Mathematical Models and Methods in Applied Sciences

Existence and Convergence Results for the Galerkin Approximation of an Electronic Density Functional

BERNHARD LANGWALLNER
University of Oxford, Mathematical Institute, 24-29 St Giles', Oxford OX1 3LB, UK.
langwallner@maths.ox.ac.uk

CHRISTOPH ORTNER
University of Oxford, Mathematical Institute, 24-29 St Giles', Oxford OX1 3LB, UK.
ortner@maths.ox.ac.uk

ENDRE SÜLI
University of Oxford, Mathematical Institute, 24-29 St Giles', Oxford OX1 3LB, UK.
suli@maths.ox.ac.uk

Received (Day Month Year)
Revised (Day Month Year)
Communicated by (xxxxxxxxxx)

We formulate and analyze a model for the study of finite clusters of atoms or localized defects in infinite crystals based on orbital-free density functional theory. We show that the resulting constrained optimization problem has a minimizer and we provide a careful analysis of the solubility 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.

Keywords: Nonlinear eigenvalue problem, Galerkin discretization, Thomas–Fermi type functionals

AMS Subject Classification: 65N60, 65N25, 65Z05

1. Introduction

Quantum mechanics is the accepted microscopic theory of atoms, molecules, and solids. Quantum mechanical simulations have therefore become an indispensable tool in physics, chemistry, and materials science. In particular, they are considered *ab initio*, that is, no empirical input is necessary. Computations based on the Schrödinger equation (the governing equation of quantum mechanics) require the solution of a linear partial differential equation in a high-dimensional configuration space. This fact dramatically limits the size of the problems that can be tackled.[22]

Possibly the most widely employed remedy is known as *Density Functional Theory* (DFT).[21,23] Its theoretical justification rests on the fact that all physical ground state properties of a quantum mechanical system can directly be computed from the ground state electron density rather than the electronic wave function. The former is defined on $\mathbb{R}^3$ whereas the latter is defined on $\mathbb{R}^{3N}$, N being the number of electrons.

The most demanding task in DFT calculations is the computation of the kinetic energy. It involves repeated numerical solution of single electron eigenvalue problems and subsequent orthogonalization of the resulting wave functions. In order to avoid this huge cost, approximate forms have also been developed for the kinetic energy. The resulting models are commonly referred to as orbital-free density functional theory.[13,26,30] Early models of that class include Thomas–Fermi type functionals; see for example Finnis.[13] Orbital-free DFT has been observed to work well for systems with weakly varying electron density, most prominently aluminium, see Wang, Govind & Carter.[33,34]

A simple example of a density functional is given by the Thomas–Fermi–Dirac–von Weizsäcker (TFDW) functional. Let ρ be the electron density, $R = \{R_i,\ i = 1,\dots,N_{\rm nuc}\}$ the set of nucleus positions and $Z_i \in \mathbb{N}$, $i = 1,\dots,N_{\rm nuc}$, the charges of the nuclei. Then the energy is written as

$$E(\rho, R) = T_s(\rho) + E_{\rm xc}(\rho) + E_{\rm H}(\rho) + E_{\rm ext}(\rho, R) + E_{\rm zz}(R). \tag{1.1}$$

Here, T_s is the kinetic energy

$$T_s(\rho) = \frac{3}{10}(3\pi^2)^{2/3}\int_{\mathbb{R}^3}\rho^{5/3}(x)\,{\rm d}x + \frac{\lambda}{8}\int_{\mathbb{R}^3}\frac{|\nabla\rho(x)|^2}{\rho(x)}\,{\rm d}x,$$

where $\lambda > 0$ is a parameter. The exchange correlation energy $E_{\rm xc}$ in the so-called local density approximation[9,31] has the form

$$E_{\rm xc}(\rho) = \int_{\mathbb{R}^3}\big(-C_{\rm x}\rho(x)^{4/3} + \varepsilon_{\rm c}(\rho(x)\big)\rho(x)\,{\rm d}x,$$

where $\varepsilon_{\rm c}$ is a phenomenological correction term. We note that this term renders $E(\cdot, R)$ non-convex.

The remaining terms are of electrostatic nature. The Hartree term

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

incorporates interaction among electrons. The energy of electrons in the electrostatic field generated by the nuclei is accounted for by

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

Finally, the electrostatic repulsion energy of the nuclei is

$$E_{\rm zz}(R) = \frac{1}{2}\sum_{i=1}^{N_{\rm nuc}}\sum_{\substack{j=1\\ j\neq i}}^{N_{\rm nuc}}\frac{Z_iZ_j}{|R_i-R_j|}.$$

For numerical or modeling reasons the nuclear point charges Z_i at R_i may be replaced by a smooth charge density ρ_n (see Gavini et al.[18]). For example, ρ_n may be written as a sum of compactly supported smooth functions centered at the positions R_i, $i = 1, \ldots, N_{\text{nuc}}$, (the dependence of ρ_n on R will be suppressed)

$$\rho_n(x) = \sum_{i=1}^{N_{\text{nuc}}} Z_i \tilde{\rho}_0(x - R_i),$$

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

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

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

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

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

$$E_{\text{H}}(\rho) + E_{\text{ext}}(\rho, R) + E_{\text{zz}}(R) = \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{(\rho(x) - \rho_n(x))\,(\rho(y) - \rho_n(y))}{|x - y|}\,\mathrm{d}y\,\mathrm{d}x.$$

The nonlocal nature of this term represents a numerical challenge for all density functional calculations. The given density functional has to be minimized subject to the constraints

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

A survey of theoretical results in connection with Thomas–Fermi type models can be found in an article by Lieb.[25] 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.

Several numerical approximations of the functional (1.1) or related models have been proposed in the literature. For Galerkin discretizations, the crucial question is the choice of basis. The most popular basis sets are plane waves (i.e. Fourier modes[7,35]) and finite elements.[18] 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.[17] Calculating the convolution in the electrostatic terms remains a challenge. Finite difference approximations were suggested in References 14, 15, 28, 29.

Little work can be found in the literature on the numerical analysis of these approximations. The only convergence result in a finite element context we are aware

of is a proof of Γ-convergence in Gavini et al.[18], which contains no information about convergence rates. The latter, however, are crucial for understanding the efficiency of the numerical method. In Cancès, Chakir & Maday[6,7] the authors perform a rigorous numerical analysis of a planewave discretization of a density functional. That work also includes convergence rates.

In our present and ongoing work[24] we aim to fill this gap and provide a complete convergence theory for Galerkin discretizations of orbital-free electronic density functionals. In Section 2.1 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. Our reformulation of the optimality system is particularly suitable for Galerkin finite element discretizations. The main result in Section 2.3 connects the stability of the minimization problem to the stability of this system. This result allows us, in Section 3, to prove the existence and convergence of Galerkin discretizations of the optimality system.

The present work is of, predominantly, theoretical/analytical nature. The important practical issues of optimal convergence rates and numerical integration will be addressed in a forthcoming article.[24] All of the results discussed herein carry over to the formulation *with* numerical integration.

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 numerical approximation.

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:

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

We also need to address the evaluation of the electrostatic term. The double integral can in principle be computed explicitly using the Fourier transform.[14,16,35] In Gavini et al.[17,18] 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 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 Gavini et al.[18]

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). For both computational and analytical reasons, we would like to formulate the problem in a bounded open domain $\Omega\subset\mathbb{R}^3$ rather than in the whole of $\mathbb{R}^3$. To this end, we will assume that we know the square root density u as well as the electrostatic potential ϕ in $\mathbb{R}^3\setminus\Omega$, 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 $\mathbb{R}^3\setminus\Omega$. 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 Blanc, Le Bris & Lions[2] and Catto, Le Bris & Lions.[8])

To make this concrete, we assume that we are given functions $u_{\mathrm{ex}},\phi_{\mathrm{ex}}\in\mathrm{H}^2_{\mathrm{loc}}(\mathbb{R}^3)$ and define the admissible set

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

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

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

where

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

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 model the function F is given by

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

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 $c : \mathrm{H}^1(\Omega) \to \mathbb{R}$,

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

Our goal is to solve the minimization problem

$$\min_{u \in A_u} E(u). \tag{2.7}$$

Remark 2.1. Our justification for this approach is that, if $\bar{u}$ is a minimizer of the original problem, with associated electrostatic potential $\bar{\phi}$, then, setting $u_{\rm ex} = \bar{u}$ and $\phi_{\rm ex} = \bar{\phi}$, $\bar{u}|_\Omega$ 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 $\mathbb{R}^3$.

