Iteration acceleration for the Kohn-Sham system of DFT for semiconductor devices

vorgelegt von
Diplom-Mathematiker
Kurt Hoke
aus Berlin

Von der Fakultät II - Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
-Dr.rer.nat.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Peter Friz
Betreuer/Gutachter:	Prof. Dr. Reinhold Schneider
Gutachter:	Prof. Dr. Anton Arnold (Technische Universität Wien, Österreich)
Zus. Gutachter:	Prof. Dr. Vidar Gudmundsson (University of Iceland, Island)

Tag der wissenschaftlichen Aussprache
2. Juli 2010

Berlin, August 2010

D 83

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass die vorliegende Dissertationsschrift mit dem Titel *Iteration acceleration for the Kohn-Sham system of DFT for semiconductor devices* von mir selbständig und ohne Verwendung anderer als der angegebenen Hilfsmittel angefertigt wurde. Ferner versichere ich, dass alle Stellen, die im Wortlaut oder dem Sinn nach aus wissenschaftlichen Publikationen oder anderen Quellen entnommen sind, von mir als solche kenntlich gemacht wurden.

Datum

Unterschrift

Abstract

In Density Functional Theory the main object of relevance is the systems particle density, which, thanks to Hohenberg and Kohn [42], is known to uniquely determine the systems ground-state and all other properties. It was due to Kohn and Sham [57] to benefit from this information, by introducing a single-particle formulation for describing the multi-particle problem. For incorporating interaction effects, they included an exchange and correlation term in the effective potential. The resulting Kohn-Sham system is a nonlinearly coupled system of partial differential equations, that has to be solved self-consistently. In this work we are mainly interested in the quantum mechanical description of semiconductor heterostructures. Thus, we will look at the involved Schrödinger operator in *effective mass approximation* and have to deal with discontinuous coefficients and potentials.

For numerically treating the Kohn-Sham system, we will use the fixed point formulation on basis of the particle density. A commonly used scheme for solving this problem is the well-known *linear mixing* scheme, that corresponds to a damped *Picard* (or *Banach*) iteration. However, this method is known to suffer from slow convergence and thus the use of acceleration methods is advised. Using well-established acceleration schemes based on the *Newton*-method, is possible, but the numerical costs for computing the needed information about the Jacobian are quite big.

The aim of this work is the introduction of a fast and efficient acceleration method that generalises the *linear mixing* scheme to higher dimensions. The basis of our approach will be the *direct inversion in the iterative subspace* (DIIS) method from quantum chemistry. In *Hartree-Fock* and *Coupled Cluster* calculations DIIS is used to accelerate the calculation of electron orbitals. However, a straight forward transfer to our problem is dangerous. This is due to the extrapolation ability of the DIIS scheme, leading to negative mixing coefficients. Applied to our density approach, this may result in a negative density, meaning an iterate lying outside of the solution space. Thus, when applying the DIIS scheme to our problem, we have to ensure positivity of the produced density. We do this by introducing further constraints on the coefficients, that ensure positivity of the computed iterates. The resulting *convex* DIIS (CDIIS) scheme is then tested on exciton calculation in a three-dimensional quantum dot example. The results show, that the CDIIS method considerably accelerates the *linear mixing* approach, while in every step only a single function evaluation is performed. Thus, the CDIIS scheme accelerates the calculation while keeping the computational costs low and ensuring the quality of the iterates.

Zusammenfassung

Die Partikeldichte spielt in der Dichte-Funktional Theorie eine wesentliche Rolle und dank Hohenberg und Kohn [42] ist bekannt, dass diese eindeutig ist und das zugrundeliegende System bereits vollständig beschreibt. Mit dieser Grundlage, waren Kohn und Sham [57] anschließend in der Lage, das schwierig zu lösende Vielteilchen-Problem durch eine einfachere Einteilchen-Formulierung zu ersetzen. Wobei Wechselwirkungen durch Einfügen eines Austausch-Korrelations Terms in das effektive Potential eingebunden wurden. Ergebnis daraus war das bekannte Kohn-Sham System, das ein nichtlinear gekoppeltes partielles Differentialgleichungssystem darstellt, welches nun in selbst-konsistenter Weise gelöst werden muss.

In dieser Arbeit steht die quantenmechanische Beschreibung von Halbleiterbauelementen im Vordergrund, so dass wir den beteiligten Schrödinger Operator in der *Effektivmassen-Approximation* betrachten und auf Grund der Heterostrukturen springende Koeffizienten und Potentiale behandeln müssen.

Grundlage der numerischen Behandlung ist eine Fixpunkt-Formulierung auf Basis der Partikeldichte. Ein gängiges Verfahren zur Lösung dieses Problems ist das *linear mixing* Verfahren, welches einer gedämpften *Picard* oder *Banach* Iteration entspricht. Die schlechten Konvergenzeigenschaften von Dämpfungsverfahren, führt dann zu der Notwendigkeit Beschleunigungsverfahren einzusetzen. Etablierte Verfahren auf Basis eines *Newton*-Verfahrens können hier Abhilfe schaffen, allerdings ist der numerische Aufwand zur Berechnung von Informationen über den Jacobian immens.

Ziel dieser Arbeit ist es, ein effizientes Beschleunigungsverfahren zu entwickeln, welches eine hochdimensionale Verallgemeinerung des *linear mixing* Verfahrens darstellt. Ansatzpunkt dafür ist das *direct inversion in the iterative subspace* (DIIS) Verfahren aus der Quantenchemie. Allerdings birgt ein direkter Übertrag dieses Verfahrens auf unser Problem deutliche Gefahren. Grund dafür ist die Extrapolationseigenschaft des DIIS Verfahrens, welche negative Koeffizienten erzeugt. Angewandt auf unser Problem bedeutet dies, dass die zusammengesetzte Dichte möglicherweise negativ ist und somit bereits außerhalb des Lösungsraums liegt. Daher muss sichergestellt sein, dass bei Anwendung des DIIS Verfahrens nur positive Dichten erzeugt werden. Dies wird durch das Einführen zusätzlicher Bedingungen an die Koeffizienten erreicht, die die Positivität der Dichte garantieren.

Das resultierende Verfahren wird als *convex* DIIS bezeichnet. Als Testbeispiel wird die Exzitonenlokalisation in einem dreidimensional gerechneten Quantenpunkt verwendet. Die Ergebnisse zeigen, dass das CDIIS Verfahren eine wesentliche Beschleunigung gegenüber dem einfachen *linear mixing* Verfahren bedeutet und zudem lediglich eine einzige Funktionsauswertung pro Schritt erforderlich ist. Zusammenfassend kann gesagt werden, dass CDIIS die Fixpunktrechnung beschleunigt, aber gleichzeitig die Kosten pro Schritt gering hält und zudem die Qualität der berechneten Dichten garantiert.

Vorwort

In dieser Dissertation sind die Ergebnisse meiner Arbeit in der Forschungsgruppe *Partielle Differentialgleichungen* am *Weierstraß-Institut für angewandte Analysis und Stochastik* (WIAS) in Berlin zusammengefasst. Die Resultate sind dabei im Rahmen meiner Tätigkeit im Projekt D4 *quantum mechanical and macroscopic models for optoelectronic devices* des *DFG-Forschungszentrums MATHEON* entstanden. Die im analytischen Teil zusammengetragenen Aussagen enthalten Resultate zweier gemeinsamer Veröffentlichungen ([15, 43]), die u.a. mit Kollegen der Forschungsgruppe entstanden sind. Insbesondere mit Dr. Hans-Christoph Kaiser und Dr. Joachim Rehberg. Grundlage des im Rahmen meiner Tätigkeit entstandenen Programms zur Lösung des Kohn-Sham Systems in bis zu drei Raum-Dimensionen war die Toolbox $p\partial e$lib2, welche am WIAS entwickelt wurde. Alle numerischen Ergebnisse beziehen sich auf damit durchgeführte Rechnungen.

Danksagung: An erster Stelle möchte ich mich sehr herzlich bei Prof. Dr. Reinhold Schneider bedanken, der bereitwillig als Betreuer dieser Arbeit eingetreten ist und der mich mit dem DIIS Verfahren bekannt gemacht hat. Bei den Herren Prof. Dr. Anton Arnold und Prof. Dr. Vidar Gudmundsson möchte ich mich dafür bedanken, dass sie ohne Zögern bereit waren, Gutachten für diese Arbeit zu erstellen.

Dr. Hans-Christoph Kaiser und Dr. Joachim Rehberg standen mir im Rahmen des gemeinsamen Projektes D4 und auch darüber hinaus stets mit viel Rat zur Seite, wofür ich ihnen sehr danken möchte. Danken möchte ich zudem Dr. Hagen Neidhardt, Dr. Paul Racec und Dr. Thomas Koprucki, die meinem Verständniss zugrundeliegender physikalischer Prozesse durch zahlreiche Gespräche und Diskussionen auf die Sprünge halfen. Vielen Dank auch allen Kollegen des WIAS und hier insbesondere der Forschungsgruppe 1 für die immer angenehme und freundliche Atmosphäre am Institut. Es war eine Freude in den letzten Jahren in einer so hilfsbereiten und inspirierenden Umgebung zu arbeiten. Vielen herzlichen Dank.

Besonders bedanken möchte ich mich bei meiner Mutter Eleonore Hoke, die mir immer die Freiheit ließ mich zu entwickeln und meinen Weg zu finden. Sie unterstützte stets all meine Ziele; sei es der Besuch des Gymnasiums, die Aufnahme des Studiums oder die angestrebte Promotion. Mein besonderer Dank schließlich gilt meiner Freundin Anja Gursinsky, die mir während der letzten Jahre zur Seite stand und mich immer nach Kräften unterstützte. Ich danke ihr für ihre tiefe Liebe und die immense Gedult. Vielen lieben Dank dafür.

Berlin, im April 2010 Kurt Hoke

Contents

1 Introduction

The conceptual origin of the Kohn-Sham system is the fundamental paper written by Hohenberg and Kohn [42] in 1964. There, the authors showed that the ground state of a quantum mechanical multi-particle system is completely described by its ground state density and all properties can be viewed as functionals of this density. This result subsequently was carried over to finite-temperature situations (equilibrium state) by Mermin, [71]. Thus, in principle the wavefunctions for the ground state and all excited states of the interacting many-particle system are determined by one scalar function depending on the space coordinate only. However, the proofs for the so called Hohenberg-Kohn theorems stated in [42] are pure existence proofs and thus, beside the knowledge about the existence of a unique density, there was no guidance for constructing these functionals. Hence, one was still left with the task of solving the high dimensional many-particle Schrödinger equation describing the behaviour of the quantum system.

It was due to the *ansatz* made by Kohn and Sham [57] that provided a possibility of benefitting from this information. The major idea was to replace the many-particle equation by a single-particle formulation. Meanwhile escaping the curse of dimensionality of the full interacting system. Kohn and Sham introduced an effective potential that was designed to produce the same ground state density for the single-particle system as for the interacting many-particle system. In order to incorporate interaction effects, Kohn and Sham inserted the so called exchange-correlation (xc) term. Although known to exist, the precise appearance of the xc-term in its dependence on the particle density is still unknown.

Hence, making this ansatz practical one needs to establish approximations to the exchange-correlation term, which remains the major task in the Kohn-Sham approach. The most important type of approximation already used by Kohn and Sham, is the *local-density approximation* (LDA). LDA uses the exactly known exchange term for the homogeneous electron gas as a description for the xc-term in the Kohn-Sham equations. Even though LDA is known to be a poor approximation in situation where the density is strongly inhomogeneous, it shows a surprisingly success in actual calculations, cf. [21, 76, 19]. Beyond LDA that solely uses information about the density, numerous advanced approximations were developed over the years that aim on improving the exchange-correlation term. Among others, these are *generalised gradient approximations* (GGA), *meta-GGA's* or *hybrid* strategies. GGA's additionally incorporate information about the gradient of the density, whereas *meta-GGA's* use even more terms coming from the Taylor expansion. The mentioned *hybrid* strategies also include interaction terms from other *ab-initio* methods such as *Hartree-Fock*.

In our considerations we are mainly interested in the modelling of semiconductor devices which especially means a bounded spatial domain for the problem, which is given by the device domain. Furthermore, when regarding semiconductor devices, one is usually interested in semiconductor heterostructures resulting in jumping material parameters. The Schrödinger operator will be regarded in *effective mass approximation*, [96], leading to discontinuous effective potentials as well.

From a mathematical point of view the Kohn-Sham system is a stationary Schrödinger-Poisson system with self-consistent effective Kohn-Sham potential. The equations are coupled via the electrostatic potential and the particle densities. The occurring partial differential equations have to be supplemented by in general mixed boundary conditions of Dirichlet and Neumann type, cf. e.g. [26, 24]. For the special cases of only one kind of particles, homogeneous boundary conditions on both Poisson's and Schrödinger's equation and without exchange-correlation potential results on the existence of solutions are carried out in the work by Kaiser and Rehberg [46] and Nier [72, 73]. Without exchange-correlation potential the Schrödinger-Poisson system is a nonlinear Poisson equation in the dual of a Sobolev space determined by the boundary conditions imposed on the electrostatic potential. The involved operator is strongly monotone and boundedly Lipschitz continuous, [45]. Hence, the operator equation has a unique solution, cf. [46, 73]. In presence of an exchange-correlation term, the system can no longer be written as a monotone operator. Instead, the result for the system without exchange and correlation can be used to set up a fixed point mapping, which meets the conditions of Schauder's fixed point theorem giving existence of solutions, cf. Kaiser and Rehberg [47, 49, 50, 48]. The results just stated include mixed Dirichlet-Neumann boundary conditions as well as jumping coefficients that model heterointerfaces. The underlying statistics, e.g. represented by Fermi's function cf. [71], are assumed differentiable. Cornean, H., Neidhardt, Racec and Rehberg showed in [15] that for effectively one-dimensional problems this assumption can be softened to only asking for continuity which then gives existence of solutions for the zero-temperature case. Motivated by a result from Gajewski and Griepentrog [27] about a descent method for the free energy of a multi-component system, there are indications that analyticity might be an advantageous property for setting up steadily converging iteration schemes. The corresponding result about analyticity of the particle density operator was shown by H., Kaiser and Rehberg in [43].

Concerning the numerical treatment of the Kohn-Sham system, there are mainly two possibilities: i) iterative methods for finding the self-consistent solution by a fixed point procedure, ii) direct approaches for determining the minimum of the total-energy functional. The latter are based on the fact that the Kohn-Sham energy functional is minimal at the ground (equilibrium) state. Minimising schemes usually are conjugate gradient (CG) type approaches, meaning minimisation along a given search direction that is conjugate to previous directions. A main problem in these kind of approaches remains the incorporation of the orthonormality conditions for the orbitals, cf. [60]. In this work we will focus on methods of type i), the self-consistent solution by means of fixed point iterations. The basis of such methods is a reformulation of the Kohn-Sham system as a fixed point mapping. Given an input-density n_i^{in} at iteration i, the corresponding effective potential is calculated. With this, the eigenstates of the resulting Hamiltonian are computed which then are used for the composition of an output-density n_i^{out}. This procedure is called the *self-consistent loop* (SCL). In terms of fixed point iterations this just means a *Picard* (or *Banach*) mapping. The simplest consequential approach following the SCL idea is the *linear mixing* method, cf. [68], where an improved input-density is computed as a fixed linear combination of the

previous input- and output-densities:

$$n_{i+1}^{in} = \alpha n_i^{out} + (1-\alpha) n_i^{in} = n_i^{in} + \alpha (n_i^{out} - n_i^{in}) \,, \quad \alpha \in (0,1] \,. \tag{1.1}$$

This corresponds to a fixed damping of the underlying *Picard* iteration and as such it often suffers from slow convergence, due to strong damping, i.e. $\alpha \ll 1$. However, a merit of the *linear mixing scheme* is the non-expensiveness of a single iteration step. In absence of further information $n_i^{out} - n_i^{in}$ is considered the best choice for a *steepest descent* direction. More advanced mixing schemes additionally use information about the Jacobian, e.g. *Broyden* [8, 93] or *Newton*-type [53] approaches. However, this includes storage of the Jacobian itself or at least a growing amount of information about it during the procedure making it storage expensive.

In this work we develop a high dimensional generalisation of the *linear mixing* scheme which on the one hand leads to a considerable acceleration of the process, making it comparable to Newton-type schemes, and on the other hand keeping the computational cost of a single step low.

The starting point of our considerations beside the iteration in a SCL, will be the *direct inversion in the iterative subspace* (DIIS) procedure invented by Pulay in 1980, [80]. This subspace acceleration method is used in *ab-initio* calculations like *Hartree-Fock*, [40], or *Coupled Cluster*, [40, 97], from quantum chemistry. It is used there for accelerating the calculation of orbital sets. In our formulation of the fixed point mapping for the Kohn-Sham system the main object of relevance is the density n instead of the orbitals, which only appear in the composition of the density. Thus, we have to deal with a certain speciality of the DIIS method. Namely its extrapolation ability, allowing the mixed states to lie outside of the convex hull. This however, is devastating for our density approach since the produced densities may be located outside of the solution space, i.e. be negative at some point in real space. To make use of the DIIS approach anyway, we thus have to make sure the mixing scheme yields a density that is positive in every point. We will do this by imposing further constraints on the mixing coefficients. This results in an increase of computational effort for determining the coefficients but this is a low dimensional problem compared to the dimension of the original problem. With this we were able to transform the orbital-based DIIS procedure in an adequate density-based one which is quite different in nature, e.g. introducing further constraints on the coefficients but dropping the orthonormality constraint of the mixed objects. The use of the pure DIIS scheme in DFT is sometimes done in quantum chemistry and is then denoted as *Pulay mixing*. Actually, this approach may work, but it does not respect the nature of the underlying problem, e.g. positivity of the density. Though used this attempt is known to sometimes suffer from slow convergence or even fail, cf. Harrison [39]. In this situations one tries to help out by introducing a damping like in (1.1) on the just calculated optimal iterate. This however, leads to similar problems in the convergence as in simple *linear mixing* and additionally another solution of the Schrödinger eigenvalue problem has to be calculated in every step, making it even more expensive.

We call our resulting acceleration scheme for the density based iteration procedure *convex* DIIS (CDIIS). The CDIIS scheme only takes a single SCL iteration and thus is as cheap as *linear mixing*. Furthermore, the numerical tests show a performance similar to a comparison scheme, [54, 56, 53, 55, 52, 53], based on a *Newton*-type acceleration. Since the computationally most expensive part of the SCL is the solution of the Schrödinger eigenvalue problem, we will judge the performance of the schemes by the number of solved eigenvalue problems. We thus end up with a cheap but fast scheme for finding the self-consistent solution to the Kohn-Sham system by simultaneously ensuring the iterates lying in the solution space.

The organisation of the work is as follows. In Section 2 we will give an overview of the origin of the Kohn-Sham system with a focus on semiconductor devices. Section 3 is devoted to analytical considerations like existence of solutions and properties of the particle density operator. A detailed description of the physical example used for the numerical tests can be found in Section 4. Finally, Section 5 deals with the numerical treatment of the Kohn-Sham system. There, we will present some mathematical background of the DIIS scheme and embed it in a general acceleration framework based on a main iteration and an error rating. Furthermore, we will introduce the CDIIS method and compare it to the simple *linear mixing* scheme and a Newton-type accelerated scheme.

2 The Kohn-Sham System

In the first section we will have a look on the origin of the Kohn-Sham system with special interest in its appearance in semiconductor device modelling. The Kohn-Sham theory is an approach of describing a quantum mechanical many-particle problem in a single-particle formulation, meanwhile providing a possibility of escaping the curse of dimensionality. There are many textbooks giving a detailed introduction to the topic, e.g. [76, 61, 19, 21, 10, 68]. Hence, we will only give a short overview, aiming on the application we have in mind.

To start with, we will describe the full many-particle problem and the corresponding operators arising in quantum mechanics. For comparison we then describe the well-known *Hartree-Fock* method, that uses *Slater determinants* in order to reach a single-particle formulation of the problem. By use of the *Hohenberg-Kohn theorems*, we then get to the Kohn-Sham system, which is an effective single-particle formulation similar in structure to the *canonical* Hartree-Fock equations. Afterwards, the *local density approximation* (LDA) will be used to represent the *exchange-correlation* potential. And finally, the resulting Schrödinger-Poisson system that is adequate for describing semiconductor heterostructures is introduced.

2.1 Modelling Aspects

The electronic Schrödinger Equation

In quantum mechanics the dynamics of a many-particle system is described by the electronic Schrödinger equation. When regarding a time-independent N-particle system in *Born-Oppenheimer* approximation, the corresponding eigenvalue equation for the systems wavefunction $\Psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ reads

$$H\Psi = E\Psi\,. \tag{2.1}$$

Where the Hamilton operator is given by

$$H = \sum_{i=1}^{N}(-\frac{1}{2}\nabla_i^2) + \sum_{i=1}^{N} v(x_i) + \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N} \frac{1}{|x_i - x_j|}\,. \tag{2.2}$$

Here, E denotes the electronic energy. The degrees of freedom $\mathbf{x_i}$ split up in a space coordinate $x_i \in \mathbb{R}^3$ and a spin degree of freedom s_i, i.e. $\mathbf{x_i} = (x_i, s_i)$. The function $v(x_i)$ is the potential of the external field acting on electron i due to the nuclei α of charges Z_α at positions a_α

$$v(x_i) = -\sum_{\alpha} \frac{Z_\alpha}{|x_i - a_\alpha|}\,. \tag{2.3}$$

Remark 2.1. *Throughout this work we will use atomic units. These are the Bohr radius a_0 ($= 0.5292\mathring{A}$) as a length unit, the elementary charge q ($= 1.6022 * 10^{-19}C$) as a charge unit and the mass of the electron m_e ($= 9,1094 * 10^{-31}kg$) as a mass unit. See Appendix A for a complete discussion of the scaling. Furthermore, we will assume the orbitals to be doubly occupied and thus, neglect the spin variable s_i in the following.*

More compactly we write

$$H = T + V_{ne} + V_{ee}\,,$$

the individual terms represent the kinetic energy operator

$$T = \sum_{i=1}^{N}(-\frac{1}{2}\nabla_i^2)\,, \tag{2.4}$$

the nucleus-electron attraction operator

$$V_{ne} = \sum_{i=1}^{N} v(x_i)\,, \tag{2.5}$$

and the electron-electron repulsion operator

$$V_{ee} = \sum_{\substack{i,j=1\\ i\neq j}}^{N} \frac{1}{|x_i - x_j|}\,. \tag{2.6}$$

To get the systems total energy W we lastly need to add the nucleus-nucleus repulsion energy

$$V_{nn} = \sum_{\alpha<\beta} \frac{Z_\alpha Z_\beta}{|a_\alpha - a_\beta|}\,. \tag{2.7}$$

That is,

$$W = E + V_{nn}\,. \tag{2.8}$$

Remark 2.2. *It makes no difference, whether (2.1) is solved using E and adding W afterwards or solving the Schrödinger equation with W instead of E, cf. [21, 76].*

Solutions to (2.1) are given by the eigenfunctions Ψ_k together with the corresponding eigenvalues E_k of the operator H. The set $\{\Psi_k\}$ is complete and may be taken orthogonal and normalised (in $L^2((\mathbb{R}^3)^N)$),

$$\int \Psi_k^*(x)\Psi_l(x)\,\mathrm{d}x = \langle \Psi_k, \Psi_l\rangle = \delta_{kl}\,. \tag{2.9}$$

Moreover, we expect the particles to have an indistinguishability property. This means, that we are not able to discriminate between two particles, and thus, have to look for a

solution in a symmetry constraint subspace of L^2. More precisely, we look at fermions whose wavefunction Ψ has to be *anti-symmetric* in the sense that for every permutation P of particle coordinates $\mathbf{x} = (x_1, \ldots, x_N)$ we have

$$\Psi(P\mathbf{x}) = \operatorname{sign}(P)\Psi(\mathbf{x})\,.$$

For an arbitrary wavefunction Ψ the expectation value of the energy is given by

$$E[\Psi] = \frac{\langle \Psi, H\Psi\rangle}{\langle \Psi, \Psi\rangle}\,.$$

In quantum mechanics one is usually interested in the systems ground state, i.e. a wavefunction Ψ_0 and the corresponding energy $E[\Psi_0]$ solving the minimisation problem

$$E[\Psi_0] = \min_{\Psi} E[\Psi] = \min_{\Psi}\{\langle \Psi, H\Psi\rangle \,:\, \langle \Psi, \Psi\rangle = 1\}\,.$$

Using calculus of variations the latter formulation can be written as

$$d[\langle \Psi, H\Psi\rangle - E(\langle \Psi, \Psi\rangle - 1)] = 0\,,$$

where the energy E, serves as the Lagrangian multiplier associated to the normalisation constraint $\langle \Psi, \Psi\rangle = 1$. Since we are interested in states of finite energy we have to assume L^2 integrability of the first derivative in order to give the kinetic energy operator T a proper sense. Hence, we further restrict the solution space from $L^2((\mathbb{R}^3)^N)$ to $W^{1,2}((\mathbb{R}^3)^N)$.

Unfortunately, this problem cannot be tackled for more than a few particles (small N), due to the high dimensionality. Thus, in order to handle this task one has to find ways of either decreasing the dimension itself (e.g. replacing the problem by an easier one) or choosing an adequate low-dimensional subspace to search the solution in. The latter approach is use by the *Hartree-Fock* (HF) method, which will be described in the following. After that we will present the idea of using *density functional theory* (DFT) to replace the $3N$ dimensional problem by a three dimensional one. Although the ideas behind HF and DFT are quite different in nature, the resulting systems are of similar structure.

The Hartree-Fock Approximation

One of the most commonly used procedures in approximately calculating the systems ground-state, is the *Hartree-Fock* (HF) approximation. The basic idea is to use a separation of variables in the form

$$\Psi(x_1, \ldots, x_N) = \varphi(x_1)\ldots\varphi(x_N)\,.$$

This however, does not respect the anti-symmetry constraint (the Pauli principle) for fermions. To fulfil this condition the wavefunction is written as a *Slater determinant*

$$\Psi_{HF}(x_1, \ldots, x_N) = \frac{1}{\sqrt{N!}}\det(\varphi_i(x_j)) = \frac{1}{\sqrt{N!}}\sum_{\sigma\in S_n}\operatorname{sign}(\sigma)\prod_{i=1}^{N}\varphi_i(x_{\sigma(i)})\,, \tag{2.10}$$

where the sum goes over all permutations of the numbers $\{1,\dots,N\}$ (Leibniz formula). The *ansatz*-functions φ_i, called orbital functions, essentially are one-particle wavefunctions lying in $W^{1,2}(\mathbb{R}^3)$. When further applying an orthogonality condition on the orbitals, i.e.

$$\langle \varphi_i, \varphi_j \rangle = \delta_{ij}\,, \tag{2.11}$$

then $\Psi_{HF} \in W^{1,2}((\mathbb{R}^3)^N)$ fulfils

$$\langle \Psi_{HF}, \Psi_{HF} \rangle = 1\,.$$

Remark 2.3. *This choice of determinantal form is appropriate for the non-interacting case, i.e. $V_{ee} = 0$, but for interacting systems it is a real constraint. This is due to the intrinsic assumption of the electron positions to be independent variables, which is not true. We thus loose some correlation between the particles, see [10] for further details. Nevertheless, the computed energy is allways an upper bound to the true energy. The error is the so called correlation energy.*

Inserting (2.10) in (2.1) and using the *Slater-Condon rules* (cf. [40]) for the occurring single- and two-particle operators, yields the representation of the energy (for the *closed shell restricted Hartree-Fock* method, [76])

$$E_{HF} = \langle \Psi_{HF}, H\Psi_{HF} \rangle = \sum_{i=1}^{N} \left(2H_i + \sum_{j=1}^{N} (J_{ij} - \frac{1}{2} K_{ij}) \right) \tag{2.12}$$

with

$$H_i = \int \frac{1}{2} |\nabla \varphi_i(x)|^2 + V_{ne}(r) |\varphi_i(x)|^2 \,\mathrm{d}x \tag{2.13}$$

$$J_{ij} = \iint \frac{|\varphi_i(x)|^2 |\varphi_j(y)|^2}{|x-y|} \,\mathrm{d}y\,\mathrm{d}x \tag{2.14}$$

$$K_{ij} = \iint \frac{|\varphi_i(x)\varphi_i(y)||\varphi_j(x)\varphi_j(y)|}{|x-y|} \,\mathrm{d}y\,\mathrm{d}x\,. \tag{2.15}$$

The J_{ij} are called *Coulomb (energy) integrals* and the K_{ij} *exchange (energy) integrals*. Minimisation of (2.12) subject to the orthonormality conditions (2.11) gives the Hartree-Fock differential equation

$$H_{HF}\varphi_i(x) = \sum_{j=1}^{N} \lambda_{ij} \varphi_j(x)\,, \tag{2.16}$$

where the Hartree-Fock operator H_{HF} is given by

$$H_{HF} = -\frac{1}{2}\nabla^2 + V_{ne} + g \tag{2.17}$$

with the Coulomb-exchange operator $g(x) = j(x) - k(x)$ given by

$$j(x)u(x) = \sum_{i=1}^{N} \frac{|\varphi_i(y)|^2}{|x-y|} u(x)\,\mathrm{d}y \tag{2.18}$$

$$k(x)u(x) = \frac{1}{2}\sum_{i=1}^{N} \frac{|\varphi_i(x)\varphi_i(y)|}{|x-y|} u(y)\,\mathrm{d}y\,, \tag{2.19}$$

here u is an arbitrary function. The first term, also called the *Hartree-term*, describes the electrostatic (or *Coulomb*) interaction of the charge distribution with itself. The second is due to the Pauli exclusion principle, i.e. the anti-symmetry of Ψ_{HF}, and is of purely quantum nature. The matrix λ in (2.16) contains the Lagrange multipliers associated with the orthonormality constraints (2.11). Since the Hartree-Fock wavefunction Ψ_{HF} is invariant under unitary transformation of the orbitals $\{\varphi_i\}$, we can assume the matrix λ to be diagonal. This results in the *canonical Hartree-Fock* equations which, have the form of effective single particle Schrödinger equations,

$$H_{HF}\psi_i(x) = \lambda_i \phi_i(x)\,. \tag{2.20}$$

These equations are much more convenient for calculation than (2.16).

Density matrix function and electron density

In what follows we want to write the Hartree-Fock approximation in a slightly different (and more compact) way that will make it easier to compare it with the upcoming equations originated in DFT. For a given set of orbitals $\{\psi_i\}$ the density matrix function and the electron density are defined by

$$\rho(x,y) = \sum_{i=1}^{N} \psi_i(x)\psi_i(y)\,, \tag{2.21}$$

$$n(x) := \rho(x,x) = \sum_{i=1}^{N} |\psi_i(x)|^2\,. \tag{2.22}$$

Since the orbitals are normalised, integrating n over $\mathbb{R}$, yields the total number of electrons in the system

$$\int n(x)\,\mathrm{d}x = N\,.$$

Using (2.21) and (2.22), we firstly rewrite the energy (2.12) in the form

$$\frac{1}{2}E_{HF} = \sum_{i=1}^{N} \frac{1}{2}\langle \nabla\varphi_i, \nabla\varphi_i\rangle + \langle V_{ne}\varphi_i, \varphi_i\rangle + \frac{1}{2}\langle V_H\varphi_i, \varphi_i\rangle - \frac{1}{4}\langle W\varphi_i, \varphi_i\rangle\,. \tag{2.23}$$

Remark 2.4. *Note that we will use the same symbol for the potential and the operator indicated by it, e.g. $V_{ne} \in X$ and $V_{ne} : Y \mapsto Y'$ with proper spaces X and Y. It should be clear from the context how the symbol has to be understood.*

The terms V_H and W are called *Hartree potential* and *exchange energy* term, respectively. Both of which can be written in terms of n and ρ. To see this we first regard the sum of the Coulomb integrals (2.14)

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N} J_{ij} &= \sum_{i=1}^{N}\sum_{j=1}^{N} \iint \frac{|\varphi_i(x)|^2|\varphi_j(y)|^2}{|x-y|}\,\mathrm{d}y\,\mathrm{d}x \\ &= \sum_{i=1}^{N} \int \left(\int \frac{\sum_{j=1}^{N} |\varphi_j(y)|^2}{|x-y|}\,\mathrm{d}y \right) |\varphi_i(x)|^2\,\mathrm{d}x\,.\end{aligned}$$

Defining

$$V_H(x) := \int \frac{n(y)}{|x-y|}\,\mathrm{d}y = \int \frac{\sum_{j=1}^{N} |\varphi_j(y)|^2}{|x-y|}\,\mathrm{d}y \tag{2.24}$$

we further get

$$\sum_{i=1}^{N} \int V_H(x)|\varphi_i(x)|^2\,\mathrm{d}x = \int V_H(x) \sum_{i=1}^{N} |\varphi_i(x)|^2\,\mathrm{d}x = \int V_H(x) n(x)\,\mathrm{d}x\,. \tag{2.25}$$

Remark 2.5. *Defining the Hartree potential by* (2.24) *is only one possibility. Since it describes electrostatic interaction effects (*Coulomb interaction*) of the electrons, V_H can be as well calculated by solving Poisson's equation in terms of the charge distribution given by n, see* (2.37)*.*

Secondly we regard the exchange integrals (2.15)

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j=1}^{N} K_{ij} &= \iint \frac{|\varphi_i(x)\varphi_i(y)||\varphi_j(x)\varphi_j(y)|}{|x-y|}\,\mathrm{d}y\,\mathrm{d}x \\ &= \sum_{i=1}^{N} \int \left(\int \frac{\sum_{j=1}^{N} |\varphi_j(x)\varphi_j(y)|}{|x-y|} \varphi_j(y)\,\mathrm{d}y \right) \varphi_i(x)\,\mathrm{d}x\,.\end{aligned}$$

When comparing this with the last term in (2.23), we see that W is the mapping

$$Wu(x) := \int \frac{\sum_{j=1}^{N} |\varphi_j(x)\varphi_j(y)|}{|x-y|} u(y)\,\mathrm{d}y = \int \frac{\rho(x,y)}{|x-y|} u(y)\,\mathrm{d}y\,.$$

Rewriting $\sum_{i=1}^{N} \langle V_{ne}\varphi_i, \varphi_i\rangle$ analogous to (2.25) we get a representation of the Hartree-Fock energy in terms of ρ and n

$$\frac{1}{2}E_{HF} = \sum_{i=1}^{N} \frac{1}{2}\langle \nabla\varphi_i, \nabla\varphi_i\rangle + \int V_{ne}(x)n(x)\,\mathrm{d}x + \frac{1}{2}\int V_H(x)n(x)\,\mathrm{d}x - \frac{1}{4}\iint \frac{|\rho(x,y)|^2}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y\,. \tag{2.26}$$

which means

$$H_{HF} = -\frac{1}{2}\nabla^2 + V_{ne} + V_H - \frac{1}{2}W\,. \tag{2.27}$$

The Hohenberg-Kohn Theorems

The following theorems, first stated by Hohenberg and Kohn [42] in 1964, yield the main reasoning behind density functional theory. Formally it is an exact theory for many-particle systems and applies to any system of interacting particles. In this approach the basic variable is the electron density (2.22), from which in principle all other quantities can be deduced.

Theorem 2.6 (HK Theorem I [68])**.** *For any system of interacting particles in an external potential $V_{ext}(x)$, the potential $V_{ext}(x)$ is determined uniquely, except for a constant, by the ground state density $n_0(x)$.*

Theorem 2.7 (HK Theorem II [68])**.** *A* universal functional *of the energy $E[n]$ in terms of the density $n(x)$ can be defined, valid for any external potential $V_{ext}(x)$. For any particular $V_{ext}(x)$, the exact ground state energy of the system is the global minimum value of this functional, and the density $n(x)$ that minimises the functional is the exact ground state density $n_0(x)$.*

Although the proofs of these theorems are quite simple, we refer to [76, 68] for details. Let us just mention that they work by contradiction.

Since, together with the external potential $V_{ext}(x)$, the Hamiltonian is fully determined except for a constant energy-shift, and all states are determined as well. Therefore, by Theorem 2.6 all properties of the system are completely described by the ground state density $n_0(x)$. Further using Theorem 2.7 we find that the energy functional $E[n]$ alone is sufficient to determine the ground state energy and density.

Following the advanced formulation of Lieb [66] we can thus reformulate the functional dependence on the (external) potential V_{ne} into a functional dependence on the density n (by substituting $V_{ne}[n]$) and define the Hohenberg-Kohn functional by

$$F_{HK}[n] = E[n] - \int V_{ne}[n] n \,\mathrm{d}x\,, \quad n \in \mathcal{A}_n$$

where $\mathcal{A}_n$ denotes the class of *pure-state v-representable densities*, i.e.

$$\mathcal{A}_n := \{n(x) \,:\, x \in \mathbb{R}^3\,;\, n \text{ comes from an } N\text{-particle ground state}\}\,.$$

And thus get the energy $E[V_{ne}]$ for a given potential as

$$E[V_{ne}] = \min_{n \in \mathcal{A}_n} \left\{ F_{HK}[n] + \int V_{ne}[n] n \,\mathrm{d}x \right\}$$

Remark 2.8. *Note, that for a non-degenerate ground state the definition of $F_{HK}[n]$ just reduces to $T[n]+V_{ee}[n]$, i.e. the kinetic and electron-repulsion operator which are uniquely determined in this case. This formulation was originally used by Hohenberg and Kohn, who had to incorporate the* non-degeneracy *condition. Using the formulation of Lieb this condition can be dropped, since by definition of F_{HK}, constant shifts of the potential cancel out and simultaneously F_{HK} finds the correct minimum, cf. [21, Ch. 4] for details.*

Unfortunately, as was mentioned, the proofs of the theorems are not constructive and none of the occurring objects ($\mathcal{A}_n$, F_{HK}, $E[n]$) is known explicitly. So, even though we know that $n_0(x)$ determines $V_{ext}(x)$ uniquely, we are still left with the task of solving the many-particle problem. It was due to Kohn and Sham in 1965 to find a way of benefitting from this information. This was the starting point of the famous Kohn-Sham system, which we will present now.

