

NEURAL NETWORK APPROXIMATION OF REGULARIZED DENSITY FUNCTIONALS

MIHÁLY A. CSIRIK¹ AND ANDRE LAESTADIUS^{1,2} AND MATHIAS OSTER³

¹*Department of Computer Science, Oslo Metropolitan University, Norway*

²*Hylleraas Centre for Quantum Molecular Sciences, Department of Chemistry, University of Oslo, Norway*

³*Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany*

ABSTRACT. Density-functional theory is one of the most efficient and widely used computational methods of quantum mechanics, especially in fields such as solid state physics and quantum chemistry. From the theoretical perspective, its central object is the universal density functional which contains all intrinsic information about the quantum system in question. Once the external potential is provided, in principle one can obtain the exact ground-state energy via a simple minimization. However, the universal density functional is a very complicated mathematical object and almost always it is replaced with its approximate variants. So far, no “first principles”, mathematically consistent and convergent approximation procedure has been devised that has general applicability. In this paper, we propose such a procedure by first applying Moreau–Yosida regularization to make the exact functionals continuous (even differentiable) and then approximate the regularized functional by a neural network. The resulting neural network preserves the positivity and convexity of the exact functionals. More importantly, it is differentiable, so it can be directly used in a Kohn–Sham calculation.

1. INTRODUCTION

Density-functional theory (DFT) has its origins in Thomas–Fermi theory, one of the first methods that made ground-state energy calculations for many-particle systems in quantum mechanics viable. The main theoretical underpinning of DFT is the Hohenberg–Kohn theorem [13], which says that for a many-electron Hamiltonian, the ground-state (electron) density determines the external potential up to a constant. This implies that the ground-state density in turn determines the ground-state *wavefunction*. Hence, it is enough to consider the density as the main variable for ground-state energy computations, as all relevant physical properties, at least in principle, can be reconstructed from a ground-state density.

The rigorous mathematical foundations of DFT were laid down by Lieb [19], who also introduced the so-called Lieb universal density functional as the convex conjugate of the ground-state energy as a function of the potential. This convenient

E-mail address: oster@igpm.rwth-aachen.de.

convex analysis setting is the starting point of many mathematics articles about DFT, and we will also adopt this setting.

The Lieb universal functional lives on L^1 , and is weakly lower semicontinuous there. Clearly, we do not have control over the kinetic energy of a weakly convergent sequence of densities in L^1 and it was shown in [17] that the Lieb functional is everywhere discontinuous in that topology. To remedy the situation, [14] introduced the use of Moreau–Yosida regularization into DFT. This way, one obtains in an invertible manner a convex and differentiable functional on the whole density space. The benefit of this procedure is that the corresponding Kohn–Sham potentials are now meaningful [14], and one may devise convergent algorithms [15, 24, 25] as well as a rigorous formulation of density-potential inversion [22, 11, 5, 23]. In all these works, the full space $\mathbb{R}^3$ needs to be truncated to a bounded domain in order to get a reflexive L^p density space, so that the dual space contains Coulombic potentials. We address this shortcoming in this work, so that no domain truncation is necessary for the application of the Moreau–Yosida regularization.

It remains one of the main challenges of many-body quantum physics to devise a consistent and convergent approximation to the universal density functionals (specifically the exchange–correlation functional). All the current approximate functionals, except for the local density approximation (LDA), contain tuning parameters to be determined based on the use cases [26]. In contrast, the neural network approach furnishes a first principles approximation to the density functionals.

Outline. The main results of this paper are the following. First, we generalize the universal approximation property on separable Banach spaces to also respect constraints. Thereafter, we will show that one can extend the concepts of Moreau–Yosida regularization to the case of non-reflexive separable Banach spaces. Lastly, these two findings are combined into an error estimate on the ground-state energy.

Acknowledgments. AL and MACs have received funding from the ERC-2021-STG under grant agreement No. 101041487 REGAL. AL were funded by the Research Council of Norway through CoE Hylleraas Centre for Quantum Molecular Sciences Grant No. 262695. MO has received funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 442047500 – through the Collaborative Research Center “Sparsity and Singular Structures” (SFB 1481).

2. PRELIMINARY NOTIONS

In the following, we will quickly review the basics of density-functional theory and approximation theory for neural networks. Our discussion will be restricted to 3 spatial dimensions, since we are mainly concerned here with density-functional theory in the quantum chemistry setting. We attempt to give a self-contained presentation when possible, however for brevity a lot of details will be left out. The interested reader can consult [7, 19, 18] for a more detailed introduction to DFT.

2.1. Schrödinger Hamiltonian. For the model Hamiltonian, we will consider the so-called Schrödinger Hamiltonian which describes (for simplicity) *spinless* electrons in an external electric potential. More precisely, the Hamiltonian is given by the Friedrichs extension of the densely defined and form lower semibounded operator

on the N -electron antisymmetric Hilbert space $L_a^2(\mathbb{R}^{3N})$ defined as

$$H_N^v = H_N^0 + V, \quad V = \sum_{j=1}^N v(x_j),$$

where the external potential v is in the space $L^\infty(\mathbb{R}^3) + L^{3/2}(\mathbb{R}^3)$ and the internal Hamiltonian is given by

$$H_N^0 = -\sum_{j=1}^N \Delta_{x_j} + \sum_{1 \leq j < k \leq N} \frac{1}{|x_j - x_k|},$$

with form domain $Q(H_N^v) = Q(H_N^0) = H_a^1(\mathbb{R}^{3N})$, the Sobolev space $H^1(\mathbb{R}^{3N})$ intersected with $L_a^2(\mathbb{R}^{3N})$. The (canonical) ground-state energy as a function of the potential is given by the variational principle

$$E_N^C(v) = \inf_{\substack{\Psi \in H_a^1(\mathbb{R}^{3N}) \\ \|\Psi\|=1}} \langle \Psi, H_N^v \Psi \rangle. \quad (1)$$

As an infimum of affine functions, $v \mapsto E_N^C(v)$ is concave and since H_N^v is lower semibounded, we have $E_N^C(v) > -\infty$. Moreover, $E_N^C(v)$ is Lipschitz continuous and order preserving in the sense that whenever $v_1 \leq v_2$ a.e., then $E_N^C(v_1) \leq E_N^C(v_2)$ (see [19, Theorem 3.1] for a proof of these properties). Note that a minimizing Ψ might not exist.

