v) + C(u,v) + G(u,v). \tag{3.3.29}$$

Note that A and G are already symmetric, so to get a symmetric operator, we add $C(v,u)$. By the definition (3.3.19) we have

$$\begin{aligned} L(u,v) := C(v,u) &= \Big(\rho_1 \frac{dv_1}{dx} u_1\Big)(d_0) - \Big(\rho_N \frac{dv_N}{dx} u_N\Big)(d_N) \\ &+ \sum_{i=1}^{N-1} \frac{1}{2}\Big(\Big[\rho_{i+1}\frac{dv_{i+1}}{dx} + \rho_i \frac{dv_i}{dx}\Big](u_{i+1} - u_i)\Big)(d_i) = \sum_{i=0}^{N} L_i(u,v), \end{aligned} \tag{3.3.30}$$

where

$$L_i(u,v) := C_i(v,u) = \begin{cases} \Big(\rho_1 \frac{dv_1}{dx} u_1\Big)(d_0), & i = 0, \\ \frac{1}{2}\Big(\Big[\rho_{i+1}\frac{dv_{i+1}}{dx} + \rho_i \frac{dv_i}{dx}\Big](u_{i+1} - u_i)\Big)(d_i), & i \in \{1,\dots,N-1\}, \\ -\Big(\rho_N \frac{dv_N}{dx} u_N\Big)(d_N), & i = N, \end{cases} \tag{3.3.31}$$

for $i \in \{0,\dots,N\}$. Then we have to introduce an additional expression to f_{CIIPG}. Referring to the internal Dirichlet conditions (3.3.1) we define

$$M(v) := \sum_{i=1}^{N-1} \frac{1}{2}\bar{u}_i \Big[\rho_{i+1}\frac{dv_{i+1}}{dx} + \rho_i \frac{dv_i}{dx}\Big](d_i), \tag{3.3.32}$$

Finally, we can write the symmetric problem as

$$b_{\mathrm{CSIPG}}(u,v) := f_{\mathrm{CSIPG}}(v), \tag{3.3.33}$$

where

$$b_{\mathrm{CSIPG}}(u,v) := A(u,v) + C(u,v) + G(u,v) + L(u,v), \tag{3.3.34}$$

$$f_{\mathrm{CSIPG}}(v) := B(v) + D(v) + H(v) + M(v). \tag{3.3.35}$$

3.3.2.3 CWOPSIP

For the CWOPSIP problem, we need over-penalized operators. We call them O, P and they are defined as

$$O(u,v) := \sum_{i=0}^{N} O_i(u,v), \tag{3.3.36}$$

$$P(v) := \sum_{i=0}^{N} P_i(v), \tag{3.3.37}$$

where

$$O_i(u,v) = \begin{cases} \delta \frac{\rho_1(d_0)}{h_1^2} u_1(d_0) v_1(d_0), & i = 0, \\ \delta \Big(\frac{\rho_i(d_i)}{h_i^2} + \frac{\rho_{i+1}(d_i)}{h_{i+1}^2}\Big)\Big(u_{i+1}(d_i) - u_i(d_i)\Big)\Big(v_{i+1}(d_i) - v_i(d_i)\Big), & i = 1,\dots,N-1, \\ \delta \frac{\rho_N(d_N)}{h_N^2} u_N(d_N) v_N(d_N), & i = N, \end{cases} \tag{3.3.38}$$

$$P_i(v) = \begin{cases} \delta \frac{\rho_1(d_0)}{h_1^2} \bar{u}_0 v_1(d_0), & i = 0, \\ \delta \Big(\frac{\rho_i(d_i)}{h_i^2} + \frac{\rho_{i+1}(d_i)}{h_{i+1}^2} \Big) \bar{u}_i \Big(v_{i+1}(d_i) - v_i(d_i) \Big), & i = 1, \dots, N-1, \\ \delta \frac{\rho_N(d_N)}{h_N^2} \bar{u}_N v_N(d_N), & i = N. \end{cases} \tag{3.3.39}$$

As before we start with CIIPG problem: find $u \in X_h(\Omega)$ such that

$$\forall v \in X_h(\Omega) \quad b_{\text{CIIPG}}(u, v) = f_{\text{CIIPG}}(v). \tag{3.3.40}$$

We have thus

$$A(u, v) + C(u, v) + G(u, v) = B(v) + D(v) + H(v). \tag{3.3.41}$$

First we ignore operator C, as in CWOPSIP method introduced in section 1.3.2.2 and we replace the penalty operators G, H with over-penalized operators O, P. We have

$$A(u, v) + O(u, v) = B(v) + D(v) + P(v). \tag{3.3.42}$$

At the first glance it seems natural to remove also the operator D from the above equation, as it is a right-hand-side counterpart of operator C. However, this is the only element imposing boundary conditions on the flux. Also, while lack of the operator C is balanced by increased penalty, this is not the case for the operator D.

Therefore we define CWOPSIP problem: find $u \in X_h$ such that

$$\forall v \in X_h \quad b_{\text{CWOPSIP}}(u, v) = f_{\text{CWOPSIP}}(v), \tag{3.3.43}$$

where

$$b_{\text{CWOPSIP}}(u, v) := A(u, v) + O(u, v), \quad f_{\text{CWOPSIP}}(v) := B(v) + D(v) + P(v). \tag{3.3.44}$$

3.3.3 Simulations

3.3.3.1 Examples

To test the formulations presented in the section 3.3.2, we have performed several simulations.

We used arbitrarily chosen set of five examples, where $\rho|_{\Omega_i}, f|_{\Omega_i} \in \mathcal{C}^\infty(\overline{\Omega_i})$, but in general $\rho, f \notin \mathcal{C}^0(\overline{\Omega})$.

Example 1 is a reference, as the unknown function and fluxes are continuous at interfaces. In example 3, fluxes are continuous, while the unknown function is not. In other examples, both unknown functions as well as fluxes are discontinuous at interfaces. Examples 3, 4, 5 are chosen such that the differential problem is nonlinear, with nonlinearities both in the coefficient ρ as well as in the right hand side f. For comparison, we use these examples also for linear simulations with nonlinearities in ρ, f substituted by known solutions.

3.3.3.2 Linear equations

The results of the computations for the linear case for the respective discrete formulations are presented in tables 3.22, 3.23,3.24, for CIIPG, CSIPG and CWOPSIP, respectively. Except of the isolated cases, error values are of similar order. We generally observe linear convergence of H^1-error and quadratic convergence in L_2-error. Examples 2 and 4 exhibit quadratic convergence also for H^1-error, but this is due to L_2-norm dominating H^1-seminorm.

	1				2			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
3	5.0×10^{-2}		2.7×10^{-1}		3.1×10^{-1}		3.1×10^{-1}	
6	1.0×10^{-2}	*(4.9)*	1.2×10^{-1}	*(2.2)*	7.9×10^{-2}	*(3.9)*	8.0×10^{-2}	*(3.9)*
12	2.3×10^{-3}	*(4.5)*	6.1×10^{-2}	*(2.0)*	2.0×10^{-2}	*(3.9)*	2.0×10^{-2}	*(3.9)*
24	5.3×10^{-4}	*(4.3)*	3.0×10^{-2}	*(2.0)*	5.1×10^{-3}	*(4.0)*	5.1×10^{-3}	*(4.0)*
48	1.3×10^{-4}	*(4.1)*	1.5×10^{-2}	*(2.0)*	1.3×10^{-3}	*(4.0)*	1.3×10^{-3}	*(4.0)*
96	3.1×10^{-5}	*(4.1)*	7.6×10^{-3}	*(2.0)*	3.2×10^{-4}	*(4.0)*	3.2×10^{-4}	*(4.0)*
192	7.7×10^{-6}	*(4.0)*	3.8×10^{-3}	*(2.0)*	8.0×10^{-5}	*(4.0)*	8.0×10^{-5}	*(4.0)*
384	1.9×10^{-6}	*(4.0)*	1.9×10^{-3}	*(2.0)*	2.0×10^{-5}	*(4.0)*	2.0×10^{-5}	*(4.0)*

	3				4			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
3	2.9×10^{-1}		1.9		7.0×10^{4}		7.2×10^{4}	
6	6.8×10^{-2}	*(4.3)*	9.4×10^{-1}	*(2.1)*	1.9×10^{4}	*(3.8)*	1.9×10^{4}	*(3.8)*
12	1.7×10^{-2}	*(4.0)*	4.6×10^{-1}	*(2.0)*	4.7×10^{3}	*(3.9)*	4.8×10^{3}	*(3.9)*
24	4.2×10^{-3}	*(4.0)*	2.3×10^{-1}	*(2.0)*	1.2×10^{3}	*(4.0)*	1.2×10^{3}	*(4.0)*
48	1.1×10^{-3}	*(4.0)*	1.2×10^{-1}	*(2.0)*	3.0×10^{2}	*(4.0)*	3.1×10^{2}	*(4.0)*
96	2.7×10^{-4}	*(4.0)*	5.8×10^{-2}	*(2.0)*	7.4×10^{1}	*(4.0)*	7.8×10^{1}	*(3.9)*
192	6.6×10^{-5}	*(4.0)*	2.9×10^{-2}	*(2.0)*	1.9×10^{1}	*(4.0)*	2.1×10^{1}	*(3.7)*
384	1.7×10^{-5}	*(4.0)*	1.4×10^{-2}	*(2.0)*	4.8	*(3.9)*	6.6	*(3.2)*

	5			
K	$L_2(\Omega)$		$H^1(\Omega)$	
3	3.6×10^{-1}		1.5	
6	1.0×10^{-1}	*(3.6)*	4.7×10^{-1}	*(3.1)*
12	2.6×10^{-2}	*(3.9)*	1.5×10^{-1}	*(3.1)*
24	6.6×10^{-3}	*(4.0)*	5.5×10^{-2}	*(2.7)*
48	1.6×10^{-3}	*(4.0)*	2.4×10^{-2}	*(2.3)*
96	4.1×10^{-4}	*(4.0)*	1.2×10^{-2}	*(2.1)*
192	1.0×10^{-4}	*(4.0)*	5.8×10^{-3}	*(2.0)*
384	2.6×10^{-5}	*(4.0)*	2.9×10^{-3}	*(2.0)*

Table 3.22: Absolute errors, approximated orders of the method and computation times for linear equations solved using CIIPG.

	1				2			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
3	5.4×10^{-2}		3.5×10^{-1}		3.3×10^{-1}		3.3×10^{-1}	
6	1.0×10^{-2}	*(5.2)*	1.5×10^{-1}	*(2.4)*	8.2×10^{-2}	*(4.0)*	8.3×10^{-2}	*(4.0)*
12	2.3×10^{-3}	*(4.5)*	6.8×10^{-2}	*(2.2)*	2.0×10^{-2}	*(4.0)*	2.1×10^{-2}	*(4.0)*
24	5.4×10^{-4}	*(4.2)*	3.2×10^{-2}	*(2.1)*	5.1×10^{-3}	*(4.0)*	5.2×10^{-3}	*(4.0)*
48	1.3×10^{-4}	*(4.0)*	1.6×10^{-2}	*(2.1)*	1.3×10^{-3}	*(4.0)*	1.3×10^{-3}	*(4.0)*
96	3.4×10^{-5}	*(4.0)*	7.7×10^{-3}	*(2.0)*	3.2×10^{-4}	*(4.0)*	3.2×10^{-4}	*(4.0)*
192	8.4×10^{-6}	*(4.0)*	3.8×10^{-3}	*(2.0)*	8.0×10^{-5}	*(4.0)*	8.1×10^{-5}	*(4.0)*
384	2.1×10^{-6}	*(4.0)*	1.9×10^{-3}	*(2.0)*	2.0×10^{-5}	*(4.0)*	2.0×10^{-5}	*(4.0)*

	3				4			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
3	7.4×10^{-1}		2.6		7.0×10^{4}		7.2×10^{4}	
6	1.7×10^{-1}	*(4.4)*	1.1	*(2.5)*	1.9×10^{4}	*(3.8)*	1.9×10^{4}	*(3.8)*
12	4.1×10^{-2}	*(4.2)*	4.9×10^{-1}	*(2.2)*	4.7×10^{3}	*(3.9)*	4.8×10^{3}	*(3.9)*
24	1.0×10^{-2}	*(4.1)*	2.4×10^{-1}	*(2.1)*	1.2×10^{3}	*(4.0)*	1.2×10^{3}	*(4.0)*
48	2.5×10^{-3}	*(4.0)*	1.2×10^{-1}	*(2.0)*	3.0×10^{2}	*(4.0)*	3.1×10^{2}	*(4.0)*
96	6.1×10^{-4}	*(4.0)*	5.8×10^{-2}	*(2.0)*	7.4×10^{1}	*(4.0)*	7.8×10^{1}	*(3.9)*
192	1.5×10^{-4}	*(4.0)*	2.9×10^{-2}	*(2.0)*	1.9×10^{1}	*(4.0)*	2.1×10^{1}	*(3.7)*
384	3.8×10^{-5}	*(4.0)*	1.4×10^{-2}	*(2.0)*	4.6	*(4.1)*	6.4	*(3.3)*

	5			
K	$L_2(\Omega)$		$H^1(\Omega)$	
3	3.4×10^{-1}		1.4	
6	9.8×10^{-2}	*(3.5)*	4.5×10^{-1}	*(3.1)*
12	2.5×10^{-2}	*(3.8)*	1.5×10^{-1}	*(3.1)*
24	6.4×10^{-3}	*(4.0)*	5.5×10^{-2}	*(2.7)*
48	1.6×10^{-3}	*(4.0)*	2.4×10^{-2}	*(2.2)*
96	4.0×10^{-4}	*(4.0)*	1.2×10^{-2}	*(2.1)*
192	1.0×10^{-4}	*(4.0)*	5.8×10^{-3}	*(2.0)*
384	2.5×10^{-5}	*(4.0)*	2.9×10^{-3}	*(2.0)*

Table 3.23: Absolute errors, approximated orders of the method and computation times for linear equations solved using CSIPG.

	1				2			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
3	5.9×10^{-2}		3.1×10^{-1}		1.1×10^{1}		1.1×10^{1}	
6	1.5×10^{-2}	*(4.1)*	1.3×10^{-1}	*(2.3)*	2.9	*(4.0)*	2.9	*(4.0)*
12	3.6×10^{-3}	*(4.0)*	6.2×10^{-2}	*(2.1)*	7.2×10^{-1}	*(4.0)*	7.2×10^{-1}	*(4.0)*
24	9.0×10^{-4}	*(4.0)*	3.0×10^{-2}	*(2.0)*	1.8×10^{-1}	*(4.0)*	1.8×10^{-1}	*(4.0)*
48	2.3×10^{-4}	*(4.0)*	1.5×10^{-2}	*(2.0)*	4.5×10^{-2}	*(4.0)*	4.5×10^{-2}	*(4.0)*
96	5.6×10^{-5}	*(4.0)*	7.6×10^{-3}	*(2.0)*	1.1×10^{-2}	*(4.0)*	1.1×10^{-2}	*(4.0)*
192	1.4×10^{-5}	*(4.0)*	3.8×10^{-3}	*(2.0)*	2.8×10^{-3}	*(4.0)*	2.8×10^{-3}	*(4.0)*
384	3.5×10^{-6}	*(4.0)*	1.9×10^{-3}	*(2.0)*	7.0×10^{-4}	*(4.0)*	7.0×10^{-4}	*(4.0)*

	3				4			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
3	1.8		2.9		7.0×10^{4}		7.2×10^{4}	
6	4.6×10^{-1}	*(4.0)*	1.1	*(2.7)*	1.9×10^{4}	*(3.8)*	1.9×10^{4}	*(3.8)*
12	1.1×10^{-1}	*(4.0)*	4.8×10^{-1}	*(2.2)*	4.7×10^{3}	*(3.9)*	4.8×10^{3}	*(3.9)*
24	2.9×10^{-2}	*(4.0)*	2.3×10^{-1}	*(2.1)*	1.2×10^{3}	*(4.0)*	1.2×10^{3}	*(4.0)*
48	7.1×10^{-3}	*(4.0)*	1.2×10^{-1}	*(2.0)*	3.0×10^{2}	*(4.0)*	3.1×10^{2}	*(4.0)*
96	1.8×10^{-3}	*(4.0)*	5.8×10^{-2}	*(2.0)*	7.4×10^{1}	*(4.0)*	7.7×10^{1}	*(3.9)*
192	4.5×10^{-4}	*(4.0)*	2.9×10^{-2}	*(2.0)*	2.0×10^{1}	*(3.7)*	2.2×10^{1}	*(3.5)*
384	1.1×10^{-4}	*(4.0)*	1.4×10^{-2}	*(2.0)*	1.5×10^{1}	*(1.3)*	1.6×10^{1}	*(1.4)*

	5			
K	$L_2(\Omega)$		$H^1(\Omega)$	
3	3.6×10^{-1}		1.4	
6	1.0×10^{-1}	*(3.5)*	4.6×10^{-1}	*(3.1)*
12	2.6×10^{-2}	*(3.8)*	1.5×10^{-1}	*(3.1)*
24	6.6×10^{-3}	*(4.0)*	5.5×10^{-2}	*(2.7)*
48	1.7×10^{-3}	*(4.0)*	2.4×10^{-2}	*(2.2)*
96	4.2×10^{-4}	*(4.0)*	1.2×10^{-2}	*(2.1)*
192	1.0×10^{-4}	*(4.0)*	5.8×10^{-3}	*(2.0)*
384	2.6×10^{-5}	*(4.0)*	2.9×10^{-3}	*(2.0)*

Table 3.24: Absolute errors, approximated orders of the method and computation times for linear equations solved using CWOPSIP.

3.3.3.3 Nonlinear equations

We specify three nonlinear tests. Results of simulations are presented in tables 3.25, 3.26 and 3.27. Generally conclusions are similar as in the linear case, i.e. linear convergence of H^1-error and quadratic convergence in L_2-error. There are some anomalies, probably due to the nonlinear solver. Also we note that error values in example 4 are generally better in nonlinear setting, which is probably due to good initial approximation. Since these numerical experiments were conducted to check the convergence rate instead of to verify the nonlinear solver, we used initial approximations close to (known) solutions to prevent divergence problems.

	3		4	
K	$L_2(\Omega)$	$H^1(\Omega)$	$L_2(\Omega)$	$H^1(\Omega)$
3	3.1×10^{-1}	1.9	3.2×10^{3}	3.4×10^{3}
6	6.6×10^{-2} *(4.7)*	9.4×10^{-1} *(2.1)*	1.1×10^{3} *(2.8)*	1.2×10^{3} *(2.7)*
12	1.7×10^{-2} *(4.0)*	4.6×10^{-1} *(2.0)*	3.1×10^{2} *(3.7)*	3.8×10^{2} *(3.3)*
24	4.3×10^{-3} *(3.9)*	2.3×10^{-1} *(2.0)*	5.4×10^{1} *(5.7)*	1.9×10^{2} *(2.0)*
48	1.1×10^{-3} *(3.9)*	1.2×10^{-1} *(2.0)*	9.6×10^{1} *(0.6)*	2.2×10^{2} *(0.8)*
96	2.7×10^{-4} *(4.0)*	5.8×10^{-2} *(2.0)*	4.6 *(20.9)*	1.9×10^{1} *(12.0)*
192	6.9×10^{-5} *(4.0)*	2.9×10^{-2} *(2.0)*	1.2 *(3.8)*	8.8 *(2.1)*
384	1.7×10^{-5} *(4.0)*	1.4×10^{-2} *(2.0)*	3.1×10^{-1} *(3.9)*	4.3 *(2.0)*

	5	
K	$L_2(\Omega)$	$H^1(\Omega)$
3	2.5×10^{-1}	1.1
6	7.7×10^{-2} *(3.3)*	3.9×10^{-1} *(2.8)*
12	2.1×10^{-2} *(3.8)*	1.3×10^{-1} *(2.9)*
24	5.2×10^{-3} *(3.9)*	5.3×10^{-2} *(2.5)*
48	1.3×10^{-3} *(4.0)*	2.4×10^{-2} *(2.2)*
96	3.3×10^{-4} *(4.0)*	1.2×10^{-2} *(2.1)*
192	8.2×10^{-5} *(4.0)*	5.8×10^{-3} *(2.0)*
384	2.0×10^{-5} *(4.0)*	2.9×10^{-3} *(2.0)*

Table 3.25: Absolute errors, approximated orders of the method and computation times for nonlinear equations solved using CIIPG.

	3				4			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
3	1.2		2.9		3.2×10^{3}		3.4×10^{3}	
6	2.7×10^{-1}	*(4.5)*	1.1	*(2.7)*	1.1×10^{3}	*(2.8)*	1.2×10^{3}	*(2.7)*
12	6.4×10^{-2}	*(4.2)*	4.9×10^{-1}	*(2.2)*	3.1×10^{2}	*(3.7)*	3.8×10^{2}	*(3.3)*
24	1.6×10^{-2}	*(4.1)*	2.4×10^{-1}	*(2.1)*	4.7×10^{1}	*(6.6)*	1.4×10^{2}	*(2.8)*
48	3.9×10^{-3}	*(4.1)*	1.2×10^{-1}	*(2.0)*	1.6×10^{1}	*(2.9)*	4.2×10^{1}	*(3.2)*
96	9.6×10^{-4}	*(4.0)*	5.8×10^{-2}	*(2.0)*	4.6	*(3.5)*	1.9×10^{1}	*(2.3)*
192	2.4×10^{-4}	*(4.0)*	2.9×10^{-2}	*(2.0)*	1.2	*(3.8)*	8.8	*(2.1)*
384	5.9×10^{-5}	*(4.0)*	1.4×10^{-2}	*(2.0)*	3.1×10^{-1}	*(3.9)*	4.3	*(2.0)*

	5			
K	$L_2(\Omega)$		$H^1(\Omega)$	
3	2.4×10^{-1}		1.0	
6	7.4×10^{-2}	*(3.2)*	3.8×10^{-1}	*(2.7)*
12	2.0×10^{-2}	*(3.7)*	1.3×10^{-1}	*(2.9)*
24	5.1×10^{-3}	*(3.9)*	5.3×10^{-2}	*(2.5)*
48	1.3×10^{-3}	*(4.0)*	2.4×10^{-2}	*(2.2)*
96	3.2×10^{-4}	*(4.0)*	1.2×10^{-2}	*(2.1)*
192	8.0×10^{-5}	*(4.0)*	5.8×10^{-3}	*(2.0)*
384	2.0×10^{-5}	*(4.0)*	2.9×10^{-3}	*(2.0)*

Table 3.26: Absolute errors, approximated orders of the method and computation times for nonlinear equations solved using CSIPG.