The Kohn-Sham System

The idea of Kohn and Sham was to replace the interacting many-particle problem by a non-interacting one which yields the same ground state density. Following the idea of Hohenberg and Kohn, Kohn and Sham stated a similar uniqueness result for non-interacting particles. Thus, there is an *effective potential* V_{eff} for a non-interacting system which yields the same ground state density as the interacting system. Of course the corresponding potentials V_{ext} and V_{eff} cannot be the same, rather the non-interacting potential somehow has to incorporate the many-particle effects, in order to produce the correct density.

When treating a non-interacting system, the HK-functional is just the corresponding kinetic energy

$$F^0_{HK}[n] = T^0[n]\,,$$

where the superscript 0 indicates the non-interacting case. Of course even this functional is not known explicitly, but according to the HK-Theorems its existence is guaranteed.

Remark 2.9. *For interaction-free particle systems of fermions, there always exist a determinantal ground-state in form of a Slater-determinant* (2.10)*, cf. Remark 2.10. In case of degeneracy, linear combinations of determinantal states may be used as ground states as well. But, densities derived from linear combinations may not be reproducible by single determinantal states. Thus a slight change in the domain of T^0 to*

$$\mathcal{A}^0_n := \{n(x) \,:\, x \in \mathbb{R}^3\,;\, n \text{ comes from a determinantal } N\text{-particle ground state}\}$$

is necessary, cf. [66, 21].

The kinetic energy for a given density from $\mathcal{A}^0_n$ is then given by

$$T^0[n = \sum_{i=1}^{N} |\psi_i(x)|^2] = \frac{1}{2}\sum_{i=1}^{N} \langle \nabla\psi_i(x), \nabla\psi_i(x)\rangle\,.$$

Furthermore, we have to assume orthonormality of the orbitals, i.e.

$$T^0[n] = \min\left\{ \frac{1}{2}\sum_{i=1}^{N}\langle\nabla\psi_i(x),\nabla\psi_i(x)\rangle \,:\, \langle\psi_i,\psi_j\rangle = \delta_{ij}\,,\, n = \sum_{i=1}^{N}|\psi_i(x)|^2 \right\}.$$

The corresponding energy for an arbitrary potential V is then given by

$$\begin{aligned} E^0[V] &= \min\left\{ T^0[n] + \int V[n]n\,\mathrm{d}x \right\} \\ &= \min\left\{ \sum_{i=1}^{N}\left(\frac{1}{2}\langle\nabla\psi_i(x),\nabla\psi_i(x)\rangle + \langle V\psi_i(x),\psi_i(x)\rangle \right) \,:\, \langle\psi_i,\psi_j\rangle = \delta_{ij} \right\}. \end{aligned}$$

Introducing Langrangian multipliers ϵ_i representing the side conditions, we end up with a set of one-particle Schrödinger equations

$$\left[-\frac{1}{2}\nabla^2 + V \right]\psi_i = \epsilon_i\psi_i\,, \tag{2.28}$$

for the N orbitals, that are lowest in energy.

Remark 2.10. *When assuming the potential V to be known and fixed, the problem is separable and thus the solution is given as a product of solutions ψ_i of* (2.28)*. Hence, Ψ is a single Slater determinant.*

Knowing this, the HK-functional of the interacting system can be decompose to

$$F_{HK}[n] = T^0[n] + E_H[n] + E_{xc}[n]\,.$$

where E_H denotes the Hartree term according to the Hartree potential, cf. (2.26). This equation is the defining relation for the *exchange-correlation* energy E_{xc}, which has to exist, since all other terms are known to be functionals of n. It holds

$$E_{xc}[n] = T[n] - T^0[n] + V_{ee}[n]\,.$$

This decomposition can be understood in the following way. The density n belonging to the interacting system is described as a density belonging to a special system of non-interacting particles. The omitted particle interactions are put into the new term E_{xc}, which is known to exist due to Hohenberg and Kohn. All together, the systems energy is now given by

$$\begin{aligned} E[V_{ne}] = \min\Bigg\{ \sum_{i=1}^{N}\left(\frac{1}{2}\langle\nabla\psi_i(x),\nabla\psi_i(x)\rangle + \langle V_{ne}\psi_i(x),\psi_i(x)\rangle \right) + E_H[n] + E_{xc}[n] \\ : \langle\psi_i,\psi_j\rangle = \delta_{ij}\,,\, n = \sum_{i=1}^{N}|\psi_i(x)|^2 \Bigg\} \end{aligned}$$

Again treating the side conditions with Lagrangian multipliers, the resulting Schrödinger equation reads

$$\left[-\frac{1}{2}\nabla^2 + V_{eff}\right]\psi_i = \lambda_i \psi_i \,, \tag{2.29}$$

with

$$V_{eff} = V_{ne} + V_H + V_{xc} \,, \tag{2.30}$$

where V_H is again the Hartree-potential belonging to the electrostatic charge distribution and the *exchange-correlation potential* V_{xc} is defined by the variation of E_{xc} with respect to n, i.e.

$$V_{xc}(x) = \frac{\partial E_{xc}}{\partial n(x)} \,.$$

Equations (2.29) and (2.30) together with (2.22) form the famous Kohn-Sham system. Comparing (2.29) with (2.27) from HF theory, we formally see the only difference in the terms V_{xc} and $-\frac{1}{2}W$. Both attempts lead to one-particle formulations, but the basic ideas behind are completely different. In HF the assumption is that the electron positions are independent variables, which is not the case. Roughly speaking, HF tries to find an interaction-free solution to an interacting problem, meanwhile accepting a certain intrinsic error. In contrast to that, the Kohn-Sham approach is formally exact by treating an interaction-free reference system that yield the correct density of the system with interaction. This reference system, or more precisely the potential leading to this system, is known to exist. The task now lies in the precise appearance of the potential term representing the interaction effects, i.e. exchange and correlation, which then are fully incorporated. Further note that the (physical) meaning of the orbital functions ψ_i in the HF and KS approach are quite different. The HF orbitals represent the real orbitals belonging to the occupied states, whereas the orbital functions in the KS approach are those of the reference system and as such only the density produced has a physical meaning. Compare [19, 10, 76, 62, 68] for more details on physical interpretation of the KS approach.

The greatest challenge in the Kohn-Sham approach is to find an explicit form (or at least a good approximation) of the exchange-correlation term. In the following the simplest approach already used by Kohn and Sham is described. The ansatz is to assume the same structure for the exchange-correlation part as for the nucleus-electron and Hartree-potential, cf. (2.27), which are given as integrals of the form

$$\int V(x)n(x)\,\mathrm{d}x \,.$$

Using the simplified view as uniform electron-gas we get to the so called *local density approximation* (LDA) for the exchange and correlation energy.

$$E_{xc}^{LDA}[n] = \int V_{xc}[n](x)n(x)\,\mathrm{d}x \,. \tag{2.31}$$

$V_{xc}[n]$ indicates exchange and correlation per particle of the uniform electron gas of density n. It can be divided into an exchange and a correlation contribution by

$$V_{xc}(n) = V_x(n) + V_c(n)\,. \tag{2.32}$$

When using *Dirac's exchange-energy formula*, cf. [76, Ch. 6], an explicit form of the exchange part is found to be given by

$$V_x(n) = -C_x n(x)^{1/3}\,, \qquad C_x = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\,. \tag{2.33}$$

This yields the exchange energy

$$E_x^{LDA}[n](x) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int n(x)^{4/3}\,\mathrm{d}x \tag{2.34}$$

and the exchange potential

$$V_x^{LDA}[n](x) = -\left(\frac{3}{\pi}n(x)\right)^{1/3}\,. \tag{2.35}$$

The term $V_c[n]$ is more complicated, since even for the uniform electron-gas only few analytic expressions are known (or suggested) and in addition these are limiting cases.

The uniform electron-gas assumption is applicable to systems of slowly varying densities, but for atoms and molecules it cannot formally be justified. Nevertheless, numerical applications showed a surprisingly well agreement with experimental data. One of the reasons for this is a systematic cancellation of errors that can be shown even for general systems. A more detailed discussion of the LDA method and its success can be found in [19, Ch. 7] and [11].

In LDA the xc-energy is given by an integral of the form

$$E_{xc}[n] = \int f(n(x))\,\mathrm{d}x\,.$$

A generalisation to this can be found by allowing the function f to depend on the (higher order) gradients of the density n as well, e.g. $f(n(x), \nabla n(x), \ldots)$. This leads to *generalised gradient approximations* (GGA). The choices of f are usually parameter-fitted to experimental data, such that a certain GGA approach usually has a stronger restriction on the problem at hand, cf. [11]. However, common LDA or GGAs still fail for describing *van der Waals* interactions. To overcome this, several other approaches came up, such as *hybrids* or *Meta-GGAs*, cf. [11, 19].

Finite Temperature Extension

The presented formalism is built to describe the zero-temperature case. For systems at finite temperature the equilibrium state is the analogon to the ground-state. This generalisation was first done by Mermin [71] in 1965. The associated electron density is then given in the form

$$n(x) = \sum_{i=1}^{\infty} f_i |\psi_i(x)|^2 \tag{2.36}$$

where the $f_i = f(\lambda_i - \mu)$ are called occupation numbers. The function f yields values in $[0, 1]$ and represents fractional occupancy according to the underlying statistics that describes the distribution of particles over energy states. The real number μ denotes the Fermi-level, the position of which determines the carrier densities of the semiconductor in thermodynamical equilibrium.

In what follows we want to model semiconductor devices in *effective mass* approximation, cf. [90, Ch. 13]. In case of a three dimensional bulk material the nanostructure can be characterised by the number d of band discontinuities. For $d = 0$ there is no nanostructure, for $d = 1$ there is a two dimensional electron gas in a quantum well, for $d = 2$ its a one-dimensional gas in a quantum wire and $d = 3$ describes a quantum dot.

According to the number of discontinuities d the distribution function f takes different forms, related to Fermi's integrals $\mathcal{F}_\alpha$, cf. [26]. More precisely (cf. [47, Appendix]), for an ensemble in a quantum dot, i.e. a number of band discontinuities $d = 3$, f is given by Fermi's function

$$f(s) = c_3 \mathcal{F}_{-1}(-\frac{s}{k_B T}) = \frac{c_3}{1 + \exp(\frac{s}{k_B T})} \,.$$

k_B denotes Boltzmann's constant and T the temperature of the carrier gas. For the one-dimensional carrier-gas in a quantum-wire, $d = 2$, it is

$$f(s) = c_2 \mathcal{F}_{-\frac{1}{2}}(-\frac{s}{k_B T}) = c_2 = \int_0^\infty \frac{\xi^{-\frac{1}{2}}}{1 + \exp(\xi + \frac{s}{k_B T})} \mathrm{d}\xi$$

and for the two-dimensional carrier-gas in a quantum-well, $d = 1$, f is given by

$$f(s) = c_1 \mathcal{F}_0(1 + \exp(-\frac{s}{k_B T})) = c_1 \ln\left(1 + \exp(-\frac{s}{k_B T})\right) \,.$$

The constants c_1, c_2 and c_3 depend on the semiconductor material.

Remark 2.11. *These expressions for the distribution function apply for both, electrons and holes. For electrons the energy is scaled on the usual axis whereas for holes it is counted on the negative axis.*

Poisson's Equation

Finally we have a closer look on the Hartree potential V_H describing the electron-electron repulsion. As was already mentioned in Remark 2.5, V_H describes the effect of electrostatic interactions of the particles. Thus, V_H is given as the solution of

$$-\nabla^2 V_H(x) = 4\pi \sum_\sigma e_\sigma n_\sigma \,, \tag{2.37}$$

which is Poisson's equation in atomic units, since the dielectricity constant ϵ_0 then takes the value $\frac{1}{4\pi}$. The sum goes over all occurring types of charged carriers σ and the corresponding densities are denoted by n_σ. The factors e_σ describe the charge of a single σ-type carrier, i.e. -1 for electrons and 1 for holes. Knowing this, we replace V_H by φ, which is the solution of Poisson's equation

$$-\varepsilon \nabla^2 \varphi = \sum_\sigma e_\sigma n_\sigma \,. \tag{2.38}$$

In this way we can describe semiconductor heterostructures by making the dielectricity position dependent. Additionally, an effective doping D of the semiconductor material can be included by adding it to the right-hand side of (2.38)

$$-\nabla \varepsilon \nabla \varphi = D + \sum_\sigma e_\sigma n_\sigma \,. \tag{2.39}$$

2.2 Semiconductor Heterostructures

The main application we have in mind is the modelling of semiconductor heterostructures. This forecloses some adjustment of the Kohn-Sham system just presented. First of all, semiconductor devices have a predetermined device domain $\Omega \in \mathbb{R}^n$ that is bounded. Hence, we will search for solutions not on the whole of $\mathbb{R}^n$ but instead on the bounded set Ω. Inside the device we allow for different types of charged carriers indicated by $\xi \in \{1, \ldots, \sigma\}$ with a charge e_ξ, e.g. $e_\xi = -1$ for electrons and $e_\xi = 1$ for holes. Each species has a fixed number of particles N_ξ that is conserved. Furthermore, we allow for an effective doping, given by a profile D over the device domain.
The Kohn-Sham system now has to be solved for a vector of carrier densities $n = (n_1, \ldots, n_\xi)$ and the electrostatic potential φ.

The densities are given by the expression

$$n(x) = \sum_{i=1}^{\infty} f_\xi(\lambda_{i,\xi} - \mathcal{E}_{F,\xi})|\psi_{i,\xi}(x)|^2 \,,$$

with occupation factors $f_\xi(\lambda_{i,\xi} - \mathcal{E}_{F,\xi})$ and Fermi levels $\mathcal{E}_{F,\xi}$. The $\lambda_{i,\xi}$ are the eigenvalues (counting multiplicity) and $\psi_{i,\xi}$ the eigenfunctions of the corresponding Schrödinger operators in *effective mass* approximation. This means we incorporate the periodic crystal structure of the semiconductor material by adjusting the mass of the particle. In this way the influence of the crystal on the mobility of the particle is taken into account. However, since we deal with heterostructures, this effective mass differs throughout the device domain Ω. Thus, the Hamiltonian is given by

$$\left[-\frac{1}{2}\nabla\frac{1}{m_\xi}\nabla + V_{eff,\xi}\right]\psi_{i,\xi} = \lambda_{i,\xi}\psi_{i,\xi} \quad \text{on} \quad \Omega \,,$$

with m_ξ the material dependent effective mass and $V_{eff,\xi}$ the effective Kohn-Sham potential depending on the carrier densities

$$V_{eff,\xi}[n] = -e_\xi V_{0,\xi} + V_{xc,\xi}[n] + e_\xi\varphi[n] \,.$$

Here $\varphi[n]$ denotes the electrostatic potential, which is the solution to Poisson's equation

$$-\nabla\epsilon\nabla\varphi = D + \sum_\xi e_\xi n_\xi \quad \text{on} \quad \Omega \,,$$

where ϵ is the material dependent dielectricity.

The given external potentials $V_{0,\xi}$ are the piecewise constant band-edge offsets, that represent the heterostructure in the effective mass approximation. $V_{xc,\xi}[n]$ is the exchange-correlation potential describing the particle interactions. It is given by the LDA expression, cf. Appendix A

$$V_{xc,\xi}[n](x) = -\left(\frac{3}{\pi}n(x)\right)^{1/3} \,.$$

The real numbers $\mathcal{E}_{F,\xi}$ are the Fermi-levels of the ξ-type carriers. They are defined by the conservation law for the corresponding carriers

$$\int_\Omega n_\xi(x)\,\mathrm{d}x = \sum_{i=1}^{\infty} f_\xi(\lambda_{i,\xi} - \mathcal{E}_{F,\xi}) = N_\xi\,.$$

$\mathcal{E}_{F,\xi}$ is well defined, due to the decaying properties of the distribution functions f_ξ, which take different forms depending on the dimension d of the carrier gas, cf. [47, Appendix].

$$f(s) = c\mathcal{F}_\alpha(-\frac{s}{k_BT})\,, \quad \alpha = \begin{cases} -1 & \text{if } d = 3 \\ -\frac{1}{2} & \text{if } d = 2 \\ 0 & \text{if } d = 1 \end{cases}.$$

$\mathcal{F}_\alpha$ denotes Fermi's integral, cf. previous section.

Finally we have to set appropriate boundary conditions. Concerning the electrostatic potential φ we regard the following ones

$$\varphi = \tilde{\varphi}_1 \quad \text{on } \Gamma\,, \quad -\langle \nu, \varepsilon\nabla\varphi\rangle = b(\varphi - \tilde{\varphi}_2) \quad \text{on } \partial\Omega \setminus \Gamma$$

where Γ is a closed subset of the boundary $\partial\Omega$. The Dirichlet conditions on Γ model Ohmic contacts and the conditions of third kind on $\partial\Omega \setminus \Gamma$ covers interfaces between the semiconductor device and insulators (with capacity $b \geq 0$) or homogenous Neumann boundary conditions ($b = 0$).
About the Schrödinger operator we assume that the device confines the charged particles. This results in boundary conditions of the form

$$\psi = 0 \quad \text{on } \Gamma\,, \quad \langle \nu, m_\xi^{-1}\nabla\psi\rangle = 0 \quad \text{on } \partial\Omega \setminus \Gamma\,.$$

We regard mixed boundary conditions to be able to model cuts through symmetric nanostructures with homogenous Dirichlet conditions on the physical boundary.

3 Analytical Considerations

In this section we analyse the previously defined Schrödinger-Poisson system from the mathematical point of view. Of special interest will be those properties that will give us solvability of the system. Therefor, requirements on the domain and the spaces are made here as well as assumptions on the involved functions.

Concerning existence of solutions, we will follow the lines of the early works from Kaiser and Rehberg [47, 49, 50] by using Schauder's fixed point theorem, cf. Theorem 3.48. However, the presented result is a little more general in that only continuity of the distribution function f is assumed, instead of differentiability. Thus, in particular the zero-temperature case as described in [15] by Cornean, H., Neidhardt, Racec and Rehberg is included. Furthermore, the analyticity of the particle density operator is treated, cf. H., Kaiser and Rehberg [43]. Even though analyticity may not be essential for the existence result, there are indications that this property might be gainful for establishing steadily converging iteration schemes. In fact it was used by Gajewski and Griepentrog in [27] for setting up a descent method for the free energy of a multicomponent system that is comparable in structure to the Kohn-Sham system.

Notation

In this work we are interested in a statistical ensembles of one-particle systems. These systems will be examined in the real space representation on a bounded up to three dimensional domain Ω, i.e. $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$. During these considerations different function spaces over Ω are involved. In order to simplify notation we omit the indication for Ω in the function space symbol, e.g. writing L^2 instead of $L^2(\Omega)$. Further, if necessary, the distinction between real and complex space will be indicated as a subscript, e.g. $L^2_{\mathbb{R}}$ or $L^2_{\mathbb{C}}$. The space of linear continuous operators from one Banach space X into another Y will be denoted by $\mathcal{B}(X;Y)$. If $X = Y$, we use the abbreviation $\mathcal{B}(X) := \mathcal{B}(X;X)$. Because of the numerous use of $X = L^2$, we once more abbreviate $\mathcal{B} := \mathcal{B}(L^2)$.
The ideal of compact operators within $\mathcal{B}$ we denote by $\mathcal{B}_\infty$ and the Schatten-class with index $r \in [1, \infty]$ in $\mathcal{B}_\infty$ will be denoted by $\mathcal{B}_r$. Without further mentioning we identify a function from L^∞ with the multiplication operator from L^2 to L^2 induced by this function. In this sense L^∞ is embedded into $\mathcal{B}$.

3.1 Domain and Spaces

About the spatial domain Ω we make the following general assumption.

Assumption 3.1. $\Omega \in \mathbb{R}^d$, $d \in \{1, 2, 3\}$, is a bounded Lipschitz domain, cf. e.g. [70, Ch. 1.1.9], [35, Defn. 1.2.1.2]. Let Γ be an arbitrary closed subset of the boundary $\partial\Omega$. Γ

and $\partial\Omega \setminus \Gamma$ fulfil the regularity property of Gröger [36], i.e. are separated by a Lipschitzian hypersurface of $\partial\Omega$.

Concerning the solution spaces for Poisson's- and Schrödinger's-problem we regard Sobolev spaces with integrability index 2 and differentiability index 1, i.e. $W^{1,2}_{\mathbb{R}}$ and $W^{1,2}_{\mathbb{C}}$. We further introduce

Definition 3.2. Let Ω and Γ fulfil Assumption 3.1. Let $W^{1,2}_{\mathbb{C},\Gamma}$ be the $W^{1,2}_{\mathbb{C}}$-closure of the set

$$\{\psi|_\Omega \,:\, \psi \in C_0^\infty(\Omega)\,,\; supp(\psi) \cap \Gamma = \emptyset\}\,.$$

and $W^{1,2}_{\mathbb{R},\Gamma}$ its real part. Further $W^{-1,2}_{\mathbb{C},\Gamma}$ denotes the space of continuous anti-linear forms on $W^{1,2}_{\mathbb{C},\Gamma}$ and $W^{-1,2}_{\mathbb{R},\Gamma}$ the space of continuous linear forms on $W^{1,2}_{\mathbb{R},\Gamma}$.

3.2 The Schrödinger Operator

First, we make the following assumption on the coefficient function of the Schrödinger operator, i.e. on the effective mass tensor.

Assumption 3.3. Let m be a bounded Lebesgue measurable function on Ω with values in the set of real, symmetric, positive definite $d \times d$ matrices, such that m^{-1} is bounded as well.

With this we define the Schrödinger operator by

Definition 3.4. Let m be given as in Assumption 3.3 . Define the Schrödinger operator with zero potential $H_0 : W^{1,2}_{\mathbb{C},\Gamma} \to W^{-1,2}_{\mathbb{C},\Gamma}$ by

$$\langle H_0 v, w\rangle = \frac{1}{2}\int_\Omega \langle m^{-1}(x)\nabla v(x), \nabla w(x)\rangle \,\mathrm{d}x\,, \quad v, w \in W^{1,2}_{\mathbb{C},\Gamma}\,. \tag{3.1}$$

Remark 3.5. *H_0, seen as a mapping into L^2, is the self-adjoint operator corresponding to the quadratic form*

$$a_0[\psi] := \int_\Omega m^{-1}\nabla\psi \cdot \nabla\overline{\psi}\;\mathrm{d}x\,, \tag{3.2}$$

by the representation theorem for forms, [51].

The boundary conditions associated with the restriction of H_0 to L^2 are homogeneous Dirichlet conditions on the boundary part Γ and homogeneous Neumann conditions on the remaining part $\partial\Omega \setminus \Gamma$.

H_0 has a discrete spectrum that lies on the real axis. More precisely, the eigenvalues of H_0 lie between the corresponding eigenvalues of the operator $-\nabla m^{-1}\nabla$ once with pure

(homogeneous) Dirichlet and and once with pure (homogeneous) Neumann boundary conditions, see [16, Ch. VI, Sec. 2+4].

In the following we consider Schrödinger operators H_0+V, which we will denote shortly by H_V. The applied potential V will be taken from the real space $L^2_{\mathbb{R}}$. This choice has mainly two reasons. First, the application we have in mind includes modelling of heterostructures of semiconductor material in an effective mass approximation. Thus the potential V includes the band-edge offsets, which are discontinuous, when regarding heterogeneities, and so spaces of continuous functions only, are not adequate. Furthermore, adding the potential V means adding a multiplication operator working on $\mathrm{dom}(H_0)$. To make sure this operation is well defined, meaning the integral

$$\int_\Omega V(x)v(x)w(x)\,\mathrm{d}x\,, \quad v,w \in W^{1,2}_{\mathbb{C},\Gamma} \tag{3.3}$$

to be finite, we chose V to be from L^2. Finally, the spectra of the operators H_V shall not expand away from the real axis, i.e. V as a multiplication operator shall be symmetric in $\mathrm{dom}(H_0)$. Hence, we restrict our considerations to the real space $L^2_{\mathbb{R}}$.

Remark 3.6. *Note that in the one-dimensional case $d = 1$ taking V from $L^1_{\mathbb{R}}$ would be sufficient, see [49].*

Concerning properties of H_0 we first state a result that is fundamental for the subsequent perturbation theory

Theorem 3.7. *For every $\theta \in]\frac{d}{4}, 1]$, the operator $(H_0+1)^{-\theta}$ maps L^2 continuously into L^∞.*

The proof of this requires the following auxiliary results:

Proposition 3.8. *[75, Cor. 4.10] $H_0 + 1$ generates a contraction semigroup on $L^\infty(\Omega)$.*

Proposition 3.9. *[75, Thm. 6.2] Let $\mathfrak{a}$ be a densely defined, closed, symmetric and non-negative sesquilinear form on L^2 and let A be the self-adjoint operator which corresponds to $\mathfrak{a}$. Assume that the semigroup $\left(e^{-tA}\right)_{t>0}$ is contractive on $L^\infty(\Omega)$. Suppose that for one $q \in]2, \infty[$ with $d\frac{q-2}{2q} < 1$ the following Gagliardo-Nirenberg-type inequality*

$$\|\psi\|_{L^q} \leq c\,\mathfrak{a}[\psi,\psi]^{d\frac{q-2}{4q}}\ \|\psi\|_{L^2}^{1-d\frac{q-2}{2q}} \tag{3.4}$$

is satisfied for a constant c and every $\psi \in Dom(\mathfrak{a})$. Then there is a constant $c_0 > 0$ such that

$$\|e^{-tA}\|_{\mathcal{B}(L^2;L^\infty)} \leq c_0 t^{-\frac{d}{4}}. \tag{3.5}$$

Proposition 3.10. *If Ω is a Lipschitz domain and the form $\mathfrak{a}$ is symmetric, coercive and continuous, then (3.4) holds.*

Proof. Under our suppositions, $\mathfrak{a}^{1/2}$ may be estimated from above and below by the $H^1(\Omega)$ norm. Thus, (3.4) is equivalent to the classical Gagliardo-Nirenberg inequality on Ω. This latter is implied by the following two statements:

i) For every Lipschitz domain there is an extension operator (see [31])

$$\mathfrak{E} \in \mathcal{B}(H^1(\Omega); H^1(\mathbb{R}^d) \cap \mathcal{B}(L^r(\Omega); L^r(\mathbb{R}^d)$$

for every $r \in [1, \infty[$.

ii) The same Gagliardo-Nirenberg inequality is valid on whole $\mathbb{R}^d$.

□

It follows the proof of Theorem 3.7: using the representation formula

$$A^{-\theta} = \frac{1}{\Gamma(\theta)} \int_0^\infty t^{\theta-1} e^{-tA} \, \mathrm{d}t \,,$$

cf. [77, Ch. 2.6]. According to (3.5), we can estimate

$$\|(H_0+2)^{-\theta}\|_{\mathcal{B}(L^2;L^\infty)} \leq \frac{1}{\Gamma(\theta)} \int_0^\infty \frac{t^{\theta-1}}{e^t} \left\|e^{-t(H_0+1)}\right\|_{\mathcal{B}(L^2;L^\infty)} \, \mathrm{d}t \leq \frac{c_0}{\Gamma(\theta)} \int_0^\infty \frac{t^{\theta-1-\frac{d}{4}}}{e^t} \, dt \qquad (3.6)$$

Since by the spectral properties of H_0 and functional calculus the operator $[(H_0+1)^{-1}(H_0+2)]^\theta \,:\, L^2 \mapsto L^2$ is finite, the right-hand side is finite for $\theta \in]\frac{d}{4}, 1]$. This finishes the proof.

Remark 3.11. *Theorem 3.7 even holds true in case of nonsmooth domains, discontinuous coefficient functions and mixed boundary conditions. Thus, realistic geometries as well as heterostructures are covered by it.*

The upcoming results make sure, that the perturbed operator H_V inherits gainful properties from H_0, e.g. self-adjointness and discreteness of the spectrum.

Lemma 3.12. *For every $V \in L^2_{\mathbb{R}}$ the corresponding multiplication operator from L^2 into L^2 induced by V, is infinitesimally small with respect to $H_0 + 1$.*

Proof. Due to Theorem 3.7 we can estimate

$$\begin{aligned} \|V\psi\|_{L^2} &\leq \|V\|_{L^2} \|\psi\|_{L^\infty} \\ &= \|V\|_{L^2} \|(H_0+1)^{-\frac{4}{5}} (H_0+1)^{\frac{4}{5}} \psi\|_{L^\infty} \\ &\leq c \, \|V\|_{L^2} \|(H_0+1)^{\frac{4}{5}} \psi\|_{L^2} \end{aligned}$$

for all $\psi \in \mathcal{D} = \mathrm{dom}(H_0)$. Since $H_0 + 1$, like H_0, is selfadjoint and positive, the right hand side can be further estimated by

$$c\,\|V\|_{L^2}\|\psi\|_{L^2}^{\frac{1}{5}}\|(H_0+1)\psi\|_{L^2}^{\frac{4}{5}},$$

cf. [77, Ch. 2.6 Th. 6.10].
According to Young's inequality, this is not larger than

$$\epsilon\|(H_0+1)\psi\|_{L^2} + \left(\frac{1}{\epsilon}\right)^4 (c\|V\|_{L^2})^5\,\|\psi\|_{L^2}$$

for any $\epsilon > 0$. □

Corollary 3.13. *For every potential $V \in L^2_{\mathbb{R}}$ the operator $H_0 + V$ inherits the essential properties from the unperturbed operator H_0, namely*

i) is self-adjoint

ii) has the same domain as H_0

iii) has a compact resolvent and, hence, a discrete spectrum

iv) is semi-bounded from below and the corresponding lower form bounds may be taken uniformely with respect to bounded sets in $L^2_{\mathbb{R}}$.

Proof. The statements follow from Lemma 3.12 by classical perturbation theory, see [51] Ch. IV Thm. 1.1+3.17 and Ch. V Thm. 4.11 . □

The next corollary provides the possibility of proving that the resolvent of a general Schrödinger operator with potential V from $L^2_{\mathbb{R}}$ has the same summability properties as the resolvent of H_0, see [49] and [50]. This will be of use when investigating the particle density operator.

Corollary 3.14. *For every $V \in L^2_{\mathbb{R}}$ and $\rho \notin spec(H_0 + V)$ both operators $(H_0+1)(H_0+V-\rho)^{-1}$ and $(H_0+V-\rho)^{-1}(H_0+1)$ are topological isomorphisms of L^2 onto itself.*

Proof. Since H_0 is self-adjoint and positive -1 belongs to its resolvent set. Thus, $(H_0 + V - \rho)(H_0+1)^{-1}$ is a bijection of L^2 onto itself. Continuity of this mapping can be seen by

$$\begin{aligned}\|(H_0+V-\rho)(H_0+1)^{-1}\|_{\mathcal{B}} &= \|\mathbb{1} + (V-\rho-1)(H_0+1)^{-1}\|_{\mathcal{B}} \\ &\leq 1 + \|(V-\rho-1)(H_0+1)^{-1}\|_{\mathcal{B}} \\ &\leq 1 + \|V-\rho-1\|_{L^2}\|(H_0+1)^{-1}\|_{\mathcal{B}(L^2,L^\infty)}\end{aligned}$$

thanks to Theorem 3.7. The assertion of this corollary for $(H_0+1)(H_0+V-\rho)^{-1}$ follows now by an application of the open mapping principle, which yields that the inverse of $(H_0+V-\rho)(H_0+1)^{-1}$ is a bounded linear map. For $(H_0+V-\rho)^{-1}(H_0+1)$ the assertion follows from this by taking the adjoint. □

Finally, we will require summability properties of the resolvent of the occurring Schrödinger operators. The central result for this is as follows:

Theorem 3.15. *For the operator H_0 from Definition 3.4 the resolvent is in a Schatten-class. More precisely:*

$$(H_0 + 1)^{-1} \in \mathcal{B}_r \quad \textit{for every} \quad r > \frac{d}{2}\,.$$

Proof. For a Schrödinger operator H_0 with homogeneous Dirichlet boundary conditions the assertion has been proved by Birman and Solomyak even for arbitrary domains Ω, cf. [2], [3]. The case of Neumann boundary conditions has been dealt with in [2, 3, 4] as well, provided that the underlying domain Ω is an $W^{1,2}$ extension domain, i.e. if there is a linear, continuous extension operator from $W^{1,2}(\Omega)$ to $W^{1,2}(\mathbb{R}^d)$. In fact, the result holds true for Lipschitz domains too, cf. [33] and [70, Ch. 1.1.16]. Having the Dirichlet and Neumann case at hand, the result can be carried over to the mixed boundary conditions case by the classical comparison principle, cf. [16, Ch. VI, Sec. 2]. □

Corollary 3.16. *For every $V \in L^2$ the operator $V(H_0 + 1)^{-1} : L^2 \to L^2$ is not only bounded, but compact and belongs to the Schatten class $\mathcal{B}_7$. More precisely, one can estimate*

$$\begin{aligned}\|V(H_0 + 1)^{-1}\|_{\mathcal{B}} &\leq \|V(H_0 + 1)^{-1}\|_{\mathcal{B}_7} \\ &\leq \|V\|_{L^2}\|(H_0 + 1)^{-10/13}\|_{\mathcal{B}(L^2;L^\infty)}\|(H_0 + 1)^{-3/13}\|_{\mathcal{B}_7} < \infty\,. \quad (3.7)\end{aligned}$$

Proof. $\|(H_0+1)^{-10/13}\|_{\mathcal{B}(L^2;L^\infty)}$ is finite since $10/13 > 3/4 \geq d/4$, cf. Theorem 3.7. Further, according to Theorem 3.15, $(H_0+1)^{-1}$ belongs to the Schatten class $\mathcal{B}_r$ for every $r > 3/2 \geq d/2$, in particular $(H_0+1)^{-1} \in \mathcal{B}_{21/13}$. Hence, $(H_0+1)^{-3/13}$ is in the Schatten class $\mathcal{B}_7$. □

3.3 The Quantum Mechanical Particle Density Operator

In this section the particle density operator will be defined and analysed. Of special interest will be continuity and monotonicity properties, which we need for showing the existence of solutions to the Kohn-Sham system, see Section 3.5.

In order to define the quantum mechanical particle density operator, we need to specify the occurring thermodynamic distribution function f, representing the underlying statistics of the ensemble of (identical) quantum particles, cf. e. g. [82, Ch. 1.12] or [44, Ch. 6.3]. For an ensemble in a three dimensional bulk material the precise appearance depends on the number d of band discontinuities, cf. Section 2. First we will state assumptions on the decaying properties of f. These are necessary to ensure the summability properties of the occurring operators of the form $f(H_V)$.

Assumption 3.17. The statistical distribution function f is a positive, real-valued, continuous function that is strictly monotonously decreasing and in addition obeys for $\rho \in \mathbb{R}$

$$\sup_{s\in[0,\infty[} |f(s)(s+\rho)^4| < \infty\,,$$

3.3.1 Definition

For a given potential $V \in L^2_{\mathbb{R}}$ and a distribution function f the quantum mechanical particle density is given by the expression

$$\mathcal{N}(V)(x) = \sum_{l=1}^{\infty} f(\lambda_l(V) - \mathcal{E}_F(V))|\psi_l(V)(x)|^2 \tag{3.8}$$

where $\{\lambda_l(V)\}$ and $\{\psi_l(V)\}$ are the eigenvalues (counting multiplicity) and L^2-normalised eigenfunctions, respectively, of the Schrödinger operator H_V. The numbers $N_l(V) := f(\lambda_l(V) - \mathcal{E}_F(V))$ are called occupation numbers. The real number $\mathcal{E}_F(V)$ is called the Fermi-level. It is defined by the conservation law

$$\int_\Omega \mathcal{N}(V)(x)\,\mathrm{d}x = \sum_{l=1}^{\infty} f(\lambda_l(V) - \mathcal{E}_F(V)) = N\,. \tag{3.9}$$

N being the fixed total number of carriers in the device domain.

Remark 3.18. *The strict monotonicity of the distribution function f together with the asymptotics of the eigenvalues of H_V, yield that the Fermi-level $\mathcal{E}_F(V)$ is uniquely determined. Moreover $f(H_V - \mathcal{E}_F(V))$ is a nuclear operator, see [47, 49, 50, 72, 73], and hence the duality between $\mathcal{N}(V)$ and test functions from L^∞ is expressed by*

$$\int_\Omega \mathcal{N}(V)W\,\mathrm{d}x = \operatorname{tr}\left(W f(H_V - \mathcal{E}_F(V))\right)\,. \tag{3.10}$$

The expression for the particle density as defined in (3.8) includes the real shift $\mathcal{E}_F(V)$, which, subject to the potential V, is fixed. Thus, the eigenvalues of the corresponding operator $H_{V-\mathcal{E}_F(V)}$ are those of H_V, though shifted by $\mathcal{E}_F(V)$. Let us therefore introduce the (unshifted) pseudo-particle operator $\widetilde{\mathcal{N}} : L^2_{\mathbb{R}} \to L^1_{\mathbb{R}}$ by putting

$$\widetilde{\mathcal{N}}(V) := \sum_{l=1}^{\infty} f(\lambda_l)|\psi_l|^2\,.$$

It is sufficient to analyse properties of $\widetilde{\mathcal{N}}$, provided $V \mapsto \mathcal{E}_F(V)$ is bounded on bounded sets and continuous. Indeed, this is the case, as shown by Lemma 3.24, cf. [47], [49], [50].