Since we do not normally know the exact electron density and electostatic 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 $u_{\rm ex} = \phi_{\rm ex} = 0$ in $\mathbb{R}^3 \setminus \Omega$. 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 $\mathbb{R}^3$. 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 a first step towards an analysis of a combined TFDW / Quasicontinuum model in the spirit of Garcia-Cervera, Lu & E[16] and Gavini et al.[17] 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. *Existence of a minimizer*

We assume from now on 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

$$|u|_{\mathrm{H}^2} \leq C_{\mathrm{reg}} \|f\|_{\mathrm{L}^2}.$$

This is the case, for example, if Ω is convex[20] or if Ω is C^2-regular.[12] 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

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

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

$$|F''(t_1) - F''(t_2)| \leq C(1 + |t_1|^{q-3} + |t_2|^{q-3})(|t_1 - t_2|^\alpha + |t_1 - t_2|), \tag{2.10}$$

where $0 < \alpha \leq 1$. 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.

It is also straightforward to verify that, under condition (2.10), the functional $X : \mathrm{H}^1(\Omega) \to \mathbb{R}$ defined by (2.5) is twice continuously Fréchet differentiable with

$$\langle X'(u), h\rangle = \int_\Omega F'(u)h \,\mathrm{d}x, \qquad \text{and} \qquad \langle X''(u)h_1, h_2\rangle = \int_\Omega F''(u)h_1 h_2 \,\mathrm{d}x,$$

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

$$\|X''(u) - X''(v)\| \leq C(1 + \|u\|^{q-3}_{\mathrm{H}^1} + \|v\|^{q-3}_{\mathrm{H}^1})(\|u - v\|^\alpha_{\mathrm{H}^1} + \|u - v\|_{\mathrm{H}^1}), \tag{2.11}$$

for all $u, v \in \mathrm{H}^1(\Omega)$. To prove this we begin as follows: for $u, v, h_1, h_2 \in \mathrm{H}^1(\Omega)$ we have that

$$\begin{aligned} |\langle (X''(u) - X''(v))h_1, h_2\rangle| &\leq \int_\Omega |F''(u) - F''(v)||h_1||h_2| \,\mathrm{d}x \qquad (2.12)\\ &\leq C \int_\Omega |h_1||h_2|(1+|u|^{q-3}+|v|^{q-3})(|u-v|^\alpha+|u-v|) \,\mathrm{d}x \\ &\leq C \,\|h_1\|_{\mathrm{L}^4} \,\|h_2\|_{\mathrm{L}^4} \,(1 + \big\||u|^{q-3}\big\|_{\mathrm{L}^4} + \big\||v|^{q-3}\big\|_{\mathrm{L}^4}) \\ &\qquad \cdot (\big\||u-v|^\alpha\big\|_{\mathrm{L}^4} + \|u-v\|_{\mathrm{L}^4}), \end{aligned}$$

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

$$\||u|^\gamma\|^k_{\mathrm{L}^k} = \int_\Omega 1 \cdot |u|^{\gamma k}\,\mathrm{d}x \leq C\,\big\||u|^{\gamma k}\big\|_{\mathrm{L}^{1/\gamma}} = C\,\|u\|^{\gamma k}_{\mathrm{L}^k}.$$

Applying this inequality to (2.12) and using the embedding $\mathrm{H}^1(\Omega)$ in $\mathrm{L}^4(\Omega)$ we deduce (2.11).

Since X is strongly continuous in $\mathrm{L}^q(\Omega)$ and since $\mathrm{H}^1(\Omega)$ is compactly embedded in $\mathrm{L}^p(\Omega)$, for $1 \leq p < 6$, see Adams,[1] it follows also that X is sequentially weakly continuous on $\mathrm{H}^1(\Omega)$.

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

Lemma 2.1. *The functional $\Phi : \mathrm{L}^4(\Omega) \to \mathbb{R}$ is well-defined and continuous; Φ is twice continuously Fréchet differentiable and the derivatives are, for $u \in \mathrm{L}^4(\Omega)$, given by*

$$\langle \Phi'(u), h\rangle = 2\int_\Omega u\phi h\,\mathrm{d}x \quad \textit{for } h \in \mathrm{L}^4(\Omega) \textit{ and,}$$
$$\langle \Phi''(u)h_1, h_2\rangle = 2\int_\Omega \phi h_1 h_2 + 8\pi u h_1(-\Delta_0)^{-1}(uh_2)\,\mathrm{d}x \quad \textit{for } h_1, h_2 \in \mathrm{L}^4(\Omega),$$

where ϕ 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 $\tilde\Phi(u,\cdot)$ from (2.5) is clearly continuous, convex, coercive and weakly lower semicontinuous on $V_\phi = \{\psi \in \mathrm{H}^1(\Omega) : \psi|_{\partial\Omega} = \phi_{\mathrm{ex}}\}$. 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

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

and, moreover, after straightforward algebraic manipulations, that

$$\begin{aligned}\Phi(u) = -\tilde\Phi(u,\phi_u) = 2\pi\int_\Omega (u^2 - \rho_n)(-\Delta_0)^{-1}(u^2 - \rho_n)\,\mathrm{d}x \\ + \int_\Omega (u^2 - \rho_n)\xi\,\mathrm{d}x - \frac{1}{8\pi}\int_\Omega |\nabla\xi|^2\,\mathrm{d}x.\end{aligned} \tag{2.13}$$

Differentiating with respect to u yields the expressions for Φ' and Φ'' as given above.

Finally, we show that $\Phi|_{A_u}$ is bounded below. Clearly, the first term on the right-hand side in (2.13) is non-negative, and the last term is a constant depending only on ϕ_{ex}. Since we assumed that $\phi_{\mathrm{ex}} \in \mathrm{H}^2_{\mathrm{loc}}(\mathbb{R}^3)$ and that the Poisson problem is

H^2-regular, it follows that $\xi \in L^\infty(\Omega)$. Therefore, the second term on the right-hand side of (2.13) can be bounded as follows:

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

Hence, we can deduce that $\Phi(u) \geq C(N, \rho_n, \phi_{\mathrm{ex}})$ for all $u \in A_u$.

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

Theorem 2.1. *$E : A_u \to \mathbb{R}$ has at least one minimizer.*

Proof. We apply the direct method of the calculus of variations.[10] First, we observe that E is coercive on A_u, which can be seen as follows:

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

Here we have used the growth condition (2.9) on F, the lower bound on $\Phi(u)$, established in Lemma 2.1, and Poincaré's inequality for $u - u_{\mathrm{ex}}$. Hence, minimizing sequences of E are bounded in $H^1(\Omega)$, 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 Dacorogna.[10] □

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 u minimizes E, then so does $-u$. The question whether there can be more than two minimizers of the energy (2.4) is beyond the scope of this article.

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, an article by Maurer & Zowe[27] provides necessary and sufficient optimality conditions for a large class of optimization problems in Banach spaces.

Throughout this section, the derivatives $E'(\bar{u})$ and $c'(\bar{u})$ are understood as elements of $H^{-1}(\Omega) := H_0^1(\Omega)^*$, that is, they operate on functions from $H_0^1(\Omega)$. This means, for example, that $\ker c'(\bar{u}) \subset H_0^1(\Omega)$.