2.2. Canonical universal density functionals. The crucial fact that starts the development of a “density-functional theory” for the ground-state problem of H_N^v is that due to the special structure of the Hamiltonian, its quadratic form $\langle \Psi, H_N^v \Psi \rangle$ splits into two parts,

$$\langle \Psi, H_N^v \Psi \rangle = \langle \Psi, H_N^0 \Psi \rangle + \langle v, \rho_\Psi \rangle. \quad (2)$$

Here, the second term only depends on the *density*

$$\rho_\Psi(x) = N \int_{\mathbb{R}^{3(N-1)}} |\Psi(x, x_2, \dots, x_N)|^2 dx_2 \dots dx_N$$

of the wavefunction Ψ . Notice that $\int_{\mathbb{R}^3} \rho_\Psi = N$ precisely if $\|\Psi\| = 1$. Hence, we may write (1) using (2),

$$\begin{aligned} E_N^C(v) &= \inf_{\substack{\Psi \in H_a^1(\mathbb{R}^{3N}) \\ \|\Psi\|=1}} \left[\langle \Psi, H_N^0 \Psi \rangle + \langle v, \rho_\Psi \rangle \right] = \inf_{\rho \in \mathcal{I}_N} \inf_{\rho_\Psi = \rho} \left[\langle \Psi, H_N^0 \Psi \rangle + \langle v, \rho \rangle \right] \\ &= \inf_{\rho \in \mathcal{I}_N} \left[\inf_{\rho_\Psi = \rho} \langle \Psi, H_N^0 \Psi \rangle + \langle v, \rho \rangle \right] = \inf_{\rho \in \mathcal{I}_N} [F_{\text{LL}}(\rho) + \langle v, \rho \rangle], \end{aligned} \quad (3)$$

where we introduced the N -representable set $\mathcal{I}_N$ which collects all the functions $\rho \in L^1(\mathbb{R}^3, \mathbb{R}_+)$ with $\int_{\mathbb{R}^3} \rho = N$ for which there is a (normalized) wavefunction $\Psi \in H_a^1(\mathbb{R}^{3N})$ such that $\rho_\Psi = \rho$. We also introduced the *canonical Levy–Lieb functional*

$$F_{\text{LL}}(\rho) = \inf_{\substack{\Psi \in H_a^1(\mathbb{R}^{3N}) \\ \rho_\Psi = \rho}} \langle \Psi, H_N^0 \Psi \rangle$$

for every $\rho \in \mathcal{I}_N$. We see that the use of the adjective “universal” is justified as $F_{\text{LL}}(\rho)$ is independent of the external potential v . Lieb also showed that the “inf” in the definition of $F_{\text{LL}}(\rho)$ is attained. The N -representable set $\mathcal{I}_N$ admits an easy and very useful description due to Lieb [19], which we will quickly sketch below.

First, we recall the *Hoffmann-Ostenhof inequality* [12], which says that for *any* self-adjoint positive operator γ on $L^2(\mathbb{R}^3)$ the kinetic energy bound

$$\mathrm{Tr}(-\Delta\gamma) \geq \int_{\mathbb{R}^3} |\nabla \sqrt{\rho_\gamma}|^2 \quad (4)$$

holds, where the density ρ_γ of γ may be defined via duality. This implies that for states γ with finite kinetic energy, we have $\sqrt{\rho_\gamma} \in H^1(\mathbb{R}^3)$.

The converse of this last statement is essentially the characterization of $\mathcal{I}_N$, due to Harriman [9] and Lieb [19, Theorem 1.2], which says that

$$\mathcal{I}_N = \left\{ \rho \in L^1(\mathbb{R}^3; \mathbb{R}_+) : \nabla \sqrt{\rho} \in L^2(\mathbb{R}^3), \int_{\mathbb{R}^3} \rho = N \right\}, \quad (5)$$

for any $N \in \mathbb{N}$. We note that $\mathcal{I}_N$ is convex, due the convexity of $\rho \mapsto \int_{\mathbb{R}^3} |\nabla \sqrt{\rho}|^2$, see [20]. Moreover, $\mathcal{I}_N \subset L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3)$ by a Sobolev inequality.

Even though the chain of equalities (3) suggests that $E_N^C(v)$ and $F_{\mathrm{LL}}(\rho)$ are Legendre transform pairs, as expected from a statistical physics point of view, this is unfortunately not the case, as the canonical Levy–Lieb functional is *not* convex [19, Theorem 3.4]. To remedy this, Lieb considered the convex hull (equivalently the double Legendre transform) of $F_{\mathrm{LL}}(\rho)$, which is nowadays called the (canonical) Lieb functional, and is given by

$$F_{\mathrm{L}}(\rho) = \sup_{v \in X^*} [E_N^C(v) - \langle v, \rho \rangle] \quad (6)$$

for all $\rho \in X$, where we let

$$X = L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3), \quad \text{and} \quad X^* = L^\infty(\mathbb{R}^3) + L^{3/2}(\mathbb{R}^3)$$

denote the space of *quasidensities* and of potentials, respectively. The former is equipped with the norm $\|\rho\|_X = \max\{\|\rho\|_1, \|\rho\|_3\}$ and the latter with

$$\|v\|_{X^*} = \inf\{\|v_1\|_\infty + \|v_2\|_{3/2} : v = v_1 + v_2, v_1 \in L^\infty(\mathbb{R}^3), v_2 \in L^{3/2}(\mathbb{R}^3)\}.$$

Because of its convexity, the Lieb functional and its generalizations are preferred over the Levy–Lieb functional by many authors [23] as the machinery of convex analysis can be applied to study them. Moreover, the “Lieb variational principle” (6) may be used in numerical computations, see [10] for a review.

Due to technical reasons explained below, we will consider the grand-canonical extension of the Lieb functional.

2.3. The grand-canonical universal density functional. We begin by recalling that the Lieb functional can also be written as a minimization over mixed states as

$$F_{\mathrm{L}}(\rho) = \inf_{\substack{0 \leq \Gamma = \Gamma^\dagger \leq 1 \\ \mathrm{Tr} \Gamma = 1 \\ \rho_\Gamma = \rho}} \mathrm{Tr}_{L^2(\mathbb{R}^{3N})} H_N^0 \Gamma,$$

for $\rho \in \mathcal{I}_N$ and $F_{\mathrm{L}}(\rho) \equiv +\infty$ otherwise. By [19, Corollary 4.5] the r.h.s. actually defines a weakly lower semicontinuous functional in the L^1 -topology, and the infimum is attained.

In the grand-canonical case we do not have a definite particle number N , so the N -representability set becomes simply

$$\mathcal{I} = \left\{ \rho \in L^1(\mathbb{R}^3; \mathbb{R}_+) : \nabla \sqrt{\rho} \in L^2(\mathbb{R}^3) \right\}.$$

For every $\rho \in \mathcal{I}$, we define

$$F_{\text{GC}}(\rho) = \inf_{\substack{\Gamma \in \mathcal{D} \\ \rho_\Gamma = \rho}} \sum_{n \geq 1} \text{Tr}_{L^2(\mathbb{R}^{3n})} H_n^0 \Gamma_n \quad (7)$$

and extend $F_{\text{GC}}(\rho) \equiv +\infty$ for $\rho \in X \setminus \mathcal{I}$. Here, $\mathcal{D}$ stands for the space of Fock space density matrices which commute with the number operator [21, 18]. By dropping the Coulomb interaction from the definition we obtain the kinetic energy functional $T(\rho)$.

Taking the Legendre transform of $F_{\text{GC}}(\rho)$ at some fixed (possibly fractional) particle number $\lambda \in \mathbb{R}_+$, we may define the corresponding grand-canonical ground-state energy

$$E_\lambda^{\text{GC}}(v) = \inf_{\substack{\rho \in \mathcal{I} \\ \int_{\mathbb{R}^3} \rho = \lambda}} [F_{\text{GC}}(\rho) + \langle v, \rho \rangle]. \quad (8)$$

Here, we remark that while $E_N^{\text{GC}}(v) \leq E_N^{\text{C}}(v)$ certainly holds, the reverse inequality is probably not true in general [6, 2].

Furthermore, it is straightforward to generalize the methods of [19] to the grand-canonical case to obtain the following.