Electrons and Holes

In what follows we want to model the spatial distribution of electrons (indicated by a subscript n) and holes (subscript p) in semiconductor heterostructures. In order to calculate the density for each species, the eigenvalues and eigenfunctions of the corresponding Schrödinger operators, $H_{n,V} := H_{n,0} + V_n$ and $H_{p,V} := H_{p,0} + V_p$, have to be computed.
In addition to the difference in the effective masses, m_n and m_p, the operators differ as well in the way the effective potentials V_n and V_p are built. In case of electrons and holes the effective potentials split up as follows:

$$V_\xi = -e_\xi V_{\xi,0} + V_{xc,\xi} + e_\xi \varphi \,.$$

The factor e_ξ is a sign, 1 for holes and -1 for electrons. The constant external potential $V_{\xi,0}$ is the band-edge offset of the species ξ. The term $V_{xc,\xi}$ is the exchange correlation potential for the species ξ. Note that this term might depend not only on species ξ, but on all occurring species, i.e. the exchange-correlation expressions for electrons and holes might each depend on both the electron density and the hole density. Finally, φ describes the electrostatic potential in the device domain.

The reason for introducing the signs e_ξ is the difference in sign of the charges of electrons and holes and the comparison to the nuclei-charges. This results in the fact that concerning the band-edge potential, the electrons will, visually speaking, fall from 'above' into potential valleys, whereas the holes will rise from 'below' into potential hills. Thus the electron eigenvalues are 'above' the band-edges and the hole eigenvalues 'below'. Since the construction of the Schrödinger operators H_V is such that eigenvalues tend to infinity, we need to rotate the hole band-edges to get the correct eigenvalues and eigenfunctions. This explaines the sign in front of $V_{\xi,0}$.
The sign in front of the electrostatic potential φ has a similar reason. The potential φ represents a field produced by the different species. Since electrons will be attracted by positive charges, these have to be valleys in the landscape of the effective potential. Contrary, the electrons will be rejected by negative charges, resulting in potential hills. Therefore, φ has to be equipped with a negative sign for the electron effective potential. Similar arguments for holes result in a positive sign for φ in V_p.

3.3.2 Continuity Properties and Monotonicity

Let us now recall some properties of $\widetilde{\mathcal{N}}$ (and $\mathcal{N}$, respectively), that are fundamental for the analysis of the Kohn-Sham system. There are three main properties we focus on: continuity, monotonicity and analyticity.

Continuity

First we want to expand the representation (3.10) to test functions from L^2. This will be achieved by the next theorem.

Theorem 3.19.

i) $\widetilde{\mathcal{N}}$ takes its values in the space $L^\infty_{\mathbb{R}}$ and is a bounded mapping from $L^2_{\mathbb{R}}$ into $L^\infty_{\mathbb{R}}$.

ii) Assume the domain of H_0 embeds into a Hölder space C^α. Then $\widetilde{\mathcal{N}}$ takes its values in this space and is a bounded mapping from $L^2_{\mathbb{R}}$ into C^α.

Proof. Let $\mathcal{M}$ be a bounded set in L^2. Further, let $\tau \in \mathbb{R}$ be such, that $(H_V - \tau)^{-1}$ exists for all $V \in \mathcal{M}$ (see Corollary 3.13 iv)).
Theorem 3.7 together with Corollary 3.13 and Corollary 3.14 show, that the domain of any Schrödinger operator H_V ($V \in L^2_{\mathbb{R}}$) continously embeds into L^∞. Thus, we can estimate

$$\sup_{V\in\mathcal{M}} \|\widetilde{\mathcal{N}}(V)\|_{L^\infty} \leq \sum_{l=1}^{\infty} f(\lambda_l) \||\psi_l|^2\|_{L^\infty} \leq \sum_{l=1}^{\infty} f(\lambda_l) \|\psi_l\|^2_{L^\infty} .$$

Using Theorem 3.7 and Corollary 3.14 we get

$$\|\psi_l\|_{L^\infty} \leq c\|(H_0+1)\psi_l\|_{L^2} \leq c\|(H_0+1)(H_V-\tau)^{-1}\|_{\mathcal{B}} \|(H_V-\tau)\psi_l\|_{L^2} .$$

The term $\|(H_0+1)(H_V-\tau)^{-1}\|_{\mathcal{B}}$ is finite, thanks to Corollary 3.14, and can be estimated by $\sup_{V\in\mathcal{M}} \|(H_0+1)(H_V-\tau)^{-1}\|_{\mathcal{B}}$, see [50, Prop. 5.3] for details.
With

$$\|(H_V-\tau)\psi_l\|_{L^2} = \|(\lambda_l-\tau)\psi_l\|_{L^2} = |\lambda_l - \tau|$$

we finally estimate

$$\sup_{V\in\mathcal{M}} \|\widetilde{\mathcal{N}}(V)\|_{L^\infty} \leq \sup_{V\in\mathcal{M}} \|(H_0+1)(H_V-\tau)^{-1}\|^2_{\mathcal{B}} \sum_{l=1}^{\infty} f(\lambda_l)(\lambda_l-\tau)^2$$

where the sum is finite, due to the decaying properties of the distribution function f, cf. Assumption 3.17. This proves i).
The proof of ii) follows the same lines after estimating

$$\|\psi_l\|_{C^\alpha} \leq c\|(H_0+1)^{-1}\|_{\mathcal{B}(L^2,C^\alpha)} \|(H_0+1)\psi_l\|_{L^2} .$$

□

Remark 3.20. *Note, that* $\mathrm{dom}(H_0)$ *always embeds into a Hölder space if* $\Omega \cup \Gamma$ *is regular in the sense of Gröger, [38].*

With this (3.10) extends to $W \in L^2$

Corollary 3.21. *For every* $V \in L^2_{\mathbb{R}}$, (3.10) *leads to*

$$\int_\Omega W\mathcal{N}(V)\,\mathrm{d}x = \mathrm{tr}\,(Wf(H_0+V))$$

for all $W \in L^2$.

Proof. For $V \in L^2$ it follows from Theorem 3.19 that $\mathcal{N}(V)$ is in L^∞. Thus for $V \in L^2$ the left hand side is a bounded linear functional on L^2.
To see that the same is true for the right hand side, we expand

$$\mathrm{tr}\,(Wf(H_0+V)) = \mathrm{tr}\,\big(W(H_V-\tau)^{-1}(H_V-\tau)^{-2}(H_V-\tau)^3 f(H_0+V)\big)$$

The term $(H_V-\tau)^3 f(H_0+V)$ is bounded, due to the asymptotics of f (Assumption 3.17). Corollary 3.14 together with Theorem 3.15 and Theorem 3.7 yield $\|(H_V-\tau)^{-2}\|_{\mathcal{B}_1} < \infty$ and $\|W(H_V-\tau)^{-1}\|_{\mathcal{B}} < \infty$, respectively. Thus, the statement holds true by extending from L^∞ to L^2. □

With this at hand Lipschitz continuity of $\widetilde{\mathcal{N}}$ can be shown.

Theorem 3.22 (cf. [49, 50])**.** $\widetilde{\mathcal{N}}$*, regarded as a mapping from* $L^2_{\mathbb{R}}$ *into itself, is boundedly Lipschitz continuous.*

Remark 3.23. *Note that this implies not only continuity but boundedness for the operator* $\widetilde{\mathcal{N}}$ *as well.*

To carry the foregoing results over to $\mathcal{N}$ the following lemma is needed.

Lemma 3.24. *The function* $\mathcal{E}_F$*, assigning its Fermi-level* $\mathcal{E}_F(V)$ *to its potential* $V \in L^2_{\mathbb{R}}$*, is bounded on bounded sets and continuous.*

Remark 3.25. *Details and further properties of* $\mathcal{E}_F(V)$ *can be found in [47, 49, 50].*

This yields

Corollary 3.26. *The operator* $\mathcal{N} : L^2_{\mathbb{R}} \to L^2_{\mathbb{R}}$ *is continuous and bounded.*

For a proof of these results see [50].

Monotonicity

The following theorems about monotonicity of the particle density operator play an essential role in the existence proof.

Theorem 3.27 (cf. [47, 49, 50]). *The mapping $-\widetilde{\mathcal{N}} : L^2_{\mathbb{R}} \to L^2_{\mathbb{R}}$ is monotone.*

Corollary 3.28 (cf. [47, 49, 50]).

i) The operator $-\widetilde{\mathcal{N}}$ is also monotone, if regarded as a mapping $W^{1,2}_{\mathbb{R},\Gamma} \to W^{1,2}_{\mathbb{R},\Gamma}$.

ii) The operator $-\mathcal{N} : W^{1,2}_{\mathbb{R},\Gamma} \to W^{-1,2}_{\mathbb{R},\Gamma}$ is monotone as well.

For the proof, one first shows that, for any $V, W \in L^\infty_{\mathbb{R}}$, the positivity relation

$$tr\big((f(H_0+V) - f(H_0+W))(W-V)\big) \geq 0, \tag{3.11}$$

holds true, see [45]. Afterwards one extends this by continuity to all $V, W \in L^2_{\mathbb{R}}$ c.f. Corollary 3.21.

Remark 3.29. *Inequality (3.11) is a consequence of the formula*

$$tr([f(H_0+U) - f(H_0+V)](U-V)) = \sum_{k,l=1}^{\infty} (f(\lambda_k) - f(\mu_l))(\lambda_k - \mu_l)|(\psi_k, \xi_l)|^2$$

where the sets $\{\lambda_k, \psi_k\}$ and $\{\mu_l, \xi_l\}$ denote the eigenvalues and corresponding eigenfunctions of the operators $H_0 + U$ and $H_0 + V$, respectively. Together with the decay properties of the distribution function f, this yields the anti-monotonicity of the pseudo particle density operator $\widetilde{\mathcal{N}}$. The assertion follows for $\mathcal{N}$ as well, see [45] for details.

Analyticity

Lastly, it is possible to show another regularity property of $\widetilde{\mathcal{N}}$ ($\mathcal{N}$, respectively). There are indications that analyticity of the particle density operator $\widetilde{\mathcal{N}}$, which is equivalent to the analyticity of the operator mapping $V \mapsto f(H_V)$, may be a gainful property for establishing steadily converging iteration schemes for the Kohn-Sham system (cf. [27]). Gajewski and Griepentrog used analyticity to proof a generalised Łojasiewicz–Simon inequality (cf. [13], [23], [27]), which was then taken to setup a descent method for the free energy of a multicomponent system, cf. [27], similar in structure to the Kohn-Sham system.
Furthermore, analyticity plays an important role in bifurcation analysis, cf. [95]. In fact there are indications, in analysis (eg. [79]) as well as numerics (eg. [69]), that the Kohn-Sham system may have multiple solutions. However, under special conditions the solution is unique, cf. [49], [50], [78].

First note, that the distribution functions we mainly have in mind are given by Fermi-integrals. As such, they have singularities in the closed left half plane. Thus, we cannot ask f to be holomorphic on the whole complex plane. Instead, we make the following assumptions, cf. [50] and [43], by expanding Assumption 3.17 in the following way.

Assumption 3.30. Denote for every $\alpha > 0$ by Υ_α the contour

$$\{\lambda \,:\, \lambda = s \pm i\alpha s,\, s \geq 0\}$$

with positive orientation. $\mathcal{P}_\alpha$ then stands for the set of points in $\mathbb{C}$ that are enclosed by Υ_α, i. e.

$$\mathcal{P}_\alpha := \{\lambda_1 + i\lambda_2 \,:\, \lambda_1 > 0, |\lambda_2| < \alpha\lambda_1\}.$$

Assume, that for every $t \in \mathbb{R}$ there is an $\alpha > 0$ such that f is defined and holomorphic on $\mathcal{P}_\alpha - t$. Moreover, there is an $\alpha > 0$ such that

$$\sup_{\lambda \in \mathcal{P}_\alpha - t} |f(\lambda)\lambda^9| < \infty\,.$$

Remark 3.31. *From Assumption 3.30 follows in particular that for every $t \in \mathbb{R}$ there is an $\alpha > 0$ such that*

$$\sup_{\lambda \in \mathcal{P}_\alpha - t} |\lambda^9 f(\lambda)| < \infty \quad \textit{and} \quad \int_\Upsilon |\lambda|^7 |f(\lambda)|\, d|\lambda| < \infty,$$

where Υ is the contour corresponding to $\mathcal{P}_\alpha - t$ in the sense of Assumption 3.30. This comes to bear in the proof of Lemma 3.37.

Let us now define, what we mean by 'analyticity' of a mapping from one Banach space into the other.

Definition 3.32. Following Vaĭnberg[95, Ch. 22], cf. also [12], [41, Ch. III.3], we call a mapping $F_j : X \to Y$, $j \in \mathbb{N}$, between two Banach spaces a *j-power mapping*, if there is a continuous mapping $G_j : X \oplus \ldots \oplus X \to Y$ which is linear in each of its j arguments, such that $F_j(\mathfrak{x}) = G_j(\mathfrak{x}, \ldots, \mathfrak{x})$. A mapping $F : X \to Y$ is called *analytic* in a point $\mathfrak{x}_0 \in X$ if there is a ball $B \subset X$ around zero and a sequence $\{F_j\}_{j \in \mathbb{N}}$ of j-power mappings such that

$$F(\mathfrak{x}_0 + \mathfrak{x}) = F(\mathfrak{x}_0) + \sum_{j=1}^{\infty} F_j(\mathfrak{x}) \quad \text{for all } \mathfrak{x} \in B,$$

and the series converges in Y uniformly for $\mathfrak{x} \in B$.

Analytic mappings possess many properties analogous to those of classical holomorphic functions, cf. [95, Ch. 22] for details. The following theorem states the main result concerning analyticity.

Theorem 3.33 (cf. [43])**.** *Let us assume the distribution function f fulfils Assumption 3.17 and 3.30. Then the mapping $L^2_{\mathbb{R}} \ni V \mapsto \widetilde{\mathcal{N}}(V) \in L^2_{\mathbb{R}}$ (and $\mathcal{N}$, respectively) is analytic in every point $V \in L^2_{\mathbb{R}}$.*

In what follows we will point out the main ideas for the proof of Theorem 3.33. Further details can be found in [43].
A key to the proof is the next proposition, cf. e.g. [6].

Proposition 3.34. *If F is a mapping between the Banach spaces X and Y, then F is analytic iff it is weakly analytic, i.e. for every $y^* \in Y^*$ the mapping $X \ni x \to \langle F(x), y^* \rangle$ is an analytic mapping into the corresponding field $\mathbb{C}$ or $\mathbb{R}$.*

This means, having the relation

$$\int_\Omega W\mathcal{N}(V)\,\mathrm{d}x = \operatorname{tr}\big(Wf(H_0+V)\big), \quad W \in L^2 \tag{3.12}$$

at hand (see Corollary 3.21) and considering the linearity and continuity of the trace, we only need to prove that for every $W \in L^2_{\mathbb{R}}$, the mapping

$$L^2_{\mathbb{R}} \ni V \mapsto Wf(H_0+V) \in \mathcal{B}_1$$

is analytic.

We will thus proceed in the following way. Choose a number $\rho \in \mathbb{R}$, such that 1 is a lower form bound for each of the operators $H_0 + V + \rho$, provided $V \in L^2_{\mathbb{R}}$ with $\|V\|_{L^2} \leq 1$ (cf. Corollary 3.13). Define $H := H_0 + \rho$ and $g := f(\cdot - \rho)$ and write $f(H_0+V) = g(H+V)$ in the usual way as a Dunford integral (see [20, Ch. VII.9])

$$g(H+V) = -\frac{1}{2\pi i}\int_\Upsilon g(\lambda)(H+V-\lambda)^{-1}\,d\lambda\,. \tag{3.13}$$

Υ denotes the contour corresponding to $\mathcal{P}_\alpha$ such that the function g is holomorphic on the set $\mathcal{P}_\alpha - 1$, cf. Assumption 3.30, and $\sup_{\lambda\in\mathcal{P}_\alpha-1}|\lambda^9 g(\lambda)| < \infty$.

Remark 3.35. *Note that then $\sup_{\lambda\in\mathcal{P}_\alpha-1}|\lambda^9 g(\lambda)| < \infty$ and by definition of ρ, Υ encloses the spectrum of $H+V$ for all $V \in L^2_{\mathbb{R}}$ with $\|V\|_{L^2} \leq 1$.*

Next we expand $(H+V-\lambda)^{-1}$ into a Neumann series

$$(H+V-\lambda)^{-1} = (H-\lambda)^{-1}\sum_{j=0}^{\infty}(-1)^j\big(V(H-\lambda)^{-1}\big)^j \tag{3.14}$$

and define the j-linear mapping T_j

$$T_j(V) := \frac{(-1)^{j+1}}{2\pi i}\int_\Upsilon g(\lambda)(H-\lambda)^{-1}\big(V(H-\lambda)^{-1}\big)^j\,\mathrm{d}\lambda\,. \tag{3.15}$$

These mappings fulfil the following properties

Lemma 3.36. *1. For every $V \in L^2_{\mathbb{R}}$, the operator $T_j(V)$ is bounded and selfadjoint.*

2. For every $V, W \in L^2_{\mathbb{R}}$, one has $WT_j(V) \in \mathcal{B}_1$.

Finally, we have to show that for $W \in L^2_{\mathbb{R}}$ and V from a sufficiently small ball in $L^2_{\mathbb{R}}$ around 0 one has

$$Wf(H+V) = \sum_{j=0}^{\infty} WT_j(V), \tag{3.16}$$

where the series on the right hand converges in $\mathcal{B}_1$. In particular this means that interchanging integration in the Dunford integral and summation in the Neumann series (3.14) is possible.

In order to prove Lemma 3.36 and Theorem 3.33 we need some technical assertions. Namely, a comparison principle that allows to replace the term $(H-\lambda)^{-1}$ by the λ-independent term H^{-1} and a continuity assertion for the trace operator in (3.12).

Lemma 3.37. *If A is a selfadjoint operator on a Hilbert space $\mathfrak{H}$ the spectrum of which is contained in $[1, \infty[$, then*

$$\sup_{\lambda \in \Upsilon} \|A(A-\lambda)^{-1}\|_{\mathcal{B}(\mathfrak{H})} \leq \frac{1}{\operatorname{dist}(1, \Upsilon)} \tag{3.17}$$

for all $\Upsilon = \Upsilon_\alpha$ with $\alpha > 0$, cf. Assumption 3.30.

The proof of this lemma runs by classical arguments from the theory of linear operators, cf. e.g. [51, Ch. V.3.5], and the special choice of Υ, see [43] for details.

Lemma 3.38. *If A is a selfadjoint operator on L^2 such that $L^2_{\mathbb{R}} \ni W \mapsto WA \in \mathcal{B}_1$ is continuous. Then the linear form $L^2_{\mathbb{R}} \ni W \mapsto tr(WA)$ is continuous and takes real values. Hence, it may be identified with an element from $L^2_{\mathbb{R}}$.*

Proof. Continuity is clear by continuity of the trace on $\mathcal{B}_1$. Concerning the second assertion, one has by splitting $W \in L^\infty_{\mathbb{R}}$ into its positive and negative part, $W = W_+ - W_-$,

$$\operatorname{tr}\big(WA\big) = \operatorname{tr}\big(W_+^{1/2} A\, W_+^{1/2}\big) - \operatorname{tr}\big(W_-^{1/2} A\, W_-^{1/2}\big). \tag{3.18}$$

Both addends on the r.h.s. are real, because the operators $W_+^{1/2} A\, W_+^{1/2}$ and $W_-^{1/2} A\, W_-^{1/2}$ are selfadjoint. □

The proof of the second assertion in Lemma 3.36 in case of $j \leq 7$ requires only some algebraic manipulations in the integrand. Exemplarily, we show the assertion for $WT_2(V)$. Making use of the resolvent equation

$$(H-\lambda)^{-1} = H^{-1} + \lambda H^{-1}(H-\lambda)^{-1} \tag{3.19}$$

we get

$$\begin{aligned} WT_2(V) &= \frac{-W}{2\pi i}\int\limits_{\Upsilon} g(\lambda)(H-\lambda)^{-1}V(H-\lambda)^{-1}V(H-\lambda)^{-1}\,d\lambda \\ &= \frac{-W}{2\pi i}\int\limits_{\Upsilon} g(\lambda)(H-\lambda)^{-1}\Big[VH^{-1}VH^{-1} \\ &\qquad + \lambda VH^{-1}(H-\lambda)^{-1}VH^{-1} + \lambda VH^{-1}VH^{-1}(H-\lambda)^{-1} \\ &\qquad + \lambda^2 VH^{-1}(H-\lambda)^{-1}VH^{-1}(H-\lambda)^{-1}\Big]\,d\lambda\,. \end{aligned}$$

Using the resolvent equation (3.19) again in those summands where $(H-\lambda)^{-1}$ appears exactly once as a factor, yields

$$\begin{aligned} WT_2(V) = \frac{-W}{2\pi i}\int\limits_{\Upsilon} g(\lambda)(H-\lambda)^{-1}\Big[&(VH^{-1})^2 + \lambda V(H^{-2}VH^{-1} + H^{-1}VH^{-2}) \\ &+ \lambda^2 VH^{-2}(H-\lambda)^{-1}VH^{-1} + \lambda^2 VH^{-1}VH^{-2}(H-\lambda)^{-1} \\ &+ \lambda^2 VH^{-1}(H-\lambda)^{-1}VH^{-1}(H-\lambda)^{-1}\Big]\,d\lambda\,. \end{aligned}$$

We discuss the summands separately. For the first term we get

$$-\frac{1}{2\pi i}W\int\limits_{\Upsilon} g(\lambda)(H-\lambda)^{-1}(VH^{-1})^2\,d\lambda = Wg(H)(VH^{-1})^2$$

which belongs to $\mathcal{B}_1$ and admits the estimate

$$\|Wg(H)(VH^{-1})^2\|_{\mathcal{B}_1} \le \|Wg(H)\|_{\mathcal{B}_1}\|V\|_{L^2}^2\|H^{-1}\|_{\mathcal{B}(L^2;L^\infty)}^2 \le c\|V\|_{L^2}^2$$

according to Theorem 3.7, Remark 3.18 and Corollary 3.14. If $\tilde{g}$ denotes the function $\lambda \mapsto \lambda g(\lambda)$, then

$$\begin{aligned} -\frac{1}{2\pi i}W\int\limits_{\Upsilon} \lambda g(\lambda)(H-\lambda)^{-1}VH^{-2}VH^{-1}\,d\lambda &= W\tilde{g}(H)VH^{-2}VH^{-1}\,, \\ -\frac{1}{2\pi i}W\int\limits_{\Upsilon} \lambda g(\lambda)(H-\lambda)^{-1}VH^{-1}VH^{-2}\,d\lambda &= W\tilde{g}(H)VH^{-1}VH^{-2}\,, \end{aligned}$$

and one can estimate

$$\begin{aligned} &\|W\tilde{g}(H)VH^{-2}VH^{-1}\|_{\mathcal{B}_1} + \|W\tilde{g}(H)VH^{-1}VH^{-2}\|_{\mathcal{B}_1} \\ &\le 2\|W\tilde{g}(H)\|_{\mathcal{B}}\|V\|_{L^2}^2\|H^{-1}\|_{\mathcal{B}(L^2;L^\infty)}\|H^{-1}\|_{\mathcal{B}_2}^2 \\ &\le 2\|W\|_{L^2}\|V\|_{L^2}^2\|H^{-1}\|_{\mathcal{B}(L^2;L^\infty)}\|H^{-1}\|_{\mathcal{B}_2}^2 \sup_{s\in\,\mathrm{spec}(H)}|s\,g(s)| < \infty. \end{aligned}$$

In order to estimate the first of the terms with λ^2 we note that the integral

$$\begin{aligned}
&\int_\Upsilon |\lambda^2 g(\lambda)| \|(H-\lambda)^{-1} V H^{-2} (H-\lambda)^{-1} V H^{-1}\|_{\mathcal{B}(L^2;\mathcal{D})}\, d|\lambda| \\
&\leq c \int_\Upsilon |\lambda^2 g(\lambda)| \|H(H-\lambda)^{-1} V H^{-2} (H-\lambda)^{-1} V H^{-1}\|_{\mathcal{B}}\, d|\lambda| \\
&\leq c \sup_{\lambda\in\Upsilon} \|H(H-\lambda)^{-1}\|_{\mathcal{B}}^2 \|V\|_{L^2}^2 \|H^{-2}\|_{\mathcal{B}} \|H^{-1}\|^2_{\mathcal{B}(L^2;L^\infty))} \int_\Upsilon |\lambda^2 g(\lambda)|\, d|\lambda|
\end{aligned}$$

is finite. Hence, one has

$$\begin{aligned}
&-\frac{1}{2\pi i} W \int_\Upsilon \lambda^2 g(\lambda) (H-\lambda)^{-1} V H^{-2} (H-\lambda)^{-1} V H^{-1}\, d\lambda \\
&\qquad\qquad = -\frac{1}{2\pi i} \int_\Upsilon \lambda^2 g(\lambda) W (H-\lambda)^{-1} V H^{-2} (H-\lambda)^{-1} V H^{-1}\, d\lambda \in \mathcal{B}.
\end{aligned}$$

Actually, this integral is a nuclear operator and can be estimated as follows:

$$\begin{aligned}
&\frac{1}{2\pi} \Big\| \int_\Upsilon \lambda^2 g(\lambda) W (H-\lambda)^{-1} V H^{-2} (H-\lambda)^{-1} V H^{-1}\, d\lambda \Big\|_{\mathcal{B}_1} \\
&\leq c \int_\Upsilon |\lambda^2 g(\lambda)| \|W(H-\lambda)^{-1} V H^{-1} H^{-2} H (H-\lambda)^{-1} V H^{-1}\|_{\mathcal{B}_1}\, d|\lambda| \\
&\leq c \sup_{\lambda\in\Upsilon} \|H(H-\lambda)^{-1}\|_{\mathcal{B}}^2 \|W\|_{L^2} \|V\|_{L^2}^2 \|H^{-1}\|^3_{\mathcal{B}(L^2;L^\infty)} \|H^{-1}\|^2_{B_2} \int_\Upsilon |\lambda^2 g(\lambda)|\, d|\lambda|\,.
\end{aligned}$$

This is finite, due to Lemma 3.37, Corollary 3.14, Theorem 3.15, and Assumption 3.30. The terms

$$\begin{aligned}
-\frac{1}{2\pi i} W \int_\Upsilon \lambda^2 g(\lambda) (H-\lambda)^{-1} V H^{-1} V H^{-2} (H-\lambda)^{-1}\, d\lambda, \\
-\frac{1}{2\pi i} W \int_\Upsilon \lambda^2 g(\lambda) (H-\lambda)^{-1} V H^{-1} (H-\lambda)^{-1} V H^{-1} (H-\lambda)^{-1}\, d\lambda
\end{aligned}$$

can be treated analogously.

Remark 3.39. *In the proof we used that $\int_\Upsilon |\lambda|^2 |f(\lambda)|\, d|\lambda|$ is finite. Analogously, one uses that the integral $\int_\Upsilon |\lambda|^7 |f(\lambda)|\, d|\lambda|$ is finite to prove the assertion for $WT_7(V)$. This is why we asked for $|\lambda|$ to the power of 9 in the supremum condition of Assumption 3.30, see Remark 3.31.*

The more crucial point is to obtain estimates for the summands with index larger than 7 which, additionally, shall allow the interchange of integration and summation. Assume $j > 7$, then we can estimate:

$$\|W(H-\lambda)^{-1}\big(V(H-\lambda)^{-1}\big)^j\|_{\mathcal{B}_1} \leq$$

$$\leq \|WH^{-1}\|_{\mathcal{B}_7}\|VH^{-1}\|^6_{\mathcal{B}_7}\|VH^{-1}\|^{j-6}_{\mathcal{B}} \sup_{\lambda\in\Upsilon}\|H(H-\lambda)^{-1}\|^{j+1}.$$

Now, taking into account (3.7) and Lemma 3.37, we can further estimate

$$\leq \gamma\|W\|_{L^2}\|V\|^j_{L^2}\|H^{-1}\|^{j-6}_{\mathcal{B}(L^2;L^\infty)}\frac{1}{\big(dist(1,\Upsilon)\big)^{j+1}}.$$

Thus, choosing V from the ball $\|V\|_{L^2} < min(1, \frac{dist(1,\Upsilon)}{\|H^{-1}\|_{\mathcal{B}(L^2;L^\infty)}})$ one recognises that the absolute value of the corresponding terms in (3.15) behaves as the addends of a geometric series, which at the end allows to interchange summation and integration, c.f. [91, Ch. IV.4 Thm. 45].

3.4 The Poisson Operator

In this section we introduce the Poisson operator governing the electrostatic potential φ, which is a constituent of the Kohn-Sham system. Let us start by fixing the boundary conditions and formulating assumptions on the dielectric permittivity function.

Assumption 3.40. The function ε, representing the dielectric permittivity on Ω, takes its values in the set of symmetric, positive definite $d \times d$-matrices. We assume that ε is Lebesgue measurable and bounded, such that ε^{-1} is bounded as well.

Concerning the electrostatic potential we regard the following boundary conditions

$$\varphi = \tilde{\varphi}_1 \quad \text{on } \Gamma\,, \quad -\langle\varepsilon\nabla\varphi,\nu\rangle - b\varphi = -b\tilde{\varphi}_2 \quad \text{on } \partial\Omega\setminus\Gamma\,,$$

where $\tilde{\varphi}_1$ are the boundary values given on Γ and $\tilde{\varphi}_2$ those for the inhomogeneous boundary conditions of third kind on $\partial\Omega\setminus\Gamma$. ν denotes the outer unit normal at the boundary $\partial\Omega$. The Dirichlet conditions on Γ model Ohmic contacts and the part $\partial\Omega\setminus\Gamma$ covers interfaces between the semiconductor device and insulators (with capacity $b > 0$) or homogeneous Neumann boundary conditions ($b = 0$).
About the boundary conditions we further assume:

Assumption 3.41. Let $b \geq 0$ be from $L^\infty(\partial\Omega\setminus\Gamma)$ (with respect to the surface measure) and let either the surface measure of Γ be nonzero or b be strictly positive on a subset of $\partial\Omega\setminus\Gamma$ with positive surface measure.

In what follows we want to solve the inhomogeneous partial differential equation

$$-\nabla \cdot \varepsilon \nabla \varphi^* = p - n + D \tag{3.20}$$

equipped with the above described inhomogeneous boundary conditions. We fix a function $\varphi_0 \in H^1$, which fulfils

$$\varphi_0 = \tilde{\varphi}_1 \text{ on } \Gamma\,. \quad \text{and } \varphi_0 = 0 \text{ on } \partial\Omega \setminus \Gamma$$

Then, the solution φ^* to (3.20) splits up as

$$\varphi^* = \varphi_0 + \varphi\,,$$

with $\varphi \in H^1_\Gamma$.

Definition 3.42. We define the linear Poisson operator $-\nabla \cdot \varepsilon \nabla : W^{1,2}_{\mathbb{R}} \to W^{-1,2}_{\mathbb{R},\Gamma}$ by

$$\langle -\nabla \cdot \varepsilon \nabla \varphi, \phi \rangle := \int_\Omega \varepsilon \nabla \varphi \cdot \nabla \psi \ \mathrm{d}x + \int_{\partial\Omega\setminus\Gamma} b(\tau)\varphi(\tau)\phi(\tau) d\tau, \quad \varphi \in W^{1,2}_{\mathbb{R}},\, \phi \in W^{1,2}_{\mathbb{R},\Gamma}. \tag{3.21}$$

The definition is correct, because $W^{1,2}_{\mathbb{R}}$ embeds continuously into $L^2(\partial\Omega \setminus \Gamma)$, cf. [28].

Remark 3.43. *Note that choosing $W^{1,2}_{\mathbb{R}}$ instead of $W^{1,2}_{\mathbb{R},\Gamma}$ as the domain of the Poisson operator defined above is necessary to have the possibility of allowing for inhomogeneous Dirichlet boundary conditions as introduced by insertion of φ_0.*

Some fundamental properties of the operator just introduced are the following:

Lemma 3.44.

i) $-\nabla \cdot \varepsilon \nabla$ *is continuous*

ii) Assume $\varphi_0 \in W^{1,2}_{\mathbb{R}}$*, then the mapping*

$$W^{1,2}_{\mathbb{R},\Gamma} \ni \varphi \mapsto -\nabla \cdot \epsilon \nabla(\varphi_0 + \varphi) \in W^{-1,2}_{\mathbb{R},\Gamma}$$

is strongly monotone.

iii) $(-\nabla \cdot \varepsilon \nabla)^{-1}$ *maps* $L^1_{\mathbb{R}}$ *continuously into* $L^2_{\mathbb{R}}$

Proof.
i) follows from the boundedness of ε and b.

ii) Due to the assumptions on ε^{-1} and b, we can estimate

$$\begin{aligned}
&\langle\big(-\nabla\cdot\varepsilon\nabla(\varphi_0+\varphi_1)+\nabla\cdot\varepsilon\nabla(\varphi_0+\varphi_2)\big),\varphi_1-\varphi_2\rangle\\
=\;&\int\limits_\Omega \varepsilon\nabla(\varphi_1-\varphi_2)\cdot\nabla(\varphi_1-\varphi_2)\,\mathrm{d}x+\int\limits_{\partial\Omega\setminus\Gamma} b(\varphi_1-\varphi_2)^2 d\tau\\
\geq\;& c_1\int\limits_\Omega \nabla(\varphi_1-\varphi_2)\cdot\nabla(\varphi_1-\varphi_2)\,\mathrm{d}x+c_2\left|\int\limits_{\partial\Omega\setminus\Gamma}(\varphi_1-\varphi_2)d\tau\right|^2\\
\geq\;& c\int\limits_\Omega |\nabla(\varphi_1-\varphi_2)|^2\,\mathrm{d}x\,.
\end{aligned}$$

iii) As for the resolvent of the Schrödinger operator H_0+1, one can show $\big(-\nabla\cdot\varepsilon\nabla\big)^{-1}\in\mathcal{B}(L^2_{\mathbb{R}};L^\infty_{\mathbb{R}})$. From this, the assertion follows by the selfadjointness of $-\nabla\cdot\varepsilon\nabla$ on $L^2_{\mathbb{R}}$ and duality, since for $\psi\in L^2_{\mathbb{R}}\subset L^1_{\mathbb{R}}$ we have

$$\begin{aligned}
\|(-\nabla\varepsilon\nabla)^{-1}\psi\|_{L^2_{\mathbb{R}}} &= \sup_{\|\varphi\|_{L^2_{\mathbb{R}}}=1}|\langle(-\nabla\varepsilon\nabla)^{-1}\psi,\varphi\rangle|\\
&= \sup_{\|\varphi\|_{L^2_{\mathbb{R}}}=1}|\langle\psi,(-\nabla\varepsilon\nabla)^{-1}\varphi\rangle|\leq\|\psi\|_{L^1_{\mathbb{R}}}\|(-\nabla\varepsilon\nabla)^{-1}\|_{\mathcal{B}(L^2_{\mathbb{R}},L^\infty_{\mathbb{R}})}
\end{aligned}$$

□

Corollary 3.45. *Assume $\varphi_0\in W^{1,2}_{\mathbb{R}},V_n,V_p\in L^2_{\mathbb{R}}$. Then the operator*

$$W^{1,2}_{\mathbb{R},\Gamma}\ni\varphi\mapsto-\nabla\cdot\varepsilon\nabla(\varphi_0+\varphi)+\mathcal{N}_n(V_n-\varphi)-\mathcal{N}_p(V_p+\varphi)\in W^{-1,2}_{\mathbb{R},\Gamma}$$

is strongly monotone and its monotonicity constant is not smaller than that of $-\nabla\cdot\varepsilon\nabla: W^{1,2}_{\mathbb{R},\Gamma}\to W^{-1,2}_{\mathbb{R},\Gamma}$. $\mathcal{N}_n$ and $\mathcal{N}_p$ denote the particle density operators corresponding to electrons and holes, as described in Section 3.3.1.

Proof. The proof results easily from Lemma 3.44, Corollary 3.28 and the fact that the $W^{1,2}_{\mathbb{R},\Gamma}\leftrightarrow W^{-1,2}_{\mathbb{R},\Gamma}$ duality is the extended L^2 duality. □

3.5 The Kohn-Sham System

In this section we finally come to analyse the Kohn-Sham system. First we define what we will call a solution.

Definition 3.46. We call $(\varphi_0+\varphi,n,p)\in H^1_\Gamma\times L^2\times L^2$ a solution of the Kohn-Sham system, iff

$$\begin{aligned}
-\nabla\cdot\varepsilon\nabla\varphi &= \widetilde{D}+p-n\\
n &= \mathcal{N}_n(V_n+V_{xc,n}(n,p)-\varphi)\\
p &= \mathcal{N}_p(V_p+V_{xc,p}(n,p)+\varphi)\,,
\end{aligned}$$

where $\widetilde{D} = D + \nabla \cdot \varepsilon \nabla \varphi_0$.

Remark 3.47.

i) As described previously, the function φ_0 represents the inhomogeneous Dirichlet boundary conditions for Poisson's problem. Thus, the problem we need to solve for getting φ is

$$-\nabla \cdot \varepsilon \nabla(\varphi_0 + \varphi) = D + p - n$$

or

$$-\nabla \cdot \varepsilon \nabla(\varphi) = \widetilde{D} + p - n\,, \text{ with } \widetilde{D} = D + \nabla \cdot \varepsilon \nabla(\varphi_0)$$

ii) As described in Section 3.3.1 $\mathcal{N}_n$ and $\mathcal{N}_p$ are the electron and hole densities, respectively.

iii) We chose the requirement $n, p \in L^2_{\mathbb{R}}$ in order to have the right-hand side of Poisson's equation in a space, which embeds into $W^{-1,2}_{\mathbb{R},\Gamma}$, cf. Definition 3.42.