Theorem 3.1 in Maurer & Zowe[27] yields the first-order necessary optimality condition

$$\langle E'(\bar{u}), v \rangle = 0 \quad \forall v \in \ker c'(\bar{u}) \subset H_0^1(\Omega),$$

provided that $0 \in \operatorname{int}\left\{\langle c'(\bar{u}), v\rangle : v \in \mathrm{H}^1_0(\Omega)\right\} \subseteq \mathbb{R}$. Since $\langle c'(\bar{u}), v\rangle = (\bar{u}, v)$ and $\bar{u} \neq 0$, this condition is always satisfied. Theorem 3.2 in Maurer & Zowe[27] yields the existence of a *Lagrange multiplier* $\bar{\mu} \in \mathbb{R}$ such that

$$E'(\bar{u}) + \bar{\mu} c'(\bar{u}) = 0 \in \mathrm{H}^{-1}(\Omega), \quad \text{and} \quad c(\bar{u}) = 0. \tag{2.14}$$

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 $\bar{u} \in A_u$ 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 2.1 implies that $\bar{u}, \bar{\phi}$ and $\bar{\mu} \in \mathbb{R}$ solve the nonlinear system

$$\begin{aligned} \lambda(\nabla u, \nabla v) + (F'(u), v) + 2(\phi u, v) + \mu(u, v) = 0 \qquad & \forall v \in \mathrm{H}^1_0(\Omega), \\ \frac{1}{4\pi}(\nabla\phi, \nabla\psi) - (u^2 - \rho_n, \psi) = 0 \qquad & \forall \psi \in \mathrm{H}^1_0(\Omega), \\ \frac{\nu}{2}\left(\int_\Omega u^2 \,\mathrm{d}x - N\right) = 0 \qquad & \forall \nu \in \mathbb{R} \end{aligned} \tag{2.15}$$

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 (2.14) or the minimization problem (2.7). It has to be pointed out that solving (2.15) (or even (2.14)) is not equivalent to solving the minimization problem since the functional is non-convex. However, we will focus on those solutions of (2.15) which correspond to local minimizers of (2.7), making use of the second-order optimality condition (2.19).

We define the function spaces

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

so that the system (2.15) defines an operator $\mathcal{F} : \mathcal{Y}_D \to \mathcal{Y}_0^*$, rewritten as

$$\big\langle \mathcal{F}(u, \phi, \mu), (v, \psi, \nu)\big\rangle = 0 \quad \forall (v, \psi, \nu) \in \mathcal{Y}_0. \tag{2.16}$$

For future reference, we split $\mathcal{F}$ into

$$\mathcal{F}(u, \phi, \mu) = \begin{pmatrix} -\lambda\Delta u \\ -\frac{1}{4\pi}\Delta\phi \\ 0 \end{pmatrix} + \mathcal{G}(u, \phi, \mu),$$

where $\mathcal{G}$ contains all terms without derivatives, and where 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 $\mathcal{F}'(u, \phi, \mu) : \mathcal{Y}_0 \to \mathcal{Y}_0^*$,

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

and that $\mathcal{F}'$ is locally Hölder continuous: there is a continuous function $L_{\mathcal{F}'} : \mathbb{R} \to \mathbb{R}$ and $\alpha \in (0, 1)$ such that

$$\|\mathcal{F}'(y_1) - \mathcal{F}'(y_2)\| \leq L_{\mathcal{F}'}(\|y_1\|_{\mathcal{Y}} + \|y_2\|_{\mathcal{Y}}) \, (\|y_1 - y_2\|_{\mathcal{Y}}^{\alpha} + \|y_1 - y_2\|_{\mathcal{Y}}) \tag{2.17}$$

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 is a continuous function $L_{\mathcal{F}} : \mathbb{R} \to \mathbb{R}$ such that

$$\|\mathcal{F}(y_1) - \mathcal{F}(y_2)\|_{\mathcal{Y}_0^*} \leq L_{\mathcal{F}}(\|y_1\|_{\mathcal{Y}} + \|y_2\|_{\mathcal{Y}}) \, \|y_1 - y_2\|_{\mathcal{Y}}, \tag{2.18}$$

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

At this point, we make some observations concerning the regularity of solutions to (2.15). The functions $\bar{u}$ and $\bar{\phi}$ solve 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 $\bar{u}|_{\partial\Omega} = u_{\text{ex}}$, $\bar{\phi}|_{\partial\Omega} = \phi_{\text{ex}}$, respectively. Here, the functions g_1 and g_2 obviously depend on $\bar{u}$, $\bar{\phi}$ and $\bar{\mu}$. From embedding theorems and the growth properties of F' we know that $g_1, g_2 \in \mathrm{L}^2(\Omega)$. Since we assumed that the Poisson problem (2.8) is H^2-regular, we can deduce that $\bar{u}$, $\bar{\phi} \in \mathrm{H}^2(\Omega)$. The Sobolev embedding theorem[1] 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 $\bar{u}$, $\bar{\phi} \in \mathrm{L}^{\infty}(\Omega)$.

2.4. *Second-order optimality conditions*

The function $\mathcal{L}(u, \mu) = E(u) + \mu c(u)$ 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 Maurer & Zowe[27] 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

$$\langle \nabla_{uu} \mathcal{L}(\bar{u}, \bar{\mu}) v, v \rangle \geq 0 \qquad \forall v \in \ker c'(\bar{u})$$

holds. Conversely, if $(\bar{u}, \bar{\mu})$ satisfies (2.14) as well as the sufficient second-order optimality condition

$$\langle \nabla_{uu} \mathcal{L}(\bar{u}, \bar{\mu}) v, v \rangle \geq \gamma \, \|\nabla v\|_{\mathrm{L}^2}^2 \qquad \forall v \in \ker c'(\bar{u}), \tag{2.19}$$

for some constant $\gamma > 0$, then $\bar{u}$ is an isolated local minimizer of E in A_u, see Maurer & Zowe.[27] We call a critical point $\bar{u} \in A_u$ that satisfies (2.19) a *uniform minimizer* of (2.7).

Written out explicitly, (2.19) reads

$$\lambda(\nabla v, \nabla v) + \big((F''(\bar{u}) + 2\bar{\phi} + \bar{\mu})v, v\big) + 16\pi(\bar{u}v, (-\Delta_0)^{-1}\bar{u}v) \geq \gamma \, \|\nabla v\|_{\mathrm{L}^2}^2 , \tag{2.20}$$

for all $v \in \ker c'(\bar{u})$, 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 $\mathcal{F}'(\bar{u}, \bar{\phi}, \bar{\mu})$ is an

isomorphism. This will be an important tool for the convergence analysis in the next section.

Proposition 2.1. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu}) \in \mathcal{Y}_D$ such that (2.20) holds with $\gamma > 0$. Then, $\mathcal{F}'(\bar{y}) : \mathcal{Y}_0 \to \mathcal{Y}_0^*$ is an isomorphism.*

Proof. We need to show that the equation

$$\mathcal{F}'(\bar{u}, \bar{\phi}, \bar{\mu}) \cdot (v, \psi, \nu) = (f, g, \kappa) \tag{2.21}$$

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

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

Then, equation (2.21) takes the form of a saddle-point problem,

$$\begin{aligned} a_{\bar{y}}\big((v,\psi),(w,\chi)\big) + b_{\bar{y}}\big((w,\chi),\nu\big) &= \langle f, w\rangle + \langle g, \chi\rangle \qquad \forall w, \chi \in \mathrm{H}_0^1(\Omega), \\ b_{\bar{y}}\big((v,\psi),\eta\big) &= \eta\kappa \qquad\qquad\qquad\qquad \forall \eta \in \mathbb{R}. \end{aligned} \tag{2.22}$$

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

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

For a saddle-point problem such as (2.22) there are well-known sufficient conditions for solvability; see Theorem 1.1 in Brezzi & Fortin.[4] The bilinear form $b_{\bar{y}}$ has to satisfy an inf-sup condition of the form

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

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

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

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 principle: *find* $(v,\psi) \in \ker b_{\bar{y}}$ *such that*

$$a_{\bar{y}}\big((v,\psi),(w,\chi)\big) = \langle f, w\rangle + \langle g, \chi\rangle \quad \forall (w,\chi) \in \ker b_{\bar{y}}, \tag{2.23}$$

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

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 (2.23) and testing with $w = v$ and $\chi = 0$ we obtain

$$\lambda(\nabla v, \nabla v) + \big((F''(\bar{u}) + 2\bar{\phi} + \bar{\mu})v, v\big) + 16\pi(\bar{u}v, (-\Delta_0)^{-1}\bar{u}v) = 0. \tag{2.24}$$

Since $\bar{u}$ is assumed to be a uniform minimizer we know from (2.20) that the bilinear form on the left-hand side of (2.24) 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 $v \in \ker c'(\bar{u})$ be the unique solution of

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

This equation is uniquely solvable by the Lax–Milgram theorem,[3] 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 (2.23). 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 Brezzi & Fortin[4] we can now conclude that $\mathcal{F}'(\bar{y})$ is indeed an isomorphism from $\mathcal{Y}_0$ to $\mathcal{Y}_0^*$. □

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

$$\begin{aligned}\inf_{(v,\psi)\in\ker b_{\bar{y}}} \sup_{(w,\chi)\in\ker b_{\bar{y}}} \frac{a_{\bar{y}}\big((v,\psi),(w,\chi)\big)}{\|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}\,\|(\nabla w, \nabla\chi)\|_{\mathrm{L}^2}} \geq \kappa_a, \quad \text{and} \\ \inf_{(w,\chi)\in\ker b_{\bar{y}}} \sup_{(v,\psi)\in\ker b_{\bar{y}}} \frac{a_{\bar{y}}\big((v,\psi),(w,\chi)\big)}{\|(\nabla v, \nabla\psi)\|_{\mathrm{L}^2}\,\|(\nabla w, \nabla\chi)\|_{\mathrm{L}^2}} \geq \kappa_a.\end{aligned} \tag{2.25}$$

For this equivalence see for example Proposition 1.2 in Brezzi & Fortin,[4] respectively the discussion following Remark 1.6 in Brezzi & Fortin.[4]

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 (2.15). 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 future work.