Theorem 2.1. *F_{GC} is weakly lower semicontinuous in the topology given by $(X, \|\cdot\|)$ and bounded subsets of level sets are weakly sequentially compact.*

Remark 1. We observe that Theorem 2.1 is true if the norm $\|\cdot\|$ is replaced by any equivalent norm.

The proof of the preceding theorem involves a useful property of the grand-canonical Levy–Lieb functional.

Theorem 2.2. [18, Theorem 5] *For any sequence $\{\sqrt{\rho_j}\} \subset H^1(\mathbb{R}^3)$ such that $\nabla \sqrt{\rho_j} \rightharpoonup \nabla \sqrt{\rho}$ in $\dot{H}^1(\mathbb{R}^3)$, there holds*

$$F_{\text{GC}}(\rho) \leq \liminf_{j \rightarrow \infty} F_{\text{GC}}(\rho_j).$$

Note that such a property does not hold true for the canonical Lieb functional because it might happen that $\int_{\mathbb{R}^3} \rho_j \rightarrow \int_{\mathbb{R}^3} \rho \neq N$, which would make the l.h.s. of the above inequality $+\infty$.

2.4. Neural Networks on Fréchet spaces. In this section, we briefly summarize and specialize the main results of [4] about the approximation qualities of neural networks relevant to us. In this section, X denotes an arbitrary (real) separable Banach space that admits a Schauder basis $\{e_n\}_{n \in \mathbb{N}}$. Fix a continuous map $\sigma : X \rightarrow X$, called *activation function*. For any bounded linear functional $\ell \in X^*$, bounded linear operator $A \in \mathcal{L}(X)$ and vector $b \in X$ define the (nonlinear) functional $\mathcal{N}_{\ell, A, b} : X \rightarrow \mathbb{R}$ called a *neuron* via

$$\mathcal{N}_{\ell, A, b}(u) = \langle \ell, \sigma(Au + b) \rangle \quad \text{for all } u \in X.$$

The linear hull of all neurons is denoted as

$$\mathfrak{N}(\sigma) = \text{Span}\{\mathcal{N}_{\ell, A, b} : \ell \in X^*, A \in \mathcal{L}(X), b \in X\}.$$

We call an activation function $\sigma : X \rightarrow X$ *discriminatory*, if for any fixed compact $K \subset X$ and any regular (positive) Borel measure μ on K , the relation

$$\int_K \langle \ell, \sigma(Au + b) \rangle \mu(du) = 0 \quad \text{for all } \ell \in X^*, A \in \mathcal{L}(X), b \in X,$$

implies that $\mu \equiv 0$.

The space $C(X, \mathbb{R})$ of all continuous functionals on X is equipped with the inductive topology of uniform convergence on compacts. More precisely, the topology is induced by the family of semi-norms

$$|F|_K = \sup_{u \in K} |F(u)|, \quad K \subset X \text{ compact.}$$

Then $C(X, \mathbb{R})$ is a Fréchet space.

The following generalization of Cybenko's famous theorem was proved in [4].

Theorem 2.3 (Universal approximation property). *Suppose that $\sigma : X \rightarrow X$ is discriminatory. Then the subspace $\mathfrak{N}(\sigma)$ is dense in $C(X, \mathbb{R})$. More explicitly, given $F \in C(X, \mathbb{R})$, for every $\varepsilon > 0$ and every compact $K \subset X$ there exists $\mathcal{NN} \in \mathfrak{N}(\sigma)$ such that $|F - \mathcal{NN}|_K < \varepsilon$.*

It is important to note that the neural network $\mathcal{NN}$ depends on the compact set K , and that the theorem does not provide any information about the behavior of $\mathcal{NN}$ outside K .

3. MAIN RESULTS

In this section, we describe the main results of our paper. The proofs can be found in Section 4 below.

3.1. Constrained approximation property of neural networks in separable Banach spaces. As we recalled above in Section 2.4 for infinite dimensional state spaces, it was shown in Ref. [4] that single-layer neural networks are dense in the Fréchet space of continuous functions with respect to the inductive topology. However, for many applications, the sought-after approximations should fulfill some additional constraints.

The next result is a simple generalization of Theorem 2.3, which allows the inclusion of some interesting classes of constraints that are relevant for our purposes.

Theorem 3.1. *Suppose that $\sigma : X \rightarrow X$ is discriminatory. Also, let $\mathfrak{C} \subset C(X, \mathbb{R})$ such that $\text{cl}(\mathfrak{C}) = \text{cl}(\text{int } \mathfrak{C})$. Then $\mathfrak{N}(\sigma) \cap \mathfrak{C}$ is dense in $\mathfrak{C}$. More explicitly, given $F \in C(X, \mathbb{R})$, for every $\varepsilon > 0$ and compact $K \subset X$ there exists $\mathcal{NN} \in \mathfrak{N}(\sigma)$ such that $|F - \mathcal{NN}|_K < \varepsilon$.*

The primary example of a constraint set fulfilling the assumptions of Theorem 3.1 is the following.

Proposition 3.2. *Let $K \subset X$ be compact. Then the set $\{F \in C(X, \mathbb{R}) : F(u) > 0 \text{ for all } u \in K\}$ is open in $C(X, \mathbb{R})$ fulfilling $\text{cl}(\mathfrak{C}) = \text{cl}(\text{int } \mathfrak{C})$.*

Remark 2. Notice that by [4, Proposition 4.1] we have for separable Banach spaces with normalized Schauder basis $\{e_k\}_{k \in \mathbb{N}}$ and Lipschitz continuous activation functions that we can approximate a continuous functional F on any compact set to arbitrary precision by a finite dimensional neural network. Observe that the linear operators $\Pi_d A_j \Pi_d$ therein might be non-local in accordance with the findings in operator learning theory.

Besides positivity, we also want to ensure convexity of the approximation. However, this constraint does not fulfill the assumptions of Theorem 3.1 as an arbitrarily small perturbation of a convex function is not convex in general. Nevertheless, for

the ReLU activation function $\sigma(x) = \max\{0, x\}$, one can exploit results from convex analysis to obtain convex neural networks. In the general separable Banach space setting, the sufficient conditions in Ref. [4] for an activation function to be discriminatory are not met for ReLU. It remains open if an appropriate generalization of ReLU to the Banach space setting is discriminatory. This is why, in the construction presented below, the ReLU is used only on finite dimensional subspaces.

Theorem 3.3. *Let $F : X \rightarrow \mathbb{R}$ be a convex, continuous functional on X . Then for every compact set $K \subset X$ and any $\varepsilon > 0$ there exists a convex, multi-layer neural network $\mathcal{NN}$ with ReLU and linear activation functions such that $|F - \mathcal{NN}|_K < \varepsilon$.*

The proof of this theorem is based on the fact that convex functions can be approximated as the maximum of supporting affine hyperplanes. As the maximum operation can be written as a neural network with ReLU activation function and as the maximum of affine functions is convex the claim follows. We observe that this hyperplane construction also allows us to estimate the number of neurons needed in terms of ε , $\text{diam}(K)$, the covering number of K and the upper and lower bounds M, m of F on K . Furthermore, if we have a class of convex functionals with uniform bounds on m and M for fixed K , all these functionals can be approximated by neural networks with the same width and depth, justifying the use of the term “universal approximation”.