3.5.1 Existence of solutions and fixed point formulation

There will be two statements in this section. Firstly, we will present the main existence theorem concerning solutions to the Kohn-Sham system. Secondly, a theorem will allow for a fixed point formulation, which will be interesting from a numerical point of view.

Theorem 3.48 (Existence of solutions)**.** *Suppose that $\Omega \cup \Gamma$ is regular in the sense of Gröger. Further assume that the operators $V_{xc,n}, V_{xc,p} : L^1_{\mathbb{R}} \times L^1_{\mathbb{R}} \to L^2_{\mathbb{R}}$ are bounded and continuous.*
Then under the general assumptions on ε, m_ξ and b, made above, the Kohn-Sham system has a solution for every $D \in W^{-1,2}_{\mathbb{R},\Gamma}$.

Proof. We will apply Schauder's Fixed Point Theorem. As the required closed, convex set we take

$$K := \{(n,p) \,:\, n,p \in L^1_{\mathbb{R}}\,,\, n,p \geq 0\,,\, \int_\Omega n \,\mathrm{d}x = N_n\,,\, \int_\Omega p \,\mathrm{d}x = N_p\}\,,$$

where N_n and N_p are the fixed numbers of electrons and holes, respectively, in the semiconductor device domain. The mapping Ψ will be defined as follows:

$$\Psi : (n,p) \mapsto (\mathcal{N}_n(V_n + V_{xc,n}(n,p) - \Phi(n,p)), \mathcal{N}_p(V_p + V_{xc,p}(n,p) + \Phi(n,p)))\,, \tag{3.22}$$

where

$$\Phi(n,p) := (-\nabla \cdot \varepsilon \nabla)^{-1}\,(p - n + \widetilde{D})\,. \tag{3.23}$$

Ψ takes its values in K by definition of $\mathcal{N}_n$ and $\mathcal{N}_p$. Moreover, according to Lemma 3.44, $\Phi : L^1_{\mathbb{R}} \times L^1_{\mathbb{R}} \to L^2$ is continuous and bounded. Hence, Ψ is continuous by Corollary 3.26. Finally, the image of K is precompact by Theorem 3.19 because the set of potentials

occurring in the argument of $\mathcal{N}_n$ and $\mathcal{N}_p$ in (3.22) is $L^2_{\mathbb{R}}$-bounded. Thus, by Schauder's theorem, Ψ has a fixed point (n, p) which is from $C^\alpha \times C^\alpha \hookrightarrow L^2_{\mathbb{R}} \times L^2_{\mathbb{R}}$, cf. Theorem 3.19. Defining φ by the right-hand side of (3.23) one gets a solution $(\varphi_0 + \varphi, n, p)$ of the Kohn-Sham system, cf. Definition 3.46 □

Remark 3.49. *According to Theorem 3.19 we already know that the electron and hole densities of a solution are a priori Hölder continuous. Then, in dependence on the regularity of the doping profile D and on the inhomogeneity φ_0, the electrostatic potential φ often has better properties than only $W^{1,2}_{\mathbb{R}}$ due to elliptic regularity, see e.g. [38].*

With the upcoming numerics in mind we now present another formulation of the fixed point mapping. An important part of this construction will be the nonlinear Poisson equation

$$\mathcal{P}_{\mathfrak{V}}(\varphi) := -\nabla \cdot \varepsilon \nabla \varphi + \mathcal{N}_n(V_n - \varphi) - \mathcal{N}_p(V_p + \varphi) = \widetilde{D}\,, \quad \mathfrak{V} = (V_n, V_p)\,, \tag{3.24}$$

to which the system reduces when the exchange-correlation potential is omitted. Due to Corollary 3.45, this operator is strongly monotone.
Basic for this approach is the fundamental result about monotone operator equations, cf. [28].

Proposition 3.50. *Let T be a strongly monotone and boundedly Lipschitz continuous operator between the Hilbert space $\mathfrak{H}$ and its dual $\mathfrak{H}^*$. Then the equation*

$$Tu = f \tag{3.25}$$

admits for any $\mathrm{f} \in \mathfrak{H}^*$ exactly one solution. Let $\mathcal{J} : \mathfrak{H} \to \mathfrak{H}^*$ be the duality mapping, m_T the monotonicity constant of T and M_T be the local Lipschitz constant of T belonging to a centred ball K in $\mathfrak{H}$ with radius not smaller than

$$\frac{2}{m_T}\|T(0) - \mathrm{f}\|_{\mathfrak{H}^*}. \tag{3.26}$$

Then the operator

$$\mathrm{u} \to \mathrm{u} - \frac{m_T}{M_T^2}\mathcal{J}^{-1}\left(Tu - \mathrm{f}\right) \tag{3.27}$$

maps the ball K strictly contractive into itself and its contractivity constant does not exceed

$$\sqrt{1 - \frac{m_T^2}{M_T^2}}. \tag{3.28}$$

The fixed point of (3.27) *is identical with the solution of* (3.25).

Having this at hand, we can define the operator

$$\mathcal{L} : L^2_{\mathbb{R}} \times L^2_{\mathbb{R}} \to W^{1,2}_{\mathbb{R},\Gamma}$$

assigning the solution φ of

$$\mathcal{P}_{\mathfrak{V}}\varphi = \widetilde{D}$$

to the given potentials $\mathfrak{V} = (V_n, V_p)$.

Theorem 3.51. *Let V_n and V_p be from $L^2_{\mathbb{R}}$. Further, let $\varphi := \mathcal{L}(V_n, V_p)$ be the solution to (3.24). Then $\mathcal{L} : L^2_{\mathbb{R}} \times L^2_{\mathbb{R}} \to W^{1,2}_{\mathbb{R},\Gamma}$ is continuous.*

Proof. Let φ be from $W^{1,2}_{\mathbb{R},\Gamma}$ and m_A be the monotonicity constant of the operator $\mathcal{P}_{\mathfrak{V}}$, compare Corollary 3.45. By definition of the dual-norm and of strong monotonicity we have

$$\begin{aligned} & \|\mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0)\|_{H^{-1}_{\Gamma}} && = \sup_{\psi \in H^1_{\Gamma}} \frac{\langle \mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0), \psi \rangle}{\|\psi\|_{H^1_{\Gamma}}} \\ \Rightarrow \quad & \|\mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0)\|_{H^{-1}_{\Gamma}} && \geq \frac{\langle \mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0), \varphi \rangle}{\|\varphi\|_{H^1_{\Gamma}}} \\ \Leftrightarrow \quad & \|\mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0)\|_{H^{-1}_{\Gamma}} \|\varphi\|_{H^1_{\Gamma}} && \geq \langle \mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0), \varphi \rangle \end{aligned}$$

and

$$\begin{aligned} & \langle \mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0), \varphi \rangle && \geq m_A \|\varphi\|^2_{H^1_{\Gamma}} \\ \Leftrightarrow \quad & \frac{1}{m_A} \frac{\langle \mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0), \varphi \rangle}{\|\varphi\|_{H^1_{\Gamma}}} && \geq \|\varphi\|_{H^1_{\Gamma}} \end{aligned}$$

Thus, we get the estimate

$$\|\varphi\|_{H^1_{\Gamma}} \leq \frac{1}{m_A} \|\mathcal{P}_{\mathfrak{V}}\varphi - \mathcal{P}_{\mathfrak{V}}(0)\|_{H^{-1}_{\Gamma}} = \frac{1}{m_A} \|\widetilde{D} - \mathcal{N}_n(V_n) + \mathcal{N}_p(V_p)\|_{H^{-1}_{\Gamma}}, \tag{3.29}$$

In particular, this gives an a priori estimate uniform for V_n, V_p from bounded sets in $L^2_{\mathbb{R}}$, see Theorem 3.19.
Assume now that the assertion is false. Then there is an $\epsilon > 0$ and a sequence $\{\mathfrak{V}_k\}_k$ in $L^2_{\mathbb{R}} \times L^2_{\mathbb{R}}$ converging in this space to $\mathfrak{V} = (V_n, V_p)$, such that $\|\varphi_k - \varphi\|_{W^{1,2}_{\mathbb{R},\Gamma}} = \|\mathcal{L}(\mathfrak{V}_k) - \mathcal{L}(\mathfrak{V})\|_{W^{1,2}_{\mathbb{R},\Gamma}} > \epsilon$, φ and φ_k being the corresponding solutions of (3.24). Since the sequence $\{\varphi_k\}_k$ is bounded in $W^{1,2}_{\mathbb{R},\Gamma}$, due to (3.29), there is a subsequence $\{\varphi_l\}_l$ which converges strongly in $L^2_{\mathbb{R}}$ to an element $\hat{\varphi} \in L^2_{\mathbb{R}}$. Thanks to the continuity properties of the operators $\mathcal{N}_n, \mathcal{N}_p$ (c.f. Theorem 3.22), this gives

$$\mathcal{N}_n(V_{n,l} + \varphi_l) - \mathcal{N}_p(V_{p,l} - \varphi_l) + E \longrightarrow \mathcal{N}_n(V_n + \hat{\varphi}) - \mathcal{N}_p(V_p - \hat{\varphi}) + E$$

in $W^{-1,2}_{\mathbb{R},\Gamma}$ for $l \to \infty$. Hence,

$$\{\varphi_l\}_l = \{(-\nabla \cdot \varepsilon \nabla)^{-1} \left(\mathcal{N}_n(V_{n,l} + \varphi_l) - \mathcal{N}_p(V_{p,l} - \varphi_l) + E\right)\}_l$$

converges in $W^{1,2}_{\mathbb{R},\Gamma}$ to an element $\check{\varphi}$. But $\check{\varphi}$ must coincide with $\hat{\varphi}$ by the injectivity of the embedding $W^{1,2}_{\mathbb{R},\Gamma} \hookrightarrow L^2$. Thus, $\hat{\varphi}$ then also satisfies (3.24). Since the solution of (3.24) is unique, due to Proposition 3.50, this means $\varphi = \hat{\varphi}$, what is a contradiction to $\|\varphi_k - \varphi\|_{W^{1,2}_{\mathbb{R},\Gamma}} > \epsilon$. □

Let us now define the second variant of a fixed point mapping

$$\begin{aligned}\Psi(n,p) = \Big(&\mathcal{N}_n(V_n + V_{xc,n}(n,p) - \mathcal{L}(V_n + V_{xc,n}(n,p), V_p + V_{xc,p}(n,p))),\\ &\mathcal{N}_p(V_p + V_{xc,p}(n,p) + \mathcal{L}(V_n + V_{xc,n}(n,p), V_p + V_{xc,p}(n,p)))\Big).\end{aligned} \tag{3.30}$$

Continuity and compactness follow from the previous considerations, cf. Theorem 3.51 and Theorem 3.19. Thus, Ψ has a fixed point giving a solution of the Kohn-Sham system.

So in view of the numerical considerations we have two possible fixed point mappings at hand. The first, (3.22), is quite natural in that it produces self-consistency of the solution to the Kohn-Sham system directly. The second (3.30) replaces the (simple) solution φ of Poisson's equation by the electrostatic potential of the self-consistent solution to the Kohn-Sham system without exchange-correlation potential. In this way, exchange-correlation effects are included in the electrostatic potential. That means, V_{xc} and φ belong to each other, which is not the case for (3.22). This might be important for the iterative procedure.

4 Cylindrical Quantum Dot

In the following sections we will deal with the numerical investigation of the presented Kohn-Sham system. This chapter introduces the main example to which the numerical results correspond to. Namely, this will be the exciton localisation in a cylindrical quantum dot (QD) embedded in a quantum well (QW). After schematically describing the device setup the chapter deals with the spectral properties of the resulting Hamiltonian. For a better understanding of the spectral properties of the quantum dot example, we will introduce a reference configuration on a square-box domain describing the quantum-well region only, i.e. the quantum box with infinite barriers. This reference system qualitatively shows the same behaviour.

4.1 Device Configuration

As a main example we use a cylindrical quantum dot within a quantum well. In Figure 1 a schematically illustration of the treated structure is shown. It is made of a thin quantum well layer sandwiched between thick bulk layers. Embedded in the quantum well layer is a cylindrical quantum box, representing the quantum dot. We are especially interested in the localisation of excitons inside the dot. On the basis of Kaiser et al. [69] this configuration models a region of phase-segragated Indium within a $(In_{0.2}, Ga_{0.8})N/GaN$ quantum well. The 2D calculations performed in [69] observed stable (bi-)exciton in physically relevant ranges of parameters, i.e. quantum box radius between 0.5 nm and 3 nm and a potential depth up to 1 eV.

For our calculations we choose a width of the quantum well of 2 nm and a radius of the quantum box of 2 nm, as well. The band-edge offsets at the hetero-interface $(In, Ga)N/GaN$ are $\Delta E_c = 0.15$ eV and $\Delta E_v = 0.23$ eV (cf. [74, 94, 69]) for the conduction and valence band, respectively, see Figure 2. The potential depth of the quantum box is set to 0.2 eV.

Since $(In_{0.2}, Ga_{0.8})N$ is a mixture on the basis of 20% InN and 80% GaN, the remaining material parameters (dielectricity, effective masses) are assumed to be constantly those of GaN throughout the whole domain. They are

$$\varepsilon = 9.5\,, \quad m_e = 0.2\,m_0\,, \quad m_h = 0.8\,m_0\,,$$

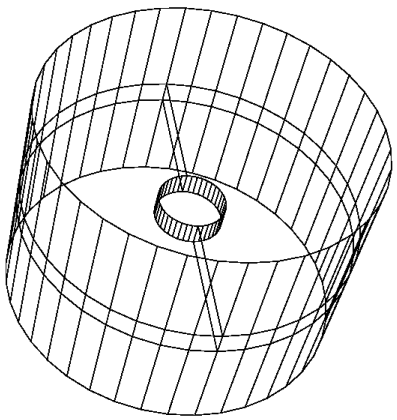

Figure 1: schematical quantum dot structure: cylindrical quantum box embedded in a quantum well.

where m_0 denotes the electron mass. The boundary conditions will be mixed Dirichlet

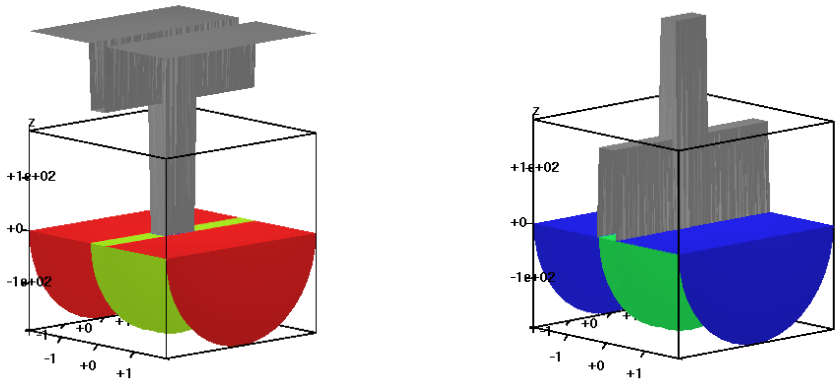

Figure 2: Variation of the band-edge offsets for the quantum dot on cross-section. Left: conduction band offset; right: valence band offset.

and Neumann conditions. Homogeneous Dirichlet on top and bottom and homogeneous Neumann on the sides.

4.2 Quantum Box with Infinite Barrier: a reference system

To understand the spectral properties of the QD Hamiltonian, we will analyse a reference system for which the analytic solutions are known explicitly. The considered problem reads

$$H\psi = E\psi$$

with $H = -\frac{1}{2m^*}\nabla^2 + V$. Where the potential V is fixed.

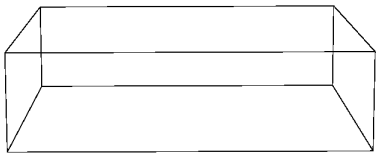

Figure 3: Reference structure: square box, modelling the quantum well region

The regarded domain will be a flat square box as shown in Figure 3, which shall represent the quantum well region of the original device with infinite barriers at the bulk interfaces. Thus, according to the original device the boundary conditions are mixed homogeneous Dirichlet on top and bottom and homogeneous Neumann on the lateral sides. The size of the square box is determined by the side lengths L_x, L_y and L_z, where the base area shall be a square, i.e. $L_x = L_y =: L_\parallel$, with $L_\parallel > L_z$. For our reference calculations we have chosen a ratio $\frac{L_\parallel}{L_z} = 2.5$.

4.2.1 Quantum Well States

First we will have a look on the quantum well states, i.e. $V = const.$
The problem now reads

$$-\frac{1}{2m^*}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)\psi = (E-V)\,\psi = \tilde{E}\psi\,.$$

By separation of variables the wavefunction ψ can be written as a product

$$\psi_{n,m,l}(x,y,z) = \psi_n(x)\psi_m(y)\psi_l(z)\,,$$

where the factors are solutions to the ordinary differential equations

$$\begin{aligned}
-\frac{1}{2m^*}\frac{\partial^2}{\partial x^2}\psi_n &= E_n\psi_n\,, \quad \Omega_x = [0,L_x]\,, \quad \psi_n'(0) = \psi_n'(L_x) = 0 \text{ (Neumann)}\\
-\frac{1}{2m^*}\frac{\partial^2}{\partial y^2}\psi_m &= E_m\psi_m\,, \quad \Omega_y = [0,L_y]\,, \quad \psi_m'(0) = \psi_m'(L_y) = 0 \text{ (Neumann)}\\
-\frac{1}{2m^*}\frac{\partial^2}{\partial z^2}\psi_l &= E_l\psi_l\,, \quad \Omega_z = [0,L_z]\,, \quad \psi_l(0) = \psi_l(L_z) = 0 \text{ (Dirichlet)}\,.
\end{aligned}$$

The eigenvalues and eigenfunctions are thus given by

$$E_n = \frac{1}{2m^*}\left(\frac{n\pi}{L_x}\right)^2\,, \quad E_m = \frac{1}{2m^*}\left(\frac{m\pi}{L_y}\right)^2\,, \quad E_l = \frac{1}{2m^*}\left(\frac{l\pi}{L_z}\right)^2$$

and

$$\psi_n(x) = \cos\left(\frac{n\pi}{L_x}\,x\right)\,, \quad \psi_m(y) = \cos\left(\frac{m\pi}{L_y}\,y\right)\,, \quad \psi_l(z) = \sin\left(\frac{l\pi}{L_z}\,z\right)$$

for integer values $n \geq 0$, $m \geq 0$ and $l \geq 1$, which are the quantum numbers. Every state will be represented by these quantum numbers in the form (n,m,l). We thus get

$$\psi_{n,m,l}(x,y,z) = \cos\left(\frac{n\pi}{L_x}\,x\right)\cos\left(\frac{m\pi}{L_y}\,y\right)\sin\left(\frac{l\pi}{L_z}\,z\right) \tag{4.1}$$

and the energy is then given by

$$\tilde{E}_{n,m,l} = E_n + E_m + E_l = \frac{1}{2m^*}\left(\left(\frac{n\pi}{L_x}\right)^2 + \left(\frac{m\pi}{L_y}\right)^2 + \left(\frac{l\pi}{L_z}\right)^2\right) = \frac{\pi^2}{2m^*}\left(\frac{n^2+m^2}{L_\parallel^2} + \frac{l^2}{L_z^2}\right)\,. \tag{4.2}$$

Due to the square base area we expect a degeneration for states (n_i,m_i,l_i) and (n_j,m_j,l_j) with $n_i^2 + m_i^2 = n_j^2 + m_j^2$. The energies

$$\frac{\pi^2}{2m^*L_\parallel^2} := \Delta E_\parallel \quad \text{and} \quad \frac{\pi^2}{2m^*L_z^2} := \Delta E_z$$

are called quantisation energies. They primarily determine the energy shifts between the states, due to the size quantisation.

Figure 4 shows a comparison of the first 50 eigenvalues computed by use of the explicit formula (4.2) and by solution of the corresponding 3D eigenvalue problem. As a solution method we used the Finite-Volume based solver introduced in [59], cf. Section 5.1. We clearly can see the predicted multiplicity of states. Since the first eigenvalue equals the quantisation energy ΔE_z, we can calculate the (energetic) position of the state $(0, 0, 2)$, i.e. $4\Delta E_z$, which is indicated by the horizontal line in Figure 4. Below this level all states belong to the quantum number $l = 1$. States with $l > 1$ can only occur above this line.

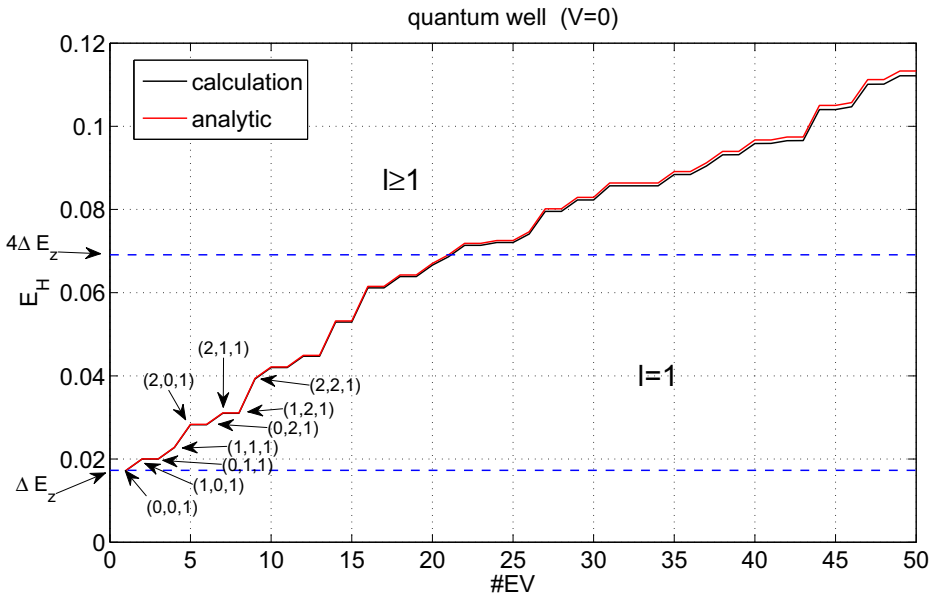

Figure 4: Quantum well eigenvalue spectrum: Calculation by use of (4.2) and numerically solution of the eigenvalue problem, cf. Section 5.1.

The included quantum numbers correspond to those eigenstates shown in Figure 5, where the probability densities of the corresponding first nine eigenfunctions are shown, i.e. $|\psi|^2$. Below the pictures the corresponding quantum numbers are written. Note that the unexpected shapes of the states $\{(2, 0, 1), (0, 2, 1)\}$ and $\{(2, 1, 1), (1, 2, 1)\}$ are due to the multiplicity of the states. In this situation, the solver for the Schrödinger problem calculates a basis of the degenerated subspace. However, this basis need not be the canonical one represented by $\psi_{n,m,l}$, cf. (4.1). In our case the found basis $\{\tilde{\psi}_{2,0,1}, \tilde{\psi}_{0,2,1}\}$ arises from $\{\psi_{2,0,1}, \psi_{0,2,1}\}$ by rotation, i.e.

$$R = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

and

$$\begin{pmatrix} \tilde{\psi}_{2,0,1} \\ \tilde{\psi}_{0,2,1} \end{pmatrix} = R \begin{pmatrix} \psi_{2,0,1} \\ \psi_{0,2,1} \end{pmatrix},$$

with $\alpha = 45°$. The same applies for the eigenfunctions $\tilde{\psi}_{2,1,1}$ and $\tilde{\psi}_{1,2,1}$.

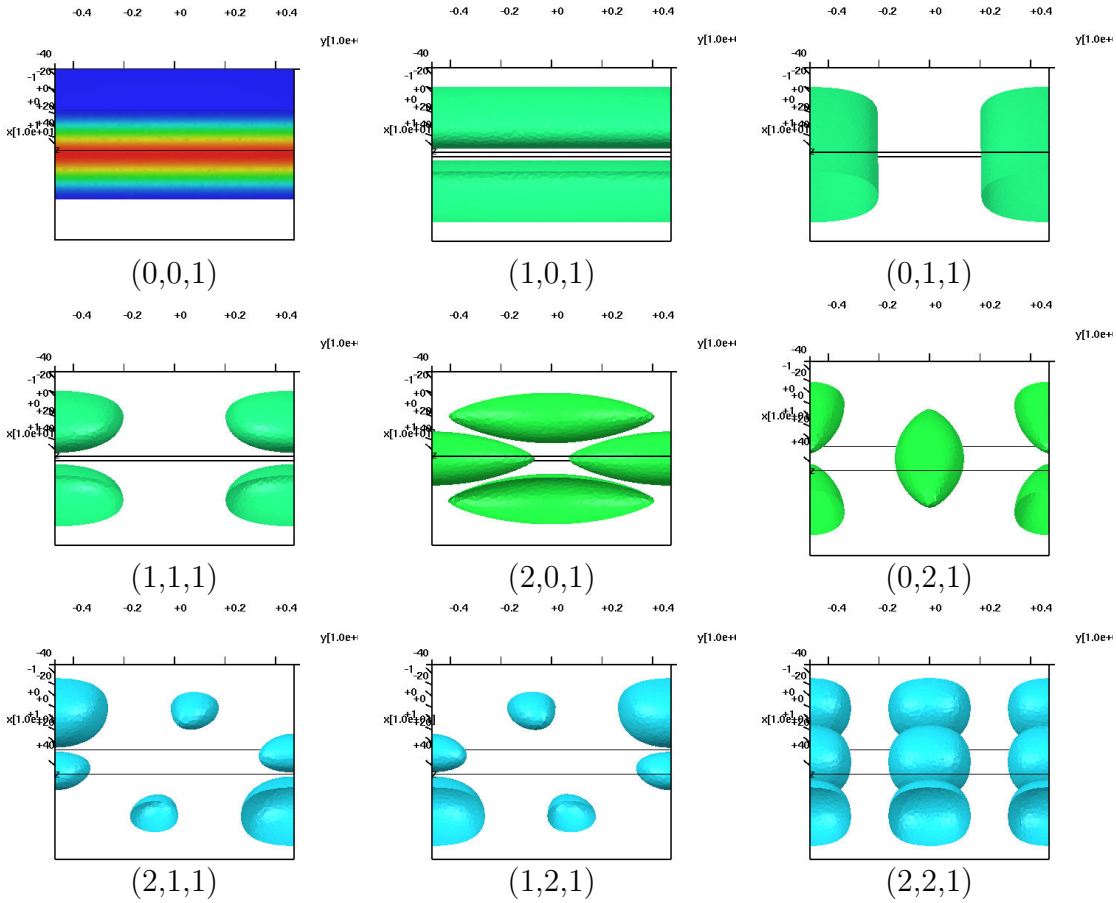

Figure 5: Quantum well states: probability density of eigenfunctions 1 to 9, cf. Figure 4. Corresponding quantum numbers given as triplet (n, m, l).

4.2.2 Harmonic Potential Valley

The second configuration for which we know the solutions explicitly is the two dimensional harmonic oscillator (HO), i.e.

$$V(x, y, z) = \frac{1}{2} m^* \omega^2 (x^2 + y^2) \,, \tag{4.3}$$

inside the quantum well. Figure 6 shows the potential (4.3) applied to the square box problem. It is constant along the z-direction and parabolic in the (x, y)-plane.

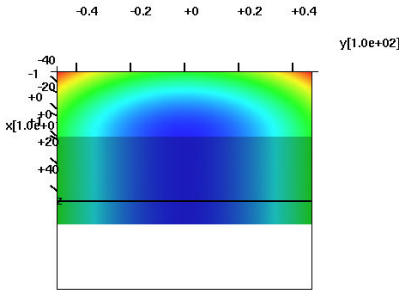

Figure 6: Harmonic potential in x-y-plane for reference square box system.

The wavefunction again separates into

$$\psi_{n,m,l}(x, y, z) = \psi_{n,m}(x, y)\psi_l(z) \,.$$

The z-component of the eigenstate is still given by

$$E_l = \frac{1}{2m^*} \left(\frac{l\pi}{L_z} \right)^2 \,, \quad \psi_l(z) = \sin\left(\frac{l\pi}{L_z} \, z \right) \,.$$

For the (x, y)-component we get the eigenvalue contribution

$$E_{n,m} = \omega(n + m + 1) \,, \tag{4.4}$$

and the corresponding eigenfunction is given by, cf. [14, Ch. 5],

$$\psi_{n,m}(x, y) = \left(\frac{m^*\omega}{\pi} \right)^{\frac{1}{2}} \frac{1}{2^{n+m} n! m!} e^{-\frac{1}{2} m^* \omega (x^2 + y^2)} H_n(\sqrt{m^*\omega} x) H_m(\sqrt{m^*\omega} y) \,,$$

where $H_i(s)$ denotes the i-th *Hermite polynomial*,

$$H_i(s) = (-1)^i e^{s^2} \frac{d^i}{ds^i} e^{-s^2} \,.$$

The real value ω is called the *angular* or *orbital frequency* and together with the quantisation energy in z-direction it determines the lowest eigenvalue of the corresponding 2D operator, i.e. $n = m = 0$. It is chosen, such that the eigenvalue range is comparable to that of the quantum well with constant potential V. The numerical value for the presented calculation is $\omega = 0.02485$

Figure 7 and Figure 8 show the eigenvalues and the corresponding probability densities of the eigenfunctions, respectively, from the 2D harmonic potential valley in the square box. Again we expect multiplicities due to the symmetric dependence of $E_{n,m}$ on the quantum numbers n and m. Degenerated states occur for all states sharing the same summation result $n + m = const.$ The horizontal lines in Figure 7 indicate the first and second quantisation in z-direction. Regarding Figure 8 we have the same effect as already seen in Figure 5. The states $\{(1, 0, 1), (0, 1, 1)\}$ and $\{(2, 0, 1), (1, 1, 1), (0, 2, 1)\}$ form bases of the corresponding degenerated subspaces. Note that the shown eigenvalues 8 and 9 are only one half of the basis set needed to span the subspace for $n + m = 3$.

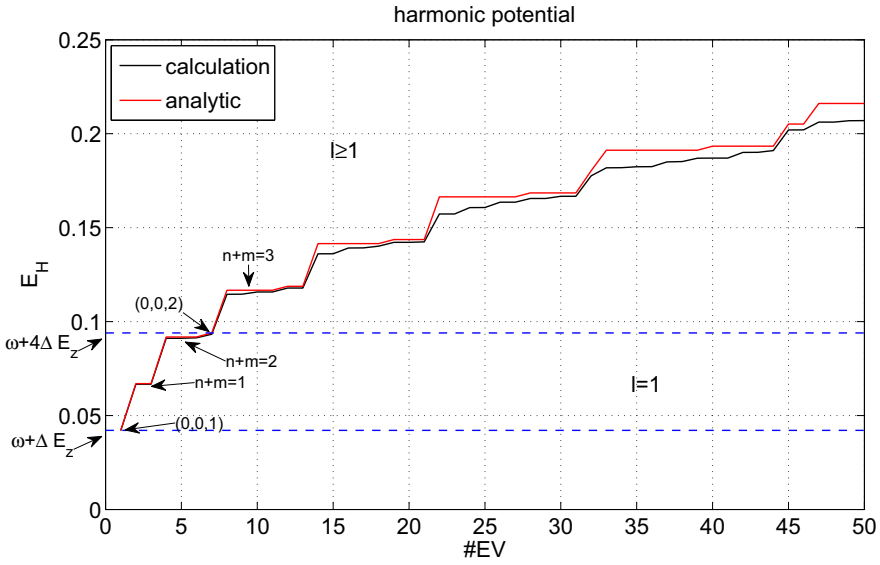

Figure 7: quantum well with harmonic potential ($\omega = 0.02485$): Eigenvalue spectrum calculated using (4.4) and numerically solution of the eigenvalue problem, cf. Section 5.1.

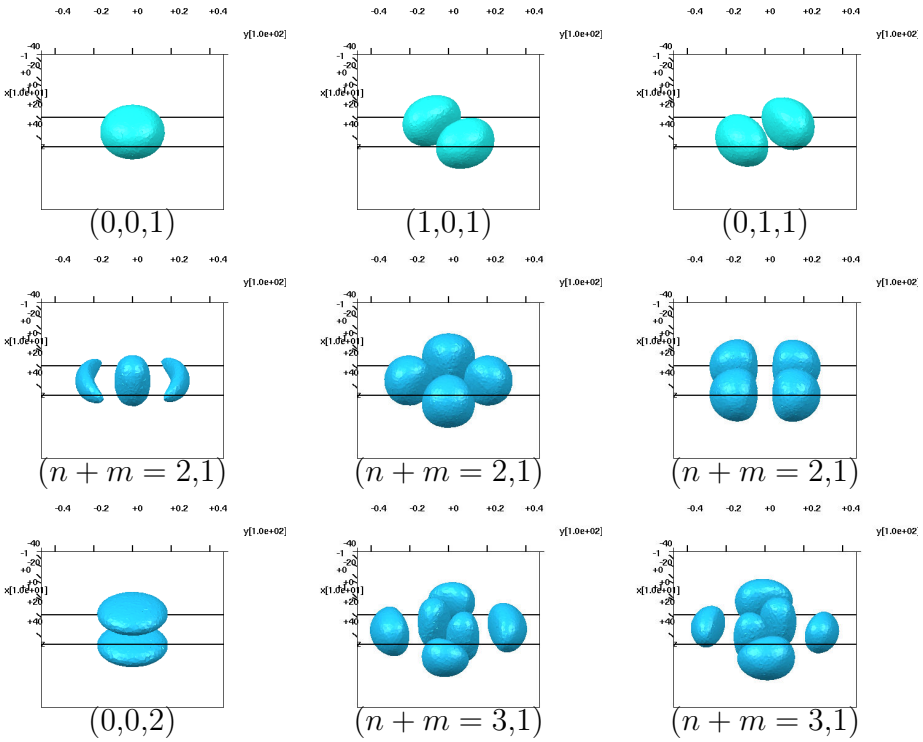

Figure 8: States in quantum well with harmonic potential: probability density of eigenfunctions 1 to 9, cf. Figure 7. Corresponding quantum numbers given as triplet (n, m, l).

4.2.3 Harmonic Potential Cutoff

Using the foregoing results we can introduce a simple quantum dot model, cf. [67], consisting of a harmonic potential that is cut at a certain value and continued constantly, see Figure 9. There the potentials V for electrons and holes are shown. The depth of the harmonic valleys as well as the *orbital frequency* $\omega(= 0.02485)$ are the same for both, electrons and holes. The different radii of the dots result from the different particle masses $m_h = 4m_e$.

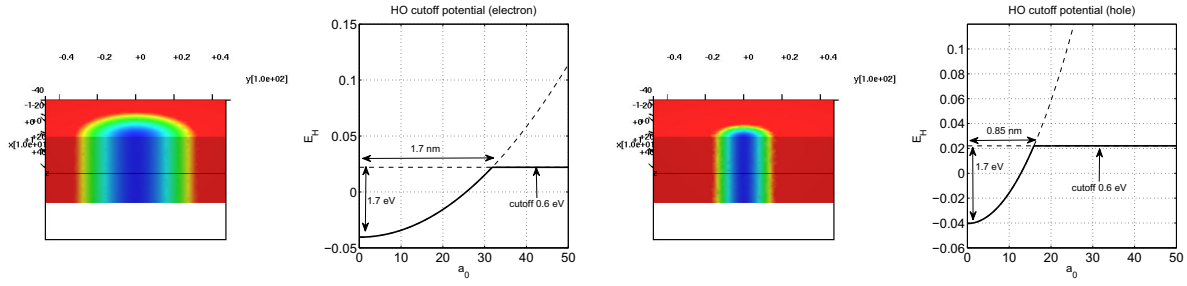

Figure 9: Cutoff of harmonic potential in quantum well. Left: potential for electron calculation (3D view and cross-section); right: potential for hole calculation (3D view and cross-section).

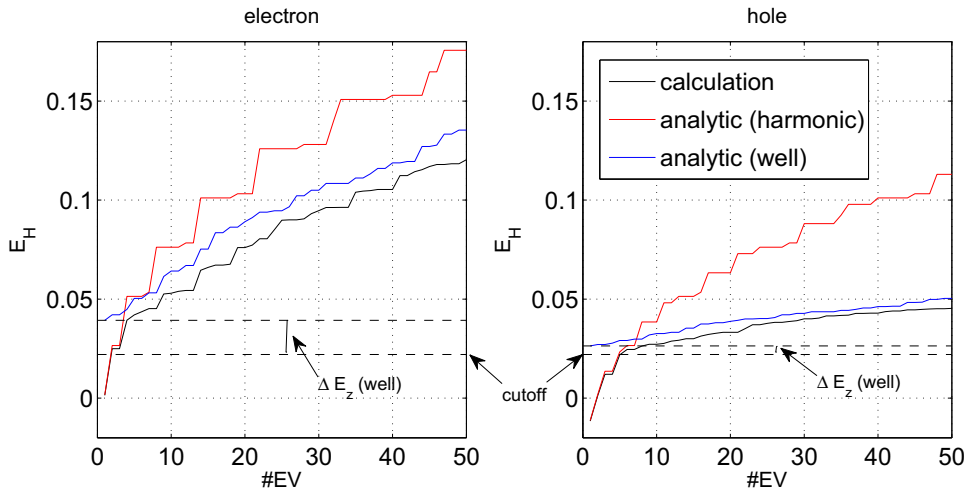

Figure 10: Eigenvalue spectrum for electrons (left) and holes (right) for quantum well with harmonic cutoff potential. Comparison of numerically calculated eigenvalues to shifted eigenvalue spectra of pure quantum well and quantum well with harmonic potential.