	3				4			
K	$L_2(\Omega)$		$H^1(\Omega)$		$L_2(\Omega)$		$H^1(\Omega)$	
3	4.0		5.1		3.2×10^{3}		3.4×10^{3}	
6	9.1×10^{-1}	*(4.4)*	1.4	*(3.6)*	1.1×10^{3}	*(2.8)*	1.2×10^{3}	*(2.7)*
12	2.2×10^{-1}	*(4.1)*	5.3×10^{-1}	*(2.7)*	3.1×10^{2}	*(3.7)*	3.8×10^{2}	*(3.3)*
24	5.5×10^{-2}	*(4.0)*	2.4×10^{-1}	*(2.2)*	1.0×10^{2}	*(3.1)*	1.7×10^{2}	*(2.2)*
48	1.4×10^{-2}	*(4.0)*	1.2×10^{-1}	*(2.1)*	2.2×10^{1}	*(4.5)*	1.3×10^{2}	*(1.3)*
96	3.5×10^{-3}	*(4.0)*	5.8×10^{-2}	*(2.0)*	4.6	*(4.9)*	1.9×10^{1}	*(6.8)*
192	8.6×10^{-4}	*(4.0)*	2.9×10^{-2}	*(2.0)*	1.2	*(3.8)*	8.8	*(2.1)*
384	2.2×10^{-4}	*(4.0)*	1.4×10^{-2}	*(2.0)*	3.1×10^{-1}	*(3.9)*	4.3	*(2.0)*

	5			
K	$L_2(\Omega)$		$H^1(\Omega)$	
3	2.4×10^{-1}		1.1	
6	7.6×10^{-2}	*(3.2)*	3.8×10^{-1}	*(2.7)*
12	2.0×10^{-2}	*(3.7)*	1.3×10^{-1}	*(2.9)*
24	5.2×10^{-3}	*(3.9)*	5.3×10^{-2}	*(2.5)*
48	1.3×10^{-3}	*(4.0)*	2.4×10^{-2}	*(2.2)*
96	3.2×10^{-4}	*(4.0)*	1.2×10^{-2}	*(2.1)*
192	8.1×10^{-5}	*(4.0)*	5.8×10^{-3}	*(2.0)*
384	2.0×10^{-5}	*(4.0)*	2.9×10^{-3}	*(2.0)*

Table 3.27: Absolute errors, approximated orders of the method and computation times for nonlinear equations solved using CWOPSIP.

Chapter 4

Appendix

Contents

4.A Theorems **193**
4.B Lemmas **195**
4.C Existence of discrete solutions in one dimension **199**
4.C.1 Operator T 200
4.C.2 Discrete operator $\tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$ 201
4.C.3 Discrete operator $v_h(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h)$ 210
4.C.4 Discrete operator $w_h(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h)$ 216
4.C.5 Discretization of the van Roosbroeck system 216
4.C.6 Analysis 217
4.D List of assumptions **222**
4.E Physical constants **223**

4.A Theorems

Theorem 4.A.1. (Schauder fixed point theorem) *Let X be a topological vector space and let K be a convex and compact subset of X. Let $T : K \mapsto K$ be a continuous function. Then there exists at least one $x \in K$, so that $T(x) = x$.*

Theorem 4.A.2. (Green's formula) *Let Ω be an open subset of $\mathbb{R}^n$ with a Lipschitz-continuous boundary Γ. Let $u \in H^1(\Omega)$ and $v \in H^1(\Omega)$. Then*

$$\int_\Omega \frac{\partial u}{\partial x_i} v \, dx = \int_\Gamma u v \nu_i \, ds - \int_\Omega u \frac{\partial v}{\partial x_i} \, dx. \tag{4.A.1}$$

Proof. See [84], special case of theorem 1.1, section 3.1.2. □

Theorem 4.A.3. (Green's formula) *Let Ω be an open subset of $\mathbb{R}^n$ with a Lipschitz-continuous boundary Γ. Let $u \in H^2(\Omega)$, $v \in H^1(\Omega)$ and $a_{ij} : \Omega \to \mathbb{R}$ such that $a_{ij} \frac{\partial u}{\partial x_i} \in H^1(\Omega)$ (for example, $a_{ij} \in \mathcal{C}^1(\overline{\Omega})$). Then*

$$\int_\Omega \sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \, dx = - \int_\Omega \sum_{i,j=1}^n \frac{\partial}{\partial x_j} \left(a_{ij} \frac{\partial u}{\partial x_i} \right) v \, dx + \int_\Gamma \sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_i} v \nu_j \, ds. \tag{4.A.2}$$

For $n = 1$, $\Omega = [x_0, x_1]$, it has a simple form

$$\int_{x_0}^{x_1} a(x) u'(x) v'(x) \, dx = - \int_{x_0}^{x_1} \left(a(x) u'(x) \right)' v(x) \, dx + a(x_1) u'(x_1) v(x_1) - a(x_0) u'(x_0) v(x_0). \tag{4.A.3}$$

Proof. For statement, see [28], section 1.2, page 22. □

Theorem 4.A.4. *Let* Ω *be an open subset of* $\mathbb{R}^n$ *with a Lipschitz-continuous boundary. Then if* $\frac{n}{p} < m$*, then* $W^{m,p}(\Omega) \overset{c}{\subset} \mathcal{C}^0(\overline{\Omega})$.

Proof. For statement, see [28], section 3.1, page 114, equation (3.1.4). □

Theorem 4.A.5. (a variant of the Brouwer theorem) *Let* $P : \mathbb{R}^n \to \mathbb{R}^n$ *be a continuous function, such that for suitable* $\rho > 0$ *we have*

$$\langle P(x)|x\rangle \geq 0 \quad \forall \|x\|_2 = \rho, \tag{4.A.4}$$

where

$$\langle x|y\rangle := \sum_{i=0}^{n} x_i y_j, \quad \|x\|_2 := \sqrt{\langle x|x\rangle}. \tag{4.A.5}$$

Then there exists $x \in \mathbb{R}^n$, $\|x\| \leq \rho$ *such that*

$$P(x) = 0. \tag{4.A.6}$$

Proof. See [69], lemma 4.3. □

Theorem 1.4.4. *Let* $P : X \to X^*$ *be a continuous function on a finite-dimensional normed real vector space* X*, such that for suitable* $\rho > 0$ *we have*

$$P(x)x \geq 0 \quad \forall \|x\| \geq \rho. \tag{1.4.6}$$

Then there exists $x \in X$ *such that*

$$P(x) = 0. \tag{1.4.7}$$

Proof. Let $\{x_i\}_{i=1}^n$ be a base of X. We define $Q : \mathbb{R}^n \to \mathbb{R}^n$ as

$$[Q(\alpha)]_j = P\Big(\sum_{i=1}^{n} \alpha_i x_i\Big) x_j. \tag{4.A.7}$$

Then Q is continuous as a composition of continuous operators. Note that $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ defined as

$$\|\alpha\| := \Big\|\sum_{i=1}^{n} \alpha_i x_i\Big\|_X, \tag{4.A.8}$$

is also a norm in $\mathbb{R}^n$. Therefore there is some $C > 0$ such that

$$C^{-1}\|\alpha\| \leq \|\alpha\|_2 \leq C\|\alpha\|. \tag{4.A.9}$$

Take $r := C\rho$ and let $\|\alpha\|_2 = r$. Then

$$\Big\|\sum_{i=1}^{n} \alpha_i x_i\Big\|_X = \|\alpha\| \geq C^{-1}\|\alpha\|_2 = C^{-1} r = \rho. \tag{4.A.10}$$

So $\|\sum_{i=1}^{n} \alpha_i x_i\|_X \geq \rho$ and by assumptions of this theorem, we have that

$$P\Big(\sum_{i=1}^{n} \alpha_i x_i\Big) \sum_{i=1}^{n} \alpha_i x_i \geq 0, \tag{4.A.11}$$

for any $\alpha \in \mathbb{R}^n$ such that $\|\alpha\|_2 = r$. Thus we have

$$\langle Q(\alpha)|\alpha\rangle = \sum_{j=1}^n \alpha_j [Q(\alpha)]_j = \sum_{j=1}^n \alpha_j P\Big(\sum_{i=1}^n \alpha_i x_i\Big) x_j. = P\Big(\sum_{i=1}^n \alpha_i x_i\Big) \sum_{j=1}^n \alpha_j x_j \geq 0, \tag{4.A.12}$$

for any $\|\alpha\|_2 = r$.

Using theorem 4.A.5 for Q we deduce that there is some α^*, $\|\alpha^*\|_2 \leq r$, such that $Q(\alpha^*) = 0$. Note that it implies for $x^* := \sum_{i=1}^n \alpha_i^* x_i$ that

$$P(x^*) x_i = 0 \quad \forall i = 1, \dots, n. \tag{4.A.13}$$

Thus $P(x^*) = 0$. Note also that

$$\|x^*\|_X = \Big\| \sum_{i=1}^n \alpha_i^* x_i \Big\|_X = \|\alpha^*\| \leq C \|\alpha^*\|_2 = C^2 \rho. \tag{4.A.14}$$

□

4.B Lemmas

Lemma 4.B.1. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a monotone increasing function and $c, d \in \mathbb{R}$, $c \in \operatorname{im} f$. Then for any $y \in \mathbb{R}$ we have*

$$\Big(f(y) - c\Big)\Big(y - d\Big) \geq \Big(f(d) - c\Big)\Big(f^{-1}(c) - d\Big). \tag{4.B.1}$$

Proof. A complete proof of this lemma may be found in article [60]. We will, however, present a proof based on a different approach. First we prove the following: let $a, b \in \mathbb{R}$. Then the following inequalities are satisfied:

$$\big(f(y) - f(a)\big)(b - y) \quad \leq \big(f(b) - f(a)\big)(b - a), \tag{4.B.2}$$

$$\big(f(y) - f(b)\big)(a - y) \quad \leq \big(f(b) - f(a)\big)(b - a). \tag{4.B.3}$$

First we note that RHS ≥ 0 as f is monotone. Then without loss of generality we assume that $a \leq b$ (otherwise we interchange a and b). Due to monotonicity

$$\begin{aligned} \operatorname{sign} \mathrm{LHS}_1 &= \operatorname{sign} \Big(f(y) - f(a)\Big)(b - y) = \operatorname{sign}(y - a)(b - y), \\ \operatorname{sign} \mathrm{LHS}_2 &= \operatorname{sign} \Big(f(y) - f(b)\Big)(a - y) = \operatorname{sign}(y - b)(a - y) = \operatorname{sign}(b - y)(y - a). \end{aligned} \tag{4.B.4}$$

Thus if $y \notin [a, b]$ then $\mathrm{LHS}_{1,2} < 0$ and then

$$\mathrm{LHS}_{1,2} < 0 \leq \mathrm{RHS}\,. \tag{4.B.5}$$

On the other hand, if $y \in [a, b]$, then $\mathrm{LHS}_1, \mathrm{LHS}_2$ and RHS are nonnegative, and it is easy to observe that they represent areas of rectangles (see figure 4.1) parallel to the Cartesian axes spanned by vertices $\big(a, f(y)\big), \big(y, f(b)\big)$ (right-hatched rectangle), $\big(y, f(a)\big), \big(b, f(y)\big)$ (left-hatched rectangle) and respectively $\big(a, f(a)\big), \big(b, f(b)\big)$ (gray rectangle). Since $a \leq y \leq b$, then first two rectangles are subsets of the last rectangle and thus

$$\mathrm{LHS}_{1,2} \leq \mathrm{RHS}\,. \tag{4.B.6}$$

Returning to our original inequality (4.B.1), we take:

$$a := f^{-1}(c), \quad b := d, \quad y := y. \tag{4.B.7}$$

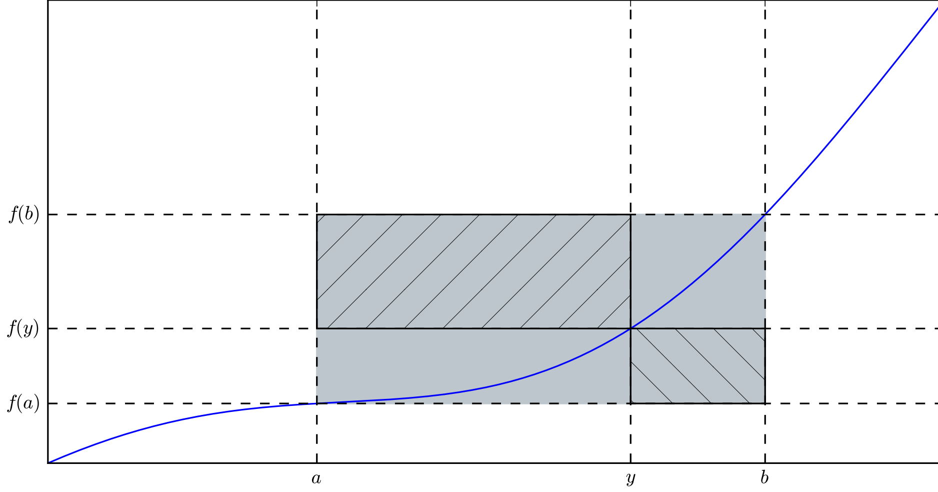

Figure 4.1: Geometric intuition for lemma 4.B.1. The blue line represents an example of a monotone increasing real function. Description of rectangles is in the proof of the lemma.

Then by inequality (4.B.2) we have

$$\Big(f(y) - c\Big)(d - y) \leq \Big(f(d) - c\Big)\Big(d - f^{-1}(c)\Big). \tag{4.B.8}$$

Multiplying this inequality by -1 we obtain inequality (4.B.1)

$$\Big(f(y) - c\Big)(y - d) \geq \Big(f(d) - c\Big)\Big(f^{-1}(c) - d\Big). \tag{4.B.9}$$

□

Lemma 4.B.2. *Let $\Omega \subset \mathbb{R}^n$ be an open set and let $f, g \in L_2(\Omega)$. Let $k \in \{1, 2, \ldots\}$. Then the following conditions are equivalent:*

1. $\int_\Omega f\phi\, dx = \int_\Omega g\phi\, dx \quad \forall \phi \in L_2(\Omega)$,
2. $\int_\Omega f\phi\, dx = \int_\Omega g\phi\, dx \quad \forall \phi \in H^k(\Omega)$,
3. $\int_\Omega f\phi\, dx = \int_\Omega g\phi\, dx \quad \forall \phi \in C_0^\infty(\Omega)$.

Proof. Implications $1 \Rightarrow 2 \Rightarrow 3$ are trivial, as $C_0^\infty(\Omega) \subset H^k(\Omega) \subset L_2(\Omega)$.

We will show $3 \Rightarrow 1$. Assume 3 is true. Let us fix $\phi \in L_2(\Omega)$. $C_0^\infty(\Omega)$ is dense in $L_2(\Omega)$, so let us take $\{\phi_i\}_i$ such that $\phi_i \xrightarrow[i\to\infty]{L_2(\Omega)} \phi$. Then we have

$$\begin{aligned} \Big|\int_\Omega (f-g)\phi\, dx\Big| &\leq \Big|\int_\Omega (f-g)\phi_i\, dx\Big| + \Big|\int_\Omega (f-g)(\phi-\phi_i)\, dx\Big| \\ &= \Big|\int_\Omega (f-g)(\phi-\phi_i)\, dx\Big| \leq \|f-g\|_{L_2(\Omega)}\|\phi-\phi_i\|_{L_2(\Omega)} \xrightarrow[i\to\infty]{} 0, \end{aligned} \tag{4.B.10}$$

as $\|\phi - \phi_i\|_{L_2(\Omega)} \to 0$. Thus

$$\int_\Omega f\phi\, dx = \int_\Omega g\phi\, dx. \tag{4.B.11}$$

This is true for any $\phi \in L_2(\Omega)$, so 1 is proven. □

Lemma 4.B.3. *Let $\Omega \subset \mathbb{R}^n$ be an open set and let $u \in H^1(\Omega)$, $f \in L_2(\Omega)$, $\varepsilon \in L_\infty(\Omega)$, $0 < \varepsilon \le \varepsilon_M$. Let $\partial\Omega_D \subset \partial\Omega$ be a subset of positive boundary measure. Then the following conditions are equivalent:*

1. $\int_\Omega \varepsilon \nabla u \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx \quad \forall \phi \in \mathcal{C}^\infty_{0,\partial\Omega_D}(\Omega)$,
2. $\int_\Omega \varepsilon \nabla u \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx \quad \forall \phi \in H^1_{0,\partial\Omega_D}(\Omega)$.

Proof. Implication $1 \Rightarrow 2$ is trivial. To prove implication $2 \Rightarrow 1$, take any $H^1_{0,\partial\Omega_D}(\Omega)$ and let $\{\phi_i\}_i \subset \mathcal{C}^\infty_{0,\partial\Omega_D}(\Omega)$, $\phi_i \to \phi$ in $H^1(\Omega)$. Then using condition 1 for Ω_i we obtain

$$\begin{aligned}\int_\Omega \varepsilon \nabla u \cdot \nabla \phi \, dx &= \int_\Omega \varepsilon \nabla u \cdot \nabla(\phi - \phi_i) \, dx + \int_\Omega \varepsilon \nabla u \cdot \nabla \phi_i \, dx \\ &= \int_\Omega \varepsilon \nabla u \cdot \nabla(\phi - \phi_i) \, dx + \int_\Omega f \phi_i \, dx \\ &= \int_\Omega \varepsilon \nabla u \cdot \nabla(\phi - \phi_i) \, dx + \int_\Omega f(\phi_i - \phi) \, dx + \int_\Omega f \phi \, dx.\end{aligned} \tag{4.B.12}$$

Then passing to the limit with i we get

$$\int_\Omega \varepsilon \nabla u \cdot \nabla \phi \, dx = \int_\Omega f \phi \, dx, \tag{4.B.13}$$

as

$$\left| \int_\Omega \varepsilon \nabla u \cdot \nabla(\phi - \phi_i) \, dx \right| \le \varepsilon_M \|u\|_{H^1(\Omega)} \|\phi - \phi_i\|_{H^1(\Omega)} \xrightarrow[i \to \infty]{} 0, \tag{4.B.14}$$

and

$$\left| \int_\Omega f(\phi_i - \phi) \, dx \right| \le \|f\|_{L_2(\Omega)} \|\phi - \phi_i\|_{L_2(\Omega)} \xrightarrow[i \to \infty]{} 0. \tag{4.B.15}$$

□

Lemma 4.B.4. *Let U be a rectangle in $\mathbb{R}^2$ or an interval in $\mathbb{R}$. Let $\Gamma_1 \subset \partial U$ be some edge of that rectangle (vertex in $\mathbb{R}$). Let $g \in \mathcal{C}^\infty_0(\Gamma_1)$($g \in \mathbb{R}$ in one dimension). Then there exists a family of functions $\{\phi_\epsilon\}_\epsilon$,$0 < \epsilon \le \epsilon_0$, such that*

- $\phi_\epsilon|_e = g$,
- $\operatorname{supp}(\phi_\epsilon) \subset U \cup \Gamma_1$,
- $\nabla \phi_\epsilon \cdot \nu = 0$ *on* ∂U,
- $\mu\big(\operatorname{supp}(\phi_\epsilon)\big) \le c\epsilon$.

Proof. First let $U = (0,1)$ and $\Gamma_1 = \{0\}$. In one dimension, Take

$$f(x) = \begin{cases} 1, & \text{if } x \in \left[0, \frac{1}{3}\right], \\ \exp\left(\frac{\left(\frac{1}{3}\right)^2 - x^2}{x^2 - \left(\frac{1}{3}\right)^2}\right), & \text{if } x \in \left(\frac{1}{3}, \frac{2}{3}\right), \\ 0, & \text{if } x \in \left[\frac{2}{3}, 1\right]. \end{cases} \tag{4.B.16}$$

Then define for $0 < \epsilon \leq 1$

$$\phi_\epsilon(x) := \begin{cases} gf(x/\epsilon) & \text{if } x \leq \epsilon, \\ 0 & \text{otherwise.} \end{cases} \tag{4.B.17}$$

These functions satisfy conditions of the lemma.

It is clear by appropriately translating and scaling x in above definition we can get ϕ_ϵ functions for any $U = (u_0, u_1)$. Also we can take $-x$ to deal with the right vertex of the interval.

For U being a rectangle, let us initially assume that $U = (0,1) \times (a,b)$ for some $a < b$ and that Γ_1 is the edge corresponding to $x = 0$. Then we take

$$\phi_\epsilon(x) := \begin{cases} g(y)f(x/\epsilon, y), & \text{if } x \leq \epsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{4.B.18}$$

Since $\operatorname{supp}(g) \Subset (a,b)$, then $\operatorname{supp}(g) \subset [e,f] \subset (a,b)$ and

$$\operatorname{supp}(\psi_\epsilon) \in [0,\epsilon] \times [e,f] \subset [0,1) \times (a,b) = U \cup \Gamma_1. \tag{4.B.19}$$

Therefore $\mu\big(\operatorname{supp}(\psi_\epsilon)\big) \leq (a-b)\epsilon$. Also $\nabla\phi_\epsilon \cdot \nu = 0$ on ∂U, as ϕ_ϵ is constant in x near Γ_1 and it is zero near $\partial U \backslash \Gamma_1$. Thus $\{\psi_\epsilon\}_\epsilon$ satisfy conditions of the lemma.

