Numerical Analysis of Variational Problems in Atomistic Interaction Models

Bernhard Langwallner
The Queen's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Hilary Term 2011

Numerical Analysis of Variational Problems in Atomistic Interaction Models

Bernhard Langwallner, The Queen's College, University of Oxford
A thesis submitted for the degree of Doctor of Philosophy
Hilary Term 2011

The present thesis consists of two parts. The first part is devoted to the analysis of discretizations of a class of basic electronic density functionals. In the second part we suggest and analyze Quasicontinuum Methods for an atomistic interaction potential that is based on a field.

We begin by formulating and analyzing a model for the study of finite clusters of atoms or localized defects in infinite crystals based on a version of the classical Thomas–Fermi–Dirac–von Weizsäcker density functional. We show that the resulting constrained optimization problem has a minimizer and we provide a careful analysis of the solvability of the associated Euler–Lagrange equation. Based on these results, and using tools from saddle-point theory and nonlinear analysis, we then show that a Galerkin discretization has a solution that converges to the correct limit (in the case of Dirichlet as well as periodic boundary conditions).

Furthermore, we investigate the issue of optimal convergence rates. Using appropriate dual problems, we can show faster convergence for the energy, the Lagrange multiplier of the underlying minimization problem, and the L^2-errors of the solutions. We also look at the dependence of the density functional on the nucleus coordinates and show a convergence result for minimizing nucleus configurations.

These results are subsequently generalized to the case of discretizations with numerical integration. Existence and convergence of solutions, as well as optimal convergence rates can be established if quadrature rules of sufficiently high order are applied.

In the second part of the thesis we consider an atomistic interaction potential in one dimension given through a minimization problem, which gives rise to a field. The forces on atoms are in this case given by local expressions involving this field. A convenient feature of this model is the existence of a weak formulation for the forces, which provides a natural connection point for the coupling with a continuum model. We suggest Quasicontinuum-like coupling mechanisms that are based on a decomposition of the domain into an atomistic and a continuum region. In the continuum region we use an approximation based on the Cauchy–Born rule. In the atomistic subdomain a version of the atomistic model with Dirichlet boundary conditions is applied. Special attention has to be paid to the dependence of the atomistic subproblem on the boundary and the boundary conditions. Applying concepts from nonlinear analysis we show existence and convergence of solutions to the Quasicontinuum approximation.

Acknowledgements

After the last couple of years of work it is my wish to thank people who have, each in their own way, contributed to this thesis.

First of all, I would like to express deep gratitude to my supervisors Dr Christoph Ortner and Professor Endre Süli. Working with them was truly inspiring and taught me a lot about mathematics and research in general. They both devoted a lof ot their precious time to meetings and gave me guidance while still leaving me a lot of freedom in my research. I thank them for their patience and for checking, and thus improving, many drafts of papers and talks.

The Numerical Analysis Group provided a tremendously supportive and inspiring working environment. I would like to thank all group members for contributing to this friendly and productive atmosphere. In particular, I thank my office mates Leonardo Figueroa, Jaroslaw Fowkes, and Ricardo Pachón for great companionship and enjoyable discussions about mathematics or in fact anything else.

A special thanks also has to go to the initiators, managers, and all other members of the OxMOS project, the source of my financial support. I greatly profited from regularly presenting updates on my work and the numerous and diverse workshops we had in Oxford.

Oxford is not only a fantastic place to work but also to follow other passions or simply meet interesting people with the most different backgrounds and ideas. I thank all of my friends from college and student societies. It was good to sometimes get away from the office and experience other rewarding aspects of this university.

Last, but certainly not least, I wish to thank my parents, Anneliese and Helmut, as well as my sister Verena for their continuous love and support before and especially during my time in England.

Contents

Chapter 1

Introduction

Computational simulations have become ubiquitous tools across many disciplines in science and technology. Mathematical models of situations or phenomena are developed and implemented on computers to gain insight. From a mathematical point of view this gives rise to a large amount and diversity of challenges. Materials science represents an ideal field of activity for computational scientists since it provides a wealth of different questions ranging from electromagnetic properties to complex phenomena like fracture. Moreover, there are time and length scales with several orders of magnitude difference involved.

Ultimately, matter is composed of atoms. Phenomena involving nontrivial physics at the atomic scale should therefore be simulated using accurate microscopic models that take into account this discrete nature. However, even with modern computer technology and very efficient algorithms, purely atomistic simulations are limited to specimen sizes that are not sufficient for macroscopic applications. For example, 24 grams of aluminium consist of roughly 10^{24} atoms, which is far beyond what is feasible with atomistic models.

As a matter of fact, in many situations atomistic detail is not actually needed throughout the whole specimen under consideration. Sufficiently far away from the regions of special interest (crack tips, defects etc.) the material behaviour can be described well with continuum models. This observation led to the development of methods based on *atomistic/continuum coupling*. These methods aim to retain atomistic accuracy while leveraging the computational efficiency of continuum models. One class of models with this philosophy goes under the name of Quasicontinuum Methods, which we will introduce in more detail in Section 1.2.

The present thesis is divided into two parts: the first and bigger part (Chapters 2 and 3) is devoted to the numerical analysis of a class of Thomas–Fermi type electronic density functionals, which represent basic, but qualitatively valuable, quantum mechanical models. The second part (Chapter 4) deals with Quasicontinuum Methods for an interaction that is mediated by a scalar field. Although this model is even more basic than Thomas–Fermi type functionals, we believe its analysis to be a relevant contribution to a thorough mathematical understanding of Quasicontinuum Methods in the presence of fields like the electron density

and the electrostatic potential.

It has to be pointed out that in this thesis we focus on static problems at zero temperature.

1.1 Electronic Density Functionals

In this section we introduce some basic concepts of quantum mechanics and density functional theory. We do not aim to be exhaustive but provide a self-contained description of topics that are relevant for this thesis and its immediate context.

1.1.1 Quantum Mechanics

Quantum mechanics is the accepted microscopic theory of atoms, molecules, and solids. It is therefore highly desirable to perform accurate simulations based on this theory. Although the equations of quantum mechanics are compact and elegant, they involve high-dimensional configuration spaces, which makes their direct discretization an immense challenge. Scientists and engineers have over the last decades developed approximations and simplifications that have made simulations tractable and hence turned them into indispensable tools for physics, chemistry, and materials science.

We consider a physical system consisting of n_{at} atoms or, more precisely, n_{at} nuclei and n_{el} electrons. In principle, both electrons and nuclei are quantum mechanical particles whose behaviour is governed by the *Schrödinger equation*. However, since the nucleus masses are several order of magnitude larger than the electron masses, it is in most cases valid to treat the nuclei as classical particles whose positions $\{R_i\}_{i=1,\dots,n_{\mathrm{at}}}$ are known. This is referred to as the *Born–Oppenheimer approximation*.

It is a basic statement of quantum mechanics that positions of particles are not known exactly. For the electrons in our situation there exists a probability distribution $P : \mathbb{R}^{3n_{\mathrm{el}}} \to \mathbb{R}$ for the positions, which satisfies $\int_{\mathbb{R}^{3n_{\mathrm{el}}}} P \,\mathrm{d}x = 1$. According to quantum mechanics $P = |\Psi|^2$, where the so called *wave function* $\Psi : \mathbb{R}^{3n_{\mathrm{el}}} \to \mathbb{C}$ satisfies the Schrödinger equation[1]

$$H\Psi := \left(-\sum_{i=1}^{n_{\mathrm{el}}} \frac{1}{2}\Delta_i + \sum_{\substack{i,j=1 \\ i\neq j}}^{n_{\mathrm{el}}} \frac{1}{|x_i - x_j|} + V_{\mathrm{nuc}} \right) \Psi = E_0 \Psi. \tag{1.1}$$

Here, $x = (x_1, \dots, x_{n_{\mathrm{el}}}) \in \mathbb{R}^{3n_{\mathrm{el}}}$, Δ_i denotes the Laplace operator with respect to the coordinate x_i, and $Z_i \in \mathbb{N}$, $i = 1, \dots, n_{\mathrm{at}}$, are the charges of the nuclei in units of the electron charge. The potential V_{nuc} is given by the Coulomb potential of the nuclei:

$$V_{\mathrm{nuc}}(x) = -\sum_{i=1}^{n_{\mathrm{el}}} \sum_{j=1}^{n_{\mathrm{at}}} \frac{Z_j}{|x_i - R_j|}.$$

[1] Note that the equation is written in atomic units [55] and we neglect spin throughout the whole thesis.

The operator H is called the Schrödinger operator and its lowest eigenvalue E_0, is called ground state energy. The Born–Oppenheimer approximation implies that it is sufficient to determine E_0 and the respective ground state wave function Ψ_0. Even for dynamical simulations higher eigenvalues (so-called excited states) are usually physically irrelevant since the electrons relax to the ground state on a much faster scale than the nucleus movement takes place.

Electrons have spin $\frac{1}{2}$ and therefore belong to the class of fermions. The major implication of this property is that their wave function Ψ has to satisfy the *Pauli exclusion principle*, which can be expressed as $\Psi(x) = -\Psi(\pi_{i,j}x)$, where $\pi_{i,j}$ is the permutation operator that exchanges the coordinates x_i and x_j. This has some implications on potential approximations of quantum mechanical equations and energies.

The Schrödinger equation (1.1) has undoubtedly generated a lot of physical, mathematical and computational interest. From a numerical point of view we can say that the large dimension $3n_{\mathrm{el}}$ of the configuration space dramatically limits the size of problems that can be tackled; see [73] for a vivid demonstration of this fact.

Since H is self-adjoint, the Schrödinger equation (1.1) can be interpreted as the optimality condition for a variational problem. Multiplying from the left with the complex conjugate Ψ^* and integrating over $\mathbb{R}^{3n_{\mathrm{el}}}$ we obtain $E_0 = \int_{\mathbb{R}^{3n_{\mathrm{el}}}} \Psi^* H \Psi \,\mathrm{d}x$. Hence, (1.1) is equivalent to the minimization problem

$$E_0 = \min_{\Psi} \left\{ \int_{\mathbb{R}^{3n_{\mathrm{el}}}} \Psi^* H \Psi \,\mathrm{d}x \;:\; \int_{\mathbb{R}^{3n_{\mathrm{el}}}} |\Psi|^2 \,\mathrm{d}x = 1 \right\}. \tag{1.2}$$

This is the so-called *variational principle* [55].

1.1.2 Density Functional Theory

The variational principle (1.2) is a useful starting point for the construction of numerical methods. Minimizing over all possible Ψ is infeasible but an approximate energy and wave function can be found by restricting the minimization space (Ritz–Galerkin method).

The Hartree method [55, 72], for example, consists in using the product ansatz

$$\Psi(x) = \psi_1(x_1) \cdots \psi_{n_{\mathrm{el}}}(x_{n_{\mathrm{el}}})$$

with mutually orthogonal single particle wave functions $\psi_i : \mathbb{R}^3 \to \mathbb{R}$, $i = 1, \ldots, n_{\mathrm{el}}$. This separation of variables amounts to assuming that the particles do not interact. The Hartree–Fock method extends this choice to a so-called *Slater determinant* $\Psi(x) = \frac{1}{\sqrt{n_{\mathrm{el}}!}} \det[\psi_i(x_j)]_{i,j}$, thus satisfying the Pauli principle. Both methods yield nonlinear systems of equations (with nonlocal coefficients in the Hartree–Fock case [55, 72]) for the single particle wave functions ψ_i. These systems are solved by self-consistent iteration, see for example the review articles [23, 77, 78], which also contain an overview of available convergence results.

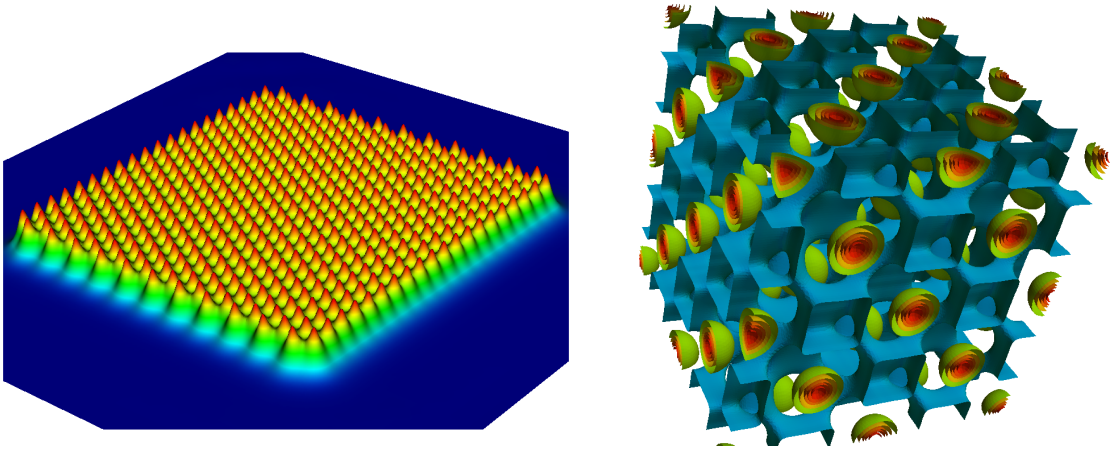

Figure 1.1: Examples of electron densities. On the left hand side the square root of the density of a two-dimensional crystal is shown. On the right-hand side contour surfaces of the square root density in a crystal with face-centered cubic lattice structure are shown.

A different approach, *Density Functional Theory (DFT)*, has emerged from two articles by Hohenberg and Kohn [68], respectively, Kohn and Sham [74]. The basic idea is to reduce the dimensionality by using the electron density $\rho : \mathbb{R}^3 \to \mathbb{R}$ instead of the wave function. Formally, the density is defined by

$$\rho(x) = n_{\mathrm{el}} \int_{\mathbb{R}^{3(n_{\mathrm{el}}-1)}} \Psi^*(x, x_2, \ldots, x_{n_{\mathrm{el}}}) \Psi(x, x_2, \ldots, x_{n_{\mathrm{el}}}) \mathrm{d}x_2 \ldots \mathrm{d}x_{n_{\mathrm{el}}}. \tag{1.3}$$

It satisfies the normalization condition $\int_{\mathbb{R}^3} \rho(x)\,\mathrm{d}x = n_{\mathrm{el}}$.

The *Hohenberg–Kohn Theorem* [68] states that there is a one-to-one correspondence between the potential V_{nuc} and the ground state density ρ_0. In other words, if the ground state density ρ_0 is known, the potential, and hence the wave function Ψ_0, can in principle be reconstructed. An immediate consequence of this is the existence of a density functional E_{HK} such that

$$E_0 = E_{\mathrm{HK}}(\rho_0) = \min\left\{ E_{\mathrm{HK}}(\rho) : \quad \rho \geq 0, \quad \int_{\mathbb{R}^3} \rho\,\mathrm{d}x = n_{\mathrm{el}} \right\}.$$

Levy and Lieb independently motivated the functional by carrying out the minimization in the variational principle (1.2) in two steps:

$$\min_{\Psi} \int \Psi^* H \Psi\,\mathrm{d}x = \min_{\rho} \left[\min_{\Psi|\rho} \int \Psi^* H \Psi\,\mathrm{d}x \right] =: \min_{\rho} E_{\mathrm{HK}}(\rho),$$

where by $\Psi|\rho$ we mean that the minimization takes place over all Ψ that have ρ as underlying density, that is they satisfy (1.3). Inserting the Schrödinger operator from (1.1) we obtain

$$E_{\mathrm{HK}}(\rho) = F_{\mathrm{HK}}(\rho) + \int V_{\mathrm{nuc}} \rho\,\mathrm{d}x\,,$$

where

$$F_{\mathrm{HK}}(\rho) = \min_{\Psi|\rho} \int \Psi^* \Big(-\sum_{i=1}^{n_{\mathrm{el}}} \frac{1}{2}\Delta_i + \sum_{\substack{i,j=1 \\ i\neq j}}^{n_{\mathrm{el}}} \frac{1}{|x_i - x_j|} \Big) \Psi \, \mathrm{d}x \,.$$

Here, F_{HK} is a universal functional in the sense that it is independent of V_{nuc}. It contains the kinetic energy of the electrons, the Coulomb interaction between electrons and quantum mechanical effects like the Pauli principle in the form of the so-called exchange-correlation. The precise form of F_{HK} is unknown but certain to be rather complex. However, a lot of work has gone into finding good approximations for F_{HK}.

Today there is a large number of different density functionals available. They can roughly be divided into two families: *orbital-free density functionals* and *Kohn–Sham functionals*. Introductions to DFT can be found in [88, 100].

Orbital-Free Density Functional Theory. Long before the theoretical justification for DFT was established, physicists used explicit functionals of the density to model physical systems. Thomas [116] and Fermi [54] independently of each other proposed models of the form

$$E_{\mathrm{TF}}(\rho) = T_{\mathrm{TF}}(\rho) + E_{\mathrm{ee}}(\rho) + E_{\mathrm{en}}(\rho). \tag{1.4}$$

The kinetic energy density of a homogeneous, noninteracting electron gas with density ρ is given by $\frac{3}{10}(3\pi^2)^{2/3}\rho^{5/3}$. This led to the following approximation of the kinetic energy

$$T_{\mathrm{TF}}(\rho) = \frac{3}{10}(3\pi^2)^{2/3} \int_{\mathbb{R}^3} \rho^{5/3}(x) \, \mathrm{d}x \,.$$

The term

$$E_{\mathrm{ee}}(\rho) = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)}{|x-y|} \, \mathrm{d}x \, \mathrm{d}y$$

represents the electrostatic repulsion between the electrons and

$$E_{\mathrm{en}}(\rho, R) = \int_{\mathbb{R}^3} \rho(x) V_{\mathrm{nuc}}(x) \, \mathrm{d}x \,, \qquad \text{where} \quad V_{\mathrm{nuc}}(x) = \sum_i \frac{-Z_i}{|x - R_i|},$$

the electrostatic attraction between electrons and nuclei. Dirac [40] suggested the addition of an exchange term to account for the quantum mechanical nature of the system:

$$E_{\mathrm{TFD}}(\rho) = E_{\mathrm{TF}}(\rho) + E_{\mathrm{x}}(\rho), \quad \text{where} \quad E_{\mathrm{x}}(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3} \int_{\mathbb{R}^3} \rho^{4/3}(x) \, \mathrm{d}x \,.$$

The resulting functionals enabled some physical insight [71, 80, 111], but showed some serious limitations. Most gravely, they failed to predict binding of atoms.

Von Weizsäcker [121] introduced the gradient correction term

$$T_{\mathrm{vW}}(\rho) = \frac{\lambda}{8} \int_{\mathbb{R}^3} \frac{|\nabla \rho(x)|^2}{\rho(x)} \, \mathrm{d}x$$

to the kinetic energy. Here, λ is a parameter that can be adjusted according to the application under consideration. Another popular addition to the functional is the correlation energy E_{c} in the so-called local density approximation

$$E_{\mathrm{c}}(\rho) = \int_{\mathbb{R}^3} \varepsilon_{\mathrm{c}}(\rho(x))\rho(x)\,\mathrm{d}x,$$

where ε_{c} is obtained in a purely phenomenological way [31, 102]. Summarizing the Thomas–Fermi–Dirac–von Weizsäcker functional is given by

$$E_{\mathrm{TFDW}}(\rho) = T_{\mathrm{TF}} + T_{\mathrm{vW}} + E_{\mathrm{x}} + E_{\mathrm{c}} + E_{\mathrm{ee}} + E_{\mathrm{en}} + E_{\mathrm{nn}}. \tag{1.5}$$

Note that we have added the electrostatic repulsion energy of the nuclei

$$E_{\mathrm{nn}}(R) = \frac{1}{2}\sum_{i=1}^{n_{\mathrm{at}}}\sum_{\substack{j=1\\ j\neq i}}^{n_{\mathrm{at}}} \frac{Z_i Z_j}{|R_i - R_j|}.$$

For the purely electronic problem this is an irrelevant constant. It will, however, be important when we study the dependence of the functional with respect to the nucleus coordinates. The given density functional (1.5) has to be minimized subject to the constraints

$$\int_{\mathbb{R}^3} \rho(x)\,\mathrm{d}x \;= N, \qquad \text{and} \qquad \rho \geq 0.$$

One of the most difficult tasks in DFT is the computation of the kinetic energy. This part of the energy is approximated more accurately in Kohn–Sham functionals (see below) but researchers have also developed kinetic energy functionals involving convolution integrals [118, 119] of the form

$$T_{\mathrm{K}}(\rho) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} f_1(\rho(x))K(|x-y|,\rho(x),\rho(y))f_2(\rho(y))\,\mathrm{d}x\;\mathrm{d}y. \tag{1.6}$$

with continuous functions f_1, f_2 and a convolution kernel K. The resulting, more sophisticated models are now referred to as orbital-free DFT. Reviews can be found in [34, 82]. Orbital-free DFT has been observed to work well for systems with weakly varying electron density, most prominently aluminium, see [118, 119]. In the present thesis, we will focus on the TFDW functional (1.5).

Kohn–Sham Density Functional Theory. The density functionals we have discussed so far only depend on the density ρ itself, which is precisely in the spirit of the Hohenberg–Kohn theory. This situation changes in Kohn–Sham DFT: here, n_{el} *fictitious non-interacting electrons* with pairwise orthogonal wave functions ψ_j are introduced. The density ρ is the sum of the probability distributions of these noninteracting particles $\rho(x) = \sum |\psi_i(x)|^2$. Moreover, the kinetic energy for these particles takes the simple form

$$T_{\mathrm{kin}} = \frac{1}{2}\sum_{i=1}^{n_{\mathrm{el}}} \int_{\mathbb{R}^d} \psi_i^*(-\Delta)\psi_i\,\mathrm{d}x.$$

This leads to single particle eigenvalue problems for the so-called *orbitals* ψ_j in an *effective, external potential* containing electrostatic interaction as well as quantum mechanical effects:

$$H_{\text{eff}}\psi_i(x) = \Big(-\tfrac{1}{2}\Delta + V_{\text{eff}}(x,\rho)\Big)\psi_i(x) = \varepsilon_i\psi_i(x).$$

The effective potential V_{eff} depends on ρ and therefore the ψ_i. The resulting system is usually solved in an iterative way. For fixed ρ the ψ_i are calculated, which leads to an updated density ρ. This process is iterated until convergence is obtained. In every step the n_{el} lowest eigenvalues and corresponding mutually orthogonal eigenfunctions of the effective Hamiltonian H_{eff} have to be found. This is computationally very demanding but leads to very accurate models. A recent mathematical review of DFT can be found in [105].

Even with the efficient algorithms and the computer technology available today DFT-based simulations are limited to $10^3 - 10^4$ atoms. For this reason a lot of work has gone into the development of methods that combine DFT or other atomistic models with computationally less demanding models (e.g. molecular mechanics or even continuum mechanics). In the following section we discuss one typical way of coupling phenomenological atomistic models with continuum mechanics.

1.2 The Quasicontinuum Method

We now consider a physical body in three space dimensions. The displacement on parts of the boundary is prescribed and the body is exposed to external forces. We think of the body as composed of n_{at} atoms rather than as a continuum and our aim is to find an equilibrium configuration. In many situations the number of atoms is too large to work with a purely atomistic model. The Quasicontinuum (QC) method [91, 108, 114] represents an attempt to couple an atomistic material description in regions of special interest with a continuum model in the rest of the body. This results in a significant reduction of complexity.

The underlying idea is that one starts off with an atomistic model for the whole body and uses continuum, or more precisely smoothness, assumptions to reduce the degrees of freedom and the complexity in regions that do not need full atomistic accuracy.

A crucial assumption we have to make is the existence of a reference configuration of the n_{at} atoms described by the lattice $\mathcal{L}$. Without external forces or given boundary displacements, the atoms of the body occupy lattice sites:

$$X_i = X_0 + \ell_1^{(i)}A_1 + \ell_2^{(i)}A_2 + \ell_3^{(i)}A_3, \quad \ell_1^{(i)}, \ell_2^{(i)}, \ell_3^{(i)} \in \mathbb{Z}.$$

Here, $X_0 \in \mathbb{R}^3$ is a reference point and $A_1, A_2, A_3 \in \mathbb{R}^3$ are the linearly independent lattice vectors. If the body is deformed, the atomic positions become y_i, $i = 1, \ldots, n_{\text{at}}$. We call the y_i deformed positions and define $\boldsymbol{y} = (y_1, \ldots, y_{n_{\text{at}}}) \in \mathbb{R}^{3n_{\text{at}}}$.

The atomistic interactions are often modelled using semi-empirical interatomic potentials. Examples include pair-potentials, embedded atom potentials or more general multibody interactions, see [38, 91]. Many of these semi-empirical models can be written in the form

$$\mathcal{E}(\boldsymbol{y}) = \sum_{i=1}^{n_{\mathrm{at}}} \mathcal{E}_i(\boldsymbol{y}),$$

where the energies $\mathcal{E}_i$ depend on relative positions of the atoms. This means that each individual atom is assigned an energy. In general, this is not possible for quantum mechanical models. The most basic examples are given by so-called pair-potentials

$$\mathcal{E}_i(\boldsymbol{y}) = \frac{1}{2} \sum_{\substack{j=1 \\ j \neq i}}^{n_{\mathrm{at}}} V(|y_i - y_j|),$$

where $V : \mathbb{R} \to \mathbb{R}$ is a two-body potential, e.g., the Lennard–Jones or Morse potential. These two-body potentials combine short-range repulsion and long-range attraction and define an equilibrium distance between two atoms. The following presentation will be restricted to the case of pair-potentials.

To obtain an equilibrium state of the body the total energy

$$E_{\boldsymbol{f}}(\boldsymbol{y}) = \mathcal{E}(\boldsymbol{y}) - \boldsymbol{f} \cdot \boldsymbol{y} \tag{1.7}$$

has to be minimized. Here, $\boldsymbol{f}$ represents external forces (e.g. gravity).

The challenges in solving (1.7) lie in the large number of degrees of freedom ($3n_{\mathrm{at}}$) and the complexity of calculating the energy $\mathcal{E}_i$ for each individual atom ($\mathcal{E}_i$ involves the summation over all other atoms). The QC method provides efficient means to overcome both of these difficulties. We present it in its most basic form. This does not immediately lead to a practical method but it makes the principles clear. The important steps of interface or ghost force corrections will be discussed subsequently.

There are two main steps: first, *coarse-graining* is used to reduce the number of degrees of freedom and, second, the *Cauchy–Born approximation* leads to a reduction of the number of necessary atomic energy evaluations.

Coarse-Graining. First, we introduce a continuous deformation y: y is a continuous function defined on the reference configuration such that $y(X_i) = y_i$ for all $i = 1, \ldots, n_{\mathrm{at}}$. The deformation gradient $F(X)$ is defined by $F(X) = \nabla y(X) - \mathrm{id}$.

The idea underlying coarse-graining is the following: in regions where the deformation gradient F varies slowly, it is not necessary to know the displacement of each individual atom. Instead, a set of n_{rep} representative atoms, so-called *repatoms*, is chosen. Their coordinates act as the degrees of freedom. Moreover, the repatoms are used to construct a triangulation

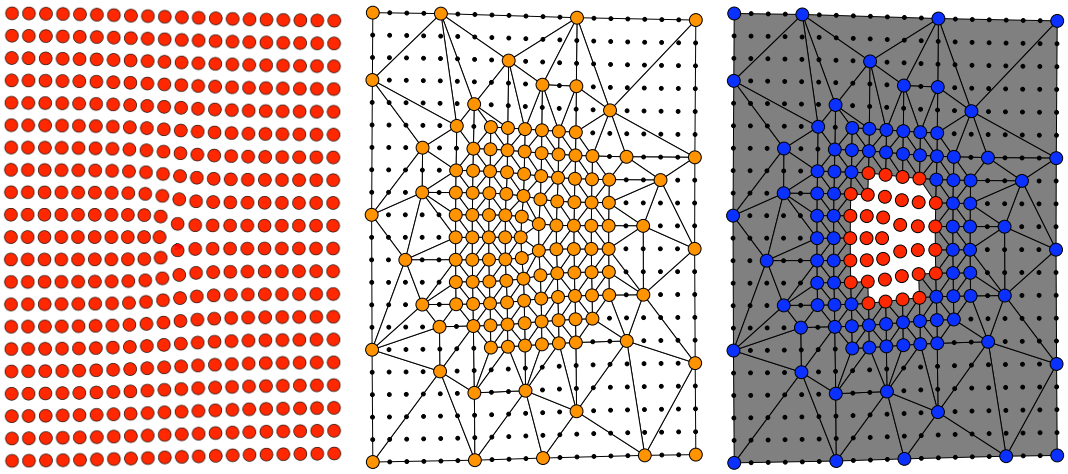

Figure 1.2: Sketch of three stages in the derivation of the QC method. Left: atoms forming a body. Centre: repatoms (orange) have been chosen and a triangulation has been constructed. The black dots depict atoms whose positions are obtained by linear interpolation. Right: division into atomistic repatoms $\mathcal{A}$ (red) and continuum repatoms $\mathcal{C}$ (blue). (Pictures courtesy of M. Dobson and M. Luskin)

$\mathcal{T}$ with $n_{\mathcal{T}}$ elements, see Figure 1.2. The positions of non-repatoms can then be approximated by linear interpolation utilizing the standard basis functions of the finite element method. In order to retain accuracy, the density of repatoms needs to be higher where the displacement varies quickly (for example around the dislocation in Figure 1.2). However, low repatom density in regions where y is smooth leads to a significant reduction of degrees of freedom.

Similarly to the finite element analysis of continuous minimization problems this gives rise to a Galerkin approximation of the problem (1.7):

$$\min_{\boldsymbol{y}_h} E_{\boldsymbol{f}}(\boldsymbol{y}_h) = \sum_{i=1}^{n_{\mathrm{at}}} \mathcal{E}_i(\boldsymbol{y}_h) - \boldsymbol{f}\cdot\boldsymbol{y}_h,$$

where the minimization takes place over all $\boldsymbol{y}_h$ that can be described by a continuous displacement y_h which is piecewise affine over the triangulation $\mathcal{T}$. Note that the number of degrees of freedom has been reduced to $3n_{\mathrm{rep}}$.

Summarizing we have reduced the number of degrees of freedom by introducing kinematic constraints on some atoms. However, since the energy $\mathcal{E}_i$ of each atom has to be computed the complexity of the evaluation of $\mathcal{E}(\boldsymbol{y}_h)$ is still prohibitive.

The Cauchy Born Approximation. To obtain the approximation that will give rise to the continuum model used in the QC method, we briefly change our perspective and look at the energy of elements $T \in \mathcal{T}$ instead of energies of individual atoms. Since y_h is piecewise affine, the deformation gradient on every element $T \in \mathcal{T}$ is constant, say F_T. The Cauchy–

Born approximation of the energy of T now consists in computing the energy of T if it is considered as a section of the infinite deformed lattice $F_T \cdot \mathcal{L}$. For this, we define an energy density $\mathcal{W} : \mathbb{R}^{3\times 3} \to \mathbb{R}$: $\mathcal{W}(F)$ is the energy per unit volume in the deformed configuration if the body is deformed according to $X \mapsto F \cdot X$ and hence given by

$$\mathcal{W}(F) = \lim_{m\to\infty} \left(\frac{1}{|m\Omega|} \sum_{y_i, y_j \in F\cdot\mathcal{L}\cap m\Omega} V(y_i - y_j) \right),$$

where Ω is any nonempty convex set in $\mathbb{R}^3$. The interpretation of this definition of $\mathcal{W}$ is straightforward: the energy content of the body $m\Omega$ consisting of atoms arranged on the lattice $F \cdot \mathcal{L}$ is computed and normalized to unit volume. The limit of a large body $m\Omega$ leads to an asymptotic energy density $\mathcal{W}(F)$.

We can now define the Cauchy–Born energy of the element $T \in \mathcal{T}$ by

$$\mathcal{E}^{\mathrm{cb}}(T) = |T|\mathcal{W}(F_T).$$

Note that instead of computing $\mathcal{E}_i(\boldsymbol{y}_h)$ for all atoms in the element T, $\mathcal{E}^{\mathrm{cb}}(T)$ only involves one evaluation of $\mathcal{W}$, which means a significant computational saving. We stress that $\mathcal{E}^{\mathrm{cb}}(T)$ is independent of the neighbouring elements of T. In fact, the derivation of $\mathcal{E}^{\mathrm{cb}}$ assumes that the deformation gradient is the same in neighbouring elements. Hence, the Cauchy–Born approximation is justified in regions where F varies sufficiently slowly.

The Cauchy–Born approximation for the whole body is obtained by summing over all elements $T \in \mathcal{T}$:

$$\mathcal{E}^{\mathrm{cb}}(\boldsymbol{y}^h) = \sum_{T\in\mathcal{T}} |T|\mathcal{W}(F_T) = \sum_{j=1}^{n_{\mathrm{rep}}} \omega_j \mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y}_h).$$

Note that in the second step we have converted the sum over elements into a sum over repatoms, by dividing the energy of each element T evenly among its repatom vertices. This process involving Voronoi tesselations gives rise to the weights ω_j and is described in more detail in [91].

Quasicontinuum Coupling. Equipped with the concept of coarse-graining and the Cauchy–Born approximation we are now ready to derive the most basic Quasicontinuum (QC) method. We start with a choice of repatoms and the implied triangulation $\mathcal{T}$. The repatoms are then divided into the two sets $\mathcal{A}$ and $\mathcal{C}$. The set $\mathcal{A}$ comprises repatoms in the region where atomistic detail is needed. We assume there is no coarse-graining in this region, which means every atom is a repatom. The repatoms in $\mathcal{C}$ lie in the region of the body where the deformation gradient varies slowly, see Figure 1.2. The QC energy $\mathcal{E}^{\mathrm{qc}}$ is defined as

$$\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}_h) = \sum_{i\in\mathcal{A}} \mathcal{E}_i(\boldsymbol{y}_h) + \sum_{j\in\mathcal{C}} \omega_j \mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y}_h).$$

This way, the individual atomistic energies $\mathcal{E}_i$ only have to be calculated for the small subset $\mathcal{A}$, whereas the computational efficiency introduced by the Cauchy–Born approximation makes $\mathcal{E}^{\mathrm{qc}}$ a tractable approximation of $\mathcal{E}$.

Ghost Forces. Depending on the smoothness properties of $\boldsymbol{y}_h$ the energy $\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}_h)$ is a satisfactory approximation of $\mathcal{E}(\boldsymbol{y}_h)$. However, the coupling results in unwanted features, the so-called *ghost forces*. These are the result of an asymmetry in the coupling and the nonlocal nature of the atomic interaction.

The energy $\mathcal{E}_i(\boldsymbol{y}_h)$ of an atom $i \in \mathcal{A}$ is the sum of $V(|y_i - y_j|)$ over *all* atoms j. In particular, this means that $\mathcal{E}_i(\boldsymbol{y}_h)$ explicitly depends on the positions of the continuum repatoms. On the other hand $\mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y}_h)$ for $j \in \mathcal{C}$ is only determined by the positions of repatoms that are also vertices of the elements neighbouring repatom j (localization). More concisely: in general $D_{y_j}\mathcal{E}_i(\boldsymbol{y}_h) \neq 0$ for all $i \in \mathcal{A}$, $j \in \mathcal{C}$. However, there are pairs $i \in \mathcal{A}$, $j \in \mathcal{C}$ such that $D_{y_i}\mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y}_h) = 0$.

If $\boldsymbol{y}_h$ describes a homogeneous deformation there are no physical forces on the atoms due to symmetry, i.e., $D\mathcal{E}(\boldsymbol{y}_h) = 0$. However, as an artefact of the coupling nonphysical forces on atoms arise near the interface between the atomistic part $\mathcal{A}$ and the continuum part $\mathcal{C}$: $D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}_h) \neq 0$. This clearly is a major problem.

A popular solution to the problem of ghost forces is the introduction of dead loads [91,108]. Given a configuration $\boldsymbol{y}_h$, the ghost forces $\boldsymbol{f}_G$ are computed and subtracted from the total energy $E_{\boldsymbol{f}}$ as dead loads $\boldsymbol{f}_G \cdot \boldsymbol{y}_h$. Thus, during a minimization process the ghost forces have to be calculated and the energy function has to be updated repeatedly. Other approaches are based on the introduction of interface atoms, whose interactions are redefined in a way that prevents ghost forces [50, 110].

For important practical aspects like the choice of repatoms and meshing we refer to the QC literature [91]. This article also gives an overview of applications of QC methods.

A 1D Example. To illustrate the construction of the QC energy and the origin of ghost forces we look at a particularly simple model problem in one space dimension. We consider an infinite chain of atoms with positions $\boldsymbol{y} = (y_i)_{i\in\mathbb{Z}}$. To keep things simple we assume that each atom i only interacts with its nearest and next nearest neighbours $\{i-2, i-1, i+1, i+2\}$. Moreover, we do not apply coarse-graining and hence every atom is a repatom. The triangulation $\mathcal{T}$ therefore consists of the elements (y_{i-1}, y_i), for $i \in \mathbb{Z}$.

The atomistic energy takes the form[2]

$$\mathcal{E}(\boldsymbol{y}) = \frac{1}{2}\sum_{i\in\mathbb{Z}} \mathcal{E}_i(\boldsymbol{y}) = \frac{1}{2}\sum_{i\in\mathbb{Z}} \big(V(y_i - y_{i-2}) + V(y_i - y_{i-1}) + V(y_{i+1} - y_i) + V(y_{i+2} - y_i)\big).$$

[2]Strictly speaking this energy is not well-defined because the chain is infinite. However, we are only interested in the area near the interface so the chain may simply be thought of as finite but sufficiently long.

Next we derive the Cauchy–Born energy $\mathcal{E}_i^{\mathrm{cb}}(\boldsymbol{y})$. The atom i located at y_i acts as a node of the elements (y_{i-1}, y_i) and (y_i, y_{i+1}). In the Cauchy–Born approximation the energies of both elements are computed as if they were part of infinite equidistant chains with atomic distances $(y_i - y_{i-1})$, respectively, $(y_{i+1} - y_i)$. We deduce that the Cauchy–Born energy of atom i is

$$\mathcal{E}_i^{\mathrm{cb}}(\boldsymbol{y}) = \frac{1}{2}\big[V\big(2(y_i - y_{i-1})\big) + V\big(y_i - y_{i-1}\big) + V\big(y_{i+1} - y_i\big) + V\big(2(y_{i+1} - y_i)\big)\big].$$

Let now $\mathcal{A} = \{-K, \dots, K\}$ for some $K \in \mathbb{N}$ and $\mathcal{C} = \mathbb{Z}\backslash\mathcal{A}$. Then, we define the QC energy $\mathcal{E}^{\mathrm{qc}}$ as

$$\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) = \sum_{i\in\mathcal{A}} \mathcal{E}_i(\boldsymbol{y}) + \sum_{j\in\mathcal{C}} \mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y}).$$

If we assume that there are no external forces, i.e., $\boldsymbol{f} = 0$, then an equilibrium is given by any infinite equidistant chain $\boldsymbol{y}$ defined by $y_i = i\Delta y$ for all $i \in \mathbb{Z}$. An elementary computation shows that $D_{y_i}\mathcal{E}(\boldsymbol{y}) = 0$ for all $i \in \mathbb{Z}$ and also $D_{y_i}\mathcal{E}^{\mathrm{cb}}(\boldsymbol{y}) = 0$ for all $i \in \mathbb{Z}$. However,

$$\begin{aligned} D_{y_{K-1}}\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) = D_{y_{K+2}}\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) &= V'(2\Delta y)/2, \\ D_{y_K}\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) = D_{y_{K+1}}\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) &= -V'(2\Delta y)/2. \end{aligned}$$

These are the ghost forces. We point out that the QC energy in this case is exact: $\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) = \mathcal{E}(\boldsymbol{y})$ but the energy can be lowered by breaking the homogeneity of the chain near the interface.

Alternative Methods An alternative version of the QC method is based directly on forces rather than the energy. In force-based QC, forces on repatoms in $\mathcal{A}$ are calculated as if there were no continuum region, whereas forces on repatoms in $\mathcal{C}$ are calculated as if the whole body were a continuum. By construction this method does not exhibit ghost forces but the resulting forces are not derived from an energy, which can be a disadvantage in certain circumstances.

The QC method is only one example of a larger number of models based on atomistic to continuum coupling. An overview and comparison of several methods is provided in the review articles [92] and [38]. The main differences lie in the choice of continuum model and the treatment of the interface. Some methods introduce overlap regions [4,5,56,122] to make the transition between atomistic and continuum description less abrupt. The issue of ghost forces is present in all methods.

We briefly recapitulate the structure of the remainder of this thesis. In Chapter 2 we analyze discretizations of a version of the density functional (1.5). Chapter 3 is devoted to the study of the effects of numerical integration and interpolation on these discretizations. In Chapter 4 we present and analyze Quasicontinuum Methods for an interaction involving a field. We point out that extensive literature reviews are given in the individual chapters.

Chapter 2

Galerkin Discretization of an Electronic Density Functional

This chapter is devoted to the analysis of a Galerkin discretization of a density functional derived from the Thomas–Fermi–Dirac–von Weizsäcker model (1.5). The Euler–Lagrange equations for the Thomas–Fermi–Dirac–von Weizsäcker functional can be rewritten as a non-monotone, semilinear elliptic system with a nonlinear constraint. In this chapter, we combine arguments used for linear saddle-point problems and linearization techniques based on the Inverse Function Theorem, to establish the existence and convergence of the sequence of solutions of a Galerkin discretization of this system.

In Section 2.1 we review theoretical and numerical work on Thomas–Fermi type functionals and orbital-free DFT. In Section 2.2 we first formulate a model in a bounded domain that allows the simulation of finite clusters or isolated defects in an infinite medium. We prove existence of minimizers and give a careful analysis of first- and second-order optimality conditions in the remainder of Section 2.2. Our reformulation of the optimality system is particularly suitable for subsequent Galerkin finite element discretizations. The main result in Section 2.2.3 connects the stability of the minimization problem to the stability of this optimality system. This result allows us, in Section 2.3, to prove the existence and convergence of Galerkin discretizations of the optimality system. The case of periodic boundary conditions will be addressed in Section 2.4. In Section 2.5 we take a closer look at convergence rates. At the end of the chapter we analyze the dependence of the minimization problem on the coordinates of the atoms; see Section 2.6.

We view the present work as a preliminary step in the development of a theory for coarse-graining the TFDW functional in the spirit of [59] and [61].

The content of the present chapter is of, predominantly, theoretical nature. The important practical issues of numerical integration will be addressed in the following chapter. All of the results discussed in the present chapter carry over to the formulation *with numerical integration*. Computational examples will also be shown in Chapter 3.

2.1 Literature Review

A survey of theoretical results in connection with Thomas–Fermi type models can be found in an article by Lieb [80]. The author considers different models of increasing complexity in $\mathbb{R}^3$ and assesses their mathematical structure and physical validity as well as their relations to quantum mechanics. An analysis of the Thomas–Fermi–von Weizsäcker functional (without the exchange correlation part) is carried out in [9]. The introduction of the Dirac term or, more generally, the exchange correlation, renders the functional nonconvex with respect to the density ρ, which makes it difficult to analyze over $\mathbb{R}^3$; see the discussion in Sections VI and VIII in [80]. Physical questions addressed include molecular binding, behaviour of ρ near nuclei, the limit $Z_i \to \infty$ and dependence on the nucleus positions. The Thomas–Fermi model (functional (1.5) without the gradient term and exchange correlation) is shown to be convex and so a unique minimizer can be found. The situation changes if exchange correlation is introduced since the resulting functional lacks convexity. The Thomas–Fermi–von Weizsäcker functional, which is obtained by adding the $\nabla\rho$ term to the Thomas–Fermi functional, is also analyzed in [9]. It can be shown to have a unique minimizer and to reproduce physical properties better than the simple Thomas–Fermi model. It should be mentioned here that the analysis of the functionals on bounded domains in $\mathbb{R}^3$ is significantly easier. In the present work the functionals will be exclusively analyzed on bounded domains.

Several numerical approximations of the functional (1.5) or related models have been proposed in the literature. We will distinguish between finite difference methods and Galerkin discretizations.

One numerical approach to Thomas–Fermi type functionals is suggested in [2, 94, 95]. The functional considered does not include a $\nabla\rho$ term. The discretization is based on a regular cartesian grid and convolution kernels are explicitly represented by matrices. Calculations of the electronic structure are only performed in certain subcells of the domain. The density in adjacent regions is subsequently obtained by interpolation. This procedure is referred to as electronic density reconstruction. Regarding the nuclei, a similar idea is invoked. A reduced set of so-called *representative nuclei* is chosen. Only the positions of these nuclei are actual degrees of freedom. All remaining coordinates are calculated by interpolation. The authors also introduce a way of adaptively refining the mesh close to nuclei, which improves the efficiency and accuracy of the method.

A density functional including a convolution term in the kinetic energy as well as pseudo-potentials for the electron nucleus interaction is considered in [57]. Instead of the electron density, its square-root $u = \sqrt{\rho}$ is the unknown, which naturally takes care of the constraint $\rho \geq 0$. In [58] the author completes the proof of existence of a minimizer over a bounded domain $\Omega \subset \mathbb{R}^3$. The Euler–Lagrange equation of the minimization problem is analyzed briefly. Subsequently, the minimization problem is discretized using a finite difference scheme.

Hence, convolutions can be computed efficiently by the Fast Fourier Transform. For the energy minimization the author applies a truncated Newton method adapted to equality constrained optimization. The Hessian and the gradient are replaced with projected versions that take into account the constraint $\int u^2 \, \mathrm{d}x = 1$. In particular, every iterate is chosen to satisfy this constraint. The article does not include convergence results

For Galerkin discretizations, the crucial question is the choice of basis. The most popular basis sets are plane waves (i.e. Fourier modes) and finite elements. Plane waves can only be applied to periodic systems, which means, for example, that no defects can be simulated (usually, periodic arrays of defects are considered instead). On the other hand, the implementation of the Coulomb interaction kernel can be done very efficiently. Finite elements are not inherently periodic and allow for adaptivity in space, which is particularly useful for additional coarse-graining approximations. Calculating the convolution in the electrostatic terms remains a challenge.

An implementation of an orbital-free kinetic energy density model, more general than (1.5), using plane waves is described in [120] or more recently [67]. The main difference to (1.5) is the enhancement of the kinetic energy by a density-dependent convolution term motivated by linear-response theory, see also [82, 117]. Furthermore, the described model uses pseudopotentials instead of the Coulomb potential of the nuclei, which accounts for the inaccurate form of ρ close to nuclei in simpler models. Loosely speaking, the electrons are all treated as valence electrons, for which the kinetic energy can be reproduced more easily than for core electrons. The method explained in [120] uses a real-space and a Fourier-space representation at the same time, which are transformed into each other using the Fast Fourier Transform. Convolutions therefore do not pose a problem. These methods have recently been used to simulate up to one million atoms [69]. Applications in material science can be found in [65, 66]. There, periodic computations of density functionals on unit cells are used to evaluate continuum properties of a solid and to predict the breakdown of the continuum description in certain situations.

A finite element approximation of the functional (1.5) on a bounded domain $\Omega \subset \mathbb{R}^3$ is described in the papers [61,62]. In order to overcome the difficulty of calculating a convolution, the authors introduce the electrostatic potential ϕ as an additional variable. This means that a Poisson equation is coupled to the minimization problem, which then takes a saddle point form. As in [57], the square root density $u = \sqrt{\rho}$ is the unknown of interest. Using standard methods of the calculus of variations, the authors show existence of a minimizer for a slightly more general class of functionals. The minimization problem is discretized by a P1-Galerkin finite element method. The sequence of approximations is then shown to Γ-converge to the continuous functional even if numerical quadrature is applied. The convergence result does not include convergence rates, which are, however, crucial for understanding the efficiency of

the numerical method.

To solve the overall minimization problem, an alternating procedure is chosen: for a given set of nucleus positions the electron wave function u is relaxed using the nonlinear conjugate gradient method, which is followed by an update of the electrostatic potential. The authors pay special attention to the choice and evolution of triangulations in calculations with dynamical nuclei. Since the electron density is expected to be localized near nuclei, the initial grid is refined there. As the nuclei move, the mesh is convected with them, which in turn ensures good starting values for the u relaxation for updated R. If the mesh quality deteriorates, local remeshing takes place. In [62], the authors also present ideas how pseudopotentials and kinetic energy functionals with convolution terms can be implemented within their finite element framework. An extension of these results to the Kohn–Sham functional is discussed in [113].

In the second work [61] the model is enhanced by including ideas from the Quasicontinuum method. Instead of treating all nuclei as free particles, a set of representative atoms is chosen, which also defines a triangulation of the atomic lattice. The positions of the remaining atoms are reconstructed using linear interpolation. For the u and ϕ computations, two nested finite element meshes are introduced. The finer one has subatomic resolution everywhere and is used to compute the periodic parts u_p and ϕ_p by calculations on unit cells of the atomic lattice; these parts, called predictors, are supposed to be accurate away from defects. To account for local variations in the electronic structure due to defects, corrector fields u_c, ϕ_c are computed on the coarser finite element mesh. This mesh only needs to be subatomic close to defects and can be chosen coarser where the deformation varies slowly. In this approach the full flexibility of the finite element method shows. The use of a plane wave basis does not allow this kind of nonperiodic simulation in conjunction with ideas from the nonlocal quasicontinuum method.

Because of the normalization constraint $\int \rho \,\mathrm{d}x = n_{\mathrm{el}}$, the minimization problem with energy (1.5) can be interpreted as a nonlinear eigenvalue problem for $u = \sqrt{\rho}$:

$$\text{minimize } E_{\mathrm{TFDW}}(u^2), \quad \text{subject to } \int_\Omega u^2 \,\mathrm{d}x = N.$$

Although the discretization of linear eigenvalue problems has been studied in great detail (see, e.g., [3]), relatively little work exists on the nonlinear situation.

Zhou [126] studied the finite element discretization of the functional $\int_\Omega (|\nabla u|^2 + Vu^2 + \beta u^4)\,\mathrm{d}x$ subject to the constraint $\int_\Omega u^2 = 1$. Here, $\beta > 0$, $V \geq 0$ is an external potential and the domain Ω is not assumed to be bounded. The problem has a unique minimizer $\bar{u}$ (up to the sign) and the optimality conditions for this problem are given by the so-called Gross-Pitaevskii equation. The author considers a Galerkin discretization using finite-dimensional approximation spaces X_n. Convergence of the disrete minimizers $\bar{u}_h \in X_h$ is shown using

continuity properties of the energy and approximation properties of X_n. Convergence rates for $\|\bar{u}-\bar{u}_h\|_{\mathrm{H}^1}$ are not obtained. However, the bound $\|\bar{u}-\bar{u}_h\|_{\mathrm{H}^1} \leq C\big(\|\bar{u}-\bar{u}_h\|_{\mathrm{L}^2}+\inf_{v_h\in X_h}\|\bar{u}-v_h\|_{\mathrm{H}^1}\big)$ is shown using a careful rearrangement of the optimality equations. Assuming that the L^2-error converges faster than the H^1-error this implies a quasi-optimal convergence rate.

In [127] the same author generalizes these results to the Thomas–Fermi–von Weizsäcker functional. The techniques are very similar to [126], however, additional complexity comes from the nonlocal nature of the electrostatic terms and the lack of convexity of the functional if formulated in terms of $u = \sqrt{\rho}$. The author deals with the lack of convexity by looking at the functional for both ρ and u. Since there is a unique minimizing ρ, the only minimizing square root densities are $\pm\sqrt{\rho}$.

Recently, these results were further generalized to a class of functionals including convolution terms like (1.6) and the classical Coulomb interaction [33]. Since the uniqueness of minimizers is not clear, the authors work with the sets of continuous and discrete global minimizers U, respectively, U_h. Continuity properties of the energy are used to show that the H^1-distance $\sup_{u_h\in U_h}\inf_{u\in U}\|u-u_h\|$ goes to zero as $h\to 0$. Moreover, rearrangements of the optimality systems lead to a bound on the H^1-distance of U and U_h in terms of the L^2-distance and the H^1-distance between U and the approximation spaces. The article concludes with numerical examples indicating that optimal convergence rates are indeed observed. In [32], the authors show convergence of an adaptive finite element discretization for a class of not necessarily convex nonlinear eigenvalue problems.

Another rigorous study of Galerkin discretizations of a class of nonlinear eigenvalue problems is carried out in [21]. The energy functional under consideration is given by

$$E(u) = \frac{1}{2}\int_\Omega \nabla u^T A \nabla u \,\mathrm{d}x + \frac{1}{2}\int_\Omega V u^2 \,\mathrm{d}x + \frac{1}{2}\int_\Omega F(u^2)\,\mathrm{d}x\,,$$

where A is a matrix-valued function, V is an external potential and F is a convex, nonlinear function. The domain Ω is either a bounded domain in $\mathbb{R}^d$ or the unit cell of a periodic lattice in $\mathbb{R}^d$. Under certain conditions on F and V there exist a minimizer $\bar{u}$ that is unique up to its sign and a Lagrange multiplier $\bar{\mu}$ such that $DE(\bar{u}) + \bar{\mu}\bar{u} = 0$. After proving convergence for a generic Galerkin discretization, both a Fourier and finite element discretization are discussed in detail. The proof of convergence is based on the observation that the bilinear form $D^2E(\bar{u}) - \bar{\mu}$ is elliptic. Optimal convergence rates for the energy and the Lagrange multiplier are shown. For the latter the authors use a rearrangement of the optimality equations and a dual problem. The effect of numerical integration is discussed in the one-dimensional case.

In [22] the same authors rigorously analyze a Fourier discretization of the Thomas–Fermi–von Weizsäcker functional. With similar techniques as in [21] convergence of a Galerkin Fourier discretization is shown and precise convergence rates are given. Moreover, a careful analysis of the errors introduced by interpolation is performed. For this, the authors work

with two different grids: one for the discretization space and a finer grid for numerical integration. The second part of [22] deals with the Kohn–Sham functional with periodic boundary conditions.

In our present work we aim to provide a complete convergence theory for Galerkin discretizations of Thomas–Fermi-type electronic density functionals including the electrostatic potential as an additional unknown.

2.2 Existence and Analysis of Minimizers

In this section we will suggest a mathematical model based on the functional described in the introduction and study the resulting minimization problem as well as the associated optimality conditions. First we discuss a few ideas that simplify the TFDW functional with regard to subsequent finite element approximation; see also [62].

For numerical reasons the nuclear point charges Z_i at the positions R_i may be replaced with a smooth charge density ρ_n (see Gavini et al. [62, Eq. (9)]). For example, ρ_n may be written as a sum of compactly supported smooth functions centered at the positions R_i, $i = 1, \dots, n_{\mathrm{at}}$, (the dependence of ρ_n on R will be suppressed in our notation)

$$\rho_n(x) = \sum_{i=1}^{n_{\mathrm{at}}} Z_i \widetilde{\rho}_0(x - R_i),$$

where $\widetilde{\rho}_0 \in \mathrm{C}_0^\infty(\mathbb{R}^3)$, $\widetilde{\rho}_0 \geq 0$, and $\int \widetilde{\rho}_0(x)\,\mathrm{d}x = 1$. Then, the repulsion energy of the nuclei takes the form

$$E_{\mathrm{nn}}(R) = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_n(x)\rho_n(y)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y.$$

Note that this expression includes the potential energy of every nucleus in its own electrostatic field. This is, however, only a constant contribution to the overall energy and may easily be subtracted. The smoothed potential for the nucleus-electron interaction is given by

$$V_{\mathrm{en}}(x) = -\int_{\mathbb{R}^3} \frac{\rho_n(y)}{|x-y|}\,\mathrm{d}y.$$

This allows for a symmetric expression for the sum of all electrostatic terms

$$E_{\mathrm{ee}}(\rho) + E_{\mathrm{en}}(\rho, R) + E_{\mathrm{nn}}(R) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\big(\rho(x) - \rho_n(x)\big)\big(\rho(y) - \rho_n(y)\big)}{|x-y|}\,\mathrm{d}y\,\mathrm{d}x\,.$$

The nonlocal nature of this term represents a numerical challenge for all density functional calculations.

The constraint $\rho \geq 0$ can be enforced by setting $\rho = u^2$. This substitution also has the

advantage that the term involving $\nabla\rho$ becomes easier to evaluate:

$$E(u,R) = \frac{\lambda}{2}\int_{\mathbb{R}^3} |\nabla u|^2 \,\mathrm{d}x + C_{\mathrm{TF}}\int_{\mathbb{R}^3} |u|^{10/3}\,\mathrm{d}x - C_{\mathrm{x}}\int_{\mathbb{R}^3} |u|^{8/3}\,\mathrm{d}x \\ + \int_{\mathbb{R}^3} \varepsilon_{\mathrm{c}}(u^2)u^2\,\mathrm{d}x - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\big(u^2(x)-\rho_n(x)\big)\big(u^2(y)-\rho_n(y)\big)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y. \tag{2.1}$$

Note that we do not include a convolution term in the kinetic energy. The reason for this is that some of these functionals are mathematically not well posed or badly understood [11,20]. Moreover, these terms are difficult to approximate in a finite element context. One method was suggested in [35] and used in [61,62]. The authors expand the convolution kernel in Fourier space and include the terms into the energy via elliptic minimization problems similar to our treatment of the Coulomb energy below.

We now need to address the evaluation of the electrostatic term. The double integral can in principle be computed explicitly using the Fourier transform [57, 59, 120]. In [61, 62] the authors suggest a different approach that makes use of the special structure, respectively the physics, represented by the term. Note that the integral kernel in the last term in (2.1) is the Green's function of the Poisson equation in $\mathbb{R}^3$. Therefore, the electrostatic potential

$$\phi(x) := \int_{\mathbb{R}^3} \frac{u^2(y)-\rho_n(y)}{|x-y|}\,\mathrm{d}y$$

is simply the solution of the equation

$$-\frac{1}{4\pi}\Delta\phi = u^2 - \rho_n,$$

subject to a homogeneous Dirichlet boundary condition at infinity. From this, it can be deduced that

$$\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\big(u^2(x)-\rho_n(x)\big)\big(u^2(y)-\rho_n(y)\big)}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y = \frac{1}{2}\int_{\mathbb{R}^3}(u^2-\rho_n)\phi\,\mathrm{d}x\,.$$

Formally, the right-hand side can also be written as

$$\frac{1}{2}\int_{\mathbb{R}^3}(u^2-\rho_n)\phi\,\mathrm{d}x = -\inf_{\varphi}\left[\int_{\mathbb{R}^3}\frac{1}{8\pi}|\nabla\varphi|^2 - (u^2-\rho_n)\varphi\,\mathrm{d}x\right]. \tag{2.2}$$

This equality is referred to as the direct Coulomb formula in [62].

2.2.1 Artificial Boundary Conditions

We now introduce a model to simulate finite clusters of atoms or localized phenomena in infinite crystals (e.g. vacancies, interstitials or dislocation loops) in d dimensions. For both computational and analytical reasons, we would like to formulate the problem in a bounded open domain $\Omega \subset \mathbb{R}^d$ rather than in the whole of $\mathbb{R}^d$. To this end, we will assume that we know the square root density u as well as the electrostatic potential ϕ in $\mathbb{R}^d \setminus \Omega$, which induces

artificial Dirichlet boundary conditions for u and ϕ on $\partial\Omega$. We discuss potential pitfalls of this approach in Remark 2.1 below.

There are two specific examples that we have in mind. If a finite cluster of atoms is studied, then we set $u = \phi = 0$ in $\mathbb{R}^d \setminus \Omega$. If we study a localized defect in an infinite crystal (e.g., a vacancy or a dislocation), then we let u and ϕ on $\partial\Omega$ be the square root density and electrostatic potential of a perfect crystal. (Minimizers of Thomas–Fermi type functionals for perfect crystals have been studied in [14] and [26]).

To make this concrete, we assume that we are given functions $u_{\rm ex}, \phi_{\rm ex} \in \mathrm{H}^2_{\rm loc}(\mathbb{R}^d)$ and define the admissible set

$$\mathrm{A}_u = \left\{ u \in u_{\rm ex} + \mathrm{H}^1_0(\Omega) : \|u\|^2_{\mathrm{L}^2} = n_{\rm el} \right\}, \tag{2.3}$$

and the energy functional (suppressing the nuclei positions, which are held fixed)

$$E(u) = T(u) + X(u) + \Phi(u), \tag{2.4}$$

where

$$\begin{aligned} T(u) &= \frac{\lambda}{2} \|\nabla u\|^2_{\mathrm{L}^2(\Omega)}, \\ X(u) &= \int_\Omega F(u)\,\mathrm{d}x\,, \qquad \text{and} \\ \Phi(u) &= -\inf_{\phi \in \phi_{\rm ex} + \mathrm{H}^1_0(\Omega)} \Psi(u,\phi) \\ &= -\inf_{\phi \in \phi_{\rm ex} + \mathrm{H}^1_0(\Omega)} \left[\frac{1}{8\pi} \int_\Omega |\nabla\phi|^2\,\mathrm{d}x \;-\; \int_\Omega (u^2 - \rho_n)\phi\,\mathrm{d}x \right]. \end{aligned} \tag{2.5}$$

We have split the energy functional, in a way that is convenient for the analysis, into a quadratic, a nonlinear local, and a nonlocal part. In the original Thomas–Fermi–Dirac–von Weizsäcker model the function F is given by

$$F(u) = C_{\rm TF}|u|^{10/3} - C_{\rm x}|u|^{8/3} + \varepsilon_{\rm c}(u^2)u^2 \tag{2.6}$$

and combines a portion of the kinetic energy and the exchange-correlation energy. However, our results are independent of the precise form of F and only require a certain degree of smoothness and growth conditions, which we make precise below.

It is also worth remarking that in the case of homogeneous boundary conditions on u and ϕ ($u_{\rm ex} = \phi_{\rm ex} = 0$), the energy Φ reduces to $\Phi(u) = \frac{1}{2}\int_\Omega (u^2 - \rho_n)\phi\,\mathrm{d}x$, where ϕ is the weak solution of the equation $-\Delta\phi = 4\pi(u^2 - \rho_n)$ with homogeneous Dirichlet boundary condition. In this case Φ is easily recognizable as the potential energy of the charge density $(u^2 - \rho_n)$ in its own electrostatic field.

For future reference, we also define the constraint functional $c : \mathrm{H}^1(\Omega) \to \mathbb{R}$,

$$c(u) = \frac{1}{2}\left(\|u\|^2_{\mathrm{L}^2} - n_{\rm el} \right).$$

Our goal is to solve the minimization problem

$$\min_{u \in \mathrm{A}_u} E(u). \tag{2.7}$$

Remark 2.1. Our justification for this choice of boundary conditions is that, if $\bar{u}$ is a minimizer of the original problem, with associated electrostatic potential $\bar{\phi}$, then, setting $u_{\mathrm{ex}} = \bar{u}$ and $\phi_{\mathrm{ex}} = \bar{\phi}$, $\bar{u}|_\Omega$ is a solution to (2.7) with associated electrostatic potential $\bar{\phi}|_\Omega$. In that sense, the problem with artificial boundary conditions is *consistent* with the original problem posed in $\mathbb{R}^d$.

Since we do not normally know the exact electron density and electrostatic potential outside of Ω, we are essentially forced to make our 'best guess' for the problem at hand. This creates an error in the system, which cannot be controlled by adjusting the discretization. One usually hopes that, by choosing domains of increasing size, this error will shrink to zero, however, it is far from straightforward to establish this rigorously for any system other than a (near-)perfect crystal.

For simplicity, let us dicuss the case of a finite cluster where we set $u_{\mathrm{ex}} = \phi_{\mathrm{ex}} = 0$ in $\mathbb{R}^d \setminus \Omega$. Hence, the question arises, if $(\bar{u}, \bar{\phi})$ are the *exact* square root density and electrostatic potential of the system, how fast $|\bar{u}(x)|$ and $|\bar{\phi}(x)|$ decay as $|x| \to \infty$.

Whereas quantum mechanics suggests that the decay of u is exponential, this is less clear for ϕ. Note, in particular, that variations of u, well inside Ω, can in general create comparatively large variations of ϕ in all of $\mathbb{R}^d$. For rather special configurations of the nuclei, we may hope that lower order multipole moments vanish, assuring a sufficiently fast decay of ϕ (e.g., if there is a point symmetry).

In the present work we will simply assume that these deviations decay sufficiently fast and concentrate on aspects of numerical approximation theory. We are planning to study the issues pointed out in this remark in future work. In particular, we consider the present work as a first step towards an analysis of a combined TFDW / Quasicontinuum model in the spirit of [59] and [61]. For this type of analysis it will also be necessary to study deviations of u and ϕ from the perfect crystal case introduced by localized defects. □

2.2.2 Existence of a Minimizer

Before turning to the analysis of the minimization problem (2.7), we briefly introduce a classical notion of differentiability in Banach spaces that will be important throughout the thesis.

Definition 2.2. Let Y, Z be Banach spaces. Let $y \in Y$ and $U \subset Y$ be an open neighbourhood of y. A map $\Phi : Y \to Z$ is called *Fréchet differentiable in* y if there exists a linear operator $T \in \mathrm{Lin}(Y, Z)$ such that

$$\frac{\|\Phi(y+h) - \Phi(y) - Th\|_Z}{\|h\|_Y} \to 0 \quad \text{as } \|h\|_Y \to 0.$$

If T exists, it is called *Fréchet derivative* of Φ in y and will also be denoted by $D\Phi(y)$. A function $\Phi : Y \to Z$ is called *continuously differentiable* in an open set $U \subset Y$ if $D\Phi(y)$ exists for all $y \in U$ and $D\Phi : y \mapsto D\Phi(y)$ is a continuous mapping from U to $\mathrm{Lin}(Y, Z)$.

Let Y_1, Y_2 be Banach spaces and $\Phi : Y_1 \times Y_2 \to Z$. Let y_2 be fixed and define $\Psi : Y_1 \to Z$ through $\Psi(y_1) := \Phi(y_1, y_2)$. If Ψ has a Fréchet derivative in y_1, then the *partial derivative of* Φ *with respect to* y_1 at (y_1, y_2) is defined as $D_{y_1}\Phi(y_1, y_2) = D\Psi(y_1)$. □

Higher order derivatives are defined analogously. In particular, the second derivative $D^2\Phi(y)$ of Φ in y is a linear operator in $\mathrm{Lin}(Y, \mathrm{Lin}(Y, Z))$. As commonly done, we will use the abbreviation $D^2\Phi(y) \cdot [\eta_1, \eta_2] = (D^2\Phi(y) \cdot \eta_1) \cdot \eta_2$ for $\eta_1, \eta_2 \in Y$. If Φ is an operator from $Y_1 \times Y_2 \times Y_3$ to Z, the partial derivative with respect to the two components $i, j \in \{1, 2, 3\}$ will be written as $D_{(y_i, y_j)}\Phi(y_1, y_2, y_3) \in \mathrm{Lin}(Y_i \times Y_j, Z)$. Analogous notation is used for higher order derivatives. If $\Phi : Y \to \mathbb{R}$ is a real-valued mapping, the derivatives canonically define symmetric multilinear forms. For example, $D^2\Phi(y)$ defines a continuous symmetric bilinear form on $Y \times Y$.

From now on we assume that the homogeneous Dirichlet problem is H^2-regular, that is, if $f \in \mathrm{L}^2(\Omega)$ and if $v \in \mathrm{H}^1_0(\Omega)$ is the solution of

$$(\nabla v, \nabla w) = (f, w) \qquad \forall w \in \mathrm{H}^1_0(\Omega), \tag{2.8}$$

then $v \in \mathrm{H}^2(\Omega)$, and there exists a constant C_{reg}, independent of f, such that

$$|v|_{\mathrm{H}^2} \le C_{\mathrm{reg}} \|f\|_{\mathrm{L}^2}.$$

This is the case, for example, if Ω is convex [64] or if Ω is C^2-regular [51, 6.3.2]. For notational convenience we define the solution operator of the Poisson equation with homogeneous Dirichlet boundary condition: $(-\Delta_0)^{-1} : \mathrm{L}^2(\Omega) \to \mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega)$, $f \mapsto v$, where v solves (2.8).

Throughout the analysis, the positions R_i of the nuclei are fixed. We assume that $F \in \mathrm{C}^2(\mathbb{R})$ and that it satisfies the growth condition

$$a_1 \le F(t) \le c_2 t^{q_F} + a_2 \qquad \forall t \in \mathbb{R}, \tag{2.9}$$

with constants a_1, $a_2 \in \mathbb{R}$, $c_2 \ge 0$, and $3 \le q_F < 6$. Moreover, we assume that F'' is locally Hölder continuous; more precisely, there exists a positive constant C, such that, for all $t_1, t_2 \in \mathbb{R}$,

$$|F''(t_1) - F''(t_2)| \le C\big(|t_1 - t_2|^{\alpha_F} + (1 + |t_1|^{q_F - 3} + |t_2|^{q_F - 3})|t_1 - t_2|\big), \tag{2.10}$$

where $0 < \alpha_F \le 1$. We note that these conditions are satisfied if F is defined by (2.6).

The functional $T : \mathrm{H}^1(\Omega) \to \mathbb{R}$ is strongly continuous, twice continuously Fréchet differentiable and weakly lower semicontinuous.

Lemma 2.3. *Under condition (2.10), the functional $X : \mathrm{H}^1(\Omega) \to \mathbb{R}$ defined by (2.5) is twice continuously Fréchet differentiable with*

$$DX(u)\cdot h = \int_\Omega F'(u)h \,\mathrm{d}x\,, \qquad \text{and} \qquad D^2X(u)\cdot[h_1,h_2] = \int_\Omega F''(u)h_1h_2 \,\mathrm{d}x\,,$$

for h, h_1, $h_2 \in \mathrm{H}^1(\Omega)$. Furthermore, X and DX are locally Lipschitz continuous, and D^2X is locally Hölder continuous, with

$$\begin{aligned} \big\|D^2X(u) - D^2X(v)\big\| &\le C\big(\|u-v\|_{\mathrm{L}^2}^{\alpha_F} + (1+\|u\|_{\mathrm{H}^1}^{q_F-3} + \|v\|_{\mathrm{H}^1}^{q_F-3})\,\|u-v\|_{\mathrm{H}^1}\big) \qquad (2.11)\\ &\le C\big(1+\|u\|_{\mathrm{H}^1}^{q_F-3} + \|v\|_{\mathrm{H}^1}^{q_F-3}\big)\big(\|u-v\|_{\mathrm{H}^1}^{\alpha_F} + \|u-v\|_{\mathrm{H}^1}\big), \qquad (2.12)\end{aligned}$$

for all $u, v \in \mathrm{H}^1(\Omega)$. X is also sequentially weakly continuous on $\mathrm{H}^1(\Omega)$.

Proof. From the assumptions it can be deduced that $|F(u)| \le c_0(1+|u|^{q_F})$ for all $u \in \mathbb{R}$ with a $c_0 > 0$, so X is well-defined. Continuity is established alongside differentiability by the following calculation. Let $u, h \in \mathrm{L}^{q_F}(\Omega)$. Using elementary calculus, we know that

$$\int_\Omega \big(F(u+h) - F(u) - F'(u)h\big)\,\mathrm{d}x \;= \int_\Omega \int_0^1 \big(F'(u+th) - F'(u)\big)\;\mathrm{d}t\; h \,\mathrm{d}x\,.$$

From the assumption (2.10) on F'' it can also be deduced that

$$|F'(u) - F'(v)| \le c_1(1+|u|^{q_F-2} + |v|^{q_F-2})|u-v| \quad \forall u, v \in \mathbb{R}$$

for a $c_1 > 0$. Consequently,

$$\begin{aligned}\int_\Omega |F(u+h) - F(u) - F'(u)h| \,\mathrm{d}x \;&\le C\int_\Omega (1+|u|^{q_F-2} + |h|^{q_F-2})h^2 \,\mathrm{d}x\\ &\le C(u)\,\|h\|_{\mathrm{L}^{q_F}}^2 + C\,\|h\|_{\mathrm{L}^{q_F}}^{q_F}\,.\end{aligned}$$

This shows differentiability of X. Existence of the second derivative follows similarly.

The stated continuity property of D^2X is shown by a similar calculation using the assumptions on F'' and $\||u|^{\alpha_F}\|_{\mathrm{L}^k} \le C\,\|u\|_{\mathrm{L}^k}^{\alpha_F}$. To prove this, we begin as follows: for $u, v, h_1, h_2 \in \mathrm{H}^1(\Omega)$ we have that

$$\begin{aligned}\big|(D^2X(u) - D^2X(v))\cdot[h_1,h_2]\big| \le\; &\int_\Omega |F''(u) - F''(v)||h_1||h_2| \,\mathrm{d}x \qquad (2.13)\\ &\le C\int_\Omega |h_1||h_2|\big(|u-v|^{\alpha_F} + (1+|u|^{q_F-3}+|v|^{q_F-3})|u-v|\big)\,\mathrm{d}x\\ &\le C\,\|h_1\|_{\mathrm{L}^4}\,\|h_2\|_{\mathrm{L}^4}\,\||u-v|^{\alpha_F}\|_{\mathrm{L}^2} + C\,\|h_1\|_{\mathrm{L}^6}\,\|h_2\|_{\mathrm{L}^6}\\ &\qquad\cdot\big(1 + \big\||u|^{q_F-3}\big\|_{\mathrm{L}^2} + \big\||v|^{q_F-3}\big\|_{\mathrm{L}^2}\big)\,\|u-v\|_{\mathrm{L}^6}\,,\end{aligned}$$

where we have used a generalized version of Hölder's inequality. Now we know that $\||u|^\gamma\|_{\mathrm{L}^k} \le C\,\|u\|_{\mathrm{L}^k}^\gamma$ for all $0 \le \gamma < 1$ and all $k > 1$:

$$\||u|^\gamma\|_{\mathrm{L}^k}^k = \int_\Omega 1\cdot|u|^{\gamma k}\,\mathrm{d}x \;\le C\big\||u|^{\gamma k}\big\|_{\mathrm{L}^{1/\gamma}} = C\,\|u\|_{\mathrm{L}^k}^{\gamma k}\,. \qquad (2.14)$$

A very similar calculation leads to $\big\||u|^{q_F-3}\big\|_{\mathrm{L}^2} \le C\,\|u\|_{\mathrm{L}^6}^{q_F-3}$ for $3 < q_F < 6$. Applying these inequalities to (2.13) and using the embedding of $\mathrm{H}^1(\Omega)$ in $\mathrm{L}^p(\Omega)$ for $p \le 6$ we deduce (2.12).

Since X is strongly continuous in $\mathrm{L}^{q_F}(\Omega)$ and $\mathrm{H}^1(\Omega)$ is compactly embedded in $\mathrm{L}^p(\Omega)$, for $1 \le p < 6$, see [1, Th 6.3], it follows also that X is sequentially weakly continuous on $\mathrm{H}^1(\Omega)$, which concludes the proof. □

Next we consider the functional Φ defined in (2.5).

Lemma 2.4. *The functional $\Phi : \mathrm{L}^4(\Omega) \to \mathbb{R}$ as defined in (2.5) is well-defined and continuous; Φ is twice continuously Fréchet differentiable and the derivatives are, for all $u \in \mathrm{L}^4(\Omega)$, given by*

$$\begin{aligned} D\Phi(u)\cdot h &= 2\int_\Omega u\phi h \,\mathrm{d}x \quad \forall h \in \mathrm{L}^4(\Omega) \text{ and,} \\ D^2\Phi(u)\cdot[h_1,h_2] &= 2\int_\Omega \big(\phi h_1 h_2 + 8\pi u h_1 (-\Delta_0)^{-1}(u h_2)\big)\,\mathrm{d}x \quad \forall h_1, h_2 \in \mathrm{L}^4(\Omega), \end{aligned}$$

where $\phi \in \mathrm{H}^1(\Omega)$ is the weak solution of $-\Delta\phi = 4\pi(u^2 - \rho_n)$, $\phi|_{\partial\Omega} = \phi_{\mathrm{ex}}$. Furthermore, Φ is bounded below on the set A_u defined in (2.3), and the restriction $\Phi|_{\mathrm{H}^1(\Omega)}$ is sequentially weakly continuous in $\mathrm{H}^1(\Omega)$.

Proof. For every $u \in \mathrm{L}^4(\Omega)$, the functional $\Psi(u,\cdot)$ from (2.5) is clearly continuous, convex, coercive and weakly lower semicontinuous on $\{\psi \in \mathrm{H}^1(\Omega) : \psi|_{\partial\Omega} = \phi_{\mathrm{ex}}\}$. Thus, there exists a unique minimizer ϕ_u. This minimizer satisfies the equation

$$(\nabla\phi_u, \nabla\psi) = 4\pi(u^2 - \rho_n, \psi) \qquad \forall \psi \in \mathrm{H}^1_0(\Omega),$$

with boundary condition $\phi_u|_{\partial\Omega} = \phi_{\mathrm{ex}}$. The auxiliary function

$$\xi = \phi_{\mathrm{ex}} - (-\Delta_0)^{-1}(-\Delta)\phi_{\mathrm{ex}} \ \in \phi_{\mathrm{ex}} + \mathrm{H}^1_0(\Omega)$$

satisfies $(\nabla\xi, \nabla\psi) = 0$ for all $\psi \in \mathrm{H}^1_0(\Omega)$. From this, it follows that

$$\phi_u = 4\pi(-\Delta_0)^{-1}(u^2 - \rho_n) + \xi \tag{2.15}$$

and, moreover, after straightforward algebraic manipulations, that

$$\begin{aligned} \Phi(u) = -\Psi(u,\phi_u) = 2\pi \int_\Omega (u^2 - \rho_n)(-\Delta_0)^{-1}(u^2 - \rho_n)\,\mathrm{d}x \\ + \int_\Omega (u^2 - \rho_n)\xi \,\mathrm{d}x - \frac{1}{8\pi}\int_\Omega |\nabla\xi|^2 \,\mathrm{d}x\,. \end{aligned} \tag{2.16}$$

Differentiating with respect to u yields the expressions for $D\Phi$ and $D^2\Phi$ as given above.

Finally, we show that $\Phi|_{\mathrm{A}_u}$ is bounded below. Clearly, the first term on the right-hand side in (2.16) is nonnegative, and the last term is a constant depending only on ϕ_{ex}. Since

we assumed that $\phi_{\text{ex}} \in \mathrm{H}^2_{\text{loc}}(\mathbb{R}^d)$ and that the Poisson problem is H^2-regular, it follows that $\xi \in \mathrm{L}^\infty(\Omega)$. Therefore, the second term on the right-hand side of (2.16) can be bounded as follows:

$$\left| \int_\Omega (u^2 - \rho_n)\xi \,\mathrm{d}x \right| \leq \|\xi\|_{\mathrm{L}^\infty} \left(\|u\|^2_{\mathrm{L}^2} + \|\rho_n\|_{\mathrm{L}^1}\right) = \|\xi\|_{\mathrm{L}^\infty} \left(n_{\text{el}} + \|\rho_n\|_{\mathrm{L}^1}\right).$$

Hence, we can deduce that $\Phi(u) \geq C(n_{\text{el}}, \rho_n, \phi_{\text{ex}})$ for all $u \in \mathrm{A}_u$.

The sequential weak continuity of $\Phi|_{\mathrm{H}^1(\Omega)}$ is a direct consequence of the compact embedding of $\mathrm{H}^1(\Omega)$ in $\mathrm{L}^4(\Omega)$ (see [1, Th 6.3]) and the strong continuity of Φ in $\mathrm{L}^4(\Omega)$. □

Theorem 2.5. *The functional $E : \mathrm{A}_u \to \mathbb{R}$ defined in (2.4) has at least one minimizer.*

Proof. We apply the direct method of the calculus of variations [39]. First, we observe that E is coercive on A_u, which can be seen as follows:

$$\begin{aligned} E(u) &\geq \frac{\lambda}{2} \|\nabla u\|^2_{\mathrm{L}^2} + \int_\Omega a_1 \,\mathrm{d}x + \Phi(u) \\ &\geq \frac{\lambda}{2} \|\nabla u\|^2_{\mathrm{L}^2} + a_1|\Omega| - C(n_{\text{el}}, \rho_n, \phi_{\text{ex}}) \\ &\geq C_1 \|u\|^2_{\mathrm{H}^1} - C_2 \|u_{\text{ex}}\|^2_{\mathrm{H}^1} - C_3(\Omega, F, n_{\text{el}}, \rho_n, \phi_{\text{ex}}). \end{aligned}$$

Here we have used the growth condition (2.9) on F, the lower bound on $\Phi(u)$, established in Lemma 2.4, and Poincaré's inequality for $u - u_{\text{ex}}$. Hence, minimizing sequences of E are bounded in $\mathrm{H}^1(\Omega)$, and we can find a weakly convergent subsequence. Since E is weakly lower semicontinuous as a sum of weakly lower semicontinuous functions, it follows that the weak limit of the subsequence is a minimizer; see for example [39, Th 3.30]. □

Remark 2.6. We mention at this point that, since u was defined to be the square root of ρ, the minimizer in the case of homogeneous boundary conditions cannot be unique: if $\bar{u}$ locally minimizes E, then so does $-\bar{u}$. The question whether there can be more than two minimizers of the energy (2.4) is beyond the scope of this thesis.

If the function $t \mapsto F(\sqrt{t})$ is convex, then the minimizer is unique up to the sign. This is well-known for TFDW type functionals, see [80]. The functional $\rho \mapsto T(\sqrt{\rho})$ is strictly convex (Theorem 7.1 in [80]) and so is $\rho \mapsto \Phi(\sqrt{\rho})$. Then, the minimization problem

$$\min\left\{ E(\sqrt{\rho}) : \ \rho \geq 0, \ \sqrt{\rho} \in \mathrm{H}^1(\Omega), \ \rho|_{\partial\Omega} = u^2_{\text{ex}}, \ \int_\Omega \rho \,\mathrm{d}x \ = 1 \right\}$$

has a unique solution since the objective function is strictly convex and the admissible set is convex. □

2.2.3 The Euler–Lagrange Equations

So far, we have shown existence of solutions to the minimization problem (2.7). Next, we are interested in the characterization of such points, i.e., in optimality conditions. For example, the article [89] provides necessary and sufficient optimality conditions for a large class of optimization problems in Banach spaces.

Throughout this section, the derivatives $DE(\bar{u})$ and $Dc(\bar{u})$ are understood as elements of $\mathrm{H}^{-1}(\Omega) := \mathrm{H}_0^1(\Omega)^*$, that is, they operate on functions from $\mathrm{H}_0^1(\Omega)$. This means, for example, that $\ker Dc(\bar{u}) \subset \mathrm{H}_0^1(\Omega)$.

Theorem 3.1 in [89] yields the first-order necessary optimality condition

$$DE(\bar{u})\cdot v = 0 \quad \forall v \in \ker Dc(\bar{u}) \subset \mathrm{H}_0^1(\Omega),$$

provided that $0 \in \operatorname{int}\left\{Dc(\bar{u}) \cdot v : v \in \mathrm{H}_0^1(\Omega)\right\} \subseteq \mathbb{R}$. Since $Dc(\bar{u}) \cdot v = (\bar{u}, v)$ and $\bar{u} \neq 0$, this condition is always satisfied. Theorem 3.2 in [89] yields the existence of a *Lagrange multiplier* $\bar{\mu} \in \mathbb{R}$ such that

$$DE(\bar{u}) + \bar{\mu}\, Dc(\bar{u}) = 0 \in \mathrm{H}^{-1}(\Omega), \quad \text{and} \quad c(\bar{u}) = 0. \tag{2.17}$$

We mention that as the Lagrange multiplier for the normalization constraint $\int_\Omega u^2 \,\mathrm{d}x = n_{\mathrm{el}}$, $\bar{\mu}$ is interpreted as a chemical potential, which means that it is the derivative of the energy with respect to the number of electrons:

$$\bar{\mu} = D_N E(\bar{u}).$$

This can be proved with the methods we will use in Section 2.6 to analyze the dependence of E on the coordinates R of the nuclei.

Using the definitions and results of the previous section we now rewrite the first-order optimality system in a way that is more suitable for numerical approximation. Let $\bar{u} \in \mathrm{A}_u$ be a local minimizer with associated Lagrange multiplier $\bar{\mu} \in \mathbb{R}$ and let $\bar{\phi} \in \mathrm{H}^1(\Omega)$ be the associated electrostatic potential, then Lemma 2.4 implies that $\bar{u}$, $\bar{\phi}$, and $\bar{\mu}$ solve the nonlinear system

$$\begin{aligned} \lambda(\nabla u, \nabla v) + (F'(u), v) + 2(\phi u, v) + \mu(u, v) = 0 \qquad & \forall v \in \mathrm{H}_0^1(\Omega), \\ \frac{1}{4\pi}(\nabla\phi, \nabla\psi) - (u^2 - \rho_n, \psi) = 0 \qquad & \forall \psi \in \mathrm{H}_0^1(\Omega), \\ \frac{\nu}{2}\left(\int_\Omega u^2 \,\mathrm{d}x - n_{\mathrm{el}}\right) = 0 \qquad & \forall \nu \in \mathbb{R}, \end{aligned} \tag{2.18}$$

with the boundary conditions

$$u|_{\partial\Omega} = u_{\mathrm{ex}}, \quad \text{and} \qquad \phi|_{\partial\Omega} = \phi_{\mathrm{ex}}.$$

We will focus on solving this system instead of (2.17) or the minimization problem (2.7). It has to be pointed out that solving (2.18) (or even (2.17)) is not equivalent to solving the

minimization problem since the functional is nonconvex. However, we will focus on those solutions of (2.18) which correspond to local minimizers of (2.7), making use of the second-order optimality condition (2.22).

We define the function spaces

$$\begin{aligned} \mathcal{Y} &= \mathrm{H}^1(\Omega)\times\mathrm{H}^1(\Omega)\times\mathbb{R}, \qquad & \mathcal{Y}_D &= (u_{\mathrm{ex}} + \mathrm{H}^1_0(\Omega))\times(\phi_{\mathrm{ex}} + \mathrm{H}^1_0(\Omega))\times\mathbb{R}, \\ \mathcal{Y}_0 &= \mathrm{H}^1_0(\Omega)\times\mathrm{H}^1_0(\Omega)\times\mathbb{R}, \qquad & \mathcal{Y}_0^* &= \mathrm{H}^{-1}(\Omega)\times\mathrm{H}^{-1}(\Omega)\times\mathbb{R}, \end{aligned}$$

so that the system (2.18) defines an operator $\mathcal{F} : \mathcal{Y}_D \to \mathcal{Y}_0^*$, rewritten as

$$\big\langle \mathcal{F}(u,\phi,\mu), (v,\psi,\nu)\big\rangle = 0 \quad \forall (v,\psi,\nu) \in \mathcal{Y}_0. \tag{2.19}$$

Here, the Laplacian Δ is understood as a linear map from $\mathrm{H}^1(\Omega)$ to $\mathrm{H}^{-1}(\Omega)$ in the following way: $\langle -\Delta u, v\rangle = (\nabla u, \nabla v)$ for all $v \in \mathrm{H}^1_0(\Omega)$.