The eigenvalue evolutions of this configuration will be a mixture of the (properly shifted) quantum well states and the HO-states for the different particles. Thus, below the cutoff

we will find pure HO-eigenstates and above a mixture with the quantum-well states will occur. Hence, it is sometimes not possible to identify a state with the corresponding quantum numbers (n, m, l) of a QW or HO state. In Figure 11 and Figure 12 we therefore omit the quantum numbers.

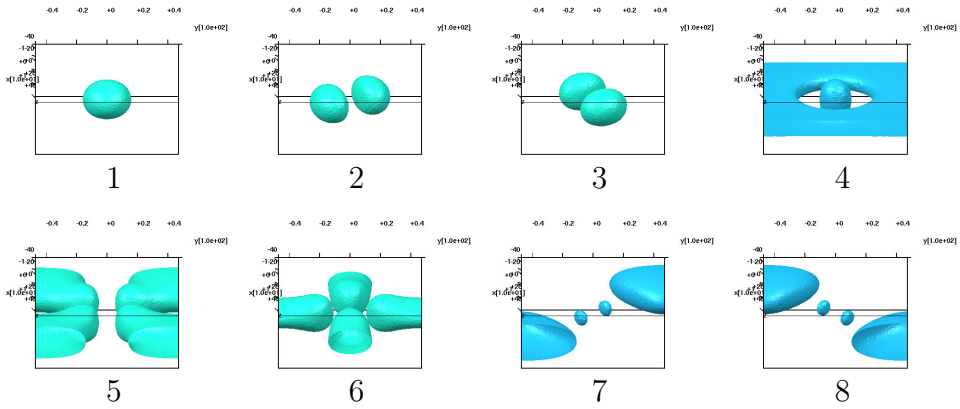

Figure 11: Probability density for electron eigenfunctions 1 to 8 in quantum well with harmonic cutoff potential.

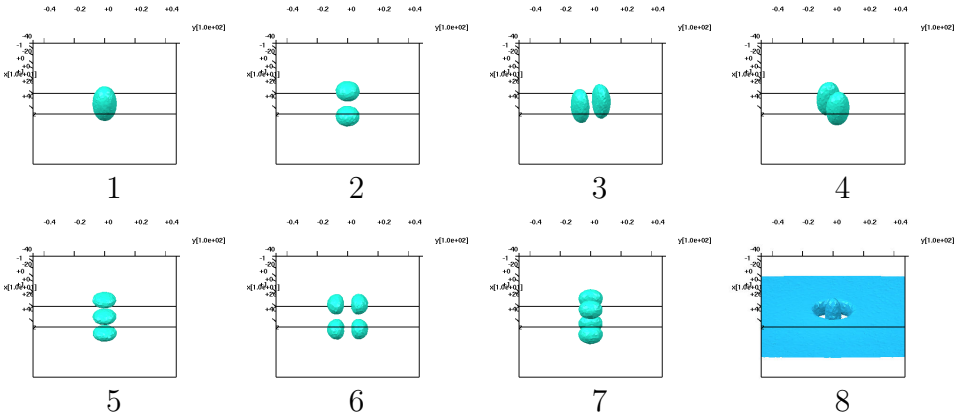

Figure 12: Probability density for hole eigenfunctions 1 to 8 in quantum well with harmonic cutoff potential.

In Figure 10 the eigenvalue spectrum for electrons and holes are shown. In both cases, we see that the spectrum starts with HO-states and continues with QW-states above the cutoff energy. Or more precisely, above the quantum well z-quantisation energy ΔE_z. However, the states above the cutoff are not pure QW-states but rather mixed HO-QW-states. This can be clearly observed in Figure 11 and Figure 12. For electrons, eigenstate four allready is a mixed state of quantum dot and well states; for holes the mixture of states starts with eigenstate eight. Even though energetically located in the quantum well, the found states do respect the potential valley at the dot, in that there is an accumulation in the dot region.

4.2.4 Multi-Particle States

Since we are mainly interested in exciton calculations, we will show results on the localisation of up to three excitons in the reference square-box system.

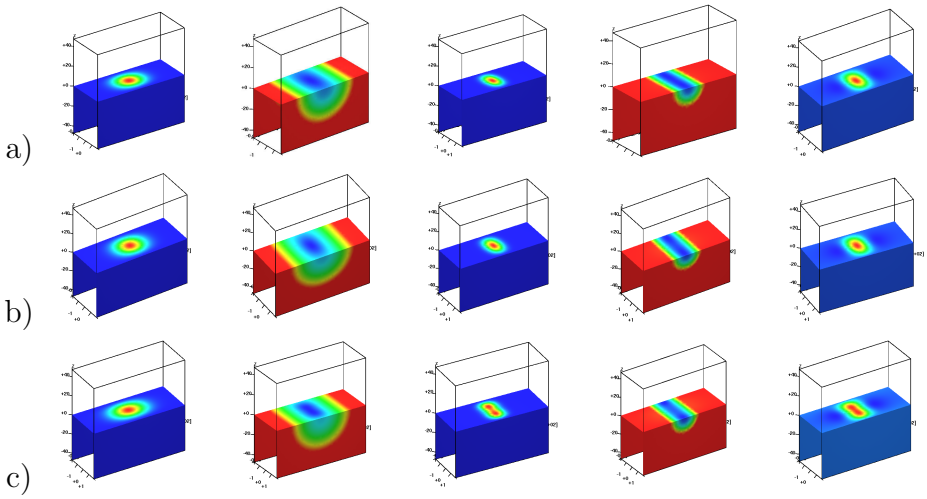

Figure 13: Exciton calculations for the quantum well reference system with harmonic cutoff potential; a) exciton (X), b) bi-exciton (XX), c) tri-exciton (XXX). From left to right: solution (electron), eff. potential (electron), solution (hole), eff. potential (hole), electrostatic potential.

#EV	1	2	3	4	5	6	7	8
ref.	0.0015	0.0250	0.0250	0.0394	0.0420	0.0436	0.0452	0.0452
X	**-0.0026**	0.0228	0.0228	0.0383	0.0414	0.0426	0.0447	0.0447
XX	**-0.0043**	0.0222	0.0222	0.0380	0.0413	0.0425	0.0446	0.0446
XXX	**-0.0057**	**0.0217**	**0.0217**	0.0375	0.0411	0.0423	0.0443	0.0443
#EV	1	2	3	4	5	6	7	8
ref.	-0.0116	0.0011	0.0121	0.0121	0.0218	0.0246	0.0246	0.0263
X	**-0.0141**	-0.0013	0.0099	0.0099	0.0195	0.0226	0.0226	0.0259
XX	**-0.0133**	-0.0008	0.0100	0.0100	0.0200	0.0226	0.0226	0.0257
XXX	**-0.0117**	**0.0007**	0.0107	0.0107	0.0213	0.0233	0.0233	0.0255

Table 1: Eigenvalues of exciton calculations for the quantum well reference system with harmonic cutoff potential. Electron (above) and hole (below) eigenvalues for exciton, bi-exciton and tri-exciton calculations. Occupied states are indicated in bold font.

In Figure 13 the resulting solutions for exciton, bi-exciton and tri-exciton calculations are shown together with the effective and electrostatic potentials. In Table 1 and Figure 14 we see the corresponding first 8 eigenvalues of the solutions compared to those of the pure reference system. For every exciton we indicated the eigenstates that play a role when

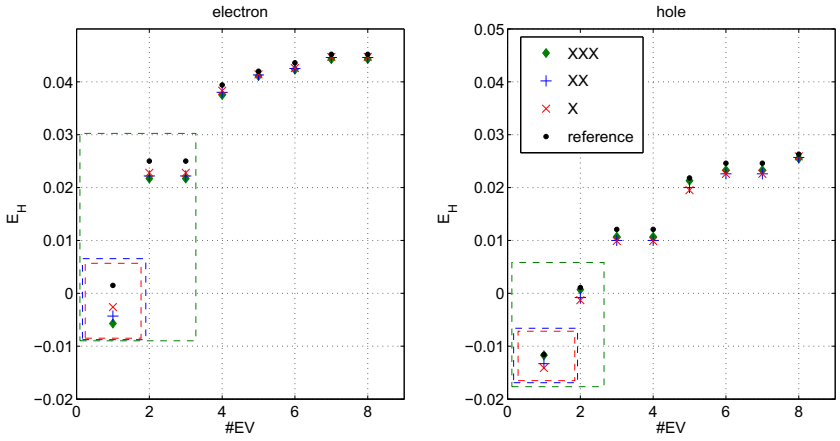

Figure 14: Comparison of exciton eigenvalue evolution for harmonic cutoff potential. Eigenvalues for exciton (red), bi-exciton (blue) and tri-exciton (green). Corresponding dashed areas mark occupied states.

composing the densities n_e and n_h, by writing them in bold font. For the single- and the bi-exciton only the first eigenstate is essential. Starting with the tri-exciton higher states get important as well.

Due to the fact that the effective hole mass is four times greater than the effective electron mass, the holes are much more localised than the electrons. This causes a peak of positive charge in the dot surrounded by a wider area of negative charge, which can be observed in the electrostatic potential, cf. Figure 13. Thus, the effective potentials are deformed such that the potential valleys get deeper with a further lowering for the electrons in the dot region. Concerning the holes, an additional peak inside the valley is generated. This effect is even stronger for a larger number of particles, cf. Figure 13, and causes the lower eigenstates to increase with the number of excitons.

The values in Table 1 and Figure 14 illustrate that the negative charges cause an overall decrease in the eigenvalues, due to the widening of the potential valleys. The exciton eigenstates are thus energetically lower than those of the reference system. However, when increasing the number of excitons, the energies of the electron states decrease because of the positive charged peak, while the hole states increase, due to the same effect. But still this configuration easily allows a localisation of three excitons inside the dot. The dashed regions in Figure 14 indicate the occupied states involved in the composition of the densities of the excitons, cf. bold font in Table 1.

Beside pure exciton states we can as well consider ionised excitons. In Figure 15 the solutions for XX- and XX+ calculations are shown. Depending on the type of ionisation, the solutions show a stronger localisation for the electrons or holes. Furthermore, the electrostatic potential reacts sensitively on the occurring net charge distribution.

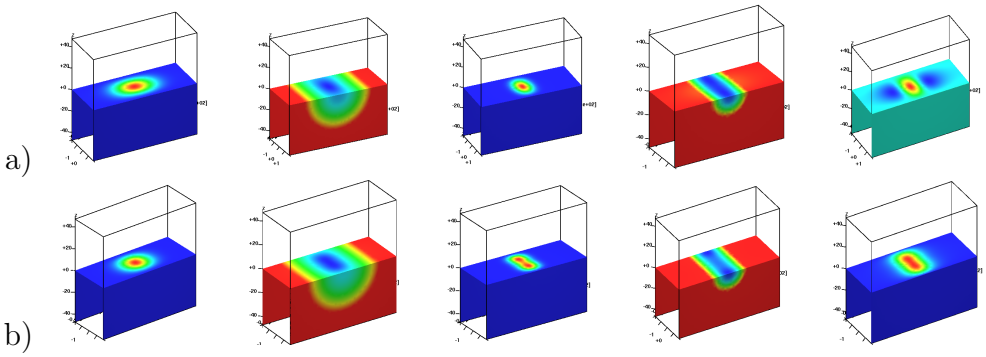

Figure 15: Ionised exciton calculations for the quantum well reference system with harmonic cutoff potential. a) XX-, b) XX+. From left to right: solution (electron), eff. potential (electron), solution (hole), eff. potential (hole), electrostatic potential.

4.3 3D Exciton Localisation in Cylindrical Quantum Dot

Let us now come to the previously described cylindrical quantum dot structure, cf. Section 4.1 and Figure 1. The device radius is set to $10\,nm$ an the thickness of the well is $2\,nm$. The radius of the embedded dot is set to $2\,nm$.

4.3.1 Single-Particle States

As for the reference system, we can now calculate single-particle states for the cylindrical quantum dot structure. This is what we are essentially interested in. In Figure 16 we see the eigenvalue spectrum for electrons and holes.

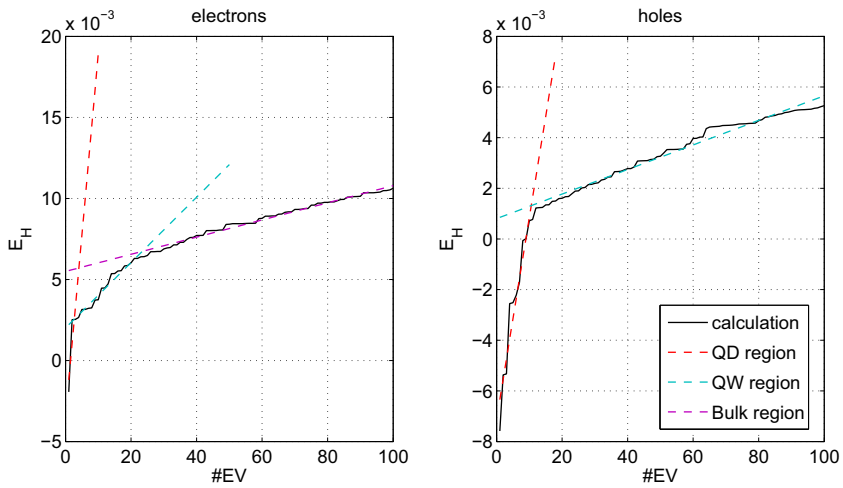

Figure 16: Eigenvalue spectrum of cylindrical quantum dot structure. Dashed lines mark spectrum essentially located in quantum dot (red), quantum well (cyan) and bulk material (violet).

The regions the eigenstates live in are indicated by the dashed lines; quantum dot, quantum well or bulk material. For the electrons we can clearly observe the transitions in the spectrum from quantum dot to quantum well (around eigenvalue #4) and from quantum well to the bulk (around eigenvalue #25). However, since the quantisation energy of the holes is much smaller, due to the bigger mass, the spectrum is much denser. And hence, the calculated first 100 eigenvalues all belong to quantum dot or quantum well states. Nevertheless, the main observation is that there is a sufficient number states inside the quantum dot to allow for a localisation of excitons.

4.3.2 Multi-Particle States

The results for single, bi- and tri-excitons are shown in Figure 17. As we can see, all the excitons localise in the dot as expected and again we observe for the holes a stronger localisation, due to the higher mass. As was observed for the reference system, the effective potential for the holes comprises a peak inside the dot region. This will prevent a decreasing of the (hole) eigenstates, when increasing the number of excitons.

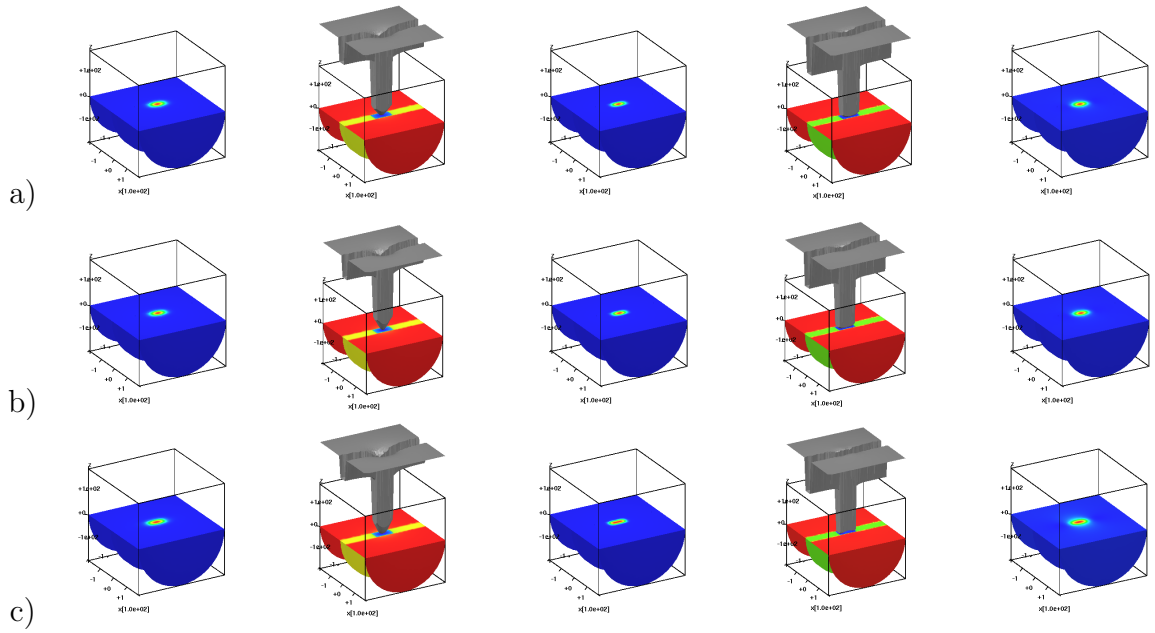

Figure 17: Exciton calculations for the cylindrical quantum dot structure; a)exciton (X), b)bi-exciton (XX), c) tri-exciton (XXX). From left to right: solution (electron), eff. potential (electron), solution (hole), eff. potential (hole), electrostatic potential.

The eigenvalues in Table 2 show this behaviour even more clearly. The electron eigenvalues belonging to the exciton states decrease with the number of excitons, whereas the hole states do not, due to the positive charged peek. Additionally we can identify degenerated states by their values. Again we indicate the occupied states that are involved in the composition of the densities by bold font.

#EV	1	2	3	4	5	6	7	8
ref.	-0.0019	0.0025	0.0025	0.0027	0.0031	0.0031	0.0032	0.0032
X	**-0.0037**	0.0018	0.0018	0.0026	0.0029	0.0029	0.0031	0.0031
XX	**-0.0044**	0.0015	0.0015	0.0026	0.0029	0.0029	0.0031	0.0031
XXX	**-0.0052**	**0.0009**	**0.0009**	0.0025	0.0027	0.0027	0.0030	0.0030
#EV	1	2	3	4	5	6	7	8
ref.	-0.0076	-0.0053	-0.0053	-0.0026	-0.0026	-0.0022	-0.0017	-0.0001
X	**-0.0089**	-0.0066	-0.0066	-0.0036	-0.0036	-0.0034	-0.0029	-0.0011
XX	**-0.0090**	-0.0067	-0.0067	-0.0038	-0.0038	-0.0035	-0.0029	-0.0013
XXX	**-0.0086**	**-0.0065**	**-0.0065**	-0.0036	-0.0036	-0.0032	-0.0026	-0.0010

Table 2: Eigenvalues of exciton calculations for the cylindrical quantum dot structure. Electron (above) and hole (below) eigenvalues for exciton, bi-exciton and tri-exciton calculations. Occupied states are indicated in bold font.

The results just presented in Figure 17 and Table 2 are the exciton states we are interested in, calculated by solving the Kohn-Sham system self-consistently. In what follows, we want to deal with the question of how to calculate these results fast and efficiently. We will deal with this topic in the next section, which is about the numerical treatment of the Kohn-Sham system.

5 Numerical Treatment

In this section the presented example of a quantum dot within a quantum well from Section 4 is dealt with from the numerical point of view. Particularly, this means numerically solving the Kohn-Sham system in three space dimensions. Our goal is to establish an efficient algorithm for iteratively finding the self-consistent solution to the Kohn-Sham system.

Since the most time-consuming part of the calculation is the solution of the Schrödinger eigenvalue problem, we evaluate efficiency by means of the total number of eigenvalue problems solved during the iterative process. As already described in Section 3 the used representation leads to a fixed point formulation based on the particle density. But this fixed point mapping generally is not a contraction and thus a straight forward iteration, such as *Picard* (or *Banach*) does not work. Simple damping strategies like *linear mixing*, cf. [68], are usually used to get a convergent scheme. However, such schemes mostly suffer from slow convergence rates making them too costly. Hence, acceleration procedures have to be used, e.g. *Newton*-like schemes, which are known to converge quadratically. Unfortunately, although successful when regarding the number of iteration steps only, these acceleration methods are very expensive or even impossible to apply, due to the necessity of computing the (approximated) Jacobian in every step.

In quantum chemistry iterative procedures for *Hartree-Fock* or *Coupled Cluster* [40, 60, 97] calculations are often accelerated using the *direct inversion in the iterative subspace* (DIIS) scheme, invented by Pulay in 1980, cf. [80]. This scheme mixes a larger number of previous iterates to create a better guess. Due to the extrapolation ability, i.e. negative coefficients, of the original DIIS scheme, it is not safely applicable to our Kohn-Sham fixed point iteration that is based on the, necessarily, positive particle density. In order to apply this acceleration method we add further constraints when calculating the coefficients. We thus ensure positivity of the composed density.

In this way the idea behind DIIS is carried over in a secure way to our density-based iteration process. The invented scheme will be called *convex* DIIS (CDIIS) method and it represents a high-dimensional generalisation of the simple *linear mixing* scheme, [68]. The performance of CDIIS is tested on the exciton calculation presented in Section 4. It turns out to be faster than *linear mixing* and more efficient (i.e. less total number of function evaluations) than the *Newton*-like scheme, since it only needs a fixed number of function evaluations per step.

The section is organised in the following way. Technical and environmental settings are described in the first part. Then, in the second part, we describe the iterative procedure for setting up the fixed point iteration schemes and simple self-consistent iterations are presented. Damping strategies for these iterations are dealt with in the third part, including adaptivity. The fourth part is denoted to acceleration schemes for the iterative procedure.

5.1 Environmental Settings

Let us first fix the system under consideration. Assume N_e and N_h to be given positive integers describing the total number of electrons and holes, respectively, in the device domain. The Kohn-Sham system for the electron- and hole-density, n_e and n_h, and the electrostatic potential φ then reads

$$\begin{aligned} \text{Schrödinger (electron):} \quad & \left[\tfrac{1}{2}\nabla\cdot(m_e)^{-1}\nabla + V_e(n_e,n_h)\right]\psi_{e,i} = \mathcal{E}_{e,i}\psi_{e,i} \\ \text{Schrödinger (hole):} \quad & \left[\tfrac{1}{2}\nabla\cdot(m_h)^{-1}\nabla + V_h(n_e,n_h)\right]\psi_{h,i} = \mathcal{E}_{h,i}\psi_{h,i} \\ \text{Poisson:} \quad & \tfrac{1}{4\pi}\nabla\cdot\varepsilon\nabla\varphi = n_h - n_e \end{aligned}$$

with

$$\begin{aligned} V_e(n_e,n_h) &= V_{0,e} + V_{xc,e}(n_e,n_h) - \varphi \\ V_h(n_e,n_h) &= -V_{0,h} + V_{xc,h}(n_e,n_h) + \varphi \\ n_e(x) &= 2\sum_{k=0}^{\infty} f(\mathcal{E}_{e,k} - \mathcal{E}_{e,F})|\psi_{e,k}(x)|^2 \\ n_h(x) &= 2\sum_{k=0}^{\infty} f(\mathcal{E}_{h,k} - \mathcal{E}_{h,F})|\psi_{h,k}(x)|^2 \end{aligned}$$

$$2\textstyle\sum_{k=0}^{\infty} f(\mathcal{E}_{e,k} - \mathcal{E}_{e,F}) = N_e\,, \quad 2\sum_{k=0}^{\infty} f(\mathcal{E}_{h,k} - \mathcal{E}_{h,F}) = N_h$$
$$f(s) = \tfrac{1}{1+\exp(\frac{s}{k_B T})}\,.$$

The device domain will be the cylindrical quantum dot structure described in Section 4. For the Schrödinger operators we assume homogeneous Dirichlet boundary conditions on top and bottom of the cylinder and homogeneous Neumann conditions on the sides. Concerning Poisson's equation we regard homogeneous Dirichlet conditions on the whole boundary. The external potentials $V_{0,e}$ and $V_{0,h}$ are the band-edge offsets originated in the effective mass approximation as specified in Section 4. The remaining material parameters are chosen in accordance to the previously described example as well. For the exchange-correlation $V_{xc,e}$ and $V_{xc,h}$ we will use the local density approximation (LDA) coming from the homogeneous electron gas, see Appendix A for a precise description,

$$\begin{aligned} V_{xc,e}(n_e,n_h) &= \frac{1}{\varepsilon}\left(\frac{3}{\pi}n_e\right)^{1/3} \\ V_{xc,h}(n_e,n_h) &= \frac{1}{\varepsilon}\left(\frac{3}{\pi}n_h\right)^{1/3}\,. \end{aligned}$$

As working temperatures for the device calculation we choose either 4K, 77K or 300K. The last one means operating at room temperature, whereas the first two are chosen in accordance to physical cooling processes, which often are used to analyse material properties.

Note that the $4K$-case is the most singular among the three and thus we will use this case when testing the algorithms.

Because of the decreasing property of Fermi's function f, summation over eigenstates are assumed to be effectively finite. The number of eigenstates taken into account depend on the Temperature T and the number of particles N_e and N_h of the problem, usually about six to nine.

As a convergence indicator of the fixed point procedures we will use the following relative error term of the particle density $n_i = (n_{e,i}, n_{h,i})$ at iteration i

$$r_i = \frac{\|\mathcal{N}(n_i) - n_i\|_\infty}{\|n_i\|_\infty},$$

where we speak of convergence, when the relative error r_i is smaller than or equal to 10^{-8}.

Convergence rates will be measured on the basis of the reduction factors $\varrho_{m+1,m} = \frac{r_{m+1}}{r_m}$. More precisely, we use the geometric mean

$$\varrho_{m+k,m} = (\varrho_{m+k,m+k-1} \cdot \ldots \cdot \varrho_{m+1,m})^{\frac{1}{k}} = \left(\frac{r_{m+k}}{r_m}\right)^{\frac{1}{k}},$$

which (in linear problems) is known to be a good approximation of the spectral radius of the underlying scheme for large k, cf. [37].

The implementation was done in the framework of WIAS-$p\partial e$lib2, which is a collection of software components for creating simulators based on solving partial differential equations. This toolbox was developed and implemented at the Weierstrass Institute for Applied Analysis and Stochastics (WIAS, Berlin), [25].

Single Particle States

For the single particle states we used the Finite Volume ([22]) based solver included in the $p\partial e$lib2-toolbox by J. Fuhrmann and T. Koprucki, cf. [59]. The authors showed the effectiveness for one-, two- and three-dimensional problems, to which exact analytic solutions are known. The solver is capable of treating jumping coefficients in the effective masses and the potentials correctly and thus is an adequate tool for treating our quantum dot example.

Grid Generation

For creating adequate grids on the computational domain we used the mesh generator TetGen, written by H. Si [92]. It produces quality tetrahedral meshes for any three dimensional polyhedral domain. Furthermore, the grids created this way are Delaunay triangulations ensuring high quality tetrahedrons that are suitable for solving partial differential equations. Moreover, the produced grid is unstructured and can easily be adapted and refined locally, according to the requested accuracy of solutions to the partial differential equations that have to be solved. The notably higher complexity in organising the grid structure is

supported by the WIAS-$p\partial e$lib2 toolbox.

Linear Solver

As a solver for linear systems of equations, e.g. Poisson's equation, we will use the sparse direct solver package PARDISO written by Schenk, Gärtner et al. [88, 85, 89, 86, 87]. PARDISO supports a wide range of problems and is shown to be highly efficient, [34]. For symmetric problems, which we mainly deal with, it essentially performs a Cholesky factorisation $PAP^T = LL^T$, where P denotes a symmetric fill-in reducing permutation. Additionally, different types of pivoting strategies are used to increase performance and accuracy.

Eigenvalue Solver

The Schrödinger eigenvalue problems will be solved using the numerical software library ARPACK (ARnoldi PACKage) for large sparse problems, written by Lehoucq, Maschhoff, Sorensen and Yang, [63, 64, 65]. It is based on an algorithmic variant of the Arnoldi process called the Implicitly Restarted Arnoldi Method (IRAM). In case of a symmetric matrix the algorithm reduces to the Lanczos variant, called the Implicitly Restarted Lanczos Method (IRLM). In essence, these methods combine the Arnoldi/Lanczos process with the Implicitly Shifted QR technique, which is appropriate especially for structured sparse matrices. We will use ARPACK in the shift-invert mode with zero shift such that the requested eigenvalues are the smallest ones.

5.2 Self-Consistent Iteration (Picard Iteration)

The system just presented will be solved using a fixed point representation. To this end we introduce the function $\mathcal{N}$, mapping a given density $n = (n_e, n_h)$ to another density $\mathcal{N}(n)$ such that a fixed point of $(\tilde{n}_e, \tilde{n}_h) = \mathcal{N}$ provides a solution (n, φ) to the Kohn-Sham system. φ denotes the solution to Poisson's equation. When evaluating $\mathcal{N}(n)$ the following set of operations is done.

Algorithm 5.1 (Fixed Point Mapping $\mathcal{N}$)**.**

- *Given:* n
- *get* φ *by solving Poisson's equation with right-hand side* $n_h - n_e$
- *compute exchange-correlation potentials* $V_{xc,e}$ *and* $V_{xc,h}$ *using* n
- *compute effective potentials* V_e *and* V_h
- *solve Schrödinger's problem for the different species and get* $\{\mathcal{E}_{e,k}\}$, $\{\psi_{e,k}\}$, $\{\mathcal{E}_{h,k}\}$ *and* $\{\psi_{h,k}\}$
- *compute Fermi levels* $\mathcal{E}_{e,F}$ *and* $\mathcal{E}_{h,F}$ *according to* N_e *and* N_h*, respectively*
- *compute new densities* $\tilde{n}_e$ *and* $\tilde{n}_h$

Remark 5.2. *Note that the mapping* $\mathcal{N}$ *defined in this way allways has a fixed point, since they coincide with solutions to the Kohn-Sham system. And according to Section 3 the Kohn-Sham system allways has a solution.*

With this procedure we can set up a simple Picard Iteration, meaning to repeatedly apply $\mathcal{N}$.

Algorithm 5.3 (Picard Iteration)**.**

- **Given:** density n_0
- $i := 0$
- **while** 'not converged' **do**

  $n_{i+1} = \mathcal{N}(n_i)$

$i \leftarrow i + 1$

- **end**

In comparison to that, we can introduce another algorithm according to the second fixed point mapping described in Section 3 on analytical considerations. There, the exchange-correlation potentials are built using the electrostatic potential φ coming from the self-consistent solution of the corresponding Kohn-Sham system without exchange and correlation effects. We denote by $\widetilde{\mathcal{N}}$ the fixed point mapping similar to Algorithm 5.1, but assuming the exchange-correlation potentials $V_{xc,e}$, $V_{xc,h}$ to be given. We thus get

Algorithm 5.4 (alternative Iteration)**.**

- **given:** density n_0
- $i := 0$
- **while** 'not converged' **do**

  $V_{xc,e} = V_{xc,e}(n_i)$**;** $V_{xc,h} = V_{xc,h}(n_i)$

  $\tilde{n}_0 := n_i$**;** $j := 0$

  while 'not converged' **do**

    $\tilde{n}_{j+1} = \widetilde{\mathcal{N}}(\tilde{n}_j)$

    $j \leftarrow j + 1$

  end

  $n_{i+1} = \mathcal{N}(\tilde{n}_j)$

  $i \leftarrow i + 1$

- **end**

Before going further, let us take a closer look on the differences of Algorithms 5.3 and 5.4. First note that the essential difference between both algorithms is the number of iterations the procedure takes before adapting the xc-potentials. Algorithm 5.4 adapts only when the self-consistent solution according to the actual potentials is reached. Whereas Algorithm 5.3 adapts immediately. In this way Algorithm 5.4 makes sure, that when exiting the inner loop, the xc-potentials and the electrostatic potential fit together. This might have a positive effect on the overall iteration in the outer loop.

Let us point out that the Kohn-Sham system without exchange-correlation potential can be reformulated to a non-linear Poisson equation, cf. Corollary 3.45. As shown by Kaiser and Rehberg [47, 49, 50] the involved operator is monotone and continuous, and thus the problem yields a unique solution for every given external potential. Therefore, the inner loop of Algorithm 5.4 yields a unique solution $\tilde{n}_j$.

As shown in Section 3 the particle density operator $\mathcal{N}$ is Lipschitz continuous but not necessarily a contraction. Hence, in order to get a converging fixed point iteration we need to introduce a damping factor $\alpha \in (0, 1]$. Changing the corresponding lines in Algorithm 5.3 and 5.4 to

$$n_{i+1} = \mathcal{N}(n_i) \;\rightsquigarrow\; n_{i+1} = (1-\alpha)n_i + \alpha\mathcal{N}(n_i)$$
$$n_{i+1} = \mathcal{N}(\tilde{n}_j) \;\rightsquigarrow\; n_{i+1} = (1-\alpha)\tilde{n}_j + \alpha\mathcal{N}(\tilde{n}_j)\,,$$

yields the damped versions of these algorithms, cf. Algorithm 5.5.

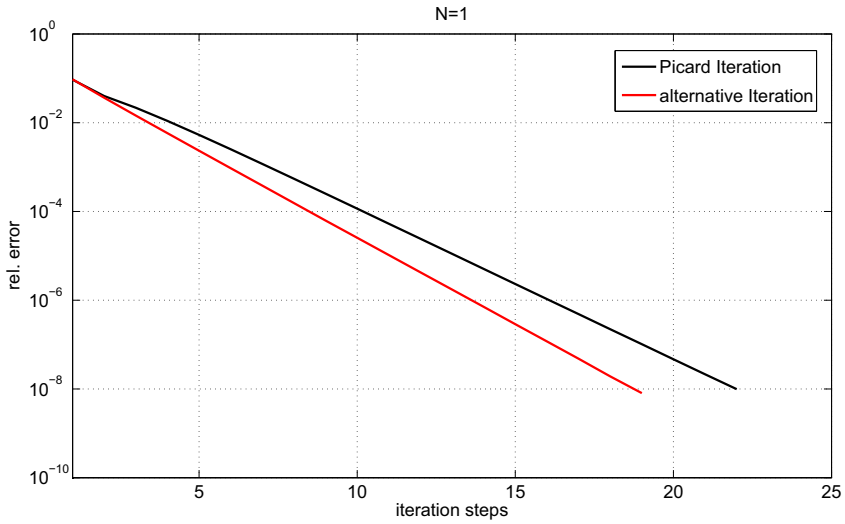

Figure 18: Comparing residual evolution for the damped Picard iteration (Algorithms 5.3) and the damped alternative iteration loop (Algorithm 5.4). The damping factor is set to $\alpha = 0.6$.

Figure 18 shows a comparison of the residual evolution for Algorithm 5.3 and 5.4 (damped versions) of single-exciton calculations for the quantum dot structure at a temperature of $4K$.

As an initial value n_0 for starting the iteration we chose the densities that result from solving the Schrödinger problems with effective potentials consisting only of the band-edge offsets, $V_e = V_{0,e}$ and $V_h = V_{0,h}$.

As can be seen the advanced variant of the simple (damped) Picard iteration given by Algorithm 5.4 shows a better performance than the original (damped) Picard iteration from Algorithm 5.3. Thus, the adjustment of the xc-potentials and the electrostatic potential carried out by the inner loop of Algorithm 5.4 indeed is beneficial when regarding the error in the outer loop only. Table 3 shows the corresponding convergence rates $\varrho_{k,0}$, which of course are better for the alternative iteration scheme. However, the price for

	Algorithm 5.3	Algorithm 5.4
$N = 1$	0.4655	0.4048

Table 3: Convergence rates $\varrho_{k,0}$ for the damped Picard iteration (Algorithm 5.3) and the damped alternative iteration loop (Algorithm 5.4) from Figure 18.

this performance increase is way too big. This can be understood by regarding Table 4, showing the number of solved eigenvalue problems of the different approaches.

	Algorithm 5.3	Algorithm 5.4
$N = 1$	46	924

Table 4: Number of solved eigenvalue problems for the damped Picard iteration (Algorithm 5.3) and the damped alternative iteration loop (Algorithm 5.4) from Figure 18.

In fact Algorithm 5.4 solves about 20 times more eigenvalue problems than Algorithm 5.3. Since the main computational cost is the solution of the eigenvalue problem, the time needed to solve the problem is 20 times as big as well. Thus, the improvement in performance by adjusting the (xc- and electrostatic) potentials is not worth the effort. In what follows we will solely use the fixed point mapping indicated by Algorithm 5.3, meaning that it will be the basis of all upcoming strategies.

5.2.1 Fixed Damping (Krasnoselskij Iteration)

Let us now have a closer look on the fixed damping strategy. In a first description damping strategies can be considered as strategies for increasing the radius of convergence for the corresponding Picard-iteration. However, this would imply that damping strategies are only feasible for contractive (or non-expansive) mappings. This is by far not the case. In fact it ensures convergence for a much wider class of fixed point operators, such as (generalised) pseudo contractive an φ-contractive ones, cf. [1]. Additionally the requirements on the underlying spaces are weaker compared to Picard-operators. Thus, it is worth using damped iterations as a first step towards fast and robust schemes. Let us recall that the non-linear Poisson operator that results when omitting the xc-potentials is monotonous. Hence, there is a close connection to fixed point iteration schemes, cf. [28], that gives another justification for applying damped schemes on our problem.

Let us now catch up for the algorithmic variant indicated in the previous part.