Again it is clear that by appropriate scaling and rotation we can generalize this procedure to any rectangle. ☐

Lemma 1.4.5. *Let $\Omega \subset \mathbb{R}^d$ be bounded. Let $f \in \mathcal{C}^1(\mathbb{R})$, $g \in L_\infty(\Omega)$. Let $P : X_h(\Omega) \to X_h^*(\Omega)$ be defined as*

$$P(u_h)\phi_h := \int_\Omega g(x) f\big(u_h(x)\big)\phi_h(x)\,dx. \tag{1.4.8}$$

Then P is continuous.

Proof. For fixed $u_h \in X_h(\Omega)$ the value $P(u_h)\phi_h$ is well-defined, as $g, u_h, \phi_h \in L_\infty(\Omega)$. Thus $P(u_h)$ is linear.

Take any $u_h, v_h \in X_h(\Omega)$. Then by equivalence of norms in finite-dimensional spaces

$$\begin{aligned} \|P(u_h) - P(v_h)\|_{X_h^*(\Omega)} &= \sup_{\|\phi_h\|_h = 1} |P(u_h)\phi_h - P(v_h)\phi_h| = \sup_{\|\phi_h\|_h=1} \left| \int_\Omega g(x)\big(f(u_h) - f(v_h)\big)\phi_h\,dx \right| \\ &\leq \sup_{\|\phi_h\|_h=1} \|g\|_{L_\infty(\Omega)} \|f(u_h) - f(v_h)\|_{L_2(\Omega)} \|\phi_h\|_{L_2(\Omega)} \\ &\leq c(h)\|g\|_{L_\infty(\Omega)} \|f(u_h) - f(v_h)\|_{L_2(\Omega)}. \end{aligned} \tag{4.B.20}$$

Then by the mean value theorem

$$\begin{aligned} \|f(u_h) - f(v_h)\|^2_{L_2(\Omega)} &= \int_\Omega \Big(f\big(u_h(x)\big) - f\big(v_h(x)\big)\Big)^2 dx \leq \int_\Omega \big(f'(\xi_x)\big)^2 \Big(u_h(x) - v_h(x)\Big)^2 dx \\ &\leq \|f'\|^2_{L_\infty(I(u_h,v_h))} \|u_h - v_h\|^2_{L_2(\Omega)}, \end{aligned} \tag{4.B.21}$$

where

$$\xi_x \in \big(u_h(x), v_h(x)\big) \cup \big(v_h(x), u_h(x)\big), \tag{4.B.22}$$

and

$$I\big(u_h, v_h\big) := [-M\big(u_h, v_h\big), M\big(u_h, v_h\big)], \quad M\big(u_h, v_h\big) := \max\{\|u_h\|_{L_\infty(\Omega)}, \|v_h\|_{L_\infty(\Omega)}\}. \tag{4.B.23}$$

Thus

$$\|P(u_h) - P(v_h)\|_{X_h^*(\Omega)} \le c(h)\|g\|_{L_\infty(\Omega)}\|f'\|^2_{L_\infty(I(u_h,v_h))}\|u_h - v_h\|^2_{L_2(\Omega)}. \tag{4.B.24}$$

Assume that a sequence $u_{(n)} \to u$ in $L_2(\Omega)$ as $n \to \infty$. This sequence converges also in $L_\infty(\Omega)$, as $X_h(\Omega)$ is finite-dimensional. Thus possibly ignoring some initial elements of $\{u_{(n)}\}_n$, for some $\epsilon > 0$ we have

$$I\big(u_h, u_{h,(n)}\big) \subset I(u_h) := \Big[-\big(\|u_h\|_{L_\infty(\Omega)} + \varepsilon\big), \big(\|u_h\|_{L_\infty(\Omega)} + \varepsilon\big)\Big]. \tag{4.B.25}$$

So we obtain

$$\|P(u_h) - P(v_h)\|_{X_h^*(\Omega)} \le c(h)\|g\|_{L_\infty(\Omega)}\|f'\|^2_{L_\infty(I(u_h))}\|u_h - v_h\|^2_{L_2(\Omega)}. \tag{4.B.26}$$

Therefore $u_h \mapsto P(u_h)$ is continuous. □

4.C Existence of discrete solutions in one dimension

In this section, we would like to discuss existence of one-dimensional CWOPSIP discretization of all equations of problem 1.2.1.

Problem 1.2.1. *Let $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2\}$ be an interval or polygon. Let $\hat{u}, \hat{v}, \hat{w} \in H^1(\Omega) \cap L_\infty(\Omega)$ be some given functions. We say that $(u^*, v^*, w^*) \in (\hat{u}, \hat{v}, \hat{w}) + \Big(H_0^1(\Omega)\Big)^3$ is a weak solution of (1.2.1) if $\forall \phi \in H_0^1(\Omega)$*

$$\begin{aligned}\int_\Omega \varepsilon(x)\nabla u^*(x)\nabla\phi(x)\,dx &= \int_\Omega \Big(k_1(x) - e^{u^*(x)-v^*(x)} + e^{w^*(x)-u^*(x)}\Big)\phi(x)\,dx,\\ \int_\Omega \mu_n(x)e^{u^*(x)-v^*(x)}\nabla v^*(x)\nabla\phi(x)\,dx &= \int_\Omega Q(u^*(x), v^*(x), w^*(x))(e^{w^*(x)-v^*(x)} - 1)\phi(x)\,dx,\\ \int_\Omega \mu_p(x)e^{w^*(x)-u^*(x)}\nabla w^*(x)\nabla\phi(x)\,dx &= -\int_\Omega Q(u^*(x), v^*(x), w^*(x))(e^{w^*(x)-v^*(x)} - 1)\phi(x)\,dx.\end{aligned} \tag{1.2.2}$$

In this section we use the following broken norm

$$\|u_h\|^2_{h,\Sigma_r} := \sum_{i=1}^N \int_{\Omega_i} \Big(\nabla u_{h,i}\Big)^2 dx + \sum_{e\in\Gamma_{DI}} \eta_{r,e}\int_e [u_h]^2\,ds. \tag{4.C.1}$$

In sections 4.C.2, 4.C.3, 4.C.4, we present discretization of the operator T (4.C.5), whose fixed points coincide with solutions of the van Roosbroeck equations. In [54], the existence of weak solutions is shown by the Schauder theorem. The idea behind the operator T is to decouple the drift-diffusion equations and to look for the solution in a Banach iteration manner, starting from some initial approximation. Then it is possible to decouple the equations by replacing unknown functions with solutions obtained from a previous iteration.

This kind of Banach iterations may be used not only in theoretical analysis, but also in numerical simulations. In the semiconductor nomenclature they are known as Gummel method [47, 101, 76], and the operator T is referred to as the Gummel's map. Theoretical results for convergence of Gummel's method are available for the drift-diffusion system when no recombination is assumed [62, 63, 30]. For certain devices, like transistors, such approximation is reasonable and simulations agree with physical experiments. However, this assumption is poor for light-emitting diodes and lasers, as the recombination in those devices is the effect leading to emission of the light. Unfortunately, the Gummel's

method convergence rate may considerably drop due to large recombination term [97], as then coupling between drift-diffusion equations increases.

Afterwards we pass to maximum principles for the discrete solutions which will be required for the Schauder theorem. As opposed to the existence, the proof for the continuous case cannot be adopted here. In general finite dimensional test space is too scarce to rule out extremes inside Ω such as for the weak solutions. We will then proceed differently.

For Poisson equation, we take an approach such as in [60]. We therefore define a generalized weak problem with severe assumptions, which are satisfied by the discretized Poisson equation. Then the boundedness proof is based on the bound of $H^1(\mathcal{E})$-seminorm of a solution. It is a consequence of the ellipticity of the equation, lower bounds on the right hand side and nonnegativity of the discrete part for certain choice of the test function. Then the bounds of the discrete solution is shown by combination of explicit expressions for its derivatives with the $|\cdot|_{H^1(\mathcal{E})}$ bounds. Finally we obtain a maximum principle for the discrete solution of the Poisson equation, which is however dependent not only on the boundary conditions, but also on the $L_\infty(\Omega)$-bounds on the solutions of the remaining two equations.

Therefore it is crucial to get maximum principles for the continuity equations, which are independent of the solution of the Poisson equation, as it is the case for the weak problem [54]. To do so, we assume the recombination to be zero. It is a severe restraint, however it is frequently used in analysis of the van Roosbroeck system [56, 76, 30]. This assumption is satisfied for a semiconductor device in the equilibrium state, when there is no current. When the recombination is zero, the continuity equations (in the sense of operator T) become linear elliptic equations with zero right hand side. Therefore we have to deal with the linear elliptic part and the CDGM part only. Then we show that linear systems corresponding to these problems have such a property that their solutions must be monotone: increasing or decreasing. Therefore we obtain maximum principles for the continuity equations with the bounding values dependent only on the boundary conditions.

Unfortunately this approach is not feasible for the CSIPG discretization or for CWOPSIP discretization in two dimensions. In these cases, we cannot prove the maximum principles analogous to the instances discussed above. For this reason, this analysis is presented only for one-dimensional CWOPSIP discretization.

Last step of this part is to show uniqueness of the decoupled equations. For the continuity equations, it follows from the diagonally-dominant form of the mentioned matrices. For the Poisson equation, we get the uniqueness by standard argument, as we demonstrate that a difference of any two hypothetical solutions must be zero. This may be achieved by subtraction of respective equations with these solutions and appropriate choice of the test function.

Results discussed so far allows us to conclude that T is a well-defined $L_\infty(\Omega)$-bounded operator. Using these bounds we may establish a closed convex bounded set, which is T-invariant. Due to finite-dimensionality of the discrete space, this set is automatically compact. To use the Schauder theorem, we still need T to be continuous. The continuity of T is derived from the decoupled equations, by taking appropriate test functions and by Schwarz inequality. The proof is technical, it also make use of $L_\infty(\Omega)$-estimates and equivalence of norms in discrete spaces. Finally, by the Schauder theorem, we obtain a fixed point of operator T, which is also a solution of the CWOPSIP discrete var Roosbroeck system.

4.C.1 Operator T

To prove the existence of a solution of problem 1.2.1, one can use the operator T defined as in [54].

Definition 4.C.1. *Let $\tilde{v}, \tilde{w} \in H^1(\Omega) \cap L_\infty(\Omega)$. We will define the operator $T : \left(H^1(\Omega) \cap L_\infty(\Omega)\right)^2 \to$*

$\left(H^1(\Omega)\right)^2$ *in the following manner. First we define* $\tilde{u} \in H^1(\Omega)$ *as a solution of the problem*

$$\forall \phi \in H^1(\Omega) \quad \int_\Omega \varepsilon(x) \nabla \tilde{u}(x) \nabla \phi(x)\, dx + \int_\Omega \left(e^{\tilde{u}(x)-\tilde{v}(x)} - e^{\tilde{w}(x)-\tilde{u}(x)} - k_1(x) \right) \phi(x)\, dx = 0. \qquad (4.C.2)$$

Then we define $v, w \in H^1(\Omega)$ *as solutions of problems*

$$\forall \phi \in H^1(\Omega) \quad \int_\Omega \mu_n(x) e^{\tilde{u}(x)-\tilde{v}(x)} \nabla v(x) \nabla \phi(x)\, dx - \int_\Omega Q(\tilde{u}(x), v(x), \tilde{w}(x)) (e^{\tilde{w}(x)-v(x)} - 1) \phi(x)\, dx = 0, \quad (4.C.3)$$

$$\forall \phi \in H^1(\Omega) \quad \int_\Omega \mu_p(x) e^{\tilde{w}(x)-\tilde{u}(x)} \nabla w(x) \nabla \phi(x)\, dx + \int_\Omega Q(\tilde{u}(x), \tilde{v}(x), w(x)) (e^{w(x)-\tilde{v}(x)} - 1) \phi(x)\, dx = 0. \quad (4.C.4)$$

The operator T *is defined as*

$$T(\tilde{v}, \tilde{w}) := (v, w). \qquad (4.C.5)$$

It is shown in [54] that in fact $T : \left(H^1(\Omega) \cap L_\infty(\Omega)\right)^2 \to \left(H^1(\Omega) \cap L_\infty(\Omega)\right)^2$ and it has a fixed point. We will follow this way using discretized equations instead.

4.C.2 Discrete operator $\tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$

To obtain a discretization of equation (4.C.2), we use discrete problem 1.3.3.

Problem 4.C.2. *Let* $\hat{u} \in H^1(\Omega) \cap L_\infty(\Omega)$ *and* $\tilde{v}_h, \tilde{w}_h \in X_h(\Omega)$ *be some given functions. Find* $\tilde{u}_h \in X_h(\Omega)$ *such that*

$$a_{u,h}(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) + b_u(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) = f_{u,h}(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h), \qquad (4.C.6)$$

where

$$\begin{aligned} a_{u,h}(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) &= \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon_i(x) \nabla \tilde{u}_{h,i}(x) \cdot \nabla \phi_{h,i}(x)\, dx + \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [\tilde{u}_h][\phi_h]\, ds, \\ b_u(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) &= \int_\Omega \left(e^{\tilde{u}_h(x)-\tilde{v}_h(x)} - e^{\tilde{w}_h(x)-\tilde{u}_h(x)} \right) \phi_h(x)\, dx, \\ f_{u,h}(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) &= \int_\Omega k_1(x) \phi_h(x)\, dx + \sum_{e \in \Gamma_D} \eta_{2,e} \int_e [\hat{u}][\phi_h]\, ds. \end{aligned} \qquad (4.C.7)$$

Our aim is to show the existence of discrete solution of the drift-diffusion system. First we would like to establish an existence theorem and maximum principles for solutions of problem 4.C.2.

Theorem 4.C.3. *Let* $\tilde{v}_h, \tilde{w}_h \in X_h(\Omega)$ *be given functions. Assume that there exist* $\alpha_h, \beta_h \in \mathbb{R}$*, such that*

$$\alpha_h \le \tilde{v}_h \le \beta_h, \quad \alpha_h \le \tilde{w}_h \le \beta_h. \qquad (4.C.8)$$

Then $\tilde{u}_h$ *of problem 4.C.2 is well-defined, i.e. it exists and it is unique. Moreover* $\tilde{u}_h$ *is bounded independently of* $\tilde{v}_h, \tilde{w}_h$*.*

Existence and uniqueness of $\tilde{u}_h$ is shown in section 1.4 if we perform the following substitutions

$$u_h^* := \tilde{u}_h, \quad \hat{v} := \tilde{v}_h, \quad \hat{w} := \tilde{w}_h. \qquad (4.C.9)$$

Thus we would like to show boundedness.

4.C.2.1 Bounds

In this section we would like to show that $\tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$ is bounded independently of $\tilde{v}_h, \tilde{w}_h$. To do so, we will prove slightly more general lemma, as in [60], for the discrete problem defined in section 1.3.2.2. Therefore we consider the following generalized problem.

4.C.2.1.1 Generalized problem

Problem 4.C.4. *Let us consider the equation*

$$A(u_h, \phi_h) + J(u_h, \phi_h) + B(u_h, \phi_h) = I(\phi_h), \tag{4.C.10}$$

for all $\phi_h \in X_h(\Omega)$, where

$$\begin{aligned}
A(u_h, \phi_h) &:= \sum_{i=1}^{N} \int_{\Omega_i} \varepsilon(x) \nabla u_h(x) \cdot \nabla \phi_h(x)\, dx, \\
B(u_h, \phi_h) &:= \int_{\Omega} \Big(f\big(x, u_h(x)\big) - g(x) \Big) \phi_h(x)\, dx, \\
J(u_h, \phi_h) &:= \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [u_h][\phi_h]\, ds, \\
I(\phi_h) &:= \sum_{e \in \Gamma_D} \eta_{2,e} \int_e [\hat{u}][\phi_h]\, ds.
\end{aligned} \tag{4.C.11}$$

Moreover, we make the following assumptions related to problem 4.C.4

Assumption A8.

- $f : \Omega \times \mathbb{R} \mapsto \mathbb{R}$.
- $g \in L_\infty(\Omega)$.
- *Let $\tilde{f}_x(y) := f(x, y)$ for fixed $x \in \Omega$. Then $\tilde{f}_x$ is a monotone increasing function for almost all $x \in \Omega$ (thus $\tilde{f}_x^{-1}$ exists for a.e. $x \in \Omega$ and it is monotone increasing).*
- $\operatorname{rg}(g) \subset \operatorname{dom} \tilde{f}_x^{-1}$ *and* $\tilde{f}_x^{-1}\big(\operatorname{rg}(g)\big)$ *is uniformly bounded set for almost all $x \in \Omega$.*
- *Let B be a bounded subset of $\mathbb{R}$. Then $\tilde{f}_x(B)$ is uniformly bounded set for almost all $x \in \Omega$.*
- $\hat{u} \in H^1(\Omega) \cap X_h(\Omega)$.
- $P \equiv 0$.

Note that in this case, since we consider one-dimensional domain Ω, function $\hat{u}$ is related to boundary conditions on two extreme points. So the last assumption of A8 is not too restrictive, as one can take linear function for example.

Theorem 4.C.5. *Under assumptions A1 to A8, if $u_h \in X_h(\Omega)$ is any solution of problem 4.C.4, then it is bounded:*

$$\gamma_h \le u_h \le \delta_h, \tag{4.C.12}$$

where

$$\gamma_h := \min\{\gamma_h' - c\sqrt{h}, \inf_{x \in \delta\Omega} \hat{u}(x)\}, \qquad \delta_h := \max\{\delta_h' + c\sqrt{h}, \sup_{x \in \delta\Omega} \hat{u}(x)\}, \tag{4.C.13}$$

$$\gamma_h' := \inf_{x \in \Omega} \tilde{f}_x^{-1}\Big(\inf_{\tilde{x} \in \Omega} g(\tilde{x}) \Big), \qquad \delta_h' := \sup_{x \in \Omega} \tilde{f}_x^{-1}\Big(\sup_{\tilde{x} \in \Omega} g(\tilde{x}) \Big), \tag{4.C.14}$$

and the constant $c > 0$ depends on $\hat{u}, \varepsilon_m, \varepsilon_M, f, g$.

For simplicity of notation, we will assume, relying on the above conditions, that there exist a constant $c_g > 0$, so that

$$\begin{aligned} -c_g \le \gamma_h' \le \hat{u} \le \delta_h' \le & c_g \text{ (a. e.)}, \\ -c_g \le \tilde{f}_x(\hat{u}(x)) \le & c_g \text{ (a. e.)}, \\ -c_g \le g(x) \le & c_g \text{ (a. e.)}, \\ -c_g \le \tilde{f}_x^{-1}(g(x)) \le & c_g \text{ (a. e.)}, \\ 0 < c_g^{-1} \le \varepsilon_m \le \varepsilon(x) \le \varepsilon_M \le & c_g \text{ (a. e.)}, \\ |\hat{u}|_{H^1(\Omega)} \le & c_g, \\ |\Omega| \le & c_g^3. \end{aligned} \tag{4.C.15}$$

4.C.2.1.2 Lemmas

Lemma 4.C.6. *If $u_h \in X_h(\Omega)$ is any solution of problem 4.C.4, then*

$$|u_h|_{H^1(\mathcal{E})} \le c_2, \tag{4.C.16}$$

where $c \ge 0$ depends on $\hat{u}, f, g, \Omega, \varepsilon_m, \varepsilon_M$.

Proof. Let u_h be a solution of problem 4.C.4. Then we can define the test function to be

$$\phi_h := u_h - \hat{u}. \tag{4.C.17}$$

so

$$A(u_h, u_h) = I(u_h - \hat{u}) - J(u_h, u_h - \hat{u}) + A(u_h, \hat{u}) - B(u_h, u_h - \hat{u}). \tag{4.C.18}$$

We have that $A(u_h, u_h) \ge \varepsilon_m |u_h|^2_{H^1(\mathcal{E})}$. Thus the lemma will be proven if we show that

$$I(u_h - \hat{u}) - J(u_h, u_h - \hat{u}) + A(u_h, \hat{u}) - B(u_h, u_h - \hat{u}) \le c|u_h|_{H^1(\mathcal{E})}. \tag{4.C.19}$$

First, using the Schwarz inequality

$$\begin{aligned} A(u_h, \hat{u}) &\le \int_\Omega \left| \varepsilon(x) \nabla u_h(x) \cdot \nabla \hat{u}(x) \right| dx \\ &\le \varepsilon_M |u_h|_{H^1(\mathcal{E})} |\hat{u}|_{H^1(\Omega)} \le c_g |u_h|_{H^1(\mathcal{E})} |\hat{u}|_{H^1(\Omega)}. \end{aligned} \tag{4.C.20}$$

Then we have

$$\begin{aligned} I(u_h - \hat{u}) - J(u_h, u_h - \hat{u}) &= \sum_{e \in \Gamma_D} \eta_{2,e} \int_e [\hat{u}][u_h - \hat{u}]\, ds - \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [u_h][u_h - \hat{u}]\, ds \\ &= \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [\hat{u}][u_h - \hat{u}]\, ds - \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [u_h][u_h - \hat{u}]\, ds \\ &= \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [\hat{u} - u_h][u_h - \hat{u}]\, ds \le 0, \end{aligned} \tag{4.C.21}$$

as $[\hat{u}] = 0$ for $e \in \Gamma_I$ since $\hat{u} \in H^1(\Omega)$. It is therefore clear that

$$\begin{aligned} A(u_h, u_h) &= I(u_h - \hat{u}) - J(u_h, u_h - \hat{u}) + A(u_h, \hat{u}) - B(u_h, u_h - \hat{u}) \\ &\le c_g |u_h|_{H^1(\mathcal{E})} |\hat{u}|_{H^1(\Omega)} - B(u_h, u_h - \hat{u}). \end{aligned} \tag{4.C.22}$$

Then $B(u_h, u_h - \hat{u})$ may be negative or not. Assume first it is nonnegative. Then

$$\begin{aligned} \varepsilon_m |u_h|^2_{H^1(\mathcal{E})} \leq A(u_h, u_h) &\leq c_g |u_h|_{H^1(\mathcal{E})} |\hat{u}|_{H^1(\Omega)} - B(u_h, u_h - \hat{u}) \\ &\leq c_g |u_h|_{H^1(\mathcal{E})} |\hat{u}|_{H^1(\Omega)}, \end{aligned} \tag{4.C.23}$$

so $|u_h|_{H^1(\mathcal{E})} \leq c_2$ with $c_2 = c_g |\hat{u}|_{H^1(\Omega)}/\varepsilon_m$.