It can be shown easily that $\mathcal{F}$ is Fréchet differentiable with derivative $D\mathcal{F}(u,\phi,\mu) : \mathcal{Y}_0 \to \mathcal{Y}_0^*$,

$$D\mathcal{F}(u,\phi,\mu)\cdot(v,\psi,\nu) = \begin{pmatrix} -\lambda\Delta v + (F''(u) + 2\phi + \mu)v + 2u\psi + \nu u \\ -\frac{1}{4\pi}\Delta\psi - 2uv \\ (u,v) \end{pmatrix} \quad \forall (v,\psi,\nu) \in \mathcal{Y}_0,$$

and that $D\mathcal{F}$ is locally Hölder continuous in the following sense: there exists a continuous function $L_{D\mathcal{F}} : \mathbb{R} \to \mathbb{R}$ and $\alpha_F \in (0,1)$ such that

$$\|D\mathcal{F}(y_1) - D\mathcal{F}(y_2)\| \le L_{D\mathcal{F}}(\|y_1\|_{\mathcal{Y}} + \|y_2\|_{\mathcal{Y}})\,(\|y_1 - y_2\|_{\mathcal{Y}}^{\alpha_F} + \|y_1 - y_2\|_{\mathcal{Y}}) \tag{2.20}$$

for all $y_1, y_2 \in \mathcal{Y}$. This follows immediately from (2.10). A direct consequence of the differentiability is that $\mathcal{F}$ is locally Lipschitz continuous: there exists a continuous function $L_{\mathcal{F}} : \mathbb{R} \to \mathbb{R}$ such that

$$\|\mathcal{F}(y_1) - \mathcal{F}(y_2)\|_{\mathcal{Y}_0^*} \le L_{\mathcal{F}}(\|y_1\|_{\mathcal{Y}} + \|y_2\|_{\mathcal{Y}})\,\|y_1 - y_2\|_{\mathcal{Y}}, \tag{2.21}$$

for all $y_1, y_2 \in \mathcal{Y}$.

At this point, we make some observations concerning the regularity of solutions to (2.18). The functions $\bar{u}$ and $\bar{\phi}$ solve elliptic equations of the type

$$-\lambda\Delta\bar{u} = g_1, \quad \text{and} \quad -\frac{1}{4\pi}\Delta\bar{\phi} = g_2,$$

subject to the boundary conditions $\bar{u}|_{\partial\Omega} = u_{\mathrm{ex}}$, $\bar{\phi}|_{\partial\Omega} = \phi_{\mathrm{ex}}$, respectively. Here, the functions g_1 and g_2 obviously depend on $\bar{u}$, $\bar{\phi}$ and $\bar{\mu}$. From the embedding $\mathrm{H}^1(\Omega) \subset \mathrm{L}^4(\Omega)$ we know that $g_2 \in \mathrm{L}^2(\Omega)$. Since we assumed that the Poisson problem (2.8) is H^2-regular, we can deduce that $\bar{\phi} \in \mathrm{H}^2(\Omega)$. For $\bar{u}$ to be in $\mathrm{H}^2(\Omega)$ we need to tighten the growth assumption on F. If $d = 3$ and $q_F = 4$, we have $|F'(\bar{u})| \le C(1 + |\bar{u}|^3)$ and hence $F'(\bar{u}) \in \mathrm{L}^2(\Omega)$, from which we

deduce $\bar{u} \in \mathrm{H}^2(\Omega)$. If $d \leq 2$, any polynomial growth bound on F and F' implies $\bar{u} \in \mathrm{H}^2(\Omega)$. We will from now on assume that

$$q_F = 4 \ \text{ if } \ d = 3, \quad q_F < \infty \ \text{ if } \ d < 3.$$

For $d = 3$ the Sobolev embedding theorem [1, Th. 4.12] states that $\mathrm{H}^2(\Omega) \subset \mathrm{C}^{0,\gamma}(\bar{\Omega})$ for all $0 < \gamma \leq 1/2$. Hence, $\bar{u}, \bar{\phi} \in \mathrm{C}^{0,\gamma}(\bar{\Omega})$ for every $\gamma \leq 1/2$, and in particular $\bar{u}$, $\bar{\phi} \in \mathrm{L}^\infty(\Omega)$. If $d < 3$, we get $\bar{u}, \bar{\phi} \in \mathrm{C}^{0,\gamma}(\bar{\Omega})$ for every $\gamma \leq 1$.

2.2.4 Second-order Optimality Conditions

The functional $\mathcal{L} : \mathrm{H}^1(\Omega) \times \mathbb{R} \to \mathbb{R}$, defined by $\mathcal{L}(u, \mu) = E(u) + \mu c(u)$, is called a *Lagrangian*. Since both E and c are twice continuously Fréchet differentiable, the same holds for $\mathcal{L}$.

From Theorem 3.3 in [89] we deduce that, if $\bar{u}$ is a solution of (2.7), and $\bar{\mu}$ is its associated Lagrange multiplier, then the necessary second-order optimality condition

$$D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}) \cdot [v, v] \geq 0 \qquad \forall v \in \ker Dc(\bar{u})$$

holds. Conversely, if $(\bar{u}, \bar{\mu})$ satisfies (2.17) as well as the sufficient second-order optimality condition

$$D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}) \cdot [v, v] \geq \gamma \, \|\nabla v\|^2_{\mathrm{L}^2} \qquad \forall v \in \ker Dc(\bar{u}), \tag{2.22}$$

for some constant $\gamma > 0$, then $\bar{u}$ is an isolated local minimizer of E in A_u, see [89, Th. 5.6]. We call a critical point $\bar{u} \in \mathrm{A}_u$ that satisfies (2.22) a *uniform minimizer* of (2.7).

Written out explicitly, (2.22) reads

$$\lambda(\nabla v, \nabla v) + \big((F''(\bar{u}) + 2\bar{\phi} + \bar{\mu})v, v\big) + 16\pi\big(\bar{u}v, (-\Delta_0)^{-1}\bar{u}v\big) \geq \gamma \, \|\nabla v\|^2_{\mathrm{L}^2}, \tag{2.23}$$

for all $v \in \ker Dc(\bar{u})$, where $\bar{\phi}$ is the electrostatic potential associated with $\bar{u}$.

The next step is to prove that, if $\bar{u}$ is a uniform local minimizer with associated electrostatic potential $\bar{\phi}$ and Lagrange multiplier $\bar{\mu}$, then $D\mathcal{F}(\bar{u}, \bar{\phi}, \bar{\mu})$ is an isomorphism. This will be an important tool for the convergence analysis in the next section.

Proposition 2.7. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_D$ such that (2.22) holds with $\gamma > 0$. Then, $D\mathcal{F}(\bar{y}) : \mathcal{Y}_0 \to \mathcal{Y}_0^*$ is an isomorphism.*

Proof. We need to show that the equation

$$D\mathcal{F}(\bar{u}, \bar{\phi}, \bar{\mu}) \cdot (v, \psi, \nu) = (f, g, \kappa) \tag{2.24}$$

is uniquely solvable in $\mathcal{Y}_0$ for every $(f, g, \kappa) \in \mathcal{Y}_0^*$. To this end we define two bilinear forms $a_{\bar{y}} : \mathrm{H}^1_0(\Omega)^2 \times \mathrm{H}^1_0(\Omega)^2 \to \mathbb{R}$ and $b_{\bar{y}} : \mathrm{H}^1_0(\Omega)^2 \times \mathbb{R} \to \mathbb{R}$,

$$\begin{aligned} a_{\bar{y}}\big((v, \psi), (w, \chi)\big) &= \lambda(\nabla v, \nabla w) + \big((F''(\bar{u}) + 2\bar{\phi} + \bar{\mu})v, w\big) + (2\bar{u}\psi, w) \\ &\quad + \frac{1}{4\pi}(\nabla\psi, \nabla\chi) - 2(\bar{u}v, \chi), \\ b_{\bar{y}}\big((v, \psi), \eta\big) &= (\bar{u}, v)\eta. \end{aligned}$$

Then, equation (2.24) takes the form of a saddle-point problem,

$$\begin{aligned} a_{\bar{y}}\big((v,\psi),(w,\chi)\big) + b_{\bar{y}}\big((w,\chi),\nu\big) &= \langle f, w\rangle + \langle g, \chi\rangle \qquad \forall w, \chi \in \mathrm{H}^1_0(\Omega), \\ b_{\bar{y}}\big((v,\psi),\eta\big) &= \eta\kappa \qquad\qquad\qquad\quad \forall \eta \in \mathbb{R}. \end{aligned} \tag{2.25}$$

The bilinear forms $a_{\bar{y}}$ and $b_{\bar{y}}$ are continuous on $\mathrm{H}^1_0(\Omega)^2 \times \mathrm{H}^1_0(\Omega)^2$ and $\mathrm{H}^1_0(\Omega)^2 \times \mathbb{R}$, respectively. We define

$$\begin{aligned} \ker b_{\bar{y}} &:= \big\{(v,\psi) \in \mathrm{H}^1_0(\Omega)^2 : b_{\bar{y}}((v,\psi),\eta) = 0 \quad \forall \eta \in \mathbb{R}\big\} \\ &= \big\{(v,\psi) \in \mathrm{H}^1_0(\Omega)^2 : v \in \ker Dc(\bar{u})\big\}. \end{aligned}$$

For a saddle-point problem such as (2.25) there are well-known sufficient conditions for solvability; see Theorem 1.1 in [18]. The bilinear form $b_{\bar{y}}$ has to satisfy an inf-sup condition of the form

$$\inf_{\nu\in\mathbb{R}} \sup_{(v,\psi)\in \mathrm{H}^1_0(\Omega)^2} \frac{b_{\bar{y}}((v,\psi),\nu)}{|\nu| \ \|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}} \geq \kappa_b > 0,$$

and the linear operator associated with $a_{\bar{y}}$ has to be invertible on $\ker b_{\bar{y}}$.

Step 1. Inf-sup condition for $b_{\bar{y}}$. Since $\|\bar{u}\|^2_{\mathrm{L}^2} = n_{\mathrm{el}}$, we have $\bar{u} \neq 0$, and therefore $b_{\bar{y}}$ obeys an inf-sup condition on $\mathrm{H}^1_0(\Omega)^2 \times \mathbb{R}$:

$$\begin{aligned} \inf_{\nu\in\mathbb{R}} \sup_{(v,\psi)\in \mathrm{H}^1_0(\Omega)^2} \frac{b_{\bar{y}}((v,\psi),\nu)}{|\nu| \ \|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}} &= \inf_{\nu\in\mathbb{R}} \sup_{(v,\psi)\in \mathrm{H}^1_0(\Omega)^2} \frac{\nu \int_\Omega \bar{u} v \,\mathrm{d}x}{|\nu| \ \|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}} \\ &\geq \frac{(\bar{u}, (-\Delta_0)^{-1}\bar{u})}{(\bar{u}, (-\Delta_0)^{-1}\bar{u})^{1/2}} =: \kappa_b > 0, \end{aligned}$$

where for a given $\nu \neq 0$ we have chosen $v = \mathrm{sign}(\nu)(-\Delta_0)^{-1}\bar{u}$, and $\psi = 0$.

Step 2. Invertibility of $a_{\bar{y}}$ on $\ker b_{\bar{y}}$. The proof of the unique solvability of the variational problem: *find $(v,\psi) \in \ker b_{\bar{y}}$ such that*

$$a_{\bar{y}}\big((v,\psi),(w,\chi)\big) = \langle f, w\rangle + \langle g, \chi\rangle \quad \forall (w,\chi) \in \ker b_{\bar{y}}, \tag{2.26}$$

where f, $g \in \mathrm{H}^{-1}(\Omega)$, requires more work and really relies on the assumption that $\bar{u}$ is a uniform minimizer of E.

First we show that solutions are unique. For $f = g = 0$ we want to prove that the only possible solution is $(v,\psi) = (0,0)$. Looking at the definition of $a_{\bar{y}}$ we see that $g = 0$ leads to $\psi = 8\pi(-\Delta_0)^{-1}\bar{u}v$. Substituting this into (2.26) and testing with $w = v$ and $\chi = 0$ we obtain

$$\lambda(\nabla v, \nabla v) + \big((F''(\bar{u}) + 2\bar{\phi} + \bar{\mu})v, v\big) + 16\pi\big(\bar{u}v, (-\Delta_0)^{-1}\bar{u}v\big) = 0. \tag{2.27}$$

Since $\bar{u}$ is assumed to be a uniform minimizer, we know from (2.23) that the bilinear form on the left-hand side of (2.27) is coercive on $\{w \in \mathrm{H}^1_0(\Omega) : (w,\bar{u}) = 0\}$, so we get $v = 0$ and hence also $\psi = 0$.

Next we prove that for every $f,\, g \in \mathrm{H}^{-1}(\Omega)$ there exists a solution. Let $v \in \ker Dc(\bar{u})$ be the unique solution of

$$\begin{aligned}\lambda(\nabla v, \nabla w) + \big((F''(\bar{u}) + 2\bar{\phi} + \bar{\mu})v, w\big) + 16\pi\big(\bar{u}w, (-\Delta_0)^{-1}\bar{u}v\big) \\ = \langle f, w\rangle - \big(4\pi\bar{u}(-\Delta_0)^{-1}g, w\big) \qquad \forall w \in \ker Dc(\bar{u}).\end{aligned}$$

This equation is uniquely solvable by the Lax–Milgram theorem [17, Th. 2.7.7], again since the bilinear form (as a function of v and w) on the left-hand side is coercive. Let

$$\psi = 4\pi(-\Delta_0)^{-1}(2\bar{u}v + g).$$

Combining the last two equations shows that v, ψ indeed solve equation (2.26). Thus, we have shown unique solvability in $\ker b_{\bar{y}}$ for every $f, g \in \mathrm{H}^{-1}(\Omega)$.

From Theorem 1.1 in [18] we can now deduce that $D\mathcal{F}(\bar{y})$ is indeed an isomorphism from $\mathcal{Y}_0$ to $\mathcal{Y}_0^*$. □

Remark 2.8. For future reference, we mention that the invertibility of $a_{\bar{y}}$ on $\ker b_{\bar{y}}$ shown in the proof is equivalent to the existence of a constant $\kappa_a > 0$ such that

$$\begin{aligned}\inf_{(v,\psi)\in\ker b_{\bar{y}}} \sup_{(w,\chi)\in\ker b_{\bar{y}}} \frac{a_{\bar{y}}\big((v,\psi),(w,\chi)\big)}{\|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}\,\|(\nabla w, \nabla\chi)\|_{\mathrm{L}^2}} \geq \kappa_a, \quad \text{and} \\ \inf_{(w,\chi)\in\ker b_{\bar{y}}} \sup_{(v,\psi)\in\ker b_{\bar{y}}} \frac{a_{\bar{y}}\big((v,\psi),(w,\chi)\big)}{\|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}\,\|(\nabla w, \nabla\chi)\|_{\mathrm{L}^2}} \geq \kappa_a.\end{aligned} \tag{2.28}$$

For this equivalence see for example [18, Proposition 1.2] or the discussion following Remark 1.6 in [18]. Similar conditions hold for $D\mathcal{F}(\bar{y})$:

$$\inf_{\eta_1\in\mathcal{Y}_0} \sup_{\eta_2\in\mathcal{Y}_0} \frac{\langle D\mathcal{F}(\bar{y})\cdot\eta_1, \eta_2\rangle}{\|\eta_1\|_{\mathcal{Y}}\|\eta_2\|_{\mathcal{Y}}} \geq \kappa_{\mathcal{F}}, \quad \inf_{\eta_1\in\mathcal{Y}_0} \sup_{\eta_2\in\mathcal{Y}_0} \frac{\langle D\mathcal{F}(\bar{y})\cdot\eta_2, \eta_1\rangle}{\|\eta_1\|_{\mathcal{Y}}\|\eta_2\|_{\mathcal{Y}}} \geq \kappa_{\mathcal{F}} \tag{2.29}$$

for some $\kappa_{\mathcal{F}} > 0$. □

2.3 Galerkin Discretization

In this section, we propose a discretization of the minimization problem (2.7), which corresponds to a Galerkin discretization of the optimality system (2.18). We will show that, for sufficiently small values of the discretization parameter, the discretized problem has a solution and that as the discretization parameter tends to zero a sequence of numerical solutions converges to the continuous solution. Optimal convergence rates will be addressed in later sections.

2.3.1 The Discretized Functional

Let $(\mathrm{S}_h)_{h\in(0,1]}$ be a family of finite-dimensional subspaces of $\mathrm{H}^1(\Omega)$ with the approximation property

$$\inf_{v\in\mathrm{S}_h} \|u-v\|_{\mathrm{H}^1} \le Ch|u|_{\mathrm{H}^2} \qquad \forall\, u\in \mathrm{H}^2(\Omega), \tag{2.30}$$

and define $\mathrm{S}_{h,0} = \mathrm{S}_h \cap \mathrm{H}^1_0(\Omega)$. Let $\mathcal{I}_h : \mathrm{H}^2(\Omega) \to \mathrm{S}_h$ be an interpolation operator with

$$\|\phi - \mathcal{I}_h\phi\|_{\mathrm{H}^1} \le Ch|\phi|_{\mathrm{H}^2} \quad \text{for all } \phi\in \mathrm{H}^2(\Omega). \tag{2.31}$$

Moreover, we define

$$u_{\mathrm{ex},h} = \mathcal{I}_h u_{\mathrm{ex}}|_{\bar\Omega} \quad \text{and} \quad \phi_{\mathrm{ex},h} = \mathcal{I}_h\phi_{\mathrm{ex}}|_{\bar\Omega}.$$

We introduce an approximation of the energy functional (2.4) defined on S_h of the following form

$$E_h(u_h) = \frac{\lambda}{2}\int_\Omega |\nabla u_h|^2\,\mathrm{d}x \,+ \int_\Omega F(u_h)\,\mathrm{d}x \,+ \Phi_h(u_h), \tag{2.32}$$

where

$$\Phi_h(u_h) = -\inf_{\phi_h\in\phi_{\mathrm{ex},h}+\mathrm{S}_{h,0}} \left[\frac{1}{8\pi}\int_\Omega |\nabla\phi_h|^2\,\mathrm{d}x \,- \int_\Omega \phi_h(u_h^2-\rho_n)\,\mathrm{d}x\right].$$

Let

$$\mathrm{A}_{u,h} := \{u_h \in u_{\mathrm{ex},h} + \mathrm{S}_{h,0} : \|u_h\|^2_{\mathrm{L}^2} = n_{\mathrm{el}}\}$$

be the set of discrete admissible functions. We consider the discretized minimization problem

$$\min_{u_h\in\mathrm{A}_{u,h}} E_h(u_h). \tag{2.33}$$

Note that this does not represent a Galerkin discretization of (2.7) since the electrostatic term Φ was replaced by the approximation Φ_h.

As in the continuous case, we get the following optimality conditions: if $\bar u_h$ is a (local) minimizer of E_h in $\mathrm{A}_{u,h}$, then there exists a discrete electrostatic potential $\bar\phi_h \in \mathrm{S}_h$ and a Lagrange multiplier $\bar\mu_h \in \mathbb{R}$ such that

$$\begin{aligned}
\lambda(\nabla\bar u_h, \nabla v) + (F'(\bar u_h), v) + 2(\bar\phi_h\bar u_h, v) + \mu_h(\bar u_h, v) = 0 \quad &\forall v\in\mathrm{S}_{h,0},\\
\frac{1}{4\pi}(\nabla\bar\phi_h, \nabla\psi) - (\bar u_h^2 - \rho_n, \psi) = 0 \quad &\forall \psi\in\mathrm{S}_{h,0},\\
\frac{\nu}{2}\left(\int_\Omega \bar u_h^2\,\mathrm{d}x \,- n_{\mathrm{el}}\right) = 0 \quad &\forall \nu\in\mathbb{R},
\end{aligned} \tag{2.34}$$

where $\bar u_h$ and $\bar\phi_h$ satisfy the boundary conditions

$$\bar u_h|_{\partial\Omega} = u_{\mathrm{ex},h}, \qquad \bar\phi_h|_{\partial\Omega} = \phi_{\mathrm{ex},h}.$$

These discrete optimality conditions turn out to be the Galerkin discretization of the optimality system (2.18). We introduce the discrete function spaces

$$\mathcal{Y}_h = \mathrm{S}_h \times \mathrm{S}_h \times \mathbb{R}, \quad \mathcal{Y}_{h,D} = (u_{\mathrm{ex},h} + \mathrm{S}_{h,0}) \times (\phi_{\mathrm{ex},h} + \mathrm{S}_{h,0}) \times \mathbb{R},$$
$$\mathcal{Y}_{h,0} = \mathrm{S}_{h,0} \times \mathrm{S}_{h,0} \times \mathbb{R}, \quad \mathcal{Y}_{h,0}^* = \mathrm{S}_{h,0}^* \times \mathrm{S}_{h,0}^* \times \mathbb{R}.$$

In analogy to the continuous case we write the system (2.34) in the more compact form

$$\big\langle \mathcal{F}_h(\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h), (v, \psi, \nu) \big\rangle = 0 \qquad \forall (v, \psi, \nu) \in \mathcal{Y}_{h,0},$$

where $\mathcal{F}_h : \mathcal{Y}_{h,D} \to \mathcal{Y}_{h,0}^*$.

The discrete Laplacian $(-\Delta_h) : \mathrm{S}_h \to \mathrm{S}_{h,0}^*$ is defined by $\langle -\Delta_h v_h, w_h \rangle = (\nabla v_h, \nabla w_h)$ for $v_h \in \mathrm{S}_h$ and all $w_h \in \mathrm{S}_{h,0}$. The operator $(-\Delta_{h,0})^{-1} : \mathrm{S}_{h,0}^* \to \mathrm{S}_{h,0}$ maps f to the solution $\phi_h \in \mathrm{S}_{h,0}$ of $(\nabla \phi_h, \nabla v_h) = \langle f, v_h \rangle$ for all $v_h \in \mathrm{S}_{h,0}$.

Differentiability of $\mathcal{F}_h$ is easily shown. The derivative $D\mathcal{F}_h$ is again Hölder continuous and takes the form

$$D\mathcal{F}_h(u_h, \phi_h, \mu_h) \cdot (v, \psi, \nu) = \begin{pmatrix} -\lambda \Delta_h v + (F''(u_h) + 2\phi_h + \mu_h) v + 2u_h \psi + \nu u_h \\ -\frac{1}{4\pi} \Delta_h \psi - 2u_h v \\ (u_h, v) \end{pmatrix}, \tag{2.35}$$

for $(v, \psi, \nu) \in \mathcal{Y}_h$. Just as in the continuous case, this linear operator has saddle-point structure.

At this point we note that $D\mathcal{F}_h$ may in fact be extended to the whole of $\mathcal{Y}$. Slightly abusing notation, we will write $D\mathcal{F}_h(y) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$ for any $y \in \mathcal{Y}$, but we stress that this is still an operator between the discrete function spaces. The next result is the discrete counterpart of Proposition 2.7.

Proposition 2.9. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_D$ such that $\bar{u}$ satisfies (2.22) with $\gamma > 0$. Then, there exist $h_0 \in (0, 1]$ and $\delta > 0$ such that $D\mathcal{F}_h(y) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$ is an isomorphism for every $h \le h_0$ and for any $y \in B_\delta(\bar{y}) \subset \mathcal{Y}$. Moreover, there exists a constant $M > 0$ such that*

$$\|D\mathcal{F}_h(y)^{-1}\| \le M \qquad \forall y \in B_\delta(\bar{y}) \quad \forall h \le h_0. \tag{2.36}$$

Proof. We begin by showing invertibility of $D\mathcal{F}_h(\bar{y}) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$. To this end, we again interpret the problem in saddle-point form and prove an inf-sup inequality for $b_{\bar{y}}$ and the invertibility of $a_{\bar{y}}$ on $\mathrm{S}_{h,0}^2 \cap \ker b_{\bar{y}}$.

Step 1. Inf-sup condition for $b_{\bar{y}}$. We have

$$\begin{aligned} \inf_{\nu \in \mathbb{R}} \sup_{(v,\psi) \in \mathrm{S}_{h,0}^2} \frac{b_{\bar{y}}((v,\psi),\nu)}{|\nu| \, \|(\nabla v, \nabla \psi)\|_{\mathrm{L}^2}} &= \inf_{\nu \in \mathbb{R}} \sup_{(v,\psi) \in \mathrm{S}_{h,0}^2} \frac{\nu \int_\Omega \bar{u} v \, \mathrm{d}x}{|\nu| \, \|(\nabla v, \nabla \psi)\|_{\mathrm{L}^2}} \\ &\ge \frac{(\bar{u}, (-\Delta_{h,0})^{-1} \bar{u})}{\|\nabla (-\Delta_{h,0})^{-1} \bar{u}\|_{\mathrm{L}^2}} =: \kappa_{b,h}, \end{aligned}$$

where, for given $\nu \neq 0$, we have chosen $v = \operatorname{sign}(\nu)(-\Delta_{h,0})^{-1}\bar{u}$ and $\psi = 0$. If h is sufficiently small, then

$$\kappa_{b,h} = \frac{(\bar{u}, (-\Delta_{h,0})^{-1}\bar{u})}{\|\nabla(-\Delta_{h,0})^{-1}\bar{u}\|_{\mathrm{L}^2}} \geq \tfrac{1}{2}\left(\bar{u}, (-\Delta_0)^{-1}\bar{u}\right)^{1/2} = \kappa_b/2 > 0.$$

In particular, we deduce that the inf-sup constant $\kappa_{b,h}$ is bounded away from zero if h is sufficiently small.

Step 2. Considering now $a_{\bar{y}}$, we have to prove that the system

$$a_{\bar{y}}\big((v,\psi),(w,\chi)\big) = \langle f, w\rangle + \langle g, \chi\rangle \quad \forall (w,\chi) \in \ker b_{\bar{y}} \cap \mathrm{S}^2_{h,0} \tag{2.37}$$

has a unique solution $(v,\psi) \in \ker b_{\bar{y}} \cap \mathrm{S}^2_{h,0}$ for every $f,\ g \in \mathrm{H}^{-1}(\Omega)$. If the trial and test spaces were simply $\mathrm{S}^2_{h,0}$ instead of the constraint space $\ker b_{\bar{y}} \cap \mathrm{S}^2_{h,0}$, then this would follow from a classical argument by Schatz, see [106]. The present case requires some modifications, which we study in detail in the Appendix. Lemmas A.7 and A.6 provide the regularity and approximation results in $\ker b_{\bar{y}} \cap \mathrm{S}^2_{h,0}$ necessary for an application of the Schatz argument, which is carried out in Theorem A.4. Hence, we deduce that (2.37) is uniquely solvable, provided that h is small enough. Theorem A.4 also implies the existence of an inf-sup constant $\kappa_{a,h}$ for $a_{\bar{y}}$ on $K_h := \mathrm{S}^2_{h,0} \cap \ker b_{\bar{y}}$, similarly as in (2.28), in the continuous case:

$$\inf_{(v,\psi)\in K_h} \sup_{(w,\chi)\in K_h} \frac{a_{\bar{y}}\big((v,\psi),(w,\chi)\big)}{\|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}\,\|(\nabla w, \nabla\chi)\|_{\mathrm{L}^2}} \geq \kappa_{a,h} > 0, \quad \text{and}$$

$$\inf_{(w,\zeta)\in K_h} \sup_{(v,\chi)\in K_h} \frac{a_{\bar{y}}\big((v,\psi),(w,\chi)\big)}{\|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}\,\|(\nabla w, \nabla\chi)\|_{\mathrm{L}^2}} \geq \kappa_{a,h} > 0.$$

Furthermore, Theorem A.4 guarantees that $\kappa_{a,h}$ is bounded away from zero as $h \to 0$.

Step 3. We have shown that for sufficiently small h, $h \leq h_0$ say, $D\mathcal{F}_h(\bar{y})$ is an isomorphism. This also means that for every $h \leq h_0$, $D\mathcal{F}_h(\bar{y})$ satisfies the inf-sup conditions

$$\inf_{y_h\in\mathcal{Y}_{h,0}} \sup_{z_h\in\mathcal{Y}_{h,0}} \frac{\langle D\mathcal{F}_h(\bar{y})\cdot y_h, z_h\rangle}{\|y_h\|_{\mathcal{Y}}\|z_h\|_{\mathcal{Y}}} \geq \kappa_h \quad \text{and} \quad \inf_{z_h\in\mathcal{Y}_{h,0}} \sup_{y_h\in\mathcal{Y}_{h,0}} \frac{\langle D\mathcal{F}_h(\bar{y})\cdot y_h, z_h\rangle}{\|y_h\|_{\mathcal{Y}}\|z_h\|_{\mathcal{Y}}} \geq \kappa_h,$$

with $\kappa_h > 0$. Theorem 1.1 in [18] shows a way of bounding the inf-sup constant κ_h in terms of the inf-sup constants $\kappa_{a,h}$ for $a_{\bar{y}}$ and $\kappa_{b,h}$ for $b_{\bar{y}}$: let $f \in \mathcal{Y}^*_{h,0}$ and let $y_h \in \mathcal{Y}_{h,0}$ satisfy $\langle D\mathcal{F}_h(\bar{y})\cdot y_h, z_h\rangle = (f, z_h)$ for all $z_h \in \mathcal{Y}_{h,0}$. Then, Theorem 1.1 in [18] implies

$$\|y_h\|_{\mathcal{Y}} \leq \|D\mathcal{F}_h(\bar{y})^{-1}\|\|f\|_{\mathcal{Y}^*_{h,0}} \leq M(\kappa_{a,h},\kappa_{b,h})\|f\|_{\mathcal{Y}^*_{h,0}}.$$

The constant $M(\kappa_{a,h},\kappa_{b,h})$ is bounded uniformly in h since $\kappa_{a,h}$ and $\kappa_{b,h}$ are bounded away from zero. Thus,

$$\|y_h\|_{\mathcal{Y}} \leq M(\kappa_{a,h},\kappa_{b,h}) \sup_{z_h\in\mathcal{Y}_{h,0}} \frac{\langle f, z_h\rangle}{\|z_h\|_{\mathcal{Y}}} = M(\kappa_{a,h},\kappa_{b,h}) \sup_{z_h\in\mathcal{Y}_{h,0}} \frac{\langle D\mathcal{F}_h(\bar{y})\cdot y_h, z_h\rangle}{\|z_h\|_{\mathcal{Y}}}.$$

We deduce that $\kappa_h \geq M(\kappa_{a,h}, \kappa_{b,h})^{-1}$ and hence κ_h is bounded away from zero.

Since $D\mathcal{F}_h$ satisfies the discrete equivalent of the Hölder condition (2.20), it follows by Lemma A.2 that there exists a neighbourhood $B_\delta(\bar{y}) \subset \mathcal{Y}$, where $\delta > 0$ is chosen sufficiently small, such that, for $h \leq h_0$ and $y \in B_\delta(\bar{y})$, $D\mathcal{F}_h(y)$ is an isomorphism and such that the norm of $D\mathcal{F}_h(y)^{-1}$ is uniformly bounded. This implies (2.36). ◻

2.3.2 Existence and Convergence

The following convergence theorem constitutes the main result of this section. The proof uses ideas commonly used in the finite element literature on nonlinear problems; see for example [19] and [41].

Theorem 2.10. *Let $\bar{u}$ be a minimizer of (2.7) that satisfies (2.22). Let $\bar{\phi} \in \mathrm{H}^1(\Omega)$ and $\bar{\mu} \in \mathbb{R}$ be, respectively, the associated electrostatic potential and Lagrange multiplier. Then, there exist $h_0 \in (0, 1], \delta > 0$ such that, for all $h < h_0$, the discretized problem (2.34) has a unique solution $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,D}$ in the neighbourhood $B_\delta(\bar{y}) \subset \mathcal{Y}$. Furthermore, there exists a constant C such that*

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{H}^1} + \left\|\bar{\phi} - \bar{\phi}_h\right\|_{\mathrm{H}^1} + |\bar{\mu} - \bar{\mu}_h| \leq Ch.$$

Proof. The proof is divided into four steps.

Step 1. We begin by showing that for an approximation $\Pi_h \bar{y} \in \mathcal{Y}_{h,D}$ of $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu})$, we have

$$||\mathcal{F}_h(\Pi_h \bar{y})||_{\mathcal{Y}_{h,0}^*} \leq C_1 h$$

for sufficiently small h where C_1 is independent of h.

Let u_h, $\phi_h \in \mathrm{S}_h$ be the Ritz projections of $\bar{u}$ and $\bar{\phi}$, respectively, i.e., the solutions of the equations

$$(\nabla u_h, \nabla v) = (\nabla \bar{u}, \nabla v) \quad \forall v \in \mathrm{S}_{h,0} \quad \text{and} \quad (\nabla \phi_h, \nabla v) = (\nabla \bar{\phi}, \nabla v) \quad \forall v \in \mathrm{S}_{h,0},$$

with boundary conditions

$$u_h|_{\partial\Omega} = u_{\mathrm{ex},h}, \quad \phi_h|_{\partial\Omega} = \phi_{\mathrm{ex},h}.$$

In other words, $\Delta_h u_h = \Delta \bar{u}|_{\mathrm{S}_{h,0}}$ and $\Delta_h \phi_h = \Delta \bar{\phi}|_{\mathrm{S}_{h,0}}$. We define $\Pi_h \bar{y} = (u_h, \phi_h, \bar{\mu})$. Standard convergence theory for the Poisson equation, $\bar{u}$, $\bar{\phi} \in \mathrm{H}^2(\Omega)$, and the approximation property (2.31) of $\mathcal{I}_h$ then yield

$$\|\bar{y} - \Pi_h \bar{y}\|_{\mathcal{Y}} \leq Ch.$$

Using the fact that $\mathcal{F}(\bar{y})|_{\mathcal{Y}_{h,0}} = 0$ and $\mathcal{F}_h(\Pi_h \bar{y}) = \mathcal{F}(\Pi_h \bar{y})|_{\mathcal{Y}_{h,0}}$ we proceed as follows:

$$\|\mathcal{F}_h(\Pi_h \bar{y})\|_{\mathcal{Y}_{h,0}^*} = \left\|\mathcal{F}(\Pi_h \bar{y})|_{\mathcal{Y}_{h,0}} - \mathcal{F}(\bar{y})|_{\mathcal{Y}_{h,0}}\right\|_{\mathcal{Y}_{h,0}^*} \leq \|\mathcal{F}(\Pi_h \bar{y}) - \mathcal{F}(\bar{y})\|_{\mathcal{Y}_0^*}.$$

From the local Lipschitz continuity (2.21) of $\mathcal{F}$ we deduce that

$$\|\mathcal{F}_h(\Pi_h\bar{y})\|_{\mathcal{Y}_{h,0}^*} \leq C\|\bar{y} - \Pi_h\bar{y}\|_{\mathcal{Y}} \leq C_1 h.$$

Step 2. In Proposition 2.9 we have shown that there is an open neighbourhood $B_\delta(\bar{y}) \subset \mathcal{Y}$ and $h_0 \in (0,1]$ such that $D\mathcal{F}_h(y) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$ is an isomorphism for all $y \in B_\delta(\bar{y})$, $h \leq h_0$ and $\|D\mathcal{F}_h(y)^{-1}\| \leq M$, uniformly for $y \in B_\delta(\bar{y})$. Moreover, we observe that $D\mathcal{F}_h$ satisfies a Hölder continuity property similar to (2.20): there is $L_{\bar{y},\delta}$ and $\alpha_F \in (0,1)$ such that

$$\|D\mathcal{F}_h(y_1) - D\mathcal{F}_h(y_2)\|_{\mathcal{Y}_{h,0}^*} \leq L_{\bar{y},\delta}(\|y_1 - y_2\|_{\mathcal{Y}}^{\alpha_F} + \|y_1 - y_2\|_{\mathcal{Y}}) \qquad \forall y_1, y_2 \in B_\delta(\bar{y}).$$

Step 3. Existence and local uniqueness of a solution. We want to show that there exists a locally unique solution of $\mathcal{F}_h(y_h) = 0$. The idea is to construct a contractive mapping whose fixed point is the desired solution y_h. To this end, we rewrite this equation as

$$\mathcal{F}_h(y_h) - \mathcal{F}_h(y_0) = -\mathcal{F}_h(y_0),$$

and choose $y_0 = \Pi_h\bar{y}$ so that the right-hand side is "small". Linearization leads to

$$D\mathcal{F}_h(y_0)(y_h - y_0) = -\mathcal{F}_h(y_0) - \int_0^1 \big(D\mathcal{F}_h(y_0 + t(y_h - y_0)) - D\mathcal{F}_h(y_0)\big)\mathrm{d}t\ (y_h - y_0).$$

We recall that $D\mathcal{F}_h(y_0)$ is an isomorphism if h is sufficiently small. Let us assume in what follows that h is small enough such that $\|\bar{y} - y_0\|_{\mathcal{Y}} \leq \delta/2$. Then, for $R < \delta/2$, we define the map $\mathcal{N} : B_R(y_0) \to \mathcal{Y}_{h,D}$ by

$$D\mathcal{F}_h(y_0)(\mathcal{N}(y) - y_0) = -\mathcal{F}_h(y_0) - \int_0^1 \big(D\mathcal{F}_h(y_0 + t(y - y_0)) - D\mathcal{F}_h(y_0)\big)\mathrm{d}t\ (y - y_0).$$

We will show that $\mathcal{N}$ is a contraction from $B_R(y_0)$ into $B_R(y_0)$ if R is chosen sufficiently small.

First, we prove that $\mathcal{N}$ maps $B_R(y_0)$ to $B_R(y_0)$ for sufficiently small R. For each $y \in B_R(y_0)$ we have, with $\alpha_F \in (0,1)$, that

$$\begin{aligned} M^{-1}\|\mathcal{N}(y) - y_0\|_{\mathcal{Y}} &\leq \|\mathcal{F}_h(y_0)\|_{\mathcal{Y}_0^*} + R\int_0^1 \big\|D\mathcal{F}_h(y_0 + t(y - y_0)) - D\mathcal{F}_h(y_0)\big\|\,\mathrm{d}t \\ &\leq C_2\big(h + RL_{\bar{y},\delta}(R + R^{\alpha_F})\big), \end{aligned}$$

where we have used the stability property (2.36). To ensure that $\mathcal{N}(y) \in B_R(y_0)$, we need to bound $C_2(h + RL_{\bar{y},\delta}(R + R^{\alpha_F}))$ by R/M. If R and h are sufficiently small, this obviously holds. It is also clear that R can be chosen independently of h.

Next, we show that $\mathcal{N}$ is a contraction on $B_R(y_0)$. If $\eta_1, \eta_2 \in B_R(\bar{y})$, then

$$\begin{aligned} D\mathcal{F}_h(y_0)(\mathcal{N}(\eta_1) - \mathcal{N}(\eta_2)) &= \mathcal{F}_h(\eta_2) - \mathcal{F}_h(\eta_1) + D\mathcal{F}_h(y_0)(\eta_1 - \eta_2) \\ &= \int_0^1 \big[D\mathcal{F}_h(y_0) - D\mathcal{F}_h(\eta_1 + t(\eta_2 - \eta_1))\big](\eta_1 - \eta_2)\,\mathrm{d}t. \end{aligned}$$

Thus, $\|\mathcal{N}(\eta_1) - \mathcal{N}(\eta_2)\|_{\mathcal{Y}}$ can be estimated as follows:

$$\begin{aligned} M^{-1}\|\mathcal{N}(\eta_1) - \mathcal{N}(\eta_2)\|_{\mathcal{Y}} &\leq \int_0^1 \big\|D\mathcal{F}_h(y_0) - D\mathcal{F}_h(\eta_1 + t(\eta_2 - \eta_1))\big\|\,\mathrm{d}t\ \|\eta_1 - \eta_2\|_{\mathcal{Y}} \\ &\leq L(R + R^{\alpha_F}) \cdot \|\eta_1 - \eta_2\|_{\mathcal{Y}}. \end{aligned}$$

For sufficiently small R we obtain $L(R+R^{\alpha_F})M < 1$ and hence $\mathcal{N}$ is a contraction on $B_R(y_0)$.

We can now use Banach's Fixed Point Theorem [124, Th. 1.A] to obtain the existence and uniqueness of a fixed point $\bar{y}_h$ of the map $\mathcal{N} : B_R(y_0) \to B_R(y_0)$. This fixed point $\bar{y}_h$ is a solution of $\mathcal{F}_h(y) = 0$. For sufficiently small h this solution is in the neighbourhood $B_{2R}(\bar{y})$:

$$\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq \|\bar{y} - \Pi_h \bar{y}\|_{\mathcal{Y}} + \|\Pi_h \bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq Ch + R.$$

Step 4. Finally, convergence can be obtained by a minor modification of the above argument. If we let $R = C_R h$ and $C_R > MC_2$ we can repeat the previous steps and deduce $\|\Pi_h \bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq C_R h$. This shows

$$\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq Ch + C_R h,$$

which concludes the proof. □

Proposition 2.11. *Under the same assumptions as in Theorem 2.10 and for sufficiently small h, the discrete solution $\bar{u}_h \in \mathrm{A}_{u,h}$ is a uniform minimizer of the discretized functional* (2.32) *over* $\mathrm{A}_{u,h}$.

Proof. We define the two Lagrangians $\mathcal{L} : \mathrm{H}^1(\Omega) \times \mathbb{R} \to \mathbb{R}$, respectively, $\mathcal{L}_h : \mathrm{S}_h \times \mathbb{R} \to \mathbb{R}$ through

$$\mathcal{L}(u, \mu) = E(u) + \mu c(u) \quad \text{and} \quad \mathcal{L}_h(u_h, \mu_h) = E_h(u_h) + \mu_h c(u_h).$$

Since $\bar{u}$ is a uniform minimizer, we have $D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}) \cdot [v, v] \geq \gamma \|\nabla v\|^2_{\mathrm{L}^2}$ for all $v \in \ker Dc(\bar{u})$. Given $v_h \in \ker Dc(\bar{u}_h) \cap \mathrm{S}_{h,0}$, we carry out the following rearrangements:

$$\begin{aligned} D^2_{uu}\mathcal{L}_h(\bar{u}_h, \bar{\mu}_h) \cdot [v_h, v_h] = {} & D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}) \cdot [v, v] + \big(D^2_{uu}\mathcal{L}_h(\bar{u}_h, \bar{\mu}_h) - D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu})\big) \cdot [v_h, v_h] \\ & + 2D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}) \cdot [v, v_h - v] \\ & + D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}) \cdot [v_h - v, v_h - v] \end{aligned} \tag{2.38}$$

for arbitrary $v \in \ker Dc(\bar{u})$.

Next, we prove that every $v_h \in \ker Dc(\bar{u}_h) \cap \mathrm{S}_{h,0}$ can be approximated by $v \in \ker Dc(\bar{u}) \subset \mathrm{H}^1_0(\Omega)$ since $\|\bar{u} - \bar{u}_h\|_{\mathrm{H}^1} \leq C_1 h$. Let $\varphi = (-\Delta_0)^{-1}\bar{u} \in \mathrm{H}^1_0(\Omega)$ and define

$$v = v_h - \frac{(\nabla\varphi, \nabla v_h)}{\|\nabla\varphi\|^2_{\mathrm{L}^2}}\,\varphi.$$

It follows immediately that $(v, \bar{u}) = 0$, i.e., $v \in \ker Dc(\bar{u})$. A quick calculation using $(\bar{u}_h, v_h) = 0$ leads to $\|\nabla(v - v_h)\|_{\mathrm{L}^2} \leq Ch\,\|\nabla v_h\|_{\mathrm{L}^2}$:

$$\begin{aligned}\|\nabla(v - v_h)\|_{\mathrm{L}^2} &= \frac{|(\nabla\varphi, \nabla v_h)|}{\|\nabla\varphi\|_{\mathrm{L}^2}} = \frac{|(\bar{u}, v_h)|}{\|\nabla\varphi\|_{\mathrm{L}^2}} = \frac{|(\bar{u} - \bar{u}_h, v_h)|}{\|\nabla\varphi\|_{\mathrm{L}^2}} \\ &\leq C\,\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2}\,\frac{\|\nabla v_h\|_{\mathrm{L}^2}}{\|\nabla\varphi\|_{\mathrm{L}^2}} \leq Ch\,\|\nabla v_h\|_{\mathrm{L}^2}\,,\end{aligned}$$

where $\|\nabla\varphi\|_{\mathrm{L}^2} > 0$ has been absorbed in the generic constant C. Here, we have used the Cauchy–Schwarz inequality and Poincaré's inequality $\|v_h\|_{\mathrm{L}^2} \leq C\|\nabla v_h\|_{\mathrm{L}^2}$ for $v_h \in \mathrm{S}_{h,0}$. Based on this result we can easily derive $\|\nabla v\|_{\mathrm{L}^2} \geq (1 - Ch)\,\|\nabla v_h\|_{\mathrm{L}^2}$.

With this choice of v we see that the first term on the right-hand side of (2.38) satisfies

$$D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu})\cdot[v, v] \geq \gamma(1 - Ch)\,\|\nabla v_h\|^2_{\mathrm{L}^2}.$$

Since $D^2\mathcal{L}(\bar{u}, \bar{\mu})$ is bounded, the third and fourth term on the right-hand side of (2.38) can be bounded by $Ch\,\|\nabla v_h\|^2_{\mathrm{L}^2}$. The remaining term has the form

$$\begin{aligned}\big(D^2_{uu}\mathcal{L}_h(\bar{u}_h, \bar{\mu}_h) - D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu})\big)\cdot[v_h, v_h] &= \big(D^2X(\bar{u}) - D^2X(\bar{u}_h)\big)\cdot[v_h, v_h] \\ &\quad + \big(D^2\Phi(\bar{u}) - D^2\Phi_h(\bar{u}_h)\big)\cdot[v_h, v_h] + (\bar{\mu}_h - \bar{\mu})\|v_h\|^2_{\mathrm{L}^2}.\end{aligned}$$

The part involving the nonlinear local functional X can obviously be bounded by $C(h + h^{\alpha_F})\,\|\nabla v_h\|^2_{\mathrm{L}^2}$; see (2.12). Using the expression for $D^2\Phi$ presented in Lemma 2.4 and its discrete analogue, as well as the convergence of $\bar{\phi}_h$ to $\bar{\phi}$, we see that also

$$\big|(D^2\Phi(\bar{u}) - D^2\Phi_h(\bar{u}_h))\cdot[v_h, v_h]\big| \leq Ch\,\|\nabla v_h\|^2_{\mathrm{L}^2}\,.$$

Finally, since $|\bar{\mu}_h - \bar{\mu}| \leq Ch$ we get $|\bar{\mu}_h - \bar{\mu}|\,\|v_h\|^2_{\mathrm{L}^2} \leq Ch\,\|\nabla v_h\|^2_{\mathrm{L}^2}$ by Poincaré's inequality.

Summarising, we have established that

$$D^2_{uu}\mathcal{L}_h(\bar{u}_h, \bar{\mu}_h)\cdot[v_h, v_h] \geq \tfrac{\gamma}{2}\,\|\nabla v_h\|^2_{\mathrm{L}^2} \qquad \forall v_h \in \ker Dc(\bar{u}_h) \cap \mathrm{S}_h,$$

for sufficiently small h. From Theorem 5.6 [89] we deduce that $\bar{u}_h$ is a uniform minimizer of E_h subject to $c(\bar{u}_h) = 0$. ◻

Remark 2.12. Looking at the proof of Theorem 2.10 we see that the convergence rate is determined by $\|\mathcal{F}_h(\Pi_h\bar{y})\|_{\mathcal{Y}_{h,0}}$, which in turn depends on the approximation error $\|\bar{y} - \Pi_h\bar{y}\|_{\mathcal{Y}}$.

If $\bar{u}$, $\bar{\phi}$ have higher regularity, say $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}(\Omega)$, and S_h has the approximation property $\inf_{v_h \in \mathrm{S}_h} \|u - v_h\|_{\mathrm{H}^1} \leq C h^{\mathrm{p}} |u|_{\mathrm{H}^{\mathrm{p}+1}}$ for all $u \in \mathrm{H}^{\mathrm{p}+1}(\Omega)$, then the proof yields

$$\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq C h^{\mathrm{p}}.$$

The Hölder continuity of F'' in the proof only affects the size of the neighbourhood in which $\mathcal{N}$ is contractive. □

Remark 2.13. The analysis outlined above works without modifications if the energy functional E from (2.4) is extended by a term $\int_\Omega V(x)u^2(x)\,\mathrm{d}x$, where $V \in \mathrm{L}^q(\Omega)$, $q > \max(d/2, 1)$, is an external potential. The term is obviously well-defined for $u \in \mathrm{H}^1(\Omega)$: by Hölder's inequality

$$\left| \int_\Omega V(x)u^2(x)\,\mathrm{d}x \right| \leq \|V\|_{\mathrm{L}^q} \left\|u^2\right\|_{\mathrm{L}^{q'}} \leq C \|V\|_{\mathrm{L}^q} \|u\|^2_{\mathrm{L}^{2q'}} \leq \|V\|_{\mathrm{L}^q} \|u\|^2_{\mathrm{H}^1},$$

where we have used that the dual index $q' = \frac{q}{q-1}$ satisfies $2q' = \frac{2q}{q-1} < \frac{2d}{d-2}$ and hence $\|u\|_{\mathrm{L}^{2q'}} \leq C \|u\|_{\mathrm{H}^1}$ by the Sobolev Imbedding Theorem. The introduction of this term does not destroy the boundedness and coercivity of E from below on A_u (see proof of existence of a minimizer in Theorem 2.5). As shown in the proof of Lemma 1 in [21] we get

$$-\|V\|_{\mathrm{L}^q} \|u\|^2_{\mathrm{L}^{2q'}} \geq -C \|V\|_{\mathrm{L}^q} \|u\|^{2-3/q}_{\mathrm{L}^2} \|u\|^{3/q}_{\mathrm{H}^1}.$$

Since $\|u\|_{\mathrm{L}^2} = N$ on A_u and $3/q < 2$, this term is still dominated by $\|u\|^2_{\mathrm{H}^1}$ and hence coercivity of the energy functional is sustained.

In many practical applications a term of the form $\int_\Omega V u^2\,\mathrm{d}x$ is used to represent a pseudo-potential. In our case, V would be the difference between the Coulomb potential of ρ_n and the actual pseudo-potential. A pseudo-potential V_{ps} replaces the Coulomb potential $V_{\mathrm{nuc}}(x) = 4\pi\rho_n * |x|^{-1}$ of the nuclei in the nucleus electron interaction term: $\int_\Omega V_{\mathrm{nuc}}(x)u^2(x)\,\mathrm{d}x$. So, we would get

$$\begin{aligned} E_{nn} + E_{ee} + \int_\Omega V_{\mathrm{ps}}\, u^2\,\mathrm{d}x &= E_{nn} + E_{ee} + \int_\Omega V_{\mathrm{nuc}}\, u^2\,\mathrm{d}x + \int_\Omega (V_{\mathrm{ps}} - V_{\mathrm{nuc}})u^2\,\mathrm{d}x \\ &\approx \Phi(u) + \int_\Omega (V_{\mathrm{ps}} - V_{\mathrm{nuc}})u^2\,\mathrm{d}x. \end{aligned}$$

The electrostatic potential V_{nuc} can be easily computed as the sum of potentials of spherically symmetric charge distributions $Z_i\widetilde{\rho}_0(\cdot - R_i)$, see for example [53] for more details in a finite element context. □

We point out that the analysis described above carries over to the case when numerical integration of sufficiently high order is used. This will be shown in Chapter 3.

Before turning to an analysis of the TFDW functional with periodic boundary conditions, we highlight the main differences of our approach to similar works in the literature. As in [62]

an important step for our analysis was the introduction of the electrostatic potential ϕ as an explicit degree of freedom. This made the functional amenable to a finite element analysis. In [33,127] as well as in [22] (where a Fourier basis is used) the authors retain the electrostatic terms in their original form with convolution integrals.

The convergence proof above relies critically on the fact that $D\mathcal{F}(\bar{y})$ is an isomorphism. For this to be satisfied, we assume that $\bar{u}$ is a uniform minimizer of the functional E on A_u. Because of this assumption we can admit fairly general functions F in the functional E. In [22] the authors work with a smaller class of (convex) functions F but derive the necessary stability conditions by hand. In [33] the nonlinear function is fairly general but the authors do not obtain convergence rates.

In Section 2.5 we present a detailed analysis of convergence rates for the energy, the Lagrange multiplier and the L^2-errors of $\bar{u}_h$ and $\bar{\phi}_h$. Similar results are given in [21, 22].

2.4 The Functional on a Periodic Domain

Many DFT based simulations in quantum chemistry or materials science utilize periodic boundary conditions because of the efficiency of Fourier mode based spectral methods [67, 75, 76, 120]. For calculations of electronic structure on unit cells this is the right setting. However, also molecules are often simulated on periodic domains. On choosing the domain Ω sufficiently large, it is assumed that a molecule does not interact with the mirror images introduced by the periodicity.

Let $\Omega = (0, L_\Omega)^d \subset \mathbb{R}^d$ now be a hypercube of edge length $L_\Omega > 0$. Then, Ω is also a unit cell of the lattice $\mathcal{L} = L_\Omega \mathbb{Z}^d$. We define the reciprocal lattice of $\mathcal{L}$ by $\mathcal{L}^* = \frac{2\pi}{L_\Omega}\mathbb{Z}^d$. The following analysis could be done for any $\Omega = B\cdot(0,1)^d$ where $B \in \mathbb{R}^{d\times d}$ is an invertible transformation matrix but the formulas would be more involved.

For $j \in \mathbb{N}_0$ we introduce the periodic Sobolev space $\mathrm{H}^j_{\#}(\Omega)$ [24, Section A.11]:

$$\mathrm{H}^j_{\#}(\Omega) = \big\{ v|_\Omega : \ v \in \mathrm{H}^j_{\mathrm{loc}}(\mathbb{R}^d), \ \ v \text{ is } \mathcal{L}\text{-periodic} \big\}.$$

Let the functions ω_k for $k \in \mathbb{R}^d$ be defined as

$$\omega_k(x) = |\Omega|^{-1/2} \mathrm{e}^{\mathrm{i}k\cdot x}.$$

Then, the family $\{\omega_k : k \in \mathcal{L}^*\}$ forms an orthonormal basis of $\mathrm{L}^2(\Omega)$: every $u \in \mathrm{L}^2(\Omega)$ can uniquely be written in the form

$$u = \sum_{k\in\mathcal{L}^*} \widehat{u}_k \omega_k \qquad \text{where } \widehat{u}_k = (\omega_k, u).$$

If the function u is real-valued, the Fourier coefficients satisfy $\widehat{u}_{-k} = \bar{\widehat{u}}_k$ for all $k \in \mathcal{L}^*$. This

gives rise to an equivalent definition of the periodic Sobolev spaces:

$$\mathrm{H}^j_{\#}(\Omega) = \Big\{ u = \sum_{k\in\mathcal{L}^*} \widehat{u}_k \omega_k : \quad \sum_{k\in\mathcal{L}^*} (1+|k|^2)^j |\widehat{u}_k|^2 < \infty, \quad \widehat{u}_{-k} = \bar{\widehat{u}}_k \ \forall k \in \mathcal{L}^* \Big\}.$$

In the present section we analyze a periodic version of the minimization problem (2.7). The density functional $E_{\#}$ takes the form

$$E_{\#}(u) = \frac{\lambda}{2} \int_\Omega |\nabla u| \,\mathrm{d}x + \int_\Omega F(u) \,\mathrm{d}x + \Phi_{\#}(u). \tag{2.39}$$

In the periodic case the charge density ρ_n belongs to $\mathrm{C}^\infty_{\#}(\Omega)$: it is the charge density produced by an Ω-periodic extension of the configuration R to $\mathbb{R}^d$. In order for the periodic model or, more specifically, the Coulomb energy to be physically sensible the net charge of the system has to be zero: $\int_\Omega (u^2 - \rho_n) \,\mathrm{d}x = 0$. The Coulomb energy is the given by [80]

$$\begin{aligned}\Phi_{\#}(u) &= \frac{1}{2} \int_\Omega \int_\Omega \big(u^2(x) - \rho_n(x)\big) G_{\#}(x-z) \big(u^2(z) - \rho_n(z)\big) \,\mathrm{d}x \,\mathrm{d}z \\ &= \frac{1}{2} \int_\Omega \phi(x) \big(u^2(x) - \rho_n(x)\big) \,\mathrm{d}x \,,\end{aligned}$$

where

$$\phi(x) = \int_\Omega (u^2(z) - \rho_n(z)) G_{\#}(x-z) \,\mathrm{d}z$$

is the periodic electrostatic potential. Here, $G_{\#} \in \mathrm{L}^2(\Omega)$ denotes the periodic Coulomb potential defined by

$$-\Delta G_{\#} = 4\pi\big(\delta - \tfrac{1}{|\Omega|}\big)$$

in the sense of distributions and $\int_\Omega G_{\#} \,\mathrm{d}x = 0$. Here δ denotes the Dirac distribution. $G_{\#}$ is only defined up to an additive constant and it satisfies $G_{\#}(x) - \frac{1}{|x|} \to \text{const.}$ for $x \to 0$ [81]. The most natural choice of $G_{\#}$ is [80, 81]

$$G_{\#}(x) = \frac{4\pi}{|\Omega|} \sum_{\substack{k\in\mathcal{L}^* \\ k\neq 0}} |k|^{-2} \mathrm{e}^{\mathrm{i}k\cdot x}.$$

On introducing the space of mean-value-free periodic Sobolev functions

$$\mathrm{H}^1_{\#,0}(\Omega) = \{\phi \in \mathrm{H}^1_{\#}(\Omega) : \int_\Omega \phi \,\mathrm{d}x = 0\},$$

it follows in analogy to the Dirichlet case that we can write $\Phi_{\#}$ as the minimum of the u-dependent functional $\Psi(u,\cdot)$ over $\mathrm{H}^1_{\#,0}(\Omega)$:

$$\Phi_{\#}(u) = -\inf_{\phi\in\mathrm{H}^1_{\#,0}(\Omega)} \Psi(u,\phi) = -\inf_{\phi\in\mathrm{H}^1_{\#,0}(\Omega)} \left[\int_\Omega \frac{1}{8\pi} |\nabla\phi|^2 - (u^2 - \rho_n)\phi \,\mathrm{d}x \right].$$

The periodic minimization problem takes the form

$$\min_{u\in \mathrm{A}_{u,\#}} E_{\#}(u), \tag{2.40}$$

where the set of admissible periodic Sobolev functions is defined as

$$\mathrm{A}_{u,\#} = \left\{u \in \mathrm{H}^1_{\#}(\Omega) : \|u\|^2_{\mathrm{L}^2} = n_{\mathrm{el}}\right\}.$$

Just as in Theorem 2.5 we can show that the functional $E_{\#}$ has a minimizer in $\mathrm{A}_{u,\#}$ if F satisfies (2.9), (2.10), and

$$c_1 t^2 + a_1 < F(t) < c_2 t^{q_F} + a_2 \tag{2.41}$$

for all $t \in \mathbb{R}$ where $a_1, a_2 \in \mathbb{R}$, $c_1, c_2 > 0$ and $3 \le q_F < 6$ if $d = 3$ or any q_F if $d < 3$. In order to ensure H^2-regularity of $\bar{u}$ we will, however, assume that $q_F = 4$ if $d = 3$; see the discussion at the end of Section 2.2.3.

We will now outline the convergence analysis for two different discretizations of the minimization problem that are again based on the optimality system. Since the analysis follows essentially the same lines as in the case of Dirichlet boundary conditions, we will only highlight the steps that need modification.

Let $\bar{u}$ be a minimizer with corresponding electrostatic potential $\bar{\phi}$ and Lagrange multiplier $\bar{\mu}$. Set $\mathcal{L}_{\#}(u, \mu) = E_{\#}(u) + \mu c(u)$ and assume that $\bar{u}$ satisfies the uniformity condition

$$D^2_{uu}\mathcal{L}_{\#}(\bar{u}, \bar{\mu})\cdot[v, v] \ge \gamma \|v\|^2_{\mathrm{H}^1} \quad \forall v \in \ker Dc(\bar{u}), \tag{2.42}$$

with $\gamma > 0$, or, more explicitly

$$\lambda(\nabla v, \nabla v) + \left((F''(\bar{u}) + 2\bar{\phi} + \bar{\mu})v, v\right) + 16\pi(\bar{u}v, (-\Delta_{\#,0})^{-1}\bar{u}v) \ge \gamma \|v\|^2_{\mathrm{H}^1}. \tag{2.43}$$

Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu})$ and define

$$\mathcal{Y}_{\#} = \mathrm{H}^1_{\#}(\Omega) \times \mathrm{H}^1_{\#,0}(\Omega) \times \mathbb{R}.$$

The first order optimality conditions of (2.40) can again be written in the (local) form

$$\begin{aligned}
\lambda(\nabla\bar{u}, \nabla v) + (F'(\bar{u}), v) + 2(\bar{\phi}\bar{u}, v) + \bar{\mu}(\bar{u}, v) = 0 \qquad & \forall v \in \mathrm{H}^1_{\#}(\Omega), \\
\frac{1}{4\pi}(\nabla\bar{\phi}, \nabla\psi) - (\bar{u}^2 - \rho_n, \psi) = 0 \qquad & \forall \psi \in \mathrm{H}^1_{\#,0}(\Omega), \\
\frac{\nu}{2}\left(\int_\Omega \bar{u}^2 \,\mathrm{d}x - n_{\mathrm{el}}\right) = 0 \qquad & \forall \nu \in \mathbb{R}.
\end{aligned} \tag{2.44}$$

As in the Dirichlet case, we will abbreviate this nonlinear system of equations as

$$\mathcal{F}_{\#}(\bar{y}) = 0 \ \in \mathcal{Y}^*_{\#},$$

where $\mathcal{F}_{\#}$ is a nonlinear operator from $\mathcal{Y}_{\#}$ to $\mathcal{Y}^*_{\#}$.

Since $\bar{u}$ is a uniform minimizer, $D\mathcal{F}_{\#}(\bar{y})$ is an isomorphism from $\mathcal{Y}_{\#}$ to $\mathcal{Y}_{\#}^*$. Proving this works exactly as in the Dirichlet case. We observe that the derivative $D\mathcal{F}_{\#}$ has saddle-point structure where $a_{\bar{y}}$ and $b_{\bar{y}}$ have the same form as in Section 2.2.4. The inf-sup condition for $b_{\bar{y}}$ follows easily

$$\begin{aligned}\inf_{\nu\in\mathbb{R}} \sup_{(v,\psi)\in \mathrm{H}^1_{\#}\times\mathrm{H}^1_{\#,0}} \frac{b_{\bar{y}}((v,\psi),\nu)}{|\nu|\,(\|v\|^2_{\mathrm{H}^1}+\|\nabla\psi\|^2_{\mathrm{L}^2})^{1/2}} &= \inf_{\nu\in\mathbb{R}} \sup_{(v,\psi)\in \mathrm{H}^1_{\#}\times\mathrm{H}^1_{\#,0}} \frac{\nu\int_\Omega \bar{u}v\,\mathrm{d}x}{|\nu|\,(\|v\|^2_{\mathrm{H}^1}+\|\nabla\psi\|^2_{\mathrm{L}^2})^{1/2}} \\ &\geq \frac{(\bar{u},(-\Delta+\mathrm{id})^{-1}_{\#}\bar{u})}{(\bar{u},(-\Delta+\mathrm{id})^{-1}_{\#}\bar{u})^{1/2}} =: \kappa_b > 0.\end{aligned}$$

The invertibility of $a_{\bar{y}}$ on $\ker b_{\bar{y}} = \{(v,\psi)\in \mathrm{H}^1_{\#}(\Omega)\times\mathrm{H}^1_{\#,0}(\Omega) : v\in\ker Dc(\bar{u})\}$ follows exactly as in the proof of Proposition 2.7.

For the discretization we consider two cases: periodic finite elements and the more classical Fourier basis. Periodic finite elements seem like a rather unconventional tool but have been used in the context of DFT in [112].

2.4.1 Discretization with Periodic Finite Elements

Before proving convergence of a finite element discretization of the periodic problem (2.44) we need to specify some properties of the meshes involved.

2.4.1.1 Mesh Regularity

Let $(\mathcal{T}_h)_{h\in(0,1]}$ be a family of nested periodic triangulations of $\Omega = (0,L_\Omega)^d$ in the sense that for every $\mathcal{T}_h$ the $(d-1)$-dimensional meshes on opposite surfaces of Ω are identical (vertices and edges are identical). The most natural uniform tetrahedral meshes satisfy this property.

For every element $T\in\mathcal{T}_h$ we define the set $P_{\mathrm{p}}(T)$ of polynomials of degree smaller or equal to p over T. We consider a family of conforming finite element spaces $(\mathrm{S}_h)_{0<h\leq 1}$ of p-th order over the triangulations $(\mathcal{T}_h)_{h\in(0,1]}$: every $u_h\in\mathrm{S}_h$ satisfies $u_h|_T\in P_{\mathrm{p}}(T)$ for all elements $T\in\mathcal{T}_h$ and $u_h\in\mathrm{C}^0(\Omega)$. From this it follows that $\mathrm{S}_h\subset\mathrm{H}^1(\Omega)$ for all $h\in(0,1]$; see, for example, [17, Chapter 4].

Let $\widehat{T}\subset\mathbb{R}^d$ be the respective reference element. For every element $T\in\mathcal{T}_h$ there exists an affine mapping

$$F_T:\widehat{T}\to T,\quad \widehat{x}\mapsto x = F_T(\widehat{x}) = B_T\widehat{x}+x_T, \tag{2.45}$$

where $B_T\in\mathbb{R}^{d\times d}$ is the transformation matrix and $x_T\in\mathbb{R}^d$ is a reference point for T, e.g., a vertex. For every function f defined on $T\in\mathcal{T}_h$, we denote by $\widehat{f}$ the corresponding function on the reference element: $\widehat{f}(\widehat{x}) = f(F_T(\widehat{x}))$ for all $\widehat{x}\in\widehat{T}$.

In the previous sections we merely assumed that S_h had the approximation property (2.30). In the remainder of this thesis we also need some geometrical properties of the meshes $(\mathcal{T}_h)_{h\in(0,1]}$ over which the spaces S_h are defined. Let h_T be the diameter of the element $T\in\mathcal{T}_h$

and ρ_T the diameter of the largest ball fitting inside T. For every $\mathcal{T}_h$ the mesh size $h \in (0,1]$ satisfies

$$h \geq \max_{T \in \mathcal{T}_h} h_T.$$

We assume in the following that the family of triangulations $(\mathcal{T}_h)_{h\in(0,1]}$ is *quasi-uniform* [17, Definition 4.4.13.]: there exists $\sigma > 0$ such that for all $h \in (0,1]$

$$\min_{T \in \mathcal{T}_h} \rho_T \geq \sigma h. \tag{2.46}$$

This implies (see [37, Section 3.1]) the existence of $c, C > 0$ such that for all $T \in \mathcal{T}_h$ and all $h \in (0,1]$

$$\|B_T\| \leq Ch, \quad \|B_T^{-1}\| \leq Ch^{-1}, \quad ch^d \leq \det B_T \leq Ch^d, \quad ch^d \leq |T| \leq Ch^d. \tag{2.47}$$

We will frequently use the transformation properties [37, Theorem 3.1.2]

$$\begin{aligned} |\widehat{v}|_{\mathrm{W}^{m,q}(\widehat{T})} &\leq C\|B_T\| \, |\det B_T|^{-1/q} |v|_{\mathrm{W}^{m,q}(T)} \quad \forall v \in \mathrm{W}^{m,q}(T), \\ |v|_{\mathrm{W}^{m,q}(T)} &\leq C\|B_T^{-1}\| \, |\det B_T|^{1/q} |\widehat{v}|_{\mathrm{W}^{m,q}(\widehat{T})} \quad \forall v \in \mathrm{W}^{m,q}(T), \end{aligned} \tag{2.48}$$

for all $T \in \mathcal{T}_h$.

2.4.1.2 Existence and Convergence

In the case of Dirichlet boundary conditions we know that $\bar{u},\ \bar{\phi} \in \mathrm{H}^2(\Omega)$. We could in general not obtain higher regularity because of possible corner singularities. This situation changes in the case of periodicity. From $F \in \mathrm{C}^{2,\alpha_F}(\mathbb{R})$ and $\bar{u},\ \bar{\phi} \in \mathrm{H}^2_{\#}(\Omega)$ it follows that $F'(\bar{u}),\ \bar{u}\bar{\phi},\ \bar{u}^2 \in \mathrm{H}^1_{\#}(\Omega)$ (see [52, Section 4.2.2]) and hence by elliptic regularity $\bar{u}, \bar{\phi} \in \mathrm{H}^3_{\#}(\Omega)$, which in turn implies $\nabla\bar{u},\ \nabla\bar{\phi} \in \mathrm{L}^\infty_{\#}(\Omega;\mathbb{R}^d)$. This argument can be iterated as many times as the differentiability of F allows. Let $j \in \mathbb{N}$, $j \geq 3$. If $F \in \mathrm{C}^j(I_{\bar{u}})$ for an open interval $I_{\bar{u}}$ such that $\bar{u}(\bar{\Omega}) \subset I_{\bar{u}}$ then $\bar{u}, \bar{\phi} \in \mathrm{H}^{j+1}_{\#}(\Omega)$.

Let $\mathrm{p} \in \mathbb{N}$ and S_h be a finite element space of p-th order over $\mathcal{T}_h$ and introduce the following discretization spaces of periodic finite elements

$$\mathrm{S}_{h,\#} = \mathrm{S}_h \cap \mathrm{H}^1_{\#}(\Omega), \qquad \mathrm{S}_{h,\#,0} = \mathrm{S}_{h,\#} \cap \mathrm{H}^1_{\#,0}(\Omega). \tag{2.49}$$

If $\mathcal{I}_h : \mathrm{H}^{\mathrm{p}+1}(\Omega) \to \mathrm{S}_h$ is the nodal interpolation operator, then we have $\mathcal{I}_h u \in \mathrm{S}_{h,\#}$ for all $u \in \mathrm{H}^{\mathrm{p}+1}_{\#}(\Omega)$. Furthermore, under condition (2.46)

$$\|u - \mathcal{I}_h u\|_{\mathrm{H}^1} \leq Ch^{\mathrm{p}} |u|_{\mathrm{H}^{\mathrm{p}+1}} \qquad \forall u \in \mathrm{H}^{\mathrm{p}+1}_{\#}(\Omega),$$

see [17, Chapter 4]. From this we immediately infer the approximation property

$$\inf_{v \in \mathrm{S}_{h,\#}} \|u - v\|_{\mathrm{H}^1} \leq Ch^{\mathrm{p}} |u|_{\mathrm{H}^{\mathrm{p}+1}} \qquad \forall\, u \in \mathrm{H}^{\mathrm{p}+1}_{\#}(\Omega).$$

The Galerkin discretization of (2.44) is given by

$$\begin{aligned} \lambda(\nabla \bar{u}_h, \nabla v) + (F'(\bar{u}_h), v) + 2(\bar{\phi}_h \bar{u}_h, v) + \mu_h(\bar{u}_h, v) = 0 \quad & \forall v \in \mathrm{S}_{h,\#}, \\ \frac{1}{4\pi}(\nabla \bar{\phi}_h, \nabla \psi) - (\bar{u}_h^2 - \rho_n, \psi) = 0 \quad & \forall \psi \in \mathrm{S}_{h,\#,0}, \\ \frac{\nu}{2}\left(\int_\Omega \bar{u}_h^2 \,\mathrm{d}x - n_{\mathrm{el}}\right) = 0 \quad & \forall \nu \in \mathbb{R}. \end{aligned} \tag{2.50}$$

On introducing the space

$$\mathcal{Y}_{h,\#} = \mathrm{S}_{h,\#} \times \mathrm{S}_{h,\#,0} \times \mathbb{R}$$

this translates into the nonlinear equation

$$\langle \mathcal{F}_{h,\#}(\bar{y}_h), \eta \rangle = 0 \quad \forall \eta \in \mathcal{Y}_{h,\#},$$

where $\mathcal{F}_{h,\#} : \mathcal{Y}_{h,\#} \to \mathcal{Y}_{h,\#}^*$ and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,\#}$ is the sought-after solution.

Theorem 2.14. *Let $\bar{u} \in \mathrm{A}_{u,\#}$ be a minimizer of (2.40) that satisfies the uniformity condition (2.42). Let $\bar{\phi} \in \mathrm{H}^1_\#(\Omega)$ and $\bar{\mu} \in \mathbb{R}$ be, respectively, the associated electrostatic potential and Lagrange multiplier. If $\mathrm{p} \in \{1, 2\}$, or $\mathrm{p} \geq 3$ and $F \in \mathrm{C}^{\mathrm{p}}(I_{\bar{u}})$, there exist $h_0 \in (0, 1], \delta > 0$ such that the discretized problem (2.50) has a unique solution $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,\#}$ in the neighbourhood $B_\delta(\bar{y}) \cap \mathcal{Y}_{h,\#}$ for all $h < h_0$. Furthermore, there exists a constant C such that*

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{H}^1} + \left\|\bar{\phi} - \bar{\phi}_h\right\|_{\mathrm{H}^1} + |\bar{\mu} - \bar{\mu}_h| \leq C h^{\mathrm{p}}.$$

Proof. As explained above, it follows from the assumed smoothness properties of F that $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}_\#(\Omega)$. The existence and convergence analysis can then proceed as in the Dirichlet case. Having established that $D\mathcal{F}_{h,\#}(\bar{y}) : \mathcal{Y}_\# \to \mathcal{Y}_\#^*$ is an isomorphism, the same proof as for Proposition 2.9 implies that $D\mathcal{F}_{h,\#}(y) : \mathcal{Y}_{h,\#} \to \mathcal{Y}_{h,\#}^*$ is an isomorphism for every y in a neighbourhood $B_\delta(\bar{y}) \subset \mathcal{Y}_\#$ and $D\mathcal{F}_{h,\#}(y)^{-1}$ is uniformly bounded in y and h. The necessary version of Schatz' argument is discussed in Section A.2.2 in the Appendix. Note that because of $\phi \in \mathrm{H}^1_{\#,0}(\Omega)$ there are now two linear constraints on the space $\mathbf{V}$.

The actual proof of convergence from Theorem 2.10 is sufficiently generic and transfers without modifications. ◻

A straightforward generalization of Proposition 2.11 implies that for sufficiently small h the discrete solution $\bar{u}_h$ is a uniform minimizer of the discrete energy $E_{h,\#}$, which is canonically given by

$$E_{h,\#}(u_h) = \frac{\lambda}{2}\int_\Omega |\nabla u_h| \,\mathrm{d}x + \int_\Omega F(u_h) \,\mathrm{d}x - \inf_{\phi_h \in \mathrm{S}_{h,\#,0}} \Psi(u_h, \phi_h). \tag{2.51}$$

2.4.2 Discretization with Fourier Basis

Let $N \in \mathbb{N}$ and define the index set

$$K_N = \left\{ \frac{2\pi j}{L_\Omega} : \ j \in \{-N, \dots, N\} \right\}.$$

We consider the following one-dimensional approximation space of Fourier modes[1]

$$\mathrm{S}_N^{(1)} = \left\{ \sum_{k \in K_N} \widehat{u}_k \omega_k^{(1)} : \ \widehat{u}_{-k} = \overline{\widehat{u}}_k \ \ \forall k \in K_N \right\},$$

where $\omega_k^{(1)}(x) = L_\Omega^{-1/2} \mathrm{e}^{\mathrm{i}kx}$, for all $x \in (0, L_\Omega)$, $k \in K_N$. From this one-dimensional space we derive the approximation spaces

$$\mathrm{S}_N = \underbrace{\mathrm{S}_N^{(1)} \otimes \ldots \otimes \mathrm{S}_N^{(1)}}_{d \text{ times}} \subset \mathrm{H}^1_{\#}(\Omega)$$

and

$$\mathrm{S}_{N,0} = \left\{ u \in \mathrm{S}_N : \int_\Omega u \, \mathrm{d}x \ = 0 \right\} \subset \mathrm{H}^1_{\#,0}.$$

The projection operator $\Pi_N : \mathrm{H}^1_{\#}(\Omega) \to \mathrm{S}_N$ is defined as follows: for $u = \sum_{k \in \mathcal{L}^*} \widehat{u}_k \omega_k \in \mathrm{H}^1_{\#}(\Omega)$ set

$$\Pi_N u = \sum_{k \in K_N^d} \widehat{u}_k \omega_k.$$

We have the following classical approximation property (see [25] or [24, Section 5.8.1])

$$\|\Pi_N u - u\|_{\mathrm{H}^r} \leq C_{r,s} N^{r-s} \|u\|_{\mathrm{H}^s} \quad \forall u \in \mathrm{H}^s_{\#}(\Omega),$$

for all $r, s \in \mathbb{Z}$ such that $s > r$. The choice of basis functions implies the following set of interpolation nodes

$$X_N = \frac{L_\Omega}{2N+1} \{0, \dots, 2N\}^d.$$

Furthermore, we define the standard interpolation operator $\mathcal{I}_N : \mathrm{H}^2_{\#}(\Omega) \to \mathrm{S}_N$ by $(\mathcal{I}_N u)(x) = u(x)$ for all nodes $x \in X_N$. The operator $\mathcal{I}_N$ has the approximation property

$$\|\mathcal{I}_N u - u\|_{\mathrm{H}^r} \leq C_{r,s} N^{r-s} \|u\|_{\mathrm{H}^s} \quad \forall u \in \mathrm{H}^s_{\#}(\Omega), \tag{2.52}$$

for all $s \in \mathbb{N}$, $s \geq 2$ and $r \in \mathbb{N}_0$, $0 \leq r \leq s$, see [25, Section 3] or [24, Section 5.8.1].

[1] From a computational point of view it is advantageous to work with a space of even dimension since the Fast Fourier Transform is fastest for powers of 2. For simplicity of notation we work with an odd dimension but the procedure works without modification in the even case.

The Galerkin discretization of (2.44) utilizing the Fourier basis S_N is evident: find $\bar{y}_N = (\bar{u}_N, \bar{\phi}_N, \bar{\mu}_N) \in \mathcal{Y}_N = \mathrm{S}_N \times \mathrm{S}_{N,0} \times \mathbb{R}$ such that

$$\begin{aligned} \lambda(\nabla \bar{u}_N, \nabla v) + (F'(\bar{u}_N), v) + 2(\bar{\phi}_N \bar{u}_N, v) + \bar{\mu}_N(\bar{u}_N, v) = 0 \quad & \forall v \in \mathrm{S}_N, \\ \frac{1}{4\pi}(\nabla \bar{\phi}_N, \nabla \psi) - (\bar{u}_N^2 - \rho_n, \psi) = 0 \quad & \forall \psi \in \mathrm{S}_{N,0}, \\ \frac{\nu}{2}\left(\int_\Omega \bar{u}_N^2 \,\mathrm{d}x - n_{\mathrm{el}} \right) = 0 \quad & \forall \nu \in \mathbb{R}, \end{aligned} \tag{2.53}$$

or $\mathcal{F}_N(\bar{y}_N) = 0$, where $\mathcal{F}_N : \mathcal{Y}_N \to \mathcal{Y}_N^*$.

Theorem 2.15. *Let $\bar{u}$ be a minimizer of (2.40) that satisfies (2.42). Let $\bar{\phi} \in \mathrm{H}^1_{\#}(\Omega)$ and $\bar{\mu} \in \mathbb{R}$ be, respectively, the associated electrostatic potential and Lagrange multiplier. Let $\mathrm{p} = 2$, or $\mathrm{p} \geq 3$ and $F \in \mathrm{C}^{\mathrm{p}}(I_{\bar{u}})$. Then, there exist $N_0 \in \mathbb{N}, \delta > 0$ such that the discretized problem (2.53) has a unique solution $\bar{y}_N = (\bar{u}_N, \bar{\phi}_N, \bar{\mu}_N)$ in the neighbourhood $B_\delta(\bar{y}) \cap \mathcal{Y}_N$ for all $N \geq N_0$. Furthermore, there exists a constant C such that*

$$\|\bar{u} - \bar{u}_N\|_{\mathrm{H}^1} + \left\|\bar{\phi} - \bar{\phi}_N\right\|_{\mathrm{H}^1} + |\bar{\mu} - \bar{\mu}_N| \leq C N^{-\mathrm{p}}.$$

Proof. The structure of the proof is clear. The crucial step involving the Schatz' argument carries over to the context of a Fourier basis; see the discussion in Section A.2.2 in the Appendix. □

2.5 Optimal Convergence Rates

In this section we investigate convergence rates for the energy $E_h(\bar{u}_h)$, the Lagrange multiplier $\bar{\mu}_h$ and the L^2-errors of $\bar{u}_h$, $\bar{\phi}_h$ more closely. We will find that under certain conditions the convergence order can be improved compared with the $\mathcal{O}(h)$ in the Dirichlet and the $\mathcal{O}(h^{\mathrm{p}})$, respectively, $\mathcal{O}(N^{-\mathrm{p}})$ in the periodic case, which we proved above. The energy E and the Lagrange multiplier μ, which can be interpreted as a chemical potential ($\mu = D_N E$) are important quantities for practitioners.

The main tool in the analysis of convergence rates will be duality ideas as described for example in the monograph of Bangerth & Rannacher [6]. Although they apply these techniques in the construction of *a posteriori* error estimators, *a priori* convergence orders can be derived equally well. In unconstrained optimization problems it is evident that under some smoothness assumptions the energy converges at twice the rate as the minimizer. This is a direct consequence of the fact that the gradient of the objective function vanishes in a minimizer. In the present case, we have to cope with the constraint $c(u) = 0$ and the fact that the nonlocal Coulomb energy Φ is approximated by Φ_h. Both issues are efficiently dealt with in the duality setting. We will also use a generalization of the "Aubin–Nitsche trick" [16,

Lemma II.7.6] to prove that the convergence rate for the L^2-errors is higher than for the H^1-errors. Moreover, we will see that under certain conditions the Lagrange multiplier converges significantly faster than we saw in Theorem 2.10, which is a well-studied phenomenon in the discretization of linear eigenvalue problems, see for example [3].

The factors governing the convergence rates in the present case are the Hölder continuity of the second derivative of E (respectively F) and the regularity of certain *dual problems* on Ω. In order to highlight the relationships between differentiability of $\mathcal{F}$ and regularity of the elliptic dual problems on the one side and convergence orders on the other side, we will distinguish different cases here. We shall begin with the case of periodic boundary conditions and a finite element discretization. This case allows for a clean and complete analysis. Then, we will address the case of Dirichlet conditions, where the regularity of certain elliptic problems will play a role.

Throughout this chapter, we assume that the family $(\mathcal{T}_h)_{h\in(0,1]}$ of triangulations is quasi-uniform as discussed in Section 2.4.1.1.

2.5.1 Periodic Boundary Conditions

We first show that the convergence rate of the discrete minimum energy $E_{h,\#}(\bar{u}_h)$ is twice the rate of $\bar{u}_h$, respectively, $\bar{y}_h$.

Proposition 2.16. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_\#$ be a solution of the continuous periodic problem (2.44) and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,\#}$ the corresponding solution to the discrete periodic problem (2.50). Then,*

$$|E_\#(\bar{u}) - E_{h,\#}(\bar{u}_h)| \leq C\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}^2.$$

Proof. We begin by defining the following Lagrangian functional $\mathcal{L}_{E,\#} : \mathcal{Y}_\# \to \mathbb{R}$:

$$\mathcal{L}_{E,\#}(y) = \frac{\lambda}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}x + \int_\Omega F(u)\,\mathrm{d}x - \Psi(u,\phi) + \mu c(u),$$

where $y = (u, \phi, \mu) \in \mathcal{Y}_\#$. This definition implies $\mathcal{L}_{E,\#}(\bar{y}) = E_\#(\bar{u})$ and $\mathcal{L}_{E,\#}(\bar{y}_h) = E_{h,\#}(\bar{u}_h)$ since $c(\bar{u}) = c(\bar{u}_h) = 0$. The optimality conditions (2.44) read

$$D\mathcal{L}_{E,\#}(\bar{y}) = 0 \quad \text{in } \mathcal{Y}_\#^*.$$

In other words $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu})$ is a stationary point of $\mathcal{L}_{E,\#}$ (a saddle point to be more precise). Taylor's Theorem [124, Theorem 4.A] with remainder term now gives

$$\begin{aligned}
|E_\#(\bar{u}) - E_{h,\#}(\bar{u}_h)| &= |\mathcal{L}_{E,\#}(\bar{y}) - \mathcal{L}_{E,\#}(\bar{y}_h)| \\
&\leq \left|D\mathcal{L}_{E,\#}(\bar{y})\cdot(\bar{y} - \bar{y}_h)\right| + \tfrac{1}{2}\|D^2\mathcal{L}_{E,\#}(y_h^*)\|\,\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}^2 \\
&= \tfrac{1}{2}\|D^2\mathcal{L}_{E,\#}(y_h^*)\|\,\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}^2,
\end{aligned}$$

where $y_h^* \in \operatorname{conv}\{\bar{y}, \bar{y}_h\} \subset \mathcal{Y}$ such that $\|\bar{y} - y_h^*\|_{\mathcal{Y}} \leq \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}$. ◻

We stress that the convergence of the energy is not affected by the continuity properties of $D^2 E_\#$, or more precisely, F. The second derivative of $E_\#$ only enters the constant multiplying $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}^2$ in the form of $\|D^2 \mathcal{L}_{E,\#}(y_h^*)\|$, which is bounded uniformly in h.

Next we take a look at optimal convergence rates of $\bar{u}_h$ and $\bar{\phi}_h$ in the L^2-norm. Here, the Hölder continuity of the energy functional $E_\#$, respectively, the nonlinear operator $\mathcal{F}_\#$ is reflected in the additional exponent of h.

Proposition 2.17. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_\#$ be a solution of the continuous periodic problem (2.44) with $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}_\#(\Omega)$ and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,\#}$ the corresponding solution to the discrete periodic problem (2.50) satisfying $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq C h^{\mathrm{p}}$. Then, there exists $C > 0$ such that*

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} + \big\|\bar{\phi} - \bar{\phi}_h\big\|_{\mathrm{L}^2} \leq C\big(h^{\mathrm{p}+1} + h^{\mathrm{p}/(1-\alpha_F)}\big),$$

if $\alpha_F < 1$. If $\alpha_F = 1$, that is $D\mathcal{F}_{h,\#}$ is locally Lipschitz continuous, $h^{\mathrm{p}/(1-\alpha_F)}$ can be replaced with $h^{\mathrm{p}+1}$.

Proof. The idea for this proof is to introduce a dual variable $\bar{z} \in \mathcal{Y}_\#$ [6, Ch. 6] that solves an equation containing information on the error under consideration: $\bar{u} - \bar{u}_h$, respectively, $\bar{\phi} - \bar{\phi}_h$ in this case.