Algorithm 5.5 (Damped Iteration)**.**

- **Given:** density n_0**;** damping factor $\alpha \in (0, 1]$
- $i := 0$
- **while** 'not converged' **do**

  $n_{i+1} = (1 - \alpha)n_i + \alpha\mathcal{N}(n_i)$

  $i \leftarrow i + 1$
- **end**

Apart from theory about fixed point iteration schemes this method is known in DFT as *linear mixing*, cf. [68], meaning to generate a new approximated input-density n_{i+1}^{in} by linearly mixing the input and output density from step i, n_i^{in} and n_i^{out} respectively. The latter is just $\mathcal{N}(n_i^{in})$ in our notation. Rewriting the scheme yields

$$n_{i+1} = (1 - \alpha)n_i + \alpha\mathcal{N}(n_i) = n_i + \alpha(\mathcal{N}(n_i) - n_i)\,. \tag{5.1}$$

In absence of further information the direction $\mathcal{N}(n_i) - n_i$ is the best *steepest descent* direction available. To get an upper bound on the steplength α we take a look on the following linearised problem for the error function $R(n) = \mathcal{N}(n) - n$,

$$\begin{aligned} 0 = R(n_{i+1}) &= R(n_i) + J_R[n_i](n_{i+1} - n_i) \\ \Leftrightarrow n_{i+1} = n_i - J_R^{-1}[n_i]R(n_i) &= n_{i+1} = n_i - J_R^{-1}[n_i](\mathcal{N}(n_i) - n_i)\,. \end{aligned}$$

Thus, the damped iteration (5.1) corresponds to a relaxed Richardson iteration with the inverse Jacobian $J_R^{-1}[n_i]$ approximated by the identity. For such an iteration it can be shown (cf. [37, Ch. 4]) that the scheme converges if and only if the steplength α does not exceed $2/\lambda_{\max}(J_R[n_i])$, where $\lambda_{\max}(J_R[n_i])$ denotes the maximal eigenvalue of $J_R[n_i]$. A thorough analysis concerning linear mixing procedures can be found in [17]. Thus, one cannot usually expect the simple Picard iteration ($\alpha = 1$) to converge.

	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$	$\alpha = 0.7$	$\alpha = 0.8$	$\alpha = 0.9$
$N = 1$	0.8226	0.7337	0.6445	0.5550	0.4655	0.3762	0.2880	0.3814
$N = 2$	0.8217	0.7327	0.6440	0.5560	0.4669	0.7302	*div.*	*div.*
$N = 3$	0.8027	0.7049	0.6094	0.5141	0.5824	0.8904	*div.*	*div.*

Table 5: Convergence rates $\varrho_{k,0}$ for the damped Picard iteration with various damping factors α from Figure 19.

Figure 19 shows calculations for Algorithm 5.5 with different damping factors. We can see that for each problem there is an optimal damping value α for which the iterative process is the fastest among all converging processes. This is easily motivated by the consideration,

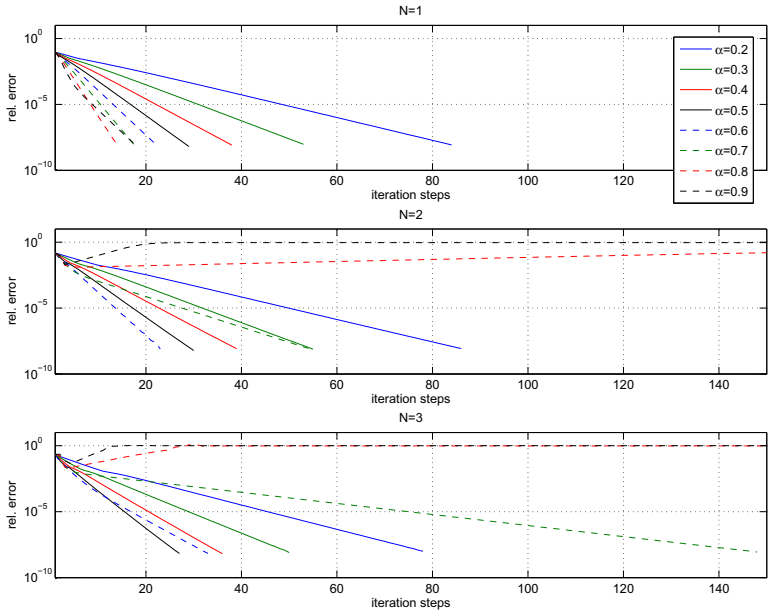

Figure 19: Comparing residual evolution for the damped Picard iteration (Algorithm 5.5) with various damping factors α.

that a somewhat too small steplength α may produce a converging iteration but takes too many iterations before approaching the limit. On the other side, a big steplength may converge faster but risks to leave the radius of convergence. Hence, the main task in these kind of approaches is the correct choice of the steplength α. The corresponding convergence rates to Figure 19 are summarised in Table 5.

Furthermore, the problem of finding an optimal α can be written as an 1D optimisation problem similar to

$$\alpha_{opt} = \underset{\alpha}{\operatorname{argmin}}\{\|\mathcal{N}(n_{i+1}(\alpha)) - n_{i+1}(\alpha)\| \ : \ n_{i+1}(\alpha) = (1-\alpha)n_i + \alpha\mathcal{N}(n_i)\,,\ \alpha \in (0,1]\}\,. \quad (5.2)$$

Typically the analytic solution to this problem is unknown and solving it approximately with high accuracy may be very expensive. Thus, one usually is satisfied with *a priori* estimates giving upper bounds on α, cf. [17]. Note that it is usually not possible to find a single optimal damping factor for a whole class of problems. This is due to its strong dependence on different input values like the geometry, the initial value or other input data such as source terms. Another alternative is the adaption of the steplength in each iteration step which will be the topic of the next part.

5.2.2 Adaptive Damping (Kerkhoven Stabilisation)

As seen previously it may be disadvantageous to fix the steplength at the beginning of the calculation and keeping it throughout the iteration process. Depending on the operator and the current iterate it might be beneficial to take an adjusted steplength α_n. For example at the end of the iterative process it is likely to be in the contractive area around the fixed point of the functional, which makes it worth trying bigger steplengths. Conversely, at the beginning of the calculation it is rather unlikely (depending on the quality of the initial value) to be steadily successful in each iteration step when using big steplengths. As mentioned before solving the problem (5.2) for an optimal α_n might be too expensive. Therefore we will use an adaption strategy which means using information gained in previous steps to estimate a reliable damping factor. This (particular) strategy was introduced by Kerkhoven et al. ([54, 55, 52, 53]) for the simulation of quantum wires.

The algorithm presented below is a variant of the *stabilisation by adaptive underrelaxation* approach used by Kerkhoven et al.

Algorithm 5.6 (Kerkhoven I (stabilised) Iteration)**.**

- **Given:** density n_0
- **set:** $n_{-1} := 0$; $n_{-2} := n_0$
- $\alpha := 1$; $i := 0$
- **while** 'not converged' **do**
  if $\frac{\|\mathcal{N}(n_i)-n_i\|}{\|\mathcal{N}(n_{i-1})-n_{i-1}\|} > \frac{\|\mathcal{N}(n_{i-1})-n_{i-1}\|}{\|\mathcal{N}(n_{i-2})-n_{i-2}\|}$
  $\alpha \leftarrow \alpha * 0.8$; $\alpha' := \min\{\alpha, \frac{\|\mathcal{N}(n_{i-1})-n_{i-1}\|}{\|\mathcal{N}(n_i)-n_i\|}\}$
  else
  $\alpha \leftarrow \max\{\alpha/0.8, 1\}$; $\alpha' := \alpha$
  end
  $n_{i+1} = (1-\alpha')n_i + \alpha'\mathcal{N}(n_i)$; $i \leftarrow i+1$
- **end**

The adaption criterion uses the improvement ratio τ_i

$$\tau_i := \frac{\|\mathcal{N}(n_i) - n_i\|}{\|\mathcal{N}(n_{i-1}) - n_{i-1}\|}\,. \tag{5.3}$$

It gives an idea of the actual performance of the iterative process. If $\tau_i < 1$, then the actual iterate n_i seems to be a better approximation to the true solution n^* than n_{i-1} was. Nevertheless, the criterion $\tau_i > \tau_{i-1}$ is mild, since it does not demand a reduction of the error norm $\|\mathcal{N}(n_i) - n_i\|$ itself. Instead, it keeps track on the overall performance of the

process in that it reacts when the improvement ratio gets worse. Note, that this allows for a growing residual as long as the magnitude of increase does not grow itself. Thus, the criterion gives the process the opportunity to get out of a local (error) valley where it would get stuck in a gradient approach.

Only in the first step the improvement-criterion reduces to an error consideration, since no further information are available. The criterion then reads

$$\|\mathcal{N}(n_0) - n_0\|^2 > \|\mathcal{N}(0)\|^2 .$$

Meaning to lower the stepsize when the zero vector is a better approximation than the initial value n_0.
Further note that the improvement-criterion adapts the iteration by reducing the damping factor when there is an indication that the bottom of a valley is reached. This is due to the assumption that the iteration progress is fast as long as the solution is sufficiently far away. Contrariwise, when approaching the solution cautiousness is advised.

As soon as the improvement-criterion is violated the algorithm chooses an even more restricted steplength by

$$\alpha' = \min\{\alpha, \tau_i^{-1}\} .$$

This is reasonable, since the foregoing approach of giving the procedure the chance of leaving a possible valley, led to a situation where the iteration procedure seems to diverge. Thus, when an unexpected large change in the error of the actual approximation occurs, an even stronger damping is applied.

In Figure 20 wee see the results of a numerical test demonstrating the performance of this approach compared to the (best corresponding) simple damping strategy, cf. Figure 19. As we can see the performance of the adaption strategy has several stages. At the beginning of the calculation it performs equally good or better compared to the damping with the best choice of α (taken according to Table 5). During the calculation it might get much worse. The reason for this is that the algorithm adapts α_n down near the minimal steplength of 0.05. When approaching the solution at the end (or a valley in between), the steplength may recover and the improvement in every step gets big again. In Table 6 the corresponding convergence rates are presented.

	damped Iteration	Kerkhoven (stabilised)
$N = 1$	0.2880	0.9324
$N = 2$	0.4669	0.9312
$N = 3$	0.5141	0.5899

Table 6: Convergence rates $\varrho_{k,0}$ for the damped iteration (Algorithm 5.5) compared to the Kerkhoven (stabilised) scheme (Algorithm 5.6) from Figure 20.

Thus, Algorithm 5.6 shows the typical behaviour of an adaption strategy applied to non-linear problems. At the beginning the algorithm is optimistic and adapts the steplength

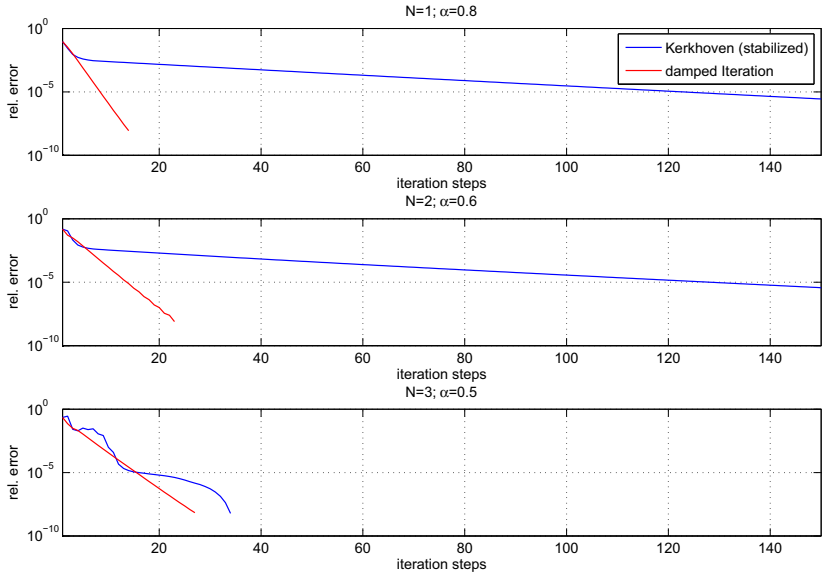

Figure 20: comparing residual evolution for the damped Picard iteration (Algorithm 5.5) and the stabilised Kerkhoven iterations (Algorithm 5.6). Damping factors chosen in accordance to the best result from Figure 19.

according to the complicated (error) landscape. With this, the algorithm tries to safely stay in the region of convergence. Unfortunately, this (somewhat too) careful bearing leads to a large and time-wasting number of steps which cannot be compensated by a possibly good performance at the end. Thus, we will deal in the following section with possibilities of accelerating the iteration procedure so that it passes the time consuming part without getting stuck with too many small steps.

5.3 Quasi-Newton Scheme

With the methods described in the foregoing sections we mainly focused on safely finding the solution. Now we want to deal with the question of how to find this solution fast and efficiently. In the following we look at two different acceleration approaches. The first is a Newton-like method which is known to show a good convergence behaviour, assumed that the initial value is close enough to the solution. This means reformulating the fixed point problem to a root finding problem and solving a linear system consisting (essentially) of the Jacobian J of $\mathcal{N}$. Since J is typically a dense system, explicit inversion is not recommended. In the following we describe and use a nonlinear version of the GMRES method (cf. [84]) introduced by Kerkhoven et al. [56, 54]. This scheme provides us with the possibility to apply a Newton-like scheme without ever generating the Jacobian.

As the iterative procedure makes progress, we necessarily get close to the solution where reliable acceleration techniques can be applied. We will apply the *nonlinear* GMRES (NLGMR) method described in [56, 54]. To do this we have to reformulate the fixed point problem into a (nonlinear) root finding problem

$$n - \mathcal{N}(n) = 0\,.$$

Newton's method requires at each step i the solution of the linear system

$$(I - J[n_i])\,\delta_i = -\,(n_i - \mathcal{N}(n_i)) \tag{5.4}$$

to get the next iterate

$$n_{i+1} = n_i + \delta_i\,.$$

The matrix $I - J(n_i)$ is typically dense and thus a direct inversion is not advised. Instead, a nonlinear version of the GMRES method (NLGMR) is used that does not need to generate the Jacobian, cf. [56, 54].

Let us first recall the idea of the GMRES algorithm. Solving (5.4) is equivalent to the Euclidean minimisation problem

$$\|\,(I - \mathcal{N})\,n_i + (I - J[n_i])\,\delta_i\|_2 \tag{5.5}$$

for finding a new approximation n_{i+1}. The vector $\delta_i^{(m)}$ is represented in the form

$$\delta_i^{(m)} = \sum_{j=1}^{m} a_j v_j\,, \tag{5.6}$$

where the set $\{v_j\,:\,j = 1, \ldots, m\}$ forms an orthonormal basis of the Krylov subspace

$$K^m = \mathrm{span}\{v_1 := n_i, (I - J[n_i])\,v_1, \ldots, (I - J[n_i])^{m-1}\,v_1\}\,.$$

Having the operation $x \mapsto (I - J[n_i])\,v$ at hand these vectors can easily be calculated by an Arnoldi process. The coefficients a_i are to be determined for composing the solution (5.6).

Minimisation of (5.5) is then done by applying the GMRES algorithm on equation (5.4). Note, that solving this equation exact would give the Newton direction

$$-\,(I - J[n_i])^{-1}\,(n_i - \mathcal{N}(n_i))\;.$$

Thus, this procedure is an inexact Newton method. As mentioned before the only operation needed in the Arnoldi process is given by the directional derivative $(I - J[n_i])\,v$, meaning a matrix-vector multiplication. Fortunately we do not need $J[n_i]$ explicitly, since we can approximate $J[n_i]v$ by a commonly used forward difference quotient

$$J[n_i]v \approx \frac{\mathcal{N}(n_i + hv) - \mathcal{N}(n_i)}{h}\,. \tag{5.7}$$

The only factor left to be adjusted is the dimension m of the Krylov subspace K^m. Since (5.4) is a linear model of a nonlinear operator it is appropriate to ask, whether it is useful to solve it with a high accuracy. If the nonlinearity is strong it might be sufficient to keep the dimension m small. Let n_i be an approximate solution to $n - \mathcal{N}(n) = 0$ and let $n_i^{(m)}$ be the improved approximation after m steps of GMRES. To measure the nonlinearity we now look at the nonlinear residual $\mathrm{res}_{\mathrm{nl}}$ and the linear residual $\mathrm{res}_{\mathrm{lin}}$ coming from the GMRES approximation

$$\begin{aligned} \mathrm{res}_{\mathrm{nl}} &= n_i^{(m)} - \mathcal{N}(n_i^{(m)}) \\ \mathrm{res}_{\mathrm{lin}} &= n_i - \mathcal{N}(n_i) + (I - J[n_i])\,(n_i^{(m)} - n_i)\,. \end{aligned}$$

If the nonlinearity is only mild the Euclidean norms of these expressions should be approximately equal. If not, the linearisation is not an adequate local approximation to the original problem and thus it is probably wasteful to solve (5.4) very accurately. Hence, m should be kept small in this case. Conversely, in case of good agreement, m should be increased. In this way the dimension of the Newton-subspace is adaptively changed in every step. To ensure an error reduction in every step a simple *linesearch* is performed in direction of δ_i, for minimising

$$\|(n_i + \lambda\delta_i^{(m)}) - \mathcal{N}(n_i + \lambda\delta_i^{(m)})\|^2\,.$$

Starting with a full Newton step the stepsize λ is halved until either $\lambda < 0.1$ or the error at n_{i+1} is smaller than the error at n_i.

Remark 5.7. *Note that with this acceleration procedure a single step might get very expensive. Firstly, the procedure has to evaluate $\mathcal{N}$ m-times in order to create the basis of the Krylov subspace K^m and with this the approximation of the Jacobian at the actual iterate n_i. And worse, calculated points from previous steps cannot be used and thus all m evaluations really have to be done. Secondly, performing* linesearch *results in further evaluations of $\mathcal{N}$. In comparison to that, the stabilisation procedure by adaptive damping takes only a single function evaluation per step. The hope of course is that the quadratic convergence behaviour of Newton's method makes only few steps of NLGMR necessary, such that the overall number of evaluations stays small. But this cannot be guaranteed in the first place and the possibly good performance might be very costly.*

The complete algorithm for Kerkhoven's method with stabilisation and acceleration now reads as follows.

Algorithm 5.8 (Kerkhoven II Iteration)**.**

- **Given:** density n_0
- **set:** $n_{-1} := 0$; $n_{-2} := n_0$; $\alpha := 1$; $i := 0$
- **do** (stabilisation)

if $\frac{\|\mathcal{N}(n_i)-n_i\|}{\|\mathcal{N}(n_{i-1})-n_{i-1}\|} > \frac{\|\mathcal{N}(n_{i-1})-n_{i-1}\|}{\|\mathcal{N}(n_{i-2})-n_{i-2}\|}$

$\alpha \leftarrow \alpha * 0.8$**;** $\alpha' := \min\{\alpha, \frac{\|\mathcal{N}(n_{i-1})-n_{i-1}\|}{\|\mathcal{N}(n_i)-n_i\|}\}$

end

$n_{i+1} = (1-\alpha')n_i + \alpha'\mathcal{N}(n_i)$**;** $i \leftarrow i+1$

- **until** 'convergence' **or** 'α decreases 5 times in a row' **or** 'α remains constant for 10 iterations'
- $m := 2$
- **while** 'not converged' **do** (acceleration)

  $n_i^{(0)} := n_i$

  do m steps of NLGMR **to get** $\delta_i^{(m)}$

  if $\frac{2}{3} \leq \frac{\|\mathbf{res_{nl}}\|}{\|\mathbf{res_{lin}}\|} \leq \frac{3}{2}$ **then** $m \leftarrow \min\{25, m*2\}$

  else

  if $\frac{3}{2} \leq \frac{\|\mathbf{res_{nl}}\|}{\|\mathbf{res_{lin}}\|} \leq 5$ **then** $m \leftarrow m$

  else $m \leftarrow \max\{2, m/2\}$

  do 'linesearch' in direction $\delta_i^{(m)}$ **for** $\lambda \in [0.01, 1]$ $\|(n_i + \lambda\delta_i^{(m)}) - \mathcal{N}(n_i + \lambda\delta_i^{(m)})\|^2$

  $n_{i+1} = (n_i + \lambda\delta_i^{(m)})$**;** $i \leftarrow i+1$

- **end**

Figure 21 shows a comparison of the residual evolution between the stabilised Kerkhoven, the stabilised-accelerated (full) Kerkhoven and the fixed damping scheme. The full Kerkhoven scheme shows a similar behaviour as the damped iteration at the beginning of the calculation. It is easy to see by the kink where the scheme changes from the stabilised iteration to the Newton-like acceleration, cf. Figure 21. Thus, with this strategy we outperform the damping approach. Unfortunately, we still need (when accelerating) a possibly (and a priori unpredictable) large number of function evaluations, i.e. solution of eigenvalue problems. This is due to the adapted dimension of the Krylov subspace and the linesearch in the Newton step, see Remark 5.7.

In Table 7 the corresponding convergence rates can be found. Of course the values for the full Kerkhoven scheme have to be understood as mean values between the adaptive damping approach at the beginning and the Newton-type approach at the end of the calculation. To get the rates for the Newton acceleration one could consider only the reduction rates $\varrho_{m^*+k,m*}$ starting from the kink which would be much better, but we are interested in the overall performance and thus the whole sequence of iterates is considered, i.e. $\varrho_{k,0}$.

In the final section we deal with a possibility of accelerating the scheme with a similar performance as the Newton-like Kerkhoven-acceleration and simultaneously ensuring a fixed (and low) number of function evaluations per step.

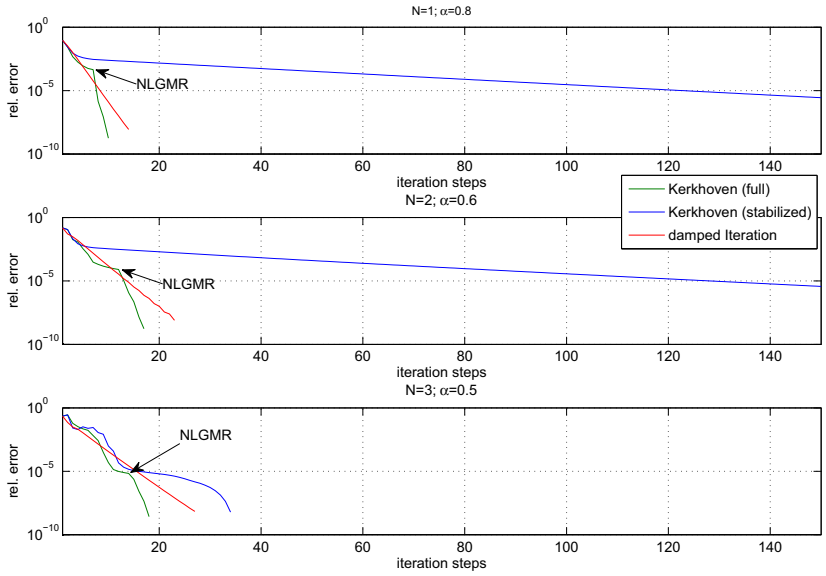

Figure 21: comparing residual evolution for the damped Picard iteration (Algorithm 5.5), the stabilised Kerkhoven scheme (Algorithm 5.6) and the full Kerkhoven scheme (Algorithm 5.8).

	damped Iteration	Kerkhoven (stabilised)	Kerkhoven (full)
$N = 1$	0.2880	0.9324	0.1387
$N = 2$	0.4669	0.9312	0.3187
$N = 3$	0.5141	0.5899	0.3417

Table 7: Convergence rates $\varrho_{k,0}$ for the damped Picard iteration (Algorithm 5.5) compared to the stabilised Kerkhoven (Algorithm 5.6) and the full Kerkhoven scheme (Algorithm 5.8) from Figure 21.

5.4 DIIS Acceleration

The second acceleration approach can be seen as a high dimensional generalisation of the simple damping strategies of Section 5.2. Here the subspace of all previous iterates is searched for an optimal (error minimising) vector. In two dimensions it reduces to an optimal damping strategy similar in structure to the *linear mixing* scheme. The approach we are going to describe is a convex variant of the *direct inversion in the iterative subspace* (DIIS) method invented by Pulay [80] in 1980. It is a subspace acceleration method for minimising the value of an error function belonging to some (nonlinear) problem. Since its first description the DIIS scheme enjoys a good reputation in the quantum chemistry community and has proven to be a powerful tool for accelerating Coupled Cluster and self-consistent field (SCF) calculations, [40, 60, 97, 39]. Moreover, it may even turn a non-converging iteration into a converging one. Although it is successful in practice, the mathematical analysis does not seemed to be traced up so far.

After describing the main idea of DIIS we focus on the relation of DIIS to GMRES in linear situations and we will point out the connection to (Broyden-like) secant methods for the non-linear case. Unfortunately the original DIIS method found in literature cannot be used for our problem, due to extrapolation effects that cause the iterates to leave the solution space. Based on a formulation of the DIIS procedure motivated by Weijo et al. in [97] we will introduce a new DIIS variant that includes an important convexity constraint. We will call this method *convex* DIIS or CDIIS, due to the convexity constraints, that ensures the produced densities to stay in the solution space, cf. Section 3. The CDIIS can be seen as a high dimensional generalisation of the *linear mixing* scheme. However, due to its flexibility it outperforms *linear mixing* and even the Newton-like Kerkhoven scheme.

5.4.1 The basic DIIS Algorithm

Let us consider an equation of the form

$$F(x^*) = 0\,,$$

which we want to solve using the main iteration scheme I together with the DIIS acceleration technique.
Assume a sequence $X^i = \{\tilde{x}_0, \ldots, \tilde{x}_i\}$ of iterates to be already computed. The DIIS algorithm now finds an optimised iterate x_n parametrised as

$$x_i = \sum_{l=0}^{i} c_l \tilde{x}_l\,, \tag{5.8}$$

in X^i where the coefficient vector $c = (c_0, \ldots, c_i)$ fulfils the constraint $\sum_{l=0}^{i} c_l = 1$. This constraint makes sure the trivial zero-solution is excluded. The optimisation is now performed with respect to the residual norm $\|F(x_i)\|_2$. Rewriting x_i as

$$x_i = \tilde{x}_i + \sum_{l=0}^{i-1} c_l(\tilde{x}_l - \tilde{x}_i) = \tilde{x}_i + \delta_i \tag{5.9}$$

and using the following Taylor polynomials of degree one

$$\begin{aligned} F(x_i) &\approx F(\tilde{x}_i) + J_F[\tilde{x}_i]\delta_i \\ F(\tilde{x}_l) &\approx F(\tilde{x}_i) + J_F[\tilde{x}_i](\tilde{x}_l - \tilde{x}_i)\,, \end{aligned} \tag{5.10}$$

we get

$$\begin{aligned} F(x_i) &\approx F(\tilde{x}_i) + \sum_{l=0}^{i-1} c_l \left(F(\tilde{x}_l) - F(\tilde{x}_i)\right) \\ &= \sum_{l=0}^{i} c_l F(\tilde{x}_l) \end{aligned}$$

Thus, the coefficients $(c_0, \ldots, c_i)$ are determined by the minimisation problem

$$c = \underset{c=(c_0,\ldots,c_i)}{\operatorname{argmin}} \{\| \sum_{l=0}^{i} c_l F(\tilde{x}_l)\| \,:\, \sum_{l=0}^{i} c_l = 1\}\,. \tag{5.11}$$

Note that for (affine) linear F the replacement of $\|F(x_i)\|$ by $\|\sum_{l=0}^{i} c_l F(\tilde{x}_l)\|$ is exact. Introducing the Lagrange multiplier λ, the solution to (5.11) can be computed by solving the linear system

$$\begin{pmatrix} B & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{5.12}$$

where $B \in \mathbb{R}^{i+1\times i+1}$ with matrix entries given by $b_{kl} = \langle F(\tilde{x}_{k-1}), F(\tilde{x}_{l-1})\rangle$. Finally, the next approximate iterate $\tilde{x}_{i+1}$ is given by applying a single step of the main Iteration I on the optimal value x_i

$$\tilde{x}_{i+1} = I(x_i)\,.$$

All in all we have an iterative procedure illustrated by

$$\tilde{x}_0 =: x_0 \overset{\text{I}}{\to} \tilde{x}_1 \overset{\text{DIIS}}{\to} x_1 \overset{\text{I}}{\to} \tilde{x}_2 \overset{\text{DIIS}}{\to} \ldots \overset{\text{I}}{\to} \tilde{x}_i \overset{\text{DIIS}}{\to} x_i \overset{\text{I}}{\to} \tilde{x}_{i+1} \overset{\text{DIIS}}{\to} x_{i+1} \ldots\,.$$

Thus, the DIIS scheme really is a family of procedures varying in the choice of the main iteration I.

Algorithm 5.9 (DIIS Iteration)**.**

- **Given:** $\tilde{x}_0$**;** I**;** F
- **set:** $X = \{\tilde{x}_0\}$**;** $i := 0$
- **while** 'not converged' **do**

  add $FX_i = F(\tilde{x}_i)$ **to** $FX : \{FX_0, \ldots, FX_{i-1}\}$

  solve $\begin{pmatrix} B & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ **for** $c = (c_0, \ldots, c_i)$

get $x_i = \sum_{l=0}^{i} c_l \tilde{x}_l$
get $\tilde{x}_{i+1} = I(x_i)$
insert $\tilde{x}_{i+1}$ **in** X**;** $i \leftarrow i+1$

- **end**

By default the main iteration I is given by

$$\tilde{x}_{i+1} = I(x_i) = x_i - F(x_i)\,, \tag{5.13}$$

which means it is the simple Picard iteration for solving the fixed point problem $(\mathbb{1}-F)x = x$.

Remark 5.10. *Note, that for $F(x) = Ax - b$ this approach uses just the gradient direction of the corresponding quadratic minimisation problem* $\min \frac{1}{2}x^t Ax - b^T x$.

Of course, regarding the previous sections we can think of many other choices of I, e.g.

$$\begin{aligned} \text{damping:} \quad & I(x) = ((1-\alpha)\mathbb{1} - \alpha F) \\ \text{adaptive damping:} \quad & I(x) = ((1-\alpha_i)\mathbb{1} - \alpha_i F) \end{aligned}$$

or any other procedure returning a new approximation $\tilde{x}_i$.

5.4.2 Equivalence to GMRES (the linear case)

Let us in this section focus on the (affine) linear case with the default iteration (5.13), meaning

$$\begin{aligned} F(x^*) &= Ax^* - b = 0 \\ I(x) &= x - F(x) = x - (Ax - b) \end{aligned}$$

with $A \in \mathbb{R}^{N\times N}$ and $b \in \mathbb{R}^N$. Throughout this section we assume A to be positive definite. Denote with $K^i := K^i(\tilde{x}_0)$ the Krylov subspace

$$K^i = \operatorname{span}\{b - A\tilde{x}_0, A(b - A\tilde{x}_0), \ldots, A^{i-1}(b - A\tilde{x}_0)\}\,. \tag{5.14}$$

Recall, that GMRES finds the residual minimising solution x_i in the affine subspace $\mathcal{K}^i = \tilde{x}_0 \oplus K^i$. Define the DIIS solution space $\mathcal{D}^i := \mathcal{D}^i(\tilde{x}_0, \ldots, \tilde{x}_i)$ according to (5.8) and (5.9) by

$$\begin{aligned} \mathcal{D}^i &= \{y \in \mathbb{R}^N \,:\, y = \sum_{l=0}^{i} c_l \tilde{x}_l\,,\, \sum_{l=0}^{i} c_l = 1\} && (5.15) \\ &= \tilde{x}_i \oplus \operatorname{span}\{\tilde{x}_l - \tilde{x}_i \,:\, l = 0, \ldots, i-1\}\,. && (5.16) \end{aligned}$$

Assumption 5.11. Let $X^i : \{\tilde{x}_0, \ldots, \tilde{x}_i\}$ be a sequence of iterates lying in $\mathcal{K}^i$ and spanning it.

With this we immediately get

$$K^i = \operatorname{span}\{\tilde{x}_l - \tilde{x}_0 \,:\, l = 1, \ldots, i\}\,. \tag{5.17}$$

Before stating the equivalence theorem we need to prove two preparative lemmas. The first gives the connection between the affine Krylov space $\mathcal{K}^i$ and the constraint DIIS space $\mathcal{D}^i$. The second shows that the DIIS scheme indeed minimises the residual norm when solving the DIIS system (5.12).

Lemma 5.12. *Let Assumption 5.11 hold true.*
Then for the spaces $\mathcal{K}^i$ and $\mathcal{D}^i$ defined above, there holds

$$\mathcal{K}^i = \mathcal{D}^i\,.$$

In particular the optimised iterate x_i calculated by the DIIS scheme is in $\mathcal{K}^i$ for all $i \in \mathbb{N}$.

Proof. As can be seen by the definitions of $\mathcal{K}^i$ and $\mathcal{D}^i$ there holds

$$\dim \mathcal{K}^i = \dim \mathcal{D}^i\,.$$

Let $y \in \mathbb{R}^N$ be from $\mathcal{K}^i$. We then get

$$y \;=\; \tilde{x}_0 + \sum_{l=1}^{i} c_l(\tilde{x}_l - \tilde{x}_0) = \sum_{l=0}^{i} c_l \tilde{x}_l$$

where we use the representation indicated by (5.17) and the definition $c_0 := 1 - \sum_{l=1}^{i} c_l$. Hence the coefficients sum up to one, $\sum_{l=0}^{i} c_l = 1$, and we proved the inclusion $\mathcal{K}^i \subset \mathcal{D}^i$. For the other direction take $y \in \mathbb{R}^N$ from $\mathcal{D}^i$ and compute

$$\begin{aligned} y \;&=\; \sum_{l=0}^{i} c_l \tilde{x}_l = (1 - \sum_{l=1}^{i} c_l)\tilde{x}_0 + \sum_{l=1}^{i} c_l \tilde{x}_l \\ &=\; \tilde{x}_0 + \sum_{l=1}^{i} c_l(\tilde{x}_l - \tilde{x}_0)\,, \end{aligned}$$

where we used the summation property $\sum_{l=0}^{i} c_l = 1$. Hence y is given in the representation (5.17) and we have $\mathcal{D}^i \subset \mathcal{K}^i$. This finishes the proof of $\mathcal{D}^i = \mathcal{K}^i$. □

Lemma 5.13. *A vector $y \in \mathcal{D}^i$ minimises the residual $F(y) = Ay - b$ in the least square sense, iff the coefficient vector $c = (c_0, \ldots, c_i)$ together with the Lagrange multiplier λ solves the (DIIS) system of equations*

$$\begin{pmatrix} B & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}\,. \tag{5.18}$$

The matrix B is determined by the entries $B_{lk} = \langle F(\tilde{x}_{l-1}), F(\tilde{x}_{k-1})\rangle$.

Proof. The minimisation problem

$$\min\{\|F(y)\|_2 \,:\, y \in \mathcal{D}^i\} \tag{5.19}$$

coincides with the problem of minimising the functional

$$G(c) = \frac{1}{2}\langle A\sum_{l=0}^{i} c_l\tilde{x}_l - b, A\sum_{l=0}^{i} c_l\tilde{x}_l - b\rangle\,,$$

over $\mathbb{R}^{i+1}$ with the constraint $\sum_{l=0}^{i} c_l = 1$. Standard Lagrangian calculus applied to this problem yields the system of equations (5.18) as a necessary condition. Since A is assumed positive definite, (5.18) has a unique solution. And thus $y = \sum_{l=0}^{i} c_l\tilde{x}_l$ is the unique solution to (5.19) in $\mathcal{D}^i$. □

Note that for $v \in \mathcal{K}^i$ it holds $F(v) \in K^{i+1}$ and thus the updated set $X^{i+1} : \{\tilde{x}_0, \ldots, \tilde{x}_i, \tilde{x}_{i+1}\}$ again spans the next Krylov subspace $\mathcal{K}^{i+1}$. With this we can now show equivalence of the sequence computed by the DIIS accelerated procedure and the GMRES method.

Theorem 5.14. *Let A be a positive definite matrix in $\mathbb{R}^{N\times N}$ and $b \in \mathbb{R}^N$. Let $F(x^*) = 0$ be the affine linear problem*

$$Ax^* - b = 0$$

and let $I(x)$ be the operation

$$I(x) = x - F(x) = x - (Ax - b)\,.$$

Further let $\tilde{x}_0 \in \mathbb{R}^N$ be some initial value. Then the DIIS accelerated procedure described in Algorithm 5.9 produces the same optimised iterates $x_i = \sum_{l=0}^{i} c_l\tilde{x}_l$, $\sum_{l=0}^{i} c_l = 1$, as the GMRES method.

Proof. The assertion is shown by induction over the iteration steps.

For $i = 0$ we have $\mathcal{K}^0 = \mathcal{D}^0 = \{\tilde{x}_0\}$. In particular the set of previous iterates $X^0 : \{\tilde{x}_0\}$ lies in the space $\mathcal{K}^0$, thus fulfilling Assumption 5.11. As can be seen, a DIIS and GMRES step yield the same minimisation problem and thus produce the same iterate $x_0 = \tilde{x}_0$. Since by definition of I the next approximated iterate $\tilde{x}_1$ is given by

$$\tilde{x}_1 = x_0 - F(x_0) = x_0 + (b - Ax_0)\,,$$

the updated set $X^2 : \{\tilde{x}_0, \tilde{x}_1\}$ fulfils Assumption 5.11 as well.