Otherwise using lemma 4.B.1, we can estimate almost everywhere in Ω

$$\begin{aligned} B(u_h, u_h - \hat{u}) &= \int_\Omega \Big(f\big(x, u_h(x)\big) - g(x) \Big) \Big(u_h(x) - \hat{u}(x) \Big)\, dx \\ &= \int_\Omega \Big(\tilde{f}_x\big(u_h(x)\big) - g(x) \Big) \Big(u_h(x) - \hat{u}(x) \Big)\, dx \\ &\geq \int_\Omega \Big(\tilde{f}_x\big(\hat{u}(x)\big) - g(x) \Big) \Big(\tilde{f}_x^{-1}\big(g(x)\big) - \hat{u}(x) \Big)\, dx. \end{aligned} \tag{4.C.24}$$

The integrand of the latter expression is nonpositive, because $\tilde{f}_x$ is monotone increasing, and then for almost all $x \in \Omega$ we have

$$\begin{aligned} &\Big(\tilde{f}_x\big(\hat{u}(x)\big) - g(x) \Big) \Big(\tilde{f}_x^{-1}\big(g(x)\big) - \hat{u}(x) \Big) \\ &\qquad = -\Big(\tilde{f}_x\big(\hat{u}(x)\big) - \tilde{f}_x\big(\tilde{f}_x^{-1}\big(g(x)\big)\big) \Big) \Big(\hat{u}(x) - \tilde{f}_x^{-1}\big(g(x)\big) \Big) \leq 0. \end{aligned} \tag{4.C.25}$$

By assumptions A8, functions $\hat{u}, g$ are bounded by c_g, and $\tilde{f}_x$ and $\tilde{f}_x^{-1}$ are also bounded uniformly by c_g for all $x \in \Omega$, if their arguments are bounded. Then there exists $c_1 > 0$, such that

$$\begin{aligned} \Big(\tilde{f}_x\big(\hat{u}(x)\big) - g(x) \Big) \Big(\tilde{f}_x^{-1}\big(g(x)\big) - \hat{u}(x) \Big) &\geq -\Big(\big|\tilde{f}_x\big(\hat{u}(x)\big)\big| + \big|g(x)\big| \Big) \Big(\big|\tilde{f}_x^{-1}\big(g(x)\big)\big| + \big|\hat{u}(x)\big| \Big) \\ &> -4c_g^2 =: -c_1. \end{aligned} \tag{4.C.26}$$

Then we may conclude

$$B(u_h, u_h - \hat{u}) \geq -c_1. \tag{4.C.27}$$

So we may estimate

$$\varepsilon_m |u_h|^2_{H^1(\mathcal{E})} \leq A(u_h, u_h) \leq c_g |u_h|_{H^1(\mathcal{E})} |\hat{u}|_{H^1(\Omega)} + c_1. \tag{4.C.28}$$

Then again there are two possibilities: either $|u_h|_{H^1(\mathcal{E})} < c_1$ and then the lemma is proven with $c_2 = c_1$, or $c_1 \leq |u_h|_{H^1(\mathcal{E})}$. In the latter case

$$\varepsilon_m |u_h|^2_{H^1(\mathcal{E})} \leq c_g |u_h|_{H^1(\mathcal{E})} |\hat{u}|_{H^1(\Omega)} + c_1 \leq c_g |u_h|_{H^1(\mathcal{E})} |\hat{u}|_{H^1(\Omega)} + |u_h|_{H^1(\mathcal{E})}, \tag{4.C.29}$$

so we obtain

$$\varepsilon_m |u_h|_{H^1(\mathcal{E})} \leq c_g |\hat{u}|_{H^1(\Omega)} + 1. \tag{4.C.30}$$

In this case $c := \varepsilon_m^{-1}(c_g |\hat{u}|_{H^1(\Omega)} + 1)$ Summarizing all possibilities, the lemma is proven with

$$c_2 := \max\{\varepsilon_m^{-1}(c_g |\hat{u}|_{H^1(\Omega)} + 1), 4c_g^2\}. \tag{4.C.31}$$

□

4.C.2.1.3 Theorem

Proof. Outline of the proof:

1. For any adjacent grid nodes x_j, x_{j+1} we have $\big(u_h(x_{j+1}) - u_h(x_j)\big)^2 \leq c_2^2 h$.
2. Upper bound:
   (a) If $u_h \in X_h(\Omega)$ has a maximum in node x_k, then $\sum_{i=1}^N \int_{\Omega_i} \Big(f\big(x, u_h(x)\big) - g(x)\Big)\varphi_{(k)}(x)\, dx \leq 0$.
   (b) For some $x \in \Omega$ we have $f\big(x, \sup_{y\in\Omega} u_h(y) - c_2\sqrt{h}\big) - \sup_{y\in\Omega} g(y) \leq 0$.
   (c) $\sup_{y\in\Omega} u_h(y) \leq \delta_h' + c_2\sqrt{h} \leq \delta_h$.
3. Lower bound:
   (a) If $u_h \in X_h(\Omega)$ has a maximum in node x_k, then $\sum_{i=1}^N \int_{\Omega_i} \Big(f\big(x, u_h(x)\big) - g(x)\Big)\varphi_{(k)}(x)\, dx \geq 0$.
   (b) For some $x \in \Omega$ we have $f\big(x, \inf_{y\in\Omega} u_h(y) + c_2\sqrt{h}\big) - \inf_{y\in\Omega} g(y) \geq 0$.
   (c) $\inf_{y\in\Omega} u_h(y) \geq \gamma_h' - c_2\sqrt{h} \geq \gamma_h$.

First note that using piecewise linearity of u_h, for any $\tau \in \mathcal{T}$, $\tau =: (x_j, x_{j+1})$ we may express $\nabla u_h|_\tau$ with the explicit formula

$$\nabla u_h|_\tau = \frac{u_h(x_{j+1}) - u_h(x_j)}{x_{j+1} - x_j}. \tag{4.C.32}$$

Using lemma 4.C.6 we obtain

$$|u_h|^2_{H^1(\mathcal{E})} = \sum_{i=1}^N \sum_{\tau\in\mathcal{T}_{h_i}} \int_\tau \Big(\nabla u_h(x)\Big)^2 dx \leq c_2^2. \tag{4.C.33}$$

Note that all elements of the above sum are nonnegative, so for a fixed $\tau \in \mathcal{T}$ we may estimate

$$\int_\tau \Big(\nabla u_h\Big)^2 dx = \frac{\big(u_h(x_{j+1}) - u_h(x_j)\big)^2}{x_{j+1} - x_j} \leq |u_h|^2_{H^1(\mathcal{E})} \leq c_2^2. \tag{4.C.34}$$

Therefore we finally obtain

$$\big(u_h(x_{j+1}) - u_h(x_j)\big)^2 \leq c_2^2 h. \tag{4.C.35}$$

We may rewrite the equation (4.C.10)

$$A(u_h, \phi_h) + J(u_h, \phi_h) + B(u_h, \phi_h) = I(\phi_h), \tag{4.C.36}$$

as

$$\begin{aligned}
&\sum_{i=1}^N \int_{\Omega_i} \Big(f\big(x, u_h(x)\big) - g(x)\Big)\phi_h(x)\, dx = \\
&\quad = -\sum_{i=1}^N \int_{\Omega_i} \varepsilon(x)\nabla u_h(x)\cdot\nabla\phi_h(x)\, dx - \sum_{e\in\Gamma_{DI}} \eta_{2,e} \int_e [u_h][\phi_h]\, ds + \sum_{e\in\Gamma_D} \eta_{2,e} \int_e [\hat{u}][\phi_h]\, ds \\
&\quad = -\sum_{i=1}^N \int_{\Omega_i} \varepsilon(x)\nabla u_h(x)\cdot\nabla\phi_h(x)\, dx - \sum_{e\in\Gamma_I} \eta_{2,e} \int_e [u_h][\phi_h]\, ds - \sum_{e\in\Gamma_D} \eta_{2,e} \int_e [u_h - \hat{u}][\phi_h]\, ds.
\end{aligned} \tag{4.C.37}$$

Since $u_h \in X_h(\Omega)$ is a piecewise linear function, its extremes lie in the nodal points. Assume $x_k \in \mathcal{N}$ is such a point that $u_h(x_k) = \sup_{x\in\Omega} u_h(x)$. Then

$$\forall_{x_j \in \mathcal{N}_h} \quad u_h(x_k) - u_h(x_j) \geq 0. \tag{4.C.38}$$

Let us take $\varphi_{(k)}$ as a test function in (4.C.37). Therefore we obtain

$$\begin{aligned}\sum_{i=1}^{N} \int_{\Omega_i} \Big(f(x, u_h(x)) - g(x)\Big)\varphi_{(k)}(x)\,dx = \\ = -\sum_{i=1}^{N} \int_{\Omega_i} \varepsilon(x)\nabla u_h(x)\cdot\nabla\varphi_{(k)}(x)\,dx - \sum_{e\in\Gamma_I} \eta_{2,e}\int_e [u_h][\varphi_{(k)}]\,ds - \sum_{e\in\Gamma_D} \eta_{2,e}\int_e [u_h - \hat{u}][\varphi_{(k)}]\,ds.\end{aligned} \tag{4.C.39}$$

We will show that the right hand side of this equation is negative, and therefore so is the left hand side.

Note that the support of $\varphi_{(k)}$ is contained within triangulation elements adjacent to the node x_k. Also $\varphi_{(k)}$ has a maximum in x_k, as u_h has. Therefore signs of $\nabla\varphi_{(k)}$ and ∇u_h agree in supp $\varphi_{(k)}$ unless one of them is zero. Thus $\nabla u_h \cdot \nabla\varphi_{(k)} \geq 0$ in supp $\varphi_{(k)}$ a.e., and it is zero in $\Omega \setminus \operatorname{supp} \varphi_{(k)}$. Then we may estimate

$$\sum_{i=1}^{N} \int_{\Omega_i} \varepsilon(x)\nabla u_h(x)\cdot\nabla\varphi_{(k)}(x)\,dx \geq \varepsilon_m \int_{\operatorname{supp}\varphi_{(k)}} \nabla u_h(x)\cdot\nabla\varphi_{(k)}(x)\,dx \geq 0. \tag{4.C.40}$$

Then we deal with the expression $\sum_{e\in\Gamma_I}\eta_{2,e}\int_e [u_h][\varphi_{(k)}]ds$. For $e \in \Gamma_I$, $e \in \partial\Omega_i \cap \partial\Omega_{i-1}$ we have

$$[u_h][\varphi_{(k)}] \geq 0, \tag{4.C.41}$$

because either $[\varphi_{(k)}] = 0$ when the maximum is not in e or it is there and we have

$$[u_h][\varphi_{(k)}] = \Big(u_{h,i}(e) - u_{h,i-1}(e)\Big)\cdot\Big(\varphi_{(k),i}(e) - \varphi_{(k),i-1}(e)\Big) \geq 0, \tag{4.C.42}$$

as both functions have maxima in the same node x_k. Then since $\eta_{2,e} \geq 0$ for any $e \in \Gamma_{DI}$, we obtain

$$\sum_{e\in\Gamma_I} \eta_{2,e}\int_e [u_h][\varphi_{(k)}]\,ds \geq 0. \tag{4.C.43}$$

Finally sum $\sum_{e\in\Gamma_D}\eta_{2,e}\int_e [u_h - \hat{u}][\varphi_{(k)}]\,ds$ is non-zero only if $x_k \in \Gamma_D$. Note that only $[u_h - \hat{u}] = u_h(x_k) - \hat{u}(x_k)$ may be negative. Therefore we have two possibilities:

1. $u_h(x_k) \geq \hat{u}(x_k)$. Then $\sum_{e\in\Gamma_D}\eta_{2,e}\int_e [u_h - \hat{u}][\varphi_{(k)}]\,ds \geq 0$.
2. $u_h(x_h) < \hat{u}(x_k)$. Then $u_h(x_k) = \sup_\Omega u_h < \hat{u}(x_k) \leq \sup_{\delta\Omega}\hat{u} = \delta_h$.

So in the latter case the theorem is proven. We then follow the first possibility, where we have

$$\sum_{e\in\Gamma_D} \eta_{2,e}\int_e [u_h - \hat{u}][\varphi_{(k)}]\,ds \geq 0. \tag{4.C.44}$$

On the other hand, since $\varphi_{(k)}$ is positive, we have

$$\begin{aligned}\sum_{i=1}^{N} \int_{\Omega_i} \Big(f(x, u_h(x)) - g(x)\Big)\varphi_{(k)}(x)\,dx &\geq \sum_{i=1}^{N} \int_{\Omega_i} \Big(f(x, u_h(x)) - \sup_{y\in\Omega} g(y)\Big)\varphi_{(k)}(x)\,dx \\ &\geq \sum_{i=1}^{N} \int_{\Omega_i} \Big(f(x, \sup_{y\in\Omega} u_h(y) - c_3\sqrt{h}) - \sup_{y\in\Omega} g(y)\Big)\varphi_{(k)}(x)\,dx,\end{aligned} \tag{4.C.45}$$

where the latter inequality is explained as follows. Using inequality (4.C.35) for $u_h(x_k)$, we obtain

$$\big(u_h(x_k) - u_h(x_{k\pm1})\big)^2 \leq c_2^2 h, \tag{4.C.46}$$

where we restrict to $k \pm 1 \in \{1, \ldots, J\}$. Taking square root of the both sides and using the fact that $u_h(x_k)$ is maximal, we obtain

$$u_h(x_k) - u_h(x_{k\pm1}) \leq c_2\sqrt{h}. \tag{4.C.47}$$

Then by assumption A8 we have that $f(x,\cdot)$ is monotonically increasing, and the integration may be restricted to $\operatorname{supp}\varphi_{(k)}$, i.e. to the neighboring nodes. Since u_h is piecewise linear,

$$u_h|_{\operatorname{supp}\varphi_{(k)}} \geq u_h(x_k) - c_2\sqrt{h} = \sup_{y\in\Omega} u_h(y) - c_2\sqrt{h}. \tag{4.C.48}$$

Then using the monotonicity of $f(x,\cdot)$ we obtain inequality (4.C.45).

Finally using inequalities (4.C.40), (4.C.43), (4.C.44), (4.C.45) we may estimate

$$\sum_{i=1}^{N}\int_{\Omega_i}\Big(f\big(x, \sup_{y\in\Omega} u_h(y) - c_2\sqrt{h}\big) - \sup_{y\in\Omega} g(y)\Big)\varphi_{(k)}(x)\,dx \leq 0. \tag{4.C.49}$$

Since $\varphi_{(k)}$ is nonnegative, there exists $x \in \Omega$, such that

$$f\big(x, \sup_{y\in\Omega} u_h(y) - c_2\sqrt{h}\big) - \sup_{y\in\Omega} g(y) \leq 0. \tag{4.C.50}$$

Rearranging the elements and using the notation $\tilde{f}_x(y) = f(x,y)$ we obtain

$$\tilde{f}_x\big(\sup_{y\in\Omega} u_h(y) - c_2\sqrt{h}\big) \leq \sup_{y\in\Omega} g(y). \tag{4.C.51}$$

By assumption A8 function $\tilde{f}_x^{-1}$ exists and it is also monotone increasing, so we may apply it to the inequality to get

$$\sup_{y\in\Omega} u_h(y) - c_2\sqrt{h} \leq \tilde{f}_x^{-1}\left(\sup_{y\in\Omega} g(y)\right). \tag{4.C.52}$$

Therefore we finally obtain

$$\sup_{y\in\Omega} u_h(y) \leq \delta_h' + c_2\sqrt{h} \leq \delta_h. \tag{4.C.53}$$

For the lower bound the analysis is similar, but we describe it for the completeness. Assume that $u_h(x_k) = \inf_{x\in\Omega} u_h(x)$. Let us take $\varphi_{(k)}$ as a test function. Then $\varphi_{(k)}$ attains a maximum in x_k and u_h attains a minimum. Therefore signs of $\nabla\varphi_{(k)}$ and ∇u_h do not agree in supp $\varphi_{(k)}$ unless one of them is zero. Thus $\nabla u_h \cdot \nabla\varphi_{(k)} \leq 0$ in supp $\varphi_{(k)}$ a.e., and it is zero in $\Omega \setminus$ supp $\varphi_{(k)}$. Then we may estimate

$$\sum_{i=1}^{N}\int_{\Omega_i}\varepsilon(x)\nabla u_h(x)\cdot\nabla\varphi_{(k)}(x)\,dx \leq \varepsilon_m \int_{\operatorname{supp}\varphi_{(k)}} \nabla u_h(x)\cdot\nabla\varphi_{(k)}(x)\,dx \leq 0. \tag{4.C.54}$$

Then we deal with the expression $\sum_{e\in\Gamma_I}\eta_{2,e}\int_e [u_h][\varphi_{(k)}]\,ds$. For $e \in \Gamma_I$, $e = \partial\Omega_i \cap \partial\Omega_{i-1}$ we have

$$[u_h][\varphi_{(k)}] \leq 0, \tag{4.C.55}$$

because either $[\varphi_{(k)}] = 0$ when $x_k \neq e$ or

$$[u_h][\varphi_{(k)}] = \Big(u_{h,i}(e) - u_{h,i-1}(e)\Big)\cdot\Big(\varphi_{(k),i}(e) - \varphi_{(k),i-1}(e)\Big) \leq 0, \tag{4.C.56}$$

as both functions have extremes in the same node x_k, but one is a minimum and the other is maximum. Then since $\eta_{2,e} \geq 0$ for $e \in \Gamma_{DI}$, we obtain

$$\sum_{e\in\Gamma_I} \eta_{2,e} \int_e [u_h][\varphi_{(k)}]\, ds \leq 0. \tag{4.C.57}$$

The last element $\sum_{e\in\Gamma_D} \eta_{2,e} \int_e [u_h - \hat{u}][\varphi_{(k)}]\, ds$ is not zero only if $x_k = e \in \Gamma_D$. Thus we have two possibilities:

1. $u_h(x_k) \leq \hat{u}(x_k)$. Then $\sum_{e\in\Gamma_D} \eta_{2,e} \int_e [u_h - \hat{u}][\varphi_{(k)}]\, ds \leq 0$.
2. $u_h(x_k) > \hat{u}(x_k)$. Then $u_h(x_k) = \inf_\Omega u_h > \hat{u}(x_k) \geq \inf_{\delta\Omega} \hat{u} = \gamma_h$.

So in the latter case the theorem is proven. Otherwise we have

$$\sum_{e\in\Gamma_D} \eta_{2,e} \int_e [u_h - \hat{u}][\varphi_{(k)}]\, ds \leq 0. \tag{4.C.58}$$

On the other hand, since $\varphi_{(k)}$ is positive, we have

$$\begin{aligned}\sum_{i=1}^{N} \int_{\Omega_i} \Big(f\big(x, u_h(x)\big) - g(x)\Big)\varphi_{(k)}(x)\, dx &\leq \sum_{i=1}^{N} \int_{\Omega_i} \Big(f\big(x, u_h(x)\big) - \inf_{y\in\Omega} g(y)\Big)\varphi_{(k)}(x)\, dx \\ &\leq \sum_{i=1}^{N} \int_{\Omega_i} \Big(f\big(x, \inf_{y\in\Omega} u_h(y) + c_2\sqrt{h}\big) - \inf_{y\in\Omega} g(y)\Big)\varphi_{(k)}(x)\, dx,\end{aligned} \tag{4.C.59}$$

where the latter inequality is explained as follows. We again use the inequality (4.C.35) for $u_h(x_k)$

$$\big(u_h(x_k) - u_h(x_{k\pm1})\big)^2 \leq c_2^2 h, \tag{4.C.60}$$

where we consider $k \pm 1 \in \{1, \dots, J\}$. Taking square root of the both sides and using the fact that $u_h(x_k)$ is minimal, we obtain

$$-u_h(x_k) + u_h(x_{k\pm1}) \leq c_2\sqrt{h}, \tag{4.C.61}$$

$$-u_h(x_k) - c_2\sqrt{h} \leq -u_h(x_{k\pm1}), \tag{4.C.62}$$

$$u_h(x_k) + c_2\sqrt{h} \geq u_h(x_{k\pm1}). \tag{4.C.63}$$

Then it is assumed that $f(x, \cdot)$ is monotonically increasing, and the integration may be restricted to $\operatorname{supp}\varphi_{(k)}$, so to the neighboring nodes. Since u_h is piecewise linear,

$$u_h|_{\operatorname{supp}\varphi_{(k)}} \leq u_h(x_k) + c_2\sqrt{h} = \inf_{y\in\Omega} u_h(y) + c_2\sqrt{h}. \tag{4.C.64}$$

Then using the monotonicity of $f(x, \cdot)$ we obtain inequality (4.C.59).