Remark 3. In the finite dimensional case, we can replace the continuity assumption by demanding lower semicontinuity and properness. However, we then need to assume $K \subset \text{dom} F$ for the compact set. Note that in infinite dimensions, this is no longer true.

3.2. Non-reflexive Moreau–Yosida regularization. The neural network approximation theory cannot be applied directly to the DFT functionals, as they are everywhere discontinuous in X , see [17]. To remedy this, we replace the density functionals with their Moreau–Yosida regularization. This operation is “lossless”, i.e. the resulting regularized functionals produce the same ground-state energies up to simple shift (see (9) below). In the original truncated setting of Ref. [14], one takes $X = L^2(\Lambda)$, where $\Lambda \subset \mathbb{R}^3$ is a box and apply the standard Hilbert space Moreau–Yosida regularization [3] to the *canonical* Lieb functional F_L .

However, for non-reflexive Banach spaces, the standard theory of Moreau–Yosida regularization falls short. This especially holds true for $X = L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3)$. In the following, we suggest two remedies that will allow us to define Moreau–Yosida regularization with similar properties as in the standard setting. First, it turns out that what is needed is a strictly convex, Gâteaux differentiable norm. Second, we need stricter assumptions on the functional that should be regularized. In particular, one needs that subsets of the level sets of the functional are weakly compact.

Define the regularization of a convex functional $F : X \rightarrow \mathbb{R}$ via the infimal convolution

$$F^\varepsilon(\rho) = \inf_{\sigma \in X} \left[F(\sigma) + \frac{1}{2\varepsilon} \|\rho - \sigma\|_X^2 \right].$$

Here, the norm $\|\cdot\|_X$ is equivalent to $\|\cdot\|_X$, strictly convex and Gâteaux differentiable.

First, we consider the question of existence of such a norm $\|\cdot\|_X$. For the case $X = L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3)$, we begin with a negative result from Banach space geometry: we cannot equip X with a Fréchet differentiable equivalent norm.

Theorem 3.4. [8, Theorem 8.24] *Let X be a separable Banach space. Then X admits an equivalent Fréchet differentiable norm if and only if X^* is separable.*

Here, $X^* = L^\infty + L^{3/2}$ is not separable because $L^\infty \subset X^*$ is not. This shows that it is *impossible* to obtain a Fréchet differentiable regularization in the non-reflexive space X via infimal convolution of the functional with some equivalent norm. We *can*, however, equip X with an equivalent strictly convex, Gâteaux differentiable norm. In fact, more is true.

Theorem 3.5. [8, Theorem 8.20] *Let X be a separable Banach space. Then X admits an equivalent locally uniformly rotund, Gâteaux differentiable norm.*

Here, “locally uniformly rotund” implies “strictly convex” [8]. Therefore, we can equip $X = L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3)$ with such a norm. For applications to DFT, Gâteaux differentiability will prove to be sufficient.

Before considering the basic properties of the (non-reflexive) Moreau–Yosida regularized functionals, we need to give a few definitions. Define the *duality mapping* $\mathcal{J} : X \rightarrow 2^{X^*}$ via $\mathcal{J} = \partial(\frac{1}{2}\|\cdot\|_X^2)$, so

$$\partial(\frac{1}{2}\|\cdot\|_X^2)(\rho) = \{v \in X^* : \langle \rho, v \rangle = \|\rho\|_X^2 = \|v\|_{X^*}^2\}$$

for any $\rho \in X$.

Next, we define the proximal mapping Π_F^ε associated to F^ε . A minimizer (whenver it exists) of

$$\sigma \mapsto F(\sigma) + \frac{1}{2\varepsilon}\|\rho - \sigma\|_X^2$$

is called a proximal point of ρ and defines the (in general set-valued mapping) $\Pi_F^\varepsilon : X \rightrightarrows X$, $\rho \mapsto \Pi_F^\varepsilon(\rho)$. Because the norm $\|\cdot\|$ is strictly convex, the set of proximal points is a singleton, or empty.

This equivalent, locally uniformly rotund, Gâteaux differentiable norm ensures the usual properties of the Moreau–Yosida envelope to hold true in the non-reflexive setting, under some additional structural hypothesis on the functional.

Theorem 3.6. *Suppose that $F : X \rightarrow \mathbb{R} \cup \{+\infty\}$ is a proper convex, nonnegative and lower semicontinuous functional. In addition, we assume either that X is reflexive or that bounded subsets of the level sets of F are weakly sequentially compact. Then the following properties hold true.*

- (i) F^ε is convex, continuous and Gâteaux differentiable everywhere.
- (ii) (Domination) $\inf_X F \leq F^\varepsilon(\rho) \leq F^\delta(\rho) \leq F(\rho)$ for all $\rho \in X$ and $0 \leq \delta < \varepsilon$. In particular, $\inf_X F^\varepsilon = \inf_X F$.
- (iii) (Pointwise convergence) $F^\varepsilon(\rho) \nearrow F(\rho)$ for all $\rho \in X$ as $\varepsilon \rightarrow 0$.
- (iv) The convex subdifferential $\partial F^\varepsilon : X \rightarrow 2^{X^*}$ of F^ε is the singleton: $\partial F^\varepsilon(\rho) = \{(F^\varepsilon)'(\rho)\}$, where $(F^\varepsilon)'(\rho) \in X^*$ is the Gâteaux derivative of F^ε at $\rho \in X$.
- (v) The proximal mapping Π_F^ε is singleton-valued everywhere and for $\rho \in \text{dom } F$ we have $\|\Pi_F^\varepsilon(\rho) - \rho\| = \mathcal{O}(\sqrt{\varepsilon})$.
- (vi) (Derivative) $(F^\varepsilon)'(\rho) = \frac{1}{\varepsilon}\mathcal{J}(\rho - \Pi_F^\varepsilon(\rho))$.
- (v) (Subdifferential) For any $\rho, \rho_\varepsilon \in X$ the relation $\rho_\varepsilon = \Pi_F^\varepsilon(\rho)$ is equivalent to $\frac{1}{\varepsilon}\mathcal{J}(\rho - \rho_\varepsilon) \in \partial F(\rho_\varepsilon)$.

Remark 4. The assumption of F being nonnegative can be relaxed using the fact that F has a continuous affine minorant.

Via the Legendre transform at a fixed particle number $\lambda \in \mathbb{R}_+$, we may associate the corresponding regularized ground-state energy to F^ε , i.e.

$$E_\lambda^\varepsilon(v) = \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = \lambda}} [F^\varepsilon(\rho) + \langle v, \rho \rangle].$$

It is well-known that the Legendre transform of an infimal convolution is simply the sum of the functions, so we obtain that the regularized energy is simply shifted according to

$$E_\lambda^\varepsilon(v) = E_\lambda(v) - \frac{\varepsilon}{2} \|v\|^2 \quad (9)$$

for all $v \in X^*$, where $E_\lambda(v)$ was defined in (8). We observe that $v \mapsto E_\lambda^\varepsilon(v)$ is *strictly* concave and locally Lipschitz.