Step 1. The dual solution. Let $f_u, f_\phi \in \mathrm{L}^2(\Omega)$ and f be an element of $\mathcal{Y}_\#^*$ defined through

$$\langle f, y\rangle = (u, f_u) + (\phi, f_\phi) \quad \forall y = (u, \phi, \mu) \in \mathcal{Y}_\#.$$

We then define the Lagrangian functional $\mathcal{L}_\# : \mathcal{Y}_\# \times \mathcal{Y}_\# \to \mathbb{R}$ by

$$\mathcal{L}_\#(y, z) = \langle f, y\rangle - \langle \mathcal{F}_\#(y), z\rangle . \tag{2.54}$$

$\mathcal{L}_\#$ is continuously Fréchet differentiable with first partial derivatives given by:

$$\begin{aligned} \langle D_y \mathcal{L}_\#(y, z), \eta\rangle &= \langle f, \eta\rangle - \langle D\mathcal{F}_\#(y)\cdot\eta, z\rangle \quad \text{for } \eta \in \mathcal{Y}_\#, \\ \langle D_z \mathcal{L}_\#(y, z), \zeta\rangle &= \langle \mathcal{F}_\#(y), \zeta\rangle \qquad\qquad\quad \text{for } \zeta \in \mathcal{Y}_\#. \end{aligned}$$

The derivative $D_y \mathcal{L}_\#$ is Hölder continuous with degree α_F in a similar sense as $D\mathcal{F}_\#$; see (2.20). We already know that for all $z \in \mathcal{Y}_\#$

$$\langle D_z \mathcal{L}_\#(\bar{y}, z), \zeta\rangle = \langle \mathcal{F}_\#(\bar{y}), \zeta\rangle = 0 \quad \forall \zeta \in \mathcal{Y}_\#.$$

Let us now look at the derivative of $\mathcal{L}_\#$ with respect to y. Since $D\mathcal{F}_\#(\bar{y}) : \mathcal{Y}_\# \to \mathcal{Y}_\#^*$ satisfies the inf-sup conditions (2.29), the same inf-sup conditions hold for the adjoint $D\mathcal{F}_\#(\bar{y})^*$:

$\mathcal{Y}_\# \to \mathcal{Y}_\#^*$, which consequently is also an isomorphism. Hence, there exists a unique *dual* solution $\bar{z} \in \mathcal{Y}_\#$ to the equation

$$\langle f, \eta \rangle - \langle D\mathcal{F}_\#(\bar{y})\cdot\eta, \bar{z} \rangle = 0 \quad \forall \eta \in \mathcal{Y}_\#, \tag{2.55}$$

with $\|\bar{z}\|_\mathcal{Y} \leq C\|f\|_{\mathcal{Y}_\#^*}$. We have thus constructed $\bar{z} \in \mathcal{Y}_\#$ such that $(\bar{y}, \bar{z}) \in \mathcal{Y}_\# \times \mathcal{Y}_\#$ is a stationary point of $\mathcal{L}_\#$:

$$D_y\mathcal{L}_\#(\bar{y}, \bar{z}) = 0 \quad \text{and} \quad D_z\mathcal{L}_\#(\bar{y}, \bar{z}) = 0.$$

Step 2. Existence and convergence of the discrete dual solution. Next, we look at the restriction of $\mathcal{L}_\#$ to the finite-dimensional space $\mathcal{Y}_{h,\#} \times \mathcal{Y}_{h,\#}$. Using identical reasoning as in the proof that $D\mathcal{F}_h(\bar{y}_h)$ is an isomorphism in Proposition 2.9, we can show the existence of a unique discrete dual solution $\bar{z}_h \in \mathcal{Y}_{h,\#}$ to

$$\langle f, \eta_h \rangle - \langle D\mathcal{F}_{h,\#}(\bar{y}_h)\cdot\eta_h, \bar{z}_h \rangle = 0 \quad \forall \eta_h \in \mathcal{Y}_{h,\#}. \tag{2.56}$$

This also implies that $(\bar{y}_h, \bar{z}_h)$ is a stationary point of $\mathcal{L}_\#|_{\mathcal{Y}_{h,\#}\times\mathcal{Y}_{h,\#}}$. Moreover, we get convergence of $\bar{z}_h$ to $\bar{z}$:

$$\|\bar{z} - \bar{z}_h\|_\mathcal{Y} \leq C\left(h + \|\bar{y} - \bar{y}_h\|_\mathcal{Y}^{\alpha_F} + \|\bar{y} - \bar{y}_h\|_\mathcal{Y}\right). \tag{2.57}$$

This can be seen as follows. $D\mathcal{F}_{h,\#}(\bar{y}_h)$ satisfies two inf-sup conditions, see Proposition 2.9. For any $\zeta_h \in \mathcal{Y}_{h,\#}$ we then have:

$$\begin{aligned}
\kappa_h\|\zeta_h - \bar{z}_h\|_\mathcal{Y} &\leq \sup_{\eta_h\in\mathcal{Y}_{h,\#}} \frac{\left\langle D\mathcal{F}_{h,\#}(\bar{y}_h)\cdot\eta_h, \zeta_h - \bar{z}_h \right\rangle}{\|\eta_h\|_\mathcal{Y}} \\
&\leq \sup_{\eta_h\in\mathcal{Y}_{h,\#}} \frac{\left\langle D\mathcal{F}_{h,\#}(\bar{y}_h)\cdot\eta_h, \zeta_h - \bar{z} \right\rangle}{\|\eta_h\|_\mathcal{Y}} + \sup_{\eta_h\in\mathcal{Y}_{h,\#}} \frac{\left\langle (D\mathcal{F}_{h,\#}(\bar{y}_h) - D\mathcal{F}_\#(\bar{y}))\cdot\eta_h, \bar{z} \right\rangle}{\|\eta_h\|_\mathcal{Y}} \\
&\quad + \sup_{\eta_h\in\mathcal{Y}_{h,\#}} \frac{\left\langle D\mathcal{F}_\#(\bar{y})\cdot\eta_h, \bar{z} \right\rangle - \left\langle D\mathcal{F}_{h,\#}(\bar{y}_h)\cdot\eta_h, \bar{z}_h \right\rangle}{\|\eta_h\|_\mathcal{Y}}.
\end{aligned} \tag{2.58}$$

The third term on the right-hand side vanishes because of the definitions of $\bar{z}$ and $\bar{z}_h$. Applying the continuity property of $D\mathcal{F}_\#$ and $D\mathcal{F}_{h,\#}$ we get

$$\kappa_h\|\zeta_h - \bar{z}_h\|_\mathcal{Y} \leq \|D\mathcal{F}_{h,\#}(\bar{y}_h)\|\|\zeta_h - \bar{z}\|_\mathcal{Y} + C\|\bar{z}\|_\mathcal{Y}\left(\|\bar{y} - \bar{y}_h\|_\mathcal{Y}^{\alpha_F} + \|\bar{y} - \bar{y}_h\|_\mathcal{Y}\right).$$

With the triangle inequality $\|\bar{z} - \bar{z}_h\|_\mathcal{Y} \leq \|\bar{z} - \zeta_h\|_\mathcal{Y} + \|\zeta_h - \bar{z}_h\|_\mathcal{Y}$ we then obtain:

$$\begin{aligned}
\|\bar{z} - \bar{z}_h\|_\mathcal{Y} &\leq \left(1 + \frac{\|D\mathcal{F}_{h,\#}(\bar{y}_h)\|}{\kappa_h}\right) \inf_{\zeta_h\in\mathcal{Y}_{h,\#}} \|\zeta_h - \bar{z}\|_\mathcal{Y} + C\|\bar{z}\|_\mathcal{Y}\left(\|\bar{y} - \bar{y}_h\|_\mathcal{Y}^{\alpha_F} + \|\bar{y} - \bar{y}_h\|_\mathcal{Y}\right) \\
&\leq Ch\left(1 + \frac{\|D\mathcal{F}_{h,\#}(\bar{y}_h)\|}{\kappa_h}\right) + C\|\bar{z}\|_\mathcal{Y}\left(\|\bar{y} - \bar{y}_h\|_\mathcal{Y}^{\alpha_F} + \|\bar{y} - \bar{y}_h\|_\mathcal{Y}\right).
\end{aligned}$$

Here, we have used that $\bar{z} \in \mathrm{H}^2(\Omega) \times \mathrm{H}^2(\Omega) \times \mathbb{R}$ by elliptic regularity and the approximation property of $\mathcal{Y}_{h,\#}$. Thus, we obtain (2.57). Note, that the regularity of $\bar{z}$ is limited by the smoothness of the right-hand sides $f_u,\ f_\phi \in \mathrm{L}^2(\Omega)$ as well as the Hölder continuous part involving F''.

Step 3. Taylor expansion of the Lagrangian. From the definition of $\mathcal{L}_\#$ it directly follows that

$$\langle f, \bar{y} - \bar{y}_h \rangle = \mathcal{L}_\#(\bar{y}, \bar{z}) - \mathcal{L}_\#(\bar{y}_h, \bar{z}_h). \tag{2.59}$$

Then using Taylor's theorem we get

$$|\langle f, \bar{y} - \bar{y}_h \rangle| \leq \left|\left\langle D_y \mathcal{L}_\#(\bar{y}, \bar{z}), \bar{y} - \bar{y}_h \right\rangle\right| + \left|\left\langle D_z \mathcal{L}_\#(\bar{y}, \bar{z}), \bar{z} - \bar{z}_h \right\rangle\right| + R_{1+\alpha_F,h}. \tag{2.60}$$

The remainder term $R_{1+\alpha_F,h}$ from the Taylor expansion satisfies (see Lemma A.1)

$$\begin{aligned} R_{1+\alpha_F,h} &\leq C\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \|\bar{z} - \bar{z}_h\|_{\mathcal{Y}} \\ &\qquad + \int_0^1 \left\| D_y \mathcal{L}_\#(\bar{y} + t(\bar{y}_h - \bar{y}), \bar{z}) - D_y \mathcal{L}_\#(\bar{y}, \bar{z}) \right\| \mathrm{d}t \, \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \\ &\leq C \left(\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \|\bar{z} - \bar{z}_h\|_{\mathcal{Y}} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}^{1+\alpha_F} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}^2 \right), \end{aligned} \tag{2.61}$$

where we have used the Hölder continuity property of $D\mathcal{L}_\#$, respectively, $D\mathcal{F}_\#$. Note that $D^2_{zz}\mathcal{L}_\# = 0$ since the second component only enters $\mathcal{L}_\#$ linearly.

Since $(\bar{y}, \bar{z})$ is a stationary point of $\mathcal{L}_\#$ and $(\bar{u} - \bar{u}_h), (\bar{\phi} - \bar{\phi}_h) \in \mathrm{H}^1_\#(\Omega)$, the terms involving first derivatives in (2.60) vanish leaving us with

$$|\langle f, \bar{y} - \bar{y}_h \rangle| \leq C\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \left(h + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}^{\alpha_F} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \right).$$

Choosing the right-hand side

$$f = \frac{(\bar{u} - \bar{u}_h, \bar{\phi} - \bar{\phi}_h, 0)}{\left(\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2}^2 + \|\bar{\phi} - \bar{\phi}_h\|_{\mathrm{L}^2}^2 \right)^{1/2}} \in \mathrm{L}^2_\# \times \mathrm{L}^2_\# \times \mathbb{R}$$

(with $\|f\|_{\mathcal{Y}_\#^*} \leq 1$) proves

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} + \|\bar{\phi} - \bar{\phi}_h\|_{\mathrm{L}^2} \leq C\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \left(h + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}^{\alpha_F} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \right).$$

If $\mathrm{p}\alpha_F \geq 1$, we already get $\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} \leq Ch\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}$ and the proof is complete. If $\mathrm{p}\alpha_F < 1$, we note that

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} + \|\bar{\phi} - \bar{\phi}_h\|_{\mathrm{L}^2} \leq Ch^{\mathrm{p}(1+\alpha_F)}.$$

Step 4. Iterating the argument. Assume $\mathrm{p}\alpha_F < 1$. Having shown the increased convergence order for the L^2-errors, we can now go through the above steps again. Using the improved continuity property (2.11) of the term involving F'' and $\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} \leq Ch^{\mathrm{p}(1+\alpha_F)}$ as shown in the previous step we get for the second term on the right-hand side of (2.58):

$$\begin{aligned} \sup_{\eta_h \in \mathcal{Y}_{h,\#}} \frac{\langle (D\mathcal{F}_{h,\#}(\bar{y}_h) - D\mathcal{F}_\#(\bar{y}))\cdot\eta_h, \bar{z} \rangle}{\|\eta_h\|_{\mathcal{Y}}} &\leq C\|z\|_{\mathcal{Y}} \big(\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2}^{\alpha_F} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \big) \\ &\leq C\|z\|_{\mathcal{Y}} \big(h^{\mathrm{p}\alpha_F(1+\alpha_F)} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \big). \end{aligned}$$

This implies with (2.58) as in Step 2 that

$$\|\bar{z} - \bar{z}_h\|_{\mathcal{Y}} \leq C\big(h + h^{\mathrm{p}\alpha_F(1+\alpha_F)} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}\big).$$

Next we note that the bound (2.61) on the remainder term can be refined in the following way

$$R_{1+\alpha_F,h} \leq C\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \left(\|\bar{z} - \bar{z}_h\|_{\mathcal{Y}} + \|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2}^{\alpha_F} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}\right),$$

Inserting the convergence result for $\bar{z}_h$ and $\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} \leq Ch^{\mathrm{p}(1+\alpha_F)}$ in this equation yields

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} \leq C\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}\big(h + h^{\mathrm{p}\alpha_F(1+\alpha_F)} + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}\big) \leq C\big(h^{\mathrm{p}+1} + h^{\mathrm{p}(1+\alpha_F+\alpha_F^2)}\big).$$

If $\mathrm{p}(\alpha_F + {\alpha_F}^2) \geq 1$, the proof now is complete. If $\mathrm{p}(\alpha_F + \alpha_F^2) < 1$, we can go through Step 4 again, obtaining $\|\bar{z} - \bar{z}_h\|_{\mathcal{Y}} \leq C\big(h + h^{\mathrm{p}\alpha_F(1+\alpha_F+\alpha_F^2)}\big)$ and hence also

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} \leq C\big(h + h^{\mathrm{p}(1+\alpha_F+\alpha_F^2+\alpha_F^3)}\big).$$

Iterating this argument yields $\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} \leq Ch^{\mathrm{p}}\big(h + h^{\mathrm{p}\alpha_F/(1-\alpha_F)}\big)$ as stated. □

The Hölder exponent in the original TFDW model (1.5) is $\alpha_F = 2/3$. Even if $\mathrm{p} = 1$ the previous result shows that the L^2-errors converge at the rate $\mathcal{O}(h^2)$ in this case. The same is true for all $\alpha_F \geq \frac{1}{2}$.

2.5.1.1 Convergence in Hölder Spaces

In this short but rather technical section we provide some results that will be useful for the further study of convergence rates and also for the analysis of the discretization (3.1) of the TFDW functional with numerical integration. The main goal is a (suboptimal) convergence result for $\bar{u}_h$ and $\bar{\phi}_h$ with respect to Hölder norms.

Let us introduce the family of Hölder spaces $\mathrm{C}^{j,\gamma}(\Omega)$: for $j \in \mathbb{N}_0$ and $0 \leq \gamma < 1$ we define

$$\mathrm{C}^{j,\gamma}(\Omega) = \big\{u \in \mathrm{C}^j(\Omega) : \|u\|_{\mathrm{C}^{j,\gamma}(\Omega)} < \infty\big\},$$

with the norm

$$\|u\|_{\mathrm{C}^{j,\gamma}(\Omega)} = \|u\|_{\mathrm{C}^j(\Omega)} + |u|_{\mathrm{C}^{j,\gamma}(\Omega)},$$

where the seminorm $|u|_{\mathrm{C}^{j,\gamma}(\Omega)}$ is given by

$$|u|_{\mathrm{C}^{j,\gamma}(\Omega)} = \max_{\substack{\beta \in \mathbb{N}^d \\ |\beta|_1 = j}} \sup_{\substack{x,y \in \Omega \\ x \neq y}} \frac{|\nabla^\beta u(x) - \nabla^\beta u(y)|}{|x - y|^\gamma}.$$

A consequence of the definition of the Hölder norms and the assumed quasi-uniformity of the mesh family $(\mathcal{T}_h)_{h \in (0,1]}$ is the existence of a constant $C > 0$, independent of h and $T \in \mathcal{T}_h$, such that

$$\begin{aligned} |u|_{\mathrm{C}^{j,\gamma}(T)} &\leq Ch^{j+\gamma}|\widehat{u}|_{\mathrm{C}^{j,\gamma}(\widehat{T})} \\ |\widehat{u}|_{\mathrm{C}^{j,\gamma}(\widehat{T})} &\leq Ch^{-j-\gamma}|u|_{\mathrm{C}^{j,\gamma}(T)} \end{aligned} \tag{2.62}$$

for all $u \in \mathrm{C}^{j,\gamma}(T)$.

We point out that for this part it is not necessary to differentiate between the Dirichlet and the periodic case. Hence, we will throughout assume that $\bar{u},\ \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}(\Omega)$. The finite element space S_h is always assumed to be of order p such that the discrete solutions $\bar{u}_h$ and $\bar{\phi}_h$ satisfy $\|\bar{u} - \bar{u}_h\|_{\mathrm{H}^1} + \|\bar{\phi} - \bar{\phi}_h\|_{\mathrm{H}^1} \leq Ch^{\mathrm{p}}$.

First we show an approximation result for the interpolation operator $\mathcal{I}_h$ with respect to Hölder norms on individual elements.

Lemma 2.18. *Let $j \geq 1$ and $0 \leq \gamma < 2 - \frac{d}{2}$. There exists $C > 0$ such that*

$$\|u - \mathcal{I}_h u\|_{\mathrm{C}^{j-1,\gamma}(T)} \leq Ch_T^{2-\gamma-d/2}|u|_{\mathrm{H}^{j+1}(T)} \leq Ch_T^{2-\gamma-d/2}|u|_{\mathrm{H}^{j+1}(\Omega)},$$

for all $u \in \mathrm{H}^{j+1}(\Omega)$, $T \in \mathcal{T}_h$ and $h \in (0, 1]$.

Proof. Proposition A.12 implies the existence of $C > 0$ such that on the reference element $\widehat{T}$:

$$\|\widehat{u} - \widehat{\mathcal{I}}\widehat{u}\|_{\mathrm{C}^{j-1,\gamma}(\widehat{T})} \leq C|\widehat{u}|_{\mathrm{H}^{j+1}(\widehat{T})} \quad \forall \widehat{u} \in \mathrm{H}^{j+1}(\widehat{T}).$$

Transforming from T to the reference element, using $\widehat{\mathcal{I}_h u} = \widehat{\mathcal{I}}\widehat{u}$ and the previous equation, and transforming back to T we get with (2.62)

$$\begin{aligned}
\|u - \mathcal{I}_h u\|_{\mathrm{C}^{j-1,\gamma}(T)} &\leq Ch_T^{-j+1-\gamma}\|\widehat{u} - \widehat{\mathcal{I}_h u}\|_{\mathrm{C}^{j-1,\gamma}(\widehat{T})} \\
&\leq Ch_T^{-j+1-\gamma}|\widehat{u}|_{\mathrm{H}^{j+1}(\widehat{T})} \\
&\leq Ch_T^{-j+1-\gamma}(\det B_T)^{-1/2}\|B_T\|^{j+1}|u|_{\mathrm{H}^{j+1}(T)},
\end{aligned}$$

where we have used the transformation rule (2.48) in the last step. Since $ch_T^d \leq \det B_T \leq Ch_T^d$ and $\|B_T\| \leq Ch_T$, the result now easily follows:

$$\|u - \mathcal{I}_h u\|_{\mathrm{C}^{j-1,\gamma}(T)} \leq Ch_T^{2-\gamma-d/2}|u|_{\mathrm{H}^{j+1}(T)} = Ch_T^{2-\gamma-d/2}|u|_{\mathrm{H}^{j+1}(\Omega)}.$$

□

The next step is to prove an inverse inequality between a Hölder norm and a classical Sobolev norm over the finite element space S_h.

Lemma 2.19. *Let $0 \leq \gamma < 1$. Then, we have the inverse inequality*

$$\|v_h\|_{\mathrm{C}^{j,\gamma}(T)} \leq Ch^{-d/2-j-\gamma}\|v_h\|_{\mathrm{L}^2(T)} \leq Ch^{-d/2-j-\gamma}\|v_h\|_{\mathrm{L}^2(\Omega)},$$

for all $v_h \in \mathrm{S}_h$.

Proof. The proof is similar to [17, Lemma 4.5.3]. As usual we transform back to the reference element, use equivalence of norms over the finite dimensional polynomial space on $\widehat{T}$ and transform back:

$$\begin{aligned}\|v_h\|_{\mathrm{C}^{j,\gamma}(T)} &\leq Ch^{-j-\gamma}\|\widehat{v}_h\|_{\mathrm{C}^{j,\gamma}(\widehat{T})}\\ &\leq Ch^{-j-\gamma}\|\widehat{v}_h\|_{\mathrm{L}^2(\widehat{T})}\\ &\leq Ch^{-j-\gamma}(\det B_T)^{-1/2}\|v_h\|_{\mathrm{L}^2(T)}\\ &\leq Ch^{-j-\gamma-d/2}\|v_h\|_{\mathrm{L}^2(T)},\end{aligned}$$

where we have used condition (2.47) on B_T. ◻

Now we show that the Galerkin solutions $\bar{u}_h$, $\bar{\phi}_h$ also converge to their continuous counterparts with respect to certain Hölder norms. We mention at this point that the convergence rates stated in the lemma are certainly not optimal. They are, however, sufficient for our purposes. A detailed study of L^∞-error estimates for the present problem is beyond the scope of this thesis.

Lemma 2.20. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_\#$ be a solution of the continuous periodic problem (2.44) with $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}_\#(\Omega)$ and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,\#}$ the corresponding solution to the discrete periodic problem (2.50) satisfying $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}$. For every $\gamma \geq 0$ such that $\gamma < \min\bigl(1 + \frac{\mathrm{p}\alpha_F}{1-\alpha_F} - \frac{d}{2}, 2 - \frac{d}{2}\bigr)$ there exists a constant $C_\gamma > 0$ independent of $T \in \mathcal{T}_h$ and h such that*

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} + \|\bar{\phi} - \bar{\phi}_h\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} \leq C_\gamma\Bigl(h^{1+\frac{\mathrm{p}\alpha_F}{1-\alpha_F}-\frac{d}{2}-\gamma} + h^{2-\frac{d}{2}-\gamma}\Bigr).$$

In particular there exists $C_\gamma > 0$ independent of $T \in \mathcal{T}_h$ and h such that

$$\|\bar{u}_h\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} + \|\bar{\phi}_h\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} \leq \bar{C}_\gamma.$$

Proof. The proof is similar to an argument in [16, II.§7]. Lemma 2.18 implies

$$\|\bar{u} - \mathcal{I}_h\bar{u}\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} \leq Ch^{2-d/2-\gamma}|\bar{u}|_{\mathrm{H}^{\mathrm{p}+1}}.$$

The approximation properties of $\mathcal{I}_h$ and the convergence result of Proposition 2.17 lead to

$$\|\bar{u}_h - \mathcal{I}_h\bar{u}\|_{\mathrm{L}^2} \leq \|\bar{u}_h - \bar{u}\|_{\mathrm{L}^2} + \|\bar{u} - \mathcal{I}_h\bar{u}\|_{\mathrm{L}^2} \leq C\Bigl(h^{\mathrm{p}+1} + h^{\frac{\mathrm{p}}{1-\alpha_F}}\Bigr).$$

Using the inverse inequality $\|v_h\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} \leq Ch^{-\mathrm{p}+1-\frac{d}{2}-\gamma}\|v_h\|_{\mathrm{L}^2}$ for $v_h \in \mathrm{S}_h$ from Lemma 2.19 we then observe that

$$\begin{aligned}\|\bar{u} - \bar{u}_h\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} &\leq \|\bar{u} - \mathcal{I}_h\bar{u}\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} + Ch^{-\mathrm{p}+1-\frac{d}{2}-\gamma}\|\bar{u}_h - \mathcal{I}_h\bar{u}\|_{\mathrm{L}^2}\\ &\leq Ch^{2-\frac{d}{2}-\gamma}|\bar{u}|_{\mathrm{H}^{\mathrm{p}+1}} + C\Bigl(h^{1+\frac{\mathrm{p}\alpha_F}{1-\alpha_F}-\frac{d}{2}-\gamma} + h^{2-\frac{d}{2}-\gamma}\Bigr).\end{aligned}$$

The result for $\bar{\phi}$ follows in the same way. ◻

This result also implies that $\bar{u}_h|_T$ and $\bar{\phi}_h|_T$ are Hölder continuous with degree α_u for all

$$\alpha_u < \min\big(1 + \tfrac{\mathrm{p}\alpha_F}{1-\alpha_F} - \tfrac{d}{2}, 2 - \tfrac{d}{2}\big),$$

uniformly in $T \in \mathcal{T}_h$ and h: there exists $C > 0$ such that

$$\|\bar{u}_h\|_{\mathrm{C}^{0,\alpha_u}(T)} + \|\bar{\phi}_h\|_{\mathrm{C}^{0,\alpha_u}(T)} \leq C$$

for all $T \in \mathcal{T}_h$ and all sufficiently small h.

2.5.1.2 Convergence Order of the Lagrange Multiplier

Equipped with the L^∞-bounds just obtained we can now improve on some convergence rates if we assume higher differentiability of F on the set of function values of $\bar{u}$. In the original TFDW model, for example, $F \in \mathrm{C}^\infty(\mathbb{R}\backslash\{0\})$. Hence, if $\bar{u} > 0$, the singularity of F'' in zero has no consequences on convergence rates as soon as $\bar{u}_h \to \bar{u}$ uniformly as $h \to 0$. The next result constitutes an improved version of Proposition 2.17 given F is three-times differentiable on $I_{\bar{u}}$. Note that the condition $1 + \frac{\mathrm{p}\alpha_F}{1-\alpha_F} - \frac{d}{2} > 0$ is satisfied for all $\alpha_F > 0$ if $d = 2$ and for all $\alpha_F > \frac{1}{2\mathrm{p}}$ if $d = 3$.

Proposition 2.21. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_\#$ be a solution of the continuous periodic problem (2.44) with $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}_\#(\Omega)$ and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,\#}$ the corresponding solution to the discrete periodic problem (2.50) satisfying $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}$. Let $F \in \mathrm{C}^3(I_{\bar{u}})$ and assume $1 + \frac{\mathrm{p}\alpha_F}{1-\alpha_F} - \frac{d}{2} > 0$. Then, the L^2-errors of $\bar{u}_h$ and $\bar{\phi}_h$ satisfy*

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} + \big\|\bar{\phi} - \bar{\phi}_h\big\|_{\mathrm{L}^2} \leq Ch^{\mathrm{p}+1},$$

for sufficiently small mesh size h.

Proof. Under the given conditions Lemma 2.20 implies that for sufficiently small mesh size h we get $\bar{u}_h(\bar{\Omega}) \subset I_{\bar{u}}$. Looking at the proof of Proposition 2.17, we can now even show that $\|\bar{z} - \bar{z}_h\|_{\mathcal{Y}} \leq C(h + \|\bar{y} - \bar{y}_h\|_{\mathcal{Y}})$ instead of (2.57) since $D\mathcal{F}_{h,\#}$ (containing F'') is Lipschitz continuous on the convex hull of $\{\bar{y}, \bar{y}_h\}$. A Taylor expansion similar to (2.60) gives

$$|\langle f, \bar{y} - \bar{y}_h\rangle| \leq R_{2,h},$$

where

$$R_{2,h} \leq C\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}}(\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} + \|\bar{z} - \bar{z}_h\|_{\mathcal{Y}}) \leq Ch^{\mathrm{p}+1}.$$

The above choice of f concludes the proof. ◻

Finally, we turn to the convergence rate of the Lagrange multiplier μ_h in the periodic setting. Since neither regularity of the elliptic systems, nor Hölder continuity of the derivative $D\mathcal{F}_\#$ pose problems anymore, we get the optimal convergence rate 2p for $\bar{\mu}_h$.

Proposition 2.22. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_\#$ be a solution of the continuous periodic problem (2.44) with $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}_\#(\Omega)$ and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,\#}$ the corresponding solution to the discrete periodic problem (2.50) satisfying $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}$. Let $\mathrm{p} \in \{1, 2\}$ and $F \in \mathrm{C}^3(I_{\bar{u}})$, or $\mathrm{p} \geq 3$ and $F \in \mathrm{C}^{\mathrm{p}+1}(I_{\bar{u}})$. Moreover, let $1 + \frac{\mathrm{p}\alpha_F}{1-\alpha_F} - \frac{d}{2} > 0$. Then,*

$$|\bar{\mu} - \bar{\mu}_h| \leq Ch^{2\mathrm{p}}$$

for sufficiently small h.

Proof. The proof is only a minor modification of Proposition 2.17. This time we choose $f = (0, 0, 1) \in \mathrm{L}^2(\Omega) \times \mathrm{L}^2(\Omega) \times \mathbb{R}$ in the definition of $\mathcal{L}_\#$ from (2.54). This means that the first two components of the dual equation

$$\langle D\mathcal{F}_\#(\bar{y}) \cdot \eta, \bar{z} \rangle = \langle f, \eta \rangle$$

for $\bar{z} = (\bar{z}_u, \bar{z}_\phi, \bar{z}_\mu) \in \mathcal{Y}_\#$ represent linear elliptic equations with smooth right-hand sides:

$$\begin{aligned} -\lambda \Delta \bar{z}_u &= -\big(F''(\bar{u}) + 2\bar{\phi} + \bar{\mu}\big)\bar{z}_u + 2\bar{u}\bar{z}_\phi - \bar{z}_\mu \bar{u}, \\ -\tfrac{1}{4\pi}\Delta \bar{z}_\phi &= -2\bar{u}\bar{z}_u. \end{aligned} \tag{2.63}$$

Note that in the case of the L^2-errors the right-hand side of the dual equation contained finite element functions ($\bar{u} - \bar{u}_h$ and $\bar{\phi} - \bar{\phi}_h$), which restricted the regularity of the arising elliptic equations.

Using $\bar{u}, \bar{\phi}, \bar{z}_u, \bar{z}_\phi \in \mathrm{H}^2(\Omega) \subset \mathrm{L}^\infty$ and the assumed differentiability of F'' on the function values of $\bar{u}$, we can deduce that the right-hand sides in (2.63) belong to $\mathrm{H}^1(\Omega)$ (see product and chain rules for Sobolev functions in [52, Section 4.2.2]). Since we are considering periodic boundary conditions, this implies $\bar{z}_u, \bar{z}_\phi \in \mathrm{H}^3(\Omega)$. Iterating this argument, we deduce that $\bar{z}_u, \bar{z}_\phi \in \mathrm{H}^{\mathrm{p}+1}(\Omega)$, which in turn yields

$$\inf_{\zeta_h \in \mathcal{Y}_{h,\#}} \|\zeta_h - \bar{z}\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}.$$

Together with (2.58) this leads to

$$\|\bar{z} - \bar{z}_h\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}.$$

This time the Taylor expansion gives

$$|\bar{\mu} - \bar{\mu}_h| = |\langle f, \bar{y} - \bar{y}_h \rangle| \leq R_{2,h} \leq C\|\bar{y} - \bar{y}_h\|^2 + C\|\bar{y} - \bar{y}_h\|\|\bar{z} - \bar{z}_h\| \leq Ch^{2\mathrm{p}},$$

which is what we wanted to prove. ∎

2.5.2 Dirichlet Boundary Conditions

We will now look at the case of Dirichlet boundary conditions. Throughout this Section 2.5.2 we set $\mathrm{p} = 1$, since we can only expect $\bar{u},\, \bar{\phi} \in \mathrm{H}^2(\Omega)$.

Proposition 2.23. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}$ and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_h$ be the solutions of the continuous problem (2.18) and the discrete problem (2.34), respectively. Assume that $u_{\mathrm{ex},h} = \mathcal{I}_h u_{\mathrm{ex}}$, $\phi_{\mathrm{ex},h} = \mathcal{I}_h \phi_{\mathrm{ex}}$ and u_{ex} and ϕ_{ex} are in $\mathrm{H}^2(\Gamma)$ for all affine parts Γ of $\partial\Omega$. Then,*

$$|E(\bar{u}) - E_h(\bar{u}_h)| \leq Ch^2.$$

Proof. As in the proof of Proposition 2.16 we deduce that $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu})$ is a stationary point of $\mathcal{L}_E : \mathcal{Y}_D \to \mathbb{R}$:

$$\mathcal{L}_E(y) = \frac{\lambda}{2} \int_\Omega |\nabla u|^2 \,\mathrm{d}x + \int_\Omega F(u) \,\mathrm{d}x - \Psi(u, \phi) + \mu c(u).$$

This means that $\langle D_u \mathcal{L}_E(\bar{y}), v\rangle = 0$ and $\langle D_\phi \mathcal{L}_E(\bar{y}), v\rangle = 0$ for all $v \in \mathrm{H}^1_0(\Omega)$. The next step is the application of Taylor's theorem. Since $c(\bar{u}) = c(\bar{u}_h) = 0$, we only have to expand with respect to u and ϕ. Since $\bar{u} - \bar{u}_h \in (u_{\mathrm{ex}} - u_{\mathrm{ex},h}) + \mathrm{H}^1_0(\Omega)$ and $\bar{\phi} - \bar{\phi}_h \in (\phi_{\mathrm{ex}} - \phi_{\mathrm{ex},h}) + \mathrm{H}^1_0(\Omega)$ we now get

$$\begin{aligned} |E(\bar{u}) - E(\bar{u}_h)| &= \left|\mathcal{L}_E(\bar{y}) - \mathcal{L}_E(\bar{y}_h)\right| \\ &\leq |\langle D_u \mathcal{L}_E(\bar{y}), u_{\mathrm{ex}} - u_{\mathrm{ex,h}}\rangle| + |\langle D_\phi \mathcal{L}_E(\bar{y}), \phi_{\mathrm{ex}} - \phi_{\mathrm{ex,h}}\rangle| \\ &\quad + \tfrac{1}{2} \|D^2 \mathcal{L}_E(y_h^*)\| \, \|\bar{y} - \bar{y}_h\|^2_{\mathcal{Y}}, \end{aligned}$$

where $y_h^* \in \operatorname{conv}\{\bar{y}, \bar{y}_h\} \subset \mathcal{Y}$ such that $\|\bar{y} - y_h^*\|_{\mathcal{Y}} \leq Ch$. Here, $D_u \mathcal{L}_E(\bar{y})$ and $D_\phi \mathcal{L}_E(\bar{y})$ are interpreted as elements of $\mathrm{H}^1(\Omega)^*$. In the case of homogeneous Dirichlet boundary conditions we in fact have $(\bar{u} - \bar{u}_h), (\bar{\phi} - \bar{\phi}_h) \in \mathrm{H}^1_0(\Omega)$, and hence

$$\langle D_u \mathcal{L}_E(\bar{y}), u_{\mathrm{ex}} - u_{\mathrm{ex,h}}\rangle = \langle D_\phi \mathcal{L}_E(\bar{y}), \phi_{\mathrm{ex}} - \phi_{\mathrm{ex,h}}\rangle = 0$$

by the optimality system (2.18).

Now assume that $u_{\mathrm{ex}}|_{\partial\Omega} \neq 0$ or $\phi_{\mathrm{ex}}|_{\partial\Omega} \neq 0$. For a thorough study of the effect of boundary data approximation on the L^2-errors in the case of the Poisson equation see [7], in particular Section 7. Since $\bar{u},\, \bar{\phi} \in \mathrm{H}^2(\Omega)$, we have

$$-\lambda \Delta \bar{u} + F'(\bar{u}) + 2\bar{\phi}\bar{u} + \bar{\mu}\bar{u} = 0 \in \mathrm{L}^2(\Omega), \quad -\tfrac{1}{4\pi}\Delta\bar{\phi} - (\bar{u}^2 - \rho_n) = 0 \in \mathrm{L}^2(\Omega).$$

Hence, integration by parts yields

$$\langle D_u \mathcal{L}_E(\bar{y}), \bar{u} - \bar{u}_h\rangle = \int_{\partial\Omega} (\bar{u} - \bar{u}_h) \partial_n \bar{u} \,\mathrm{d}s, \quad \langle D_\phi \mathcal{L}_E(\bar{y}), \bar{\phi} - \bar{\phi}_h\rangle = \int_{\partial\Omega} (\bar{\phi} - \bar{\phi}_h) \partial_n \bar{\phi} \,\mathrm{d}s.$$

Using the Cauchy–Schwarz inequality and the continuity of the trace operator $\mathrm{H}^1(\Omega) \to \mathrm{L}^2(\partial\Omega)$, we get

$$\big|\langle D_u\mathcal{L}_E(\bar{y}), \bar{u} - \bar{u}_h\rangle\big| \leq \|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2(\partial\Omega)}\|\partial_n\bar{u}\|_{\mathrm{L}^2(\partial\Omega)} \leq \|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2(\partial\Omega)}\|\bar{u}\|_{\mathrm{H}^2(\Omega)}.$$

Since $\bar{u}$ was assumed H^2 regular on flat parts of $\partial\Omega$, we have the interpolation estimate $\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2(\partial\Omega)} \leq Ch^2\|\bar{u}\|_{\mathrm{H}^2(\partial\Omega)}$ and thus

$$\big|\langle D_u\mathcal{L}_E(\bar{y}), \bar{u} - \bar{u}_h\rangle\big| \leq Ch^2.$$

The same is true for $|\langle D_\phi\mathcal{L}_E(\bar{y}), \bar{\phi} - \bar{\phi}_h\rangle|$. □

Finally, we look at the L^2-errors and the Lagrange multiplier error.

Proposition 2.24. *Under the conditions of Proposition 2.23 the L^2-errors of $\bar{u}_h$ and $\bar{\phi}_h$ and the error $|\bar{\mu} - \bar{\mu}_h|$ satisfy the following:*

$$\begin{aligned}\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2} + \big\|\bar{\phi} - \bar{\phi}_h\big\|_{\mathrm{L}^2} &\leq C\big(h^2 + h^{1/(1-\alpha_F)}\big),\\ |\bar{\mu} - \bar{\mu}_h| &\leq C\big(h^2 + h^{1/(1-\alpha_F)}\big),\end{aligned}$$

where α_F is the Hölder exponent of $D\mathcal{F}$.

Proof. The proof works exactly as in Proposition 2.17 up to the equation (2.60):

$$|\langle f, \bar{y} - \bar{y}_h\rangle| \leq \big|\langle D_y\mathcal{L}(\bar{y}, \bar{z}), \bar{y} - \bar{y}_h\rangle\big| + \big|\langle D_z\mathcal{L}(\bar{y}, \bar{z}), \bar{z} - \bar{z}_h\rangle\big| + R_{1+\alpha_F,h},$$

where $\bar{z} \in \mathcal{Y}_{h,0}$ is the dual solution. The second term $\big|\langle D_z\mathcal{L}(\bar{y}, \bar{z}), \bar{z} - \bar{z}_h\rangle\big|$ on the right-hand side vanishes because $\bar{z}, \bar{z}_h \in \mathcal{Y}_0$. The term $\big|\langle D_y\mathcal{L}(\bar{y}, \bar{z}), \bar{y} - \bar{y}_h\rangle\big|$ again contains an error from the approximation of the boundary conditions. This error can be dealt with as in the proof of Proposition 2.23. The choice of the linear functional f in the Lagrange functional this time is

$$f = \frac{(\bar{u} - \bar{u}_h, \bar{\phi} - \bar{\phi}_h, 1)}{\big(\|\bar{u} - \bar{u}_h\|_{\mathrm{L}^2}^2 + \|\bar{\phi} - \bar{\phi}_h\|_{\mathrm{L}^2}^2 + 1\big)^{1/2}}.$$

The iteration process from Proposition 2.17 can be carried out in this case, too. □

Lemma 2.25. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_\#$ be a solution of the continuous problem (2.18) and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,\#}$ the corresponding solution to the discrete problem (2.34) satisfying $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq Ch$. For every $\gamma \geq 0$ such that $\gamma < \min(\frac{1}{1-\alpha_F} - \frac{d}{2}, 2 - \frac{d}{2})$ there exists a constant $C > 0$ independent of $T \in \mathcal{T}_h$ and h such that*

$$\|\bar{u} - \bar{u}_h\|_{\mathrm{C}^{0,\gamma}(T)} + \|\bar{\phi} - \bar{\phi}_h\|_{\mathrm{C}^{0,\gamma}(T)} \leq C\Big(h^{\frac{1}{1-\alpha_F} - \frac{d}{2} - \gamma} + h^{2 - \frac{d}{2} - \gamma}\Big).$$

In particular there exists $C_\gamma > 0$ independent of $T \in \mathcal{T}_h$ and h such that

$$\|\bar{u}_h\|_{\mathrm{C}^{0,\gamma}(T)} \leq C_\gamma, \qquad \|\bar{\phi}_h\|_{\mathrm{C}^{0,\gamma}(T)} \leq C_\gamma.$$

Proof. The proof works exactly as outlined in Section 2.5.1.1 for the periodic case. □

Remark 2.26. This result also implies that $\bar{u}_h|_T$ and $\bar{\phi}_h|_T$ are Hölder continuous with degree α_u for all

$$\alpha_u < \min\left(\tfrac{1}{1-\alpha_F} - \tfrac{d}{2}, 2 - \tfrac{d}{2}\right) \tag{2.64}$$

uniformly in $T \in \mathcal{T}_h$ and h: there exists $C > 0$ such that

$$\|\bar{u}_h\|_{\mathrm{C}^{0,\alpha_u}(T)} + \|\bar{\phi}_h\|_{\mathrm{C}^{0,\alpha_u}(T)} \leq C$$

for all $T \in \mathcal{T}_h$ and all sufficiently small h.

If $\frac{1}{1-\alpha_F} \geq \frac{d}{2}$ and $F \in \mathrm{C}^3(I_{\bar{u}})$, the previous result implies that $\bar{u}_h(\bar{\Omega}) \subset I_{\bar{u}}$ for sufficiently small h. Going through the proofs again as in the periodic case implies that the exponent $\frac{1}{1-\alpha_F}$ in Proposition 2.24 can be replaced with 2. □

2.6 Dependence on Nucleus Coordinates

Having studied the Thomas–Fermi–Dirac–von Weizsäcker functional for fixed nucleus coordinates, we now look at the dependence of the minimization problem (2.7) on the nucleus coordinates. Our perspective will be that the electronic density functional generates an n_{at}-body potential V for the nuclei. The physics behind this is that the nucleus-nucleus interaction is the sum of electrostatic repulsion and an attractive interaction mediated by electrons. Some theoretical results in the case of the Thomas–Fermi functional are available in [8,10]. Rigorous stability analysis for different quantum mechanical models is carried out in [27–30].

We first study continuity and differentiability properties of the many-body potential that is generated by the density functional with Dirichlet boundary conditions. Then, we address its approximability by the Galerkin discretized functional. The generalization of these ideas to a periodic domain is straightforward since there are no boundary conditions that need to be approximated.

Let the potential $V : \Omega^{n_{\mathrm{at}}} \to \mathbb{R}$ be defined by

$$V(R) = \inf_{u \in \mathrm{A}_u} E(u, R), \tag{2.65}$$

where $E(\cdot, R)$ is the energy functional (2.4) and A_u the admissible set (2.3). Note that we have reintroduced the dependence of E on the nuclear coordinates in our notation.

2.6.1 Analysis of the Potential

In the following we will analyze the potential V with regard to continuity and differentiability properties. Using very little structure of the density functional we first show that V is continuous.

Lemma 2.27. *The potential $V : \Omega^{n_{\mathrm{at}}} \to \mathbb{R}$ defined in (2.65) is upper semicontinuous.*

Proof. For every fixed $u \in \mathrm{A}_u$, the functional E is continuously partially differentiable with respect to R with the derivative

$$D_R E(u,R) = D_R \Phi(u,R) = -\int_\Omega \phi(x) D_R \rho_n(x,R)\,\mathrm{d}x\,, \tag{2.66}$$

where ϕ is the electrostatic potential corresponding to u, which satisfies $-\Delta\phi = 4\pi(u^2 - \rho_n)$, $\phi|_{\partial\Omega} = \phi_{\mathrm{ex}}$. This can be proved using ideas from Lemma 2.4 and the fact that the function $\rho_n : \mathbb{R}^{dn_{\mathrm{at}}} \to \mathrm{L}^2(\Omega)$, $R \mapsto \rho_n(\cdot, R)$ is differentiable. In particular, $E(u,\cdot)$ is continuous for every fixed $u \in \mathrm{A}_u$. From that it directly follows that V is upper semicontinuous. To show this, let $R \in \Omega^{n_{\mathrm{at}}}$ and $(R_i)_{i\in\mathbb{N}} \subset \Omega^{n_{\mathrm{at}}}$ be a sequence converging to R. Furthermore, let $u \in \arg\inf_{v\in\mathrm{A}_u} E(v,R)$ and $u_i \in \arg\inf_{v\in\mathrm{A}_u} E(v,R_i)$ for all $i \in \mathbb{N}$. Then,

$$V(R_i) = E(u_i, R_i) \le E(u, R_i).$$

Taking the lim-sup over i and recalling the continuity of $E(u,\cdot)$ for fixed u, we see that

$$\limsup_{i\in\mathbb{N}} V(R_i) \le V(R),$$

i.e., V is upper semicontinuous. ◻

Lemma 2.28. *The potential V is locally Lipschitz continuous on $\Omega^{n_{\mathrm{at}}}$.*

Proof. Let $K \subset \Omega^{n_{\mathrm{at}}}$ be a closed set. Then, V is bounded above on K, say $V(R) \le M_K$ for all $R \in K$, since V is upper semicontinuous by Lemma 2.27 and upper semicontinuous functions are bounded above on compact sets. Denote by

$$U_K = \big\{u : u \in \arg\inf_{v\in\mathrm{A}_u} E(v,R),\ R \in K\big\}$$

the set of all u that are minimizers of $E(\cdot,R)$ for an $R \in K$. Since V is bounded in K, we have $E(u,R) \le C_K$ for all $u \in U_K$ and all $R \in K$ with some $C_K > 0$. The coercivity of E for fixed R and the Sobolev Imbedding Theorem then imply the existence of $C_K > 0$ such that

$$\sup_{u\in U_K} \|u\|_{\mathrm{H}^1} \le C_K, \quad \text{and hence} \quad \sup_{u\in U_K} \big\|u^2\big\|_{\mathrm{L}^2} = \sup_{u\in U_K} \|u\|_{\mathrm{L}^4}^2 \le C \sup_{u\in U_K} \|u\|_{\mathrm{H}^1}^2 \le CC_K.$$

The norm $\|\rho(\cdot,R)\|_{\mathrm{L}^2}$ can be bounded uniformly in $R \in K$. We can deduce that the set of all solutions $\phi \in \mathrm{H}^1(\Omega)$ to $-\Delta\phi = 4\pi(u^2 - \rho_n(\cdot,R))$, $\phi|_{\partial\Omega} = \phi_{\mathrm{ex}}$, for $u \in U_K$ and $R \in K$ is bounded in $\mathrm{H}^1(\Omega)$ and consequently $\mathrm{L}^2(\Omega)$. To see this, we use the decomposition $\phi = 4\pi(-\Delta_0)^{-1}(u^2 - \rho_n(\cdot,R)) + \xi$ with $\xi \in \phi_{\mathrm{ex}} + \mathrm{H}_0^1(\Omega)$ from (2.15). Then,

$$\begin{aligned}
\|\phi\|_{\mathrm{H}^1} &\le \|4\pi(-\Delta_0)^{-1}(u^2 - \rho_n(\cdot,R))\|_{\mathrm{H}^1} + \|\xi\|_{\mathrm{H}^1} \\
&\le C\left\|u^2 - \rho_n(\cdot,R)\right\|_{\mathrm{L}^2} + C(\phi_{\mathrm{ex}}) \\
&\le C(\|u\|_{\mathrm{L}^4}^2 + \|\rho_n(\cdot,R)\|_{\mathrm{L}^2}) + C(\phi_{\mathrm{ex}}) \le C.
\end{aligned}$$

This leads to the uniform boundedness of $D_R E(u, R)$:

$$|D_R E(u, R)| \leq C \quad \text{for all } u \in U_K \text{ and } R \in K.$$

We deduce that for $u \in U_K$, $E(u, \cdot)$ is Lipschitz continuous in K with a Lipschitz constant L_K that is uniform in $u \in U_K$:

$$|E(u, R_1) - E(u, R_2)| \leq L_K |R_1 - R_2| \quad \forall u \in U_K.$$

This observation can now be used to prove that V is Lipschitz continuous on K. For the following calculation it is convenient to use the $\mathcal{O}$-notation. Let $R_1, R_2 \in K$ and $u_i \in \arg\min_{u \in \mathrm{A}_u} E(u, R_i)$, $i = 1, 2$, which means $V(R_i) = E(u_i, R_i)$. Then,

$$\begin{aligned} V(R_1) &= E(u_1, R_1) \leq E(u_2, R_1) = E(u_2, R_2) + \mathcal{O}(|R_1 - R_2|) \\ &\leq E(u_1, R_2) + \mathcal{O}(|R_1 - R_2|) = E(u_1, R_1) + \mathcal{O}(|R_1 - R_2|) \\ &= V(R_1) + \mathcal{O}(|R_1 - R_2|), \end{aligned}$$

from which we conclude that $|V(R_1) - V(R_2)| = \mathcal{O}(|R_1 - R_2|)$. Summarizing, we have shown that V is locally Lipschitz continuous on $\Omega^{n_{\mathrm{at}}}$. □

It is worth pointing out that the continuity of E with respect to u was not used in the previous proofs.

We will now show that under certain circumstances V is locally differentiable. As we will see in the proof, what can prevent V from being differentiable is a sudden jump of the global minimizer(s) of $E(\cdot, R)$. If $E(\cdot, R)$ were convex, this would be impossible. For the following result we assume that there are no multiple global minimizers apart from $\bar{u}$ and $-\bar{u}$. The function values in possible local minima of $E(\cdot, \bar{R})$ are bounded away from $E(\bar{u}, \bar{R})$. From now on we will not distinguish between u and $-u$.

Proposition 2.29. *Assume that for $\bar{R} \in \Omega^{n_{\mathrm{at}}}$, $\bar{u} \in \arg\min_{u \in \mathrm{A}_u} E(u, \bar{R})$ is a uniform minimizer, i.e., (2.22) holds with a Lagrange multiplier $\bar{\mu}$. Furthermore, assume that $\bar{u}$ and $-\bar{u}$ are the only global minimizers in the following sense: there exist $\varepsilon > 0$ and $\delta > 0$ such that*

$$E(u, \bar{R}) \geq E(\bar{u}, \bar{R}) + \delta \quad \forall u \in \mathrm{A}_u \backslash (B_\varepsilon(\bar{u}) \cup B_\varepsilon(-\bar{u})). \tag{2.67}$$

Then, V is differentiable in a neighbourhood $B_\tau(\bar{R})$ of $\bar{R}$ and the derivative is given by

$$DV(R) = D_R E(u_R, R) = -\int_\Omega \phi(x) D_R \rho_n(x, R) \, \mathrm{d}x \tag{2.68}$$

for all $R \in B_\tau(\bar{R})$, where $u_R \in \mathrm{A}_u$ is a global minimizer of $E(\cdot, R)$.

Proof. *Step 1.* We define the functional $\mathcal{L} : \mathrm{A}_u \times \mathbb{R} \times \Omega^{n_{\mathrm{at}}} \to \mathbb{R}$ by

$$\mathcal{L}(u, \mu, R) = E(u, R) + \mu c(u).$$

Then, the necessary optimality conditions (2.18) of first order in $\bar{u}$ read

$$D_{(u,\mu)}\mathcal{L}(\bar{u}, \bar{\mu}, \bar{R}) = 0.$$

Since $\bar{u}$ is a uniform minimizer, we know that $D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}, \bar{R})$ defines a coercive bilinear form on $\{v \in \mathrm{H}^1_0(\Omega) : (v, \bar{u}) = 0\}$:

$$D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}, \bar{R})\cdot[v, v] \geq \gamma_{\bar{R}} \|\nabla v\|^2_{\mathrm{L}^2} \quad \text{for all } v \in \mathrm{H}^1_0(\Omega) \text{ with } (v, \bar{u}) = 0.$$

Since $\bar{u} \neq 0$, it follows again from the main theorem on saddle point problems [18, Th. 1.1] that $D^2_{(u,\mu),(u,\mu)}\mathcal{L}(\bar{u}, \bar{\mu}, \bar{R})$ is an isomorphism from $\mathrm{H}^1_0(\Omega) \times \mathbb{R}$ to $\mathrm{H}^{-1}(\Omega) \times \mathbb{R}$. Applying the implicit function theorem [124, Th. 4.B] to $D_{(u,\mu)}\mathcal{L}$, we deduce that there are continuously differentiable functions $u : \Omega^{n_{\mathrm{at}}} \to \mathrm{A}_u$ and $\mu : \Omega^{n_{\mathrm{at}}} \to \mathbb{R}$ defined on a neighbourhood $B_{\varepsilon_1}(\bar{R})$ of $\bar{R}$ such that $u(\bar{R}) = \bar{u}$, $\mu(\bar{R}) = \bar{\mu}$, and

$$D_{(u,\mu)}\mathcal{L}\big(u(R), \mu(R), R\big) = 0 \quad \text{for } R \in B_{\varepsilon_1}(\bar{R}). \tag{2.69}$$

The idea now is to show that $u(R)$ is the global minimizer of $E(\cdot, R)$ if R is sufficiently close to $\bar{R}$, in other words $V(R) = E(u(R), R)$. To this end, we will use sufficient optimality conditions.

Step 2. Let $R_0 \in \Omega^{n_{\mathrm{at}}}$ be sufficiently close to $\bar{R}$ in a sense to be made more precise. We will first show that $u(R_0)$ is a *local* minimizer of $E(\cdot, R_0)$. Define $u_0 = u(R_0)$ and $\mu_0 = \mu(R_0)$. From the continuity of $D^2_{uu}\mathcal{L}$ we deduce that

$$\big\|D^2_{uu}\mathcal{L}(\bar{u}, \bar{\mu}, \bar{R}) - D^2_{uu}\mathcal{L}(u_0, \mu_0, R_0)\big\| \leq \gamma_{\bar{R}}/4 \tag{2.70}$$

provided $|R_0 - \bar{R}|$ is sufficiently small. Let $\bar{V}$ and V_0 denote the tangent spaces corresponding to the constraint c in $\bar{u}$, respectively, u_0:

$$\bar{V} = \{v \in \mathrm{H}^1_0(\Omega) : \ \|\nabla v\|_{\mathrm{L}^2} = 1, \ (v, \bar{u}) = 0\},$$
$$V_0 = \{v \in \mathrm{H}^1_0(\Omega) : \ \|\nabla v\|_{\mathrm{L}^2} = 1, \ (v, u_0) = 0\}.$$

Elements of V_0 can be approximated by elements of $\bar{V}$ in the following sense: there exists a constant $C(u_0)$ such that

$$\inf_{w \in \bar{V}} \|\nabla(v_0 - w)\|_{\mathrm{L}^2} \leq C(u_0) \|\nabla(\bar{u} - u_0)\|_{\mathrm{L}^2} \quad \text{for all } v_0 \in V_0. \tag{2.71}$$

This can be shown with similar methods as used in the proof of Lemma A.7. Now, let $v_0 \in V_0$. We can make the following rearrangements

$$\begin{aligned} D^2_{uu}\mathcal{L}(R_0)\cdot[v_0, v_0] = \ & \big(D^2_{uu}\mathcal{L}(R_0) - D^2_{uu}\mathcal{L}(\bar{R})\big)\cdot[v_0, v_0] \\ & + D^2_{uu}\mathcal{L}(\bar{R})\cdot[\bar{v} + v_0 - \bar{v}, \bar{v} + v_0 - \bar{v}] \end{aligned}$$

with $\bar{v} \in \bar{V}$, where we have compressed the arguments of $\mathcal{L}$ in an obvious way. The first term on the right-hand side is small by (2.70). Using (2.71) we get for the second term

$$D^2_{uu}\mathcal{L}(\bar{R})\cdot[\bar{v}+v_0-\bar{v},\bar{v}+v_0-\bar{v}] = D^2_{uu}\mathcal{L}(\bar{R})\cdot[\bar{v},\bar{v}] + \mathcal{O}(\|\nabla(u_0-\bar{u})\|_{\mathrm{L}^2}),$$

where $\bar{v}$ was chosen appropriately. Hence, we can deduce that for sufficiently small $|\bar{R}-R_0|$

$$D^2_{uu}\mathcal{L}(u_0,\mu_0,R_0)\cdot[v_0,v_0] \geq \gamma_{\bar{R}}/2\,\|\nabla v\|^2_{\mathrm{L}^2} \quad \forall v_0 \in V_0.$$

Then, Theorem 5.6 in [89] about sufficient optimality conditions guarantees that $u_0 = u(R_0)$ is a uniform local minimizer of $E(\cdot,R_0)$, for all R_0 with $|R_0-\bar{R}|$ sufficiently small. Moreover, looking at the proof of the theorem in more detail, we observe that there exist $\beta > 0$ and $\varepsilon_2 > 0$ such that

$$E(v,R_0) \geq E(u_0,R_0) + \beta\,\|\nabla(u_0-v)\|^2_{\mathrm{L}^2} \quad \forall v \in B_{\varepsilon_2}(u_0)\cap \mathrm{A}_u, \tag{2.72}$$

for all R_0 with sufficiently small $|R_0-\bar{R}|$. For $\bar{R}$ we can combine this with condition (2.67) to obtain

$$E(v,\bar{R}) \geq E(\bar{u},\bar{R}) + \min\bigl(\bar{\delta},\beta\|\nabla(\bar{u}-v)\|^2_{\mathrm{L}^2}\bigr) \quad \forall v \in \mathrm{A}_u, \tag{2.73}$$

where $\bar{\delta} > 0$ is sufficiently small.

Step 3. We still need to show that $u(R)$ is in fact the global minimizer of $E(\cdot,R)$ if $|R-\bar{R}|$ is sufficiently small. Let $R_0 \in \Omega^{n_{\mathrm{at}}}$ and $u_0 \in \arg\inf_{v\in\mathrm{A}_u} E(\cdot,R_0)$. From the local Lipschitz continuity of V we know that

$$E(u_0,R_0) = V(R_0) = V(\bar{R}) + \mathcal{O}(|R_0-\bar{R}|) = E(\bar{u},\bar{R}) + \mathcal{O}(|R_0-\bar{R}|).$$

On the other hand we have

$$E(u_0,R_0) = E(u_0,\bar{R}) + \mathcal{O}(|R_0-\bar{R}|),$$

which implies $E(u_0,\bar{R}) \to E(\bar{u},\bar{R})$ as $R_0 \to \bar{R}$. Note that the constants entering $\mathcal{O}(|R_0-\bar{R}|)$ are locally bounded in R_0 as shown in Lemma 2.28.

From (2.73) we now deduce that $u_0 \to \bar{u}$ as $R_0 \to \bar{R}$. However, it was obtained above that $u(R_0)$ is a uniform local minimizer of $E(\cdot,R_0)$ and, using (2.72), we deduce that $u_0 = u(R_0)$ for sufficiently small $|R_0-\bar{R}|$. Summarizing, we get

$$V(R) = E(u(R),R)$$

if $|R-\bar{R}|$ is sufficiently small, say $R \in B_\tau(\bar{R})$.

Step 4. The potential V is differentiable in $B_\tau(\bar{R})$ since E and $u(R)$ are. We interpret the derivative $D_R u(R)$ as a vector of dn_{at} functions in $\mathrm{H}^1_0(\Omega)$: $D_R u(R) \in \mathrm{H}^1_0(\Omega,\mathbb{R}^{dn_{\mathrm{at}}})$. To calculate the derivative DV in R we note that

$$E(u(R),R) = E(u(R),R) + \mu(R)c(u(R)) = \mathcal{L}(u(R),\mu(R),R) =: \widetilde{\mathcal{L}}(R),$$

because $c(u(R)) = 0$. From (2.69) we know for sufficiently small distance $|\bar{R} - R_0|$ that $D_u\mathcal{L}(u(R), \mu(R), R) = 0 \in \mathrm{H}^{-1}(\Omega)$ and $D_\mu\mathcal{L}(u(R), \mu(R), R) = 0$. Therefore

$$\begin{aligned} DV(R) &= D\widetilde{\mathcal{L}}(R) \\ &= D\mathcal{L}(u(R), \mu(R), R) \cdot [D_R u(R)\ D_R\mu(R)\ 1]^T \\ &= D_R\mathcal{L}(u(R), \mu(R), R) \\ &= D_R E(u(R), R), \end{aligned}$$

which is what we wanted to prove. □

Remark 2.30. In the previous result we saw that if it exists, the derivative of V equals the partial derivative of E with respect to R. This is in fact a version of the Hellman–Feynman theorem from quantum mechanics, see [55, Chapter 3]. Since an energy minimum of $E(\cdot, R)$ is calculated, the variation of E with respect to u vanishes. □

Under higher differentiability assumptions on F the proof above can be extended to show that the second derivative of V is given by

$$D^2V(R) = \int_\Omega \phi(R) D^2_{RR}\rho_n(x; R)\,\mathrm{d}x + \int_\Omega D_R\phi(R) \otimes D_R\rho_n(x; R)\,\mathrm{d}x\,. \tag{2.74}$$

The derivative $D_R y(R) = \big[D_R u(R),\, D_R\phi(R),\, D_R\mu(R)\big]^T$ can be calculated using the Implicit Function Theorem [124, Theorem 4.B]. Informally, by differentiating $\mathcal{F}(y(R), R) = 0$ we arrive at

$$D_y\mathcal{F}(y(R), R) D_R y(R) = -D_R\mathcal{F}(y(R), R). \tag{2.75}$$

The derivative $D_R y(R)$ is a linear operator in $\mathrm{Lin}(\mathbb{R}^{dn_{\mathrm{at}}}, \mathcal{Y}_0)$, which we will identify with $\mathcal{Y}_0^{dn_{\mathrm{at}}}$.

In direct analogy with the potential V from (2.65) we can define the discretized potential $V_h : \Omega^{n_{\mathrm{at}}} \to \mathbb{R}$ by

$$V_h(R) = \inf_{u \in \mathrm{A}_{u,h}} E_h(u, R).$$

All of the results we have obtained for V hold for V_h in the same form. The following result shows that the sufficient condition (2.67) for the differentiability of V in a point $\bar{R}$ implies an analogous condition for V_h given that h is sufficiently small.

Proposition 2.31. *Let $\bar{R} \in \Omega^{n_{\mathrm{at}}}$ and assume that $\bar{u} \in \mathrm{H}^2(\Omega)$ is a uniform, unique global minimizer of $E(\cdot, \bar{R})$ that satisfies the condition* (2.67)*. Then, for sufficiently small h there exists a uniform minimizer $\bar{u}_h \in \mathrm{A}_{u,h}$ of $E_h(\cdot, \bar{R})$ with $\|\bar{u} - \bar{u}_h\|_{\mathrm{H}^1} \leq Ch$. Moreover, $\bar{u}_h$ is the unique global minimizer of $E_h(\cdot, \bar{R})$ and*

$$E_h(u, \bar{R}) \geq E_h(\bar{u}_h, \bar{R}) + \delta_d \quad \forall u \in \mathrm{A}_{u,h} \backslash \big(B_{\varepsilon_d}(\bar{u}_h) \cup B_{\varepsilon_d}(-\bar{u}_h)\big) \tag{2.76}$$

for sufficiently small h, where $\varepsilon_d > 0$ and $\delta_d > 0$ are independent of h.

Proof. Throughout the proof we will suppress the argument $\bar{R}$ in E and E_h. The existence and uniformity of $\bar{u}_h$ were already proved in Theorem 2.10 and Proposition 2.11. The goal of this proof is to show a discrete analogue of (2.73), which will imply that $\bar{u}_h$ is the unique global minimizer in the sense of (2.76).

Let $u_h^* \in \mathrm{A}_{u,h}$. Then by (2.73)

$$\begin{aligned} E(u_h^*) &\geq E(u_h^* - u_{\mathrm{ex},h} + u_{\mathrm{ex}}) - \mathcal{O}(h) \\ &\geq E(\bar{u}) + \min\bigl(\beta \bigl\|\bar{u} - u_h^*\bigr\|_{\mathrm{H}^1}^2 - \mathcal{O}(h), \bar{\delta}\bigr) - \mathcal{O}(h) \\ &\geq E_h(\bar{u}_h) - \mathcal{O}(h^2) + \min\bigl(\beta \bigl\|\bar{u}_h - u_h^*\bigr\|_{\mathrm{H}^1}^2 - \mathcal{O}(h), \bar{\delta}\bigr) - \mathcal{O}(h) \\ &\geq E_h(\bar{u}_h) - \mathcal{O}(h) + \min\bigl(\beta \bigl\|\bar{u}_h - u_h^*\bigr\|_{\mathrm{H}^1}^2, \bar{\delta}/2\bigr) \end{aligned} \tag{2.77}$$

for sufficiently small h. On the other hand we have

$$\bigl|E(u_h^*) - E_h(u_h^*)\bigr| \leq \frac{1}{2} \int_\Omega \bigl|(u_h^*)^2 - \rho_n\bigr| \, \bigl|\phi^* - \phi_h^*\bigr| \, \mathrm{d}x \,, \tag{2.78}$$

where ϕ^* satisfies $-\Delta\phi^* = 4\pi\bigl((u_h^*)^2 - \rho_n\bigr)$, $\phi^*|_{\partial\Omega} = \phi_{\mathrm{ex}}$ and ϕ_h^* satisfies $-\Delta_h\phi_h^* = 4\pi\bigl((u_h^*)^2 - \rho_n\bigr)$, $\phi_h^*|_{\partial\Omega} = \phi_{\mathrm{ex},h}$. It follows that

$$\begin{aligned} \bigl\|\phi_h^* - \phi^*\bigr\|_{\mathrm{H}^1} &\leq Ch\bigl(\|u_h^*\|_{\mathrm{L}^4}^2 + \|\rho_n\|_{\mathrm{L}^2}\bigr) + C\|\phi_{\mathrm{ex}} - \phi_{\mathrm{ex},h}\|_{\mathrm{H}^1} \\ &\leq Ch\bigl(\|u_h^*\|_{\mathrm{H}^1}^2 + \|\rho_n\|_{\mathrm{L}^2}\bigr) + Ch. \end{aligned} \tag{2.79}$$

Since E_h is coercive, $\|u_h^*\|_{\mathrm{H}^1}$ is bounded for all $u_h^* \in \mathrm{S}_h$ such that, say, $E_h(u_h^*) \leq E(\bar{u}) + \bar{\delta}$. This holds uniformly in h. Combining this with (2.77), (2.78), and (2.79) we obtain

$$E_h(u_h^*) \geq E_h(\bar{u}_h) - \mathcal{O}(h) + \min\bigl(\beta \bigl\|\bar{u}_h - u_h^*\bigr\|_{\mathrm{H}^1}^2, \bar{\delta}/2\bigr) \quad \forall u_h^* \in \mathrm{A}_{u,h},$$

which implies (2.76) for sufficiently small δ_d. □

2.6.2 Convergence of Minimizing Configurations

In this section we will look at the approximability of minimizers of V by minimizers of V_h. In order to avoid some technical difficulties we will restrict the analysis to the periodic setting. In a Dirichlet setting, for example, the translation symmetry of V is lost, which leads to complications. We will prove that if V has a minimizer $\bar{R} \in \Omega^{n_{\mathrm{at}}}$, then V_h has a minimizer $\bar{R}_h \in \Omega^{n_{\mathrm{at}}}$ with $|\bar{R} - \bar{R}_h| \leq Ch^{\mathrm{p}}$. Most of the techniques are very similar to the ones used previously, so we will keep the presentation rather tight. Since we need second derivatives of V, we assume for simplicity that $F \in \mathrm{C}^{\min(3,\mathrm{p})}(\mathbb{R})$.

We will from now on read every $R \in \Omega^{n_{\mathrm{at}}}$ as a vector from $\mathbb{R}^{dn_{\mathrm{at}}}$ with the n_{at} first components of the nucleus positions first, then the second components and so forth:

$$R = \bigl[R_{1,1} \, \ldots \, R_{n_{\mathrm{at}},1} \, R_{1,2} \, \ldots \, R_{n_{\mathrm{at}},2} \ldots R_{n_{\mathrm{at}},d}\bigr]^T.$$

Because of the periodic nature of the problem, the potential V has a translation symmetry, which can be written as

$$DV(R)\cdot E_i = 0 \quad \forall i \in \{1,\dots,d\},$$

where $E_i \in \mathbb{R}^{dn_{\mathrm{at}}}$, $i \in \{1,\dots,d\}$ is the vector with ones in all i-components and zeros elsewhere. Because of this symmetry we will look for minimizers of V in the space

$$\mathcal{R}_m = \left\{ R \in \Omega^{n_{\mathrm{at}}} : \sum_{j=1}^{n_{\mathrm{at}}} R_{j,i} = m_i \quad \forall i \in \{1,\dots,d\} \right\}$$

of configurations whose center of mass is the center $m = (m_1,\dots,m_d)$ of Ω. We denote by

$$\mathcal{R}_0 = \left\{ R \in \mathbb{R}^{n_{\mathrm{at}}} : \sum_{j=1}^{n_{\mathrm{at}}} R_{j,i} = 0 \quad \forall i \in \{1,\dots,d\} \right\}$$

the tangential space of $\mathcal{R}_m$. The affine constraints on $R \in \mathcal{R}_m$ can also be written more compactly as $AR = m$, where $A \in \mathbb{R}^{d\times dn_{\mathrm{at}}}$ is an appropriate matrix of full rank containing only ones and zeros:

$$A = \begin{bmatrix} E_1^T \\ \vdots \\ E_d^T \end{bmatrix}.$$

Theorem 2.32. *Let $\bar{R} \in \mathcal{R}_m$. Moreover, let (2.67) be satisfied and $\bar{R} \in \mathcal{R}_m$ be a uniform minimizer of $V : \mathcal{R}_m \to \mathbb{R}$. Then, for sufficiently small h there exists a uniform minimizer $\bar{R}_h \in \mathcal{R}_m$ of $V_h|_{\mathcal{R}_m}$ such that*

$$|\bar{R} - \bar{R}_h| \le Ch^{\mathrm{p}}.$$

Proof. Since very similar arguments were given above, we only sketch the proof. If $\bar{R} \in \mathcal{R}_m$ is a uniform minimizer of V, then it satisfies $DV(\bar{R}) = 0$, which we can rewrite as the saddle-point problem

$$\left.\begin{array}{rl} DV(\bar{R}) + A^T\bar{s} & = 0, \\ A\bar{R} & = m \end{array}\right\} \qquad \mathcal{G}(\bar{R},\bar{s}) = 0,$$

with the Lagrange multiplier $\bar{s} = 0 \in \mathbb{R}^d$ for the linear constraint. Note that A has full rank. Moreover, the uniformity condition reads $D^2V(\bar{R})\cdot[H,H] \ge \gamma|H|^2$ for all $H \in \mathcal{R}_0$, where $\gamma > 0$.

The goal is now to show the existence of a minimizer $\bar{R}_h \in \mathcal{R}_m$ of $V_h|_{\mathcal{R}_m}$ that approximates $\bar{R}$. This discrete minimizer will satisfy the optimality system

$$\left.\begin{array}{rl} DV_h(\bar{R}_h) + A^T\bar{s}_h & = 0, \\ A\bar{R}_h & = m \end{array}\right\} \qquad \mathcal{G}_h(\bar{R}_h,\bar{s}_h) = 0,$$

with an appropriate Lagrange multiplier $\bar{s}_h \in \mathbb{R}^d$. The discretized potential V_h is not necessarily translation symmetric due to the discreteness of the underlying computational mesh. Hence, we expect that $\bar{s}_h \neq 0$ but small.

The procedure of the existence and convergence proof is clear. We begin by observing that $|\mathcal{G}_h(\bar{R}, \bar{s})| \leq Ch^{\mathrm{p}}$. Indeed,

$$\left|DV_h(\bar{R}) + A^T\bar{s}\right| = \left|DV_h(\bar{R})\right| = \left|DV_h(\bar{R}) - DV(\bar{R})\right| \leq Ch^{\mathrm{p}}$$

and $|A\bar{R}| = 0$. Moreover, by the form (2.74) of D^2V we have

$$\begin{aligned}|D^2V(\bar{R}) - D^2V_h(\bar{R})| \leq\ & \left\|\phi(\bar{R}) - \phi_h(\bar{R})\right\|_{\mathrm{L}^2} \left\|D_R^2\rho_n(\cdot, \bar{R})\right\|_{\mathrm{L}^2} \\ & + \left\|D_R\phi(\bar{R}) - D_R\phi_h(\bar{R})\right\|_{\mathrm{L}^2} \left\|D_R\rho_n(\cdot, \bar{R})\right\|_{\mathrm{L}^2}.\end{aligned}$$

From the equation (2.75) and its discrete counterpart we can derive that

$$\left\|D_Ry(R) - D_Ry_h(R)\right\|_{\mathcal{Y}^{dn_{\mathrm{at}}}} \leq Ch^{\mathrm{p}}.$$

This follows by an argument very similar to the convergence proof for the dual variables $\bar{z}$, $\bar{z}_h$ in Proposition 2.17. Hence, we deduce that

$$|D^2V(\bar{R}) - D^2V_h(\bar{R})| \leq Ch^{\mathrm{p}},$$

which, with a finite dimensional version of Theorem 1.1 in [18] on saddle-point problems, implies that the matrix

$$D\mathcal{G}_h(\bar{R}, \bar{s}) = \begin{bmatrix} D^2V_h(\bar{R}) & A^T \\ A & 0 \end{bmatrix}$$

is invertible for sufficiently small h and in fact in an open neighbourhood of $\bar{R}$. A proof along the lines of the existence proof in Theorem 2.10 then shows the existence of a minimizer $\bar{R}_h \in \mathcal{R}$ of V_h such that $|\bar{R} - \bar{R}_h| \leq Ch^{\mathrm{p}}$. □

This concludes our analysis of the discretization of the TFDW functional assuming exactly computed integrals. In the following chapter we investigate effects of numerical integration and interpolation to obtain practically relevant convergence results.

Chapter 3

Discretization of the Density Functional with Quadrature

In the previous chapter we analyzed discretizations of the TFDW functional that were of rather theoretical nature in that all integrals involved were assumed to be computed exactly. In the present chapter we remove this assumption and look at discretizations that take into account numerical integration in the finite element case, respectively, interpolation in the Fourier case. After showing existence and convergence of a solution for this discretization we study optimal convergence rates. The chapter closes with some numerical examples.

If we use approximations for the integrals in the discretized system (2.34) that can not easily be computed exactly, we obtain a system of the form

$$\begin{aligned} \lambda(\nabla\widetilde{u}_h, \nabla v) + \mathcal{Q}_h\big[F'(\widetilde{u}_h)v\big] + 2\mathcal{Q}_h\big[\widetilde{\phi}_h\widetilde{u}_h v\big] + \mu_h\mathcal{Q}_h[\widetilde{u}_h v] = 0 \quad \forall v \in \mathrm{S}_{h,0}, \\ \frac{1}{4\pi}(\nabla\widetilde{\phi}_h, \nabla\psi) - \mathcal{Q}_h\big[(\widetilde{u}_h^2 - \rho_n)\psi\big] = 0 \quad \forall \psi \in \mathrm{S}_{h,0}, \\ \frac{\nu}{2}\left(\mathcal{Q}_h\big[\widetilde{u}_h^2\big] - n_{\mathrm{el}}\right) = 0 \quad \forall \nu \in \mathbb{R} \end{aligned} \tag{3.1}$$

for $\widetilde{y}_h = (\widetilde{u}_h, \widetilde{\phi}_h, \widetilde{\mu}_h) \in \mathcal{Y}_{h,D}$, where $\widetilde{u}_h|_{\partial\Omega} = u_{\mathrm{ex},h}$ and $\widetilde{\phi}_h|_{\partial\Omega} = \phi_{\mathrm{ex},h}$. Here, $\mathcal{Q}_h[\cdot]$ represents a quadrature rule that operates on each element $T \in \mathcal{T}_h$ independently. The precise form of $\mathcal{Q}_h$ will be given below. We write (3.1) as

$$\widetilde{\mathcal{F}}_h(\widetilde{u}_h, \widetilde{\phi}_h, \widetilde{\mu}_h) = 0, \tag{3.2}$$

with the nonlinear map $\widetilde{\mathcal{F}}_h : \mathcal{Y}_{h,D} \to \mathcal{Y}_{h,0}^*$.

For $u_h \in \mathrm{S}_h$ we define the discrete energy with quadrature in the following way:

$$\widetilde{E}_h(u_h) = \frac{\lambda}{2}\int_\Omega |\nabla u_h|^2 \,\mathrm{d}x + \mathcal{Q}_h[F(u_h)] + \widetilde{\Phi}_h(u_h). \tag{3.3}$$

The Coulomb energy $\widetilde{\Phi}$ has the familiar form

$$\begin{aligned}\widetilde{\Phi}_h(u_h) = & -\inf_{\phi_h \in \phi_{\mathrm{ex},h} + \mathrm{S}_{h,0}} \widetilde{\Psi}_h(u_h, \phi_h) \\ = & -\inf_{\phi_h \in \phi_{\mathrm{ex},h} + \mathrm{S}_{h,0}} \Big[\frac{1}{8\pi} \int_\Omega |\nabla \phi_h|^2 \,\mathrm{d}x - \mathcal{Q}_h\big[(u_h^2 - \rho_n)\phi_h\big] \Big].\end{aligned}$$

Similarly as in the continuous and the Galerkin case we get $\widetilde{\Phi}(u_h) = \widetilde{\Psi}(u_h, \phi_h^*)$ where ϕ_h^* solves the variational problem: find $\phi_h \in \mathrm{S}_h$ such that

$$\begin{aligned}(\nabla \phi_h, \nabla v_h) &= 4\pi \mathcal{Q}_h\big[(u_h^2 - \rho_n) v_n\big] \quad \forall v_h \in \mathrm{S}_{h,0}, \\ \phi_h^*|_{\partial\Omega} &= \phi_{\mathrm{ex},h}.\end{aligned}$$

The constraint functional c is approximated by $\widetilde{c}_h : \mathrm{S}_h \to \mathbb{R}$,

$$\widetilde{c}_h(u) = \frac{1}{2}\big(\mathcal{Q}_h\big[u^2\big] - n_{\mathrm{el}}\big). \tag{3.4}$$

It follows as in the Galerkin case that (3.1) is the optimality condition for the minimization problem

$$\min\big\{\widetilde{E}_h(u) : \quad u \in u_{\mathrm{ex},h} + \mathrm{S}_{h,0}, \quad \widetilde{c}_h(u) = 0\big\}.$$

Analogous approximations of integrals in the periodic finite element discretization (2.44) lead to the equation

$$\widetilde{\mathcal{F}}_{h,\#}(\widetilde{y}_h) = 0, \quad \text{where } \widetilde{\mathcal{F}}_{h,\#} : \mathcal{Y}_{h,\#} \to \mathcal{Y}_{h,\#}^*. \tag{3.5}$$

The periodic energy with quadrature takes the form

$$\widetilde{E}_{h,\#}(u_h) = \frac{\lambda}{2} \int_\Omega |\nabla u_h|^2 \,\mathrm{d}x + \mathcal{Q}_h[F(u_h)] + \widetilde{\Phi}_{h,\#}(u_h), \tag{3.6}$$

where

$$\begin{aligned}\widetilde{\Phi}_{h,\#}(u_h) = & -\inf_{\phi_h \in \mathrm{S}_{h,\#,0}} \widetilde{\Psi}_h(u_h, \phi_h) \\ = & -\inf_{\phi_h \in \mathrm{S}_{h,\#,0}} \Big[\frac{1}{8\pi} \int_\Omega |\nabla \phi_h|^2 \,\mathrm{d}x - \mathcal{Q}_h\big[(u_h^2 - \rho_n)\phi_h\big] \Big].\end{aligned}$$

Since some of the following considerations will be quite technical, we give a brief overview of the essential steps. The strategy for the existence and convergence proof of the discretization (3.1) is in fact a purely discrete analogue of the previous chapter. From the existence of the Galerkin solution $\bar{y}_h \in \mathcal{Y}_{h,D}$ to $\mathcal{F}_h(\bar{y}_h) = 0$ we will deduce the existence of a solution $\widetilde{y}_h \in \mathcal{Y}_{h,D}$ to $\widetilde{\mathcal{F}}_h(\widetilde{y}_h) = 0$. The main difference is that we now have to deal with an approximation $\widetilde{\mathcal{F}}_h$ of the nonlinear operator $\mathcal{F}_h$ (variational crimes) while the space $\mathcal{Y}_{h,D}$ both operators are defined on is the same. In the Galerkin case in Chapter 2 the operators $\mathcal{F}$, $\mathcal{F}_h$ were essentially of the same form but defined on different spaces ($\mathcal{Y}_{h,D}$ approximated $\mathcal{Y}_D$).

There are again three main steps involved: 1. showing that the Galerkin solution $\bar{y}_h$ is an approximate solution of (3.2), that is, $\|\widetilde{\mathcal{F}}_h(\bar{y}_h)\| \to 0$ as $h \to 0$; 2. ensuring that $D\widetilde{\mathcal{F}}_h$ is stable in a sufficiently large neighbourhood of $\bar{y}_h$; 3. applying the Inverse Function Theorem to deduce the existence of $\widetilde{y}_h$ such that $\widetilde{\mathcal{F}}_h(\widetilde{y}_h) = 0$ and $\|\bar{y} - \widetilde{y}_h\| \leq Ch^{\mathrm{p}}$.

3.1 Quadrature Rules

Let $\widehat{\mathcal{Q}} : \mathrm{C}^0(\overline{\widehat{T}}) \to \mathbb{R}$ be a quadrature rule on the reference element $\widehat{T}$. We will assume that $\widehat{\mathcal{Q}}$ has the form

$$\widehat{\mathcal{Q}}[\widehat{v}] = \sum_{\nu=1}^{n_{\mathcal{Q}}} \omega_\nu \widehat{v}(\widehat{x}_\nu) \quad \text{for } v \in \mathrm{C}^0(\overline{\widehat{T}}),$$

where $\{\omega_\nu\}_{\nu\in\{1,\dots,n_{\mathcal{Q}}\}}$ are the weights and $\{\widehat{x}_\nu\}_{\nu\in\{1,\dots,n_{\mathcal{Q}}\}}$ the nodes of the quadrature rule. The weights satisfy

$$\sum_{\nu=1}^{n_{\mathcal{Q}}} \omega_\nu = |\widehat{T}|$$

and we will assume that $\omega_\nu > 0$ for all $\nu \in \{1, \dots, n_{\mathcal{Q}}\}$, since negative weights can, for example, lead to negative integral approximations for strictly positive integrands.

From $\widehat{\mathcal{Q}}$ we can now easily derive quadrature rules $\mathcal{Q}_T : \mathrm{C}^0(\overline{T}) \to \mathbb{R}$ for all $T \in \mathcal{T}_h$ by

$$\mathcal{Q}_T[v] = \det B_T \, \widehat{\mathcal{Q}}\big[v \circ F_T\big] = \det B_T \sum_{\nu=1}^{n_{\mathcal{Q}}} \omega_\nu v(x_\nu^{(T)}).$$

The nodes $x_\nu^{(T)} = F_T(\widehat{x}_\nu)$, $\nu \in \{1, \dots, n_{\mathcal{Q}}\}$, on T are obtained from the reference nodes via the mapping $F_T : \widehat{T} \to T$ defined in (2.45). For convenience, we will assume that the $\mathcal{Q}_T$ are all derived from the same quadrature rule $\widehat{\mathcal{Q}}$.

The quadrature operator $\mathcal{Q}_h : \mathrm{C}^0(\overline{\Omega}) \to \mathbb{R}$ for functions defined in the whole of Ω is now constructed by applying the quadrature rules $\mathcal{Q}_T$ on every element $T \in \mathcal{T}_h$. Let $v \in \mathrm{C}^0(\overline{\Omega})$, then

$$\int_\Omega v \,\mathrm{d}x \approx \mathcal{Q}_h[v] = \sum_{T\in\mathcal{T}_h} \mathcal{Q}_T[v] = \sum_{T\in\mathcal{T}_h} \det B_T \sum_{\nu=1}^{n_{\mathcal{Q}}} \omega_\nu v(x_\nu^{(T)}).$$

Let us define error functionals on an element T of $\mathcal{T}_h$ and on Ω by:

$$e_T[g] = \int_T g \,\mathrm{d}x - \mathcal{Q}_T[g]\,, \qquad e_h[g] = \int_\Omega g \,\mathrm{d}x - \mathcal{Q}_h[g]\,, \tag{3.7}$$

for any $g \in \mathrm{C}^0(\overline{\Omega})$. On the reference element we define for $g \in \mathrm{C}^0(\overline{\widehat{T}})$

$$\widehat{e}\,[\widehat{g}] = \int_{\widehat{T}} \widehat{g} \,\mathrm{d}x - \widehat{\mathcal{Q}}\,[\widehat{g}]\,. \tag{3.8}$$

A collection of useful results regarding quadrature errors can be found in Section A.3 in the Appendix.

3.2 Existence and Convergence of Numerical Solutions

The goal of this section is to prove existence and convergence of solutions to the systems (3.1) and (3.5). The proofs rely mainly on the careful analysis of quadrature errors. For these, it is not essential to distinguish between the Dirichlet and the periodic case. We will state all results for both cases but the proofs will frequently be kept generic. Differences between the two cases will be pointed out when necessary. We will use the following standard assumptions, which will not be stated every time:

Dirichlet case (indicated by **(D)**): $\bar{u} \in \mathrm{A}_u$ is a uniform minimizer of (2.7). $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_D$ is the corresponding solution to (2.18) and $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,D}$ is the respective solution to (2.34) such that $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq Ch$. Moreover, we assume $\frac{1}{1-\alpha_F} \geq \frac{d}{2}$ such that the $\mathrm{C}^0(\Omega)$-convergence result Lemma 2.25 holds. The reference quadrature rule $\widehat{\mathcal{Q}}$ is exact for all polynomials in $P_1(\widehat{T})$.

Periodic case (indicated by **(P)**): $\bar{u} \in \mathrm{A}_{u,\#}$ is a uniform minimizer of (2.40). $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_\#$ is the corresponding solution to (2.44) with $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}_\#(\Omega)$. $\bar{y}_h \in \mathcal{Y}_{h,\#}$ is the corresponding discrete solution to (2.50) satisfying $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}$. Moreover, we assume $F \in \mathrm{C}^{\mathrm{p}+1}(I_{\bar{u}})$ for an open interval $I_{\bar{u}}$ such that $\bar{u}(\bar{\Omega}) \subset I_{\bar{u}}$. The quadrature rule $\widehat{\mathcal{Q}}$ is exact for all polynomials $P_{2\mathrm{p}-1}(\widehat{T})$.

The discretized system $\widetilde{\mathcal{F}}_h(\widetilde{y}_h) = 0$, respectively, $\widetilde{\mathcal{F}}_{h,\#}(\widetilde{y}_h) = 0$ involves approximated integrals of functions of finite element functions. In order to get error estimates for these approximations, we need element-wise bounds of certain Sobolev and Hölder norms of the Galerkin solutions $\bar{u}_h$ and $\bar{\phi}_h$. The necessary results for Hölder norms were proved in Section 2.5.1.1. The next lemma deals with the convergence of $(\bar{u}_h, \bar{\phi}_h)$ to $(\bar{u}, \bar{\phi})$ with respect to element-wise Sobolev norms of higher order.

If $q = \infty$, we interpret

$$\Big(\sum_{T \in \mathcal{T}_h} \|f\|^q_{\mathrm{W}^{j,q}(T)} \Big)^{1/q} \quad \text{as} \quad \max_{T \in \mathcal{T}_h} \|f\|_{\mathrm{W}^{j,\infty}(T)}.$$

Lemma 3.1.