3.1. *The Discretized Functional*

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

$$\inf_{v\in S_h} \|\nabla(u-v)\|_{\mathrm{L}^2} \leq Ch|u|_{\mathrm{H}^2} \qquad \forall\, u \in \mathrm{H}^2(\Omega),$$

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

$$\|\phi - \mathcal{I}_h\phi\|_{\mathrm{H}^1} \leq Ch|\phi|_{\mathrm{H}^2} \quad \text{for all } \phi \in \mathrm{H}^2(\Omega). \tag{3.1}$$

Moreover, let

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

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

$$E_h(u_h) = \frac{\lambda}{2}\int_\Omega |\nabla u_h|^2\,\mathrm{d}x + \int_\Omega F(u_h)\,\mathrm{d}x + \Phi_h(u_h), \tag{3.2}$$

where

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

Let $A_{u,h} := \{u_h \in u_{\mathrm{ex},h} + S_{h,0} : \|u_h\|^2_{\mathrm{L}^2} = N\}$ be the set of discrete admissible functions. We consider the discretized minimization problem

$$\min_{u_h\in A_{u,h}} E_h(u_h). \tag{3.3}$$

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

$$\begin{aligned}
\lambda(\nabla\bar u_h, \nabla v) + (F'(\bar u_h), v) + 2(\bar\phi_h\bar u_h, v) + \mu_h(\bar u_h, v) = 0 \quad &\forall v \in S_{h,0},\\
\frac{1}{4\pi}(\nabla\bar\phi_h, \nabla\psi) - (\bar u_h^2 - \rho_n, \psi) = 0 \quad &\forall \psi \in S_{h,0},\\
\frac{\nu}{2}\left(\int_\Omega \bar u_h^2\,\mathrm{d}x - N\right) = 0 \quad &\forall \nu \in \mathbb{R},
\end{aligned} \tag{3.4}$$

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 (2.15). We introduce the discrete function spaces

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

In analogy to the continuous case we write the system (3.4) in the more compact form

$$\left\langle \mathcal{F}_h(\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h), (v, \psi, \nu) \right\rangle = 0 \qquad \forall (v, \psi, \nu) \in \mathcal{Y}_{h,0}, \tag{3.5}$$

where $\mathcal{F}_h : \mathcal{Y}_{h,D} \to \mathcal{Y}_{h,0}^*$. The operator $\mathcal{F}_h$ will sometimes be split into two parts,

$$\mathcal{F}_h(u_h, \phi_h, \mu_h) = \begin{pmatrix} -\lambda \Delta_h u_h \\ -\frac{1}{4\pi} \Delta_h \phi_h \\ 0 \end{pmatrix} + \mathcal{G}(u_h, \phi_h, \mu_h),$$

where $\mathcal{G}$ has the same form as in the continuous case but it is now restricted to $\mathcal{Y}_h$, and where the discrete Laplacian $(-\Delta_h) : S_h \to S_{h,0}^*$ is defined by $\langle -\Delta_h v_h, w_h \rangle = (\nabla v_h, \nabla w_h)$ for $v_h \in S_h$ and all $w_h \in S_{h,0}$. The operator $(-\Delta_{h,0})^{-1} : S_{h,0}^* \to S_{h,0}$ maps f to the solution $\phi_h \in S_{h,0}$ of $(\nabla \phi_h, \nabla v_h) = \langle f, v_h \rangle$ for all $v_h \in S_{h,0}$.

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

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

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 $\mathcal{F}_h'$ may in fact be extended to the whole of $\mathcal{Y}$. Slightly abusing notation, we will write $\mathcal{F}_h'(y) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$ 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 2.1.

Proposition 3.1. *Let $\bar{y} = (\bar{u}, \bar{\phi}, \bar{\mu})$ be a solution of (2.15) such that $\bar{u}$ is a uniform minimizer of (2.7) that satisfies (2.19). Then, there exist $h_0 \in (0, 1]$ and $\delta > 0$ such that $\mathcal{F}_h'(y) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$ is an isomorphism for every $h \leq h_0$ and for any $y \in B_\delta(\bar{y}) \subset \mathcal{Y}$. Moreover, there is a constant $M > 0$ such that*

$$\|\mathcal{F}_h'(y)^{-1}\| \leq M \qquad \forall y \in B_\delta(\bar{y}) \quad \forall h \leq h_0. \tag{3.6}$$

Proof. We begin by showing invertibility of $\mathcal{F}_h'(\bar{y}) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$. 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 $S_{h,0}^2 \cap \ker b_{\bar{y}}$.

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

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

where, for given $\nu \neq 0$, we have chosen $v = \mathrm{sign}(\nu)(-\Delta_{h,0})^{-1}\bar{u}$ 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

$$a_{\bar{y}}\big((v,\psi),(w,\chi)\big) = \langle f, w\rangle + \langle g, \chi\rangle \quad \forall (w,\chi) \in \ker b_{\bar{y}} \cap S^2_{h,0} \tag{3.7}$$

has a unique solution $(v,\psi) \in \ker b_{\bar{y}} \cap S^2_{h,0}$ for every $f,\ g \in \mathrm{H}^{-1}(\Omega)$. If the trial and test spaces were simply $S^2_{h,0}$ instead of the constraint space $\ker b_{\bar{y}} \cap S^2_{h,0}$, then this would follow from a classical argument by Schatz.[32] The present case requires some modifications, which we study in detail in the Appendix. Lemmas 5.2 and 5.1 provide the regularity and approximation results in $\ker b_{\bar{y}} \cap S^2_{h,0}$ necessary for an application of the Schatz argument, which is carried out in Theorem 5.1. Hence, we deduce that (3.7) is uniquely solvable, provided that h is small enough. Theorem 5.1 also implies the existence of an inf-sup constant $\kappa_{a,h}$ for $a_{\bar{y}}$ on $K_h := S^2_{h,0} \cap \ker b_{\bar{y}}$, similarly as in (2.25), 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 5.1 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, $\mathcal{F}'_h(\bar{y})$ is an isomorphism. This also means that for every $h \leq h_0$, $\mathcal{F}'_h(\bar{y})$ satisfies the inf-sup conditions

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

with $\kappa_h > 0$. Theorem 1.1 in Brezzi & Fortin[4] 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 $b_{\bar{y}}$. Since the latter are uniformly bounded away from zero as $h \to 0$, we deduce that the same holds for κ_h.

Since $\mathcal{F}'_h$ satisfies the discrete equivalent of the Hölder condition (2.17), it follows 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 $y \in B_\delta(\bar{y})$, $\mathcal{F}'_h(y)$ is an isomorphism and such that the inf-sup constants for $\mathcal{F}'_h(y)^{-1}$ are uniformly bounded away from zero. This implies (3.6). □

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 Brezzi, Rappaz & Raviart[5] and Dobrowolski & Rannacher.[11]

Theorem 3.1. *Let $\bar{u}$ be a minimizer of (2.7) that satisfies (2.19). Let $\bar{\phi}$ and $\bar{\mu}$ be, respectively, the associated electrostatic potential and Lagrange multiplier. Then, there exist $h_0 \in (0,1], \delta > 0$ such that the discretized problem (3.4) has a unique solution $\bar{y}_h = (\bar{u}_h, \bar{\phi}_h, \bar{\mu}_h) \in \mathcal{Y}_{h,D}$ in the neighbourhood $B_\delta(\bar{y}) \subset \mathcal{Y}$ for all $h < h_0$. Furthermore, there exists a constant C such that*

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

Proof. The proof is divided into four steps.