To summarize, we may choose $X = L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3)$ equipped with $\|\cdot\|$ coming from Theorem 3.5 and $F = F_{\text{GC}}$, the grand-canonical Levy–Lieb functional defined in Section 2.3. It is also possible to take $X = L^2(\Lambda)$, with $\Lambda \subset \mathbb{R}^3$ a box, and $F = F_{\text{L}}$ the canonical Lieb functional. In addition, $X = L^3(\Lambda)$ was considered in Ref. [15], beneficial for the application of Moreau–Yosida regularization to current-density-functional theory [16].

3.3. An error estimate for the approximate ground-state energy. In this section, we combine our results regarding neural networks and regularization. As we saw, the neural network approximation theorems only work for a continuous functional, so we need to apply Moreau–Yosida regularization to the grand-canonical Levy–Lieb functional to make it continuous.

Recall that in order to invoke the universal approximation theorems for neural networks we need to restrict the possible densities ρ on which we want to approximate our functionals to a compact set K .

Theorem 3.7. *Suppose that X and $F : X \rightarrow \mathbb{R} \cup \{+\infty\}$ fulfills the assumptions of Theorem 3.6. Let $\varepsilon > 0$, $N \in \mathbb{N}$, and $K \subset \{\int_{\mathbb{R}^3} \cdot = N\} \subset X$ be compact, such that the set*

$$V_K = \left\{ v \in X^* : \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle] = \inf_{\rho \in K} [F^\varepsilon(\rho) + \langle v, \rho \rangle] \right\}$$

is non-empty.

Then for any $\delta > 0$ there exists a neural network $\mathcal{NN} \in \mathfrak{N}(\sigma)$ such that for all $v \in V_K$

$$|E_N(v) - E_{\mathcal{NN},K}(v)| \leq \frac{\varepsilon}{2} \|v\|^2 + \delta,$$

where

$$E_N(v) = \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F(\rho) + \langle v, \rho \rangle]$$

and

$$E_{\mathcal{NN},K}(v) = \inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle].$$

One can observe that there exists at least one such a compact set K such that V_K is non-empty by considering a potential that allows for a ground state.

Remark 5. An interesting class of compact sets can be deduced from the exponential bounds of ground states. For example, for potentials stemming from electronic structure calculations with fixed number of nuclei with uniformly bounded pairwise distance, one would expect a uniform exponential decay behavior of the ground-state densities implying that the ground states with uniformly bounded energy lie in one compact set. To be more precise, for some $\lambda > 0$ let

$$U = \{\rho \in \mathcal{I} : \exp(\lambda|x|)\rho(x) \in L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3), \exp(\lambda|x|)\nabla\sqrt{\rho(x)} \in L^2(\mathbb{R}^3)\}.$$

Then U is compact in X , see Theorem 4.2. The associated set

$$V = \{v \in X^* \mid \inf_{\rho \in X} [F(\rho) + \langle v, \rho \rangle] = \inf_{\rho \in U} [F(\rho) + \langle v, \rho \rangle]\}$$

would be the set of Coulomb potentials with uniform bounded gap between the ground-state energy and the ionization threshold and, thus, is non-empty by construction.

Remark 6. Observe that the actual complexity of the neural network depends on K and δ .

Proposition 3.8. *Let $\mathcal{NN}$ and compact $K \subset X$ be as in Theorem 3.7. Then the associated energy $E_{\mathcal{NN},K} : X^* \rightarrow \mathbb{R}$ is concave, locally Lipschitz and fulfills $E_{\mathcal{NN},K}(v_1) \leq E_{\mathcal{NN},K}(v_2)$ if $v_1(x) \leq v_2(x)$ for almost all $x \in \mathbb{R}^3$.*

4. PROOFS

The rest of the paper is devoted to proofs.

4.1. Proof of Section 2.

Proof of Theorem 2.1. It is well-known that for a convex functional, weak lower semicontinuity is equivalent to strong lower semicontinuity. Hence, let $\rho_j \rightarrow \rho$ strongly. Without loss of generality, we may assume that $\lim_{j \rightarrow \infty} F_{\text{GC}}(\rho_j)$ exists and is finite. We may also assume that $\rho_j \geq 0$ for all j sufficiently large. The Hoffmann–Ostenhof inequality (4) implies that $\{\nabla\sqrt{\rho_j}\} \subset L^2$ is bounded since $F_{\text{GC}}(\rho_j)$ is assumed finite (and that bounds the kinetic energy of the density sequence). Furthermore, $\{\rho_j\} \subset X$ is also bounded (by the norm convergence), i.e. $\|\rho_j\|_{L^1} \leq C$. So $\{\sqrt{\rho_j}\} \subset H^1$ is bounded. By the Banach–Alaoglu theorem, $\sqrt{\rho_j} \rightharpoonup \tau$ weakly in H^1 up to a subsequence for some $\tau \in H^1(\mathbb{R}^3)$. Using $\tau \geq 0$, an elementary argument shows that we must have $\tau = \sqrt{\rho}$. Using Theorem 2.2 shows lower semicontinuity of F_{GC} .

For the second part of the theorem, let ρ_j be bounded and in the C -level set of F_{GC} . Similarly, we can again use the Hoffmann–Ostenhof inequality and the argument above to extract a subsequence weakly converging in H^1 . Together with the weak lower semicontinuity of F_{GC} and of the norm we conclude that the limit is also in the level set and thus proving weak sequential compactness. $\square$

4.2. Proofs of Section 3.1.

Proof of Theorem 3.1. First, we observe that $\text{cl}(\mathfrak{C}) = \text{cl}(\text{int}\mathfrak{C})$ implies that the interior of $\mathfrak{C}$ is non-empty and that every point in $\mathfrak{C}$ can be arbitrarily well approximated by interior points.

Now let K be compact and $\varepsilon > 0$. Furthermore, let $f \in \text{int}(\mathfrak{C})$. Then there is $\varepsilon_0 > 0$ s.t. for all $\varepsilon < \varepsilon_0$ we have that for any $g \in C(X, \mathbb{R})$ with $|g - f|_K < \varepsilon$ it

holds $g \in \text{int}(\mathfrak{C})$ as the interior is open. By the UAP ([4]) or Theorem 3.3 there is a $\mathcal{NN} \in C(X, \mathbb{R})$ such that $|\mathcal{NN} - f|_K < \varepsilon$. Thus, $\mathcal{NN} \in \text{int}(\mathfrak{C})$.

Next, take a function $f \in \partial\mathfrak{C} \cap \mathfrak{C}$. As $\text{int}(\mathfrak{C})$ is dense in $\mathfrak{C}$, there is a sequence $f_n \in \text{int}(\mathfrak{C})$ such that $f_n \rightarrow f$, i.e. there is some $N \in \mathbb{N}$ such that $|f - f_n|_K < \frac{\varepsilon}{2}$. Again, by UAP with the arguments of the previous paragraph, there is a $\mathcal{NN} \in \mathfrak{N}(\sigma) \cap \text{int}\mathfrak{C}$ such that $|f_n - \mathcal{NN}|_K < \frac{\varepsilon}{2}$. Thus, $|f - \mathcal{NN}|_K < \varepsilon$. $\square$

Proof of Proposition 3.2. Let $f_1 \in \{f \in C(X, \mathbb{R}) : f(x) > 0, x \in K\}$. As $f(x) > 0$ and K compact there is $\delta > 0$ such that $f(x) > \delta$ for all $x \in K$. Let $f_2 \in C(X, \mathbb{R})$ with $|f_1 - f_2|_K < \frac{\delta}{2}$. Then we have

$$f_2(x) = f_2(x) - f_1(x) + f_1(x) \geq f_1(x) - \frac{\delta}{2} \geq \frac{\delta}{2} > 0.$$

Hence, $f_2 \in \{f \in C(X, \mathbb{R}) : f(x) > 0, x \in K\}$. Therefore, there exists a open neighborhood of f_1 contained in $\{f \in C(X, \mathbb{R}) : f(x) > 0, x \in K\}$. $\square$

In order to obtain convex neural network approximations, we will first show that any continuous, convex functional can be approximated on some compact set K by finitely many supporting hyperplanes.