(D) *Let $2 < q < \frac{2d}{d-2}$. Then, there exists $C > 0$ such that for sufficiently small $h > 0$*

$$\Big(\sum_{T \in \mathcal{T}_h} \|\bar{u}_h - \bar{u}\|^q_{\mathrm{W}^{1,q}(T)} \Big)^{1/q} + \Big(\sum_{T \in \mathcal{T}_h} \|\bar{\phi}_h - \bar{\phi}\|^q_{\mathrm{W}^{1,q}(T)} \Big)^{1/q} \leq Ch^{1+\frac{d}{q}-\frac{d}{2}}.$$

(P) *If $j \in \mathbb{N}$ and $q > 0$ satisfy either $1 \leq j \leq \mathrm{p}-1$ and $2 < q \leq \infty$, or $j = \mathrm{p}$ and $2 < q < \frac{2d}{d-2}$, then there exists $C > 0$ such that for sufficiently small $h > 0$*

$$\Big(\sum_{T \in \mathcal{T}_h} \|\bar{u}_h - \bar{u}\|^q_{\mathrm{W}^{j,q}(T)} \Big)^{1/q} + \Big(\sum_{T \in \mathcal{T}_h} \|\bar{\phi}_h - \bar{\phi}\|^q_{\mathrm{W}^{j,q}(T)} \Big)^{1/q} \leq Ch^{\mathrm{p}-j+1+\frac{d}{q}-\frac{d}{2}}.$$

Proof. Since there is no need to distinguish formally between the Dirichlet and the periodic case, we present the proof in a generic way. Let $q > 2$. We first look at $\|\mathcal{I}_h\bar{u}-\bar{u}_h\|_{\mathrm{W}^{j,q}(T)}$. Since the family $(\mathcal{T}_h)_{h\in(0,1]}$ is quasi-uniform, we can apply the inverse inequality from [17, Theorem 4.5.11]:

$$\Big(\sum_{T\in\mathcal{T}_h}\|\mathcal{I}_h\bar{u}-\bar{u}_h\|^q_{\mathrm{W}^{j,q}(T)}\Big)^{1/q} \le Ch^{\frac{d}{q}-\frac{d}{2}}\Big(\sum_{T\in\mathcal{T}_h}\|\mathcal{I}_h\bar{u}-\bar{u}_h\|^2_{\mathrm{H}^{j}(T)}\Big)^{1/2},$$

for every j, where the constant C is independent h. The application of another inverse inequality yields

$$\Big(\sum_{T\in\mathcal{T}_h}\|\mathcal{I}_h\bar{u}-\bar{u}_h\|^2_{\mathrm{H}^{j}(T)}\Big)^{1/2} \le Ch^{1-j}\Big(\sum_{T\in\mathcal{T}_h}\|\mathcal{I}_h\bar{u}-\bar{u}_h\|^2_{\mathrm{H}^{1}(T)}\Big)^{1/2} \le Ch^{\mathrm{p}-j+1}.$$

Here, we have used the triangle inequality and the convergence of $\mathcal{I}_h\bar{u}$ and $\bar{u}_h$ to $\bar{u}$. Combining these two bounds we get

$$\Big(\sum_{T\in\mathcal{T}_h}\|\mathcal{I}_h\bar{u}-\bar{u}_h\|^q_{\mathrm{W}^{j,q}(T)}\Big)^{1/q} \le Ch^{\mathrm{p}-j+1+\frac{d}{q}-\frac{d}{2}}, \tag{3.9}$$

where the exponent of h is positive under the given assumptions. From [37, Th. 3.1.6] we know (using assumption (2.47) on the inverse and determinant of B_T) that

$$\Big(\sum_{T\in\mathcal{T}_h}\|\bar{u}-\mathcal{I}_h\bar{u}\|^q_{\mathrm{W}^{j,q}(T)}\Big)^{1/q} \le Ch^{\mathrm{p}-j+1+\frac{d}{q}-\frac{d}{2}}|\bar{u}|_{\mathrm{H}^{\mathrm{p}+1}(\Omega)}.$$

Combining this with (3.9) and applying the triangle inequality we get the desired inequality uniformly in $T \in \mathcal{T}_h$ and h. A similar argument shows the equivalent result for $\bar{\phi}_h$. ◻

As a direct consequence of the previous result we get the following Lemma.

Lemma 3.2.

(D) *For $2 < q < \frac{2d}{d-2}$ there exist $C_q, C_\infty > 0$ such that for sufficiently small $h > 0$*

$$\|\bar{u}_h\|_{\mathrm{L}^\infty(T)} + \|\bar{\phi}_h\|_{\mathrm{L}^\infty(T)} \le C_\infty \qquad \forall T \in \mathcal{T}_h,$$
$$\Big(\sum_{T\in\mathcal{T}_h}\|\bar{u}_h\|^q_{\mathrm{W}^{1,q}(T)}\Big)^{1/q} + \Big(\sum_{T\in\mathcal{T}_h}\|\bar{\phi}_h\|^q_{\mathrm{W}^{1,q}(T)}\Big)^{1/q} \le C_q.$$

(P) *For $2 < q < \frac{2d}{d-2}$ there exist $C_q, C_\infty > 0$ such that for sufficiently small $h > 0$*

$$\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)} + \|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)} \le C_\infty \qquad \forall T \in \mathcal{T}_h,$$
$$\Big(\sum_{T\in\mathcal{T}_h}\|\bar{u}_h\|^q_{\mathrm{W}^{\mathrm{p},q}(T)}\Big)^{1/q} + \Big(\sum_{T\in\mathcal{T}_h}\|\bar{\phi}_h\|^q_{\mathrm{W}^{\mathrm{p},q}(T)}\Big)^{1/q} \le C_q.$$

Proof. If $\mathrm{p} = 1$, the $\mathrm{L}^\infty(\Omega)$ bounds in both cases follow from Lemma 2.20, respectively, 2.25, where convergence with respect to Hölder norms was shown. If $\mathrm{p} > 1$, the $\mathrm{W}^{\mathrm{p}-1,\infty}(T)$-bounds in the periodic case follow directly from Lemma 3.1 with $q = \infty$.

The $\mathrm{W}^{\mathrm{p},q}$-boundedness in both cases follows from the previous Lemma 3.1 and the triangle inequality since $\|\bar{u}\|_{\mathrm{W}^{\mathrm{p},q}} \leq \infty$ for $2 < q < \frac{2d}{d-2}$ by Sobolev's Imbedding Theorem and $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}(\Omega)$. □

The existence proof for discrete solutions $\widetilde{y}_h$ to (3.2) and (3.5) will be heavily based on local regularity of the derivative $D\widetilde{\mathcal{F}}$ in a sufficiently large neighbourhood of $\bar{y}_h$. As in the Galerkin analysis, we need a continuity property for $D\widetilde{\mathcal{F}}_h$ and invertibility of $D\widetilde{\mathcal{F}}_h(y)$ and $D\widetilde{\mathcal{F}}_{h,\#}(y)$ in a neighbourhood of $\bar{y}_h$ that will contain a solution. The idea is to first prove invertibility in $\bar{y}_h$ and then use continuity to infer invertibility in a neighbourhood of $\bar{y}_h$.

Lemma 3.3.
(D) *(i) The nonlinear operator $\widetilde{\mathcal{F}}_h : \mathcal{Y}_{h,D} \to \mathcal{Y}_{h,0}^*$ defined in* (3.1) *is Fréchet differentiable at every $y = (u, \phi, \mu) \in \mathcal{Y}_h$ with derivative given by*

$$\begin{aligned}\big\langle D\widetilde{\mathcal{F}}_h(y)\cdot\eta_1, \eta_2\big\rangle = {} & \lambda(\nabla v_1, \nabla v_2) + \mathcal{Q}_h\big[(F''(u) + 2\phi)v_1v_2\big] + \mu\mathcal{Q}_h[v_1v_2] \\ & + 2\mathcal{Q}_h[u\psi_1v_2] + \nu_1\mathcal{Q}_h[uv_2] + \tfrac{1}{4\pi}(\nabla\psi_1, \nabla\psi_2) - 2\mathcal{Q}_h[uv_1\psi_2] \\ & + \nu_2\mathcal{Q}_h[uv_1]\,,\end{aligned} \tag{3.10}$$

for all $\eta_1 = (v_1, \psi_1, \nu_1)$, $\eta_2 = (v_2, \psi_2, \nu_2) \in \mathcal{Y}_{h,0}$. The derivative $D\widetilde{\mathcal{F}}_h$ is Hölder continuous in $\mathcal{Y}_{h,D}$ in the following sense: for all $y_1, y_2 \in \mathcal{Y}_h$,

$$\begin{aligned}\|D\widetilde{\mathcal{F}}_h(y_1) - D\widetilde{\mathcal{F}}_h(y_2)\| &\leq C(\|y_1\|_{\mathcal{Y}}, \|y_2\|_{\mathcal{Y}})(\|u_1 - u_2\|_{\mathrm{L}^2}^{\alpha_F} + \|y_1 - y_2\|_{\mathcal{Y}}) \\ &\leq C(\|y_1\|_{\mathcal{Y}}, \|y_2\|_{\mathcal{Y}})(\|y_1 - y_2\|_{\mathcal{Y}}^{\alpha_F} + \|y_1 - y_2\|_{\mathcal{Y}}).\end{aligned}$$

(ii) There exist numbers $\delta, M, h_0 > 0$ such that for all $h < h_0$ the derivative $D\widetilde{\mathcal{F}}_h(y_h) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$ is an isomorphism for all $y_h \in B_\delta(\bar{y}_h) \subset \mathcal{Y}_h$ and $\|D\widetilde{\mathcal{F}}_h(y)^{-1}\| \leq M$, uniformly in h and $y \in B_\delta(\bar{y}_h)$.

(P) *The operator $\widetilde{\mathcal{F}}_h$ in part* **(D)** *can be replaced with $\widetilde{\mathcal{F}}_{h,\#} : \mathcal{Y}_{h,\#} \to \mathcal{Y}_{h,\#}^*$ from* (3.5) .

Proof. We present the proof for the Dirichlet case. The generalization to the periodic case is evident.

(i): Since the quadrature rule $\mathcal{Q}_h[\cdot]$ is a linear operator involving only point evaluations of the integrands, the form (3.10) of the derivative $D\widetilde{\mathcal{F}}_h$ is easy to obtain. The fact that for every $v \in \mathrm{S}_h$ the operator $D\widetilde{\mathcal{F}}_h(y)\cdot v$ is well-defined as an element of $\mathrm{S}_{h,0}^*$ follows by a calculation similar to the one showing continuity, which we provide below.

Step 1. Continuity. We begin by showing Hölder continuity of $\mathcal{Q}_h[F''(u)vw)]$ with respect to $u \in \mathrm{S}_h$ for fixed $v, w \in \mathrm{S}_h$. First, we look at a single element $T \in \mathcal{T}_h$:

$$\begin{aligned}\big|\mathcal{Q}_T\big[F''(u_1)vw\big] - \mathcal{Q}_T\big[F''(u_2)vw\big]\big| \\ \leq \det B_T \sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu \big|F''(u_1(x_\nu^{(T)})) - F''(u_2(x_\nu^{(T)}))\big|\,|v(x_\nu^{(T)})|\,|w(x_\nu^{(T)})|.\end{aligned}$$

Since, for the moment, we only look at one element T, we will use the abbreviations $u_{i,\nu} = u_i(x_\nu^{(T)})$, $v_\nu = v(x_\nu^{(T)})$, and $w_\nu = w(x_\nu^{(T)})$. Using the continuity condition on F'' we get:

$$\begin{aligned}&\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |F''(u_{1,\nu}) - F''(u_{2,\nu})||v_\nu||w_\nu| \qquad (3.11)\\ &\quad\leq \sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu \big(|u_{1,\nu} - u_{2,\nu}|^{\alpha_F} + (1 + |u_{1,\nu}|^{q_F-3} + |u_{2,\nu}|^{q_F-3})|u_{1,\nu} - u_{2,\nu}|\big)|v_\nu|\,|w_\nu|.\end{aligned}$$

Let us first look at the part with α_F:

$$\begin{aligned}\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |u_{1,\nu} - u_{2,\nu}|^{\alpha_F} |v_\nu|\,|w_\nu| \leq C\bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |u_{1,\nu} - u_{2,\nu}|^{2\alpha_F}\bigg)^{1/2} \\ \cdot \bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |v_\nu|^4\bigg)^{1/4} \bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |w_\nu|^4\bigg)^{1/4},\end{aligned}$$

where we have used $\omega_\nu > 0$ for all $\nu \in \{1, \dots, n_\mathcal{Q}\}$ and the generalized Hölder inequality. Since the weights ω_ν, $\nu \in \{1, \dots, n_\mathcal{Q}\}$ were assumed positive, we can prove that

$$\bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |u_{1,\nu} - u_{2,\nu}|^{2\alpha_F}\bigg)^{1/2} \leq C\bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |u_{1,\nu} - u_{2,\nu}|^{2}\bigg)^{\alpha_F/2},$$

using the discrete Hölder inequality (see (2.14) for a similar calculation in the continuous case). For the second part on the right-hand side of (3.11), we get

$$\begin{aligned}&\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu \big(1 + |u_{1,\nu}|^{q_F-3} + |u_{2,\nu}|^{q_F-3}\big)|u_{1,\nu} - u_{2,\nu}|\,|v_\nu|\,|w_\nu| \\ &\qquad\leq C\bigg(1 + \bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |u_{1,\nu}|^6\bigg)^{1/2} + \bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |u_{2,\nu}|^6\bigg)^{1/2}\bigg) \\ &\qquad\quad\cdot \bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |u_{1,\nu} - u_{2,\nu}|^6\bigg)^{1/6} \bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |v_\nu|^6\bigg)^{1/6} \bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |w_\nu|^6\bigg)^{1/6}.\end{aligned}$$

In order to bound the discrete expressions above by Sobolev norms we go the usual way of transforming to the reference triangle and using the equivalence of all norms on the finite dimensional space $P_\mathrm{p}(\widehat{T})$:

$$\begin{aligned}\bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu |v_{T,\nu}|^q\bigg)^{1/q} &\leq \|v\|_{\mathrm{L}^\infty(T)} \bigg(\sum_{\nu=1}^{n_\mathcal{Q}} \omega_\nu\bigg)^{1/q} = |\widehat{T}|^{1/q} \|\widehat{v}\|_{\mathrm{L}^\infty(\widehat{T})} \\ &\leq C\|\widehat{v}\|_{\mathrm{L}^q(\widehat{T})} \leq C(\det B_T)^{-1/q} \|v\|_{\mathrm{L}^q(T)}\end{aligned}$$

for all $v \in P_{\mathrm{p}}(T)$ and all $q \geq 1$. Hence, we conclude that

$$\begin{aligned}\big|\mathcal{Q}_T[F''(u_1)vw] - \mathcal{Q}_T[F''(u_2)vw]\big| \leq C\|v\|_{\mathrm{L}^6(T)}\|w\|_{\mathrm{L}^6(T)}\big(\|u_1-u_2\|^{\alpha_F}_{\mathrm{L}^2(T)} + \\ \big(1+\|u_1\|^3_{\mathrm{L}^6(T)} + \|u_2\|^3_{\mathrm{L}^6(T)}\big)\|u_1-u_2\|_{\mathrm{L}^6(T)}\big).\end{aligned}$$

Summing over $T \in \mathcal{T}_h$ and using the imbedding of $\mathrm{H}^1(\Omega)$ into $\mathrm{L}^6(\Omega)$ then leads to

$$\big|\mathcal{Q}_h\big[F''(u_1)vw\big] - \mathcal{Q}_h\big[F''(u_2)vw\big]\big| \leq C(u_1,u_2)\left(\|u_1-u_2\|^{\alpha_F}_{\mathrm{L}^2} + \|u_1-u_2\|_{\mathrm{H}^1}\right)\|v\|_{\mathrm{H}^1}\|w\|_{\mathrm{H}^1},$$

for all $v, w \in \mathrm{S}_h$.

Similar continuity properties can be obtained for terms of the forms $\mathcal{Q}_h[uvw]$, $\mathcal{Q}_h[\phi vw]$, and $\mathcal{Q}_h[uv]$ in (3.10). The only difference is that these terms are in fact Lipschitz continuous with respect to u, respectively, ϕ. For example, a similar calculation to the one described above yields

$$\big|\mathcal{Q}_h[u_1vw] - \mathcal{Q}_h[u_2vw]\big| \leq C\,\|u_1-u_2\|_{\mathrm{H}^1}\|v\|_{\mathrm{H}^1}\|w\|_{\mathrm{H}^1}$$

for all $u_1, u_2, v, w \in \mathrm{S}_h$ with a constant C that depends on $(\|u_1\|_{\mathrm{H}^1} + \|u_2\|_{\mathrm{H}^1})$.

Step 2. $D\widetilde{\mathcal{F}}(\bar{y}_h)$ is an isomorphism. We already know that the derivative $D\mathcal{F}_h(\bar{y}_h) : \mathcal{Y}_{h,0} \to \mathcal{Y}^*_{h,0}$ of the Galerkin operator is an isomorphism (Proposition 2.9) and satisfies the inf-sup conditions

$$\inf_{\eta_1\in\mathcal{Y}_{h,0}}\sup_{\eta_2\in\mathcal{Y}_{h,0}}\frac{\langle D\mathcal{F}_h(\bar{y}_h)\cdot\eta_1,\eta_2\rangle}{\|\eta_1\|_{\mathcal{Y}}\|\eta_2\|_{\mathcal{Y}}} \geq \kappa_h, \qquad \inf_{\eta_1\in\mathcal{Y}_{h,0}}\sup_{\eta_2\in\mathcal{Y}_{h,0}}\frac{\langle D\mathcal{F}_h(\bar{y}_h)\cdot\eta_2,\eta_1\rangle}{\|\eta_1\|_{\mathcal{Y}}\|\eta_2\|_{\mathcal{Y}}} \geq \kappa_h, \tag{3.12}$$

where $\kappa_h > 0$ is bounded away from zero for sufficiently small h. We will now show that

$$\big\|D\mathcal{F}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\bar{y}_h)\big\| = \sup_{\eta_1\in\mathcal{Y}_{h,0}}\sup_{\eta_2\in\mathcal{Y}_{h,0}}\frac{\big|\big\langle\big(D\mathcal{F}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\bar{y}_h)\big)\cdot\eta_1,\eta_2\big\rangle\big|}{\|\eta_1\|_{\mathcal{Y}}\|\eta_2\|_{\mathcal{Y}}} \leq Ch^{\alpha_{F,h}} \tag{3.13}$$

for some $0 < \alpha_{F,h} < 1$, uniformly for sufficiently small $h > 0$.

Comparing (3.10) and (2.35) we obtain (with $\eta_1 = (v_1,\psi_1,\nu_1)$, $\eta_2 = (v_2,\psi_2,\nu_2)$):

$$\begin{aligned}\big\langle\big(D\mathcal{F}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\bar{y}_h)\big)\cdot\eta_1,\eta_2\big\rangle = e_h\big[F''(\bar{u}_h)v_1v_2\big] + 2e_h\big[\bar{\phi}_hv_1v_2\big] + \bar{\mu}_he_h[v_1v_2] \\ + 2e_h[\bar{u}_h\psi_1v_2] + \nu_1e_h[\bar{u}_hv_2] - 2e_h[\bar{u}_hv_1\psi_2] \\ + \nu_2e_h[\bar{u}_hv_1]\,.\end{aligned}$$

Again we outline the necessary steps using the example of the term $\mathcal{Q}_h[F''(\bar{u}_h)v_1v_2]$ and its counterpart $(F''(\bar{u}_h)v_1,v_2)$ in $D\mathcal{F}_h(\bar{y}_h)$. We know that F'' is Hölder continuous with degree α_F. Moreover, we have shown in Lemma 2.25 that $\bar{u}_h|_T$ is also Hölder continuous with degree α_u from (2.64), uniformly in T. Therefore $F''(\bar{u}_h)|_T$ is Hölder continuous with degree $\alpha_{F,u} = \alpha_F\alpha_u$ for every T as the following calculation shows:

$$\begin{aligned}\big|F''(\bar{u}_h)\big|_{\mathrm{C}^{0,\alpha_{F,u}}(T)} &= \sup_{x_1,x_2\in T}\frac{|F''(\bar{u}_h(x_1)) - F''(\bar{u}_h(x_2))|}{|x_1-x_2|^{\alpha_F\alpha_u}} \\ &\leq \sup_{x_1,x_2\in T}\frac{|F''(\bar{u}_h(x_1)) - F''(\bar{u}_h(x_2))|}{|\bar{u}_h(x_1)-\bar{u}_h(x_2)|^{\alpha_F}}\frac{|\bar{u}_h(x_1)-\bar{u}_h(x_2)|^{\alpha_F}}{|x_1-x_2|^{\alpha_F\alpha_u}} \\ &\leq C_F(|\bar{u}_h|_{\mathrm{C}^{0,\alpha_u}(T)})^{\alpha_F}.\end{aligned} \tag{3.14}$$

As we have seen in Lemma 2.20, $|\bar{u}_h|_{\mathrm{C}^{0,\alpha_u}(T)}$ is bounded uniformly in T and h. Since $F''(\bar{u}_h)$ is obviously bounded in $\mathrm{L}^\infty(T)$ we deduce that $\|F''(\bar{u}_h)\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)}$ is also bounded uniformly in $T \in \mathcal{T}_h$ and h for

$$\alpha_{F,u} = \alpha_F \alpha_u < \alpha_F \min\big(\tfrac{1}{1-\alpha_F} - \tfrac{d}{2}, 2 - \tfrac{d}{2}\big), \tag{3.15}$$

where we have used α_u from (2.64). It now follows from Proposition A.15(ii) on errors for quadrature rules that

$$\big|e_T\big[F''(\bar{u}_h)v_1v_2\big]\big| \le Ch^{\alpha_{F,u}} \|F''(\bar{u}_h)\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)} \|v_1\|_{\mathrm{H}^1(T)} \|v_2\|_{\mathrm{H}^1(T)},$$

for all $v, w \in \mathrm{S}_{h,0}$ and after summing over all elements $T \in \mathcal{T}_h$:

$$\big|e_h\big[F''(\bar{u}_h)v_1v_2\big]\big| \le Ch^{\alpha_{F,u}} \|v_1\|_{\mathrm{H}^1} \|v_2\|_{\mathrm{H}^1},$$

for all $v, w \in \mathrm{S}_{h,0}$. Similar results hold for the other terms in $D\widetilde{\mathcal{F}}_h(\bar{y}_h)$ that involve numerical integration:

$$\begin{aligned}
\big|e_h\big[\bar{\phi}_h v_1 v_2\big]\big| + \big|e_h[\bar{u}_h \psi_1 v_2]\big| + \big|e_h[\bar{u}_h v_1 \psi_2]\big| &\le Ch^{\alpha_u}\big(\|v_1\|_{\mathrm{H}^1}\|v_2\|_{\mathrm{H}^1} + \|v_2\|_{\mathrm{H}^1}\|\psi_1\|_{\mathrm{H}^1} \\
&\qquad + \|v_1\|_{\mathrm{H}^1}\|\psi_2\|_{\mathrm{H}^1}\big), \\
\big|e_h[\bar{u}_h v_1]\big| + \big|e_h[\bar{u}_h v_2]\big| &\le Ch\big(\|v_1\|_{\mathrm{H}^1} + \|v_2\|_{\mathrm{H}^1}\big), \\
\big|e_h[v_1 v_2]\big| &\le Ch\|v_1\|_{\mathrm{H}^1}\|v_2\|_{\mathrm{H}^1},
\end{aligned}$$

by Proposition A.15 (i) and Theorem A.11. Here, we have also used that $\sum_{T\in\mathcal{T}_h} \|\bar{u}_h\|^q_{\mathrm{W}^{1,q}(T)} \le C_q$ by Lemma 3.2.

We deduce that (3.13) holds with $\alpha_{F,h} = \alpha_{F,u}$ which then, together with (3.12), implies

$$\inf_{\eta_1 \in \mathcal{Y}_{h,0}} \sup_{\eta_2 \in \mathcal{Y}_{h,0}} \frac{\big\langle D\widetilde{\mathcal{F}}_h(\bar{y}_h)\cdot\eta_1, \eta_2\big\rangle}{\|\eta_1\|_{\mathcal{Y}}\|\eta_2\|_{\mathcal{Y}}} \ge \frac{\kappa_h}{2}, \tag{3.16}$$

for sufficiently small h. Similarly, we can prove the second inf-sup condition (with exchanged η_1 and η_2, compare (3.12)) that is necessary. Hence, $D\widetilde{\mathcal{F}}_h(\bar{y}_h)$ is an isomorphism for sufficiently small h.

Step 3. The boundedness of $\big\|D\widetilde{\mathcal{F}}_h(y_h)^{-1}\big\|$ for y_h in a neighbourhood of $\bar{y}_h$ follows from continuity. Assume that $\|D\widetilde{\mathcal{F}}_h(\bar{y}_h)^{-1}\| \le \frac{M}{2}$. Then, since $D\widetilde{\mathcal{F}}_h$ is Hölder continuous, we can find $\delta > 0$ (independent of h) such that $\big\|D\widetilde{\mathcal{F}}_h(y_h)^{-1}\big\| \le M$ for all $y_h \in B_\delta(\bar{y}_h)$ and all h sufficiently small, see Lemma A.2. ◻

We now show that the Galerkin solution $(\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,D}$ to (2.34) solves (3.1) approximately. The proof consists in estimating quadrature errors to show that $\|\widetilde{\mathcal{F}}(\bar{y}_h)\|_{\mathcal{Y}^*_{h,0}}$ is small.

Lemma 3.4.

(D) *There exists $C > 0$ such that, for sufficiently small h,*

$$\big\|\widetilde{\mathcal{F}}_h(\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h)\big\|_{\mathcal{Y}^*_{h,0}} \le Ch.$$

(P) *There exists $C > 0$ such that, for sufficiently small h,*

$$\big\|\widetilde{\mathcal{F}}_{h,\#}(\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h)\big\|_{\mathcal{Y}^*_{h,\#}} \leq Ch^{\mathrm{p}}.$$

Proof. We present the proof in a generic fashion for arbitrary p without formally distinguishing between the Dirichlet and the periodic case.

Comparing (3.1) with (2.34), respectively, (2.50) we see that it is necessary to estimate the quadrature errors of products of $F'(\bar{u}_h)$, $\bar{u}_h\bar{\phi}_h$ and $(\bar{u}_h^2 - \rho_n)$ with finite element test functions. We begin by establishing some regularity properties of these nonlinear terms.

Step 1. We show that $F'(\bar{u}_h)$, $\bar{u}_h\bar{\phi}_h$ and $\bar{u}_h^2$ belong to $\mathrm{W}^{\mathrm{p},q}(T)$ for $3 < q < 6$, for all $T \in \mathcal{T}_h$, with norms bounded uniformly in $T \in \mathcal{T}_h$ and h. For $\mathrm{p} = 1$, recalling that $\bar{u}_h$, $\bar{\phi}_h$ are uniformly bounded in $\mathrm{L}^\infty(T)$ (Lemma 3.2) we obtain:

$$\begin{aligned}
\|\nabla(F'(\bar{u}_h))\|_{\mathrm{L}^q(T)} &\leq \|F''(\bar{u}_h)\|_{\mathrm{L}^\infty(\Omega)}\|\nabla\bar{u}_h\|_{\mathrm{L}^q(T)},\\
\|\nabla(\bar{u}_h\bar{\phi}_h)\|_{\mathrm{L}^q(T)} &\leq \|\bar{u}_h\|_{\mathrm{L}^\infty(\Omega)}\|\nabla\bar{\phi}_h\|_{\mathrm{L}^q(T)} + \|\bar{\phi}_h\|_{\mathrm{L}^\infty(\Omega)}\|\nabla\bar{u}_h\|_{\mathrm{L}^q(T)},\\
\|\nabla\bar{u}_h^2\|_{\mathrm{L}^q(T)} &\leq 2\|\bar{u}_h\|_{\mathrm{L}^\infty(\Omega)}\|\nabla\bar{u}_h\|_{\mathrm{L}^q(\Omega)},
\end{aligned}$$

for any $3 < q < 6$. More generally, if $\mathrm{p} \geq 2$ and $F \in \mathrm{C}^{\mathrm{p}+1}(I_{\bar{u}})$ we have, for every $j \leq \mathrm{p}$:

$$\begin{aligned}
|\nabla^j F'(\bar{u}_h)| \leq C(j)\Big(&\big|F^{(j+1)}(\bar{u}_h)\big|\,|\nabla\bar{u}_h|^j + |F^{(j)}(\bar{u}_h)||\nabla^2\bar{u}_h||\nabla\bar{u}_h|^{j-2} + \dots\\
&+ |F^{(3)}(\bar{u}_h)|\big(|\nabla\bar{u}_h|\big|\nabla^{j-1}\bar{u}_h\big| + |\nabla^2\bar{u}_h|\big|\nabla^{j-2}\bar{u}_h\big| + \dots\big)\\
&+ |F''(\bar{u}_h)|\big|\nabla^j\bar{u}_h\big|\Big).
\end{aligned} \tag{3.17}$$

Similarly,

$$\big|\nabla^j(\bar{u}_h\bar{\phi}_h)\big| \leq C(j)\Big(\big|\nabla^j\bar{u}_h\big|\,|\bar{\phi}_h| + \big|\nabla^{j-1}\bar{u}_h\big|\,|\nabla\bar{\phi}_h| + \dots + |\bar{u}_h|\big|\nabla^j\bar{\phi}_h\big|\Big).$$

The uniform boundedness of $\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)}$ and $\|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)}$ implies that

$$\begin{aligned}
\|F'(\bar{u}_h)\|_{\mathrm{W}^{\mathrm{p},q}(T)} &\leq C(\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)} + \|F'(\bar{u}_h)\|_{\mathrm{L}^q(T)}),\\
\|\bar{u}_h\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)} &\leq C(\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)} + \|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)}),\\
\|\bar{u}_h^2\|_{\mathrm{W}^{\mathrm{p},q}(T)} &\leq C\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)}.
\end{aligned}$$

Here, the constants C depend on the derivatives of F on $I_{\bar{u}}$, $\max_{T\in\mathcal{T}_h}\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)}$, and $\max_{T\in\mathcal{T}_h}\|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)}$, which are bounded uniformly in T and h by Lemma 3.2. Hence, there exists $C > 0$ such that

$$\sum_{T\in\mathcal{T}_h}\|F'(\bar{u}_h)\|^q_{\mathrm{W}^{\mathrm{p},q}(T)} \leq C, \quad \sum_{T\in\mathcal{T}_h}\|\bar{u}_h\bar{\phi}_h\|^q_{\mathrm{W}^{\mathrm{p},q}(T)} \leq C, \quad \sum_{T\in\mathcal{T}_h}\|\bar{u}_h^2 - \rho_n\|^q_{\mathrm{W}^{\mathrm{p},q}(T)} \leq C, \tag{3.18}$$

uniformly in h by Lemma 3.2.

Step 2. After these preparations we can estimate $\big\|\widetilde{\mathcal{F}}_h(\bar{y}_h)\big\|_{\mathcal{Y}_{h,0}^*}$. Theorem A.11 on quadrature errors together with the bounds (3.18) now yields:

$$\begin{aligned}
\sup_{v\in \mathrm{S}_{h,0}} \frac{|(F'(\bar{u}_h),v) - \mathcal{Q}_h[F'(\bar{u}_h)v]|}{\|v\|_{\mathrm{H}^1}} &\le Ch^{\mathrm{p}},\\
\sup_{v\in \mathrm{S}_{h,0}} \frac{\big|(\bar{u}_h\bar{\phi}_h,v) - \mathcal{Q}_h\big[\bar{u}_h\bar{\phi}_h v\big]\big|}{\|v\|_{\mathrm{H}^1}} &\le Ch^{\mathrm{p}},\\
\sup_{v\in \mathrm{S}_{h,0}} \frac{|(\bar{u}_h,v) - \mathcal{Q}_h[\bar{u}_h v]|}{\|v\|_{\mathrm{H}^1}} &\le Ch^{\mathrm{p}},\\
\sup_{\psi\in \mathrm{S}_{h,0}} \frac{\big|(\bar{u}_h^2-\rho_n,\psi) - \mathcal{Q}_h\big[(\bar{u}_h^2-\rho_n)\psi\big]\big|}{\|\psi\|_{\mathrm{H}^1}} &\le Ch^{\mathrm{p}},\\
\big|\|\bar{u}_h\|_{\mathrm{L}^2} - \mathcal{Q}_h\big[\bar{u}_h^2\big]\big| &\le Ch^{\mathrm{p}},
\end{aligned}$$

where C is independent of h. Since $\mathcal{F}_h(\bar{y}_h)=0$ we then get

$$\big\|\widetilde{\mathcal{F}}_h(\bar{y}_h)\big\|_{\mathcal{Y}_{h,0}^*} = \sup_{\eta\in\mathcal{Y}_{h,0}} \frac{\big|\langle\widetilde{\mathcal{F}}_h(\bar{y}_h),\eta\rangle\big|}{\|\eta\|_{\mathcal{Y}}} = \sup_{\eta\in\mathcal{Y}_{h,0}} \frac{\big|\big\langle\big(\widetilde{\mathcal{F}}_h(\bar{y}_h)-\mathcal{F}_h(\bar{y}_h)\big),\eta\big\rangle\big|}{\|\eta\|_{\mathcal{Y}}} \le Ch^{\mathrm{p}},$$

as desired. □

We are now ready to prove existence and convergence of a solution $\widetilde{y}_h = (\widetilde{u}_h,\widetilde{\phi}_h,\widetilde{\mu}_h)\in\mathcal{Y}_{h,D}$ to the discretized system (3.1).

Theorem 3.5.
(D) *There exist $h_0\in(0,1],\delta>0$ such that the discretized problem* (3.1) *has a unique solution $\widetilde{y}_h=(\widetilde{u}_h,\widetilde{\phi}_h,\widetilde{\mu}_h)\in\mathcal{Y}_{h,D}$ in the neighbourhood $B_\delta(\bar{y})\subset\mathcal{Y}$ for all $h<h_0$. Furthermore, there exists a constant $C>0$ such that*

$$\|\bar{u}-\widetilde{u}_h\|_{\mathrm{H}^1} + \|\bar{\phi}-\widetilde{\phi}_h\|_{\mathrm{H}^1} + |\bar{\mu}-\widetilde{\mu}_h| \le Ch.$$

(P) *There exist $h_0\in(0,1],\delta>0$ such that the discretized periodic problem* (3.5) *has a unique solution $\widetilde{y}_h=(\widetilde{u}_h,\widetilde{\phi}_h,\widetilde{\mu}_h)\in\mathcal{Y}_{h,\#}$ in the neighbourhood $B_\delta(\bar{y})\subset\mathcal{Y}$ for all $h<h_0$. Furthermore, there exists $C>0$ such that*

$$\|\bar{u}-\widetilde{u}_h\|_{\mathrm{H}^1} + \|\bar{\phi}-\widetilde{\phi}_h\|_{\mathrm{H}^1} + |\bar{\mu}-\widetilde{\mu}_h| \le Ch^{\mathrm{p}}.$$

Proof. We will outline the proof for the Dirichlet case. The generalization to the periodic case is straightforward.

Similarly to Theorem 2.10 the idea of this proof is to construct a contractive mapping whose fixed point is the desired solution $\widetilde{y}_h$. For $R>0$ we define the map $\widetilde{\mathcal{N}}: B_R(\bar{y}_h)\to\mathcal{Y}_{h,D}$ by

$$D\widetilde{\mathcal{F}}_h(\bar{y}_h)\cdot\big(\widetilde{\mathcal{N}}(y)-\bar{y}_h\big) = -\widetilde{\mathcal{F}}_h(\bar{y}_h) - \int_0^1 \Big(D\widetilde{\mathcal{F}}_h(\bar{y}_h+t(y-\bar{y}_h)) - D\widetilde{\mathcal{F}}_h(\bar{y}_h)\Big)\,\mathrm{d}t\cdot(y-\bar{y}_h).$$

It is easy to see that a fixed point y of $\widetilde{\mathcal{N}}$ satisfies $\widetilde{\mathcal{F}}_h(y) = 0$, and vice versa.

First, we prove that $\widetilde{\mathcal{N}}$ maps $B_R(\bar{y}_h)$ to $B_R(\bar{y}_h)$ for sufficiently small R. For each $y_h \in B_R(\bar{y}_h)$ we have

$$\begin{aligned} M^{-1}\big\|\widetilde{\mathcal{N}}(y_h) - \bar{y}_h\big\|_{\mathcal{Y}} &\le \big\|\widetilde{\mathcal{F}}_h(\bar{y}_h)\big\|_{\mathcal{Y}_0^*} + R\int_0^1 \big\|D\widetilde{\mathcal{F}}_h(\bar{y}_h + t(y_h - \bar{y}_h)) - D\widetilde{\mathcal{F}}_h(\bar{y}_h)\big\|\,\mathrm{d}t \\ &\le C\big(h + RL_{\bar{y},\delta}(R + R^{\alpha_F})\big). \end{aligned}$$

Here we have used Lemma 3.3 and that $\big\|D\widetilde{\mathcal{F}}_h(\bar{y}_h)^{-1}\big\| < M$ (see Lemma 3.4). For sufficiently small h and R we see that $C(h + RL_{\bar{y},\delta}(R + R^{\alpha_F})) < RM^{-1}$ and hence $\widetilde{\mathcal{N}}$ maps $B_R(\bar{y}_h)$ to $B_R(\bar{y}_h)$.

Next, we show that $\mathcal{N}$ is a contraction on $B_R(\bar{y}_h)$. If $\eta_1, \eta_2 \in B_R(\bar{y}_h)$, then

$$D\widetilde{\mathcal{F}}_h(\bar{y}_h)\cdot\big(\widetilde{\mathcal{N}}(\eta_1) - \widetilde{\mathcal{N}}(\eta_2)\big) = \int_0^1 \big[D\widetilde{\mathcal{F}}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\eta_1 + t(\eta_2 - \eta_1))\big]\cdot(\eta_1 - \eta_2)\mathrm{d}t.$$

Thus, $\big\|\widetilde{\mathcal{N}}(\eta_1) - \widetilde{\mathcal{N}}(\eta_2)\big\|_{\mathcal{Y}}$ can be estimated as follows:

$$\begin{aligned} M^{-1}\big\|\widetilde{\mathcal{N}}(\eta_1) - \widetilde{\mathcal{N}}(\eta_2)\big\|_{\mathcal{Y}} &\le \int_0^1 \big\|D\widetilde{\mathcal{F}}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\eta_1 + t(\eta_2 - \eta_1))\big\|_{\mathcal{Y}}\,\mathrm{d}t\ \|\eta_1 - \eta_2\| \\ &\le L_{\bar{y},\delta}(R + R^{\alpha_F})\cdot\|\eta_1 - \eta_2\|_{\mathcal{Y}}. \end{aligned}$$

For sufficiently small R we obtain $L_{\bar{y},\delta}(R + R^{\alpha_F})M < 1$ and hence $\mathcal{N}$ is a contraction on $B_R(y_0)$.

We can now use Banach's Fixed Point Theorem [124, Th. 1.A] to obtain the existence and uniqueness of a fixed point $\widetilde{y}_h$ of the map $\widetilde{\mathcal{N}} : B_R(\bar{y}_h) \to B_R(\bar{y}_h)$. This fixed point $\widetilde{y}_h$ is a solution of $\widetilde{\mathcal{F}}_h(y) = 0$ and $\|\bar{y}_h - \widetilde{y}_h\|_{\mathcal{Y}} \le R$.

As in the Galerkin case convergence can now be obtained by a minor modification of the above argument. If we let $R = C_R h$ and C_R is sufficiently large, we can repeat the previous steps and deduce $\|\bar{y}_h - \widetilde{y}_h\|_{\mathcal{Y}} \le C_R h$. Since $\|\bar{y} - \bar{y}_h\|_{\mathcal{Y}} \le Ch$ this yields the desired result. □

Before proving that $\widetilde{u}_h$ is a uniform minimizer of the discrete energy with numerical quadrature we introduce some notation. By $(-\widetilde{\Delta}_{h,0})^{-1} : \mathrm{C}^0(\overline{\Omega}) \to \mathrm{S}_{h,0}$ we denote the solution operator of the problem: for $f \in \mathrm{C}^0(\overline{\Omega})$ find $\phi = (-\widetilde{\Delta}_{h,0})^{-1} f \in \mathrm{S}_{h,0}$ such that $(\nabla\phi, \nabla v) = \mathcal{Q}_h[fv]$ for all $v \in \mathrm{S}_{h,0}$.

It is relatively straightforward to show that

$$\big\|(-\widetilde{\Delta}_{h,0})^{-1}(vw)\big\|_{\mathrm{H}^1} \le C\,\|v\|_{\mathrm{H}^1}\,\|w\|_{\mathrm{H}^1} \quad \forall v, w \in \mathrm{S}_h. \tag{3.19}$$

For this let $\psi = (-\widetilde{\Delta}_{h,0})^{-1}(vw)$. Starting with $\|\nabla\psi\|_{\mathrm{L}^2}^2 = \mathcal{Q}_h[vw\psi]$ we can proceed as in the continuity proof of $\widetilde{\mathcal{F}}_h$ in Lemma 3.3.

Proposition 3.6.

(D) *Under the assumptions of Theorem 3.5,* $\widetilde{u}_h \in \mathrm{A}_{u,h}$ *is a uniform local minimizer of the functional* $\widetilde{E}_h$ *from* (3.3).

(P) *Under the assumptions of Theorem 3.5,* $\widetilde{u}_h \in \mathrm{A}_{u,h,\#}$ *is a uniform local minimizer of the functional* $\widetilde{E}_{h,\#}$ *from* (3.6).

Proof. Again the proof is given for the Dirichlet case. It transfers almost verbatim to the periodic setting.

Step 1. The idea of the proof is the same as in the Galerkin case analyzed in Proposition 2.11. We introduce the Lagrangian $\widetilde{\mathcal{L}}_h : \mathcal{Y}_h \times \mathbb{R} \to \mathbb{R}$ by

$$\widetilde{\mathcal{L}}_h(u,\mu) = \widetilde{E}_h(u) + \mu \widetilde{c}_h(u)$$

with $\widetilde{E}_h$ from (3.3) and $\widetilde{c}_h$ from (3.4). It suffices to show the existence of $\widetilde{\gamma} > 0$ such that

$$D^2_{uu}\widetilde{\mathcal{L}}_h(\widetilde{u}_h,\widetilde{\mu}_h)\cdot[\widetilde{v}_h,\widetilde{v}_h] \geq \widetilde{\gamma}\,\|\widetilde{v}_h\|^2_{\mathrm{H}^1} \qquad \forall \widetilde{v}_h \in \ker D\widetilde{c}_h(\widetilde{u}_h), \tag{3.20}$$

uniformly in h. For arbitrary $\bar{v}_h \in \ker Dc(\bar{u}_h) \cap \mathrm{S}_{h,0}$, we proceed with a similar rearrangement of $D^2_{uu}\widetilde{\mathcal{L}}_h(\widetilde{u}_h,\widetilde{\mu}_h)\cdot[\widetilde{v}_h,\widetilde{v}_h]$ as in (2.38):

$$\begin{aligned} D^2_{uu}\widetilde{\mathcal{L}}_h(\widetilde{u}_h,\widetilde{\mu}_h)\cdot[\widetilde{v}_h,\widetilde{v}_h] = {} & D^2_{uu}\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)\cdot[\bar{v}_h,\bar{v}_h] \\ & + \big(D^2_{uu}\widetilde{\mathcal{L}}_h(\widetilde{u}_h,\widetilde{\mu}_h) - D^2_{uu}\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)\big)\cdot[\widetilde{v}_h,\widetilde{v}_h] \\ & + 2D^2_{uu}\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)\cdot[\bar{v}_h,\widetilde{v}_h-\bar{v}_h] \\ & + D^2_{uu}\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)\cdot[\widetilde{v}_h-\bar{v}_h,\widetilde{v}_h-\bar{v}_h]. \end{aligned} \tag{3.21}$$

In Proposition 2.11 we established that

$$D^2_{uu}\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)\cdot[\bar{v}_h,\bar{v}_h] \geq \frac{\gamma}{2}\,\|\nabla\bar{v}_h\|^2_{\mathrm{L}^2} \qquad \forall \bar{v}_h \in \ker Dc(\bar{u}_h) \cap \mathrm{S}_{h,0},$$

for a $\gamma > 0$.

Step 2. First, we need to prove that for every $\widetilde{v}_h \in \ker D\widetilde{c}_h(\widetilde{u}_h)$ there exists $\bar{v}_h \in \ker Dc(\bar{u}_h)$ such that $\|\nabla(\widetilde{v}_h - \bar{v}_h)\|_{\mathrm{L}^2} \leq Ch$ and $\|\nabla\bar{v}_h\|_{\mathrm{L}^2} \geq (1-Ch)\,\|\nabla\widetilde{v}_h\|_{\mathrm{L}^2}$. It is straightforward to show that such a $\bar{v}_h \in \ker Dc(\bar{u}_h) \cap \mathrm{S}_{h,0}$ is given by

$$\bar{v}_h = \widetilde{v}_h - \frac{(\nabla\varphi_h, \nabla\widetilde{v}_h)}{\|\nabla\varphi_h\|^2_{\mathrm{L}^2}}\,\varphi_h, \quad \text{where} \quad \varphi_h = (-\Delta_{h,0})^{-1}\bar{u}_h.$$

Step 3. Next, we will show for the second term on the right-hand side of (3.21) that

$$\big|\big(D^2_{uu}\widetilde{\mathcal{L}}_h(\widetilde{u}_h,\widetilde{\mu}_h) - D^2_{uu}\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)\big)\cdot[\widetilde{v}_h,\widetilde{v}_h]\big| \leq Ch^{\alpha_{F,u}} \quad \forall \widetilde{v}_h \in \mathrm{S}_h.$$

For this we have to look at quadrature errors. Drawing from the techniques used in Lemmas 3.3 and 3.4 and the triangle inequality, we immediately see that

$$\begin{aligned}\left|\left(F''(\bar{u}_h)\widetilde{v}_h,\widetilde{v}_h\right)-\mathcal{Q}_h\left[F''(\widetilde{u}_h)\widetilde{v}_h^2\right]\right| &\leq \left|e_h\left[F''(\bar{u}_h)\widetilde{v}_h^2\right]+\left|\mathcal{Q}_h\left[F''(\bar{u}_h)\widetilde{v}_h^2\right]-\mathcal{Q}_h\left[F''(\widetilde{u}_h)\widetilde{v}_h^2\right]\right|\right.\\ &\leq C\left(h^{\alpha_{F,u}}+h^{\alpha_F}\right)\|\nabla\widetilde{v}_h\|_{\mathrm{L}^2}^2\,.\end{aligned}$$

Similarly, we obtain

$$\left|\bar{\mu}_h\|\widetilde{v}_h\|_{\mathrm{L}^2}^2-\widetilde{\mu}_h\mathcal{Q}_h\left[\widetilde{v}_h^2\right]\right|\leq Ch\|\nabla\widetilde{v}_h\|_{\mathrm{L}^2}^2\,.$$

The difference arising from the electrostatic terms requires more work. First we note that the second derivative of $\widetilde{\Phi}_h$ is given by:

$$D_{uu}^2\widetilde{\Phi}_h(\widetilde{u}_h)\cdot[\widetilde{v}_h,\widetilde{v}_h]=2\mathcal{Q}_h\left[\widetilde{\phi}_h\widetilde{v}_h^2\right]+16\pi\mathcal{Q}_h\left[\widetilde{u}_h\widetilde{v}_h\left(-\widetilde{\Delta}_{h,0}\right)^{-1}(\widetilde{u}_h\widetilde{v}_h)\right].$$

Hence, we need to estimate the error

$$\begin{aligned}\left|\left(D_{uu}^2\Phi_h(\bar{u}_h)-D_{uu}^2\widetilde{\Phi}_h(\widetilde{u}_h)\right)\cdot[\widetilde{v}_h,\widetilde{v}_h]\right| &\leq 2\left|\int_\Omega\bar{\phi}_h\widetilde{v}_h^2\,\mathrm{d}x-\mathcal{Q}_h\left[\widetilde{\phi}_h\widetilde{v}_h^2\right]\right| \\ &+16\pi\left|\int_\Omega\bar{u}_h\widetilde{v}_h\left(-\Delta_{h,0}\right)^{-1}(\bar{u}_h\widetilde{v}_h)\,\mathrm{d}x-\mathcal{Q}_h\left[\widetilde{u}_h\widetilde{v}_h\left(-\widetilde{\Delta}_{h,0}\right)^{-1}(\widetilde{u}_h\widetilde{v}_h)\right]\right|.\end{aligned}\tag{3.22}$$

Looking at the first term we note that

$$\begin{aligned}\left|\int_\Omega\bar{\phi}_h\widetilde{v}_h^2\,\mathrm{d}x-\mathcal{Q}_h\left[\widetilde{\phi}_h\widetilde{v}_h^2\right]\right| &\leq \left|\int_\Omega\bar{\phi}_h\widetilde{v}_h^2\,\mathrm{d}x-\mathcal{Q}_h\left[\bar{\phi}_h\widetilde{v}_h^2\right]\right|+\left|\mathcal{Q}_h\left[\bar{\phi}_h\widetilde{v}_h^2\right]-\mathcal{Q}_h\left[\widetilde{\phi}_h\widetilde{v}_h^2\right]\right| \\ &\leq\left(Ch^{\alpha_u}+Ch\right)\|\widetilde{v}_h\|_{\mathrm{H}^1}^2\,,\end{aligned}$$

with Proposition A.15, the Hölder continuity of $\bar{\phi}_h$ and because of $\left\|\widetilde{\phi}_h-\bar{\phi}_h\right\|_{\mathrm{H}^1}\leq Ch$.

Now we take care of the second term on the right-hand side of (3.22). For given $\widetilde{v}_h\in\mathrm{S}_{h,0}$ we define

$$\bar{\psi}_h=\left(-\Delta_{h,0}\right)^{-1}(\bar{u}_h\widetilde{v}_h),\quad \widetilde{\psi}_h=\left(-\widetilde{\Delta}_{h,0}\right)^{-1}(\bar{u}_h\widetilde{v}_h),\quad \widetilde{\widetilde{\psi}}_h=\left(-\widetilde{\Delta}_{h,0}\right)^{-1}(\widetilde{u}_h\widetilde{v}_h).$$

Testing the weak equations for $\bar{\psi}_h$ and $\widetilde{\psi}_h$ with $(\bar{\psi}_h-\widetilde{\psi}_h)$, we then have with Proposition A.15

$$\begin{aligned}\left\|\nabla\left(\bar{\psi}_h-\widetilde{\psi}_h\right)\right\|_{\mathrm{L}^2}^2 &=\left|\left(\bar{u}_h\widetilde{v}_h,\bar{\psi}_h-\widetilde{\psi}_h\right)-\mathcal{Q}_h\left[\bar{u}_h\widetilde{v}_h\left(\bar{\psi}_h-\widetilde{\psi}_h\right)\right]\right| \\ &\leq Ch^{\alpha_u}\|\widetilde{v}_h\|_{\mathrm{H}^1}\left\|\bar{\psi}_h-\widetilde{\psi}_h\right\|_{\mathrm{H}^1},\end{aligned}$$

which implies $\left\|\bar{\psi}_h-\widetilde{\psi}_h\right\|_{\mathrm{H}^1}\leq Ch^{\alpha_u}$. Using (3.19) we also get

$$\left\|\widetilde{\widetilde{\psi}}_h-\widetilde{\psi}_h\right\|_{\mathrm{H}^1}\leq C\|\bar{u}_h-\widetilde{u}_h\|_{\mathrm{H}^1}\|\widetilde{v}_h\|_{\mathrm{H}^1}\leq Ch\|\widetilde{v}_h\|_{\mathrm{H}^1}\,.$$

Thus, we deduce

$$\begin{aligned}
&\left|\int_\Omega \bar{u}_h\widetilde{v}_h\big(-\Delta_{h,0}\big)^{-1}\,(\bar{u}_h\widetilde{v}_h)\,\mathrm{d}x \;-\; \mathcal{Q}_h\Big[\widetilde{u}_h\widetilde{v}_h\big(-\widetilde{\Delta}_{h,0}\big)^{-1}(\widetilde{u}_h\widetilde{v}_h)\Big]\right| \\
&\qquad\le \int_\Omega \big|\bar{u}_h\widetilde{v}_h(\bar{\psi}_h-\widetilde{\psi}_h)\big|\,\mathrm{d}x \;+\; \left|\int_\Omega \bar{u}_h\widetilde{v}_h\widetilde{\psi}_h\,\mathrm{d}x \;-\; \mathcal{Q}_h\Big[\bar{u}_h\widetilde{v}_h\widetilde{\psi}_h\Big]\right| \\
&\qquad\quad + \big|\mathcal{Q}_h\big[(\bar{u}_h-\widetilde{u}_h)\widetilde{v}_h\widetilde{\psi}_h\big]\big| + \big|\mathcal{Q}_h\big[\widetilde{u}_h\widetilde{v}_h\big(\widetilde{\bar{\psi}}_h-\widetilde{\psi}_h\big)\big]\big| \\
&\qquad\le\; \big(Ch^{\alpha_u}+Ch^{\alpha_u}+Ch+Ch\big)\,\|\widetilde{v}_h\|_{\mathrm{H}^1}^2\,.
\end{aligned}$$

Summarizing we have shown

$$\big|\big(D^2_{uu}\Phi_h(\bar{u}_h)-D^2_{uu}\widetilde{\Phi}_h(\widetilde{u}_h)\big)\cdot[\widetilde{v}_h,\widetilde{v}_h]\big|\le Ch^{\alpha_u}.$$

This shows that the second term on the right-hand side of (3.21) goes to zero as $h\to 0$. The third and fourth term are small since $\|\bar{v}_h-\widetilde{v}_h\|_{\mathrm{H}^1}\le Ch$. This proves the existence of $\widetilde{\gamma}>0$ such that (3.20) is satisfied for sufficiently small h. □

Remark 3.7. We note that for the discretization of an elliptic equation of the form

$$-\Delta u + a\,u = f,\quad u|_{\partial\Omega}=0,$$

with sufficiently smooth coefficient function a and right-hand side f it also takes a quadrature rule that is exact for polynomials of degree $2\mathrm{p}-1$ to get the optimal convergence order h^{p} [37, Section 4.1]. We conclude that the nonlinearity of the TFDW functional does not lead to increased quadrature demands on its discretization. □

3.3 Optimal Convergence Rates

Having established the existence of a solution $\widetilde{u}_h$ to the discretized minimization problem with quadrature, we will look at the convergence rates of the energy, the Lagrange multiplier and the L^2-errors. The techniques we will use are very similar to the study of convergence rates in the Galerkin case but quadrature errors have to be addressed.

In the periodic case we are, under some additional differentiability assumptions on F and for $\mathrm{p}\ge 2$, able to show the optimal convergence rates of $2\mathrm{p}$ for the energy and the Lagrange multiplier, as well as $\mathrm{p}+1$ for the L^2-errors of $\widetilde{u}_h$ and $\widetilde{\phi}_h$. If $\mathrm{p}=1$, the situation is more involved. The convergence rate 2 can still be shown for the energy. However, the functions $\bar{u}$, $\bar{\phi}$ and their finite element counterparts are generally only Hölder continuous, which shows in the convergence rates of $\widetilde{\mu}_h$ and the L^2-errors.

Where necessary, we will consider the Dirichlet and the periodic case separately and provide two proofs. The discretization parameter h is implicitly assumed to be sufficiently small.

Proposition 3.8.
(D) *Let $\widetilde{y}_h = (\widetilde{u}_h, \widetilde{\phi}_h, \widetilde{\mu}_h) \in \mathcal{Y}_{h,D}$ be a solution of the discrete problem with quadrature* (3.1) *such that $\|\bar{y} - \widetilde{y}_h\|_{\mathcal{Y}} \le Ch$. Assume that $u_{\mathrm{ex},h} = \mathcal{I}_h u_{\mathrm{ex}}$, $\phi_{\mathrm{ex},h} = \mathcal{I}_h \phi_{\mathrm{ex}}$ and u_{ex} and ϕ_{ex} belong to $\mathrm{H}^2(\Gamma)$ for all affine parts Γ of $\partial\Omega$. Then,*

$$|E(\bar{u}) - \widetilde{E}_h(\widetilde{u}_h)| \le Ch^2.$$

(P) *Let $\widetilde{y}_h = (\widetilde{u}_h, \widetilde{\phi}_h, \widetilde{\mu}_h) \in \mathcal{Y}_{h,\#}$ be a solution of the discrete periodic problem with quadrature* (3.5) *such that $\|\bar{y} - \widetilde{y}_h\|_{\mathcal{Y}} \le Ch^{\mathrm{p}}$. Moreover, assume that $F \in \mathrm{C}^{2\mathrm{p}}(I_{\bar{u}})$. Then,*

$$|E_{\#}(\bar{u}) - \widetilde{E}_{h,\#}(\widetilde{u}_h)| \le Ch^{2\mathrm{p}}.$$

Proof. In Propositions 2.23 and 2.16 we saw that under the given conditions

$$\begin{aligned} |E(\bar{u}) - E_h(\bar{u}_h)| &\le Ch^2, \quad \text{respectively,} \\ |E_{\#}(\bar{u}) - E_{h,\#}(\bar{u}_h)| &\le Ch^{2\mathrm{p}}. \end{aligned}$$

Since $\widetilde{u}_h$ minimizes $\widetilde{E}_h$, respectively, $\widetilde{E}_{h,\#}$, we can show similarly that

$$\begin{aligned} \big|\widetilde{E}_h(\bar{u}_h) - \widetilde{E}_h(\widetilde{u}_h)\big| &\le C\|\bar{y}_h - \widetilde{y}_h\|_{\mathcal{Y}}^2 \le Ch^2, \\ \big|\widetilde{E}_{h,\#}(\bar{u}_h) - \widetilde{E}_{h,\#}(\widetilde{u}_h)\big| &\le C\|\bar{y}_h - \widetilde{y}_h\|_{\mathcal{Y}}^2 \le Ch^{2\mathrm{p}}. \end{aligned}$$

Here we have also used that in the Dirichlet case the discrete boundary conditions for $\bar{u}_h$ and $\widetilde{u}_h$ are the same.

Hence, we now need to prove that $\big|E_h(\bar{u}_h) - \widetilde{E}_h(\bar{u}_h)\big| \le Ch^2$, respectively, $\big|E_{h,\#}(\bar{u}_h) - \widetilde{E}_{h,\#}(\bar{u}_h)\big| \le Ch^{2\mathrm{p}}$. The remainder of the proof consists in a thorough study of quadrature errors, which is the same for both cases. We present the proof using the Dirichlet case but admit arbitrary $\mathrm{p} \ge 1$.

We start by noting that

$$\big|E_h(\bar{u}_h) - \widetilde{E}_h(\bar{u}_h)\big| \le \left| \int_\Omega F(\bar{u}_h)\,\mathrm{d}x \, - \mathcal{Q}_h[F(\bar{u}_h)] \right| + \big|\Phi_h(\bar{u}_h) - \widetilde{\Phi}_h(\bar{u}_h)\big|. \tag{3.23}$$

Step 1. First, we aim to prove $e_h[F(\bar{u}_h)] \le Ch^{2\mathrm{p}}$. The appropriate tool for this is provided by Proposition A.14. We thus need a bound on $\sum_{T\in\mathcal{T}_h} \|F(\bar{u}_h)\|^q_{\mathrm{W}^{2\mathrm{p},q}(T)}$, which we now derive.

If $\mathrm{p} = 1$, we directly get

$$\begin{aligned} |F(\bar{u}_h)|_{\mathrm{W}^{1,q}(T)} &\le \|F'(\bar{u}_h)\|_{\mathrm{L}^\infty} \|\nabla \bar{u}_h\|_{\mathrm{L}^q(T)}, \\ |F(\bar{u}_h)|_{\mathrm{W}^{2,q}(T)} &\le \|F''(\bar{u}_h)\|_{\mathrm{L}^\infty} \|\nabla \bar{u}_h \otimes \nabla \bar{u}_h\|_{\mathrm{L}^q(T)} \le C\|\nabla \bar{u}_h\|^2_{\mathrm{L}^{2q}(T)}, \end{aligned}$$

for $1 < q < 3$ and all $T \in \mathcal{T}_h$. More generally, for $m \geq 2$,

$$\begin{aligned}\big|\nabla^m F(\bar{u}_h)\big| \leq C(m)\Big(&|F^{(m)}(\bar{u}_h)||\nabla\bar{u}_h|^m \\ &+ \big|F^{(m-1)}(\bar{u}_h)\big||\nabla\bar{u}_h|^{m-2}|\nabla^2\bar{u}_h| \\ &+ \big|F^{(m-2)}(\bar{u}_h)\big|\big(|\nabla\bar{u}_h|^{m-3}|\nabla^3\bar{u}_h| + |\nabla\bar{u}_h|^{m-4}|\nabla^2\bar{u}_h|^2\big) \\ &+ \big|F^{(m-3)}(\bar{u}_h)\big|\big(|\nabla\bar{u}_h|^{m-4}|\nabla^4\bar{u}_h| + |\nabla\bar{u}_h|^{m-5}|\nabla^2\bar{u}_h||\nabla^3\bar{u}_h| \\ &\qquad\qquad + |\nabla\bar{u}_h|^{m-6}|\nabla^2\bar{u}_h|^2\big) \\ &\vdots \\ &+ |F'(\bar{u}_h)||\nabla^m\bar{u}_h|\Big).\end{aligned}$$

We know that $\nabla^j\bar{u}_h = 0$ for $j > \mathrm{p}$ since $\bar{u}_h$ is a piecewise polynomial of degree p. Moreover, we know from Lemma 3.2 that for $m < \mathrm{p}$ and for sufficiently small h the derivatives $|\nabla^m\bar{u}_h|$ are uniformly bounded in $\mathrm{L}^\infty(T)$ for all $T \in \mathcal{T}_h$. From this we deduce that, for all $0 \leq m \leq 2\mathrm{p}$,

$$\big|\nabla^m F(\bar{u}_h)\big| \leq C\big(1 + |\nabla^{\mathrm{p}}\bar{u}_h| + |\nabla^{\mathrm{p}}\bar{u}_h|^2\big),$$

where C depends on F and its derivatives as well as $\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)}$ but is independent of T and h. Hence, summing over m and integrating over T leads to

$$\|F(\bar{u}_h)\|_{\mathrm{W}^{2\mathrm{p},q}(T)} \leq C\left(\|F(\bar{u}_h)\|_{\mathrm{L}^q(T)} + \|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)} + \|\bar{u}_h\|^2_{\mathrm{W}^{\mathrm{p},2q}(T)}\right).$$

Summing over $T \in \mathcal{T}_h$ yields

$$\sum_{T\in\mathcal{T}_h}\|F(\bar{u}_h)\|^q_{\mathrm{W}^{2\mathrm{p},q}(T)} \leq C\left(\|F(\bar{u}_h)\|^q_{\mathrm{L}^q(\Omega)} + \sum_{T\in\mathcal{T}_h}\|\bar{u}_h\|^q_{\mathrm{W}^{\mathrm{p},q}(T)} + \sum_{T\in\mathcal{T}_h}\|\bar{u}_h\|^{2q}_{\mathrm{W}^{\mathrm{p},2q}(T)}\right) \leq C,$$

uniformly in h. Applying Proposition A.14 we obtain

$$\left|\int_\Omega F(\bar{u}_h)\,\mathrm{d}x \,-\, \mathcal{Q}_h[F(\bar{u}_h)]\right| \leq Ch^{2\mathrm{p}},$$

as desired.

Step 2. We now address the second term on the right-hand side of equation (3.23). The goal is to prove that

$$\big|\Phi_h(\bar{u}_h) - \widetilde{\Phi}_h(\bar{u}_h)\big| = \big|\Psi(\bar{u}_h,\bar{\phi}_h) - \widetilde{\Psi}_h(\bar{u}_h,\bar{\phi}_h^*)\big| \leq Ch^{2\mathrm{p}}.$$

Here, $\bar{\phi}_h^* \in \mathrm{S}_h$ solves the discrete boundary value problem

$$\begin{aligned}(\nabla\bar{\phi}_h^*, \nabla v_h) &= 4\pi\mathcal{Q}_h\big[(\bar{u}_h^2 - \rho_n)v_n\big] \quad \forall v_h \in \mathrm{S}_{h,0}, \\ \bar{\phi}_h^*|_{\partial\Omega} &= \phi_{\mathrm{ex},h}.\end{aligned}$$

With the triangle inequality we get

$$\big|\Psi(\bar{u}_h,\bar{\phi}_h) - \widetilde{\Psi}_h(\bar{u}_h,\bar{\phi}_h^*)\big| \leq \big|\Psi(\bar{u}_h,\bar{\phi}_h) - \widetilde{\Psi}_h(\bar{u}_h,\bar{\phi}_h)\big| + \big|\widetilde{\Psi}_h(\bar{u}_h,\bar{\phi}_h) - \widetilde{\Psi}_h(\bar{u}_h,\bar{\phi}_h^*)\big|. \quad (3.24)$$

Let us start with the second term on the right-hand side of (3.24). To begin with, we estimate the error $\|\bar{\phi}_h - \bar{\phi}_h^*\|_{\mathrm{H}^1(\Omega)}$. Testing the equations that $\bar{\phi}_h$ and $\bar{\phi}_h^*$ satisfy with $\bar{\phi}_h^* - \bar{\phi}_h \in \mathrm{S}_{h,0}$, we get

$$\|\nabla(\bar{\phi}_h^* - \bar{\phi}_h)\|_{\mathrm{L}^2(\Omega)}^2 \leq 4\pi \left| \int_\Omega (\bar{u}_h^2 - \rho_n)(\bar{\phi}_h^* - \bar{\phi}_h)\,\mathrm{d}x \, - \mathcal{Q}_h\big[(\bar{u}_h^2 - \rho_n)(\bar{\phi}_h^* - \bar{\phi}_h)\big] \right|.$$

Since the reference quadrature rule $\widehat{\mathcal{Q}}$ is exact for all polynomials of degree $2\mathrm{p}-1$ we then get

$$\big\|\nabla(\bar{\phi}_h^* - \bar{\phi}_h)\big\|_{\mathrm{L}^2(\Omega)}^2 \leq Ch^{\mathrm{p}} \left(\|\rho_n\|_{\mathrm{W}^{\mathrm{p},q}} + \Big(\sum_{T\in\mathcal{T}_h} \|\bar{u}_h^2\|_{\mathrm{W}^{\mathrm{p},q}(T)}^q \Big)^{1/q} \right) \|\bar{\phi}_h^* - \bar{\phi}_h\|_{\mathrm{H}^1(\Omega)}$$

for $q > d/\mathrm{p}$ by Proposition A.11. In Lemma 3.2 we have shown that $\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)} \leq C$ and $\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)} \leq C$ for all $T \in \mathcal{T}_h$ and $2 < q < \frac{2d}{d-2}$. Treating the terms $\|\bar{u}_h^2\|_{\mathrm{W}^{\mathrm{p},q}(T)}^q$ similarly as in the proof of Lemma 3.4 we deduce that

$$\begin{aligned}
\big\|\bar{\phi}_h^* - \bar{\phi}_h\big\|_{\mathrm{H}^1} &\leq C\|\nabla(\bar{\phi}_h^* - \bar{\phi}_h)\|_{\mathrm{L}^2(\Omega)} \\
&\leq Ch^{\mathrm{p}} \left(\|\rho_n\|_{\mathrm{W}^{\mathrm{p},q}} + \Big(\sum_{T\in\mathcal{T}_h} \|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)}^q \Big)^{1/q} \right) \\
&\leq Ch^{\mathrm{p}},
\end{aligned}$$

uniformly in h for $3 < q < 6$. Since $\bar{\phi}_h^*$ is a minimizer of $\widetilde{\Psi}(\bar{u}_h, \cdot)$ and $\|\bar{\phi}_h - \bar{\phi}_h^*\|_{\mathrm{H}^1} \leq Ch^{\mathrm{p}}$ as just shown, we obtain

$$\Big| \widetilde{\Psi}_h(\bar{u}_h, \bar{\phi}_h) - \widetilde{\Psi}_h(\bar{u}_h, \bar{\phi}_h^*) \Big| \leq Ch^{2\mathrm{p}}. \tag{3.25}$$

Finally, we turn our attention to the first term on the right-hand side of (3.24):

$$\Big| \Psi(\bar{u}_h, \bar{\phi}_h) - \widetilde{\Psi}_h(\bar{u}_h, \bar{\phi}_h) \Big| = \left| \int_\Omega (\bar{u}_h^2 - \rho_n)\bar{\phi}_h\,\mathrm{d}x \, - \mathcal{Q}_h\big[(\bar{u}_h^2 - \rho_n)\bar{\phi}_h\big] \right|.$$

Proposition A.14 implies that this term is of order $h^{2\mathrm{p}}$ provided

$$\sum_{T\in\mathcal{T}_h} \|(\bar{u}_h^2 - \rho_n)\bar{\phi}_h\|_{\mathrm{W}^{2\mathrm{p},q}(T)}^q \leq C, \tag{3.26}$$

for $q > 3/(2\mathrm{p})$, uniformly in h. Similarly as in the proof of Lemma 3.4 or in Step 1 above we can deduce that

$$\begin{aligned}
&\big\|\bar{u}_h^2\bar{\phi}_h\big\|_{\mathrm{W}^{2\mathrm{p},q}(T)} \leq C\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)}\big(\|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)} + \|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)}\big), \quad \text{and} \\
&\|\rho_n\bar{\phi}_h\|_{\mathrm{W}^{2\mathrm{p},q}(T)} \leq C\|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p},q}(T)},
\end{aligned}$$

where the constants depend on $\|\rho_n\|_{\mathrm{W}^{2\mathrm{p},\infty}(T)}$, $\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)}$, and $\|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p}-1,\infty}(T)}$. This and Lemma 3.1 directly imply equation (3.26) for $2 < q < \frac{d}{d-2}$, which together with (3.24) and (3.25) concludes the proof. ◻

Next, we look at convergence rates for the L^2-errors and the Lagrange multiplier. The proofs will use duality concepts as in the Galerkin case and rely on the analysis of certain quadrature errors. First, we address the Dirichlet case.

Proposition 3.9. (D) *Let $\widetilde{y}_h = (\widetilde{u}_h, \widetilde{\phi}_h, \widetilde{\mu}_h) \in \mathcal{Y}_{h,D}$ be a solution of the discretization (3.1) satisfying $\|\bar{y} - \widetilde{y}_h\|_{\mathcal{Y}} \leq Ch$. Assume that $\alpha_F \geq \frac{1}{2}$. Then, there exists a constant $C > 0$ such that*

$$\|\bar{u}_h - \widetilde{u}_h\|_{\mathrm{L}^2} + \|\bar{\phi}_h - \widetilde{\phi}_h\|_{\mathrm{L}^2} + |\bar{\mu}_h - \widetilde{\mu}_h| \leq Ch^{1+\alpha_{F,u}},$$

and hence

$$\|\bar{u} - \widetilde{u}_h\|_{\mathrm{L}^2} + \|\bar{\phi} - \widetilde{\phi}_h\|_{\mathrm{L}^2} + |\bar{\mu} - \widetilde{\mu}_h| \leq Ch^{1+\alpha_{F,u}}.$$

Proof. Let $f = (f_1, f_2, f_3) \in \mathrm{L}^2(\Omega) \times \mathrm{L}^2(\Omega) \times \mathbb{R} \subset \mathcal{Y}_{h,0}^*$ be a linear functional given through $\langle f, v\rangle = (f_1, u) + (f_2, \phi) + f_3\mu$ for all $y = (u, \phi, \mu) \in \mathcal{Y}_{h,0}$. As in the corresponding proof in the Galerkin case (see Proposition 2.17 for the periodic case) we define the Lagrangian functionals $\mathcal{L}_h, \widetilde{\mathcal{L}}_h : \mathcal{Y}_h \times \mathcal{Y}_{h,0} \to \mathbb{R}$ by

$$\begin{aligned}\mathcal{L}_h(y_h, z_h) &= \langle f, y_h\rangle - \langle \mathcal{F}_h(y_h), z_h\rangle,\\ \widetilde{\mathcal{L}}_h(y_h, z_h) &= \langle f, y_h\rangle - \langle \widetilde{\mathcal{F}}_h(y_h), z_h\rangle.\end{aligned}$$

From Proposition 2.24 we already know that $(\bar{y}_h, \bar{z}_h) \in \mathcal{Y}_{h,D} \times \mathcal{Y}_{h,0}$, where $\bar{z}_h$ is the solution of the dual equation

$$\langle D\mathcal{F}_h(\bar{y}_h)\cdot\eta_h, \bar{z}_h\rangle = \langle f, \eta_h\rangle \quad \forall \eta_h \in \mathcal{Y}_{h,0},$$

is a stationary point of $\mathcal{L}_h|_{\mathcal{Y}_{h,D}\times\mathcal{Y}_{h,0}}$. Since $D\widetilde{\mathcal{F}}_h(\widetilde{y}_h) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$ is an isomorphism (see Lemma 3.3), there exists $\widetilde{z}_h \in \mathcal{Y}_{h,0}$ such that

$$\langle D\widetilde{\mathcal{F}}_h(\widetilde{y}_h)\cdot\eta_h, \widetilde{z}_h\rangle = \langle f, \eta_h\rangle \quad \forall \eta_h \in \mathcal{Y}_{h,0}.$$

In other words, $(\widetilde{y}_h, \widetilde{z}_h) \in \mathcal{Y}_{h,D} \times \mathcal{Y}_{h,0}$ is a stationary point of $\widetilde{\mathcal{L}}_h|_{\mathcal{Y}_{h,D}\times\mathcal{Y}_{h,0}}$.

Step 1. First, we need to prove that $\|\widetilde{z}_h - \bar{z}_h\|_{\mathcal{Y}} \leq Ch^\gamma$ for some exponent $\gamma > 0$. Using one of the inf-sup conditions for $D\widetilde{\mathcal{F}}_h(\widetilde{y}_h)$ we get

$$\begin{aligned}\widetilde{\kappa}_h\|\widetilde{z}_h - \bar{z}_h\|_{\mathcal{Y}} &\leq \sup_{\eta_h\in\mathcal{Y}_{h,0}} \frac{\langle D\widetilde{\mathcal{F}}_h(\widetilde{y}_h)\cdot\eta_h, \widetilde{z}_h - \bar{z}_h\rangle}{\|\eta_h\|_{\mathcal{Y}}}\\ &\leq \sup_{\eta_h\in\mathcal{Y}_{h,0}} \frac{\big\langle \big(D\mathcal{F}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\bar{y}_h)\big)\cdot\eta_h, \bar{z}_h\big\rangle}{\|\eta_h\|_{\mathcal{Y}}}\\ &\quad + \sup_{\eta_h\in\mathcal{Y}_{h,0}} \frac{\big\langle \big(D\widetilde{\mathcal{F}}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\widetilde{y}_h)\big)\cdot\eta_h, \bar{z}_h\big\rangle}{\|\eta_h\|_{\mathcal{Y}}},\end{aligned} \tag{3.27}$$

where we have used $\langle D\widetilde{\mathcal{F}}_h(\widetilde{y}_h)\cdot\eta_h, \widetilde{z}_h\rangle = \langle D\mathcal{F}_h(\bar{y}_h)\cdot\eta_h, \bar{z}_h\rangle$ for all $\eta_h \in \mathcal{Y}_{h,0}$ by the definition of $\bar{z}_h$ and $\widetilde{z}_h$.

For the second term on the right-hand side of (3.27) we have

$$\sup_{\eta_h\in\mathcal{Y}_{h,0}} \frac{\langle \big(D\widetilde{\mathcal{F}}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\widetilde{y}_h)\big)\cdot\eta_h, \bar{z}_h\rangle}{\|\eta_h\|_{\mathcal{Y}}} \leq C(h^{\alpha_F}+h)\|\bar{z}_h\|_{\mathcal{Y}}$$

because of the Hölder continuity of $D\widetilde{\mathcal{F}}_h$ and $\|\bar{y}_h-\widetilde{y}_h\|\leq Ch$. The first term on the right-hand side of (3.27) satisfies

$$\sup_{\eta_h\in\mathcal{Y}_{h,0}} \frac{\langle \big(D\mathcal{F}_h(\bar{y}_h) - D\widetilde{\mathcal{F}}_h(\bar{y}_h)\big)\cdot\eta_h, \bar{z}_h\rangle}{\|\eta_h\|_{\mathcal{Y}}} \leq Ch^{\alpha_{F,u}}\|\bar{z}_h\|_{\mathcal{Y}},$$

as shown in Lemma 3.3 (see (3.13) and the following discussion). Summarizing, we have established the following error bound for $\widetilde{z}_h$ and $\bar{z}_h$:

$$\|\widetilde{z}_h-\bar{z}_h\|_{\mathcal{Y}}\leq C(h^{\alpha_F}+h^{\alpha_{F,u}}+h)\leq Ch^{\alpha_{F,u}}.$$

Step 2. We point out that under the assumption $\alpha_F\geq\frac{1}{2}$, we get the optimal convergence rate $\|\bar{z}-\bar{z}_h\|_{\mathcal{Y}}\leq Ch$ for the discrete (Galerkin) dual solution $\bar{z}_h$ from Proposition 2.24. Following the same lines as in Section 2.5 we can show that

$$\|\bar{z}_u-\bar{z}_{u,h}\|_{\mathrm{L}^2}+\|\bar{z}_\phi-\bar{z}_{\phi,h}\|_{\mathrm{L}^2}\leq Ch^2.$$

This implies (compare Section 2.5.1.1) the convergence, and hence uniform boundedness, of $\bar{z}_{u,h}$ and $\bar{z}_{\phi,h}$ with respect to Hölder norms:

$$\|\bar{z}_u-\bar{z}_{u,h}\|_{\mathrm{C}^{0,\alpha_z}(T)}+\|\bar{z}_\phi-\bar{z}_{\phi,h}\|_{\mathrm{C}^{0,\alpha_z}(T)}\leq Ch^{2-\frac{d}{2}-\alpha_z}$$

for all Hölder indices $0<\alpha_z<2-\frac{d}{2}$.

Step 3. From the definitions of $\mathcal{L}_h$ and $\widetilde{\mathcal{L}}_h$ and $\mathcal{F}_h(\bar{y}_h)=0$, $\widetilde{\mathcal{F}}_h(\widetilde{y}_h)=0$ we derive that

$$\begin{aligned}\langle f,\bar{y}_h-\widetilde{y}_h\rangle &= \mathcal{L}_h(\bar{y}_h,\bar{z}_h)-\widetilde{\mathcal{L}}_h(\widetilde{y}_h,\widetilde{z}_h)\\ &= \big(\mathcal{L}_h(\bar{y}_h,\bar{z}_h)-\widetilde{\mathcal{L}}_h(\bar{y}_h,\bar{z}_h)\big)+\big(\widetilde{\mathcal{L}}_h(\bar{y}_h,\bar{z}_h)-\widetilde{\mathcal{L}}_h(\widetilde{y}_h,\widetilde{z}_h)\big).\end{aligned}$$

With a Taylor expansion argument as in the proof of Proposition 2.17 for the Galerkin case we can show for the second term on the right-hand side that

$$\begin{aligned}\Big|\widetilde{\mathcal{L}}_h(\bar{y}_h,\bar{z}_h)-\widetilde{\mathcal{L}}_h(\widetilde{y}_h,\widetilde{z}_h)\Big| &\leq C\left(\|\bar{y}_h-\widetilde{y}_h\|_{\mathcal{Y}}^2+\|\bar{y}_h-\widetilde{y}_h\|_{\mathcal{Y}}^{1+\alpha_F}+\|\bar{z}_h-\widetilde{z}_h\|_{\mathcal{Y}}\|\bar{y}_h-\widetilde{y}_h\|_{\mathcal{Y}}\right)\\ &\leq C\big(h^2+h^{1+\alpha_F}+h^{1+\alpha_{F,u}}\big),\end{aligned}$$

since $(\widetilde{y}_h,\widetilde{z}_h)$ is a stationary point of $\widetilde{\mathcal{L}}_h|_{\mathcal{Y}_{h,D}\times\mathcal{Y}_{h,0}}$.

The estimation of $\big|\mathcal{L}_h(\bar{y}_h,\bar{z}_h)-\widetilde{\mathcal{L}}_h(\bar{y}_h,\bar{z}_h)\big|$ once again comes down to the study of quadrature errors:

$$\begin{aligned}\big|\mathcal{L}_h(\bar{y}_h,\bar{z}_h)-\widetilde{\mathcal{L}}_h(\bar{y}_h,\bar{z}_h)\big| &= \big|\langle\mathcal{F}_h(\bar{y}_h)-\widetilde{\mathcal{F}}_h(\bar{y}_h),\bar{z}_h\rangle\big|\\ &\leq C\Big(\big|e_h\big[F'(\bar{u}_h)\bar{z}_{u,h}\big]\big|+\big|e_h\big[\bar{\phi}_h\bar{u}_h\bar{z}_{u,h}\big]\big|+\big|e_h[\bar{u}_h\bar{z}_{u,h}]\big|\\ &\qquad+\big|e_h\big[(\bar{u}_h^2-\rho_n)\bar{z}_{\phi,h}\big]\big|+\big|e_h\big[\bar{u}_h^2\big]\big|\Big).\end{aligned}\tag{3.28}$$

Note that we have absorbed some constants and Lagrange multipliers into the constant C.

Similarly to the treatment of some terms in the proof of Proposition 3.8 we can estimate Sobolev norms of $\bar\phi_h\bar u_h\bar z_{u,h}$ for $2<q<\frac{d}{d-2}$:

$$\begin{aligned}\|\bar\phi_h\bar u_h\bar z_{u,h}\|_{\mathrm{W}^{2,q}(T)} &\le C\|\bar z_{u,h}\|_{\mathrm{L}^\infty(T)}\|\bar u_h\|_{\mathrm{W}^{1,2q}(T)}\|\bar\phi_h\|_{\mathrm{W}^{1,2q}(T)}\\ &\quad + C\|\bar z_{u,h}\|_{\mathrm{W}^{1,2q}(T)}\big(\|\bar u_h\|_{\mathrm{W}^{1,2q}(T)}+\|\bar\phi_h\|_{\mathrm{W}^{1,2q}(T)}\big),\end{aligned}$$

where we have used that $\bar u_h,\bar\phi_h,\bar z_{u,h}\in\mathrm{L}^\infty(T)$ uniformly in $T\in\mathcal{T}_h$ and h by Lemma 3.2. Therefore,

$$\bigg(\sum_{T\in\mathcal{T}_h}\|\bar\phi_h\bar u_h\bar z_{u,h}\|^q_{\mathrm{W}^{2,q}(T)}\bigg)^{1/q}\le C\bigg(1+\sum_{T\in\mathcal{T}_h}\|\bar z_{u,h}\|^{2q}_{\mathrm{W}^{1,2q}(T)}\bigg)^{1/2q}\le C.$$

Analogous estimates can be proved for $(\bar u_h^2-\rho_n)\bar z_{\phi,h}$, $\bar u_h\bar z_{u,h}$, and $\bar u_h^2$ in (3.28). Therefore, by Proposition A.14

$$\big|e_h\big[\bar\phi_h\bar u_h\bar z_{u,h}\big]\big|+\big|e_h[\bar u_h\bar z_{u,h}]\big|+\big|e_h\big[(\bar u_h^2-\rho_n)\bar z_{\phi,h}\big]\big|+\big|e_h\big[\bar u_h^2\big]\big|\le Ch^2.$$

The error $e_h[F'(\bar u_h)\bar z_{u,h}]$ in (3.28) needs a slightly different treatment. We will prove that

$$F'(\bar u_h)\bar z_{u,h}\in\mathrm{C}^{1,\alpha_{F,u}}(T),\quad\text{uniformly in }\ T\in\mathcal{T}_h\text{ and }h,\tag{3.29}$$

and use Proposition A.14 to show that $e_h[F'(\bar u_h)\bar z_{u,h}]\le Ch^{1+\alpha_{F,u}}$.

To begin with, we have

$$\|F'(\bar u_h)\bar z_{u,h}\|_{\mathrm{C}^0(T)}\le C,$$

uniformly in $T\in\mathcal{T}_h$ and h, because F' is locally Lipschitz continuous and $\bar u_h$ and $\bar z_h$ are bounded in $\mathrm{C}^0(\Omega)$. Differentiating gives

$$\nabla\big(F'(\bar u_h)\bar z_{u,h}\big)=F''(\bar u_h)\bar z_{u,h}\nabla\bar u_h+F'(\bar u_h)\nabla\bar z_{u,h}.$$

Thus, since $\nabla\bar u_h$ and $\nabla\bar z_{u,h}$ are constant on every $T\in\mathcal{T}_h$,

$$\begin{aligned}|F'(\bar u_h)\bar z_{u,h}|_{\mathrm{C}^{1,\alpha_{F,u}}(T)}&\le\|\nabla\bar u_h\|_{\mathrm{L}^\infty(T)}\|F''(\bar u_h)\bar z_{u,h}\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)}\\&\quad+\|\nabla\bar z_{u,h}\|_{\mathrm{L}^\infty(T)}\|F'(\bar u_h)\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)}.\end{aligned}$$

We have $\|F''(\bar u_h)\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)}<C$, see the proof of Lemma 3.3. Moreover, $\|\bar z_{u,h}\|_{\mathrm{C}^{0,\alpha_z}(T)}<C$ as seen in Step 2. From this we deduce that

$$\begin{aligned}\|F''(\bar u_h)\bar z_h\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)}&\le\|F''(\bar u_h)\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)}\|\bar z_{u,h}\|_{\mathrm{L}^\infty(T)}+\|F''(\bar u_h)\|_{\mathrm{L}^\infty(T)}\|\bar z_{u,h}\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)}\\&\le C.\end{aligned}$$

Since F' is locally Lipschitz continuous and $\bar{u}_h$ is uniformly Hölder continuous with exponent $\alpha_u \geq \alpha_{F,u}$ we have

$$\|F'(\bar{u}_h)\|_{\mathrm{C}^{0,\alpha_{F,u}}(T)} < C,$$

uniformly in $T \in \mathcal{T}_h$ and h. Combining the previous equations, we obtain (3.29).

Since $\mathrm{p} = 1$, $\nabla \bar{u}_h$ and $\nabla \bar{z}_{u,h}$ are constant on every $T \in \mathcal{T}_h$. This leads to

$$\begin{aligned}\|\nabla \bar{u}_h\|_{\mathrm{L}^\infty(T)} &= C(\det B_T)^{-1/2}\|\nabla \bar{u}_h\|_{\mathrm{L}^2(T)},\\ \|\nabla \bar{z}_{u,h}\|_{\mathrm{L}^\infty(T)} &= C(\det B_T)^{-1/2}\|\nabla \bar{z}_{u,h}\|_{\mathrm{L}^2(T)},\end{aligned}$$

where, by quasi-uniformity of the meshes, the constant C is independent of h and $T \in \mathcal{T}_h$. Hence, using Proposition A.14 and (3.29), we get

$$\begin{aligned}\big|e_h\big[F'(\bar{u}_h)\bar{z}_{1,h}\big]\big| &\leq C\sum_{T\in\mathcal{T}_h} h_T^{1+\alpha_{F,u}} \det B_T (\det B_T)^{-1/2}(\|\nabla \bar{u}_h\|_{\mathrm{L}^2(T)} + \|\nabla \bar{z}_{u,h}\|_{\mathrm{L}^2(T)})\\ &\leq Ch^{1+\alpha_{F,u}}|\Omega|^{1/2}(\|\nabla \bar{u}_h\|_{\mathrm{L}^2} + \|\nabla \bar{z}_{u,h}\|_{\mathrm{L}^2}).\end{aligned}$$

Note that we have absorbed the Hölder norms $\|F'(\bar{u}_h)\bar{z}_{u,h}\|_{\mathrm{C}^{1,\alpha_{F,u}}(T)}$ into the constant C.

Summarizing, we have shown that

$$|\langle f, \bar{y}_h - \widetilde{y}_h\rangle| \leq C\big(h^2 + h^{1+\alpha_F} + h^{1+\alpha_u} + h^{1+\alpha_{F,u}}\big) \leq Ch^{1+\alpha_{F,u}}.$$

Choosing

$$f = \frac{(\bar{u}_h - \widetilde{u}_h, \bar{\phi}_h - \widetilde{\phi}_h, \bar{\mu}_h - \widetilde{\mu}_h)}{\big(\|\bar{u}_h - \widetilde{u}_h\|_{\mathrm{L}^2}^2 + \|\bar{\phi}_h - \widetilde{\phi}_h\|_{\mathrm{L}^2}^2 + |\bar{\mu}_h - \widetilde{\mu}_h|^2\big)^{1/2}} \in \mathrm{S}_h \times \mathrm{S}_h \times \mathbb{R}$$

completes the proof. ◻

Next, we consider the periodic case when $\mathrm{p} > 1$. Here, we do not have to deal with Hölder exponents.