Finally using inequalities (4.C.54), (4.C.57), (4.C.58), (4.C.59) we may estimate

$$\sum_{i=1}^{N} \int_{\Omega_i} \Big(f\big(x, \inf_{y\in\Omega} u_h(y) + c_2\sqrt{h}\big) - \inf_{y\in\Omega} g(y)\Big)\varphi_{(k)}(x)\, dx \geq 0. \tag{4.C.65}$$

Since $\varphi_{(k)}$ is nonnegative, there exists $x \in \Omega$, such that

$$f\big(x, \inf_{y\in\Omega} u_h(y) + c_2\sqrt{h}\big) - \inf_{y\in\Omega} g(y) \geq 0. \tag{4.C.66}$$

Rearranging the elements and using the notation $\tilde{f}_x(y) = f(x, y)$ we obtain

$$\tilde{f}_x\big(\inf_{y\in\Omega} u_h(y) + c_2\sqrt{h}\big) \geq \inf_{y\in\Omega} g(y). \tag{4.C.67}$$

By assumptions $\tilde{f}_x^{-1}$ exists. Since $\tilde{f}_x$ is monotonically increasing, then $\tilde{f}_x^{-1}$ is also monotonically increasing and we may apply it to the inequality to get

$$\inf_{y\in\Omega} u_h(y) + c_2\sqrt{h} \geq \tilde{f}_x^{-1}\left(\inf_{y\in\Omega} g(y)\right). \tag{4.C.68}$$

Therefore we finally obtain

$$\inf_{y\in\Omega} u_h(y) \geq \gamma_h' - c_2\sqrt{h} \geq \gamma_h. \tag{4.C.69}$$

□

Thus summarizing (4.C.53) and (4.C.69) we obtain

$$\begin{aligned} &\gamma_h \leq \gamma_h' - 3c_g^3\sqrt{h} \leq \inf_{\tilde{x}\in\Omega} u_h(\tilde{x}), \\ &\sup_{\tilde{x}\in\Omega} u_h(\tilde{x}) \leq \delta_h' + 3c_g^3\sqrt{h} \leq \delta_h. \end{aligned} \tag{4.C.70}$$

4.C.2.1.4 Conclusions We want to apply theorem 4.C.5 to the problem (4.C.6) to obtain bounds on $\tilde{u}_h$ independent of $\tilde{v}_h, \tilde{w}_h$. Therefore we do the following substitutions in (4.C.10):

- $f(x, y) := e^{y-\tilde{v}_h(x)} - e^{\tilde{w}_h(x)-y}$,
- $g \leftarrow k_1$,
- $u_h \leftarrow \tilde{u}_h$,
- $\hat{u} \leftarrow \hat{u}$.

We have to check the assumptions on f and $\tilde{f}_x$. Let $x \in \Omega$. Since assumptions on $\tilde{f}_x$ may be satisfied almost everywhere, we assume that $x \notin \Gamma_I$, as at these points there are two candidates for values of $\tilde{v}_h, \tilde{w}_h$. We may then compute

$$\tilde{f}_x'(y) = e^{y-\tilde{v}_h(x)} + e^{\tilde{w}_h(x)-y} > 0. \tag{4.C.71}$$

Therefore $\tilde{f}_x$ is a differentiable function for a.e. $x \in \Omega$ and moreover it is monotone increasing. Note that $\tilde{f}_x(y)$ is defined for every $y \in \mathbb{R}$. By assumptions of the problem (4.C.6), we have that $\alpha_h \leq \tilde{v}_h, \tilde{w}_h \leq \beta_h$. Therefore we may estimate

$$\begin{aligned} \tilde{f}_x(y) &= e^{y-\tilde{v}_h(x)} - e^{\tilde{w}_h(x)-y} \geq e^{y-\beta_h} - e^{\beta_h - y} =: f_1(y), \\ \tilde{f}_x(y) &= e^{y-\tilde{v}_h(x)} - e^{\tilde{w}_h(x)-y} \leq e^{y-\alpha_h} - e^{\alpha_h - y} =: f_2(y). \end{aligned} \tag{4.C.72}$$

So function $\tilde{f}_x$ is bounded a.e. by

$$f_1 \leq \tilde{f}_x \leq f_2, \tag{4.C.73}$$

where f_1, f_2 are independent of x.

Let $B \subset \mathbb{R}$ be a bounded set. Let us fix x. Then

$$\inf \tilde{f}_x(B) \geq \inf f_1(B) > -\infty, \quad \sup \tilde{f}_x(B) \leq \sup f_2(B) < \infty, \tag{4.C.74}$$

as $f_1, f_2 \in \mathcal{C}(\mathbb{R})$, so they preserve boundedness. Thus $\tilde{f}_x(B)$ is bounded uniformly for almost all x.

Still we have to prove similar result for f^{-1}. First goes the existence. Note that

$$\begin{aligned} \lim_{y\to-\infty} \tilde{f}_x(y) &= \lim_{y\to-\infty} e^{y-\tilde{v}_h(x)} - e^{\tilde{w}_h(x)-y} = [0-\infty] = -\infty, \\ \lim_{y\to\infty} \tilde{f}_x(y) &= \lim_{y\to\infty} e^{y-\tilde{v}_h(x)} - e^{\tilde{w}_h(x)-y} = [\infty-0] = \infty, \end{aligned} \tag{4.C.75}$$

so $\operatorname{rg} \tilde{f}_x = \mathbb{R}$. Analogously $\operatorname{rg} f_1 = \operatorname{rg} f_2 = \mathbb{R}$. We may therefore conclude that

$$f_2^{-1} \le \tilde{f}_x^{-1} \le f_1^{-1}. \tag{4.C.76}$$

Moreover $\operatorname{rg} \tilde{f}_x = \mathbb{R} = \operatorname{dom} \tilde{f}_x^{-1}$, so $\operatorname{rg}\ g \subset \operatorname{dom} \tilde{f}_x^{-1}$. Then analogously as for $\tilde{f}_x(B)$, we obtain that for almost all x, $\tilde{f}_x^{-1}(\operatorname{rg}\ g)$ is bounded uniformly, as $\operatorname{rg}\ g$ is a bounded set (because $g \in L^\infty(\Omega)$).

Then γ_h', δ_h' are finite, as $k_1 \in L_\infty(\Omega)$, so are γ_h, δ_h. Therefore we apply theorem 4.C.5 and we get bounds on $\tilde{u}_h$, independent of $\tilde{v}_h, \tilde{w}_h$, where

$$\begin{aligned} c_g := \max\{&|\max_\Omega k_1|, |\min_\Omega k_1|, \max_\Omega \varepsilon, \max_\Omega \varepsilon^{-1}, \|\nabla \hat{u}\|_{L_2(\Omega)}, |f_1(\gamma_h')|, |f_1(\delta_h')|, |f_2(\gamma_h')|, |f_2(\delta_h')|, \\ &|f_1^{-1}(\min_\Omega k_1(x))|, |f_1^{-1}(\max_\Omega k_1(x))|, |f_2^{-1}(\min_\Omega k_1(x))|, |f_2^{-1}(\max_\Omega k_1(x))|, |\gamma_h'|, |\delta_h'|, (|\Omega|)^{1/3}\}. \end{aligned} \tag{4.C.77}$$

Note that functions f_1, f_2 depend on α_h, β_h. However none of the above values depend on γ_h, δ_h, so the definitions of γ_h, δ_h (4.C.13) are well-posed and $c(\hat{u}, \varepsilon_m, \varepsilon_M, f, g) = c(c_g)$.

4.C.3 Discrete operator $v_h(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h)$

To obtain a discretization of the second equation (4.C.3), we also use the general discrete problem 1.3.3.

Problem 4.C.7. *Let $\hat{v} \in H^1(\Omega) \cap L_\infty(\Omega)$ and $\tilde{u}_h, \tilde{v}_h, \tilde{w}_h \in X_h(\Omega)$ be given functions. Find $v_h \in X_h(\Omega)$ such that*

$$a_{v,h}(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) + b_v(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) = f_{v,h}(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h), \tag{4.C.78}$$

where

$$\begin{aligned} a_{v,h}(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) &= \sum_{i=1}^N \int_{\Omega_i} \mu_n(x) e^{\tilde{u}_{h,i}(x) - \tilde{v}_{h,i}(x)} \nabla v_{h,i}(x) \nabla \phi_{h,i}(x)\, dx \\ &\quad + \sum_{e\in\Gamma_{DI}} \eta_{2,e} \int_e [v_h]_e [\phi_h]_e \, ds, \\ b_v(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) &= -\int_\Omega Q(\tilde{u}_h(x), v_h(x), \tilde{w}_h(x))(e^{\tilde{w}_h(x) - v_h(x)} - 1)\phi_h(x)\, dx, \\ f_{v,h}(v_h, \phi_h) &= \sum_{e\in\Gamma_D} \eta_{2,e} \int_e [\hat{v}]_e [\phi_h]_e \, ds. \end{aligned} \tag{4.C.79}$$

We would like to show the existence and uniqueness of the solution of problem 4.C.7. We are going to prove the following theorem.

Theorem 4.C.8. *Under assumptions A1 to A8, let $\tilde{u}_h, \tilde{v}_h, \tilde{w}_h \in X_h(\Omega)$ be functions as in theorem 4.C.3. Then the solution v_h of the problem 4.C.7 exists and it is unique. Moreover v_h is bounded:*

$$\alpha_h \le v_h \le \beta_h, \tag{4.C.80}$$

where constants $\alpha_h, \beta_h \in \mathbb{R}$ do not depend on $\tilde{v}_h, \tilde{w}_h$.

4.C.3.1 Existence

We would like to show the existence of the function v_h by Brouwer theorem 1.4.4. In this part we do not need to assume that $P \equiv 0$, thus we will prove the existence in the general case. We define $P : X_h(\Omega) \to X_h^*(\Omega)$ as

$$P(v_h)\phi_h := a_{v,h}(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) + b_v(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) + f_{v,h}(\phi_h). \tag{4.C.81}$$

Then

$$P(v_h)v_h := a_{v,h}(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, v_h) + b_v(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, v_h) + f_{v,h}(v_h). \tag{4.C.82}$$

First we note that

$$a_{v,h}(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, v_h) \geq c\|v_h\|^2_{h,\Sigma_2}, \tag{4.C.83}$$

where the constant c depends on $\alpha_h, \beta_h, \gamma_h, \delta_h$. By Schwarz inequality we have

$$\Big|f_{v,h}(v_h)\Big| \leq \|\hat{v}\|_{h,\Sigma_2}\|v_h\|_{h,\Sigma_2}. \tag{4.C.84}$$

Then

$$\begin{aligned} b_v(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, v_h) &= -\int_\Omega Q(\tilde{u}_h(x), v_h(x), \tilde{w}_h(x))(e^{\tilde{w}_h(x)-v_h(x)} - 1)v_h(x)\,dx \\ &= \int_\Omega Q(\tilde{u}_h(x), v_h(x), \tilde{w}_h(x))v_h(x)dx \\ &+ \int_\Omega Q(\tilde{u}_h(x), v_h(x), \tilde{w}_h(x))e^{\tilde{w}_h(x)}\Big(e^{-v_h(x)}\big(-v_h(x)\big)\Big)dx. \end{aligned} \tag{4.C.85}$$

We recall that $0 \leq Q \leq Q_M$. Thus

$$\Big|\int_\Omega Q(\tilde{u}_h(x), v_h(x), \tilde{w}_h(x))v_h(x)dx\Big| \leq Q_M\|v_h\|_{L_1(\Omega)} \leq C\|v_h\|_{h,\Sigma_2}. \tag{4.C.86}$$

On the other hand, note that $\forall x \in \mathbb{R} \quad xe^x \geq -e$, so

$$\int_\Omega Q(\tilde{u}_h(x), v_h(x), \tilde{w}_h(x))e^{\tilde{w}_h(x)}\Big(e^{-v_h(x)}\big(-v_h(x)\big)\Big)dx \geq -Q_M e^{\beta_h+1}|\Omega| \geq -c. \tag{4.C.87}$$

Therefore we have that

$$P(v_h)v_h \geq c(\|v_h\|^2_{h,\Sigma_2} - \|v_h\|_{h,\Sigma_2} - 1), \tag{4.C.88}$$

for some $c > 0$.

Then we show that P is continuous by element by element approach. Let

$$P(v_h) = P_a(v_h) + P_b(v_h) + P_f(v_h), \tag{4.C.89}$$

where

$$\begin{aligned} P_a(v_h)\phi_h &:= a_{v,h}(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h), \\ P_b(v_h)\phi_h &:= b_v(v_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h), \\ P_f(v_h)\phi_h &:= f_{v,h}(\phi_h). \end{aligned} \tag{4.C.90}$$

By Schwarz inequality

$$\begin{aligned} |P_a(v_h)\phi_h| &:= |a_{v,h}(v_h, \phi_h)| \\ &= \Big|\sum_{i=1}^N \int_{\Omega_i} \mu_n(x)e^{\tilde{u}_{h,i}(x)-\tilde{v}_{h,i}(x)}\nabla v_{h,i}(x)\nabla\phi_{h,i}(x)\,dx + \sum_{e\in\Gamma_{DI}} \eta_{2,e}\int_e [v_h]_e[\phi_h]_e\,ds\Big| \\ &\leq C\|v_h\|_{h,\Sigma_2}\|\phi_h\|_{h,\Sigma_2} \end{aligned} \tag{4.C.91}$$

Thus

$$\|P_a(v_h)\| = \sup_{\|\phi_h\|_{h,\Sigma_2}=1} |P_a(v_h)\phi_h| \leq C \sup_{\|\phi_h\|_{h,\Sigma_2}=1} \|v_h\|_{h,\Sigma_2}\|\phi_h\|_{h,\Sigma_2} = C\|v_h\|_{h,\Sigma_2}, \tag{4.C.92}$$

so P_a is bounded. It is linear, so then it is continuous. For P_f we have

$$P_f(v_h)\phi_h := f_{v,h}(\phi_h), \tag{4.C.93}$$

so it does not depend on v_h, so trivially it is continuous. Finally we have

$$P_b(v_h)\phi_h := -\int_\Omega Q(\tilde{u}_h(x), v_h(x), \tilde{w}_h(x))(e^{\tilde{w}_h(x)-v_h(x)} - 1)v_h(x)\,dx. \tag{4.C.94}$$

Take any $v_{h,(n)} \to v_h$ in $X_h(\Omega)$. We will show that $P(v_{h,(n)}) \to P(v_h)$. Since $X_h(\Omega)$ is a finite space, $\{v_{h,(n)}\}_n$ also converges in $\|\cdot\|_{L_2(\Omega)}$ and $\|\cdot\|_{L_\infty(\Omega)}$. Therefore $\|v_{h,(n)}\|_{L_\infty(\Omega)}$ are bounded uniformly.

Due to assumption A1, function $P(u,v,w) = Q(u,v,w)(e^{w-v} - 1)$ is locally Lipschitz-continuous. Thus since $\tilde{u}_h, \tilde{w}_h, v_h, v_{h,(n)}$ are all bounded, we may use Lipschitz-continuity to estimate

$$\begin{aligned}\Big|[P(v_{h,(n)}) - P(v_h)]\phi_h\Big| &= \Big|\int_\Omega Q(\tilde{u}_h(x), v_h(x), \tilde{w}_h(x))(e^{\tilde{w}_h(x)-v_h(x)} - 1)\phi_h(x)\,dx \\ &\quad - \int_\Omega Q(\tilde{u}_h(x), v_{h,(n)}(x), \tilde{w}_h(x))(e^{\tilde{w}_h(x)-v_{h,(n)}(x)} - 1)\phi_h(x)\,dx\Big| \\ &\leq \int_\Omega C|v_h(x) - v_{h,(n)}||\phi_h(x)|\,dx \\ &\leq C\|v_h(x) - v_{h,(n)}\|_{L_2(\Omega)}\|\phi_h(x)\|_{L_2(\Omega)}.\end{aligned} \tag{4.C.95}$$

Therefore

$$\|P(v_{h,(n)}) - P(v_h)\| = \sup_{\|\phi_h\|_{h,\Sigma_2}=1} \Big|[P(v_{h,(n)}) - P(v_h)]\phi_h\Big| \leq C\|v_h(x) - v_{h,(n)}\|_{L_2(\Omega)} \to 0. \tag{4.C.96}$$

Thus P_b is also continuous.

Therefore by theorem 1.4.4 there is some v_h, such that $P(v_h) = 0$. Existence is now proven.

4.C.3.2 Bounds and uniqueness

We begin with two abstract lemmas.

Lemma 4.C.9. *Let $n \in \mathbb{N}$, $\hat{y}_0, \hat{y}_{n+1} \in \mathbb{R}$ and $a = [a_1, \ldots, a_n], b = [b_1, \ldots, b_n]$ be given, so that $a_i > 0, b_i > 0$ for every $i \in \{1, \ldots, n\}$. Let $y = [y_1, \ldots, y_n] \in \mathbb{R}^n$ be a solution of equation*

$$\begin{bmatrix} a_1+b_1 & -b_1 & 0 & 0 & 0 & \ldots & 0 \\ -a_2 & a_2+b_2 & -b_2 & 0 & 0 & \ldots & 0 \\ 0 & -a_3 & a_3+b_3 & -b_3 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ldots & \vdots \\ 0 & \ldots & 0 & -a_{n-2} & a_{n-2}+b_{n-2} & -b_{n-2} & 0 \\ 0 & \ldots & 0 & 0 & -a_{n-1} & a_{n-1}+b_{n-1} & -b_{n-1} \\ 0 & \ldots & 0 & 0 & 0 & -a_n & a_n+b_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-2} \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} a_1\hat{y}_0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ b_n\hat{y}_{n+1} \end{bmatrix}. \tag{4.C.97}$$

Then either

$$\hat{y}_0 \leq y_1 \leq y_2 \leq \ldots \leq y_{n-1} \leq y_n \leq \hat{y}_{n+1}, \tag{4.C.98}$$

or

$$\hat{y}_0 \geq y_1 \geq y_2 \geq \ldots \geq y_{n-1} \geq y_n \geq \hat{y}_{n+1}. \tag{4.C.99}$$

Proof. To simplify the syntax, let us define

$$y_0 := \hat{y}_0, \quad y_{n+1} := \hat{y}_{n+1}. \tag{4.C.100}$$

For any $i \in \{1, \dots, n\}$ we will show that if $y_{i-1} \le y_i$ (resp. $y_{i-1} \ge y_i$), then $y_i \le y_{i+1}$ ($y_i \ge y_{i+1}$). By i-th line of the equation (4.C.97), we have

$$-a_i y_{i-1} + (a_i + b_i) y_i - b_i y_{i+1} = 0. \tag{4.C.101}$$

Rearranging elements, we obtain

$$a_i(y_i - y_{i-1}) + b_i y_i = b_i y_{i+1}. \tag{4.C.102}$$

Dividing both sides by $b_i > 0$ we obtain

$$y_{i+1} = y_i + \frac{a_i}{b_i}(y_i - y_{i-1}). \tag{4.C.103}$$

Note that $\frac{a_i}{b_i} > 0$. If $y_{i-1} \le y_i$, then $y_i - y_{i-1} \ge 0$, so y_{i+1} is not less than y_i. Similarly when $y_{i-1} \ge y_i$, then $y_i - y_{i-1} \le 0$, so y_{i+1} is not greater than y_i.

We will proceed by induction. First let $y_0 \le y_1$. Then, as we have shown already, $y_i \le y_{i+1}$ for any $i \in \{1, \dots, n\}$, so we obtain

$$y_0 \le y_1 \le y_2 \le \dots \le y_{n-1} \le y_n \le y_{n+1}. \tag{4.C.104}$$

On the other hand, if $y_0 \ge y_1$, then $y_i \ge y_{i+1}$ for any $i \in \{1, \dots, n\}$ and we have

$$y_0 \ge y_1 \ge y_2 \ge \dots \ge y_{n-1} \ge y_n \ge y_{n+1}. \tag{4.C.105}$$

Substituting $y_0 \leftarrow \hat{y}_0, y_{n+1} \leftarrow \hat{y}_{n+1}$ proves the lemma. □

Lemma 4.C.10. *Let $n \in \mathbb{N}$, $a = [a_2, \dots, a_n], b = [b_1, \dots, b_{n-1}], c = [c_1, \dots, c_n]$ be given, so that $a_i \neq 0, b_i \neq 0$ for every $i \in \{2, \dots, n-1\}$ and*

$$|c_i| \ge |a_i| + |b_i| \quad \forall i \in \{2, \dots, n-1\}, \tag{4.C.106}$$

$$|c_1| \ge |b_1|, \quad |c_n| \ge |a_n|, \tag{4.C.107}$$

and

$$|c_1| > |b_1| \text{ or } |c_n| > |a_n|. \tag{4.C.108}$$

Then the matrix A

$$A = \begin{bmatrix} c_1 & -b_1 & 0 & 0 & 0 & \dots & 0 \\ -a_2 & c_2 & -b_2 & 0 & 0 & \dots & 0 \\ 0 & -a_3 & c_3 & -b_3 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \dots & \vdots \\ 0 & \dots & 0 & -a_{n-2} & c_{n-2} & -b_{n-2} & 0 \\ 0 & \dots & 0 & 0 & -a_{n-1} & c_{n-1} & -b_{n-1} \\ 0 & \dots & 0 & 0 & 0 & -a_n & c_n \end{bmatrix}, \tag{4.C.109}$$

is non-singular.

$$\begin{bmatrix} 2.2\times10^{38} & -9.2\times10^{29} & 0 & 0 & 0 & 0 & 0 & 0 \\ -9.2\times10^{29} & 1.2\times10^{30} & -3.2\times10^{29} & 0 & 0 & 0 & 0 & 0 \\ 0 & -3.2\times10^{29} & 3.2\times10^{29} & -1.2\times10^{20} & 0 & 0 & 0 & 0 \\ 0 & 0 & -1.2\times10^{20} & 1.2\times10^{20} & -1.2\times10^{-9} & 0 & 0 & 0 \\ 0 & 0 & 0 & -1.2\times10^{-9} & 1.2\times10^{-9} & -1.6\times10^{-18} & 0 & 0 \\ 0 & 0 & 0 & 0 & -1.6\times10^{-18} & 1.6\times10^{-18} & -4.1\times10^{-23} & 0 \\ 0 & 0 & 0 & 0 & 0 & -4.1\times10^{-23} & 4.9\times10^{-23} & -7.9\times10^{-24} \\ 0 & 0 & 0 & 0 & 0 & 0 & -7.9\times10^{-24} & 9.7\times10^{-17} \end{bmatrix}$$

Figure 4.2: Example of a matrix A constructed for a numerical solution of v_h in a simulation of a simple p-n GaN diode.