Lemma 4.1. *Let $F : X \rightarrow \mathbb{R}$ be a convex, continuous functional on X . Then for any compact set K and any $\varepsilon > 0$ there exists an $N \in \mathbb{N}$ and $a_n \in X^*$ and $b_n \in \mathbb{R}$ for $n \in \{1, \dots, N\}$ such that for all $u \in K$ it holds*

$$F(u) \geq \max_{n \in \{1, \dots, N\}} [\langle a_n, \rho \rangle + b_n] \quad \text{and} \quad \left| F(u) - \max_{n \in \{1, \dots, N\}} [\langle a_n, \rho \rangle + b_n] \right| < \varepsilon.$$

Proof of Lemma 4.1. It is standard that

$$F(u) = \sup[\langle a, u \rangle + b : a \in X^*, b \in \mathbb{R}, \langle a, u \rangle + b \leq F(u)].$$

By definition of the subdifferential (and its non-emptiness for convex functions) given u' there are $a_{u'} \in \partial F \subset X^*$ and $\langle a_{u'}, u' \rangle + b_{u'} = F(u') \in \mathbb{R}$ such that $F(u') = \langle a_{u'}, u' \rangle + b_{u'}$ with $F(u) \geq \langle a_{u'}, u \rangle + b_{u'}$. Here, we used the continuity assumption on F . Therefore, we can assume that the supremum is attained for continuous convex functionals and we get

$$F(u) = \max_{a_u \in X^*, b_u \in \mathbb{R}} [\langle a_u, u \rangle + b_u].$$

Then by continuity, we have $m \leq F(u) \leq M$ for some real numbers $m, M \in \mathbb{R}$ for any $u \in K$. Therefore, F is Lipschitz on K with Lipschitz constant $L = \frac{M-m}{\text{diam}(K)}$. Furthermore, as the subdifferentials of continuous convex functions are bounded, we can define $\sup_{u \in K} \|a_u\| < \infty$.

Now let $U_u \subset K$ be an open cover K such that $\text{diam}(U_u) < \frac{\varepsilon}{L + \sup_{u \in K} \|a_u\|}$. By compactness, we can choose $N \in \mathbb{N}$ and a finite subcover $U_n, n \in \{1, \dots, N\}$. Then we have

$$\begin{aligned} 0 &\leq F(u) - \max_{n \in \{1, \dots, N\}} [\langle a_n, u \rangle + b_n] \\ &= F(u) - \max_{n: u \in U_n} [\langle a_n, u \rangle + b_n] = F(u) - \langle a_{n^*}, u \rangle - b_{n^*} \end{aligned}$$

with n^* realizing the last maximum. Then

$$\begin{aligned}
& \left| F(u) - \max_{n \in \{1, \dots, N\}} [\langle a_n, u \rangle + b_n] \right| \\
& \leq |F(u) - F(u_{n^*})| + |\langle a_{n^*}, u \rangle + b_{n^*} - F(u_{n^*})| \\
& \leq L \|u - u_{n^*}\| + |\langle a_{n^*}, u - u_{n^*} \rangle| \\
& \leq L \|u - u_{n^*}\| + \|a_{n^*}\| \|u - u_{n^*}\| \\
& \leq (L + \max_{n \in \{1, \dots, N\}} \|a_n\|) \frac{\varepsilon}{L + \sup_{u \in K} \|a_u\|} < \varepsilon
\end{aligned}$$

as stated. $\square$

Proof of Theorem 3.3. First we observe that for some $N \in \mathbb{N}$, the function $g : \mathbb{R}^N \rightarrow \mathbb{R}$ defined as

$$g(u_1, \dots, u_N) = \max\{\ell_1, \dots, \ell_N\}$$

can be written as ReLU network of depth $\log(N) + 1$ and width $9N - 4$, [1].

By Lemma 4.1, there exists $N \in \mathbb{N}$, $a_n \in X^*$ and $b_n \in \mathbb{R}$ such that

$$\left| F(u) - \max_{n \in \{1, \dots, N\}} [\langle a_n, u \rangle + b_n] \right| < \varepsilon.$$

We observe, that for arbitrary but fixed $u \in K$ we have $\ell_n = \langle a_n, u \rangle + b_n \in \mathbb{R}$. Therefore, there exists a (standard) ReLU network $G_{\mathcal{NN}}(\ell_1, \dots, \ell_N) = \Lambda_L \circ \dots \circ \Lambda_2(\ell_1, \dots, \ell_N)$ where $\Lambda_j(y) = \text{ReLU}(A_j y + b_j)$ such that

$$G_{\mathcal{NN}}(\ell_1, \dots, \ell_N) = \max_{n \in \{1, \dots, N\}} \{\ell_1, \dots, \ell_n\}.$$

Lastly, we observe that $u \mapsto (\langle a_n, u \rangle)_{n=1}^N$ can be understood as bounded linear operator on X by defining $Au = \sum_{n=1}^N \langle a_n, u \rangle e_n$ given a Schauder basis $\{e_n\}_{n \in \mathbb{N}}$, which exists by assumption. Then defining $\Lambda_1(x) = Ax$ allows us to write

$$\max_{n \in \{1, \dots, N\}} [\langle a_n, x \rangle + b_n] = \Lambda_L \circ \dots \circ \Lambda_2 \circ \Lambda_1(x).$$

Therefore, $\max_{n \in \{1, \dots, N\}} [\langle a_n, x \rangle + b_n]$ can be exactly written by a neural network with depth $\log(N) + 2$ and with width $9N - 4$. Note that by construction the neural network is convex. $\square$

4.3. Proof of Theorem 3.6. Part (i) follows from Theorem 2.1.3, Theorem 2.2.14 and Corollary 2.4.8 in Ref. [27] using the Gâteaux differentiability of $\|\cdot\|_X^2$. We obtain

$$\text{int}(\text{dom } F^\varepsilon) = \text{dom } F + \text{int dom } \|\cdot\|_X^2 = \text{dom } F + X = X.$$

For part (ii), note that

$$\inf_{\sigma \in X} F(\sigma) \leq F^\varepsilon(\rho) = \inf_{\sigma \in X} \left[F(\sigma) + \frac{1}{2\varepsilon} \|\rho - \sigma\|_X^2 \right] \leq F(\rho) \quad \text{for all } \rho \in X.$$

Part (iii) trivially follows from (ii). Part (iv) follows from (i).