Proposition 3.10. (P) *Let $\alpha_F \geq \frac{1}{2}$ and $\mathrm{p} \geq 2$. Assume that $F \in \mathrm{C}^{2\mathrm{p}+1}(I_{\bar{u}})$. Let $\widetilde{y}_h = (\widetilde{u}_h, \widetilde{\phi}_h, \widetilde{\mu}_h) \in \mathcal{Y}_{h,\#}$ be a solution of the discretization (3.5) satisfying $\|\bar{y} - \widetilde{y}_h\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}$. Then, there exists a constant $C > 0$ such that*

$$\|\bar{u}_h - \widetilde{u}_h\|_{\mathrm{L}^2} + \|\bar{\phi}_h - \widetilde{\phi}_h\|_{\mathrm{L}^2} + |\bar{\mu}_h - \widetilde{\mu}_h| \leq Ch^{\mathrm{p}+1}$$

and

$$\|\bar{u} - \widetilde{u}_h\|_{\mathrm{L}^2} + \|\bar{\phi} - \widetilde{\phi}_h\|_{\mathrm{L}^2} + |\bar{\mu} - \widetilde{\mu}_h| \leq Ch^{\mathrm{p}+1}.$$

Proof. The structure of the proof is identical to the previous one for the Dirichlet case. Since the details of convergence rates and quadrature errors differ, we still present the whole argument.

Let $f = (f_1, f_2, f_3) \in \mathrm{L}^2(\Omega) \times \mathrm{L}^2(\Omega) \times \mathbb{R} \subset \mathcal{Y}_{h,\#}^*$ represent a linear functional. We define the Lagrangian functionals $\mathcal{L}_{h,\#}, \widetilde{\mathcal{L}}_{h,\#} : \mathcal{Y}_{h,\#} \times \mathcal{Y}_{h,\#} \to \mathbb{R}$ by

$$\mathcal{L}_{h,\#}(y_h, z_h) = \langle f, y_h \rangle - \langle \mathcal{F}_{h,\#}(y_h), z_h \rangle,$$
$$\widetilde{\mathcal{L}}_{h,\#}(y_h, z_h) = \langle f, y_h \rangle - \langle \widetilde{\mathcal{F}}_{h,\#}(y_h), z_h \rangle.$$

Let $\bar{z} \in \mathcal{Y}_\#$ be the solution to the dual equation (2.55) and $\bar{z}_h = (\bar{z}_{u,h}, \bar{z}_{\phi,h}, \bar{z}_{\mu,h})$ the corresponding solution of the discrete dual equation (2.56) satisfying $\|\bar{z} - \bar{z}_h\|_{\mathcal{Y}} \leq Ch$. In Proposition 2.17 we showed that $(\bar{y}_h, \bar{z}_h) \in \mathcal{Y}_{h,\#} \times \mathcal{Y}_{h,\#}$ is a stationary point of $\mathcal{L}_{h,\#}$.

Since $D\widetilde{\mathcal{F}}_{h,\#}(\widetilde{y}_h) : \mathcal{Y}_{h,\#} \to \mathcal{Y}_{h,\#}^*$ is an isomorphism (see Lemma 3.3), there exists $\widetilde{z}_h \in \mathcal{Y}_{h,\#}$ such that

$$\langle D\widetilde{\mathcal{F}}_{h,\#}(\widetilde{y}_h) \cdot \eta_h, \widetilde{z}_h \rangle = \langle f, \eta_h \rangle \quad \forall \eta_h \in \mathcal{Y}_{h,\#}.$$

This is equivalent to $(\widetilde{y}_h, \widetilde{z}_h)$ being a saddle point of $\widetilde{\mathcal{L}}_{h,\#}$.

We recall that duality arguments as given in Section 2.5 and an analysis analogous to the one given in Lemmas 3.1 and 3.2 imply that $\bar{z}_{u,h}, \bar{z}_{\phi,h} \in \mathrm{L}^\infty(\Omega)$ and, for $2 < q < \frac{2d}{d-2}$,

$$\sum_{T \in \mathcal{T}_h} \|\bar{z}_{u,h}\|_{\mathrm{W}^{1,q}(T)}^q + \sum_{T \in \mathcal{T}_h} \|\bar{z}_{\phi,h}\|_{\mathrm{W}^{1,q}(T)}^q \leq C,$$

uniformly in h.

From $\alpha_F \geq \frac{1}{2}$ and Proposition 2.17 it follows that $\|\bar{u}_h - \widetilde{u}_h\|_{\mathrm{L}^2} \leq Ch^{\mathrm{p}+1}$. Hence, we get by an inverse inequality [17, Theorem 4.5.11] that

$$\|\bar{u}_h - \widetilde{u}_h\|_{\mathrm{L}^\infty} \leq Ch^{-\frac{d}{2}} \|\bar{u}_h - \widetilde{u}_h\|_{\mathrm{L}^2} \leq Ch^{\mathrm{p}+1-\frac{d}{2}}.$$

This implies $\widetilde{u}_h(\bar{\Omega}) \subset I_{\bar{u}}$ for sufficiently small h.

Step 1. In the present periodic case we can in fact show that $\|\widetilde{z}_h - \bar{z}_h\|_{\mathcal{Y}} \leq Ch$. With an inf-sup condition for $D\widetilde{\mathcal{F}}_{h,\#}(\widetilde{y}_h)$ we get

$$\begin{aligned}\widetilde{\kappa}_h \|\widetilde{z}_h - \bar{z}_h\|_{\mathcal{Y}} \leq & \sup_{\eta_h \in \mathcal{Y}_{h,\#}} \frac{\left\langle \left(D\mathcal{F}_{h,\#}(\bar{y}_h) - D\widetilde{\mathcal{F}}_{h,\#}(\bar{y}_h)\right) \cdot \eta_h, \bar{z}_h \right\rangle}{\|\eta_h\|_{\mathcal{Y}}} \\ & + \sup_{\eta_h \in \mathcal{Y}_{h,\#}} \frac{\left\langle \left(D\widetilde{\mathcal{F}}_{h,\#}(\bar{y}_h) - D\widetilde{\mathcal{F}}_{h,\#}(\widetilde{y}_h)\right) \cdot \eta_h, \bar{z}_h \right\rangle}{\|\eta_h\|_{\mathcal{Y}}}. \end{aligned} \tag{3.30}$$

Now, the second term on the right-hand side satisfies

$$\sup_{\eta_h \in \mathcal{Y}_{h,\#}} \frac{\left\langle \left(D\widetilde{\mathcal{F}}_{h,\#}(\bar{y}_h) - D\widetilde{\mathcal{F}}_{h,\#}(\widetilde{y}_h)\right) \cdot \eta_h, \bar{z}_h \right\rangle}{\|\eta_h\|_{\mathcal{Y}}} \leq Ch^{\mathrm{p}} \|\bar{z}_h\|_{\mathcal{Y}}$$

because $D\widetilde{\mathcal{F}}_{h,\#}$ is Lipschitz continous on the convex hull of $\{\bar{y}_h, \widetilde{y}_h\}$.

From Lemma 3.2 we know that $\bar{u}_h, \bar{\phi}_h \in \mathrm{W}^{1,\infty}(T)$ uniformly for all $T \in \mathcal{T}_h$. A calculation as in (3.14) then shows that $F''(\bar{u}_h)$ is uniformly Lipschitz continuous on all $T \in \mathcal{T}_h$. We can

then repeat Step 2 from the proof of Lemma 3.3 (using (A.28) in Proposition A.15 for the quadrature errors) to deduce that

$$\sup_{\eta_h \in \mathcal{Y}_{h,\#}} \frac{\langle \big(D\mathcal{F}_{h,\#}(\bar{y}_h) - D\widetilde{\mathcal{F}}_{h,\#}(\bar{y}_h)\big)\cdot \eta_h, \bar{z}_h\rangle}{\|\eta_h\|_{\mathcal{Y}}} \leq Ch\|\bar{z}_h\|_{\mathcal{Y}}.$$

Applying this to (3.30) we obtain

$$\|\widetilde{z}_h - \bar{z}_h\|_{\mathcal{Y}} \leq Ch.$$

Step 2. As in the previous proof we have

$$\langle f, \bar{y}_h - \widetilde{y}_h\rangle = \big(\mathcal{L}_{h,\#}(\bar{y}_h, \bar{z}_h) - \widetilde{\mathcal{L}}_{h,\#}(\bar{y}_h, \bar{z}_h)\big) + \big(\widetilde{\mathcal{L}}_{h,\#}(\bar{y}_h, \bar{z}_h) - \widetilde{\mathcal{L}}_{h,\#}(\widetilde{y}_h, \widetilde{z}_h)\big). \tag{3.31}$$

The usual Taylor expansion argument around the saddle point $(\widetilde{y}_h, \widetilde{z}_h)$ of $\widetilde{\mathcal{L}}_{h,\#}$ gives for the second term on the right-hand side

$$\begin{aligned}\Big|\widetilde{\mathcal{L}}_h(\bar{y}_h, \bar{z}_h) - \widetilde{\mathcal{L}}_h(\widetilde{y}_h, \widetilde{z}_h)\Big| &\leq C\left(\|\bar{y}_h - \widetilde{y}_h\|_{\mathcal{Y}}^2 + \|\bar{z}_h - \widetilde{z}_h\|_{\mathcal{Y}}\|\bar{y}_h - \widetilde{y}_h\|_{\mathcal{Y}}\right)\\ &\leq Ch^{\mathrm{p}+1}.\end{aligned} \tag{3.32}$$

For $\big|\mathcal{L}_h(\bar{y}_h, \bar{z}_h) - \widetilde{\mathcal{L}}_h(\bar{y}_h, \bar{z}_h)\big|$ we need to look at quadrature errors

$$\begin{aligned}\big|\mathcal{L}_h(\bar{y}_h, \bar{z}_h) - \widetilde{\mathcal{L}}_h(\bar{y}_h, \bar{z}_h)\big| \leq C\Big(&\big|e_h\big[F'(\bar{u}_h)\bar{z}_{u,h}\big]\big| + \big|e_h\big[\bar{\phi}_h\bar{u}_h\bar{z}_{u,h}\big]\big| + \big|e_h[\bar{u}_h\bar{z}_{u,h}]\big|\\ &+ \big|e_h\big[(\bar{u}_h^2 - \rho_n)\bar{z}_{\phi,h}\big]\big| + \big|e_h\big[\bar{u}_h^2\big]\big|\Big)\end{aligned} \tag{3.33}$$

and show that each of them is of order $\mathcal{O}(h^{\mathrm{p}+1})$.

For example, let us look at the term $\big|e_h\big[\bar{\phi}_h\bar{u}_h\bar{z}_{u,h}\big]\big|$. Using the L^∞-boundedness of $\bar{z}_{u,h}$ and the uniform boundedness of $\bar{u}_h$ and $\bar{\phi}_h$ in $\mathrm{W}^{\mathrm{p}-1,\infty}(T)$ we can show that, for all $2 < q < \frac{d}{d-2}$,

$$\begin{aligned}\|\bar{\phi}_h\bar{u}_h\bar{z}_{u,h}\|_{\mathrm{W}^{2\mathrm{p},q}(T)} \leq C\big(&\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)}\|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)} + \|\bar{z}_{u,h}\|_{\mathrm{W}^{\mathrm{p},2q}(T)}\|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)}\\ &+ \|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)}\|\bar{z}_{u,h}\|_{\mathrm{W}^{\mathrm{p},2q}(T)}\big).\end{aligned}$$

Summing over $T \in \mathcal{T}_h$ then yields

$$\begin{aligned}\left(\sum_{T\in\mathcal{T}_h}\|\bar{\phi}_h\bar{u}_h\bar{z}_{u,h}\|_{\mathrm{W}^{2\mathrm{p},q}(T)}^q\right)^{1/q} &\leq C + C\left(1 + \sum_{T\in\mathcal{T}_h}\|\bar{z}_{u,h}\|_{\mathrm{W}^{\mathrm{p},2q}(T)}^{2q}\right)^{1/2q}\\ &\leq C + C(1 + h^{1-\mathrm{p}})\|\bar{z}_{u,h}\|_{\mathrm{W}^{1,2q}(\Omega)},\end{aligned} \tag{3.34}$$

where we have used a standard inverse estimate [17, Theorem 4.5.11] to get from the $\mathrm{W}^{\mathrm{p},2q}$-norm to the $\mathrm{W}^{1,2q}$-norm. The constants C depend on $\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)}$ and $\|\bar{\phi}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)}$. An application of Proposition A.14 leads to

$$\big|e_h\big[\bar{\phi}_h\bar{u}_h\bar{z}_{u,h}\big]\big| \leq Ch^{\mathrm{p}+1}.$$

Similarly we obtain

$$\big|e_h\big[(\bar{u}_h^2-\rho_n)\bar{z}_{\phi,h}\big]\big|+\big|e_h\big[\bar{u}_h^2\big]\big|+\big|e_h[\bar{u}_h\bar{z}_{u,h}]\big|\le Ch^{\mathrm{p}+1}.$$

Lastly, we have

$$\big\|F'(\bar{u}_h)\bar{z}_{u,h}\big\|_{\mathrm{W}^{2\mathrm{p},q}(T)}\le C\|\bar{u}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)}\|\bar{z}_h\|_{\mathrm{W}^{\mathrm{p},2q}(T)},$$

where the constant depends on the derivatives of F up to order $2\mathrm{p}+1$. Hence,

$$\begin{aligned}\left(\sum_{T\in\mathcal{T}_h}\|F'(\bar{u}_h)\bar{z}_{u,h}\|^q_{\mathrm{W}^{2\mathrm{p},q}(T)}\right)^{1/q}&\le C\left(1+\sum_{T\in\mathcal{T}_h}\|\bar{z}_{u,h}\|^{2q}_{\mathrm{W}^{\mathrm{p},2q}(T)}\right)^{1/2q}\\&\le C+C(1+h^{1-\mathrm{p}})\|\bar{z}_{u,h}\|_{\mathrm{W}^{1,2q}(\Omega)},\end{aligned}\tag{3.35}$$

and Proposition A.14 implies that

$$\big|e_h\big[F'(\bar{u}_h)\bar{z}_{u,h}\big]\big|\le Ch^{\mathrm{p}+1}.$$

Summarizing, we have established the bound

$$\big|\mathcal{L}_h(\bar{y}_h,\bar{z}_h)-\widetilde{\mathcal{L}}_h(\bar{y}_h,\bar{z}_h)\big|\le Ch^{\mathrm{p}+1},$$

and therefore, with (3.32) and (3.31), also

$$|\langle f,\bar{y}_h-\widetilde{y}_h\rangle|\le Ch^{\mathrm{p}+1}.$$

Choosing

$$f=\frac{(\bar{u}_h-\widetilde{u}_h,\bar{\phi}_h-\widetilde{\phi}_h,\bar{\mu}_h-\widetilde{\mu}_h)}{\big(\|\bar{u}_h-\widetilde{u}_h\|^2_{\mathrm{L}^2}+\|\bar{\phi}_h-\widetilde{\phi}_h\|^2_{\mathrm{L}^2}+|\bar{\mu}_h-\widetilde{\mu}_h|^2\big)^{1/2}}\in\mathrm{S}_{h,\#}\times\mathrm{S}_{h,\#}\times\mathbb{R}$$

completes the proof. □

Finally, we show that in the periodic case, $\widetilde{\mu}_h$ converges at the optimal rate.

Proposition 3.11. (P) *Let $\alpha_F\ge\frac{1}{2}$ and $\mathrm{p}\ge 2$. Assume that $F\in\mathrm{C}^{2\mathrm{p}+1}(I_{\bar{u}})$. Let $\widetilde{y}_h=(\widetilde{u}_h,\widetilde{\phi}_h,\widetilde{\mu}_h)\in\mathcal{Y}_{h,\#}$ be a solution of the discretization* (3.5) *satisfying $\|\bar{y}-\widetilde{y}_h\|_{\mathcal{Y}}\le Ch^{\mathrm{p}}$. Then, there exists a constant $C>0$ such that*

$$|\bar{\mu}-\widetilde{\mu}_h|\le Ch^{2\mathrm{p}}.$$

Proof. Yet again the idea of the proof is the same as previously. We therefore only highlight the necessary modifications. This time we choose $f=(0,0,1)\in\mathcal{Y}^*_{h,\#}$ in the definitions of $\mathcal{L}_{h,\#}$ and $\widetilde{\mathcal{L}}_{h,\#}$. As we saw in the proof of Proposition 2.22, the dual solutions $\bar{z}_u$, $\bar{z}_\phi$ then belong to $\mathrm{H}^{\mathrm{p}+1}(\Omega)$. Moreover, the discrete dual solution $\bar{z}_h=(\bar{z}_{u,h},\bar{z}_{\phi,h},\bar{z}_{\mu,h})\in\mathcal{Y}_{h,\#}$ satisfies

$\|\bar{z}-\bar{z}_h\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}$. From this we derive the uniform boundedness of $\bar{z}_{u,h}$ and $\bar{z}_{\phi,h}$ in $\mathrm{W}^{\mathrm{p}-1,\infty}(T)$ for all $T \in \mathcal{T}_h$ and h and

$$\sum_{T\in\mathcal{T}_h} \|\bar{z}_{u,h}\|^q_{\mathrm{W}^{\mathrm{p},q}(T)} + \sum_{T\in\mathcal{T}_h} \|\bar{z}_{\phi,h}\|^q_{\mathrm{W}^{\mathrm{p},q}(T)} \leq C, \tag{3.36}$$

uniformly in h. These observations can be used to show (with Proposition A.11 for the quadrature errors)

$$\sup_{\eta_h\in\mathcal{Y}_{h,\#}} \frac{\left\langle \left(D\mathcal{F}_{h,\#}(\bar{y}_h) - D\widetilde{\mathcal{F}}_{h,\#}(\bar{y}_h)\right)\cdot\eta_h, \bar{z}_h\right\rangle}{\|\eta_h\|_{\mathcal{Y}}} \leq Ch^{\mathrm{p}}\|\bar{z}_h\|_{\mathcal{Y}},$$

from which, with (3.30), we deduce that

$$\|\widetilde{z}_h - \bar{z}_h\|_{\mathcal{Y}} \leq Ch^{\mathrm{p}}.$$

This increased order of $\|\bar{z}_h - \widetilde{z}_h\|_{\mathcal{Y}}$ gives, instead of (3.31),

$$\left|\widetilde{\mathcal{L}}_h(\bar{y}_h, \bar{z}_h) - \widetilde{\mathcal{L}}_h(\widetilde{y}_h, \widetilde{z}_h)\right| \leq Ch^{2\mathrm{p}}.$$

The quadrature error analysis necessary to bound $\left|\mathcal{L}_h(\bar{y}_h, \bar{z}_h) - \widetilde{\mathcal{L}}_h(\bar{y}_h, \bar{z}_h)\right|$ (see (3.33)) needs little modification. The application of inverse inequalities in (3.34) and (3.35) is, however, unnecessary due to (3.36). Thus, we obtain

$$\left|\mathcal{L}_h(\bar{y}_h, \bar{z}_h) - \widetilde{\mathcal{L}}_h(\bar{y}_h, \bar{z}_h)\right| \leq Ch^{2\mathrm{p}}$$

and therefore with (3.31)

$$|\bar{\mu}_h - \widetilde{\mu}_h| = |\langle f, \bar{y}_h - \widetilde{y}_h\rangle| \leq Ch^{2\mathrm{p}}.$$

Since $|\bar{\mu} - \bar{\mu}_h| \leq Ch^{2\mathrm{p}}$ this concludes the proof. □

Remark 3.12. In Remark 2.13, we showed that under certain integrability conditions on V, the convergence results in the Galerkin case still hold true if the energy functional E is extended by a potential term $\int_\Omega Vu^2 \,\mathrm{d}x$. We now analyze which differentiability conditions on V are necessary such that the above results for the discretization with numerical integration still hold if $\int_\Omega Vu^2 \,\mathrm{d}x$ is part of the energy.

Looking at Lemma 3.3 and its proof, we obtain by Proposition A.15 that $V \in \mathrm{W}^{1,\infty}(\Omega)$ is sufficient for $\widetilde{\mathcal{F}}$ to be differentiable and for $D\widetilde{\mathcal{F}}$ to be an isomorphism in a neighbourhood of $\bar{y}_h$ ($e_h[Vvw] \leq Ch\,\|v\|_{\mathrm{H}^1}\,\|w\|_{\mathrm{H}^1}$ for all $v, w \in \mathrm{S}_h$). Lemma 3.4 can be extended if $V \in \mathrm{W}^{\mathrm{p},q}(\Omega)$ for some $q > \frac{d}{\mathrm{p}}$ since then $e_h[V\bar{u}_h v] \leq Ch^{\mathrm{p}}\,\|v\|_{\mathrm{H}^1}$ by Theorem A.11. Hence $V \in \mathrm{W}^{1,\infty}(\Omega) \cap \mathrm{W}^{\mathrm{p},q}(\Omega)$ is sufficient to show convergence of $\bar{u}_h$ in this case.

To sustain the optimal convergence order 2p for the energy, V has to belong to $\mathrm{W}^{2\mathrm{p},q}(\Omega)$ for some $q > \frac{d}{2p}$. This follows by a direct application of Proposition A.14. □

3.4 Fourier Discretization with Interpolation

In this section we briefly discuss the effects of interpolation on the Fourier discretization (2.53). A comprehensive study for the Thomas–Fermi–von Weizsäcker functional has been carried out in [20–22]. We will therefore only highlight the ideas needed to prove convergence in our framework.

To obtain a practical Fourier discretization, interpolation is used to approximate the integrals in the Galerkin discretization (2.53) that can not be easily evaluated. This results in

$$\begin{aligned} \lambda(\nabla\widetilde{u}_N, \nabla v) + \big(\mathcal{I}_N(F'(\widetilde{u}_N)), v\big) + 2\big(\mathcal{I}_N(\widetilde{\phi}_N\widetilde{u}_N), v\big) + \mu_N(\widetilde{u}_N, v) = 0 \quad \forall v \in \mathrm{S}_N, \\ \frac{1}{4\pi}(\nabla\widetilde{\phi}_N, \nabla\psi) - \big(\mathcal{I}_N(\widetilde{u}_N^2 - \rho_n), \psi\big) = 0 \quad \forall \psi \in \mathrm{S}_{N,0}, \\ \frac{\nu}{2}\left(\int_\Omega \widetilde{u}_N^2 \,\mathrm{d}x - n_{\mathrm{el}}\right) = 0 \quad \forall \nu \in \mathbb{R}, \end{aligned} \tag{3.37}$$

or $\widetilde{\mathcal{F}}_N(\widetilde{y}_N) = 0 \in \mathcal{Y}_N^*$, where $\widetilde{y}_N = (\widetilde{u}_N, \widetilde{\phi}_N, \widetilde{\mu}_N) \in \mathcal{Y}_N$. We will from now on assume that $\mathrm{p} \geq 3$ and $F \in \mathrm{C}^{\mathrm{p}+1}(I_{\bar{u}})$ for an open interval $I_{\bar{u}} \supset \bar{u}(\bar{\Omega})$.

Theorem 3.13. *Let $\bar{u} \in \mathrm{A}_{u,\#}$ be a uniform minimizer of (2.40) and $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_\#$ the corresponding solution to (2.44) with $\bar{u}, \bar{\phi} \in \mathrm{H}^{\mathrm{p}+1}_\#(\Omega)$. Moreover, assume that $F \in \mathrm{C}^{\mathrm{p}+1}(I_{\bar{u}})$. Then, there exist $N_0 \in \mathbb{N}$ and $C > 0$ such that for all $N \geq N_0$ there exists a unique solution $\widetilde{y}_N \in \mathcal{Y}_N$ to (3.37) such that*

$$\big\|\bar{y} - \widetilde{y}_N\big\|_{\mathcal{Y}} \leq CN^{-\mathrm{p}}.$$

Proof. Showing existence and convergence of a solution $\widetilde{y}_N$ of this system follows the same lines as in the finite element case with quadrature. Hence, we will only outline how to deal with the errors introduced by interpolation.

From $\bar{u} \in \mathrm{H}^{\mathrm{p}+1}_\#(\Omega)$ and $\|\bar{u} - \bar{u}_N\|_{\mathrm{H}^1} \leq CN^{-\mathrm{p}}$ as seen in Theorem 2.15 it follows with an inverse inequality [24, (5.8.8)] or [25, Proposition 1.1]) that

$$\|\bar{u} - \bar{u}_N\|_{\mathrm{H}^2} \leq \|\bar{u} - \mathcal{I}_N\bar{u}\|_{\mathrm{H}^2} + \|\mathcal{I}_N\bar{u} - \bar{u}_N\|_{\mathrm{H}^2} \leq CN^{-(\mathrm{p}-1)}\|\bar{u}\|_{\mathrm{H}^{\mathrm{p}+1}} + CN^{-\mathrm{p}+1}.$$

This implies $\|\bar{u} - \bar{u}_N\|_{\mathrm{L}^\infty} \leq CN^{-\mathrm{p}+1}$ and hence $\bar{u}_N(\bar{\Omega}) \subset I_{\bar{u}}$ for sufficiently large N. We note that the optimality system (2.53) implies (see [21]) that $\bar{u}_N$ and $\bar{\phi}_N$ are weak solutions to the equations

$$\begin{aligned} -\lambda\Delta\bar{u}_N + \Pi_N\big(F'(\bar{u}_N) + 2\bar{\phi}_N\bar{u}_N + \bar{\mu}_N\bar{u}_N\big) = 0, \\ -\tfrac{1}{4\pi}\Delta\bar{\phi}_N - \Pi_N\big(\bar{u}_N^2 - \rho_n\big) = 0, \end{aligned}$$

and hence uniformly bounded in $\mathrm{H}^2_\#(\Omega)$ by elliptic regularity. Again, this argument can be iterated to deduce that $\bar{u}_N$, $\bar{\phi}_N$ are uniformly bounded in $\mathrm{H}^{\mathrm{p}}_\#(\Omega)$. This also implies that $\bar{u}_N, \bar{\phi}_N$ are uniformly bounded in $\mathrm{W}^{\mathrm{p}-2,\infty}(\Omega)$ as well as $\mathrm{W}^{\mathrm{p}-1,q}(\Omega)$ for all $2 \leq q < \frac{2d}{d-2}$.

We now turn to the interpolation error estimates necessary for the convergence proof. In order to show that $\big\|\widetilde{\mathcal{F}}_N(\bar{y}_N)\big\|_{\mathcal{Y}_N^*} \le CN^{-\mathrm{p}}$, we have to assure that

$$\begin{aligned} \big|\big(\mathcal{I}_N F'(\bar{u}_N) - F'(\bar{u}_N), v\big)\big| &\le CN^{-\mathrm{p}}, \\ \big|\big(\mathcal{I}_N(\bar{\phi}_N \bar{u}_N) - \bar{\phi}_N \bar{u}_N, v\big)\big| &\le CN^{-\mathrm{p}}, \\ \big|\big(\mathcal{I}_N(\bar{u}_N^2 - \rho_n) - (\bar{u}_N^2 - \rho_n), v\big)\big| &\le CN^{-\mathrm{p}} \end{aligned} \tag{3.38}$$

for all $v \in \mathrm{S}_N$. Using the standard approximation result for the interpolation operator $\mathcal{I}_N$ we get, for example,

$$\begin{aligned} \big|\big(\mathcal{I}_N F'(\bar{u}_N) - F'(\bar{u}_N), v\big)\big| &\le C\big\|\mathcal{I}_N F'(\bar{u}_N) - F'(\bar{u}_N)\big\|_{\mathrm{L}^2}\|v\|_{\mathrm{L}^2} \\ &\le CN^{-\mathrm{p}}\big\|F'(\bar{u}_N)\big\|_{\mathrm{H}^{\mathrm{p}}}\|v\|_{\mathrm{L}^2}. \end{aligned}$$

With (3.17) we can show that

$$\big\|F'(\bar{u}_N)\big\|_{\mathrm{H}^{\mathrm{p}}} \le C\|\bar{u}_N\|^2_{\mathrm{W}^{\mathrm{p}-1,4}} + C\|\bar{u}_N\|_{\mathrm{H}^{\mathrm{p}}} \le C$$

uniformly in N. The constant C depends on the derivatives of F up to order $\mathrm{p}+1$ on $I_{\bar{u}}$, as well as $\|\bar{u}_N\|_{\mathrm{W}^{\mathrm{p}-2,\infty}}$. As we saw above $\|\bar{u}_N\|_{\mathrm{H}^{\mathrm{p}}}$ is bounded uniformly in N. The other errors in (3.38) are treated similarly.

For invertibility of $D\widetilde{\mathcal{F}}_N(\bar{y}_N)$ for large N it suffices to show that $\big\|D\widetilde{\mathcal{F}}_N(\bar{y}_N) - D\mathcal{F}_N(\bar{y}_N)\big\| \to 0$ as $N \to \infty$. To prove this, we need to show among others that

$$\big\|\mathcal{I}_N\big(F''(\bar{u}_N)v_N\big) - F''(\bar{u}_N)v_N\big\|_{\mathrm{L}^2} \le CN^{-1}\|v_N\|_{\mathrm{H}^1}$$

for all $v_N \in \mathrm{S}_N$. A quick calculation leads to

$$\begin{aligned} \big\|F''(\bar{u}_N)v_N\big\|_{\mathrm{H}^2} &\le C\|\bar{u}_N\|_{\mathrm{H}^2}\|v_N\|_{\mathrm{L}^\infty} + C\|\bar{u}_N\|_{\mathrm{H}^1}\|v_N\|_{\mathrm{H}^1} + C\|\bar{u}_N\|_{\mathrm{L}^\infty}\|v_N\|_{\mathrm{H}^2} \\ &\le CN\|v_N\|_{\mathrm{H}^1}, \end{aligned}$$

where $C = C(F, \|\bar{u}_N\|_{\mathrm{H}^{\mathrm{p}}})$. Here we have used $\|v_N\|_{\mathrm{L}^\infty} \le C\|v_N\|_{\mathrm{H}^2}$ and the inverse inequality $\|v_N\|_{\mathrm{H}^2} \le CN\|v_N\|_{\mathrm{H}^1}$ for all $v_N \in \mathrm{S}_N$. Then, we have

$$\big\|\mathcal{I}_N\big(F''(\bar{u}_N)v_N\big) - F''(\bar{u}_N)v_N\big\|_{\mathrm{L}^2} \le CN^{-2}\big\|F''(\bar{u}_N)v_N\big\|_{\mathrm{H}^2} \le CN^{-1}\|v_N\|_{\mathrm{H}^1}.$$

Applying similar ideas to the other interpolation errors appearing in $\big\|D\widetilde{\mathcal{F}}_N(\bar{y}_N) - D\mathcal{F}_N(\bar{y}_N)\big\|$ we deduce that

$$\big\|D\widetilde{\mathcal{F}}_N(\bar{y}_N) - D\mathcal{F}_N(\bar{y}_N)\big\| \le CN^{-1},$$

Since $D\mathcal{F}_N(\bar{y}_N)$ is an isomorphism from $\mathcal{Y}_N$ to $\mathcal{Y}_N^*$ this means that $D\widetilde{\mathcal{F}}_N(\bar{y}_N)$ is an isomorphism for sufficiently large N. By continuity of $D\widetilde{\mathcal{F}}_N$ this also holds for $D\widetilde{\mathcal{F}}_N(y_N)$ for y_N in a neighbourhood of $\bar{y}_N$. The proof of convergence can be completed along the lines of Theorem 2.10. ∎

With methods similar to the ones used in the finite element case we can show that $|E_\#(\bar{u}) - E_\#(\bar{u}_N)| \leq C\|\bar{y} - \bar{y}_N\|_{\mathcal{Y}}^2$ and even $|E_\#(\bar{u}) - E_\#(\widetilde{u}_N)| \leq C\|\bar{y} - \widetilde{y}_N\|_{\mathcal{Y}}^2$. However, obtaining the doubled convergence order if the energy is computed using numerical integration is not as straightforward as in the finite element case. A natural way of defining the energy $\widetilde{E}_N : \mathrm{S}_N \to \mathbb{R}$ with interpolation is

$$\widetilde{E}_N(u_N) = \frac{\lambda}{2}\int_\Omega |\nabla u_N|^2 \,\mathrm{d}x + \int_\Omega \mathcal{I}_N\big(F(u_N)\big)\,\mathrm{d}x + \frac{1}{2}\int_\Omega \mathcal{I}_N\big(u_N^2 - \rho_n\big)\phi_N \,\mathrm{d}x$$

where $\phi_N \in \mathrm{S}_N$ satisfies

$$(\nabla\phi_N, \nabla v_N) = 4\pi\big(\mathcal{I}_N(u_N^2 - \rho_n), v_N\big) \quad \forall v_N \in \mathrm{S}_N.$$

(Note that the Fourier discretization (3.37) does not represent the optimality system of a discrete minimization problem with $\widetilde{E}_N$.) In the finite element case we showed that the quadrature errors in $E_h(\bar{u}_h) - \widetilde{E}_h(\bar{u}_h)$ are of order $\mathcal{O}(h^{2\mathrm{p}})$, for which it is sufficient that the quadrature rule is of order $2\mathrm{p} - 1$. However, in the Fourier case we generally only get, for example

$$\left|\int_\Omega \big(\mathcal{I}_N(F(\bar{u}_N)) - F(\bar{u}_N)\big)\,\mathrm{d}x\right| \leq CN^{-\mathrm{p}}. \tag{3.39}$$

At this point we do not proceed further but refer to [22], where a rigorous study of convergence rates for the Thomas–Fermi–von Weizsäcker functional is provided. There, the authors work on two different grids: one for the definition of the discretization space and a finer one for the computation of integrals.

3.5 Numerical Examples

Despite the nonconvexity of the optimization problem (2.7) and its discretizations, the numerical solution turns out to be straightforward. To solve the nonlinear system (3.1) we can directly apply Newton's method. This choice is justified because of the availability of good initial guesses for u and ϕ.

In the case of homogeneous boundary conditions (an isolated cluster of atoms), for example, we expect the electron density near the nuclei to be close to the one of an isolated atom. We therefore solve the spherically symmetric TFDW problem for a single atom first. The initial value for u is then the square root of the sum of spherically symmetric single atom electron densities centered around the nuclei. An initial guess for ϕ is successively obtained by solving $-\Delta\phi = 4\pi(u^2 - \rho_n)$ subject to the right boundary conditions. In all computations, ρ_n is the sum of Gauss functions with variance σ_0 centered at nucleus positions. Given a sufficiently fine mesh, we observe that the Newton iteration enters the regime of local quadratic convergence immediately. We deduce that there is no need to apply globalization strategies.

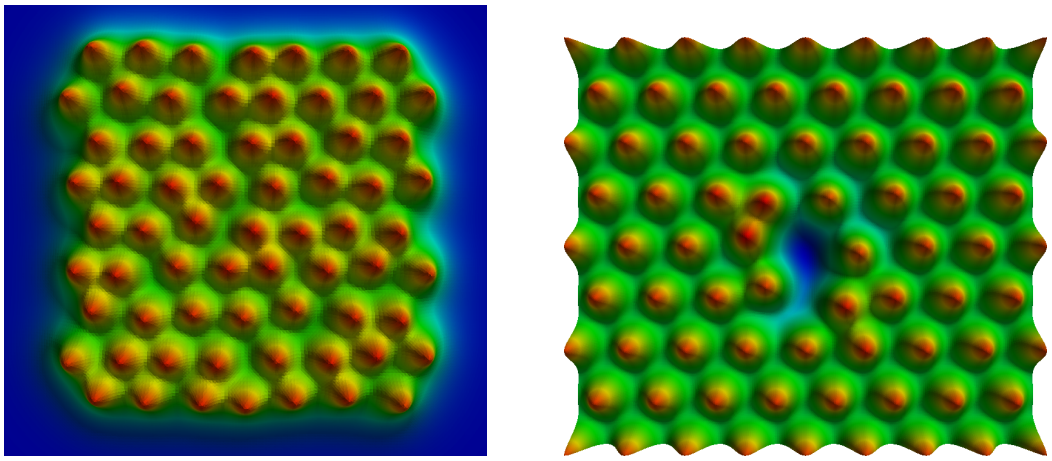

Figure 3.1: Outline of the solution $\bar{u}_h$ for two numerical examples. Left: a cluster of 67 atoms in two dimensions is simulated. Note that only a close-up of the interesting features is shown in the plot. The computational domain Ω is larger than this. Right: a vacancy in a two dimensional lattice with slightly distorted atoms around the defect. The boundary conditions were obtained from a perfect lattice.

When simulating isolated defects in an infinite lattice, the boundary conditions for the electron density and the electrostatic potential are taken from a perfect crystal. These perfect crystal solutions can be obtained by solving the periodic problem on a unit cell of the lattice. Starting values for the Newton iteration are also given by the perfect crystal solutions subject to straightforward corrections for defects like vacancies or impurities

We now briefly report on some numerical results. In particular, the examples shown include finite element simulations of a cluster and a defect in 2D, and a Fourier based simulation of a periodic cell problem in 3D.

A Cluster of 67 Atoms in 2D. We have implemented a two-dimensional version of the discretization (3.1) described above using piecewise linear (p = 1), respectively, cubic (p = 3) Lagrange finite elements. In principle, the two-dimensional functional is obtained by replacing 8π with 4π in the definition of Φ (2.5) and the function F with $F(u) = \frac{\pi}{2}u^4 - \frac{4\sqrt{2}}{3\pi^{5/2}}|u|^3$, see [63]. As we saw above, the Hölder coefficient of F'' can influence some of the convergence rates. Since this two-dimensional version of F is $\mathrm{C}^3(\mathbb{R})$, and therefore $\alpha_F = 1$, we still use the three-dimensional version of F given in (2.6).

Results of a computation involving a cluster of 67 atoms in two dimensions are documented in Figure 3.2. The boundary conditions in this case were homogeneous and the configuration was obtained by slightly perturbing a section of a hexagonal lattice. The data used in this computation were $\Omega = (-30, 30)^2$, $\lambda = 3.2$, $\sigma_0 = 0.4$ and $Z_i = 6$ for all $i = 1, \ldots, 67$. Relative errors of E, μ, and DV are plotted against the mesh size h for linear, respectively, cubic finite

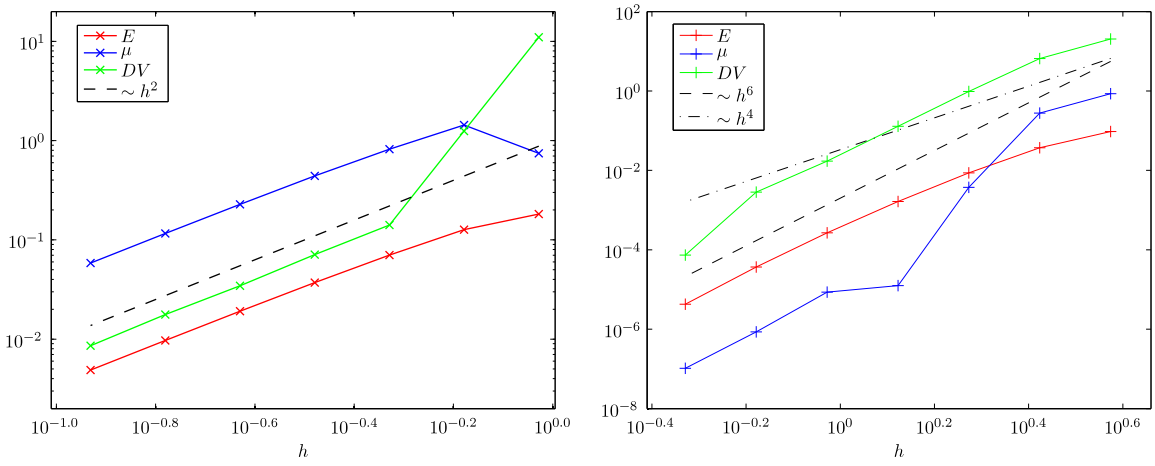

Figure 3.2: Convergence rates of calculations for a cluster of 67 atoms in two dimensions: relative errors of E, μ, and DV are plotted along with lines indicating the relevant orders of h. Left: linear finite elements ($\mathrm{p} = 1$), right: cubic finite elements ($\mathrm{p} = 3$). The reference values were obtained from a computation with $\mathrm{p} = 3$ on a finer mesh ($h = 0.331 = 10^{-0.479}$).

elements on a series of uniform grids. We observe that both the energy E and the Lagrange multiplier μ converge with order 2p if p is the order of the finite element space.

These observed convergence rates exceed what we proved in Section 3.3. Because of the homogeneous boundary conditions a mirroring argument as given in [21] can be used to show $\bar{u}, \bar{\phi} \in \mathrm{H}^3(\Omega)$. However, the plot for the case $\mathrm{p} = 3$ suggest even higher regularity. A possible explanation for this is that the domain Ω is rather large, giving the solutions $\bar{u}$, $\bar{\phi}$ space to decay such that singularities in the corners of the domain do not feature too prominently. We also point out that the convergence rate for the derivatives DV of the implied many-body potential are better than the expected $\mathrm{p} + 1$ (from the L^2-convergence rate of $\bar{\phi}_h$).

A Defect in 2D. As a second example we consider a vacancy, that is a missing atom, in two dimensions. The configuration was obtained by cutting out a section of an infinite hexagonal lattice and removing the atom in the centre. We also slightly displaced several atoms near the vacancy. The data used in this computation were $\Omega = (-13.84, 13.84) \times (-11.98, 11.98)$, $\lambda = 3.2$, $\sigma_0 = 0.25$ and $Z_i = 6$ for all i. The boundary conditions u_{ex}, ϕ_{ex} were obtained by solving the periodic cell problem (with a Fourier discretization) and interpolating the obtained cell solutions into the respective finite element spaces.

Results are documented in Figure 3.3. For $\mathrm{p} = 1$ we obtain the expected behaviour: energy, Lagrange multiplier, and forces converge at the rate $\mathcal{O}(h^2)$. In this case, we are not able to theoretically obtain a higher regularity for $\bar{u}$ and $\bar{\phi}$ than H^2. Nevertheless, the $\mathrm{p} = 3$ results do indicate higher smoothness. The energy and the Lagrange multiplier show

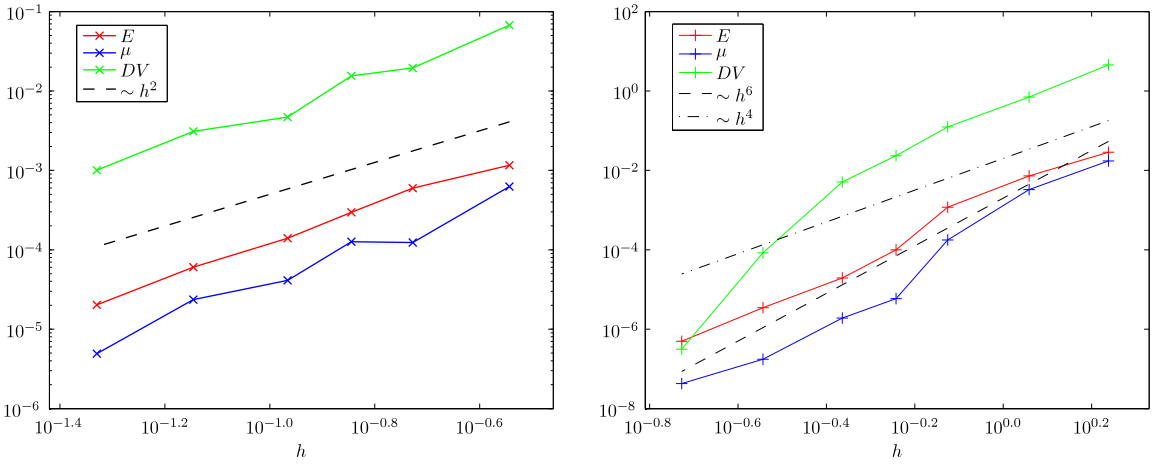

Figure 3.3: Convergence rates of calculations for a vacancy in two dimensions: relative errors of E, μ, and DV are plotted along with lines indicating the relevant orders of h. Left: linear finite elements, right: cubic finite elements. The reference values were obtained from a computation with p = 3 on a finer mesh ($h = 0.143 = 10^{-0.845}$).

convergence order $\mathcal{O}(h^4)$. A doubling of the convergence orders, as we saw in the case of homogeneous boundary conditions, is prevented by the interpolation error in obtaining the boundary data.

An Face-Centered Cubic Unit Cell in 3D. Finally, we consider the unit cell of a face-centered cubic crystal in three dimensions. The computational domain is $\Omega = (0,1)^3$. There are 14 atoms in the cell: 8 of them in the corners and 6 in the midpoints of the faces of the cube, see Figure 3.4. The data were $\lambda = 1/10$, $\sigma_0 = 0.1$, $Z = 5$. As expected, the solution $\bar{u}$ turns out to be strictly positive on $\bar{\Omega}$. Therefore F belongs to $\mathrm{C}^\infty(I_{\bar{u}})$ and we get $\bar{u}$, $\bar{\phi} \in \mathrm{C}^\infty(\Omega)$. This is confirmed by the convergence rates shown in Figure 3.4, where we plot $|\widetilde{E}_N(\widetilde{u}_N) - E(\bar{u})|$ and $|\widetilde{\mu}_N - \bar{\mu}|$ as functions of the parameter N. Both energy and Lagrange multiplier converge exponentially, that is, faster than any inverse power of N.

We conclude that the convergence rates that were proved in Sections 3.3 and 3.4 can indeed be observed experimentally. We again stress that in the finite element case a reference quadrature rule that is exact for all polynomials of degree $2\mathrm{p}-1$ is sufficient to get the same convergence rates as in the Galerkin case.

Finite element simulations in three space-dimensions are expected to give similar results as we have seen in two dimensions. However, detailed convergence studies using uniform meshes in 3D are computationally expensive, especially for larger numbers of atoms and therefore larger domains Ω. In particular, the efficient solution of the linear systems arising in Newton's method would require more technology in the form of appropriate preconditioners.

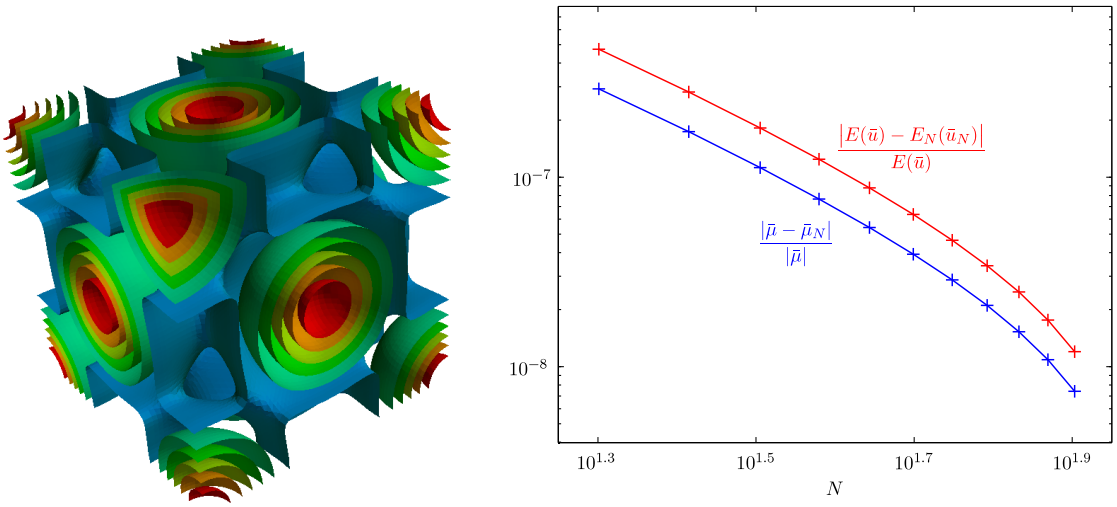

Figure 3.4: Square root density u and convergence rates for the unit cell of a face-centered cubic lattice. Since u is positive and F is therefore effectively smooth, the convergence rates on the right-hand side show the expected exponential behaviour.

As outlined in the Introduction, it is unfeasible to simulate physical phenomena at a macroscopic scale using atomistic models of interactions. This is especially true for density functional based models since the involved fields (e.g. the electronic density and the electrostatic potential) have to be resolved on subatomic meshes. It is therefore important to develop consistent multiscale models, like the Quasicontinuum Method, for atomistic interactions involving fields. In the next chapter we propose and analyze Quasicontinuum-like methods for a very basic field-based interaction potential in one dimension.

Chapter 4

Quasicontinuum Coupling for a Field-Based Interaction Potential

In the present chapter we formulate and analyze one-dimensional QC methods for an atomistic interaction that is mediated by a field.

The chapter is structured as follows. In Section 4.1 we give a literature review, motivate the atomistic model, and introduce the necessary notation. In Section 4.2 we formulate the model in a more precise mathematical way and derive a "weak formulation" for the resulting forces on the particles. Section 4.3 is devoted to the analysis of the model in a bounded domain when the fields are subjected to Dirichlet boundary conditions. The respective continuum model is derived and analyzed in Section 4.4 using the Cauchy–Born approximation. Finally, in Section 4.5 we propose different possibilities for constructing QC methods that are based on exchange of boundary conditions and prove convergence. The chapter closes with an outlook on possible extensions and open problems in Section 4.6.

4.1 Introduction

4.1.1 Literature Review

Some applications of the QC method can be found in [90, 109, 114, 115]. Phenomena investigated include defects, fracture, grain boundaries, and nano-indentation.

As mentioned in the Introduction, the most direct energy-based way of QC coupling leads to inconsistencies in the form of ghost forces. Naturally, a lot of work has therefore gone into the design of methods that do not exhibit these unphysical forces. Most of these approaches are based on a more careful treatment of interactions between atoms in the atomistic and the continuum part. For example, in [110] the quasi-nonlocal QC method was suggested. There, a layer of so-called quasi-nonlocal atoms is introduced between the atomistic and the continuum region. These quasi-nonlocal atoms interact normally with neighbours in the atomistic region, whereas interactions with atoms in the continuum region are replaced with

virtual atoms whose positions are obtained by extrapolating nearest neighbour positions. A conceptually similar but more general philosophy based on reconstruction schemes for atomic environments is followed in [50].

In [79] the quasi-nonlocal QC method was extended to arbitrary finite-range interactions in one space dimension. A similar, bond-based approach was suggested in 1D and 2D in [107]. A method directly based on the coarse-graining idea is presented in [85].

Although the QC method was originally developed in the 1990s, attempts at its analysis have started only recently. Early results addressed the coarse-graining step in 1D [83, 98] and 2D [84]. In [98] *a priori* and *a posteriori* error bounds for the resulting Galerkin approximation are proved. An analysis of the decay of ghost force induced errors away from the interface is provided in [43].

To obtain actual error bounds for QC-like methods, it is in general not sufficient to show the absence of ghost forces. Following a classical paradigm of numerical analysis, many rigorous approaches have focused on the issues of consistency and stability of a QC method. The issue of stability was investigated for the one-dimensional standard energy-based QC method and the quasi-nonlocal QC methods in [45]. The authors show that besides its inconsistency the standard energy-based QC method also has unsatisfactory stability properties compared with the original atomistic model and the quasi-nonlocal QC method. Rigorous error analysis for the quasi-nonlocal QC method was performed in one space dimension [44, 79, 93, 96]. The quasi-nonlocal QC method has excellent consistency and stability properties and convergence can be obtained, see [79, 96].

Methods based on summation rules instead of the Cauchy–Born approximation to reduce the complexity were analyzed in [87]. Theoretical results in connection with force-based QC methods can be found in [42, 46–48]. The standard force-based QC method has excellent consistency properties. However, the analysis of stability is more involved. Linearizations of the involved operators are nonnormal and generally not positive definite. The choice of topology turns out to be crucial for obtaining stability [46]. Convergence of the force-based QC method in 1D is proved in [47].

A way of coupling a density functional based atomistic model with a semi-empirical simulation was suggested in [36]. The authors independently use a DFT model in a subdomain and an embedded atom potential (EAM) in the remainder of the domain. The actual coupling is achieved by introducing an interaction energy, which involves a phenomenological electron density in the EAM region as input. These ideas have also been combined with a standard QC method resulting in a model with a quantum mechanical, a classical atomistic and a continuum region [86]. A very similar approach is given in [101, 125]. There, phenomenological electron densities in a patch region are used as boundary conditions for the density functional simulation. The Cauchy–Born approximation of OFDFT provides the continuum model.

Some rigorous mathematical results concerning the continuum and thermodynamic limits of different atomistic models are provided in [14, 15, 26]. In [14] the authors rigorously derive continuum models from pair-potentials and Thomas–Fermi type models. In [15] also a limiting process based on Γ-convergence is analyzed. In [12,13] the authors analyze models that couple an atomistic nearest neighbour and a continuum energy in one space dimension. The domain is divided into two regions and there is no underlying QC mesh. The message of the articles is that the natural way of coupling the two models leads to failure in the sense that if fracture arises, it does so in the continuum part rather than the atomistic part. The authors then propose a modified coupling that leads to the correct behaviour.

In [49] the authors derive the continuum limits for the Thomas–Fermi–von Weizsäcker and the Kohn–Sham functionals by separating the two scales involved: the scale of the macroscopic displacement field and the scale of the electron density. In the second part of the article two different versions of coupling between the TFW functional and its continuum limit are suggested. Both are based on decomposing the computational domain into a nonsmooth part (where atomistic detail is needed) and a smooth part (where the approximation by the continuum limit is thought to be accurate). In the first coupling method, the TFW model is used in the whole domain, however, the electron density in the smooth domain is obtained from local cell problems. This approach is shown not to give ghost forces. The second coupling method is obtained by replacing the energy of the smooth region by its Cauchy–Born approximation. This time there are ghost forces due to the unsymmetric treatment of the Coulomb interaction.

4.1.2 Outline of the Field-Based Model

We now motivate a basic atomistic interaction that is mediated by a field. The following ideas were first outlined in [70]. There, a coarse-grained version of the model was suggested as a potential alternative to classical QC coupling.

We start our considerations with a simple atomistic energy based on a pair-potential V in one dimension. Let $\boldsymbol{y} = (y_1, \dots, y_N) \in \mathbb{R}^N$ represent the coordinates of N particles. We consider the energy

$$\mathcal{E}(\boldsymbol{y}) = \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} V(|y_i - y_j|).$$

Obviously, the force on particle i is given by

$$-D_{y_i}\mathcal{E}(\boldsymbol{y}) = -\sum_{\substack{j=1 \\ j \neq i}}^{N} \operatorname{sign}(y_i - y_j) V'(|y_i - y_j|).$$

We note that the forces are nonlocal expressions in the sense that their computation involves the summation over the other $N - 1$ particles.

Next, we make a few modifications to this model. First, we replace the pointwise particles with smooth, nonnegative, and compactly supported particle densities $\delta_\varepsilon(\cdot - y_i)$ (such that $\int_{\mathbb{R}} \delta_\varepsilon(x)\,\mathrm{d}x = 1$). This leads to

$$\mathcal{E}(\boldsymbol{y}) \approx \frac{1}{2} \sum_{\substack{i,j=1 \\ i \neq j}}^{N} \int_{\mathbb{R}} \int_{\mathbb{R}} \delta_\varepsilon(z - y_i) V(z - x) \delta_\varepsilon(x - y_j)\,\mathrm{d}z\,\mathrm{d}x\,.$$

To simplify the presentation further, we include the self-energies of the individual particle densities and define

$$\mathcal{E}_\varepsilon(\boldsymbol{y}) = \frac{1}{2} \sum_{i,j=1}^{N} \int_{\mathbb{R}} \int_{\mathbb{R}} \delta_\varepsilon(z - y_i) V(|z - x|) \delta_\varepsilon(x - y_j)\,\mathrm{d}z\,\mathrm{d}x\,.$$

This additional self-energy contribution does not affect the forces. It can be computed explicitly and subtracted from the energy later on. We introduce the field $\phi : \mathbb{R} \to \mathbb{R}$ through

$$\phi(x) = \int_{\mathbb{R}} \rho_{\boldsymbol{y}}(z) V(|x - z|)\,\mathrm{d}z, \quad \text{where} \quad \rho_{\boldsymbol{y}}(z) = \sum_{i=1}^{N} \delta_\varepsilon(z - y_i). \tag{4.1}$$

Then, the energy $\mathcal{E}_\varepsilon(\boldsymbol{y})$ can be written in the following form

$$\mathcal{E}_\varepsilon(\boldsymbol{y}) = \frac{1}{2} \int_{\mathbb{R}} \rho_{\boldsymbol{y}}(x) \phi(x)\,\mathrm{d}x\,.$$

It is easy to see that the forces are now given by the *local* expression

$$-D_{\boldsymbol{y}} \mathcal{E}_\varepsilon(\boldsymbol{y}) = -\int_{\mathbb{R}} D_{\boldsymbol{y}} \rho_{\boldsymbol{y}}(z) \phi(z)\,\mathrm{d}z.$$

By knowing the field ϕ it is unnecessary to compute the force on one particle by adding up the forces that are exerted by all other particles. The nonlocality of the interaction has been encoded in the field ϕ. However, we have replaced the problem of nonlocality with the necessity to calculate the field ϕ, which is defined on the whole of $\mathbb{R}$, via the convolution (4.1).

If the pair-potential V is the Green's function belonging to a linear partial differential operator $L_V(\nabla)$, then ϕ can be computed by solving an equation with right-hand side $\rho_{\boldsymbol{y}}$:

$$L_V(\nabla)\phi = \rho_{\boldsymbol{y}}.$$

As an example we consider the so-called Yukawa potential in one space dimension

$$V(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{1}{k^2 + m^2}\, \mathrm{e}^{\mathrm{i}kx}\,\mathrm{d}k = \frac{1}{2m} \mathrm{e}^{-m|x|}.$$

In this case ϕ can be obtained as the solution to

$$-\Delta\phi + m^2 \phi = \rho_{\boldsymbol{y}}$$

or, equivalently, the minimization problem

$$\phi = \arg\min_{\varphi} \left\{ \frac{1}{2} \int_{\mathbb{R}} |\nabla \varphi|^2 + m^2 \varphi^2 \, \mathrm{d}x \, - \int_{\mathbb{R}} \rho_{\boldsymbol{y}} \varphi \, \mathrm{d}x \right\}.$$

The interaction potential $\mathcal{E}_\varepsilon$ takes the form

$$\mathcal{E}_\varepsilon(\boldsymbol{y}) = -\min_{\varphi} \left\{ \frac{1}{2} \int_{\mathbb{R}} |\nabla \varphi|^2 + m^2 \varphi^2 \, \mathrm{d}x \, - \int_{\mathbb{R}} \rho_{\boldsymbol{y}} \varphi \, \mathrm{d}x \right\}. \tag{4.2}$$

The interaction defined by (4.2) is purely repulsive. A purely attractive interaction can be obtained by changing the outer minus sign in the definition of $\mathcal{E}_\varepsilon$ to a plus sign. We could combine two energies $\mathcal{E}_{\varepsilon,+}$, $\mathcal{E}_{\varepsilon,-}$ of the form (4.2) with different values for m to model an interaction similar to Morse's potential $V(|x|) = \mathrm{e}^{-2|x|} - 2\mathrm{e}^{-|x|}$. Note that this would give rise to two fields ϕ_+ and ϕ_-. Forces and weak formulations could simply be added.

The present chapter is devoted to the analysis of QC approximations of (4.2) in a periodic one-dimensional setting. The basic idea of QC coupling in this case is as follows. The computational domain is divided into an atomistic and a continuum region. In the continuum region, we use the standard Cauchy–Born approximation of $\mathcal{E}_\varepsilon$. For the atomistic part we use a version of (4.2) on a bounded domain Ω^{at} subject to certain boundary conditions. Both the boundary and the boundary data will be allowed to depend on the configuration $\boldsymbol{y}$.

4.1.3 Notation

When working with atomistic models, boundaries have to be treated carefully. Strictly speaking they represent defects that lead to boundary layers in the displacement. To avoid these difficulties we look at an infinite chain of atoms on the one-dimensional lattice $\widehat{\boldsymbol{X}} = \varepsilon\mathbb{Z}$, where $\varepsilon > 0$ is the reference lattice spacing. Moreover, to keep the functional analysis simple, we consider only $(2N+1)$-periodic displacements from the reference lattice (see also [96]). Let

$$\mathcal{U} = \left\{ \boldsymbol{u} \in \mathbb{R}^{\mathbb{Z}} : u_{j+(2N+1)} = u_j \;\; \forall j \in \mathbb{Z}, \quad \sum_{j=-N}^{N} u_j = 0 \right\}$$

and define

$$\mathcal{Y} = F\widehat{\boldsymbol{X}} + \mathcal{U}$$

with a macroscopic deformation gradient $F > 0$. As a computational domain we use the interval

$$\Omega = (y_{-N-1}, y_N).$$

To keep the reference length of the interval constant we set $\varepsilon = 2/(2N+1)$.

We define the finite differences $\boldsymbol{y}', \boldsymbol{y}'' \in \mathcal{U}$ for $\boldsymbol{y} \in \mathcal{Y}$ or $\mathcal{U}$ by their respective components

$$y'_j = \frac{y_j - y_{j-1}}{\varepsilon}, \qquad y''_j = \frac{y_{j+1} - 2y_j + y_{j-1}}{\varepsilon^2}.$$

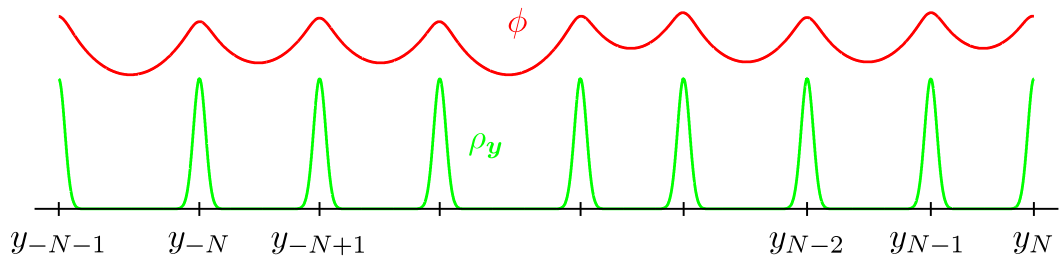

Figure 4.1: Sketch of the basic atomistic problem: the field ϕ is periodic in $\Omega = (y_{-N-1}, y_N)$ and $\rho_{\boldsymbol{y}}$ is a smooth particle density representing the atoms with positions given by $\boldsymbol{y} \in \mathcal{Y}$.

Let us also define the weighted ℓ^2-scalar product and norm by

$$(\boldsymbol{u}, \boldsymbol{v})_\varepsilon = \varepsilon \sum_{\nu=-N}^{N} u_\nu v_\nu \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}, \qquad \|\boldsymbol{u}\|_{\ell^2_\varepsilon} := (\boldsymbol{u}, \boldsymbol{u})_\varepsilon^{1/2} \quad \forall \boldsymbol{u} \in \mathcal{U}. \tag{4.3}$$

The ℓ^∞-norm is defined in the obvious way

$$\|\boldsymbol{u}\|_{\ell^\infty} = \max_{\nu=-N,\dots,N} |u_\nu| \quad \forall \boldsymbol{u} \in \mathcal{U}.$$

The space $\mathcal{U}$ equipped with the Sobolev-like norm $\|\boldsymbol{u}\|_{\mathcal{U}^{1,2}} = \|\boldsymbol{u}'\|_{\ell^2_\varepsilon}$ will be denoted by $\mathcal{U}^{1,2}$ and its topological dual space by $\mathcal{U}^{-1,2}$. The norm on $\mathcal{U}^{-1,2}$ is given by

$$\|T\|_{\mathcal{U}^{-1,2}} = \sup_{\boldsymbol{u} \in \mathcal{U}^{1,2}} \frac{T\boldsymbol{u}}{\|\boldsymbol{u}\|_{\mathcal{U}^{1,2}}}.$$

For monotonously increasing $\boldsymbol{y} \in \mathcal{Y}$ (which we will write as $\boldsymbol{y}' > 0$) we denote by $\mathrm{S}(\boldsymbol{y}) \subset \mathrm{H}^1(\Omega)$ the space of continuous functions that are linear on every interval $Q_i = (y_{i-1}, y_i)$, $i \in \{-N, \dots, N\}$. Furthermore, we define $\mathrm{S}_\#(\boldsymbol{y}) = \mathrm{S}(\boldsymbol{y}) \cap \mathrm{H}^1_\#(\Omega)$ to be the subset of all periodic functions in $\mathrm{S}(\boldsymbol{y})$.

4.2 The Model in a Periodic Setting

We now put the field-based interaction potential that was outlined above in a precise mathematical framework. For this, we define the functional $I : \mathrm{H}^1_\#(\Omega) \times \mathcal{Y} \to \mathbb{R}$ by

$$I(\varphi, \boldsymbol{y}) = \int_\Omega \left(\tfrac{1}{2}\varepsilon^2 |\nabla\varphi|^2 + \tfrac{1}{2} m^2 \varphi^2 \right) \mathrm{d}x \; - \int_\Omega \rho_{\boldsymbol{y}} \varphi \, \mathrm{d}x,$$

where

$$\rho_{\boldsymbol{y}}(x) = \varepsilon \sum_{j \in \mathbb{Z}} Z_j \delta_\varepsilon(x - y_j), \quad \text{and} \quad \delta_\varepsilon(x) = \varepsilon^{-1} \delta_1(x/\varepsilon).$$

Here, δ_1 is a symmetric, nonnegative, regularized delta distribution with compact support $\left[-\frac{\varsigma_0}{2}, \frac{\varsigma_0}{2}\right]$, where $\varsigma_0 > 0$ and $\int_{\mathbb{R}} \delta_1 \, \mathrm{d}x = 1$, see Figure 4.1. To avoid cluttering we set $Z_j = 1$ for all $j \in \mathbb{Z}$.

We then define the interaction potential $\mathcal{E} : \mathcal{Y} \to \mathbb{R}$ by

$$\mathcal{E}(\boldsymbol{y}) = - \min_{\varphi \in \mathrm{H}^1_\#(\Omega)} I(\varphi, \boldsymbol{y}). \tag{4.4}$$

The respective minimizer (see Figure 4.1)

$$\phi = \arg \min_{\varphi \in \mathrm{H}^1_\#(\Omega)} I(\varphi, \boldsymbol{y})$$

is obviously the periodic solution to the Euler–Lagrange equation

$$-\varepsilon^2 \Delta \phi + m^2 \phi = \rho_{\boldsymbol{y}} \quad \text{in } \Omega. \tag{4.5}$$

Although ϕ depends on $\boldsymbol{y}$, we will usually suppress this in our notation. It will always be clear from the context, which configuration ϕ belongs to. It follows immediately from (4.5) and integration by parts that

$$\mathcal{E}(\boldsymbol{y}) = \frac{1}{2} \int_\Omega \phi \rho_{\boldsymbol{y}} \, \mathrm{d}x \, .$$

To determine equilibrium configurations subject to a given external force $\boldsymbol{f} \in \mathcal{U}^{-1,2}$ we need to minimize the total potential energy $E_{\boldsymbol{f}} : \mathcal{Y} \to \mathbb{R}$ defined by

$$E_{\boldsymbol{f}}(\boldsymbol{y}) = \mathcal{E}(\boldsymbol{y}) + (\boldsymbol{f}, \boldsymbol{y})_\varepsilon. \tag{4.6}$$

A minimizer $\bar{\boldsymbol{y}} \in \mathcal{Y}$ of (4.6) will satisfy the Euler–Lagrange equation

$$DE_{\boldsymbol{f}}(\bar{\boldsymbol{y}}) = D\mathcal{E}(\bar{\boldsymbol{y}}) + \boldsymbol{f} = 0 \quad \in \mathcal{U}^{-1,2}.$$

In the following we address the derivatives of $\mathcal{E}$. Our main observation is a "weak" formulation for the first derivative $D\mathcal{E}$ that acts as a natural connection point for the coupling with a continuum model. The proof of the following result uses the same ideas involving the Implicit Function Theorem as we applied in Section 2.6. The situation is, however, significantly easier since the functional $I(\cdot, \boldsymbol{y})$ is convex and quadratic for every $\boldsymbol{y} \in \mathcal{Y}$.

Proposition 4.1. *The potential $\mathcal{E} : \mathcal{Y} \to \mathbb{R}$ defined by (4.4) is twice continuously Fréchet differentiable. The components of the first derivative are given by*

$$D_{y_j} \mathcal{E}(\boldsymbol{y}) = -\varepsilon \int_\Omega \nabla \delta_\varepsilon(x - y_j) \phi(x) \, \mathrm{d}x \tag{4.7}$$

for $j \in \{-N, \dots, N-1\}$ and by

$$D_{y_N} \mathcal{E}(\boldsymbol{y}) = -\varepsilon \int_\Omega \big(\nabla \delta_\varepsilon(x - y_{-N-1}) + \nabla \delta_\varepsilon(x - y_N) \big) \phi(x) \, \mathrm{d}x \, . \tag{4.8}$$

Proof. First we note that the map $\boldsymbol{y} \mapsto \rho_{\boldsymbol{y}}(x)$ from $\mathcal{Y}$ to $\mathbb{R}$ is continuously differentiable for all x and $D_{\boldsymbol{y}}\rho_{\boldsymbol{y}}(x)$ is uniformly bounded in x. Since $\Omega = (y_{-N-1}, y_N)$ depends on $\boldsymbol{y}$ but $|\Omega| = 2F$ is fixed, we will only look at the internal atoms represented by $\widetilde{\boldsymbol{y}} = (y_{-N}, \dots, y_{N-1})$. The derivative with respect to y_N follows from periodicity or by simply shifting Ω to the right by one atom.

For every fixed $\boldsymbol{y} \in \mathcal{Y}$ there is a unique minimizer $\phi(\boldsymbol{y})$ of $I(\cdot, \boldsymbol{y})$ (we are slightly abusing notation here and briefly interpret ϕ as a function from $\mathcal{Y}$ to $\mathrm{H}^1_\#(\Omega)$). For every $\boldsymbol{y} \in \mathcal{Y}$ the function $\phi(\boldsymbol{y}) \in \mathrm{H}^1_\#(\Omega)$ satisfies the Euler–Lagrange equation $D_\phi I(\phi(\boldsymbol{y}), \boldsymbol{y}) = 0$. Since $D_{\phi\phi} I(\phi, \boldsymbol{y}) = -\varepsilon^2\Delta + m^2\mathrm{id}$ is positive definite for all ϕ and all $\boldsymbol{y}$, the function $\widetilde{\boldsymbol{y}} \mapsto \phi(\boldsymbol{y})$ is differentiable by Theorem 4.B in [124]. We interpret the derivative $D_{\widetilde{\boldsymbol{y}}}\phi(\boldsymbol{y}) = \big(D_{y_{-N}}\phi(\boldsymbol{y}), \dots, D_{y_{N-1}}\phi(\boldsymbol{y})\big)$ as a vector of $2N$ functions from $\mathrm{H}^1_\#(\Omega)$. Using the chain rule we then calculate the derivative $D_{\widetilde{\boldsymbol{y}}}\mathcal{E}(\boldsymbol{y})$ to be

$$D_{\widetilde{\boldsymbol{y}}}\mathcal{E}(\boldsymbol{y}) = D_\phi I(\phi(\boldsymbol{y}), \boldsymbol{y}) D_{\widetilde{\boldsymbol{y}}}\phi(\boldsymbol{y}) + D_{\widetilde{\boldsymbol{y}}} I(\phi(\boldsymbol{y}), \boldsymbol{y}) = D_{\widetilde{\boldsymbol{y}}} I(\phi(\boldsymbol{y}), \boldsymbol{y}).$$

Because ϕ is a minimizer of $I(\cdot, \boldsymbol{y})$ (and therefore $D_\phi I(\phi(\boldsymbol{y}), \boldsymbol{y}) = 0$) to calculate the derivative of $\mathcal{E}$ it is sufficient to calculate the partial derivative of I with respect to $\widetilde{\boldsymbol{y}}$. By uniform differentiability of $\rho_{\boldsymbol{y}}(x)$ with respect to $\boldsymbol{y}$ and continuity of ϕ we can differentiate under the integral sign [103, Theorem 9.42] and arrive at

$$D_{\widetilde{\boldsymbol{y}}}\mathcal{E}(\boldsymbol{y}) = D_{\widetilde{\boldsymbol{y}}} \int_\Omega \rho_{\boldsymbol{y}}(x)\phi(x)\,\mathrm{d}x = \int_\Omega D_{\widetilde{\boldsymbol{y}}}\rho_{\boldsymbol{y}}(x)\phi(x)\,\mathrm{d}x\,.$$

The expression (4.7) for $j = -N, \dots, N-1$ then follows directly from

$$D_{y_j}\rho_{\boldsymbol{y}}(x) = \varepsilon D_{y_j}\delta_\varepsilon(x - y_j) = -\varepsilon\nabla\delta_\varepsilon(x - y_j)$$

for all $x \in \Omega$. □

We again stress the fact that the forces $-D_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y})$ are local expressions. To calculate the force on atom j it is necessary to know ϕ in $\operatorname{supp}\delta(\cdot - y_j)$ but there is no need to sum over all remaining atoms. This nonlocality is encoded in the field ϕ.

Next we establish the *weak formulation* for the forces on particles. This very much resembles the structure of the continuum equations and will be the basis for the QC coupling in Section 4.5. A version of this calculation was already shown in [60]. There, the author worked with an interpolant that was assumed constant on the support of every $\delta_\varepsilon(\cdot - y_j)$.

For simplicity we assume that the supports of the densities of different particles do not intersect:

$$\operatorname{supp}\delta_\varepsilon(\cdot - y_i) \cap \operatorname{supp}\delta_\varepsilon(\cdot - y_j) = \emptyset \quad \forall i, j \in \mathbb{Z}, \quad i \neq j.$$

Since, $|\operatorname{supp}\delta_\varepsilon(\cdot - y_i)| = \varepsilon\varsigma_0$, this is equivalent to $|y_j - y_i| > \varepsilon\varsigma_0$ for $i \neq j$ or, if $\boldsymbol{y}$ is monotonously increasing $y'_j > \varsigma_0$ for all $j \in \mathbb{Z}$.

Lemma 4.2. *Let $\boldsymbol{y} \in \mathcal{Y}$ satisfy $\boldsymbol{y}' > \varsigma_0$ and let $\phi \in \mathrm{H}^1_{\#}(\Omega)$ be the corresponding field defined by (4.5). Moreover, let $\boldsymbol{u} = (u_j)_{j\in\mathbb{Z}} \in \mathcal{U}$ be a test vector and $u \in \mathrm{S}_{\#}(\boldsymbol{y})$ a periodic piecewise linear interpolant of $\boldsymbol{u}$ in the sense that*

$$u(y_j) = u_j \quad \forall j \in \{-N-1, \dots, N\}. \tag{4.9}$$

Then,

$$D\mathcal{E}(\boldsymbol{y})\cdot\boldsymbol{u} = \sum_{j=-N}^{N} D_{y_j}\mathcal{E}(\boldsymbol{y})\cdot u_j = \int_\Omega \sigma_{\boldsymbol{y}}(x)\nabla u(x)\,\mathrm{d}x\,, \tag{4.10}$$

where $\sigma_{\boldsymbol{y}} = \sigma_{\boldsymbol{y},1} + \sigma_{\boldsymbol{y},2}$ and

$$\begin{aligned} \sigma_{\boldsymbol{y},1}(x) &= \tfrac{1}{2}\varepsilon^2|\nabla\phi|^2 - \tfrac{1}{2}m^2\phi^2 + \rho_{\boldsymbol{y}}\phi, \\ \sigma_{\boldsymbol{y},2}(x) &= \varepsilon \sum_{j=-N-1}^{N} \phi(x)\nabla\delta_\varepsilon(x-y_j)(x-y_j). \end{aligned} \tag{4.11}$$

Proof. Let $u \in \mathrm{S}_{\#}(\boldsymbol{y})$ be the interpolant of $\boldsymbol{u}$ satisfying (4.9). We start by multiplying the derivative (4.7) for $j \in \{-N, \dots, N-1\}$ by the component u_j:

$$\begin{aligned} D_{y_j}\mathcal{E}(\boldsymbol{y})u_j =& -\varepsilon u_j \int_\Omega \nabla\delta_\varepsilon(x-y_j)\phi(x)\,\mathrm{d}x \\ =& -\varepsilon\int_\Omega u(x)\nabla\delta_\varepsilon(x-y_j)\phi(x)\,\mathrm{d}x + \varepsilon\int_\Omega (u(x)-u_j)\nabla\delta_\varepsilon(x-y_j)\phi(x)\,\mathrm{d}x \\ =& \varepsilon\int_\Omega \delta_\varepsilon(x-y_j)u(x)\nabla\phi(x)\,\mathrm{d}x + \varepsilon\int_\Omega \delta_\varepsilon(x-y_j)\phi(x)\nabla u(x)\,\mathrm{d}x \\ &+ \varepsilon\int_\Omega (u(x)-u_j)\nabla\delta_\varepsilon(x-y_j)\phi(x)\,\mathrm{d}x =: T_1^{(j)} + T_2^{(j)} + T_3^{(j)}. \end{aligned}$$

Here we have used integration by parts but there are no boundary terms since u, ϕ and $\rho_{\boldsymbol{y}}$ are periodic on Ω. Using (4.8) we obtain a similar expression for $D_{y_N}\mathcal{E}(\boldsymbol{y})u_N$. Summing over $j = -N, \dots, N$ we obtain

$$D\mathcal{E}(\boldsymbol{y})\cdot\boldsymbol{u} = \sum_{j=-N}^{N} D_{y_j}\mathcal{E}(\boldsymbol{y})\cdot u_j = T_1 + T_2 + T_3, \tag{4.12}$$

where $T_i = \sum_{j=-N}^{N} T_i^{(j)}$, $i \in \{1,2,3\}$. From $\rho_{\boldsymbol{y}} = \varepsilon\sum_{j\in\mathbb{Z}}\delta_\varepsilon(\cdot - y_j)$ it immediately follows that

$$T_2 = \int_\Omega \rho_{\boldsymbol{y}}(x)\phi(x)\nabla u(x)\,\mathrm{d}x\,.$$

For T_1 we can carry out the following rearrangements

$$\begin{aligned} T_1 &= \int_\Omega \rho_{\boldsymbol{y}} u \nabla\phi \,\mathrm{d}x \\ &= \int_\Omega \big(-\varepsilon^2\Delta\phi + m^2\phi\big) u\nabla\phi \,\mathrm{d}x \\ &= \int_\Omega \big(-\varepsilon^2\nabla\phi\Delta\phi + m^2\phi\nabla\phi\big) u \,\mathrm{d}x \\ &= \frac{1}{2}\int_\Omega \nabla\big(-\varepsilon^2|\nabla\phi|^2 + m^2\phi^2\big) u \,\mathrm{d}x \\ &= \frac{1}{2}\int_\Omega \big(\varepsilon^2|\nabla\phi|^2 - m^2\phi^2\big)\nabla u \,\mathrm{d}x\,. \end{aligned}$$

Here, we have again used integration by parts and the periodicity of all functions involved. We deduce that

$$T_1 + T_2 = \int_\Omega \sigma_{\boldsymbol{y},1}(x)\nabla u(x)\,\mathrm{d}x$$

with $\sigma_{\boldsymbol{y},1}$ as defined in (4.11).

Before turning to T_3 we first note that since u is piecewise linear

$$u(x) = u_j + \frac{x-y_j}{y_j - y_{j-1}}(u_j - u_{j-1}) = u_j + (x-y_j)\nabla u(x) \quad \text{for } x \in Q_j = (y_{j-1}, y_j),$$
$$u(x) = u_j + \frac{x-y_j}{y_{j+1} - y_j}(u_{j+1} - u_j) = u_j + (x-y_j)\nabla u(x) \quad \text{for } x \in Q_{j+1} = (y_j, y_{j+1}).$$

Hence, T_3 in the above equation (4.12) can be written as

$$\begin{aligned} T_3 &= \varepsilon \sum_{j=-N-1}^{N} \int_\Omega \phi(x)\nabla\delta_\varepsilon(x-y_j)(u(x)-u_j)\,\mathrm{d}x \\ &= \varepsilon \sum_{j=-N-1}^{N} \int_\Omega \phi(x)\nabla\delta_\varepsilon(x-y_j)(x-y_j)\nabla u(x)\,\mathrm{d}x \\ &= \varepsilon \int_\Omega \sigma_{\boldsymbol{y},2}(x)\nabla u\,\mathrm{d}x\,, \end{aligned}$$

with $\sigma_{\boldsymbol{y},2}$ as defined in (4.11), which concludes the proof. ◻

Remark 4.3. In more than one space dimension the above calculations can be generalized if a triangular, respectively, tetrahedral mesh with the atomic positions as nodes is constructed. For example, this leads to

$$\sigma_{\boldsymbol{y},1}(x) = \big(-\tfrac{1}{2}\varepsilon^2|\nabla\phi|^2 - \tfrac{1}{2}m^2\phi^2 + \rho_{\boldsymbol{y}}\phi\big)\,\mathrm{id} \; + \varepsilon^2\nabla\phi\otimes\nabla\phi.$$

A closer look at the calculations in the above proof also shows that the weak form can be obtained for semilinear models $-\varepsilon^2\Delta\phi + F'(\phi) = \rho_{\boldsymbol{y}}$ with any convex function F. Even a fourth-order model of the form $\varepsilon^4\Delta^2\phi - \varepsilon^2\Delta\phi + F'(\phi) = \rho_{\boldsymbol{y}}$ admits a weak formulation in a similar vein. ◻

As already suggested in the introduction to this chapter the Green's function for the differential operator $-\varepsilon^2\Delta + m^2\mathrm{id}$ acting on functions defined on $\mathbb{R}$ is given by

$$G_\varepsilon(x) = \frac{1}{2\varepsilon m}\,\mathrm{e}^{-\frac{m}{\varepsilon}|x|}. \tag{4.13}$$

This will now be proved rigorously by deriving an explicit formula for the function values $\phi(x)$ and $\nabla\phi(x)$ for $x \in \Omega$.

Proposition 4.4. *Let $\boldsymbol{y} \in \mathcal{Y}$ be given and $\phi = \arg\min_{\varphi\in \mathrm{H}^1_\#(\Omega)} I(\varphi, \boldsymbol{y})$ be the corresponding interaction field. Then, for every $x \in \Omega$,*

$$\phi(x) = \int_{\mathbb{R}} G_\varepsilon(x-z)\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z = \frac{1}{2m}\sum_{k\in\mathbb{Z}}\int_{\mathbb{R}} \delta_\varepsilon(z-y_k)\,\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z, \tag{4.14}$$

$$\nabla\phi(x) = \int_{\mathbb{R}} G_\varepsilon(x-z)\nabla\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z = \frac{1}{2m}\sum_{k\in\mathbb{Z}}\int_{\mathbb{R}} \nabla\delta_\varepsilon(z-y_k)\,\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z. \tag{4.15}$$

Proof. The proof is similar to the one given for Theorem 2.1 in [51]. We start by constructing the solution $\phi_0 : \mathbb{R} \to \mathbb{R}$ to $-\varepsilon^2\Delta\phi_0 + m^2\phi_0 = \rho^c_{\boldsymbol{y}}$ in $\mathbb{R}$ for the compactly supported right-hand side $\rho^c_{\boldsymbol{y}} = \varepsilon\sum_{j=-N}^{N}\delta_\varepsilon(\cdot - y_j) \in \mathrm{C}^\infty_0(\mathbb{R})$, i.e., the particle density of the atoms $\{-N,\dots,N\}$. The periodic solution ϕ will be obtained by adding shifted versions of ϕ_0.

Let $\phi_0 : \mathbb{R} \to \mathbb{R}$ be defined by

$$\phi_0(x) = \int_{\mathbb{R}} G_\varepsilon(z)\rho^c_{\boldsymbol{y}}(x-z)\,\mathrm{d}z. \tag{4.16}$$

Moreover, let $\delta > 0$ be small. Since $\rho^c_{\boldsymbol{y}} \in \mathrm{C}^\infty_0(\mathbb{R})$ and G_ε is continuous we can differentiate under the integral sign [103, Theorem 9.42]:

$$\begin{aligned}\Delta\phi_0(x) &= \int_{\mathbb{R}} G_\varepsilon(z)\Delta\rho^c_{\boldsymbol{y}}(x-z)\,\mathrm{d}z\\ &= \int_{B_\delta(0)} G_\varepsilon(z)\Delta\rho^c_{\boldsymbol{y}}(x-z)\,\mathrm{d}z + \int_{\mathbb{R}\setminus B_\delta(0)} G_\varepsilon(z)\Delta\rho^c_{\boldsymbol{y}}(x-z)\,\mathrm{d}z.\end{aligned}$$

The first term on the right-hand side is $\mathcal{O}(\delta)$ as G_ε is bounded. For the second term we have

$$\begin{aligned}\int_{\mathbb{R}\setminus B_\delta(0)} G_\varepsilon(z)\Delta\rho^c_{\boldsymbol{y}}(x-z)\,\mathrm{d}z = &-\int_{\mathbb{R}\setminus B_\delta(0)} \nabla G_\varepsilon(z)\nabla\rho^c_{\boldsymbol{y}}(x-z)\,\mathrm{d}z\\ &+ G_\varepsilon(\delta)\big(\nabla\rho^c_{\boldsymbol{y}}(x-\delta) - \nabla\rho^c_{\boldsymbol{y}}(x+\delta)\big).\end{aligned}$$

The second term on the right-hand side of this equation is $\mathcal{O}(\delta)$ since $\nabla\rho^c_{\boldsymbol{y}}$ is globally Lipschitz continuous. Continuing with integration by parts yields

$$\begin{aligned}-\int_{\mathbb{R}\setminus B_\delta(0)} \nabla G_\varepsilon(z)\nabla\rho^c_{\boldsymbol{y}}(x-z)\,\mathrm{d}z = &\int_{\mathbb{R}\setminus B_\delta(0)} \Delta G_\varepsilon(z)\rho^c_{\boldsymbol{y}}(x-z)\,\mathrm{d}z\\ &+ \nabla G_\varepsilon(-\delta)\rho^c_{\boldsymbol{y}}(x-\delta) - \nabla G_\varepsilon(\delta)\rho^c_{\boldsymbol{y}}(x+\delta)\\ = &-\frac{m^2}{\varepsilon^2}\phi_0(x) + \frac{1}{\varepsilon^2}\rho^c_{\boldsymbol{y}}(x) + \mathcal{O}(\delta).\end{aligned}$$

Here, we have used that $-\varepsilon^2\Delta G_\varepsilon(x) + m^2 G_\varepsilon(x) = 0$ for $x \neq 0$ and $\nabla G_\varepsilon(\pm\delta) = \mp\frac{1}{2\varepsilon^2}\mathrm{e}^{\mp\frac{m}{\varepsilon}\delta}$. Letting $\delta \to 0$ shows that $-\varepsilon^2\Delta\phi_0 + m^2\phi_0 = \rho^c_{\boldsymbol{y}}$ in $\mathbb{R}$.

Next, we need to construct the $|\Omega|$-periodic solution ϕ. Because of the exponential decay of G_ε it is straightforward to verify that the series

$$\phi(x) = \sum_{j\in\mathbb{Z}} \phi_0(x + j|\Omega|)$$

converges uniformly on every compact subset of $\mathbb{R}$. Moreover, ϕ is Ω-periodic and solves the equation $-\varepsilon^2\Delta\phi + m^2\phi = \rho_{\boldsymbol{y}}$ in Ω. A simple change of coordinates in the integral (4.16) defining ϕ_0 implies (4.14).