Proof. See [36], lemma 10.10. □

Theorem 4.C.11. *Let us consider the general CWOPSIP discretization (see problem 1.3.3) in one dimension, with $f \equiv 0$:*

$$a_h(u_h^*, \phi_h) = f_h(\phi_h), \quad \forall \phi_h \in X_h(\Omega), \tag{4.C.110}$$

where

$$\begin{aligned} a(u_h, \phi_h) =& \sum_{i=1}^{N} \int_{\Omega_i} a \nabla u_{h,i} \cdot \nabla \phi_{h,i} \, dx, \\ a_h(u_h, \phi_h) =& a(u_h, \phi_h) + \sum_{e \in \Gamma_{DI}} \eta_e \int_e [u_h] \cdot [\phi_h] \, ds, \\ f_h(\phi_h) =& \sum_{e \in \Gamma_D} \eta_e \int_e [\hat{u}] \cdot [\phi_h] \, ds. \end{aligned} \tag{4.C.111}$$

Then it has an unique solution $u_h^* \in X_h(\Omega)$ such that

$$\min\{\hat{u}(e_0), \hat{u}(e_1)\} \leq u_h^* \leq \max\{\hat{u}(e_0), \hat{u}(e_1)\}, \tag{4.C.112}$$

for $\Gamma_D = \{e_0, e_1\}$, $\Omega = (e_0, e_1)$.

Proof. Since the general discrete problem 1.3.3 is linear, then it can be written in a matrix form. We will then use lemmas 4.C.10, 4.C.9. We make the following substitutions:

- $n \leftarrow J$,
- $\hat{y}_0 \leftarrow \hat{u}(e_0)$,
- $\hat{y}_{n+1} \leftarrow \hat{u}(e_1)$,
- $y \leftarrow u_h^*$.

Then we will show that the matrix has the form required by the lemmas. Since $f \equiv 0$, we solve the following problem: find $u_h \in X_h$ that for every $\phi_h \in X_h$

$$a_h(u_h, \phi_h) = a(u_h, \phi_h) + \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [u_h]_e [\phi_h]_e \, ds = \sum_{e \in \Gamma_D} \eta_{2,e} \int_e [\hat{u}]_e [\phi_h]_e \, ds, \tag{4.C.113}$$

where

$$a(u_h, \phi_h) := \sum_{i=1}^{N} \int_{\Omega_i} a \nabla u_{h,i} \cdot \nabla \phi_{h,i}\, dx. \tag{4.C.114}$$

Thus the matrix for this problem reads

$$A = \Big[a_h\big(\varphi_{(j)}, \varphi_{(k)}\big) \Big]_{j,k \in \{1,\dots,J\}}. \tag{4.C.115}$$

By definition, $a\Big(\varphi_{(j)}, \varphi_{(k)}\Big)$ may be non-zero only for $k \in \{j-1, j, j+1\} \cap \{1, \dots, J\}$. Also $\sum_{e \in \Gamma_{DI}} \eta_e [\varphi_{(j)}]_e [\varphi_{(k)}]_e$ is non-zero only for basis functions corresponding to the same interface, so $|j-k| \le 1$. Thus A is tridiagonal. We will check it has the required form by row-by-row verification.

For $j > 1$ we define

$$a_j := -a_h(\varphi_{(j)}, \varphi_{(j-1)}), \tag{4.C.116}$$

and

$$a_1 := a_h(\varphi_{(1)}, \varphi_{(1)}) - a(\varphi_{(1)}, \varphi_{(1)}). \tag{4.C.117}$$

For $j < J$ we define

$$b_j := -a_h(\varphi_{(j)}, \varphi_{(j+1)}), \tag{4.C.118}$$

and

$$b_J := a_h(\varphi_J, \varphi_J) - a(\varphi_J, \varphi_J). \tag{4.C.119}$$

Take any $j \in \{2, \dots, J\}$. Then

$$a_h(\varphi_{(j-1)}, \varphi_{(j)}) = a_h(\varphi_{(j)}, \varphi_{(j-1)}) = a(\varphi_{(j)}, \varphi_{(j-1)}) + \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [\varphi_{(j)}][\varphi_{(j-1)}]\, ds. \tag{4.C.120}$$

We have that

$$a(\varphi_{(j)}, \varphi_{(j-1)}) = \sum_{\tau \in \mathcal{T}} \int_\tau a \frac{d}{dx}\varphi_{(j)} \frac{d}{dx}\varphi_{(j-1)}. \tag{4.C.121}$$

Any element of this sum is zero unless $\tau = \mathrm{supp}(\varphi_{(j)}) \cap \mathrm{supp}(\varphi_{(j-1)})$ and then

$$-\int_\tau a \frac{d}{dx}\varphi_{(j)} \frac{d}{dx}\varphi_{(j-1)} = \int_\tau a \frac{d}{dx}\varphi_{(j)} \frac{d}{dx}\varphi_{(j)} = \int_\tau a \frac{d}{dx}\varphi_{(j-1)} \frac{d}{dx}\varphi_{(j-1)}. \tag{4.C.122}$$

Note that this case may only happen if x_{j-1} and x_j are nodes for some Ω_i, and $a(\varphi_{(j)}, \varphi_{(j-k)}) = a(\varphi_{(j-k)}, \varphi_{(j)}) = 0$ for any $k > 1$. On the other hand $\eta_{2,e} \int_e [\varphi_{(j)}][\varphi_{(j-1)}]\, ds$ is nonzero only if $\varphi_{(j)}$ and $\varphi_{(j-1)}$ correspond to interface between some Ω_i and Ω_{i+1}. In this case

$$-\eta_{2,e} \int_e [\varphi_{(j)}][\varphi_{(j-1)}]\, ds = \eta_{2,e} \int_e [\varphi_{(j)}][\varphi_{(j)}]\, ds = \eta_{2,e} \int_e [\varphi_{(j-1)}][\varphi_{(j-1)}]\, ds. \tag{4.C.123}$$

Note also that $\eta_{2,e} \int_e [\varphi_{(j)}][\varphi_{(j)}]\, ds$ is nonzero only if x_j lies on the boundary of some $\Omega_i \in \mathcal{E}$.

It is therefore clear that $a_j + b_j = a_h(\varphi_{(j)}, \varphi_{(j)})$. Also note that for $e = x_1 \in \Gamma_D$

$$a_1 = \eta_{2,e} \int_e [\varphi_{(1)}][\varphi_{(1)}]\, ds = \eta_{2,e} \int_e \varphi_{(1)}(x_1)\varphi_{(1)}(x_1)\, ds = \eta_{2,e}. \tag{4.C.124}$$

Thus

$$\eta_{2,e} \int_e [\hat{u}][\varphi_{(1)}]\, ds = \eta_{2,e} \int_e \hat{u}(e)\, ds = a_1 \hat{u}(e). \tag{4.C.125}$$

Analogously for $e = x_J \in \Gamma_D$

$$\eta_{2,e} \int_e [\hat{u}][\varphi_J]\, ds = b_J \hat{u}(e). \tag{4.C.126}$$

Thus indeed we see that the matrix A has a desired form. Then by lemma 4.C.10 we have that A is non-singular, as we substitute $c_i := a_i + b_i$ and $a_i, b_i > 0$, so the inequality $c_1 > b_1$ is obvious. Therefore the solution is unique. Then by lemma 4.C.9 we have that u_h is monotone and it is bounded by values of $\hat{u}$ on $\partial\Omega$, as stated in this theorem. □

In section 4.C.3 we have defined the discrete problem on v_h, as a special case of the general discrete problem (1.3.20). Thus we can use theorem 4.C.11 to obtain bounds as in theorem 4.C.8.

4.C.4 Discrete operator $w_h(\tilde{u}_h, \tilde{v}_h, \tilde{w}_h)$

In this case we proceed analogously to section 4.C.3. The discrete problem corresponding to equation (4.C.4) is as follows.

Problem 4.C.12. *Let $\hat{w} \in H^1(\Omega) \cap L_\infty(\Omega)$ and $\tilde{u}_h, \tilde{v}_h, \tilde{w}_h \in X_h(\Omega)$ be given functions. Find $w_h \in X_h(\Omega)$ such that*

$$a_{w,h}(w_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) + b_w(w_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) = f_{w,h}(w_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h), \tag{4.C.127}$$

where

$$\begin{aligned} a_{w,h}(w_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) &= \sum_{i=1}^{N} \int_{\Omega_i} \mu_p(x) e^{\tilde{w}_{h,i}(x) - \tilde{u}_{h,i}(x)} \frac{d}{dx} w_{h,i}(x) \frac{d}{dx} \phi_{h,i}(x)\, dx \\ &\quad + \sum_{e \in \Gamma_{DI}} \eta_{2,e} \int_e [w_h][\phi_h]\, ds, \\ b_w(w_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) &= \int_\Omega Q(\tilde{u}_h(x), \tilde{v}_h(x), w_h(x))(e^{w_h(x) - \tilde{v}_h(x)} - 1)\phi_h(x)\, dx, \\ f_{w,h}(w_h, \tilde{u}_h, \tilde{v}_h, \tilde{w}_h, \phi_h) &= \sum_{e \in \Gamma_D} \eta_{2,e} \int_e [\hat{w}][\phi_h]\, ds. \end{aligned} \tag{4.C.128}$$

Theorem 4.C.13. *Under assumptions A1 to A8, let $\tilde{u}_h, \tilde{v}_h, \tilde{w}_h \in X_h(\Omega)$ be functions as in theorem 4.C.3. Then the solution w_h of problem 4.C.127 exists and it is unique. Moreover w_h is bounded by*

$$\alpha_h \le w_h \le \beta_h. \tag{4.C.129}$$

where constants α_h, β_h are the same as in theorem 4.C.8.

Proof of this theorem is completely analogous to the proof of theorem 4.C.8.

4.C.5 Discretization of the van Roosbroeck system

For the discretization of the coupled van Roosbroeck system (problem 1.2.1), we will use operators introduced in sections 4.C.2, 4.C.3 and 4.C.4. Therefore the discrete problem is as follows.

Find $(u_h^*, v_h^*, w_h^*) \in \left(X_h(\Omega)\right)^3$, such that for every $\phi_h \in X_h(\Omega)$ we have

$$\begin{aligned} a_{u,h}(u_h^*, v_h^*, w_h^*, \phi_h) + b_u(u_h^*, v_h^*, w_h^*, \phi_h) &= f_{u,h}(v_h^*, u_h^*, v_h^*, w_h^*, \phi_h), \\ a_{v,h}(v_h^*, u_h^*, v_h^*, w_h^*, \phi_h) + b_v(v_h^*, u_h^*, v_h^*, w_h^*, \phi_h) &= f_{v,h}(v_h^*, u_h^*, v_h^*, w_h^*, \phi_h), \\ a_{w,h}(w_h^*, u_h^*, v_h^*, w_h^*, \phi_h) + b_w(w_h^*, u_h^*, v_h^*, w_h^*, \phi_h) &= f_{w,h}(w_h^*, u_h^*, v_h^*, w_h^*, \phi_h). \end{aligned} \tag{4.C.130}$$

We would like to prove existence of solutions of this system by arguing that operator $T : X_h^2(\Omega) \to X_h^2(\Omega)$ defined as

$$T(\tilde{v}_h, \tilde{w}_h) = \Big(v_h\big(\tilde{u}_h(\tilde{v}_h, \tilde{w}_h), \tilde{v}_h, \tilde{w}_h\big), w_h\big(\tilde{u}_h(\tilde{v}_h, \tilde{w}_h), \tilde{v}_h, \tilde{w}_h\big)\Big), \tag{4.C.131}$$

has a fixed point.

4.C.6 Analysis

So far we have proven the following results. Assume there is no recombination ($P \equiv 0$). Then there exist constants $\alpha_h, \beta_h, \gamma_h, \delta_h$ such that if $\alpha_h \le \tilde{v}_h, \tilde{w}_h \le \beta_h$, then $\alpha_h \le v_h, w_h \le \beta_h$ and $\gamma_h \le \tilde{u}_h \le \delta_h$, for $(v_h, w_h) := T(\tilde{v}_h, \tilde{w}_h)$ and $\tilde{u}_h = \tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$. Thus if we define $K := \{(u_h, v_h) \in X_h(\Omega) : \alpha_h \le u_h, v_h \le \beta_h\}$, then we may state that $T : K \mapsto K$.

K is a bounded closed subset of $X_h(\Omega)$, so it is compact as $X_h(\Omega)$ is a finite dimensional space. Also it is convex. Therefore we would like to use the Schauder fixed point theorem (theorem 4.A.1) to show that T has a fixed point. To do so, we must prove that T is a continuous function.

4.C.6.1 Continuity of $\tilde{u}_h$

First want to show that $\tilde{u}_h = \tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$ is a continuous operator.

Idea: let $\tilde{u}_h := \tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$ and $u_h := \tilde{u}_h(v_h, w_h)$. We will show that $\|\tilde{u}_h - u_h\| \le c\Big(\|\tilde{v}_h - v_h\| + \|\tilde{w}_h - w_h\|\Big)$, where $\|\cdot\|$ will be an appropriate norm (not necessarily the same for all the elements). Equivalence of norms in $X_h(\Omega)$ would be very helpful in these analysis.

Thus let $\tilde{u}_h := \tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$ and $u_h := \tilde{u}_h(v_h, w_h)$. By definition for every $\phi_h \in X_h(\Omega)$ we have

$$\begin{aligned}
\int_\Omega \varepsilon(x)\nabla\tilde{u}_h(x)\cdot\nabla\phi_h(x) + \Big(e^{\tilde{u}_h(x)-\tilde{v}_h(x)} - e^{\tilde{w}_h(x)-\tilde{u}_h(x)}\Big)\phi_h(x)dx + \sum_{e\in\Gamma_{DI}} \eta_{2,e}\int_e [\tilde{u}_h][\phi_h]\,ds \\
= \int_\Omega k_1(x)\phi_h(x)\,dx + \sum_{e\in\Gamma_D}\eta_{2,e}\int_e [\hat{u}][\phi_h]\,ds, \\
\int_\Omega \varepsilon(x)\nabla u_h(x)\cdot\nabla\phi_h(x) + \Big(e^{u_h(x)-v_h(x)} - e^{w_h(x)-u_h(x)}\Big)\phi_h(x)dx + \sum_{e\in\Gamma_{DI}} \eta_{2,e}\int_e [u_h][\phi_h]\,ds \\
= \int_\Omega k_1(x)\phi_h(x)\,dx + \sum_{e\in\Gamma_D}\eta_{2,e}\int_e [\hat{u}][\phi_h]\,ds.
\end{aligned} \tag{4.C.132}$$

Subtracting these equations we obtain

$$\begin{aligned}
&\int_\Omega \varepsilon(x)\Big(\nabla\tilde{u}_h(x) - \nabla u_h(x)\Big)\cdot\nabla\phi_h(x)\,dx \\
&\quad + \int_\Omega \Big(e^{\tilde{u}_h(x)-\tilde{v}_h(x)} - e^{\tilde{w}_h(x)-\tilde{u}_h(x)} - e^{u_h(x)-v_h(x)} + e^{w_h(x)-u_h(x)}\Big)\phi_h(x)\,dx \\
&\quad + \sum_{e\in\Gamma_{DI}}\eta_{2,e}\int_e [\tilde{u}_h - u_h][\phi_h]\,ds = 0.
\end{aligned} \tag{4.C.133}$$

If we then take $\phi_h := \tilde{u}_h - u_h$ and we substitute it into (4.C.133)

$$\|\tilde{u}_h - u_h\|_{h,\Sigma_2}^2 = \int_\Omega \Big(-e^{\tilde{u}_h(x)-\tilde{v}_h(x)} + e^{\tilde{w}_h(x)-\tilde{u}_h(x)} + e^{u_h(x)-v_h(x)} - e^{w_h(x)-u_h(x)}\Big)\Big(\tilde{u}_h(x) - u_h(x)\Big)dx. \tag{4.C.134}$$

Note that for any $a, b, c, d \in \mathbb{R}$ by monotonicity of the exponential function we have

$$\begin{aligned}\Big(\exp(c-d)-\exp(a-b)\Big)(a-c) &= \Big(\exp(c-d)-\exp(a-b)\Big)\Big((a-b)-(c-d)\Big)\\ &+\Big(\exp(c-d)-\exp(a-b)\Big)(b-d)\\ &\le \Big(\exp(c-d)-\exp(a-b)\Big)(b-d),\end{aligned} \tag{4.C.135}$$

and

$$\begin{aligned}\Big(\exp(c-d)-\exp(a-b)\Big)(d-b) &= \Big(\exp(c-d)-\exp(a-b)\Big)\Big((a-b)-(c-d)\Big)\\ &+\Big(\exp(c-d)-\exp(a-b)\Big)(c-a)\\ &\le \Big(\exp(c-d)-\exp(a-b)\Big)(c-a).\end{aligned} \tag{4.C.136}$$

Thus

$$\begin{aligned}&\int_\Omega\Big(-e^{\tilde u_h(x)-\tilde v_h(x)}+e^{\tilde w_h(x)-\tilde u_h(x)}+e^{u_h(x)-v_h(x)}-e^{w_h(x)-u_h(x)}\Big)\Big(\tilde u_h(x)-u_h(x)\Big)dx\\ &\quad=\int_\Omega\Big(e^{u_h(x)-v_h(x)}-e^{\tilde u_h(x)-\tilde v_h(x)}\Big)\Big(\tilde u_h(x)-u_h(x)\Big)dx\\ &\quad+\int_\Omega\Big(e^{\tilde w_h(x)-\tilde u_h(x)}-e^{w_h(x)-u_h(x)}\Big)\Big(\tilde u_h(x)-u_h(x)\Big)dx\\ &\quad\le\int_\Omega\Big(e^{u_h(x)-v_h(x)}-e^{\tilde u_h(x)-\tilde v_h(x)}\Big)\Big(\tilde v_h(x)-v_h(x)\Big)dx\\ &\quad+\int_\Omega\Big(e^{\tilde w_h(x)-\tilde u_h(x)}-e^{w_h(x)-u_h(x)}\Big)\Big(\tilde w_h(x)-w_h(x)\Big)dx\\ &\le\|e^{u_h-v_h}-e^{\tilde u_h-\tilde v_h}\|_{L_2(\Omega)}\|\tilde v_h-v_h\|_{L_2(\Omega)}+\|e^{\tilde w_h-\tilde u_h}-e^{w_h-u_h}\|_{L_2(\Omega)}\|\tilde w_h-w_h\|_{L_2(\Omega)}\\ &\le 2\max\Big\{\|e^{\delta_h-\alpha_h}\|_{L_2(\Omega)},\|e^{\beta_h-\gamma_h}\|_{L_2(\Omega)}\Big\}\Big(\|\tilde v_h-v_h\|_{L_2(\Omega)}+\|\tilde w_h-w_h\|_{L_2(\Omega)}\Big).\end{aligned} \tag{4.C.137}$$

Therefore together with (4.C.134) we obtain

$$\|\tilde u_h-u_h\|^2_{h,\Sigma_2}\le c\Big(\|\tilde v_h-v_h\|_{L_2(\Omega)}+\|\tilde w_h-w_h\|_{L_2(\Omega)}\Big), \tag{4.C.138}$$

where c depends on $\Omega, \alpha_h, \beta_h, \delta_h, \gamma_h$. Therefore due to equivalence of norms in $X_h(\Omega)$, we have that if $(v_h, w_h) \to (\tilde v_h, \tilde w_h)$, then $u_h \to \tilde u_h$. Therefore $\tilde u_h : K \mapsto X_h(\Omega)$ is a continuous operator.

4.C.6.2 Continuity of v_h and w_h

To prove continuity of the operators $v_h(\tilde u_h, \tilde v_h)$ and $w_h(\tilde u_h, \tilde w_h)$ we will use some estimates derived for the discrete operator of problem 1.3.3 with $f \equiv 0$.

4.C.6.2.1 Generalized case We would like to prove the following lemma.

Lemma 4.C.14. *Let $a, b \in L_\infty(\Omega) \cap L_2(\Omega)$. Let $u_h, v_h \in X_h(\Omega)$ be the solutions of*

$$\int_\Omega a(x)\frac{d}{dx}u_h\frac{d}{dx}\phi_h+\sum_{e\in\Gamma_{DI}}\eta_{2,e}\int_e[u_h][\phi_h]\,ds = \sum_{e\in\Gamma_D}\eta_{2,e}\int_e[\hat u][\phi_h]\,ds, \tag{4.C.139}$$

$$\int_\Omega b(x)\frac{d}{dx}v_h\frac{d}{dx}\phi_h+\sum_{e\in\Gamma_{DI}}\eta_{2,e}\int_e[v_h][\phi_h]\,ds = \sum_{e\in\Gamma_D}\eta_{2,e}\int_e[\hat u][\phi_h]\,ds, \tag{4.C.140}$$

for every $\phi_h \in X_h(\Omega)$. If $0 < a_m \le a, b \le a_M$ and $\|\frac{d}{dx} v_h\|_{L_\infty(\Omega)} < c$, where c is independent of b, and $b \to a$ in $L_2(\Omega)$, then $v_h \to u_h$ in $X_h(\Omega)$.