Inductive step $i \to i+1$. By induction hypothesis the set $X^i = \{\tilde{x}_0, \ldots, \tilde{x}_i\}$ fulfils Assumption 5.11. As shown in Lemma 5.13 the DIIS step minimises the residual $\|F(x)\|_2$ in the DIIS construction space $\mathcal{D}^i$. The GMRES method minimises the residual in the Krylov

subspace $\mathcal{K}^i$. But, by Lemma 5.12, $\mathcal{K}^i$ and $\mathcal{D}^i$ are the same. Thus one step of DIIS and one step of GMRES yield the same optimised solution x_i. Finally, we see from

$$\begin{aligned} \tilde{x}_{i+1} &= x_i + b - Ax_i \\ &= x_i + b - A\left(\tilde{x}_0 + \sum_{k=0}^{i-1} a_k A^k (b - A\tilde{x}_0)\right) \\ &= x_i + (b - A\tilde{x}_0) - \sum_{k=0}^{i-1} a_k A^{k+1} (b - A\tilde{x}_0) \end{aligned} \tag{5.20}$$

that the updated set of iterates $X^{i+1} : \{\tilde{x}_0, \ldots, \tilde{x}_i, \tilde{x}_{i+1}\}$ fulfils Assumption 5.11. In (5.20) we used the representation indicated by the definition of the Krylov subspace (5.14).
This finishes the proof. □

So, for the (affine) linear case with a definite matrix the DIIS procedure computes exactly the same optimised iterates as the GMRES scheme and thus it can be seen as a special implementation of it. Of course this implementation is quite unfavourable, since it requires to memorise the complete history of computed iterates. However, if the matrix A is in addition symmetric Weijo et al. [97] showed that the DIIS scheme reduces to the *conjugate residual* (CR) method when effectively replacing the set of approximated iterates $X^i : \{\tilde{x}_0, \ldots, \tilde{x}_i\}$ by the optimised ones $\{x_0, \ldots, x_i\}$. With this strategy they can omit a large part of the history and restrict themselves to the last three iterates without loss of accuracy. Further note that, like the GMRES scheme, the DIIS procedure is exact after (at most) N steps.
Actually and fortunately, this procedure still makes sense for the nonlinear case and it is as well easy to apply. Thus, it gives a version of the GMRES scheme that is appropriate for nonlinear problems. We will deal with this case in the following.

5.4.3 Nonlinear Problems

Let us now look at the rootfinding problem

$$F(x^*) = 0\,,$$

where F is some nonlinear function. Since the DIIS procedure does not require explicitly linearity of F it is directly applicable to this problem. Meaning to minimise the residual $F(x_i)$ of the optimised subspace solution $x_i \in \operatorname{span}\{\tilde{x}_0, \ldots, \tilde{x}_i\}$ parametrised by $x_i = \sum_{l=0}^{i} c_l \tilde{x}_l$. To compute the coefficient vector $c = (c_0, \ldots, c_i)$ we replace the function $F(x_i)$ by the (subspace) linearisation $\sum_{l=0}^{i} c_l F(\tilde{x}_l)$. Thus, we end up with the same system of (DIIS) equations (5.12) as in the linear case. When using the standard main iteration I the next approximation is then given by

$$\tilde{x}_{i+1} = x_i - F(x_i)\,. \tag{5.21}$$

Of course, unlike the linear case, exactness of the found solution cannot be guaranteed after a finite number of steps. But still the DIIS scheme (and variants of it) has proven a favourable performance in application, cf. [60, 39].

Before describing some variations to the original scheme that enhance the performance in the nonlinear case, we demonstrate the relation of a general DIIS step to secant methods, cf. as well [68].

As seen by (5.10) the Jacobian J_F of F at the actual approximate iterate $\tilde{x}_i$ is characterised by the Taylor polynomials

$$F(\tilde{x}_l) \approx F(\tilde{x}_i) + J_F[\tilde{x}_i](\tilde{x}_l - \tilde{x}_i)\,, \quad l = 0, \ldots, i-1\,.$$

With this we can define an approximated Jacobian $M_i := \tilde{J}_F[\tilde{x}_i] \approx J_F[\tilde{x}_i]$ by requiring

$$M_i(\tilde{x}_l - \tilde{x}_i) = F(\tilde{x}_l) - F(\tilde{x}_i)\,. \tag{5.22}$$

for every $l = 0, \ldots, i-1$ in the subspace spanned by the differences $\tilde{x}_l - \tilde{x}_i$. The equations (5.22) are called secant conditions. Let us assume that both sets

$$\begin{aligned} \Delta X^i &: \{\tilde{x}_0 - \tilde{x}_i, \ldots, \tilde{x}_{i-1} - \tilde{x}_i\} \\ \Delta Y^i &: \{F(\tilde{x}_0) - F(\tilde{x}_i), \ldots, F(\tilde{x}_{i-1}) - F(\tilde{x}_i)\} \end{aligned}$$

are linearly independent. In the DIIS scheme the optimal approximation is now given by

$$x_i = \tilde{x}_i + \sum_{l=0}^{i-1} c_l(\tilde{x}_l - \tilde{x}_i) = \tilde{x}_i + M_i^{-1}\left(\sum_{l=0}^{i-1} c_l(F(\tilde{x}_l) - F(\tilde{x}_i))\right)\,, \tag{5.23}$$

where the coefficients are chosen, such that the following minimisation task is solved

$$\begin{aligned} & \min\{\|F(\tilde{x}_i) + \sum_{l=0}^{i-1} c_l(F(\tilde{x}_l) - F(\tilde{x}_i))\|_2 \,:\, c \in \mathbb{R}^i\} \\ = \;& \min\{\|F(\tilde{x}_i) + M_i \sum_{l=0}^{i-1} c_l(\tilde{x}_l - \tilde{x}_i)\|_2 \,:\, c \in \mathbb{R}^i\} \\ = \;& \min\{\|M_i\delta_i - b_i\|_2 \,:\, \delta_i \in X^i\} \end{aligned} \tag{5.24}$$

where in the last step we introduced $\delta_i := \sum_{l=0}^{i-1} c_l(\tilde{x}_l - \tilde{x}_i)$ and $b_i = -F(\tilde{x}_i)$. Regarding the last minimisation problem, it is necessary and sufficient for δ_i to be the minimiser of (5.24), that $M_i\delta_i - b_i$ fulfils the Petrov-Galerkin condition, cf. [83, Prop. 5.3]. Meaning it is orthogonal to all vectors from $M_iX^i = Y^i$, i.e.

$$\langle M_i\delta_i - b_i, w\rangle = 0\,, \quad \forall w \in Y^i\,.$$

Since $\delta_i \in X^i$, this means

$$M_i\delta_i = P_Y^i b_i = -P_Y^i F(\tilde{x}_i)\,,$$

where P_Y^i denotes the projector on ΔY^i. Thus (5.23) can be written in the form

$$x_i = \tilde{x}_i - M_i^{-1} P_Y^i F(\tilde{x}_i)\,. \tag{5.25}$$

Comparing (5.25) to classical secant methods, the optimised iterate should be given in the form

$$x_i = x_{i-1} - H_i F(x_{i-1})\,, \tag{5.26}$$

meaning, to compute the x_i by use of x_{i-1}. Analogous, by inserting the main iteration I, we can formulate (5.25) in a generalised form of (5.26) by setting

$$x_i = I(x_{i-1}) - M_i^{-1} P_Y^i F(I(x_{i-1}))\,. \tag{5.27}$$

If I is taken to be the identity the generalised equation (5.27) reduces to the classical secant method (5.26).

Remark 5.15. *Note that choosing $I(x) = x$ is a regular possibility. While this choice does not make sense in the linear case, it indeed does for nonlinear problems, when assuming a set of linearly independent values $\{x_0, \ldots, x_i\}$ to be given. However, of course, with this choice one will only find an approximation in the subspace spanned by these values.*

Furthermore, recall that according to the Dennis-Moré theorem (1974), cf. [18, Thm. 8.2.4], a sequence of iterates that converges and is generated by a rule like (5.26) converges superlinearly, iff the sequence $(M_i - J(x^*))s_i$, with $s_i = x_{i+1} - x_i$, converges superlinearly. Where the only additional assumptions are regularity-type assumptions on F, J_F and the matrices H_i. Thus, a sequence produced by a DIIS procedure is likely to behave similar when choosing the main iteration I carefully.

Remark 5.16. *As mentioned at the beginning of this section we used the DIIS formulation motivated in [97]. There are (at least) two further formulations closely related to this one, which we will shortly introduce. The first is the original description of Pulay [80] for accelerating self-consistent field iterations. He assumed a set of iterates $\{p^1, \ldots, p^m\}$ to be given that come from a quasi-Newton-Raphson procedure of the form*

$$p^{i+1} = p^i - H_0^{-1} g^i\,,$$

where g^i is the gradient $\partial E/\partial p$ at p^i and H_0^{-1} an approximation to the inverse Hessian matrix. Pulay called this procedure simple relaxation *(SR) after its application in geometry optimisation. His idea was to accelerate the SR procedure by finding a better approximant in the current subspace of known iterates, i.e. $p = \sum_{i=1}^{m-1} c_i p_i$. Denoting the true solution by p^f this can be written as $p = \sum_{i=1}^{m-1} c_i(p^f + e^i) = p^f \sum_{i=1}^{m-1} c_i + \sum_{i=1}^{m-1} c_i e^i$. Thus, in order to minimise the actual error $\|p - p^f\|^2$, the second term has to vanish under the condition $\sum_{i=1}^{m-1} c_i = 1$. Since the true solution is unknown, Pulay replaced the error e^i by $p^{i+1} - p^i$. This immediately leads to the famous DIIS system of equations. Having the subspace optimum p, Pulay then performed another step of the SR method and added the new iterate to his set. It is clear, how his formulation relates to our notation. Firstly,*

as an error measure e^i we use the distance between input- and output-density instead of the distance between two procedure-iterates. Secondly, Pulay's formulation describes the special case of I being his SR procedure. And third, Pulay's SR-Iterates p^i correspond to our approximated iterates $\tilde{x}_i$.

The second (further) formulation can be found in [40]. There the authors abstain from defining an intermediate step like $x_i \rightsquigarrow \tilde{x}_{i+1}$ (or $p \rightsquigarrow p^{i+1}$) and instead directly go to a new approximation by setting $x_{i+1} = \sum_i c_i(x_i + e_i) = \sum_i c_i x_i + \sum_i c_i e_i$. Here e_i describes the actual error at x_i. The coefficients are then determined in the usual (DIIS) way. The relation to our formulation is quite clear. The term $\sum_i c_i x_i$ still describes the optimal subspace solution x_{opt}. But, instead of checking convergence at this optimum, the authors directly look at a new approximation which is near the optimum when the vanishing of the second term is assumed. Thus, this choice of monitored iterates is reasonable. Note, that the application of the underlying main iteration $I(x_{opt}) = x_{opt} + \sum_i c_i e_i$ in this formulation is not constant but is adapted according to the coefficient vector c. Furthermore, this formulation is as well variable in the choice of the error measure e.

Looking at typical applications, one evaluation of F is usually quite expensive. Thus it is favourable to have as few function evaluations as possible. Regarding the DIIS procedure just presented we have two function evaluations per iteration step $F(\tilde{x}_i)$ and $F(x_i)$. To avoid the evaluation of $F(x_i)$ we can replace it by the minimised sum $\sum_{l=1}^{i} c_l F(\tilde{x}_l)$, which are equal in the linear case. For the nonlinear case it is assumed to be a good approximation to it. The algorithm then reads

Algorithm 5.17 (DIIS Iteration (single evaluation))**.**

- **Given:** $\tilde{x}_0$**;** I**;** F
- **set:** $X = \{\tilde{x}_0\}$**;** $i := 0$
- **while** 'not converged' **do**

  add $FX_i = F(\tilde{x}_i)$ **to** $FX : \{FX_0, \ldots, FX_{i-1}\}$

  solve $\begin{pmatrix} B & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ **for** $c = (c_0, \ldots, c_i)$

  get $x_i = \sum_{l=0}^{i} c_l \tilde{x}_l$

  get $\tilde{x}_{i+1} = x_i - \sum_{l=0}^{i} c_l F(\tilde{x}_l)$

  insert $\tilde{x}_{i+1}$ **in** X**;** $i \leftarrow i + 1$
- **end**

In particular this means, choosing the main iteration as an adaption to the actual history of iterates, cf. Remark 5.16. Of course there is no indication, whether this variant performs better or worse than the original one. But it needs one evaluation less.

Besides the freedom of choice concerning the computation of the next approximated iterate $\tilde{x}_{n+1}$, i.e. the choice of I, we can as well adapt the set of memorised iterates. An example of this strategy can be found in [97]. There the authors memorised the actual optimal iterate x_n instead of the approximations $\tilde{x}_n$. The corresponding algorithm reads

Algorithm 5.18 (DIIS Iteration (optimised memory))**.**

- **Given:** $\tilde{x}_0$**;** I**;** F
- **set:** $X = \{\tilde{x}_0\}$**;** $i := 0$
- **while** 'not converged' **do**

  add $FX_i = F(\tilde{x}_i)$ **to** $FX : \{FX_0, \ldots, FX_{i-1}\}$

  solve $\begin{pmatrix} B & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ **for** $c = (c_0, \ldots, c_i)$

  get $x_i = \sum_{l=0}^{i} c_l \tilde{x}_l$

  replace $\tilde{x}_i$ **by** x_i **in** X

  replace FX_i **by** $F(x_i)$ **in** FX

  get $\tilde{x}_{i+1} = I(x_i)$

  insert $\tilde{x}_{i+1}$ **in** X**;** $i \leftarrow i + 1$
- **end**

As shown by Weijo et al. [97] in the linear case, this version allows to shorten the history down to three iterates in the past without loss of accuracy. Of course, again, it is not guaranteed that this variant will work better than the original one, but for nearly linear or mildly nonlinear problems it should perform good since the information from the optimal subspace solutions x_i are kept explicitly. However this variant still takes two function evaluations per step.

5.4.4 History Shortening

Another important variation is the possibility of history shortening. This means considering only a certain number of previous iterates at every step. This number can be chosen fixed in advance or be adapted during the iterative process (trust region strategy). For linear problems shortening the history of the original DIIS scheme (Algorithm 5.9) is counterproductive since every information stored in the iterates is needed in order to guarantee convergence after (at most) N steps. Only with further changes in the method that ensure maintaining of gained information, a shortened history might be advantageous.

In contrast to that, shortening the history seems to be essential for nonlinear problems, since it provides the opportunity of getting rid of wrong information and preventing it

from getting too memory expensive. Furthermore, for nonlinear problems the DIIS scheme tends to find a solution in a linearised subspace spanned by the previous iterates. The bigger this subspace is, the stronger is the assumption of linearity of F in this subspace. Thus, in order to prevent the scheme from assuming the whole problem to be linear, the history has to be shortened.

On the other side shortening the history too much troubles the convergence, since needed information might be lost. Hence, it seems valuable to carfully choose the length of the history for the problem at hand.

A general algorithm incorporating short histories is given below

Algorithm 5.19 (DIIS Iteration (short history))**.**

- **Given:** $\tilde{x}_0$**;** I**;** F
- **set:** $X = \{\tilde{x}_0\}$**;** $i := 0$
- **while** 'not converged' **do**

  add $FX_i = F(\tilde{x}_i)$ **to** $FX : \{FX_0, \ldots, FX_{i-1}\}$

  solve $\begin{pmatrix} B & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ **for** $c = (c_0, \ldots, c_i)$

  get $x_i = \sum_{l=0}^{i} c_l \tilde{x}_l$

  get $\tilde{x}_{i+1} = I(x_i)$

  delete elements from X

  insert $\tilde{x}_{i+1}$ **in** X**;** $i \leftarrow i + 1$
- **end**

Certainly all of the mentioned variations can be applied simultaneously in an appropriate way. For example, when using a shortened history one can use a reordering of the set X^i, such that the corresponding sequence of residuals is monotonously decreasing. In this way the adaption process for X^i keeps good approximations and deletes bad ones.

5.5 Convex DIIS

As described before, a standard method when solving the Kohn-Sham system from density functional theory is the two dimensional *linear mixing* scheme

$$n_{i+1}^{in} = \alpha n_i^{out} + (1-\alpha) n^{in}\,, \tag{5.28}$$

where $n_i^{out} = \mathcal{N}(n^{in})$ and $\alpha > 0$.

Beside the known general upper bounds on the steplength α (cf. [17]), there is another effect causing low convergence rates, namely *charge sloshing*, cf. [60, 58, 81]. This instability effect mainly is a problem in metallic systems, but can occur in inhomogeneous non-metallic systems as well. It is caused by vibrational changes between several configurations that result in a likewise vibrating Hartree-term which then dominates the effective potential. Breaking the sloshing when using the *linear mixing* scheme is only possible by use of short steplengths (typical range $\alpha \in [0.01, 0.1]$).

As shown in [17] the convergence criterion for (5.28) is

$$|1 - \alpha\mu_i| < 1\,,$$

where μ_i denotes the i-th eigenvalue of the Jacobian J of $\mathcal{N}$. When approaching a stable solution the eigenvalues fulfil $\mu_i > 0$ and thus an adequate choice of α is necessarily positive. However, when using (badly chosen) approximations to the Jacobian during some iteration process or one is interested in unstable solutions, it might be $\mu_{min} \leq 0$. In this case the solution process diverges. In principle, to get convergence for the μ_{min} component, a negative α is then needed. This however, would destroy convergence of all remaining (positive) components. Hence, unstable solutions cannot be computed with this approach and the constraint $\alpha > 0$ is applied in *linear mixing*.

In contrast to that, the DIIS method was originally introduced for accelerating self-consistent field (SCF) or Hartree-Fock calculations. There the mixing procedure was used to combine electron orbitals. It is used in this way, for example, in the wavelet based code package *BigDFT*, [30]. Since wavefunction at certain points in real space can be both, positive or negative, there are no further constraints on the coefficients except the requirement of summing up to one. Particularly, negative coefficients are possible. Indeed, Kresse and Furthmüller in [60] explicitly allowed for negative coefficients when the DIIS subspace is of dimension two. This was needed in order to recognise a false moving direction.

When considering the Kohn-Sham system from density functional theory the situation changes completely since the object of interest (the charged particle density) is related to the orbital wavefunction by $|\psi|^2$. Hence, together with the occupation numbers the density is a positive function and negative values will never occur. Thus, unlike mixing schemes for orbital calculations, density mixing reacts sensitively to negative coefficients. This is due to the structure of the solution space (cf. Section 3)

$$K := \{(n,p) \,:\, n,p \in L^1_{\mathbb{R}}\,,\, n,p \geq 0\,,\, \int_\Omega n\,\mathrm{d}x = N_n\,,\, \int_\Omega p\,\mathrm{d}x = N_p\}\,. \tag{5.29}$$

Hence, in order to stay in the solution space the new density built by mixing previous densities has to be positive in every point. But this cannot be guaranteed when allowing negative mixing coefficients to occur. In what follows we will therefore introduce a further constraint ensuring the linear combination to be, additionally, convex.

Using the pure DIIS scheme for combining the densities in DFT calculations is of course possible, and sometimes done. In quantum chemistry this is then called *Pulay mixing*. However note that this approach is known to sometimes suffer from slow convergence or even fails to converge at all, cf. [39]. As well instabilities at convergence are known, cf. [7]. For enhancing the behaviour of pure DIIS in this situation one usually goes back to *linear mixing* which is additionally applied to the optimised iterates. This however, brings in the known problems for this scheme again.

Beside the *linear mixing* strategy, several other authors introduced convexity constraints in their schemes. For example Cancès [9] used a convexity constraint when combining density matrices in his *optimal damping* strategy which essentially is a direct minimisation method transformed such that it is adequate for solving the Kohn-Sham equations. Another example is the *general descent method for the free energy of multicomponent systems* introduced by Gajewski and Griepentrog in [27]. There the minimisation is performed under the general constraint of mass (or charge) conservation and the energy functional is assumed to consist of a strong convex and a non-convex part that has a Lipschitz continuous Fréchet derivative. The actual iterate is then (convex) mixed with the solution of a constrained minimum problem for a partially linearised (energy) functional. Hence, Cancès as well as Gajewski/Griepentrog use an energy formulation to choose a search direction together with a stepsize. In particular, Cancès can guarantee a decreasing energy during his iteration. For the DIIS version we introduced, this is not possible. Rather, we know from density functional theory that the found self-consistent solution minimises the energy but the convergence process need not produce a descending sequence of corresponding energies. Note, that the just described schemes both are more related to direct minimisation methods whereas we want to use the convexity argument in order to get a generalised density mixing scheme for a fixed point iteration that ensures the iterates lying inside the solution space.

5.5.1 DIIS and Kohn-Sham

Concerning the Kohn-Sham system, the nonlinear rootfinding problem $F(x^*)$ on which the DIIS procedure is applied has the form

$$n^* - \mathcal{N}(n^*) = 0\,.$$

Due to this structure the memorised sets will be

$$\begin{aligned} X^i &: \{\tilde{n}_0, \ldots, \tilde{n}_i\}\,, \\ Y^i &: \{\mathcal{N}(\tilde{n}_0), \ldots, \mathcal{N}(\tilde{n}_i)\} \end{aligned}$$

and the DIIS minimisation problem then reads

$$\min\{\|\sum_{l=0}^{i} c_l(\tilde{n}_l - \mathcal{N}(\tilde{n}_i))\|_2 \,:\, \sum_{l=0}^{i} c_l = 1\}\,.$$

The necessity of the constraint $\sum_{l=0}^{i} c_l = 1$ is here twofold, firstly it makes sure that the total charge is conserved and secondly it prevents from finding the trivial zero solution. To lower the computational cost of one iteration step we will use the variant indicated by Algorithm 5.17. Meaning to use the standard main iteration $I(x_i) = x_i - F(x_i)$ and replacing $F(x_i)$ by $\sum_{l=0}^{i} c_l F(\tilde{x}_l)$. Thus, the next approximated iterate $\tilde{n}_{i+1}$ is computed by

$$\begin{aligned}\tilde{n}_{i+1} &= n_i - \sum_{l=0}^{i} c_l(\tilde{n}_l - \mathcal{N}(\tilde{n}_i)) \\ &= \sum_{l=0}^{i} c_l\tilde{n}_l - \sum_{l=0}^{i} c_l(\tilde{n}_l - \mathcal{N}(\tilde{n}_i)) = \sum_{l=0}^{i} c_l\mathcal{N}(\tilde{n}_i)\,.\end{aligned}$$

In what follows we prefix a DIIS version that uses the main iteration $I(x_i) = x_i - F(x_i)$, and thus taking two function evaluations per step, with the term 'original'. In comparison to that a DIIS version using $I(x_i) = x_i - \sum_{l=0}^{i} c_l F(\tilde{x}_l)$ (only one evaluation) is prefixed by 'KS', standing for Kohn-Sham.

As already mentioned there is no indication, that the single evaluation versions perform better or worse, compared to the original scheme. But anyway the computational costs for a single step are only half as big.

Unfortunately, we cannot use the error term $r_i = \frac{\|\mathcal{N}(n_i)-n_i\|}{\|n_i\|}$ for the optimal value n_i anymore in the convergence criterion in case of a 'KS' version, since $\mathcal{N}(n_i)$ is not computed anymore. We will instead use the error at the approximation $\tilde{n}_i$

$$\tilde{r}_i = \frac{\|\mathcal{N}(\tilde{n}_i) - \tilde{n}_i\|}{\|\tilde{n}_i\|}\,.$$

Note, that the convergence behaviour will essentially be the same assuming the sequence of iterates n_i converges. This is due to the fact that $\min\{\|\sum_{l=0}^{i} c_l(\tilde{n}_l - \mathcal{N}(\tilde{n}_i))\|_2 \,:\, \sum_{l=0}^{i} c_l = 1\}$ decreases monotonically.

5.5.2 Positivity Constraint: CDIIS

As mentioned before we have to introduce another constraint in order to stay in the solution space K, (5.29), when applying the DIIS scheme for accelerating solution procedures for

the Kohn-Sham system. Namely, this is the positivity constraint on the coefficients c_l. Recall that n is computed by the quantum mechanical expression

$$n(x) = \sum_{j=1}^{N} f(\mathcal{E}_j - \mathcal{E}_F)|\psi_j(x)|^2 \,,$$

and thus the outcome of the evaluation $\mathcal{N}(n)$ is necessarily a positive function. And the same is true for the sought solution n^*. However, when performing a general DIIS step, meaning to minimise

$$\min\{\| \sum_{l=0}^{i} c_l(\tilde{n}_l - \mathcal{N}(\tilde{n}_i))\|_2 \,:\, \sum_{l=0}^{i} c_l = 1\} \tag{5.30}$$

for the coefficient vector c, there is no guarantee that the components c_l are positive. And thus, by setting

$$n_i = \sum_{l=0}^{i} c_l \mathcal{N}(\tilde{n}_i) \,, \tag{5.31}$$

it is not sure that $n(x)$ is a positive value for every $x \in \mathbb{R}^3$. Hence, the density produced by the DIIS step might lie outside of the solution space K and thus be unphysical, which might corrupt the computation. This is additionally fatal, since from analysis in Section 3 we cannot guarantee uniqueness of the solution in the first place. Thus, appearance of unphysical solutions cannot be excluded.

To prevent the scheme from either finding those unphysical solutions or diverging, we have to make sure the density given by (5.31) is positive in every point, i.e. lies in K. To this end we replace the minimisation problem (5.30) by

$$\min\{\| \sum_{l=0}^{i} c_l(\tilde{n}_l - \mathcal{N}(\tilde{n}_i))\|_2 \,:\, \sum_{l=0}^{i} c_l = 1 \,;\, c_l \geq 0 \,;\, l = 0, \ldots, i\} \,. \tag{5.32}$$

Hence, in every step we now have to solve a quadratic minimisation problem with equality and inequality constraints. We do this, using the *active set* method from quadratic programming, cf. [5, 32]. The main idea behind this iterative procedure is that at the solution to (5.32) a certain set of inequality constraints is active, i.e. satisfied with equality. If this set was known a priori, the problem reduces to an optimisation problem with equality constraints only.

In our situation the method takes the following form of a quadratic programming problem

$$\min g(c) = c^T B c \,, \quad Ac = 1 \,, \quad c_l \geq 0 \tag{5.33}$$

with $A = (1, \ldots, 1) \in \mathbb{R}^{i+1}$. The procedure updates the actual solution

$$c \to \tilde{c} = c + t\, d$$

and the corresponding active set $\mathcal{K} \to \mathcal{L}$, ($c_{\mathcal{K}} = 0$, $\tilde{c}_{\mathcal{L}} = 0$). The following steps are then performed iteratively.

i) Denote with $\mathcal{I}$ the index set complementary to the active set $\mathcal{K}$. Minimise $g(c+d)$ subject to the active constraints, i.e. $d_{\mathcal{K}} = 0$, $A_{\mathcal{I}} d_{\mathcal{I}} = 0$. The non-trivial solution components of d are given as the solution of the system

$$\begin{pmatrix} B_{\mathcal{I}\mathcal{I}} & A_{\mathcal{I}}^T \\ A_{\mathcal{I}} & 0 \end{pmatrix} \begin{pmatrix} d_{\mathcal{I}} \\ \lambda \end{pmatrix} = \begin{pmatrix} -Q_{\mathcal{I}\mathcal{I}} c_{\mathcal{I}} \\ 0 \end{pmatrix}, \tag{5.34}$$

where λ is the Lagrange multiplier belonging to the equality constraint and the subscript denotes the corresponding submatrices.

ii) If $d_{\mathcal{I}} = 0$, check wether c is optimal for the starting problem (5.33). This is the case, if the gradient vector

$$\rho_{\mathcal{K}} = Q_{\mathcal{K}\mathcal{I}} c_{\mathcal{I}} + A_{\mathcal{K}}^T \lambda$$

is non-negative. If not, remove the index k of the smallest component (which is negative) from $\mathcal{K}$ and proceed with $\tilde{c} = c$, $\mathcal{L} = \mathcal{K} \backslash k$.

iii) If $d_{\mathcal{I}} \neq 0$, determine $t \in (0, 1]$ such that $\tilde{c} = c + td$ is feasible, i.e.

$$t = \min\{1, -\frac{c_l}{d_l} \text{ for } d_l < 0\}.$$

Add indeces $j \in \mathcal{I}$ with $\tilde{c}_j = 0$ to $\mathcal{K}$ and proceed with i).

As a feasible initial guess we use $c = (0, \ldots, 0, 1)$. Meaning that we assume the actual approximation $\tilde{x}_n$ to be the optimal solution in X^n.

Note, that the computational costs for finding the coefficient vector c with the iterative *active set* procedure are negligible, since the dimension of the (non-linear) Krylov subspace $\text{span}\{n_l - \mathcal{N}(n_l) : l = 0, \ldots, i\}$ is much smaller than the dimension of the underlying real-space problem.

Our algorithm for finding the fixed point of the Kohn-Sham system with only a single evolution now reads

Algorithm 5.20 (KS-CDIIS Iteration)**.**

- **Given:** $\tilde{n}_0$
- **set:** $X = \{\tilde{x}_0\}$**;** $i := 0$
- **while** 'not converged' **do**

  add $\mathcal{N}(\tilde{x}_i)$ **to** $Y : \{\mathcal{N}(\tilde{n}_0), \ldots, \mathcal{N}(\tilde{n}_{i-1})\}$

  solve $\min\{\| \sum_{l=0}^{i} c_l (\tilde{n}_l - \mathcal{N}(\tilde{n}_i)) \|_2 : \sum_{l=0}^{i} c_l = 1\,;\, c_l \geq 0\}$ **for** $c = (c_0, \ldots, c_i)$

  get $\tilde{n}_{i+1} = \sum_{l=0}^{i} c_l \mathcal{N}(\tilde{n}_l)$

  insert $\tilde{n}_{i+1}$ **in** X**;** $i \leftarrow i + 1$
- **end**

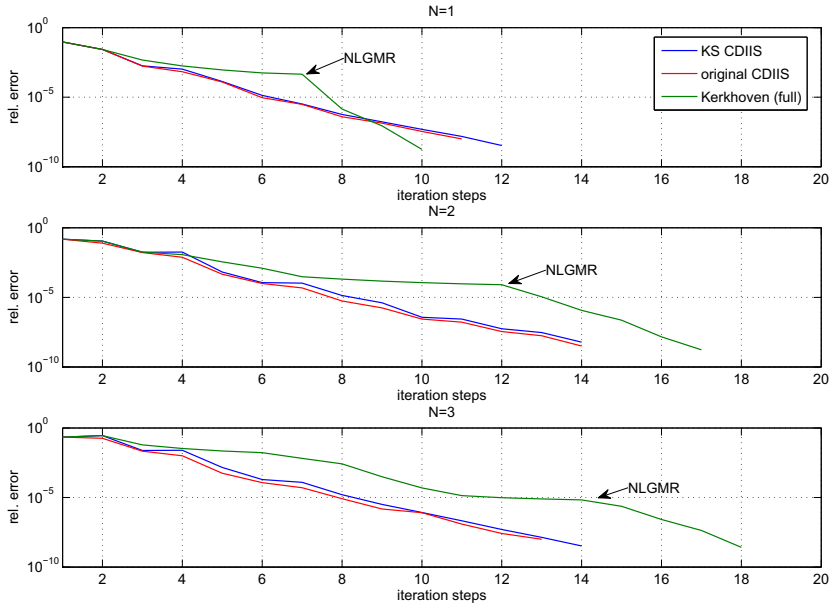

Figure 22: comparing residual evolution of the original CDIIS scheme (Algorithm 5.9), the KS CDIIS scheme (Algorithm 5.20) and the full Kerkhoven scheme (Algorithm 5.8).

In Figure 22 we see that the CDIIS versions usually perform equally good or better compared to the (full) Kerkhoven scheme, when taking the number of steps as a basis of performance rating only. The original CDIIS scheme, with two function evaluations performs slightly better than the KS CDIIS version (only one evaluation), as we expected. Thus we see that the Newton-type approach and the CDIIS mixing approach perform equally good. This can be observed as well from the convergence rates in Table 8. For comparison reasons the convergence rate of the (best) *linear mixing* strategy are given as well, cf. Table 5. Note that all accelerated strategies perform better than the *linear mixing* when regarding the number of iteration steps only.

	Kerkhoven	original CDIIS	KS CDIIS	(best) *lin. mix.*
$N = 1$	0.1387	0.2009	0.2106	0.2880
$N = 2$	0.3187	0.2569	2699	0.4669
$N = 3$	0.3417	0.2441	2501	0.5141

Table 8: Convergence rates $\varrho_{k,0}$ for the full Kerkhoven (Algorithm 5.8), the original CDIIS (Algorithm 5.9) and the KS CDIIS scheme (Algorithm 5.20) from Figure 22. For comparison reasons the rates for the (best) *linear mixing* strategy (Algorithm 5.3) are given as well.

	Kerkhoven	original CDIIS	KS CDIIS	(best) *lin. mix.*
$N = 1$	52	46	26	30
$N = 2$	70	58	30	48
$N = 3$	64	54	30	56

Table 9: Number of solved eigenvalue problems for the original CDIIS scheme (Algorithm 5.9), the KS CDIIS scheme (Algorithm 5.20) and the full Kerkhoven scheme (Algorithm 5.8) from Figure 22. The corresponding numbers for the (best) *linear mixing* scheme (Algorithm 5.3) are given for comparison.

However, when regarding Table 9 we see that the number of function evaluations (i.e. solved eigenvalue problems) is lowest for the KS CDIIS algorithm. Hence, even though the original CDIIS scheme as well as the Kerkhoven scheme may need less iteration steps to approach the solution, the KS CDIIS algorithm is more efficient, since it needs less time to do so. Thus, among the three acceleration schemes, the KS CDIIS scheme is preferable. Again the values for the *linear mixing* scheme are given. Note that under this aspect *linear mixing* is not automatically the worst choice. It needs less evaluations than the Newton-like acceleration and the original CDIIS scheme, making the better (in terms of iteration steps) performance questionable. But the KS CDIIS outperforms all the other schemes from this point of view. Thus, the fact that only one function evaluation is done during one step of the KS CDIIS iteration is rather essential. But this is true as well for *linear mixing*. Hence, there must be another reasoning for the good performance. And indeed, the KS CDIIS scheme additionally benefits from the higher dimensional subspace mixing ability. Both aspects together (single evaluation, mixing ability) clarify the strength of the KS CDIIS scheme besides the guarantee of safely staying in the solution space.

Temperature	4K	77K	300K
$N = 1$	0.1984	0.1985	0.1795
$N = 2$	0.3318	0.3025	0.2655
$N = 3$	0.2896	0.2897	0.2560

Table 10: Convergence rates $\varrho_{k,0}$ for the KS DIIS scheme (Algorithm 5.20) with a history of 5 for temperatures ($4K$, $77K$ and $300K$) from Figure 23

The graphs in Figure 23 lastly show calculations for different temperatures using the KS CDIIS scheme with a history of length 5. As we expected, the performance slightly increases when staying away from the singular $0K$ problem. The corresponding convergence rates can be found in Table 10. This shows that the KS CDIIS scheme seems to be quite insensitive to changes in the temperature and even when approaching zero temperature, the good performance is kept up.

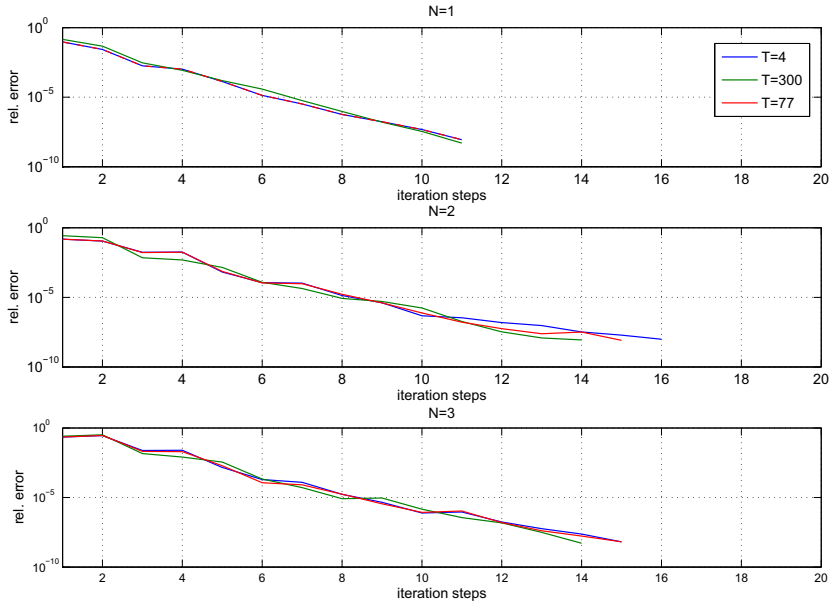

Figure 23: comparing residual evolution for the KS CDIIS scheme (Algorithm 5.20) with a history of 5 at temperatures ($4K$, $77K$ and $300K$).

5.5.3 History Length and Occupation Pattern

As was seen by the previous calculations, the KS DIIS scheme outperforms the *linear mixing* scheme in the number of iteration steps and solved eigenvalue problems. However, there is another important criterion for rating the performance of iteration schemes. Namely, this is the storage used during the calculation. In KS CDIIS we have to keep track on every iterate calculated so far, which of course makes it quite storage expensive. Especially, for problems coming from discretised partial differential equations. Here, the *linear mixing* clearly has an advantage since only two iterates are needed during the process. Thus, it is worth analysing the behaviour of KS CDIIS when shortening the number of memorised iterates.

Figure 24 shows a comparison of the KS CDIIS scheme (Algorithm 5.20) for different lengths of histories. We see, that shortening the history of course has an effect, but it seems that a lot of gainful informations are already contained in low dimensional subspaces. This allows a shortening of the history without too much performance losings. However shortening the history excessively clearly troubles the calculation, as can be seen from the computation with a history of three. Nevertheless, we can say that a history of length 5 to 10, in any case, already produces satisfactorily results. See Table 11 for a summary of the corresponding convergence rates.