Due to the exponential decay of the Green's function we can differentiate under the integral sign to get

$$\nabla\phi(x) = \nabla\int_{\mathbb{R}} \rho_{\boldsymbol{y}}(x-z)G_\varepsilon(z)\,\mathrm{d}z = \int_{\mathbb{R}} \nabla\rho_{\boldsymbol{y}}(x-z)G_\varepsilon(z)\,\mathrm{d}z = \int_{\mathbb{R}} \nabla\rho_{\boldsymbol{y}}(z)G_\varepsilon(x-z)\,\mathrm{d}z$$

for all $x \in \Omega$, which is equivalent to (4.15). □

Using the fact that $\rho_{\boldsymbol{y}}$ is $|\Omega|$-periodic we can write the integral (4.14) over $\mathbb{R}$ as an integral over Ω by introducing a periodic Green's function $G^{\#}_{\varepsilon,\Omega}$:

$$\phi(x) = \int_{\mathbb{R}} \rho_{\boldsymbol{y}}(z)G_\varepsilon(x-z)\,\mathrm{d}z = \int_\Omega \rho_{\boldsymbol{y}}(z)G^{\#}_{\varepsilon,\Omega}(x-z)\,\mathrm{d}z,$$

where $G^{\#}_{\varepsilon,\Omega}$ is given by

$$G^{\#}_{\varepsilon,\Omega}(x) = \frac{1}{2m\varepsilon}\sum_{\nu\in\mathbb{Z}} \mathrm{e}^{-\frac{m}{\varepsilon}|x-\nu|\Omega||}.$$

The energy $\mathcal{E}(\boldsymbol{y})$ then takes the form

$$\mathcal{E}(\boldsymbol{y}) = \frac{1}{2}\int_\Omega \rho_{\boldsymbol{y}}(x)\phi(x)\,\mathrm{d}x = \frac{1}{2}\int_\Omega\int_\Omega \rho_{\boldsymbol{y}}(x)G^{\#}_{\varepsilon,\Omega}(x-z)\rho_{\boldsymbol{y}}(z)\,\mathrm{d}x\,\mathrm{d}z. \tag{4.17}$$

A consequence of the simple exponential form of the Yukawa potential and some elementary properties of the exponential function in one dimension is the following. Let $y_i, y_j \in \mathbb{R}$ satisfy $y_j > y_i + \varepsilon\varsigma_0$ such that the supports of particle densities representing the atoms i and j do not intersect. Then,

$$\begin{aligned}
\int_{\mathbb{R}}\int_{\mathbb{R}} \delta_\varepsilon(z-y_j)\,\mathrm{e}^{-\frac{m}{\varepsilon}|z-x|}\delta_\varepsilon(x-y_i)\,\mathrm{d}x\,\mathrm{d}z &= \int_{\mathbb{R}}\int_{\mathbb{R}} \delta_\varepsilon(z-y_j)\,\mathrm{e}^{-\frac{m}{\varepsilon}(z-x)}\delta_\varepsilon(x-y_i)\,\mathrm{d}x\,\mathrm{d}z \\
&= \mathrm{e}^{-\frac{m}{\varepsilon}(y_j-y_i)}\int_{\mathbb{R}} \mathrm{e}^{-\frac{m}{\varepsilon}(z-y_j)}\delta_\varepsilon(z-y_j)\,\mathrm{d}z \\
&\qquad\cdot\int_{\mathbb{R}} \mathrm{e}^{-\frac{m}{\varepsilon}(y_i-x)}\delta_\varepsilon(y_i-x)\,\mathrm{d}x \\
&= \mu^2\,\mathrm{e}^{-\frac{m}{\varepsilon}(y_j-y_i)},
\end{aligned} \tag{4.18}$$

where we have defined

$$\mu = \int_{\mathbb{R}} \delta_\varepsilon(x) \mathrm{e}^{-\frac{m}{\varepsilon}x} \,\mathrm{d}x \;= \int_{\mathbb{R}} \delta_\varepsilon(x) \mathrm{e}^{\frac{m}{\varepsilon}x} \,\mathrm{d}x \;= \int_{\mathbb{R}} \delta_1(x) \mathrm{e}^{mx} \,\mathrm{d}x\,.$$

Although we will frequently use this property, it is not essential for our reasoning. It merely makes some calculations more convenient. We point out that $\mu = \mathcal{O}(\mathrm{e}^m)$.

The following result establishes L^∞-bounds on ϕ and $\nabla\phi$ that only depend on $m \min \boldsymbol{y}'$.

Lemma 4.5. *Let $\boldsymbol{y} \in \mathcal{Y}$, $\boldsymbol{y}' > \varsigma_0$, and let $\phi = \arg\min_{\varphi \in \mathrm{H}^1_\#(\Omega)} I(\varphi, \boldsymbol{y})$ be the corresponding field. Then, there are continuous functions K_0, K_1, independent of ε, such that*

$$\begin{aligned} \|\phi\|_{\mathrm{L}^\infty(\Omega)} &\leq K_0(m \min \boldsymbol{y}'), \\ \varepsilon \|\nabla\phi\|_{\mathrm{L}^\infty(\Omega)} &\leq K_1(m \min \boldsymbol{y}'). \end{aligned}$$

Proof. Let $x \in [y_{j-1}, y_j]$. Then, (4.14) and (4.18) imply that

$$\begin{aligned} |\phi(x)| &= \frac{1}{2m} \sum_{k \in \mathbb{Z}} \int_{\mathbb{R}} \delta_\varepsilon(z - y_k)\, \mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} \,\mathrm{d}z, \\ &= \frac{1}{2m} \int_{\mathbb{R}} \big(\delta_\varepsilon(z - y_{j-1}) + \delta_\varepsilon(z - y_j)\big)\, \mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} \,\mathrm{d}z + \frac{\mu}{2m} \sum_{\substack{k \in \mathbb{Z} \\ k \neq j-1, j}} \mathrm{e}^{-\frac{m}{\varepsilon}|x-y_k|}. \end{aligned}$$

Since $|x - y_k| \geq (k-j)\varepsilon \min \boldsymbol{y}'$ for all $k > j$, and $|x - y_k| \geq (j-1-k)\varepsilon \min \boldsymbol{y}'$ for all $k < j-1$,

$$\begin{aligned} \phi(x) &\leq \frac{1}{m} \int_{\mathbb{R}} \delta_\varepsilon(z)\, \mathrm{e}^{-\frac{m}{\varepsilon}|z|} \,\mathrm{d}z + \frac{\mu}{m} \sum_{\nu=1}^{\infty} \mathrm{e}^{-\nu m \min \boldsymbol{y}'} \\ &= \frac{1}{m} \int_{\mathbb{R}} \delta_\varepsilon(z)\, \mathrm{e}^{-\frac{m}{\varepsilon}|z|} \,\mathrm{d}z + \frac{\mu}{m} \frac{\mathrm{e}^{-m \min \boldsymbol{y}'}}{1 - \mathrm{e}^{-m \min \boldsymbol{y}'}} \leq K_0(m \min \boldsymbol{y}'). \end{aligned}$$

The integral term here is bounded since $\mathrm{e}^{-\frac{m}{\varepsilon}|z|}$ is bounded and $\int_{\mathbb{R}} \delta_\varepsilon(z) \,\mathrm{d}z = 1$. Similarly we obtain, with (4.15), that

$$\varepsilon |\nabla\phi(x)| \leq \frac{1}{m} \int_{\mathbb{R}} \varepsilon |\nabla\delta_\varepsilon(z)| \mathrm{e}^{-\frac{m}{\varepsilon}|z|} \,\mathrm{d}z + \mu \frac{\mathrm{e}^{-m \min \boldsymbol{y}'}}{1 - \mathrm{e}^{-m \min \boldsymbol{y}'}} \leq K_1(m \min \boldsymbol{y}'),$$

where we have used that $\int_{\mathbb{R}} \varepsilon |\nabla\delta_\varepsilon(z)| \,\mathrm{d}z$ is uniformly bounded in ε. ∎

Note that both K_0 and K_1 also implicitly depend on m. However, we think of m as fixed and therefore suppress this dependence. The parameter m determines the range of the interaction and therefore also $\boldsymbol{y}'$. Instances of $m\boldsymbol{y}'$ will appear frequently in the analysis of the QC method.

The following result shows coercivity of $D^2\mathcal{E}(\boldsymbol{y})$ if $\boldsymbol{y}$ is monotonously increasing. The coercivity constant decreases as the maximal distance of atoms (in the form of $\max \boldsymbol{y}'$) increases.

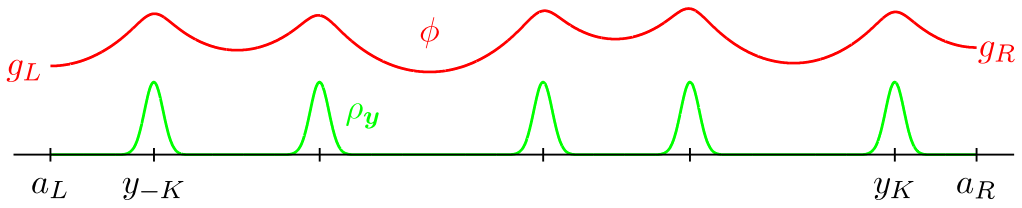

Figure 4.2: The atomistic model in the domain Ω_a with Dirichlet boundary conditions $g = [g_L\ g_R]^T$.

Lemma 4.6. *Let $\boldsymbol{y} \in \mathcal{Y}$ satisfy $\boldsymbol{y}' > \varsigma_0$. Then,*

$$D^2\mathcal{E}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq \frac{m\mu^2}{2}\mathrm{e}^{-m\max\boldsymbol{y}'}\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2 \quad \forall \boldsymbol{u} \in \mathcal{U}.$$

Proof. From the formula (4.17) we can deduce with (4.18) that

$$\begin{aligned}\mathcal{E}(\boldsymbol{y}) &= \frac{\varepsilon^2\mu^2}{2}\sum_{\substack{i,j=-N\\ i\neq j}}^{N} G^{\#}_{\varepsilon,\Omega}(y_i - y_j) + \mathcal{E}_{\mathrm{self}} \\ &= \frac{\varepsilon\mu^2}{4m}\sum_{\substack{i,j=-N\\ i\neq j}}^{N}\sum_{\nu\in\mathbb{Z}} \mathrm{e}^{-\frac{m}{\varepsilon}|y_i - y_j - \nu|\Omega||} + \mathcal{E}_{\mathrm{self}},\end{aligned}$$

where $\mathcal{E}_{\mathrm{self}}$ contains interactions of atoms with themselves and their periodic images and only depends on N and $|\Omega|$. Differentiating this twice with respect to $\boldsymbol{y}$ we obtain

$$D^2\mathcal{E}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] = \frac{m\varepsilon\mu^2}{4}\sum_{\substack{i,j=-N\\ i\neq j}}^{N}\sum_{\nu\in\mathbb{Z}} \mathrm{e}^{-\frac{m}{\varepsilon}|y_i - y_j - \nu|\Omega||}\frac{(u_i - u_j)^2}{\varepsilon^2}.$$

Since all terms in the sum are positive, only retaining nearest neighbour interactions (i.e., pairs (i,j) such that $|i-j| = 1$ as well as $(i,j) = (N,-N)$ and $(i,j) = (-N,N)$) leads to

$$D^2\mathcal{E}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq \frac{m\mu^2}{2}\mathrm{e}^{-m\max\boldsymbol{y}'}\varepsilon\sum_{i=-N}^{N}|u_i'|^2 = \frac{m\mu^2}{2}\mathrm{e}^{-m\max\boldsymbol{y}'}\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2.$$

This concludes the proof. ∎

4.3 The Model with Dirichlet Boundary Conditions

In this section we consider a version of the model (4.4) in the domain $\Omega_a = (a_L, a_R) \subset \mathbb{R}$ subject to Dirichlet instead of periodic boundary conditions. This concept will later be used as the atomistic subproblem of QC methods. We set $a = [a_L\ a_R]^T \in \mathbb{R}^2$ and $\Delta a = a_R - a_L$.

Throughout the present Section 4.3 we think of $\boldsymbol{y} = (y_{-K}, \dots, y_K)$ as an ordered element of Ω_a^{2K+1} such that $a_L < y_{-K} < \cdots < y_K < a_R$. The particle density $\rho_{\boldsymbol{y}}$ is canonically defined by

$$\rho_{\boldsymbol{y}} = \varepsilon \sum_{j=-K}^{K} \delta_\varepsilon(\cdot - y_j).$$

For simplicity we assume that the y_j lie well inside Ω_a in the sense that $\operatorname{supp}\rho_{\boldsymbol{y}} \cap \partial\Omega_a = \emptyset$ or, equivalently, $y_{-K} - a_L > \varsigma_0/2$ and $a_R - y_K > \varsigma_0/2$.

We impose the following boundary conditions on the resulting field $\phi : \Omega_a \to \mathbb{R}$:

$$\phi(a_L) = g_L, \qquad \phi(a_R) = g_R,$$

and write $g = [g_L \; g_R]^T \in \mathbb{R}^2$. Let $\xi_{a,g} \in \mathrm{H}^1(\Omega_a)$ be the function that satisfies the boundary value problem

$$\begin{aligned} -\varepsilon^2 \Delta \xi_{a,g} + m^2 \xi_{a,g} &= 0 \quad \text{in } \Omega_a, \\ \xi_{a,g}|_{\partial\Omega_a} &= g. \end{aligned}$$

Then we define the interaction potential $\mathcal{E}_{a,g} : \Omega_a^{2K+1} \to \mathbb{R}$ via

$$\mathcal{E}_{a,g}(\boldsymbol{y}) = - \min_{\varphi \in \xi_{a,g} + \mathrm{H}_0^1(\Omega_a)} I_a(\varphi, \boldsymbol{y}), \tag{4.19}$$

where the functional $I_a : \mathrm{H}^1(\Omega_a) \times \Omega_a^{2K+1} \to \mathbb{R}$ is given by

$$I_a(\varphi, \boldsymbol{y}) = \int_{a_L}^{a_R} \left(\tfrac{1}{2}\varepsilon^2 |\nabla\varphi|^2 + \tfrac{1}{2} m^2 \varphi^2\right) \mathrm{d}x \; - \int_{a_L}^{a_R} \rho_{\boldsymbol{y}} \varphi \, \mathrm{d}x \,. \tag{4.20}$$

For given $\boldsymbol{y}$ the minimizer $\phi = \arg\min_{\varphi \in \xi_{a,g} + \mathrm{H}_0^1(\Omega_a)} I_a(\varphi, \boldsymbol{y})$ is the weak solution to

$$\begin{aligned} -\varepsilon^2 \Delta \phi + m^2 \phi &= \rho_{\boldsymbol{y}} \quad \text{in } \Omega_a, \\ \phi|_{\partial\Omega_a} &= g. \end{aligned} \tag{4.21}$$

We will frequently use the decomposition

$$\phi = \phi_0 + \xi_{a,g}, \tag{4.22}$$

where $\phi_0 \in \mathrm{H}_0^1(\Omega_a)$ solves the boundary value problem

$$\begin{aligned} -\varepsilon^2 \Delta \phi_0 + m^2 \phi_0 \; &= \rho_{\boldsymbol{y}} \quad \text{in } \Omega_a, \\ \phi_0|_{\partial\Omega_a} \; &= 0. \end{aligned}$$

It is easy to show that $\xi_{a,g}$ has the form

$$\xi_{a,g}(x) = c_L(a,g) \mathrm{e}^{-\frac{m}{\varepsilon}(x - a_L)} + c_R(a,g) \mathrm{e}^{-\frac{m}{\varepsilon}(a_R - x)}, \tag{4.23}$$

where the coefficients $c_L(a,g)$ and $c_R(a,g)$ are given by

$$c(a,g) = \begin{bmatrix} c_L(a,g) \\ c_R(a,g) \end{bmatrix} = \begin{bmatrix} \frac{1}{1-\tau^2} & -\frac{\tau}{1-\tau^2} \\ -\frac{\tau}{1-\tau^2} & \frac{1}{1-\tau^2} \end{bmatrix} \begin{bmatrix} g_L \\ g_R \end{bmatrix} = T_a \cdot g \tag{4.24}$$

and we have defined

$$\tau = \mathrm{e}^{-\frac{m}{\varepsilon}\Delta a}.$$

We note that if $|\Omega_a| = \Delta a = a_R - a_L$ is large, we get $\tau \approx 0$ and therefore $c(a,g) \approx g$ because of the exponential decay of the Green's function.

Next, we compute the derivative of $\mathcal{E}_{a,g}$ with respect to the atomic coordinates. For these derivatives, we obtain a "weak" formulation of the same shape as in the periodic case (see Proposition 4.1).

If $\boldsymbol{y}$ is monotonously increasing in the sense that $y_{-K} < \cdots < y_K$, we denote by $\mathrm{S}(\boldsymbol{y} \cup a)$ the set of continuous, piecewise affine functions over the mesh given by the nodes $a_L, y_{-K}, \ldots, y_K, a_R$. Moreover, $\mathrm{S}_0(\boldsymbol{y} \cup a) = \mathrm{S}(\boldsymbol{y} \cup a) \cap \mathrm{H}^1_0(\Omega)$.

Proposition 4.7. *Let $a, g \in \mathbb{R}^2$ and $\boldsymbol{y} \in \mathcal{Y}$ be given. The potential $\mathcal{E}_{a,g} : \mathcal{Y} \to \mathbb{R}$ defined by* (4.19) *is continuously Fréchet differentiable.*
(i) The components of the first derivative are given by

$$D_{y_j}\mathcal{E}_{a,g}(\boldsymbol{y}) = -\varepsilon \int_{\Omega_a} \nabla\delta_\varepsilon(x - y_j)\phi(x)\,\mathrm{d}x \tag{4.25}$$

for $j = -K, \ldots, K$.
(ii) Let $\boldsymbol{u} \in \mathcal{U}$ be a test vector and $u \in \mathrm{S}_0(\boldsymbol{y} \cup a)$ an interpolant of $\boldsymbol{u}$ in the sense that

$$u(a_L) = u(a_R) = 0 \quad \text{and} \quad u(y_j) = u_j \quad \forall j \in \{-K, \ldots, K\}.$$

Then, if $y_{i+1} - y_i > \varsigma_0$ for all $i \in \{-K+1, \ldots, K\}$, $a_R - y_K > \varsigma_0/2$, and $y_{-K} - a_L > \varsigma_0/2$, we have

$$D_{\boldsymbol{y}}\mathcal{E}_{a,g}(\boldsymbol{y}) \cdot \boldsymbol{u} = \int_{\Omega_a} \sigma_{\boldsymbol{y}}(x) \nabla u(x)\,\mathrm{d}x\,,$$

where $\sigma_{\boldsymbol{y}}$ is given by (4.11).

Proof. The derivatives with respect to the coordinates $\boldsymbol{y}$ are easy to calculate along the same lines as in the proof of Proposition 4.1. The weak formulation can be obtained as in the periodic case (Lemma 4.2) using the fact that the interpolant u vanishes on $\partial\Omega_a$. □

It is worth pointing out that for $g \neq 0$ in general

$$\mathcal{E}_{a,g}(\boldsymbol{y}) \neq \frac{1}{2} \int_{\Omega_a} \rho_{\boldsymbol{y}} \phi\,\mathrm{d}x\,.$$

However, we will see below that $\mathcal{E}_{a,g}(\boldsymbol{y})$ can be written as the sum of a boundary data contribution and a term that is independent of g.

With a view to the subsequent derivation of QC methods we will from now on interpret a and g as arguments to $\mathcal{E}_{a,g}$ rather than fixed parameters entering its definition. In other words we consider the mapping $\Omega_a^{2K+1} \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$, $(\boldsymbol{y}, a, g) \mapsto \mathcal{E}_{a,g}(\boldsymbol{y})$. For future reference we derive the derivatives of this mapping with respect to the boundary a and the boundary data g.

4.3.1 Dependence on the Boundary

When formulating QC methods in Section 4.5 we will let the boundary a of the atomistic subdomain depend on the configuration $\boldsymbol{y}$. Therefore, we need to understand the dependence of the energy $\mathcal{E}_{a,g}(\boldsymbol{y})$ on a. Our main result is that the derivative $D_a\mathcal{E}_{a,g}(\boldsymbol{y})$ can be combined with $D_{\boldsymbol{y}}\mathcal{E}_{a,g}(\boldsymbol{y})$ into a weak formulation reminiscent of (4.10). This will be a central building block for QC methods.

Proposition 4.8. *Let $\boldsymbol{y} \in \mathcal{Y}$ satisfy $y_{i+1} - y_i > \varepsilon\varsigma_0$ for all $i \in \{-K+1, \dots, K\}$, $a_R - y_K > \varepsilon\varsigma_0/2$, and $y_{-K} - a_L > \varepsilon\varsigma_0/2$. Let $\boldsymbol{u} = (u_{-K}, \dots, u_K) \in \mathbb{R}^{2K+1}$ and $h = [h_L\ h_R]^T \in \mathbb{R}^2$ be test vectors. Moreover, let $u \in \mathrm{S}(\boldsymbol{y} \cup a)$ be the interpolant of $\boldsymbol{u}$ and h in the sense that*

$$u(a_L) = h_L, \quad u(a_R) = h_R, \quad \text{and} \quad u(y_j) = u_j \quad \forall j \in \{-K, \dots, K\}.$$

Then,

$$D_a\mathcal{E}_{a,g}(\boldsymbol{y})\cdot h + D_{\boldsymbol{y}}\mathcal{E}_{a,g}(\boldsymbol{y})\cdot \boldsymbol{u} = \int_{\Omega_a} \sigma_{\boldsymbol{y}}(x)\nabla u(x)\,\mathrm{d}x\,.$$

Proof. This is a direct consequence of the following two lemmas. □

For the first auxiliary lemma it is convenient to define a derivative of $\mathcal{E}_{a,g}(\boldsymbol{y})$ with respect to $a = [a_L\ a_R]^T$ when the relative distances between the atoms are kept constant. In other words we consider the change in $\mathcal{E}_{a,g}(\boldsymbol{y})$ when the whole domain Ω_a is stretched with the atom positions following this stretching. For given $\boldsymbol{y} \in \Omega_a^{2K+1}$ let $\boldsymbol{X} = (X_{-K}, \dots, X_K) \in (0,1)^{2K+1}$ be determined by $y_j = a_L + (a_R - a_L)X_j$ for all $j \in \{-K, \dots, K\}$. Then, for fixed g and $\boldsymbol{X}$ we define

$$\widetilde{\mathcal{E}}(a) := \mathcal{E}_{a,g}(a_L + (a_R - a_L)\boldsymbol{X}) \tag{4.26}$$

and

$$\widetilde{D}_{a_R}\mathcal{E}_{a,g}(\boldsymbol{y}) := D_{a_R}\widetilde{\mathcal{E}}(a).$$

Here, we interpret the sum $a_L + (a_R - a_L)\boldsymbol{X}$ in a componentwise manner: $(a_L + \Delta a\boldsymbol{X})_j = a_L + \Delta a X_j$ for all $j \in \{-K, \dots, K\}$. The derivative $\widetilde{D}_{a_L}\mathcal{E}_{a,g}(\boldsymbol{y})$ is defined analogously, from which it is immediately clear that $\widetilde{D}_{a_L}\mathcal{E}_{a,g}(\boldsymbol{y}) = -\widetilde{D}_{a_R}\mathcal{E}_{a,g}(\boldsymbol{y})$.

Lemma 4.9. *Let $\boldsymbol{y} \in \Omega_a^{2K+1}$. Then*

$$\widetilde{D}_{a_R}\mathcal{E}_{a,g}(\boldsymbol{y}) = \frac{1}{\Delta a}\int_{\Omega_a} \sigma_{\boldsymbol{y}}(x)\,\mathrm{d}x\,. \tag{4.27}$$

Proof. First, we set $\boldsymbol{\eta}(a) = a_L + \Delta a \boldsymbol{X}$. We begin by transforming the problem to the unit interval $(0,1)$ using the transformation $x \mapsto X(x) = (x-a_L)/(a_R-a_L)$:

$$\begin{aligned}\widetilde{\mathcal{E}}(a) = \mathcal{E}_{a,g}(\boldsymbol{\eta}(a)) &= \int_{\Omega_a}\Big(-\tfrac{1}{2}\varepsilon^2|\nabla\phi|^2 - \tfrac{1}{2}m^2\phi^2 + \rho_{\boldsymbol{\eta}(a)}\phi\Big)\,\mathrm{d}x \\ &= \Delta a\int_0^1\left(-\frac{\varepsilon^2}{2\Delta a^2}|\nabla\widehat{\phi}|^2 - \frac{m^2}{2}\widehat{\phi}^2 + \widehat{\rho}_{\boldsymbol{\eta}(a)}\widehat{\phi}\right)\mathrm{d}X.\end{aligned}$$

Here, $\widehat{\phi}(X) = \phi(x(X))$ and $\widehat{\rho}_{\boldsymbol{\eta}(a)}(X) = \rho_{\boldsymbol{\eta}(a)}(x(X))$. It follows as in Proposition 4.1 that to calculate $D_a\widetilde{\mathcal{E}}(a)$ it is sufficient to calculate the partial derivatives of the right-hand side with respect to a_R (the derivative of ϕ or $\widehat{\phi}$ with respect to a_R does not appear since ϕ is a minimizer of $I_a(\cdot,\boldsymbol{y})$). This leads to

$$\begin{aligned}D_{a_R}\widetilde{\mathcal{E}}(a) = &\int_0^1\left(-\frac{\varepsilon^2}{2\Delta a^2}|\nabla\widehat{\phi}|^2 - \frac{m^2}{2}\widehat{\phi}^2 + \widehat{\rho}_{\boldsymbol{\eta}(a)}\widehat{\phi}\right)\mathrm{d}X + \Delta a\int_0^1\frac{\varepsilon^2}{\Delta a^3}|\nabla\widehat{\phi}|^2\,\mathrm{d}X \\ &+ \Delta a\int_0^1\widehat{\phi}D_{a_R}\widehat{\rho}_{\boldsymbol{\eta}(a)}\,\mathrm{d}X\,.\end{aligned}$$

Transforming the first two integrals on the right-hand side back to the interval Ω_a we arrive at

$$\frac{1}{\Delta a}\mathcal{E}_{a,g}(\boldsymbol{y}) + \frac{\varepsilon^2}{\Delta a}\int_{\Omega_a}|\nabla\phi|^2\,\mathrm{d}x = \frac{1}{\Delta a}\int_{\Omega_a}\sigma_{\boldsymbol{y},1}(x)\,\mathrm{d}x\,,$$

where $\sigma_{\boldsymbol{y},1}$ was given in (4.11).

What remains to be done is differentiating $\widehat{\rho}_{\boldsymbol{\eta}(a)}$ with respect to a_R. By the definition of the transformation $x \mapsto X(x)$ we have

$$\begin{aligned}D_{a_R}\widehat{\rho}_{\boldsymbol{\eta}(a)}(X) &= \varepsilon D_{a_R}\sum_{j=-K}^{K}\delta_\varepsilon\big((a_R-a_L)(X-X_j)\big) \\ &= \varepsilon\sum_{j=-K}^{K}(X-X_j)\nabla\delta_\varepsilon\big((a_R-a_L)(X-X_j)\big).\end{aligned}$$

Using $\Delta a(X-X_j) = (x-y_j)$ we therefore get

$$\begin{aligned}\Delta a\int_0^1\widehat{\phi}D_{a_R}\widehat{\rho}_{\boldsymbol{\eta}(a)}\,\mathrm{d}X &= \frac{\varepsilon}{\Delta a}\sum_{j=-K}^{K}\int_{\Omega_a}(x-y_j)\nabla\delta_\varepsilon(x-y_j)\phi(x)\,\mathrm{d}x \\ &= \frac{1}{\Delta a}\int_{\Omega_a}\sigma_{\boldsymbol{y},2}(x)\,\mathrm{d}x\end{aligned}$$

with $\sigma_{\boldsymbol{y},2}(x)$ as given in (4.11). ∎

We can now derive a weak form for the derivative $D_a\mathcal{E}_{a,g}(\boldsymbol{y})$. Therefore, we define $\theta_R \in \mathrm{S}(\boldsymbol{y}\cup a)$ to be the piecewise linear function with

$$\theta_R(a_R) = 1, \quad \theta_R(a_L) = 0, \quad \theta_R(y_j) = 0 \ \text{ for all } j \in \{-K, \dots, K\}.$$

The function $\theta_L \in \mathrm{S}(\boldsymbol{y} \cup a)$ is defined analogously.

Lemma 4.10. *Let $\boldsymbol{y} \in \Omega_a^{2K+1}$ satisfy $y_{i+1} - y_i > \varepsilon\varsigma_0$ for all $i \in \{-K+1, \dots, K\}$, $a_R - y_K > \varepsilon\varsigma_0/2$, and $y_{-K} - a_L > \varepsilon\varsigma_0/2$. The derivatives of $\mathcal{E}_{a,g}(\boldsymbol{y})$ with respect to a_L, a_R (for fixed $\boldsymbol{y}$ and g) satisfy*

$$D_{a_L}\mathcal{E}_{a,g}(\boldsymbol{y}) = \int_{\Omega_a} \sigma_{\boldsymbol{y}}(x)\nabla\theta_L(x)\,\mathrm{d}x\,,$$
$$D_{a_R}\mathcal{E}_{a,g}(\boldsymbol{y}) = \int_{\Omega_a} \sigma_{\boldsymbol{y}}(x)\nabla\theta_R(x)\,\mathrm{d}x\,.$$

Proof. Let Θ_R be the affine function defined on Ω_a with $\Theta_R(a_L) = 0$, $\Theta_R(a_R) = 1$. Since $\nabla\Theta_R(x) = \frac{1}{\Delta a}$, Lemma 4.9 yields

$$\widetilde{D}_{a_R}\mathcal{E}_{a,\boldsymbol{y}}(\boldsymbol{y}) = \int_{\Omega_a} \sigma_{\boldsymbol{y}}\nabla\Theta_R\,\mathrm{d}x = \int_{\Omega_a} \sigma_{\boldsymbol{y}}\nabla(\Theta_R - \theta_R)\,\mathrm{d}x + \int_{\Omega_a} \sigma_{\boldsymbol{y}}\nabla\theta_R\,\mathrm{d}x\,. \tag{4.28}$$

Now, we have $\Theta_R - \theta_R \in \mathrm{S}_0(\boldsymbol{y} \cup a)$ and hence, by Proposition 4.7,

$$\int_{\Omega_a} \sigma_{\boldsymbol{y}}(x)\nabla(\Theta_R - \theta_R)\,\mathrm{d}x = \sum_{j=-K}^{K} D_{y_j}\mathcal{E}_{a,g}(\boldsymbol{y})\Theta_R(y_j). \tag{4.29}$$

However, $\widetilde{D}_{a_R}\mathcal{E}_{a,g}(\boldsymbol{y})$ was defined as derivative with respect to a_R, while the relative distances of the atoms are kept constant. This can be formulated as

$$\widetilde{D}_{a_R}\mathcal{E}_{a,g}(\boldsymbol{y}) = D_{a_R}\mathcal{E}_{a,g}(\boldsymbol{y}) + \sum_{j=-K}^{K} D_{y_j}\mathcal{E}_{a,g}(\boldsymbol{y})\Theta_R(y_j).$$

Inserting this into (4.28) and using (4.29) then gives

$$\int_{\Omega_a} \sigma_{\boldsymbol{y}}(x)\nabla\theta_R\,\mathrm{d}x = D_{a_R}\mathcal{E}_{a,g}(\boldsymbol{y}).$$

Similarly, we can show the expression stated for $D_{a_L}\mathcal{E}_{a,g}(\boldsymbol{y})$. □

4.3.2 Dependence on the Boundary Conditions

Next, we deal with the derivative of $\mathcal{E}_{a,g}(\boldsymbol{y})$ with respect to the boundary conditions g when the configuration $\boldsymbol{y}$ and the boundary a are kept fixed. Knowing this derivative is important

for subsequent QC coupling because the boundary data will in general depend on $\boldsymbol{y}$. We define

$$\begin{aligned}\gamma_L(\boldsymbol{y},a) &= 2\int_{\Omega_a} \rho_{\boldsymbol{y}}(x) G_\varepsilon(x-a_L)\,\mathrm{d}x\,,\\ \gamma_R(\boldsymbol{y},a) &= 2\int_{\Omega_a} \rho_{\boldsymbol{y}}(x) G_\varepsilon(a_R-x)\,\mathrm{d}x\,.\end{aligned} \tag{4.30}$$

Lemma 4.11. *The partial derivative of $\mathcal{E}_{a,g}(\boldsymbol{y})$ with respect to g is given by:*

$$D_g\mathcal{E}_{a,g}(\boldsymbol{y}) = -m\varepsilon\left((1-\tau^2)\begin{bmatrix}c_L(a,g)\\ c_R(a,g)\end{bmatrix} - \begin{bmatrix}\gamma_L(\boldsymbol{y},a)\\ \gamma_R(\boldsymbol{y},a)\end{bmatrix}\right)^T\cdot T_a,$$

where T_a and $c(a,g)=[c_L(a,g)\ c_R(a,g)]^T$ were given in (4.24) and $\tau = \mathrm{e}^{-\frac{m}{\varepsilon}\Delta a}$.

Proof. Throughout the proof we suppress the arguments of γ_L, γ_R, and c for readability. We recall the additive decomposition $\phi=\phi_0+\xi_{a,g}$ from (4.22). It follows from $-\varepsilon^2\Delta\xi_{a,g}+m^2\xi_{a,g}=0$ and $\phi_0\in \mathrm{H}^1_0(\Omega)$ that $\varepsilon^2(\nabla\xi_{a,g},\nabla\phi_0)+m^2(\xi_{a,g},\phi_0)=0$. Hence, a quick calculation shows that the energy $\mathcal{E}_{a,g}(\boldsymbol{y})$ can be written in the following additive way:

$$\mathcal{E}_{a,g}(\boldsymbol{y}) = -I_a(\phi,\boldsymbol{y}) = -I_a(\phi_0,\boldsymbol{y}) - I_a(\xi_{a,g},\boldsymbol{y}). \tag{4.31}$$

The first term on the right-hand side does not depend on the boundary conditions g and the second term is known explicitly: using $-\varepsilon^2\Delta\xi_{a,g}+m^2\xi_{a,g}=0$, integration by parts and the explicit formula (4.23) for $\xi_{a,g}$, we get

$$\begin{aligned}I_a(\xi_{a,g},\boldsymbol{y}) &= \int_{\Omega_a}\tfrac{1}{2}\big(\varepsilon^2|\nabla\xi_{a,g}|^2+m^2\xi_{a,g}^2\big)\,\mathrm{d}x \;-\int_{\Omega_a}\rho_{\boldsymbol{y}}\xi_{a,g}\,\mathrm{d}x\\ &= \frac{\varepsilon^2}{2}\big(-\xi_{a,g}(a_L)\nabla\xi_{a,g}(a_L)+\xi_{a,g}(a_R)\nabla\xi_{a,g}(a_R)\big) - \int_{\Omega_a}\rho_{\boldsymbol{y}}\xi_{a,g}\,\mathrm{d}x\\ &= \frac{\varepsilon m}{2}\big(c_L^2+c_R^2\big)\big(1-\mathrm{e}^{-2\frac{m}{\varepsilon}\Delta a}\big) - \int_{\Omega_a}\rho_{\boldsymbol{y}}\big(c_L\mathrm{e}^{-\frac{m}{\varepsilon}(x-a_L)}+c_R\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-x)}\big)\,\mathrm{d}x\\ &= m\varepsilon\left(\frac{c_L^2+c_R^2}{2}\big(1-\tau^2\big) - \frac{2}{2m\varepsilon}\int_{\Omega_a}\rho_{\boldsymbol{y}}\big(c_L\mathrm{e}^{-\frac{m}{\varepsilon}(x-a_L)}+c_R\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-x)}\big)\,\mathrm{d}x\right)\\ &= m\varepsilon\left(\frac{c_L^2+c_R^2}{2}\big(1-\tau^2\big) - \big(c_L\gamma_L+c_R\gamma_R\big)\right).\end{aligned}$$

Here we have used the Green's function G_ε from (4.13). Differentiating this expression with respect to c_L and c_R and applying the chain rule with $D_g c = T_a$ yield the result. ∎

Remark 4.12. We remark that $D_g\mathcal{E}_{a,g}(\boldsymbol{y})=0$ if and only if $c_L(a,g)=\gamma_L(\boldsymbol{y},a)/(1-\tau^2)$ and $c_R(a,g)=\gamma_R(\boldsymbol{y},a)/(1-\tau^2)$. According to (4.24) this corresponds to the boundary conditions

$$g_L^*(\boldsymbol{y},a) = \frac{1}{1-\tau}\frac{\gamma_L(\boldsymbol{y},a)+\tau\gamma_R(\boldsymbol{y},a)}{1+\tau}, \qquad g_R^*(\boldsymbol{y},a) = \frac{1}{1-\tau}\frac{\tau\gamma_L(\boldsymbol{y},a)+\gamma_R(\boldsymbol{y},a)}{1+\tau}. \tag{4.32}$$

In other words the boundary conditions are weighted averages of the values $\frac{1}{1-\tau}\gamma_L(\boldsymbol{y},a)$ and $\frac{1}{1-\tau}\gamma_R(\boldsymbol{y},a)$. It is quite instructive to derive an interpretation of these values. For example, we get

$$\begin{aligned}\frac{1}{1-\tau}\gamma_R(\boldsymbol{y},a) &= \frac{2}{1-\mathrm{e}^{-m\Delta a/\varepsilon}}\int_{a_L}^{a_R}\rho_{\boldsymbol{y}}(x)G_\varepsilon(a_R-x)\,\mathrm{d}x\\ &= 2\sum_{j=0}^{\infty}\mathrm{e}^{-j\frac{m}{\varepsilon}\Delta a}\int_{a_L}^{a_R}\rho_{\boldsymbol{y}}(x)G_\varepsilon(a_R-x)\,\mathrm{d}x\\ &= 2\sum_{j=0}^{\infty}\int_{a_L}^{a_R}\rho_{\boldsymbol{y}}(x)G_\varepsilon(a_R-(x-j\Delta a))\,\mathrm{d}x\,.\end{aligned}$$

Here we have used that

$$\mathrm{e}^{-j\frac{m}{\varepsilon}\Delta a}G_\varepsilon(a_R-x) = \frac{1}{2m\varepsilon}\mathrm{e}^{-j\frac{m}{\varepsilon}\Delta a}\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-x)} = \frac{1}{2m\varepsilon}\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-(x-j\Delta a))}.$$

On defining $\rho_{\boldsymbol{y},R}:(-\infty,a_R]\to\mathbb{R}$ to be the Δa-periodic continuation of $\rho_{\boldsymbol{y}}$ to the left of Ω_a, we obtain

$$\frac{1}{1-\tau}\gamma_R(\boldsymbol{y},a) = 2\int_{-\infty}^{a_R}\rho_{\boldsymbol{y},R}(x)G_\varepsilon(a_R-x)\,\mathrm{d}x\,.$$

Next, we introduce $\rho^{\mathrm{refl}}_{\boldsymbol{y},R}:\mathbb{R}\to\mathbb{R}$ by reflecting $\rho_{\boldsymbol{y},R}$ across a_R:

$$\rho^{\mathrm{refl}}_{\boldsymbol{y},R}(x) = \begin{cases}\rho_{\boldsymbol{y},R}(x), & \text{if } x\le a_R,\\ \rho_{\boldsymbol{y},R}(2a_R-x), & \text{if } x>a_R.\end{cases}$$

Together with the symmetry of G_ε this immediately leads to

$$\frac{1}{1-\tau}\gamma_R(\boldsymbol{y},a) = \int_{\mathbb{R}}\rho^{\mathrm{refl}}_{\boldsymbol{y},R}(x)G_\varepsilon(a_R-x)\,\mathrm{d}x\,.$$

Hence, we have obtained an interpretation of $\frac{1}{1-\tau}\gamma_R(\boldsymbol{y},a)$: it is the value of a field generated by a charge distribution that is point symmetric with respect to a_R and periodic on both half lines $(-\infty,a_R]$ and $[a_R,\infty)$. □

Remark 4.13. As seen in Lemma 4.11 the boundary data contribution $I_a(\xi_{a,g},\boldsymbol{y})$ to the energy $\mathcal{E}_{a,g}(\boldsymbol{y})$ is quadratic in g. For fixed configuration $\boldsymbol{y}$ and domain Ω_a the boundary conditions $g=g^*(\boldsymbol{y},a)$ minimize the boundary data contribution $I_a(\xi_{a,g},\boldsymbol{y})$ to the energy $\mathcal{E}_{a,g}(\boldsymbol{y})$. This is equivalent to minimizing $I_a(\cdot,\boldsymbol{y})$ over $\mathrm{H}^1(\Omega_a)$ and therefore leads to homogeneous Neumann boundary conditions $\nabla\phi=0$ on $\partial\Omega_a$. □

If the domain Ω_a is large and hence $\tau\approx 0$, then $\gamma_L(\boldsymbol{y},a)\approx g^*_L(\boldsymbol{y},a)$ and $\gamma_R(\boldsymbol{y},a)\approx g^*_R(\boldsymbol{y},a)$. We can then simplify the expression for $I_a(\xi_{a,g},\boldsymbol{y})$ given in the proof of Lemma 4.11:

$$\begin{aligned}I_a(\xi_{a,g},\boldsymbol{y}) &= m\varepsilon\bigg(\frac{c_L(a,g)^2+c_R(a,g)^2}{2}\big(1-\tau^2\big)-\big(c_L(a,g)\gamma_L(\boldsymbol{y},a)+c_R(a,g)\gamma_R(\boldsymbol{y},a)\big)\bigg)\\ &= m\varepsilon\bigg(\frac{g_L^2+g_R^2}{2}-\big(g_Lg^*_L(\boldsymbol{y},a)+g_Rg^*_R(\boldsymbol{y},a)\big)\bigg)+\mathcal{O}(\tau).\end{aligned}\tag{4.33}$$

The next result addresses the differentiability of γ_L and γ_R. We show that the derivatives satisfy certain bounds. Calculations like these are typical for this type of atomistic model.

Lemma 4.14. *Let $\boldsymbol{y} \in \Omega_a^{2K+1}$ satisfy $y_{i+1} - y_i > \varepsilon\varsigma_0$ for all $i \in \{-K+1, \dots, K\}$, $a_R - y_K > \varepsilon\varsigma_0/2$, and $y_{-K} - a_L > \varepsilon\varsigma_0/2$. Then,*

$$\gamma_L(\boldsymbol{y}, a) \leq \frac{\mu}{m} \frac{1}{1 - \mathrm{e}^{-m \min \boldsymbol{y}'}}, \qquad \gamma_R(\boldsymbol{y}, a) \leq \frac{\mu}{m} \frac{1}{1 - \mathrm{e}^{-m \min \boldsymbol{y}'}}.$$

Moreover, $\gamma_L(\boldsymbol{y})$ is twice continuously differentiable with respect to $\boldsymbol{y}$ and a and there exists $C(m \min \boldsymbol{y}')$ (independent of ε) such that

$$\begin{aligned}
\big|D\gamma_L(\boldsymbol{y}, a)\cdot(\boldsymbol{u}, h)\big| &\leq C(m \min \boldsymbol{y}')\bigg(\bigg(\frac{u_{-K} - h_L}{\varepsilon}\bigg)^2 + \sum_{k=-K+1}^{K} (u_k')^2\bigg)^{1/2}, \\
\big|D^2\gamma_L(\boldsymbol{y}, a)\cdot\big[(\boldsymbol{u}, h), (\boldsymbol{u}, h)\big]\big| &\leq C(m \min \boldsymbol{y}')\bigg(\bigg(\frac{u_{-K} - h_L}{\varepsilon}\bigg)^2 + \sum_{k=-K+1}^{K} (u_k')^2\bigg)
\end{aligned}$$

for all $\boldsymbol{u} \in \mathcal{U}$ and $h \in \mathbb{R}^2$. Analogous bounds hold for $\gamma_R(\boldsymbol{y}, a)$.

Proof. We start with the following observation

$$\gamma_L(\boldsymbol{y}, a) = \frac{1}{m} \sum_{j=-K}^{K} \int_{\Omega_a} \mathrm{e}^{-\frac{m}{\varepsilon}(x - a_L)} \delta_\varepsilon(x - y_j)\, \mathrm{d}x = \frac{\mu}{m} \mathrm{e}^{-\frac{m}{\varepsilon}(y_{-K} - a_L)} \sum_{j=-K}^{K} \mathrm{e}^{-\frac{m}{\varepsilon}(y_j - y_{-K})}.$$

Using $y_j - y_{-K} \leq (j + K) \min \boldsymbol{y}'$ for $j = -K, \dots, K$ we directly infer that

$$\gamma_L(\boldsymbol{y}, a) \leq \frac{\mu}{m} \sum_{j=-K}^{K} \mathrm{e}^{-\frac{m}{\varepsilon}(y_j - y_{-K})} \leq \frac{\mu}{m} \frac{1 - \mathrm{e}^{-(2K+1)m \min \boldsymbol{y}'}}{1 - \mathrm{e}^{-m \min \boldsymbol{y}'}} \leq \frac{\mu}{m} \frac{1}{1 - \mathrm{e}^{-m \min \boldsymbol{y}'}}.$$

Differentiating gives

$$\begin{aligned}
D_{a_L}\gamma_L(\boldsymbol{y}, a) h_L + D_{\boldsymbol{y}}\gamma_L(\boldsymbol{y}, a)\cdot\boldsymbol{u} &= -\mu \sum_{j=-K}^{K} \mathrm{e}^{-\frac{m}{\varepsilon}(y_j - a_L)} \frac{u_j - h_L}{\varepsilon}, \\
D^2\gamma_L(\boldsymbol{y}, a)\cdot\big[(\boldsymbol{u}, h), (\boldsymbol{u}, h)\big] &= m\mu \sum_{j=-K}^{K} \mathrm{e}^{-\frac{m}{\varepsilon}(y_j - a_L)} \frac{(u_j - h_L)^2}{\varepsilon^2}.
\end{aligned}$$

We show the stated bound for the second derivative. The one for the first derivative is obtained similarly. We have

$$\begin{aligned}
\bigg(\frac{u_j - h_L}{\varepsilon}\bigg)^2 &= \bigg(\frac{u_{-K} - h_L}{\varepsilon} + \sum_{k=-K+1}^{j} u_k'\bigg)^2 \\
&\leq 2\bigg(\frac{u_{-K} - h_L}{\varepsilon}\bigg)^2 + 2(j + K) \sum_{k=-K+1}^{j} (u_k')^2.
\end{aligned}$$

Therefore,

$$D^2\gamma_L(\boldsymbol{y},a)\cdot\big[(\boldsymbol{u},h),(\boldsymbol{u},h)\big] \le 2m\mu \sum_{j=-K}^{K}(j+K)\mathrm{e}^{-(j+K)m\min \boldsymbol{y}'} \cdot\left(\left(\frac{u_{-K}-h_L}{\varepsilon}\right)^2+\sum_{k=-K+1}^{K}(u_k')^2\right),$$

which is the desired bound. ◻

A useful fact for the analysis of QC methods is the global Lipschitz continuity of the solution ϕ in the L^∞-norm with respect to the boundary conditions g for fixed $\boldsymbol{y}$ and a. This is direct result of the exponential decay of the Green's function.

Lemma 4.15. *Let $\phi_1,\phi_2\in \mathrm{H}^1(\Omega_a)$ be minimizers of $I_a(\cdot,\boldsymbol{y})$ subject to the boundary conditions $g_1\in\mathbb{R}^2$, respectively, $g_2\in\mathbb{R}^2$. Then,*

$$\|\phi_1-\phi_2\|_{\mathrm{L}^\infty}\le\sqrt{2}|T_a||g_1-g_2|,$$
$$\varepsilon\|\nabla\phi_1-\nabla\phi_2\|_{\mathrm{L}^\infty}\le\sqrt{2}m|T_a||g_1-g_2|.$$

Proof. We write both functions in the form $\phi_i=\phi_0+\xi_{a,g_i}$, $i\in\{1,2\}$. Let $c_1=[c_{1,L}\ c_{1,R}]^T, c_2=[c_{2,L}\ c_{2,R}]^T$ be the respective coefficients entering ξ_{a,g_i}, $i\in\{1,2\}$. Hence, we get

$$\begin{aligned}|\phi_1(x)-\phi_2(x)| &= |\xi_{a,g_1}(x)-\xi_{a,g_2}(x)|\\ &\le |c_{1,L}-c_{2,L}|\,\mathrm{e}^{-\frac{m}{\varepsilon}(x-a_L)}+|c_{1,R}-c_{2,R}|\,\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-x)}\\ &\le \sqrt{2}|g_1-g_2||T_a|\end{aligned}$$

uniformly in x. Taking the supremum over $x\in\Omega_a$ yields the bound for $\|\phi_1-\phi_2\|_{\mathrm{L}^\infty}$. The bound for the derivatives is obtained similarly. ◻

The bound given in the previous result is rather crude. The effects from the boundary data decay exponentially away from $\partial\Omega_a$, and hence $|\phi_1(x)-\phi_2(x)|$ is much smaller well inside Ω_a.

4.3.3 The Green's Function on a Bounded Domain

As we saw in Proposition 4.4, the Green's function for the equation $-\varepsilon^2\Delta\phi+m^2\phi=\rho_{\boldsymbol{y}}$ in $\mathbb{R}$ is given by $G_\varepsilon(x,y)=\frac{1}{2m\varepsilon}\mathrm{e}^{-\frac{m}{\varepsilon}|x-y|}$. This could be used to obtain an explicit formula for $\phi(x)$ in the periodic case. We will now construct the Green's function $G_{\varepsilon,a}$ for the operator $-\varepsilon^2\Delta+m^2\mathrm{id}$ subject to homogeneous Dirichlet conditions on $\partial\Omega_a$. In the present 1D setting it is straightforward to determine $G_{\varepsilon,a}$, see for example [51, Chapter 2.2.4]. We have

$$G_{\varepsilon,a}(x,z)=G_\varepsilon(x,z)-H_{\varepsilon,a}(x,z),$$

where, for every fixed x, $H_{\varepsilon,a}(x,\cdot)$ solves the boundary value problem

$$\begin{aligned} -\varepsilon^2\Delta_z H_{\varepsilon,a}(x,\cdot) + m^2 H_{\varepsilon,a}(x,\cdot) &= 0 \ \text{ in } \Omega, \\ H_{\varepsilon,a}(x,a_L) &= G_\varepsilon(a_L - x), \\ H_{\varepsilon,a}(x,a_R) &= G_\varepsilon(a_R - x). \end{aligned}$$

The same ideas that led to formula (4.23) for $\xi_{a,g}$ yield

$$\begin{aligned} H_{\varepsilon,a}(x,z) &= \begin{bmatrix} \mathrm{e}^{-\frac{m}{\varepsilon}(z-a_L)} & \mathrm{e}^{-\frac{m}{\varepsilon}(a_R-z)} \end{bmatrix} \cdot T_a \cdot \begin{bmatrix} \frac{1}{2m\varepsilon}\,\mathrm{e}^{-\frac{m}{\varepsilon}(x-a_L)} \\ \frac{1}{2m\varepsilon}\,\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-x)} \end{bmatrix} \\ &= \frac{1}{2m\varepsilon}\frac{1}{1-\tau^2}\big(\mathrm{e}^{-\frac{m}{\varepsilon}(x+z-2a_L)} + \mathrm{e}^{-\frac{m}{\varepsilon}(2a_R-x-z)} \\ &\qquad - \tau\mathrm{e}^{-\frac{m}{\varepsilon}(x-z+a_R-a_L)} - \tau\mathrm{e}^{-\frac{m}{\varepsilon}(z-x+a_R-a_L)}\big). \end{aligned}$$

A rigorous proof now follows.

Lemma 4.16. *Let $\phi_0 \in \mathrm{H}^1_0(\Omega_a)$ satisfy $-\varepsilon^2\Delta\phi_0 + m^2\phi_0 = \rho_{\boldsymbol{y}}$ in Ω_a. Then,*

$$\phi_0(x) = \int_{\Omega_a} G_{\varepsilon,a}(x,z)\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z \quad \forall x \in \Omega_a. \tag{4.34}$$

Proof. It follows immediately from its definition that $H_{\varepsilon,a}$ satisfies

$$-\varepsilon^2\Delta_x H_{\varepsilon,a}(\cdot,z) + m^2 H_{\varepsilon,a}(\cdot,z) = 0$$

in Ω_a for all fixed z. The proof of Proposition 4.4 can then easily be generalized to show that the function ζ defined by

$$\zeta(x) = \int_{\Omega_a} \big(G_\varepsilon(x,z) - H_{\varepsilon,a}(x,z)\big)\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z$$

for $x \in \Omega_a$ satisfies $-\varepsilon^2\Delta\zeta + m^2\zeta = \rho_{\boldsymbol{y}}$ in Ω_a. It remains to show that ζ attains the appropriate values on the boundary $\partial\Omega_a$. We note that

$$\begin{aligned} H_{\varepsilon,a}(a_R,z) &= \frac{1}{2m\varepsilon}\frac{1}{1-\tau^2}\big(\tau\mathrm{e}^{-\frac{m}{\varepsilon}(z-a_L)} + \mathrm{e}^{-\frac{m}{\varepsilon}(a_R-z)} - \tau^2\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-z)} - \tau\mathrm{e}^{-\frac{m}{\varepsilon}(z-a_L)}\big) \\ &= \frac{1}{2m\varepsilon}\,\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-z)}. \end{aligned}$$

With $G_\varepsilon(a_R,z) = \frac{1}{2m\varepsilon}\,\mathrm{e}^{-\frac{m}{\varepsilon}(a_R-z)}$ this implies $G_{\varepsilon,a}(a_R,z) = 0$. Similarly we get $G_{\varepsilon,a}(a_L,z) = 0$. Therefore, $\zeta(a_R) = 0$ and $\zeta(a_L) = 0$ and we conclude that $\phi_0 = \zeta$. ∎

Looking at $H_{\varepsilon,a}$ we observe terms that vanish as $\tau \to 0$. The other terms decay exponentially away from $\partial\Omega_a$. For future reference we divide the Green's function $G_{\varepsilon,a}$ into two parts:

$$G_{\varepsilon,a}(x,z) = G^{(1)}_{\varepsilon,a}(x,z) + \tau G^{(2)}_{\varepsilon,a}(x,z), \tag{4.35}$$

where

$$G^{(1)}_{\varepsilon,a}(x,z) = \frac{1}{2m\varepsilon}\Big(\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} - \mathrm{e}^{-\frac{m}{\varepsilon}(x+z-2a_L)} - \mathrm{e}^{-\frac{m}{\varepsilon}(2a_R-x-z)}\Big),$$
$$G^{(2)}_{\varepsilon,a}(x,z) = -\frac{1}{2m\varepsilon}\frac{1}{1-\tau^2}\Big(\tau\mathrm{e}^{-\frac{m}{\varepsilon}(x+z-2a_L)} + \tau\mathrm{e}^{-\frac{m}{\varepsilon}(2a_R-x-z)} - \mathrm{e}^{-\frac{m}{\varepsilon}(x-z+a_R-a_L)} - \mathrm{e}^{-\frac{m}{\varepsilon}(z-x+a_R-a_L)}\Big).$$

If the domain Ω_a is large compared with ε, that is $\Delta a \gg \varepsilon$, we have $G_{\varepsilon,a} \approx G^{(1)}_{\varepsilon,a}$. To make the following formulas more readable we suppress the arguments of γ_L and γ_R.

Lemma 4.17. *For given $\boldsymbol{y} \in \Omega_a^{2K+1}$ let $\phi_0 \in \mathrm{H}^1_0(\Omega_a)$ satisfy $-\varepsilon^2\Delta\phi_0 + m^2\phi_0 = \rho_{\boldsymbol{y}}$ in Ω_a. Then,*

$$I_a(\phi_0,\boldsymbol{y}) = -\frac{1}{2}\int_{\Omega_a}\int_{\Omega_a}\rho_{\boldsymbol{y}}(x)G_\varepsilon(x,z)\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z\,\mathrm{d}x + \frac{m\varepsilon}{4}\big(\gamma_L^2+\gamma_R^2\big) + \frac{m\varepsilon}{4}\frac{\tau}{1-\tau^2}\big(\tau\gamma_L^2+\tau\gamma_R^2-2\gamma_L\gamma_R\big).$$

Proof. Since the function ϕ_0 is a minimizer of $I_a(\cdot,\boldsymbol{y})$ over $\mathrm{H}^1_0(\Omega)$, we have with the expression (4.34) for $\phi_0(x)$ that

$$I_a(\phi_0,\boldsymbol{y}) = -\frac{1}{2}\int_{\Omega_a}\int_{\Omega_a}\rho_{\boldsymbol{y}}\phi_0\,\mathrm{d}x = -\frac{1}{2}\int_{\Omega_a}\int_{\Omega_a}\rho_{\boldsymbol{y}}(x)G_{\varepsilon,a}(x,z)\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z\,\mathrm{d}x\,. \tag{4.36}$$

By the definition (4.30) of γ_L and γ_R we have

$$\begin{aligned}
\frac{1}{4m\varepsilon}\int_{\Omega_a}\int_{\Omega_a}\rho_{\boldsymbol{y}}(x)\,\mathrm{e}^{-\frac{m}{\varepsilon}(2a_R-x-z)}\rho_{\boldsymbol{y}}(z)\,\mathrm{d}x\,\mathrm{d}z &= \frac{m\varepsilon}{4}\gamma_R^2,\\
\frac{1}{4m\varepsilon}\int_{\Omega_a}\int_{\Omega_a}\rho_{\boldsymbol{y}}(x)\,\mathrm{e}^{-\frac{m}{\varepsilon}(x+z-2a_L)}\rho_{\boldsymbol{y}}(z)\,\mathrm{d}x\,\mathrm{d}z &= \frac{m\varepsilon}{4}\gamma_L^2,\\
\frac{1}{4m\varepsilon}\int_{\Omega_a}\int_{\Omega_a}\rho_{\boldsymbol{y}}(x)\,\mathrm{e}^{-\frac{m}{\varepsilon}(z-x+a_R-a_L)}\rho_{\boldsymbol{y}}(z)\,\mathrm{d}x\,\mathrm{d}z &= \frac{m\varepsilon}{4}\gamma_L\gamma_R.
\end{aligned} \tag{4.37}$$

Inserting the expression (4.35) for $G_{\varepsilon,a}$ into (4.36) and using these equalities yields the stated expression for $I_a(\phi_0,\boldsymbol{y})$. ∎

4.3.4 A Special Case

In this short section we take a look at the interaction potential $\mathcal{E}_{a,g}$ from (4.19) with the $\boldsymbol{y}$-dependent boundary conditions $g = g^*(\boldsymbol{y},a)$ from Remark 4.12. This will be a useful starting point for the design of QC methods in Section 4.5.

We have already seen in (4.31) that for any choice of boundary data $g \in \mathbb{R}^2$ the energy $\mathcal{E}_{a,g}(\boldsymbol{y})$ can be written as the sum of two terms

$$\mathcal{E}_{a,g}(\boldsymbol{y}) = -I_a(\phi_0,\boldsymbol{y}) - I_a(\xi_{a,g},\boldsymbol{y}),$$

where $I_a(\phi_0, \boldsymbol{y})$ is independent of the boundary conditions and was calculated in Lemma 4.17. The special choice $g^*(\boldsymbol{y}, a)$ of boundary conditions allows us to calculate $I_a(\xi_{a,g^*(\boldsymbol{y},a)}, \boldsymbol{y})$ easily, too.

Lemma 4.18. *For $\boldsymbol{y} \in \Omega_a^{2K+1}$ let $\xi_{a,g^*(\boldsymbol{y},a)}$ be given by (4.23) with $g = g^*(\boldsymbol{y}, a)$. Then,*

$$I_a(\xi_{a,g^*(\boldsymbol{y},a)}, \boldsymbol{y}) = -\frac{m\varepsilon}{2}\big(\gamma_L(\boldsymbol{y}, a)^2 + \gamma_R(\boldsymbol{y}, a)^2\big) - \frac{m\varepsilon}{2}\frac{\tau^2}{1-\tau^2}\big(\gamma_L(\boldsymbol{y}, a)^2 + \gamma_R(\boldsymbol{y}, a)^2\big). \quad (4.38)$$

Proof. Again, we suppress the arguments of γ_L, γ_R, and c for readability. From Lemma 4.11 we know that for general g

$$I_a(\xi_{a,g}, \boldsymbol{y}) = m\varepsilon\bigg(\frac{c_L^2 + c_R^2}{2}\big(1-\tau^2\big) - \big(c_L\gamma_L + c_R\gamma_R\big)\bigg).$$

If $g = g^*(\boldsymbol{y}, a)$, then $c_L = \gamma_L/(1-\tau^2)$ and $c_R = \gamma_R/(1-\tau^2)$ as seen in Remark 4.12. Hence,

$$I_a(\xi_{a,g^*(\boldsymbol{y},a)}, \boldsymbol{y}) = -\frac{m\varepsilon}{2}\frac{1}{1-\tau^2}\big(\gamma_L^2 + \gamma_R^2\big).$$

Isolating the dependence on τ gives the desired form of $I(\xi_{a,g^*(\boldsymbol{y},a)}, \boldsymbol{y})$. □

Adding $-I_a(\xi_{a,g^*(\boldsymbol{y},a)}, \boldsymbol{y})$ as just obtained and $-I_a(\phi_0, \boldsymbol{y})$ from Lemma 4.17 we arrive at

$$\begin{aligned}\mathcal{E}_{a,g^*(\boldsymbol{y},a)}(\boldsymbol{y}) = \frac{1}{4m\varepsilon}\int_{\Omega_a}\int_{\Omega_a} \rho_{\boldsymbol{y}}(x)\,\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z\,\mathrm{d}x + \frac{m\varepsilon}{4}\big(\gamma_L^2 + \gamma_R^2\big) \\ - \frac{m\varepsilon}{4}\frac{\tau}{1-\tau^2}\big(\tau\gamma_L^2 + 2\gamma_L\gamma_R + \tau\gamma_R^2\big).\end{aligned} \quad (4.39)$$

With (4.37) we can rewrite this as

$$\mathcal{E}_{a,g^*(\boldsymbol{y},a)}(\boldsymbol{y}) = \frac{1}{4m\varepsilon}\int_{\Omega_a}\int_{\Omega_a} \rho_{\boldsymbol{y}}(x)\big(\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} + \mathrm{e}^{-\frac{m}{\varepsilon}(2a_R - x - z)} + \mathrm{e}^{-\frac{m}{\varepsilon}(x+z-2a_L)}\big)\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z\,\mathrm{d}x$$

up to a term of order $\mathcal{O}(\tau)$. This can be interpreted as the energy of the atoms represented by $\boldsymbol{y}$ interacting with each other plus the interaction with mirror atoms outside Ω_a. This mirror interaction was introduced by means of the boundary conditions.

4.4 The Cauchy–Born Approximation

The next building block we need for the design of QC methods based on the model (4.4) is the respective continuum model, which will be derived using the Cauchy–Born approximation. Let $\boldsymbol{y} \in \mathcal{Y}$ satisfy $\min \boldsymbol{y}' > \varsigma_0$. As outlined in the Introduction the Cauchy–Born approximation consists in considering the cells $Q_j = (y_{j-1}, y_j)$ one by one and computing their energy as if they were part of an infinite chain with homogeneous deformation. This is equivalent to

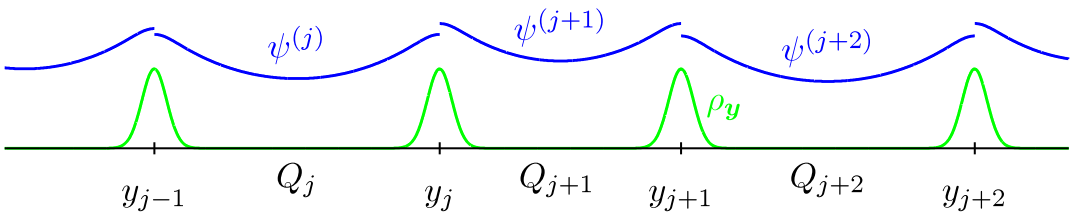

Figure 4.3: The Cauchy–Born approximation: independent periodic problems are solved on the cells $Q_j = (y_{j-1}, y_j)$ leading to locally defined fields $\psi^{(j)}$.

solving the periodic minimization problem restricted to Q_j (see Figure 4.3). We therefore define the Cauchy–Born energy of the cell Q_j by

$$\mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y}) = -\min_{\psi \in \mathrm{H}^1_{\#}(Q_j)} \left(\int_{Q_j} \left(\tfrac{1}{2}\varepsilon^2 |\nabla \psi|^2 + \tfrac{1}{2} m^2 \psi^2 \right) \mathrm{d}x - \int_{Q_j} \rho_{\boldsymbol{y}} \psi \, \mathrm{d}x \right). \tag{4.40}$$

Note that this energy only depends on the distance $(y_j - y_{j-1})$. The minimizer $\psi^{(j)}$ of the above problem (4.40) obviously satisfies the equation $-\varepsilon^2 \Delta \psi^{(j)} + m^2 \psi^{(j)} = \rho_{\boldsymbol{y}}$ in Q_j and its $|Q_j|$-periodic extension to $\mathbb{R}$:

$$-\varepsilon^2 \Delta \psi^{(j)} + m^2 \psi^{(j)} = \rho_{\boldsymbol{y}^{(j)}} \quad \text{in } \mathbb{R}. \tag{4.41}$$

Here we have defined the positions $\boldsymbol{y}^{(j)} = (y_k^{(j)})_{k \in \mathbb{Z}}$ of an infinite chain of equidistant atoms by

$$y_k^{(j)} = y_j + (k - j)(y_j - y_{j-1}) \quad \forall k \in \mathbb{Z}.$$

The Cauchy–Born approximation $\mathcal{E}^{\mathrm{cb}}(\boldsymbol{y})$ of the atomistic energy $\mathcal{E}(\boldsymbol{y})$ is then given by the sum over all cells

$$\mathcal{E}^{\mathrm{cb}}(\boldsymbol{y}) = \sum_{j=-N}^{N} \mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y}) = \frac{1}{2} \sum_{j=-N}^{N} \int_{Q_j} \rho_{\boldsymbol{y}} \psi^{(j)} \, \mathrm{d}x \,. \tag{4.42}$$

Whether $\mathcal{E}^{\mathrm{cb}}(\boldsymbol{y})$ is a good approximation of $\mathcal{E}(\boldsymbol{y})$ strongly depends on the regularity properties of $\boldsymbol{y}$. As we will see below, if $\boldsymbol{y}$ is smooth, i.e., the second difference $\boldsymbol{y}''$ is small, then $|\mathcal{E}^{\mathrm{cb}}(\boldsymbol{y}) - \mathcal{E}(\boldsymbol{y})|$ is small.

Let $\boldsymbol{u} \in \mathcal{U}$ be a test vector and $u \in \mathrm{S}_{\#}(\boldsymbol{y})$ an interpolant of $\boldsymbol{u}$ in the sense of (4.9). It follows as in Lemma 4.9 that the derivative of $\mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y})$ can be written in the form

$$D_{\boldsymbol{y}} \mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y}) \cdot \boldsymbol{u} = \frac{u_j - u_{j-1}}{y_j - y_{j-1}} \int_{Q_j} \sigma_{j,\boldsymbol{y}}^{\mathrm{cb}}(x) \, \mathrm{d}x \; = \int_{Q_j} \sigma_{j,\boldsymbol{y}}^{\mathrm{cb}}(x) \nabla u(x) \, \mathrm{d}x \,,$$

where the local continuum stress function $\sigma_{j,\boldsymbol{y}}^{\mathrm{cb}}$, in direct correspondence with (4.11), is

$$\begin{aligned} \sigma_{j,\boldsymbol{y}}^{\mathrm{cb}}(x) = {} & \tfrac{1}{2}\varepsilon^2 |\nabla \psi^{(j)}(x)|^2 - \tfrac{1}{2} m^2 \psi^{(j)}(x)^2 + \rho_{\boldsymbol{y}}(x) \psi^{(j)}(x) \\ & + \varepsilon \sum_{j=-N-1}^{N} \psi^{(j)}(x) \nabla \delta_\varepsilon(x - y_j)(x - y_j). \end{aligned} \tag{4.43}$$

Furthermore, we define the Cauchy–Born stress function $\sigma^{\mathrm{cb}}_{\boldsymbol{y}} : \Omega \to \mathbb{R}$ by

$$\sigma^{\mathrm{cb}}_{\boldsymbol{y}}(x) = \sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}(x) \quad \text{if} \quad x \in \Omega_j$$

for all $x \in \Omega$.

Remark 4.19. We note that our derivation of the Cauchy–Born energy is equivalent to the following procedure, see for example [15, 26]. The definition of the atomistic energy can be used to derive a strain energy density $W : \mathbb{R} \to \mathbb{R}$ (in the Lagrange picture):

$$W(A) = -\frac{A}{|\widehat{Q}|} \inf_{\widehat{\psi} \in \mathrm{H}^1_{\#}(\widehat{Q})} \left[\frac{1}{2} \int_{\widehat{Q}} \left(\frac{\varepsilon^2}{A^2} |\nabla \widehat{\psi}|^2 + m^2 \widehat{\psi}^2 \right) \mathrm{d}\xi - \int_Q \rho_{\boldsymbol{y}}(A\xi) \widehat{\psi}(\xi) \, \mathrm{d}\xi \right], \tag{4.44}$$

for some reference unit cell $\widehat{Q}$. This leads to the elastic energy $\int_{\widehat{\Omega}} W(\nabla y(X)) \, \mathrm{d}X$ for a continuous deformation $y : \widehat{\Omega} \to \Omega$. If one now uses a linear finite element discretization over the mesh given by the reference configuration $\widehat{X}$ and transforms this energy to Ω (that is, the Euler picture), one obtains exactly the energy (4.42). ◻

4.4.1 Consistency

Next, we turn to the consistency analysis of the Cauchy–Born approximation, for which we thoroughly analyze the modelling error incurred. From the previous sections we deduce that

$$\begin{aligned} \left| \left(D\mathcal{E}(\boldsymbol{y}) - D\mathcal{E}^{\mathrm{cb}}(\boldsymbol{y}) \right) \cdot \boldsymbol{u} \right| &\leq \int_\Omega \left| \sigma_{\boldsymbol{y}}(x) - \sigma^{\mathrm{cb}}_{\boldsymbol{y}}(x) \right| |\nabla u(x)| \, \mathrm{d}x \\ &= \sum_{j=-N}^{N} \int_{Q_j} \left| \sigma_{\boldsymbol{y}}(x) - \sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}(x) \right| |\nabla u(x)| \, \mathrm{d}x \,, \end{aligned}$$

where the stress functions $\sigma_{\boldsymbol{y}}$ and $\sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}$ are given by (4.11) and (4.43), respectively. The difference between $\sigma^{\mathrm{cb}}_{\boldsymbol{y}}$ and $\sigma_{\boldsymbol{y}}$ is that the fields $\psi^{(j)}$ entering $\sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}$ are calculated independently on every Q_j. To investigate the modelling error $\left| \sigma_{\boldsymbol{y}}(x) - \sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}(x) \right|$ incurred by going from the atomistic description to the Cauchy–Born approximation it is hence sufficient to analyze $|\phi - \psi^{(j)}|$ and $|\nabla \phi - \nabla \psi^{(j)}|$ in Q_j for every $j \in \{-N, \dots, N\}$.

First, we provide a technical lemma.

Lemma 4.20. *Let $\boldsymbol{y} \in \ell^\infty(\mathbb{Z})$ and define $\boldsymbol{y}^{(j)} = (y^{(j)}_k)_{k \in \mathbb{Z}}$ by $y^{(j)}_k = y_j + \varepsilon y'_j(k-j)$ for all $k \in \mathbb{Z}$. Then, for $n > j$:*

$$\left| y_n - y^{(j)}_n \right| \leq (n-j)^2 \varepsilon^2 \|\boldsymbol{y}''\|_{\ell^\infty([j,n-1])}.$$

If $n < j-1$, then

$$\left| y_n - y^{(j)}_n \right| \leq (j-1-n)^2 \varepsilon^2 \|\boldsymbol{y}''\|_{\ell^\infty([n+1,j-1])}.$$

Proof. Since $y_{j-1} = y_{j-1}^{(j)}$ and $y_j = y_j^{(j)}$ we get for $n > j$:

$$y_n - y_n^{(j)} = \varepsilon \sum_{k=j+1}^{n} \big(y_k' - (y_k^{(j)})'\big) = \varepsilon^2 \sum_{k=j+1}^{n} \sum_{l=j}^{k-1} \big(y_l'' - (y_l^{(j)})''\big) = \varepsilon^2 \sum_{k=j+1}^{n} \sum_{l=j}^{k-1} y_l'',$$

where have used that $(\boldsymbol{y}^{(j)})'$ is constant. Changing the summation order we get

$$y_n - y_n^{(j)} = \varepsilon^2 \sum_{l=j}^{n-1} \sum_{k=l+1}^{n} y_l'' = \varepsilon^2 \sum_{l=j}^{n-1} (n-l) y_l''.$$

So, we deduce that

$$\big|y_n - y_n^{(j)}\big| \le (n-j)^2 \varepsilon^2 \|\boldsymbol{y}''\|_{\ell^\infty([j,n-1])}.$$

If $n < j-1$, we obtain with similar steps

$$y_n - y_n^{(j)} = -\varepsilon \sum_{k=n+1}^{j-1} (y_k' - (y_k^{(j)})') = \varepsilon^2 \sum_{k=n+1}^{j-1} \sum_{l=k}^{j-1} y_l'',$$

which implies that

$$\big|y_n - y_n^{(j)}\big| \le (j-n-1)^2 \varepsilon^2 \|\boldsymbol{y}''\|_{\ell^\infty([n+1,j-1])},$$

as desired. ◻

The next result addresses the errors $|\phi(x) - \psi^{(j)}(x)|$, $|\nabla\phi(x) - \nabla\psi^{(j)}(x)|$ for x in the cell Q_j. As anticipated by Lemma 4.20 it depends on the second difference $\boldsymbol{y}''$.

Lemma 4.21. *Let $\boldsymbol{y} \in \mathcal{Y}$ satisfy $\min \boldsymbol{y}' > \varsigma_0$. Let $\phi \in \mathrm{H}^1_{\#}(\Omega)$ satisfy (4.5) and $\psi^{(j)} \in \mathrm{H}^1_{\#}(Q_j)$ satisfy (4.41), respectively. Then,*

$$\big\|\phi - \psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)} \le \mu\varepsilon \sum_{n=1}^{\infty} \|\boldsymbol{y}''\|_{\ell^\infty([j-n,j+n-1])} n^2 \mathrm{e}^{-mn \min \boldsymbol{y}'}, \quad \textit{and}$$
$$\big\|\varepsilon\nabla\phi - \varepsilon\nabla\psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)} \le m\mu\varepsilon \sum_{n=1}^{\infty} \|\boldsymbol{y}''\|_{\ell^\infty([j-n,j+n-1])} n^2 \mathrm{e}^{-mn \min \boldsymbol{y}'}.$$

Proof. From Proposition 4.4 we immediately deduce that, for all $x \in Q_j$,

$$\begin{aligned} \phi(x) &= \frac{1}{2m} \int_{\mathbb{R}} \sum_{k\in\mathbb{Z}} \delta_\varepsilon(z - y_k) \mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} \,\mathrm{d}z, \\ \psi^{(j)}(x) &= \frac{1}{2m} \int_{\mathbb{R}} \sum_{k\in\mathbb{Z}} \delta_\varepsilon(z - y_k^{(j)}) \mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} \,\mathrm{d}z. \end{aligned} \tag{4.45}$$

Since $y_j^{(j)} = y_j$ and $y_{j-1}^{(j)} = y_{j-1}$, the respective terms in the sums cancel. Hence, we get for $x \in Q_j$:

$$\phi(x) - \psi^{(j)}(x) = \frac{1}{2m} \sum_{\substack{k\in\mathbb{Z} \\ k\neq j-1,j}} \int_{\mathbb{R}} \big(\delta_\varepsilon(z - y_k) - \delta_\varepsilon(z - y_k^{(j)})\big) \mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} \,\mathrm{d}z.$$

We now derive bounds on the individual terms in the sum. Note that (4.18) simplifies the following calculations but due to the smoothness of the Green's function similar bounds can be obtained without it.

Step 1. Let $k > j$. Then we have $|x - z| = z - x$ for all $z \in \operatorname{supp}\delta_\varepsilon(\cdot - y_k)$ and all $z \in \operatorname{supp}\delta_\varepsilon(\cdot - y_k^{(j)})$. Thus, with (4.18),

$$\frac{1}{2m}\int_{\mathbb{R}} \big(\delta_\varepsilon(z - y_k) - \delta_\varepsilon(z - y_k^{(j)})\big)\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z = \frac{\mu}{2m}\big(\mathrm{e}^{-\frac{m}{\varepsilon}(y_k - x)} - \mathrm{e}^{-\frac{m}{\varepsilon}(y_k^{(j)} - x)}\big). \tag{4.46}$$

If $y_k^{(j)} \geq y_k$, then

$$\begin{aligned}\left|\frac{1}{2m}\int_{\mathbb{R}} \big(\delta_\varepsilon(z - y_k) - \delta_\varepsilon(z - y_k^{(j)})\big)\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z\right| &\leq \frac{\mu}{2m}\mathrm{e}^{-\frac{m}{\varepsilon}(y_k - x)}\big(1 - \mathrm{e}^{-\frac{m}{\varepsilon}(y_k^{(j)} - y_k)}\big)\\ &\leq \frac{\mu}{2m}\mathrm{e}^{-\frac{m}{\varepsilon}(y_k - x)}\frac{m}{\varepsilon}(y_k^{(j)} - y_k).\end{aligned}$$

Using $(y_k - x) \geq (k - j)\varepsilon \min \boldsymbol{y}'$ for all $x \in Q_j$ and applying Lemma 4.20 leads to

$$\frac{\mu}{2\varepsilon}\mathrm{e}^{-\frac{m}{\varepsilon}(y_k - x)|}\big|y_k - y_k^{(j)}\big| \leq \frac{\mu\varepsilon}{2}\|\boldsymbol{y}''\|_{\ell^\infty([j,k-1])}(k - j)^2\mathrm{e}^{-(k-j)m\min \boldsymbol{y}'}.$$

The same bound on (4.46) can be obtained if $y_k^{(j)} \leq y_k$.

Step 2. For any $k < j - 1$ we can use the same techniques to obtain that

$$\begin{aligned}&\left|\frac{1}{2m}\int_{\mathbb{R}} \big(\delta_\varepsilon(z - y_k) - \delta_\varepsilon(z - y_k^{(j)})\big)\ \mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z\right|\\ &\qquad\leq \frac{\mu\varepsilon}{2}\|\boldsymbol{y}''\|_{\ell^\infty([k+1,j-1])}(j - k - 1)^2\mathrm{e}^{-(j-k-1)m\min \boldsymbol{y}'}.\end{aligned}$$

Step 3. Summing over all $k \in \mathbb{Z}\backslash\{j - 1, j\}$ we deduce that

$$\big|\phi(x) - \psi^{(j)}(x)\big| \leq \mu\varepsilon\sum_{n=1}^{\infty}\|\boldsymbol{y}''\|_{\ell^\infty([j-n,j+n-1])}n^2\mathrm{e}^{-mn\min \boldsymbol{y}'}.$$

Step 4. The derivatives of ϕ, $\psi^{(j)}$ in $Q_j = [y_{j-1}, y_j]$ are given by

$$\begin{aligned}\nabla\phi(x) &= \frac{1}{2m}\sum_{k\in\mathbb{Z}}\int_{\mathbb{R}}\nabla\delta_\varepsilon(z - y_k)\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z,\\ \nabla\psi^{(j)}(x) &= \frac{1}{2m}\sum_{k\in\mathbb{Z}}\int_{\mathbb{R}}\nabla\delta_\varepsilon(z - y_k^{(j)})\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z.\end{aligned}$$

For $k > j$, respectively, $k < j - 1$ the exponential factor can in both cases be replaced with $\mathrm{e}^{-\frac{m}{\varepsilon}(z-x)}$, respectively, $\mathrm{e}^{-\frac{m}{\varepsilon}(x-z)}$. Integration by parts leads to the same situation as above with one power of ε less and a factor of m more. ◻

The next result is only a slight modification of the previous one. The idea is to only treat the modelling error from a finite number of neighbouring atoms explicitly and to find an upper bound on the contribution from atoms that are further away from Q_j.

Lemma 4.22. *Let $\phi \in \mathrm{H}^1_\#(\Omega)$ satisfy (4.5) and $\psi^{(j)} \in \mathrm{H}^1_\#(Q_j)$ satisfy (4.41), respectively. Then, for any $M \in \mathbb{N}$:*

$$\big\|\phi - \psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)} \leq \varepsilon\mu \sum_{n=1}^{M-1} \|\boldsymbol{y}''\|_{\ell^\infty(j-n,j+n-1)} n^2 \mathrm{e}^{-mn\min \boldsymbol{y}'} + \frac{2\mu}{m} \frac{\mathrm{e}^{-mM\min \boldsymbol{y}'}}{1-\mathrm{e}^{-m\min \boldsymbol{y}'}}$$

and

$$\varepsilon\big\|\nabla\phi - \nabla\psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)} \leq \varepsilon m\mu \sum_{n=1}^{M-1} \|\boldsymbol{y}''\|_{\ell^\infty(j-n,j+n-1)} n^2 \mathrm{e}^{-mn\min \boldsymbol{y}'} + 2\mu \frac{\mathrm{e}^{-mM\min \boldsymbol{y}'}}{1-\mathrm{e}^{-m\min \boldsymbol{y}'}}.$$

Proof. The first error part is derived as in the previous lemma. We then look at the contribution to $\phi(x)$ from the atoms in y_k for $k \geq j + M$:

$$\begin{aligned}
\frac{1}{2m}\int_{\mathbb{R}} \delta_\varepsilon(z-y_k)\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z &= \frac{1}{2m}\int_{\mathbb{R}} \delta_\varepsilon(z-y_k)\mathrm{e}^{-\frac{m}{\varepsilon}(z-y_k)}\mathrm{e}^{-\frac{m}{\varepsilon}(y_k-x)}\,\mathrm{d}z \\
&= \frac{\mu}{2m}\mathrm{e}^{-\frac{m}{\varepsilon}(y_k-x)} \\
&\leq \frac{\mu}{2m}\mathrm{e}^{-m(k-j)\min \boldsymbol{y}'} \\
&\leq \frac{\mu}{2m}\mathrm{e}^{-mM\min \boldsymbol{y}'}\mathrm{e}^{-m(k-M-j)\min \boldsymbol{y}'}.
\end{aligned}$$

Summing over $k \geq M + j$ we hence get

$$\frac{1}{2m}\sum_{k=j+M}^{\infty}\int_{\mathbb{R}} \delta_\varepsilon(z-y_k)\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z \leq \frac{\mu}{2m}\frac{\mathrm{e}^{-mM\min \boldsymbol{y}'}}{1-\mathrm{e}^{-m\min \boldsymbol{y}'}}.$$

The same bound can be obtained for the sum over all $k \leq j - M - 1$. Similarly we deal with the contributions of these k to $\psi^{(j)}$. The triangle inequality then shows the bound on $\big\|\phi - \psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)}$. The proof for the bound on $\varepsilon\big\|\nabla\phi - \nabla\psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)}$ works analogously. ◻

The quadratic nature of the model (4.4) results in stress functions $\sigma_{\boldsymbol{y}}$ and $\sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}$ that are quadratic in the fields ϕ and $\psi^{(j)}$, respectively. Together with L^∞-bounds on ϕ and $\psi^{(j)}$ this allows us to easily bound the modelling error $\big\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}\big\|_{\mathrm{L}^\infty(Q_j)}$ in terms of $\big\|\phi - \psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)}$ and $\big\|\nabla\phi - \nabla\psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)}$.

Lemma 4.23. *Let $\sigma_{\boldsymbol{y}}$ and $\sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}$ be given by (4.11), respectively, (4.43). Then, for all $j \in \{-N,\dots,N\}$*

$$\begin{aligned}
\big\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}\big\|_{\mathrm{L}^\infty(Q_j)} \leq{}& K_1(m\min \boldsymbol{y}')\,\varepsilon\,\big\|\nabla\phi - \nabla\psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)} \\
&+ \big(m^2 K_0(m\min \boldsymbol{y}') + C\big)\big\|\phi - \psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)},
\end{aligned}$$

where the constant C only depends on δ_1.

Proof. From the definitions of the atomistic and continuum stress function we deduce that

$$\begin{aligned}\sigma_{\boldsymbol{y}}(x) - \sigma^{\mathrm{cb}}_{j,\boldsymbol{y}}(x) = &-\tfrac{1}{2}\big(\varepsilon\nabla\phi(x) - \varepsilon\nabla\psi^{(j)}(x)\big)\big(\varepsilon\nabla\phi(x) + \varepsilon\nabla\psi^{(j)}(x)\big)\\ &+\tfrac{1}{2}m^2\big(\phi(x) - \psi^{(j)}(x)\big)\big(\phi(x) + \psi^{(j)}(x)\big)\\ &-\rho_{\boldsymbol{y}}(x)\big(\phi(x) - \psi^{(j)}(x)\big)\\ &-\big(\phi(x) - \psi^{(j)}(x)\big)\sum_{i=j-1}^{j}\varepsilon\nabla\delta_\varepsilon(x-y_i)(x-y_i)\end{aligned}$$

for all $x \in Q_j$. With $\delta_\varepsilon(x) = \varepsilon^{-1}\delta_1(x/\varepsilon)$, the L^∞-bound on ϕ from Lemma 4.5, and a similar bound for $\psi^{(j)}$ we get

$$\begin{aligned}\tfrac{1}{2}\big|\varepsilon\nabla\phi(x) + \varepsilon\nabla\psi^{(j)}(x)\big| &\le K_1(m\min\boldsymbol{y}'),\\ \tfrac{m^2}{2}\big|\phi(x) + \psi^{(j)}(x)\big| &\le m^2K_0(m\min\boldsymbol{y}'),\\ \|\rho_{\boldsymbol{y}}\|_{\mathrm{L}^\infty} &\le \|\delta_1\|_{\mathrm{L}^\infty},\\ \big|\varepsilon\nabla\delta_\varepsilon(x-y_i)(x-y_i)\big| &\le \|\nabla\delta_1\mathrm{id}\|_{\mathrm{L}^\infty},\end{aligned}$$

which implies the stated result. □

4.4.2 Stability

Besides consistency, the second crucial property of an approximation to a given model is stability. In the present case, we need to determine under which conditions $D^2\mathcal{E}^{\mathrm{cb}}(\boldsymbol{y})$ is positive definite.