Step 1. We show 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 $u_h, \phi_h \in S_h$ be the Ritz projections of $\bar{u}$ and $\bar{\phi}$, respectively, i.e., the solutions of the equations

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

with boundary conditions

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

In other words, $\Delta_h u_h = \Delta \bar{u}|_{S_{h,0}}$ and $\Delta_h \phi_h = \Delta \bar{\phi}|_{S_{h,0}}$. We define $\Pi_h \bar{y} = (u_h, \phi_h, \bar{\mu})$. Convergence theory for the Poisson equation, $\bar{u}$, $\bar{\phi} \in \mathrm{H}^2(\Omega)$, and the approximation property (3.1) of $\mathcal{I}_h$ then lead to

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

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 (2.18) 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 3.1 we have shown that there is an open neighbourhood $B_\delta(\bar{y})$ of $\bar{y}$ in $\mathcal{Y}$ and $h_0 \in (0,1]$ such that $\mathcal{F}_h'(y) : \mathcal{Y}_{h,0} \to \mathcal{Y}_{h,0}^*$ is an isomorphism for all $y \in B_\delta(\bar{y})$, $h \leq h_0$ and $\|\mathcal{F}_h'(y)^{-1}\| \leq M$, uniformly for $y \in B_\delta(\bar{y})$. Moreover, we observe that $\mathcal{F}_h'$ satisfies a Hölder continuity property similar to (2.17): there is $L_{\bar{y},\delta}$ and $\alpha \in (0,1)$ such that

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

Step 3. Existence and 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 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

$$\mathcal{F}_h'(y_0)(y_h - y_0) = -\mathcal{F}_h(y_0) - \int_0^1 (\mathcal{F}_h'(y_0 + t(y_h - y_0)) - \mathcal{F}_h'(y_0))\,\mathrm{d}t\ (y_h - y_0).$$

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

$$\mathcal{F}_h'(y_0)(\mathcal{N}(y) - y_0) = -\mathcal{F}_h(y_0) - \int_0^1 (\mathcal{F}_h'(y_0 + t(y - y_0)) - \mathcal{F}_h'(y_0))\,\mathrm{d}t\ (y - y_0).$$

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 $\alpha \in (0,1)$, that

$$\begin{aligned} M^{-1}\|\mathcal{N}(y) - y_0\|_{\mathcal{Y}} &\le \|\mathcal{F}_h(y_0)\|_{\mathcal{Y}_0^*} + R\int_0^1 \|\mathcal{F}_h'(y_0 + t(y - y_0)) - \mathcal{F}_h'(y_0)\|\,\mathrm{d}t \\ &\le C_2(h + RL_{\bar{y},\delta}(R + R^\alpha)), \end{aligned}$$

where we have used the stability property (3.6). To ensure that $\mathcal{N}(y) \in B_R(y_0)$, we need to bound $C_2(h + RL_{\bar{y},\delta}(R + R^\alpha))$ 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

$$\begin{aligned} \mathcal{F}_h'(y_0)(\mathcal{N}(\eta_1) - \mathcal{N}(\eta_2)) &= \mathcal{F}_h(\eta_2) - \mathcal{F}_h(\eta_1) + \mathcal{F}_h'(y_0)(\eta_1 - \eta_2) \\ &= \int_0^1 \big[\mathcal{F}_h'(y_0) - \mathcal{F}_h'(\eta_1 + t(\eta_2 - \eta_1))\big](\eta_1 - \eta_2)\,\mathrm{d}t. \end{aligned}$$

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

$$\begin{aligned} M^{-1}\|\mathcal{N}(\eta_1) - \mathcal{N}(\eta_2)\|_{\mathcal{Y}} &\le \int_0^1 \|\mathcal{F}_h'(y_0) - \mathcal{F}_h'(\eta_1 + t(\eta_2 - \eta_1))\|_{\mathcal{Y}}\,\mathrm{d}t\ \|\eta_1 - \eta_2\| \\ &\le L(R + R^\alpha) \cdot \|\eta_1 - \eta_2\|_{\mathcal{Y}}. \end{aligned}$$

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

We can now use Banach's Fixed Point Theorem[36] 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 neighborhood $B_{2R}(\bar{y})$:

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

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 3.2. *Under the same assumptions as in Theorem 3.1 and for sufficiently small h, the discrete solution $\bar{u}_h \in A_{u,h}$ is a uniform minimizer of the discretized functional* (3.2) *over $A_{u,h}$.*

Proof. We define the two Lagrangians

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

Since $\bar{u}$ is a uniform minimizer, we have $\langle \nabla^2_{uu}\mathcal{L}(\bar{u},\bar{\mu})v, v\rangle \geq \gamma \|\nabla v\|^2_{\mathrm{L}^2}$ for all $v \in \ker c'(\bar{u})$.

Given $v_h \in \ker c'(\bar{u}_h) \cap S_{h,0}$, we have

$$\begin{aligned}\langle \nabla^2_{uu}\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)v_h, v_h\rangle &= \langle \nabla^2_{uu}\mathcal{L}(\bar{u},\bar{\mu})v, v\rangle + \langle (\nabla^2_{uu}\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h) - \nabla^2_{uu}\mathcal{L}(\bar{u},\bar{\mu}))v_h, v_h\rangle \\ &\quad + 2\langle \nabla^2_{uu}\mathcal{L}(\bar{u},\bar{\mu})v, (v_h - v)\rangle \\ &\quad + \langle \nabla^2_{uu}\mathcal{L}(\bar{u},\bar{\mu})(v_h - v), (v_h - v)\rangle \end{aligned} \tag{3.8}$$

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

Next, we prove that every $v_h \in \ker c'(\bar{u}_h)$ can be approximated by $v \in \ker c'(\bar{u})$ 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., $v \in \ker c'(\bar{u})$. 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 $v_h \in S_{h,0}$. 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 (3.8) satisfies

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

The third and fourth term on the right-hand side of (3.8) can be bounded by $Ch\,\|\nabla v_h\|_{\mathrm{L}^2}^2$. For the remaining term, we get

$$\langle(\nabla_{uu}^2\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)-\nabla_{uu}^2\mathcal{L}(\bar{u},\bar{\mu}))v_h,v_h\rangle = \langle(X''(\bar{u})-X''(\bar{u}_h))v_h,v_h\rangle + \langle(\Phi''(\bar{u})-\Phi_h''(\bar{u}_h))v_h,v_h\rangle+(\bar{\mu}_h-\bar{\mu})\|v_h\|_{\mathrm{L}^2}^2.$$

The part involving the nonlinear local functional X can obviously be bounded by $C(h+h^\alpha)\,\|\nabla v_h\|_{\mathrm{L}^2}^2$; see (2.11). Using the expression for Φ'' presented in Lemma 2.1 and its discrete analogue, as well as the convergence of $\bar{\phi}_h$ to $\bar{\phi}$, we see that also $|\langle(\Phi''(\bar{u})-\Phi_h''(\bar{u}_h))v_h,v_h\rangle| \le Ch\,\|\nabla v_h\|_{\mathrm{L}^2}^2$. Finally, since $|\bar{\mu}_h-\bar{\mu}|\le Ch$ we get $|\bar{\mu}_h-\bar{\mu}|\,\|v_h\|_{\mathrm{L}^2}^2 \le Ch\,\|\nabla v_h\|_{\mathrm{L}^2}^2$ by Poincaré's inequality.