As just mentioned, primal information might already be contained in low dimensional

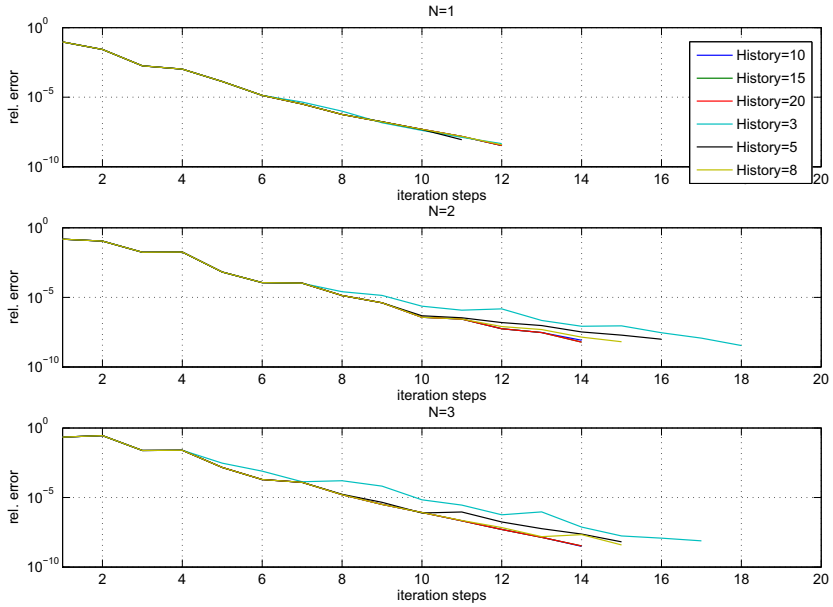

Figure 24: Comparing residual evolution of KS CDIIS scheme (Algorithm 5.20) with different lengths of history.

subspaces. When regarding a certain length of history, it is of interest which iterates exactly span the actual subspace used to find the next approximation. In case of CDIIS this can easily be analysed by looking at the occupation pattern. Meaning the segmentation of zero- and nonzero-coefficients. Especially it is of interest, whether the newest information coming from the last approximation is considered or not. If not, there is the danger of getting stuck in a subspace by calculating the same solution over and over again.

To illustrate this, we have a look on Figure 25 showing calculations of the KS CDIIS method for history length 5 and 20. There, we changed the convergence criterion down to a demanded accuracy of 10^{-15}. In this case numerical noise will get dominant and avoid the termination of the process. Of course the problem now is artificially produced but in a case when one is not sure about a trustable accuracy it is important to identify the

History	3	5	8	10	15	20
$N = 1$	0.2168	0.1984	0.2129	0.2106	0.2106	0.2106
$N = 2$	0.3553	0.3318	0.2978	0.2764	0.2699	0.2699
$N = 3$	0.3417	0.2896	0.2793	0.2487	0.2501	0.2501

Table 11: Convergence rates $\varrho_{k,0}$ for the KS DIIS scheme (Algorithm 5.20) with different lengths of history from Figure 24.

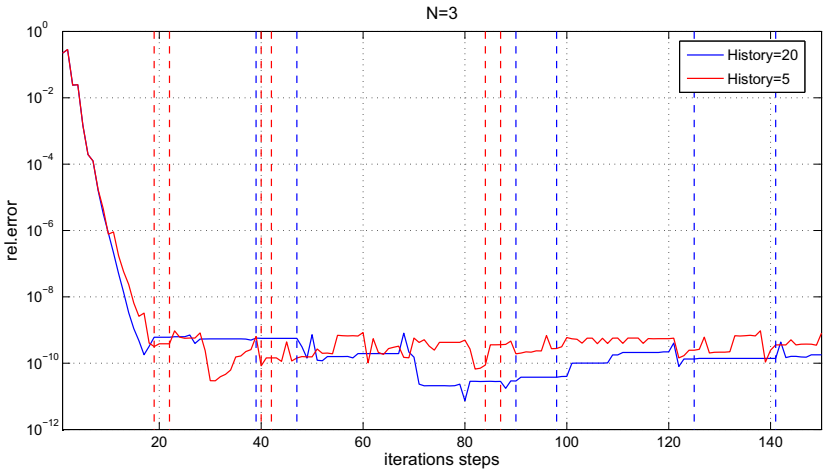

Figure 25: Comparing residual evolution of KS CDIIS scheme (Algorithm 5.20) for history lengths 5 and 20 with reduced convergence criterion.

problem. Furthermore, for a complicated error landscape it is possible that the process gets stuck in a similar way without touching noise effects.

We clearly can see the described effect. For the 20-evolution several plateau-like regions occur, whereas there are none in the 5-case. Thus, the algorithm stops improving (or even changing) the iterates for a certain number of steps. The reason can be understood when looking on the occupation pattern shown in Figure 26. There we see the pattern of the corresponding matrices $C = (c_{i,j})$, where $c_{i,j}$ is the CDIIS-coefficient of the j-th element of X^n in the i-th iteration. Non-existing elements like $c_{1,2}$ are set to zero, meaning C to be lower triangular. For the short history of 5 the main diagonal is almost completely occupied, which means that the actual approximated solution $\tilde{n}_i$ has a contribution when composing the optimal solution n_i. Hence, new information are incorporated immediately.

For a 20-history there are long ranges with only zero elements on the main diagonal. Thus, the information contained in the actual approximation are not used and instead the same subspace is taken into account. Additionally, the effective dimension of the subspace used to compose the optimal solution is small. Hence, it might happen that new information are neglected as long as the subspace that is used to compose the optimal solution is contained in the subspace spanned by X^i. This results in the vertical pattern as we can see them in the occupation patterns. Note however, that a zero-coefficient for the actual approximation does not necessarily mean there is a problem in the process evolution. It could rather mean that the non-linearity of the problem led to a false estimation of the localisation of the solution and the procedure just reacts properly. Such situation can usually be seen in the error evolution where they result in a worsening of the improvement by increasing the error.

In Figure 25 and 26 we indicated no-progress regions with dashed lines. The vertical lines

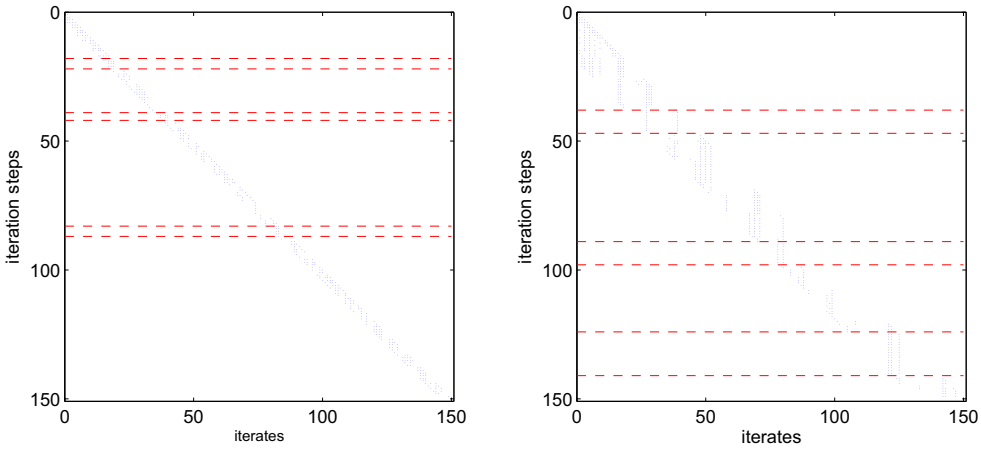

Figure 26: Occupation pattern for KS CDIIS scheme (Algorithm 5.20) with history lengths 5 and 20 from Figure 25.

in Figure 25 correspond to the horizontal lines in Figure 26.

Another important fact we can learn from this is the need for a finite history. If during the calculation the process gets stuck in a subspace spanned by certain iterates, it can only leave this impasse when deleting one of the basis vectors of the subspace. Thus, for an infinite history the process would never recover. So for large histories, when detecting vertical pattern formations, the process should properly react on this to break the subspace lock. For short histories the situation is a little more relaxed. Since the subspace is small anyway it will only take a few iterations, until one of the originator vectors expires.

5.6 Summary

In this work about the Kohn-Sham system and its numerical treatment, we introduced a high dimensional generalisation of the *linear mixing* scheme that leads to a considerably acceleration of the iterative process.

Our starting point was the commonly used damping strategy *linear mixing* from density functional theory. Even though usually successful, it often suffers from slow convergence due to strong damping which results in lengthy time-consuming calculations.

To overcome this, a switch to acceleration methods is done such as *Newton* or *Newton*-like procedures which are known for their good convergence behaviour. However, one iteration of such acceleration methods usually is quite expensive since it requires computation of information about the Jacobian. In essence we can say, a damping strategy is slow but cheap and a Newton-acceleration is fast but expensive. With the developed CDIIS method we were able to combined the good aspects from both worlds. Like *linear mixing* it requires only a singel function evaluation, while the performance is comparable to a Newton approach.

The basis of our CDIIS method is the well-known *direct inversion in the iterative subspace* (DIIS) method from quantum chemistry. As in DIIS we try to find a linear combination of iterates minimising a given error functional. However, the originally given version of DIIS cannot safely be applied to our fixed point problem. This is due to its extrapolation ability producing negative coefficients. For the Kohn-Sham density this means leaving the solution space of positive functions. To overcome this, we introduced an additional positivity constraint resulting in positive coefficients. Thus, the linear combination is ensured to be a convex one. With this we guarantee the new approximation to stay in the solution space.

The given formulation of CDIIS is embedded in a generalised formulation of DIIS-accelerated iteration procedures. The variable components are the main iteration I and the error measuring functional F. Both are essential for the nature of the procedure. The main iteration describes how one tries to basically approach to the solution, while the error functional decides about the weighting and judgement of the calculated iterates. Furthermore, this formulation can be used to handle general nonlinear problems that can be written as a fixed point problem.

By regarding the linear case, we pointed out that DIIS is equivalent to GMRES in this situation. Thus, DIIS provides an alternative implementation of the GMRES method. While disadvantageous in the linear case, the DIIS implementation of GMRES is directly applicable to nonlinear problems as well. In this way we end up with a general version of the GMRES method carried over to nonlinear problems.

Using the CDIIS scheme we successfully accelerated 3D-exciton calculations for a cylindrical quantum dot structure as described in Section 4. The analytical results from Section 3 showed existence of solutions in a rather general setting that includes zero and finite temperature configurations, cf. [15]. Furthermore, gainful properties like the analyticity of the

particle density operator are included, cf. [43], e.g. analyticity was used in the development of a steadily converging iteration scheme for the free energy of a multi-component system, cf. [27]. The calculation performed for different temperatures showed additionally robustness of the CDIIS method concerning fractional occupation of states.

5.7 Outlook

The results on the numerical behaviour of the presented CDIIS indicate a high potential for accelerating self-consistent iterations on basis of the particle density in DFT. Several topics are appropriate for a closer look towards a better understanding and further improvements of the procedure.

Convergence Analysis

The numerical calculations carried out in the last section showed the promising performance of the CDIIS method. In Section 5.4.3 we already mentioned the connection to secant methods. Using this connection one might be able to analytically show convergence of the produced iterates. A starting point of these considerations could be the exemplary treatment of *Broyden's* method ([8, 93]). The corresponding proofs of convergence for *Broyden* can be found in [18, Ch. 8] and [29]. To use these results one should work on the representation of the CDIIS method in form of a secant method comparable to the Broyden update. One could then follow the lines of the proof presented in [29] to get (local) linear convergence. Finally, similar considerations as carried out in the proof of [18, Thm. 8.2.2] about Broyden's method might lead to the application of the theorem of Dennis-Moré (1974), cf. [18, Thm. 8.2.4] giving superlinear convergence. This possible approach seems favourable since less is known about the Jacobian approximation defined by the secant conditions for the DIIS method. Thus, one should try to benefit from results about comparable looking method such as Broyden or other quasi-Newton methods like BFGS.

Energy as Error Functional

The aim of the CDIIS procedure is the minimisation of a certain error functional $F(n)$ in the convex hull of the previous iterates. The error functional used so far is considerably affected by the fixed point formulation we developed it for. More precisely, it is given by the actual residual that we tend to minimise. When going back to our main task, we realize that instead we are actually interested in minimising the systems energy $E(n)$, which defines the ground (or equilibrium) state of the problem at hand. Thus, it can be advantageous to include an energy dependence in the error functional and hence, the weighting of the iterates. Meaning to still use the same main iteration I coming from the fixed point procedure but changing the way of calculating the entries in the DIIS matrix.

Unfortunately, unlike the residual the energy at the true solution is not zero. And thus changing the minimisation task from $F(n_i) = 0$ to $E(n_i) = 0$ will hardly work. Instead, the energy error should be defined as $E(n_i) - E^* = 0$, where E^* denotes the ground (or equilibrium) state energy. But E^* is not known and one should have to work with an approximation to it. For example one could use $E(n_i) - \tilde{E}$, with $\tilde{E}$ the energy of the actual approximation. With this change, we could connect the CDIIS method with energy minimising procedures. However, it is at the one hand a priori not clear whether an energy dependent weighting is more promising than an residual approach. At the other hand the residual is cheap to calculate which is not clear for the energy. Putting together, it might be advantageous to use energy information in the composition of the next approximation, but this has to be done in an efficient way.

A Scaling

In this section the rescaling of the Kohn-Sham system from SI- to atomic-units is shown in detail. This is of special importance for finding appropriate coefficients for Poisson's equation and the local density approximation. To simplify the calculation we only consider a single species, namely electrons.

SI-Units

The Kohn-Sham system in SI-units reads

$$\begin{aligned} -\nabla\epsilon_r\epsilon_0\nabla\varphi &= q(D-n) \\ \left[-\frac{\hbar^2}{2}\nabla\frac{1}{m_r m_0}\nabla + V_{eff}(n)\right]\psi_i &= \mathcal{E}_i\psi_i \\ V_{eff}(n) &= V_0 + V_{xc}(n) - q\varphi(n) \\ n(x) &= \sum f(\mathcal{E}_i - \mathcal{E}_F)|\psi_i(x)|^2 \\ f(s) = \frac{1}{1+e^{\frac{s}{k_B T}}} \quad &, \quad \sum f(\mathcal{E}_i - \mathcal{E}_F) = N \equiv const. \end{aligned}$$

where the following constants appear:

Planck's constant	$\hbar$	$1.0596 * 10^{-34} Js$
dielectric permittivity	ϵ_0	$8.854187 * 10^{-12} Fm^{-1}$
electron mass	m_0	$9.1094 * 10^{-31} kg$
elementary charge	q	$1.6022 * 10^{-19} C$

Atomic Units

This system will be transformed into atomic units. The basic units of which are:

electron mass	$m_0 = 1$
elementary charge	$q = 1$
Planck's constant	$\hbar = 1$
dielectric permittivity	$\varepsilon_0 = \frac{1}{4\pi}$

To begin with and in view of the scaling of the exchange-correlation potential we look at the scaling to atomic units of the Schrödinger equation of the hydrogene atom.

Hydrogen Atom

The Schrödinger equation for the hydrogen atom in SI-units is given by

$$\left[-\frac{\hbar^2}{2m_0}\nabla^2 - \frac{Zq^2}{4\pi\varepsilon_0 r}\right]\psi_i = \mathcal{E}_i\psi_i$$

Introducing the length scaling $x = \lambda * x'$ we get

$$\left[-\frac{\hbar^2}{2m_0\lambda^2}(\nabla')^2 - \frac{Zq^2}{4\pi\varepsilon_0\lambda r'}\right]\psi_i' = \mathcal{E}_i\psi_i'$$

The scaling factor λ is chosen, such that

$$\frac{\hbar^2}{m_0\lambda^2} = \frac{q^2}{4\pi\varepsilon_0\lambda} \tag{A.1}$$

$$\Rightarrow\ \lambda = \frac{\hbar^2 4\pi\varepsilon_0}{m_0 q^2} = \frac{(1.0545716\ldots * 10^{-34})^2 \cdot 4\pi \cdot 8.854187\ldots * 10^{-12}}{2 \cdot 9.109382\ldots * 10^{-31} \cdot (1.6022\ldots * 10^{-19})^2} \approx 5.2918\ldots * 10^{-11}\,.$$

The quantity λ is called Bohr's radius and is denoted by a_0. From this we find the Hartree energy E_h, by use of (A.1), to be

$$E_h = 4.3597\ldots * 10^{-18} J = 27.2114 eV\,.$$

Schrödinger's Equation

The Schrödinger equation with effective potential is given by

$$\left[-\frac{\hbar^2}{2}\nabla\frac{1}{m_r m_0}\nabla + V_0 + V_{xc}(n) - q\varphi(n)\right]\psi_i = \mathcal{E}_i\psi_i\,.$$

Scaling of the length and energy to a_0 and E_h, respectively, yields

$$E_h\left[-\frac{1}{2}\nabla'\frac{1}{m_r}\nabla' + \frac{V_0}{E_h} + \frac{V_{xc}(n)}{E_h} - \frac{q}{E_h}\varphi(n)\right]\psi_i = \mathcal{E}_i\psi_i \tag{A.2}$$

$$\left[-\frac{1}{2}\nabla'\frac{1}{m_r}\nabla' + V_0' + V_{xc}'(n) - \frac{q}{E_h}\varphi(n)\right]\psi_i = \frac{\mathcal{E}_i}{E_h}\psi_i\,, \tag{A.3}$$

$$\tag{A.4}$$

where the prime indicates the corresponding quantity in atomic units. Concerning the electrostatic potential φ the unit is $\frac{E_h}{q}$. Thus we have

$$\left[-\frac{1}{2}\nabla'\frac{1}{m_r}\nabla' + V_0' + V_{xc}'(n) - \varphi'(n)\right]\psi_i = \mathcal{E}_i'\psi_i\,.$$

Note that the eigenfunctions of the Schrödinger operator are not affected.

Poisson's Equation

For Poisson's equation we analogously get

$$-\frac{\varepsilon_0}{a_0^2}\nabla\varepsilon_r\nabla\varphi = \frac{q}{a_0^3}(D' - n')$$

We thus have

$$-\frac{\varepsilon_0 a_0}{q}\nabla\varepsilon_r\nabla\varphi = D' - n'$$

A closer look on the factor $\frac{\varepsilon_0 a_0}{q}$ yields

$$\frac{\varepsilon_0 a_0}{q} = \frac{\kappa_0 a_0}{4\pi q} = \frac{a_0^2 m_0 q}{4\pi\hbar^2} = \frac{1}{4\pi}\frac{q}{E_h}$$

where we used the equations $\kappa_0 = 4\pi\varepsilon_0$ (absolute dielectricity), $a_0 = \frac{\kappa_0\hbar^2}{m_0 q^2}$ (Bohrs's radius) and $E_h = \frac{\hbar^2}{a_0^2 m_0}$ (Hartree energy). Resulting in

$$-\nabla\varepsilon_r\tfrac{1}{4\pi}\nabla\tfrac{\varphi}{\frac{E_h}{q}} = D' - n' \tag{A.5}$$

$$\Leftrightarrow \quad -\nabla\tfrac{\varepsilon_r}{4\pi}\nabla\varphi' = D' - n'\,. \tag{A.6}$$

Local Density Approximation

Finally, we treat the exchange-correlation term. In the local density approximation (LDA) this term is given by

$$V_{xc}(n) = -\left(\frac{3}{\pi}n\right)^{1/3}\,. \tag{A.7}$$

The origin of this term is the homogeneous electron gas. And in particular, it belongs to an equation of the form

$$\left[-\nabla^2 + V\right]\psi = \mathcal{E}\psi \tag{A.8}$$

instead of

$$\left[-\nabla\frac{1}{m}\nabla + V\right]\psi = \mathcal{E}\psi\,. \tag{A.9}$$

To be able to use (A.7), we need to adapt the unit system, such that the equation appears in the form (A.8). Analogue to the hydrogen atom, we get a length unit a_b and an energy unit E_b

$$\begin{aligned} a_b &= \frac{\varepsilon_r}{m_r}\cdot\frac{\hbar^2 4\pi\varepsilon_0}{m_0 q^2} = \frac{\varepsilon_r}{m_r}\cdot a_0\,, \\ E_b &= \frac{m_r}{\varepsilon_r^2}\cdot\frac{q^2}{4\pi\varepsilon_0 a_0} = \frac{m_r}{\varepsilon_r^2}\cdot E_h\,. \end{aligned} \tag{A.10}$$

With these units the term (A.7) would be correct and could then be used.

In our considerations heterostructures play an essential role. Meaning, the constants ε_r and m_r vary throughout the domain. And thus, a different unit system has to be used in every material which is not practical. Therefore we use the atomic unit-system throughout the whole domain and adapt the exchange-correlation term by a corresponding correction factor, according to (A.10). To get this correction factor, we look at the following calculation starting with the correct LDA term in the adapted units.

$$\begin{aligned} -\left(\frac{3}{\pi}n_b\right)^{1/3} E_b &= -\left(\frac{3}{\pi}n \cdot (\frac{\varepsilon_r}{m_r})^3\right)^{1/3} \cdot \frac{m_r}{\varepsilon_r} \cdot E_h \\ &= -\left(\frac{3}{\pi}n\right)^{1/3} \cdot \frac{\varepsilon_r}{m_r} \cdot \frac{m_r}{\varepsilon_r} \cdot E_h = -\frac{1}{\varepsilon_r}\left(\frac{3}{\pi}n\right)^{1/3} E_h \,. \end{aligned}$$

Thus the correction factor for the exchange-correlation term is given by $\frac{1}{\varepsilon_r}$.

List of Figures

List of Tables

References

[1] V. Berinde. *Iterative Approximation of Fixed Points*. Lecture Notes in Mathematics. Springer, Berlin, 2007.

[2] M. S. Birman and M. Z. Solomyak. Spectral asymptotics for nonsmooth elliptic operators. *Sov. Math.*, Dokl. 13:906–910, 1972 (English. Russian original).

[3] M. S. Birman and M. Z. Solomyak. Spectral asymptotics for nonsmooth elliptic operators. *Trans. Moscow Math. Soc.*, 27:1–32, 1975 (English. Russian original).

[4] M. S. Birman and M. Z. Solomyak. Spectral asymptotics for nonsmooth elliptic operators. *Trans. Moscow Math. Soc.*, 28:1–32, 1975 (English. Russian original).

[5] A. Björck. *Numerical methods for least squares problems*. SIAM, Philadelphia, 1996.

[6] J. Bochnak and J. Siciak. Analytic functions in topological vector spaces. *Studia Math.*, 39:77–112, 1971.

[7] D. R. Bowler and M. J. Gillan. An Efficient and Robust Technique for Achieving SelfConsistency in Electronic Structure Calculations. *Chemical Physics Letters*, 325(4):473–476, 2000.

[8] C. G. Broyden. A Class of Methods for Solving Nonlinear Simultaneous Equations. *Math. Comp.*, 19(92):577–593, 1965.

[9] E. Cancès. SCF algorithms for Kohn-Sham models with fractional occupation numbers. *J. Chem. Phys.*, 114:10616–10622, 2001.

[10] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. *Special Volume: Computational Chemistry*, volume 10 of *Handbook of Numerical Analysis*, chapter Computational Quantum Chemistry: A Primer, pages 3–270. Elsevier Science, 2003.

[11] K. Capelle. A Bird's-Eye View of Density-Functional Theory. *arXiv:cond-mat/0211443v5*, 2006.

[12] S. B. Chae. *Holomorphy and calculus in normed spaces*. Marcel Decker Inc., New York, 1985, (with an Appendix by Angus E. Taylor).

[13] R. Chill. On the łojasiewicz–Simon gradient inequality. *J. Funct. Anal.*, 201:572–601, 2003.

[14] C. Cohen-Tannoudji, B. Diu, and F. Laloë. *Quantenmechanik*, volume 1. de Gruyter, 3 edition, 2007.

[15] H. Cornean, K. Hoke, H. Neidhardt, P. Racec, and J. Rehberg. A Kohn-Sham system at zero temperature. *J. Phys. A: Math. Theor.*, 41:385304/1–385304/21, 2008.

[16] R. Courant and D. Hilbert. *Methoden der mathematischen Physik*. Springer-Verlag, 4. Auflage, Berlin, 1993.

[17] P. Dederichs and R. Zeller. Self-consistent iterations in electronic-structure calculations. *Physical Review B*, 28:5462–5472, 1983.

[18] J. R. Dennis Jr. and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. SIAM, Philadelphia, 1996.

[19] R. M. Dreizler and E. K. U. Gross. *Density Functional Theory: An Approach to the Quantum Many-Body Problem*. Springer-Verlag, 1990.

[20] N. Dunford and J. T. Schwartz. *Linear Operators: Part I*. Pure and Applied Mathematics: A Series of Texts and Monographs. Interscience Publishers, 1958.

[21] H. Eschrig. *The Fundamentals of Density Functional Theory, 2nd edition*. eagle Leipzig, 2003.

[22] R. Eymard, T. Gallouet, and R. Herbin. *Handbook of Numerical Analysis*, volume 7 of *Handb. of Num. Anal.*, chapter The finite volume method, pages 723–1020. North Holland, 2000.

[23] E. Feireisl, F. Issard-Roch, and H. Petzoltova. A non-smooth version of the łojasiewicz–Simon theorem with applications to non-local phase-field systems. *J. Differential Equations*, 199:1–21, 2004.

[24] W. R. Frensley. Boundary conditions for open quantum systems driven far from equilibrium. *Rev. modern Phys.*, 62, 1990.

[25] J. Fuhrmann, H. Langmach, T. Streckenbach, and M. Uhle. *p∂e*lib2. software toolbox, http://www.wias-berlin.de/software/pdelib.

[26] H. Gajewski. Analysis und Numerik von Ladungstransport in Halbleitern. *GAMM Mitteilungen*, 16(35), 1993.

[27] H. Gajewski and J. A. Griepentrog. A descent method for the free energy of multicomponent systems. *Discrete Contin. Dyn. Syst.*, 15:505–528, 2006.

[28] H. Gajewski, K. Gröger, and K. Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*. 1974.

[29] D. M. Gay and R. B. Schnabel. Solving systems of nonlinear equations by Broyden's method with projected updates. *NBER Working Paper Series*, (169), 1977.

[30] L. Genovese, S. Goedecker, A. Neelov, M. Ospici, D. Caliste, S. Ghasemi, T. Deutsch, and Q. Hill. BigDFT: A fast and precise DFT wavelet code. http://inac.cea.fr/sp2m/L_Sim/BigDFT.

[31] D. Gilbarg and N. Trudinger. *Elliptic partial differential equations of second order*. Springer, Berlin, 1983.

[32] P. Gill, W. Murray, and M. H. Wright. *Numerical Linear Algebra and Optimization*, volume 1. Addison Wesley, 1991.

[33] E. Giusti. *Direct methods in the calculus of variations*. World Scientific Publishing, 2003.

[34] N. I. Gould, Y. Hu, and J. A. Scott. A numerical evaluation of sparse direct solvers for the solution of large sparse, symmetric linear systems of equations. *Council for the Central Laboratory of the Research Councils (CCLRC) UK, Technical Report RAL-TR-2005-005*, 2005.

[35] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, London, 1985.

[36] K. Gröger. A $W^{1,p}$-estimate for solutions to mixed boundary value problems for second order elliptic differential equations. *Math. Ann.*, 283:679–687, 1989.

[37] W. Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Teubner Studienbücher, 1991.

[38] R. Haller-Dintelmann, C. Meyer, J. Rehberg, and A. Schiela. Hölder Continuity and Optimal Controle for Nonsmooth Elliptic Problems. *WIAS-Preprint No. 1316*, 2008, accepted for 'Appl. Math. Opt.'.

[39] R. J. Harrison. Krylov Subspace Accelerated Inexact Newton Method for Linear and Nonlinear Equations. *Journal of Computational Chemistry*, 25(3):328–334, 2004.

[40] T. Helgaker, P. Jorgensen, and J. Olsen. *Molecular electronic-structure theory*. Wiley, New York, 2000.

[41] E. Hille and R. S. Phillips. *Functional analysis and semi-groups*. American Mathematical Society, Providence, R.I., 1957, (rev. ed.).

[42] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):864–871, Nov 1964.

[43] K. Hoke, H.-C. Kaiser, and J. Rehberg. Analyticity for some Operator Functions from Statistical Quantum Mechanics. *Ann. Henri Poincare*, 10:749–771, 2009.

[44] J. Honerkamp. *Statistical physics*. Springer-Verlag, Berlin, 1998, (advanced approach with applications).

[45] H.-C. Kaiser, H. Neidhardt, and J. Rehberg. *WIAS-Preprint No. 1275, accepted for 'Monatshefte für Mathematik'*, 2007.

[46] H.-C. Kaiser and J. Rehberg. On stationary Schrödinger-Poisson equations modelling an electron gas with reduced dimension. *Math. Meth. Appl. Sci.*, 20:1283–1312, 1997.

[47] H.-C. Kaiser and J. Rehberg. About a stationary Schrödinger-Poisson system with Kohn-Sham potential in nanoelectronics. *WIAS Preprint 339, Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany*, 1997 (Appendix A by Udo Krause about the Kohn-Sham system).

[48] H.-C. Kaiser and J. Rehberg. About a stationary Schrödinger-Poisson system with Kohn-Sham potential in nanoelectronics. *WIAS Preprint*, 339, 1998.

[49] H.-C. Kaiser and J. Rehberg. About a one-dimensionaly stationary Schrödinger-Poisson system with Kohn-Sham potential. *Zeit. Angew. Math. Phys.*, 50:423–458, 1999.

[50] H.-C. Kaiser and J. Rehberg. About a stationary Schrödinger-Poisson system with Kohn-Sham potential in a bounded two- or three-dimensioal domain. *Nonlin. Ana.*, 41:33–72, 2000.

[51] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Second Edition, Berlin, 1984.

[52] T. Kerkhoven. Mathematical modelling of quantum wires in periodic heterojunction structures. *in: Semiconductors Part II*, 59:237–253, 1994.

[53] T. Kerkhoven. Numerical Nanostructure Modeling. *Z. angew. Math. Mech.*, 76, 1996.

[54] T. Kerkhoven, A. T. Galick, U. Ravaioli, J. H. Arends, and Y. Saad. Efficient numerical simulation of electron states in quantum wires. *J. Appl. Phys.*, 68(7), 1990.

[55] T. Kerkhoven, M. W. Raschke, and U. Ravaioli. Self-consistent simulation of quantum wires in periodic heterojunction structures. *J. Appl. Phys.*, 74(2), 1993.

[56] T. Kerkhoven and Y. Saad. On acceleration methods for coupled nonlinear elliptic systems. *Numer. Math.*, 60:525–548, 1992.

[57] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140(4A):A1133–A1138, Nov 1965.

[58] M. Kohyama. *Ab initio* calculations for SiC-Al interfaces: test of electronic-minimization techniques. *Modelling Simul. Mater. Sci. Eng.*, 4:397–408, 1996.

[59] T. Koprucki, R. Eymard, and J. Fuhrmann. Convergence of a finite volume scheme to the eigenvalues of a schrödinger operator. *WIAS Preprint*, (1260), 2007.

[60] G. Kresse and J. Furthmüller. Efficient iterative schemes of ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996.

[61] E. S. Kryachko and E. V. Ludena. *Energy Density Functional Theory of Many-Electron Systems*. Kluwer Academic Publishers, 1990.

[62] W. Kutzelnigg. Density Functional Theory (DFT) and ab-initio Quantum Chemistry (AIQC): Story of a difficult partnership. *Lecture Series on Computer and Computational Sciences*, 6:23–62, 2006.

[63] R. Lehoucq, K. Maschhoff, D. Sorensen, and C. Yang. ARPACK (ARnoldi PACKage) software. http://www.caam.rice.edu/software/ARPACK.

[64] R. Lehoucq and J. A. Scott. An evaluation of software for computing eigenvalues of sparse nonsymmetric matrices. *Preliminary proceedings, Copper Mountain Conference on Iterative Methods*, 1996.

[65] R. Lehoucq, D. C. Sorensen, and C. Yang. ARPACK Users Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods. *SIAM, Philadelphia*, 1998.

[66] E. H. Lieb. Density Functionals for Coulomb Systems. *Int. J. Quant. Chem.*, 24:243–277, 1983.

[67] E. Malic. *Many-particle theory of optical properties in low-dimensional nanostructures*. PhD thesis, Technical University Berlin, 2008. URN: urn:nbn:de:kobv:83-opus-20584.

[68] R. M. Martin. *Electronic Structure - Basic Theory and Practical Methods*. Camebridge Univ. Press, 2008.

[69] O. Mayrock, H.-J. Wünsche, F. Henneberger, O. Brandt, U. Bandelow, and H.-C. Kaiser. Calculation of localized multi-particle states in $(zn, cd)se$ and $(in, ga)n$ quantum wells. *Proceedings of the International Conference on the Physics of Semiconductors ICPS 24 (D. Gershoni, ed.)*, 1998, Art. IXB29 (1344.pdf).

[70] V. G. Mazya. *Sobolev spaces*. Springer-Verlag, Berlin, 1985.

[71] N. Mermin. Thermal properties of the inhomogeneous electron gas. *Phys. Rev.*, 137(5A):A1441–A1443, Mar 1965.

[72] F. Nier. A stationary Schrödinger-Poisson system arising from the modeling of electronic devices. *Forum Math.*, 2:489–510, 1990.

[73] F. Nier. A variational formulation of Schrödinger-Poisson systems in dimensions $d \leq 3$. *Commun. in Partial Differential Equations*, 18:1125–1147, 1993.

[74] K. P. O'Donnell, T. Breitkopf, H. Kalt, W. Van der Stricht, I. Moerman, P. Demeester, and P. G. Middleton. Optical linewidths of InGaN light emitting diodes and epilayers. *Appl. Phys. Lett.*, 70(14):1843–1845, 1997.

[75] E. M. Ouhabaz. *Analysis of Heat Equations on Domains*. Princeton University Press, 2005.

[76] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. International Series of Monographs on Chemistry (16). Oxford University Press, New York, Claredon Press, Oxford, 1989.

[77] A. Pazy. *Semigroups of linear Operators and Applications to Partial Differetnial Equations*. Springer-Verlag, Berlin, 1983.

[78] E. Prodan. Symmetry breaking in the self-sonsistent Kohn-Sham equations. *J. Phys. A*, 38:5647–5657, 2005.

[79] E. Prodan and P. Nordlander. On the Kohn-Sham equations with periodic background potentials. *J. Statist. Phys.*, 111:967–992, 2003.

[80] P. Pulay. Convergence acceleration of iterative sequences. The case of SCF iteration. *Chem. Phys. Letters*, 73(2):393–398, 1980.

[81] D. Raczkowski, A. Canning, and L. W. Wang. Thomas-Fermi charge mixing for obtaining self-consistency in density functional calculations. *Phys. Rev. B*, 64:1211011–1211014, 2001.

[82] B. K. Ridley. *Quntum processesn in semiconductors*. Clarendon Press, Oxford, 1999.

[83] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2003.

[84] Y. Saad and M. Schultz. GMRES: a Generalized Minimal RESidual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7(3):856–869, 1986.

[85] O. Schenk, M. Bollhöfer, and R. Römer. On large-scale diagonalization techniques for the Anderson model of localization. *Featured SIGEST paper in the SIAM Review selected "on the basis of its exceptional interest to the entire SIAM community". SIAM Review*, 50:91–112, 2008.

[86] O. Schenk and K. Gärtner. Solving Unsymmetric Sparse Systems of Linear Equations with PARDISO. *Journal of Future Generation Computer Systems*, 20(3):475–487, 2004.

[87] O. Schenk and K. Gärtner. On fast factorization pivoting methods for symmetric indefinite systems. *Elec. Trans. Numer. Anal.*, 23:158–179, 2006.

[88] O. Schenk, K. Gärtner, G. Karypis, S. Röllin, and M. Hagemann. PARDISO. sparse direct solver for linear systems, http://www.pardiso-project.org.

[89] O. Schenk, A. Wächter, and M. Hagemann. Matching-based Preprocessing Algorithms to the Solution of Saddle-Point Problems in Large-Scale Nonconvex Interior-Point Optimization. *Journal of Computational Optimization and Applications*, 36(2-3):321–341, 2007.

[90] U. Scherz. *Quantenmechanik: Eine Einfhrung mit Anwendungen auf Atome, Molekle und Festkrper*. Teubner Studienbücher, 1999.

[91] L. Schwartz. *Analyse Mathématique*. Hermann, Paris, 1967.

[92] H. Si. TetGen. Quality Tetrahedral Mesh Generator and 3D Delaunay Triangulator, http://tetgen.berlios.de.

[93] G. P. Srivastava. Broyden's method for self-consistent field convergence acceleration. *J. Phys. A: Math. Gen.*, 17(13):L317, 1984.

[94] C. J. Sun, M. Zubair Anwar, Q. Chen, J. W. Yang, M. Asif Khan, M. S. Shur, A. D. Bykhovski, Z. Lilienthal-Weber, C. Kisielowski, M. Smith, J. Y. Lin, and H. X. Xiang. Quantum shift of band-edge stimulated emission in InGaN-GaN multiple quantum well light-emitting diodes. *Appl. Phys. Lett.*, 70(22):2978–2980, 1997.

[95] M. M. Vaĭnberg and V. A. Trenogin. *Theory of branching of solutions of non-linear equations*. Noordhoff International Publishing, Leyden, 1974.

[96] C. Weisbuch and B. Vinter. *Quantum Semiconductor Structures: Fundamentals and Applications*. Academic Press Inc., San Diego, 1990.

[97] M. Zółkowski, V. Weijo, P. Jørgensen, and J. Olsen. An efficient algorithm for solving nonlinear equations with minimal number of trial vectors: Applications to atomic-orbital based coupled-cluster theory. *J. Chem. Phys.*, 128:204105–1 – 204105–12, 2008.