Lemma 4.24. *Let $\boldsymbol{y} \in \mathcal{Y}$ satisfy $\boldsymbol{y}' > \varsigma_0$. Then, for all $j \in \{-N,\dots,N\}$,*

$$D^2\mathcal{E}^{\mathrm{cb}}_j(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \ge \frac{m^2\mu^2}{2}\,\mathrm{e}^{-m\max\boldsymbol{y}'}\varepsilon|u_j'|^2 \quad \forall \boldsymbol{u} \in \mathcal{U}.$$

Proof. We first recall that $\mathcal{E}^{\mathrm{cb}}_j(\boldsymbol{y}) = \frac{1}{2}\int_{Q_j}\rho_{\boldsymbol{y}}\psi^{(j)}\,\mathrm{d}x$ because $\psi^{(j)}$ is a minimizer of (4.40). Extending $\psi^{(j)}$ $|Q_j|$-periodically to $\mathbb{R}$ and using the symmetry of the cell problem, we can rewrite this as

$$\mathcal{E}^{\mathrm{cb}}_j(\boldsymbol{y}) = \frac{\varepsilon}{2}\int_{\mathbb{R}}\delta_\varepsilon(x-y_j)\psi^{(j)}(x)\,\mathrm{d}x\,.$$

We now insert the explicit formula (4.45) for $\psi^{(j)}(x)$ and apply (4.18) to get

$$\begin{aligned}\mathcal{E}^{\mathrm{cb}}_j(\boldsymbol{y}) &= \frac{\varepsilon}{4m}\sum_{k\in\mathbb{Z}}\int_{\mathbb{R}}\int_{\mathbb{R}}\delta_\varepsilon(x-y_j)\delta_\varepsilon(z-y_k^{(j)})\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|}\,\mathrm{d}z\,\mathrm{d}x\\ &= \frac{\mu^2\varepsilon}{4m}\sum_{\substack{k\in\mathbb{Z}\\k\neq j}}\mathrm{e}^{-\frac{m}{\varepsilon}|y_j-y_k^{(j)}|} + \mathcal{E}_{\mathrm{self}}\\ &= \frac{\mu^2\varepsilon}{2m}\sum_{\nu=1}^{\infty}\mathrm{e}^{-m\nu y_j'} + \mathcal{E}_{\mathrm{self}},\end{aligned}$$

where the constant $\mathcal{E}_{\text{self}}$ coming from $k = j$ in the sum represents the self-energies of the atoms in the cell Q_j. Here we have also used that $\left|y_k^{(j)} - y_j\right| = |k-j|y'_j$ for all $k \in \mathbb{Z}$. Differentiating twice leads to

$$\begin{aligned} D^2\mathcal{E}_j^{\text{cb}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] &= \frac{m\mu^2}{2}\varepsilon\sum_{\nu=1}^{\infty}\nu^2\mathrm{e}^{-\nu m y'_j}|u'_j|^2 \\ &\geq \frac{m\mu^2}{2}\varepsilon|u'_j|^2\sum_{\nu=1}^{\infty}\nu^2\mathrm{e}^{-\nu m\max \boldsymbol{y}'} \\ &\geq \frac{m\mu^2}{2}\mathrm{e}^{-m\max \boldsymbol{y}'}\ \varepsilon|u'_j|^2. \end{aligned}$$

In the last step we have only kept the term for $\nu = 1$, which represents the nearest neighbour interactions. □

Finally, we prove a Lipschitz bound for the second derivatives $D^2\mathcal{E}_j^{\text{cb}}$.

Lemma 4.25. *Let $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{Y}$. If $\min \boldsymbol{y}'_1 \geq s_0 > \varsigma_0$ and $\min \boldsymbol{y}'_2 \geq s_0$, then for all $j \in \{-N, \dots, N\}$*

$$\left|\left(D^2\mathcal{E}_j^{\text{cb}}(\boldsymbol{y}_1) - D^2\mathcal{E}_j^{\text{cb}}(\boldsymbol{y}_2)\right)\cdot[\boldsymbol{u},\boldsymbol{u}]\right| \leq \frac{m^2\mu^2}{2}\left(\varepsilon|u'_j|^2\right)\left\|\boldsymbol{y}'_1 - \boldsymbol{y}'_2\right\|_{\ell^\infty}\sum_{\nu=1}^{\infty}\nu^3\mathrm{e}^{-\nu m s_0} \quad \forall \boldsymbol{u} \in \mathcal{U}.$$

Proof. The derivative $D^2\mathcal{E}_j^{\text{cb}}$ was calculated in the previous proof. Hence,

$$\left(D^2\mathcal{E}_j^{\text{cb}}(\boldsymbol{y}_1) - D^2\mathcal{E}_j^{\text{cb}}(\boldsymbol{y}_2)\right)\cdot[\boldsymbol{u},\boldsymbol{u}] = \frac{m\mu^2}{2}\varepsilon\sum_{\nu=1}^{\infty}\nu^2\left(\mathrm{e}^{-\nu m y'_{1,j}} - \mathrm{e}^{-\nu m y'_{2,j}}\right)|u'_j|^2.$$

Since by assumption $y'_{1,j} \geq s_0$ and $y'_{2,j} \geq s_0$, we get with the Mean Value Theorem

$$\left|\mathrm{e}^{-\nu m y'_{1,j}} - \mathrm{e}^{-\nu m y'_{2,j}}\right| \leq \nu m\mathrm{e}^{-\nu m s_0}|y'_{1,j} - y'_{2,j}|.$$

Inserting this in the previous equation concludes the proof. □

4.5 Quasicontinuum Coupling

The computation of the original atomistic energy $\mathcal{E}(\boldsymbol{y})$ involves the solution of the optimization problem (4.4) posed in the whole of $\Omega = (y_{-N-1}, y_N)$. Our goal is the construction of computationally cheaper, approximate energies $\mathcal{E}^{\text{qc}}(\boldsymbol{y})$ such that $\mathcal{E}(\boldsymbol{y}) \approx \mathcal{E}^{\text{qc}}(\boldsymbol{y})$ for all relevant $\boldsymbol{y}$ and minimizers $\bar{\boldsymbol{y}}^{\text{qc}} \in \mathcal{Y}$ of

$$E_{\boldsymbol{f}}^{\text{qc}}(\boldsymbol{y}) = \mathcal{E}^{\text{qc}}(\boldsymbol{y}) + (\boldsymbol{f}, \boldsymbol{y})_\varepsilon,$$

are good approximations of minimizers $\bar{\boldsymbol{y}}$ of the energy $E_{\boldsymbol{f}}$ from (4.6).

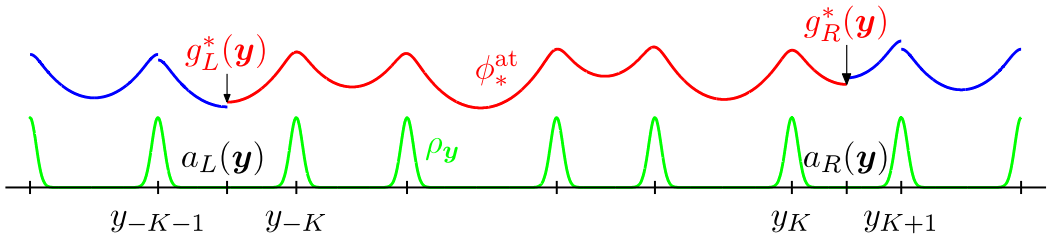

Figure 4.4: An illustration of the first QC method. In $\Omega^{\rm at} = (a_L(\boldsymbol{y}), a_R(\boldsymbol{y}))$ the atomistic problem is solved with the Dirichlet boundary conditions $g^*(\boldsymbol{y})$. Outside $\Omega^{\rm at}$ the Cauchy–Born approximation is used in all cells Q_j.

Following the basic philosophy of the QC method we need to approximate $\mathcal{E}(\boldsymbol{y})$ using the continuum model where $\boldsymbol{y}$ is smooth and a version of the atomistic model where $\boldsymbol{y}$ is nonsmooth. In the following we will implicitly assume that the configurations $\boldsymbol{y} \in \mathcal{Y}$ under consideration are smooth except in the segment $y_{-K}, \dots, y_K$ for some $K < N$. We divide Ω into an atomistic subdomain $\Omega^{\rm at}$ such that $y_j \in \Omega^{\rm at}$ for all $j \in \{-K, \dots, K\}$ and the continuum domain $\Omega^{\rm cb} = \Omega \backslash \Omega^{\rm at}$. In $\Omega^{\rm cb}$ we will use the Cauchy–Born approximation on a cell-by-cell basis. On the other hand, in $\Omega^{\rm at}$ we will use the atomistic model with Dirichlet boundary conditions as discussed in Section 4.3.

This basic outset gives rise to a variety of possibilities including the precise choice of $\partial\Omega^{\rm at}$ and the boundary conditions imposed on the atomistic subproblem. Both will in general depend on the configuration $\boldsymbol{y}$. Our main objective for $\mathcal{E}^{\rm qc}$ is the existence of a weak formulation in the sense that

$$D\mathcal{E}^{\rm qc}(\boldsymbol{y}) \cdot \boldsymbol{u} = \int_\Omega \sigma_{\boldsymbol{y}}^{\rm qc}(x) \nabla u(x) \, {\rm d}x \,,$$

where $u \in {\rm S}_\#(\boldsymbol{y})$ is a piecewise linear interpolant of $\boldsymbol{u} \in \mathcal{U}$ and $\sigma_{\boldsymbol{y}}^{\rm qc}$ is a stress function to be determined. If this weak formulation can be obtained, the consistency analysis reduces to error estimates on fields.

In this chapter we focus on the consistency properties of QC methods at the interface between the atomistic and the continuum region. We therefore do not take into account coarse-graining in the continuum region. An analysis of the coarse-graining error for classical QC methods can be found in [99].

Throughout this section $\phi \in {\rm H}^1_\#(\Omega)$ denotes the solution of the original minimization problem (4.4) for a given configuration $\boldsymbol{y} \in \mathcal{Y}$.

4.5.1 A Method With Optimal Boundary Conditions

For the first QC method we place the boundary a of the atomistic subproblem halfway between the interface atoms, that is $a = a(\boldsymbol{y}) = [a_L(\boldsymbol{y})\ a_R(\boldsymbol{y})]^T$ with[1]

$$a_L(\boldsymbol{y}) = \frac{y_{-K-1} + y_{-K}}{2}, \qquad a_R(\boldsymbol{y}) = \frac{y_K + y_{K+1}}{2}.$$

Let $\Omega^{\rm at} = (a_L(\boldsymbol{y}), a_R(\boldsymbol{y}))$ and $\Omega^{\rm cb} = \Omega \backslash \Omega^{\rm at}$. We write the QC energy $\mathcal{E}^{\rm qc}(\boldsymbol{y})$ as the sum of a continuum and an atomistic part

$$\mathcal{E}^{\rm qc}(\boldsymbol{y}) = \mathcal{E}^{\rm cb}_*(\boldsymbol{y}) + \mathcal{E}^{\rm at}_*(\boldsymbol{y}), \tag{4.47}$$

which are introduced below.

We note that because of the definition of $a(\boldsymbol{y})$ there are two half cells, $(y_{-K-1}, a_L(\boldsymbol{y}))$ and $(a_R(\boldsymbol{y}), y_{K+1})$, in the continuum region $\Omega^{\rm cb}$ (see Figure 4.4). Since the cell problems on all Q_j are symmetric, the Cauchy–Born energies of these half cells are given by $\frac{1}{2}\mathcal{E}^{\rm cb}_{-K}(\boldsymbol{y})$ and $\frac{1}{2}\mathcal{E}^{\rm cb}_{K+1}(\boldsymbol{y})$. The continuum part of the energy $\mathcal{E}^{\rm qc}$ is then defined by

$$\mathcal{E}^{\rm cb}_*(\boldsymbol{y}) = \sum_{j=-N+1}^{-K-1} \mathcal{E}^{\rm cb}_j(\boldsymbol{y}) + \tfrac{1}{2}\mathcal{E}^{\rm cb}_{-K}(\boldsymbol{y}) + \tfrac{1}{2}\mathcal{E}^{\rm cb}_{K+1}(\boldsymbol{y}) + \sum_{j=K+2}^{N} \mathcal{E}^{\rm cb}_j(\boldsymbol{y}). \tag{4.48}$$

The coordinates of the atoms in the atomistic region $\Omega^{\rm at}$ are represented by

$$\boldsymbol{y}_{\rm at} = (y_{-K}, \ldots, y_K).$$

For the definition of $\mathcal{E}^{\rm at}_*(\boldsymbol{y})$ we consider the minimization problem (4.19) on the atomistic domain $\Omega^{\rm at}$ subject to the Dirichlet boundary conditions $g^*(\boldsymbol{y}) = [g^*_L(\boldsymbol{y})\ g^*_R(\boldsymbol{y})]^T$. In correspondence with Remark 4.12 and Section 4.3.4 they are given by

$$g^*_L(\boldsymbol{y}) = \frac{1}{1-\tau}\frac{\gamma_L(\boldsymbol{y}) + \tau\gamma_R(\boldsymbol{y})}{1+\tau}, \qquad g^*_R(\boldsymbol{y}) = \frac{1}{1-\tau}\frac{\tau\gamma_L(\boldsymbol{y}) + \gamma_R(\boldsymbol{y})}{1+\tau},$$

where $\tau = \mathrm{e}^{-\frac{m}{\varepsilon}\Delta a(\boldsymbol{y})}$, and (see also (4.30))

$$\gamma_L(\boldsymbol{y}) = 2\int_{a_L(\boldsymbol{y})}^{a_R(\boldsymbol{y})} \rho_{\boldsymbol{y}}(x) G_\varepsilon(x - a_L)\,\mathrm{d}x\,, \qquad \gamma_R(\boldsymbol{y}) = 2\int_{a_L(\boldsymbol{y})}^{a_R(\boldsymbol{y})} \rho_{\boldsymbol{y}}(x) G_\varepsilon(a_R - x)\,\mathrm{d}x\,.$$

The energy contribution from the atomistic subproblem is thus given by

$$\begin{aligned}\mathcal{E}^{\rm at}_*(\boldsymbol{y}) &= \mathcal{E}_{a(\boldsymbol{y}),g^*(\boldsymbol{y})}(\boldsymbol{y}_{\rm at}) \\ &= -\inf\Big\{ I_{a(\boldsymbol{y})}(\varphi, \boldsymbol{y}_{\rm at}) : \quad \varphi \in \mathrm{H}^1(\Omega^{\rm at}), \quad \varphi|_{\partial\Omega^{\rm at}} = g^*(\boldsymbol{y}) \Big\},\end{aligned}$$

[1] The analysis presented in this section immediately carries over to the choice $a_L(\boldsymbol{y}) = y_{-K}$ and $a_R(\boldsymbol{y}) = y_K$.

where $I_{a(\boldsymbol{y})}$ is defined as in (4.20). We denote the solution of this optimization problem by $\phi^*_{\mathrm{at}} \in \mathrm{H}^1(\Omega^{\mathrm{at}})$. It satisfies the boundary value problem

$$-\varepsilon^2 \Delta \phi^*_{\mathrm{at}} + m^2 \phi^*_{\mathrm{at}} = \rho_{\boldsymbol{y}} \quad \text{in } \Omega^{\mathrm{at}},$$
$$\phi^*_{\mathrm{at}}|_{\partial\Omega^{\mathrm{at}}} = g^*(\boldsymbol{y}).$$

From a computational point of view $g^*(\boldsymbol{y})$ is also a convenient choice since this is equivalent to homogeneous Neumann boundary conditions. In Section 4.3.4 we deduced a clear interpretation of the effect of this choice of boundary data: besides the interaction among themselves, the atoms in Ω^{at} interact with mirror atoms outside Ω^{at}.

We stress that $g^*(\boldsymbol{y})$ and hence $\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})$ only depend on the components $y_{-K-1}, \dots, y_{K+1}$. Only the four components y_{-K-1}, y_{-K}, y_K, and y_{K+1} enter both $\mathcal{E}^{\mathrm{at}}_*$ and $\mathcal{E}^{\mathrm{cb}}_*$.

In analogy to (4.6) we search for minimizers of the total potential energy

$$E^{\mathrm{qc}}_{\boldsymbol{f}}(\boldsymbol{y}) = \mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) + (\boldsymbol{f}, \boldsymbol{y})_\varepsilon \tag{4.49}$$

in $\mathcal{Y}$, where $\boldsymbol{f} \in \mathcal{U}^{-1,2}$ represents an external force. A minimizer $\bar{\boldsymbol{y}}^{\mathrm{qc}}$ will satisfy the Euler–Lagrange equation

$$DE^{\mathrm{qc}}_{\boldsymbol{f}}(\boldsymbol{y}) = D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) + \boldsymbol{f} = 0 \quad \in \mathcal{U}^{-1,2}.$$

Throughout the remainder of this chapter we assume that the atomistic domain Ω^{at} is large compared with ε:

$$a_R(\boldsymbol{y}) - a_L(\boldsymbol{y}) \gg \varepsilon \quad \text{such that} \quad \tau = \mathrm{e}^{-\frac{m}{\varepsilon}\Delta a(\boldsymbol{y})} \approx 0.$$

To keep the formulas slightly more compact we therefore do not keep track of the τ-dependent terms arising from the atomistic domain explicitly but include an $\mathcal{O}(\tau)$-term where necessary.

4.5.1.1 Consistency

In order to study the consistency properties of the QC energy $\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})$ from (4.47) we first need to calculate its derivative. Having established weak formulations for the derivatives of $\mathcal{E}$, $\mathcal{E}^{\mathrm{cb}}$, as well as $\mathcal{E}_{a,g}$, we will prove that the Quasicontinuum energy $\mathcal{E}^{\mathrm{qc}}$ admits a similar reformulation of $D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) \cdot \boldsymbol{u}$. For this we have to take into account that both the boundary of the atomistic domain Ω^{at} and the boundary conditions depend on $\boldsymbol{y}$. The necessary preparations were carried out in Section 4.3.

Lemma 4.26. *Let $\boldsymbol{y} \in \mathcal{Y}$ satisfy $\min \boldsymbol{y}' > \varsigma_0$. Furthermore, let $\boldsymbol{u} \in \mathcal{U}$ be a test vector and $u \in \mathrm{S}_{\#}(\boldsymbol{y})$ an interpolant of $\boldsymbol{u}$ in the sense of* (4.9)*. Then,*

$$D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) \cdot \boldsymbol{u} = \int_\Omega \sigma^{\mathrm{qc}}_{\boldsymbol{y}}(x) \nabla u(x) \, \mathrm{d}x \,, \tag{4.50}$$

where

$$\sigma^{\rm qc}_{\boldsymbol{y}}(x) = \begin{cases} \sigma^{\rm cb}_{\boldsymbol{y}}(x) & \text{if } x \in \Omega^{\rm cb}, \\ \sigma^{\rm at}_{\boldsymbol{y},*}(x) & \text{if } x \in \Omega^{\rm at}, \end{cases}$$

and $\sigma^{\rm at}_{\boldsymbol{y},*}(x)$ *is given by* (4.11) *with* $\phi = \phi^*_{\rm at}$.

Proof. *Continuum Contribution.* From Section 4.4 we already have the equality

$$D\mathcal{E}^{\rm cb}_j(\boldsymbol{y})\cdot\boldsymbol{u} = \int_{Q_j} \sigma^{\rm cb}_{\boldsymbol{y},j}(x)\nabla u(x)\,{\rm d}x\,,$$

$j \in \{-N,\dots,-K-1\}\cup\{K+2,\dots,N\}$. For the contribution $\frac{1}{2}\mathcal{E}^{\rm cb}_{-K}(\boldsymbol{y})$ from the half cell $(y_{-K-1}, a_L(\boldsymbol{y}))$ we make use of the symmetry of the cell problems. Since $\nabla u|_{Q_{-K}}$ is constant, $a_L(\boldsymbol{y})$ is the midpoint of $Q_{-K} = (y_{-K-1}, y_{-K})$, and $\sigma^{\rm cb}_{\boldsymbol{y},-K}$ is symmetric in Q_{-K}, we deduce that

$$\tfrac{1}{2}D\mathcal{E}^{\rm cb}_{-K}(\boldsymbol{y})\cdot\boldsymbol{u} = \frac{1}{2}\int_{Q_{-K}} \sigma^{\rm cb}_{\boldsymbol{y},-K}(x)\nabla u(x)\,{\rm d}x = \int_{y_{-K-1}}^{a_L(\boldsymbol{y})} \sigma^{\rm cb}_{\boldsymbol{y},-K}(x)\nabla u(x)\,{\rm d}x\,.$$

Analogously we treat $\frac{1}{2}\mathcal{E}^{\rm cb}_{K+1}(\boldsymbol{y})$. Hence,

$$D\mathcal{E}^{\rm cb}_*(\boldsymbol{y})\cdot\boldsymbol{u} = \int_{\Omega^{\rm cb}} \sigma^{\rm cb}_{\boldsymbol{y}}(x)\nabla u(x)\,{\rm d}x$$

where $\sigma^{\rm cb}_{\boldsymbol{y}}(x) = \sigma^{\rm cb}_{\boldsymbol{y},j}(x)$ if $x \in Q_j$.

Atomistic Contribution. To calculate the derivative $D\mathcal{E}^{\rm at}_*(\boldsymbol{y})$ we use the chain rule and the derivatives that were provided in Section 4.3. Applying Proposition 4.8 (with $h_L = (u_{-K-1} + u_{-K})/2$, $h_R = (u_K + u_{K+1})/2$ because of $D_{\boldsymbol{y}}a(\boldsymbol{y})\cdot\boldsymbol{u} = a(\boldsymbol{u})$), we get

$$\begin{aligned} D\mathcal{E}^{\rm at}_*(\boldsymbol{y})\cdot\boldsymbol{u} &= D_{\boldsymbol{y}}\mathcal{E}_{a(\boldsymbol{y}),g^*(\boldsymbol{y})}(\boldsymbol{y}_{\rm at})\cdot\boldsymbol{u}_{\rm at} + D_a\mathcal{E}_{a(\boldsymbol{y}),g^*(\boldsymbol{y})}(\boldsymbol{y}_{\rm at})\cdot D_{\boldsymbol{y}}a(\boldsymbol{y})\cdot\boldsymbol{u} \\ &= \int_{\Omega^{\rm at}} \sigma^{\rm at}_{\boldsymbol{y},*}(x)\nabla u(x)\,{\rm d}x\,, \end{aligned} \tag{4.51}$$

where the stress $\sigma^{\rm at}_{\boldsymbol{y},*}$ is given by (4.11) with $\phi = \phi^*_{\rm at}$ and $\boldsymbol{u}_{\rm at} = (u_{-K},\dots,u_K) \in \mathbb{R}^{2K+1}$ is the section of $\boldsymbol{u}$ corresponding to the atoms in the atomistic region. Note that the choice of boundary conditions implies $D_g\mathcal{E}_{a(\boldsymbol{y}),g^*(\boldsymbol{y})}(\boldsymbol{y}_{\rm at}) = 0$ as seen in Remark 4.12. ◻

We point out that the weak form (4.50) of the derivative $D\mathcal{E}^{\rm qc}$ already implies that there are no ghost forces for homogeneous deformations $\boldsymbol{y}$. If the atoms are equidistant, then $g^*_L(\boldsymbol{y}) = \phi(a_L)$ and $g^*_R(\boldsymbol{y}) = \phi(a_R)$ and thus also $\phi^*_{\rm at} = \phi$ in $\Omega^{\rm at}$. It is obvious that $\psi^{(j)} = \phi$ in Q_j for all $j \in \{-N,\dots,-K-1\}\cup\{K+2,\dots,N\}$. Summarizing, we get $\sigma^{\rm qc}_{\boldsymbol{y}}(x) = \sigma_{\boldsymbol{y}}(x)$ for all $x \in \Omega$, which implies that there are no ghost forces, that is, $D\mathcal{E}^{\rm qc}(\boldsymbol{y}) = 0$ for all $\boldsymbol{y} = F\widehat{\boldsymbol{X}} \in \mathcal{Y}$ representing homogeneous deformations.

Next, we prove consistency of the QC method. Because of the structure of the weak formulation (4.50), the analysis comes down to estimating the errors between the field ϕ coming from the original atomistic model and the fields $\psi^{(j)}$, respectively, $\phi^*_{\rm at}$.

For $M \in \mathbb{N}$ we define the index set

$$\mathcal{C}_M = \{-N, \dots, -K+M-1\} \cup \{K-M+1, \dots, N\}. \tag{4.52}$$

This set represents all atoms in the continuum region plus $2M$ atoms at the two ends of the atomistic region.

Lemma 4.27. *Let $\boldsymbol{y} \in \mathcal{Y}$ be given and assume $\min \boldsymbol{y}' \geq s_0 > \varsigma_0$. Then, there exists $C > 0$ such that*

$$\big|\big(D\mathcal{E}(\boldsymbol{y}) - D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\big)\cdot\boldsymbol{u}\big| \leq C\big(\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mMs_0}\big)\|\nabla u\|_{\mathrm{L}^2},$$

for all $\boldsymbol{u} \in \mathcal{U}$, where $u \in \mathrm{S}_{\#}(\boldsymbol{y})$ denotes an interpolant of $\boldsymbol{u}$ in the sense of (4.9)*. The constant C depends only on s_0 and δ_1.*

Proof. Using the weak formulation (4.50) of $D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})$ we obtain

$$\begin{aligned}\big|\big(D_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}) - D_{\boldsymbol{y}}\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\big)\cdot\boldsymbol{u}\big| &= \left|\int_\Omega \big(\sigma_{\boldsymbol{y}}(x) - \sigma^{\mathrm{qc}}_{\boldsymbol{y}}(x)\big)\nabla u(x)\,\mathrm{d}x\right| \\ &\leq \|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{qc}}_{\boldsymbol{y}}\|_{\mathrm{L}^2(\Omega)}\,\|\nabla u\|_{\mathrm{L}^2} \\ &\leq |\Omega|^{1/2}\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{qc}}_{\boldsymbol{y}}\|_{\mathrm{L}^\infty(\Omega)}\,\|\nabla u\|_{\mathrm{L}^2}\,.\end{aligned} \tag{4.53}$$

We therefore need to find error bounds for $\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{qc}}_{\boldsymbol{y}}\|_{\mathrm{L}^\infty(\Omega)}$ both in the atomistic and the continuum region.

Continuum Contribution. Since $\min \boldsymbol{y}' \geq s_0 > \varsigma_0$, Lemma 4.22 implies that

$$\varepsilon\big\|\nabla\phi - \nabla\psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)} + \big\|\phi - \psi^{(j)}\big\|_{\mathrm{L}^\infty(Q_j)} \leq C\Big(\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mM\min\boldsymbol{y}'}\Big)$$

uniformly in j, where C only depends on ms_0. Note that compared with Lemma 4.22 we have taken the ℓ^∞-norm of $\boldsymbol{y}''$ over the index set $\mathcal{C}_M$, which contains the continuum atoms as well as $2M$ atoms at the two ends of the atomistic domain. Referring to Lemma 4.23 we deduce that

$$\big\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{qc}}_{\boldsymbol{y}}\big\|_{\mathrm{L}^\infty(\Omega^{\mathrm{cb}})} = \big\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{cb}}_{\boldsymbol{y}}\big\|_{\mathrm{L}^\infty(\Omega^{\mathrm{cb}})} \leq C\Big(\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mM\min\boldsymbol{y}'}\Big),$$

where C depends on δ_1 and $m\min\boldsymbol{y}'$, respectively, ms_0.

Atomistic Contribution. This time we need a bound on the difference $\big\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{qc}}_{\boldsymbol{y}}\big\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} = \big\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{at}}_{\boldsymbol{y},*}\big\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})}$. For this we first address the error $|\phi(x) - \phi^*_{\mathrm{at}}(x)|$, $x \in \Omega^{\mathrm{at}}$. The functions ϕ and ϕ^*_{at} satisfy the equations $-\varepsilon^2\Delta\phi + m^2\phi = \rho_{\boldsymbol{y}}$, respectively, $-\varepsilon^2\Delta\phi^*_{\mathrm{at}} + m^2\phi^*_{\mathrm{at}} = \rho_{\boldsymbol{y}}$ in Ω^{at}. However, the boundary conditions are different: $\phi(a_L)$ and $\phi(a_R)$ for ϕ, respectively, $g^*_L(\boldsymbol{y})$ and $g^*_R(\boldsymbol{y})$ for ϕ^*_{at}. According to Lemma 4.15 we thus get

$$\|\phi - \phi^*_{\mathrm{at}}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} + \varepsilon\|\nabla\phi - \nabla\phi^*_{\mathrm{at}}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} \leq C\big(\big|\phi(a_L) - g^*_L(\boldsymbol{y})\big| + \big|\phi(a_R) - g^*_R(\boldsymbol{y})\big|\big).$$

The definitions of $g^*(\boldsymbol{y})$, $\gamma_L(\boldsymbol{y})$, and $\gamma_R(\boldsymbol{y})$ imply

$$\big|\gamma_L(\boldsymbol{y}) - g_L^*(\boldsymbol{y})\big| = \mathcal{O}(\tau), \quad \big|\gamma_R(\boldsymbol{y}) - g_R^*(\boldsymbol{y})\big| = \mathcal{O}(\tau).$$

For the value $\gamma_R(\boldsymbol{y})$, for example, we obtained in Remark 4.12 the equality

$$\gamma_R(\boldsymbol{y}) = \int_{\mathbb{R}} \rho_{\boldsymbol{y},R}^{\mathrm{refl}}(x) G_\varepsilon(a_R - x)\,\mathrm{d}x + \mathcal{O}(\tau).$$

where $\rho_{\boldsymbol{y},R}^{\mathrm{refl}}$ is a reflected and periodized extension of $\rho_{\boldsymbol{y}}|_{\Omega^{\mathrm{at}}}$. This then leads to

$$\begin{aligned}\phi(a_R) - g_R^*(\boldsymbol{y}) &= \phi(a_R) - \gamma_L(\boldsymbol{y}) + \mathcal{O}(\tau)\\ &= \frac{1}{2m\varepsilon}\int_{\mathbb{R}} \big(\rho_{\boldsymbol{y}}(z) - \rho_{\boldsymbol{y},R}^{\mathrm{refl}}(z)\big)\,\mathrm{e}^{-\frac{m}{\varepsilon}|a_R - z|}\,\mathrm{d}z + \mathcal{O}(\tau).\end{aligned}$$

Using the same ideas as in Lemmas 4.21 and 4.22 we can then show that

$$|\phi(a_R) - g_R^*(\boldsymbol{y})| \leq C\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + C\mathrm{e}^{-mM\min \boldsymbol{y}'} + \mathcal{O}(\tau),$$

where the constants only depend on $m\min \boldsymbol{y}'$. The same bound can be obtained for $|\phi(a_L) - g_L^*(\boldsymbol{y})|$. This then implies that

$$\|\phi - \phi_{\mathrm{at}}^*\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} + \varepsilon\|\nabla\phi - \nabla\phi_{\mathrm{at}}^*\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} \leq C\big(\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mM\min \boldsymbol{y}'} + \tau\big)$$

and hence

$$\big\|\sigma_{\boldsymbol{y}} - \sigma_{\boldsymbol{y}}^{\mathrm{qc}}\big\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} = \big\|\sigma_{\boldsymbol{y}} - \sigma_{\boldsymbol{y},*}^{\mathrm{at}}\big\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} \leq C\big(\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mMs_0} + \tau\big),$$

where C only depends on $m\min \boldsymbol{y}'$. Together with (4.53) and $\tau \leq \mathrm{e}^{-mMs_0}$ this completes the proof. ∎

4.5.1.2 Stability

The special choice $g^*(\boldsymbol{y})$ of boundary conditions for the atomistic subproblem allows for an elementary stability analysis of $\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})$ that draws from the ideas we used in Section 4.3.4. We recall that

$$\begin{aligned}\mathcal{E}_*^{\mathrm{at}}(\boldsymbol{y}) = \frac{1}{4m\varepsilon}\int_{\Omega^{\mathrm{at}}}\int_{\Omega^{\mathrm{at}}} \rho_{\boldsymbol{y}}(x)\big(&\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} + \mathrm{e}^{-\frac{m}{\varepsilon}(2a_R(\boldsymbol{y})-x-z)}\\ &+ \mathrm{e}^{-\frac{m}{\varepsilon}(x+z-2a_L(\boldsymbol{y}))}\big)\rho_{\boldsymbol{y}}(z)\,\mathrm{d}z\,\mathrm{d}x + \mathcal{O}(\tau).\end{aligned}$$

The τ-dependent terms in $\mathcal{E}_*^{\mathrm{at}}(\boldsymbol{y}) = \mathcal{E}_{a(\boldsymbol{y}),g^*(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}})$ from (4.39) only contain $\gamma_L(\boldsymbol{y})$ and $\gamma_R(\boldsymbol{y})$, whose derivatives are bounded by Lemma 4.14. The derivatives of these τ-dependent terms are therefore still of order $\mathcal{O}(\tau)$ and will be neglected in the proof of the following result.

Lemma 4.28. *Let $\boldsymbol{y} \in \mathcal{Y}$ satisfy $\min \boldsymbol{y}' > \varsigma_0$. Then*

$$D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq \Big(\frac{m\mu^2}{2}\,\mathrm{e}^{-m\max \boldsymbol{y}'} - \mathcal{O}(\tau)\Big)\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2 \quad \forall \boldsymbol{u} \in \mathcal{U}.$$

Proof. We treat continuum and atomistic contributions independently and start with the former. Lemma 4.24 states that

$$D^2\mathcal{E}_j^{\mathrm{cb}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq \frac{m^2\mu^2}{2}\,\mathrm{e}^{-m\max \boldsymbol{y}'}\varepsilon|u_j'|^2$$

for all $j = -N,\dots,N$. Hence, the definition (4.48) of $\mathcal{E}_*^{\mathrm{cb}}$ directly implies that

$$D^2\mathcal{E}_*^{\mathrm{cb}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq \mathrm{e}^{-m\max \boldsymbol{y}'}\frac{m^2\mu^2}{2}\varepsilon\Big(\sum_{j=-N}^{-K-1}|u_j'|^2 + \frac{1}{2}\big(|u_{-K}'|^2 + |u_{K+1}'|^2\big) + \sum_{j=K+2}^{N}|u_j'|^2\Big).$$

Let us now turn to the atomistic part $\mathcal{E}_*^{\mathrm{at}}(\boldsymbol{y})$. From Section 4.3.4 we know that for the given choice of boundary conditions and $a(\boldsymbol{y})$ we can write the energy of the atomistic part as

$$\begin{aligned}\mathcal{E}_*^{\mathrm{at}}(\boldsymbol{y}) &= \frac{\varepsilon}{4m}\sum_{i,j=-K}^{K}\int_{\Omega^{\mathrm{at}}}\int_{\Omega^{\mathrm{at}}}\delta_\varepsilon(x-y_i)\big(\mathrm{e}^{-\frac{m}{\varepsilon}|x-z|} + \mathrm{e}^{-\frac{m}{\varepsilon}(x+z-y_{-K-1}-y_{-K})}\\ &\qquad + \mathrm{e}^{-\frac{m}{\varepsilon}(y_{K+1}+y_K-x-z)}\big)\delta_\varepsilon(z-y_j)\,\mathrm{d}z\,\mathrm{d}x\\ &= \frac{\varepsilon\mu^2}{4m}\sum_{i,j=-K}^{K}\big(\mathrm{e}^{-\frac{m}{\varepsilon}|y_i-y_j|} + \mathrm{e}^{-\frac{m}{\varepsilon}(y_i+y_j-y_{-K}-y_{-K-1})}\\ &\qquad + \mathrm{e}^{-\frac{m}{\varepsilon}(y_K+y_{K+1}-y_i-y_j)}\big) + \mathcal{E}_{\mathrm{self}} + \mathcal{O}(\tau),\end{aligned} \tag{4.54}$$

where the constant $\mathcal{E}_{\mathrm{self}}$ accounts for the self-energies of the atoms $\{-K,\dots,K\}$. Differentiating twice and keeping only contributions from nearest neighbour interactions leads directly leads to

$$D^2\mathcal{E}_*^{\mathrm{at}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq \mathrm{e}^{-m\max \boldsymbol{y}'}\frac{m\mu^2}{2}\,\varepsilon\left(\frac{1}{2}|u_{-K}'|^2 + \sum_{i=-K+1}^{K}|u_i'|^2 + \frac{1}{2}|u_{K+1}'|^2\right) - \mathcal{O}(\tau)$$

Adding the lower bounds for $D^2\mathcal{E}_*^{\mathrm{cb}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}]$ and $D^2\mathcal{E}_*^{\mathrm{at}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}]$ we arrive at

$$\begin{aligned}D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] &= \big(D^2\mathcal{E}_*^{\mathrm{cb}}(\boldsymbol{y}) + D^2\mathcal{E}_*^{\mathrm{at}}(\boldsymbol{y})\big)\cdot[\boldsymbol{u},\boldsymbol{u}]\\ &\geq \Big(\mathrm{e}^{-m\max \boldsymbol{y}'}\frac{m\mu^2}{2} - \mathcal{O}(\tau)\Big)\,\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2\end{aligned}$$

for all $\boldsymbol{u} \in \mathcal{U}$, as desired. ∎

Next, we provide a Lipschitz continuity result for the second derivative $D^2\mathcal{E}^{\mathrm{qc}}$.

Lemma 4.29. *Let $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathcal{Y}$. If $\min \boldsymbol{y}_1' \geq s_0 > \varsigma_0$ and $\min \boldsymbol{y}_2' \geq s_0$, then*

$$\big\|D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}_1) - D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}_2)\big\| \leq L(s_0)\|\boldsymbol{y}_1' - \boldsymbol{y}_2'\|_{\ell^\infty}.$$

Proof. The Lipschitz continuity of $D^2\mathcal{E}_*^{\mathrm{cb}}$ follows from Lemma 4.25. Thus, we only have to consider the atomistic part $\mathcal{E}_*^{\mathrm{at}}$ given in a convenient form in (4.54). We present the proof for the part

$$\mathcal{E}_0^{\mathrm{at}}(\boldsymbol{y}) = \frac{\varepsilon\mu^2}{4m} \sum_{i,j=-K}^{K} \mathrm{e}^{-\frac{m}{\varepsilon}|y_i - y_j|} + \mathcal{E}_{\mathrm{self}}.$$

The parts involving $\mathrm{e}^{-\frac{m}{\varepsilon}(y_i+y_j-y_{-K}-y_{-K-1})}$ and $\mathrm{e}^{-\frac{m}{\varepsilon}(y_K+y_{K+1}-y_i-y_j)}$ can be treated analogously. Differentiating $\mathcal{E}_0^{\mathrm{at}}(\boldsymbol{y})$ twice leads to

$$\big(D^2\mathcal{E}_0^{\mathrm{at}}(\boldsymbol{y}_1) - D^2\mathcal{E}_0^{\mathrm{at}}(\boldsymbol{y}_2)\big)\cdot[\boldsymbol{u},\boldsymbol{u}] = \frac{m\varepsilon\mu^2}{4} \sum_{\substack{i,j=-K \\ i\neq j}}^{K} \big(\mathrm{e}^{-\frac{m}{\varepsilon}|y_{1,i}-y_{1,j}|} - \mathrm{e}^{-\frac{m}{\varepsilon}|y_{2,i}-y_{2,j}|}\big)\frac{(u_i-u_j)^2}{\varepsilon^2}.$$

Next, we analyze the first factor inside the sum. Since $y_{1,i}' \geq s_0$ and $y_{2,i}' \geq s_0$ for all $i \in \{-N,\dots,N\}$ we have $|y_{1,i}-y_{1,j}| \geq |i-j|\varepsilon s_0$, $|y_{2,i}-y_{2,j}| \geq |i-j|\varepsilon s_0$ for all $i,j \in \{-K,\dots,K\}$ and therefore by the Mean Value Theorem

$$\begin{aligned}
\big|\mathrm{e}^{-\frac{m}{\varepsilon}|y_{1,i}-y_{1,j}|} - \mathrm{e}^{-\frac{m}{\varepsilon}|y_{2,i}-y_{2,j}|}\big| &\leq m\mathrm{e}^{-m|i-j|s_0}\varepsilon^{-1}\big|y_{1,i}-y_{1,j}-(y_{2,i}-y_{2,j})\big| \\
&= m\mathrm{e}^{-m|i-j|s_0}\bigg|\sum_{\nu=i+1}^{j}\big(y_{1,\nu}'-y_{2,\nu}'\big)\bigg| \\
&\leq m\mathrm{e}^{-m|i-j|s_0}|j-i|\,\big\|\boldsymbol{y}_1'-\boldsymbol{y}_2'\big\|_{\ell^\infty}.
\end{aligned}$$

Hence,

$$\big|\big(D^2\mathcal{E}_0^{\mathrm{at}}(\boldsymbol{y}_1) - D^2\mathcal{E}_0^{\mathrm{at}}(\boldsymbol{y}_2)\big)\cdot[\boldsymbol{u},\boldsymbol{u}]\big| \leq \frac{m^2\varepsilon\mu^2}{4}\big\|\boldsymbol{y}_1'-\boldsymbol{y}_2'\big\|_{\ell^\infty} \sum_{\substack{i,j=-K \\ i\neq j}}^{K} \mathrm{e}^{-m|i-j|s_0}|j-i|\frac{(u_i-u_j)^2}{\varepsilon^2}.$$

The idea now is to show that the sum on the right-hand side is bounded by a function of s_0 times $\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2$. Elementary rearrangments lead to

$$\begin{aligned}
\sum_{\substack{i,j=-K \\ i\neq j}}^{K} \mathrm{e}^{-m|i-j|s_0}|j-i|\frac{(u_i-u_j)^2}{\varepsilon^2} &= 2\sum_{i=-K}^{K}\sum_{j=i+1}^{K} \mathrm{e}^{-m|i-j|s_0}|j-i|\frac{(u_i-u_j)^2}{\varepsilon^2} \\
&= 2\sum_{i=-K}^{K}\sum_{\nu=1}^{K-i} \mathrm{e}^{-m\nu s_0}\nu\frac{(u_{i+\nu}-u_i)^2}{\varepsilon^2} \\
&= 2\sum_{\nu=1}^{2K} \mathrm{e}^{-m\nu s_0}\nu \sum_{i=-K}^{K-\nu}\frac{(u_{i+\nu}-u_i)^2}{\varepsilon^2}.
\end{aligned}$$

We note that, for all $\nu \in \mathbb{N}$,

$$\frac{(u_{i+\nu}-u_i)^2}{\varepsilon^2} = \Big(\sum_{\kappa=i+1}^{i+\nu} u'_\kappa\Big)^2 \le \nu \sum_{\kappa=i+1}^{i+\nu} |u'_\kappa|^2.$$

Using this we obtain

$$\begin{aligned} 2\sum_{\nu=1}^{2K} \mathrm{e}^{-m\nu s_0}\nu \sum_{i=-K}^{K-\nu} \frac{(u_{i+\nu}-u_i)^2}{\varepsilon^2} &\le 2\sum_{\nu=1}^{2K} \mathrm{e}^{-m\nu s_0}\nu^2 \sum_{i=-K}^{K-\nu}\sum_{\kappa=i+1}^{i+\nu} |u'_\kappa|^2 \\ &\le 2\varepsilon^{-1}\sum_{\nu=1}^{\infty} \mathrm{e}^{-m\nu s_0}\nu^3 \|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2, \end{aligned}$$

where in the last step we have changed the summation order over j and κ and summed over $j = -K, \dots, K$. Summarizing, we have shown that

$$\big|\big(D^2\mathcal{E}_0^{\mathrm{at}}(\boldsymbol{y}_1) - D^2\mathcal{E}_0^{\mathrm{at}}(\boldsymbol{y}_2)\big)\cdot[\boldsymbol{u},\boldsymbol{u}]\big| \le \frac{2m^2\mu^2}{4}\big\|\boldsymbol{y}'_1 - \boldsymbol{y}'_2\big\|_{\ell^\infty}\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2 \sum_{\nu=1}^{\infty}\nu^3\mathrm{e}^{-m\nu s_0}.$$

Taking the supremum over $\boldsymbol{u} \in \mathcal{U}$ concludes the proof. ◻

4.5.1.3 Existence and Convergence

Before stating and proving the main result, we provide a lemma that relates the difference of $\boldsymbol{u} \in \mathcal{U}$ to the derivative of its interpolant $u \in \mathrm{S}_{\#}(\boldsymbol{y})$.

Lemma 4.30. *Let $\boldsymbol{u} \in \mathcal{U}$ be a test vector and $u \in \mathrm{S}_{\#}(\boldsymbol{y})$ an interpolant of $\boldsymbol{u}$ in the sense of* (4.9)*. Then,*

$$\|\nabla u\|_{\mathrm{L}^2} \le \frac{1}{(\min \boldsymbol{y}')^{1/2}}\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}.$$

Proof. By the definition of the interpolant u we have

$$\begin{aligned} \int_\Omega |\nabla u|^2 \,\mathrm{d}x &= \sum_{i=-N}^{N}\int_{y_{i-1}}^{y_i}\left(\frac{u_i-u_{i-1}}{y_i-y_{i-1}}\right)^2 \mathrm{d}x \\ &= \varepsilon\sum_{i=-N}^{N}\frac{(u_i-u_{i-1})^2}{\varepsilon^2}\frac{\varepsilon}{y_i-y_{i-1}} = \varepsilon\sum_{i=-N}^{N}\frac{|u'_i|^2}{y'_i}. \end{aligned}$$

Taking the square root concludes the proof. ◻

We are finally ready to prove an existence and convergence result. The proof consists in showing that all conditions in Lemma A.3 are satisfied, see for example [96, Theorem 8]. The parameter $M \in \mathbb{N}_0$ provides some flexibility. It can be adjusted so that the conditions are satisfied. The theorem can informally be paraphrased as follows: if the minimizer $\bar{\boldsymbol{y}} \in \arg\min E_{\boldsymbol{f}}$ of the original atomistic problem is sufficiently smooth in the neighbourhood $\mathcal{C}_M$ of the atoms in the continuum region and M is sufficiently large, then there exists a solution $\bar{\boldsymbol{y}}_{\mathrm{qc}} \in \mathcal{Y}$ to the QC approximated problem that is a good approximation to $\bar{\boldsymbol{y}}$.

Theorem 4.31. *Let $\bar{\boldsymbol{y}} \in \arg\min E_{\boldsymbol{f}}$ satisfy $\min \bar{\boldsymbol{y}}' \geq s_0 > \varsigma_0$ and $\max \bar{\boldsymbol{y}}' \leq S_0$. Then, there exists $\lambda(s_0, S_0) > 0$ such that, if, for some $M \in \mathbb{N}$,*

$$\varepsilon^{1/2}\|\bar{\boldsymbol{y}}''\|_{\ell^\infty(\mathcal{C}_M)} + \varepsilon^{-1/2}\mathrm{e}^{-mMs_0} \leq \lambda(s_0, S_0), \tag{4.55}$$

then there exists a solution $\bar{\boldsymbol{y}}_{\mathrm{qc}} \in \mathcal{Y}$ to

$$DE_{\boldsymbol{f}}^{\mathrm{qc}}(\bar{\boldsymbol{y}}^{\mathrm{qc}}) = 0 \quad \textit{in} \ \ \mathcal{U}^{-1,2},$$

that satisfies

$$\big\|\bar{\boldsymbol{y}}' - \bar{\boldsymbol{y}}'_{\mathrm{qc}}\big\|_{\ell^2_\varepsilon} \leq C(s_0, S_0)\big(\varepsilon\|\bar{\boldsymbol{y}}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mMs_0} + \tau\big).$$

Proof. First, we define $\mathcal{F} : \mathcal{U}^{1,2} \to \mathcal{U}^{-1,2}$ by $\mathcal{F}(\boldsymbol{w}) = DE_{\boldsymbol{f}}^{\mathrm{qc}}(\bar{\boldsymbol{y}} + \boldsymbol{w})$ for $\boldsymbol{w} \in \mathcal{U}$, where the spaces $\mathcal{U}^{1,2}$ and $\mathcal{U}^{-1,2}$ were introduced in Section 4.1.3. We need to show that $\mathcal{F}(\boldsymbol{w}) = 0$ has a solution $\boldsymbol{w} \in A = \{\boldsymbol{w} \in \mathcal{U} : \min(\bar{\boldsymbol{y}}' + \boldsymbol{w}') \geq \varsigma_0\}$.

Step 1. Consistency. The analysis from Lemma 4.27 together with Lemma 4.30 shows that

$$\|\mathcal{F}(0)\|_{\mathcal{U}^{-1,2}} = \big\|DE_{\boldsymbol{f}}^{\mathrm{qc}}(\bar{\boldsymbol{y}})\big\|_{\mathcal{U}^{-1,2}} = \big\|D\mathcal{E}^{\mathrm{qc}}(\bar{\boldsymbol{y}}) - D\mathcal{E}(\bar{\boldsymbol{y}})\big\|_{\mathcal{U}^{-1,2}} \leq \eta,$$

where

$$\eta = \frac{C(ms_0)}{s_0}\big(\varepsilon\|\bar{\boldsymbol{y}}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mMs_0} + \tau\big).$$

Step 2. Stability. Lemma 4.28 states that

$$D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\cdot[\boldsymbol{u}, \boldsymbol{u}] \geq \Big(\frac{m\mu^2}{2}\mathrm{e}^{-mS_0} - \mathcal{O}(\tau)\Big)\|\boldsymbol{u}'\|^2_{\ell^2_\varepsilon} =: \vartheta\|\boldsymbol{u}'\|^2_{\ell^2_\varepsilon} \quad \forall \boldsymbol{u} \in \mathcal{U},$$

which immediately translates to

$$\big\|D\mathcal{F}(0)^{-1}\big\|_{\mathrm{Lin}(\mathcal{U}^{-1,2}, \mathcal{U}^{1,2})} \leq \vartheta^{-1}.$$

Step 3. Lipschitz bound. The next step is to show the existence of a Lipschitz constant for $D\mathcal{F}$ in the neighbourhood $B_{2\eta\vartheta}(0)$. For all $\boldsymbol{w} \in \mathcal{U}$ with $\|\boldsymbol{w}'\|_{\ell^2_\varepsilon} \leq 2\eta\vartheta$ we get with an inverse inequality

$$\|\boldsymbol{w}'\|_{\ell^\infty} \leq \varepsilon^{-1/2}\|\boldsymbol{w}'\|_{\ell^2_\varepsilon} \leq 2\varepsilon^{-1/2}\eta\vartheta.$$

Let $0 < \delta < 1$. If $2\varepsilon^{-1/2}\eta\vartheta \leq (1-\delta)s_0$, we hence have $\min(\bar{\boldsymbol{y}}' + \boldsymbol{w}') \geq \delta s_0$ for all $\boldsymbol{w}$ with $\|\boldsymbol{w}'\|_{\ell^2_\varepsilon} \leq 2\eta\vartheta$. Knowing that $\bar{\boldsymbol{y}}' + \boldsymbol{w}'$ is bounded below, we can apply the Lipschitz bound from Lemma 4.29:

$$\begin{aligned}\big\|D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y} + \boldsymbol{w}_1) - D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y} + \boldsymbol{w}_2)\big\| &\leq L(\delta s_0)\|\boldsymbol{w}'_1 - \boldsymbol{w}'_2\|_{\ell^\infty} \\ &\leq L_\varepsilon\|\boldsymbol{w}'_1 - \boldsymbol{w}'_2\|_{\ell^2_\varepsilon},\end{aligned}$$

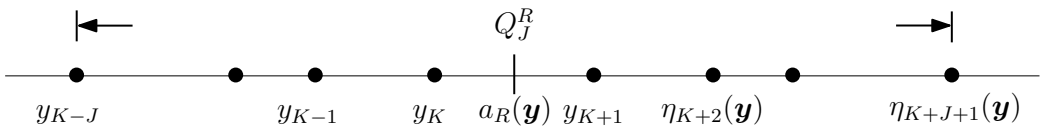

Figure 4.5: Illustration of the problem in the interval $Q_J^R = (y_{K-J}, 2a_R(\boldsymbol{y}) - y_{K-J})$ used to compute $g_R(\boldsymbol{y})$.

where $L_\varepsilon = \varepsilon^{-1/2} L(\delta s_0)$.

Step 4. Conclusion. What remains to be done is to ensure that $2L_\varepsilon \eta \vartheta < 1$. Looking at the product of these values, we see that for sufficiently small $\varepsilon^{-1/2}(\varepsilon \|\bar{\boldsymbol{y}}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mMs_0} + \tau)$, this can be satisfied. Lemma A.3 then guarantees the existence of $\bar{\boldsymbol{y}}_{\mathrm{qc}} \in \mathcal{Y}$ such that $DE_{\boldsymbol{f}}^{\mathrm{qc}}(\bar{\boldsymbol{y}}_{\mathrm{qc}}) = 0$. The configuration $\bar{\boldsymbol{y}}_{\mathrm{qc}}$ is a minimizer of $E_{\boldsymbol{f}}^{\mathrm{qc}}$ since $D^2 E_{\boldsymbol{f}}^{\mathrm{qc}}(\bar{\boldsymbol{y}}_{\mathrm{qc}})$ is positive definite. □

Referring to (4.55) we note that the magnitude of M depends on ε. The condition (4.55) can be satisfied if, for example, $M \sim -\log \varepsilon$.

4.5.2 Boundary Conditions From Cell Problems

The boundary conditions $g^*(\boldsymbol{y})$ we imposed on the atomistic subproblem in Section 4.5.1 gave rise to a method whose analysis turned out to be straightforward. The reasons for this lie in the clean weak formulation (4.50) of $D\mathcal{E}^{\mathrm{qc}}$ and the convenient stability properties established in Lemma 4.28. We now investigate how this situation changes if the boundary conditions chosen are approximations of $g^*(\boldsymbol{y})$ that can be computed easily. The following construction may also be easier to generalize to higher dimensions. We still set $a_L(\boldsymbol{y}) = \frac{1}{2}(y_{-K-1} + y_{-K})$ and $a_R(\boldsymbol{y}) = \frac{1}{2}(y_K + y_{K+1})$.

We recall from Remark 4.12 that $g_R^*(\boldsymbol{y})$ could (up to $\mathcal{O}(\tau)$) be interpreted as the field value in a_R produced by the symmetric particle density $\rho_{\boldsymbol{y},R}^{\mathrm{refl}}$. Loosely speaking, we now cut off this distribution and extend it periodically so that the new boundary conditions $g(\boldsymbol{y}) = [g_L(\boldsymbol{y})\ g_R(\boldsymbol{y})]^T$ can be obtained by solving periodic problems on certain domains. The presentation will be kept more informal than before.

A computationally cheap option to obtain $g(\boldsymbol{y})$ is given by

$$g_L(\boldsymbol{y}) = \psi^{(-K)}(a_L), \qquad g_R(\boldsymbol{y}) = \psi^{(K+1)}(a_R), \tag{4.56}$$

where $\psi^{(-K)}$ and $\psi^{(K+1)}$ are the fields from the Cauchy–Born approximations in the cells Q_{-K}, respectively, Q_{K+1}. These two cell problems have to be solved in any case to compute the energy of the continuum part. This particular choice of $g(\boldsymbol{y})$ would therefore not cause additional computational costs.

More generally, for $J \in \mathbb{N}_0$ we define the computational cell

$$Q_J^R = \big(y_{K-J},\, 2a_R - y_{K-J}\big).$$

For $J = 0$ this is just the cell Q_{K+1}. We will see that the magnitude of J is unimportant for the consistency of the method, but J does enter the stability analysis.

Let us introduce the linear operator $\boldsymbol{\eta} : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$ mapping $\boldsymbol{y}$ to $\boldsymbol{\eta}(\boldsymbol{y}) = (\eta_j(\boldsymbol{y}))_{j\in\mathbb{Z}}$. We define the components

$$\eta_{K-J}(\boldsymbol{y}) = y_{K-J}, \quad \ldots, \quad \eta_{K+1}(\boldsymbol{y}) = y_{K+1},$$

and (see Figure 4.5)

$$\eta_{K+2}(\boldsymbol{y}) = 2a_R(\boldsymbol{y}) - y_{K-1}, \quad \ldots, \quad \eta_{K+J+1}(\boldsymbol{y}) = 2a_R(\boldsymbol{y}) - y_{K-J}.$$

Note that the components $\eta_{K+2}(\boldsymbol{y})$, ... , $\eta_{K+J+1}(\boldsymbol{y})$ are mirror images of the coordinates $\eta_{K-1}(\boldsymbol{y})$, ... , $\eta_{K-J}(\boldsymbol{y})$ across $a_R(\boldsymbol{y})$.

Next, we define the missing components of $\boldsymbol{\eta}(\boldsymbol{y})$ by periodic extension:

$$\eta_{K+(2J+2)\nu+j}(\boldsymbol{y}) = \eta_{K+j}(\boldsymbol{y}) + \nu|Q_J^R| \quad \forall j \in \{-J, \ldots, J+1\}, \quad \forall \nu \in \mathbb{Z}.$$

The boundary condition $g_R(\boldsymbol{y})$ is now obtained by solving the periodic problem

$$-\varepsilon^2 \Delta \widetilde{\psi}_R + m^2 \widetilde{\psi}_R = \rho_{\boldsymbol{\eta}(\boldsymbol{y})} \quad \text{in } Q_J^R$$

or, equivalently,

$$-\varepsilon^2 \Delta \widetilde{\psi}_R + m^2 \widetilde{\psi}_R = \rho_{\boldsymbol{\eta}(\boldsymbol{y})} \quad \text{in } \mathbb{R} \tag{4.57}$$

and setting

$$g_R(\boldsymbol{y}) = \widetilde{\psi}_R(a_R). \tag{4.58}$$

The left-hand boundary condition $g_L(\boldsymbol{y})$ is defined analogously. We set $g(\boldsymbol{y}) = [g_L(\boldsymbol{y})\ g_R(\boldsymbol{y})]^T$.

We then define a second Quasicontinuum energy $\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})$ as follows

$$\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y}) = \mathcal{E}_*^{\mathrm{cb}}(\boldsymbol{y}) + \mathcal{E}^{\mathrm{at}}(\boldsymbol{y}), \tag{4.59}$$

where $\mathcal{E}_*^{\mathrm{cb}}(\boldsymbol{y})$ is the same as in the method discussed in Section 4.5.1 (see (4.48)) and

$$\begin{aligned}\mathcal{E}^{\mathrm{at}}(\boldsymbol{y}) &= \mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}}) \\ &= -\inf\Big\{ I_{a(\boldsymbol{y})}(\varphi, \boldsymbol{y}_{\mathrm{at}}) : \quad \varphi \in \mathrm{H}^1(\Omega^{\mathrm{at}}), \quad \varphi|_{\partial\Omega^{\mathrm{at}}} = g(\boldsymbol{y}) \Big\}.\end{aligned}$$

We denote the minimizer for given $\boldsymbol{y}$ by $\phi_{\mathrm{at}} \in \mathrm{H}^1(\Omega^{\mathrm{at}})$.

4.5.2.1 Consistency Analysis

A crucial difference between the QC energy (4.59) and the energy from Section 4.5.1 is that now the derivative of the atomistic energy with respect to the boundary conditions in general does not vanish. We therefore have to ensure that the term $D_g \mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}}) D_{\boldsymbol{y}} g(\boldsymbol{y}) \cdot \boldsymbol{u}$ emerging in $D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})$ can still be included in the weak formulation.

Weak Formulation. Let $\boldsymbol{u} \in \mathcal{U}$ be a test vector and $u \in \mathrm{S}_{\#}(\boldsymbol{y})$ an interpolant of $\boldsymbol{u}$. The goal now is to show that

$$D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\cdot\boldsymbol{u} = \int_\Omega \sigma^{\mathrm{qc}}_{\boldsymbol{y}}(x)\nabla u(x)\,\mathrm{d}x + \int_\Omega \sigma^{\mathrm{qc}}_{g(\boldsymbol{y})}\nabla u\,\mathrm{d}x\,, \tag{4.60}$$

where

$$\sigma^{\mathrm{qc}}_{\boldsymbol{y}}(x) = \begin{cases} \sigma^{\mathrm{cb}}_{\boldsymbol{y}}(x) & \text{if } x \in \Omega^{\mathrm{cb}}, \\ \sigma^{\mathrm{at}}_{\boldsymbol{y}}(x) & \text{if } x \in \Omega^{\mathrm{at}}, \end{cases}$$

and $\sigma^{\mathrm{at}}_{\boldsymbol{y}}(x)$ is given by (4.11) with $\phi = \phi_{\mathrm{at}}$. The additional term $\sigma^{\mathrm{qc}}_{g(\boldsymbol{y})}$ in (4.60) satisfies

$$\big\|\sigma^{\mathrm{qc}}_{g(\boldsymbol{y})}\big\|_{\mathrm{L}^\infty} \le C\big|g(\boldsymbol{y}) - g^*(\boldsymbol{y})\big|, \tag{4.61}$$

where C depends on $m \min \boldsymbol{y}'$ and $\max \boldsymbol{y}'$.

Since the continuum contribution to $D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\cdot\boldsymbol{u}$ is the same as in Section 4.5.1 we only need to analyze $D\mathcal{E}^{\mathrm{at}}(\boldsymbol{y})$. Using the chain rule we obtain

$$\begin{aligned} D\mathcal{E}^{\mathrm{at}}(\boldsymbol{y})\cdot\boldsymbol{u} = {} & D_{\boldsymbol{y}_{\mathrm{at}}}\mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}})\cdot\boldsymbol{u}_{\mathrm{at}} + D_a\mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}})\cdot a(\boldsymbol{u}) \\ & + D_g\mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}})\cdot D_{\boldsymbol{y}}g(\boldsymbol{y})\cdot\boldsymbol{u}. \end{aligned}$$

The same reasoning as in Section 4.5.1 gives for the first two terms on the right-hand side

$$D_{\boldsymbol{y}_{\mathrm{at}}}\mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}})\cdot\boldsymbol{u}_{\mathrm{at}} + D_a\mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}})\cdot a(\boldsymbol{u}) = \int_{\Omega^{\mathrm{at}}} \sigma^{\mathrm{at}}_{\boldsymbol{y}}(x)\nabla u(x)\,\mathrm{d}x\,, \tag{4.62}$$

where $\sigma^{\mathrm{at}}_{\boldsymbol{y}}(x)$ is given by (4.11) with $\phi = \phi_{\mathrm{at}}$.

Next, we turn our attention to the term $D_g\mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}}) \cdot D_{\boldsymbol{y}}g(\boldsymbol{y}) \cdot \boldsymbol{u}_{\mathrm{at}}$ and start by considering $D_{\boldsymbol{y}}g_R(\boldsymbol{y}) \cdot \boldsymbol{u}$. Going back to Remark 4.12 we recall that

$$\begin{aligned} g^*_R(\boldsymbol{y}) = \gamma_R(\boldsymbol{y}) + \mathcal{O}(\tau) &= \frac{1}{m}\int_\Omega \sum_{j=-K}^{K} \delta_\varepsilon(x - y_j)\mathrm{e}^{-\frac{m}{\varepsilon}(a_R(\boldsymbol{y})-x)}\,\mathrm{d}x + \mathcal{O}(\tau) \\ &= \frac{\mu}{m}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K + y_{K+1} - 2y_{K-j})} + \mathcal{O}(\tau). \end{aligned} \tag{4.63}$$

Here, we have extended the sum to infinity for convenience. This gives rise to an additional error of order $\mathcal{O}(\tau)$. Since we still assume that the atomistic domain $\Omega^{\mathrm{at}} = \big(a_L(\boldsymbol{y}), a_R(\boldsymbol{y})\big)$ is large, the error thus incurred is small. Similarly, it follows from (4.57) that the definition (4.58) of $g_R(\boldsymbol{y})$ is equivalent to

$$g_R(\boldsymbol{y}) = \frac{\mu}{m}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K + y_{K+1} - 2\eta_{K-j}(\boldsymbol{y}))}. \tag{4.64}$$

Depending on the properties of $\boldsymbol{y}$ we have $g(\boldsymbol{y}) \approx g^*(\boldsymbol{y})$ but g only depends on $2J+4$ entries of $\boldsymbol{y}$ whereas g^* depends on $\{y_{-K-1}, \dots, y_{K+1}\}$. We note that $\eta_{K-j}(\boldsymbol{y})$ can, for $j \in \mathbb{N}_0$, be expressed in the form

$$\eta_{K-j}(\boldsymbol{y}) = y_K - \varepsilon\sum_{i=0}^{J} k_i^{(j)} y'_{K+1-i},$$

where

$$k_i^{(j)} \in \mathbb{N}_0 \quad \text{for all } i \in \{0,\dots,J\} \ \text{ and all } j \in \mathbb{N}_0, \ \text{ and } \sum_{i=0}^{J} k_i^{(j)} = j.$$

In words: the distance between $\eta_{K-j}(\boldsymbol{y})$ and y_K is the sum of multiples of the distances $y_{K+1-J} - y_{K-J}$, $\dots$, $y_{K+1} - y_K$. This is a direct consequence of the definition of $\boldsymbol{\eta}(\boldsymbol{y})$ in terms of reflection and periodization. Differentiating (4.64) then leads to

$$\begin{aligned} D_{\boldsymbol{y}} g_R(\boldsymbol{y})\cdot\boldsymbol{u} &= \frac{-\mu}{2}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2\eta_{K-j}(\boldsymbol{y}))}\,\frac{u_K+u_{K+1}-2\eta_{K-j}(\boldsymbol{u})}{\varepsilon} \\ &= \frac{-\mu}{2}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2\eta_{K-j}(\boldsymbol{y}))}\left(u'_{K+1} + 2\sum_{i=0}^{J} k_i^{(j)} u'_{K+1-i}\right) \\ &= \frac{-\mu}{2}\sum_{i=0}^{J}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2\eta_{K-j}(\boldsymbol{y}))}\left(\tfrac{1}{J+1}u'_{K+1} + 2k_i^{(j)} u'_{K+1-i}\right) \\ &= \sum_{i=0}^{J} \mu^R_{K+1-i}(\boldsymbol{y}) u'_{K+1-i}. \end{aligned} \tag{4.65}$$

Here, we have defined

$$\mu^R_{K+1-i}(\boldsymbol{y}) = -\frac{\mu}{2}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2\eta_{K-j}(\boldsymbol{y}))}\left(2k_i^{(j)} + \delta_{i,0}\tfrac{1}{J+1}\right) \quad \forall j \in \{0,\dots,J\}.$$

It is clear from their definition that $k_i^{(j)} \le j$ for all $i \in \{0,\dots,J\}$ and all $j \in \mathbb{N}_0$. Thus, we get the following bound for $\mu^R_{K+1-i}(\boldsymbol{y})$:

$$\begin{aligned} |\mu^R_{K+1-i}(\boldsymbol{y})| &= \left|\frac{\mu}{2}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2\eta_{K-j}(\boldsymbol{y}))}\left(2k_i^{(j)} + \delta_{i,0}\tfrac{1}{J+1}\right)\right| \\ &\le \frac{\mu}{2}\sum_{j=0}^{\infty}(2j+1)\mathrm{e}^{-m(j+1/2)\min \boldsymbol{y}'}. \end{aligned} \tag{4.66}$$

Similar considerations can be carried out at the left-hand boundary for $g_L(\boldsymbol{y})$.

We recall from Lemma 4.11 that (for $\tau \approx 0$)

$$D_g \mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}}) = -m\varepsilon\big[g_L(\boldsymbol{y}) - g_L^*(\boldsymbol{y}),\ \ g_R(\boldsymbol{y}) - g_R^*(\boldsymbol{y})\big] + \mathcal{O}(\tau).$$

Multiplying this with (4.65) we deduce that the contribution to $D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})$ from the boundary data takes the form

$$D_g \mathcal{E}_{a(\boldsymbol{y}),g(\boldsymbol{y})}(\boldsymbol{y}_{\mathrm{at}})\cdot D_{\boldsymbol{y}} g(\boldsymbol{y})\cdot\boldsymbol{u} = \int_\Omega \sigma^{\mathrm{qc}}_{g(\boldsymbol{y})}(x)\nabla u(x)\,\mathrm{d}x + \mathcal{O}(\tau), \tag{4.67}$$

where the additional stress function $\sigma^{\mathrm{qc}}_{g(\boldsymbol{y})}$ is piecewise constant and nonzero in neighbourhoods of the atomistic/continuum interface $\partial\Omega^{\mathrm{at}}$:

$$\sigma^{\mathrm{qc}}_{g(\boldsymbol{y})}(x) = \begin{cases} y'_i m\mu_i^L(\boldsymbol{y})\big(g_L^*(\boldsymbol{y}) - g_L(\boldsymbol{y})\big) & \text{if } x \in Q_i, \ \ i \in \{-K,\dots,-K+J\}, \\ y'_i m\mu_i^R(\boldsymbol{y})\big(g_R^*(\boldsymbol{y}) - g_R(\boldsymbol{y})\big) & \text{if } x \in Q_i, \ \ i \in \{K-J+1,\dots,K+1\}, \\ 0, & \text{otherwise.} \end{cases}$$

Here, we have used that $u_i' = y_i' \nabla u|_{Q_i}$. Adding (4.62) and (4.67) completes the proof of (4.60). The L^∞-bound (4.61) follows immediately from (4.66) and the definition of $\sigma^{\mathrm{qc}}_{g(\boldsymbol{y})}(x)$.

Consistency. As in Lemma 4.27 the consistency proof for given $\boldsymbol{y} \in \mathcal{Y}$ consists in finding an appropriate bound on $\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{qc}}_{\boldsymbol{y}}\|_{\mathrm{L}^\infty(\Omega)}$. In the atomistic region Ω^{at} we therefore need to estimate the errors $\|\phi - \phi_{\mathrm{at}}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})}$ and $\|\nabla\phi - \nabla\phi_{\mathrm{at}}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})}$. Lemma 4.15 gives

$$\begin{aligned}\|\phi - \phi_{\mathrm{at}}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} + \varepsilon\|\nabla\phi - \nabla\phi_{\mathrm{at}}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} &\leq C\big(|\phi(a_L) - \phi_{\mathrm{at}}(a_L)| + |\phi(a_R) - \phi_{\mathrm{at}}(a_R)|\big)\\ &= C\big(|\phi(a_L) - g_L(\boldsymbol{y})| + |\phi(a_R) - g_R(\boldsymbol{y})|\big).\end{aligned}$$

Using the definitions of $g_L(\boldsymbol{y})$ and $g_R(\boldsymbol{y})$, and techniques similar to the ones used in the proof of Lemma 4.21 it can be shown that

$$\|\phi - \phi_{\mathrm{at}}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} + \varepsilon\|\nabla\phi - \nabla\phi_{\mathrm{at}}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} \leq C\big(\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mM\min\boldsymbol{y}'} + \tau\big)$$

as well as

$$|g(\boldsymbol{y}) - g^*(\boldsymbol{y})| \leq C\big(\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mM\min\boldsymbol{y}'}\big)$$

for $M \in \mathbb{N}$ and $\mathcal{C}_M$ as defined in (4.52). From this we then deduce consistency in the sense that

$$\|\sigma_{\boldsymbol{y}} - \sigma^{\mathrm{qc}}_{\boldsymbol{y}}\|_{\mathrm{L}^\infty(\Omega)} + \|\sigma^{\mathrm{qc}}_{g(\boldsymbol{y})}\|_{\mathrm{L}^\infty(\Omega^{\mathrm{at}})} \leq C\big(\varepsilon\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)} + \mathrm{e}^{-mM\min\boldsymbol{y}'} + \tau\big).$$

Note that we have used (4.61).

4.5.2.2 Stability Analysis

The stability analysis for the QC energy (4.59) is slightly more involved than for the method discussed in Section 4.5.1. Let $\boldsymbol{y} \in \mathcal{Y}$ be given. Our main observation is that for sufficiently large J there exists a constant $C(m\min\boldsymbol{y}', \max\boldsymbol{y}')$ such that

$$D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq C(m\min\boldsymbol{y}', \max\boldsymbol{y}')\|\boldsymbol{u}'\|^2_{\ell^2_\varepsilon} \qquad \forall \boldsymbol{u} \in \mathcal{U}^{1,2}.$$

Since the continuum part of the energy is the same as in the first method, we only have to look at stability of the atomistic subproblem with the given choice of boundary data. The idea is to write the second derivative of the energy $\mathcal{E}^{\mathrm{at}}$ in the form

$$D^2\mathcal{E}^{\mathrm{at}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] = D^2\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] + \big(D^2\mathcal{E}^{\mathrm{at}}(\boldsymbol{y}) - D^2\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})\big)\cdot[\boldsymbol{u},\boldsymbol{u}]$$

and use the coercivity of $D^2\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})$: we know from Lemma 4.28 that

$$D^2\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq \mathrm{e}^{-m\max\boldsymbol{y}'}\frac{m\mu^2}{2}\,\varepsilon\left(\frac{1}{2}|u'_{-K}|^2 + \sum_{i=-K+1}^{K}|u_i'|^2 + \frac{1}{2}|u'_{K+1}|^2\right) - \mathcal{O}(\tau)$$

for all $\boldsymbol{u} \in \mathcal{U}$. Hence, our aim is to show that the difference $\|D^2\mathcal{E}^{\mathrm{at}}(\boldsymbol{y}) - D^2\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})\|$ is sufficiently small not to break stability.

The difference between the energies $\mathcal{E}^{\mathrm{at}}(\boldsymbol{y})$ and $\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})$ only consists in effects from the boundary conditions and we have, by (4.33),

$$\begin{aligned}\mathcal{E}^{\mathrm{at}}(\boldsymbol{y}) - \mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y}) =& -I_{a(\boldsymbol{y})}(\xi_{a(\boldsymbol{y}),g(\boldsymbol{y})},\boldsymbol{y}) + I_{a(\boldsymbol{y})}(\xi_{a(\boldsymbol{y}),g^*(\boldsymbol{y})},\boldsymbol{y})\\ =& \frac{m\varepsilon}{2}\big|g(\boldsymbol{y}) - g^*(\boldsymbol{y})\big|^2 + \mathcal{O}(\tau).\end{aligned}$$

This implies that

$$\begin{aligned}\big(D^2\mathcal{E}^{\mathrm{at}}(\boldsymbol{y}) - D^2\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})\big)\cdot[\boldsymbol{u},\boldsymbol{u}] =& \, m\varepsilon\big(g(\boldsymbol{y}) - g^*(\boldsymbol{y})\big)^T\big[\big(D^2g(\boldsymbol{y}) - D^2g^*(\boldsymbol{y})\big)\cdot[\boldsymbol{u},\boldsymbol{u}]\big]\\ &+ 2m\varepsilon\big|\big(Dg(\boldsymbol{y}) - Dg^*(\boldsymbol{y})\big)\cdot\boldsymbol{u}\big|^2 + \mathcal{O}(\tau).\end{aligned} \tag{4.68}$$

We will show that $\varepsilon\big|\big(D^2g(\boldsymbol{y}) - D^2g^*(\boldsymbol{y})\big)\cdot[\boldsymbol{u},\boldsymbol{u}]\big|$ is bounded so that the first term on the right-hand side is bounded by $C\big|g(\boldsymbol{y}) - g^*(\boldsymbol{y})\big|$. The second term, $2m\varepsilon\big|(Dg(\boldsymbol{y}) - Dg^*(\boldsymbol{y}))\cdot\boldsymbol{u}\big|$, is positive for all $\boldsymbol{u}\in\mathcal{U}$ and therefore does not affect the positive definiteness of $D^2\mathcal{E}^{\mathrm{at}}(\boldsymbol{y})$. If, however, two energies $\mathcal{E}_1$ and $\mathcal{E}_2$ generating a purely repulsive, respectively, a purely attractive interaction are combined to obtain a Morse-like interaction potential, then these terms $\big|(Dg(\boldsymbol{y}) - Dg^*(\boldsymbol{y}))\cdot\boldsymbol{u}\big|$ are relevant for the overall stability analysis. For this reason we provide a bound below. We show that $\big|(Dg(\boldsymbol{y}) - Dg^*(\boldsymbol{y}))\cdot\boldsymbol{u}\big|$ decreases as the sizes of the cells Q^L_J, Q^R_J, on which the boundary conditions $g(\boldsymbol{y})$ are computed, increases.

First, we address the first term on the right-hand side of (4.68). Differentiating gives

$$\begin{aligned}D^2g^*_R(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] =& \frac{m\mu}{4}\sum_{j=0}^{\infty}\mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2y_{K-j})}\left(\frac{u_K + u_{K+1} - 2u_{K-j}}{\varepsilon}\right)^2\\ =& \frac{m\mu}{4}\sum_{j=0}^{\infty}\mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2y_{K-j})}\left(u'_{K+1} - 2\sum_{i=0}^{j}u'_{K-j}\right)^2\end{aligned}$$

and, similarly,

$$D^2g_R(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] = \frac{m\mu}{4}\sum_{j=0}^{\infty}\mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2\eta_{K-j}(\boldsymbol{y}))}\left(u'_{K+1} - 2\sum_{i=0}^{J}k_i^{(j)}u'_{K+1-j}\right)^2.$$

A calculation very similar to the one given in the proof of Lemma 4.14 leads to

$$\varepsilon\big(\big|D^2g_R(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}]\big| + \big|D^2g^*_R(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}]\big|\big) \le C(m\min\boldsymbol{y}')\|\boldsymbol{u}'\|^2_{\ell^2_\varepsilon},$$

which implies, for the first term on the right-hand side of (4.68)

$$m\varepsilon\big|\big(g(\boldsymbol{y}) - g^*(\boldsymbol{y})\big)^T\big(D^2(g(\boldsymbol{y}) - g^*(\boldsymbol{y}))\cdot[\boldsymbol{u},\boldsymbol{u}]\big)\big| \le C(m\min\boldsymbol{y}')\big|g(\boldsymbol{y}) - g^*(\boldsymbol{y})\big|\|\boldsymbol{u}'\|^2_{\ell^2_\varepsilon}$$

for all $\boldsymbol{u}\in\mathcal{U}$.

Now we analyze the second term $2m\varepsilon\big(D_{\boldsymbol{y}}(g(\boldsymbol{y}) - g^*(\boldsymbol{y})) \cdot \boldsymbol{u}\big)^2$ on the right-hand side of (4.68). We recall from (4.63), respectively, (4.64) that

$$Dg_R(\boldsymbol{y})\cdot\boldsymbol{u} = \frac{-\mu}{2}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2\eta_{K-j}(\boldsymbol{y}))}\frac{u_K + u_{K+1} - 2\eta_{K-j}(\boldsymbol{u})}{\varepsilon},$$
$$Dg_R^*(\boldsymbol{y})\cdot\boldsymbol{u} = \frac{-\mu}{2}\sum_{j=0}^{\infty} \mathrm{e}^{-\frac{m}{2\varepsilon}(y_K+y_{K+1}-2y_{K-j})}\frac{u_K + u_{K+1} - 2u_{K-j}}{\varepsilon}.$$

As a direct result of the construction of $\boldsymbol{\eta}$ the first $J+1$ terms in the above sums are equal. A quick calculation then shows that

$$\varepsilon\big|\big(Dg(\boldsymbol{y}) - Dg^*(\boldsymbol{y})\big)\cdot\boldsymbol{u}\big|^2 \leq C(m\min\boldsymbol{y}')\mathrm{e}^{-m(2J+1)\min\boldsymbol{y}'}\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2.$$

Summarizing, we have shown that

$$\big\|D^2\mathcal{E}^{\mathrm{at}}(\boldsymbol{y}) - D^2\mathcal{E}^{\mathrm{at}}_*(\boldsymbol{y})\big\| \leq C(m\min\boldsymbol{y}')\big(|g(\boldsymbol{y}) - g^*(\boldsymbol{y})| + \mathrm{e}^{-m(2J+1)\min\boldsymbol{y}'}\big),$$

from which we deduce that

$$D^2\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})\cdot[\boldsymbol{u},\boldsymbol{u}] \geq C(m\min\boldsymbol{y}')\|\boldsymbol{u}'\|_{\ell^2_\varepsilon}^2 \quad \forall \boldsymbol{u}\in\mathcal{U}^{1,2}$$

for sufficiently small $\|\boldsymbol{y}''\|_{\ell^\infty(\mathcal{C}_M)}$ and sufficiently large M and J.

A convergence result analogous to Theorem 4.31 can be proven using the same techniques based on the Implicit Function Theorem.

4.6 Conclusions and Outlook

In this chapter we have presented a rigorous analysis of a particular way of Quasicontinuum like coupling for a linear, field-based interaction potential in one space dimension. The starting point for the design of coupling methods was a weak formulation of the forces arising from the atomistic model. This provided a natural connection point to the corresponding continuum model. We believe that the present work in a comparably simple setting addresses several important questions relevant for QC coupling in the presence of fields: most prominently the dependence of minimization problems on the boundary and boundary data.

We close the chapter with some comments on open questions. This model being evidently basic from the outset, there is a lot of scope for further work.

Throughout the chapter we only considered lattices with one species of atoms. The definition of the energy $\mathcal{E}(\boldsymbol{y})$, the weak formulation and the outlined construction of QC methods do, however, generalize to lattices with more than one species, i.e., complex lattices. In this case, the unit cells obviously contain more than one atom. The particle positions can be represented by $\boldsymbol{y} = (y_i^\alpha)_{i=-N,\dots,N}^{\alpha=1,\dots,n_{\mathrm{uc}}}$, where n_{uc} is the number of atoms in the unit cell, and

$2N+1$ is the number of unit cells under consideration. The particle density $\rho_{\boldsymbol{y}}$ then takes the form

$$\rho_{\boldsymbol{y}}(z) = \sum_{i=-N}^{N} \sum_{\alpha=1}^{n_{\mathrm{uc}}} Z_\alpha \delta_\varepsilon(z - y_i^\alpha).$$

The numbers $Z_\alpha \in \mathbb{R}, \alpha = 1, \dots, n_{\mathrm{uc}}$, represent the "charges" of the different species. The energy $\mathcal{E}(\boldsymbol{y})$ is defined in the same way as in the simple lattice case. The Cauchy–Born approximation has to take into account internal relaxation of the atoms in the unit cell. This means that the relative positions of atoms in the unit cell may change when it is deformed. Defining Quasicontinuum approximations by imposing boundary conditions on the atomistic subproblem is still possible. Homogeneous Neumann boundary conditions can only be used when the unit cell has a mirror symmetry.

For the construction of the two QC methods in Section 4.5 we chose $\boldsymbol{y}$-dependent boundaries $a(\boldsymbol{y})$ of the atomistic subdomain Ω^{at}. In other words we fixed the position of the boundary in the Lagrange picture. This led to very convenient weak formulations of $D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})$. An obvious alternative (particularly relevant for higher dimensions) is the choice of $\boldsymbol{y}$-independent a. We note that this, however, comes with additional technical difficulties. Let $\boldsymbol{y} \in \mathcal{Y}$ and assume that $y_{-K-1} < a_L < y_{-K}$. Then, the Cauchy–Born energy of the interval (y_{-K-1}, a_L) in the continuum region Ω^{cb} is given by

$$\begin{aligned} -\int_{y_{-K-1}}^{a_L} \left(\tfrac{1}{2}\varepsilon^2 |\nabla\psi^{(-K)}|^2 + \tfrac{1}{2}m^2(\psi^{(-K)})^2 - \rho_{\boldsymbol{y}}\psi^{(-K)}\right) \mathrm{d}x \\ = \frac{1}{2}\int_{y_{-K-1}}^{a_L} \rho_{\boldsymbol{y}}\psi^{(-K)}\,\mathrm{d}x - \frac{\varepsilon^2}{2}\psi^{(-K)}(a_L)\nabla\psi^{(-K)}(a_L), \end{aligned}$$

where $\psi^{(-K)}$ is the Cauchy–Born field on the cell Q_{-K}. To calculate the derivative of this energy contribution, it is hence necessary to know $D_{a_L}\psi^{(-K)}(a_L)$ and $D_{a_L}\nabla\psi^{(-K)}(a_L)$ explicitly and include the resulting terms into the weak formulation of $D\mathcal{E}^{\mathrm{qc}}(\boldsymbol{y})$. Moreover, the boundary conditions $g^*(\boldsymbol{y}, a)$ cannot be used for the atomistic subproblem. If a_L, a_R do not coincide with atomic positions or lie halfway between two atoms, Neumann boundary conditions lead to field values $\phi^*_{\mathrm{at}}(a_L) = g^*_L(\boldsymbol{y}, a)$, $\phi^*_{\mathrm{at}}(a_R) = g^*_R(\boldsymbol{y}, a)$ that are inconsistent with the exact values $\phi(a_L)$, $\phi(a_L)$. For example, if a_L is closer to y_{-K} than to y_{-K-1}, then $g^*_L(\boldsymbol{y}, a) > \phi(a_L)$ where the error is of order $\mathcal{O}(1)$. This also implies that the weak formulation will include a term coming from the derivative of the atomistic energy with respect to the boundary conditions, similar to (4.60).

It has to be stressed that our analysis heavily utilized explicitly known Green's functions (G_ε for the whole of $\mathbb{R}$, $G_{\varepsilon,a}$ for bounded domains with homogeneous Dirichlet boundary data, and $G^{\#}_{\varepsilon,\Omega}$ for periodic domains). The exponential decay of G_ε combined with explicit formulas for field values allowed a straightforward consistency analysis, see for example Lemmas 4.21 and 4.27. Moreover, the one-dimensional setting allowed us to fully understand the

dependence of certain minimization problems with respect to the domain and the boundary conditions. Knowing the Green's function $G_{\varepsilon,a}$ we derived an explicit formula for the field ϕ in a bounded domain with Dirichlet boundary conditions, which in turn gave us an explicit formula for the energy, see (4.39). This proved very convenient for the stability analysis.

Depending on the geometry of the domains Ω and Ω^{at} the construction of Green's functions in higher dimensions might be impossible. Furthermore, the use of Green's functions is restricted to linear models. In the case of nonlinear models the solution operator for the resulting partial differential equations is not given by a convolution with the Green's function. Hence, different tools have to be found in order to generalize the present analysis to higher dimensions and nonlinear models. Especially for the consistency analysis it might not be necessary to work with explicit formulas for the fields. The consistency proofs were essentially based on the decay of the effect the position y_i has on the field value $\phi(x)$ as $|x-y_i|$ increases. This decay can be expected to show in many field-based interaction models.

Both QC methods we presented were based on boundary conditions on the atomistic subproblem that led to the complete decoupling of the continuum region and the atomistic region. The atomistic energy part $\mathcal{E}^{\mathrm{at}}(\boldsymbol{y})$ only depended on the components $y_{-K-1},\dots,y_{K+1}$. In the case of the first method the effect of the boundary conditions $g^*(\boldsymbol{y})$ could elegantly be interpreted as the interaction with mirror atoms outside the atomistic subdomain Ω^{at}. A generalization of this framework to higher dimensions is likely to involve more complicated geometrical constructions, and may even turn out to be impossible.

Appendix A

Miscellaneous Results

A.1 Analysis

First, we establish a Taylor expansion with remainder term for mappings whose first derivative is only Hölder continuous. The result can be generalized to higher order expansions without problems.

Lemma A.1. *Let X, Y be Banach spaces and $T : X \to Y$ a Fréchet differentiable operator, whose derivative satisfies*

$$\|DT(x_1) - DT(x_2)\|_{\mathrm{Lin}(X,Y)} \leq M(\|x_1\|, \|x_2\|)\big(\|x_1 - x_2\|_X^\alpha + \|x_1 - x_2\|_X\big) \quad \forall x_1, x_2 \in X,$$

where $0 < \alpha < 1$ and the continuous function M is bounded on every compact subset of $\mathbb{R}^2$. Then, we have the following Taylor expansion

$$T(x_2) - T(x_1) = DT(x_1)\cdot(x_2 - x_1) + R_{1+\alpha}(x_1, x_2) \quad \forall x_1, x_2 \in X,$$

where the remainder term can be bounded as follows:

$$\|R_{1+\alpha}(x_1, x_2)\|_Y \leq C\big(\|x_1\|_X, \|x_2\|_X\big)(\|x_1 - x_2\|_X^{1+\alpha} + \|x_1 - x_2\|_X^2).$$

Proof. For fixed x_1, $x_2 \in X$ we define the mapping $g : [0, 1] \to Y$ by $g(t) = T(tx_2+(1-t)x_1)$. Since T is Fréchet differentiable, so is g. By the Mean Value Theorem [124, Proposition 3.5] we have

$$T(x_2) - T(x_1) = g(1) - g(0) = \int_0^1 g'(s)\,\mathrm{d}s \;= \int_0^1 DT(sx_2 + (1 - s)x_1)\cdot(x_2 - x_1)\,\mathrm{d}s.$$

Hence, we get the desired Taylor expansion:

$$T(x_2) - T(x_1) = DT(x_1)\cdot(x_2 - x_1) + \int_0^1 \big(DT(sx_2 + (1 - s)x_1) - DT(x_1)\big)\cdot(x_2 - x_1)\,\mathrm{d}s.$$

The bound on the remainder term can now easily be obtained as follows:

$$\begin{aligned}\|R_{1+\alpha}(x_1,x_2)\|_Y &\leq \int_0^1 \big\|DT(sx_2+(1-s)x_1)-DT(x_1)\big\|\,\mathrm{d}s\|x_2-x_1\|_X\\ &\leq C(\|x_1\|,\|x_2\|)\int_0^1 \big(s^\alpha\|x_1-x_2\|_X^{1+\alpha}+s\|x_1-x_2\|_X^2\big)\,\mathrm{d}s\\ &\leq 2C(\|x_1\|,\|x_2\|)\big(\|x_1-x_2\|_X^{1+\alpha}+\|x_1-x_2\|_X^2\big),\end{aligned}$$

where we have used the triangle inequality and the continuity property of DT. □

The next lemma deals with the local invertibility of operators depending continuously on a Banach-space-valued parameter.