Summarising, we have established that

$$\langle\nabla_{uu}^2\mathcal{L}_h(\bar{u}_h,\bar{\mu}_h)v_h,v_h\rangle \ge \gamma/2\,\|\nabla v_h\|_{\mathrm{L}^2}^2 \qquad \forall v_h\in\ker c'(\bar{u}_h)\cap S_h,$$

for sufficiently small h. From Theorem 5.6 in Maurer & Zowe[27] we deduce that $\bar{u}_h$ is a uniform minimizer of E_h subject to $c(\bar{u}_h)=0$. □

4. Numerical Example

We have implemented a two-dimensional version of the discretization described above using piecewise linear, respectively, cubic Lagrange finite elements. The two-dimensional functional is obtained by replacing 8π with 4π in the definition of Φ (2.5) and the function F with $F(u)=\pi u^4/2-\frac{4\sqrt{2}}{3\pi^{5/2}}|u|^3$, see Ghosh & Dhara.[19]

To solve the nonlinear system (3.4) we apply Newton's method. The justification for this choice emerges from the availability of good initial guesses for u. Physical insight suggests that the electron density is higher close to nuclei. In fact, close to nuclei we expect the electron density to be close to the case 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 nuclei. An initial guess for ϕ is obtained by solving $-\Delta\phi = 4\pi(u^2-\rho_n)$. 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 fast local convergence immediately. We conclude that there is no need to apply globalization strategies.

Results of a computation involving a cluster of 39 atoms in two dimensions are documented in Figure 1. The configuration was obtained by slightly perturbing a hexagonal configuration. Parameters used in this computation were $\Omega=(-30,30)^2$, $\lambda=3.2$, $\sigma_0=0.5$ and $Z_i=6$ for all $i=1,\dots,39$. The solution $\bar{u}_h$ is shown on the left. Note that the solution is only shown on a subset of Ω to highlight the interesting features. On the right-hand side relative errors of E and μ are plotted against the number of degrees of freedom, N_{DoF}, for linear (P1), respectively, cubic (P3) finite 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

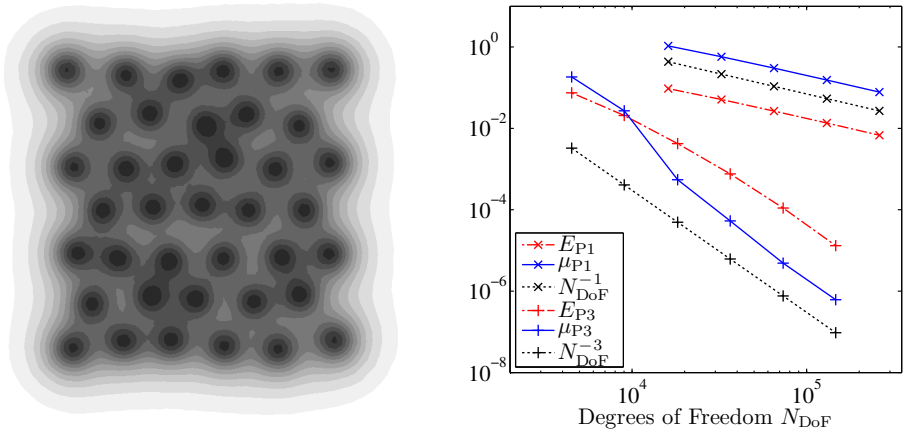

Fig. 1. Results of calculations for a cluster of 39 atoms in two dimensions: $\bar{u}_h$ is shown on the left, relative errors of E and μ are plotted on the right (along with lines indicating decay with N_{DoF}^{-1} and N_{DoF}^{-3}). Since the results are in 2D, the discretization parameter satisfies $h \sim N_{\mathrm{DoF}}^{-1/2}$. Hence, the plot suggests that the errors of E and μ behave like $\mathcal{O}(h^2)$ for P1 elements and $\mathcal{O}(h^6)$ for P3 elements, respectively.

finite element space. This has been proven for planewave and finite element discretizations of a slightly simpler class of nonlinear eigenvalue problems in Cancès, Chakir & Maday[6] (see Theorems 2 and 3). In Cancès, Chakir & Maday[7] (Theorem 3.1) the authors prove for a planewave discretization of the Thomas–Fermi–von Weizsäcker functional that the convergence order of the Lagrange multiplier is the same as for the energy. A rigorous study of convergence rates for a finite element discretization with numerical integration will be provided in our forthcoming work Langwallner, Ortner & Süli.[24] The main difference between the work in Cancès, Chakir & Maday[7] and ours is that the plane wave discretization allows to treat the electrostatic interactions directly, whereas in the finite element discretization, we solve an enlarged system.

Conclusions and Future Work

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 paper, we have combined 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. All results of the present work can be maintained if numerical integration of sufficiently high order is used. We will show this in a forthcoming article.[24]

In our current research we are developing a theory for optimal convergence rates

in situations where the Laplace operator has higher boundary regularity, and we are studying how the accuracy of the electronic density and electrostatic potential affects the accuracy of the nuclei when they are relaxed to their equilibrium positions. Another extension, which will require a careful study, is the effect of "cutting off" the electronic density and the electrostatic potential by imposing Dirichlet boundary values on $\partial\Omega$.

We view the present paper as the first in a series of papers, and as a preliminary step in our effort to develop a theory for coarse-graining the TFDW functional in the spirit of Garcia-Cervera, Lu & E[16] and Gavini et al.[17]

5. Appendix: An Indefinite Elliptic System with Constraint

In this appendix we generalize Schatz' classical result on the Galerkin discretization of indefinite elliptic equations[32] to the constrained system, which we encountered in the analysis in Section 3.1.

We assume that $\Omega \subset \mathbb{R}^d$, $d \geq 1$, 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

$$|u|_{\mathrm{H}^2} \leq C_{\mathrm{reg}} \, \|f\|_{\mathrm{L}^2}. \tag{5.1}$$

With $d = 3$ this is exactly the assumption we made in Section 2.2.

Let $m \geq 1$ and let $\boldsymbol{V} \subset \mathrm{H}^1_0(\Omega; \mathbb{R}^m)$ 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 $\boldsymbol{V}$ can be written as

$$\boldsymbol{V} = \big\{ \boldsymbol{v} \in \mathrm{H}^1_0(\Omega; \mathbb{R}^m) : (\boldsymbol{v}, \boldsymbol{g}) = 0 \big\}.$$

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

$$a(\boldsymbol{u}, \boldsymbol{v}) = (\nabla \boldsymbol{u}, \nabla \boldsymbol{v}) + (\boldsymbol{u}, \mathrm{M}\boldsymbol{v}) \qquad \text{for } \boldsymbol{u}, \boldsymbol{v} \in \mathrm{H}^1_0(\Omega; \mathbb{R}^m), \tag{5.2}$$

where $\mathrm{M} \in \mathrm{L}^\infty(\Omega; \mathbb{R}^{m \times m})$.

We immediately see that a is bounded:

$$a(\boldsymbol{u}, \boldsymbol{v}) \leq C_a \, \|\nabla \boldsymbol{u}\|_{\mathrm{L}^2} \, \|\nabla \boldsymbol{v}\|_{\mathrm{L}^2} \qquad \forall \, \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}, \tag{5.3}$$

and that there exists a constant $K \geq 0$ such that $a(\boldsymbol{u}, \boldsymbol{v}) + K(\boldsymbol{u}, \boldsymbol{v})$ is coercive,

$$a(\boldsymbol{u}, \boldsymbol{u}) + K(\boldsymbol{u}, \boldsymbol{u}) \geq \alpha \, \|\nabla \boldsymbol{u}\|^2_{\mathrm{L}^2} \qquad \forall \, \boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}. \tag{5.4}$$

We assume that for every $\boldsymbol{f} \in \mathrm{L}^2(\Omega; \mathbb{R}^m)$ there is a unique solution $u \in \boldsymbol{V}$ of the variational problem

$$a(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{V}. \tag{5.5}$$

The adjoint variational problem to (5.5) is given by

$$a(\boldsymbol{v}, \boldsymbol{u}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{V}.$$

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, and which carries over verbatim from the scalar case (see Theorem 6.2.4 in Evans[12]), that the adjoint problem also has a unique solution. This implies, in particular, the existence of $\kappa > 0$ such that

$$\inf_{\boldsymbol{u}\in\boldsymbol{V}} \sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{a(\boldsymbol{u}, \boldsymbol{v})}{\|\boldsymbol{u}\|_{\mathrm{H}^1} \|\boldsymbol{v}\|_{\mathrm{H}^1}} \geq \kappa \quad \text{and} \quad \inf_{\boldsymbol{u}\in\boldsymbol{V}} \sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{a(\boldsymbol{v}, \boldsymbol{u})}{\|\boldsymbol{u}\|_{\mathrm{H}^1} \|\boldsymbol{v}\|_{\mathrm{H}^1}} \geq \kappa.$$

From that we can infer the solvability of (5.5) for all $\boldsymbol{f} \in \boldsymbol{V}^*$.