For part (v), we need to show that the infimum in the definition of F^ε is attained at a unique point. Fix $\rho \in X$ and define the functional $Q : X \rightarrow \mathbb{R} \cup \{+\infty\}$ via

$$Q(\sigma) := F(\sigma) + \frac{1}{2\varepsilon} \|\rho - \sigma\|_X^2 \quad \text{for all } \sigma \in X.$$

Since F is convex and $\|\cdot\|_X^2$ is strictly convex, the uniqueness of the optimizer follows as Q is strictly convex. So it remains to prove existence. Let $\{\sigma_j\}$ be a minimizing

sequence. If X is reflexive we use Banach-Alaoglu to extract a weakly convergent subsequence. Otherwise, as

$$F(\sigma_j) \leq Q(\sigma_j) \leq C$$

we have that σ_j are in the C -level set of F and moreover are bounded. Thus, by assumption (for the level sets of F) we can extract a weakly convergent subsequence with limit σ . In either case, the weak lower semicontinuity of F and the norm $\|\cdot\|_X$ allows us to conclude that

$$Q(\sigma) \leq \liminf_{j \rightarrow \infty} Q(\sigma_j),$$

therefore σ is a minimizer, i.e. the proximal point. It remains to show the convergence rate bound. To see this, note that

$$F(\Pi_F^\varepsilon(\rho)) + \frac{1}{2\varepsilon} \|\rho - \Pi_F^\varepsilon(\rho)\|^2 = F^\varepsilon(\rho) \leq F(\rho),$$

where $\rho \in \text{dom } F$, so

$$\|\rho - \Pi_F^\varepsilon(\rho)\|^2 \leq 2F(\rho)\varepsilon$$

as stated.

4.4. Proof of Theorem 3.7. As F fulfills the assumptions of Theorem 3.6 we have that F^ε is continuous.

The continuity of F^ε implies by Theorem 3.1 and Theorem 3.3 that for every $\delta > 0$ there exists a $\mathcal{NN}$ approximating F^ε on K in the sense $|\mathcal{NN}(\rho) - F^\varepsilon(\rho)| < \delta$ for every $\rho \in K$. From the relation (9), we get

$$|E_N(v) - E_{\mathcal{NN},K}(v)| \leq \frac{\varepsilon}{2} \|v\|^2 + |E_{\mathcal{NN},K}(v) - E_N^\varepsilon(v)|. \quad (10)$$

We now claim that for the second term we have

$$|E_{\mathcal{NN},K}(v) - E_N^\varepsilon(v)| \leq \delta. \quad (11)$$

Recall that $v \in V_K$, i.e. the infimum is attained in the compact set K . Let us consider now the two cases:

Case 1.

$$\inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] > \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle].$$

Then there exists a γ_0 such that for all $0 < \gamma \leq \gamma_0$ we have

$$0 < \inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] - \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle] - \gamma.$$

Now, let $\{\rho_n\} \subset K$ be a minimizing sequence of $\rho \mapsto F^\varepsilon(\rho) + \langle v, \rho \rangle$ such that

$$\inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle] + \gamma > F^\varepsilon(\rho_n) + \langle v, \rho_n \rangle.$$

Then

$$\begin{aligned} 0 &< \inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] - \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle] - \gamma \\ &\leq \inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] - F^\varepsilon(\rho_n) - \langle v, \rho_n \rangle \\ &\leq \mathcal{NN}(\rho_n) + \langle v, \rho_n \rangle - F^\varepsilon(\rho_n) - \langle v, \rho_n \rangle \leq \delta, \end{aligned}$$

where we used that $\{\rho_n\} \subset K$ and thus

$$0 \leq \mathcal{NN}(\rho_n) - F^\varepsilon(\rho_n) \leq |\mathcal{NN}(\rho_n) - F^\varepsilon(\rho_n)| < \delta.$$

Case 2.

$$\inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] < \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle].$$

Then there exists a γ_0 such that for all $0 < \gamma \leq \gamma_0$ we have

$$0 < \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle] - \inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] - \gamma.$$

Now, let $\{\rho_n\} \subset K$ be a minimizing sequence of $\rho \mapsto \mathcal{NN}(\rho) + \langle v, \rho \rangle$ such that

$$\inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] + \gamma > \mathcal{NN}(\rho_n) + \langle v, \rho_n \rangle.$$

Then

$$\begin{aligned} 0 &< \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle] - \inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] - \gamma \\ &\leq \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle] - \mathcal{NN}(\rho_n) - \langle v, \rho_n \rangle \\ &\leq F^\varepsilon(\rho_n) + \langle v, \rho_n \rangle - \mathcal{NN}(\rho_n) - \langle v, \rho_n \rangle \leq \delta, \end{aligned}$$

where we used that $\{\rho_n\} \subset K$ and thus

$$0 \leq F^\varepsilon(\rho_n) - \mathcal{NN}(\rho_n) \leq |\mathcal{NN}(\rho_n) - F^\varepsilon(\rho_n)| < \delta.$$

To conclude, we have

$$0 < \gamma < \left| \inf_{\substack{\rho \in X \\ \int_{\mathbb{R}^3} \rho = N}} [F^\varepsilon(\rho) + \langle v, \rho \rangle] - \inf_{\rho \in K} [\mathcal{NN}(\rho) + \langle v, \rho \rangle] \right| < \delta + \gamma$$

for all $0 < \gamma \leq \gamma_0$ and, thus, the claim follows taking $\gamma \rightarrow 0+$.

4.5. Further proofs.

Lemma 4.2. *Let $M > 0$ and U be uniformly integrable for $p = 1, 3$ as well $\sqrt{U} = \{\nabla \sqrt{\rho} : \rho \in U\}$ be uniformly integrable for $p = 2$. Then U is compact in X .*

Proof. Let $\varepsilon > 0$. By uniform integrability there is a closed ball $B_{\mathbb{R}^3}$ such that $\sup_{\rho \in U} \int_{\mathbb{R}^3 \setminus B_{\mathbb{R}^3}} |\rho|^p(x) dx < \frac{\varepsilon}{2}$ as well as $\sup_{\rho \in U} \int_{\mathbb{R}^3 \setminus B_{\mathbb{R}^3}} |\nabla \sqrt{\rho}|^2(x) dx < \frac{\varepsilon}{2}$. Let ρ_i be a weakly convergent sequence in X with $\nabla \sqrt{\rho_i} \rightharpoonup \nabla \sqrt{\rho}$ in L^2 . We have that $\nabla \sqrt{\rho_i}|_{B_{\mathbb{R}^3}} \rightharpoonup \nabla \sqrt{\rho}|_{B_{\mathbb{R}^3}}$ implies $\sqrt{\rho_i} \rightarrow \sqrt{\rho}$ strongly in $L^2(B_{\mathbb{R}^3})$. This implies $\sqrt{\rho_i}$ converging strongly in $L^2(\mathbb{R}^3)$ and thus strong convergence of ρ_i in $L^1(\mathbb{R}^3)$.

Therefore, we have that U is totally bounded in $L^1(\mathbb{R}^3)$. We observe that complete integrable sets of functions are closed under strong convergence by dominated convergence. That means U is compact in L^1 . We also have that U is closed in $L^1 \cap L^3$ under strong convergence. Thus, U is compact in X . $\square$

Proof of Proposition 3.8. To show concavity, we observe that as infimum of affine functions $v \mapsto \langle v, \rho \rangle + \mathcal{NN}(\rho) + \chi_K(\rho)$ the energy is concave, independent of the properties of the neural network.