Remark 4.C.15. *In the above theorem in both cases $\hat{u}$ is a function used as a Dirichlet boundary condition.*

Proof. Subtracting (4.C.140) from (4.C.139) gives

$$\int_\Omega \Big(a(x)\frac{d}{dx}u_h(x) - b(x)\frac{d}{dx}v_h(x)\Big)\frac{d}{dx}\phi_h(x)\,dx + \sum_{e\in\Gamma_{DI}} \eta_{2,e}\int_e [u_h - v_h][\phi_h]\,ds = 0 \tag{4.C.141}$$

We may rewrite first element of the left hand side as

$$\begin{aligned}\int_\Omega \Big(a(x)\frac{d}{dx}u_h(x) - b(x)\frac{d}{dx}v_h(x)\Big)\frac{d}{dx}\phi_h(x)dx = &\int_\Omega a(x)\frac{d}{dx}\Big(u_h(x) - v_h(x)\Big)\frac{d}{dx}\phi_h(x)\,dx \\ &- \int_\Omega \big(b(x) - a(x)\big)\frac{d}{dx}v_h(x)\frac{d}{dx}\phi_h(x)\,dx\end{aligned} \tag{4.C.142}$$

For the purposes of this proof, if we substitute $\varepsilon := a$ in the definition of the broken norm (1.3.11), then using above result in equation (4.C.141) and taking $\phi_h := u_h - v_h$ we obtain

$$\|u_h - v_h\|^2_{h,\Sigma_2} = \int_\Omega \big(b(x) - a(x)\big)\frac{d}{dx}v_h(x)\frac{d}{dx}\big(u_h(x) - v_h(x)\big)\,dx. \tag{4.C.143}$$

Then we can estimate the right hand side using the Schwarz inequality

$$\int_\Omega \big(b(x) - a(x)\big)\frac{d}{dx}v_h(x)\frac{d}{dx}\big(u_h(x) - v_h(x)\big)\,dx \le \|b - a\|_{L_2(\Omega)}\Big\|\frac{d}{dx}v_h\Big\|_{L_\infty(\Omega)}\Big\|\frac{d}{dx}u_h - \frac{d}{dx}v_h\Big\|_{L_2(\Omega)}. \tag{4.C.144}$$

Therefore we have

$$\begin{aligned}\|u_h - v_h\|^2_{h,\Sigma_2} &\le \|b - a\|_{L_2(\Omega)}\Big\|\frac{d}{dx}v_h\Big\|_{L_\infty(\Omega)}\Big\|\frac{d}{dx}u_h - \frac{d}{dx}v_h\Big\|_{L_2(\Omega)} \\ &\le a_m^{-1/2}\|b - a\|_{L_2(\Omega)}\Big\|\frac{d}{dx}v_h\Big\|_{L_\infty(\Omega)}\|u_h - v_h\|_{h,\Sigma_2},\end{aligned} \tag{4.C.145}$$

where we used inequality

$$\|w_h\|^2_{h,\Sigma_2} = \int_\Omega a(x)\Big(\frac{d}{dx}w_h\Big)^2 + \sum_{i=1}^N \eta_{2,e}\int_e [w_h]^2\,ds \ge a_m\Big\|\frac{d}{dx}w_h\Big\|^2_{L_2(\Omega)}. \tag{4.C.146}$$

Assuming $u_h \ne v_h$ and dividing by $\|u_h - v_h\|_{h,\Sigma_2}$ we obtain

$$\|u_h - v_h\|_{h,\Sigma_2} \le a_m^{-1/2}\|b - a\|_{L_2(\Omega)}\Big\|\frac{d}{dx}v_h\Big\|_{L_\infty(\Omega)}. \tag{4.C.147}$$

Then if we fix a and thus also $u_h \in X_h(\Omega)$ and if $b \to a$ in $L_2(\Omega)$ and if $\|\frac{d}{dx}v_h\|_{L_\infty(\Omega)}$ is bounded, then $\|u_h - v_h\|_{h,\Sigma_2} \to 0$, so $v_h \to u_h$ in $X_h(\Omega)$. ☐

4.C.6.2.2 Operator v_h We would like to use lemma 4.C.14 to show the continuity of $v_h(\tilde{v}_h, \tilde{w}_h)$. Assume then that we have two functions $(\tilde{v}_h, \tilde{w}_h) \in K$ and a sequence $(\tilde{v}_{h,(n)}, \tilde{w}_{h,(n)}) \in K$ so that $\tilde{v}_{h,(n)} \to \tilde{v}_h, \tilde{w}_{h,(n)} \to \tilde{w}_h$ in $X_h(\Omega)$. Therefore we will do the following substitutions

- $a \leftarrow \mu_n \exp\big(\tilde{u}_h - \tilde{v}_h\big)$,
- $b \leftarrow \mu_n \exp\big(\tilde{u}_{h,(n)} - \tilde{v}_{h,(n)}\big)$,
- $u_h \leftarrow v_h = v_h(\tilde{v}_h, \tilde{w}_h)$,
- $v_h \leftarrow v_{h,(n)} = v_h(\tilde{v}_{h,(n)}, \tilde{w}_{h,(n)})$,
- $\tilde{u}_h = \tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$,
- $\tilde{u}_{h,(n)} = \tilde{u}_h(\tilde{v}_{h,(n)}, \tilde{w}_{h,(n)})$,
- $a_m \leftarrow \mu_m \exp(\gamma_h - \beta_h)$,
- $a_M \leftarrow \mu_M \exp(\delta_h - \alpha_h)$.

Then by definition of K and assumption A1 relative to μ_n (1.2.1), we have that $a_m \le a \le a_M$ and $a_m \le b \le a_M$.

Let $h_m := \min\{\operatorname{diam}\tau : \tau \in \mathcal{T}\}$. Then we have that

$$\Big\|\frac{d}{dx}v_h\Big\|_{L_2(\Omega)} = \Big\|\frac{d}{dx}v_{h,(n)}\|_{L_2(\Omega)} \le 2h_m^{-1}\|v_{h,(n)}\|_{L_\infty(\Omega)} \le 2h_m^{-1}\max\{|\alpha_h|, |\beta_h|\}, \tag{4.C.148}$$

thus the derivative of v_h is bounded by a constant independent of b.

Also note that since the exponential function is Lipschitz-continuous on any finite interval, we have

$$\begin{aligned} \|a-b\|^2_{L_2(\Omega)} &= \int_\Omega \mu_n^2(x)\Big(e^{\tilde{u}_h(x)-\tilde{v}_h(x)} - e^{\tilde{u}_{h,(n)}(x)-\tilde{v}_{h,(n)}(x)}\Big)^2\,dx \\ &\le \int_\Omega \mu_n^2 L_e^2\big(\tilde{u}_h(x) - \tilde{v}_h(x) - \tilde{u}_{h,(n)}(x) + \tilde{v}_{h,(n)}(x)\big)\,dx \\ &\le L_e^2\mu_M^2\|\tilde{u}_h - \tilde{v}_h - \tilde{u}_{h,(n)} + \tilde{v}_{h,(n)}\|_{L_2(\Omega)} \\ &\le L_e^2\mu_M^2\big(\|\tilde{u}_h - \tilde{u}_{h,(n)}\|_{L_2(\Omega)} + \|\tilde{v}_h - \tilde{v}_{h,(n)}\|_{L_2(\Omega)}\big), \end{aligned} \tag{4.C.149}$$

where L_e is a Lipschitz constant for exp on $[\gamma_h - \beta_h, \delta_h - \alpha_h]$. Then by the equivalence of norms in $X_h(\Omega)$ and continuity of operator $\tilde{u}_h$ (section 4.C.6.1) we have the following result. If $\tilde{v}_{h,(n)} \to \tilde{v}_h, \tilde{w}_{h,(n)} \to \tilde{w}_h$ in $X_h(\Omega)$, then $\tilde{u}_{h,(n)} \to \tilde{u}_h$ in $X_h(\Omega)$, an therefore also in $L_2(\Omega)$. Thus $b \to a$ in $L_2(\Omega)$.

Then all of the assumptions of lemma 4.C.14 are satisfied and therefore $v_{h,(n)} \to v_h$, so the operator $v_h(\tilde{v}_h, \tilde{w}_h)$ is continuous in K.

4.C.6.2.3 Operator w_h We will proceed similarly to section 4.C.6.2.2. Therefore we will do the following substitutions

- $a \leftarrow \mu_p \exp\big(\tilde{w}_h - \tilde{u}_h\big)$,
- $b \leftarrow \mu_p \exp\big(\tilde{w}_{h,(n)} - \tilde{u}_{h,(n)}\big)$,
- $u_h \leftarrow w_h = w_h(\tilde{v}_h, \tilde{w}_h)$,
- $v_h \leftarrow w_{h,(n)} = w_h(\tilde{v}_{h,(n)}, \tilde{w}_{h,(n)})$,

- $\tilde{u}_h = \tilde{u}_h(\tilde{v}_h, \tilde{w}_h)$,
- $\tilde{u}_{h,(n)} = \tilde{u}_h(\tilde{v}_{h,(n)}, \tilde{w}_{h,(n)})$,
- $a_m \leftarrow \mu_m \exp(\alpha_h - \delta_h)$,
- $a_M \leftarrow \mu_M \exp(\beta_h - \gamma_h)$.

Then by definition of K and assumption A1 relative to μ_p, we have that $a_m \leq a \leq a_M$ and $a_m \leq b \leq a_M$.

Then

$$\left\|\frac{d}{dx}v_h\right\|_{L_2(\Omega)} = \left\|\frac{d}{dx}w_{h,(n)}\right\|_{L_2(\Omega)} \leq 2h_m^{-1}\|w_{h,(n)}\|_{L_\infty(\Omega)} \leq 2h_m^{-1}\max\{|\alpha_h|, |\beta_h|\}, \tag{4.C.150}$$

and analogously to section 4.C.6.2.2

$$\begin{aligned}\|a-b\|^2_{L_2(\Omega)} &= \int_\Omega \mu_p^2(x)\Big(e^{\tilde{w}_h(x)-\tilde{u}_h(x)} - e^{\tilde{w}_{h,(n)}(x)-\tilde{u}_{h,(n)}(x)}\Big)^2 dx \\ &\leq \int_\Omega \mu_p^2 L_e^2\big(\tilde{w}_h(x) - \tilde{u}_h(x) - \tilde{w}_{h,(n)}(x) + \tilde{u}_{h,(n)}(x)\big)\,dx \\ &\leq L_e^2\mu_M^2\|\tilde{w}_h - \tilde{u}_h - \tilde{w}_{h,(n)} + \tilde{u}_{h,(n)}\|_{L_2(\Omega)} \\ &\leq L_e^2\mu_M^2\big(\|\tilde{u}_h - \tilde{u}_{h,(n)}\|_{L_2(\Omega)} + \|\tilde{w}_h - \tilde{w}_{h,(n)}\|_{L_2(\Omega)}\big).\end{aligned} \tag{4.C.151}$$

Thus if $\tilde{v}_{h,(n)} \to \tilde{v}_h, \tilde{w}_{h,(n)} \to \tilde{w}_h$ in $X_h(\Omega)$, then $\tilde{u}_{h,(n)} \to \tilde{u}_h$ in $X_h(\Omega)$, an therefore also in $L_2(\Omega)$. Thus $b \to a$ in $L_2(\Omega)$.

Then all of the assumptions of lemma 4.C.14 are satisfied and therefore $w_{h,(n)} \to w_h$, so the operator $w_h(\tilde{v}_h, \tilde{w}_h)$ is continuous in K.

4.C.6.3 Conclusions

In the previous subsections of these section we have proven that the operator T, defined as

$$T(\tilde{v}_h, \tilde{w}_h) := \Big(v_h\big(\tilde{u}_h(\tilde{v}_h, \tilde{w}_h), \tilde{v}_h\big), w_h\big(\tilde{u}_h(\tilde{v}_h, \tilde{w}_h), \tilde{w}_h\big)\Big) = \big(v_h \circ \tilde{u}_h, w_h \circ \tilde{u}_h\big)(\tilde{v}_h, \tilde{w}_h), \tag{4.C.152}$$

is a continuous mapping of convex compact $K \subset X_h(\Omega)^2$ into itself, as it is a composition of the continuous functions. Therefore using the Schauder fixed point theorem (theorem 4.A.1) we obtain existence of such functions v_h, w_h that

$$T(v_h, w_h) = (v_h, w_h). \tag{4.C.153}$$

Therefore functions $u_h := \tilde{u}_h(v_h, w_h) \in X_h(\Omega)$, $v_h \in X_h(\Omega)$ and $w_h \in X_h(\Omega)$ are a possibly non-unique solution of the system

$$a_{u,h}(u_h, v_h, w_h, \phi_h) = f_{u,h}(u_h, v_h, w_h, \phi_h), \tag{4.C.154}$$
$$a_{v,h}(v_h, u_h, v_h, w_h, \phi_h) = f_{v,h}(v_h, u_h, v_h, w_h, \phi_h), \tag{4.C.155}$$
$$a_{w,h}(w_h, u_h, v_h, w_h, \phi_h) = f_{w,h}(w_h, u_h, v_h, w_h, \phi_h). \tag{4.C.156}$$

for every $\phi_h \in X_h(\Omega)$.

4.D List of assumptions

Assumption A1.

1. *$\Omega \subset \mathbb{R}^d$ for $d \in \{1,2\}$, and it is an interval ($d = 1$) or a polygon ($d = 2$).*
2. *$0 \leq Q(u,v,w) \leq Q_M$ for any $u, v, w \in \mathbb{R}$.*
3. *$P(u,v,w)$ is monotone decreasing in v for $u, v, w \in \mathbb{R}$.*
4. *$P(u,v,w)$ is monotone increasing in w for $u, v, w \in \mathbb{R}$.*
5. *P is locally Lipschitz.*
6. *$0 < \varepsilon_m \leq \varepsilon(x) \leq \varepsilon_M$ for some $\varepsilon_m, \varepsilon_M \in \mathbb{R}$.*
7. *$k_1 \in L_\infty(\Omega)$.*
8. *μ_n, μ_p are Lipschitz continuous functions.*
9. *$0 < \mu_m \leq \mu_n(x), \mu_p(x) \leq \mu_M$ for some constants $\mu_m, \mu_M \in \mathbb{R}$.*

Assumption A2. *$\{\mathcal{T}_{i,h_i}(\Omega)\}_{h_i}$ is a quasi-uniform family of meshes (see definition 1.1.6).*

Assumption A3. *The coarse mesh $\mathcal{E}$ is chosen in such a manner so that Γ is a sum of disjoint sets Γ_D, Γ_N and Γ_I, where*

$$\begin{aligned} \Gamma_D &:= \{e \in \Gamma : e \subset \partial\Omega_D\}, \\ \Gamma_N &:= \{e \in \Gamma : e \subset \partial\Omega_N\}, \\ \Gamma_I &:= \{e \in \Gamma : e \subset \operatorname{int}(\Omega)\}. \end{aligned} \tag{1.3.6}$$

Assumption A4.

- *$\Gamma_D \neq \emptyset$.*
- *$\mathcal{T}_h$ is a shape regular mesh (see definition 1.1.5).*

Assumption A5. *$\hat{u} \in H^1(\Omega) \cap L_\infty(\Omega)$.*

Assumption A6.

- *$\Omega \subset \mathbb{R}^d$, $d \in \{1,2\}$, is an interval ($d = 1$) or a polygon ($d = 2$).*
- *$\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$.*
- *$\partial\Omega_D$ has nonzero measure.*

Assumption A7.

- *There is some $0 < h_M \leq 1$ such that for any $0 < h < h_M$ and for any $e \in \Gamma_{DI}$ we have $\eta_{e,r} \geq |e|^{-1}$ and $\eta_{e,r} \geq 1$ (cf. (1.3.12)).*
- *Constant h_M is sufficiently small, so that for any $0 < h < h_M$ lemma 1.3.7 holds.*
- (CSIPG only) *Constant $\sigma_m > 0$ is sufficiently large such that lemma 1.3.5 holds with $\alpha = 1/2$.*
- *$\varepsilon|_{\Omega_i} \in \mathcal{C}^1(\overline{\Omega_i})$ for every $\Omega_i \in \mathcal{E}$* (this assumption could be weakened, but in semiconductor simulations this function is normally constant or linear on Ω_i).

- *$u^* \in H^1(\Omega) \cap H^2(\mathcal{E})$, where u^* is a solution of problem 1.2.2.*
- *$\hat{v}, \hat{w} \in L_2(\Omega) \cap L_\infty(\Omega)$, where $\hat{v}, \hat{w}$ are defined in problem 1.2.2.*

Assumption A8.

- *$f : \Omega \times \mathbb{R} \mapsto \mathbb{R}$.*
- *$g \in L_\infty(\Omega)$.*
- *Let $\tilde{f}_x(y) := f(x, y)$ for fixed $x \in \Omega$. Then $\tilde{f}_x$ is a monotone increasing function for almost all $x \in \Omega$ (thus $\tilde{f}_x^{-1}$ exists for a.e. $x \in \Omega$ and it is monotone increasing).*
- *$\operatorname{rg}(g) \subset \operatorname{dom} \tilde{f}_x^{-1}$ and $\tilde{f}_x^{-1}\big(\operatorname{rg}(g)\big)$ is uniformly bounded set for almost all $x \in \Omega$.*
- *Let B be a bounded subset of $\mathbb{R}$. Then $\tilde{f}_x(B)$ is uniformly bounded set for almost all $x \in \Omega$.*
- *$\hat{u} \in H^1(\Omega) \cap X_h(\Omega)$.*
- *$P \equiv 0$.*

4.E Physical constants

Symbol	Value	Name
m_0	$9.109\,382\,15 \times 10^{-31}$ kg	electron rest mass
$\hbar$	$1.054\,571\,80 \times 10^{-34}$ J/s	reduced Planck constant
k_B	$1.380\,648\,52 \times 10^{-23}$ J/K	Boltzmann constant
q	$1.602\,176\,62 \times 10^{-19}$ C	elementary charge

Bibliography

[1] *SiLENSe Version 4.0 Physics Summary*. Richmond, 2009.

[2] I. Akasaki and M. Hashimoto. Infrared lattice vibration of vapour-grown AlN. *Solid State Communications*, 5(11):851–853, 1967.

[3] H. Amano, M. Kito, K. Hiramatsu, and I. Akasaki. P-type conduction in Mg-doped GaN treated with low-energy electron beam irradiation (LEEBI). *Japanese Journal of Applied Physics*, 28:L2112, 1989.

[4] H. Amano, N. Sawaki, I. Akasaki, and Y. Toyoda. Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AlN buffer layer. *Applied Physics Letters*, 48:353–355, 1986.

[5] Douglas N. Arnold. An Interior Penalty Finite Element Method with Discontinuous Elements. *SIAM Journal on Numerical Analysis*, 19(4):742–760, 1982.

[6] Douglas N. Arnold, Franco Brezzi, Bernardo Cockburn, and L. Donatella Marini. Unified Analysis of Discontinuous Galerkin Methods for Elliptic Problems. *SIAM Journal on Numerical Analysis*, 39(5):1749–1779, 2001.

[7] I. Babuska. The finite element method with penalty. *Mathematics of Computation*, 27(122):221–228, 1973.

[8] I. Babuska and M.R Dorr. Error Estimates for the Combined h and p Versions of the Finite Element Method. *Numerische Mathematik*, 37:257–278, 1981.

[9] I. Babuska, B. A. Szabo, and I. N. Katz. The p-Version of the Finite Element Method. *SIAM Journal on Numerical Analysis*, 18(3):515–545, 1981.

[10] Randolph E. Bank and Donald J. Rose. Some Error Estimates for the Box Method. *SIAM Journal of Numerical Analysis*, 24:777–787, 1987.

[11] A. T. Barker, S. C. Brenner, E.-H. Park, and L.-Y. Sung. Two-Level Additive Schwarz Preconditioners for a Weakly Over-Penalized Symmetric Interior Penalty Method. *Journal of Scientific Computing*, 47(1):27, 2011.

[12] F. Bassi and S. Rebay. A High-Order Accurate Discontinuous Finite Element Method for the Numerical Solution of the Compressible Navier–Stokes Equations. *Journal of Computational Physics*, 131(2):267–279, 1997.

[13] G. Birkhoff, M. H. Schultz, and R. S. Varga. Piecewise Hermite interpolation in one and two variables with applications to partial differential equations. *Numerische Mathematik*, 11(3):232–256, 1968.

[14] P. Bogusławski and J. Bernholc. Doping properties of C, Si, and Ge impurities in GaN and AlN. *Physical Review B*, 56(15):9496–9505, 1997.

[15] J. Borysiuk, K. Sakowski, P. Dróżdż, K. P. Korona, K. Sobczak, G. Muziol, C. Skierbiszewski, A. Kaminska, and S. Krukowski. Electric field dynamics in nitride structures containing quaternary alloy (Al, In, Ga)N. *Journal of Applied Physics*, 120:015702, 2016.

[16] R. Braunstein. Radiative Transitions in Semiconductors. *Physical Review*, 99(6):1892–1893, 1955.

[17] S. C. Brenner, L. Owens, and L.-Y. Sung. A weakly over-penalized symmetric interior penalty method. *Electronic Transactions on Numerical Analysis*, 30:107–127, 2008.

[18] Susanne C. Brenner. Poincare-Friedrichs Inequalities for Piecewise H1 Functions. *SIAM Journal on Numerical Analysis*, 41:306–324, 2004.

[19] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods.* Springer, New York, 2008.

[20] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from lagrangian multipliers. *ESAIM: Mathematical Modelling and Numerical Analysis*, 8(R2):129–151, 1974.

[21] M. Burger and R. Pinnau. A globally convergent Gummel map for optimal dopant profiling. *Mathematical Models and Methods in Applied Sciences*, 19(05):769, 2009.

[22] E. H. S. Burhop. *The Auger Effect and Other Radiationless Transitions.* Cambridge University Press, 2014.

[23] Graham F. Carey, A. L. Pardhanani, and S. W. Bova. Advanced Numerical Methods and Software Approaches for Semiconductor Device Simulation. *VLSI Design*, 10:391, 2000.

[24] Philippe Caussignac and Rachid Touzan. Solution of three-dimensional boundary layer equations by a discontinuous finite element method, part I: Numerical analysis of a linear model problem. *Computer Methods in Applied Mechanics and Engineering*, 78(3):249–271, 1990.