To define a Galerkin discretization of (5.5), let $(S_{h,0})_{h\in(0,1]}$ be a family of finite-dimensional subspaces of $\mathrm{H}^1_0(\Omega)$, which satisfy the approximation property

$$\inf_{u_h\in S_{h,0}} \|\nabla(u - u_h)\|_{\mathrm{L}^2} \leq C_{\mathrm{apx}} h |u|_{\mathrm{H}^2} \quad \text{for all } u \in \mathrm{H}^1_0(\Omega) \cap \mathrm{H}^2(\Omega), \tag{5.6}$$

for every h, where C_{apx} is independent of h. We then define the approximation space

$$\boldsymbol{V}_h = \big\{\boldsymbol{v}_h \in S^m_{h,0} : (\boldsymbol{g}, \boldsymbol{v}_h) = 0 \text{ for } i = 1, \ldots, m\big\}.$$

The Galerkin discretization of (5.5) is given by

$$a(\boldsymbol{u}_h, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{V}_h. \tag{5.7}$$

Note that it may occur that $\boldsymbol{V}_h = S^m_{h,0}$. However, it follows immediately from (5.6) that, for sufficiently small h, the co-dimension of $\boldsymbol{V}_h$ in $S^m_{h,0}$ is also one.

Our main result of this appendix ensures solvability of the Galerkin discretization (5.7), provided that h is sufficiently small. The proof parallels Theorem 5.7.6 in Brenner & Scott.[3]

Theorem 5.1. *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* (5.7)*, satisfying*

$$\|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathrm{L}^2} \leq C_1 h |\boldsymbol{u}|_{\mathrm{H}^2} \quad \textit{and} \quad \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathrm{L}^2} \leq C_2 h^2 |\boldsymbol{u}|_{\mathrm{H}^2},$$

where $\boldsymbol{u}$ is the exact solution of (5.5)*. Furthermore, there exists $\kappa_d > 0$ such that,*

$$\inf_{\boldsymbol{v}_h\in\boldsymbol{V}_h} \sup_{\boldsymbol{w}_h\in\boldsymbol{V}_h} \frac{a(\boldsymbol{v}_h, \boldsymbol{w}_h)}{\|\nabla \boldsymbol{v}_h\|_{\mathrm{L}^2} \|\nabla \boldsymbol{w}_h\|_{\mathrm{L}^2}} \geq \kappa_d \qquad \forall h \in (0, h_0]. \tag{5.8}$$

Remark 5.1. We remark that Theorem 5.1 as well as the following auxiliary Lemmas hold for any finite number of linear constraints. For example, if $\boldsymbol{V}$ is given by

$$\boldsymbol{V} = \big\{\boldsymbol{v} \in \mathrm{H}^1_0(\Omega; \mathbb{R}^m) : (\boldsymbol{g}_i, \boldsymbol{v}) = 0,\ i = 1, \ldots, n\big\},$$

where $\boldsymbol{g}_i \in \mathrm{L}^2(\Omega; \mathbb{R}^m)$, and if the functions $\boldsymbol{g}_i$, $i = 1, \ldots, n$, 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 5.1 requires two auxiliary results, which we provide in the following two lemmas. The first result shows that the constrained variational problem (5.5) inherits the H^2-regularity of the Laplace operator.

Lemma 5.1. *Let* $\boldsymbol{f} \in \mathrm{L}^2(\Omega;\mathbb{R}^m)$ *and let* $\boldsymbol{u} \in \boldsymbol{V}$ *be the solution of* (5.5)*, 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 result follows by explicitly computing a representation of $\boldsymbol{u}$ in terms of the solution $\tilde{\boldsymbol{u}}$ of a Poisson problem in the entire space $\mathrm{H}^1_0(\Omega;\mathbb{R}^m)$:

$$(\nabla\tilde{\boldsymbol{u}}, \nabla\boldsymbol{v}) = (\boldsymbol{f} - \mathrm{M}\boldsymbol{u}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \mathrm{H}^1_0(\Omega;\mathbb{R}^m).$$

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

$$\tilde{\boldsymbol{u}} - t\boldsymbol{e}, \quad \text{where} \quad t = \frac{(\nabla\tilde{\boldsymbol{u}}, \nabla\boldsymbol{e})}{\|\nabla\boldsymbol{e}\|^2},$$

belongs to $\boldsymbol{V}$, and that

$$(\nabla(\tilde{\boldsymbol{u}} - t\boldsymbol{e}), \nabla\boldsymbol{v}) = (\nabla\tilde{\boldsymbol{u}}, \nabla\boldsymbol{v}) - t(\boldsymbol{g}, \boldsymbol{v}) = (\boldsymbol{f} - \mathrm{M}\boldsymbol{u}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{V}.$$

Hence, we deduce that

$$\boldsymbol{u} = \tilde{\boldsymbol{u}} - t\boldsymbol{e}.$$

Assumption (5.1) ensures that $\tilde{\boldsymbol{u}}$ and $\boldsymbol{e}$ belong to $\mathrm{H}^2(\Omega;\mathbb{R}^m)$, and therefore, we can deduce that $\boldsymbol{u} \in \mathrm{H}^2(\Omega)$. The existence of C'_{reg} follows from the open mapping theorem and the fact that $|\boldsymbol{u}|_{\mathrm{H}^2}$ is a norm on $\boldsymbol{V}$. □

Our second auxiliary result shows that the constrained subspace $\boldsymbol{V}_h$ inherits the approximation property (5.6) of $S_{h,0}$.

Lemma 5.2. *There exists* $h_0 > 0$ *and a constant* C'_{apx} *such that*

$$\inf_{\boldsymbol{u}_h \in \boldsymbol{V}_h} \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathrm{L}^2} \leq C'_{\mathrm{apx}} h |\boldsymbol{u}|_{\mathrm{H}^2} \quad \textit{for all } \boldsymbol{u} \in \boldsymbol{V} \cap \mathrm{H}^2(\Omega;\mathbb{R}^m) \quad \textit{and for } h \in (0, h_0].$$

Proof. Let $\boldsymbol{e}_h \in S^m_{h,0}$ be the solution of

$$(\nabla\boldsymbol{e}_h, \nabla\boldsymbol{v}) = (\boldsymbol{g}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in S^m_{h,0}.$$

It is not difficult to verify, for h sufficiently small, say $h \in (0, h_0]$, that there exists $\boldsymbol{v} \in S^m_{h,0}$ such that $(\boldsymbol{g}, \boldsymbol{v}) \neq 0$ and hence $\boldsymbol{e} \neq 0$ for $h \in (0, h_0]$.

Let $\tilde{\boldsymbol{u}}_h \in S^m_{h,0}$ be the Ritz projection of $\boldsymbol{u}$, i.e.,

$$(\nabla\tilde{\boldsymbol{u}}_h, \nabla\boldsymbol{v}) = (\nabla\boldsymbol{u}, \nabla\boldsymbol{v}) \qquad \forall\boldsymbol{v} \in S^m_{h,0}.$$

We construct the final approximant $\boldsymbol{u}_h$, as in the proof of Lemma 5.1,

$$\boldsymbol{u}_h = \tilde{\boldsymbol{u}}_h - \frac{(\nabla\tilde{\boldsymbol{u}}_h, \nabla\boldsymbol{e}_h)}{\|\nabla\boldsymbol{e}_h\|^2}\boldsymbol{e}_h \in \boldsymbol{V}_h.$$

Since $(\boldsymbol{g}, \boldsymbol{u}) = 0$, the error $\|\nabla(\boldsymbol{u}_h - \tilde{\boldsymbol{u}}_h)\|_{\mathrm{L}^2}$ can be estimated as follows:

$$\|\nabla(\boldsymbol{u}_h - \tilde{\boldsymbol{u}}_h)\|_{\mathrm{L}^2} = \frac{|(\nabla\boldsymbol{e}_h, \nabla\tilde{\boldsymbol{u}}_h)|}{\|\nabla\boldsymbol{e}_h\|_{\mathrm{L}^2}} = \frac{|(\boldsymbol{g}, \tilde{\boldsymbol{u}}_h - \boldsymbol{u})|}{\|\nabla\boldsymbol{e}_h\|_{\mathrm{L}^2}} \leq \frac{\|\boldsymbol{g}\|_{\mathrm{H}^{-1}}}{\|\nabla\boldsymbol{e}_h\|_{\mathrm{L}^2}} \|\nabla(\boldsymbol{u} - \tilde{\boldsymbol{u}}_h)\|_{\mathrm{L}^2}.$$

Since $\boldsymbol{e}_h$ converges in $\mathrm{H}^1_0(\Omega)$ 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 (5.6) we arrive at the desired approximation result. □

Proof of Theorem 5.1.