Then we can show local Lipschitz continuity as follows. It is known that if a proper convex function is upper-bounded in a neighborhood of a point in its

domain, it is locally Lipschitz. Hence, if a concave function is bounded from below, it is locally Lipschitz. First, we observe that as $\mathcal{NN}$ is continuous and K compact we have $|\langle v, \rho \rangle + \mathcal{NN}^\varepsilon(\rho) + \chi_K(\rho)| < \infty$ for all $v \in X^*$ and $\rho \in X$. Therefore, $E_{\mathcal{NN},K}$ is proper. Now it remains to show that for any $v_0 \in X^*$ the energy $E_{\mathcal{NN},K}$ is lower bounded in some neighborhood of v_0 . Observe that due to continuity of $\mathcal{NN}$ and compactness of K there is $m \in \mathbb{R}$ such that $\inf_{\rho \in K} [\mathcal{NN}(\rho) + \chi_K(\rho)] \geq m$ independent of v . Hence,

$$\langle v, \rho \rangle + \mathcal{NN}(\rho) + \chi_K(\rho) \geq m - \|v\|_{X^*} \min_{\rho \in K} \|\rho\| \geq m - (\|v_0\| + \delta) \min_{\rho \in K} \|\rho\|$$

if $\|v - v_0\| \leq \delta$ establishing a lower bound.

Lastly, we can show that the energy is order preserving. Let $v_1 \leq v_2$ almost everywhere. As $E_{\mathcal{NN},K}(v_2)$ is finite if and only if the infimum is realized for some positive density ρ_{v_2} we get $\langle v_1, \rho_{v_2} \rangle \leq \langle v_2, \rho_{v_2} \rangle$. Therefore, it holds

$$\begin{aligned} E_{\mathcal{NN},K}(v_1) &\leq \mathcal{NN}(\rho_{v_2}) + \chi_K(\rho_{v_2}) + \langle v_1, \rho_{v_2} \rangle \\ &\leq \mathcal{NN}(\rho_{v_2}) + \chi_K(\rho_{v_2}) + \langle v_2, \rho_{v_2} \rangle = E_{\mathcal{NN},K}(v_2) \end{aligned}$$

as stated. $\square$

REFERENCES

- [1] R. Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. *Electron. Colloquium Comput. Complex.*, TR17, 2016.
- [2] Paul W. Ayers. Energy is not a convex function of particle number for r^{-k} interparticle potentials with $k > \log_3 4$. *The Journal of Chemical Physics*, 160(4), January 2024.
- [3] Heinz H Bauschke and Patrick L Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.
- [4] Fred Espen Benth, Nils Detering, and Luca Galimberti. Neural networks in Fréchet spaces. *Annals of Mathematics and Artificial Intelligence*, 91(1):75–103, 2023.
- [5] Oliver M. Bohle, Maryam Lotfigolian, Andre Laestadius, and Erik I. Tellgren. Regularised density-potential inversion for periodic systems: application to exact exchange in one dimension. arXiv:2510.24330 [physics.chem-ph], 2025.
- [6] Simone Di Marino, Mathieu Lewin, and Luca Nenna. Ground state energy is not always convex in the number of electrons. *The Journal of Physical Chemistry A*, 128(49):10697–10706, November 2024.
- [7] Eberhard Engel and Reiner M Dreizler. *Density Functional Theory: An Advanced Course*. Springer Science & Business Media, 2011.
- [8] Marián Fabian, Petr Habala, Petr Hájek, Vicente Montesinos Santalucía, Jan Pelant, and Václav Zizler. *Functional analysis and infinite-dimensional geometry*, volume 8. Springer, 2001.
- [9] John E Harriman. Orthonormal orbitals for the representation of an arbitrary density. *Physical Review A*, 24(2):680, 1981.
- [10] Trygve Helgaker and Andrew M Teale. Lieb variation principle in density-functional theory. *The Physics and Mathematics of Elliott Lieb*, page 527.
- [11] Michael F. Herbst, Vebjørn H. Bakkestuen, and Andre Laestadius. Kohn–Sham inversion with mathematical guarantees. *Phys. Rev. B*, 111(20), May 2025.
- [12] Maria Hoffmann-Ostenhof and Thomas Hoffmann-Ostenhof. “Schrödinger inequalities” and asymptotic behavior of the electron density of atoms and molecules. *Physical Review A*, 16(5):1782, 1977.
- [13] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136(3B):B864–B871, 1964.
- [14] Simen Kvaal, Ulf Ekström, Andrew M Teale, and Trygve Helgaker. Differentiable but exact formulation of density-functional theory. *The Journal of chemical physics*, 140(18):18A518, 2014.

- [15] Andre Laestadius, Markus Penz, Erik I. Tellgren, Michael Ruggenthaler, Simen Kvaal, and Trygve Helgaker. Generalized Kohn–Sham iteration on Banach spaces. *J. Chem. Phys.*, 149(16):164103, 2018.
- [16] Andre Laestadius, Erik I Tellgren, Markus Penz, Michael Ruggenthaler, Simen Kvaal, and Trygve Helgaker. Kohn–Sham theory with paramagnetic currents: compatibility and functional differentiability. *J. Chem. Theory Comput.*, 15(7):4003–4020, 2019.
- [17] Paul E Lammert. Differentiability of Lieb functional in electronic density functional theory. *International Journal of Quantum Chemistry*, 107(10):1943–1953, 2007.
- [18] Mathieu Lewin, Elliott H Lieb, and Robert Seiringer. Universal functionals in density functional theory. In *Density Functional Theory: Modeling, Mathematical Analysis, Computational Methods, and Applications*, pages 115–182. Springer, 2022.
- [19] Elliott H Lieb. Density functionals for Coulomb systems. *International Journal of Quantum Chemistry*, 24(3):243–277, 1983.
- [20] Elliott H Lieb and Michael Loss. Analysis. In *Amer. Math. Soc*, 2001.
- [21] Elliott H Lieb and Robert Seiringer. *The stability of matter in quantum mechanics*. Cambridge University Press, 2010.
- [22] Markus Penz, Mihály A Csirik, and Andre Laestadius. Density-potential inversion from Moreau–Yosida regularization. *Electron. Struct.*, 5(1):014009, March 2023.
- [23] Markus Penz, Michael F Herbst, Trygve Helgaker, and Andre Laestadius. Perspective on Moreau–Yosida regularization in density-functional theory. *arXiv preprint arXiv:2511.06957*, 2025.
- [24] Markus Penz, Andre Laestadius, Erik I Tellgren, and Michael Ruggenthaler. Guaranteed convergence of a regularized Kohn–Sham iteration in finite dimensions. *Phys. Rev. Lett.*, 123(3):037401, 2019.
- [25] Markus Penz, Andre Laestadius, Erik I. Tellgren, Michael Ruggenthaler, and Paul E. Lammert. Erratum: Guaranteed convergence of a regularized Kohn–Sham iteration in finite dimensions. *Phys. Rev. Lett.*, 125:249902, Dec 2020.
- [26] Julien Toulouse. Review of approximations for the Exchange-Correlation Energy in DFT. In *Density Functional Theory: Modeling, Mathematical Analysis, Computational Methods, and Applications*, pages 1–90. Springer, 2022.
- [27] Constantin Zălinescu. *Convex analysis in general vector spaces*. World scientific, 2002.