[25] Guy Chavent and Bernardo Cockburn. The local projection $P^0 - P^1$-discontinuous-Galerkin finite element method for scalar conservation laws. *ESAIM: Mathematical Modelling and Numerical Analysis*, 23(4):565–592, 1989.

[26] Fei Chen, A. N. Cartwright, Hai Lu, and William J. Schaff. Hole transport and carrier lifetime in InN epilayers. *Applied Physics Letters*, 87(21):212104, 2005.

[27] V. W. L. Chin, T. L. Tansley, and T. Osotchan. Electron mobilities in gallium, indium, and aluminum nitrides. *Journal of Applied Physics*, 75(11):7365–7372, 1994.

[28] Philippe G. Ciarlet. *The finite element method for elliptic problems.* North-Holland Publishing Company, 1978.

[29] Bernardo Cockburn and Chi-Wang Shu. The Local Discontinuous Galerkin Method for Time-Dependent Convection-Diffusion Systems. *SIAM Journal on Numerical Analysis*, 35(6):2440–2463, 1998.

[30] R. K. Coomer and I. G. Graham. Massively Parallel Methods for Semiconductor Device Modelling. *Computing*, 56:1–27, 1996.

[31] M. H. Crawford, J. J. Wierer, A. J. Fischer, G. T. Wang, D. D. Koleske, G. S. Subramania, M. E. Coltrin, R. F. Karlicek, and J. Y. Tsao. *Solid-State Lighting: Toward Smart and Ultraefficient Materials, Devices, Lamps, and Systems*, pages 1–56. John Wiley & Sons, Inc., 2015.

[32] E. A. Davis, S. F. J. Cox, R. L. Lichti, and C. G. Van de Walle. Shallow donor state of hydrogen in indium nitride. *Applied Physics Letters*, 82(4):592–594, 2003.

[33] John E. Dennis and Robert B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. Prentiece-Hall, Inc., Englewood Cliffs, 1983.

[34] R. Dingle, W. Wiegmann, and C. H. Henry. Quantum States of Confined Carriers in Very Thin $Al_xGa_{1-x}As$-GaAs-$Al_xGa_{1-x}As$ Heterostructures. *Physical Review Letters*, 33:827–830, 1974.

[35] Jim Douglas and Todd Dupont. Interior Penalty Procedures for Elliptic and Parabolic Galerkin Methods: Second International Symposium December 15–19,1975. In R. Glowinski and J. L. Lions, editors, *Computing Methods in Applied Sciences: Second International Symposium December 15–19,1975*, pages 207–216. Springer Berlin Heidelberg, Berlin, Heidelberg, 1976.

[36] M. Dryja, J. Jankowska, and M. Jankowski. *Przegląd metod i algorytmów numerycznych. Część 2*. WNT, Warszawa, 1982.

[37] Maksymilian Dryja. On Discontinuous Galerkin Methods for Elliptic Problems with Discontinuous Coefficients. *Computational Methods in Applied Mathematics*, 3(1):76–85, 2003.

[38] Maksymilian Dryja, Juan Galvis, and Marcus Sarkis. A FETI-DP Preconditioner for a Composite Finite Element and Discontinuous Galerkin Method. *SIAM Journal on Numerical Analysis*, 51:400, 2013.

[39] J. Edwards, K. Kawabe, G. Stevens, and R. H. Tredgold. Space charge conduction and electrical behaviour of aluminium nitride single crystals. *Solid State Communications*, 3(5):99–100, 1965.

[40] Lawrence C. Evans. *Partial Differential Equations*. American Mathematical Society, 1998.

[41] W. J. Fan, M. F. Li, T. C. Chong, and J. B. Xia. Electronic properties of zinc-blende GaN, AlN, and their alloys $Ga_{1-x}Al_xN$. *Journal of Applied Physics*, 79(1):188, 1996.

[42] C. P. Foley and T. L. Tansley. Pseudopotential band structure of indium nitride. *Physical Review B*, 33(2):1430–1433, 1986.

[43] Stefano Giani. Solving elliptic eigenvalue problems on polygonal meshes using discontinuous Galerkin composite finite element methods. *Applied Mathematics and Computation*, 267:618–631, 2015.

[44] Vivette Girault and Pierre-Arnaud Raviart. *Finite Element Methods for Navier-Stokes Equations*. Springer, Berlin Heidelberg, 1986.

[45] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[46] Izabella Grzegory, Stanisław Krukowski, Michael Leszczyński, Piotr Perlin, Tadeusz Suski, and Sylwester Porowski. High-Pressure Crystallization of GaN. In Pierre Ruterana, Martin Albrecht, and Jorg Neugebauer, editors, *Nitride Semiconductors: Handbook on Materials and Devices*, pages 1–43. Wiley, 2006.

[47] H. K. Gummel. A Self-Consistent Iterative Scheme for One-Dimensional Steady State Transistor Calculations. *IEEE Transactions on Electron Devices*, 11:455–465, 1964.

[48] R. N. Hall. Electron-hole recombination in germanium. *Physical Review*, 87:387, 1952.

[49] R. N. Hall, G. E. Fenner, J. D. Kingsley, T. J. Soltys, and R. O. Carlson. Coherent Light Emission From GaAs Junctions. *Physical Review Letters*, 9:366–368, 1962.

[50] H. J. Hovel and J. J. Cuomo. Electrical and Optical Properties of rf-Sputtered GaN and InN. *Applied Physics Letters*, 20(2):71–73, 1972.

[51] G. A. M. Hurkx, D. B. M. Klaassen, and M. P. G. Kmuvers. A New Recombination Model for Device Simulation Including Tunneling. *IEEE Transactions on Electron Devices*, 39:331–338, 1992.

[52] A. A. Huurdeman. *The Worldwide History of Telecommunications*. A Wiley-interscience publication. Wiley, 2003.

[53] Justin Iveland, Lucio Martinelli, Jacques Peretti, James S. Speck, and Claude Weisbuch. Direct Measurement of Auger Electrons Emitted from a Semiconductor Light-Emitting Diode under Electrical Injection: Identification of the Dominant Mechanism for Efficiency Droop. *Physical Review Letters*, 110:177406, 2013.

[54] Joseph W. Jerome. Consistency of Semiconductor Modeling: An Existence/Stability Analysis for the Stationary van Roosbroeck System. *SIAM Journal of Applied Mathematics*, 45(4):565–590, 1985.

[55] Joseph W. Jerome. The Role of Semiconductor Device Diameter and Energy-Band Bending in Convergence of Picard Iteration for Gummel's Map. *IEEE Transactions on Electron Devices*, 32:2045–2051, 1985.

[56] Joseph W. Jerome and Thomas Kerkhoven. A Finite Element approximation theory for the Drift Diffusion semiconductor model. *SIAM Journal of Numerical Analysis*, 28(2):403–422, 1991.

[57] C. Johnson and J. Pitkaranta. An Analysis of the Discontinuous Galerkin Method for a Scalar Hyperbolic Equation. *Mathematics of Computation*, 46(173):1–26, 1986.

[58] Nick Holonyak Jr. and S. F. Bevacqua. Coherent (visible) light emission from $Ga(As_{1-x}P_x)$ junctions. *Applied Physics Letters*, 1:82–83, 1962.

[59] Carl T. Kelley. *Iterative methods for optimization*. Society for Industrial and Applied Mathematics, 1999.

[60] T. Kerkhoven and J. W. Jerome. L_∞ stability of finite element approximations of elliptic gradient equations. *Numerische Mathematik*, 57:561, 1990.

[61] T. Kerkhoven and Y. Saad. On acceleration methods for coupled nonlinear elliptic systems. *Numerische Mathematik*, 60:525–548, 1992.

[62] Thomas Kerkhoven. A Proof of Convergence of Gummel's Algorithm for Realistic Device Geometries. *SIAM Journal on Numerical Analysis*, 23(6):1121–1137, 1986.

[63] Thomas Kerkhoven. A Spectral Analysis of the Decoupling Algorithm for Semiconductor Simulation. *SIAM Journal on Numerical Analysis*, 25(6):1299–1312, 1988.

[64] E. Knapp, R. Hausermann, H. U. Schwarzenbach, and B. Ruhstaller. Numerical simulation of charge transport in disordered organic semiconductor devices. *Journal of Applied Physics*, 108(5):054504, 2010.

[65] K. Lehovec, C. A. Accardo, and E. Jamgochian. Injected Light Emission of Silicon Carbide Crystals. *Physical Review*, 83(3):603–607, 1951.

[66] P. Lesaint and P. A. Raviart. On a Finite Element Method for Solving the Neutron Transport Equation. *Publications mathématiques et informatique de Rennes*, (S4):1–40, 1974.

[67] M. Leszczynski, T. Suski, J. Domagala, and P. Prystawko. Lattice parameters of the group III nitrides. In J. H. Edgar, S. Strite, I. Akasaki, H. Amano, and C. Wetzel, editors, *Properties, Processing and Applications of Gallium Nitride and Related Semiconductors*, pages 6–10. INSPEC, London, 1999.

[68] M.E. Levinshteĭn, S.L. Rumyantsev, and M. Shur. *Properties of advanced semiconductor materials: GaN, AlN, InN, BN, SiC, SiGe*. Wiley-Interscience publication. Wiley, 2001.

[69] J. L. Lions. *Quelques méthodes de résolution des problèmes aux limites non linéaires*. Dunod/Gauthier-Villars, Paris, 1969.

[70] E. Litwin-Staszewska, T. Suski, R. Piotrzkowski, I. Grzegory, M. Bockowski, J. L. Robert, L. Kończewicz, D. Wasik, E. Kamińska, D. Cote, and B. Clerjaud. Temperature dependence of electrical properties of gallium-nitride bulk single crystals doped with Mg and their evolution with annealing. *Journal of Applied Physics*, 89(12):7960, 2001.

[71] A.F.D. Loula, M.R. Correa, J.N.C. Guerreiro, and E.M. Toledo. On finite element methods for heterogeneous elliptic problems. *International Journal of Solids and Structures*, 45(25-26):6436, 2008.

[72] John L. Lyons, Audrius Alkauskas, Anderson Janotti, and Chris G. Van de Walle. First-principles theory of acceptors in nitride semiconductors. *physica status solidi (b)*, 252(5):900–908, 2015.

[73] L. Machiels. A posteriori finite element bounds for output functionals of discontinuous Galerkin discretizations of parabolic problems. *Computer Methods in Applied Mechanics and Engineering*, 190:3401–3411, 2001.

[74] P. A. Markowich, C. A. Ringhofer, and C. Schmeiser. *Semiconductor Equations*. Springer-Verlag, Wien, 1990.

[75] J. J. H. Miller, W. H. A. Schilders, and S. Wang. Application of finite element methods to the simulation of semiconductor devices. *Reports on Progress in Physics*, 62:277–353, 1999.

[76] M. S. Mock. On equations describing steady-state carrier distributions in a semiconductor device. *Communications on Pure and Applied Mathematics*, 25:781–792, 1972.

[77] B. Monemar. Fundamental energy gap of GaN from photoluminescence excitation spectra. *Physical Review B*, 10:676–681, 1974.

[78] B. Monemar, J. P. Bergman, I. A. Buyanova, H. Amano, I. Akasaki, T. Detchprohm, K. Hiramatsu, and N. Sawaki. The excitonic bandgap of GaN: Dependence on substrate . *Solid-State Electronics* , 41(2):239–241, 1997. Proceedings of the Topical Workshop on III-V Nitrides .

[79] H. Morkoç. *Handbook of Nitride Semiconductors and Devices: Electronic and optical processes in nitrides.* Handbook of Nitride Semiconductors and Devices. Wiley-VCH, 2008.

[80] Abdeljalil Nachaoui. Iterative solution of the drift-diffusion equations. *Numerical Algorithms*, 21:323–341, 1999.

[81] Shuji Nakamura, Takashi Mukai, and Masayuki Senoh. Candela-class high-brightness InGaN/AlGaN double-heterostructure blue-light-emitting diodes. *Applied Physics Letters*, 64:1687–1689, 1994.

[82] Shuji Nakamura, Masayuki Senoh, Shin-ichi Nagahama, Naruhito Iwasa, Takao Yamada, Toshio Matsushita, Hiroyuki Kiyoku, and Yasunobu Sugimoto. InGaN-based multi-quantum-well-structure laser diodes. *Japanese Journal of Applied Physics*, 35(1 B):L74, 1996.

[83] Marshall I. Nathan, William P. Dumke, Gerald Burns, Frederick H. Dill Jr. , and Gordon Lasher. Stimulated emission of radiation from GaAs p-n junctions. *Applied Physics Letters*, 1(3):62–64, 1962.

[84] Jindrich Necas. *Direct Methods in the Theory of Elliptic Equations.* Springer-Verlag, Berlin Heidelberg, 2012.

[85] J. Neugebauer and C. G. Van de Walle. Theory of point defects and complexes in GaN. *Materials Research Society Symposium Proceedings*, 395:645, 1996.

[86] Jörg Neugebauer and Chris G. Van de Walle. Gallium vacancies and the yellow luminescence in GaN. *Applied Physics Letters*, 69:503–505, 1996.

[87] B. Neuschl, K. Thonke, M. Feneberg, R. Goldhahn, T. Wunderer, Z. Yang, N. M. Johnson, J. Xie, S. Mita, A. Rice, R. Collazo, and Z. Sitar. Direct determination of the silicon donor ionization energy in homoepitaxial AlN from photoluminescence two-electron transitions. *Applied Physics Letters*, 103(12):122105, 2013.

[88] H. Obloh, K. H. Bachem, U. Kaufmann, M. Kunzer, M. Maier, A. Ramakrishnan, and P. Schlotter. Self-compensation in Mg doped p-type GaN grown by MOCVD. *Journal of Crystal Growth*, 195:270–273, 1998.

[89] P. Perlin, E. Litwin-Staszewska, B. Suchanek, W. Knap, J. Camassel, T. Suski, R. Piotrzkowski, I. Grzegory, S. Porowski, E. Kaminska, and J. C. Chervin. Determination of the effective mass of GaN from infrared reflectivity and Hall effect. *Applied Physics Letters*, 68(8):1114–1116, 1996.

[90] Daniele A. Di Pietro and Alexandre Ern. *Mathematical aspects of Discontinuous Galerkin Methods.* Springer, Berlin, 2012.

[91] S. J. Polak, C. den Heijer, and W. H. A. Schilders. Semiconductor device modelling from the numerical point of view. *International Journal for Numerical Methods in Engineering*, 24:763–838, 1987.

[92] S. J. Polak, C. den Heijer, W. H. A. Schilders, and P. Markowich. Semiconductor device modelling from the numerical point of view. *Journal for Numerical Methods in Engineering*, 24:763–838, 1987.

[93] W.H. Reed and T.R. Hill. Triangular mesh methods for the neutron transport equation. Technical Report LA UR-73-479, Los Alamos Laboratory, 1973.

[94] Beatrice Riviere. *Discontinuous Galerkin methods for solving elliptic and parabolic equations: theory and implementation*. Society for Industrial and Applied Mathematics, Philadelphia, 2008.

[95] D. L. Rode and D. K. Gaskill. Electron Hall mobility of n-GaN. *Applied Physics Letters*, 66(15):1972–1973, 1995.

[96] W. V. Van Roosbroeck. Theory of Flow of Electrons and Holes in Germanium and Other Semiconductors. *The Bell System Technical Journal*, 29:560–607, 1950.

[97] R. Sacco S. Micheletti, A. Quarteroni. Current-Voltage Characteristics Simulation of Semiconductor Devices Using Domain Decomposition. *Journal of Computational Physics*, 119:46–61, 1995.

[98] Konrad Sakowski, Leszek Marcinkowski, Stanislaw Krukowski, Szymon Grzanka, and Elzbieta Litwin-Staszewska. Simulation of trap-assisted tunneling effect on characteristics of gallium nitride diodes. *Journal of Applied Physics*, 111(12):123115, 2012.

[99] B. Šantí. On the determination of the statistical characteristics of the magnesium acceptor in GaN. *Superlattices and Microstructures*, 36:445–453, 2004.

[100] Leonard I. Schiff. *Mechanika Kwantowa*. PWN, Warszawa, 1977.

[101] Siegfried Selberherr. *Analysis and Simulation of Semiconductor Devices*. Springer-Verlag, Wien, 1984.

[102] Y. C. Shen, G. O. Mueller, S. Watanabe, N. F. Gardner, A. Munkholm, and M. R. Krames. Auger recombination in InGaN measured by photoluminescence. *Applied Physics Letters*, 91:141101, 2007.

[103] W. Shockley. Problems releated to p-n junctions in silicon. *Solid-State Electronics*, 2:35–67, 1961.

[104] W. Shockley and W. T. Read. Statistics of the recombinations of holes and electrons. *Physical Review*, 87(5):835–842, 1952.

[105] E. Silveira, J. A. Freitas, S. B. Schujman, and L. J. Schowalter. AlN bandgap temperature dependence from its optical properties . *Journal of Crystal Growth* , 310(17):4007–4010, 2008. Special issue IWBNS-5 International Workshop on Bulk Nitride Semiconductors V .

[106] J. Smalc-Koziorowska, S. Grzanka, E. Litwin-Staszewska, R. Piotrzkowski, G. Nowak, M. Leszczynski, P. Perlin, E. Talik, J. Kozubowski, and S. Krukowski. Ni-Au contacts to p-type GaN - structure and properties. *Solid-State Electronics*, 54:701–709, 2010.

[107] C. Stampfl, C. G. Van de Walle, D. Vogel, P. Krüger, and J. Pollmann. Native defects and impurities in InN: First-principles studies using the local-density approximation and self-interaction and relaxation-corrected pseudopotentials. *Physical Review B*, 61(12):R7846–R7849, 2000.

[108] T. Suski, P. Perlin, H. Teisseyre, M. Leszczyński, I. Grzegory, J. Jun, M. Boćkowski, S. Porowski, and T. D. Moustakas. Mechanism of yellow luminescence in GaN. *Applied Physics Letters*, 67(15):2188–2190, 1995.

[109] Masakatsu Suzuki and Takeshi Uenoyama. Strain effect on electronic and optical properties of GaN/AlGaN quantum-well lasers. *Journal of Applied Physics*, 80(12):6868–6874, 1996.

[110] T. Tanaka, A. Watanabe, H. Amano, Y. Kobayashi, I. Akasaki, S. Yamazaki, and M. Koike. P-type conduction in Mg-doped GaN and $Al_{0.08}Ga_{0.92}N$ grown by metalorganic vapor phase epitaxy. *Applied Physics Letters*, 65(5):593–594, 1994.

[111] T. L. Tansley and R. J. Egan. Defects, optical absorption and electron mobility in indium and gallium nitrides. *Physica B: Condensed Matter*, 185(1–4):190–198, 1993.

[112] Vidar Thomée. From finite differences to finite elements: A short history of numerical analysis of partial differential equations . *Journal of Computational and Applied Mathematics* , 128:1–54, 2001.

[113] A. Toselli and O. Widlund. *Domain Decomposition Methods - Algorithms and Theory.* Springer-Verlag, Berlin Heidelberg, 2005.

[114] M. J. Turner, R. W. Clough, H. C. Martin, and L. J. Topp. Stiffness and Deflection Analysis of Complex Structures. *Journal of the Aeronautical Sciences*, 23(9):805–823, 1956.

[115] Mary F. Wheeler Vivette Girault, Beatrice Riviere. A Discontinuous Galerkin Method with Nonoverlapping Domain Decomposition for the Stokes and Navier-Stokes Problems. *Mathematics of Computation*, 74(249):53–84, 2005.

[116] Tianhu Wang, Jinliang Xu, and Xiaodong Wang. Self-heating dependent characteristic of GaN-based light-emitting diodes with and without AlGaInN electron blocking layer. *Chinese Science Bulletin*, 59(20):2460–2469, 2014.

[117] Peter Wilkes. *Solid State Theory in Metallurgy.* Cambridge University Press, Cambridge, 1973.

[118] David Wells Windston. *SimWindows16 and SimWindows32 Version 1.4.2 User's Manual*, 1995. `http://ecee.colorado.edu/~bart/ecen6355/simwindows/manual.pdf`.

[119] J. Wu, W. Walukiewicz, W. Shan, K. M. Yu, J. W. Ager III, S. X. Li, E. E. Haller, Hai Lu, and William J. Schaff. Temperature dependence of the fundamental band gap of InN. *Journal of Applied Physics*, 94(7):4457–4460, 2003.

[120] Yong-Nian Xu and W. Y. Ching. Electronic, optical, and structural properties of some wurtzite crystals. *Physical Review B*, 48(7):4335–4351, 1993.

[121] H. Yamashita, K. Fukui, S. Misawa, and S. Yoshida. Optical properties of AlN epitaxial thin films in the vacuum ultraviolet region. *Journal of Applied Physics*, 50(2):896–898, 1979.

[122] Y. C. Yeo, T. C. Chong, and M. F. Li. Electronic band structures and effective-mass parameters of wurtzite GaN and InN. *Journal of Applied Physics*, 83(3):1429–1436, 1998.

[123] A. Yoshida. Bandedge and optical functions of AlN. In J. H. Edgar, S. Strite, I. Akasaki, H. Amano, and C. Wetzel, editors, *Properties, Processing and Applications of Gallium Nitride and Related Semiconductors*, pages 31–34. INSPEC, London, 1999.

[124] K. M. Yu, Z. Liliental-Weber, W. Walukiewicz, W. Shan, J. W. Ager, S. X. Li, R. E. Jones, E. E. Haller, Hai Lu, and William J. Schaff. On the crystalline structure, stoichiometry and band gap of InN thin films. *Applied Physics Letters*, 86(7):071910, 2005.

[125] N. Zheludev. The life and times of the LED - A 100-year history. *Nature Photonics*, 1(4):189–192, 2007.