Step 1: The Schatz argument. The main part of the proof follows an argument originally given by Schatz.[32] Our presentation is largely analogous to Theorem 5.7.6 in Brenner & Scott.[3] 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 $\boldsymbol{v} \in \boldsymbol{V}_h$. Then, using (5.4) we deduce that

$$\alpha \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|^2_{\mathrm{L}^2} \leq C_a \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathrm{L}^2} \|\nabla(\boldsymbol{u} - \boldsymbol{v})\|_{\mathrm{L}^2} + K \|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{\mathrm{L}^2}, \qquad (5.9)$$

for every $\boldsymbol{v} \in \boldsymbol{V}_h$, where we have used the continuity of a. Considering the adjoint problem

$$a(\boldsymbol{z}, \boldsymbol{w}) = (\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{z}) \qquad \forall\boldsymbol{z} \in \boldsymbol{V}$$

we can show that

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|^2_{\mathrm{L}^2} \leq C'_{\mathrm{reg}} h \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathrm{L}^2} |\boldsymbol{w}|_{\mathrm{H}^2} \leq C'_{\mathrm{reg}} h \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathrm{L}^2} \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathrm{L}^2},$$

where Lemma 5.1 was used to obtain H^2-regularity for $\boldsymbol{w}$. This results in

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathrm{L}^2} \leq C'_{\mathrm{reg}} h \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathrm{L}^2}. \qquad (5.10)$$

Applying this bound to (5.9) and choosing h sufficiently small, we obtain the bound

$$\|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{\mathrm{L}^2} \leq C \|\nabla(\boldsymbol{u} - \boldsymbol{v})\|_{\mathrm{L}^2} \qquad \forall\boldsymbol{v} \in \boldsymbol{V}_h. \qquad (5.11)$$

$\boldsymbol{V}_h$ 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 (5.11) 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 $\boldsymbol{v} \in \boldsymbol{V}_h$ in (5.11) and using Lemma 5.2 yields the first error bound stated in the theorem. Combining this bound with (5.10) provides the second error bound.

Step 2. Uniform Inf-Sup Constant. Unique solvability of (5.5) for $\boldsymbol{f} = \boldsymbol{0}$ implies that a satisfies the inf-sup condition

$$\inf_{\boldsymbol{u}\in\boldsymbol{V}} \sup_{\boldsymbol{v}\in\boldsymbol{V}} \frac{a(\boldsymbol{u},\boldsymbol{v})}{\|\nabla\boldsymbol{u}\|_{\mathrm{L}^2}\,\|\nabla\boldsymbol{v}\|_{\mathrm{L}^2}} = \kappa > 0. \tag{5.12}$$

Our aim now is to prove the validity of a corresponding condition for $\boldsymbol{V}_h \times \boldsymbol{V}_h$, which is uniform in h.

To obtain $\mathcal{O}(h)$-convergence of $\boldsymbol{u}_h$ to $\boldsymbol{u}$ it had to be assumed that $\boldsymbol{f} \in \mathrm{L}^2(\Omega)^m$. However, both the continuous and discrete solution $\boldsymbol{u} \in \boldsymbol{V}$, respectively $\boldsymbol{u}_h \in \boldsymbol{V}_h$, exist and are unique if $\boldsymbol{f} \in \mathrm{H}^{-1}(\Omega)^m$,

$$a(\boldsymbol{u},\boldsymbol{v}) = \langle \boldsymbol{f},\boldsymbol{v}\rangle \quad \forall v \in \boldsymbol{V} \quad \text{and} \quad a(\boldsymbol{u}_h,\boldsymbol{v}) = \langle \boldsymbol{f},\boldsymbol{v}\rangle \quad \forall v \in \boldsymbol{V}_h.$$

We note that, to prove (5.11), 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 (5.11) and using a triangle inequality and the inf-sup condition (5.12) we deduce that,

$$\|\nabla\boldsymbol{u}_h\|_{\mathrm{L}^2} \le (1+C)\,\|\nabla\boldsymbol{u}\|_{\mathrm{L}^2} \le (1+C)\kappa^{-1}\|\boldsymbol{f}\|_{\boldsymbol{V}^*} \qquad \forall h \in (0,h_0], \tag{5.13}$$

where $h_0 > 0$ is chosen sufficiently small, and where C is the constant from (5.11). If $\boldsymbol{f} \in \boldsymbol{V}_h^*$ then, by the Hahn–Banach theorem, $\boldsymbol{f}$ can be extended to an element of $\boldsymbol{V}^*$ while preserving its norm, and hence, we obtain

$$\left\|\nabla L_h^{-1}\boldsymbol{f}\right\|_{\mathrm{L}^2} \le (1+C)\kappa^{-1}\|\boldsymbol{f}\|_{\boldsymbol{V}_h^*} \qquad \forall \boldsymbol{f} \in \boldsymbol{V}_h^* \quad \forall h \in (0,h_0],$$

where L_h^{-1} denotes the solution operator for (5.7). This statement is equivalent to the uniform inf-sup condition (5.8) with $\kappa_d = \kappa/(1+C)$. □

Acknowledgment

The authors were supported by the EPSRC project 'New Frontiers in the Mathematics of Solids'.

References

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. X. Blanc, C. Le Bris, and P.-L. Lions. From molecular models to continuum mechanics. *Arch. Ration. Mech. Anal.*, 164(4):341–381, 2002.
3. S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002.
4. F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

5. 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.
6. E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of nonlinear eigenvalue problems. *arXiv:0905.1645v2*, 2009.
7. 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.
8. 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.
9. 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.
10. B. Dacorogna. *Direct methods in the calculus of variations*, volume 78 of *Applied Mathematical Sciences*. Springer, New York, second edition, 2008.
11. M. Dobrowolski and R. Rannacher. Finite element methods for nonlinear elliptic systems of second order. *Math. Nachr.*, 94:155–172, 1980.
12. L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1998.
13. M. Finnis. *Interatomic Forces in Condensed Matter*. Oxford University Press, USA, 2003.
14. C. J. García-Cervera. An efficient real space method for orbital-free density-functional theory. *Commun. Comput. Phys.*, 2(2):334–357, 2007.
15. 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.
16. 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.
17. 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.
18. 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.
19. 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.
20. P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.
21. P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev*, 136(3B):B864–B871, 1964.
22. W. Kohn. Nobel Lecture: Electronic structure of matterwave functions and density functionals. *Reviews of Modern Physics*, 71(5):1253–1266, 1999.
23. W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev*, 140(4A):A1133–A1138, 1965.
24. B. Langwallner, C. Ortner, and E. Süli. Optimal convergence rates for the finite element discretization of an electronic density functional. *In preparation.*
25. E. H. Lieb. Thomas–Fermi and related theories of atoms and molecules. *Rev. Modern Phys.*, 53(4):603–641, 1981.
26. L. Lignère and A. C. Carter. An introduction to orbital-free density functional theory. In S. Yip, editor, *Handbook of Materials Modeling*, volume 15 of *Texts in Applied Mathematics*, chapter 1.8. Springer-Verlag, second edition, 2005.

27. 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.
28. 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.
29. D. Negrut, M. Anitescu, T. Munson, and P. Zapol. Simulating nanoscale processes in solids using DFT and the quasicontinuum method (IMECE2005-81755. *Proceedings of IMECE*, 2005.
30. R.G. Parr and W. Yang. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, USA, 1989.
31. 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.
32. A. H. Schatz. An observation concerning Ritz–Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28:959–962, 1974.
33. 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.
34. 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.
35. 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.
36. 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.