Lemma A.2. *Let P, X, Y be Banach spaces and $T : P \to \mathrm{Lin}(X,Y)$ a continuous, linear-operator-valued mapping. Let T satisfy*

$$\|T(p_2)-T(p_1)\|_{\mathrm{Lin}(X,Y)} \leq M(\|p_1\|_P,\|p_2\|_P)\,g(\|p_1-p_2\|_P) \quad \forall p_1,p_2\in P,$$

where $g:\mathbb{R}\to\mathbb{R}$ is a continuous, monotonously increasing function with $g(0)=0$ and $M:\mathbb{R}\times\mathbb{R}\to\mathbb{R}^+$ is continuous. Furthermore, let $T(p^*)$ be invertible with $\big\|T(p^*)^{-1}\big\|_{\mathrm{Lin}(Y,X)}\leq K$. Then, there exists a neighbourhood $B_\delta(p^*)\subset P$ such that $T(p)$ is invertible and

$$\big\|T(p)^{-1}\big\|_{\mathrm{Lin}(Y,X)} \leq 2K \quad \forall p\in B_\delta(p^*).$$

Proof. Let $p\in P$. We have

$$\begin{aligned}\big\|\mathrm{id}-T(p^*)^{-1}T(p)\big\|_{\mathrm{Lin}(X,X)} &\leq \big\|T(p^*)^{-1}\big\|_{\mathrm{Lin}(Y,X)}\,\|T(p^*)-T(p)\|_{\mathrm{Lin}(X,Y)}\\ &\leq KM(\|p^*\|_P,\|p\|_P)g(\|p^*-p\|_P).\end{aligned}$$

Let $\delta>0$ be sufficiently small such that $KM(\|p^*\|_P,\|p^*\|_P+\delta)g(\delta)<1/2$. Then, we get, using the Neumann series [123, Th. 2, Sec. II.1],

$$\big(T(p^*)^{-1}T(p)\big)^{-1}=\big[\mathrm{id}-\big(\mathrm{id}-T(p^*)^{-1}T(p)\big)\big]^{-1}=\sum_{k=0}^{\infty}\big(\mathrm{id}-T(p^*)^{-1}T(p)\big)^k,$$

for every $p\in B_\delta(p^*)$ with convergence of the series in the strong sense. This implies invertibility of $T(p)$ for all $p\in B_\delta(p^*)$ and furthermore

$$\begin{aligned}\big\|T(p)^{-1}\big\|_{\mathrm{Lin}(Y,X)} &\leq \sum_{k=0}^{\infty}\big\|\mathrm{id}-T(p^*)^{-1}T(p)\big\|_{\mathrm{Lin}(X,X)}^k\big\|T(p^*)^{-1}\big\|_{\mathrm{Lin}(Y,X)}\\ &\leq \frac{K}{1-\big\|\mathrm{id}-T(p^*)^{-1}T(p)\big\|_{\mathrm{Lin}(X,X)}}\leq 2K,\end{aligned}$$

as desired. □

Next, we state a useful, general existence result from [96, 97]. This represents a practical version of the Inverse Function Theorem.

Lemma A.3. *Let X, Y be Banach spaces, A an open subset of X, and let $\mathcal{F} : A \to Y$ be Fréchet differentiable. Suppose that $x_0 \in A$ satisfies the conditions*

$$\begin{aligned} \|\mathcal{F}(x_0)\|_Y &\le \eta, \\ \left\|D\mathcal{F}(x_0)^{-1}\right\|_{\mathrm{Lin}(X,Y)} &\le \vartheta, \\ \overline{B_{2\eta\vartheta}(x_0)} &\subset A, \\ \|D\mathcal{F}(x_1) - D\mathcal{F}(x_2)\|_{\mathrm{Lin}(X,Y)} &\le L\|x_1 - x_2\|_X \quad \textit{for} \quad \|x_i - x_0\|_X \le 2\eta\vartheta, \\ 2L\vartheta^2\eta &< 1. \end{aligned}$$

Then, there exists $x \in X$ such that $\mathcal{F}(x) = 0$ and $\|x - x_0\| \le 2\eta\vartheta$. □

A.2 An Indefinite Elliptic System with Constraint

In this section we generalize Schatz' classical result on the Galerkin discretization of indefinite elliptic equations [106] to the constrained system, which we first encountered in the analysis in Section 2.3.1; see the proof of Proposition 2.9.

A.2.1 The Dirichlet Case

We assume that $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, is a bounded, open domain in which the Poisson problem is $\mathrm{H}^2(\Omega)$-regular, that is, for any right-hand side $f \in \mathrm{L}^2(\Omega)$, the solution $u \in \mathrm{H}^1_0(\Omega)$ of

$$(\nabla u, \nabla v) = (f, v) \qquad \forall v \in \mathrm{H}^1_0(\Omega)$$

belongs to $\mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega)$ and

$$|u|_{\mathrm{H}^2} \le C_{\mathrm{reg}} \|f\|_{\mathrm{L}^2}. \tag{A.1}$$

With $d = 3$ this is exactly the assumption we made in Section 2.2.2.

Let $m \ge 1$ and let $\mathbf{V} \subset \mathrm{H}^1_0(\Omega; \mathbb{R}^m)$ be a subspace with co-dimension one. We assume that the linear constraint defining the subspace is given by a nonzero L^2-function $\boldsymbol{g} \in \mathrm{L}^2(\Omega; \mathbb{R}^m)$, so that $\mathbf{V}$ can be written as

$$\mathbf{V} = \big\{\boldsymbol{v} \in \mathrm{H}^1_0(\Omega; \mathbb{R}^m) : (\boldsymbol{v}, \boldsymbol{g}) = 0\big\}.$$

Let a be a bilinear form on $\mathrm{H}^1_0(\Omega; \mathbb{R}^m) \times \mathrm{H}^1_0(\Omega; \mathbb{R}^m)$ defined by

$$a(\boldsymbol{u}, \boldsymbol{v}) = (\nabla \boldsymbol{u}, \nabla \boldsymbol{v}) + (\boldsymbol{u}, \mathrm{M}\boldsymbol{v}) \qquad \text{for } \boldsymbol{u}, \boldsymbol{v} \in \mathrm{H}^1_0(\Omega; \mathbb{R}^m), \tag{A.2}$$

where $\mathrm{M} \in \mathrm{L}^\infty(\Omega; \mathbb{R}^{m \times m})$. We immediately see that a is bounded:

$$a(\boldsymbol{u}, \boldsymbol{v}) \le C_a \|\nabla \boldsymbol{u}\|_{\mathrm{L}^2} \|\nabla \boldsymbol{v}\|_{\mathrm{L}^2} \qquad \forall\, \boldsymbol{u}, \boldsymbol{v} \in \mathbf{V}, \tag{A.3}$$

and that there exists a constant $K \geq 0$ such that $a(\boldsymbol{u}, \boldsymbol{v}) + K(\boldsymbol{u}, \boldsymbol{v})$ is coercive,

$$a(\boldsymbol{u}, \boldsymbol{u}) + K\|\boldsymbol{u}\|_{\mathrm{L}^2}^2 \geq \alpha \, \|\nabla \boldsymbol{u}\|_{\mathrm{L}^2}^2 \qquad \forall \, \boldsymbol{u}, \boldsymbol{v} \in \mathbf{V}. \tag{A.4}$$

We assume that for every $\boldsymbol{f} \in \mathrm{L}^2(\Omega; \mathbb{R}^m)$ there is a unique solution $\boldsymbol{u} \in \mathbf{V}$ of the variational problem

$$a(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathbf{V}. \tag{A.5}$$

The adjoint variational problem to (A.5) is given by

$$a(\boldsymbol{v}, \boldsymbol{u}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathbf{V}.$$

Note that the adjoint problem has the same form except that M is replaced by M^T. It follows by an argument involving the Fredholm alternative[1], and which carries over verbatim from the scalar case (see Theorem 6.2.4 in [51]), that the adjoint problem also has a unique solution. This implies, in particular, the existence of $\kappa > 0$ such that

$$\inf_{\boldsymbol{u} \in \mathbf{V}} \sup_{\boldsymbol{v} \in \mathbf{V}} \frac{a(\boldsymbol{u}, \boldsymbol{v})}{\|\nabla \boldsymbol{u}\|_{\mathrm{L}^2} \, \|\nabla \boldsymbol{v}\|_{\mathrm{L}^2}} \geq \kappa \quad \text{and} \quad \inf_{\boldsymbol{u} \in \mathbf{V}} \sup_{\boldsymbol{v} \in \mathbf{V}} \frac{a(\boldsymbol{v}, \boldsymbol{u})}{\|\nabla \boldsymbol{u}\|_{\mathrm{L}^2} \, \|\nabla \boldsymbol{v}\|_{\mathrm{L}^2}} \geq \kappa.$$

From these inf-sup conditions we can infer the following bound on the solution $\boldsymbol{u}$ to (A.5):

$$\|\nabla \boldsymbol{u}\|_{\mathrm{L}^2} \leq \kappa^{-1} \sup_{\boldsymbol{v} \in \mathbf{V}} \frac{a(\boldsymbol{u}, \boldsymbol{v})}{\|\nabla \boldsymbol{v}\|_{\mathrm{L}^2}} \leq \kappa^{-1} \sup_{\boldsymbol{v} \in \mathbf{V}} \frac{(\boldsymbol{f}, \boldsymbol{v})}{\|\nabla \boldsymbol{v}\|_{\mathrm{L}^2}} \leq \kappa^{-1} \|\boldsymbol{f}\|_{\mathbf{V}^*}. \tag{A.6}$$

To obtain a Galerkin discretization of (A.5), let $(\mathrm{S}_{h,0})_{h \in (0,1]}$ be a family of finite-dimensional subspaces of $\mathrm{H}_0^1(\Omega)$, which satisfy the approximation property

$$\inf_{u_h \in \mathrm{S}_{h,0}} \|\nabla (u - u_h)\|_{\mathrm{L}^2} \leq C_{\mathrm{apx}} h |u|_{\mathrm{H}^2} \quad \text{for all } u \in \mathrm{H}_0^1(\Omega) \cap \mathrm{H}^2(\Omega), \tag{A.7}$$

for every h, where C_{apx} is independent of h. We then define the approximation space

$$\mathbf{V}_h = \big\{ \boldsymbol{v}_h \in \mathrm{S}_{h,0}^m : (\boldsymbol{g}, \boldsymbol{v}_h) = 0 \big\}.$$

The Galerkin discretization of (A.5) is given by

$$a(\boldsymbol{u}_h, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathbf{V}_h. \tag{A.8}$$

Note that it may occur that $\mathbf{V}_h = \mathrm{S}_{h,0}^m$. However, it follows immediately from (A.7) that, for sufficiently small h, the co-dimension of $\mathbf{V}_h$ in $\mathrm{S}_{h,0}^m$ is also one.

Our main result in this appendix ensures solvability of the Galerkin discretization (A.8), provided that h is sufficiently small. The proof parallels Theorem 5.7.6 in [17].

[1]Since $a(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v})$ for all $\boldsymbol{v}$ has a unique solution $\boldsymbol{u}$ for every $\boldsymbol{f} \in \mathrm{L}^2(\Omega, \mathbb{R}^m)$, one can deduce (using the Fredholm alternative for compact operators) that there is no solution $\boldsymbol{u} \neq 0$ of $a(\boldsymbol{u}, \boldsymbol{v}) = 0$ for all $\boldsymbol{v}$. Hence there is no solution $\boldsymbol{u} \neq 0$ of $a(\boldsymbol{v}, \boldsymbol{u}) = 0$ for all $\boldsymbol{v}$. This is equivalent to the existence and uniqueness of a solution to $a(\boldsymbol{v}, \boldsymbol{u}) = (\boldsymbol{f}, \boldsymbol{v})$ for all v.

Theorem A.4. *There exists $h_0 > 0$ such that, for every $h \in (0, h_0]$ and for every $\boldsymbol{f} \in \mathrm{L}^2(\Omega;\mathbb{R}^m)$, there is a unique solution $\boldsymbol{u}_h \in \boldsymbol{V}_h$ of the Galerkin discretization* (A.8)*, satisfying*

$$\|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\|_{\mathrm{L}^2} \le C_1 h |\boldsymbol{u}|_{\mathrm{H}^2} \quad \textit{and} \quad \|\boldsymbol{u}-\boldsymbol{u}_h\|_{\mathrm{L}^2} \le C_2 h^2 |\boldsymbol{u}|_{\mathrm{H}^2},$$

where $\boldsymbol{u}$ is the exact solution of (A.5)*. Furthermore, there exists $\kappa_d > 0$ such that,*

$$\inf_{\boldsymbol{v}_h \in \mathbf{V}_h} \sup_{\boldsymbol{w}_h \in \mathbf{V}_h} \frac{a(\boldsymbol{v}_h, \boldsymbol{w}_h)}{\|\nabla \boldsymbol{v}_h\|_{\mathrm{L}^2} \|\nabla \boldsymbol{w}_h\|_{\mathrm{L}^2}} \ge \kappa_d \qquad \forall h \in (0, h_0]. \tag{A.9}$$

Remark A.5. We remark that Theorem A.4 as well as the following auxiliary Lemmas hold for any finite number of linear constraints. For example, if $\mathbf{V}$ is given by

$$\mathbf{V} = \big\{ \boldsymbol{v} \in \mathrm{H}^1_0(\Omega;\mathbb{R}^m) : (\boldsymbol{g}_i, \boldsymbol{v}) = 0,\ i = 1, \dots, n \big\},$$

where $\boldsymbol{g}_i \in \mathrm{L}^2(\Omega;\mathbb{R}^m)$, and if the functions $\boldsymbol{g}_i$, $i = 1, \dots, n$, are linearly independent, then either minor modifications of the proofs, or simply a successive application of the results for a single constraint can establish the results for a finite number of constraints. □

The proof of Theorem A.4 requires two auxiliary results, which we provide in the following two lemmas. The first one shows that the constrained variational problem (A.5) inherits the H^2-regularity of the Laplace operator.

Lemma A.6. *Let $\boldsymbol{f} \in \mathrm{L}^2(\Omega;\mathbb{R}^m)$ and let $\boldsymbol{u} \in \mathbf{V}$ be the solution of* (A.5)*. Then, $\boldsymbol{u} \in \mathrm{H}^2(\Omega;\mathbb{R}^m)$ and*

$$|\boldsymbol{u}|_{\mathrm{H}^2} \le C'_{\mathrm{reg}} \|\boldsymbol{f}\|_{\mathrm{L}^2}.$$

Proof. The result follows by explicitly computing a representation of $\boldsymbol{u}$ in terms of the solution $\widetilde{\boldsymbol{u}} \in \mathrm{H}^1_0(\Omega;\mathbb{R}^m)$ of a Poisson problem without constraint:

$$(\nabla \widetilde{\boldsymbol{u}}, \nabla \boldsymbol{v}) = (\boldsymbol{f} - \mathrm{M}\boldsymbol{u}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathrm{H}^1_0(\Omega;\mathbb{R}^m).$$

Let $\boldsymbol{e}$ be the Riesz representation of $\boldsymbol{g}$ in $\mathrm{H}^1_0(\Omega)$,

$$(\nabla \boldsymbol{e}, \nabla \boldsymbol{v}) = (\boldsymbol{g}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathrm{H}^1_0(\Omega;\mathbb{R}^m);$$

then it follows that the function

$$\widetilde{\boldsymbol{u}} - t\boldsymbol{e}, \quad \text{where} \quad t = \frac{(\nabla \widetilde{\boldsymbol{u}}, \nabla \boldsymbol{e})}{\|\nabla \boldsymbol{e}\|^2_{\mathrm{L}^2}},$$

belongs to $\mathbf{V}$, and that

$$(\nabla(\widetilde{\boldsymbol{u}} - t\boldsymbol{e}), \nabla \boldsymbol{v}) = (\nabla \widetilde{\boldsymbol{u}}, \nabla \boldsymbol{v}) - t(\boldsymbol{g}, \boldsymbol{v}) = (\boldsymbol{f} - \mathrm{M}\boldsymbol{u}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathbf{V}.$$

Hence, we deduce that

$$\boldsymbol{u} = \widetilde{\boldsymbol{u}} - t\boldsymbol{e}.$$

Assumption (A.1) ensures that $\widetilde{\boldsymbol{u}}$ and $\boldsymbol{e}$ belong to $\mathrm{H}^2(\Omega;\mathbb{R}^m)$, and therefore, we can deduce that $\boldsymbol{u} \in \mathrm{H}^2(\Omega;\mathbb{R}^m)$. Furthermore,

$$|\boldsymbol{u}|_{\mathrm{H}^2} \leq |\widetilde{\boldsymbol{u}}|_{\mathrm{H}^2} + \frac{\|-\Delta\widetilde{\boldsymbol{u}}\|_{\mathrm{L}^2}\|\boldsymbol{e}\|_{\mathrm{L}^2}}{\|\nabla\boldsymbol{e}\|_{\mathrm{L}^2}^2}|\boldsymbol{e}|_{\mathrm{H}^2} \leq C(\|\boldsymbol{e}\|_{\mathrm{H}^2})|\widetilde{\boldsymbol{u}}|_{\mathrm{H}^2}.$$

H^2-regularity, (A.6), and Poincaré's inequality imply that

$$|\widetilde{\boldsymbol{u}}|_{\mathrm{H}^2} \leq C\big(\|\boldsymbol{f}\|_{\mathrm{L}^2} + \|\mathrm{M}\|_{\mathrm{L}^\infty}\|\boldsymbol{u}\|_{\mathrm{L}^2}\big) \leq C\|\boldsymbol{f}\|_{\mathrm{L}^2}.$$

Thus we conclude that

$$|\boldsymbol{u}|_{\mathrm{H}^2} \leq C'_{\mathrm{reg}}\|\boldsymbol{f}\|_{\mathrm{L}^2}$$

for some constant $C'_{\mathrm{reg}} > 0$. □

Our second auxiliary result shows that the constrained subspace $\mathbf{V}_h$ inherits the approximation property (A.7) of $\mathrm{S}_{h,0}$.

Lemma A.7. *There exists $h_0 > 0$ and a constant C'_{apx} such that, for $h \in (0, h_0]$,*

$$\inf_{\boldsymbol{u}_h \in \mathbf{V}_h} \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathrm{L}^2} \leq C'_{\mathrm{apx}}h|\boldsymbol{u}|_{\mathrm{H}^2} \quad \forall \boldsymbol{u} \in \mathbf{V} \cap \mathrm{H}^2(\Omega;\mathbb{R}^m).$$

Proof. Let $\boldsymbol{e}_h \in \mathrm{S}_{h,0}^m$ be the solution of

$$(\nabla\boldsymbol{e}_h, \nabla\boldsymbol{v}) = (\boldsymbol{g}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathrm{S}_{h,0}^m.$$

It is not difficult to verify, for h sufficiently small, say $h \in (0, h_0]$, that there exists $\boldsymbol{v} \in \mathrm{S}_{h,0}^m$ such that $(\boldsymbol{g}, \boldsymbol{v}) \neq 0$ and hence $\boldsymbol{e}_h \neq 0$ for $h \in (0, h_0]$. Let $\widetilde{\boldsymbol{u}}_h \in \mathrm{S}_{h,0}^m$ be the Ritz projection of $\boldsymbol{u}$, i.e.,

$$(\nabla\widetilde{\boldsymbol{u}}_h, \nabla\boldsymbol{v}) = (\nabla\boldsymbol{u}, \nabla\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathrm{S}_{h,0}^m.$$

We construct the final approximant $\boldsymbol{u}_h$, as in the proof of Lemma A.6,

$$\boldsymbol{u}_h = \widetilde{\boldsymbol{u}}_h - \frac{(\nabla\widetilde{\boldsymbol{u}}_h, \nabla\boldsymbol{e}_h)}{\|\nabla\boldsymbol{e}_h\|_{\mathrm{L}^2}^2}\boldsymbol{e}_h \in \mathbf{V}_h.$$

Since $(\boldsymbol{g}, \boldsymbol{u}) = 0$, the error $\|\nabla(\boldsymbol{u}_h - \widetilde{\boldsymbol{u}}_h)\|_{\mathrm{L}^2}$ can be estimated as follows:

$$\|\nabla(\boldsymbol{u}_h - \widetilde{\boldsymbol{u}}_h)\|_{\mathrm{L}^2} = \frac{|(\nabla\boldsymbol{e}_h, \nabla\widetilde{\boldsymbol{u}}_h)|}{\|\nabla\boldsymbol{e}_h\|_{\mathrm{L}^2}} = \frac{|(\boldsymbol{g}, \widetilde{\boldsymbol{u}}_h - \boldsymbol{u})|}{\|\nabla\boldsymbol{e}_h\|_{\mathrm{L}^2}} \leq \frac{\|\boldsymbol{g}\|_{\mathrm{H}^{-1}}}{\|\nabla\boldsymbol{e}_h\|_{\mathrm{L}^2}}\,\|\nabla(\boldsymbol{u} - \widetilde{\boldsymbol{u}}_h)\|_{\mathrm{L}^2}.$$

Since $\boldsymbol{e}_h$ converges in $\mathrm{H}_0^1(\Omega)$ to the Riesz representation of $\boldsymbol{g}$, the term $\|\boldsymbol{g}\|_{\mathrm{H}^{-1}}/\|\nabla\boldsymbol{e}_h\|_{\mathrm{L}^2}$ converges to one as $h \to 0$, and in particular is uniformly bounded on $(0, h_0]$, provided h_0 is

chosen sufficiently small. Invoking the approximation property (A.7) we arrive at the desired approximation result. ◻

Proof of Theorem A.4.

Step 1: The Schatz argument. The main part of the proof follows an argument originally given by Schatz in [106]. Our presentation is largely analogous to [17, Th. 5.7.6]. First, let us simply assume the existence of a discrete solution $\boldsymbol{u}_h$. From Galerkin orthogonality we get $a(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{u}-\boldsymbol{u}_h) = a(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{u}-\boldsymbol{v})$ for every $\boldsymbol{v} \in \mathbf{V}_h$. Then, using (A.4) we deduce that

$$\begin{aligned} \alpha \left\|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\right\|_{\mathrm{L}^2}^2 &\leq a(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{u}-\boldsymbol{u}_h) + K \left\|\boldsymbol{u}-\boldsymbol{u}_h\right\|_{\mathrm{L}^2}^2 \\ &\leq C_a \left\|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\right\|_{\mathrm{L}^2} \left\|\nabla(\boldsymbol{u}-\boldsymbol{v})\right\|_{\mathrm{L}^2} + K \left\|\boldsymbol{u}-\boldsymbol{u}_h\right\|_{\mathrm{L}^2}^2 , \end{aligned} \tag{A.10}$$

for every $\boldsymbol{v} \in \mathbf{V}_h$, where we have used the continuity of a. Considering the adjoint problem

$$a(\boldsymbol{z}, \boldsymbol{w}) = (\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{z}) \qquad \forall \boldsymbol{z} \in \mathbf{V}$$

and testing with $\boldsymbol{z} = \boldsymbol{u}-\boldsymbol{u}_h$ we can show that

$$\begin{aligned} \left\|\boldsymbol{u}-\boldsymbol{u}_h\right\|_{\mathrm{L}^2}^2 &= a(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{w}) \\ &= a(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{w}-\boldsymbol{w}_h) \\ &\leq C_a \left\|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\right\|_{\mathrm{L}^2} \left\|\nabla(\boldsymbol{w}-\boldsymbol{w}_h)\right\|_{\mathrm{L}^2} \\ &\leq C_a C'_{\mathrm{reg}} h \left\|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\right\|_{\mathrm{L}^2} |\boldsymbol{w}|_{\mathrm{H}^2} \\ &\leq C_a C'_{\mathrm{reg}} h \left\|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\right\|_{\mathrm{L}^2} \left\|\boldsymbol{u}-\boldsymbol{u}_h\right\|_{\mathrm{L}^2} , \end{aligned}$$

where Lemma A.6 was used to obtain $\boldsymbol{w} \in \mathrm{H}^2(\Omega, \mathbb{R}^m)$. This results in

$$\left\|\boldsymbol{u}-\boldsymbol{u}_h\right\|_{\mathrm{L}^2} \leq C_a C'_{\mathrm{reg}} h \left\|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\right\|_{\mathrm{L}^2} . \tag{A.11}$$

Applying this bound to (A.10) and choosing h sufficiently small, we obtain the bound

$$\left\|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\right\|_{\mathrm{L}^2} \leq C \left\|\nabla(\boldsymbol{u}-\boldsymbol{v})\right\|_{\mathrm{L}^2} \qquad \forall \boldsymbol{v} \in \mathbf{V}_h. \tag{A.12}$$

$\mathbf{V}_h$ is a finite-dimensional space, so existence of a solution $\boldsymbol{u}_h$ for arbitrary right-hand sides, and its uniqueness are equivalent. Suppose, for $\boldsymbol{f} = 0$, the discrete problem had a nontrivial solution $\boldsymbol{u}_h \neq 0$. Then, equation (A.12) would produce a contradiction for h sufficiently small because $\boldsymbol{u} = 0$. Hence, there is a unique solution $\boldsymbol{u}_h$.

Taking the infimum over $\boldsymbol{v} \in \mathbf{V}_h$ in (A.12) and using Lemma A.7 yields the first error bound stated in the theorem. Combining this bound with (A.11) provides the second error bound.

Step 2. Uniform Inf-Sup Constant. Unique solvability of (A.5) for $\boldsymbol{f} = \mathbf{0}$ implies that a satisfies the inf-sup condition

$$\inf_{\boldsymbol{u}\in\mathbf{V}} \sup_{\boldsymbol{v}\in\mathbf{V}} \frac{a(\boldsymbol{u}, \boldsymbol{v})}{\left\|\nabla \boldsymbol{u}\right\|_{\mathrm{L}^2} \left\|\nabla \boldsymbol{v}\right\|_{\mathrm{L}^2}} = \kappa > 0. \tag{A.13}$$

Our aim now is to prove the validity of a corresponding condition for $\mathbf{V}_h \times \mathbf{V}_h$, which is uniform in h.

To obtain $\mathcal{O}(h)$-convergence of $\boldsymbol{u}_h$ to $\boldsymbol{u}$ it had to be assumed that $\boldsymbol{f} \in \mathrm{L}^2(\Omega;\mathbb{R}^m)$. However, both the continuous and discrete solution $\boldsymbol{u} \in \mathbf{V}$, respectively $\boldsymbol{u}_h \in \mathbf{V}_h$, exist and are unique if $\boldsymbol{f} \in \mathrm{H}^{-1}(\Omega;\mathbb{R}^m)$,

$$a(\boldsymbol{u},\boldsymbol{v}) = \langle \boldsymbol{f},\boldsymbol{v}\rangle \quad \forall v \in \mathbf{V} \quad \text{and} \quad a(\boldsymbol{u}_h,\boldsymbol{v}) = \langle \boldsymbol{f},\boldsymbol{v}\rangle \quad \forall v \in \mathbf{V}_h.$$

We note that, to prove (A.12), we only used that $\boldsymbol{u}-\boldsymbol{u}_h \in \mathrm{L}^2(\Omega;\mathbb{R}^m)$ but not $\boldsymbol{f} \in \mathrm{L}^2(\Omega;\mathbb{R}^m)$. Hence, choosing $\boldsymbol{v} = 0$ in (A.12) and using a triangle inequality and the inf-sup condition (A.13) we deduce that

$$\begin{aligned}
\|\nabla \boldsymbol{u}_h\|_{\mathrm{L}^2} &\le \|\nabla \boldsymbol{u}\|_{\mathrm{L}^2} + \|\nabla(\boldsymbol{u}-\boldsymbol{u}_h)\|_{\mathrm{L}^2} \\
&\le (1+C)\,\|\nabla \boldsymbol{u}\|_{\mathrm{L}^2} \\
&\le (1+C)\kappa^{-1} \sup_{\boldsymbol{v}\in\mathbf{V}} \frac{\langle \boldsymbol{f},\boldsymbol{v}\rangle}{\|\nabla \boldsymbol{v}\|_{\mathrm{L}^2}} \\
&\le (1+C)\kappa^{-1}\|\boldsymbol{f}\|_{\mathbf{V}^*}
\end{aligned} \tag{A.14}$$

for all $h \in (0,h_0]$, where $h_0 > 0$ is chosen sufficiently small, and where C is the constant from (A.12), which is indeed bounded as $h \to 0$. If $\boldsymbol{f} \in \mathbf{V}_h^*$ then, by the Hahn–Banach theorem [104, Theorem 3.3], $\boldsymbol{f}$ can be extended to an element of $\mathbf{V}^*$ while preserving its norm, and hence, we obtain

$$\big\|\nabla(L_h^{-1}\boldsymbol{f})\big\|_{\mathrm{L}^2} \le (1+C)\kappa^{-1}\|\boldsymbol{f}\|_{\mathbf{V}_h^*} \qquad \forall \boldsymbol{f} \in \mathbf{V}_h^* \quad \forall h \in (0,h_0],$$

where L_h^{-1} denotes the solution operator for (A.8). This statement is equivalent to the uniform inf-sup condition (A.9) with $\kappa_d = \kappa/(1+C)$. ◻

A.2.2 The Periodic Case

We now briefly consider the previous elliptic system on a domain Ω with periodic boundary conditions. Let $m \ge 1$, $n_{\mathbf{V}} \ge 1$ and let $\mathbf{V} \subset \mathrm{H}^1_{\#}(\Omega;\mathbb{R}^m)$ be given by

$$\mathbf{V}_{\#} = \big\{\boldsymbol{v} \in \mathrm{H}^1_{\#}(\Omega;\mathbb{R}^m) : (\boldsymbol{g}_\ell,\boldsymbol{v}) = 0 \text{ for } \ell = 1,\dots,n_{\mathbf{V}}\big\}.$$

Let the bilinear form (A.2) now be defined on $\mathrm{H}^1_{\#}(\Omega;\mathbb{R}^m) \times \mathrm{H}^1_{\#}(\Omega;\mathbb{R}^m)$. We can easily prove that a is bounded:

$$a(\boldsymbol{u},\boldsymbol{v}) \le C_a\,\|\boldsymbol{u}\|_{\mathrm{H}^1}\,\|\boldsymbol{v}\|_{\mathrm{H}^1} \qquad \forall\, \boldsymbol{u},\boldsymbol{v} \in \mathbf{V}_{\#}, \tag{A.15}$$

and that there exists a constant $K \ge 0$ such that $a(\boldsymbol{u},\boldsymbol{v}) + K(\boldsymbol{u},\boldsymbol{v})$ is coercive (with respect to the H^1-norm,

$$a(\boldsymbol{u},\boldsymbol{u}) + K\|\boldsymbol{u}\|^2_{\mathrm{L}^2} \ge \alpha\,\|\boldsymbol{u}\|^2_{\mathrm{H}^1} \qquad \forall\, \boldsymbol{u},\boldsymbol{v} \in \mathbf{V}_{\#}. \tag{A.16}$$

Given $\boldsymbol{f} \in \mathrm{L}^2(\Omega, \mathbb{R}^m)$ we look at the variational problem: find $\boldsymbol{u} \in \mathbf{V}_\#$ such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathbf{V}_\#. \tag{A.17}$$

The following result is a version of the regularity result from Lemma A.6 for the periodic case.

Lemma A.8. *Let $\boldsymbol{f} \in \mathrm{L}^2(\Omega; \mathbb{R}^m)$ and let $\boldsymbol{u} \in \mathbf{V}_\#$ be the solution of* (A.17)*, then $\boldsymbol{u} \in \mathrm{H}^2(\Omega; \mathbb{R}^m)$ and*

$$|\boldsymbol{u}|_{\mathrm{H}^2} \leq C'_{\mathrm{reg}} \|\boldsymbol{f}\|_{\mathrm{L}^2}.$$

Proof. The method of proof is the same as in Lemma A.6, only the Ritz projection needs to be changed. Also we only develop the proof here for one linear constraint, as the generalization to more than one constraint is straightforward. Let $\widetilde{\boldsymbol{u}} \in \mathrm{H}^1_\#(\Omega; \mathbb{R}^m)$ be such that

$$(\nabla \widetilde{\boldsymbol{u}}, \nabla \boldsymbol{v}) + (\widetilde{\boldsymbol{u}}, \boldsymbol{v}) = (\boldsymbol{f} - \mathrm{M}\boldsymbol{u} + \boldsymbol{u}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathrm{H}^1_\#(\Omega; \mathbb{R}^m).$$

Let $\boldsymbol{e}$ be the Riesz representation of $\boldsymbol{g}$ in $\mathrm{H}^1_\#(\Omega; \mathbb{R}^m)$,

$$(\nabla \boldsymbol{e}, \nabla \boldsymbol{v}) + (\boldsymbol{e}, \boldsymbol{v}) = (\boldsymbol{g}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathrm{H}^1_\#(\Omega; \mathbb{R}^m);$$

it then follows that the function

$$\widetilde{\boldsymbol{u}} - t\boldsymbol{e}, \quad \text{where} \quad t = \frac{(\nabla \widetilde{\boldsymbol{u}}, \nabla \boldsymbol{e}) + (\widetilde{\boldsymbol{u}}, \boldsymbol{e})}{\|\boldsymbol{e}\|^2_{\mathrm{H}^1}},$$

belongs to $\mathbf{V}_\#$, and that

$$\begin{aligned}(\nabla(\widetilde{\boldsymbol{u}} - t\boldsymbol{e}), \nabla \boldsymbol{v}) &= (\nabla \widetilde{\boldsymbol{u}}, \nabla \boldsymbol{v}) - t\big((\boldsymbol{g}, \boldsymbol{v}) - (\boldsymbol{e}, \boldsymbol{v})\big) \\ &= (\boldsymbol{f} - \mathrm{M}\boldsymbol{u}, \boldsymbol{v}) + (\boldsymbol{u} - \widetilde{\boldsymbol{u}}, \boldsymbol{v}) + t(\boldsymbol{e}, \boldsymbol{v})\end{aligned}$$

for all $\boldsymbol{v} \in \mathbf{V}_\#$. Hence $\widetilde{\boldsymbol{u}} - t\boldsymbol{e}$ is the unique solution of

$$(\nabla(\widetilde{\boldsymbol{u}} - t\boldsymbol{e}), \nabla \boldsymbol{v}) + (\widetilde{\boldsymbol{u}} - t\boldsymbol{e}, \boldsymbol{v}) = (\boldsymbol{f} - \mathrm{M}\boldsymbol{u}, \boldsymbol{v}) + (\boldsymbol{u}, \boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathbf{V}_\#,$$

from which we deduce that $\boldsymbol{u} = \widetilde{\boldsymbol{u}} - t\boldsymbol{e}$.

Since both $\widetilde{\boldsymbol{u}}$ and $\boldsymbol{e}$ belong to $\mathrm{H}^2(\Omega; \mathbb{R}^m)$, we deduce that $\boldsymbol{u} \in \mathrm{H}^2(\Omega)$. The existence of C'_{reg} follows as in the Dirichlet case. ◻

We now consider two ways of discretizing the variational problem (A.17).

Periodic Finite Elements. In the finite element case we then define the approximation space

$$\mathbf{V}_{h,\#} = \big\{ \boldsymbol{v}_h \in \mathrm{S}^m_{h,\#} : (\boldsymbol{g}_\ell, \boldsymbol{v}_h) = 0 \text{ for all } \ell = 1, \dots, n_{\mathbf{V}} \big\},$$

where $\mathrm{S}_{h,\#}$ was introduced in (2.49). The Galerkin discretization of (A.17) is given by: find $\boldsymbol{u} \in \mathbf{V}_{h,\#}$ such that

$$a(\boldsymbol{u}_h, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathbf{V}_{h,\#}. \tag{A.18}$$

In direct correspondence with Lemma (A.7) we can prove the following approximation result:

$$\inf_{\boldsymbol{u}_h \in \mathbf{V}_{h,\#}} \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathrm{H}^1} \leq C^{\#}_{\mathrm{apx}} h |\boldsymbol{u}|_{\mathrm{H}^2} \quad \text{for all } \boldsymbol{u} \in \mathbf{V}_{\#} \cap \mathrm{H}^2(\Omega; \mathbb{R}^m) \text{ and for } h \in (0, h_0].$$

The only necessary modification in the proof compared with Lemma A.7 is to exchange the $\mathrm{H}^1_0(\Omega)$-scalar product with the $\mathrm{H}^1(\Omega)$-scalar product.

We can now state the analogon of Theorem A.4 for the periodic case.

Theorem A.9. *There exists $h_0 > 0$ such that, for every $h \in (0, h_0]$ and for every $\boldsymbol{f} \in \mathrm{L}^2(\Omega; \mathbb{R}^m)$, there is a unique solution $\boldsymbol{u}_h$ of the Galerkin discretization* (A.18)*, satisfying*

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathrm{H}^1} \leq C_1 h |\boldsymbol{u}|_{\mathrm{H}^2} \quad \textit{and} \quad \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathrm{L}^2} \leq C_2 h^2 |\boldsymbol{u}|_{\mathrm{H}^2},$$

where $\boldsymbol{u}$ is the exact solution of (A.17)*. Furthermore, there exists $\kappa_d > 0$ such that,*

$$\inf_{\boldsymbol{v}_h \in \mathbf{V}_{h,\#}} \sup_{\boldsymbol{w}_h \in \mathbf{V}_{h,\#}} \frac{a(\boldsymbol{v}_h, \boldsymbol{w}_h)}{\|\boldsymbol{v}_h\|_{\mathrm{H}^1} \|\boldsymbol{w}_h\|_{\mathrm{H}^1}} \geq \kappa_d \qquad \forall h \in (0, h_0]. \tag{A.19}$$

Proof. The proof follows the exact same lines as in the Dirichlet case. ◻

Fourier Method. In the Fourier case we define

$$\mathbf{V}_N = \big\{ \boldsymbol{v}_N \in \mathrm{S}^m_N : (\boldsymbol{g}_\ell, \boldsymbol{v}_N) = 0 \text{ for } \ell = 1, \dots, n_{\mathbf{V}} \big\},$$

and get the following Galerkin discretization of (A.17):

$$a(\boldsymbol{u}_N, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathbf{V}_N. \tag{A.20}$$

The necessary approximation result for the constraint space $\mathbf{V}_N$ is:

$$\inf_{\boldsymbol{u}_N \in \mathbf{V}_N} \|\boldsymbol{u} - \boldsymbol{u}_N\|_{\mathrm{H}^1} \leq C^{\#}_{\mathrm{apx}} N^{-1} |\boldsymbol{u}|_{\mathrm{H}^2} \quad \text{for all } \boldsymbol{u} \in \mathbf{V}_{\#} \cap \mathrm{H}^2_{\#}(\Omega; \mathbb{R}^m)$$

and for $N \geq N_0$. Again, the proof of this result is a slight variation of Lemma A.7.

Theorem A.10. *There exists $N_0 > 0$ such that, for every $N \geq N_0$ and for every $\boldsymbol{f} \in \mathrm{L}^2(\Omega; \mathbb{R}^m)$, there is a unique solution $\boldsymbol{u}_N$ of the Galerkin discretization* (A.20)*, satisfying*

$$\|\nabla(\boldsymbol{u} - \boldsymbol{u}_N)\|_{\mathrm{L}^2} \leq C_1 N^{-1} |\boldsymbol{u}|_{\mathrm{H}^2} \quad \textit{and} \quad \|\boldsymbol{u} - \boldsymbol{u}_N\|_{\mathrm{L}^2} \leq C_2 N^{-2} |\boldsymbol{u}|_{\mathrm{H}^2},$$

where $\boldsymbol{u}$ is the exact solution of (A.17)*. Furthermore, there exists $\kappa_d > 0$ such that,*

$$\inf_{\boldsymbol{v}_N \in \mathbf{V}_N} \sup_{\boldsymbol{w}_N \in \mathbf{V}_N} \frac{a(\boldsymbol{v}_N, \boldsymbol{w}_N)}{\|\boldsymbol{v}_N\|_{\mathrm{H}^1} \|\boldsymbol{w}_N\|_{\mathrm{H}^1}} \geq \kappa_d \qquad \forall N \geq N_0. \tag{A.21}$$

A.3 Some Results on Quadrature

This section provides useful results for the analysis of the discretization of the TFDW functional with numerical integration carried out in Chapter 3. We will cite some classical results and prove certain necessary extensions. The finite element space S_h is throughout assumed to be of p-th order with an appropriate approximation result. The family of triangulations $(\mathcal{T}_h)_{h\in(0,1]}$ is assumed to satisfy the quasi-uniformity condition (2.46). We will use the notation introduced in Chapters 2 and 3.

We first state a classical result taken from [37]. The first statement is identical to [37, Theorem 4.1.5], the second one is proved in [37, Theorem 4.1.6]. We recall the definitions (3.7) and (3.8) of the error functionals $e_T[g]$, $e_h[g]$, and $\widehat{e}\,[\widehat{g}]$.

Theorem A.11. *Assume the quadrature rule $\mathcal{Q}_h[\cdot]$ satisfies $\widehat{e}\,[\widehat{\varphi}] = 0$ for all $\widehat{\varphi} \in P_{2\mathrm{p}-2}(\widehat{T})$. Let $q > 1$ satisfy $\mathrm{p} - \frac{d}{q} > 0$. Then, there exists $C > 0$ independent of $T \in \mathcal{T}_h$ and $h \in (0,1]$ such that for all $f \in \mathrm{W}^{\mathrm{p},q}(T)$, and $v \in P_{\mathrm{p}}(T)$:*

$$\big|e_T[fv]\big| \leq C h_T^{\mathrm{p}} |T|^{1/2-1/q} \|f\|_{\mathrm{W}^{\mathrm{p},q}(T)} \|v\|_{\mathrm{H}^1(T)}. \tag{A.22}$$

If $f \in \mathrm{W}^{\mathrm{p},q}(T)$ for all $T \in \mathcal{T}_h$, we get

$$\left| \int_\Omega fv \,\mathrm{d}x \, - \mathcal{Q}_h[fv] \right| \leq C h^{\mathrm{p}} |\Omega|^{1/2-1/q} \Big(\sum_{T\in\mathcal{T}_h} \|f\|^q_{\mathrm{W}^{\mathrm{p},q}(T)} \Big)^{1/q} \|v\|_{\mathrm{H}^1} \,,$$

for all $v \in \mathrm{S}_h$.

In the remainder of this section we generalize some classical results on quadrature. The necessity for that arises from the presence of the Hölder continuous function F'' in the second derivative of the energy functional $\widetilde{E}_h$, respectively, the derivative $D\widetilde{\mathcal{F}}_h$ of the nonlinear optimality system. Before we address actual quadrature results we need some preparations in the form of approximation results with respect to Hölder norms and a version of the Bramble–Hilbert Lemma.

Let $\{r_i\}_{i=1,\dots,N_{\mathrm{p}}}$ be a basis of the dual space $P_{\mathrm{p}}(\Omega)^*$, where $P_{\mathrm{p}}(\Omega)$ is the space of polynomials of degree less than or equal p and let $N_{\mathrm{p}} = \dim P_{\mathrm{p}}(\Omega)$. By the Hahn–Banach Theorem [104, Theorem 3.3] we can extend the dual basis functionals r_i to bounded functionals on $\mathrm{H}^{\mathrm{p}+1}(\Omega)$, which we will again denote by r_i. Moreover, let $\mathcal{I}_{\mathrm{p}} : \mathrm{H}^{\mathrm{p}+1}(\Omega) \to P_{\mathrm{p}}(\Omega)$ be the respective interpolation operator. That is

$$r_i(v) = r_i(\mathcal{I}_{\mathrm{p}} v),$$

for all i and all $v \in \mathrm{H}^{\mathrm{p}+1}(\Omega)$.

The following approximation result can be compared with a result on Sobolev spaces, see [37, Th. 3.1.1] or [16, Lemma 6.2]. The proof is very similar and based on compact imbeddings of Sobolev spaces.

Proposition A.12. *Let $\Omega \subset \mathbb{R}^d$ be a bounded open domain.*
(i) Let $\mathrm{p} \in \mathbb{N}_0$ and $0 < \gamma < 1$. Then, there exists $C > 0$ such that for all $u \in \mathrm{C}^{\mathrm{p},\gamma}(\Omega)$:

$$\|u - \mathcal{I}_{\mathrm{p}} u\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} \leq C\, |u|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)}. \tag{A.23}$$

(ii) Let $0 < \gamma < \min(2 - d/2, 1)$ and $\mathrm{p} \geq 0$, then there exists $C > 0$ such that

$$\|u - \mathcal{I}_{\mathrm{p}} u\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} \leq C\, |u|_{\mathrm{H}^{\mathrm{p}+2}(\Omega)}, \tag{A.24}$$

for all $u \in \mathrm{H}^{\mathrm{p}+2}(\Omega)$

Proof. *Part (i).* By the Hahn–Banach Theorem [104, Theorem 3.3] we can extend the dual basis functionals r_i to bounded functionals on $\mathrm{C}^{\mathrm{p},\gamma}(\Omega)$, which we will again denote by r_i. Also, since the r_i, $i = 1, \ldots, N_{\mathrm{p}}$ form a basis of $P_{\mathrm{p}}^*(\Omega)$ we have for $v \in P_{\mathrm{p}}(\Omega)$ that $r_i(v) = 0$ for all $i = 1, \ldots, N_{\mathrm{p}}$ if and only if $v = 0$.

We will now prove that there exists $C > 0$ such that

$$\|v\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} \leq C \left(|v|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} + \sum_{i=1}^{N_{\mathrm{p}}} |r_i(v)| \right) \quad \text{for all } v \in \mathrm{C}^{\mathrm{p},\gamma}(\Omega). \tag{A.25}$$

Suppose (A.25) is false. Then there is a sequence $(v_j)_{j \in \mathbb{N}} \subset \mathrm{C}^{\mathrm{p},\gamma}(\Omega)$ such that $\|v_j\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} = 1$ and

$$\lim_{j \to \infty} \left(|v_j|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} + \sum_{i=1}^{N_{\mathrm{p}}} |r_i(v_j)| \right) = 0. \tag{A.26}$$

Obviously the sequence $(v_j)_{j \in \mathbb{N}}$ is bounded in $\mathrm{C}^{\mathrm{p},\gamma}(\Omega)$. Since the imbedding $\mathrm{C}^{\mathrm{p},\gamma}(\Omega) \hookrightarrow \mathrm{C}^{\mathrm{p}}(\Omega)$ is compact [1, Th. 1.34] there exists a subsequence (again named $(v_j)_{j \in \mathbb{N}}$) such that $v_j \to v$ in $\mathrm{C}^{\mathrm{p}}(\Omega)$ as $j \to \infty$ for some $v \in \mathrm{C}^{\mathrm{p}}(\Omega)$.

From (A.26) we can also deduce that $|v_j|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} \to 0$ as $j \to \infty$, which in fact implies that $v_j \to v$ strongly in $\mathrm{C}^{\mathrm{p},\gamma}(\Omega)$, $|v|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} = 0$ and $\lim_{j \to \infty} \|v_j\|_{\mathrm{C}^{\mathrm{p}}(\Omega)} = 1$. Hence, $\nabla^\beta v = \text{const.}$ for all multi-indices $\beta \in \mathbb{N}^d$ with $|\beta|_1 = \mathrm{p}$. We have thus shown that v is a polynomial of degree at most p. Equation (A.26) also shows that $r_i(v) = 0$ for all $i = 1, \ldots, N_{\mathrm{p}}$, which by the choice of the r_i means that $v = 0$. This is, however, a contradiction to $\|v_j\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} = 1$ for all $j \in \mathbb{N}$. Hence, equation (A.25) holds true.

Let $u \in \mathrm{C}^{\mathrm{p},\gamma}(\Omega)$ be arbitrary and $\mathcal{I}_{\mathrm{p}} u \in P_{\mathrm{p}}(\Omega)$ be its interpolant, i.e., $r_i(u - \mathcal{I}_{\mathrm{p}} u) = 0$ for all $i = 1, \ldots, N_{\mathrm{p}}$. Then, by inserting $u - \mathcal{I}_{\mathrm{p}} u$ for v in (A.25) we deduce that

$$\|u - \mathcal{I}_{\mathrm{p}} u\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} \leq C |u|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)},$$

since clearly $|\mathcal{I}_{\mathrm{p}} u|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} = 0$. This proves part (i) of the proposition.

Part (ii). To show this, we convince ourselves that (in analogy with (A.25)) there exists $C > 0$ such that

$$\|v\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} \leq C \left(|v|_{\mathrm{H}^{\mathrm{p}+2}(\Omega)} + \sum_{i=1}^{N_{\mathrm{p}}} |r_i(v)| \right),$$

for all $v \in \mathrm{H}^{\mathrm{p}+2}(\Omega)$. Again, we assume the inequality is false. This implies the existence of $(v_j)_{j\in\mathbb{N}} \subset \mathrm{H}^{\mathrm{p}+2}(\Omega)$ with $\|v_j\|_{\mathrm{H}^{\mathrm{p}+2}(\Omega)} = 1$ and

$$\frac{|v_j|_{\mathrm{H}^{\mathrm{p}+2}(\Omega)} + \sum_{i=1}^{N_{\mathrm{p}}} |f_i(v_j)|}{\|v_j\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)}} \to 0.$$

Using the (compact) imbedding $\mathrm{H}^{\mathrm{p}+2}(\Omega) \hookrightarrow \mathrm{C}^{\mathrm{p},\gamma}(\Omega)$ (see [1, Th. 6.3 Part III]) for $0 < \gamma < 2 - d/2$, we then get in particular $\|v_j\|_{\mathrm{C}^{\mathrm{p},\gamma}} \leq C$ for all j. Hence $|v_j|_{\mathrm{H}^{\mathrm{p}+2}(\Omega)} \to 0$ and $\sum_{i=1}^{N_{\mathrm{p}}} |r_i(v_j)| \to 0$. The compact imbedding $\mathrm{H}^{\mathrm{p}+2}(\Omega) \hookrightarrow \mathrm{H}^{\mathrm{p}+1}(\Omega)$ yields $v_j \to v$ in $\mathrm{H}^{\mathrm{p}+1}(\Omega)$ for some $v \in \mathrm{H}^{\mathrm{p}+1}(\Omega)$ for a subsequence $(v_j)_{j\in\mathbb{N}}$. Now, recalling $|v_j|_{\mathrm{H}^{\mathrm{p}+2}(\Omega)} \to 0$ we in fact get $v_j \to v$ in $\mathrm{H}^{\mathrm{p}+2}(\Omega)$ and $|v|_{\mathrm{H}^{\mathrm{p}+2}(\Omega)} = 0$. Hence, $v \in P_{\mathrm{p}+1}(\Omega)$. But since $|r_i(v_j)| \to 0$, we get $v = 0$, i.e., a contradiction to $\|v\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} = 1$.

The rest of the proof is analogous to part (i). ◻

Next, we provide two versions of the well-known Bramble–Hilbert Lemma for Sobolev, respectively, Hölder spaces.

Lemma A.13. *Let $\Omega \subset \mathbb{R}^n$ be a bounded open domain with Lipschitz-continuous boundary.*
(i) For $q \in [1, \infty)$ let r be a continuous linear functional on the space $\mathrm{W}^{\mathrm{p}+1,q}(\Omega)$ with the property

$$r(\psi) = 0 \quad \forall \psi \in P_{\mathrm{p}}(\Omega).$$

Then, there exists $C(\Omega) > 0$ such that

$$|r(v)| \leq C(\Omega)\|r\|_{\mathrm{W}^{\mathrm{p}+1,q}(\Omega)^*}|v|_{\mathrm{W}^{\mathrm{p}+1,q}} \quad \forall v \in \mathrm{W}^{\mathrm{p}+1,q}(\Omega).$$

(ii) For $\gamma \in (0, 1)$ let r be a continuous linear functional on the space $\mathrm{C}^{\mathrm{p},\gamma}(\Omega)$ such that

$$r(\psi) = 0 \quad \forall \psi \in P_{\mathrm{p}}(\Omega)$$

Then, there exists $C(\Omega, \gamma) > 0$ such that

$$|r(v)| \leq C(\Omega, \gamma)\|r\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)^*}|v|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)} \quad \forall v \in \mathrm{C}^{\mathrm{p},\gamma}(\Omega).$$

Proof. Part (i) of the result is just a restatement of Theorem 4.1.3 in [37]. Let $r \in \mathrm{C}^{\mathrm{p},\gamma}(\Omega)^*$. For given $v \in \mathrm{C}^{\mathrm{p},\gamma}(\Omega)$ we have $r(v) = r(v + \psi)$ for any $\psi \in P_{\mathrm{p}}(\Omega)$. Hence,

$$|r(v)| = |r(v + \psi)| \leq \|r\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)^*}\|v + \psi\|_{\mathrm{C}^{\mathrm{p},\gamma}(\Omega)}.$$

Choosing $\psi = -\mathcal{I}_{\mathrm{p}}v \in P_{\mathrm{p}}(\Omega)$ and using Proposition A.12, we obtain the desired result. ◻

We are now ready to prove the required results on quadrature errors. The following proposition shows that a quadrature rule that is exact for polynomials of order smaller or equal $2\mathrm{p}-1$ yields an error of order $\mathcal{O}(h^{2\mathrm{p}})$ if the integrand is in the union of certain Sobolev spaces over all elements.

Proposition A.14.
(i) Assume that the reference quadrature rule $\widehat{\mathcal{Q}}$ underlying $\mathcal{Q}_h$ satisfies $\widehat{e}[\widehat{\varphi}] = 0$ for all $\widehat{\varphi} \in P_{2\mathrm{p}-1}(\widehat{T})$. Let $q > 1$ such that $2\mathrm{p} - d/q > 0$. Then, there exists $C > 0$, independent of $T \in \mathcal{T}_h$ and $h \in (0, 1]$, such that:

$$\left|e_T[f]\right| \leq C h_T^{2\mathrm{p}} |T|^{1-1/q} \|f\|_{\mathrm{W}^{2\mathrm{p},q}(T)} \quad \forall f \in \mathrm{W}^{2\mathrm{p},q}(T). \tag{A.27}$$

If, moreover, $f \in \mathrm{W}^{2\mathrm{p},q}(T)$ for all $T \in \mathcal{T}_h$, then

$$\left| \int_\Omega f \,\mathrm{d}x \, - \mathcal{Q}_h[f] \right| \leq C h^{2\mathrm{p}} |\Omega|^{1-1/q} \left(\sum_{T \in \mathcal{T}_h} \|f\|^q_{\mathrm{W}^{2\mathrm{p},q}(T)} \right)^{1/q}.$$

(ii) Assume that $\widehat{\mathcal{Q}}$ satisfies $\widehat{e}[\widehat{\varphi}] = 0$ for all $\widehat{\varphi} \in P_1(\widehat{T})$. Then, there exists $C > 0$, independent of $T \in \mathcal{T}_h$ and $h \in (0, 1]$, such that for all $f \in \mathrm{C}^{1,\gamma}(T)$, with $0 < \gamma < 1$,

$$|e_T[f]| \leq C h_T^{1+\gamma} \det B_T |f|_{\mathrm{C}^{1,\gamma}(T)}.$$

Proof. (i) First we note that $e_T[f] = \det B_T \, \widehat{e}[\widehat{f}]$ for all continuous functions f. Let $f \in \mathrm{W}^{2\mathrm{p},q}(T)$ and denote by $\widehat{f} \in \mathrm{W}^{2\mathrm{p},q}(\widehat{T})$ the function given by $\widehat{f} = f \circ F_T$. The Sobolev Imbedding Theorem [1, Theorem 4.12] guarantees the compact imbedding $\mathrm{W}^{2\mathrm{p},q}(\widehat{T}) \hookrightarrow \mathrm{L}^\infty(\widehat{T})$. Hence,

$$\left|\widehat{e}[\widehat{f}]\right| \leq C \|\widehat{f}\|_{\mathrm{L}^\infty(\widehat{T})} \leq C \|\widehat{f}\|_{\mathrm{W}^{2\mathrm{p},q}(\widehat{T})} \quad \forall f \in \mathrm{W}^{2\mathrm{p},q}(\widehat{T}).$$

Since the quadrature rule $\mathcal{Q}_T$ is exact on $P_{2\mathrm{p}-1}(\widehat{T})$, part (i) of the Bramble–Hilbert Lemma A.13 gives

$$\left|\widehat{e}[\widehat{f}]\right| \leq C |\widehat{f}|_{\mathrm{W}^{2\mathrm{p},q}(\widehat{T})}.$$

The transformation rule

$$|\widehat{f}|_{\mathrm{W}^{2\mathrm{p},q}(\widehat{T})} \leq C h_T^{2\mathrm{p}} (\det B_T)^{-1/q} |f|_{\mathrm{W}^{2\mathrm{p},q}(T)},$$

then implies

$$\left|e_T[f]\right| \leq C h_T^{2\mathrm{p}} |T|^{1-1/q} |f|_{\mathrm{W}^{2\mathrm{p},q}(T)}.$$

Summing over all $T \in \mathcal{T}_h$ and applying Hölder's inequality yields

$$|e_h[f]\,| \leq C h^{2\mathrm{p}} |\Omega|^{1-1/q} \left(\sum_{T \in \mathcal{T}_h} \|f\|^q_{\mathrm{W}^{2\mathrm{p},q}(T)} \right)^{1/q}.$$

(ii) The structure of this proof is the same as for part (i): we start by observing that

$$\left|\widehat{e}[\widehat{f}]\right| \leq C \|\widehat{f}\|_{\mathrm{L}^\infty(\widehat{T})} \leq C \|\widehat{f}\|_{\mathrm{C}^{1,\gamma}(\widehat{T})}$$

for all $\widehat{f} \in \mathrm{C}^{1,\gamma}(\widehat{T})$. Since $\mathcal{Q}_T$ is exact on $P_1(\widehat{T})$, Lemma A.13 implies $\big|\widehat{e}[\widehat{f}]\big| \leq C|\widehat{f}|_{\mathrm{C}^{1,\gamma}(\widehat{T})}$. Together with the transformation bound (2.62) this yields

$$|e_T[f]| = \det B_T \big|\widehat{e}[\widehat{f}]\big| \leq C \det B_T |\widehat{f}|_{\mathrm{C}^{1,\gamma}(\widehat{T})} \leq C h^{1+\gamma} \det B_T |f|_{\mathrm{C}^{1,\gamma}(T)}$$

as desired. □

In the analysis of the energy functional with numerical integration $\widetilde{E}_h$ in Section 3.2 we have to deal with approximations of integrals of the form $\int_\Omega avw \,\mathrm{d}x$ with $a \in \mathrm{L}^\infty(\Omega)$ and $v, w \in \mathrm{S}_h$ arising in the first derivative $D\widetilde{\mathcal{F}}_h$. In the following proposition we first state a slight modification of Theorem 4.1.4 in [37]. The second part is a generalization of this result to the case when the function a is Hölder continuous rather than in $\mathrm{W}^{\mathrm{p},\infty}(T)$.

Proposition A.15. *Let the reference quadrature rule $\widehat{\mathcal{Q}}$ satisfy $\widehat{e}\,[\widehat{\varphi}] = 0$ for all $\widehat{\varphi} \in P_{2\mathrm{p}-1}(\widehat{T})$. Then, there exists $C > 0$, independent of $T \in \mathcal{T}_h$ and $h \in (0, 1]$, such that for all $a \in \mathrm{W}^{\mathrm{p},\infty}(T)$, and $v, w \in P_{\mathrm{p}}(T)$:*

$$\big|e_T[avw]\big| \leq C h_T^{\mathrm{p}} \|a\|_{\mathrm{W}^{\mathrm{p},\infty}(T)} \|v\|_{\mathrm{H}^{\mathrm{p}}(T)} \|w\|_{\mathrm{H}^1(T)}. \tag{A.28}$$

Furthermore, for every $0 < \gamma \leq 1$ there exists $C_\gamma > 0$ such that for all $a \in \mathrm{C}^{\mathrm{p}-1,\gamma}(T)$, and $v, w \in P_{\mathrm{p}}(T)$:

$$\big|e_T[avw]\big| \leq C h_T^{\mathrm{p}-1+\gamma} \|a\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} \|v\|_{\mathrm{H}^{\mathrm{p}}(T)} \|w\|_{\mathrm{H}^1(T)}. \tag{A.29}$$

Proof. We only present the proof of equation (A.29) as (A.28) follows from (A.29) with $\gamma = 1$. The proof is similar to [37, Th. 4.1.4]. We begin by transforming from T to the reference element $\widehat{T}$. Obviously, $e_T[avw] = \det B_T\, \widehat{e}\,[\widehat{a}\widehat{v}\widehat{w}]$.

Let $\widehat{w} \in P_{\mathrm{p}}(\widehat{T})$ and $\widehat{\varphi} \in \mathrm{C}^{\mathrm{p}-1,\gamma}(\widehat{T})$. Then,

$$\begin{aligned}\big|\widehat{e}\,[\widehat{\varphi}\widehat{w}]\big| &= \left| \int_{\widehat{T}} \widehat{\varphi}\widehat{w} \,\mathrm{d}x - \sum_q \omega_q \widehat{\varphi}(\widehat{x}_q)\widehat{w}(\widehat{x}_q) \right| \\ &\leq C\|\widehat{\varphi}\|_{\mathrm{L}^\infty(\widehat{T})} \|\widehat{w}\|_{\mathrm{L}^\infty(\widehat{T})} \leq C\|\widehat{\varphi}\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(\widehat{T})} \|\widehat{w}\|_{\mathrm{L}^\infty(\widehat{T})}.\end{aligned}$$

Since $\widehat{w} \in P_{\mathrm{p}}(\widehat{T})$ and all norms are equivalent on the finite dimensional space $P_{\mathrm{p}}(\widehat{T})$, we obtain $\big|\widehat{e}\,[\widehat{\varphi}\widehat{w}]\big| \leq C\|\widehat{\varphi}\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(\widehat{T})} \|\widehat{w}\|_{\mathrm{L}^2(\widehat{T})}$.

Thus, for given $\widehat{w} \in P_{\mathrm{p}}(\widehat{T})$ the linear form $\widehat{\varphi} \mapsto \widehat{e}\,[\widehat{\varphi}\widehat{w}]$ from $\mathrm{C}^{\mathrm{p}-1,\gamma}(\widehat{T})$ to $\mathbb{R}$ is continuous with norm $\leq C\|\widehat{w}\|_{\mathrm{L}^2(\widehat{T})}$ and vanishes on $P_{\mathrm{p}-1}(\widehat{T})$ since, by assumption, the quadrature rule is exact on $P_{2\mathrm{p}-1}(\widehat{T})$. Using part (ii) of the Bramble–Hilbert Lemma A.13 we obtain

$$\big|\widehat{e}\,[\widehat{\varphi}\,\widehat{w}]\big| \leq C|\widehat{\varphi}|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(\widehat{T})} \|\widehat{w}\|_{\mathrm{L}^2(\widehat{T})}. \tag{A.30}$$

Now, let $\widehat{\varphi} = \widehat{a}\widehat{v}$ with $\widehat{v} \in P_{\mathrm{p}}(\widehat{T})$. A simple calculation shows that

$$\begin{aligned} |\widehat{a}\widehat{v}|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(\widehat{T})} &\leq C \sum_{j=0}^{\mathrm{p}-1} \Big(|\widehat{a}|_{\mathrm{C}^{j,\gamma}(\widehat{T})} |\widehat{v}|_{\mathrm{W}^{\mathrm{p}-1-j,\infty}(\widehat{T})} + |\widehat{a}|_{\mathrm{W}^{j,\infty}(\widehat{T})} |\widehat{v}|_{\mathrm{C}^{\mathrm{p}-1-j,\gamma}(\widehat{T})} \Big) \\ &\leq C \sum_{j=0}^{\mathrm{p}-1} \Big(|\widehat{a}|_{\mathrm{C}^{j,\gamma}(\widehat{T})} |\widehat{v}|_{\mathrm{H}^{\mathrm{p}-1-j}(\widehat{T})} + |\widehat{a}|_{\mathrm{W}^{j,\infty}(\widehat{T})} |\widehat{v}|_{\mathrm{H}^{\mathrm{p}-j}(\widehat{T})} \Big) . \end{aligned} \tag{A.31}$$

Here we have used that all norms on the finite dimensional spaces $P_{\mathrm{p}}(\widehat{T})$, respectively, on the quotient spaces $P_j(\widehat{T})/P_{j-1}(\widehat{T})$ are equivalent. Now we transform back to the mesh element T using the following inequalities:

$$|\widehat{a}|_{\mathrm{C}^{j,\gamma}(\widehat{T})} \leq C h^{j+\gamma} |a|_{\mathrm{C}^{j,\gamma}(T)}, \quad |\widehat{v}|_{\mathrm{H}^{j}(\widehat{T})} \leq C h^{j} (\det B_T)^{-1/2} |v|_{\mathrm{H}^{j}(T)}$$

and

$$|\widehat{a}|_{\mathrm{W}^{j,\infty}(\widehat{T})} \leq C h^{j} |a|_{\mathrm{W}^{j,\infty}(T)}, \qquad \|\widehat{w}\|_{\mathrm{L}^2(\widehat{T})} \leq C (\det B_T)^{-1/2} \|w\|_{\mathrm{L}^2(T)}.$$

Applying these estimates and (A.31) to (A.30) we arrive at

$$\big| e_T[avw] \big| = \det B_T \big| \widehat{e}\,[\widehat{a}\widehat{v}\widehat{w}] \big| \leq C h^{\mathrm{p}-1+\gamma} \|a\|_{\mathrm{C}^{\mathrm{p}-1,\gamma}(T)} \|v\|_{\mathrm{H}^{\mathrm{p}}(T)} \|w\|_{\mathrm{H}^{1}(T)},$$

just as desired. □

Bibliography

[1] R. A. Adams and J. F. Fournier. *Sobolev Spaces*. Pure and Applied Mathematics Series. Academic Press, Amsterdam, second edition, 2003. International Series in Pure and Applied Mathematics.

[2] M. Anitescu, D. Negrut, P. Zapol, and A. El-Azab. A note on the regularity of reduced models obtained by nonlocal quasi-continuum-like approaches. *Mathematical Programming*, **118**(2):207–236, 2009.

[3] I. Babuška and J. Osborn. Eigenvalue problems. In *Handbook of numerical analysis, Vol. II*, Handb. Numer. Anal., II, pages 641–787. North-Holland, Amsterdam, 1991.

[4] S. Badia, P. Bochev, J. Fish, M. Gunzburger, R. Lehoucq, M. Nuggehally, and M. L. Parks. A force-based blending model for atomistic-to-continuum coupling. *International Journal for Multiscale Computational Engineering*, **5**(5):387–406, 2007.

[5] S. Badia, M. L. Parks, P. B. Bochev, M. Gunzburger, and R. B. Lehoucq. On atomistic-to-continuum (AtC) coupling by blending. *SIAM J. Multiscale Modeling & Simulation*, **7**(1):381–406, 2008.

[6] W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003.

[7] S. Bartels, C. Carstensen, and G. Dolzmann. Inhomogeneous Dirichlet conditions in a priori and a posteriori finite element error analysis. *Numer. Math.*, **99**(1):1–24, 2004.

[8] R. Benguria. Dependence of the thomas–fermi energy on the nuclear coordinates. *Communications in Mathematical Physics*, **81**:419–428, 1981. 10.1007/BF01209076.

[9] R. Benguria, H. Brézis, and E. H. Lieb. The Thomas–Fermi–von Weizsäcker theory of atoms and molecules. *Comm. Math. Phys.*, **79**(2):167–180, 1981.

[10] R. Benguria and E. H. Lieb. Many-body atomic potentials in Thomas–Fermi theory. *Ann. Physics*, **110**(1):34–45, 1978.

[11] X. Blanc and E. Cances. Nonlinear instability of density-independent orbital-free kinetic-energy functionals. *The Journal of chemical physics*, **122**:214106, 2005.

[12] X. Blanc, C. Le Bris, and F. Legoll. Analysis of a prototypical multiscale method coupling atomistic and continuum mechanics. *M2AN Math. Model. Numer. Anal.*, **39**(4):797–826, 2005.

[13] X. Blanc, C. Le Bris, and F. Legoll. Analysis of a prototypical multiscale method coupling atomistic and continuum mechanics: the convex case. *Acta Math. Appl. Sin. Engl. Ser.*, **23**(2):209–216, 2007.

[14] X. Blanc, C. Le Bris, and P.-L. Lions. From molecular models to continuum mechanics. *Arch. Ration. Mech. Anal.*, **164**(4):341–381, 2002.

[15] X. Blanc, C. Le Bris, and P.-L. Lions. Atomistic to continuum limits for computational materials science. *M2AN Math. Model. Numer. Anal.*, **41**(2):391–426, 2007.

[16] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, third edition, 2007. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker.

[17] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, **15** of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002.

[18] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, **15** of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[19] F. Brezzi, J. Rappaz, and P.-A. Raviart. Finite-dimensional approximation of nonlinear problems. I. Branches of nonsingular solutions. *Numer. Math.*, **36**(1):1–25, 1980/81.

[20] E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of the planewave discretization of orbital-free and Kohn–Sham models part i: The Thomas–Fermi–von Weizsäcker model. *arXiv:0909.1464v1*, 2009.

[21] E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of nonlinear eigenvalue problems. *Journal of Scientific Computing*, **45**(1):90–117, 2010.

[22] E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models. *arXiv:1003.1612*, 2010.

[23] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. Computational quantum chemistry: a primer. In *Handbook of numerical analysis, Vol. X*, Handb. Numer. Anal., X, pages 3–270. North-Holland, Amsterdam, 2003.

[24] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods.* Scientific Computation. Springer-Verlag, Berlin, 2006. Fundamentals in single domains.

[25] C. Canuto and A. Quarteroni. Approximation results for orthogonal polynomials in Sobolev spaces. *Math. Comp.*, **38**(157):67–86, 1982.

[26] I. Catto, C. Le Bris, and P.-L. Lions. *The mathematical theory of thermodynamic limits: Thomas-Fermi type models.* Oxford Mathematical Monographs. The Clarendon Press Oxford University Press, New York, 1998.

[27] I. Catto and P.-L. Lions. Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories. I. A necessary and sufficient condition for the stability of general molecular systems. *Comm. Partial Differential Equations*, **17**(7-8):1051–1110, 1992.

[28] I. Catto and P.-L. Lions. Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories. II. Stability is equivalent to the binding of neutral subsystems. *Comm. Partial Differential Equations*, **18**(1-2):305–354, 1993.

[29] I. Catto and P.-L. Lions. Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories. III. Binding of neutral subsystems. *Comm. Partial Differential Equations*, **18**(3-4):381–429, 1993.

[30] I. Catto and P.-L. Lions. Binding of atoms and stability of molecules in Hartree and Thomas-Fermi type theories. IV. Binding of neutral systems for the Hartree model. *Comm. Partial Differential Equations*, **18**(7-8):1149–1159, 1993.

[31] D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, **45**(7):566–569, Aug 1980.

[32] H. Chen, X. Gong, L. He, and A. Zhou. Convergence of adaptive finite element approximations for nonlinear eigenvalue problems. *arXiv:1001.2344*, 2010.

[33] H. Chen, X. Gong, and A. Zhou. Numerical approximations of a nonlinear eigenvalue problem and applications to a density functional model. *Mathematical Methods in the Applied Sciences*, **33**(14):1723–1742, 2010.

[34] H. Chen and A. Zhou. Orbital-free density functional theory for molecular structure calculations. *Numer. Math. Theory Methods Appl.*, **1**(1):1–28, 2008.

[35] N. Choly and E. Kaxiras. Kinetic energy density functionals for non-periodic systems. *Solid State Communications*, **121**(5):281–286, 2002.

[36] N. Choly, G. Lu, W. E, and E. Kaxiras. Multiscale simulations in simple metals: A density-functional-based methodology. *Phys. Rev. B*, **71**(9):094101, Mar 2005.

[37] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam, 1978. Studies in Mathematics and its Applications, Vol. 4.

[38] W. A. Curtin and R. E. Miller. Atomistic/continuum coupling in computational materials science. *Modelling and simulation in materials science and engineering*, **11**:R33, 2003.

[39] B. Dacorogna. *Direct methods in the calculus of variations*, **78** of *Applied Mathematical Sciences*. Springer, New York, second edition, 2008.

[40] P. A. M. Dirac. Note on Exchange Phenomena in the Thomas Atom. In *Proceedings of the Cambridge Philosophical Society*, **26**, page 376, 1930.

[41] M. Dobrowolski and R. Rannacher. Finite element methods for nonlinear elliptic systems of second order. *Math. Nachr.*, **94**:155–172, 1980.

[42] M. Dobson and M. Luskin. Analysis of a force-based quasicontinuum approximation. *M2AN Math. Model. Numer. Anal.*, **42**(1):113–139, 2008.

[43] M. Dobson and M. Luskin. An analysis of the effect of ghost force oscillation on quasicontinuum error. *M2AN Math. Model. Numer. Anal.*, **43**(3):591–604, 2009.

[44] M. Dobson and M. Luskin. An optimal order error analysis of the one-dimensional quasicontinuum approximation. *SIAM J. Numer. Anal.*, **47**(4):2455–2475, 2009.

[45] M. Dobson, M. Luskin, and C. Ortner. Sharp stability estimates for the accurate prediction of instabilities by the quasicontinuum method. *Arxiv preprint arXiv:0905.2914*, 2009.

[46] M. Dobson, M. Luskin, and C. Ortner. Sharp stability estimates for the force-based quasicontinuum approximation of homogeneous tensile deformation. *Multiscale Model. Simul.*, **8**(3):782–802, 2010.

[47] M. Dobson, M. Luskin, and C. Ortner. Stability, instability, and error of the force-based quasicontinuum approximation. *Arch. Ration. Mech. Anal.*, **197**(1):179–202, 2010.

[48] M. Dobson, C. Ortner, and A. V. Shapeev. The Spectrum of the Force-Based Quasicontinuum Operator for a Homogeneous Periodic Chain. *Arxiv preprint arXiv:1004.3435*, 2010.

[49] W. E and J. Lu. The continuum limit and QM-continuum approximation of quantum mechanical models of solids. *Commun. Math. Sci.*, **5**(3):679–696, 2007.

[50] W. E, J. Lu, and J. Z. Yang. Uniform accuracy of the quasicontinuum method. *Phys. Rev. B*, **74**(21):214115, Dec 2006.

[51] L. C. Evans. *Partial differential equations*, **19** of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.

[52] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[53] J. L. Fattebert, R. D. Hornung, and A. M. Wissink. Finite element approach for density functional theory calculations on locally-refined meshes. *Journal of Computational Physics*, **223**(2):759–773, 2007.

[54] E. Fermi. Un metodo statistico per la determinazione di alcune priorieta dell'atome. *Rend. Accad. Naz. Lincei*, **6**(602-607):32, 1927.

[55] M. Finnis. *Interatomic Forces in Condensed Matter*. Oxford University Press, USA, 2003.

[56] J. Fish, M. A. Nuggehally, M. S. Shephard, C. R. Picu, S. Badia, M. L. Parks, and M. Gunzburger. Concurrent AtC coupling based on a blend of the continuum stress and the atomistic force. *Computer Methods in Applied Mechanics and Engineering*, **196**(45-48):4548–4560, 2007.

[57] C. J. García-Cervera. An efficient real space method for orbital-free density-functional theory. *Commun. Comput. Phys.*, **2**(2):334–357, 2007.

[58] C. J. García-Cervera. A remark on "an efficient real space method for orbital-free density-functional theory". *Commun. Comput. Phys.*, **3**(4):968–972, 2008.

[59] C. J. García-Cervera, J. Lu, and W. E. A sub-linear scaling algorithm for computing the electronic structure of materials. *Commun. Math. Sci.*, **5**(4):999–1026, 2007.

[60] V. Gavini. Configurational Forces in Field Formulation of Quasicontinuum. Unpublished manuscript. 2009.

[61] V. Gavini, K. Bhattacharya, and M. Ortiz. Quasi-continuum orbital-free density-functional theory: a route to multi-million atom non-periodic DFT calculation. *J. Mech. Phys. Solids*, **55**(4):697–718, 2007.

[62] V. Gavini, J. Knap, K. Bhattacharya, and M. Ortiz. Non-periodic finite-element formulation of orbital-free density functional theory. *J. Mech. Phys. Solids*, **55**(4):669–696, 2007.

[63] S. K. Ghosh and A. K. Dhara. Density-functional theory of two-dimensional electron gas in a magnetic field. *Physical Review A*, **40**(10):6103–6106, 1989.

[64] P. Grisvard. *Elliptic problems in nonsmooth domains*, **24** of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.

[65] R. L. Hayes, M. Fago, M. Ortiz, and E. A. Carter. Prediction of dislocation nucleation during nanoindentation by the orbital-free density functional theory local quasi-continuum method. *Multiscal Modeling Aan Simulation*, **4**(2):359, 2005.

[66] R. L. Hayes, G. Ho, M. Ortiz, and E. A. Carter. Prediction of dislocation nucleation during nanoindentation of Al_3Mg by the orbital-free density functional theory local quasicontinuum method. *Philosophical Magazine*, **86**(16):2343–2358, 2006.

[67] G. S. Ho, V. L. Lignères, and E. A. Carter. Introducing PROFESS: A new program for orbital-free density functional theory calculations. *Computer physics communications*, **179**(11):839–854, 2008.

[68] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev. (2)*, **136**:B864–B871, 1964.

[69] L. Hung and E. A. Carter. Accurate simulations of metals at the mesoscale: Explicit treatment of 1 million atoms with quantum mechanics. *Chemical Physics Letters*, **475**(4-6):163–170, 2009.

[70] M. Iyer and V. Gavini. A field theoretical approach to the Quasi-Continuum method. Unpublished Manuscript. 2010.

[71] R. O. Jones and O. Gunnarsson. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.*, **61**(3):689–746, 1989.

[72] E. Kaxiras. *Atomic and electronic structure of solids*. Cambridge University Press, Cambridge, 2003.

[73] W. Kohn. Nobel Lecture: Electronic structure of matterwave functions and density functionals. *Reviews of Modern Physics*, **71**(5):1253–1266, 1999.

[74] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev*, **140**(4A):A1133–A1138, 1965.

[75] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, **6**(1):15 – 50, 1996.

[76] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, **54**(16):11169–11186, 1996.

[77] C. Le Bris. Computational chemistry from the perspective of numerical analysis. *Acta Numer.*, **14**:363–444, 2005.

[78] C. Le Bris and P.-L. Lions. From atoms to crystals: a mathematical journey. *Bull. Amer. Math. Soc. (N.S.)*, **42**(3):291–363 (electronic), 2005.

[79] X.H. Li and M. Luskin. A generalized quasi-nonlocal atomistic-to-continuum coupling method with finite range interaction. *Arxiv preprint arXiv:1007.2336*, 2010.

[80] E. H. Lieb. Thomas–Fermi and related theories of atoms and molecules. *Rev. Modern Phys.*, **53**(4):603–641, 1981.

[81] E. H. Lieb and B. Simon. The Thomas–Fermi theory of atoms, molecules and solids. *Advances in Math.*, **23**(1):22–116, 1977.

[82] V. L. Lignère and A. C. Carter. An introduction to orbital-free density functional theory. In S. Yip, editor, *Handbook of Materials Modeling*, **15** of *Texts in Applied Mathematics*, chapter 1.8. Springer-Verlag, second edition, 2005.

[83] P. Lin. Theoretical and numerical analysis for the quasi-continuum approximation of a material particle model. *Math. Comp.*, **72**(242):657–675 (electronic), 2003.

[84] P. Lin. Convergence analysis of a quasi-continuum approximation for a two-dimensional material without defects. *SIAM J. Numer. Anal.*, **45**(1):313–332 (electronic), 2007.

[85] P. Lin and A. V. Shapeev. Energy-based ghost force removing techniques for the quasicontinuum method. *Arxiv preprint arXiv:0909.5437*, 2009.

[86] G. Lu, E. B. Tadmor, and E. Kaxiras. From electrons to finite elements: A concurrent multiscale approach for metals. *Phys. Rev. B*, **73**(2):024108, Jan 2006.

[87] M. Luskin and C. Ortner. An analysis of node-based cluster summation rules in the quasicontinuum method. *SIAM J. Numer. Anal.*, **47**(4):3070–3086, 2009.

[88] R. M. Martin. *Electronic structure: basic theory and practical methods*. Cambridge University Press, Cambridge, 2004.

[89] H. Maurer and J. Zowe. First and second order necessary and sufficient optimality conditions for infinite-dimensional programming problems. *Math. Programming*, **16**(1):98–110, 1979.

[90] R. Miller, E. B. Tadmor, R. Phillips, and M. Ortiz. Quasicontinuum simulation of fracture at the atomic scale. *Modelling and Simulation in Materials Science and Engineering*, **6**:607, 1998.

[91] R. E. Miller and E. B. Tadmor. The quasicontinuum method: overview, applications and current directions. *Journal of Computer-Aided Materials Design*, **9**(3):203–239, 2002.

[92] R. E. Miller and E. B. Tadmor. A unified framework and performance benchmark of fourteen multiscale atomistic/continuum coupling methods. *Modelling Simul. Mater. Sci. Eng*, **17**(053001):053001, 2009.

[93] P. Ming and J. Z. Yang. Analysis of a one-dimensional nonlocal quasi-continuum method. *Multiscale Model. Simul.*, **7**(4):1838–1875, 2009.

[94] D. Negrut, M. Anitescu, A. El-Azab, and P. Zapol. Quasicontinuum-like reduction of density functional theory calculations of nanostructures. *Journal of Nanoscience and Nanotechnology*, **8**(7):3729–3740, 2008.

[95] D. Negrut, M. Anitescu, T. Munson, and P. Zapol. Simulating nanoscale processes in solids using DFT and the quasicontinuum method (IMECE2005-81755). *Proceedings of ASME International Mechanical Engineering Congress and Exposition (IMECE) 2005.*

[96] C. Ortner. A priori and a posteriori analysis of the quasi-nonlocal Quasicontinuum Method in 1D. *Arxiv preprint arXiv:0911.0671*, 2009.

[97] C. Ortner. A posteriori existence in numerical computations. *SIAM J. Numer. Anal.*, **47**(4):2550–2577, 2009.

[98] C. Ortner and E. Süli. Analysis of a quasicontinuum method in one dimension. *M2AN Math. Model. Numer. Anal.*, **42**(1):57–91, 2008.

[99] C. Ortner and H. Wang. A priori error estimates for energy-based quasicontinuum approximations of a periodic chain. *to appear in Math. Models Methods Appl. Sci.*

[100] R. G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, USA, 1989.

[101] Q. Peng, X. Zhang, L. Hung, E. A. Carter, and G. Lu. Quantum simulation of materials at micron scales and beyond. *Physical Review B*, **78**(5):54118, 2008.

[102] J. P. Perdew and A. Zunger. Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B*, **23**(10):5048–5079, May 1981.

[103] W. Rudin. *Principles of mathematical analysis*. McGraw-Hill Book Co., New York, third edition, 1976. International Series in Pure and Applied Mathematics.

[104] W. Rudin. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., New York, second edition, 1991.

[105] Y. Saad, J. R. Chelikowsky, and S. M. Shontz. Numerical methods for electronic structure calculations of materials. *SIAM Rev.*, **52**(1):3–54, 2010.

[106] A. H. Schatz. An observation concerning Ritz–Galerkin methods with indefinite bilinear forms. *Math. Comp.*, **28**:959–962, 1974.

[107] A. V. Shapeev. Consistent Energy-Based Atomistic/Continuum Coupling for Two-Body Potential: 1D and 2D Case. *Arxiv preprint arXiv:1010.0512*, 2010.

[108] V. B. Shenoy, R. Miller, E. Tadmor, D. Rodney, R. Phillips, and M. Ortiz. An adaptive finite element approach to atomic-scale mechanics – the quasicontinuum method. *Journal of the Mechanics and Physics of Solids*, **47**(3):611–642, 1999.

[109] V. B. Shenoy, R. Miller, E. B. Tadmor, R. Phillips, and M. Ortiz. Quasicontinuum models of interfacial structure and deformation. *Physical Review Letters*, **80**(4):742–745, 1998.

[110] T. Shimokawa, J. J. Mortensen, J. Schiøtz, and K. W. Jacobsen. Matching conditions in the quasicontinuum method: Removal of the error introduced at the interface between the coarse-grained and fully atomistic region. *Phys. Rev. B*, **69**(21):214104, Jun 2004.

[111] L. Spruch. Pedagogic notes on thomas–fermi theory (and on some improvements): atoms, stars, and the stability of bulk matter. *Rev. Mod. Phys.*, **63**(1):151–209, 1991.

[112] N. Sukumar and J. E. Pask. Classical and enriched finite element formulations for Bloch-periodic boundary conditions. *International Journal for Numerical Methods in Engineering*, **77**(8):1121–1138, 2009.

[113] P. Suryanarayana, V. Gavini, T. Blesgen, K. Bhattacharya, and M. Ortiz. Non-periodic finite-element formulation of Kohn–Sham density functional theory. *Journal of the Mechanics and Physics of Solids*, **58**(2):256–280, 2010.

[114] E. B. Tadmor, M. Ortiz, and R. Phillips. Quasicontinuum analysis of defects in solids. *Philosophical Magazine A*, **73**(6):1529–1563, 1996.

[115] E. B. Tadmor, R. Phillips, and M. Ortiz. Mixed Atomistic and Continuum Models of Deformation in Solids. *Langmuir*, **12**(19):4529–4534, 1996.

[116] L. H. Thomas. The calculation of atomic fields. In *Proceedings of the Cambridge Philosophical Society*, **23**, pages 542–548, 1927.

[117] Y. A. Wang and E. A. Carter. Orbital-free kinetic-energy density functional theory. *Theoretical Methods in Condensed Phase Chemistry*, **5**:117–84, 2000.

[118] Y. A. Wang, N. Govind, and E. A. Carter. Orbital-free kinetic-energy functionals for the nearly free electron gas. *Physical Review B*, **58**(20):13465–13471, 1998.

[119] Y. A. Wang, N. Govind, and E. A. Carter. Orbital-free kinetic-energy density functionals with a density-dependent kernel. *Physical Review B*, **60**(24):16350–16358, 1999.

[120] S. C. Watson and E. A. Carter. Linear-scaling parallel algorithms for the first principles treatment of metals. *Computer Physics Communications*, **128**(1-2):67–92, 2000.

[121] C. F. Weizsäcker. Zur Theorie der Kernmassen. *Zeitschrift für Physik*, **96**(7):431–458, 1935.

[122] S. P. Xiao and T. Belytschko. A bridging domain method for coupling continua with molecular dynamics. *Computer methods in applied mechanics and engineering*, **193**(17-20):1645–1669, 2004.

[123] K. Yosida. *Functional analysis*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the sixth (1980) edition.

[124] E. Zeidler. *Nonlinear functional analysis and its applications. I*. Springer-Verlag, New York, 1986. Fixed-point theorems, Translated from the German by Peter R. Wadsack.

[125] X. Zhang and G. Lu. Quantum mechanics/molecular mechanics methodology for metals based on orbital-free density functional theory. *Phys. Rev. B*, **76**(24):245111, Dec 2007.

[126] A. Zhou. An analysis of finite-dimensional approximations for the ground state solution of Bose-Einstein condensates. *Nonlinearity*, **17**(2):541–550, 2004.

[127] A. Zhou. Finite dimensional approximations for the electronic ground state solution of a molecular system. *Mathematical Methods in the Applied Sciences*, **30**(4):429–447, 2007.